Automatic forward model parameterization with Bayesian inference of conformational populations

Cite as: APL Mach. Learn. 4, 016102 (2026); doi: 10.1063/5.0287423
Submitted: 24 June 2025 • Accepted: 26 November 2025 •
Published Online: 8 January 2026

View Online Export Citation CrossMark

Robert M. Raddi, Tim Marshall, and Vincent A. Voelz[a]

AFFILIATIONS

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

[a]Author to whom correspondence should be addressed: vvoelz@temple.edu

ABSTRACT

To quantify how well theoretical predictions of structural ensembles agree with experimental measurements, we depend on the accuracy of forward models (FMs). 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 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 [R. M. Raddi *et al.*, J. Chem. Theory Comput. **21**, 5880–5889 (2025) and R. M. Raddi and V. A. Voelz, "Automated optimization of force field parameters against ensemble-averaged measurements with Bayesian inference of conformational populations," arXiv:2402.11169 (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 ${}^3J_{H^NH^\alpha}$, ${}^3J_{H^\alpha C'}$, ${}^3J_{H^NC^\beta}$, ${}^3J_{H^NC'}$, ${}^3J_{C'C^\beta}$, and ${}^3J_{C'C'}$. Finally, we demonstrate that our framework naturally generalizes optimization to any differentiable FM, such as those constructed with neural networks. This approach provides a promising direction for training and validating neural network-based FMs.

© 2026 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0287423

I. INTRODUCTION

In the field of molecular modeling and dynamics, the accuracy of theoretical predictions is crucial. The quantitative agreement between theory and experiment is typically characterized through the use of accurate *forward models* (FMs)—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 nuclear magnetic resonance (NMR) observables critically depends on reliable forward models (FMs) that have been robustly parameterized to reduce error. In refining these FMs, it is important to consider random and systematic errors inherent to the experimental data, which in turn contribute to overall uncertainty of integrative models. It is usually challenging to obtain complete knowledge of such errors; instead, such errors typically need to be inferred from the data.

A further challenge is presented by missing or insufficient examples of known structures that 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 have been developed to address some of these challenges.[1–4] 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 is a reweighting algorithm to refine structural ensembles against sparse and/or noisy experimental observables, which has been used in many previous applications.[7–10] BICePs infer all possible sources of error by sampling the posterior distribution of these parameters directly from the data through Markov chain Monte Carlo (MCMC) sampling. BICePs also compute a free energy-like quantity called the BICePs score that can be used for model selection and model parameterization.[6,11,12]

Recently, BICePs were 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 is particularly powerful when used with specialized likelihood functions that allow for the automatic detection and down-weighting of the importance of experimental observables subject to systematic error.[6]

Beyond structural biology, Bayesian methods are widely used in survival analysis to model censored and heterogeneous discrete data. Recent work by Akhtar and colleagues shows that flexible priors and robust loss functions enable accurate parameter estimation and outlier resistance in such settings.[13,14] These developments highlight the versatility of Bayesian inference for complex, noisy datasets and provide additional examples of analogous strategies for handling heterogeneity and outliers in our domain.

To effectively refine FM parameters, we use BICePs to sample over the entire posterior distribution of FM parameters, enabling simultaneous ensemble reweighting and parameter refinement. In addition, 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 using a toy model and then apply it to optimize six distinct sets of Karplus parameters for the human protein ubiquitin. As a proof-of-concept for more complex models, we also train a neural network to learn J-coupling constants from dihedral angles using the BICePs' score as a loss function. We then discuss how this idea could be applied to neural networks for chemical shift prediction.[15–18] Of course, our framework naturally extends to other empirical forward models with tunable parameters—for example, hydrogen–deuterium exchange (HDX) models, with adjustable hydrogen-bonding and burial weights, or paramagentic relaxation enhancement (PRE) models with system-specific correlation times and scaling factors. These examples illustrate the generality and flexibility of our approach for refining a wide range of forward models to improve the agreement between simulations and experiment.

II. THEORY

In this section, we develop two different theoretical approaches to optimize forward models. In the first method, we include forward model parameters in the posterior and sample the full posterior distribution of all parameters. In the second approach, we optimize the forward model by minimizing the BICePs score, a free energy-like quantity that can be used for variational optimization. While the two methods are theoretically equivalent (see the supplementary material for a short proof of this), each has practical advantages and disadvantages.

A. Forward model optimization by posterior sampling of forward model parameters

BICePs uses a Bayesian statistical framework, inspired by Inferential Structure Determination (ISD),[19] 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.

For a specific forward model $g(X,\theta)$ with a set of FM parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$, where m is the number of parameters, we can additionally include the parameters 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 used with a replica-averaged forward model, it becomes a MaxEnt reweighting method in the limit of large numbers of replicas.[20–25] 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 the observable j estimated using the standard error of the mean (SEM):[20,25] $\sigma_j^{\mathrm{SEM}} = \sqrt{\frac{1}{N}\sum_r^N \left(g_j(X_r,\theta) - \langle g_j(\mathbf{X},\theta)\rangle\right)^2}$. Thus, σ_j^{SEM} decreases as the square root of the number of replicas.

For a single forward model with parameters θ, the joint posterior distribution for all parameters is

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

where $\mathbf{X}$ is a set of N conformation replicas. In many cases, we wish to optimize several forward models at once. For example, there are unique sets of Karplus parameters in the forward models for vicinal J-coupling constant observables ${}^3J_{H^N H^\alpha}$, ${}^3J_{H^\alpha C'}$, and ${}^3J_{H^N C^\beta}$. If there are K different forward models $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}$, each with their own set of parameters and error distributions, we can describe their full joint posterior distribution as

© Author(s) 2026

$$p(X,\sigma,\theta^{(1)},\theta^{(2)},\ldots\theta^{(K)}|D) \propto \prod_{r=1}^{N} p(X_r)\prod_{k=1}^{K} p(D_k|g(\mathbf{X},\theta^{(k)}),\sigma_k) \times p(\sigma_k)p(\theta^{(k)}), \tag{4}$$

where D_k represents the experimental observables predicted by the k^{th} forward model, $\theta^{(k)}$ represents the parameters for the k^{th} FM, and $\sigma_k = \sqrt{(\sigma_k^{\mathrm{SEM}})^2 + (\sigma_k^{\mathrm{B}})^2}$, describing the total error for the kth observable. The total error originates from both finite samplings through the standard error of the mean (σ_k^{SEM}) 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,

$$p(\theta|D) = \sum_{X} \int p(X,\sigma,\theta|D)d\sigma. \tag{5}$$

Gradients speed up convergence: In our methodology, Markov chain Monte Carlo (MCMC) is used to sample the posterior with acceptances following the Metropolis–Hastings 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 following equation:

$$\theta_{\mathrm{trial}} = \theta_{\mathrm{old}} - l_{\mathrm{rate}} \cdot \nabla u + \eta \cdot \mathcal{N}(0,1), \tag{6}$$

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 calculated gradient of BICePs energy function $u = -\ln p(X,\sigma,\theta|D)$ with respect to the parameters θ; and η scales the noise drawn from a standard normal distribution $\mathcal{N}(0,1)$.

This parameter update strategy is designed to satisfy the Metropolis–Hastings criterion, ensuring each step in the parameter space not only moves toward 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) predicting observables using perturbed FM parameters and computing the total energy, and (4) assessing the new energy against the previous one to determine acceptance based on the Metropolis–Hastings criterion. This ensures a thorough and effective search of the parameter space, taking advantage of both the landscape topology and stochastic elements to guide the exploration.

