Encoding prior knowledge in ensemble refinement

Jürgen Köfinger[1] and Gerhard Hummer[1,2]
[1)] *Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany*[a)]
[2)] *Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany*[b)]

The proper balancing of information from experiment and theory is a long-standing problem in the analysis of noisy and incomplete data. Viewed as a Pareto optimization problem, improved agreement with the experimental data comes at the expense of growing inconsistencies with the theoretical reference model. Here, we propose how to set the exchange rate *a priori* to properly balance this trade-off. We focus on gentle ensemble refinement, where the difference between the potential energy surfaces of the reference and refined models is small on a thermal scale. By relating the variance of this energy difference to the Kullback-Leibler divergence between the respective Boltzmann distributions, one can encode prior knowledge about energy uncertainties, i.e., force-field errors, in the exchange rate. The energy uncertainty is defined in the space of observables and depends on their type and number, and on the thermodynamic state. We highlight the relation of gentle refinement to free energy perturbation theory. A balanced encoding of prior knowledge increases the quality and transparency of ensemble refinement. Our findings extend to non-Boltzmann distributions, where the uncertainty in energy becomes an uncertainty in information.

I. INTRODUCTION

In the natural sciences, we frequently encounter the challenge of analyzing noisy and incomplete experimental data using theoretical models. The number of parameters in the models often exceeds the number of available data points. Image reconstruction is the textbook example of this class of ill-defined inverse problems.[1] Another important class of problems is ensemble refinement, as performed in integrative structural biology and hybrid modelling. Experimentally averaged observables are used to refine an ensemble of biomolecular structural models.[2–4] In such cases, we have to take into account statistical and systematic errors in the model and in the data, and balance the information provided by experiment and theory. While maximum entropy (MaxEnt) and Bayesian methods address these challenges, the balancing of information has proven to be a long-standing and difficult problem.[5]

The balancing of information from experiment and theory can be viewed as a Pareto optimization problem (Figure 1). In a multi-objective optimization, we seek to minimize both the mean-squared deviations from the experimental data, χ^2, with respect to the statistical weights in the refined model, and the deviation of these weights from a given reference model, as quantified by the Kullback-Leibler (KL) divergence[6] S_{KL}. In the plane spanned by S_{KL} and χ^2, the set of weights where one is optimal for a fixed value of the other defines a Pareto front. On this Pareto front, we then seek a particular solution for which we consider the trade-off between increases in S_{KL} and χ^2 to be fair. To encode the relative value given to S_{KL} and χ^2, we introduce a parameter θ that sets the so-called marginal rate of substitution, $\mathrm{d}\chi^2/\mathrm{d}S_{\mathrm{KL}} = -2\theta$. Minimization of a loss function $\chi^2/2 + \theta S_{\mathrm{KL}}$ with respect to the model weights then gives us a specific set of refined weights on the Pareto front (or L-curve[7]) with an optimal trade-off.

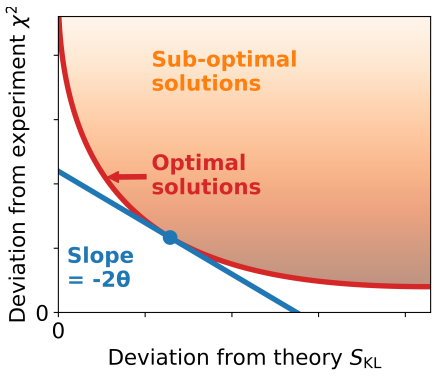

FIG. 1. Bayesian/MaxEnt ensemble refinement as a Pareto optimization problem. The curve of optimal solutions in the plane spanned by the deviations from experiment, χ^2, and the deviations from theory, S_{KL}, determines the Pareto front or L-curve (red line). Sub-optimal solutions (orange shaded area) lie above this curve. We encode our prior knowledge by choosing a value of the so-called marginal rate of substitution, given by -2θ here. We use this rate to trade off the two deviations according to our prior experience, $\mathrm{d}\chi^2 = -2\theta\,\mathrm{d}S_{\mathrm{KL}}$, which gives a unique solution (blue disk). At this point, the slope of the Pareto front is exactly the marginal rate, as indicated by the blue line.

The proper choice of the θ value[5] has been tackled by different workarounds. As a result, various methods to image reconstruction, ensemble refinement, and related problems can be distinguished by their choice of a solution parameterized by a multiplicative parameter θ of the KL divergence: solutions are chosen by L-curve analysis;[7–10] by using algorithms to determine

[a)] Electronic mail: juergen.koefinger@biophys.mpg.de
[b)] Electronic mail: gerhard.hummer@biophys.mpg.de

an elbow of the L-curve;[11,12] so that the resulting error is the most-likely according to some statistic;[1] by putting a prior on θ and integrating it out;[13,14] by selecting a perfect fit ignoring errors completely as in classical MaxEnt methods;[15,16] by not including θ as parameter at all, which corresponds to setting $\theta = 1$;[17] or by cross-validation.[18] While these workarounds are useful in practice, they generally do not balance information from experiment and theory properly in the sense of using *a priori* knowledge. As a result, the solutions obtained tend to be overfitted or underfitted. Moreover, a solution is not always guaranteed to exist for some of these workarounds.[10]

Here, we propose how to choose the θ parameter *a priori* for gentle ensemble refinement. Often, ensemble refinement is quite crude, such that the optimal ensemble deviates drastically from the reference ensemble. In gentle refinement, we take care that this is not case. In this regime, the expected KL divergence $\bar{S}_{\mathrm{KL}}$ can be approximated in terms of the mean-squared error of the potential energy surface U defining the force field used to create the reference ensemble, $\bar{S}_{\mathrm{KL}} \approx \mathrm{var}(\beta\Delta U)/2$. Here, energies are measured in units of the thermal energy, $\beta = 1/(k_{\mathrm{B}}T)$, in the equilibrium Boltzmann distributions of reference and refined ensembles, with k_{B} the Boltzmann constant and T the absolute temperature. We propose to use this physically meaningful relation to encode prior knowledge about the expected force-field error by setting

$$\theta \cong \frac{1}{\bar{S}_{\mathrm{KL}}} \approx \frac{2}{\mathrm{var}(\beta\Delta U)} \tag{1}$$

to fix the Pareto exchange rate at -2θ. The mean-squared energy error depends on the type and number of observables used for refinement, and on the thermodynamic state.

