

# Inferring Structural Ensembles from Noisy Experiments and Molecular Dynamics: Correcting Force Field Bias with Bayesian Energy Landscape Tilting

Kyle A. Beauchamp,  
Biophysics Program,  
Stanford University, Stanford, CA

Vijay S. Pande,  
Chemistry Department and Structural Biology Program,  
Stanford University, Stanford, CA

Rhiju Das  
Biochemistry and Physics Departments,  
Stanford University, Stanford, CA

September 5, 2013

## Abstract

Inferring biomolecular conformation from experiment is a fundamental goal of structural biology. Structure determination often requires the combination of modeling and experiment, but the vast majority of approaches model only a single conformation, provide limited uncertainty information, and inherit biases from assumed force fields when data are limited. Building on recent conceptual advances, we hypothesized that these biases and missing uncertainty estimates could be addressed through Bayesian Energy Landscape Tilting (BELT), a scheme that enables the systematic computation of fully Bayesian 'hyperensembles' over conformational ensembles. As a test of BELT's ability to correct force field bias, we show that conformational ensembles of trialanine derived from five different force fields (ff96, ff99, ff99sbnmr-ildn, CHARMM27, and OPLS-AA) and chemical shift and scalar coupling measurements give convergent values of the peptide's  $\alpha$ ,  $\beta$ , and  $PP_{II}$  conformational populations. Furthermore, the ensembles recover set-aside measurements not used in the fitting. BELT's principled combination of simulation and limited experimental data promises rigorous assessment of force field bias and sets a foundation for modeling ensembles and uncertainties in complex biomolecular systems.

*Key words:* Molecular Dynamics, NMR, Conformational Ensembles, Bayesian Statistics

## 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 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, models of conformational ensembles are required to understand and predict structural and equilibrium properties.

A growing body of research has sought to characterize structural ensembles. Much of it has focused on incorporating dynamical information during NMR structure determination (5, 6) or the extraction of multiple conformers from X-ray diffraction data (7, 8). While these techniques are powerful, they share difficulties in data collection, the unified treatment of heterogeneous experimental data, and the data sparseness relative to the number of degrees of freedom. In particular, conformational ensemble modeling requires the estimation of not just a single structure, but a collection of structures and their associated equilibrium populations. This highly under-determined problem involves the simultaneous estimation of approximately  $3 \times m \times N$  parameters, where  $m$  is the number of states in the ensemble and  $N$  is the number of atoms in the molecule. Inference in this regime necessarily requires additional information, which can be provided by combining measurements with an atomistic force field.

Unfortunately, simulation benchmark studies have demonstrated continuing inaccuracies in molecular dynamics force fields (9–11). Force field modifications based on direct fitting to NMR measurements have also been demonstrated (12–14), but such work has optimized only a small fraction of the required force field parameters. Thus, simulations are often unable to fully recapitulate the wide variety of measurements available on molecular systems. This inaccuracy poses a challenge when one desires atomic-scale models that are both consistent with presently available measurements and predictive of those yet to be measured.

Here we introduce a statistical approach to modeling solution ensembles of biological macromolecules. The algorithm, Bayesian Energy Landscape Tilting (BELT), uses solution experiments to reweight a collection of atomistic models. BELT extends a recent maximum entropy method for restraining simulations (15) to reweight existing simulations. Furthermore, BELT leverages Markov Chain Monte Carlo (16) to transform experimental ambiguity into error bars on arbitrary structural features. BELT modeling produces a hyperensemble or an “ensemble of ensembles”; the output of BELT is a collection of statistical samples, each of which is a conformational ensemble.

BELT allows the full characterization of posterior distributions using MCMC. Most previous approaches, however, have focused on obtaining estimates of a single best-fit conformational ensemble (17, 18). In many cases, however, ambiguous experimental data preclude such a point-estimate of the conformational ensemble. For example, we plot one measured (18) value of  $^3J(H_N H^\alpha)$  in the context of the Karplus (19) equation relating  $\phi$  to  $^3J(H_N H^\alpha)$ . The measured coupling corresponds to four different values of  $\phi$  (Fig. 1a), showing that a point estimate can be inappropriate for modeling conformations. As previously stated, a point estimate of conformational ensembles is even more ill-posed. In such cases, it is most appropriate to report a collection of ensembles that are consistent with the available measurements (Fig. 1b). The BELT approach offers a practical recipe for describing such a hyperensemble, computing the hyperensemble’s predictions for new experimental observables, and giving error estimates on these predictions.

One stringent validation of BELT is to assess the convergence of ensembles constructed from

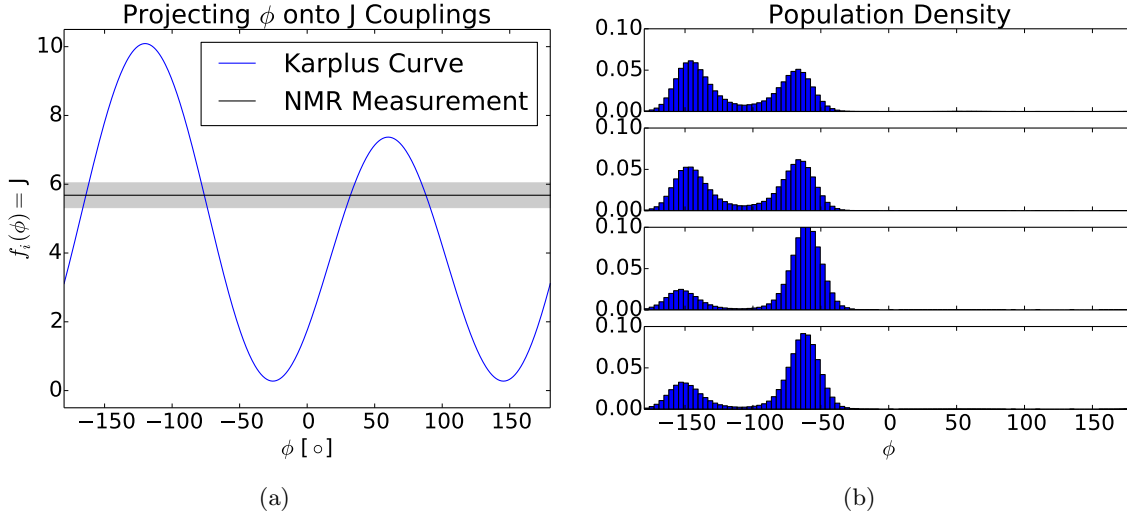


Figure 1: (a). The Karplus equation connecting the backbone torsion  $\phi$  to  ${}^3J(H_N H^\alpha)$  is ambiguous; the observed value of  ${}^3J(H_N H^\alpha)$  is consistent multiple conformations. (b). Four conformational ensembles that fit the data; each ensemble was randomly sampled using BELT as detailed below. The uncertainty ( $\sigma$ ) on  ${}^3J(H_N H^\alpha)$  was increased 2.5-fold to better illustrate differences.

force fields with radically different properties. We therefore investigated the conformational propensities of trialanine using NMR measurements (18) and MD simulations performed in five different force fields. The small size of this model system enabled assessment of BELT without complications from incomplete sampling. At the same time, trialanine populates multiple conformational states and allows a powerful illustration of ensemble modeling. Although the raw simulations show wide variations in their conformational preferences, BELT corrects force field errors to provide self-consistent estimates of the  $\alpha$ ,  $\beta$ , and  $PP_{II}$  populations. The ability to correct the biases of diverse forcefields provides a stringent and foundational test of the BELT approach for connecting simulation and equilibrium measurements.

