Supporting Information for Inferring Structural Ensembles from Noisy Experiments: Application to Trialanine

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

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

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

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

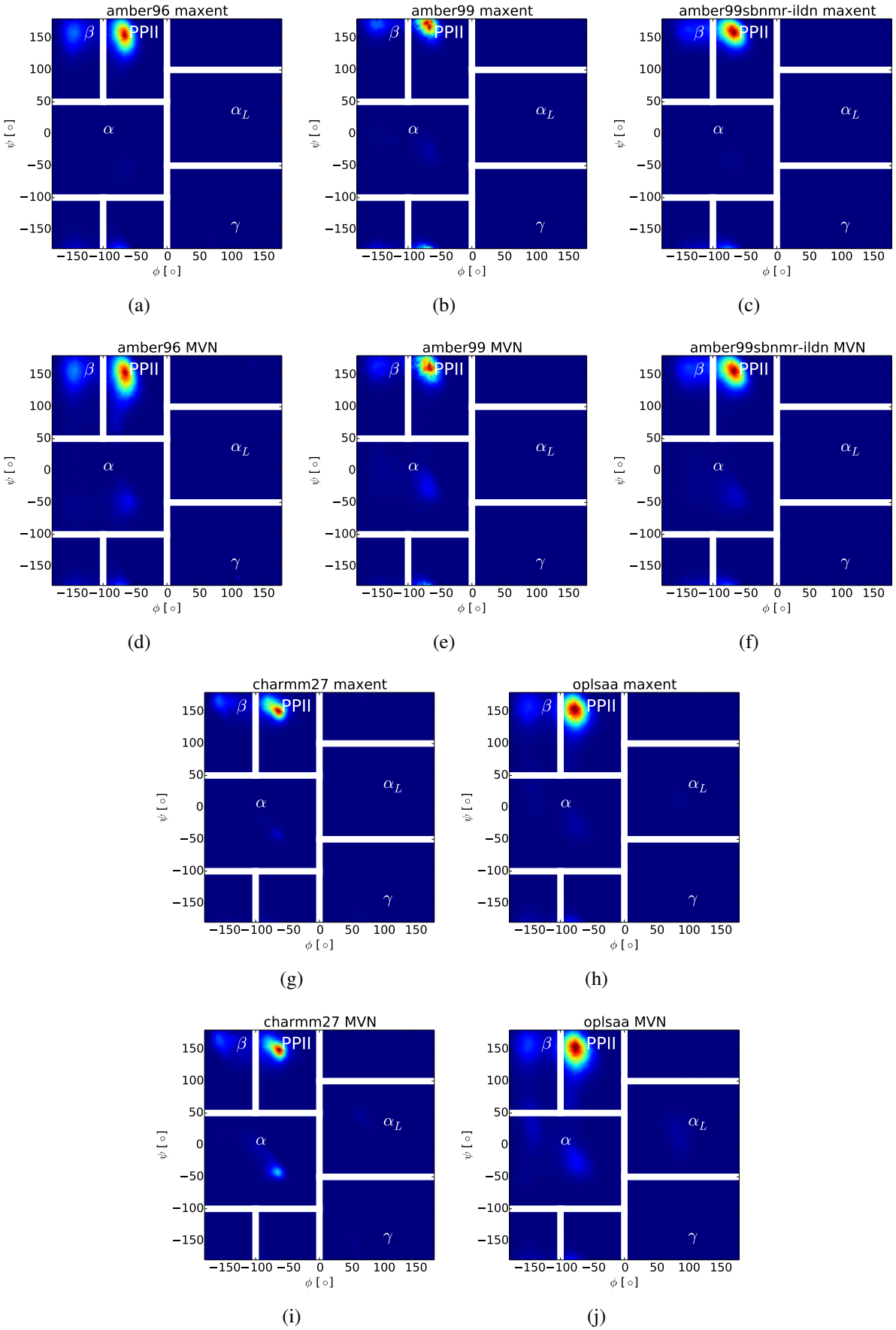


Figure S1: Ramachandran plot of BELT ensembles calculated using the maxent and MVN priors.

