Model selection using replica averaging with Bayesian inference of conformational populations

Robert M. Raddi, Tim Marshall, Yunhui Ge, and Vincent Voelz*

Department of Chemistry, Temple University, Philadelphia, PA 19122, USA

Abstract

Bayesian Inference of Conformational Populations (BICePs) is a reweighting algorithm that reconciles simulated ensembles with sparse and/or noisy observables, by sampling the full posterior distribution of conformational populations in the presence of experimental restraints. By modifying BICePs to use replica-averaging in its forward model, BICePs becomes similar to other MaxEnt approaches, but with the significant advantages of (1) being able to sample over the posterior distribution of uncertainties due to random and systematic error, with improved likelihoods to deal with outliers, and (2) having an objective score for model selection, a free energy-like quantity called the BICePs score. To demonstrate the power of our approach, we used BICePs to reweight conformational ensembles of the mini-protein chignolin simulated in nine different force fields with TIP3P water, using a set of 158 experimental measurements (139 NOE distances, 13 chemical shifts, and 6 vicinal J-coupling constants for H^N and H^α. In all cases, reweighted populations favor the correctly folded conformation. The BICePs score, which reports the free energy of "turning on" conformational populations along with experimental restraints, provides a metric to evaluate each force field. For the nine force fields tested (A14SB, A99SB-ildn, A99, A99SBnmr1-ildn, A99SB, C22star, C27, C36, OPLS-aa), we obtain results consistent with previous work that used a conventional χ^2 metric for model selection for small polypeptides and ubiquitin (Beauchamp et al 2012). These results suggest a powerful

role for BICePs in future applications requiring ensemble reweighting and model selection.

Keywords Bayesian inference, maximum entropy, molecular simulation, force fields, chignolin, conformational populations, model selection

Significance Statement

Reconciling molecular models of conformational ensembles with ensemble-averaged experimental measurements is a central problem in biophysical chemistry. Bayesian and maximum-entropy approaches for this purpose are state-of-the-art, but uncertainty in both models and measurement still make it challenging to integrate these two kinds of information, much less select optimal models. An improved version of the Bayesian Inference of Conformational Populations (BICePs) algorithm addresses both these issues by sampling distributions of conformational states with restraints to replica-averaged observables. BICePs has the ability to learn uncertainties directly from the data, using improved likelihoods to deal with outliers, and computes a score to objectively assess model quality, called the BICePs score. As demonstrated for the mini-protein chignolin, the combination of maximum-entropy reweighting and model selection makes BICePs a uniquely powerful tool.

Introduction

A central problem in molecular modeling is constructing accurate models of conformational populations that agree with ensemble-averaged experimental measurements. While molecular simulation models can provide valuable microscopic insight, they are limited by the accuracy of the chosen force field, and must be validated by comparing experimental measurements (e.g., NMR observables) with theoretical predictions of those measurements from the simulated ensembles (the "forward model"). In doing so, however, one is faced with several complications. First, the forward model has some error due to the approximate description of the measurement. Second, ensemble-averaged experimental measurements may be sparse and/or noisy, subject to random and systematic errors, which are often unknown a priori. Comparing different models

thus requires some way of integrating these multiple sources of uncertainty to perform objective model selection.

Numerous methods have been developed to address these challenges, most of which use either a maximum entropy (MaxEnt) approach, Bayesian inference, or some combination of the two to reconcile simulation predictions with ensemble-averaged experimental observables.[1–9] Bayesian inference methods estimate a posterior model of conformational populations by treating simulation predictions as prior information, weighted by a likelihood function constructed from the experimental measurements and their uncertainties. MaxEnt methods aim to maximize the relative entropy of population distributions with respect to predicted distributions, given constraints to ensemble-averaged observables. Both approaches often utilize a replica-based sampling approach, where replica-averaged forward model predictions are restrained against ensemble-averaged experimental observables; in the limit of large numbers of replicas, this produces the maximum entropy distribution.[10–14]

Proper consideration of the sources of error and their unknown uncertainties is a major challenge. The Metainference method,[5] and its combined use with MetaDynamics (M&M),[6,15] is a Bayesian inference approach that addresses this challenge by imposing replica-averaged restraints within a simulation, sampling over various restraint strengths to infer the distribution of experimental uncertainties. A drawback of this approach–and others like it[3,11,13,16,17]–is that dynamic restraints must be implemented within a simulation.

Alternatively, many methods *reweight* predicted conformational populations as a post-processing step after simulations.[3–5,8,14,18–24] MaxEnt approaches such as BioEn[14,25] and BME[16,26] reweight populations to maximize entropy with a constraint on the χ^2 metric characterizing the expected error between simulated and experimental observables. A drawback of this approach is the requirement that the expected error must be specified beforehand, determined using a heuristic procedure .

The Bayesian Inference of Conformational Populations (BICePs) algorithm[19,27–31] is a related (see SI Text) but unique MaxEnt method that addresses both of these shortcomings: it is a post-processing reweighting method that does not require knowledge of experimental errors, and instead infers this from the data.

The BICePs algorithm. BICePs uses a Bayesian statistical framework, inspired by Inferential Structure Determination (ISD),[1] to treat the extent of uncertainty in experimental observables, σ, as nuisance parameters. Previous versions of BICePs sampled conformational states X and uncertainty parameter(s) σ from the Bayesian posterior, which takes the form

$$\underbrace{p(X,\sigma|D)}_{\text{posterior}} \propto \underbrace{p(D|X,\sigma)}_{\text{likelihood}} \underbrace{p(X)p(\sigma)}_{\text{priors}}. \tag{1}$$

Here, the prior $p(X)$ comes from a theoretical model of conformational state populations (typically from a molecular simulation), $p(D|X,\sigma)$ is likelihood function quantifying how well a forward model prediction $f(X)$ agrees with the experimental data D, and $p(\sigma) \sim \sigma^{-1}$ is a non-informative Jeffreys prior.

When BICePs is equipped with a replica-averaged forward model, it becomes a MaxEnt reweighting method in the limit of large numbers of replicas.[10–14] The posterior takes the general form

$$p(\mathbf{X},\sigma|D) \propto \prod_{r=1}^{N_r}\Big\{p(X_r)\prod_{j=1}^{N_j}\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\Big[-\frac{(f_j(\mathbf{X})-d_j)^2}{2\sigma_j^2}\Big]p(\sigma_j)\Big\} \tag{2}$$

where $\mathbf{X}$ is a set of N_r conformation replicas, d_j is an observable in the set of N_j ensemble-averaged experimental measurements, and $f_j(\mathbf{X}) = \frac{1}{N_r}\sum_r^{N_r} f_j(X_r)$ is the replica-averaged forward model prediction of observable j. The σ_j values are nuisance parameters that capture uncertainty in the measurements as well as the replica-averaged forward model. In (2), a Gaussian model is used for the likelihood, but more sophisticated models can be used to capture outliers and systematic error with fewer parameters, as discussed below. Markov chain Monte Carlo (MCMC) is used to sample the posterior.

Model selection using the BICePs score. BICePs evaluates model quality by calculating a free energy-like quantity called the BICePs score. For a model k with prior populations $p^{(k)}(X)$, the BICePs score $f^{(k)}$ is computed as the negative logarithm of a Bayes factor comparing the total evidence of a given model against a well-defined reference, marginalizing over

all uncertainty,[29]

$$f^{(k)} = -\ln \frac{Z^{(k)}}{Z_0}, \tag{3}$$

where

$$Z^{(k)} = \int P^{(k)}(\mathbf{X}, \sigma \mid D) d\mathbf{X} d\sigma \tag{4}$$

is the evidence for model k, and Z_0 is the evidence for a suitable reference state. To construct the reference state, we consider a series of priors $p_\lambda(X) \sim [p(X)]^\lambda$ parameterized by $\lambda \in [0, 1]$ and likelihoods $p_\xi(D|\mathbf{X}, \sigma) \sim [p(D|\mathbf{X}, \sigma)]^\xi$ parameterized by $\xi \in [0, 1]$, and set the reference state as the thermodynamic ensemble corresponding to $\lambda = 0, \xi = 0$. The BICePs score is then calculated as the change in free energy of "turning on" experimental restraints with uniform microstate populations ($\lambda = 0, \xi = 0 \rightarrow 1$), and then scaling the prior populations in the presence of the restraints ($\xi = 1, \lambda = 0 \rightarrow 1$) (Figure 1). The calculation is performed using the MBAR free energy estimator, by sampling at several intermediates (see SI Methods).