The article is organized as follows. In Sec. II A, we first introduce ensemble refinement with a focus on the Bayesian inference of ensembles (BioEn) method.[8,9] We relate the KL divergence for Boltzmann distributions to the mean of the reduced energy difference in Sec. II B. We derive an approximation for the KL divergence as half the variance of the reduced energy difference in subsection II C. Moreover, the KL divergence is approximately symmetric with respect to its arguments. In subsection II D, we show that these approximations are exact in the Gaussian case. In subsection II F, we discuss how to use this information to encode our prior knowledge in the entropic prior used in BioEn and related methods. In subsection II G, we describe how to estimate the relevant energy uncertainty. We present three example systems in Sec. III. For these examples, we quantify the validity of the approximation of the KL divergence by the energy variance and illustrate the benefits and limits of gentle ensemble refinement in Sec. IV. In Sec. V, we discuss how we build prior knowledge about simulations. We end with Concluding Remarks in Sec. VI and the implication of our results for ill-defined inverse problems in general.

II. THEORY

A. Background

The BioEn posterior[9] for a sampled ensemble with N conformations and normalized reference weights $\mathbf{w}^{(0)} = (w_1^{(0)}, \ldots, w_N^{(0)})$ is given by

$$p(\mathbf{w}|\mathrm{data}) \propto p(\mathbf{w}|\mathbf{w}^{(0)})p(\mathrm{data}|\mathbf{w}) \tag{2}$$

where $\mathbf{w} = (w_1, \ldots, w_N)$ is the vector of the normalized weights we want to find by refinement. The so-called entropic prior[19] is given by

$$p(\mathbf{w}|\mathbf{w}^{(0)}) \propto e^{-\theta S_{\mathrm{KL}}(\mathbf{w}||\mathbf{w}^{(0)})} \tag{3}$$

where θ is the confidence parameter and $S_{\mathrm{KL}}(\mathbf{w}||\mathbf{w}^{(0)})$ is the KL divergence or relative entropy

$$S_{\mathrm{KL}}(\mathbf{w}||\mathbf{w}^{(0)}) = \sum_{\alpha=1}^{N} w_\alpha \ln \frac{w_\alpha}{w_\alpha^{(0)}} \tag{4}$$

θ encodes how much we trust our original ensemble. For independent Gaussian errors, for example, the likelihood of the measured data is given by

$$p(\mathrm{data}|\mathbf{w}) \propto e^{-\frac{\chi^2(\mathbf{w})}{2}} \tag{5}$$

where

$$\chi^2(\mathbf{w}) = \sum_{i=1}^{M} \left(\frac{\langle y_i \rangle - Y_i}{\sigma_i} \right)^2 \tag{6}$$

Here, $\langle y_i \rangle = \sum_{\alpha=1}^{N} w_\alpha y_i^\alpha$ are the averages of the calculated observables y_i^α for conformation α with indices $i = 1, \ldots, M$ for the measured averages Y_i with associated errors σ_i, both theoretical and experimental. The theoretical errors primarily reflect uncertainties in calculating the observables $\mathbf{y}(x)$ for given conformations x using simplified models of the experiment. By contrast, we account for errors in the weights in the entropic prior, Eq. (3), through the confidence parameter θ. Note that the structural ensembles do not have to come from molecular simulations[20–22] and that the data do not necessarily have to be from experiment.

To find the optimal solutions, we maximize the BioEn posterior or, equivalently, minimize the negative log-posterior given by

$$\mathcal{L} = \theta S_{\mathrm{KL}}(\mathbf{w}||\mathbf{w}^{(0)}) + \frac{\chi^2(\mathbf{w})}{2} \tag{7}$$

which is the loss function mentioned in the Introduction. This formulation has been originally developed for the EROS method[8] based on a method for image

reconstruction.[1] Since then it has entered into the BME method[23], for example. Note that any method that is equivalent to refinement by replica simulations[24,25] is also equivalent to the EROS/BioEn method with properly chosen coupling constant and in the limit of infinite numbers of replicas.[9] As discussed in Ref. 9, the dependence on the number of replicas can be removed with a subsequent BioEn refinement.

We explain next how we calculate and approximate the KL divergence for Boltzmann distributions. In the following, we focus on the continuous distributions $q(x)$ and $p(x)$ underlying the reference weights $\mathbf{w}^{(0)}$ and optimal weights $\mathbf{w}$, respectively.

B. Kullback-Leibler divergence for Boltzmann distributions

For continuous probability densities $p(x)$ and $q(x)$ the KL divergence[6] is defined as

$$S_{\mathrm{KL}}(p||q) = \int \mathrm{d}x\, p(x) \ln \frac{p(x)}{q(x)} = \left\langle \ln \frac{p(x)}{q(x)} \right\rangle_p \tag{8}$$

In the refinement of an isothermal ensemble, x represents $3N$-dimensional conformations, $q(x)$ is the reference distribution underlying simulations and $p(x)$ is the refined distribution. The angular brackets with subscript 'p' indicate the expectation value with respect to $p(x)$.

For the sake of generality, we point out here that the KL divergence in Eq. (8) is the expectation of the information difference

$$\Delta h(x) = \ln \frac{p(x)}{q(x)} = \ln \frac{m(x)}{q(x)} - \ln \frac{m(x)}{p(x)} \tag{9}$$

with respect to $p(x)$. The Jayne's measure[26] $m(x)$ guarantees invariance of information under variable transformation. Although it cancels in the expression above, $m(x)$ is needed to define a proper difference between the information of the two ensembles. As we shall see below, this difference becomes a difference in energies for Boltzmann distributions, as has been established previously.[8–10,16,17,27]

Let us assume that probability distributions are given by Boltzmann distributions. For a potential energy surface $U_q(x)$ defining the reference ensemble, we then have

$$q(x) = \frac{e^{-\beta U_q(x)}}{Q_q} \tag{10}$$

where the normalization constant Q_q is the partition function,

$$Q_q = \int \mathrm{d}x\, e^{-\beta U_q(x)} \equiv e^{-\beta F_q} \tag{11}$$

with $F_q = -k_{\mathrm{B}} T \ln Q_q$ the free energy. We analogously define $p(x)$, Q_p, and F_p for the potential energy surface $U_p(x)$ of the refined ensemble. In an isothermal ensemble at inverse temperature $\beta = 1/(k_{\mathrm{B}}T)$, the energy of conformation x is given by $U_q(x) = E_q(x)$, where $E_q(x)$ is its potential energy. In an isobaric-isothermal ensemble, $U_q(x) = E_q(x) + \mathsf{p} V_q(x)$, where p is the pressure and $V(x)$ is the box volume for conformation x.

To evaluate the KL divergence, Eq. (8), for Boltzmann distributions, we use that

$$\frac{p(x)}{q(x)} = e^{-\beta \Delta U(x)} \frac{Q_q}{Q_p} \tag{12}$$

where $\Delta U(x) \equiv U_p(x) - U_q(x)$ is the energy difference. Using that the free-energy difference is given by $\Delta F \equiv F_p - F_q$, we obtain

$$\ln \frac{p(x)}{q(x)} = -\Delta u(x) \tag{13}$$

where we introduced the reduced energy difference

$$\Delta u(x) = \beta \Delta U(x) - \beta \Delta F \tag{14}$$

between the force fields of the two Boltzmann distributions $p(x)$ and $q(x)$. Importantly, these energy differences are uniquely determined because additive constants in the energies cancel in the Boltzmann distributions due to normalization.

