Automatic Forward Model Parameterization with Bayesian Inference of Conformational Populations

Robert M. Raddi, Tim Marshall, and Vincent A. Voelz

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

(*Electronic mail: vvoelz@temple.edu)

(Dated: May 30, 2024)

To quantify how well theoretical predictions of structural ensembles agree with experimental measurements, we depend on the accuracy of forward models. These models are computational frameworks that generate observable quantities from molecular configurations based on empirical relationships linking specific molecular properties to experimental measurements. Bayesian Inference of Conformational Populations (BICePs) is a reweighting algorithm that reconciles simulated ensembles with ensemble-averaged experimental observations, even when such observations are sparse and/or noisy. This is achieved by sampling the posterior distribution of conformational populations under experimental restraints as well as sampling the posterior distribution of uncertainties due to random and systematic error. In this study, we enhance the algorithm for the refinement of empirical forward model (FM) parameters. We introduce and evaluate two novel methods for optimizing FM parameters. The first method treats FM parameters as nuisance parameters, integrating over them in the full posterior distribution. The second method employs variational minimization of a quantity called the BICePs score that reports the free energy of "turning on" the experimental restraints. This technique, coupled with improved likelihood functions for handling experimental outliers, facilitates force field validation and optimization, as illustrated in recent studies (Raddi et al. 2023, 2024). Using this approach, we refine parameters that modulate the Karplus relation, crucial for accurate predictions of J-coupling constants based on dihedral angles (ϕ) between interacting nuclei. We validate this approach first with a toy model system, and then for human ubiquitin, predicting six sets of Karplus parameters for $^3J_{H^N H\alpha}$, $^3J_{H\alpha C'}$, $^3J_{H^N C\beta}$, $^3J_{H^N C'}$, $^3J_{C'C\beta}$, $^3J_{C'C'}$. This approach, which does not rely on any predetermined parameterization, enhances predictive accuracy and can be used for many applications.

I. INTRODUCTION

In the field of molecular modeling and dynamics, the accuracy of theoretical predictions that reflect real-world observations is crucial. Quantifying the agreement between theory and experiment is highly dependent on the accuracy of forward models—computational frameworks that predict observable quantities from molecular configurations. These models often depend on empirical relationships that link specific molecular properties to experimental measurements.

Model validation and refinement of structural ensembles against NMR observables critically depends on reliable forward models (FMs) that have been robustly parameterized, so that FM error is minimal in the validation/refinement process. An important challenge in the parameterization of FMs is presented by random and systematic errors inherent to the experimental data. These errors need to be considered in the comparison and integration of experimental data with computational models for objective model selection and accurate uncertainty representation.

A further challenge is presented by missing or insufficient examples of known structures than can be used to train forward models. For NMR observables that depend on backbone ϕ-angles, such as J-coupling constants, the reference data from X-ray crystallography may be missing or dynamically averaged, creating large uncertainties in the correct ϕ-angles. Numerous approaches[1–4] have been developed to address some of these challenges. Some algorithms rely heavily on X-ray crystal structure data; others have many hyperparameters that need to be determined.

To address these challenges, we extend the Bayesian Inference of Conformational Populations (BICePs) algorithm[5,6] to refine FM parameters. BICePs, a reweighting algorithm, refines structural ensembles against sparse and/or noisy experimental observables, and has been used in many previous applications.[7–10] BICePs infers all possible sources of error by sampling the posterior distribution of these parameters directly from the data through MCMC sampling BICePs also computes a free energy-like quantity called the BICePs score that can be used for model selection and model parameterization.[6,11,12]

Recently, BICePs was enhanced with a replica-averaging forward model, making it a maximum-entropy (MaxEnt) reweighting method, and unique in that no adjustable regularization parameters are required to balance experimental information with the prior.[6] With this new approach, the BICePs score becomes a powerful objective function to parameterize optimal models. Here, we show that the BICePs score, which reflects the total evidence for a model, can be used for variational optimization of FM parameters. The BICePs score contains a form of inherent regularization, and has specialized likelihood functions that allow for the automatic detection and down-weighting of the importance of experimental observables subject to systematic error.[6]

To effectively refine FM parameters, we sample over the full posterior distribution of FM parameters. Through this approach, BICePs performs ensemble reweighting and FM parameter refinement simultaneously. Additionally, we show that by variational minimization of the BICePs score, we obtain the same result and show that the two approaches are equivalent, with each method requiring particular considerations. We first demonstrate our method's effectiveness on a toy model system, and then optimize six distinct sets of Karplus parameters for the human protein ubiquitin, and compare our findings with previously established results. Through this, we aim to showcase a systematic and robust approach to enhancing the accuracy of theoretical predictions, thereby bridging the gap between computational

models and experimental observations.

II. THEORY

Posterior sampling of forward model parameters gives reliable parameter uncertainties. BICePs uses a Bayesian statistical framework, inspired by Inferential Structure Determination (ISD)[13], to model the posterior distribution $p(X, \sigma)$, for conformational states X, and nuisance parameters σ, which characterize the extent of uncertainty in the experimental observables D:

$$p(X, \sigma | D) \propto p(D|X, \sigma)p(X)p(\sigma). \tag{1}$$

Here, $p(D|X, \sigma)$ is a likelihood function that uses a forward model to enforce the experimental restraints, $p(X)$ is a prior distribution of conformational populations from some theoretical model, and $p(\sigma) \sim \sigma^{-1}$ is a non-informative Jeffrey's prior.

We now consider a specific forward model $g(X, \theta)$ with a set of FM parameters θ that we wish to additionally include in the posterior,

$$p(X, \sigma, \theta | D) \propto p(D|X, \sigma, \theta)p(X)p(\sigma)p(\theta) \tag{2}$$

Replica-averaging. When BICePs is equipped with a replica-averaged forward model, it becomes a MaxEnt reweighting method in the limit of large numbers of replicas[14–19]. Consider a set of N replicas, $\mathbf{X} = \{X_r\}$, where X_r is the conformational state being sampled by replica r. To compare the sampled replicas with ensemble-averaged experimental observables, we define a replica-averaged forward model $g(\mathbf{X}, \theta) = \frac{1}{N}\sum_r^N g(X_r, \theta)$. This quantity is an estimator of the true ensemble-average, with an error due to finite sampling for observable j estimated using standard error of the mean (SEM):[14,19] $\sigma_j^{SEM} = \sqrt{\frac{1}{N}\sum_r^N (g_j(X_r, \theta) - \langle g_j(\mathbf{X}, \theta)\rangle)^2}$. Thus, σ_j^{SEM} decreases as the square root of the number of replicas.

In the scenario that observables can be collected into different types, e.g., a particular type of vicinal *J*-coupling, then each collection can be described with its own set of parameters and error distribution. For K distinct sets of FM parameters $\theta = \{\theta_k\}$, the joint posterior distribution for all parameters is

$$p(X, \sigma, \theta | D) \propto \prod_{r=1}^{N} p(X_r) \prod_{k=1}^{K} p(D_k|g(\mathbf{X}, \theta_k), \sigma_k)p(\sigma_k)p(\theta_k) \tag{3}$$

where $\mathbf{X}$ is a set of N conformation replicas, and θ_k is the kth set of FM parameters. The kth set has an uncertainty parameter $\sigma_k = \sqrt{(\sigma_k^{SEM})^2 + (\sigma_k^B)^2}$, that describes the total error, arising from both finite sampling $(\sigma_k^{SEM})^2$, and uncertainty in the experimental measurements, known as a Bayesian uncertainty parameter σ_k^B. The prior distribution of uncertainties $p(\sigma_k)$ is treated as a non-informative Jeffrey's prior (σ_k^{-1}) for each collection of observables, and the posterior of FM parameters $p(\theta|D)$ is recovered by marginalization over all X and σ:

$$p(\theta|D) = \sum_X \int p(X, \sigma, \theta | D)d\sigma \tag{4}$$

Gradients speed up convergence. In our methodology, Markov chain Monte Carlo (MCMC) is used to sample the posterior with acceptances following the Metropolis-Hastings (M-H) criterion. Our algorithm can be used with or without gradients. However, significantly faster convergence, especially in higher dimensions, is achieved through an integration of stochastic gradient descent approach. Our gradient descent approach allows for informed updates to the FM parameters, incorporating stochastic noise to facilitate the escape from local minima and enhance exploration of the parameter space.

The update mechanism is succinctly encapsulated in the equation:

$$\theta_{\text{trial}} = \theta_{\text{old}} - l_{\text{rate}} \cdot \nabla u + \eta \cdot \mathscr{N}(0, 1) \tag{5}$$

where θ_{trial} and θ_{old} denote the trial parameters and previous parameters, respectively. The learning rate is denoted by l_{rate}, ∇u signifies the computed gradient of BICePs energy function with respect to the parameters θ, and η scales the noise drawn from a standard normal distribution $\mathscr{N}(0, 1)$.

This strategic parameter update protocol is designed to satisfy the M-H criterion, ensuring that each step in the parameter space not only moves towards minimizing the energy of the forward model but also adheres to the probabilistic acceptance of potentially non-optimal moves to avoid local optima traps. Ergodic sampling is ensured by "turning off" the gradient after burn-in. The sampling procedure involves: (1) acquiring derivatives of the FM parameters, (2) perturbing these parameters based on the derived information, (3) predict observables using perturbed FM parameters and compute the total energy, and (4) assessing the new energy against the previous to determine acceptance based on the M-H criterion. This ensures a thorough and effective search of the parameter space, leveraging both the landscape topology and stochastic elements to guide the exploration.

The Good-Bad model accounts for systematic error due to outlier measurements BICePs now is equipped with sophisticated likelihood models that are extremely robust in the presence of systematic error[6]. Recently, we demonstrated the ability of the Student's model to account for systematic error for force field optimization[12]. In this work, we use a likelihood function called the Good-Bad model to demonstrate the validity of forward model refinement. The derivatives of the Good-Bad model are far less complicated than the Student's model.

The Good-Bad likelihood model[6] assumes 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 $p(\sigma_j|\sigma^B)$. We derive a posterior for the kth parameter set having a single uncertainty parameter σ^B by marginalizing over all σ_j

$$p(\mathbf{X}, \sigma_0, \theta_k | D) \propto \prod_{r=1}^{N} p(X_r) \prod_{j=1}^{N_d} \int_{\sigma^{SEM}}^{\infty} p(d_j|g_j(\mathbf{X}, \theta_k), \sigma_j)p(\sigma_j|\sigma_0)d\sigma_j \tag{6}$$

where $\sigma_0 = \sqrt{(\sigma^B)^2 + (\sigma^{SEM})^2}$. 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 $\varphi^2\sigma_0^2$, where $\varphi \geq 1$.

By this assignment, $p(\sigma_j|\sigma_0)$ from equation 6 becomes

$$p(\sigma_j|\sigma_0, \omega, \varphi) = \omega\delta(\sigma_j - \varphi\sigma_0) + (1-\omega)\delta(\sigma_j - \sigma_0) \tag{7}$$

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

$$
p(\mathbf{X},\sigma_0,\varphi,\theta_k|D) \propto \prod_{r=1}^{N}\left\{ p(X_r)\prod_{j=1}^{N_d}\int_0^1 d\omega \int_{\sigma^{SEM}}^{\infty} \exp\left(-\frac{(d_j - g_j(\mathbf{X},\theta_k))^2}{2\sigma_j^2}\right)\frac{\omega\delta(\sigma_j - \varphi\sigma_0) + (1-\omega)\delta(\sigma_j - \sigma_0)}{\sqrt{2\pi}\sigma_j}d\sigma_j \right\}
$$
$$
= \prod_{r=1}^{N}\left\{ p(X_r)\prod_{j=1}^{N_d}\left(\frac{\left(1 - H\left(\sigma^{SEM} - \sigma_0\right)\right)}{2\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(d_j - g_j(\mathbf{X},\theta_k))^2}{2\sigma_0^2}\right) + \frac{\left(1 - H\left(\sigma^{SEM} - \varphi\sigma_0\right)\right)}{2\varphi\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(d_j - g_j(\mathbf{X},\theta_k))^2}{2\varphi^2\sigma_0^2}\right)\right)\right\},
$$
$$\tag{8}$$

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 $\varphi = 1$, the model reverts to a Gaussian likelihood model. When considering the full posterior, this extra nuisance parameter is given a non-informative Jeffrey's prior, $p(\varphi) \sim \varphi^{-1}$.