The Good-and-Bad model accounts for systematic error due to outlier measurements. Recent work has developed sophisticated likelihood models for BICePs, designed to detect outliers that are robust in the presence of systematic error.[6] Raddi *et al.* developed two likelihood models for this purpose: the Student's model and the Good-and-Bad model. In related work using BICePs to perform automated force field optimization, we used Student's model.[12] Here, in this work, we use the Good-and-Bad model, primarily because the derivatives are less complicated than those of Student's model.

The Good-and-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 in which the uncertainties σ_j for particular observables j are distributed with respect to some typical uncertainty σ^B according to a conditional probability $p(\sigma_j|\sigma^B)$. We derive a posterior for the parameters of the kth FM having a single uncertainty parameter σ^B by marginalizing over all σ_j,

$$p(\mathbf{X},\sigma_0,\theta^{(k)}|D) \propto \prod_{r=1}^{N} p(X_r)\prod_{j=1}^{N_d} \int_{\sigma^{SEM}}^{\infty} p(d_j|g_j(\mathbf{X},\theta^{(k)}),\sigma_j)p(\sigma_j|\sigma_0)d\sigma_j, \tag{7}$$

where $\sigma_0 = \sqrt{(\sigma^B)^2 + (\sigma^{SEM})^2}$. Under the Good–Bad model, we say that the "good" data consist of observables normally distributed about their true values with effective variance σ_0^2, while the "bad" data are 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 Eq. (7) becomes

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

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

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

$$= \prod_{r=1}^{N}\left\{p(X_r)\prod_{j=1}^{N_d}\left(\frac{1-H(\sigma^{SEM}-\sigma_0)}{2}\mathcal{N}\left(g_j(\mathbf{X},\theta^{(k)})|\, d_j,\sigma_0^2\right) + \frac{1-H(\sigma^{SEM}-\varphi\sigma_0)}{2\varphi}\mathcal{N}\left(g_j(\mathbf{X},\theta^{(k)})|\, d_j,\varphi^2\sigma_0^2\right)\right)\right\}, \tag{9}$$

© Author(s) 2026

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, $u = -\log p(\mathbf{X}, \sigma_0, \varphi, \theta^{(k)}|D)$, the negative logarithm of the posterior in its full form is given by

$$u = \sum_{r=1}^{N} -\log\left(p(X_r)\right) - N\sum_{j=1}^{N_d}\log\left[\frac{1 - H(\sigma^{SEM} - \sigma_0)}{2}\mathcal{N}\left(g_j(\mathbf{X}, \theta^{(k)})|\, d_j, \sigma_0^2\right) + \frac{1 - H(\sigma^{SEM} - \varphi\sigma_0)}{2\varphi}\mathcal{N}\left(g_j(\mathbf{X}, \theta^{(k)})|\, d_j, \varphi^2\sigma_0^2\right)\right]. \tag{10}$$

When $\varphi = 1$, this energy function is the same result we would get if we used a Gaussian likelihood function,

$$u = \sum_{r=1}^{N} -\log\left(p(X_r)\right) + N\left[\sum_{j=1}^{N_d} -\log\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) + \frac{(d_j - g_j(\mathbf{X}, \theta))^2}{2\sigma_0^2} - \log\left(p(\sigma_0)\right)\right]. \tag{11}$$

The first derivative of Eq. (10) with respect to the ith FM parameter in the parameter set $\theta^{(k)}$ is

$$\frac{\partial u}{\partial \theta_i^{(k)}} = N\sum_{j=1}^{N_d}\frac{\partial g_j(\mathbf{X}, \theta^{(k)})}{\partial \theta_i^{(k)}}\frac{(d_j - g_j(\mathbf{X}, \theta^{(k)}))}{\sigma_0^2} \times \left(\frac{\varphi^2\left(1 - H\left(\sigma^{SEM} - \sigma_0\right)\right)\mathcal{N}_{\varphi^2}^{-1} + \left(1 - H\left(-\varphi\sigma_0 + \sigma^{SEM}\right)\right)\mathcal{N}_1^{-1}}{\varphi^2\left(H\left(\sigma^{SEM} - \sigma_0\right) - 1\right)\mathcal{N}_{\varphi^2}^{-1} + \varphi^2\left(H\left(-\varphi\sigma_0 + \sigma^{SEM}\right) - 1\right)\mathcal{N}_1^{-1}}\right), \tag{12}$$

where $\mathcal{N}_1^{-1} = \mathcal{N}\left(g_j(\mathbf{X}, \theta^{(k)})|\, d_j, \sigma_0^2\right)^{-1}$ and $\mathcal{N}_{\varphi^2}^{-1} = \mathcal{N}\left(g_j(\mathbf{X}, \theta^{(k)})|\, d_j, \varphi^2\sigma_0^2\right)^{-1}$. Both represent the reciprocal of Gaussian likelihoods, where $\mathcal{N}_1^{-1}$ and $\mathcal{N}_{\varphi^2}^{-1}$ denote inverse-weighted forms of normal distributions with variances σ_0^2 and $\varphi^2\sigma_0^2$, respectively.

When $\varphi = 1$, the gradient becomes

$$\frac{\partial u}{\partial \theta_i^{(k)}} = -N\sum_{j=1}^{N_d}\frac{\partial g_j(\mathbf{X}, \theta^{(k)})}{\partial \theta_i^{(k)}}\frac{(d_j - g_j(\mathbf{X}, \theta^{(k)}))}{\sigma_0^2}. \tag{13}$$

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, but the derivation is straightforward. For the forward models, we consider in the following (Karplus relations) all have second derivatives that go to zero. For more general cases, see the supplementary material, Appendix A for more details. The energy of the Good–Bad likelihood model and its first and second derivatives are shown in Fig. S1.

B. Forward model optimization by variational minimization of the BICePs score

As we show in Sec. III and Sec. IV, treating the forward model parameters as nuisance parameters and sampling them as part of the full posterior is an efficient strategy to refine FM parameters. However, in the limit of a very large number of FM parameters, the dimensionality of the posterior may become unwieldy and may suffer from the "curse of dimensionality."

As an alternative strategy, we consider optimization of forward model parameters using variational minimization of the BICePs score, an approach that we have used to perform 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, Eq. (2) becomes

$$p(X, \sigma|D, \theta) \propto p(D, \theta|X, \sigma)p(X)p(\sigma). \tag{14}$$

In this view, ensemble refinement is performed with a static set of FM parameters for each BICePs calculation. The BICePs score then enables the selection of the optimal model parameterized by θ.

BICePs evaluate 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,

$$f(\theta) = -\ln\left(Z(\theta)/Z_0\right), \tag{15}$$

where

$$Z(\theta) = \iint \exp\left(-u(\mathbf{X}, \sigma; \theta)\right)d\mathbf{X}d\sigma \tag{16}$$

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 Eq. (10). 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 free energy of "turning on" experimental restraints ($\xi = 0 \to 1$).

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 priors $p_\lambda(X) \sim [p(X)]^\lambda$ and the BICePs score is computed as the free energy of ($\lambda = 0 \to 1$) and ($\xi = 0 \to 1$) transformations. Here, since we are only interested in evaluating and/or parameterizing the likelihood functions, we set $p(X)$ to be uniform.

To optimize forward model parameters, the BICePs score is used as an objective function that is estimated given a candidate value of θ. Based on the estimate (and possibly other previous estimates), new candidates for θ are iteratively proposed and accepted or rejected, with the goal of finding the values θ^* that minimize the BICePs score.