Consequently, for Boltzmann distributions the KL divergence can be written as an average of the reduced energy difference

$$S_{\mathrm{KL}}(p||q) = -\left\langle \Delta u \right\rangle_p \tag{15}$$

The reduced energy differences tell us how we have to change the energies of the reference Boltzmann distribution $q(x)$ to sample the optimal ensemble according to Eq. (13), i.e.,

$$p(x) = q(x)\, e^{-\Delta u(x)} \tag{16}$$

as a physical interpretation of the information differences introduced in Eq. (9).

To estimate the reference and refined weights, and thus the reduced energy differences, we do not need to calculate any partition functions. On the contrary, we actually numerically estimate the free-energy difference and thus the log-ratio of partition functions. For two different force fields, we can estimate the free-energy difference from the KL divergence using Eqs. (14) and (15),

$$\Delta F = \langle \Delta U \rangle + k_B T\, S_{\mathrm{KL}}(p||q) \tag{17}$$

This expression is akin to the definitions of the Helmholtz free energy for the isothermal ensemble and to the Gibbs free energy for the isothermal-isobaric ensemble, where $k_B S_{\mathrm{KL}}(p||q)$ corresponds to the negative entropy. We obtain more familiar looking equations if we perform the average in the q-ensemble,

$$\Delta F = \langle \Delta U \rangle_q - k_B T\, S_{\mathrm{KL}}(q||p) \tag{18}$$

For finite ensembles, we estimate the KL divergence, Eq. (8), numerically using Eq. (4). In Appendix A, we

show how to properly interpret discrete reference weights $w_\alpha^{(0)}$ and refined weights w_α for ensembles sampled from arbitrary distributions. Thus, all results derived here for continuous distributions also apply to discrete ensembles. We next use that the KL divergence is given by the mean reduced energy differences and relate it to the variance for gentle ensemble refinement.

C. KL divergence approximations

In the following, we rewrite averages of exponential functions as cumulant expansions, which lead to simple expressions for Gaussian distributions of the exponent.[28,29] We calculate the mean energy change determining the KL divergence, Eq. (15), introducing the refined distribution function $p(\Delta u)$ of the energy differences,

$$p(\Delta u) = \int \mathrm{d}x\, p(x)\, \delta[\Delta u - \Delta u(x)] \tag{19}$$

$\delta[\cdot]$ is the Dirac delta function. Analogously, we define $q(\Delta u)$ for the reference distribution. Equation (16) becomes

$$q(\Delta u) = p(\Delta u) e^{\Delta u} \tag{20}$$

We integrate both sides of this equation over Δu and use that $q(\Delta u)$ is normalized, such that

$$\int \mathrm{d}\Delta u\, p(\Delta u) e^{\Delta u} = \left\langle e^{\Delta u} \right\rangle_p = 1 \tag{21}$$

Introducing the cumulant generating function,

$$G(t) = \ln \left\langle e^{t\Delta u} \right\rangle_p = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}, \tag{22}$$

we have $\ln \left\langle e^{\Delta u} \right\rangle_p = G(1) = 0$. κ_n is the nth cumulant of the p-ensemble. By solving $G(1) = 0$ for κ_1, we obtain

$$\langle \Delta u \rangle_p = -\frac{\mathrm{var}_p(\Delta u)}{2} - \sum_{n=3}^{\infty} \frac{\kappa_n}{n!} \tag{23}$$

We have used that the first two cumulants are the mean $\kappa_1 = \langle \Delta u \rangle_p = \mu$ and the variance $\kappa_2 = \mathrm{var}_p(\Delta u) = \sigma^2$.

Consequently, the KL divergence, Eq. (15), is given by a sum of the higher-order cumulants. The leading term is half the variance. For small errors in the energy, $|\Delta u| \ll 1$, the cumulants of order $n = 3$ and higher can be ignored,

$$\kappa_1 \approx -\frac{\kappa_2}{2} \tag{24}$$

allowing us to approximate the KL divergence in Eq. (15) by

$$S_{\mathrm{KL}}(p||q) = -\langle \Delta u \rangle \approx \frac{\mathrm{var}(\Delta u)}{2} \tag{25}$$

We have dropped the sub-script 'p' for average and variance. That is, these quantities without a subscript always refer to the refined ensemble $p(x)$.

The KL divergence is equally approximated by the variance of the configurational energy difference $\beta\Delta U$. The free-energy difference ΔF cancels in the variance, Eq. (14), such that $\mathrm{var}(\Delta u) = \mathrm{var}(\beta\Delta U)$ and

$$S_{\mathrm{KL}}(p||q) \approx \frac{\mathrm{var}(\beta\Delta U)}{2} \tag{26}$$

We next show that for small errors $|\Delta u|$ in the force field, where higher cumulants can be ignored, the KL divergence is approximately symmetric with respect to its arguments, $S_{\mathrm{KL}}(p||q) \approx S_{\mathrm{KL}}(q||p)$, where

$$S_{\mathrm{KL}}(q||p) = \langle \Delta u \rangle_q \tag{27}$$

We can express the average of Δu in the q-ensemble, using the cumulant expansion as before. Using that $p(\Delta u)$ is normalized, we obtain $\ln \left\langle e^{-\Delta u} \right\rangle_q = H(-1) = 0$, where we introduced the cumulant generating function for the q-ensemble

$$H(t) = \ln \left\langle e^{t\Delta u} \right\rangle_q = \sum_{n=1}^{\infty} \lambda_n \frac{t^n}{n!}, \tag{28}$$

Evaluating $H(-1) = 0$, we obtain

$$\langle \Delta u \rangle_q = \frac{\mathrm{var}_q(\Delta u)}{2} + \sum_{n=3}^{\infty} (-1)^n \frac{\lambda_n}{n!} \tag{29}$$

where we used that $\int \mathrm{d}\Delta u\, q(\Delta u) e^{-\Delta u} = \left\langle e^{-\Delta u} \right\rangle_q = 1$, Eq. (20). λ_n is the nth cumulant of the q-ensemble. For $|\Delta u| \ll 1$,

$$\lambda_1 \approx \frac{\lambda_2}{2} \tag{30}$$

such that

$$S_{\mathrm{KL}}(q||p) \approx \frac{\mathrm{var}_q(\beta\Delta U)}{2} \tag{31}$$

The cumulants of the p-ensemble can be expressed by the cumulants of the q-ensemble and the other way round.[28] From Eq. (20) follows that $G(t) = H(t-1)$. Inserting the corresponding cumulant expansions on both sides of this equation and collecting equal powers of t by applying the binomial theorem to $(t-1)^n$, we obtain

$$\kappa_k = \sum_{n=0}^{\infty} \frac{(-1)^n \lambda_{n+k}}{n!} \tag{32}$$

Correspondingly we obtain from $H(t) = G(t+1)$ that

$$\lambda_k = \sum_{n=0}^{\infty} \frac{\kappa_{n+k}}{n!} \tag{33}$$

From Eq. (32) and for $|\Delta u| \ll 1$, we obtain , $\kappa_1 = \lambda_1 - \lambda_2 \approx -\lambda_1$, where have used $\lambda_2 \approx 2\lambda_1$, Eq. (30). We find that the two definitions of the KL divergence are approximately equivalent,

$$S_{\mathrm{KL}}(p||q) = -\langle \Delta u \rangle_p \approx \langle \Delta u \rangle_q = S_{\mathrm{KL}}(q||p) \tag{34}$$

We obtain the same result using Eqs. (24) and (33).