For a single set of FM parameters (for simplicity), the BICePs energy function, $u = -\log p(\mathbf{X},\sigma_0,\varphi,\theta|D)$, the negative logarithm of the posterior in its full form is given by

$$
u = \sum_{r=1}^{N} -\log(p(X_r)) - N\sum_{j=1}^{N_d}\log\left[\frac{\left(1 - H\left(\sigma^{SEM} - \sigma_0\right)\right)}{2\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(d_j - g_j(\mathbf{X},\theta))^2}{2\sigma_0^2}\right) + \frac{\left(1 - H\left(\sigma^{SEM} - \varphi\sigma_0\right)\right)}{2\varphi\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(d_j - g_j(\mathbf{X},\theta))^2}{2\varphi^2\sigma_0^2}\right)\right],
$$
$$\tag{9}$$

and when $\varphi = 1$ our energy function becomes

$$
u = \sum_{r=1}^{N} -\log(p(X_r)) + N\left[\sum_{j=1}^{N_d} -\log\left(\frac{1}{\sqrt{2\pi}\sigma_j}\right) + \frac{(d_j - g_j(\mathbf{X},\theta))^2}{2\sigma_j^2} - \log(p(\sigma_j))\right]. \tag{10}
$$

The first derivative of equation 9 with respect to the ith FM parameter θ_i is

$$
\frac{\partial u}{\partial \theta_i} = N\sum_{j=1}^{N_d}\frac{\partial g_j(X,\theta)}{\partial \theta_i}\frac{(d_j - g_j(X,\theta))}{\varphi^2\sigma_0^2}\frac{\left\{\varphi^3\left(1 - H\left(\sigma^{SEM} - \sigma_0\right)\right)\exp\left(\frac{(d_j - g_j(X,\theta))^2}{2\varphi^2\sigma_0^2}\right) + \left(1 - H\left(-\varphi\sigma_0 + \sigma^{SEM}\right)\right)\exp\left(\frac{(d_j - g_j(X,\theta))^2}{2\sigma_0^2}\right)\right\}}{\left\{\varphi\left(H\left(\sigma^{SEM} - \sigma_0\right) - 1\right)\exp\left(\frac{(d_j - g_j(X,\theta))^2}{2\varphi^2\sigma_0^2}\right) + \left(H\left(-\varphi\sigma_0 + \sigma^{SEM}\right) - 1\right)\exp\left(\frac{(d_j - g_j(X,\theta))^2}{2\sigma_0^2}\right)\right\}},
$$
$$\tag{11}
$$

and in the case of $\varphi = 1$ the gradient becomes

$$
\frac{\partial u}{\partial \theta_i} = -N\left[\sum_{j=1}^{N_d}\frac{\partial g_j(X,\theta)}{\partial \theta_i}\frac{(d_j - g_j(\mathbf{X},\theta))}{\sigma_j^2}\right]. \tag{12}
$$

Second derivatives of the BICePs energy function and the BICePS score are useful for descent and uncertainty quantification using other forward models. We refrain from writing out the second derivative here, since the specific class of forward models we consider below all have second derivatives that go to zero. For more general cases, see Appendix A for more details. The energy of the Good-Bad likelihood model and its first and second derivatives are shown in Figure S1.

III. RESULTS/DISCUSSION

Testing algorithm performance on a toy model

To investigate the efficacy of BICePs for this optimization problem, we introduce a simplified, yet comprehensive toy model. This model is designed to mimic the complexity of protein structure elements by generating ϕ-angles from a multi-modal distribution, thereby emulating configurations characteristic of different secondary structure elements (Figure 1). This distribution encompasses three distinct modes, each characterized by a mean (μ), standard deviation (σ), and weight (w): beta sheets ($\mu = -110°$, $\sigma = 20°$, $w = 0.35$), right-handed helices ($\mu = -60°$, $\sigma = 10°$, $w = 0.5$), and left-handed helices ($\mu = 60°$, $\sigma = 5°$, $w = 0.15$). These parameters were chosen to accurately

reflect the structural variability found in proteins. Angles ϕ_i were sampled from the multi-modal distribution,

$$p(\phi|\mu,\sigma) = \sum_l w_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(-\frac{(\phi-\mu_l)^2}{2\sigma_l^2}\right). \quad (13)$$

The sampled ϕ_i were then used to calculate experimental J-coupling constants $J(\phi)$ using the Karplus relation with the *true* Karplus coefficients (A^*, B^*, C^*).

$$^3J(\phi) = A\cos^2(\phi) + B\cos(\phi) + C \quad (14)$$

Synthetic experimental J-coupling data is generated to represent a mixture of all conformational states, $d_j^{Exp} = \sum_X {}^3J(\phi_{X,j}) \cdot p(X)$, with FM parameters $\theta = \{A,B,C\}$ set to their true values ($\{A = 6.51, B = -1.76, C = 1.6\}$). The initial forward model data is generated using reference Karplus parameters (A_0, B_0, C_0) and refined through the optimization process to showcase the algorithm's adaptability and precision in parameter estimation.
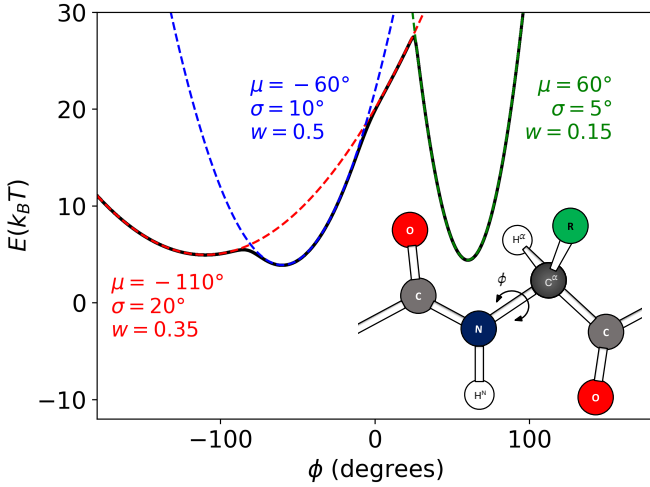

Figure 1. **A versatile toy model for measuring the performance of forward model optimization.** The ϕ-angles for each conformational state is pulled from a multi-modal distribution and corresponding energies. (a) This multi-modal distribution of ϕ-angles was intended to represent configurations with different secondary structure elements having three distinct modes described by the mean (μ), standard deviation (σ) and weight (w): beta sheets ($\mu = -110°$, $\sigma = 20°$, $w = 0.35$), right-handed helices ($\mu = -60°$, $\sigma = 10°$, $w = 0.5$), and left-handed helices ($\mu = 60°$, $\sigma = 5°$, $w = 0.15$). (b) Cartoon representation of the backbone torsion angle, ϕ.

BICePs robustly finds optimal Karplus parameters in the presence of experimental errors.

To evaluate the resilience of our algorithm against experimental inaccuracies, we introduced random and systematic errors of varying magnitudes (σ_{data}) into the synthetic experimental scalar couplings. The performance of our Good-Bad likelihood model, a Gaussian likelihood model, and singular value decomposition (SVD) was compared under these conditions.

SVD calculations. Using methods similar to previous efforts by others,[20] we derived the Karplus parameters $\theta = \{A,B,C\}$ using a weighted singular value decomposition (SVD) fitting approach to optimally fit the J-coupling values as a function of dihedral angles. For each observation j across N_d measurements, the matrix M was constructed with rows for each ϕ angle:

$$M = \begin{bmatrix} \sum_X p(X)\cos^2(\phi_{1,X}+\phi_0) & \sum_X p(X)\cos(\phi_{1,X}+\phi_0) & 1 \\ \sum_X p(X)\cos^2(\phi_{2,X}+\phi_0) & \sum_X p(X)\cos(\phi_{2,X}+\phi_0) & 1 \\ \vdots & \vdots & \vdots \\ \sum_X p(X)\cos^2(\phi_{N_d,X}+\phi_0) & \sum_X p(X)\cos(\phi_{N_d,X}+\phi_0) & 1 \end{bmatrix} \quad (15)$$

where $p(X)$ represents the true populations for state X, and ϕ_0 is the phase shift of $-60°$.

SVD was applied to decompose the matrix as $M = U\Sigma V^T$, and Karplus coefficients were derived using:

$$\theta = V^T(\Sigma + \varepsilon I)^{-1}U^T J^{\text{exp}}, \quad (16)$$

where $\varepsilon = 1e-6$ a small regularization term added to the diagonal of Σ to ensure stability of the pseudo-inverse, and J^{exp} represents the vector of experimental J-coupling values. This method ensures robust estimation of θ under ideal experimental conditions, given the true conformational populations. In practice, the true populations are not known *a priori*. The uncertainty in SVD coefficients was determined through 1k iterations of fitting, each omitting 10% of the data points chosen at random.

Typical uncertainties in NMR frequency measurements range from 0.1 to 1.0 Hz, primarily influenced by magnetic field strength, instrument quality, sample conditions, and the specifics of the pulse sequence used. In these experiments, 100 conformational states and 60 synthetic experimental scalar couplings were used. We introduced systematic error by shifting the experimental ^3J values by +2.0 Hz to +4.0 Hz for up to 20% of the data points. BICePs calculations were performed by averaging FM parameters over three chains of MCMC stating from different initial parameters ($\{A = 9, B = -1, C = 1\}, \{A = 4, B = 0, C = 3\}, \{A = 0, B = 0, C = 0\}$). Regardless of different starting parameters, posterior sampling universally converges to "true" optimal FM parameters. In these calculations, we used 32 BICePs replicas, and burned 10k steps followed by 50k steps of MCMC sampling.

We evaluated model performance by the root-mean-square error (RMSE) between the true J-coupling values with parameters $\{A^* = 6.51, B^* = -1.76, C^* = 1.6\}$ and the J-coupling values using predicted Karplus coefficients for all 60 synthetic measurements, performed over 1k independent trials of random generations of toy model data. Average RMSE results, computed over 100 BICePs calculations, highlight the algorithm's robustness and its ability to accurately predict FM parameters even in the presence of data perturbations. Error bars in our results represent the standard deviation across these calculations, providing a comprehensive measure of the algorithm's reliability under various experimental accuracy.

Our findings indicate that the Good-Bad likelihood model (red) exhibits superior resilience to experimental errors compared to a traditional Gaussian likelihood model (blue) and SVD (green) approaches (Figure 2). Predictions from SVD and the Gaussian likelihood model become notably less dependable when data incorporates errors, especially when σ_{data} exceeds 0.5

Hz. On average, error in predictions (RMSE) from the Good-Bad model does not exceed 0.1 Hz over the full range of σ_{data}.
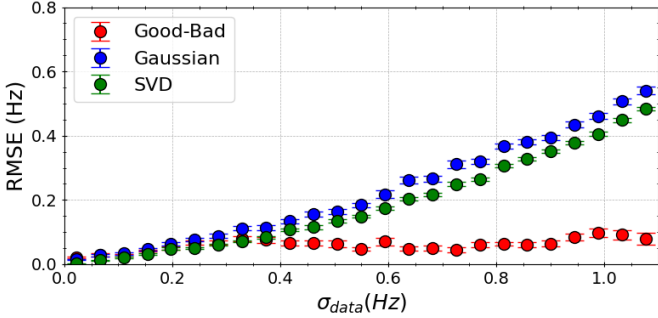

Figure 2. Comparative analysis in performance of the Good-Bad likelihood model (red), a Gaussian likelihood model (blue), and singular value decomposition (SVD) using the "true" ϕ angles with synthetic experimental data. Here, we induced random and systematic error of varying magnitude (σ_{data}) to the experimental scalar couplings. Model performance was measured by computing RMSE (Hz) between the "true" scalar couplings and the couplings generated from the Karplus relations with predicted Karplus coefficients over 1,500 random perturbations to the experimental data, and represent the average of 100 BICePs calculations. Error bars represent the standard deviation. Predictions from SVD and the Gaussian likelihood model become notably less dependable when data incorporates errors, especially when σ_{data} exceeds 0.5 Hz.