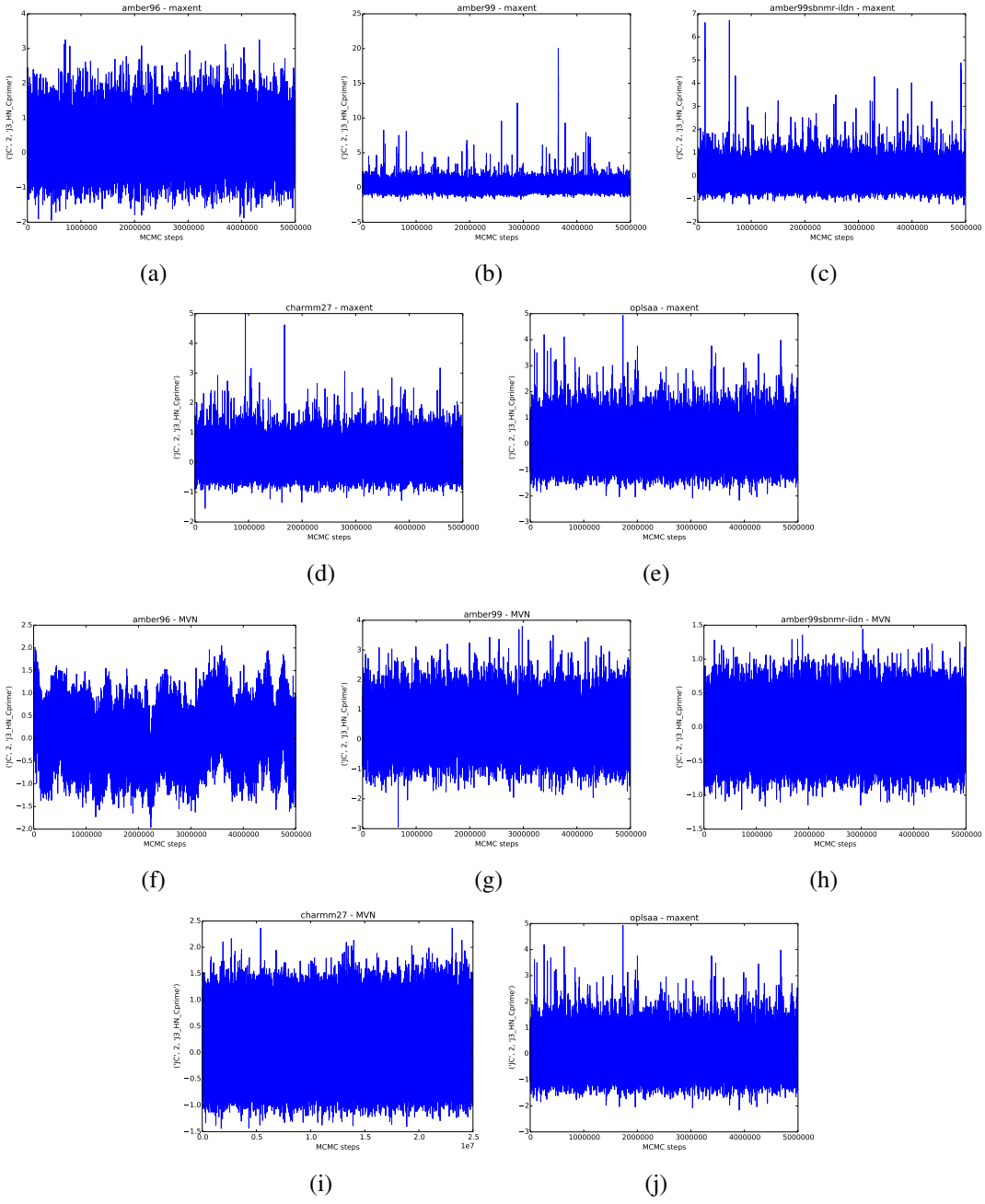
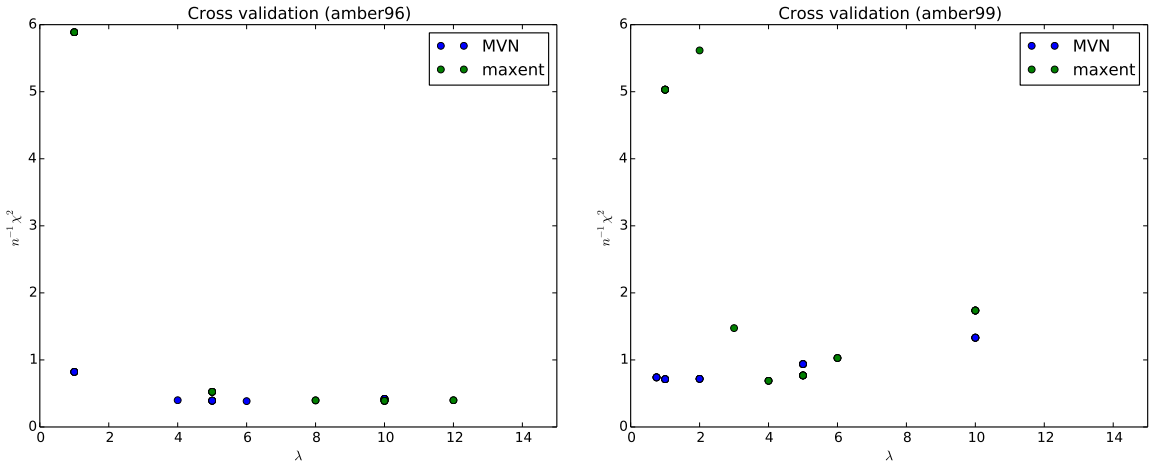
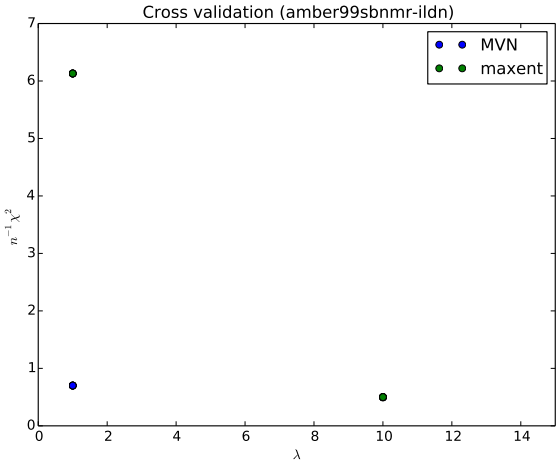