## Theory: Bayesian Energy Landscape Tilting

### Model Inputs

To model an ensemble using BELT requires three components (Fig. 2). First, we need conformations  $x_j$  ( $j = 1, \dots, m$ ) sampled from the equilibrium distribution of some physically realistic model. This model will serve as a prior on structural properties; in the absence of experimental data, the BELT model inherits the properties of the conformations  $x_j$ . In the present work, such conformations will be generated from molecular dynamics simulations. Second, we require equilibrium experimental measurements  $F_i$  ( $i = 1, \dots, n$ ) and their associated uncertainties  $\sigma_i$  ( $i = 1, \dots, n$ ). 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  $x_j$ .

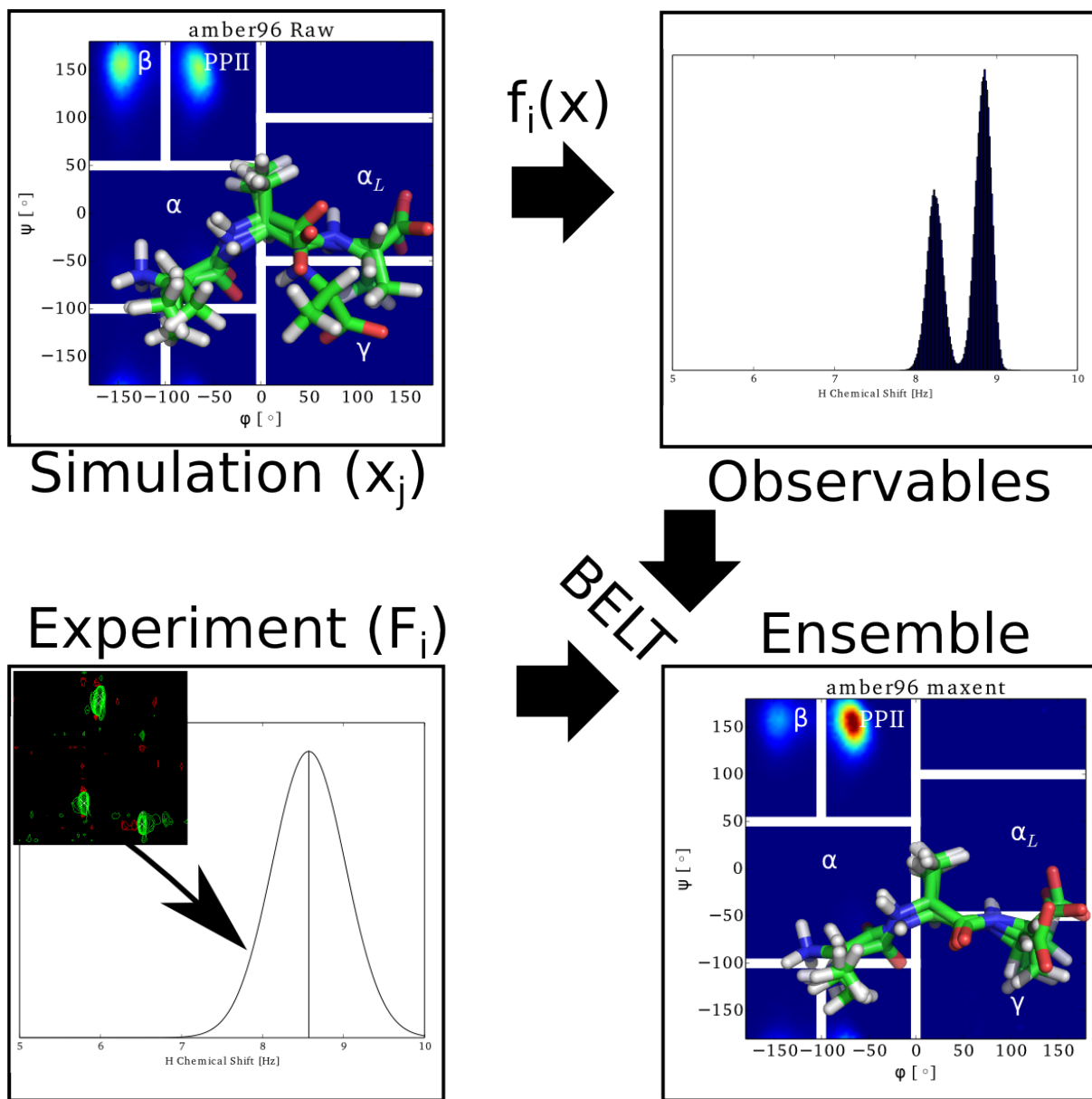


Figure 2: General scheme for BELT modeling.

## Reweightings

The next step in constructing an ensemble is to calculate the population of each conformation. Inspired by a previous method for restraining simulations (15) (see Appx. S1), we reweight individual conformations by a biasing potential that is a linear combination of the predicted observables:

$$\Delta U(x; \alpha) = \sum_{i=1}^n \alpha_i f_i(x)$$

In  $\Delta U(x; \alpha)$ , the parameters  $\alpha_i$  determine how strongly each experiment contributes to the biasing potential. As shown previously (15), such a linear biasing potential gives a maximum entropy ensemble for some set of experimental observations. The BELT strategy is to look beyond the single best such ensemble so as to estimate the uncertainty in the ensemble modeling. BELT instead samples over a distribution of such maximum entropy ensembles each parametrized by  $\alpha_i$ . This approach is connected to prior work by Crooks that proposed to model hyper ensembles in general physical problems by using an entropic prior (see derivation in Appx. S1).

The end result is a collection of ‘landscape-tilted’ ensembles. That is, each conformational ensemble is a perturbed version of the initial molecular dynamics ensemble but reweighted (see Appx. S2) according to energetic perturbations that are linear in the experimental observables  $f_i(x)$ :

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

It is informative to consider the case of a single observable  $f_1(x)$  (and therefore a single parameter  $\alpha_1$ ). Suppose the molecule of interest shows a bimodal observable with two equally populated states. If we let  $\alpha_1 = 0$ , then the biasing potential is 0 everywhere and our reweighted ensemble simply returns the results of the MD simulation (Fig. 3b). If we let  $\alpha_1 = -1$ , conformations with large values of  $f_1(x)$  are upweighted, while conformations with lower values of  $f_1(x)$  are downweighted (Fig. 3a). Finally, if  $\alpha_1 = 1$ , the ensemble shifts in the opposite direction (Fig. 3c).