D. Exact results for Gaussian energy distributions

If the refined distribution $p(\Delta u)$ of the energy error is Gaussian, then the approximations above are exact.[28,29] This case not only serves as a reference, but it is also of practical interest, as we shall see below (Ala_5).

For the Gaussian distribution, only the average and variance are non-zero. All higher-order cumulants are zero. Eq. (23) becomes

$$\mu = -\frac{\sigma^2}{2} \tag{35}$$

This constraint on the Gaussian distribution derives from Eq. (20) for normalized distribution $p(\Delta u)$ and $q(\Delta u)$.

In this Gaussian case, the approximation of the KL divergence by the variance half, Eqs. (25) and (26), is exact, i.e.,

$$S(p||q) = -\mu = \frac{\sigma^2}{2} \tag{36}$$

In the Gaussian case, the KL divergence is exactly symmetric with respect to its arguments, i.e., $S_{\mathrm{KL}}(p||q) = S_{\mathrm{KL}}(q||p)$.[30] Using Eq. (20), we can rewrite

$$S_{\mathrm{KL}}(q||p) = \langle \Delta u \rangle_q = \int \mathrm{d}\Delta u \, q(\Delta u) \Delta u \tag{37}$$

as

$$\begin{aligned} S_{\mathrm{KL}}(q||p) &= \int \mathrm{d}\Delta u \, p(\Delta u) e^{\Delta u} \Delta u \\ &= (\mu + \sigma^2) e^{\mu + \frac{\sigma^2}{2}} = -\mu = S_{\mathrm{KL}}(p||q) \end{aligned} \tag{38}$$

where we used Eq. (35).

E. Relation to free energy perturbation theory

We can now use the different expressions for the KL divergence to approximate the free-energy difference ΔF as in free energy perturbation theory.[31] Inserting Eq. (26) into Eq. (17), we obtain

$$\Delta F \approx \langle \Delta U \rangle + \beta \frac{\mathrm{var}(\Delta U)}{2} \tag{39}$$

Usually, ΔF is calculated in the q-ensemble, Eq. (18),

$$\Delta F \approx \langle \Delta U \rangle_q - \beta \frac{\mathrm{var}_q(\Delta U)}{2} \tag{40}$$

These approximations become exact for Gaussian distributions of the potential energy difference ΔU,[30] where $S_{\mathrm{KL}}(p||q) = S_{\mathrm{KL}}(q||p) = \mathrm{var}(\beta \Delta U)/2 = \mathrm{var}_q(\beta \Delta U)/2$.

F. Encoding prior knowledge in ensemble refinement

The relation of the KL divergence to the reduced energy differences allows us to relate the more abstract information difference expectation of Eq. (8) to the more physical quantity of the variance of the reduced energy difference, Eq. (26). We use the latter to choose the confidence parameter θ, which defines the rate at which the entropic prior $\exp(-\theta S_{\mathrm{KL}})$ in Eq. (3) decreases with increasing KL divergence.

The choice of the prior as an exponential function of the KL divergence can be motivated with the maximum entropy principle.[15] If we only know the expectation value of the KL divergence, $\bar{S}_{\mathrm{KL}}$, then the maximum entropy distribution of the KL divergence is given by $\exp(-S_{\mathrm{KL}}/\bar{S}_{\mathrm{KL}})$. In this case,

$$\theta = \frac{1}{\bar{S}_{\mathrm{KL}}} \tag{41}$$

In gentle ensemble refinement, we demand that $S_{\mathrm{KL}} \lesssim 1$, $S_{\mathrm{KL}}(p||q) \approx S_{\mathrm{KL}}(q||p)$, and $S_{\mathrm{KL}} \approx \mathrm{var}(\beta \Delta U)/2$ according to Eq. (26). For a given force-field error, we then expect $\bar{S}_{\mathrm{KL}} \approx \mathrm{var}(\beta \Delta U)/2$. According to Eq. (41), we set $\theta \cong 2/\mathrm{var}(\beta \Delta U)$ *a priori* as in Eq. (1). Beyond gentle ensemble refinement, where the expectation of the KL divergence is no longer determined by $\mathrm{var}(\beta \Delta U)$, we have to set $\bar{S}_{\mathrm{KL}}$ directly to encode our prior knowledge.

The choice of θ in the prior can be validated by checking its consistency with the optimal value of the KL divergence *a posteriori*. If we evaluate the prior for given θ and the corresponding optimal S_{KL} value, then we expect a reasonably high value of the prior. That is, θS_{KL} is of the order of one after refinement. If it is much larger, then the information from the experiment dominates the ensemble. We might have underestimated the quality of our ensemble or the size of the errors in the data. If it is much smaller, then the reference distribution dominates the ensemble. We might have overestimated the quality of our reference ensemble or the size of the errors in the data.

G. How to determine the energy uncertainty

In gentle ensemble refinement, we have to choose the energy uncertainty $\mathrm{var}(\beta \Delta U)$ *a priori* to properly set the confidence parameter θ according to Eq. (1). To do so, we go from configuration space to the space of observables. As introduced in the Appendix of Ref. 9, the probability density of the reference ensemble in observable space is given by

$$q(\mathbf{y}) = \int \mathrm{d}x \, q(x) \prod_{i=1}^{M} \delta[y_i(x) - y_i] \tag{42}$$

where y_i is the ith component of the observable vector $\mathbf{y}$. $\delta[\cdot]$ is Dirac's delta function. The probability density of

the refined ensemble $p(\mathbf{y})$ in observable space is defined analogously.

Refining in the space of observables is equivalent to refining in the space of conformations. By design, the average observables entering the likelihood are equal in both spaces. As we show in Appendix C, also the KL divergence in configuration space is equal to the KL divergence in observable space for the optimal solutions.