An example of a single trial of forward model parameter refinement using the toy model is shown in Figure S2, where BICePs predicts Karplus coefficients by posterior sampling over FM parameters. Both BICePs and Singular Value Decomposition (SVD) methods successfully reproduce the "true" Karplus curve. However, BICePs excels by accurately identifying the error present in the data ($\sigma_{data} = 0.471$), as indicated in the marginal posterior of uncertainty $p(\sigma_J)$. The BICePs predicted maximum a posteriori uncertainty was found to be $\sigma_J = 0.272$ with a variance scaling parameter of $\varphi_J = 1.98$. The marginal posterior distributions of FM parameters for the Good-Bad model were $\{A = 6.6 \pm 0.04, B = -1.8 \pm 0.02, C = 1.5 \pm 0.03\}$, and for SVD, $\{A = 6.11 \pm 0.06, B = -1.63 \pm 0.04, C = 1.80 \pm 0.04\}$.

In addition to the Good-bad model, we refined parameters using the Student's model ($\{A = 6.8 \pm 0.03, B = -1.9 \pm 0.03, C = 1.4 \pm 0.03\}$) to demonstrate that the Student's model yields similar performance (Figure S3). The computed Gelman-Rubin ($\hat{R}$) statistic for these calculations was found to be $\hat{R} = 1.01$ for each of the marginal posterior distributions of Karplus coefficients, which demonstrates that our chains converge to the same parameter location with similar variance.

Furthermore, we assessed model performance across varying qualities of prior structural ensembles as illustrated in Figure S4. By introducing varying levels of prior error σ_{prior} (measured in degrees) through perturbations to the "true" ϕ angles, even in the presence of random and systematic error, we observed strong correlation between the BICePs score and the quality of the structural ensemble, with a coefficient of determination R^2 of 0.99. For these calculations, we employed the Good-Bad model, utilizing 32 replicas, and conducted 1,000 random perturbations

to the ϕ angles with errors up to $\sigma_{prior} = 4°$, and perturbations to the experimental data $\sigma_{data} = 0.68 \pm 0.24$ Hz. Karplus's warning about the perils of precise angle estimation[21] underscores our approach's necessity and performance in error aware modeling in structural biology.

The comprehensive evaluation of our algorithm with this toy model underscores its efficacy in accurately determining FM parameters, reflecting scenarios commonly encountered in real-world applications. The robust performance of the algorithm, even in the face of random and systematic errors, can be attributed to BICePs' sophisticated error-handling within its likelihood models. This approach also ensures that predicted FM parameters derived from sub-optimal structural ensembles remain reliable. Additionally, our findings reveal a strong correlation between the BICePs score and the quality of the structural ensemble, demonstrating an immense utility in this context.

Variational minimization of the BICePs score to find optimal parameters.

Treating the forward model parameters as nuisance parameters, and sampling over them with the full posterior is an efficient strategy that grants the ability to include all sources of error while refining the structural ensemble with FM parameters. However, in the limit of large number of FM parameters, the dimensionality of the posterior may ultimately become unwieldy and present the curse of dimensionality. Here, we introduce an alternative strategy for refining FM parameters that has previously demonstrated to be a viable approach to automated force field optimization[12].

In this approach, the FM parameters are no longer part of the joint posterior density. Instead, the posterior is conditioned on the set of FM parameters θ, that is, equation 2 becomes

$$p(X, \sigma | D, \theta) \propto p(D, \theta | X, \sigma) p(X) p(\sigma) \qquad (17)$$

In this view, ensemble refinement is performed with a static set of FM parameters for each BICePs calculation.

BICePs evaluates model quality by calculating a free energy-like quantity called the BICePs score. For a forward model with parameters θ, the BICePs score $f(\theta)$ 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,

$$f(\theta) = -\ln\left(Z(\theta)/Z_0\right), \qquad (18)$$

where

$$Z(\theta) = \iint \exp\left(-u(\mathbf{X}, \sigma \mid D, \theta)\right) d\mathbf{X} d\sigma \qquad (19)$$

is the evidence for FM parameters θ, Z_0 is the evidence for a suitable reference state, and u is the unchanged BICePs energy function (equation 9). To construct the reference state, we consider a series of likelihoods $p_\xi(D, \theta | \mathbf{X}, \sigma) \sim [p(D | \mathbf{X}, \sigma)]^\xi$ parameterized by $\xi \in [0, 1]$, and set the reference state as the thermodynamic ensemble corresponding to $\xi = 0$. The BICePs score is then calculated as the change in free energy of "turning on" experimental restraints ($\xi = 0 \rightarrow 1$).

It should be noted that in other applications of BICePs,[6,12] the reference state for the BICePs score is defined using the $\lambda = 0$ state for a series of a priors $p_\lambda(X) \sim [p(X)]^\lambda$, and the BICePs score is computed as the free energy of ($\lambda = 0 \rightarrow 1$) and ($\xi = 0 \rightarrow 1$) transformations. Here, since we are only interested in evaluating and/or parameterizing the likelihood functions, we set $p(X)$ to be uniform. Constructing $p(X)$ is thus very straightforward: it's a collection of conformations all having equal statistical weight.

The derivative of the BICePs score with respect to the FM parameters θ reduces to the difference of Boltzmann averaged values of $\partial u / \partial \theta$ shown as

$$\frac{\partial f(\theta)}{\partial \theta_i} = \iint \frac{1}{Z(\theta)} \left[\frac{\partial u}{\partial \theta_i} \right] \exp(-u) \, d\mathbf{X} d\sigma = \left\langle \frac{\partial u}{\partial \theta_i} \right\rangle \quad (20)$$

In this study, we demonstrate our methodology using first-order optimization methods, such as L-BFGS-B. For more complex forward models, the employment of second derivatives might become necessary. Interested readers are directed to the Supporting Information for second derivatives of the BICePs score with respect to FM parameters.

Calculation of the BICePs score (a free energy difference) and its derivatives (expectation values of energy derivative observables) is performed using the MBAR free energy estimator,[22] by sampling at several intermediates $\xi = 0 \rightarrow 1$, which enables accurate estimates of all quantities.

Optimizing ξ-values. The accuracy of the BICePs score depends on converged sampling and sufficient thermodynamic overlap of intermediates ($\xi = 0 \rightarrow 1$) in the BICePs computation. To ensure strong overlap, we optimize the ξ-values by spacing ensembles equidistantly in thermodynamic length, employing a strategy akin to the "thermodynamic trailblazing" method proposed by Rizzi et al.[23] Our approach is facilitated by a custom optimization algorithm called `pylambdaopt` (Zhang et al., in preparation).

The optimization process is a two-step process: First, a preliminary BICePs calculation is performed using provisional ξ-values, yielding estimates of the thermodynamic length $|\ell(\xi_{n+1}) - \ell(\xi_n)|$ for each pair of intermediates[24,25], derived from the variance in distributions $p(\Delta u_{n,n+1})$, where $\Delta u_{n,n+1} = u_{n+1} - u_n$ represents the change in the (reduced) BICePs energy incurred by bringing a sample from thermodynamic ensemble n to thermodynamic ensemble $n+1$.

Second, cubic spline fitting is employed to derive a smooth and differentiable function $\ell(\xi)$ that accurately interpolates the computed $\ell(\xi_i)$. Optimization through steepest-descent minimization is then applied to determine new ξ_i^* values that minimize the loss function $\mathscr{L} = \sum_n |\ell(\xi_{n+1}) - \ell(\xi_n)|^2$. This results in ξ_i^* values uniformly spaced in terms of thermodynamic length, thus maximizing the thermodynamic overlap between adjacent ensembles and enhancing the precision of free energy calculations. These optimized ξ_i^* values are subsequently used in production runs. An illustration of the ξ-values pre- and post-optimization is depicted in Figure S5. Refer to figures S6&S7 for overlap matrices pre- and post- optimization.

Comparison of variational minimization of the BICePs score vs. sampling the full joint posterior

In the comparison of the two approaches for parameter estimation and optimization in our model, we utilized a toy model (Figure S2) to evaluate the efficacy of each method under the same data conditions. Prior to FM parameter refinement, 11 ξ-values were optimized from $\{1.0, 0.9, 0.8, ..., 0.0\}$ to $\{1.0, 0.7, 0.56, 0.45, 0.36, 0.28, 0.2, 0.14, 0.08, 0.04, 0.0\}$ (Figures S5-S7). Variational minimization using the Good-Bad model with 4 replicas (for reduced computational cost), where each evaluation of the objective function consisted of running 10k MCMC steps. Optimal parameters were determined to be $\{A = 6.31 \pm 0.02, B = -1.69 \pm 0.03, C = 1.69 \pm 0.01\}$, averaged over 3 independent runs with very low variance between runs, shown in Figure S8. Regardless of different starting parameters ($\{A = 9, B = -1, C = 1\}$, $\{A = 4, B = 0, C = 3\}$, $\{A = 0, B = 0, C = 0\}$), variational minimization converges to "true" optimal FM parameters. This analysis demonstrated that both the joint posterior sampling approach and variational minimization yield near equivalent performance when applied to this model.

As a method for forward model optimization, variational minimization of the BICePs score has advantages and disadvantages. This method is particularly advantageous for handling many FM parameters, offering a potential solution to the curse of dimensionality faced by Monte Carlo Markov Chain (MCMC) methods. Additionally, it is easier for users to adapt different forward models, and performs exceptionally well in convex landscapes. When landscapes are non-convex, however, the inverse Hessian may not provide a comprehensive view of the parameter space's uncertainty; instead, uncertainty estimation could be computed using the variance across multiple BICePs runs starting from different initial parameters. The variational minimization approach also requires careful consideration of disperse starting parameters to ensure global minimization.

The joint posterior sampling method, which involves sampling the joint posterior distribution of forward model (FM) parameters, has several advantages. One significant benefit is that the posterior distribution provides a direct estimate of the uncertainties in forward model parameters and their covariance. Compared to variational minimization, this method generally has a faster runtime and is particularly effective in handling non-convex landscapes, allowing for robust parameter estimation even in complex scenarios. However, it is not without drawbacks. As the number of FM parameters increases, the posterior sampling method may encounter the curse of dimensionality, which makes it computationally challenging to explore the parameter space efficiently.

In summary, while both approaches are valuable tools for parameter estimation in parameter and ensemble refinement, each has its strengths and weaknesses. The choice between these methods should be guided by the specific characteristics of the problem at hand, such as the landscape's convexity and the number of parameters involved.

Determination of optimal Karplus coefficients for ubiquitin

To evaluate the performance of our algorithm, we applied BICePs to human ubiquitin to predict Karplus coefficients for six sets of scalar coupling constants: $^3J_{H^N H^\alpha}$, $^3J_{H^\alpha C'}$, $^3J_{H^N C\beta}$, $^3J_{H^N C'}$, $^3J_{C'C\beta}$, and $^3J_{C'C'}$. To test our algorithm's robustness, we conducted a comprehensive evaluation for predicting optimal Karplus coefficients using three different structural ensembles as priors, each derived from distinct computational approaches: (1) 10 conformations from the NMR-refined structural ensemble, 1D3Z[26], (2) 144 conformations from NMR-restrained simulations, 2NR2[27], and (3) 25 conformations from the RosettaFold2 (RF2) algorithm.[28]

We then validated the forward model parameters derived from each prior using the BICePs score, R^2 and mean absolute errors (MAE) for forward model predictions. As priors for these calculations, we used three independent structural ensembles: 1D3Z, 2NR2, and a 500-state conformational ensemble derived from a millisecond-long simulation of ubiquitin using CHARMM22*.[29] For further details on these ensembles, refer to the SI methods section.

To refine the forward model (FM) parameters, we employed full joint posterior distribution sampling. This method was chosen to navigate the non-convex parameter space efficiently, given its relatively low dimensionality (18 FM parameters). BICePs calculations were executed by averaging the FM parameters over four Markov Chain Monte Carlo (MCMC) chains, each starting from distinct initial parameters: $\{A = 9, B = -1, C = 1\}$, $\{A = 4, B = 0, C = 3\}$, $\{A = 0, B = 0, C = 0\}$, and $\{A = 6, B = -1, C = 0\}$. Flexible residues were excluded from the calculations, consistent with previous studies[2,3]. As a result, a total of 346 J-couplings were used in these refinements. We used the Good-Bad model with 32 BICePs replicas, discarding the first 50k steps as burn-in, followed by 50k steps for MCMC sampling. Unlike the parameters derived from 1D3Z and RF2, the Karplus coefficients obtained by using the 2NR2 ensemble required a burn-in of 100k steps to appropriately converge due to a larger number of conformational states. The six sets of refined Karplus coefficients resulting from the 1D3Z, 2NR2 and RF2 ensembles are presented in Table I.

