Published in final edited form as:

J Phys Chem B. 2018 May 31; 122(21): 5610–5622. doi:10.1021/acs.jpcb.7b11871.

Model Selection Using BICePs: A Bayesian Approach for Force Field Validation and Parameterization

Yunhui Ge and Vincent A. Voelz*

Department of Chemistry, Temple University, Philadelphia, Pennsylvania, USA

Abstract

The Bayesian Inference of Conformational Populations (BICePs) algorithm reconciles theoretical predictions of conformational state populations with sparse and/or noisy experimental measurements. Among its key advantages is the ability to perform objective model selection through a quantity we call the BICePs score, which reflects the integrated posterior evidence in favor of a given model, computed through free energy estimation methods. Here, we explore how the BICePs score can be used for force field validation and parameterization. Using a 2D lattice protein as a toy model, we demonstrate that BICePs is able to select the correct value of an interaction energy parameter given ensemble-averaged experimental distance measurements. We show that if conformational states are sufficiently fine-grained, the results are robust to experimental noise and measurement sparsity. Using these insights, we apply BICePs to perform force field evaluations for all-atom simulations of designed beta-hairpin peptides against experimental NMR chemical shift measurements. These tests suggest that BICePs scores can be used for model selection in the context of all-atom simulations. We expect this approach to be particularly useful for computational foldamer design, as a tool for improving general-purpose force fields given sparse experimental measurements.

Graphical Abstract

*To whom correspondence should be addressed: voelz@temple.edu, Phone: +1 (215) 204 1973., Fax: +1 (215) 204 1532.

Supporting Information Available

The following files are available free of charge.

• Supporting Figures S1-S3

This material is available free of charge via the Internet at <http://pubs.acs.org/>.

Introduction

An overarching goal of computational chemistry and biophysics is to use theoretical predictions alongside experimental measurements to determine molecular structure and function. This idea underlies most of the biophysical methods for protein structure determination, as well as NMR structure determination of non-biological molecules. Recently, there has been a growing need for new methods that combine theory and experiment to characterize heterogeneous structural ensembles.¹ For proteins, this interest has been driven in part by the importance of intrinsically disordered proteins² and conformational ensembles in allostery and signaling.^{3,4} For non-biological molecules, this need has been driven by the interest in characterizing large, conformationally heterogeneous molecules such as natural products, macrocycles, and bio-inspired foldamers. The main challenge for heterogeneous conformational ensembles is that experimental observables (e.g. NMR, FRET, SAXS) are ensemble-average measurements, and the data is often sparse and noisy.

An area of research that would benefit greatly from these developments is computational foldamer design. We and others have been pursuing all-atom molecular simulation approaches to designing peptide and peptidomimetic scaffolds that are sufficiently preorganized in solution.^{5–11} Such scaffolds may be developed into potent therapeutics that disrupt protein-protein interactions,^{12,13} or enable the design of new catalysts and bio-inspired materials.^{14–17} While these applications are promising, using simulation methods to predict conformational ensembles of foldamers is challenging due to the lack of general-purpose force fields that are accurate enough for computational design. A practical solution to this problem would be to perform simulations with a reasonably accurate potential that could be augmented with information from sparse experimental measurements.

Toward this end, we have developed BICePs (Bayesian Inference of Conformational Populations), a statistically rigorous Bayesian inference method to reconcile theoretical predictions of conformational state populations with sparse and/or noisy experimental measurements. A key advantage of BICePs is that it can be used to reweight a set of conformational populations without having to perform additional simulations or modeling. This approach favors the use of high-resolution modeling efforts, which can then be refined using sparse experimental data. To date, we have exploited these features of BICePs to model a number of molecular systems with experimental data from NMR (e.g. chemical shifts, coupling constants, NOE distance restraints), including macrolide antibiotics,¹⁸ cyclic peptides,¹⁹ peptoid pentamers.²⁰

Another advantage of BICePs—which we explore more in this manuscript—is that it provides a rigorous statistical framework for model selection that can be used to quantitatively estimate how a consistent a simulated ensemble is with the experimental data. As we elaborate below, there is an unequivocal measure of model quality, herein defined as the “BICePs score”, that can potentially be used to perform force field validation and parameterization.

In this manuscript, we first review the theory underlying the BICePs algorithm and describe its key advantages. We define the BICePs score, show how it can be used for model

selection, and discuss the practical considerations of computing BICePs scores. Then, using a 2D lattice protein as a toy model, we explore the prospects of using BICePs for force field parameterization and model selection. Specifically, we explore the robustness of these model selection results with respect to the coarse-graining of conformational states, experimental noise and measurement sparsity. Using these insights, we apply BICePs to perform force field evaluations for all-atom simulations of designed beta-hairpin peptides against experimental NMR chemical shift measurements.

Theory

Bayesian inference.

Bayesian inference approaches for modeling conformational ensembles generally seek to model a *posterior* distribution $P(X|D)$ of conformational states X , given some experimental data D . According to Bayes' theorem, the posterior distribution is proportional to a product of (1) a *likelihood* function $Q(D|X)$ of observing the experimental data given a conformational state X , and (2) a *prior* distribution $P(X)$ representing any prior knowledge about conformational states.

$$P(X | D) \propto Q(D | X)P(X) \quad (1)$$

Here, the prior $P(X)$ comes from theoretical modeling, while the likelihood $Q(D|X)$ corresponds to experimental restraints, typically in the form of some error model reflecting how well a given conformation X agrees with experimental measurements. One can think of $Q(D|X)$ as a reweighting factor for the population of each state X , and the BICePs algorithm can be thought of as a way to reweight conformational populations to agree with experimental knowledge.

In BICePs, the error model reflects both uncertainty in the experimental measurements and heterogeneity in the conformational ensemble, (i.e. how likely are the observables for conformation X to be away from the ensemble-average measurement, either due to experimental noise or conformational heterogeneity). These uncertainties are usually not known *a priori*, and must be treated as nuisance parameters σ which can be modeled using some prior model $P(\sigma)$:

$$P(X, \sigma | D) \propto Q(D | X, \sigma)P(X)P(\sigma)$$

Posterior sampling over X and σ using Markov Chain Monte Carlo (MCMC) can then be used to determine the conformational populations given the experimental restraints as $P(X|D) = \int P(X, \sigma|D)d\sigma$. Similarly, the posterior distribution of the experimental uncertainty $P(\sigma|D) = \int P(X, \sigma|D)dX$ gives information about how well the posterior conformational distribution agrees with the experimental restraints.

We note that BICePs is not unique in utilizing this Bayesian framework for inference. Indeed, many similar Bayesian approaches have been developed, including Inferential Structure Determination (ISD),²¹ MELD,²² and BELT,²³ Bayesian Weighting²⁴ and

metainference.^{25,26} In addition, a number of non-Bayesian algorithms, such as NAMFIS²⁷ and DISCON^{28,29} can be used to estimate conformational populations from NMR observables. In their excellent review,¹ Bonomi et al. make the distinction between two broad categories of approaches for Bayesian structural inference: those that use a likelihood function that relates a single structure to the experimental data, and those that use a likelihood functional that relates an ensemble of structures to the experimental data (as in maximum entropy methods). As we will see below, BICePs is in the former category, with the reweighting factor $Q(D|X)$ specifying how well each conformation X agrees with ensemble-average experimental observables. BICePs is therefore more suited for use with more structured ensembles, as opposed to highly disordered and heterogeneous ensembles for which maximum entropy methods are more applicable.

With several existing Bayesian inference approaches already available, what does BICePs have to offer? BICePs uniquely improves upon existing approaches in two key ways: (1) correct implementation of reference potentials, and (2) quantitative metrics for model selection.

Reference potentials.