Figure S2: MCMC traces of first component of α.

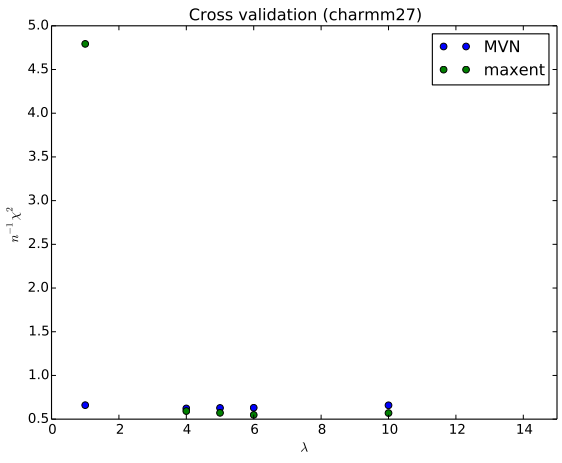


(a)

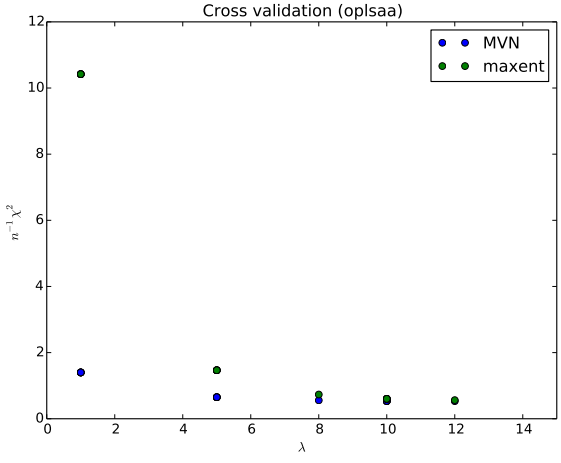
(b)



(c)



(d)



(e)

Figure S3: Cross validated reduced χ^2

Appendix S1. Connection between Maximum Entropy and BELT

Bayesian Energy Landscape Tilting generalizes a recent maximum entropy formalism[1] to include statistical uncertainty. We now show the connection between the previous maximum entropy approach[1] and BELT. The previous approach minimizes an objective function Λ; it is sufficient to consider the zeros of its gradient:

$$\frac{d\Lambda}{d\alpha_k} = -\langle f_k \rangle_\alpha + F_k$$

The obvious solution, when feasible, is to set $\langle f_i(x) \rangle_\alpha = F_i$.