Figure 3 compares the Karplus curves derived from BICePs using the 1D3Z ensemble with previously published parameters obtained from NMR refinements, showing subtle differences. Both the marginal posterior distributions of the FM parameters and the Karplus curves for each scalar coupling demonstrate significant congruence with the historical NMR refinement results[3,26]. For all six types of J-coupling, see Figure S9.

The predicted parameters, better represented by the marginal posterior distributions of the FM parameters, have large similarities across structural ensembles. BICePs-predicted coefficients using the 1D3Z ensemble (Figure S10) and predicted coefficients using the RF2 ensemble (Figure S11) are found to have very strong overlap. Furthermore, the traces of the FM parameters ober time (Figure S12) confirm convergence.

One advantage of BICePs is that as FM parameters are being sampled, the posterior densities of FM uncertainties, $p(\sigma)$, are also revealed (Figure S13). For certain sets of J-coupling constants (e.g., $^3J_{H^N H^\alpha}$ and $^3J_{H^N C'}$) the marginal posterior distribution of the variance scaling parameter $p(\varphi)$ has a sampled

Table I. Coefficients for the Karplus relation $^3J(\phi) = A\cos^2(\phi + \phi_0) + B\cos(\phi + \phi_0) + C$, determined by BICePs sampling the joint posterior of FM parameters.

		ϕ_0	A (Hz)	B (Hz)	C (Hz)
$^3J_{C'C}$	1	$0°$	1.71 ± 0.02	-0.85 ± 0.01	0.54 ± 0.00
	2	$0°$	1.30 ± 0.03	-0.91 ± 0.01	0.62 ± 0.01
	3	$0°$	1.62 ± 0.03	-0.87 ± 0.01	0.63 ± 0.01
$^3J_{C'C\beta}$	1	$60°$	1.83 ± 0.04	0.34 ± 0.05	0.41 ± 0.02
	2	$60°$	2.20 ± 0.04	0.34 ± 0.04	0.04 ± 0.02
	3	$60°$	1.81 ± 0.04	0.38 ± 0.04	0.31 ± 0.02
$^3J_{H^\alpha C'}$	1	$120°$	3.64 ± 0.02	-2.14 ± 0.02	1.27 ± 0.02
	2	$120°$	4.10 ± 0.03	-2.00 ± 0.02	0.95 ± 0.02
	3	$120°$	3.78 ± 0.02	-2.12 ± 0.02	1.21 ± 0.02
$^3J_{H^N C'}$	1	$180°$	4.33 ± 0.04	-1.17 ± 0.01	0.14 ± 0.01
	2	$180°$	4.60 ± 0.12	-0.57 ± 0.03	-0.10 ± 0.01
	3	$180°$	4.57 ± 0.09	-1.20 ± 0.03	0.13 ± 0.01
$^3J_{H^N C\beta}$	1	$60°$	2.72 ± 0.03	-0.35 ± 0.03	0.12 ± 0.01
	2	$60°$	3.00 ± 0.04	-0.26 ± 0.03	-0.28 ± 0.02
	3	$60°$	2.52 ± 0.03	-0.03 ± 0.02	-0.09 ± 0.02
$^3J_{H^N H^\alpha}$	1	$-60°$	7.11 ± 0.05	-1.38 ± 0.03	1.43 ± 0.04
	2	$-60°$	7.50 ± 0.07	-1.50 ± 0.02	1.50 ± 0.06
	3	$-60°$	6.97 ± 0.07	-1.49 ± 0.04	1.63 ± 0.05

[1] 1D3Z as the structural ensemble
[2] 2NR2 as the structural ensemble
[3] RosettaFold2 (RF2) as the structural ensemble

mean slightly larger than 1.0, indicating that the functional form of the likelihood opted for long tails to account for a few outlier data points deviating from the mean.

The BICePs free energy landscape for $^3J_{H^N C'}$ Karplus parameters. In Figure 4, we show the free energy landscape, which is also equivalent to the BICePs score landscape $f_{\xi=0\to1}$. The Karplus curve for $^3J_{H^N C'}$ was found to overlap strongly with the results obtained by SVD when using ϕ angles from the X-ray crystal structure (Figure S9). Red data points are shown using the experimental J-couplings with ϕ angles derived from X-ray crystal pose 1UBQ[30]. The joint BICePs score landscape for the six sets of parameters is too complex to visualize. In an attempt to do our best, we constructed a smooth 2-D landscape for each pair of parameters within a set of scalar couplings by training a Gaussian process on the BICePs energy trace using a radial basis function (RBF) kernel. The landscape matches the computed BICePs scores, and shows minima in the correct locations. All BICePs score landscapes for each of the six sets of Karplus coefficients are illustrated in Figure S14.

To demonstrate the transferability across different generative models and validate our parameters, we evaluated the accuracy of the back-calculated scalar couplings using the different sets of Karplus coefficients. In Figure 5, we illustrate how the various sets of parameters derived from different techniques and different structural ensembles exhibit similar performance metrics. Interestingly, applying BICePs-refined Karplus parameters to an ensemble generated by a molecular dynamics simulation (CHARMM22*),[29] some parameter sets are revealed to be more transferable than others. The mean absolute error (MAE) and coefficient of determination (R^2) for all six types of scalar couplings across different structural ensembles are shown in Figures S15-S17. On average, the BICePs-refined parameters derived
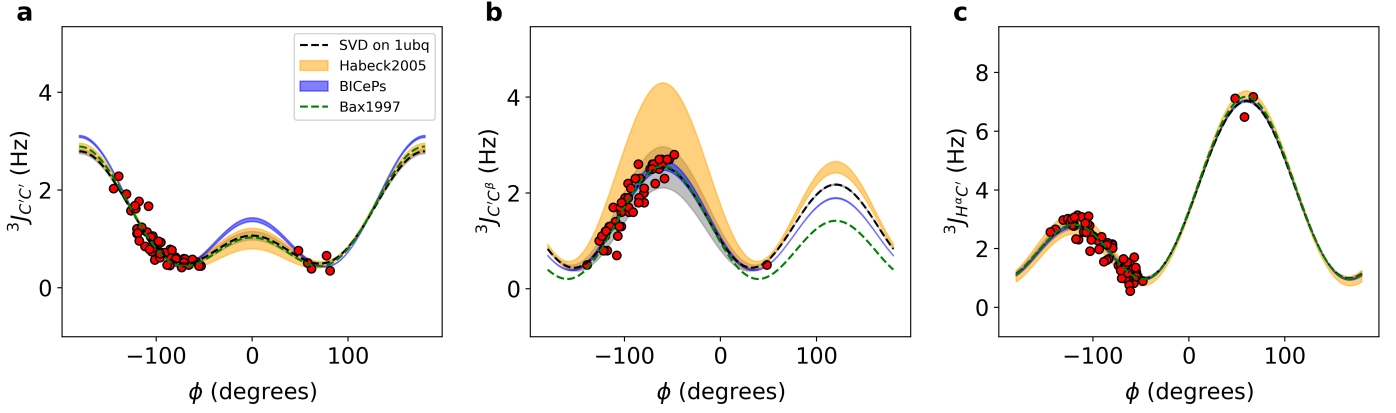
Figure 3. Karplus curves with BICePs-refined Karplus coefficients using the 1d3z ensemble for (a-c) $^3J_{C'C'}$, $^3J_{C'C^\beta}$, and $^3J_{H^\alpha C'}$. For comparison, SVD on 1ubq using experimental scalar coupling constants with ϕ-angles derived from the X-ray structure (black dashed line), and red dots correspond to the fitted data points. Additionally, parameterizations from Bax et al. 1997 (green) and parameterization from Habeck et al. 2005 (yellow) were overlaid for comparison. The thickness of the line corresponds to the uncertainty.

Figure 4. Landscapes of the BICePs score with respect to the predicted Karplus coefficients for $^3J_{H^N C'}$. Panels a, c and d illustrate the energy landscape f for pairs of Karplus coefficients when using the 1D3Z structural ensemble during refinement.

from the 2NR2 ensemble (BICePs(2NR2)) give the lowest MAE between experiment and predictions for the CHARMM22* simulated ensemble, closely followed by BICePs(RF2) parameters, whereas Habeck 2005 has the highest due to known difficulties with $^3J_{C'C^\beta}$.

To objectively quantify which parameters produce the best predictions for ubiquitin, we compute BICePs scores, $f_{\xi=0\to1}$ for each of the structural ensembles. This score directly relates to the quality of FM parameters and their predictive accuracy at reproducing experimental scalar couplings, while taking into consideration all sources of potential error. Lower BICePs scores indicate better agreement with experiment. Each row in Table II corresponds to BICePs scores using all six sets of Karplus coefficients used on different structural ensembles. The lowest score is shown in bold.

BICePs scores, $f_{\xi=0\to1}$ were computed to objectively rank the quality of FM parameters and their predictive accuracy at reproducing experimental scalar couplings (Table II). The left-most column in Table II corresponds to the parameters, where *BICePs(1D3Z)* are the parameters in Table I (set 1), which used 1D3Z ensemble to obtain Karplus coefficients. BICePs score columns, e.g., $f_{\xi=0\to1}^{1d3z}$ corresponds to BICePs scores evaluated for the 1D3Z ensemble. That is, the superscript corresponds to the structural ensemble used as a validation step. BICePs scores, f for each structural ensemble over all sets of parameters, averaged over five independent rounds of validation each. BICePs calculations burned for 1k steps, followed by 50k steps of MCMC sampling.

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, $f(\theta)/N_r$. We can confirm that the BICePs score, $f_{\xi=0\to1}^{1d3z} = 38.14 \pm 0.08$ (Table II) is equivalent (within error) with the most probable landscape basin $f = 38.15 \pm 0.19$ from sampling the energy landscape, computed as an averrage across four chains; an example for one chain is shown in Figure 4. This is additional evidence of the algorithm's reliability and quality of the BICePs score, corroborating that the results from variational minimization of the BICePs score and full joint posterior sampling are equivalent.

Table II. BICePs scores (32 replicas), f for each structural ensemble over all sets of parameters, averaged over five independent rounds of validation each.

Parameters	$f_{\xi=0\to1}^{1d3z}$	$f_{\xi=0\to1}^{2nr2}$	$f_{\xi=0\to1}^{CHARMM22*}$
Bax 1997[2,26]	61.12 ± 0.08	132.49 ± 0.09	99.27 ± 1.91
Habeck 2005[3]	135.66 ± 0.08	199.32 ± 0.24	165.64 ± 0.47
BICePs(1D3Z)	**38.14 ± 0.08**	141.69 ± 0.16	105.00 ± 0.74
BICePs(2NR2)	118.42 ± 0.14	**113.15 ± 1.34**	**76.27 ± 0.66**
BICePs(RF2)	68.07 ± 0.60	129.42 ± 0.15	88.04 ± 0.21

For both the 2NR2 and CHARMM22* structural ensembles, Bax 1997, BICePs(RF2) and BICePs(1D3Z) parameters give very similar BICePs scores, which suggests robust accu-
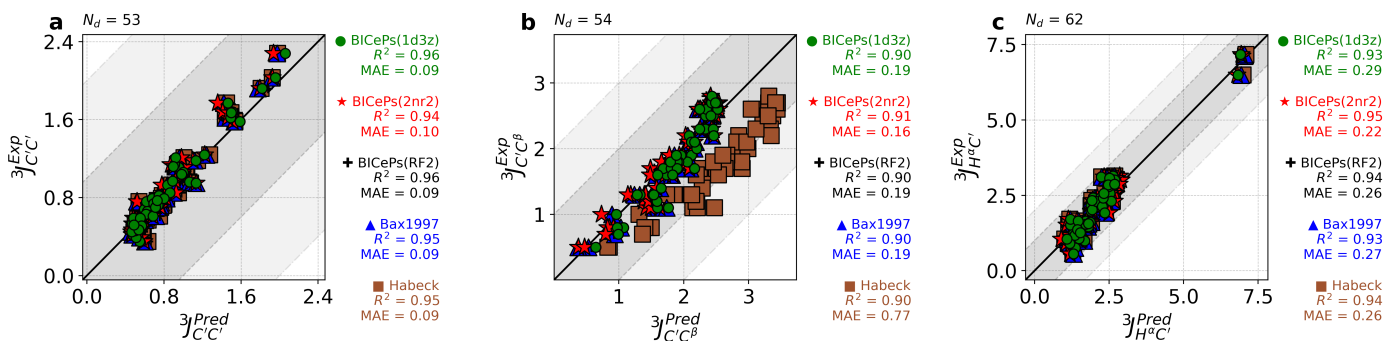
Figure 5. Validation of BICePs-predicted Karplus coefficients perform similarly to Bax1997 and achieve minor improvements over Habeck2005 for scalar coupling predictions for the simulated ensemble of CHARMM22*. Each panel for (a) $^3J_{H^\alpha C'}$, (b) $^3J_{C'C^\beta}$, and (c) $^3J_{C'C'}$ shows strong correlations between predictions and experiment. Karplus coefficients derived from BICePs using the 2NR2 ensemble gives the best performance for CHARMM22*. For the remaining sets of *J*-coupling, please see Figure S17.

racy of FM parameters in reproducing experimental scalar couplings and the transferability of FM parameters across different prior structural ensembles. Furthermore, when it comes to the CHARMM22* simulated ensemble, the BICePs(2NR2) parameters give the lowest BICePs score. However, it is important to note that the structural ensemble 2NR2 was generated using CHARMM22 force field with additional experimental restraints during simulation.