The correct implementation of reference potentials is an importance advantage of the BICePs algorithm. The experimental data used to construct the likelihood function comes from some set of ensemble-averaged experimental observables $\mathbf{r} = (r_1, r_2, \dots, r_N)$. Such observables are low-dimensional projections of some high-dimensional state space X , and therefore these restraints in the space of observables need to be treated as *potentials of mean force*.^{30–32}

$$P(X | D) \propto \left[\frac{Q(\mathbf{r}(X) | D)}{Q_{\text{ref}}(\mathbf{r}(X))} \right] P(X)$$

The bracketed weighting function is now a ratio, with the numerator $Q(\mathbf{r}|D)$ being a likelihood function enforcing experimental restraints in the space of observables, while the denominator $Q_{\text{ref}}(\mathbf{r})$ reflects some reference distribution for possible values of the observables $\mathbf{r}$ in the absence of any experimental restraint information, such that $-\ln[Q(\mathbf{r}|D)/Q_{\text{ref}}(\mathbf{r})]$ is a potential of mean force. Without reference potentials, a great deal of unnecessary bias is introduced when many non-informative restraints are used.

As an example to illustrate the importance of reference potentials, consider an experimental distance restraint applied to two residues of a polypeptide chain (Figure 1). In the absence of any experimental information, we assume a reference potential $Q_{\text{ref}}(\mathbf{r})$ corresponding to the end-to-end distance of a random-coil polymer with a chain length equal to that of the intervening residues. For residues near each other along the chain, a short-distance restraint may have $[Q(\mathbf{r}|D)/Q_{\text{ref}}(\mathbf{r})] \sim 1$, contributing little or no information to refine the conformational ensemble. If the residues are far apart along the chain, however, a short-distance restraint can be highly informative, with $[Q(\mathbf{r}|D)/Q_{\text{ref}}(\mathbf{r})]$ greatly rewarding small distances where the reference potential $Q_{\text{ref}}(\mathbf{r})$ is small.

In our previous work on BICePs, we showed that the use of reference potentials yields models superior to those without reference potentials; without reference potentials, a great deal of the unnecessary bias is introduced when many simultaneous distance restraints are used.¹⁸ While there may not be a way to determine the optimal reference potential in advance, there is some freedom to choose an appropriate reference potential in a problem-dependent way. For example, for cyclic molecules, we have used a maximum-entropy distribution parameterized from the average distances seen across the conformational state space. For a linear chain, a Gaussian reference potential may be more appropriate.

BICePs scores for quantitative model selection.

Another key advantage of the BICePs algorithm is its ability to use the sampled posterior probability function to perform quantitative model selection. This is done through a quantity we will call the *BICePs score*.

Consider K different theoretical models $P^{(k)}(X)$, $k = 1, \dots, K$, that we may wish to use as our prior distribution. Perhaps these models come from simulations using different potential energy functions. Or perhaps we want to compare a model shaped only experimental restraints, to a model using both simulation and experimental information. Which model is more consistent with the experimental data? To determine this, we compute the posterior likelihood of the model, $Z^{(k)}$, by integrating the k^{th} posterior distribution across all conformations X and values of σ .

$$Z^{(k)} = \int P^{(k)}(X, \sigma | D) dX d\sigma = \int P^{(k)}(X) Q(X) dX \quad (2)$$

$$\text{total evidence for model } P^{(k)} \text{ overlap integral} \quad (3)$$

The quantity $Z^{(k)}$ can be interpreted as the total evidence in favor of a given model. Equivalently, we can think of this term as an overlap integral between the prior $P^{(k)}(X)$ and a likelihood function $Q(X) = \int [Q(\mathbf{r}(X) | D, \sigma) / Q_{\text{ref}}(\mathbf{r}(X))] P(\sigma) d\sigma$. The overlap integral quantifies how well the theoretical modeling agrees with the experimental data (Figure 2); i.e. the value of $Z^{(k)}$ is maximal when $P^{(k)}(X)$ most closely matches the likelihood distribution $Q(X)$ specified by the experimental restraints.

In Bayesian statistics, the so-called Bayes factor, $Z^{(1)}/Z^{(2)}$ is a likelihood ratio that can be used to choose between competing models (1) and (2), similar to a likelihood-ratio test in classical statistics. In order to assign each model a unique score, we define a free energy-like quantity,

$$f^{(k)} = -\ln \frac{Z^{(k)}}{Z_0} \quad (4)$$

to compare model $P^{(k)}(X, \sigma | D)$ against some reference model Z_0 . In practice, we choose the reference model to be the posterior for a uniform prior $P(X)$, i.e. no information from theoretical modeling. We call $f^{(k)}$ the BICePs score. The BICePs score provides an unequivocal measure of model quality, to be used for objective model selection. The lower the BICePs score, the better the model (see Figure 4). This useful property means that the BICePs score can be used a metric for force field validation and parameterization. In addition, the BICePs score has a useful physical interpretation: it reflects the improvement (or disimprovement) of the posterior distribution when going from a conformation ensemble shaped only experimental constraints, to a new distribution additionally shaped by a theoretical model.

Practical aspects of BICePs.

While Bayesian posterior distributions are easy to write down, their practical calculation requires numerical sampling. The BICePs algorithm does this through MCMC sampling via the Metropolis-Hastings algorithm, over a discrete set of conformational states X . In this way, conformational states and their prior distribution $P^{(k)}(X)$ can come from high-resolution modeling efforts that do not have to be repeated in the BICePs calculation. Examples of this include: (1) a series of conformational minima and their populations estimated from *ab initio* quantum mechanical calculations, or (2) conformational clusters derived from molecular simulation trajectory data, with equilibrium populations calculated from free energy estimates. In fact, as long as the conformational state space is well-defined (and fine-grained enough to be accurate, more on this below), disparate kinds of models can be objectively compared.

Posterior sampling over X and nuisance parameters σ using MCMC requires special care, due to the potentially rugged prior and likelihood functions. We have found good results using a free energy perturbation (FEP) method: posterior sampling a series of models with priors $P^{(k)}(X) \sim \exp(-\lambda_k F(X))$ (where $0 \leq \lambda_k \leq 1$, and $F(X)$ is the free energy of conformational state X estimated from theoretical models). Enhanced sampling is achieved by MCMC sampling over a number of intermediate ensembles $P_\lambda^{(k)}(X) \sim [P^{(k)}(X)]^\lambda / \int [P^{(k)}(X)]^\lambda dX$ (Figure 2), and then using the MBAR free energy estimator of Shirts and Chodera³³ to integrate samples from each λ -ensemble to make statistically optimal estimates of $P_\lambda^{(k)}(X | D)$ and $Z_\lambda^{(k)}$ at all λ values. The final BICePs score is then calculated as $f^{(k)} = -\ln Z_{\lambda=1}^{(k)} / Z_{\lambda=0}^{(k)}$.

Summary of the key advantages of BICePs.

To put the BICePs algorithm in a larger context, we summarize the its key advantages as follows:

- BICePs can be used with molecular dynamics (MD) or quantum mechanics (QM) methods. Conformational states can be individual conformations (like single-point QM minima) or collections of conformations (e.g. from clustering of trajectory data), as long as experimental observables can be attached to each conformational state.

- BICePs performs reweighting of conformational states derived from modeling; it is currently a post-processing algorithm (MCMC) with no additional MD or QM required.
- Bayesian inference offers a rigorous statistical framework for achieving the correct balance of theoretical modeling and experimental data.
- BICePs correctly uses reference potentials, which is essential to proper weighing of experimental restraints
- With proper reference potentials, BICePs scores can be used for unambiguous, objective model selection.

Can BICePs be used for force field parameterization?

This is the central question we raise in our manuscript. While BICePs scores in theory provide an unambiguous way to select the best models, there are several practical issues that need to be addressed in order to make this approach successful. We would like to understand how several sources of potential error affect the BICePs results, so we can develop an optimal BICePs protocol for force field evaluation and parameterization. For one, the use of a set of discrete conformational states introduces some coarse-graining error. Experimental noise and sparsity affects the accuracy of the BICePs predictions, as does statistical sampling error.