For this purpose, gradient-based searches are very helpful. Conveniently, the derivative of the BICePs score $f(\theta)$ with respect to the FM parameters reduces to the Boltzmann-averaged values of the gradient $\partial u/\partial\theta$,

$$\frac{\partial f(\theta)}{\partial \theta_i} = \iint \frac{1}{Z(\theta)}\left[\frac{\partial u}{\partial \theta_i}\right]\exp\left(-u\right)d\mathbf{X}d\sigma = \left\langle\frac{\partial u}{\partial \theta_i}\right\rangle. \tag{17}$$

© Author(s) 2026

In our results in the following, we use the first-order optimization method L-BFGS-B,[26] which only requires first derivatives. For more complex forward models, second-order methods may be needed. The Hessian matrix (of second partial derivatives of the BICePs score with respect to FM parameters) can be used for uncertainty estimation when $f(\theta)$ is convex. Assuming this is true, we estimate uncertainties in the FM model parameters using covariances obtained through inversion of the Hessian. Conceptually, this means that the estimated uncertainties in the best-fit values represent the widths of the basins in 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.

Calculation of the BICePs score (a free energy difference) and its derivatives (expectation values of energy derivatives) is performed using the multistate Bennett acceptance ratio (MBAR) free energy estimator[27] by sampling at several intermediates $\xi = 0 \rightarrow 1$, which enables accurate estimates of all quantities. 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,[28,29] employing a strategy akin to the "thermodynamic trailblazing" method proposed by Rizzi.[30] This optimization is performed using an algorithm we have developed called `pylambdaopt`.[31]

III. RESULTS

A. A toy model to test the performance of forward model optimization

To investigate the performance of BICePs in optimizing forward models, we introduce a simplified, yet physical, toy model. This model is designed to mimic the complexity of protein backbone structures by generating ϕ-angles from a multi-modal distribution, thereby emulating configurations characteristic of different secondary structural elements (Fig. 1). This distribution encompasses three distinct modes, each characterized by a mean (μ), standard deviation (σ), and weight (w): beta sheets ($\mu = -110°$, $\sigma = 20°$, and $w = 0.35$), right-handed helices ($\mu = -60°$, $\sigma = 10°$, and $w = 0.5$), and left-handed helices ($\mu = 60°$, $\sigma = 5°$, and $w = 0.15$). These parameters were chosen to accurately reflect the structural variability found in proteins. Angles ϕ_i were sampled from the multi-modal distribution,

$$p(\boldsymbol{\phi}|\mu,\sigma) = \sum_l w_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(-\frac{(\phi-\mu_l)^2}{2\sigma_l^2}\right). \tag{18}$$

The sampled ϕ_i were then used to calculate synthetic ensemble-averaged experimental J-coupling constants, ${}^3J(\phi)$, using the Karplus relation,

$${}^3J(\phi) = A\cos^2(\phi) + B\cos(\phi) + C, \tag{19}$$

with the coefficients set to their chosen *true* values: ($A^* = 6.51$, $B^* = -1.76$, and $C^* = 1.6$). The synthetic J-coupling constant observables were calculated using the ensemble-average, as $\sum_i {}^3J(\phi_i)p(\phi_i)$.

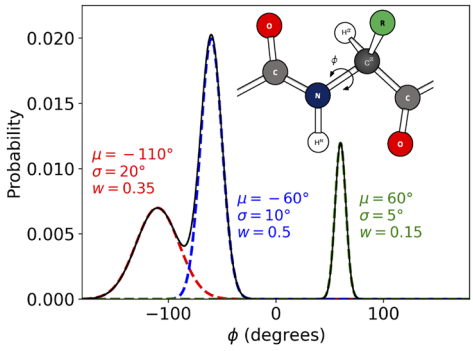


FIG. 1. Toy model for measuring the performance of forward model optimization. The ϕ-angles for each conformational state is pulled from a multi-modal distribution and corresponding energies. (a) This multi-modal distribution of ϕ-angles was intended to represent configurations with different secondary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

B. Including forward model parameters in the BICePs posterior is a robust method to optimize Karplus parameters in the presence of experimental errors

Using the toy model described above, we next perform tests to examine whether including forward model parameters in the BICePs posterior can be used to find the optimal Karplus parameters starting from an initial reference forward model with parameters (A_0, B_0, C_0). Through these tests, we evaluate the precision of the results and how robust the algorithm is in the presence of noise.

To evaluate the robustness of our algorithm in the presence of 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: Parameters of many reference Karplus relations were previously determined using a singular value determination approach.[32] To test this approach against our BICePs method, we derived 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,

$$M = \begin{bmatrix} \sum_X p(X)\cos^2(\phi_{1,X}+\phi_0) & \sum_X p(X)\cos(\phi_{1,X}+\phi_0) & 1 \\ \sum_X p(X)\cos^2(\phi_{2,X}+\phi_0) & \sum_X p(X)\cos(\phi_{2,X}+\phi_0) & 1 \\ \vdots & \vdots & \vdots \\ \sum_X p(X)\cos^2(\phi_{N_d,X}+\phi_0) & \sum_X p(X)\cos(\phi_{N_d,X}+\phi_0) & 1 \end{bmatrix}, \tag{20}$$

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

$$\theta = V(\Sigma + \varepsilon I)^{-1} U^T J^{\text{exp}}, \tag{21}$$

© Author(s) 2026

where $\varepsilon = 1 \times 10^{-6}$ is 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 1000 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 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 starting 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 accuracies.

Our findings indicate that the Good–Bad likelihood model (red) exhibits superior insensitivity to experimental errors compared to a traditional Gaussian likelihood model (blue) and SVD (green) approaches (Fig. 2). Predictions from SVD and the Gaussian likelihood model become notably less dependable when the experimental data contain 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}.

An example of a single trial of forward model parameter refinement using the toy model is shown in the supplementary material Fig. S2), where BICePs predict Karplus coefficients by posterior sampling over FM parameters. Both BICePs and SVD methods successfully reproduce the "true" Karplus curve. However, BICePs can accurately identify the error present in the data ($\sigma_{\text{data}} = 0.471$ Hz), 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 (supplementary material, Fig. S3). The computed

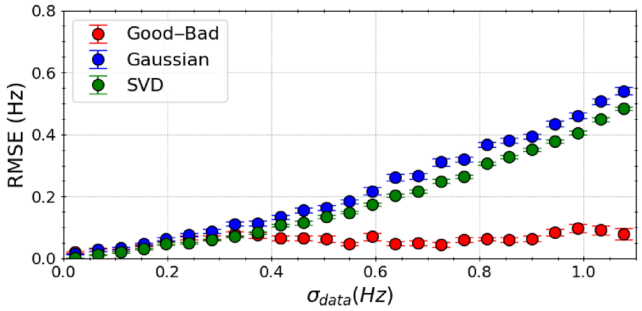


FIG. 2. Comparative analysis in the performance of the Good–Bad likelihood model (red), a Gaussian likelihood model (blue), and singular value decomposition (SVD) using the "true" ϕ angles with synthetic experimental data. Here, we induced random and systematic errors of varying magnitudes (σ_{data}) 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 1500 random perturbations to the experimental data, and represents the average of 100 BICePs calculations. The error bars represent the standard deviation. Predictions from SVD and the Gaussian likelihood model become notably less dependable when data incorporate errors, especially when σ_{data} exceeds 0.5 Hz.

Gelman–Rubin ($\hat{R}$) statistic[33] 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 (supplementary material, Fig. 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 1000 random perturbations to the ϕ angles with errors up to $\sigma_{\text{prior}} = 4^\circ$ and perturbations to the experimental data $\sigma_{\text{data}} = 0.68 \pm 0.24$ Hz.

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 predicted FM parameters derived from sub-optimal structural ensembles remain reliable. In addition, our findings reveal a strong correlation between the BICePs score and the quality of the structural ensemble, demonstrating an immense utility in this context.

C. Variational minimization of the BICePs score to optimize FM parameters and comparison with posterior sampling

Next, we performed variational minimization of the BICePs score to determine optimal FM parameters for our toy model (see the supplementary material, Fig. S2) and compared the results to the posterior sampling approach under the same conditions.