The above approach is different from previous versions of BICePs, which used the ($\lambda = 0, \xi = 1$) ensemble as the reference, assuming the conformational state definitions are identical for each model. We have found this criterion very difficult to achieve in practice for biomolecular simulations (for further discussion, see SI Text).

Note that the BICePs score is an extensive quantity that grows linearly with the number of replicas. For this reason, our results report the *reduced* BICePs score, $f^{(k)}/N_r$.

Likelihoods to better account for systematic error. Systematic error is ubiquitous in both experimental data and the forward model used to predict those same observables. The problem of accounting for such error can be approached in two ways: 1) assign higher uncertainties to particular observables, or 2) neglect those affected by systematic error and treat them as outliers.

Towards the first approach, BICePs can be used with the *Gaussian* model in Equation (2), which uses one uncertainty per data point and is able to automatically assign higher uncertainty to data affected by systematic error. However, as the number of observables grows, sampling over multiple σ_j becomes costly. As an alternative, we introduce two different models: 1) the *Good-and-Bad* error formulation and 2) the *Student's* model. Both approaches marginalize the

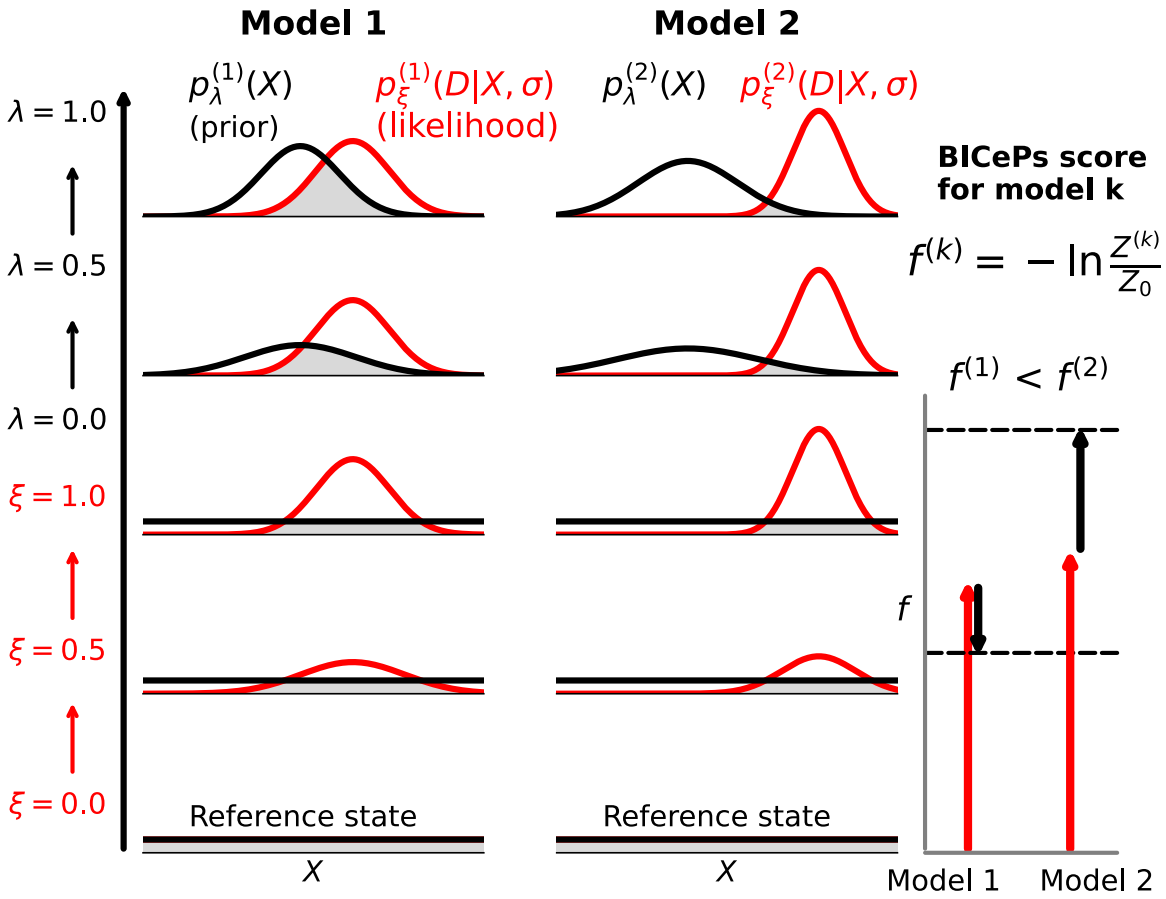

Figure 1: BICePs scores $f^{(1)}$ and $f^{(2)}$ for two theoretical models, $p^{(1)}(\mathbf{X},\sigma|D)$ and $p^{(2)}(\mathbf{X},\sigma|D)$, are calculated using a common reference state, in which the prior and the likelihood with experimental restraints are turned off. To efficiently and accurately calculate ratios of model evidences, a free energy perturbation approach is used, in which posterior sampling is performed for a series of models that scale the likelihood ($\xi = 0 \rightarrow 1$), and then the prior ($\lambda = 0 \rightarrow 1$). The MBAR estimator is then used to calculate the BICePs score as the free energy change of this transformation, $f^{(k)} = -\ln(Z^{(k)}/Z_0)$.

uncertainties for individual observables, assuming a uniform level of noise, except for a few erratic data points. This limits the number of uncertainty parameters that need to be sampled. Plots of the probability density functions for these likelihoods are found in Figures S1-S6 and details are given in Methods.

Overview. In this article, we demonstrate how BICePs can now perform robust MaxEnt reweighting of protein conformational ensembles and select optimal models. Using a simple toy model, we show how BICePs can better deal with various sources of error than existing approaches. We test different likelihood models that can efficiently and accurately detect outliers, and evaluate their performance in reweighting conformational ensembles of the mini-protein chignolin simulated in nine different force fields, against a set of 158 experimental NMR observables. In each case, we compute the BICePs score to objectively evaluate the performance of each force field model.

Results and Discussion

BICePs ensemble reweighting is robust in the presence of unknown random and systematic error. First, to demonstrate the performance of BICePs with noisy experimental data, we use a simple toy model to quantitatively compare our replica-averaging BICePs algorithm with the previous Bayesian single-replica approach, and existing MaxEnt approaches that use fixed estimates of uncertainty (Figure 2).

The toy model consists of three conformational states: a folded state S_0, intermediate state S_1, and unfolded state S_2. Each conformational state is characterized by a collection of 500 intermolecular distances x, normally distributed about means 3.0, 4.5 and 6.0, respectively. The *true* populations of each state are set to 65%, 15%, and 20%, respectively. In this toy example, we assume that experimental observables are directly averaged, such that in an ideal experiment with no sources of error, the 500 distances would be measured as the population-weighted average of each distance.