To address these issues, we explore a simple, exact model with which we can compare BICePs predictions with known results: the 2D HP lattice model of Dill and Chan.³⁴ In this “toy” model of a protein, amino acid residues are represented as a chain of hydrophobic (H) and polar (P) beads on a two-dimensional square lattice. We introduce an interaction parameter, ϵ , quantifying the strength of attraction between hydrophobic residues, and ask under what conditions can BICePs scores find the correct value of ϵ given a combination of ensemble-averaged interresidue distance measurements, and MCMC sampling for a series of models $P^\epsilon(X)$. After exploring the robustness of the BICePs results with respect to conformational state coarse-graining, experimental noise, and data sparsity, we go on to see if BICePs scores can discriminate the all-atom force field that best agrees with experimental chemical shift measurements for molecular simulations of designed beta-hairpin peptides.

Methods

A toy lattice model of protein folding

As a model system, we consider a 12-residue 2D square lattice model protein with the following sequence of hydrophobic (H) and polar (P) residues: HPHPHPHPPH (Figure 3A). This 12-mer sequence has a total of 15037 unique microstates, defined as non-crossing chain conformations on the lattice with all rotational and mirror symmetries removed. We define the potential energy of each chain conformation X to be $U(X) = \epsilon \cdot n(X)$, where $n(X)$ is the number of non-bonded hydrophobic contacts (HH) for state X and ϵ is a bond interaction strength parameter. When ϵ is sufficiently negative, there is a single lowest-energy microstate that is also the global free energy minimum, and the Boltzmann populations are dominated by the folded conformation.

From the Boltzmann populations, we obtain a series of theoretical prior models for various values of ϵ , each of which can reconcile against experimental distance measurements using the BICePs algorithm:

$$P^{(\epsilon)}(X) = \frac{\exp(-\beta[\epsilon \cdot n(X)])}{\sum_X \exp(-\beta[\epsilon \cdot n(X)])} \quad (5)$$

To obtain simulated experimental distance measurements, we consider all possible distances between pairs of hydrophobic residues. For a given pair of residues, indexed by variable j , we compute the experimental measurement of distance d_j as the population-weighted average $\langle d_j^{\text{exp}} \rangle = \sum_X d_j(X) P^{(\epsilon)}(X)$. We assume the square lattice is on a 1-Å grid, so the distance is reported in Ångströms. For a given conformation X , the sum of the squared deviations from the experimental measurements are calculated as

$$\chi_d^2(X) = \sum_j (d_j(X) - \langle d_j^{\text{exp}} \rangle)^2 \quad (6)$$

Coarse-graining of conformational states—To examine the effects of coarse-graining on BICePs prediction accuracy, the 15037 microstates were clustered into smaller numbers of macrostates based on structural similarity (root mean squared deviation of residue positions). Macrostate decompositions of 1000, 5000 and 10000 states were obtained by performing k -centers clustering using the MSMBuilder³⁵ software package. The prior distribution for a set of macrostates Y is computed as $P^{(\epsilon)}(Y) = \sum_{X \in Y} P^{(\epsilon)}(X)$, where the $P^{(\epsilon)}(X)$ are computed as in Equation 5. The interresidue distance $d_j(Y)$ for a given macrostate Y is modeled as $d_j(Y) = \sum_{X \in Y} d_j(X) \Omega(Y)^{-1}$, where $\Omega(Y)$ is the number of microstates in macrostate Y .

Reference potentials.—In previous work,¹⁸ we developed a simple exponential reference potential for distance observables, and compared its performance in relation to a uniform reference potential. The exponential reference potential was derived as the maximum-entropy distribution given a constraint on the mean value of the measured experimental distances, which we inferred from the set of distances corresponding to each microstate X .

Here, we similarly derive a Gaussian reference potential, which we expect to more realistically model reference distributions for interatomic distances in chain molecules like proteins. A Gaussian reference distribution is the maximum-entropy distribution given a constraint on both the mean and the variance of the experiment observables, which we estimate from the mean and variance across the set of microstates X .

The Gaussian reference distribution for a set of observables $\mathbf{r} = (r_1, r_2, \dots, r_j)$ takes the form

$$Q_{\text{ref},j}(\mathbf{r}) = \prod_{j=1}^J Q_{\text{ref},j}(r_j | \mu_j, \sigma_j) = \prod_{j=1}^J \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \frac{-(r_j - \mu_j)^2}{2\sigma_j^2}. \quad (7)$$

We find the reference potential mean μ_j and variance σ_j^2 for observable r_j by maximizing the posterior probability of μ_j and σ_j^2 given the values of $r_j(X)$ across microstates X . Assuming a uniform prior $P(\mu_j) \sim 1$ and a non-informative Jeffreys prior $P(\sigma_j) \sim \sigma_j^{-1}$, we maximize

$$P(\sigma_j, \mu_j | \{r_j(X)\}) \propto \left[\prod_X Q_{\text{ref},j}(r_j(X) | \mu_j, \sigma_j) \right] P(\sigma_j) P(\mu_j) \quad (8)$$

with respect to μ_j and variance σ_j^2 and find

$$\mu_j = \frac{1}{N} \sum_X r_j(X), \quad (9)$$

$$\sigma_j^2 = \frac{1}{N+1} \sum_X (r_j(X) - \mu_j)^2, \quad (10)$$

where N is the number of microstate conformations X . Here, the $1/(N+1)$ term arises from correcting for bias.

Markov Chain Monte Carlo sampling of the posterior—The Metropolis-Hastings algorithm was used to perform MCMC sampling of the full posterior distribution

$$P^{(e)}(X, \sigma_d | D) \propto \left[\frac{Q(\mathbf{r}(X) | \sigma_d, D)}{Q_{\text{ref}}(\mathbf{r}(X))} \right] P^{(e)}(X) P(\sigma_d).$$

In this expression, $Q(\mathbf{r}(X) | \sigma_d, D)$ is a likelihood function enforcing restraints on J experimental observables $D = \{\langle r_j^{\text{exp}} \rangle\}$:

$$Q(\mathbf{r}(X) | \sigma_d, D) = \prod_{j=1}^J (2\pi\sigma_d^2)^{-1/2} \exp[-(r_j(X) - \langle r_j^{\text{exp}} \rangle)^2 / 2\sigma_d^2],$$

$P(\sigma_d) \sim \sigma_d^{-1}$ is the prior distribution of the uncertainty in the experimental measurements, $Q_{\text{ref}}(\mathbf{r}(X))$ is the reference potential, and $P^{(e)}(X)$ is the prior from theoretical modeling. Here, we assume that all measurements have similar uncertainties, and can be modeled using a single value of σ_d . (While we do not do so here, we note that our scheme can easily be

extended to measurement-specific uncertainties $\sigma_d(j)$ by sampling over scaled values $\alpha\sigma_d(j)$ for a series of constants α .

Several reference potentials $Q_{\text{ref}}(\mathbf{r}(X))$ were tested: uniform, exponential, and Gaussian. For the uniform reference potential, $Q_{\text{ref}}(\mathbf{r}(X)) \sim 1$. The mean μ_j and variance σ_j of the Gaussian reference potential for each chemical shift observable was parameterized as described above. Exponential reference potentials for each chemical shift were parameterized as described previously,¹⁸ using a mean $\mu_j = (N_j + 1)^{-1} \sum_i (r_j)_i$, where i is indexed over all conformational states.