© Author(s) 2026

Variational minimization was performed using the Good–Bad model with four replicas (for reduced computational cost), where each evaluation of the objective function consisted of running 10 000 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 three independent runs with very low variance between runs (supplementary material, Fig. S6). 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 show nearly 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 basic Monte Carlo Markov chain (MCMC) methods. In addition, 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 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 of the 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.

D. Determination of optimal Karplus coefficients for ubiquitin

As an application 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 the robustness of our algorithm, we conducted a comprehensive evaluation using three different structural ensembles as priors, each derived from distinct computational approaches: (1) 10 conformations from the NMR-refined structural ensemble, 1D3Z,[34] (2) 144 conformations from NMR-restrained simulations, 2NR2,[35] and (3) 25 conformations from the RosettaFold2 (RF2) algorithm.[36] Further details about these ubiquitin structural ensembles are given in the supplementary material.

We 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 validation calculations, we used three independent structural ensembles: 1D3Z and 2NR2 (described above) and a 500-state conformational ensemble derived from a millisecond-long simulation of ubiquitin using CHARMM22*.[37] Further details about the CHARMM22* ensemble are given in the supplementary material.

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). Estimates of FM parameters were made by averaging the BICePs results over four Markov chain Monte Carlo (MCMC) sampling runs, 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

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

		ϕ_0 (deg)	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

© Author(s) 2026

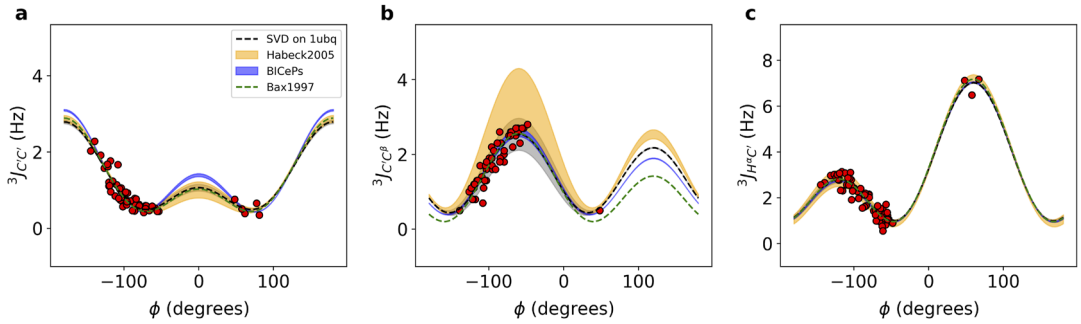


FIG. 3. Karplus curves with BICePs-refined Karplus coefficients using the 1d3z ensemble for (a)–(c) $^3J_{C'C'}$, $^3J_{C'C^\beta}$, and $^3J_{H^\alpha C'}$. For comparison, SVD on 1ubq using experimental scalar coupling constants with ϕ-angles derived from the x-ray structure (black dashed line) and the red dots corresponding to the fitted data points are shown. In addition, parameterizations from Bax *et al.* (green) and parameterization from Habeck *et al.* (yellow) were overlaid for comparison. The thickness of the line corresponds to the uncertainty.

Karplus curves for each scalar coupling demonstrate significant congruence with the historical NMR refinement results.[3,34] For all six types of J-coupling, see the supplementary material, Fig. S7.

The predicted parameters have large similarities across structural ensembles. BICePs-predicted coefficients using the 1D3Z ensemble (supplementary material, Fig. S8) and predicted coefficients using the RF2 ensemble (supplementary material, Fig. S9) are found to strongly overlap. Furthermore, traces of the FM parameters over time confirm that the sampling converges (supplementary material, Fig. S10).

A key advantage of BICePs is its ability to sample the posterior densities of forward model uncertainties, $p(\sigma)$ (supplementary material, Fig. S11). For certain sets of J-coupling constants (e.g., $^3J_{H^NH^\alpha}$ and $^3J_{H^NC'}$), the marginal posterior distribution of the variance scaling parameter $p(\varphi)$ has a sampled 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 $^3J_{H^NC'}$ Karplus parameters: In Fig. 4, we show the free energy landscape, which is also equivalent to the BICePs score landscape $f_{\xi=0\to1}$. The Karplus relation for $^3J_{H^NC'}$ was found to overlap strongly with the results obtained by SVD when using ϕ angles from the x-ray crystal structure (supplementary material, Fig. S7). The red data points are shown using the experimental J-couplings with ϕ angles derived from x-ray crystal pose 1UBQ.[38] The joint BICePs score landscape for the six sets of parameters is too complex to visualize. Instead, 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 with an additive white noise to account for observational uncertainty. The characteristic length scale was bounded within $[0.1, 10.0]$; the signal variance was set to 1.0; and the noise variance was set to 10^{-5}. 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 the supplementary material, Fig. S12.

To demonstrate the transferability across different generative models and validate our parameters, we evaluated the accuracy of the

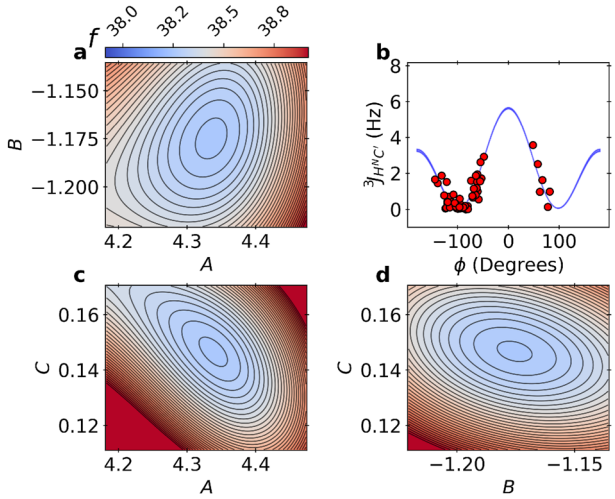


FIG. 4. Landscapes of the BICePs score with respect to the predicted Karplus coefficients for $^3J_{H^NC'}$. Panels (a), (c), and (d) illustrate the energy landscape f for pairs of Karplus coefficients when using the 1D3Z structural ensemble during refinement. Panel (b) shows the Karplus model using parameters with the lowest BICePs score, with red circles showing the experimental J-coupling values and corresponding ϕ angles from the x-ray crystal structure of ubiquitin (1UBQ).

back-calculated scalar couplings using the different sets of Karplus coefficients. In Fig. 5, we illustrate how the various sets of parameters derived from different techniques and different structural ensembles exhibit similar performance metrics. Interestingly, on applying BICePs-refined Karplus parameters to an ensemble generated by a molecular dynamics simulation (CHARMM22*),[37] 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 the supplementary material, Figs. S13–S15. On average, the BICeP-refined parameters derived from the 2NR2 ensemble [BICePs(2NR2)] give the lowest MAE between experiment and predictions for the CHARMM22* simulated ensemble, closely followed

© Author(s) 2026

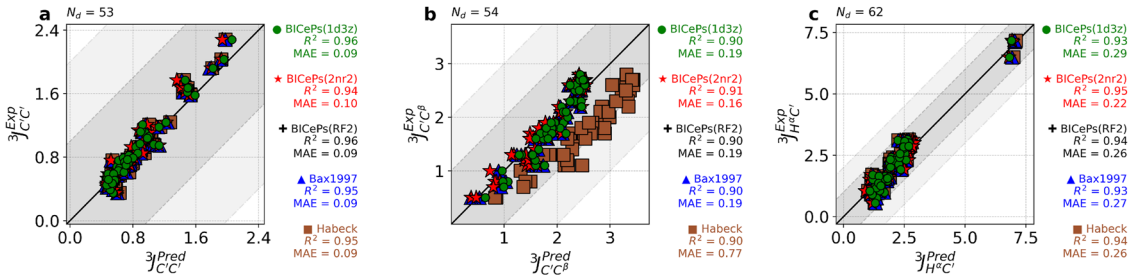


FIG. 5. Validation of BICeP-predicted Karplus coefficients perform similarly to Bax 1997 and achieve minor improvements over Habeck 2005 for scalar coupling predictions for the simulated ensemble of CHARMM22*. Each panel for (a) ${}^3J_{H^\alpha C'}$, (b) ${}^3J_{C'C^\beta}$, and (c) ${}^3J_{C'C'}$ shows 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 the supplementary material, Fig. S15.