In BELT, we instead sample the following log likelihood:

$$LL(\alpha) = -\sum_i \frac{1}{2\sigma_i^2} (\langle f_i \rangle_\alpha - F_i)^2 + \log P(\alpha)$$

If we assume a constant prior and maximize the likelihood, the problem becomes equivalent to setting the derivative of LL to zero:

$$\frac{dLL}{d\alpha_k} = -\sum_i \frac{1}{\sigma_i^2} (\langle f_i \rangle_\alpha - F_i) \frac{d\langle f_i \rangle_\alpha}{d\alpha_k} = 0$$

As before, if we find a value of α such that $\langle f_i(x) \rangle_\alpha = F_i$, we will maximize the log likelihood. Thus, under ideal conditions, we expect similar results using the maximum entropy approach and BELT.

Appendix S2. Derivation of Reweighting

Here we derive the population estimator used in BELT. As in the main text, we use subscripted angle brackets to indicate ensemble averages in reweighted ensembles: $\langle h(x)\rangle_\alpha$ is the ensemble average of $h(x)$ in an ensemble that is perturbed by a biasing potential $\Delta(x;\alpha) = \sum_i \alpha_i f_i(x)$:

$$\langle h(x)\rangle_\alpha = \frac{1}{Z(\alpha)} \int dx \exp[-U(x) + \sum_i \alpha_i f_i(x)] h(x)$$

$Z(\alpha)$ denotes the partition function for the α ensemble. To proceed, we first note a simple Zwanzig identity that allows us to relate samples taken from different ensembles:

$$\langle h(x)\rangle_\alpha = \frac{1}{Z(\alpha)} \int h(x) dx \exp[-U(x) - \Delta(x)] = \frac{Z(0)}{Z(\alpha)} \langle h(x) \exp[-\Delta(x;\alpha)]\rangle_0$$

In the above expression, $\langle\rangle_0$ denotes an unperturbed ensemble (e.g. $\alpha = 0$) and $Z(0)$ is the partition function of the unbiased ensemble ($\alpha = 0$). Now we sample from the unperturbed ensemble to statistically estimate the expectation

$$\langle h(x) \exp[-\Delta(x;\alpha)]\rangle_0 = \frac{1}{m} \sum_{j=1}^{m} \exp[-\Delta(x_j;\alpha)] h(x_j)$$

By letting $h(x) = 1$, we can estimate the partition function $Z(\alpha)$ up to the constant factor Z(0):

$$\frac{Z(\alpha)}{Z(0)} = \frac{1}{m} \sum_j \exp[-\Delta(x_j;\alpha)]$$

Combining these equations, we have

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

where $\pi_j(\alpha)$ give estimates of the conformation weights at a particular value of α:

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

Thus, BELT is essentially exponential averaging applied to a weighted combination of experiment-derived biasing potentials. However, the present work has introduced two key advances. First, the use of Markov chain Monte Carlo allows rigorous uncertainty analysis. Second, regularization reduces the high variance previously associated with exponential averaging.

Appendix S3. Alternative Error Models

The model presented in the main text assumes independent normal deviations between measurements and the predicted ensemble. This model is a useful approximation that leads to a straightforward χ^2 likelihood. However, in some situations, one might expect correlation between ensemble measurements. Detecting this correlation would require additional *experimental* measurements. However, it is possible to modify the χ^2 likelihood to account for correlations between the *predicted* observables. The net result is a modified log likelihood:

$$LL(\alpha) = \frac{1}{2} z^T P^{-1} z$$

where P is the correlation matrix of the observables: $P_{ij} = Cor(f_i(x), f_j(x))$ and z is the deviation between the α ensemble and the measurement, measured in units of the known uncertainty σ_i: $z_i = \frac{\langle f_i \rangle_\alpha - F_i}{\sigma_i}$. Using this model will likely lead to increased estimates of uncertainties.

Other possible error models involve modifying the assumption of normality. A normal model penalizes models by the squared deviation from the experimental measurements. However, expert knowledge may sometimes suggest different error models. For example, one could imagine a model where small deviations are not penalized at all. Such models could be inserted into the same MCMC framework with little extra effort.

Appendix S4. Choice of Prior

Maximum Entropy (maxent) Prior

Multivariate Normal (MVN) Prior

In the main text, we used both maximum entropy and multivariate normal (MVN) priors. Here we derive the parameters used in the MVN prior.