MCMC sampling was performed in the state space $\xi = (X, \sigma_d)$ by proposing a randomly selected move $\xi \rightarrow \xi^*$, and accepting or rejecting the move with probability $\min[1, P(\xi^*|D)/P(\xi|D)]$ (i.e. the Metropolis criterion). In practice, we treat $-\ln P^{(\epsilon)}(X, \sigma_d|D)$ like an energy function, which, ignoring constant terms, is given by:

$$-\ln P^{(\epsilon)}(X, \sigma_d | D) = (J + 1)\ln\sigma_d + \frac{\chi_d^2(X)}{2\sigma_d^2} - \sum_{j=1}^J \frac{(r_j(X) - \mu_j)^2}{2\sigma_j^2} - \ln P^{(\epsilon)}(X). \quad (11)$$

We considered ϵ values ranging from 0 to $-5.5 k_B T$, in decrements of $0.5 k_B T$. For each ϵ value, 100 rounds MCMC sampling were performed, where each round consisted of 107 steps, resulting in 109 steps of total sampling. Such extensive sampling was useful in reducing finite sampling errors. In each calculation, we perform so-called “one-shot” FEP, sampling ensembles using $[P^{(\epsilon)}(X)]^\lambda$ for $\lambda = 1$ and $\lambda = 0$ (the reference state), and use the MBAR algorithm to obtain estimates of the BICePs score $f^{(\epsilon)} = -\ln[Z_{\lambda=1}^{(\epsilon)} / Z_{\lambda=0}^{(\epsilon)}]$.

Molecular simulation and BICePs calculations for designed beta-hairpin peptides

TrpLoop-2a and TrpLoop-2b (Figure 4) are peptides designed to fold to a beta hairpin stabilized by a terminal Trp-Trp capping motif.³⁶ The GWxQ turn motif in these peptides are designed to mimic a conserved recognition sequence in bacterial protein LapD, which normally sequesters LapG, a periplasmic protease able to cleave adhesin protein LapA and disperse bacterial biofilms.³⁷ As part of previous work with experimental collaborators, we have performed extensive molecular simulations of TrpLoop-2a/-2b, and obtained NMR chemical shift measurements to test the accuracy of these simulations.⁶

All-atom molecular dynamics simulation—As described in previous work,⁶ all-atom replica-exchange molecular dynamics (REMD) simulations were performed using GROMACS 4.5.4³⁸ on the Owlsnest high performance computing cluster at Temple University. Twenty- four replicas with temperatures exponentially-spaced from 300 to 450 K were chosen to ensure broad conformational sampling. The total simulation time of each replica is around 2 μ s. Three different force fields-AMBER ff96,³⁹ AMBER ff99sb-ildn⁴⁰ and AMBER ff99sb-ildn-nmr⁴¹—were used in conjunction with the OBC GBSA implicit solvation model.⁴² We also performed molecular dynamics simulations using the AMBER ff99sb-ildn-nmr⁴¹ force field in combination with the TIP3P explicit solvent model, using the Folding@home distributed computing platform.⁴³ Aggregate explicit simulation data for

TrpLoop-2a and TrpLoop-2b were 167 μ s and 54 μ s, respectively. Average trajectory lengths for TrpLoop-2a and TrpLoop-2b were 1034 ns and 316 ns, respectively.

Conformational state decomposition—In order to make force field comparisons using BICePs, it is essential to define a unified set of conformational states. To do this, we first performed time-structure-based independent component analysis (tICA).^{44,45} We projected the data to the four largest tICA components, and then performed k -centers clustering using MSMBuild to define a set of conformational states. The implicit solvent simulation data was projected to tICA coordinates and assigned using these state definitions. As done in the lattice model analysis, we performed clustering using different numbers of states (250, 500, 1000, 5000, 10000) to test the effects of conformational coarse-graining on the BICePs results. To account for stochastic variation in the clustering results, multiple clusterings were performed, from which uncertainties could be estimated. Theoretical predictions of the prior distribution $P(X)$ come from the populations each cluster. For implicit-solvent simulations, the lowest-temperature replica was used.

Experimental restraints (NMR chemical shifts)—Experimental NMR HN and Ha chemical shifts were measured in previous work.⁶ The SHIFTX2 algorithm,⁴⁶ using wrapper functions implemented in MDTraj,⁴⁷ was used to calculate predicted HN and Ha chemical shifts from the simulation trajectory data. The predicted chemical shifts used in the BICePs calculation are ensemble-averaged SHIFTX2 predictions across the snapshots corresponding to each conformational state.

Markov Chain Monte Carlo sampling of the posterior—Similar to our lattice model sampling, multiple runs of MCMC sampling are performed to reduce finite sampling errors. Here 20 runs MCMC sampling are performed with 10 million steps for each run. Instead of distance restraints, we use chemical shifts in posterior distribution in sampling: The likelihood function used to enforce restraints on J experimental chemical shift (cs) observables $D = \{r_j^{\text{exp}}\}$ is:

$$Q(\mathbf{r}(X) | \sigma_{cs}, D) = \prod_{j=1}^J (2\pi\sigma_{cs}^2)^{-w_j/2} \exp[-w_j(r_j(X) - r_j^{\text{exp}})^2 / 2\sigma_{cs}^2],$$

where r_j are chemical shifts of protons indexed by j , and w_j is a weight factor (for example, $w_j = 1/n$ for n equivalent chemical shifts). This way, there are $J = \sum_j w_j$ independent chemical shifts, with the sum of squared deviations computed as:

$$\chi_{cs}^2(X) = \sum_{j=1}^J w_j (r_j(X) - r_j^{\text{exp}})^2. \quad (12)$$

As in the lattice model, several reference potentials $Q_{\text{ref}}(\mathbf{r}(X))$ were tested: uniform, exponential, and Gaussian. In practice, all reference distributions were fairly non-informative (i.e. flat), and we did not attempt to center the experimental data about random-

coil shift values (Figure S2-3). We treat $-\ln P(X, \sigma_{cs}|D)$ like an energy function, which, ignoring constant terms, is given by:

$$-\ln P(X, \sigma_{cs} | D) = (J + 1) \ln \sigma_d + \frac{\chi_{cs}^2(X)}{2\sigma_{cs}^2} + \ln Q_{\text{ref}}(X) - \ln P(X). \quad (13)$$

Results

Force field parameterization in a toy lattice model

Can BICePs scores be used for force field parameterization? To test this idea in a toy model, we set the parameter ϵ to $-2.0 k_B T$ and generated a mock experimental data set for eight interresidue distances (see Figure 3B) from the resulting conformational distribution $P^{(\epsilon)}(X)$ (see Equation 5).

BICePs scores correctly identify the optimal interaction strength ϵ .—We tested a series of ϵ values starting from $0.0 k_B T$ to $-5.5 k_B T$ in decrements of $-0.5 k_B T$, performing BICePs calculations using 10^9 steps of total MCMC sampling for each value of ϵ . Three different reference potentials were tested: uniform, exponential and Gaussian. Each theoretical prior $P^{(\epsilon)}(X)$ can be considered as a unique model, analogous to different force field parameterizations. In this case, the BICePs score should be able to report how well each model agrees with the experimental measurements, so we can find the optimal value of parameter ϵ .

Even though the toy lattice model is relatively simple, we emphasize that the inference problem here is quite challenging, for several reasons. For one, the large number of microstates (15,037) in the model requires a substantial amount of conformational sampling to produce converged statistically converged results. Furthermore, the mock experimental measurements suffer substantially from ensemble-averaging, and with only eight such distance measurements, it is unclear how much detail can be resolved about the underlying microscopic ensemble. Complicating matters further is the highly cooperative folding landscape of the toy lattice model when $\epsilon = -2.0 k_B T$, resulting in correlated distance restraints that do not independently confer useful information.

To illustrate this inference challenge, we show an example of the likelihood and prior distributions calculated for a particular distance observable, $d_{(2,9)}$, for several values of ϵ near $-2.0 k_B T$ (Figure 5). Recall that the optimal BICePs score occurs when there is the greatest overlap between the likelihood and prior distributions (see Figure 2).