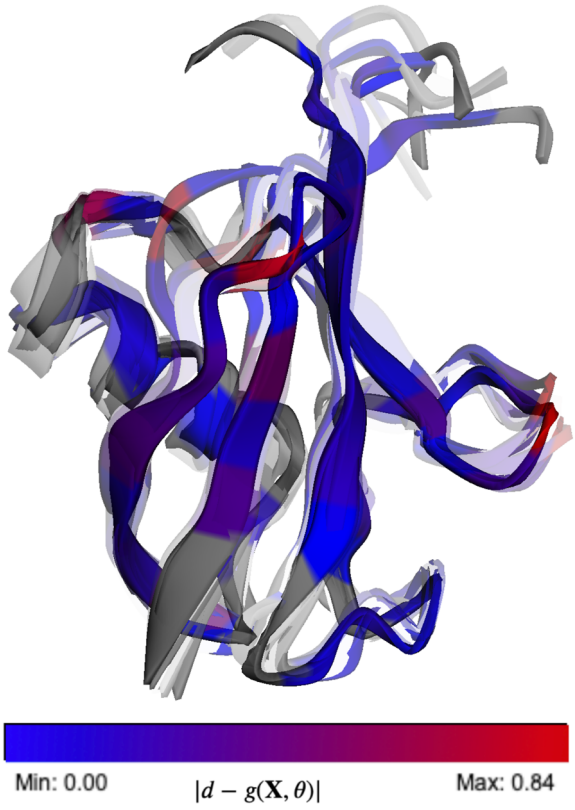


FIG. 6. Visualization of the 2NR2 conformational ensemble after joint conformational populations and Karplus parameter optimization using BICePs. Structures (144 in total) with higher transparency have lower populations. Residues are colored by the magnitude of the deviation between the replica-averaged forward model observations $g(\mathbf{X}, \theta)$ and the experimental ${}^3J_{C'C'}$ data d, where the color is normalized from blue (lowest deviation) to red (highest deviation).

by BICePs(RF2) parameters, whereas Habeck 2005 has the highest MAE due to known difficulties with ${}^3J_{C'C^\beta}$. A visualization of the BICePs-reweighted 2NR2 structural ensemble using BICePs(2NR2) parameters is shown in Fig. 6, with higher transparency representing lower populations. Residues are colored by the magnitude of the replica-averaged data deviation from the experiment in ${}^3J_{C'C'}$, where the color is normalized from blue (lowest deviation) to red (highest deviation).

To 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, integrating over all sources of error. Lower BICeP scores indicate better agreement with experiment. BICeP scores were computed using 11 intermediate ξ-values ranging between 0 and 1, which were optimized using *pylambdaopt*.[31] The results are shown in Table II. Each row in Table II corresponds to BICeP scores using all six sets of Karplus coefficients used on different structural ensembles. The lowest score is shown in bold [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, $f(\theta)/N_r$].

The leftmost column in Table II corresponds to the parameters, where *BICePs(1D3Z)* are the parameters in Table I (set 1), which used the 1D3Z ensemble to obtain Karplus coefficients. BICeP score columns, e.g., $f^{\text{1d3z}}_{\xi=0\to1}$ corresponds to BICeP scores evaluated for the 1D3Z ensemble. That is, the superscript corresponds to the structural ensemble used as a validation step. The reported BICePs scores were averaged over five independent rounds of validation each. BICeP calculations were burned for 1k MCMC steps, followed by 50k steps of production-run MCMC sampling.

The BICePs score, $f^{\text{1d3z}}_{\xi=0\to1} = 38.14 \pm 0.08$ (Table II), is equivalent (within error) with the most probable landscape basin

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

Parameters	$f^{\text{1d3z}}_{\xi=0\to1}$	$f^{\text{2nr2}}_{\xi=0\to1}$	$f^{\text{CHARMM22}*}_{\xi=0\to1}$
Bax 1997[2,34]	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
SVD(1D3Z)	44.76 ± 0.66	147.44 ± 0.70	120.63 ± 0.21
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

© Author(s) 2026

$f = 38.15 \pm 0.19$ from sampling the energy landscape, computed as an average across four chains; an example for one chain is shown in Fig. 4. This is additional evidence of the algorithm's reliability and corroborates that the results from variational minimization of the BICePs score and full joint posterior sampling are equivalent.

For both the 2NR2 and CHARMM22* structural ensembles, Bax 1997, BICePs(RF2) and BICePs(1D3Z) parameters give very similar BICePs scores, which suggests robust accuracy of FM parameters in reproducing experimental scalar couplings and the transferability of FM parameters across different prior structural ensembles. For the CHARMM22* simulated ensemble, the BICePs(2NR2) parameters give the lowest BICePs score. A partial explanation for this is that the structural ensemble 2NR2 deposited in the Protein Data Bank was generated using the CHARMM22 force field with additional experimental restraints during simulation.[35]

We obtained Karplus parameters using SVD with uniform populations across the 1D3Z ensemble (Table S1), denoted SVD(1D3Z). As shown in Table II, these parameters perform well for 1D3Z itself ($f^{1d3z}_{\xi=0\to1} = 44.76 \pm 0.66$) but generalize poorly to other ensembles, giving higher scores for 2NR2 and CHARMM22*. The MAE results in Table S2 show a similar trend: while SVD(1D3Z) predictions are competitive within 1D3Z, they are consistently worse than BICePs(1D3Z) for 2NR2 and CHARMM22*. This highlights a key limitation of SVD with uniform populations—without proper treatment of ensemble populations, the parameters risk overfitting to the training ensemble and lose transferability. By contrast, BICePs jointly optimize both parameters and populations, yielding lower errors across ensembles and more robust performance.

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 the CHARMM22* force field. 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. However, our BICePs(RF2) parameters have the best transferability across structural ensembles and have lower BICePs scores compared to Bax 1997 parameters. In a separate study, we evaluated the performance of the BICePs(RF2) parameters (denoted as "Raddi 2024" in Ref. 39) in predicting ${}^3J_{H^N H^\alpha}$ couplings for a set of non-natural and cyclic β-hairpin peptides.[39] Our results show that these parameters yield consistently lower prediction errors compared to the Bax 2007[40] and Kessler 1988[41] parameter sets. Based on all evidence, we recommend the BICePs(RF2) parameters for downstream applications, such as structural refinement and validation. Nonetheless, users should carefully assess accuracy for their specific system, as additional refinement may still be required.

IV. DISCUSSION

A. Ensembles from deep generative models such as RosettaFold2 can be used for parameter refinement

The booming field of machine learning and artificial intelligence is transforming the field of biomolecular modeling at a swift pace. Recent advancements in generative models, such as AlphaFold,[42] Rosetta-Fold,[36] 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.[43] 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 variety of structural ensembles of ubiquitin.

B. Automatic determination of unknown errors

Our method provides a notable advantage by automatically estimating all potential error sources throughout the ensemble refinement process. 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.[1,44]

In previous work, we have shown how the BICePs score is a better metric model validation than the traditional χ^2 test.[6] Unlike χ^2, which presupposes a fixed and known error, BICePs dynamically learn the complete posterior distribution of all uncertainty parameters, providing a measure of model quality that integrates over all sources of error. Our work here supports this premise.