In the MVN prior, $\alpha \sim N(\mu, \Sigma)$. We let $\mu = 0$ to center the MVN around $\alpha = 0$–this places the highest prior density on the raw simulation. To pick Σ, we note that the simple choice of $\Sigma_{ij} = \delta_{ij}$ leads to a prior that depends on the units of α; this dependence on the unit system is undesirable. However, if we choose $\Sigma_{ij} = \lambda Cov(f_i(x), f_j(x))$, the units of α_i and $f_i(x)$ cancel out, leaving a result that is unit-invariant. We have also introduced a scaling factor λ to tune our relative belief in the simulation versus experiment. In practice, we found the MVN prior to give similar results to the maximum entropy prior. Occasionally, the MVN prior weights individual frames heavily, whereas the maximum entropy prior leads to much smoother ensembles.

Dirichlet Prior

We also consider the Dirichlet prior. Dirichlet priors are commonly used as conjugate priors to multinomial random variables–that is, when dealing with counts and probabilities of categorical data. The dirichlet distribution is nonzero on the unit simplex and has the following functional form:

$$f(\pi; s) = \frac{1}{B(\alpha)} \prod_j \pi_j^{s_j - 1}$$

The Dirichlet prior is an obvious choice for BELT, because the object of interest is the probability distribution on conformations. However, in BELT, we must restrict the distribution to the subset of probability distributions that can be achieved via reweighting. Thus, instead of x_j, we have $\pi_j(\alpha)$:

$$f(\pi;s) = \frac{1}{B(\alpha)} \prod_j \pi_j(\alpha)^{s_j - 1}$$

For our MCMC calculations, we work with the *log* probability:

$$\log f(\pi;s) = -\log(B(\alpha)) + \sum_j (s_j - 1) \log \pi_j(\alpha)$$

Note that the constant term is unimportant, as MCMC relies on the *difference* in *log* probabilities:

$$\log f(\pi;s) \approx \sum_j (s_j - 1) \log \pi_j(\alpha)$$

What you should notice is that this functional form is quite similar to the maxent prior that we previously discussed. The difference between the maxent and Dirichlet priors can be explained in terms of the relative entropy between two distributions P and Q. The relative entropy is given by

$$D_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$$

The relative entropy is *not* a symmetric relationship–that is, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The maxent and Dirichlet priors are simply the relative entropy between $\pi(\alpha)$ and a reference distribution $\pi(0)$, calculated in either direction.

In practice, the maxent prior has a large number of hyperparameters–the pseudocounts s_i on each conformation. To avoid the need for many hyperparameters, we assume that

$$s_j - 1 = \lambda \pi_j(0)$$

Thus, we assume that the pseudocounts are proportional to the raw MD simulation populations, which for constant temperature MD should be a uniform distribution. The final *log* prior has the form

$$\log f(\pi;s) \approx \sum_j \lambda \pi_j(0) \log \pi_j(\alpha)$$

Jeffrey's Prior

Another choice of prior would be to use the Jeffrey's prior, which is uninformative and invariant under reparameterization. We found Jeffrey's prior to be less desirable, however, because it does not necessarily place the prior maximum at $\alpha = 0$–thus, Jeffrey's prior was unable to tune our relative belief in the forcefield versus experimental data. Regardless, we derive Jeffrey's prior below.

Derivation of Jeffrey's Prior

This section derives the Jeffrey's prior for the BELT likelihood. We do *not* recommend the use of this prior, as it is expensive to compute and unable to provide regularization. This section can be skipped by most readers; we include it only for completeness.

Jeffrey's prior dictates that

$$P(\alpha) \propto \det(I(\alpha))^{\frac{1}{2}}$$

The Fisher information matrix, $I(\alpha)$, is given by

$$I_{ab}(\alpha) = E_\alpha\left(\frac{d\log P(F|\alpha)}{d\alpha_a}\frac{d\log P(F|\alpha)}{d\alpha_b}\right)$$

First, we examine the log likelihood (dropping terms independent of α) and calculate its derivative:

$$LL = \log P(F|\alpha) = -\frac{1}{2}\sum_i \left(\frac{F_i - < F_i >_\alpha}{\sigma_i}\right)^2$$

$$\frac{d(LL)}{d\alpha_a} = \sum_i \frac{1}{\sigma_i}(F_i - \langle F_i \rangle_\alpha)\frac{d\langle F_i \rangle_\alpha}{d\alpha_a}$$