Figure 6 shows the average BICePs score for different values of ϵ , using uniform, exponential, and Gaussian reference potentials. Remarkably, we find that the lowest BICePs score corresponds to ϵ value of $-2.0 k_B T$, computed using Gaussian reference potential. This suggests that objective model selection by the lowest BICePs score can successfully be used to perform force field parameterization. The BICePs score for the best model is near -6 , which corresponds to a evidence ratio of about 400:1 favoring the $\epsilon = -2.0 k_B T$ model with

experimental restraints over a model with only experimental restraints (i.e. the $\epsilon = 0 \text{ } k_B T$ model).

These results also highlight the importance of choosing an appropriate reference potential. The uniform reference potential gives a lowest BICePs score of around -0.4 at an ϵ value of $-1.5 \text{ } k_B T$, which is close, but misses the correct answer. The exponential reference potential gives the worst results, with positive BICePs scores for all values of ϵ . The comparison of these three reference potentials suggests that it is important to choose a reference potential that accurately represents any *a priori* knowledge we have about the expected distribution of experimental observables. Because the 2D lattice model is a polymer, the Gaussian reference potential is the most physical representation, while the exponential reference potential is the least physical (Figure S1).

Furthermore, these results also reveal the limitations of finite sampling, especially for highly negative ϵ values. As the interaction strength increases, the conformational ensemble $P^{(\epsilon)}(X)$ is increasingly dominated by the folded state. As a result, there is little discriminatory power to the BICePs score beyond $\epsilon = -3.5 \text{ } k_B T$. At the same time, statistical uncertainty increases due to finite sampling error, as unfolded conformations are sampled more rarely.

Conformational coarse-graining limits the accuracy of force field parameterization.

—In its current form, BICePs performs sampling over a set of discrete conformational states, X . While in the toy lattice model we can exactly enumerate and sample individual microstates, this is not possible for most continuous-space molecular models. In practice, we must choose some discrete partitioning of conformational states, for example, by conformational clustering, or consideration of discrete energy minima. The resolution of conformational states is potentially important for several reasons. For one, the state decomposition determines the integrated posterior probability of reference potential (i.e. Z_0 , when $P(X) \sim 1$), which will affect the BICePs score. Also, while coarse-graining of conformational states may improve finite sampling error, useful information may be lost due to conformational averaging.

What is the effect of coarse-graining on the accuracy of model selection by BICePs ? To examine this, we clustered the 15,037 microstates to 1000, 5000 and 10000 macrostates (see Methods). We still use $\epsilon = -2.0 \text{ } k_B T$ to compute the experimental distances. Again we performed 10^9 steps of MCMC sampling, using a Gaussian reference potential with ϵ values ranging from 0.0 to $-5.5 \text{ } k_B T$, and compute BICePs scores as a function of ϵ (Figure 7). The results show that, for the lattice model, coarse-graining to 10000 macrostates still retains sufficient information for BICePs to find the correct value of $\epsilon = -2.0 \text{ } k_B T$, although this instigates greater statistical uncertainty, as reflected by the larger error bars (Figure 7C).

When further coarse-graining is performed, to 5000 and 1000 macrostates, the BICePs score no longer is able to select the correct model. Instead, BICePs scores keep decreasing with increasingly negative ϵ values; such that any model with $\epsilon < -3.5 \text{ } k_B T$ appears to be optimal. This loss of information due to conformational coarse-graining arises when a sufficient number of microstates become lumped together with the native folded state, making the predicted distance observables $\langle d_j \rangle$ for that state much closer to the experimental

measurements. In that case, optimal models will have the largest possible native-state population, make highly negative ϵ values most favorable. One might imagine that such artifacts could be improved with different conformation clustering protocols; for instance, kinetics- based clustering that might more finely partition conformational states near the native folded conformation.⁴⁸ Even with these improvements, however, there is likely a threshold of conformational clustering beyond which key information for parameterization is lost. Thus, it would appear that the accuracy of BICePs benefits from the most fine-grained decomposition as possible.

BICePs results are robust to experimental noise.—To examine how experimental noise affect BICePs predictions, we introduced random errors in the mock experimental distances measured for the toy lattice model when $\epsilon = -2.0k_B T$, creating various sets of noisy distance measurements d'_j . We define noise level σ added to the experimental data as

$$\sigma = \sqrt{\frac{\sum_{j=1}^J (d'_j - d_j^{\text{exp}})^2}{J}} \quad (14)$$

where d_j^{exp} is the j th interresidue distance measurement ($J = 8$, as shown in Figure 3). For each set of noisy distance measurements d'_j , we first perform BICePs calculations for different values of ϵ (0.0 to $-5.5 k_B T$), using Gaussian reference potentials, and 10^8 steps of MCMC sampling. We then select the model $P^{(\epsilon)}(X)$ with the lowest BICePs score. To assess how each BICePs prediction $P^{(\epsilon)}(X)$ varies from the theoretical distribution $P^{(\epsilon)}(X)$ (computed using Equation 5 when $\epsilon = -2.0 k_B T$), we calculate the Kullback-Leibler divergence (also known as relative entropy) between the two distributions:

$$D_{KL} = \sum_X P^{(\epsilon)}(X) \ln \frac{P^{(\epsilon)}(X)}{P^{(\epsilon)}(X)} \quad (15)$$

By repeating this procedure for many different sets of noisy distance measurements d'_j , we construct a plot of the Kullback-Leibler divergence D_{KL} as a function of the noise level σ (Figure 8). We find that when the noise level is lower than about 2 Å, BICePs is able to infer conformational populations close to theoretical distribution, as reflected by relatively small D_{KL} . However, higher levels of noise leads to a sharp increase in D_{KL} , indicating poor predictions from BICePs. The poor predictions are likely due to inaccurate populations predicted for the native conformational state, which greatly contributes to the Kullback-Leibler divergence metric. These results suggest that in the presence of some experimental noise, BICePs predictions of conformational populations remains robust. It also suggests that for BICePs to make successful predictions, sufficiently accurate experimental data must be obtained.

BICePs results are robust to data sparsity.—To test how the sparsity of experimental data affects BICePs predictions, we first ordered all eight interresidue distances in reverse

order of their sequence separation; this roughly corresponds to an ordering of most-informative to least-informative contacts. Then, one-by-one, we removed these distance measurements from the available experimental data until only three short distances were left. This generated groups of experimental data with different levels of sparsity (G0-G5 in Figure 9C).

Next, we performed 10^8 of MCMC sampling, using Gaussian reference potentials, for each group of experimental data, computing BICePs scores over range of ϵ values (0.0 to $-5.5 k_B T$). The results show that even upon losing highly informative distance measurements $d_{(0,11)}$ and $d_{(0,9)}$ (groups of experimental data G0, G1 and G2), the remaining distance measurements are sufficient for model selection by BICePs, with the lowest averaged BICePs scores corresponding to a model using $\epsilon = -2.0 k_B T$ (Figure 9). Losing more experimental measurements (groups of experimental data G3, G4 and G5) leads to incorrect predictions of the optimal model, with the lowest BICePs score occurring for models with $\epsilon = -1.5 k_B T$. Still, an error of only $0.5 k_B T$ for a parameterization problem with only three distance measurements is quite impressive, suggesting that even sparse measurements, if sufficiently accurate, can be used effectively with BICePs to infer conformational populations.

Application of BICePs: All-atom force field evaluation for designed beta-hairpin peptides

Can the BICePs approach work for parameterization of all-atom force field models? To test this idea, we used BICePs scores to evaluate four AMBER all-atom molecular force fields against experimental NMR chemical shift data for two designed β -hairpin mimics, TrpLoop-2a and TrpLoop-2b. The force fields tested included three combinations used with an GBSA implicit solvent model (ff96, ff99sb-ildn, and ff99sb-ildn-nmr), and one used with TIP3P explicit solvent (ff99sb-ildn-nmr). More details can be found in the Methods section.