The observable space is a coarse-grained representation of the configuration space.[32] We introduce a coarse-grained energy potential energy $V_q(\mathbf{y})$ using

$$q(\mathbf{y}) = \frac{e^{-\beta V_q(\mathbf{y})}}{Q_q} \tag{43}$$

which gives

$$\beta V_q(\mathbf{y}) = -\ln q(\mathbf{y}) - \ln Q_q \tag{44}$$

Analogously we introduce the coarse-grained energy $V_p(\mathbf{y})$ in the p-ensemble. Note that Q_q and Q_p and thus ΔF are the same as introduced above in the configuration space. The coarse-grained reduced energy difference becomes

$$\Delta v(\mathbf{y}) = -\ln \frac{p(\mathbf{y})}{q(\mathbf{y})} = \beta V_p(\mathbf{y}) - \beta V_q(\mathbf{y}) - \beta \Delta F \tag{45}$$

The reduced energy uncertainty in observable space is a free-energy difference. Expressing $q(x)$ in Eq. (42) by Eq. (16), we obtain for the free-energy difference in Eq. (45)

$$\begin{aligned} \Delta v(\mathbf{y}) &= \ln \frac{\int \mathrm{d}x\, p(x)\, e^{\Delta u(x)} \prod_{i=1}^{M} \delta[y_i(x) - y_i]}{\int \mathrm{d}x p(x) \prod_{i=1}^{M} \delta[y_i(x) - y_i]} \\ &= \ln \left\langle e^{\Delta u(x)} \right\rangle_{p|\mathbf{y}} = -\ln \left\langle e^{-\Delta u(x)} \right\rangle_{q|\mathbf{y}} \end{aligned} \tag{46}$$

The subscripts $p|\mathbf{y}$ and $q|\mathbf{y}$ of the angular brackets indicate the sub-ensembles of the p- and q-ensembles with fixed values of $\mathbf{y}$. Consequently, $\Delta v(\mathbf{y})$ corresponds to the free-energy difference between the constrained p-ensemble and constrained q-ensemble for a given value of $\mathbf{y}$.

For the optimal weights obtained by BioEn, the energy uncertainties calculated in the space of conformations x and of observables $\mathbf{y}$ are equal, $\langle \Delta u \rangle_p = \langle \Delta v \rangle_{p(\mathbf{y})}$ (see Appendix C) and $\mathrm{var}_p(\Delta u) \approx \mathrm{var}_{p(\mathbf{y})}(\Delta v)$ in gentle ensemble refinement. The underlying reason is that in the BioEn optimal solution, the factor scaling the relative weight of conformation x depends only on $\mathbf{y}(x)$, not on x directly. Conformations with the same value of $\mathbf{y}$ are thus treated equally.

The uncertainty in the free-energy difference $\Delta v(\mathbf{y})$ depends on the type and number of observables. For different types of observables probing different aspects of molecular conformations, we will have different expectations about the energy error. For a polymer, the expected energy uncertainties for sub-ensembles with fixed end-to-end distance will be different than for sub-ensembles with fixed carbon-hydrogen bond lengths. The sub-ensembles for these two observables are quite different and so is our energy uncertainty.

If we combine independent observables and refine against all of them at once, then our uncertainty will be larger than for each individual observable. If observables are uncorrelated, then the probability distribution of the observable vector $q(\mathbf{y})$ factorize into probability distributions for individual components, $q(\mathbf{y}) = \prod_{i=1}^{M} q(y_i)$. The same is true for $p(\mathbf{y})$. The total KL divergence then becomes a sum over the KL divergences for individual components of the observable vector,

$$S_{\mathrm{KL}}(p(\mathbf{y})||q(\mathbf{y})) = \sum_{i=1}^{M} S_{\mathrm{KL}}(p(y_i)||q(y_i)) \tag{47}$$

In this case, we add up the energy variances for the individual components

$$S_{\mathrm{KL}}(p(\mathbf{y})||q(\mathbf{y})) \approx \frac{1}{2} \sum_{i=1}^{M} \mathrm{var}_{p(y_i)}(\Delta v) \tag{48}$$

to define θ. The energy uncertainty thus depends both on the type and number of observables, and on the thermodynamic state.

The sensitivity of the distribution $p\,(\mathbf{y}|c)$ of the observables to a particular force-field parameter c determines its impact on the respective energy uncertainty, an issue examined in detail in Bayesian inference of force fields (BioFF).[33] $p\,(\mathbf{y}|c)$ is defined analogously to Eq. (42) with $p(x)$ now paramaterized by c, i.e., $p(x|c) \propto \exp[-\beta U(x|c)]$ through the potential energy $U(x|c)$. The reference ensemble is defined by $c = c_0$. To lowest order, the KL divergence $S_{\mathrm{KL}}(c)$ grows quadratically with small changes $\delta c = c - c_0$ in the force-field parameter c,

$$\begin{aligned} S_{\mathrm{KL}}(c) &\approx \frac{\delta c^2}{2} \int d\mathbf{y}\; p\,(\mathbf{y}|c_0) \left(\frac{\partial \ln p\,(\mathbf{y}|c)}{\partial c} \right)^2_{c=c_0} \\ &= \frac{\delta c^2}{2} \left\langle \left(\frac{\partial \ln p(\mathbf{y}|c)}{\partial c} \right)^2 \right\rangle_{c=c_0} \end{aligned} \tag{49}$$

with a proportionality coefficient that is given by the expectation value of the squared mean force with respect to the parameter c. This relation follows from the definition of S_{KL} and the normalization condition of $p\,(\mathbf{y}|c)$. Note that errors in the force field might not only be due to inaccurate parameters but also due to their simplified functional forms.

In summary, we have to estimate the energy uncertainty in observable space to determine θ according to Eq. (1). We can use this θ-value to directly refine in configuration space. If we refine in observable space instead, then we use the optimal generalized forces as derived in Ref. 9 to obtain the refined ensemble in configuration space.

III. METHODS

We explore the concept of gentle ensemble refinement and the proposed encoding of prior knowledge using three example systems of increasing complexity. We refine a continuous version of the double-well model presented in Ref. 10 and a simple polymer model based on the von Mises probability distribution presented in Ref. 33 using synthetic data. We also refine fully atomistic simulations of the pentapeptide Ala_5 in explicit solvent[10] using data from nuclear magnetic resonance (NMR) experiments.[34]

As a simple model system, we define the reference ensemble in terms of a continuous double-well potential given by

$$U_q(x) = a(x^2 - {x_0}^2)^2 \tag{50}$$

where x is a scalar and also serves as observable, i.e., $y(x) = x$. This energy function is symmetric with respect to $x = 0$. It has two minima at $\pm x_0$ with values $U_q(\pm x_0) = 0$. These minima are separated by a barrier at $x = 0$ of a height given by $a{x_0}^4$. The corresponding Boltzmann distribution is given by

$$q(x) = \frac{e^{-U_q(x)}}{Q_q} \tag{51}$$

with the normalization constant (partition function) given by

$$Q_q = \frac{\pi}{2} x_0 \, e^{-c} \left[I_{-\frac{1}{4}}(c) + I_{\frac{1}{4}}(c) \right] \tag{52}$$

where $c = a{x_0}^4/2$ and $I_n(c)$ is the modified Bessel function of the first kind of order n. Note that $\langle x \rangle_q = 0$ due to symmetry.