Next, we generate synthetic experimental measurements affected by random and systematic error, by adding normally-distributed error of standard deviation 0.5, and then systematic error

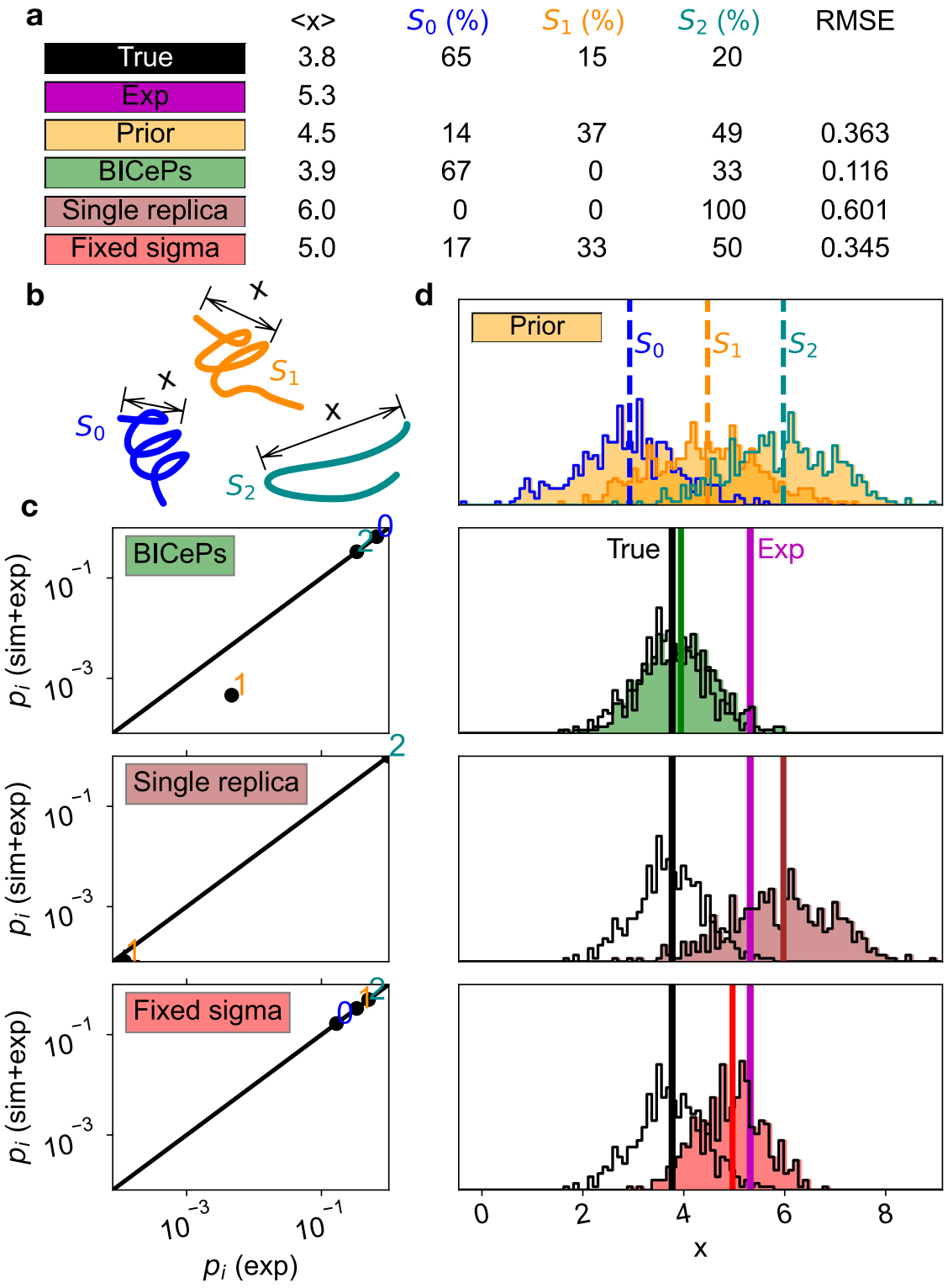

Figure 2: BICePs performs better than existing Bayesian or MaxEnt approaches in the presence of random and systematic error. In this three-state toy model, each conformational state is characterized by a collection of 500 intermolecular distance observables x, normally distributed about means 3.0, 4.5 and 6.0. (a) Table describing ensemble populations (b) Cartoon representation of the three conformational states. (c) Plots of inferred conformational state populations p_i(sim+exp) versus populations p_i(exp) inferred using only the experimental restraints (i.e. a uniform prior). (d) Posterior distributions of the 500 distance observables, with vertical lines denoting the mean of the distributions. The "true" distribution and its mean are shown in black.

by shifting 30% of the distance measurements by +3 to +5, resulting in a set of experimental measurements that vary about a mean value of 4.8. Finally, we generate a prior model of conformational populations by supposing that some theoretical method predicted populations that incorrectly estimate the populations as 13%, 41%, and 46%, respectively (an overall RMSE of 0.363).

With this input data, we proceed to test each method using 100,000 MCMC iterations of BICePs sampling (see Supplemental Methods). The single-replica approach significantly overestimates the population of S_2, resulting in an RMSE of 0.601 compared to the true populations. Due to the lack of replica-averaging in the forward model, the sampled posterior places nearly 100% of the population in this state, as it is favored by both the prior and synthetic experimental data. The fixed-sigma (MaxEnt) ($\sigma = \sigma_{\text{Exp.}} = 0.70$) method does a better job, achieving an RMSE of 0.345 compared to the true populations, but is unable to fully account for systematic error in the experimental data, minimally perturbing the prior, leading to slight over-population of S_1. Both the single-replica and fixed-sigma approaches use a Gaussian likelihood. In contrast, BICePs can use the Student's likelihood model (results shown in Figure 2) or Good-Bad model (Figure S7) that automatically handle systematic error by sampling an extra nuisance parameter. BICePs achieves the best prediction of the true populations, with an RMSE of 0.116. This is because the forward model is able to correctly compare predictions with ensemble-averaged observables, and better deal with systematic error by sampling the complete posterior distribution of the expected uncertainty. For more details and results for the three-state toy system, see the SI.

BICePs reweights conformational populations and ranks the accuracy of simulation models of the mini-protein chignolin in nine force fields. To further demonstrate its utility, we apply BICePs to a series of prior conformational populations derived from all-atom simulations of the beta-hairpin chignolin CLN001,[32] using published NMR measurements as experimental restraints.[33] From over 20 μs of aggregate simulation trajectory data for each force field in TIP3P explicit solvent, Markov state models (MSMs) of chignolin folding were constructed using various numbers $\{5, 10, 50, 75, 100, 500\}$ of microstates defined by conformational clustering (see Methods). For each MSM, microstates

were assigned forward model predictions for a set of 158 observables (139 NOE distances, 13 chemical shifts, and 6 vicinal J-coupling constants for H^N and H^α, shown in Figures S8-S11) and their populations reweighted according to the ensemble-averaged experimental measurements. For BICePs calculations using the Gaussian likelihood model, all observables are assigned individual uncertainty parameters. For other likelihood models, each type of observable is assigned an uncertainty parameter (e.g. σ_{NOE}, σ_J, σ_{cs}).

As shown by others,[34–36] MSM microstates can be structurally categorized into three macrostates: unfolded (U), misfolded (M) and folded (F) (Figure 3a). To illustrate the results of BICePs with various likelihood models, we show reweighted macrostate populations of a 500-microstate MSM (Figures S12-S13) of chignolin CLN001 simulated using A99SB-ildn. This force field overestimates the misfolded population at $\sim$70%, and underestimates the folded population at $\sim$20% (MSM, Figure 3b); the experimentally measured folded-state population is 61.0 ± 3.4 % at 300 K.[33] BICePs-reweighted populations (averaged over five independent trials of 10M MCMC steps each) are able to correctly up-weight the folded macrostate and down-weight the misfolded macrostate. While single-replica BICePs puts nearly all the population in the folded macrostate (maximum parsimony), replica-averaging BICePs (8 replicas) with Good-Bad and Student's likelihoods improves this situation, achieving results comparable to the more costly Gaussian likelihood, which assigns uncertainty parameters to each observable (see Figures S14-17 for posterior distributions of uncertainties and Figure S18 for reweighted observables). Despite the large number of parameters, BICePs sampling is well converged, as shown by traces of energy (the negative logarithm of the posterior) sampled over time (Figures S19-20) and the JSD analysis (Figures S21-22) using convergence tools from the BICePs v2.0 software package.[31]

A comparison of nine different force fields shows the robustness of the BICePs reweighting approach (Figure 4). Despite widely varying predictions of macrostate populations, BICePs correctly upweights the folded macrostate population in all cases. Single-replica BICePs consistently places $\sim$100% of the population in the folded state for all force fields, while replica-averaging (using 8 replicas) correctly achieves maximum-entropy reweighting according to the ensemble-averaged observables. Even force fields that incorrectly predict chignolin to be en-

tirely unfolded (CM36) or misfolded (A99) can be "rescued" by BICePs reweighting, provided that folded conformations are sampled in the input ensemble. The relative entropy difference between the prior and reweighted populations can be measured by the Kullback-Leibler divergence (D_{KL}); these calculations show that A99 requires the most perturbation, with a D_{KL} value of 9.3 for replica-averaging BICePs using a Student's likelihood and 500 states (Figure S23).