As described in Methods, experimental restraints came from NMR chemical shifts (Ha and HN). As with any experimental data of this type, there is some uncertainty associated with the experimental values, the extent to which BICePs is designed to infer. Additional uncertainty comes from the predicted chemical shifts from simulated conformational ensembles using SHIFTX2. The sparse number of experimental restraints, as well as the intrinsic sources of error, provide an excellent opportunity to test BICePs in a “real-word” situation.

To test the performance of BICePs across models with different extents of coarse-graining, we constructed models using different numbers of conformational states (250, 1000, 5000, 10000) obtained by k -centers clustering. Several rounds of clustering and subsequent BICePs calculations were performed to account for stochastic variation in the clustering results.

For each force field, we explored three different reference potentials (uniform, exponential, Gaussian), performing 2×10^8 total steps of MCMC sampling. The results, shown in Figure 10 and Figure 11 for TrpLoop-2a and TrpLoop-2b respectively, reveal a consistent ranking of force field accuracy for the TrpLoop-2a and -2b systems. The BICePs scores suggest that AMBER ff99sb-ildn-nmr, in combination with the TIP3P explicit solvent model, most

accurately predicts the experimental NMR chemical shifts. This result is consistent with recent tests of different force fields,⁴⁹ and makes sense given that AMBER ff99sb-ildn-nmr was parameterized specifically to better reproduce NMR observables.^{41,41}

The BICePs scores also suggest that the least-accurate force field combination is AMBER ff99sb-ildn-nmr + GBSA. This finding is consistent with the results of a more conventional force field validation study we published previously,⁶ in which we found that AMBER ff99sb-ildn-nmr + GBSA also poorly reproduced the experimental NMR chemical shifts and folded- state populations. We attributed this finding to a well-known tendency of the GBSA implicit solvent model to overstabilize helical conformations.^{50,51}

Interestingly, while the BICePs scores computed here suggest that AMBER ff99sb-ildn-nmr + TIP3P is the most accurate model, our previous work has suggested that AMBER ff96 + GBSA agrees best with experimental observables.⁶ AMBER ff96 is known to overstabilize beta-sheet conformations when used with explicit-solvent models. When used with GBSA implicit solvent models, however, AMBER ff96 benefits from a fortuitous correction of this bias due to the overstabilization of alpha-helical states by GBSA, resulting in a potential that can model small mini-proteins⁵⁰ accurately enough to be used for *ab initio* protein folding.⁵² Despite these favorable attributes, AMBER ff96 + GBSA suffers from salt-bridge overstabilization and other artifacts that should ultimately favor the use of explicit-solvent models for accurate simulation. We suspect that the BICePs approach used here, which considers conformational populations, can more finely distinguish differences between force fields than more conventional approaches. In our previous study, we compared experimental and simulated chemical shift observables ensemble-averaged over the entire conformational ensemble. This is akin to a BICePs calculation that only considers only one conformational state.

We also note that, unlike the lattice model results, the trends in BICePs scores are mostly insensitive to the reference potential used. This is likely because all three reference potentials tested (uniform, exponential, Gaussian) are similarly non-informative (i.e. flat). A plot of the distribution of ensemble averaged chemical shifts cross all conformational states, superimposed with the computed reference potentials and it clearly shows this clearly (Figure S2 and S3). As the number of coarse-grained states increases, BICeP scores tend to decrease slightly, indicating better models. At the same time, statistical variance tend to increase slightly; this trend is expected for models with larger number of states, due to finite sampling and stochastic clustering effects.

Discussion

The main purpose of BICePs is to infer conformational state populations from a combination of computational modeling and sparse experimental data. Here, for the first time, we explore the possibility of using the integrated posterior model probabilities (i.e. BICePs scores) as a quantitative metric for parameterization and validation of force field models against experimental data. The advantage of this approach is the rigorous statistical framework for model selection provided by BICePs scores, while the potential pitfalls of this approach lie in its practical application.

For this reason, we have explored the practical aspects of model selection via BICePs using a toy lattice model of protein folding. The lessons learned have implications for future work in using BICePs for model selection. Most importantly, our tests using the toy lattice model demonstrate that the BICePs algorithm can indeed be used to perform quantitative model selection, by computing BICePs scores for competing models. For this to work well, however, it appears that several practical requirements must be in place. First, an appropriate reference potential must be used; this is a non-trivial and problem-dependent choice for which we have derived some useful options (uniform, exponential, and Gaussian reference potentials) but may require more detailed consideration. Second, conformational states must be sufficiently fine-grained, so that valuable information isn't lost to conformational averaging. Our results suggest that if these requirements can be satisfied, then BICePs can accurately infer optimal models, and do so fairly robustly in the presence of sparse and noisy experimental measurements.

Using the lessons learned from the lattice model studies, we have applied BICePs to an example force field validation problem: selecting the all-atom force field that best agrees with experimental chemical shift data for two designed beta-hairpin peptides. Here, our results agree with previous assessments of force field quality, suggesting that BICePs scores are up to the task of model selection alongside more complicated, all-atom simulation models. An issue that may require more attention in the future is deriving a unified set of conformational states for all-atom simulations that can be compared across different models. We suggest it is important to use a sufficiently fine-grained state decomposition, but this may be difficult to achieve while also trying to choose a state decomposition that overlaps well across the different models. While we found that AMBER ff99sb-ildn-nmr + TIP3P is the most accurate model for TrpLoop-2a and -2b, an important caveat is that the state decomposition for model comparison came from conformational clustering of trajectory data obtained in that force field; it is therefore possible that the favorable BICePs scores computed for AMBER ff99sb-ildn-nmr + TIP3P are, in part, due to a more favorable state decomposition. In future work, we hope to look more closely at the possibility of bias arising from the state decomposition. For example, other methods of dimensionality reduction and conformational clustering may perform better than the tICA projection scheme used here. We can also alleviate such bias by implementing a version of BICePs that works with continuous-space conformational states. For all-atom simulation models, this can be performed in practice using bias-exchange molecular dynamics simulation.

Conclusions

We have presented a Bayesian inference method to parameterize and validate force fields, based on the BICePs (Bayesian Inference of Conformational Populations) algorithm. BICePs reconciles computational models with experimental information reweighed by proper reference potentials, by using MCMC to sample the posterior distribution of conformations and model parameters. We have shown how the theory underlying BICePs can be used for objective model selection, by computing a quantity we call the BICePs score, which reflects the total integrated posterior probability of a given model. This feature, along with the implementation of reference potentials, are unique advantages of the BICePs algorithm.

By studying the performance of this approach in a 2D lattice model of protein folding, we have shown that the BICePs approach can successfully infer model parameters, even in the presence of sparse and/or noisy experimental data, provided that an appropriate reference potential is used, and that conformational states are sufficiently fine-grained. As an initial step towards using BICePs to evaluate potentials used in all-atom molecular simulations, we used BICePs scores to select optimal AMBER force field combinations for two designed beta-hairpin peptides against sparse experimental chemical shift data.

In the future, we hope to improve the ability of BICePs to select optimal all-atom simulation models by (1) considering alternative conformation state decompositions, (2) developing more realistic reference potentials for particular experimental observables, and (3) incorporating measurement-specific uncertainties to better account for outliers. We plan to write an improved version of BICePs (v2.0) which will be faster, more flexible and easier for users. With these improvements, we expect BICePs to be a powerful Bayesian inference tool to reconcile molecular models against experimental data, enabling the development of accurate models for specific purposes (e.g. foldamers, peptidomimetics, macrocycles) from more general-purpose potentials. The BICePs algorithm is freely available with source code and examples at <http://github.com/vvoelz/nmr-biceps>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank the participants of Folding@home, without whom this work would not be possible. This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant CNS-09-58854 and MCB-1412508, and by NIH grants 1S10OD020095-01 and 1R01GM123296-01.