In the following, we use $x_0 = 1$ and $a = 3$ for the reference distribution $q(x)$, such that the barrier height is $a{x_0}^4 = 3$ (Figure 2). The experimentally measured expectation value of the observables x is set to $Y = 0.8$. The error is $\sigma = 0.2$. We use rejection sampling to generate an ensemble of $N = 10000$ independent random points for ensemble refinement.

For concreteness, we also work out a specific case. The reference potential $U_q(x)$ deviates from the assumed true potential by the addition of a linear term, $U_p(x) = U_q(x) - b\,x$, with $U_q(x) = 3(x^2-1)^2$ as above. A slope of $b = 1.217165$ was chosen so that $\langle x \rangle_p = 0.8 = Y$ exactly. The corresponding force-field error is $\mathrm{var}_p(\Delta u) \approx 0.51$. For this error, we expect a value of $\theta = 2/\mathrm{var}_p(\Delta u) \approx 4$ to give a balanced fit, as will be tested.

As a more realistic example, we refine a two-dimensional polymer model[33] using synthetic one-dimensional data (Figure 3). The Boltzmann distribution is given by a product of von Mises distributions acting on the angle differences between neighboring bonds. In this phantom-chain model, the beads do not interact. We set the mean values of the angle differences to zero. All bonds have length one and we apply the same stiffness parameter $\kappa = 10$. We randomly generate conformations of polymers with 100 beads. We sample $N = 10000$

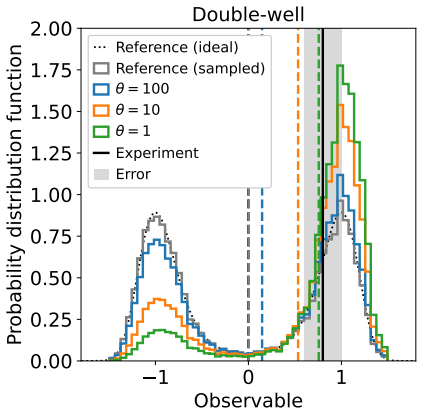

FIG. 2. Ensemble refinement of the double-well system using synthetic data. The black solid vertical line indicates the experimentally measured observable of $Y = 0.8$. The grey shaded area indicates $Y \pm \sigma$ with $\sigma = 0.2$. The reference probability distribution function of the observable for parameters $a = 3$ and $x_0 = 1$ is shown as black dotted line. The histogram of $N = 10000$ samples drawn from the reference distribution, Eq. (51), is shown as grey solid line. Histograms of x obtained with BioEn optimal weights for different θ values are shown in color. For all distribution functions, we show the average values of the observables as dashed vertical lines in the corresponding color. The calculated average values approach the experimental value for decreasing θ.

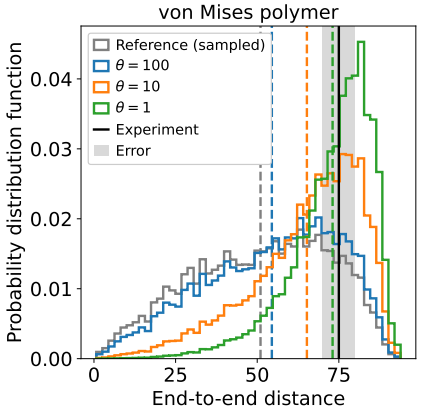

FIG. 3. Ensemble refinement of the polymer model[10] using synthetic data. The black solid vertical line indicates the experimentally measured observable of $Y = 75$. The grey shaded area indicates $Y \pm \sigma$ with $\sigma = 5$. The histogram of $N = 10000$ samples drawn from the Boltzmann distribution is shown as grey solid line. Histograms of the end-to-end distance obtained with BioEn optimal weights for different θ values are shown in color. For all distribution functions, we show the average values of the observables as dashed vertical lines in the corresponding color. The calculated average values approach the experimental value for decreasing θ.

independent conformations. As observable, we use the end-to-end distance. We set $Y = 75$ for the experimental value and an error of $\sigma = 5$.

As an example for an actual application, we refine previously published simulation data[10] of Ala_5 with experimental data in the form of NMR J-couplings.[34] The J-coupling calculation of Ref. 10 applied the Karplus equation[35] using the so-called DFT2 parameters from Ref. 36. We use $N = 50000$ conformations as in the original publication.

To find optimal solutions for the weights given a value of θ, we use the forces method[10] as implemented in the open-source BioEn software available at `https://github.com/bio-phys/BioEn`. An open-source Julia[37] implementation using the package Optim.jl[38] can be downloaded from `https://github.com/bio-phys/BioEn.jl` free of charge. To generate synthetic data, we use `https://github.com/bio-phys/RefinementModels.jl` for the double-well model and `https://github.com/bio-phys/BioFF` for the von Mises polymer model. We use uniform reference weights for all systems.

We illustrate the effects of refinement on the weights using cumulative ranked weights.[10] For the three systems, we compare numerical results for the cumulative ranked weights to analytical results for Gaussian energy distributions (see Appendix B).

IV. RESULTS

We first establish the range of validity for gentle ensemble refinement for the three example systems, as established in subsection II F. The range of validity of the approximation of the KL divergence by the energy uncertainty, Eq. (26), depends on the system under consideration (Figure 4). We find good agreement for decreasing θ values down to $(\theta, S_{\mathrm{KL}}) \approx (10, 0.17)$ for the double-well system; to $(\theta, S_{\mathrm{KL}}) \approx (10, 0.26)$ for the polymer model; and to $(\theta, S_{\mathrm{KL}}) \approx (1, 1.43)$ for Ala_5. In gentle ensemble refinement, the two definitions of the KL divergence are approximately equivalent, $S_{\mathrm{KL}}(p||q) \approx S_{\mathrm{KL}}(q||p)$ [Eq. (34) and Figure 4]. The relative residuals of $S_{\mathrm{KL}}(q||p)$ and of the approximation by half the energy variance with respect to $S_{\mathrm{KL}}(p||q)$ are of similar shape and magnitude but opposite sign (Figure 4, bottom panels). For all three systems, we have $S_{\mathrm{KL}} \lesssim 1$ for $\theta \gtrsim 10$, which overall delimits the regime of gentle refinement.