C. Comparison of joint posterior sampling and variational minimization for FM parameterization

Two strategies can be used for forward model parameter optimization in BICePs: joint posterior sampling and variational minimization. Joint posterior sampling, which we recommend is efficient with modern gradient-based samplers, such as Hamiltonian Monte Carlo (HMC), provides direct estimates of parameter uncertainties and covariances, and is fast in practice (50 000 steps in 15 s on a MacBook M1 Pro). Variational minimization of the BICePs score can be advantageous when handling many parameters and is straightforward to adapt for different forward models, especially in convex landscapes, but it requires repeated free energy calculations via MBAR at each epoch, resulting in longer runtimes, and its uncertainty estimates are less reliable in non-convex landscapes. Over all, posterior sampling is faster and more robust, while variational minimization may be preferable for very high-dimensional, convex problems.

D. Bayesian ranking of Karplus-type relations

Although here we optimize Karplus relations with a standard functional form, many possible relations can be used to accommodate the diverse characteristics of molecular structures, from rigid to flexible.[45] 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.

© Author(s) 2026

E. Application to non-natural and cyclic peptides for fine-tuning folding landscapes

Our framework for forward model optimization has recently been successfully applied to a challenging test case involving a diverse set of non-natural and cyclic β-hairpin peptides.[39] In that study, we used BICePs' forward model optimization approach to refine Karplus parameters for ${}^3J_{H^N H^\alpha}$ predictions directly against experimental NMR data. With an improved Karplus relation, BICePs reshaped the folding landscapes to better agree with NMR data, enabling more accurate predictions of folding stability for each peptide. This work highlights the resolution enhancement gained by the improved forward model predictions and population reweighting using BICePs, enabling the detection of subtle changes in folding stability—such as those introduced by a sidechain hydrogen- or halogen-bonding group, which alters stability by no more than 2 kJ/mol—offering a reliable pathway for designing foldable non-natural and cyclic peptides.

F. The BICePs score can be used as a loss function to train neural networks and other differentiable forward models

A central challenge in physical sciences is integrating domain knowledge with data-driven models. Traditional machine learning often lacks physical consistency or interpretability, while purely physics-based models may miss subtle experimental trends. One way to bridge this gap is to use the BICePs score—a free energy-based Bayesian loss—to train machine learning forward models using ensemble-averaged experimental data. This allows one to parameterize neural networks or other differentiable models in a physically grounded, uncertainty-aware fashion. In this way, BICePs provides a general-purpose learning framework for constructing interpretable, probabilistically justified forward models for complex observables. This capability may be especially valuable in contexts where observables are too complex to be modeled analytically—such as in chemical shift prediction or other nonlinear structure–observable relationships.[15,46,47]

To demonstrate the viability of using BICePs for neural network training, we constructed a toy model for J-coupling prediction. In this model, the backbone dihedral angle ϕ was the sole input variable. To appropriately encode the periodicity of angular data, each ϕ-angle was embedded into two dimensions via a sine–cosine transformation,

$$x = \begin{bmatrix} \cos(\phi) \\ \sin(\phi) \end{bmatrix} \in \mathbb{R}^2. \tag{22}$$

This representation ensures continuity across the periodic boundary at $\pm\pi$ and provides a smooth input feature space for learning.

A fully connected feedforward neural network was constructed using this 2D input, consisting of one hidden layer with 200 neurons and Gaussian Error Linear Unit (GELU) activation functions. The final output was a scalar prediction corresponding to the J-coupling value. To ensure stable training dynamics compatible with GELU, the weights were initialized using LeCun normal initialization.[48]

The network was trained using Adaptive Moment Estimation (ADAM) optimization for 2000 epochs with a learning rate of 10^{-3}, using the Good–Bad likelihood model with the BICePs score as the loss function. The toy dataset consisted of five conformational states, each with 60 synthetic J-coupling observables using the Karplus relation $J(\phi) = A\cos^2(\phi) + B\cos(\phi) + C$, with ground truth coefficients $A = 6.51$, $B = -1.76$, and $C = 1.60$. Our training dataset consisted of 300 ϕ-angles as NN inputs.

First, we test the performance without adding random or systematic noise to the synthetic experimental J-coupling observables. Figure S16(a) in the supplementary material shows the convergence behavior of the BICePs' score across five independent training runs, where a plateau is reached around step 1500. Figure S16(b) in the supplementary material shows the Karplus curves predicted by the five trained networks, along with their mean and standard deviation (blue curve). For comparison, we also include the result of singular value decomposition (SVD) applied directly to the Karplus form under the assumption that backbone ϕ angles are known precisely, as would be the case with a high-resolution crystal structure. The parameters from SVD were determined to be $A = 6.47 \pm 0.002$, $B = -1.75 \pm 0.001$, and $C = 1.62 \pm 0.001$. Our framework yields comparable predictive performance, where the predictive accuracy of J-couplings for the SVD approach achieved an RMSE of 0.01 ± 0.001 Hz, while the BICePs-trained neural network reached an RMSE of 0.04 ± 0.01 Hz.

Next, we evaluate the performance when adding random or systematic noise to the synthetic experimental J-coupling observables ($\sigma_{\text{data}} = 0.79$ Hz). Figures S17(a) and (b) in the supplementary material shows the convergence behavior of the BICePs score and gradient norm across five independent training runs, where convergence begins around 1000 epochs. The BICePs score was determined to be $f_{\xi=0\to1} = 62.29$ nats. Figure S17(c) in the supplementary material shows the Karplus curves predicted by the five trained networks, along with their mean and standard deviation (blue curve). Comparison against SVD illustrates that BICePs, when used to train neural networks from conformational ensembles alone, recovers predictive performance comparable with analytical fitting based on ground-truth dihedral angles, even in the presence of random and systematic errors. When comparing predictive accuracy of J-couplings, the SVD approach achieved an RMSE of 0.24 ± 0.03 Hz, while the BICePs-trained neural network reached an RMSE of 0.21 ± 0.01 Hz, demonstrating that our framework yields improved predictive performance. For completeness, we also used BICePs to optimize the parameters of the Karplus relation in this example, obtaining $A = 6.6 \pm 0.05$, $B = -1.7 \pm 0.023$, and $C = 1.7 \pm 0.037$, which resulted in the lowest RMSE of 0.17 ± 0.0001 Hz. This outcome is expected since the "true" values were originally generated using the Karplus relation. By contrast, the neural network achieves remarkably low error without being constrained to any functional form, highlighting its flexibility and strong predictive capability.

Convergence of the BICePs' score is typically accompanied by a reduction in the posterior uncertainty parameter, σ, which sharpens the likelihood landscape and amplifies gradients with respect to model parameters. In particular, the derivative of the BICePs' score with respect to a parameter θ contains a term proportional to $1/\sigma^2$ [see Eqs. (12) and (13)], meaning that as $\sigma \to 0$, the sensitivity of the loss to prediction errors increases.

During training of the NN-based forward model, the posterior sampling of σ plays a critical role in regularization. If σ were held fixed during training, the model would be limited in its ability to adjust the sharpness of the likelihood surface and could fail to reach an optimal fit to the data. By treating σ as a nuisance parameter

© Author(s) 2026

and marginalizing over it within the BICePs framework, the model adaptively tunes the effective strength of the loss function, yielding an implicit form of regularization. This dynamic adjustment helps stabilize training and ensures neural network parameters are optimized in a manner consistent with both the experimental data and the underlying physical ensemble.

Despite these advantages, neural networks trained on limited or biased datasets can still face overfitting risks. Several mitigation strategies can be incorporated within the BICePs training framework: (1) data augmentation via additional synthetic observables or sampling of underrepresented dihedral regions, (2) regularization through dropout layers to discourage over-parameterization, and (3) physically informed constraints on network architecture or inputs, such as torsional symmetries—as we did for J-coupling or local environment descriptors for chemical shifts. These strategies, combined with the adaptive regularization provided by marginalizing over σ, further safeguard against overfitting and improve model generalizability.