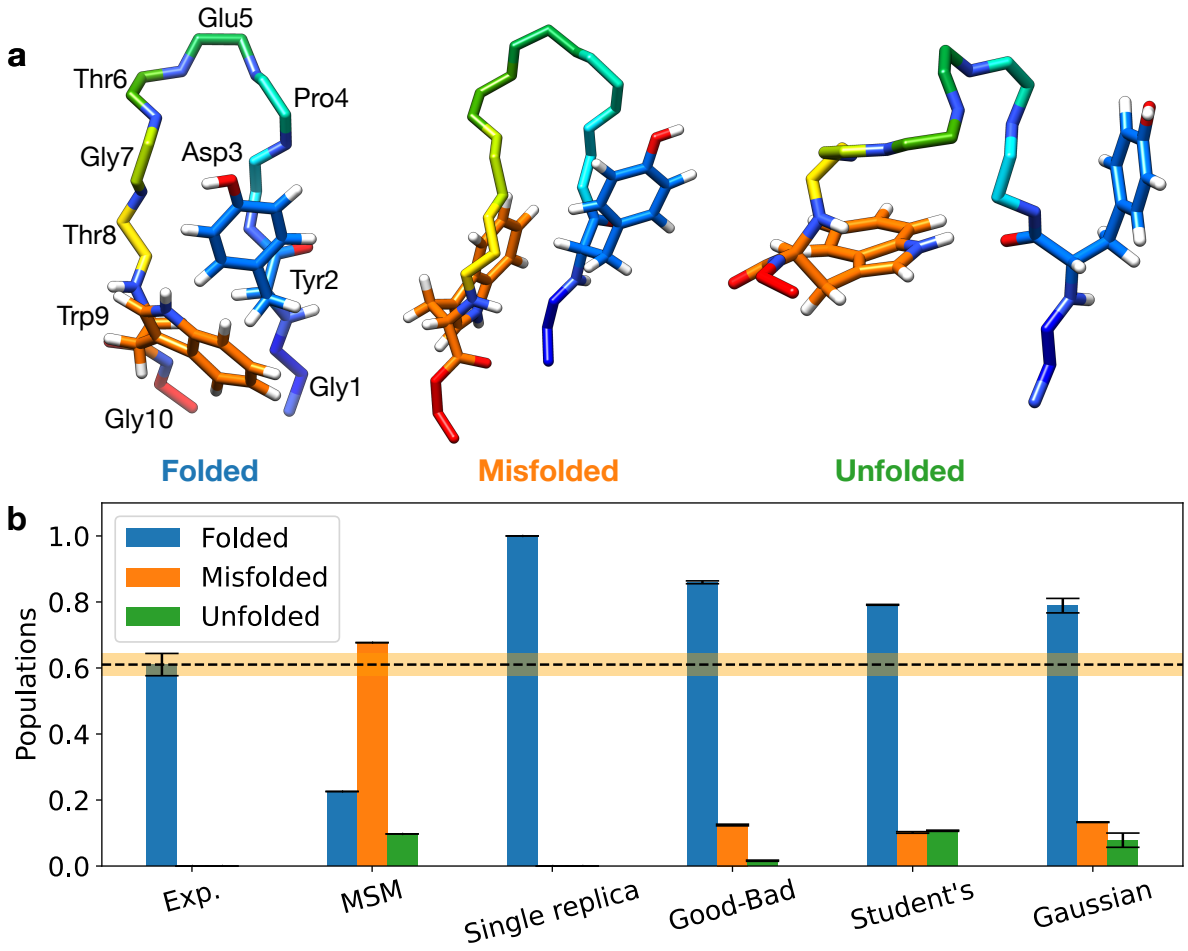

Figure 3: Reweighted macrostate populations of chignolin CLN001 for various data models. (a) Representative structures of macrostates derived from a 500-state MSM built from $\sim$ 30 μs simulations in A99SB-ildn. (b) Prior MSM macrostate populations and BICePs-reweighted populations for five different models are compared against experimental folded-state populations ($\sim$61% from Honda 2004). Three of the models use an effective uncertainty parameter for each type of observable: Single replica model, Good-Bad model (8 replicas), Student's model (8 replicas). The Gaussian model (8 replicas) uses an uncertainty parameter for each observable. Error bars correspond to the SEM over five independent rounds of sampling for 10M steps each.