Having established the limits of validity for gentle ensemble refinement, we now show that already gentle refinement substantially improves the agreement between simulation and experiment. This improvement is illustrated by the L-curves defining Pareto fronts for Bayesian ensemble refinement (Figure 5). The L-curve consists of the optimal χ^2 values divided by the number datapoints M plotted against the optimal KL divergence values S_{KL} for different θ values. The three chosen θ values (100,10,1) cover the elbow regions of the respective L-curves. Note that these elbow regions are not sharply defined. For gentle refinement at $\theta = 10$, χ^2 has drastically decreased while S_{KL} remains relatively small.

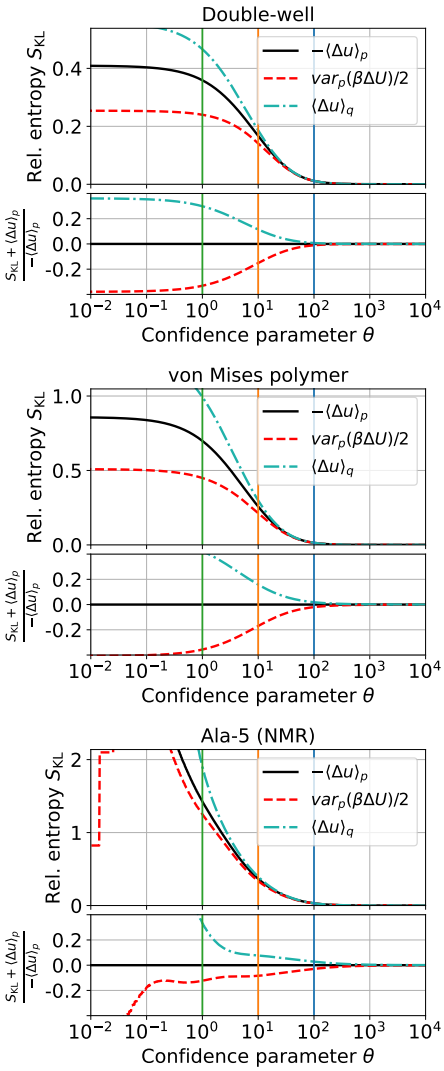

FIG. 4. Comparison of the KL divergence $S_{\mathrm{KL}}(p||q)$, Eq. (15), used in BioEn (black) to its approximation (dashed red line) given by half the variance of the reduced energy change $\beta\Delta U$, Eq. (26), and to its alternative definition $S_{\mathrm{KL}}(q||p)$ (dashed-dotted cyan line), Eq. (27). We show results for the double-well system (top), the polymer mode (middle), and Ala_5 (bottom). The bottom panels show the relative difference of $-\mathrm{var}_p(\Delta u)$ (red dashed) and $S_{\mathrm{KL}}(q||p)$ (cyan dot-dashed) with respect to $S_{\mathrm{KL}}(p||q) = -\langle\Delta u\rangle_p$. The approximations roughly start deviating from the values given by the exact expression at $(\theta, S_{\mathrm{KL}}) \approx (10, 0.17)$ for the double-well system, $(\theta, S_{\mathrm{KL}}) \approx (10, 0.26)$ for the polymer model, and $(\theta, S_{\mathrm{KL}}) \approx (1, 1.43)$ for Ala_5. For smaller θ values, we leave the regime of gentle refinement. Vertical lines indicate $\theta = 100$ (blue), 10 (orange), and 1 (green).

Overfitting and underfitting are illustrated by the double-well system (see Figure 2). We have underfitting for $\theta = 100$ as the calculated average is multiple standard

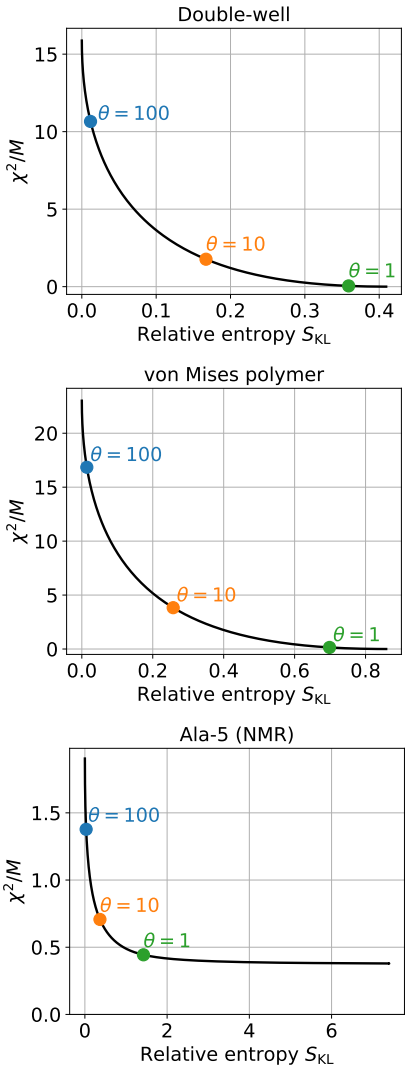

FIG. 5. The L-curve plots of the optimal χ^2/M vs. the optimal KL divergence S_{KL} for the double-well system (top), the polymer model (middle), and Ala_5 (bottom). M is the number of data points. Annotated disks indicate the optimal values for $\theta = 100$ (blue), 10 (orange), and 1 (green).

deviations away from the experimental value. We have overfitting for $\theta = 1$ as the calculated average agrees nearly perfectly with the experimental average despite the error. Both extreme cases are unlikely with respect to the Gaussian likelihood, Eq. (5). The character of the distribution of the observable can change quite substantially even when the optimal entropy values appear relatively small, as we discuss in more detail below. For the optimal ensembles, the statistical weights of the left state [with observable values $y(x) = x < 0$] are given by 42%, 23%, 12% for $(\theta, S_{KL}) = (100, 0.01), (10, 0.17)$, and $(1, 0.36)$, respectively, compared to 50% for the reference state.

For the concrete example of a linear perturbation (Methods), we obtain a balanced solution for $\theta \approx 4$, estimated using Eq. (1), between under- and overfitting. In particular, the calculated mean after refinement of $\langle x \rangle \approx 0.66$ matches the target $Y = 0.8 \pm 0.2$ within the error. $\theta S_{KL} \approx 1.05$ is close to one, as expected. In this sense, the θ estimate from Eq. (1) is indeed consistent.

The approximation of the KL divergence by the variance, Eq. (26), holds for arbitrarily shaped distribution functions of the reduced energy differences Δu (Figure 6, top). The distribution functions are bimodal for the double-well system, monomodal for the polymer model, and well approximated by Gaussian distributions for Ala_5. The approximation of the KL divergence by the variance works especially well for the latter, as it is exact for a Gaussian Δu distribution.