When we insert this equation into the expectation, only $(F_i - \langle F_i \rangle_\alpha)$ depends on F_i. The remaining terms can be pulled outside the expectation:

$$I_{ab} = \sum_{ij} \frac{d\langle F_i \rangle_\alpha}{d\alpha_a}\frac{d\langle F_j \rangle_\alpha}{d\alpha_b} E\left(\frac{1}{\sigma_i \sigma_j}(F_i - \langle F_i \rangle_\alpha)(F_j - \langle F_j \rangle_\alpha)\right)$$

Because the conditional likelihood is a diagonal multivariate normal, the expectation is simply δ_{ij}, leading to

$$I_{ab} = \sum_i \frac{d\langle F_i \rangle_\alpha}{d\alpha_a}\frac{d\langle F_i \rangle_\alpha}{d\alpha_b}$$

Now, we know that

$$\frac{d\langle F_i \rangle_\alpha}{d\alpha_a} = \sum_k f_{ak}\frac{d\pi_k}{d\alpha_a}$$

Similarly, we can show that

$$\frac{d\pi_k}{d\alpha_a} = \pi_k(\langle F_a \rangle_\alpha - f_{ak})$$

Putting all this together, we can show that

$$I = S^T S$$

Where

$$S_{ia} = \sum_k \pi_k(< F_a >_\alpha - f_{ka}) f_{ki}$$

Appendix S5. Determining Prior Strength Via Cross-Validation

The maxent and MVN priors both contain a single free parameter, λ, which controls the relative weight of simulation and experiment. At least three different approaches can help select an appropriate value of λ:

1. Cross validation on simulation data (used in main text)
2. Cross validation on experimental data
3. χ^2 analysis.

Cross validation on simulation data

We first discuss cross-validating on the simulation data. The underlying idea is that too little regularization ($\lambda = 0$) leads to models that overfit the available simulation data and generalize poorly–that is, repeating or extending the MD simulations would lead to a different result. At the other extreme, underfit models ($\lambda = \infty$) will simply report the unbiased simulations, leading to poor agreement with experiment. To perform this form of cross-validation, first separate the simulation data into several independent subsets. Mark one subset as the "test" set and fit the model on the remaining data (the "training" set). The χ^2 score is evaluated on the test data. We then repeat the process, letting the test set be equal to each of the other subsets. The final χ^2 square is averaged over each of these iterations. The value of λ is chosen to minimize the test set error.

When using MD to generate conformations, you must perform cross-validation using uncorrelated subsets of the data. This precludes the typical standard cross-validation approach that uses randomly selected subsets of your data–randomly selected folds will be tainted by correlation between the folds. As a thought experiment, suppose you do cross validation by dividing your trajectory into even and odd frames. Because of time-correlation in the data, the even and odd subsets will essentially contain the same information–ruining the cross-validation. To avoid these perilous correlations, we recommend that you split the trajectory into time-contiguous blocks. For

the present work, we divided each trajectory into two halves.

There are some situation where cross validation leads to excessive amounts of regularization. Suppose, for example, that your simulation was performed to exactly match experiment. Then cross validation will recommend infinitely strong regularization ($\lambda = \infty$). This is because infinite regularization damps the statistical fluctuations in α. Thus, in this extreme limit, the "optimal" model is the raw simulation. In this situation, the χ^2 likelihood is dwarfed by the regularization, resulting in ensembles with zero statistical uncertainty. Such a situation occurred with the amber99sbnmr-ildn force fields. Rather than use $\lambda = \infty$, we selected a moderate value of λ based on the values of the other forcefields.

Cross validation on experimental data

Cross validating on experimental data instead sets aside experimental measurements that can then be used to evaluate model quality. One key difficulty with this approach, however, is that experimental datasets are often sparse–that is, there are often only few information-rich measurements. This can lead to difficulties defining meaningful training and test sets.

χ^2 analysis

A heuristic approach to selecting λ is to choose the smallest value of λ such that $\frac{\chi^2}{n} \approx 1$; models with $\frac{\chi^2}{n} < 1$ are essentially fitting experimental noise. This approach is computationally trivial, as it requires no additional computation. However, it relies on having accurate estimates of uncertainty for each experiment.