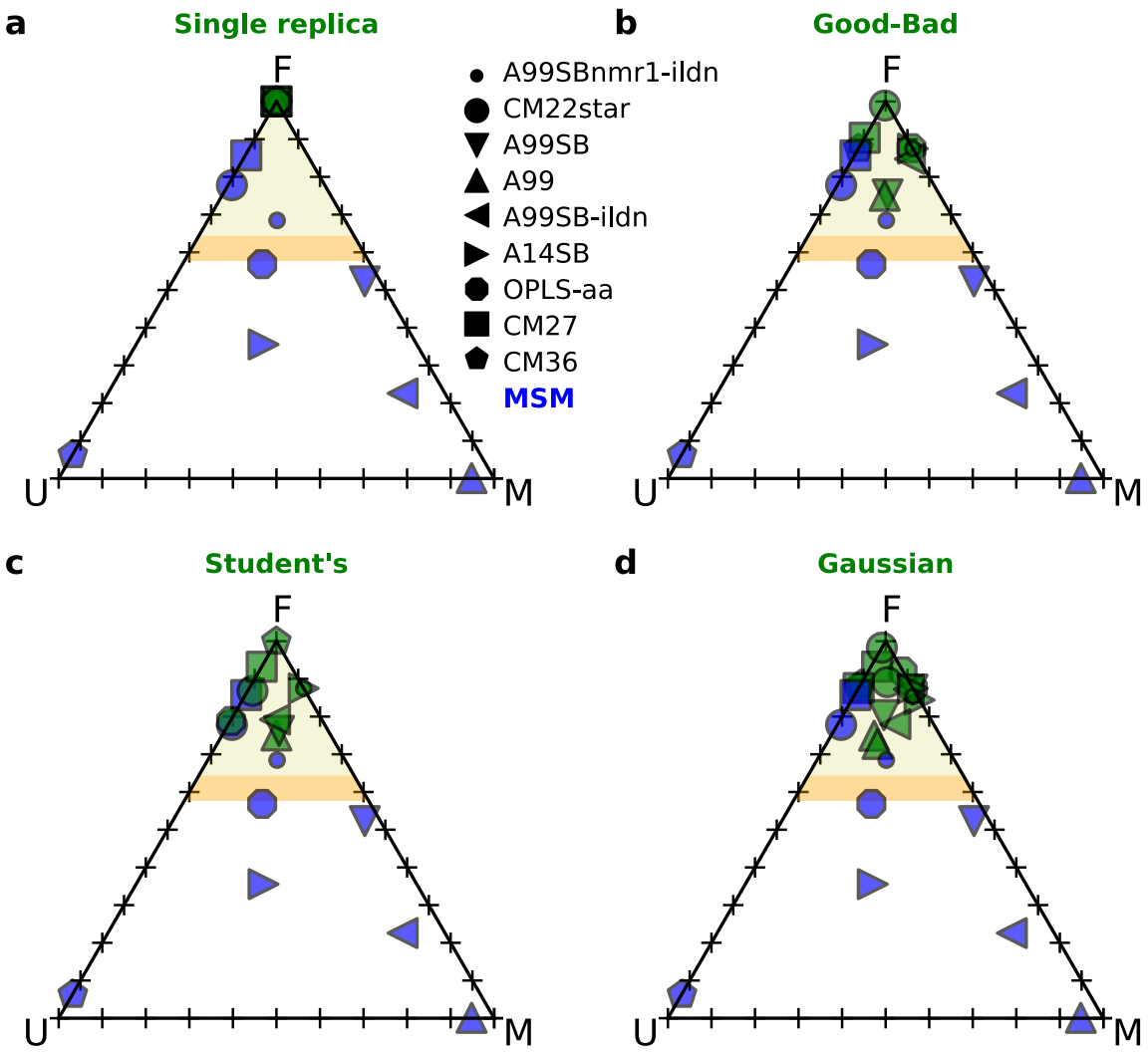

Figure 4: Triangle plots show that reweighted populations of folded (F), unfolded (U) and misfolded (M) macrostates (green markers) are shifted toward the native folded state compared to the prior populations (blue markers). All models enrich the folded state upon inclusion of experimental data, with the more sophisticated likelihood models (designed to deal with outliers) showing closer agreement to the experimental folded state population (highlighted in peach: 0.610 ± 0.034 at 300 K.)

Model selection using the BICePs score. In addition to sampling the maximum-entropy conformational ensemble of chignolin in the presence of experimental restraints, BICePs can objectively rank the quality of each force field model using the BICePs score. As described in Theory, we calculate the BICePs score by first "turning on" experimental restraints with uniform state populations ($\lambda = 0, \xi = 0 \rightarrow 1$), and then the prior $p^{(k)}(x)$ for each force field ($\lambda = 0 \rightarrow 1, \xi = 1$). For this chignolin system, the first transformation results in a large positive change in free energy $f^{(k)}_{\xi=0\rightarrow1}$ (a penalty), because of the large number of experimental observables. The magnitude of the second transformation is much smaller, and is typically a negative value $f^{(k)}_{\lambda=0\rightarrow1}$ (a reward), for force fields whose predicted distribution of conformational states overlaps well with the experimental restraints. The sum of the two free energies is the BICePs score; the lower the score, the more the simulated populations agree with the experimental data.

Since the BICePs score is calculated via stochastic sampling and free energy estimation, it is important that sampling is converged (Figures S19-22) and that intermediates used in the BICePs calculation have sufficient thermodynamic overlap (Figure S24). Another factor influencing the accuracy of the calculation is the resolution of the conformational state space. We tested how BICePs scores values vary using different numbers of MSM microstates (5, 10, 50, 75, 100, 500) and generally find that BICePs scores converge beyond 75 states (Figures S25). We also tested various numbers of replicas, and found that $N_r = 8$ replicas achieved a good balance of accuracy and computational efficiency (Figure S26).

Shown in Figure S34 is a comparison of BICePs scores for the nine different force fields, using 8 replicas and 500 conformational states (standard errors shown in Figure S27). To make these results more easily understood, we report relative BICePs scores, $f^{(k)} - \min(\{f^{(k)} \forall k\})$, so that the lowest score in each group is set to zero. Replica-averaging with Good-Bad and Student's likelihood models rank A99SBnmr1-ildn as having the best BICePs score when the entire set of experimental data is used (see Figures S28-30 for individual contributions of $f^{(k)}_{\xi=0\rightarrow1}$ and $f^{(k)}_{\lambda=0\rightarrow1}$, and their uncertainties).

This result agrees with work by Beauchamp et al., who used a reduced χ^2 metric to comprehensively rank force fields in comparison with NMR observables for short peptides and

ubiquitin, and found A99SBnmr1-ildn[37] and A99SB-ildn-phi[38] to be the most accurate when coupled with TIP3P aqueous solvent.[39] To further compare with Beauchamp et al., we computed reduced χ^2 values using the MSM populations (see SI Methods) for the six force fields tested in both studies (A99SBnmr1-ildn, A99SB-ildn, OPLS-aa, A99, A99SB, C27), and find a strong correlation (R^2=0.89) (see Figures S31-33 for χ^2 analysis).

Single-replica BICePs (which is not a true MaxEnt method) ranks A99SB slightly better than A99SBnmr1-ildn. We obtain a similar result from rankings computed by reduced χ^2 values (Figure S34). The MSM populations for A99SB heavily favor compact states (folded and misfolded, with less than 5% unfolded population) that are more compatible with the NOE distance restraints; single-replica BICePs rewards individual conformational states that are most compatible with the experimental restraints, rather than enforcing the ensemble average.

Using the Student's likelihood model, we calculated BICePs scores using only one type of experimental observable (Figure S34d,e,f). Results using only NOE distances are highly similar to the overall results, indicating that the 139 NOE distance restraints play an outsize role in the ranking due to their large number. Results using only the 6 J-coupling constant favor A99 (which incorrectly predicts a large misfolded population), while results using only the 13 chemical shift observables favor CM27. These results underscore the importance of using multiple observables to evaluate force fields amidst uncertainties in experimental measurements and forward models.

To further probe the source of variation in BICePs scores for different models, we computed $f^{(k)}_{\lambda=0\to1}$ for single-replica, Good-Bad, Student's, and Gaussian likelihood models, for various coarse-grained numbers of conformational states (Figure S35-35). We find similar patterns of variation across all likelihood models, suggesting that heterogeneity in conformational state definitions, and their computed forward model predictions, are the main source of variance. We also tested a single-prior Gaussian (GaussianSP) likelihood, in which one uncertainty parameter is used each group of observables, by setting $\phi = 1$ in the Good-Bad model; we find very similar results to the Good-Bad model(Figure S36).

Physical interpretation of free energies contributing to the BICePs score.

As mentioned in Theory, a key reason for choosing the $(\xi = 0, \lambda = 0)$ ensemble as the ref-

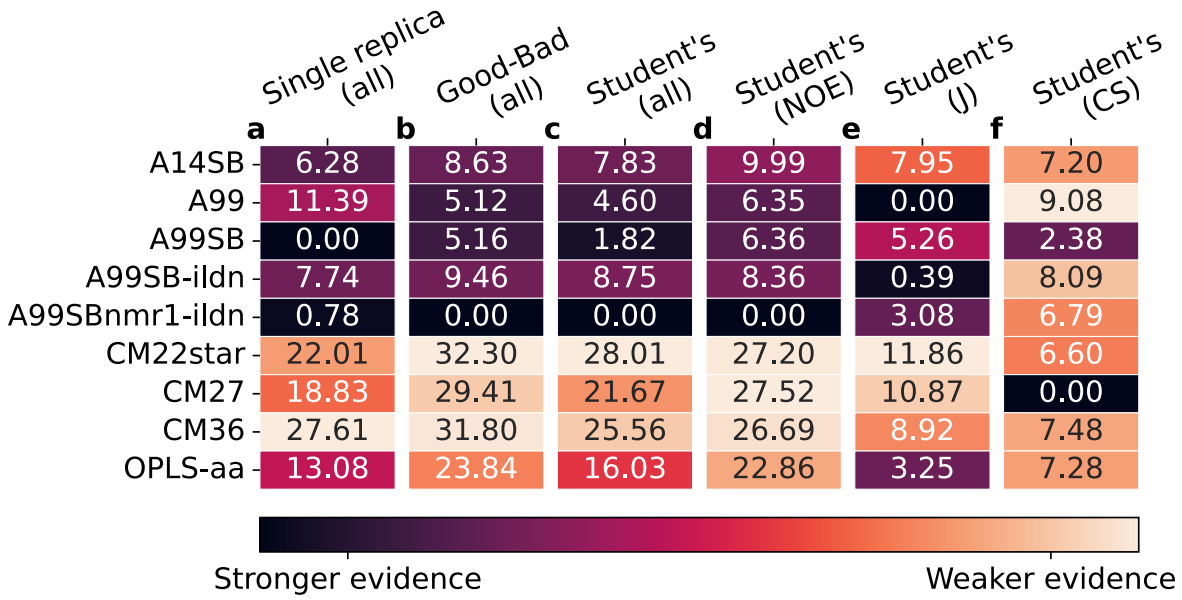

Figure 5: Ranking of force fields for chignolin CLN001 using the total BICePs score $f^{(k)}$ calculated using various likelihoods and different types of experimental observables. (500 microstates). Each column reports values relative to the minimum value.