It is difficult to say which of the model parameters are the best for ubiquitin, so we compare the top four: BICePs(RF2), Bax 1997, BICePs(1D3Z), and BICePs(2NR2). The BICePs(2NR2) parameters are objectively better at predicting *J*-couplings from structures of ubiquitin generated from simulations using CHARMM22* force field. Futhermore, our BICePs(RF2) parameters have slightly better transferability across structural ensembles and have a better BICePs score over Bax 1997 parameters. In general, when looking across structural ensembles, the lowest BICePs scores come from the 1D3Z structural ensemble ($f^{1d3z}_{\xi=0\to1}$) except for BICePs parameters derived from the 2NR2 ensemble (BICePs(2NR2)). This confirms that the 1D3Z structural ensemble gives the strongest agreement with experimental NMR observations.

Ensembles from generative models like RosettaFold2 can be used for parameter refinement. The booming field of machine learning and artificial intelligence is swiftly transforming the field of modeling structure and dynamics in biological systems. Recent advancements in generative models, such as AlphaFold[31], RosettaFold[28] and others, have heralded a new era in the accurate prediction of structural ensembles. Leveraging the predictive power of these models as structural priors is expected to help refine ensemble predictions when integrated with similar algorithms to BICePs[32]. Here, we have demonstrated that structural ensembles generated from RosettaFold2 (RF2) can be reweighted to better align with experimental measurements, while simultaneously refining Karplus parameters. Validation of these parameters by the BICePs score and other statistics demonstrates improved accuracy across a varity of structural ensembles of ubiquitin.

Automatic determination of unknown errors. Our method provides a notable advantage by automatically estimating all potential error sources throughout the ensemble refinement pro-

cess. This estimation is facilitated through the analysis of posterior distributions, which are instrumental in deriving accurate error assessments for the Karplus coefficients. Consequently, this negates the need for cross-validation techniques commonly used in other approaches[1,33].

In the context of model validation, the BICePs score emerges as a superior metric over the traditional χ^2 test. Unlike χ^2, which presupposes a fixed and known error, BICePs dynamically ascertains the level of uncertainty, thereby providing a more nuanced and accurate measure of model quality.

Bayesian ranking of Karplus-type relations The Karplus equation, a cornerstone for interpreting NMR spectroscopy data, comes in multiple forms to accommodate the diverse characteristics of molecular structures, from rigid to flexible[34]. The BICePs algorithm can determine coefficients and their uncertainties for any functional form, including those with additional parameters. Although we do pursue this aim in our current work, it is straightforward to use Bayesian model selection to objectively rank empirical models based on their BICePs scores, while automatically accounting for model complexity, thus providing a balance of model accuracy and parsimony.

Adaptive variance as simulated annealing. As an alternative approach to determine optimal FM parameters, we propose that future work might utilize an annealing approach, in which the variance parameter σ^2 is akin to the temperature. Initially, a high σ^2 would enable extensive exploration of the parameter space to circumvent local optima. This exploration phase mimics the high-temperature regime in annealing, allowing for a broad search. Subsequently, we suggest a schedule of stepwise reduction in σ^2, similar to cooling in simulated annealing, to gradually narrow the search area and determine the optimum solution. This method balances between wide-ranging search and focused refinement, potentially enhancing search efficiency and robustness in FM parameter optimization.

IV. CONCLUSION

In the quest for accurate forward model predictions, specifically for *J*-coupling, researchers often navigate the vast literature seeking Karplus parameters that align with their specific

systems, occasionally settling for less-than-ideal solutions. Our work demonstrates BICePs as a robust tool for determining forward model (FM) parameters by sampling over their full posterior distribution. We used a toy model to demonstrate that variational minimization of the BICePs score is also a valid approach for FM parameter refinement.

We have shown how the BICePs score–the free energy of "turning on" the restraints tethering the forward model predictions to the experimental values–serves as an effective validation metric for FM parameters. Using structural ensembles and experimental data for ubiquitin, BICePs determined six different sets of Karplus coefficients using different types of J-coupling measurements, while effectively addressing both random and systematic errors. From these results, one can see how this algorithm can be applied more generally to find other optimal forward model parameters. These advances not only contribute to the refinement of molecular simulations but also hold promise for a wide range of applications within the scientific community, particularly among those analyzing structural dynamics and performing model validation.

CONFLICTS OF INTEREST

Authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

REFERENCES

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

[2] A. C. Wang and A. Bax, "Determination of the backbone dihedral angles ϕ in human ubiquitin from reparametrized empirical karplus equations," Journal of the American Chemical Society **118**, 2483–2494 (1996).

[3] M. Habeck, W. Rieping, and M. Nilges, "Bayesian Estimation of Karplus Parameters and Torsion Angles from Three-Bond Scalar Couplings Constants," Journal of magnetic resonance **177**, 160–165 (2005).

[4] J. M. Schmidt, M. Blümel, F. Löhr, and H. Rüterjans, "Self-consistent 3j coupling analysis for the joint calibration of karplus coefficients and evaluation of torsion angles," Journal of biomolecular NMR **14**, 1–12 (1999).

[5] V. A. Voelz, Y. Ge, and R. M. Raddi, "Reconciling Simulations and Experiments with BICePs: a Review," Front. Mol. Biosci. **8**, 661520 (2021).

[6] R. M. Raddi, T. Marshall, Y. Ge, and V. Voelz, "Model selection using replica averaging with bayesian inference of conformational populations," chemrXiv preprint 10.26434/chemrxiv-2023-396mm (2023).

[7] V. a. Voelz and G. Zhou, "Bayesian Inference of Conformational State Populations from Computational Models and Sparse Experimental Observables," J. Comput. Chem. **35**, 2215–2224 (2014).

[8] H. Wan, Y. Ge, A. Razavi, and V. A. Voelz, "Reconciling Simulated Ensembles of Apomyoglobin with Experimental Hydrogen/Deuterium Exchange Data Using Bayesian Inference and Multiensemble Markov State Models." J. Chem. Theory Comput. **16**, 1333–1348 (2020).

[9] M. F. D. Hurley, J. D. Northrup, Y. Ge, C. E. Schafmeister, and V. A. Voelz, "Metal Cation-Binding Mechanisms of Q-Proline Peptoid Macrocycles in Solution," J. Chem. Inf. Model. **61**, 2818–2828 (2021).

[10] R. M. Raddi, Y. Ge, and V. A. Voelz, "BICePs V2. 0: Software for Ensemble Reweighting Using Bayesian Inference of Conformational Populations," Journal of chemical information and modeling **63**, 2370–2381 (2023).

[11] Y. Ge and V. A. Voelz, "Model Selection Using BICePs: a Bayesian Approach for Force Field Validation and Parameterization," J. Phys. Chem. B **122**, 5610–5622 (2018).

[12] R. M. Raddi and V. A. Voelz, "Automated optimization of force field parameters against ensemble-averaged measurements with bayesian inference of conformational populations," arXiv preprint arXiv:2402.11169 (2024).

[13] W. Rieping, M. Habeck, and M. Nilges, "Inferential Structure Determination," Science **309**, 303–306 (2005).

[14] J. W. Pitera and J. D. Chodera, "On the Use of Experimental Observations to Bias Simulated Ensembles," Journal of chemical theory and computation **8**, 3445–3451 (2012).

[15] A. Cavalli, C. Camilloni, and M. Vendruscolo, "Molecular Dynamics Simulations with Replica-Averaged Structural Restraints Generate Structural Ensembles According to the Maximum Entropy Principle," The Journal of chemical physics **138**, 03B603 (2013).

[16] A. Cesari, S. Reißer, and G. Bussi, "Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments," Computation **6**, 15 (2018).

[17] B. Roux and J. Weare, "On the Statistical Equivalence of Restrained-Ensemble Simulations with the Maximum Entropy Method," The Journal of chemical physics **138**, 02B616 (2013).

[18] G. Hummer and J. Köfinger, "Bayesian Ensemble Refinement by Replica Simulations and Reweighting," The Journal of chemical physics **143**, 12B634_1 (2015).

[19] M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo, "Metainference: A Bayesian Inference Method for Heterogeneous Systems," Sci. Adv. **2**, e1501177 (2016).

[20] J.-S. Hu and A. Bax, "Determination of ϕ and χ_1 angles in proteins from ^{13}c-^{13}c three-bond j couplings measured by three-dimensional heteronuclear nmr. how planar is the peptide bond?" (1997).

[21] M. Karplus, "Vicinal proton coupling in nuclear magnetic resonance," Journal of the American Chemical Society **85**, 2870–2871 (1963).

[22] M. R. Shirts and J. D. Chodera, "Statistically Optimal Analysis of Samples from Multiple Equilibrium States," J. Chem. Phys. **129**, 124105–11 (2008).

[23] A. Rizzi, *Improving Efficiency and Scalability of Free Energy Calculations through Automatic Protocol Optimization*, Ph.D. thesis, Weill Medical College of Cornell University (2020).

[24] D. A. Sivak and G. E. Crooks, "Thermodynamic metrics and optimal paths," Physical review letters **108**, 190602 (2012).

[25] D. K. Shenfeld, H. Xu, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Minimizing Thermodynamic Length to Select Intermediate States for Free-Energy Calculations and Replica-Exchange Simulations," Phys. Rev. E **80**, 46705 (2009).

[26] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, "Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase," Journal of the American Chemical Society **120**, 6836–6837 (1998).

[27] B. Richter, J. Gsponer, P. Várnai, X. Salvatella, and M. Vendruscolo, "The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins," Journal of biomolecular NMR **37**, 117–135 (2007).

[28] M. Baek, I. Anishchenko, I. Humphreys, Q. Cong, D. Baker, and F. DiMaio, "Efficient and accurate prediction of protein structure using rosettafold2," bioRxiv, 2023–05 (2023).

[29] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," Proc. Natl. Acad. Sci. U. S. A. **110**, 5915–5920 (2013).

[30] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 åresolution," Journal of molecular biology **194**, 531–544 (1987).

[31] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," Nature **596**, 583–589 (2021).

[32] Z. F. Brotzakis, S. Zhang, and M. Vendruscolo, "Alphafold prediction of structural ensembles of disordered proteins," bioRxiv, 2023–01 (2023).

[33] G. W. Vuister and A. Bax, "Quantitative j correlation: a new approach for measuring homonuclear three-bond j (hnh. alpha.) coupling constants in 15n-enriched proteins," Journal of the American Chemical Society **115**, 7772–7777 (1993).

[34]M. J. Minch, "Orientational dependence of vicinal proton-proton nmr coupling constants: The karplus relationship," Concepts in magnetic resonance **6**, 41–56 (1994).

[35]M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: making protein folding accessible to all," Nature methods **19**, 679–682 (2022).

[36]C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. Noé, "Introduction to markov state modeling with the pyemma software [article v1.0]," Living Journal of Computational Molecular Science **1**, 5965 (2019).

[37]H. Wu and F. Noé, "Variational approach for learning markov processes from time series data," J. Nonlinear Sci. **30**, 23–66 (2020).

[38]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," J. Chem. Phys. **139**, 015102 (2013).

[39]M. J, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (1967).

Appendix A: Variational minimization of the BICePs score to find optimal parameters.