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 the parameters  $\alpha_i$  has not yet been discussed. The key idea, however, is that the  $\alpha$  reweighted ensemble  $\langle \rangle_\alpha$  should recapitulate the experimental measurements:

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

Forcing this to be an exact equality recovers previous results (15) that can be derived from maximum entropy considerations (Appx. S1); here, however, we take into account the experimental uncertainties associated with each  $F_i$ .

## Determining $\alpha$

We now derive a Bayesian framework for determining the coefficients  $\alpha$  used in the biasing potential. An alternative derivation using the Crooks hyperensemble formalism (20) is given in Appx. S1.

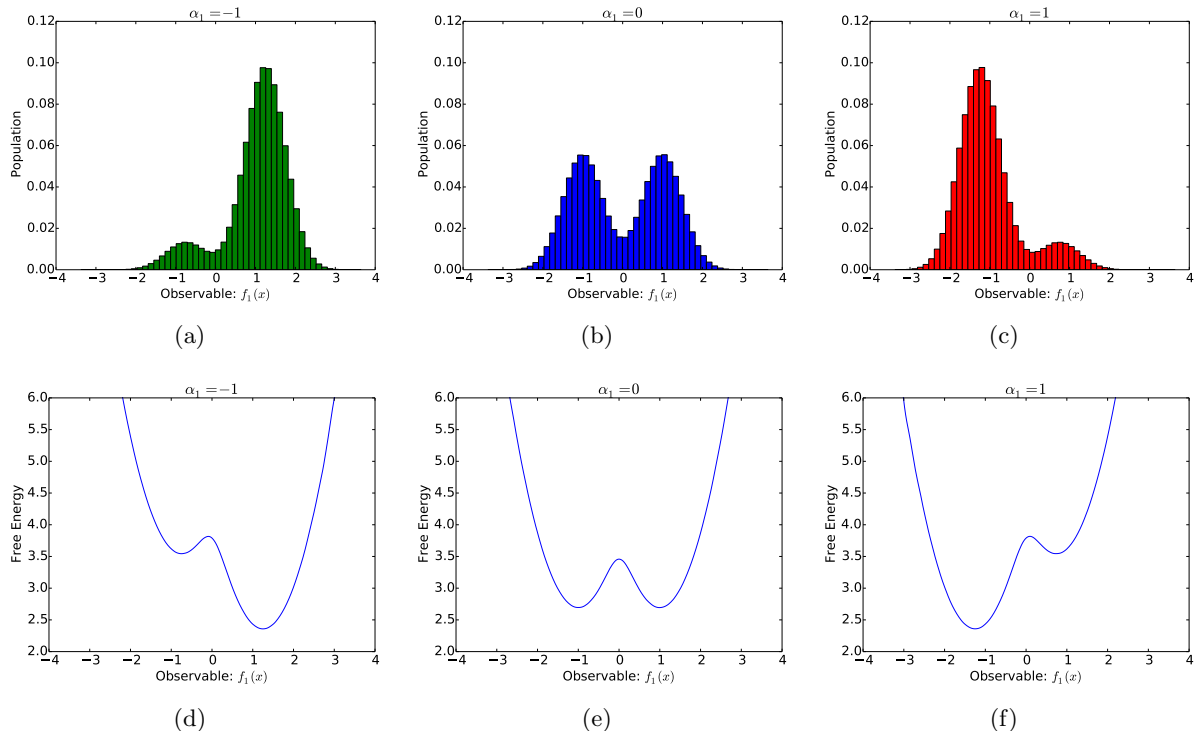


Figure 3: (a, b, c): Raw ( $\alpha_1 = 0$ ) and reweighted (e.g. tilted) histograms of a one dimensional observable. (d, e, f): The same, but plotted as free energies (e.g.  $-kT \log(p)$ ).

BELT assumes that, given the correct choice of  $\alpha$ , the predicted observables  $f_i(x)$  provide unbiased (but noisy) predictions of the measurements  $F_i$ . This assumption is codified by the following conditional probability:

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

For the current work, we model  $\sigma_i$  as the uncertainty associated with predicting chemical shifts and scalar couplings from structures; this error is quantified by the RMS uncertainty estimated during the parameterization of chemical shift and scalar coupling models. Using Bayes' Theorem, we can calculate the posterior distribution of  $\alpha$ :

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

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

$$LP(\alpha) = \log[P(\alpha|F_1, \dots, F_n)] = - \sum_i^n \frac{1}{2\sigma_i^2} (\langle f_i(x) \rangle_\alpha - F_i)^2 + \log P(\alpha) + \text{constant}$$

Note the simple form of the log posterior. The first term (i.e. the log likelihood) measures the  $\chi^2$  agreement between the reweighted ensemble and measurements. The second term is the log of the prior distribution on  $\alpha$ .

In the present work, we evaluate three different choices of prior (Appx. S4), finding similar results for each. The first is the maximum entropy (maxent) prior, which penalizes ensembles as they deviate from the raw simulation results:

$$\log P(\alpha) = -\lambda \sum_j^m \pi_j(\alpha) \log \frac{\pi_j(\alpha)}{\pi_j^0}$$

In the previous expression,  $\pi_j^0$  refers to the populations of an unweighted ensemble, which are typically  $\frac{1}{m}$ , while  $\lambda$  is a hyperparameter that controls the strength of the prior. We also consider using a Dirichlet prior, which is functionally similar to the maxent prior (Appx. S4):

$$\log P(\alpha) = -\lambda \sum_j \pi_j^0 \log \frac{\pi_j^0}{\pi_j(\alpha)}$$

The third prior we consider is a multivariate normal prior, where  $\alpha \sim N(0, \Sigma)$ . The value of  $\Sigma$  is given by  $\Sigma_{ij} = \lambda \text{Cov}(f_i(x), f_j(x))$ , as derived in Appx. S4.

These priors can be used to achieve regularization, which is a powerful technique to reduce overfitting (21). Large values of  $\lambda$  favor the raw simulation results (i.e. uniform conformational populations):  $\pi_j \approx \pi_j^0 = \frac{1}{m}$ . The value of  $\lambda$  can be chosen via cross-validation or other methods (see Appx. S5). When using the maxent prior in the limit of  $\lambda \rightarrow \infty$  and  $\sigma \rightarrow 0$ , BELT recovers the hyperensemble picture of nonequilibrium statistical mechanics as developed (20) by Crooks (See Appx. S1). The Dirichlet and Normal priors do not share the same connection to the Crooks hyperensemble formalism; however, for normally distributed observables, all three priors will give identical results (22).

## MCMC Sampling of Structural Ensembles

As noted above, because ensemble inference often presents many plausible solutions (23, 24), we avoid statistical methods that return a single solution (e.g. maximum likelihood or maximum entropy). We therefore use Markov chain Monte Carlo (MCMC), as implemented in PyMC (16), 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 or experimental observables. Similarly, examining the MCMC variances provides statistical uncertainties of equilibrium or structural features. A Bayesian bootstrapping procedure (25) can also be used to model the statistical uncertainty of the MD simulations (see Appx. S6).