References

- (1). Bonomi M; Heller GT; Camilloni C; Vendruscolo M Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 2017, 42, 106–116. [PubMed: 28063280]
- (2). Wright PE; Dyson HJ Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 2015, 16, 18–29. [PubMed: 25531225]
- (3). Hilser VJ; Wrabl JO; Motlagh HN Structural and Energetic Basis of Allostery. *Ann. Rev. Biophys* 2012, 41, 585–609. [PubMed: 22577828]
- (4). Hilser VJ; Garcia-Moreno E B; Oas TG; Kapp G; Whitten ST A Statistical Thermodynamic Model of the Protein Ensemble. *Chem. Rev.* 2006, 106, 1545–1558. [PubMed: 16683744]
- (5). Butterfoss GL; Yoo B; Jaworski JN; Chorny I; Dill KA; Zuckermann RN; Bonneau R; Kirshenbaum K; Voelz VA De Novo Structure Prediction and Experimental Characterization of Folded Peptoid Oligomers. *Proc. Natl. Acad. Sci. U. S. A. of the United States of America* 2012, 109, 14320–14325.
- (6). Ge Y; Kier BL; Andersen NH; Voelz VA Computational and Experimental Evaluation of Designed β -Cap Hairpins Using Molecular Simulations and Kinetic Network Models. *J. Chew., Inf. Model* 2017, 57, 1609–1620.
- (7). McHugh SM; Rogers JR; Solomon SA; Yu H; Lin Y-S Computational Methods to Design Cyclic Peptides. *Curr. Opin. Chew., Biol* 2016, 34, 95–102.
- (8). Mukherjee S; Zhou G; Michel C; Voelz VA Insights into Peptoid Helix Folding Cooperativity from an Improved Backbone Potential. *J. Phys. Chem. B.* 2015, 119, 15407–15417. [PubMed: 26584227]

- (9). Voelz VA; Dill KA; Chorny I Peptoid Conformational Free Energy Landscapes from Implicit-Solvent Molecular Simulations in AMBER. *Biopolymers* 2011, 96, 639–650. [PubMed: 21184487]
- (10). Razavi AM; Voelz VA Kinetic Network Models of Tryptophan Mutations in β -Hairpins Reveal the Importance of Non-Native Interactions. *J. Chem. Theor. Comput* 2015, 11, 2801–2812.
- (11). Razavi AM; Wuest WM; Voelz VA Computational Screening and Selection of Cyclic Peptide Hairpin Mimetics by Molecular Simulation and Kinetic Network Models. *J. Chem. Inf. Model* 2014, 54, 1425–1432. [PubMed: 24754484]
- (12). Lao BB; Drew K; Guarracino DA; Brewer TF; Heindel DW; Bonneau R; Arora PS Rational Design of Topographical Helix Mimics as Potent Inhibitors of Protein-Protein Interactions. *J. Am. Chem. Soc.* 2014, 136, 7877–7888. [PubMed: 24972345]
- (13). Robinson J; DeMarco S; Gombert F; Moehle K; Obrecht D The Design, Structures and Therapeutic Potential of Protein Epitope Mimetics. *Drug Discovery Today* 2008, 13, 944–951. [PubMed: 18725320]
- (14). Parker MFL; Osuna S; Bollot G; Vaddypally S; Zdilla MJ; Houk KN; Schafmeister CE Acceleration of an Aromatic Claisen Rearrangement via a Designed Spiroligozyme Catalyst that Mimics the Ketosteroid Isomerase Catalytic Dyad. *J. Am. Chem. Soc.* 2014, 136, 3817–3827.
- (15). Sun J; Zuckermann RN Peptoid Polymers: a Highly Designable Bioinspired Material. *ACS Nano* 2013, 7, 4715–4732. [PubMed: 23721608]
- (16). Yang W; Kang B; Voelz VA; Seo J Control of Porphyrin Interactions via Structural Changes of a Peptoid Scaffold. *Org. Biomol. Chem.* 2017, 15, 9670–9679. [PubMed: 29119173]
- (17). Kang B; Yang W; Lee S; Mukherjee S; Forstater J; Kim H; Goh B; Kim TY; Voelz VA; Pang Y; Seo J Precisely Tuneable Energy Transfer System Using Peptoid Helix-Based Molecular Scaffold. *Sci. Rep* 2017, 7, 20130901–10.
- (18). Voelz VA; Zhou G Bayesian Inference of Conformational State Populations from Computational Models and Sparse Experimental Observables. *J. Comp. Chem.* 2014, 35, 2215–2224. [PubMed: 25250719]
- (19). Wakefield AE; Wuest WM; Voelz V. a. Molecular Simulation of Conformational Pre-Organization in Cyclic RGD Peptides. *J. Chem. Inf. Model* 2015, 55, 806–813. [PubMed: 25741627]
- (20). Mukherjee S; Zhou G; Michel C; Voelz V a. Insights into Peptoid Helix Folding Cooperativity from an Improved Backbone Potential. *J. Phys. Chem. B.* 2015, 119, 15407–15417. [PubMed: 26584227]
- (21). Rieping W; Habeck M; Nilges M Inferential Structure Determination. *Science (New York, N.Y.)* 2005, 309, 303–306.
- (22). Perez A; MacCallum JL; Dill KA Accelerating Molecular Simulations of Proteins using Bayesian Inference on Weak Information. *Proc. Natl. Acad. Sci. U. S. A* 2015, 112, 11846–11851. [PubMed: 26351667]
- (23). Beauchamp KA; Pande VS; Das R Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. *Biophys. J.* 2014, 106, 1381–1390. [PubMed: 24655513]
- (24). Fisher CK; Huang A; Stultz CM Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* 2010, 132, 14919–14927. [PubMed: 20925316]
- (25). Bonomi M; Camilloni C; Vendruscolo M Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble using Metadynamics. *Sci. Rep* 2016, 6, 96–11.
- (26). Bonomi M; Camilloni C; Cavalli A; Vendruscolo M Metainference: A Bayesian Inference Method for Heterogeneous Systems. *Science Advances* 2016, 2, e1501177. [PubMed: 26844300]
- (27). Cicero DO; Barbato G; Bazzo R NMR Analysis of Molecular Flexibility in Solution: A New Method for the Study of Complex Distributions of Rapidly Exchanging Conformations. Application to a 13-Residue Peptide with an 8-Residue Loop. *J. Am. Chem. Soc.* 1995, 117, 1027–1033.
- (28). Smith AB III; Risatti CA; Atasoylu O; Bennett CS; Liu J; Cheng H; TenDyke K; Xu Q Design, Synthesis, and Biological Evaluation of Diminutive Forms of (+)-Spongistatin 1: Lessons Learned. *J. Am. Chem. Soc.* 2011, 133, 14042–14053. [PubMed: 21761891]