While this toy model serves as a useful proof-of-concept, the broader applicability of our framework is in modeling more complex observables such as NMR chemical shifts. These observables are typically nonlinear and high-dimensional functions of atomic coordinates, making them difficult to model analytically. In contrast, neural networks trained using BICePs can serve as flexible surrogates for such forward models. Importantly, this framework supports the simultaneous training of multiple neural networks, each targeting a distinct observable—for example, one network may predict $^{3}J_{\mathrm{HN-H}\alpha}$ couplings, another $^{3}J_{\mathrm{C'-C'}}$ couplings, and others may be used for predicting chemical shifts. In the case of backbone chemical shift prediction, a separate neural network is generally constructed for each nucleus type (e.g., Cβ, ^{1}Hα, and ^{1}HN), enabling specialized learning tailored to the distinct structural dependencies of each observable. All networks can be trained jointly using a shared conformational ensemble, with BICePs enforcing consistency with the corresponding experimental data.

Models have limited predictive reliability in regions lacking sufficient training data. A central advantage of our framework lies in its ability to incorporate physically relevant information through the choice of input features or explicit regularization. In our example, using the toy model system, we employed the backbone dihedral angle ϕ as the sole input feature (projected to the unit circle to maintain periodicity). However, neural network inputs can encompass a broader range of features for complex observables such as chemical shifts, including quantum calculations, partial charges, descriptors of local atomic environments, interatomic distances, hydrogen bonding patterns, solvent accessibility, or torsional strain energies. These diverse features encode crucial structural and electronic contexts and can be precomputed from molecular simulations or quantum calculations before training.

We did not impose physical constraints or priors when training our model to predict J-coupling constants. Users might explicitly incorporate such physical priors or constraints through architectural designs, regularization terms, or physically informed input features, especially critical when dealing with sparse or unevenly distributed training data. For instance, in our experiments, neural networks employing Rectified Linear Unit (ReLU) activations struggled to accurately predict J-couplings at backbone dihedral angles greater than +90°. In contrast, GELU activations provided smoother and more reliable extrapolation in these underrepresented regions. Thus, choosing suitable activation functions and rich input features improves predictive accuracy across the conformational landscape.

G. Relation to mixture-based outlier models and multi-level outlier severity

Our outlier-aware Good-and-Bad likelihood model uses a two-component mixture $M = 2$ of variances. One of the components serves to reduce the influence of data affected by systematic error, while the other component functions to preserve the influence of "good" data. A natural extension of the Good–Bad model replaces a single "bad" mode with multiple levels of outlier severity.

In reliability and survival analysis, censored finite-mixture models help separate heterogeneous data while accounting for censoring. For example, Akhtar and Alharthi[49] apply a censored two-component mixture of geometric distributions under a Bayesian framework, illustrating how finite mixtures capture central trends while disregarding extreme values. Although distinct from our context, it illustrates the broader principle that finite mixtures model structured heterogeneity rather than ad hoc data exclusion.

For a natural extension to our Good–Bad model, we assign $p(\sigma_j|\sigma_0)$ from Eq. (7) to be a multi-severity Good–Bad prior

$$p(\sigma_j \mid \sigma_0, \{\omega_m\}, \{\varphi_m\}) = \sum_{m=1}^{M} \omega_m \, \delta(\sigma_j - \varphi_m \sigma_0), \tag{23}$$

where $\omega_m \geq 0$ are mixture weights with $\sum_{m=1}^{M} \omega_m = 1$ and the severity inflators satisfy $\varphi_1 \equiv 1 < \varphi_2 < \cdots < \varphi_M$. Thus, φ_m scales the typical uncertainty σ_0 to represent increasingly severe outliers (mild → severe). The standard two-component Good–Bad model is recovered at $M = 2$.

Inserting Eq. (23) into the integrand of Eq. (7) gives, for each observable j,

$$\begin{aligned} &\int_{\sigma^{\mathrm{SEM}}}^{\infty} \mathcal{N}\left(d_j \mid g_j, \sigma_j^2\right) \sum_{m=1}^{M} \omega_m \, \delta(\sigma_j - \varphi_m \sigma_0) \, d\sigma_j \\ &\qquad = \sum_{m=1}^{M} \omega_m \, H\left(\varphi_m \sigma_0 - \sigma^{\mathrm{SEM}}\right) \mathcal{N}\left(d_j \mid g_j, \varphi_m^2 \sigma_0^2\right), \end{aligned} \tag{24}$$

i.e., each observation is evaluated under a finite Gaussian scale mixture with the Heaviside factor enforcing the σ^{SEM} lower bound.

If we integrate out the mixing weights ω_m as in our two-component case, a uniform Dirichlet prior that is flat on the simplex $\omega_m \geq 0, \sum_m \omega_m = 1$ implies $\mathbb{E}[\omega_m] = 1/M$ by symmetry, leading to

$$\begin{aligned} &\sum_{m=1}^{M} \left[\int_{\Delta^{M-1}} \omega_m \, p(\boldsymbol{\omega}) \, d\boldsymbol{\omega} \right] H\left(\varphi_m \sigma_0 - \sigma^{\mathrm{SEM}}\right) \mathcal{N}\left(d_j \mid g_j, \varphi_m^2 \sigma_0^2\right) \\ &\qquad = \frac{1}{M} \sum_{m=1}^{M} H\left(\varphi_m \sigma_0 - \sigma^{\mathrm{SEM}}\right) \mathcal{N}\left(d_j \mid g_j, \; \varphi_m^2 \sigma_0^2\right). \end{aligned} \tag{25}$$

Practically, this multi-severity formulation leaves a single Bayesian uncertainty parameter σ_0^B and introduces $M - 1$ free severity parameters φ_m (with $\varphi_1 = 1$ fixed) to be inferred in the posterior. This extends the single-severity Good–Bad model used

© Author(s) 2026

in our framework (where only a single φ is sampled alongside σ_0^B) and provides finer control over how aggressively outliers are down-weighted.

V. 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 that BICePs can be used as a robust tool for determining forward model (FM) parameters by sampling over the full posterior distribution of their values and uncertainties.

We used a toy model of protein dihedral angles to demonstrate that posterior sampling and variational minimization of the BICePs score are valid and robust approaches for FM parameter refinement. Using structural ensembles and experimental data for ubiquitin, we applied BICePs with posterior sampling to optimize 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 BICePs can be applied more generally to optimize a wide variety of forward models, including those represented by neural networks. By treating the BICePs score as a physically grounded loss function, our framework enables the training of data-driven forward models that are consistent with ensemble-averaged experimental observables and incorporate principled uncertainty quantification. These advances not only contribute to the refinement of molecular simulations but also provide a scalable and extensible machine learning approach for physical sciences, with applications in structural dynamics, model validation, and the construction of predictive, interpretable models across a broad spectrum of data regimes.

SUPPLEMENTARY MATERIAL

The supplementary material includes appendices describing second derivatives of the BICePs score and the equivalence of joint posterior sampling and variational minimization; a supplementary methods section describing more details about structural ensembles and posterior sampling; supplementary material, Figs. S1–S17.