Cross Validation Results

Here we summarize the values of λ used in this work. These values were determined by cross-validating on the simulation data. For the amber99sbnmr-ildn forcefield, we selected a moderate amount of regularization ($\lambda = 5.0$) (see above).

	λ	
prior	MVN	maxent
forcefield		
amber96	7.0	5.0
amber99	0.5	1.0
amber99sbnmr-ildn	5.0	5.0
charmm27	1.0	0.9
oplsaa	11.0	7.0

The corresponding cross-validated reduced χ^2 scores are given below. These scores were generated using the *training* set of experimental measurements, but done in the setting of cross-validation on the simulation data. Thus, the models were fit to *half* the trajectory data and evaluated on the other half. As before, we see similar performance with the maxent and MVN priors. Full sweeps of λ are depicted in Fig. SS3.

	$\frac{1}{n}\chi^2$ (cross-validated)	
prior	MVN	maxent
forcefield		
amber96	0.22	0.22
amber99	0.38	0.38
amber99sbnmr-ildn	0.31	0.31
charmm27	0.37	0.35
oplsaa	0.33	0.37

Appendix S6. Bayesian Bootstrapping

The BELT model presented in the main text does not directly model simulation uncertainty. This effect, however, can be introduced using a resampling technique known as Bayesian bootstrapping.[2] In Bayesian bootstrapping, every data point (e.g. conformation) is associated with a Dirichlet random variable that models the effect of resampling the given data points. In effect, each conformation is given a "prior" population that is allowed to fluctuate around its average value of $\frac{1}{n}$.

One additional complication arises when using molecular dynamics simulations, which produce a correlated time series. Because of this, it is not sufficient to simply use a Dirichlet whose dimension is the same as the number of snapshots–such a procedure will significantly underestimate uncertainties due to correlation between frames. Instead, one must first divide the trajectory into independent blocks. The Dirichlet random variable is then chosen to sample the relative weights of each of the independent blocks. Choosing the length of each block can be done by applying Bayesian bootstrapping to the un-reweighted trajectory. Given some observable of interest, O, one calculates $O(B)$ for a sequence of block lengths, choosing the value of B that maximizes the estimated uncertainty of O. The block length could also be calculated using other blocking methods[3] or by statistical inefficiency analysis.[4]

In practice, applying Bayesian bootstrapping involves repeating several BELT calculations using different values of "prior" conformational populations that were drawn from a Dirichlet random variable. The MCMC traces of each run are then pooled.

Appendix S7. Convergence Analysis and MCMC Parameters

Although more sophisticated convergence tests are available, we evaluated convergence of MCMC traces by visual analysis. A properly sampled and thinned model will appear similar to white noise, as we observed in the following traces. A few interesting features are worth noting. The charmm27 and amber99 forcefields with MVN prior seem to suffer from increased correlation in their MCMC traces.

Based on this and our other experience, we offer some suggestions for achieving converged traces. First, it seems that the maxent prior is better able to achieve independent MCMC samples than the MVN prior. Second, poorer force fields (e.g. amber99 and charmm27) seem more prone to correlated MCMC samples. This is likely because the sampler is forced to explore "extreme" models–that is, models that lie further from the raw forcefield. Finally, we find that adding additional measurements–particularly ones that are correlated to previous measurements–leads to increased correlation within the MCMC traces. We think these observations should help guide users towards achieving convergence without excessive computational resources.

Appendix S8. Data Curation

Because the BELT log likelihood weights errors quadratically, it is vital to use the highest quality experimental measurements and predictions. We recommend that users manually inspect *all* measured and predicted observables before performing BELT analysis. We discuss two examples we encountered in the current analysis.

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

Another example of data quality arose in the BPTI analysis. It has previously been observed that different chemical shift models give significantly different predictions for the millisecond simulation of BPTI.[5] Modern chemical shift prediction tools (e.g. PPM, SPARTA+, ShiftX2) are reported to have similar accuracy. We therefore chose to use the average predictions of the three chemical shift models; this averaging procedure reduces the statistical uncertainty of individual chemical shift predictions.

References

(1) Pitera, J.; Chodera, J. *J. Chem. Theory Comput.* **2012**,

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

(3) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461.

(4) Shirts, M.; Chodera, J. *J. Chem. Phys.* **2008**, *129*, 124105.

(5) Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. *J. Am. Chem. Soc.* **2012**, *134*, 2555–2562.