## Methods

### Molecular Dynamics Simulations

Trialanine was simulated in the ff96 (26), ff99 (27), ff99sbnmr-ildn (28, 29), CHARMM27 (30, 31), and OPLS-AA (32) force fields, as previously reported (11). Simulations were performed using Gromacs 4.5 (33) and run at constant temperature (300 K) and pressure (1.01 atm). Each simulation was at least 225 ns long. Conformations were stored every 1 ps.

### Chemical Shifts and Scalar Couplings

All NMR measurements in this work refer to experiments (18) probing the central residue of trialanine. Note that the experimental data was measured at pH 2, near the pKa of the carboxylate moiety of the C terminus. This indicates that the true ensemble likely requires a constant pH simulation, rather than a fixed protonation state. Because such simulations are challenging with



current force fields and simulation packages, we simulated the trialanine construct with charged termini—where the the force fields have been best calibrated and tested. We therefore focus our analysis on the central alanine residue, which should be most robust to pH dependent effects. Both pH differences and force field inaccuracies will lead to systematic differences between simulation and experiment; indeed, we assess whether BELT robustly corrects these deviations.

Chemical shifts ( $H$ ,  $H^\alpha$ ,  $C^\alpha$ ,  $C^\beta$ ) for each frame were calculated using a weighted average of ShiftX2 (34), SPARTA+ (35), and PPM (36) predictions; uncertainties for each model were estimated using their reported RMS prediction errors. Overall uncertainties were estimated as  $\sqrt{\sum w_i \sigma_i^2}$ , where  $w_i \propto \frac{1}{\sigma_i^2}$  is the weight ( $\sum_i w_i = 1$ ) of each chemical shift model and  $\sigma_i$  is the uncertainty of each chemical shift model. The J couplings were calculated using the following Karplus relations:  $^3J(H^N C')$  (19),  $^3J(H^N H^\alpha)$  (19),  $^2J(NC^\alpha)$  (18),  $^3J(H^\alpha C')$  (37),  $^1J(NC^\alpha)$  (18),  $^3J(H^N C^\beta)$  (19). J coupling uncertainties were approximated as the RMS errors reported when fitting the Karplus coefficients.

We have divided the available experimental measurements into training and test sets, with the training set consisting of the  $^3J(H^N C')$ ,  $^2J(NC^\alpha)$ , and  $^3J(H^N C^\beta)$  scalar couplings and the  $C^\alpha$ ,  $H^N$ , and  $C^\beta$  chemical shifts. The test set consists of  $^3J(H^N H^\alpha)$ ,  $^3J(H^\alpha C')$ ,  $^1J(NC^\alpha)$ , and the  $H^\alpha$  chemical shift. The division into training and test sets serves three purposes. First, it provides a test of overfitting. Second, it allows us to reduce the computational cost of BELT calculations. Third, it allows us to train on data that are approximately uncorrelated; BELT is best suited for working with uncorrelated data. Additional suggestions for data curation are provided in Discussion.

## BELT

All BELT calculations were performed using the FitEnsemble package (<https://github.com/kyleabeauchamp/FitEnsemble>). The online FitEnsemble tutorial demonstrates the use of BELT with a single experimental measurement ( $^3J(H^N H^\alpha)$ ). Source code for calculations in this work will be made available at <https://github.com/kyleabeauchamp/EnsemblePaper>.

Regularization strength was determined via cross validation on the simulation data, as described in Appx. S5; this form of cross validation reduces errors due to finite sampling of equilibrium properties. For each model, we used PyMC to sample at least 5,000,000 values of  $\alpha$ ; sampled values of  $\alpha$  were thinned 100-fold to reduce correlation. The first 5,000 samples (before thinning) were discarded as burn-in. Convergence of MCMC sampling was assessed by visual examination of MCMC traces; a well-sampled and thinned trace will appear to be white noise, without correlation between one sample and the next. MCMC traces are shown in Fig. S2 and discussed in Appx. S7. To incorporate simulation uncertainty, we used Bayesian Bootstrapping (Appx. S6). Two Bayesian bootstrap replicates were performed.

## Results

Short peptides provide crucial tests for evaluating and optimizing molecular dynamics force fields (9, 11, 14, 18, 38). 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 (39). Here, we use BELT to infer the conformational populations of trialanine from chemical shift and scalar coupling measurements (18).

## Conformational Propensities of Trialanine Simulations

Trialanine was simulated (see Methods) in five different force fields; these simulations are summarized here. The chosen force fields show considerable variation in their predicted conformational propensities. The ff96 force field shows a bias towards  $\beta$  conformations (population: 51%) (Fig. 4b, red). On the other hand, ff99 strongly favors helical conformations, with a predicted  $\alpha$  population of 80% (Fig. 4c, red). The  $PP_{II}$  state, known to be the dominant state in solution (18, 38, 40), is the dominant simulated state only in the ff99sbnmr-ildn force field (Fig. 4a, red). Low  $PP_{II}$  populations and inconsistency between force fields have been previously noted (9, 11, 14, 18).

## Agreement with NMR Measurements: MD and BELT Ensembles

Given the differences in conformational propensities, one might expect varying degrees of agreement with the available experimental measurements. This is indeed the case; four out of five force fields show large values of the reduced  $\chi^2$  (e.g.  $\frac{1}{n}\chi^2$ ) (Fig. 5a, red). Because of this considerable error, we therefore examined BELT ensembles trained on six NMR measurements. As expected, the BELT ensembles accurately recapitulate these six measurements (Fig. 5a). In a more incisive test, the BELT ensembles accurately predict four measurements that were not used to fit the models. (Fig. 5b). A table of predicted and observed NMR measurements is given in Tables 1, S1, and S2.

## Converged Conformational Propensities Observed in BELT Ensembles

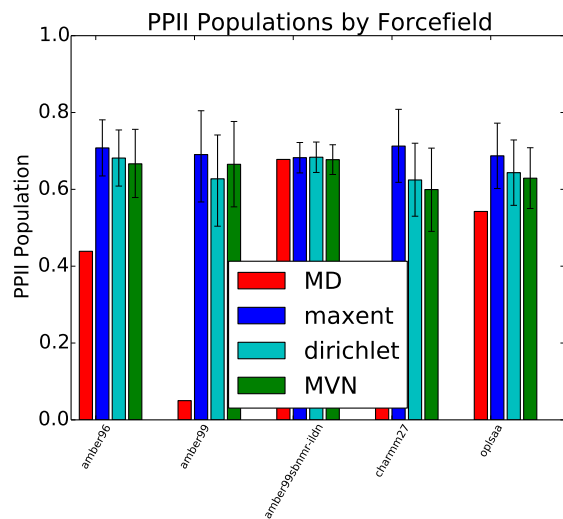
Although the raw MD simulations predicted quite different conformational propensities, BELT reweighting gives five ensembles with conformational populations that agree to within statistical uncertainty (Fig. 4). In general, we find ( $PP_{II}$ ,  $\beta$ ,  $\alpha$ ) populations of ( $67 \pm 9\%$ ,  $23 \pm 6\%$ ,  $10 \pm 8\%$ ); here the mean and uncertainty are approximated as the mean and standard deviation across all force fields and priors. Quantitative predictions and uncertainties are given in Tables S3-S6.