erence state is the general difficulty of defining a unified set of conformational states that can be used as a uniform-population reference state to compare multiple force fields (see also SI Text). This is because MSM conformational states are most easily defined using unsupervised conformational clustering, and distributions of sampled conformations for different force fields tend to be uneven (Figure S19) and have poor overlap as the number of conformational states becomes large. In this case, however, we know that the conformational landscape of chignolin has three main macrostate populations: folded, misfolded, and unfolded. Therefore, we can construct a nearly-ideal reference state for each force field by reweighting the populations of microstates in each macrostate by a constant factor, to achieve a prior $p'(X)$ whose macrostate populations are uniform.

In this way, we can achieve a physically meaningful BICePs score $f'^{(k)}_{\lambda=0\to1}$ for just the $(\lambda = 0 \to 1)$ leg of the transformation, that can be computed as a correction to our previous result:

$$f'^{(k)}_{\lambda=0\to1} = -\ln \frac{Z^{(k)}}{Z'} = f^{(k)}_{\lambda=0\to1} + \ln \frac{Z'}{Z_0}. \tag{5}$$

The $f'^{(k)}_{\lambda=0\to1}$ scores can be interpreted as the change in free energy of "turning on" a given force field's prior $p^{(k)}(X)$ from a reference with uniform macrostates. The calculated results (Table

S1, Figures S37-39) show that the CM27 force field prior gets a bigger reward (by a slight margin) than for A99SBnmr1-ildn, suggesting CM27 has the best overlap with a distribution of conformational populations shaped solely by the experimental restraints. However, the overall BICePs score for CM27 is over twenty nats larger than that of A99SBnmr1-ildn, suggesting much more frustration when enforcing experimental restraints. This is reflected in the larger overall reduced χ^2 values for CM27 compared to A99SBnmr1-ildn (Figure S34).

Force field rankings using both $f^{(k)}_{\lambda=0\to1}$ and $f'^{(k)}_{\lambda=0\to1}$ also generally agree with previous findings. We find correlations of $R^2 \approx 0.8$ and $R^2 \approx 0.76$ respectively when comparing with reduced χ^2 values from Beauchamp et al. (Figure S40).

Caveats. Our work ranking force field models for chignolin should be considered as a proof of principle; proper assessments must encompass a diverse set of experimental systems. Another important caveat to be considered and improved in future work is the treatment of NOE observables. While here we treat the observables as (r^{-6}-averaged) distances, the physical measurement is the NOE intensity, which can be better assessed using improved forward models . This caveat is important because the large number of distance restraints heavily influences the outcome of BICePs. Another related issue we do not address is the statistical independence of measurements included in the set of experimental restraints. Cross-validation (testing and training with subsets of experimental restraints) may be useful for this purpose.

Conclusion

While several existing methods can perform post-processing MaxEnt reweighting against ensemble-averaged experimental constraints, BICePs is unique in (1) treating uncertainties as nuisance parameters that can be inferred from the data, and (2) providing an objective measure of model quality through the BICePs score. As our results show, BICePs can now perform sophisticated model selection for biomolecular simulations against large sets of experimental observables. This opens up new possibilities for automated force field validation against large experimental data sets, as well as machine learning of optimal parameters for force fields[40–43] and forward model refinement,[44,45] using the BICePs score as an objective function.

Methods

Accounting for sampling error in the replica-averaged forward model Replica-averaging introduces finite sampling error, which must be considered as additional uncertainty in the forward model. Following Bonomi and Camilloni et al.,[5,6] we estimate the standard error of the mean σ_j^{SEM} for an observable j by taking a windowed average over our finite sample $f_j(\mathbf{X})$:

$$\sigma_j^{\mathrm{SEM}} = \sqrt{\frac{1}{N_r}\sum_{r=1}^{N_r}(f_j(X_r) - \langle f_j(\mathbf{X})\rangle)^2}. \tag{6}$$

This quantity decreases as the square root of the number of replicas.

The Gaussian model. In this likelihood model (see (2)), errors between a forward model $f_j(\mathbf{X})$ and experimental measurement d_j are assumed to be normally distributed with unknown uncertainties $\sigma_j = \sqrt{(\sigma_j^{\mathrm{SEM}})^2 + (\sigma_j^{\mathrm{B}})^2}$, which considers both uncertainty in the forward model, and Bayesian uncertainty σ_j^{B} in the experimental measurements.

The single-replica model. The Gaussian model reduces to the *single-replica* model when $N_r = 1$. In this case, σ^{SEM} is ignored, and only the Bayesian uncertainty $\sigma_j = \sigma_j^{\mathrm{B}}$ is required. This is equivalent to the version of BICePs used in previous work.[18,31]

Likelihoods to account for systematic error and outliers. We present the *Good-Bad* likelihood model and the *Student's* likelihood model as approaches to marginalize the uncertainty parameters for individual observables, assuming that the level of noise is mostly uniform, except for a few erratic measurements. This limits the number of uncertainty parameters that need to be sampled, while still capturing outliers.

Consider a model where uncertainties σ_j for particular observables j are distributed about some typical uncertainty σ^{B} according to a conditional probability $p(\sigma_j|\sigma^{\mathrm{B}})$. We derive a posterior with a single uncertainty parameter σ^{B} by marginalizing over all σ_j. For a single replica

(for simplicity), the posterior is given by

$$p(X_r, \sigma_0|D) \propto p(X_r) \prod_{j=1}^{N_j} \int_{\sigma^{\text{SEM}}}^{\infty} p(d_j|\mathbf{X}, \sigma_j) p(\sigma_j|\sigma_0) d\sigma_j \tag{7}$$

where $\sigma_0 = \sqrt{(\sigma^{\text{B}})^2 + (\sigma^{\text{SEM}})^2}$.

The "Good-Bad" error model Under the Good-Bad model, we say that the "good" data consists of observables normally distributed about their true values with effective variance σ_0^2, while the "bad" data is subject to systematic error, leading to a larger effective variance $\phi^2\sigma_0^2$, where $\phi \geq 1$.[46,47] By this assignment, $p(\sigma_j|\sigma_0)$ from equation 7 becomes

$$p(\sigma_j|\sigma_0, \omega, \phi) = \omega\delta(\sigma_j - \phi\sigma_0) + (1-\omega)\delta(\sigma_j - \sigma_0) \tag{8}$$

where $0 \leq \omega < 1$ describes the fraction of "bad" observables. Since the value of ω is unknown, it is treated as a nuisance parameter, and marginalized over its range. The resulting posterior is

$$\begin{aligned} p(X_r, \sigma_0, \phi|D) &\propto p(X_r) \prod_{j=1}^{N_j} \int_0^1 d\omega \int_{\sigma^{\text{SEM}}}^{\infty} \exp\left(-\frac{(d_j - f_j(\mathbf{X}))^2}{2\sigma_j^2}\right) \\ &\quad \times \frac{\omega\delta(\sigma_j - \phi\sigma_0) + (1-\omega)\delta(\sigma_j - \sigma_0)}{\sqrt{2\pi}\sigma_j} d\sigma_j \\ &= p(X_r) \prod_{j=1}^{N_j} \left[\frac{\left(1 - H\left(\sigma^{\text{SEM}} - \sigma_0\right)\right)}{2\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(d_j - f_j(\mathbf{X}))^2}{2\sigma_0^2}\right) \right. \\ &\quad \left. + \frac{\left(1 - H\left(\sigma^{\text{SEM}} - \phi\sigma_0\right)\right)}{2\phi\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(d_j - f_j(\mathbf{X}))^2}{2\phi^2\sigma_0^2}\right) \right], \end{aligned} \tag{9}$$

where H is the Heaviside step function. After marginalization, we are left with the Bayesian uncertainty parameter σ_0^B, and an additional parameter ϕ. Both parameters are sampled in the posterior. When $\phi = 1$, the model reverts to the Gaussian model. When considering the full posterior, this extra nuisance parameter is given a non-informative Jeffreys prior, $p(\phi) \sim \phi^{-1}$.