- (29). Atasoylu O; Furst G; Risatti C; Smith III AB The Solution Structure of (+)- Spongistatin 1 in DMSO. *Org. Lett.* 2010, 12, 1788–1791. [PubMed: 20297809]
- (30). Olsson S; Frellsen J; Boomsma W; Mardia KV; Hamelryck T Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. *PLoS ONE* 2013, 8, e79439. [PubMed: 24244505]
- (31). Olsson S; Boomsma W; Frellsen J; Bottaro S; Harder T; Ferkinghoff-Borg J; Hamelryck T Generative Probabilistic Models Extend the Scope of Inferential Structure Determination. *J. Mag. Res.* 2011, 213, 182–186.
- (32). Hamelryck T; Borg M; Paluszewski M; Paulsen J; Frellsen J; Andreetta C; Boomsma W; Bottaro S; Ferkinghoff-Borg J Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. *PLoS ONE* 2010, 5, e13714.
- (33). Shirts MR; Chodera JD Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* 2008, 129, 124105–11. [PubMed: 19045004]
- (34). Dill KA; Bromberg S; Yue K; Fiebig KM; Yee DP; Thomas PD; Chan HS Principles of Protein Folding-a Perspective from Simple Exact Models. *Protein Science* 1995, 4, 561–602. [PubMed: 7613459]
- (35). Beauchamp KA; Bowman GR; Lane TJ; Maibaum L; Haque IS; Pande VS MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theor. Comput.* 2011, 7, 3412–3419.
- (36). Kier BL; Shu I; Eidenschink LA; Andersen NH Stabilizing Capping Motif for β -Hairpins and Sheets. *Proc. Natl. Acad. Sci. U. S. A.* 2010, 107, 10466–10471. [PubMed: 20484672]
- (37). Boyd CD; Chatterjee D; Sondermann H; O'Toole G a. LapG, Required for Modulating Biofilm Formation by *Pseudomonas fluorescens* Pf0–1, is a Calcium-Dependent pProtease. *J. Bacteriol* 2012, 194, 4406–4414. [PubMed: 22707708]
- (38). Pronk S; Páll S; Schulz R; Larsson P; Bjelkmar P; Apostolov R; Shirts MR; Smith JC; Kasson PM; van der Spoel D; Hess B; Lindahl E GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* 2013, 29, 845–854. [PubMed: 23407358]
- (39). Kollman PA Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc. Chem. Res.* 1996, 29, 461–469.
- (40). Lindorff-Larsen K; Piana S; Palmo K; Maragakis P; Klepeis JL; Dror RO; Shaw DE Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Structure, Function, and Bioinformatics* 2010, 78, 1950–1958.
- (41). Li D-W; Bruschweiler R NMR-Based Protein Potentials. *Angew. Chem., Int. Ed.* 2010, 49, 6778–6780.
- (42). Onufriev A; Case D. a.; Bashford D Effective Born Radii in the Generalized Born Approximation: The Importance of Being Perfect. *J. Comp. Chem.* 2002, 23, 1297–1304. [PubMed: 12214312]
- (43). Shirts M Screen Savers of the World Unite! *Science* 2000, 290, 1903–1904. [PubMed: 17742054]
- (44). Schwantes CR; Pande VS Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theor. Comput* 2013, 9, 2000–2009.
- (45). Pérez-Hernández G; Paul F; Giorgino T; De Fabritiis G; Noé F Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* 2013, 139, 015102–14. [PubMed: 23822324]
- (46). Han B; Liu Y; Ginzinger SW; Wishart DS SHIFTX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* 2011, 50, 43–57. [PubMed: 21448735]
- (47). McGibbon RT; Beauchamp KA; Harrigan MP; Klein C; Swails JM; Hernández CX; Schwantes CR; Wang L-P; Lane TJ; Pande VS MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 2015, 109, 1528–1532. [PubMed: 26488642]
- (48). Schwantes CR; Pande VS Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theor. Comput.* 2013, 9, 2000–2009.

- (49). Beauchamp KA; Lin Y-S; Das R; Pande VS Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theor. Comput.* 2012, 8, 1409–1414.
- (50). Shell MS; Ritterson R; Dill KA A Test on Peptide Stability of AMBER Force Fields with Implicit Solvation. *J. Phys. Chem. B.* 2008, 112, 6878–6886. [PubMed: 18471007]
- (51). Roe DR; Okur A; Wickstrom L; Hornak V; Simmerling C Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B.* 2007, 111, 1846–1857. [PubMed: 17256983]
- (52). Voelz VA; Bowman GR; Beauchamp K; Pande VS Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1–39). *J. Am. Chem. Soc.* 2010, 132, 1526–1528. [PubMed: 20070076]

Figure 1:

A highly informative distance restraint has an experimental likelihood that differs significantly from its reference distribution.

Figure 2:

To compute unique BICePs scores $f^{(1)}$ and $f^{(2)}$ two theoretical models, $P^{(1)}(X, \sigma|D)$ and $P^{(2)}(X, \sigma|D)$, each must be calculated using a common reference state, which we take to be a model of uniform populations $P(X) \sim 1$, i.e. no information from theory. To efficiently and accurately calculate ratios of posterior model probabilities, we use a free energy perturbation approach, performing posterior sampling for a series of models $P_{\lambda}^{(k)}(X, \sigma | D) \sim [P^{(k)}(X, \sigma | D)]^{\lambda}$, for $0 \leq \lambda_k \leq 1$, and then using the MBAR free energy estimator to calculate the BICePs score as $f^{(k)} = -\ln Z_{\lambda=1}^{(k)} / Z_{\lambda=0}^{(k)}$.

Figure 3:
(A) The folded state of a 12-residue 2D HP lattice model. Highlighted in yellow are favorable non-bonded contacts between hydrophobic residues (black beads). (B) A diagram of the eight distance measurements used in this work.

Figure 4:
NMR structure and sequences of (A) TrpLoop2a and (B) TrpLoop2b. Mutated residues are shown in red.

Figure 5:

Likelihood and prior distributions calculated for distance observable, $d_{(2,9)}$ for several values of ϵ . In black is shown the prior distribution $P^{(\epsilon)}(d_j)$ predicted from the theoretical model. Note that the lattice model gives discrete values of $d_{(2,9)}$. In blue is shown a Gaussian reference potential $Q_{\text{ref}}(d_j)$, parameterized using the distribution of $d_{(2,9)}$ distances when $\epsilon = 0 k_B T$. In green is shown the distance restraint potential $Q_{\text{restraint}}(d_j) = (2\pi\sigma_d^2)^{-1/2} \exp(-(d_j - d_j^{\text{exp}})^2 / 2\sigma_d^2)$ using typical values of σ_d^2 sampled with BICePs. One can see that this distribution gets narrower for more negative values of ϵ as the

conformational ensemble becomes less heterogeneous. The green dotted line shows the location of the experimental measurement d_j^{exp} . In red is shown the overall likelihood function $Q(d_j) = Q_{\text{restraint}}(d_j)/Q_{\text{ref}}(d_j)$.

Figure 6: BICePs scores for different ϵ values (0.0 to $-5.5 k_B T$) with different reference potentials. With Gaussian reference potential implemented, BICePs correctly identifies the optimal interaction strength ($\epsilon = -2.0 k_B T$). The shaded region shows the standard deviation of BICePs scores across multiple runs of the algorithm.

Figure 7: BICePs scores for different ϵ values (0.0 to $-5.5 k_B T$) from different degrees of coarse-graining clustering. The shadow parts show the error bars. The shaded region highlights optimal models selected based on lowest BICePs scores. Panel D is identical to Figure 6

Figure 8:

The Kullback-Leibler divergence of BICePs prediction $P^{(\epsilon)}(X)$ (made in the presence of experimental noise) with respect to the actual distribution $P^\epsilon(X)$ is shown as a function of the experimental noise level, σ . Blue dots represent averages across BICePs predictions for a given set of noisy distances d'_j ; error bars display standard deviations. The red and green star represent the lower and upper bound of divergence. The dashed magenta line is to qualitatively guide the eye.

Figure 9:

(A) BICePs scores for different ϵ values (0.0 to $-5.5 k_B T$) calculated using different degrees of experimental sparsity (G0-G5). (B) The folded state of a 12-residue 2D HP lattice model and a diagram of the all available eight interresidue distances. (C) Diagrams of the different sets of experimental measurements used in each BICePs calculation.

Figure 10: BICePs scores for TrpLoop-2a with (A) uniform (B) exponential and (C) Gaussian reference potential. Averaged BICePs score for each trial is shown with error bars.

Figure 11: BICePs scores for TrpLoop-2b with (A) uniform (B) exponential and (C) Gaussian reference potential. Averaged BICePs score for each trial is shown with error bars.