ACKNOWLEDGMENTS

The authors acknowledge that substantial content in this manuscript has also been included in the doctoral dissertation of R.M.R.[50] R.M.R., T.M., and V.A.V. are supported by National Institutes of Health under Grant No. R01GM123296. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation Grant No. 1625061 and by the U.S. Army Research Laboratory under Contract No. ARL W911NF212007. The authors also thank Kim Sharp for valuable discussions on extending the Good-and-Bad likelihood framework to incorporate multiple modes of data severity and Somaiyeh Dadashi for her assistance with illustrations.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Robert M. Raddi: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Tim Marshall**: Data curation (equal); Formal analysis (equal); Investigation (equal); Resources (equal). **Vincent A. Voelz**: Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The BICePs algorithm is openly available at https://github.com/vvoelz/biceps. All calculations in this work were performed using the development version biceps_v3.0a, which is planned to be merged into the main branch. The specific version can be accessed at https://github.com/vvoelz/biceps/tree/biceps_v3.0a. All scripts, Jupyter notebook examples and analysis can be found here. Detailed documentation and tutorials can be found here: https://biceps.readthedocs.io/en/latest/. For any issues or questions, please submit the request on GitHub.

REFERENCES

[1] T. Fröhlking, M. Bernetti, and G. Bussi, "Simultaneous refinement of molecular dynamics ensembles and forward models using experimental data," J. Chem. Phys. **158**, 214120 (2023).

[2] A. C. Wang and A. Bax, "Determination of the backbone dihedral angles ϕ in human ubiquitin from reparametrized empirical Karplus equations," J. Am. Chem. Soc. **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," J. Magn. Reson. **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," J. Biomol. 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. A. Voelz, "Model selection using replica averaging with Bayesian inference of conformational populations," J. Chem. Theory Comput. **21**, 5880 (2025).

[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," J. Chem. Inf. Model. **63**, 2370–2381 (2023).

© Author(s) 2026

[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:2402.11169 (2024).
[13] N. Akhtar, S. A. Khan, E. A. A. Ismail, F. A. Awwad, A. A. Khan, T. Gul, and H. Alqahtani, "Analyzing quantitative performance: Bayesian estimation of 3-component mixture geometric distributions based on Kumaraswamy prior," Stat. Papers **65**, 4431–4451 (2024).
[14] N. Akhtar, M. F. Alharthi, S. A. Khan, A. A. Khan, and M. Amin, "Bayesian estimation strategy for multi-component geometric life testing model under doubly type-1 censoring scheme," Kuwait J. Sci. **52**, 100339 (2025).
[15] C. Han, D. Zhang, S. Xia, and Y. Zhang, "Accurate prediction of NMR chemical shifts: Integrating DFT calculations with three-dimensional graph neural networks," J. Chem. Theory Comput. **20**, 5250–5258 (2024).
[16] Z. Yang, M. Chakraborty, and A. D. White, "Predicting chemical shifts with graph neural networks," Chem. Sci. **12**, 10802–10809 (2021).
[17] J. Meiler, "PROSHIFT: Protein chemical shift prediction using artificial neural networks," J. Biomol. NMR **26**, 25–37 (2003).
[18] Y. Shen and A. Bax, "SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network," J. Biomol. NMR **48**, 13–22 (2010).
[19] W. Rieping, M. Habeck, and M. Nilges, "Inferential structure determination," Science **309**, 303–306 (2005).
[20] J. W. Pitera and J. D. Chodera, "On the use of experimental observations to bias simulated ensembles," J. Chem. Theory Comput. **8**, 3445–3451 (2012).
[21] A. Cavalli, C. Camilloni, and M. Vendruscolo, "Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle," J. Chem. Phys. **138**, 094112 (2013).
[22] A. Cesari, S. Reißer, and G. Bussi, "Using the maximum entropy principle to combine simulations and solution experiments," Computation **6**, 15 (2018).
[23] B. Roux and J. Weare, "On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method," J. Chem. Phys. **138**, 084107 (2013).
[24] G. Hummer and J. Köfinger, "Bayesian ensemble refinement by replica simulations and reweighting," J. Chem. Phys. **143**, 243150 (2015).
[25] M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo, "Metainference: A Bayesian inference method for heterogeneous systems," Sci. Adv. **2**, e1501177 (2016).
[26] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," ACM Trans. Math. Software **23**, 550–560 (1997).
[27] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105–124111 (2008).
[28] D. A. Sivak and G. E. Crooks, "Thermodynamic metrics and optimal paths," Phys. Rev. Lett. **108**, 190602 (2012).
[29] 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**, 046705 (2009).
[30] A. Rizzi, "Improving efficiency and scalability of free energy calculations through automatic protocol optimization," Ph.D. thesis, Weill Medical College of Cornell University, 2020.
[31] D. Novack, R. M. Raddi, S. Zhang, M. F. D. Hurley, and V. A. Voelz, "Simple method to optimize the spacing and number of alchemical intermediates in expanded ensemble free energy calculations," J. Chem. Inf. Model. **65**, 6089 (2025).
[32] 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?," J. Am. Chem. Soc. **119**, 6360 (1997).
[33] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," Stat. Sci. **7**, 457–472 (1992).
[34] 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," J. Am. Chem. Soc. **120**, 6836–6837 (1998).
[35] 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," J. Biomol. NMR **37**, 117–135 (2007).
[36] M. Baek, I. Anishchenko, I. Humphreys, Q. Cong, D. Baker, and F. DiMaio, "Efficient and accurate prediction of protein structure using RoseTTAFold2," bioRxiv:542179 (2023).
[37] 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).
[38] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 Å resolution," J. Mol. Biol. **194**, 531–544 (1987).
[39] T. D. Nguyen, R. M. Raddi, and V. A. Voelz, "High-resolution tuning of non-natural and cyclic peptide folding landscapes against NMR measurements using Markov models and Bayesian inference of conformational populations," J. Chem. Theory Comput. **21**, 6213 (2025).
[40] B. Vögeli, J. Ying, A. Grishaev, and A. Bax, "Limits on variations in protein backbone dynamics from precise measurements of scalar couplings," J. Am. Chem. Soc. **129**, 9377–9385 (2007).
[41] H. Kessler, C. Griesinger, J. Lautz, A. Mueller, W. F. Van Gunsteren, and H. J. C. Berendsen, "Conformational dynamics detected by nuclear magnetic resonance NOE values and J coupling constants," J. Am. Chem. Soc. **110**, 3393–3396 (1988).
[42] 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).
[43] Z. F. Brotzakis, S. Zhang, M. H. Murtada, and M. Vendruscolo, "AlphaFold prediction of structural ensembles of disordered proteins," Nat. Commun. **16**, 1632 (2025).
[44] G. W. Vuister and A. Bax, "Quantitative J correlation: A new approach for measuring homonuclear three-bond $J(H^NH^\alpha)$ coupling constants in ^{15}N-enriched proteins," J. Am. Chem. Soc. **115**, 7772–7777 (1993).
[45] M. J. Minch, "Orientational dependence of vicinal proton-proton NMR coupling constants: The Karplus relationship," Concepts Magn. Reson. **6**, 41–56 (1994).
[46] J. Fang, L. Hu, J. Dong, H. Li, H. Wang, H. Zhao, Y. Zhang, and M. Liu, "Predicting scalar coupling constants by graph angle-attention neural network," Sci. Rep. **11**, 18686 (2021).
[47] R. Sanchez-Garcia, J. Gomez-Blanco, A. Cuervo, J. M. Carazo, C. O. S. Sorzano, and J. Vargas, "DeepEMhancer: A deep learning solution for cryo-EM volume post-processing," Commun. Biol. **4**, 874 (2021).
[48] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade* (Springer, 2002), pp. 9–50.
[49] N. Akhtar and M. F. Alharthi, "Bayesian analysis of heterogeneous data outliers using a censored mixture model," Stat. Papers **66**, 56 (2025).
[50] R. M. Raddi, "A Bayesian inference/maximum entropy approach for optimization and validation of empirical molecular models," Ph.D. dissertation (Temple University, 2024), available from ProQuest Dissertations and Theses. Copyright ProQuest LLC.

© Author(s) 2026