The Student's model is an intermediate between Cauchy and Gaussian distributions Modeling $p(\sigma_j|\sigma_0)$ as a Cauchy distribution is very useful because its long tail makes it able to tolerate outliers.[5,48] In most cases, however, it is unclear *a priori* what distribution is best for modeling the input data. To improve the situation, we introduce a model with an additional nuisance parameter β, that is able to tune the extent of the distribution's tail:

$$p(\sigma_j|\sigma_0,\beta) = \frac{\Gamma((\beta+1)/2)}{\Gamma(\beta/2)} \frac{2\beta^\beta {\sigma_0}^{2\beta-1}}{\sqrt{\beta}{\sigma_j}^{2\beta}} \exp\left(-\frac{\beta{\sigma_0}^2}{\sigma_j^2}\right). \tag{10}$$

where σ_0 is defined as above, and $1 \leq \beta < \infty$. When this distribution is inserted into the posterior, and marginalized over all σ_j, the result is

$$\begin{aligned} p(X_r,\sigma_0,\beta|D) \propto p(X_r) \prod_{j=1}^{N_j} \frac{\Gamma((\beta+1)/2)}{\Gamma(\beta/2)} \frac{1}{\sqrt{2\pi\beta}\sigma_0} \\ \times \left[1+\frac{(d_j-f_j(X))^2}{2\beta\sigma_0^2}\right]^{-\beta} \gamma\left(\beta, \frac{(d_j-f_j(X))^2+2\beta\sigma_0^2}{2\left(\sigma^{\mathrm{SEM}}\right)^2}\right). \end{aligned} \tag{11}$$

Here, the marginal likelihood contains the lower incomplete gamma function, γ. We call this the Student's model because it is a variation of Student's t-distribution that can be interpolated between functional forms. When $\beta = 1$, the model is equivalent Metainference's Outliers model[5] . In the limit of $\beta \to \infty$, the likelihood becomes Gaussian. When considering the full posterior, this extra nuisance parameter is given a non-informative Jeffreys prior, $p(\beta) \sim \beta^{-1}$. For more detials regarding likelihood moels, see SI methods.

Molecular dynamics simulations of chignolin. Folding simulations of chignolin CLN001 (GYDPETGTWG) were performed using GROMACS 2020.4[49] on the Folding@home distributed computing platform,[50] using nine different force fields: Amber14SB (A14SB),[51] Amber99SB-ildn (A99SB-ildn),[52] Amber99 (A99),[53] Amber99SBnmr1-ildn (A99SBnmr1-ildn),[37] Amber99SB (A99SB),[51] CHARMM22* (C22star),[54] CHARMM27 (C27),[55] CHARMM36 (C36),[56] and OPLS-AA.[57] Chignolin was solvated in a $\sim$ 5-nm cubic periodic box with 4000 TIP3P waters and neutralizing NaCl at 100 nM. After minimization and equilibration, NPT production runs at 300 K were initiated from folded and unfolded conformations, producing 34 trajectories

with lengths between 0.2 and 1.6 μs for each force field, and a total aggregate of 200 μs.

Markov State Models. For each set of trajectories performed using a given force field, pairwise atomic inverse distances for selected atoms (all Cα, Tyr2Cζ, Trp9Cδ1, Asp3N, Thr8O and Gly7O) were used as input for dimensionality reduction using TICA.[58,59] Clustering using k-means was performed after projection to eight principal TICA components. Using these conformational state definitions, Markov State Models (MSMs) with lag time 20 ns were constructed using a maximum-likelihood estimator, with a bootstrapping procedure (randomly selecting 50% of the trajectories as input data over 5 trials) to estimate equilibrium state populations $p(X)$ and their uncertainties. Full details are described in Marshall et al.[32]

Six models of varying levels of coase graining of conformational space were constructed for each force field. Markov state models (MSM) were built , where ensemble averaged forward model data was averaged over 20 snapshots for each microstate.

Comparison to experimental observables. Forward-model observables for each MSM microstate were computed by averaging the forward model predictions of twenty randomly selected trajectory snapshots; these values were used as input to BICePs. The simulations were compared against 158 experimental measurements[33] (139 NOE distances, 13 chemical shifts, and 6 vicinal J-coupling constants for H^N and H^α. Forward model predictions for chemical shifts were computed using SHIFTX2 version 1.11,[60] forward model NOE distances were computed using MDTraj,[61] and forward model J-coupling constants were computed using the Karplus relation of Vögeli et al.,[62] a built-in function from BICePs v2.0.[31] For BICePs calculations using the Gaussian likelihood, all 158 experimental restraints were used. For all other BICePs calculations, indistinguishable protons were treated as a single restraint, resulting in a total of 138 experimental restraints.

Data and software The BICePs algorithm is freely available at github.com/vvoelz/biceps and can be installed using `pip install biceps`. All chignolin CLN001 data, input files, Jupyter notebook examples and BICePs analysis can be found here: github.com/robraddi/chignolin. Detailed documentation and tutorials can be found here: https://biceps. readthedocs.io. For any

issues or questions, please submit the request on GitHub.

Acknowledgments

RMR, TM and VAV are supported by National Institutes of Health grant R01GM123296. We thank the participants of Folding@home; this work would not be possible without their contributions. We appreciate George Doenlen for the insightful C++ programming conversations.

References

(1) Rieping, W.; Habeck, M.; Nilges, M. Inferential structure determination. *Science* **2005**, *309*, 303–306.

(2) Perez, A.; MacCallum, J. L.; Dill, K. A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proceedings of the National Academy of Sciences* **2015**, *112*, 11846–11851.

(3) Beauchamp, K. A.; Pande, V. S.; Das, R. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophysical journal* **2014**, *106*, 1381–1390.

(4) Fisher, C. K.; Huang, A.; Stultz, C. M. Modeling intrinsically disordered proteins with bayesian statistics. *Journal of the American Chemical Society* **2010**, *132*, 14919–14927.

(5) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Science Advances* **2016**, *2*, e1501177.

(6) Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic metainference: enhanced sampling of the metainference ensemble using metadynamics. *Scientific reports* **2016**, *6*, 31232.

(7) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Current opinion in structural biology* **2017**, *42*, 106–116.

(8) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining experiments and simulations using the maximum entropy principle. *PLoS computational biology* **2014**, *10*, e1003406.

(9) Orioli, S.; Larsen, A. H.; Bottaro, S.; Lindorff-Larsen, K. *Progress in Molecular Biology and Translational Science*; Elsevier, 2020; Vol. 170; pp 123–176.

(10) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *Journal of chemical theory and computation* **2012**, *8*, 3445–3451.

(11) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of chemical physics* **2013**, *138*, 03B603.

(12) Cesari, A.; Reißer, S.; Bussi, G. Using the maximum entropy principle to combine simulations and solution experiments. *Computation* **2018**, *6*, 15.

(13) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of chemical physics* **2013**, *138*, 02B616.

(14) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of chemical physics* **2015**, *143*, 12B634_1.

(15) Paissoni, C.; Camilloni, C. How to determine accurate conformational ensembles by metadynamics metainference: a chignolin study case. *Frontiers in molecular biosciences* **2021**, *8*, 694130.

(16) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. *Structural Bioinformatics*; Springer, 2020; pp 219–240.

(17) Xu, H. Molecular simulations minimally restrained by experimental data. *The Journal of chemical physics* **2019**, *150*, 154121.