## Effect of Prior

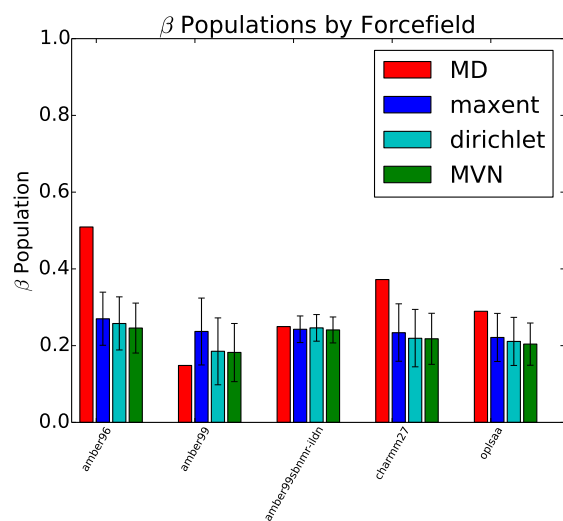
In addition to convergence between models constructed from different force fields, one can also assess the convergence between BELT models built using different priors on the parameters  $\alpha$ . In general, different priors give similar results with small quantitative differences (Figs. 4 and 5). Building BELT models with different priors could therefore be useful for bracketing uncertainties in situations with limited simulation data.

## The Resolution Limit of Trialanine BELT Ensembles

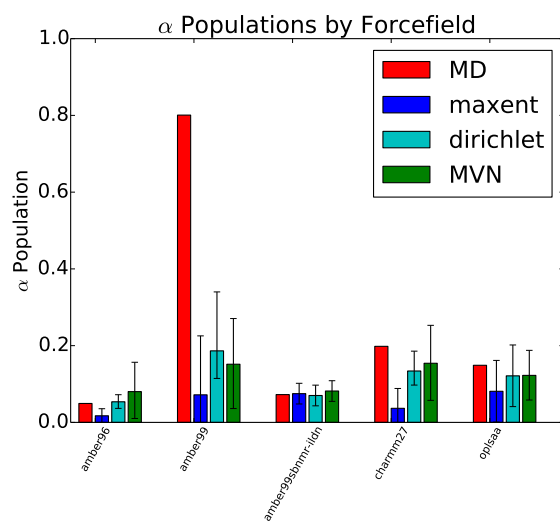
Despite the near-quantitative agreement in  $\alpha$ ,  $\beta$ , and  $PP_{II}$  populations (Fig. 4) and overall Ramachandran features (Fig. 6), the fine structure of the Ramachandran plots differs between the five models. Because all five BELT ensembles show excellent agreement with experiment (Fig. 5), we conclude that six chemical shifts and scalar couplings are insufficiently informative to resolve (and falsify) subtle force field differences. The most obvious such difference is the width, shape, and orientation of the  $PP_{II}$  basin. Most strikingly, ff96 and OPLS-AA have  $PP_{II}$  basins that are vertically oriented, while ff99, ff99sbnmr-ildn, and CHARMM27 show diagonally oriented  $PP_{II}$  basins. Two different effects contribute to this resolution limit: the information content in the experimental measurement and the uncertainty in predictors of experimental observables.



(a)



(b)



(c)

Figure 4: MD and BELT (maxent, Dirichlet, and MVN priors) conformational propensities (for central alanine residue) in each force field.

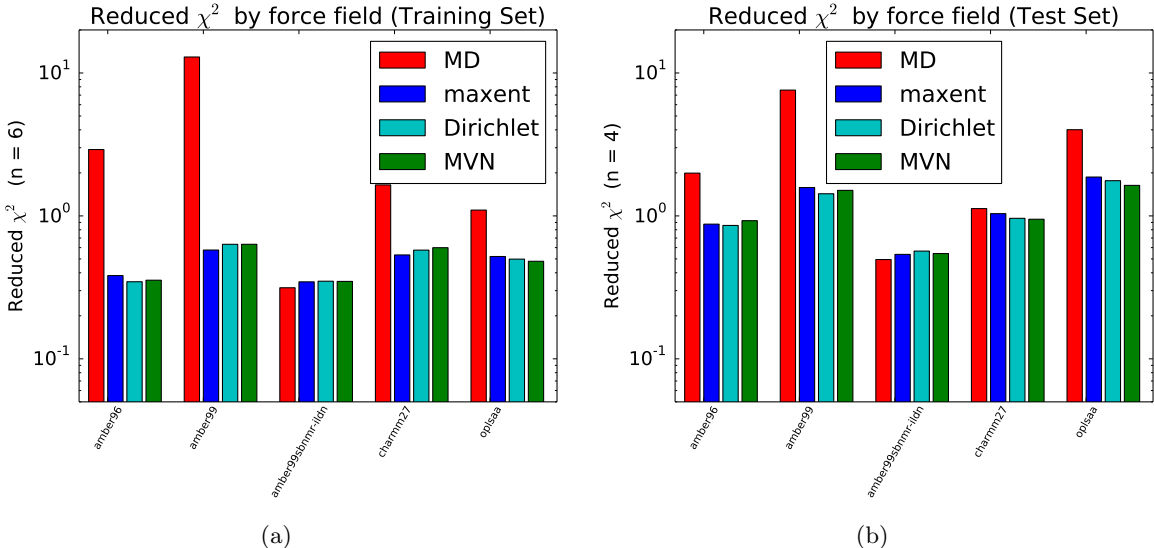


Figure 5: The reduced  $\chi^2$  error (e.g.  $\frac{\chi^2}{n}$ ) for MD and BELT (maxent, Dirichlet, and MVN priors) models. The BELT reduced  $\chi^2$  is estimated as the mean reduced  $\chi^2$  over all MCMC samples. (a). Calculated using the six measurements used to fit the BELT model. (b). Calculated using four measurements not used to fit the BELT model. See Methods for the definition of training and test sets. Note that the training and test sets are not fully independent because all measurements probe the  $(\phi, \psi)$  backbone torsions.

## 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 that fold, unfold, bind ligands, aggregate, and change conformation. Biology is controlled by the relative populations of these states. Ensembles capture aspects of these phenomena by encoding equilibrium populations with structures. A second argument for ensemble modeling is fidelity to experiment. Most solution experiments measure ensemble average equilibrium properties—chemical shifts, scalar couplings, NOEs, SAXS, and FRET can often be approximated as equilibrium properties. A truly quantitative connection to these measurements requires modeling the equilibrium ensemble. Finally, recent advances in atomistic simulation (33, 41–43), special-purpose hardware (44), and distributed computing analysis (45, 46) have enabled atomistic simulations to reach the millisecond timescale (47–50); 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 modeling approaches. However, we suggest that the number of states be inferred—not assumed.