It is likely that the second derivative w.r.t FM parameters would be useful for second-order optimization methods and uncertainty quantification. For the sake of simplicity, we refrain from writing out the second derivative for the Good-Bad model here. However, when $\varphi = 1$ the second partial derivatives of the BICePs energy with respect to parameters θ_i and θ_k are:

$$\frac{\partial^2 u}{\partial \theta_i \partial \theta_k} = N \left[\sum_{j=1}^{N_d} -\frac{\partial^2 g_j(X,\theta)}{\partial \theta_i \partial \theta_k} \frac{(d_j - g_j(\mathbf{X},\theta))}{\sigma_j^2} + \frac{\partial g_j(X,\theta)}{\partial \theta_i} \cdot \frac{\partial g_j(X,\theta)}{\partial \theta_k} \frac{1}{\sigma_j^2} \right]. \quad \text{(A1)}$$

When $i = k$, the second partial derivative of the energy is just:

$$\frac{\partial^2 u}{\partial \theta^2} = N \left[\sum_{j=1}^{N_d} \left(-\frac{\partial^2 g_j(X,\theta)}{\partial \theta^2} \frac{(d_j - g_j(\mathbf{X},\theta))}{\sigma_j^2} + \left(\frac{\partial g_j(X,\theta)}{\partial \theta} \right)^2 \frac{1}{\sigma_j^2} \right) \right]. \quad \text{(A2)}$$

The second partial derivatives of the BICePs score with respect to parameters θ_i and θ_k are:

$$\frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_k} = \left\langle \frac{\partial^2 u}{\partial \theta_i \partial \theta_k} \right\rangle - \left(\left\langle \frac{\partial u}{\partial \theta_i} \cdot \frac{\partial u}{\partial \theta_k} \right\rangle - \left\langle \frac{\partial u}{\partial \theta_i} \cdot \frac{\partial u}{\partial \theta_k} \right\rangle' \right), \quad \text{(A3)}$$

where the notation $\langle \cdot \rangle'$ denotes the ensemble average with respect to $\left(\frac{1}{Z} \exp(-u) \right)^2$. The first term on the right is the ensemble-averaged second derivative of the energy function u with respect parameters θ_i and θ_k is equation A2 (when $\varphi = 1$).

When $i = k$, the second partial derivative of the BICePs score reduces to the difference between the ensemble-averaged second derivative of the energy u and the variance of its first partial derivative:

$$\frac{\partial^2 f}{\partial \theta^2} = \left\langle \frac{\partial^2 u}{\partial \theta^2} \right\rangle - \left(\left\langle \left(\frac{\partial u}{\partial \theta} \right)^2 \right\rangle - \left\langle \frac{\partial u}{\partial \theta} \right\rangle^2 \right) \quad \text{(A4)}$$

The calculation is performed using the MBAR free energy estimator for the BICePs score and it's derivatives, by sampling at several intermediates $\xi = 0 \to 1$, which enables accurate estimates of all quantities.

In the case of convex landscapes, one can compute reported uncertainties in the set of optimized FM parameters θ by estimation of the covariance through inversion of the Hessian (the matrix of second partial derivatives of the BICePs score). Conceptually, this means that the estimated uncertainties in the best-fit values represent the widths of the basins on the BICePs score landscape. For non-convex problems, this can sometimes lead to an under-estimation of uncertainty depending upon the curvature of the basins of the local minima in the BICePs score landscape.

Appendix B: Supplemental Information

1. Methods

Structural ensembles of ubiquitin

1D3Z Ubiquitin NMR structural ensemble that consists of 10 conformations[26].

2NR2 Next, we use an ensemble (144 conformations) from Richter et al. (PDB: 2NR2), where the refinement was performed using the MUMO (minimal under-restraining minimal over-restraining) method[27]. In this approach, simulations started from the X-ray crystal pose[30]. NOEs and S^2 Lipari–Szabo order parameters were used as NMR structural restraints during a replica-averaged restrained simulation with TIP3P solvent and CHARM22 force field. The augmented potential energy function given as: $E_{total} = E_{CHARMM22} + E_{restraints}$. Scalar couplings were not used during the refinement, but were only used as a validation metric.

RosettaFold2 (RF2) Using the Colabfold notebook[35] with default parameters, 25 conformations were predicted using RosettaFold2[28].

*CHARMM22** An ensemble of conformationa states was derived from a Markov State Model (MSMs) were constructed from a 1-ms simulation of ubiquitin's native state at 300 K from Piana et al.[29]. The PyEMMA Python package[36] was used to determine appropriate backbone featurizations using the Variational Approach for Markov Processes (VAMP) scoring function VAMP-2[37]. Based on the obtained scores, inverse distances were selected as features, while torsions were excluded due to lower VAMP-2 scores and minimal contribution when paired with inverse distances. When comparing distances and inverse distances, the similarity of the average scores and relative standard deviations show that both featurizations are adequate, and inverse distances was selected purely on the higher average scores. Time-lagged independent component analysis (tICA) followed by k-means clustering was used to discretize simulations into a structural ensemble of 500 conformational states for MSM construction[38,39]. For the BICePs calculation, each of the 500 states were given equal statistical weight to enforce the uniform prior $p(X)$.

Details of parameters used in posterior sampling

The range of sampled uncertainty parameters σ on a grid of logarithmically-spaced values between 0.001 to 100, to enforce the Jeffrey's prior. Each grid value in the list was a factor of 1.02 larger than the next: [1.00e-03, 1.02e-03, 1.04e-03, 1.06e-03, ... 9.72e+01, 9.92e+01], resulting in a list of 582 values. For the Good-Bad model, sampling of the extra nuisance parameter, φ took place on a grid from 1 to 100 with 1000 equally-spaced points.

Details of ξ optimization

Optimization was performed for a maximum of 2M steps with a tolerance of 1e-7 and $\alpha = 1e - 5$. With increasing amounts of data restraint energy, the optimization problem becomes more complicated and more iterations are required to converge. In the case of insufficent iterations, the ξ-optimization might return negative ξ-values, which is incorrect and not physical. Initially, we start with 11 ξ-values {1.0, 0.9, 0.8,...,0.0} and with shift to lower values of ξ after optimization e.g., {1.0, 0.7, 0.56, 0.45, 0.36, 0.28, 0.2, 0.14, 0.08, 0.04, 0.0}.

Figure S1. The Good-Bad model properly detects outliers. The probability density function (a) computed as $p(D|\mathbf{X},\sigma_0,\varphi) = \prod_j^{N_j} p(d_j|\mathbf{X},\sigma_0,\varphi)$ and energy landscape (b) of the marginal likelihood for the Good-Bad model with respect to the replica-averaged forward model data $f(\mathbf{X})$ using multiple data points. Shown here, are three data points, two good data points $\{0.25, 0.3\}$ and one outlier $\{1.0\}$. The Good-Bad model ($\varphi = 7.0$) is centered about the mean of the two good data points, demonstrating that this model can distinguish the good and bad data. The standard Gaussian likelihood ($\varphi = 1$) is centered about the mean of all three data points. The colored curves are different values of nuisance parameter φ. Subplots (a) and (b) show how the Good-Bad model when $\varphi = 1$ is equivalent to the Gaussian likelihood and harmonic potential energy function. Subplots (c) and (d) are the first and second derivatives of the potential energy curves shown in subplot (b).

Figure S2. **BICePs predicted forward model parameters in the presence of random and systematic error ($\sigma_{data} = 0.471$) for a toy model system.** (a) Karplus curves predicted for SVD (orange) and BICePs (blue), where the "true" (black dashed line) parameters were set to be $\{A = 6.51, B = -1.76, C = 1.6\}$. The extracted parameters from the SVD fitting were found to be $\{A = 6.11 \pm 0.06, B = -1.63 \pm 0.04, C = 1.80 \pm 0.04\}$ and the BICePs was $\{A = 6.6 \pm 0.037, B = -1.8 \pm 0.016, C = 1.5 \pm 0.027\}$, averaged over three independent chains. The uncertainty is represented by the thickness of the curves. For the BICePs calculation, we used the Good-Bad likelihood model with 32 replicas and burned for 20k steps, followed by 50k steps of additional sampling. (b) The marginal posterior distribution of uncertainty $p(\sigma_J)$. The maximum a posteriori was determined to be $\hat{\sigma}_J = 0.272$, and the *a posteriori* variance scaling parameter $\hat{\phi}_J = 1.98$. (c) Landscapes of the BICePs score, f for pairs of Karplus coefficients. (d) The marginal posterior distribution of Karplus coefficients.

Figure S3. **The Student's model gives similar performance to the Good-Bad model.** Karplus coefficients predicted using the Student's likeliood model ($\{A = 6.8 \pm 0.033, B = -1.9 \pm 0.025, C = 1.4 \pm 0.033\}$) are compared against SVD when faced with random and systematic error ($\sigma_{data} = 0.471$). The "True" parameters were set to be $\{A = 6.51, B = -1.76, C = 1.6\}$, the same as Figure S2.

Figure S4. **The BICePs score is a measure of structural ensemble quality.** Using the same toy model system described in the main text. We vary the quality of the prior structural ensemble (σ_{prior}) by perturbing the "true" ϕ angles of the structural ensemble. In these experiments, we induced over 1,000 random perturbations to the prior structural ensemble, and calculated BICePs scores for each. Error bars represent the standard deviation from the mean. The top panel shows the relationship between the BICePs score and the amount of error added to the structural ensemble. Each data point is an average across of 100 BICePs calculations. In these calculations, we used the Good-Bad likelihood model with 32 replicas.

Figure S5. **Optimization of ξ-values to enhance free energy estimation**. The y-axis is thermodyamic length, $|\ell(\xi_{n+1}) - \ell(\xi_n)|$ for each pair of intermediates, derived from the variance in distributions $p(\Delta u_{n,n+1})$. The top panel shows ξ_i values before optimization. The bottom panel shows the optimized ξ_i^* values. Optimized ξ-values have shifted to smaller values to make up for the lack of overlap between these adjacent thermodynamic ensembles.

Figure S6. The overlap matrix O_{ij} between thermodynamic states i and j for intermediates along the ($\xi = 0 \rightarrow 1$) transformation, **before** optimization of ξ-values. The overlap matrix (Klimovich et al 2015) is a Monte Carlo estimate of the average probability of samples in thermodynamic state j being observed in thermodynamic state i: $O_{ij} = \left\langle \frac{N_i p_i(\mathbf{x})}{\sum_{k=1}^{K} N_k p_k(\mathbf{x})} \right\rangle_j$, where N_i are the number of samples from thermodynamic state i and $p_i(\mathbf{x})$ are posterior probabilities. Significant off-diagonal overlap provides evidence of reliable free energy estimates.

Figure S7. The overlap matrix O_{ij} between thermodynamic states i and j for intermediates along the ($\xi = 0 \rightarrow 1$) transformation, **after** optimization of ξ-values. Stronger overlap is clearly demonstrated by comparison of post- and pre-optimization (Figure S6).

Figure S8. **Variational minimization of the BICePs score can be used to optimize forward model parameters.** By variational minimzation of the BICePs score, f optimization traces converge to the "true" parameters. In these calculations, ξ-values were optimized prior to running the parameter refinement. We used the Good-Bad model with 4 replicas to minimize computational cost. Karplus coefficients predicted using the Good-Bad likeliood model ($\{6.31 \pm 0.02, -1.69 \pm 0.03, 1.69 \pm 0.01\}$) are compared against SVD when faced with random and systematic error ($\sigma_{data} = 0.471$). The "True" parameters were set to be $\{A = 6.51, B = -1.76, C = 1.6\}$, the same as Figure S2.