(18) Voelz, V. A.; Ge, Y.; Raddi, R. M. Reconciling simulations and experiments with BICePs: a review. *Frontiers in Molecular Biosciences* **2021**, *8*, 325.

(19) Voelz, V. A.; Zhou, G. Bayesian inference of conformational state populations from computational models and sparse experimental observables. *Journal of Computational Chemistry* **2014**, *35*, 2215–2224.

(20) Olsson, S.; Frellsen, J.; Boomsma, W.; Mardia, K. V.; Hamelryck, T. Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. *PLoS ONE* **2013**, *8*, e79439–7.

(21) Potrzebowski, W.; Trewhella, J.; Andre, I. Bayesian inference of protein conformational ensembles from limited structural data. *PLoS computational biology* **2018**, *14*, e1006641.

(22) Crehuet, R.; Buigues, P. J.; Salvatella, X.; Lindorff-Larsen, K. Bayesian-maximum-entropy reweighting of IDP ensembles based on NMR chemical shifts. *Entropy* **2019**, *21*, 898.

(23) Köfinger, J.; Stelzl, L. S.; Reuter, K.; Allande, C.; Reichel, K.; Hummer, G. Efficient ensemble refinement by reweighting. *Journal of chemical theory and computation* **2019**, *15*, 3390–3401.

(24) Różycki, B.; Kim, Y. C.; Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011**, *19*, 109–116.

(25) Köfinger, J.; Różycki, B.; Hummer, G. Inferring structural ensembles of flexible and dynamic macromolecules using Bayesian, maximum entropy, and minimal-ensemble refinement methods. *Biomolecular Simulations: Methods and Protocols* **2019**, 341–352.

(26) Pesce, F.; Lindorff-Larsen, K. Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. *Biophysical journal* **2021**, *120*, 5124–5135.

(27) Wakefield, A. E.; Wuest, W. M.; Voelz, V. A. Molecular simulation of conformational pre-organization in cyclic RGD peptides. *Journal of chemical information and modeling* **2015**, *55*, 806–813.

(28) Liang, H.; Zhou, G.; Ge, Y.; D'Ambrosio, E. A.; Eidem, T. M.; Blanchard, C.; Shehatou, C.; Chatare, V. K.; Dunman, P. M.; Valentine, A. M., et al. Elucidating the inhibition of peptidoglycan biosynthesis in Staphylococcus aureus by albocycline, a macrolactone isolated from Streptomyces maizeus. *Bioorganic & medicinal chemistry* **2018**, *26*, 3453–3460.

(29) Ge, Y.; Voelz, V. A. Model Selection Using BICePs: A Bayesian Approach for Force Field Validation and Parameterization. *The Journal of Physical Chemistry B* **2018**, *122*, 5610–5622.

(30) Hurley, M. F. D.; Northrup, J. D.; Ge, Y.; Schafmeister, C. E.; Voelz, V. A. Metal Cation-Binding Mechanisms of Q-Proline Peptoid Macrocycles in Solution. **2021**,

(31) Raddi, R.; Ge, Y.; Voelz, V. BICePs v2. 0: Software for Ensemble Reweighting using Bayesian Inference of Conformational Populations. **2022**,

(32) Tim Marshall, V. V., R Raddi An Evaluation of Force Field Accuracy for the Mini-Protein Chignolin using Markov State Models. *Submitted* **2023**,

(33) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **2004**, *12*, 1507–1518.

(34) Maruyama, Y.; Takano, H.; Mitsutake, A. Analysis of molecular dynamics simulations of 10-residue peptide, chignolin, using statistical mechanics: Relaxation mode analysis and three-dimensional reference interaction site model theory. *Biophysics and physicobiology* **2019**, *16*, 407–429.

(35) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS letters* **2006**, *580*, 3422–3426.

(36) Harada, R.; Kitao, A. Exploring the folding free energy landscape of a β-hairpin miniprotein, chignolin, using multiscale free energy landscape calculation method. *The Journal of Physical Chemistry B* **2011**, *115*, 8806–8812.

(37) Li, D.-W.; Brüschweiler, R. NMR-based protein potentials. *Angewandte Chemie International Edition* **2010**, *49*, 6778–6780.

(38) Nerenberg, P. S.; Head-Gordon, T. Optimizing protein- solvent force fields to reproduce intrinsic conformational preferences of model peptides. *Journal of Chemical Theory and Computation* **2011**, *7*, 1220–1230.

(39) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of chemical theory and computation* **2012**, *8*, 1409–1414.

(40) Köfinger, J.; Hummer, G. Empirical optimization of molecular simulation force fields by Bayesian inference. *The European Physical Journal B* **2021**, *94*, 245.

(41) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building force fields: An automatic, systematic, and reproducible approach. *The journal of physical chemistry letters* **2014**, *5*, 1885–1891.

(42) Madin, O. C.; Boothroyd, S.; Messerly, R. A.; Fass, J.; Chodera, J. D.; Shirts, M. R. Bayesian-inference-driven model parametrization and model selection for 2CLJQ fluid models. *Journal of Chemical Information and Modeling* **2022**, *62*, 874–889.

(43) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining simulations and solution experiments as a paradigm for RNA force field refinement. *Journal of chemical theory and computation* **2016**, *12*, 6192–6200.

(44) Habeck, M.; Rieping, W.; Nilges, M. Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar couplings constants. *Journal of magnetic resonance* **2005**, *177*, 160–165.

(45) Fröhlking, T.; Bernetti, M.; Bussi, G. Simultaneous refinement of molecular dynamics ensembles and forward models using experimental data. *The Journal of Chemical Physics* **2023**, *158*.

(46) Sivia, D. Dealing with duff data. *Maximum Entropy and Bayesian Methods* **1996**, 131–137.

(47) Box, G. E.; Tiao, G. C. A Bayesian approach to some outlier problems. *Biometrika* **1968**, *55*, 119–129.

(48) Sivia, D.; Skilling, J. *Data analysis: a Bayesian tutorial*; OUP Oxford, 2006.

(49) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(50) Voelz, V. A.; Pande, V. S.; Bowman, G. R. Folding@ home: achievements from over twenty years of citizen science herald the exascale era. *Biophysical Journal* **2023**,

(51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

(52) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 1950–1958.

(53) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry* **2000**, *21*, 1049–1074.

(54) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophysical journal* **2011**, *100*, L47–L49.

(55) MacKerell Jr, A. D.; Feig, M.; Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society* **2004**, *126*, 698–699.

(56) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

(57) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B* **2001**, *105*, 6474–6487.

(58) Schwantes, C. R.; Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of chemical theory and computation* **2013**, *9*, 2000–2009.

(59) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013**, *139*.

(60) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR* **2011**, *50*, 43–57.

(61) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.

(62) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. Limits on Variations in Protein Backbone Dynamics from Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.* **2007**, *129*, 9377–9385, PMID: 17608477.

(63) Hansen, P. C.; O'Leary, D. P. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM journal on scientific computing* **1993**, *14*, 1487–1503.

(64) Razavi, A. M.; Wuest, W. M.; Voelz, V. A. Computational screening and selection of cyclic peptide hairpin mimetics by molecular simulation and kinetic network models. *Journal of chemical information and modeling* **2014**, *54*, 1425–1432.

(65) Voelz, V. A.; Elman, B.; Razavi, A. M.; Zhou, G. Surprisal metrics for quantifying perturbed conformational dynamics in Markov state models. *Journal of chemical theory and computation* **2014**, *10*, 5716–5728.

(66) Razavi, A. M.; Voelz, V. A. Kinetic network models of tryptophan mutations in β-hairpins reveal the importance of non-native interactions. *Journal of Chemical Theory and Computation* **2015**, *11*, 2801–2812.

(67) Wan, H.; Zhou, G.; Voelz, V. A. A maximum-caliber approach to predicting perturbed folding kinetics due to mutations. *Journal of chemical theory and computation* **2016**, *12*, 5768–5776.