### Comparison to Previous Ensemble Methods

Most previous ensemble modeling efforts involve a protocol with three key ingredients: state decomposition, a  $\chi^2$  objective function, and population inference on the clusters. For example, this general recipe describes the approach used in previous analyses of homopeptides (18), the EROS

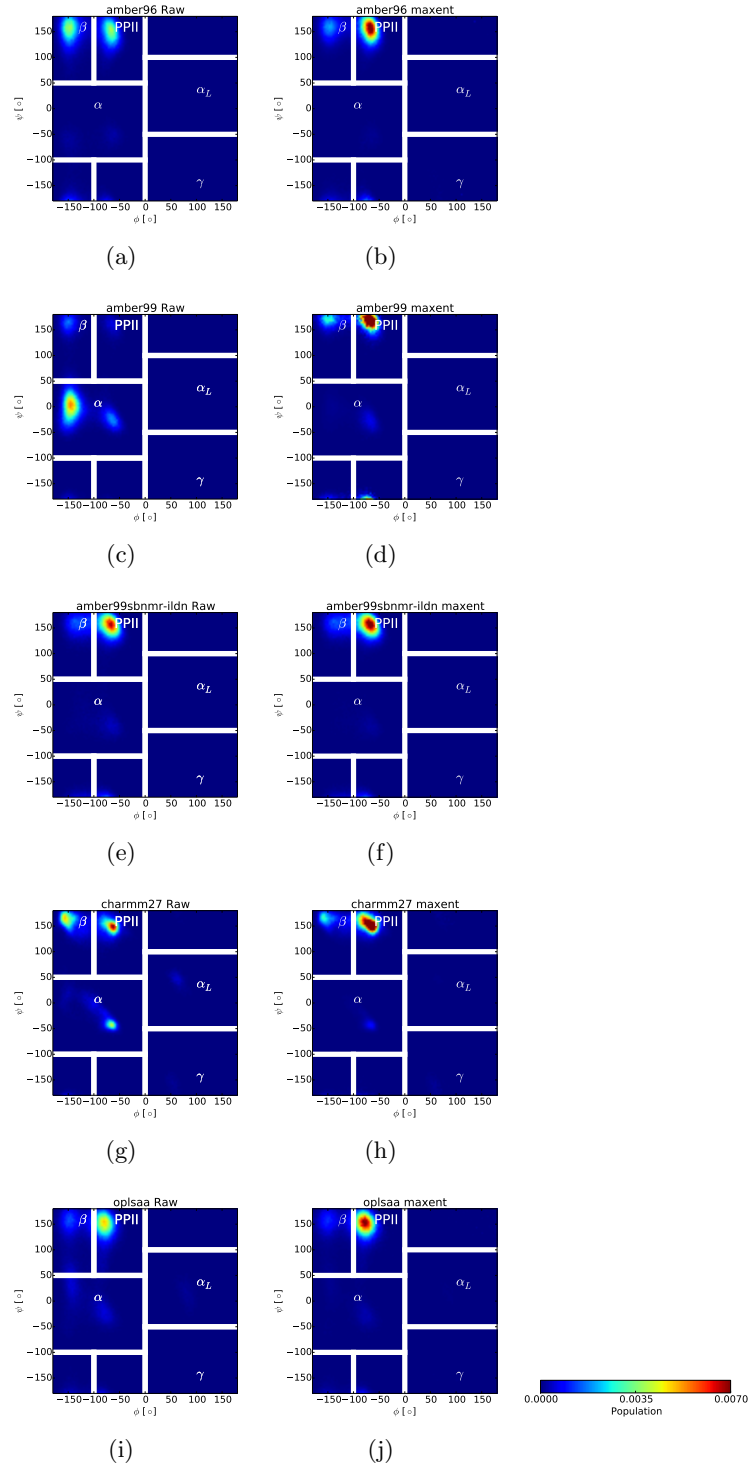


Figure 6: Ramachandran plots of MD and BELT (maxent prior) ensembles for each force field; additional priors are shown in Fig. S1. The jagged appearance of the ff99 BELT model is due to limited sampling of PPII configurations in that forcefield.

technique for SAXS modeling (17), and the Bayesian Weighting (BW) formalism (23). Note that of these three techniques, only BW goes beyond returning a single best-fit ensemble and instead characterizes the posterior distribution via MCMC; below we therefore focus our attention on BW as it is most directly comparable to BELT in scope and purpose.

The primary disadvantage of previous techniques is the need for a state decomposition, which can be defined either by hand or by clustering. Working with a given state decomposition can introduce two different errors, depending on the number and quality of states. In the limit of few states, clustering can overly coarsen the system of interest, preventing the model from reproducing multiple experimental observables. At the other extreme, too many states leads to a large number of parameters to be estimated. This will lead to poor generalization performance and large errors when predicting experiments not used to train the model, as well as reliance on a subjective choice of how many states is appropriate. One symptom of this regime is discontinuity in conformational populations. For example, imagine two nearby conformations at the boundary between two BW states—one conformation on each side of the boundary. In BW, the populations of each conformation could fluctuate dramatically with the corresponding state populations. In BELT, however, the two conformations will have nearly identical populations if the predicted observables vary smoothly.

BELT avoids arbitrary state decompositions by projecting simulations onto a basis defined by the predicted experimental observables. The advantage of working in this basis are threefold. First, in BELT, one estimates a single parameter ( $\alpha_i$ ) for each experimental observable. If the number of experiments is small, as is often the case, the inference problem involves only a few parameters. Second, the predicted observables are a natural basis for biophysical calculations, in that the predicted observables are the fundamental connection between simulation and experiment. Working in this basis allows direct connection to experiment and often provides insight into the molecular interactions driving biophysical phenomena. For example, the projection onto observables could be used to rationally infer force field parameters—essentially a Bayesian version of the ForceBalance method (51, 52). Third, in the limit of exact measurements, BELT reduces to a previous (15) maximum entropy approach (see Appx. S1).

We also point out some surprising differences between BELT and BW-like methods. BW-like methods have the property that the in-state means of features are preserved, leading to an undesirable dependence on the choice of state decomposition. More precisely, suppose that  $\chi_s(x)$  is the indicator function of a conformational state  $s$ . Then in-state averages of the form  $\langle \chi_s(x) \rangle^{-1} \langle h(x) \chi_s(x) \rangle$  do not depend on the reweighted populations. BELT, however, does not preserve the in-state averages; in fact, this property is the direct result of BELT’s connection to maximum entropy modeling (see Appx. S1 and ref. (15)). The effect of this property is that the peaks of reweighted histograms are slightly shifted relative to the raw MD results, as observed in Fig. 3.