Figure S9. Karplus curves with BICePs refined Karplus coefficients using the 1d3z ensemble for (a-f) $^3J_{C'C'}$, $^3J_{C'C^\beta}$, $^3J_{H^\alpha C'}$, $^3J_{H^N C'}$, $^3J_{H^N C^\beta}$, $^3J_{H^N H^\alpha}$. BICePs calculations were run using four chains with 32 replicas each, where we burned 50k steps, then sampled for another 50k MCMC steps. For comparison, SVD on 1ubq using experimental scalar coupling constants with ϕ-angles derived from the X-ray structure (black dashed line) and red dots correspond to the fitted data points. Additionally, parameterizations from Bax et al. 1997 (green), and parameterization from Habeck et al. 2005 (yellow) were overlaid for comparison. The thickness of the line corresponds to the uncertainty.
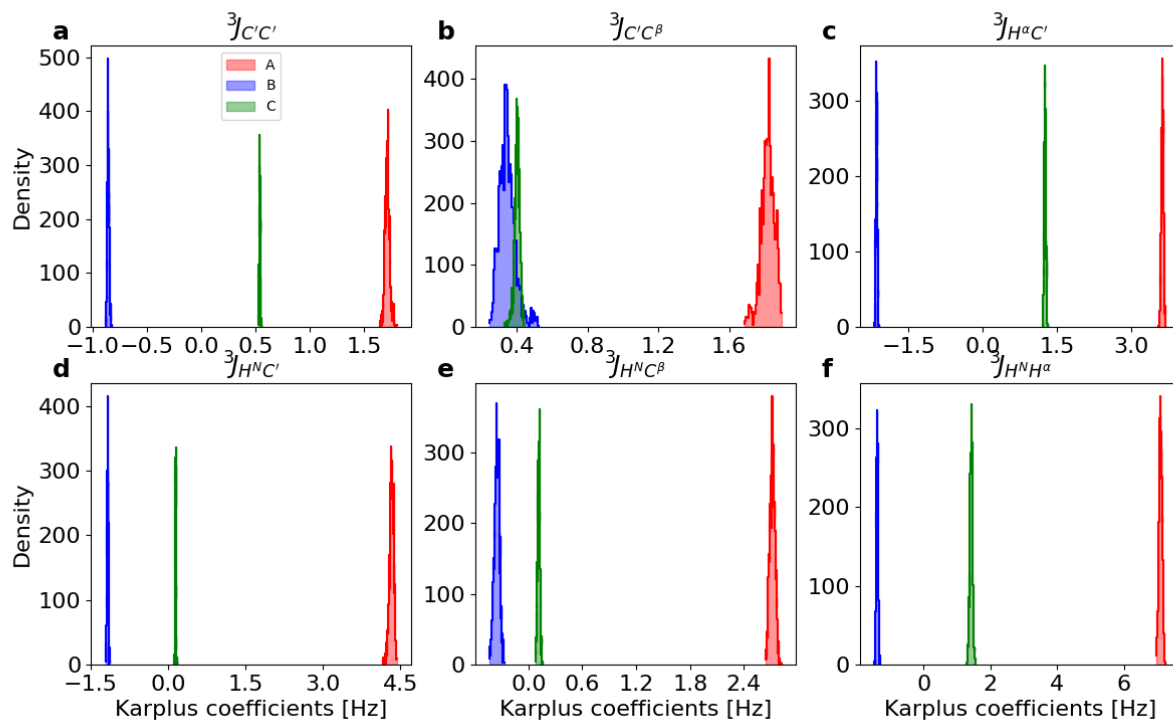
Figure S10. **Sampling the joint posterior distributions of six sets of Karplus coefficients using the Good-Bad model on the 1d3z ensemble.** BICePs calculations were run using four chains with 32 replicas each, where we burned 50k steps, then sampled for another 50k MCMC steps. These marginal posterior distributions shown here are from a randomly selected chain.
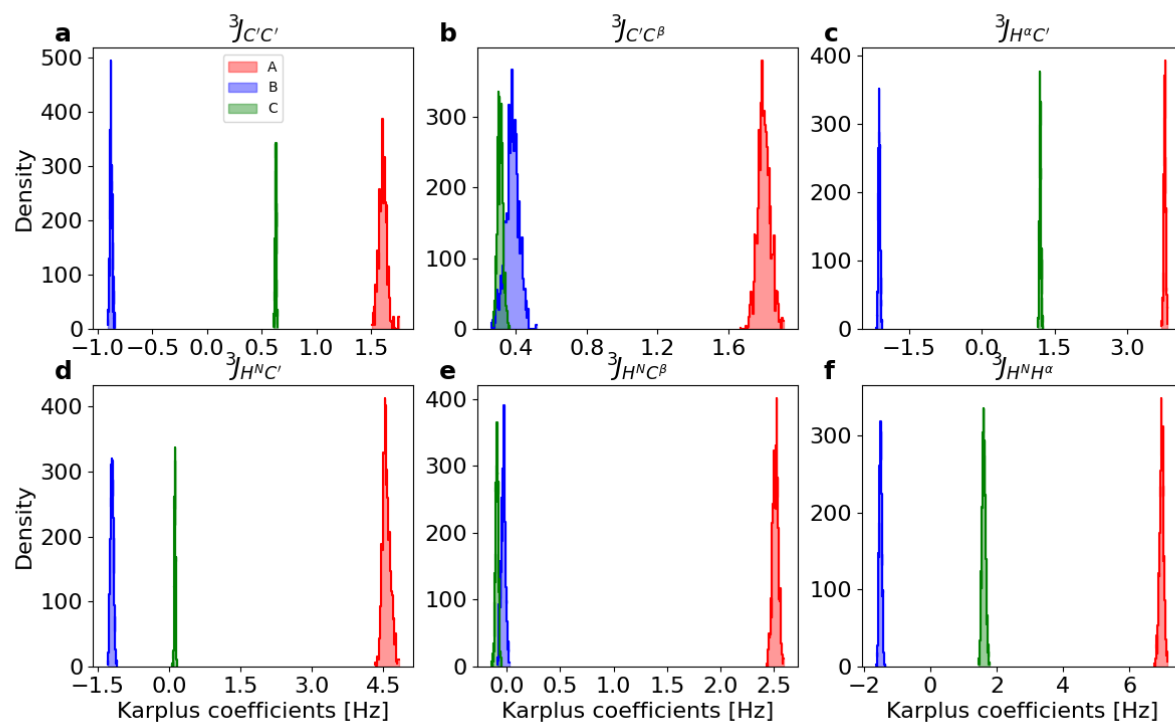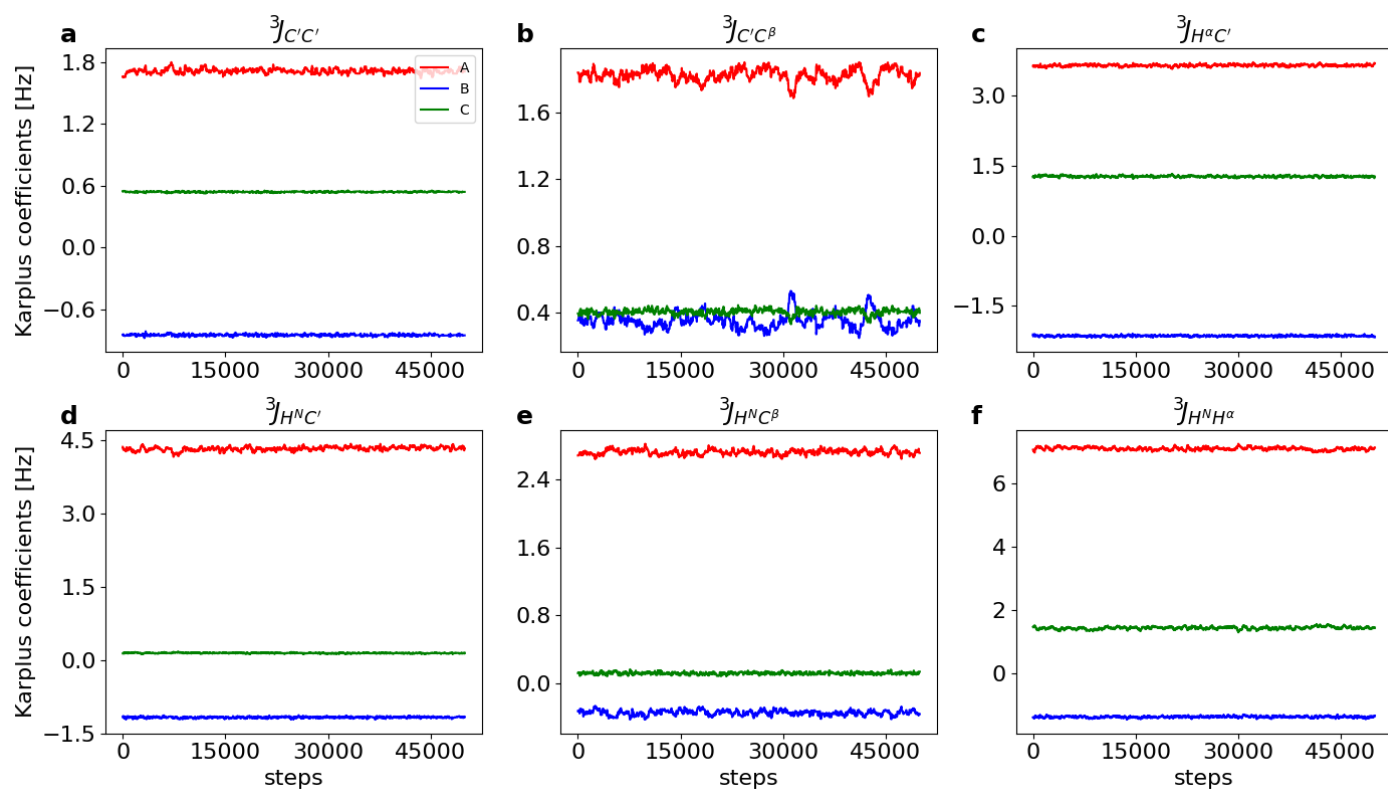
Figure S11. **Sampling the joint posterior distributions of six sets of Karplus coefficients using the Good-Bad model on the RosettaFold2 (RF2) ensemble.** BICePs calculations were run using four chains with 32 replicas each, where we burned 50k steps, then sampled for another 50k MCMC steps. Compare with Figure S10 to see similarities.

Figure S12. Traces of sampled Karplus coefficients for the 1d3z ensemble over 50k steps of MCMC, post-burn. BICePs calculations used the Good-Bad model with 32 replicas. Traces display low variance with no jumps, which demonstrates converged samples