## Comparison to a Previous Trialanine Study

Our results are in qualitative, but not quantitative, agreement with a previous study of trialanine (18). That study suggested a  $PP_{II}$  population as high as  $92 \pm 5\%$ , somewhat higher than our  $67 \pm 9\%$  and with a twofold lower estimated uncertainty. The difference can be attributed to three methodological differences. First, the previous study used likelihood maximization to directly fit the ( $PP_{II}$ ,  $\beta$ , and  $\alpha$ ) populations from a three-state decomposition of their simulations. The use of likelihood maximization may give misleading results when the likelihood surface is broad and shallowly peaked, as was found in the previous study. However, this does not appear to be the primary cause of disagreement, as maximization of the BELT likelihood recovers populations within  $\pm 5\%$  of the values obtained via MCMC sampling. Second, the previous study assumed each

scalar coupling to have an uncertainty of 1, while we approximate the uncertainties as the RMS errors determined when fitting the Karplus equations. This weights the measurements differently and will lead to quantitative differences in estimated populations. Different choices of Karplus coefficients also may lead to different predicted properties, as has been discussed elsewhere (53). Finally, differences in state decomposition may cause slight differences in estimated conformational populations.

## Curating Data and Errors for BELT

Because the BELT log likelihood weights errors quadratically, it is vital to use the highest quality experimental measurements, predictions, and error estimates. We recommend that users manually inspect all measured and predicted observables before performing BELT analysis. As discussed above, understanding the errors on different measurements affects the inferred ensemble properties. We discuss another example we encountered in the current analysis.

Scalar couplings predicted using parameterized Karplus relations will span a limited range that is determined by the Karplus coefficients. In several cases, however, experimentally measured J couplings lie outside this range—meaning that even a perfect force field would be unable to recapitulate the experimental measurements. Such measurements indicate limits in the transferability of simple Karplus prediction of scalar couplings; any such examples are best excluded from BELT analysis. Improved Karplus models for scalar couplings are clearly desirable.

## Limitations of BELT

There are several limitations of the BELT method. Most obvious is that BELT can only reweight existing simulations. The force field used must be sufficiently accurate to sample experimentally-consistent conformations. Achieving accurate predictions using poor force fields may require orders of magnitude more MD sampling. The same increased sampling requirements apply as one attempts to fit many measurements simultaneously. In the worst-case scenario, fitting  $n$  measurements could require MD datasets whose size grows exponentially with  $n$ . A second limitation in BELT is its computational cost. In the present calculations, MCMC sampling for each model took approximately 24 hours. These datasets included six measurements and approximately 250,000 conformations. Because using additional conformations is computationally prohibitive, BELT ensembles are essentially limited to resolving states that have populations of at least  $\frac{1}{250000}$ . A third limitation is the independent normal approximation used in BELT’s error model (see Appx. S3). More sophisticated error models may be possible, including the Bayesian formalisms from the Nilges group (24, 54, 55). Finally, BELT is currently formulated for use with equilibrium measurements, meaning that the present framework cannot be used to predict kinetic properties.

## Conclusion

Bayesian Energy Landscape Tilting allows the simultaneous characterization of structural and equilibrium properties. Through its use of MCMC, BELT is robust to ambiguous experiments and provides rigorous uncertainty estimates, as illustrated here in the case of a simple tripeptide system. BELT models constructed with a handful of NMR measurements correct significant force field bias and provide generalizable, force field independent alanine ensembles. The principled combination of simulation and experiment will enable robust, incisive predictions of the atomic-scale behavior of macromolecules.

## Acknowledgements

We thank John Chodera, TJ Lane, Frank Cochran, Pehr Harbury, Xuesong Shi, and Dan Herschlag for helpful discussions.



	$F_i$	$\sigma_i$	ff96		ff99		ff99sbnmr-ildn		CHARMM27		OPLS-AA	
			MD	BELT	MD	BELT	MD	BELT	MD	BELT	MD	BELT
$C^\alpha$	52.4	0.9	52.1	52.2	52.8	52.5	52.4	52.4	52.5	52.4	52.2	52.2
$C^\beta$	19.2	1.0	20.0	19.8	18.0	18.8	18.3	18.4	18.2	18.6	19.6	19.6
$H$	8.6	0.5	8.6	8.6	8.3	8.4	8.2	8.2	8.3	8.3	8.6	8.6
$H^\alpha$	4.4	0.2	4.6	4.6	4.6	4.6	4.5	4.5	4.6	4.6	4.6	4.6
$^1J(NC^\alpha)$	11.3	0.5	11.3	11.5	10.4	11.8	11.5	11.5	11.2	11.7	11.1	11.3
$^3J(H^\alpha C\prime)$	1.8	0.4	2.0	1.7	2.2	1.7	1.8	1.8	2.0	1.8	2.2	2.0
$^3J(H^N C^\beta)$	2.4	0.2	1.5	2.3	0.8	2.3	2.3	2.3	1.8	2.3	1.9	2.2
$^3J(H^N C\prime)$	1.1	0.3	1.5	1.2	1.8	1.2	1.0	1.0	1.4	1.2	0.9	0.8
$^3J(H^N H^\alpha)$	5.7	0.4	6.6	5.7	7.5	5.6	6.1	6.0	6.3	5.7	7.0	6.5
$^2J(NC^\alpha)$	8.4	0.5	8.5	8.6	6.4	8.5	8.5	8.5	8.1	8.6	8.1	8.4
$\chi^2$ (all)			2.5	0.6	10.8	1.0	0.4	0.4	1.4	0.7	2.3	1.1
$\chi^2$ (train)			2.9	0.4	12.9	0.6	0.3	0.3	1.6	0.5	1.1	0.5
$\chi^2$ (test)			2.0	0.9	7.6	1.6	0.5	0.5	1.1	1.0	4.0	1.9

Table 1: Predicted and measured observables are given. BELT predictions are calculated using the maxent prior; see Tables S1-S7 for complete table. The ‘all’, ‘training’, and ‘test’ datasets have 10, 6, and 4 measurements, respectively.

## References

- [1] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
- [2] Dethoff, E. A., K. Petzold, J. Chugh, A. Casiano-Negroni, and H. M. Al-Hashimi, 2012. Visualizing transient low-populated structures of RNA. *Nature* .
- [3] Fink, A. L., 2005. Natively unfolded proteins. *Current opinion in structural biology* 15:35–41.
- [4] Korzhnev, D. M., X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson, and L. E. Kay, 2004. Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* 430:586–590.
- [5] Lindorff-Larsen, K., R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, 2005. Simultaneous determination of protein structure and dynamics. *Nature* 433:128–132.
- [6] Lange, O. F., N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. de Groot, 2008. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 320:1471–1475.
- [7] DePristo, M. A., P. I. de Bakker, and T. L. Blundell, 2004. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12:831–838.
- [8] Lang, P. T., H.-L. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, and T. Alber, 2010. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science* 19:1420–1431.
- [9] Best, R., N. Buchete, and G. Hummer, 2008. Are current molecular dynamics force fields too helical? *Biophys. J.* 95:L07–L09.
- [10] Lindorff-Larsen, K., P. Maragakis, S. Piana, M. Eastwood, R. Dror, and D. Shaw, 2012. Systematic Validation of Protein Force Fields against Experimental Data. *PloS one* 7:e32131.
- [11] Beauchamp, K., Y. Lin, R. Das, and V. Pande, 2012. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* 8:1409.
- [12] Li, D.-W., and R. Bruschweiler, 2011. Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. *J. Chem. Theory Comput.* 7:1773.
- [13] Best, R. B., X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, 2012. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone  $\varphi$ ,  $\psi$  and side-chain  $\chi_1$  and  $\chi_2$  dihedral angles. *J. Chem. Theory Comput.* .
- [14] Nerenberg, P., and T. Head-Gordon, 2011. Optimizing Protein- Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. *J. Chem. Theory Comput.* 7:1220–1230. <http://pubs.acs.org/doi/abs/10.1021/ct2000183>.
- [15] Pitera, J., and J. Chodera, 2012. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* .
- [16] Patil, A., D. Huard, and C. J. Fonnesbeck, 2010. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software* 35:1.

- [17] Rozycki, B., Y. C. Kim, and G. Hummer, 2011. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19:109–116.
- [18] Graf, J., P. Nguyen, G. Stock, and H. Schwalbe, 2007. Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *J. Am. Chem. Soc.* 129:1179–1189.
- [19] Vögeli, B., J. Ying, A. Grishaev, and A. Bax, 2007. Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. *J. Am. Chem. Soc.* 129:9377–9385.
- [20] Crooks, G. E., 2007. Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out of equilibrium. *Physical Review E* 75:041119.
- [21] Friedman, J., T. Hastie, and R. Tibshirani, 2001. The elements of statistical learning, volume 1. Springer Series in Statistics.
- [22] Wikipedia, 2004. Kullback–Leibler divergence — Wikipedia, The Free Encyclopedia. [http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence](http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence), [Online; accessed 15-July-2013].
- [23] Fisher, C. K., A. Huang, and C. M. Stultz, 2010. Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.* 132:14919.
- [24] Rieping, W., M. Habeck, and M. Nilges, 2005. Inferential structure determination. *Science* 309:303–306.
- [25] Rubin, D., 1981. The bayesian bootstrap. *The annals of statistics* 9:130–134.
- [26] Kollman, P., 1996. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* 29:461–469.
- [27] Wang, J., P. Cieplak, and P. Kollman, 2000. How well does a restrained electrostatic potential(RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21:1049–1074.
- [28] Li, D., and R. Bruschweiler, 2010. NMR-Based Protein Potentials. *Angew. Chem.* 122:6930–6932.
- [29] Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. Klepeis, R. Dror, and D. Shaw, 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* 78:1950–1958.
- [30] Mackerell Jr, A., M. Feig, and C. Brooks III, 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25:1400–1415.
- [31] Bjelkmar, P., P. Larsson, M. Cuendet, B. Hess, and E. Lindahl, 2010. Implementation of the CHARMM force field in GROMACS: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *J. Chem. Theory Comput.* 6:459–466.
- [32] Kaminski, G., R. Friesner, J. Tirado-Rives, and W. Jorgensen, 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* 105:6474–6487.

- [33] Hess, B., C. Kutzner, D. Van Der Spoel, and E. Lindahl, 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.
- [34] Han, B., Y. Liu, S. Ginzinger, and D. Wishart, 2011. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* 1–15.
- [35] Shen, Y., and A. Bax, 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* 48:13–22.
- [36] Li, D.-W., and R. Brüschweiler, 2012. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biol. NMR* 1–9.
- [37] Schmidt, J., M. Blümel, F. Lohr, and H. Ruterjans, 1999. Self-consistent 3J coupling analysis for the joint calibration of Karplus coefficients and evaluation of torsion angles. *J. Biomol. NMR* 14:1–12.
- [38] Grdadolnik, J., V. Mohacek-Grosev, R. Baldwin, and F. Avbelj, 2011. Populations of the three major backbone conformations in 19 amino acid dipeptides. *Proc. Natl. Acad. Sci. U. S. A.* 108:1794.
- [39] Jha, A. K., A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed, 2005. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry* 44:9691–9702.
- [40] Avbelj, F., S. Grdadolnik, J. Grdadolnik, and R. Baldwin, 2006. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl. Acad. Sci. U. S. A.* 103:1272.
- [41] Pronk, S., S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, and D. van der Spoel, 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* .
- [42] Eastman, P., M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, and V. S. Pande, 2012. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* 9:461–469.
- [43] Eastman, P., and V. Pande, 2010. OpenMM: a hardware-independent framework for molecular simulations. *Comp. in Sci. Eng.* 12:34–39.
- [44] Shaw, D., M. Deneroff, R. Dror, J. Kuskin, R. Larson, J. Salmon, C. Young, B. Batson, K. Bowers, and J. Chao, 2008. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51:91–97.
- [45] Senne, M., B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé, 2012. EMMA - A software package for Markov model building and analysis. *J. Chem. Theory Comput.* .
- [46] Beauchamp, K., G. Bowman, T. Lane, L. Maibaum, I. Haque, and V. Pande, 2011. MSM-Builder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* 7:3412–3419.
- [47] Voelz, V., G. Bowman, K. Beauchamp, and V. Pande, 2010. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9 (1- 39). *J. Am. Chem. Soc.* 132:1526–1528.

- [48] Bowman, G. R., V. A. Voelz, and V. S. Pande, 2011. Atomistic folding simulations of the five helix bundle protein  $\lambda$ 6–85. *J. Am. Chem. Soc.* 133:664.
- [49] Shaw, D. E., P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, 2010. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* 330:341–346.
- [50] Lindorff-Larsen, K., S. Piana, R. Dror, and D. Shaw, 2011. How Fast-Folding Proteins Fold. *Science* 334:517–520.
- [51] Wang, L.-P., J. Chen, and T. Van Voorhis, 2012. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* 9:452–460.
- [52] Wang, L.-P., T. L. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martínez, and V. S. Pande, 2013. Systematic Improvement of a Classical Molecular Model of Water. *J. Phys. Chem. B*.
- [53] Markwick, P. R., S. A. Showalter, G. Bouvignies, R. Brüschweiler, and M. Blackledge, 2009. Structural dynamics of protein backbone  $\varphi$  angles: extended molecular dynamics simulations versus experimental 3 J scalar couplings. *Journal of biomolecular NMR* 45:17–21.
- [54] Habeck, M., W. Rieping, and M. Nilges, 2006. Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci.* 103:1756–1761.
- [55] Habeck, M., W. Rieping, and M. Nilges, 2005. Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar couplings constants. *J. Mag. Res.* 177:160–165.