Figure S13. The marginal posterior distribution of σ_J, the uncertainty parameters for each set of J-coupling in the 1d3z ensemble. Densities are a result of posterior sampling of FM parameters during ensemble refinement using the Good-Bad model with 32 replicas. The marginal posterior distributions of the variance scaling parameter φ has a sampled mean slightly larger than 1.0 for particular sets of J-coupling, indicating that the functional form of the likelihood opted for long tails to account for a few outlier data points deviating from the mean.
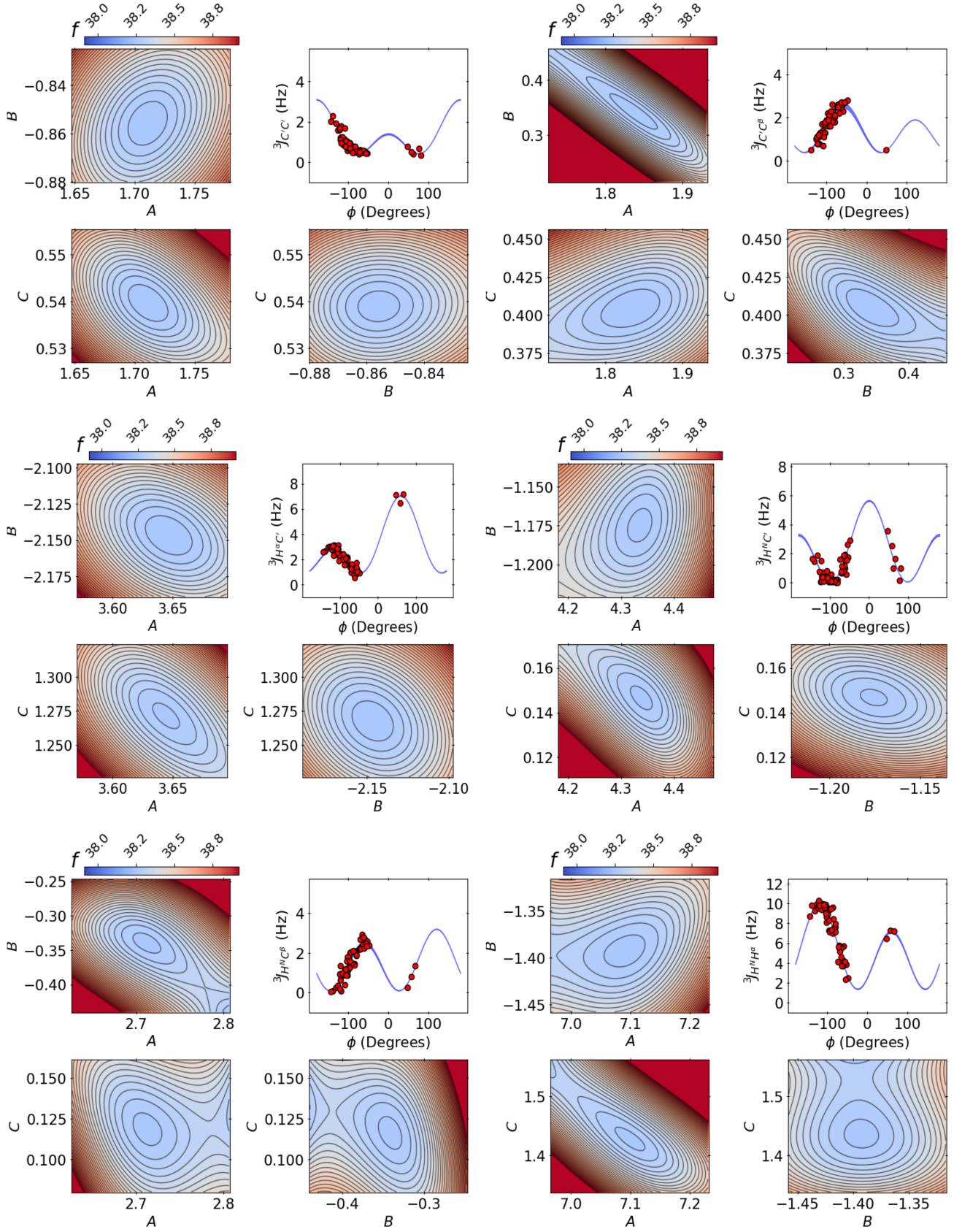
Figure S14. BICePs score landscapes of FM parameters on the 1d3z ensemble, unveiled during ensemble refinement. BICePs calcualtions used the Good-Bad model with 32 replicas. Each set of $\{A, B, C\}$ was included in the joint posterior of FM parameters.
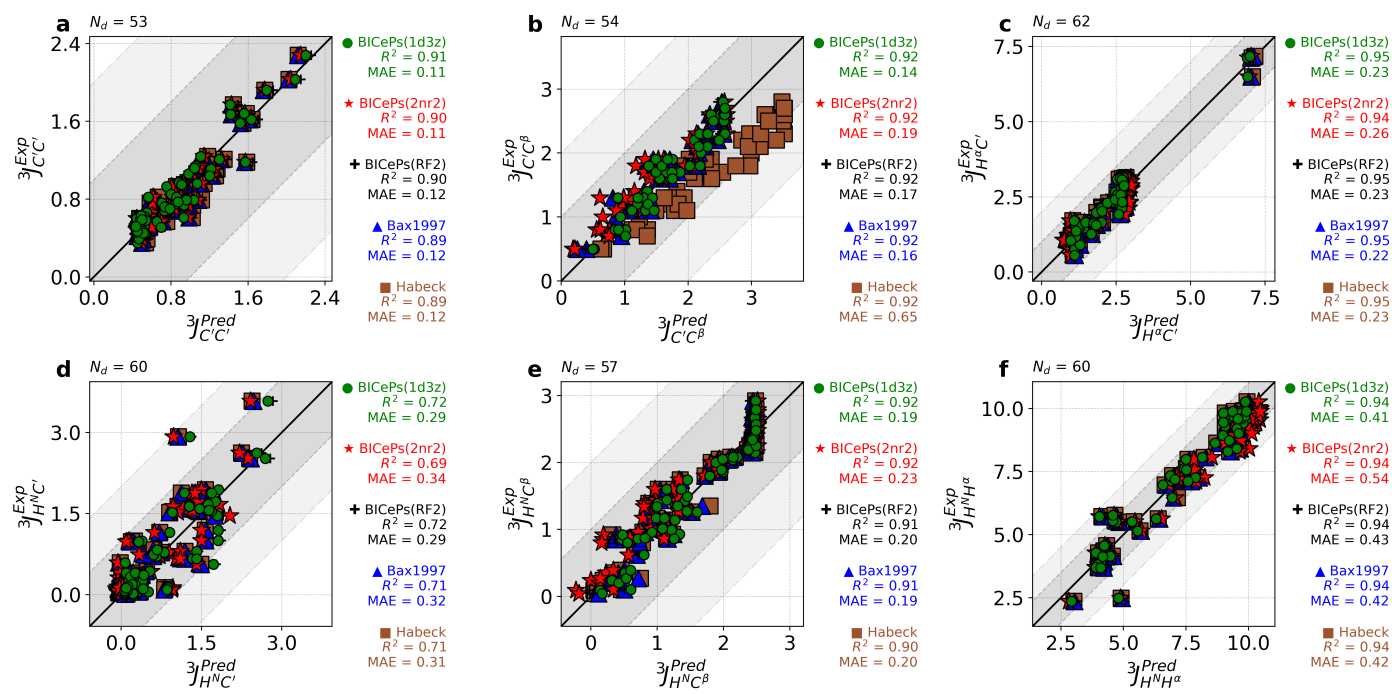
Figure S15. Validation of refined Karplus coefficients using BICePs on the 1D3Z structural ensemble show similar results to Bax1997 and minor improvements over Habeck2005 for scalar coupling predictions. Here, we compare models for predicting six sets of scalar coupling constants. Each panel shows strong correlations and relatively low error.
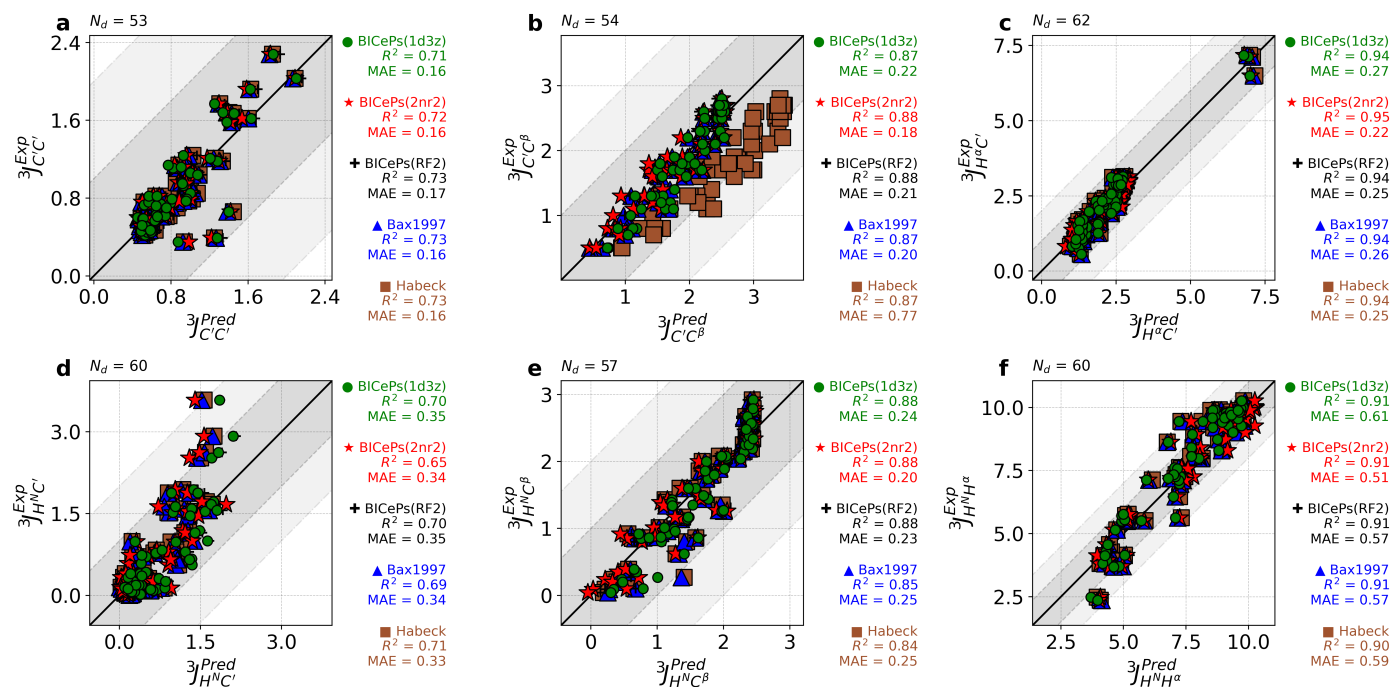

Figure S16. Validation of refined Karplus coefficients using BICePs on the 2NR2 structural ensemble show similar results to Bax1997 and minor improvements over Habeck2005 for scalar coupling predictions. Here, we compare various models for predicting six sets of scalar coupling constants. Each panel shows strong correlations between predictions and experiment.
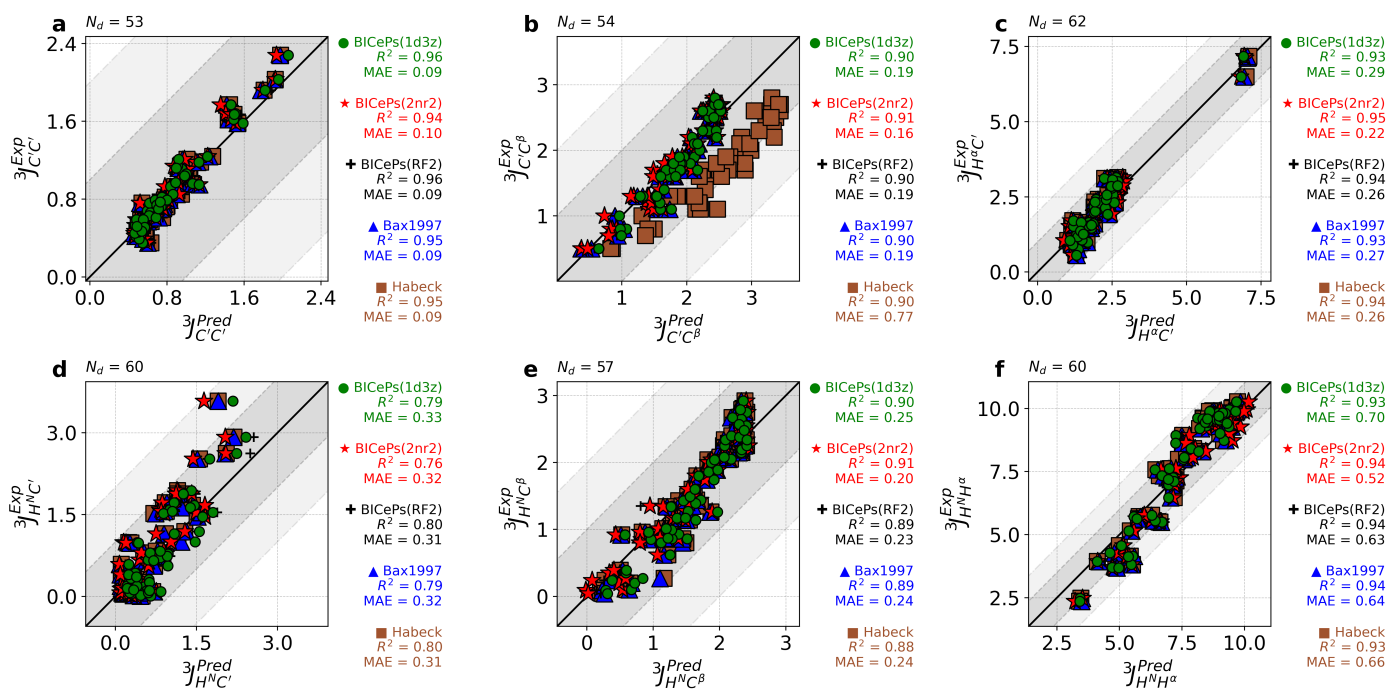
Figure S17. Validation of refined Karplus coefficients using BICePs on the CHARMM22* structural ensemble show similar results to Bax1997 and minor improvements over Habeck2005 for scalar coupling predictions. Here, we compare various sets parameters for predicting six sets of scalar coupling constants. Each panel shows strong correlations between predictions and experiment. On average, BICePs parameters derived from the 2NR2 ensemble give the lowest MAE between experiment and predictions, whereas Habeck2005 has the highest due to $^3J_{C'C^\beta}$.