Refined weights with small entropy values can already be quite different from the reference weights (Figure 7). To illustrate this point, we focus on results for $\theta = 10$ located in the elbow region of the L-curve (see Figure 5) and within the regime of gentle refinement. We show the cumulative ranked weights in Figure 7. For $\theta = 10$, the S_{KL} values are small and range from 0.17 for the double-well system, over 0.26 for the polymer model, to 0.37 for Ala_5. In these three cases, the cumulative weights show that the top half of the conformations already have a cumulative probability of 80%, which is quite a substantial change. The relative weight of top ranked conformations further increases with increasing KL divergence.

In Figure 7, we compare numerical results of the ranked-weight distributions to analytical results obtained by assuming Gaussian distributions of Δu, parameterized by $\langle \Delta u \rangle$ and $\mathrm{var}(\Delta u)$ calculated from the weights (see Appendix B). For the largest value of θ and thus the smallest values of S_{KL}, the agreement is excellent. However, for smaller values of θ, the agreement deteriorates somewhat. For the double-well system (Figure 7, left column), the analytical approximation captures the trends qualitatively but fails quantitatively because the underlying distribution of Δu is clearly non-Gaussian. For the polymer model, the distribution of Δu is skewed, but unimodal for all θ values considered (see Figure 6), resulting in near-quantitative agreement of the Gaussian-based approximations and the actual weight distributions after refinement. For the most realistic case presented here, Ala_5, the Δu distribution is well approximated by a Gaussian. Consequently, the approximations of the weight distributions are quite accurate, even for the smallest value of $\theta = 1$ with $S_{KL} \approx 1.43$.

V. DISCUSSION

When refining sufficiently gently, the KL divergence is well approximated by the expected energy uncertainty, Eq. (26). We can quantify and visualize the extent of the resulting weight changes by S_{KL} and a plot of the cumulative ranked weights, respectively. KL divergences $S_{KL} \ll 1$ and a narrow gap between the respective cumulative ranked weights indicate good overlap between

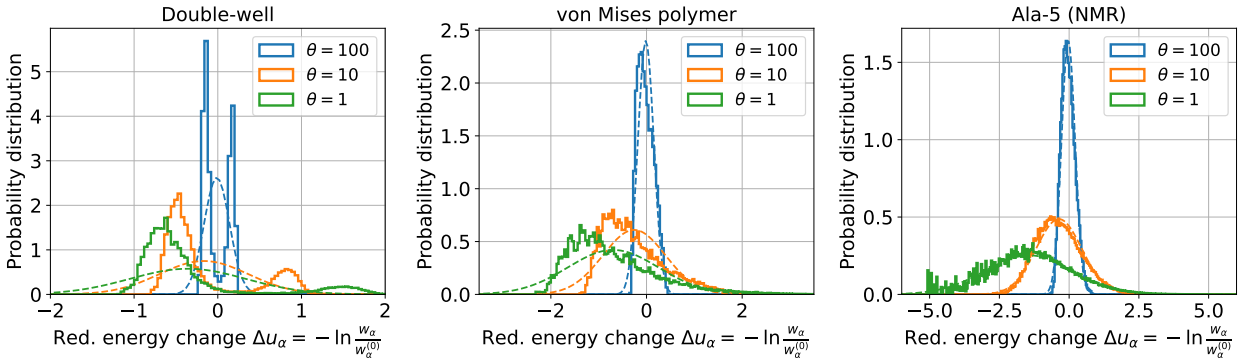

FIG. 6. Histograms (solid lines) of the energy changes Δu_α for $\theta = 100$ (blue), 10 (orange), and 1 (green) for the double-well system (left), the polymer model (center), and Ala_5 (right). Gaussian distributions of the same mean and variance are shown as dashed lines in matching colors.

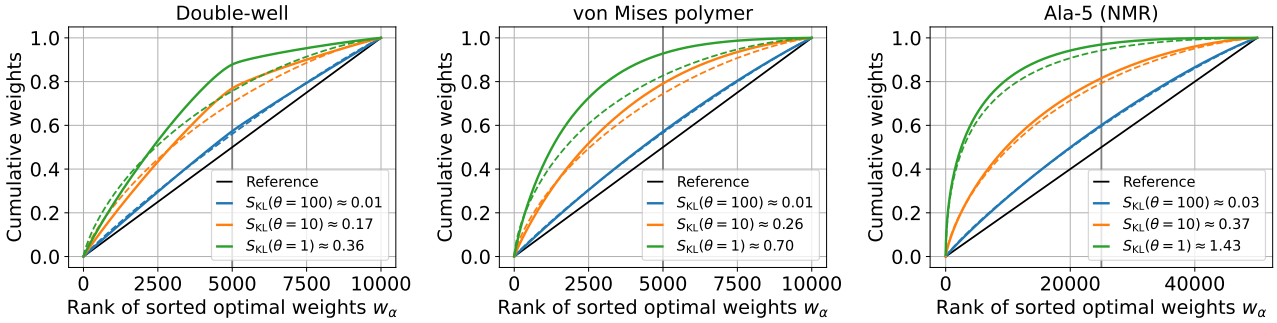

FIG. 7. The cumulative of the optimal weights sorted in descending order as function of their rank for the double-well system (left), the polymer model (center), and Ala_5 (right) for different values of the confidence parameter $\theta = 100$ (blue), 10 (orange), and 1 (green). The black lines indicate the results for the reference. The dashed lines show analytical results, Eq. (B5), for a Gaussian distribution of Δu parameterized by the mean $\mu = \langle \Delta u \rangle$ and variance $\sigma^2 = \mathrm{var}(\beta \Delta \mathrm{U})$. The grey vertical line indicates half the ensemble size for reference.

reference and refined ensembles.

Poor overlap, as indicated by large weight changes, can be improved by enhanced sampling methods that enrich the sample before refinement to better match the experimental observables, e.g., by using replica simulations.[24,25] We then reweight the sampled ensemble using MBAR or binless WHAM to produce a reference ensemble for subsequent BioEn ensemble refinement, as we have proposed in Ref. 9. By doing so, we can remove any biases, e.g., due to a finite number of replicas. One can also directly bias the degrees of freedom determining the values of the observables, e.g., by using empirical force-field refinement, as described for example in Ref. 33. Also in this case, we can generate a reference ensemble by reweighting the sampled ensembles for subsequent BioEn ensemble refinement.

Even for gentle refinement, the changes to the reference weights can be noticeable. For $\theta = 10$, we found the top 50% of the conformations to carry ∼80% of the weight in our three examples, consistent with the Gaussian approximation (Figure 7). Despite the small entropy values of $S_{\mathrm{KL}} \approx 0.2$ to 0.4, these changes to the weights are sufficient to substantially reduce the deviations from experiment (see L-curves in Figure 5).

We suggest to use our prior knowledge about the force-field accuracy in the space of observables to set the confidence parameter as $\theta \cong 2/\mathrm{var}(\beta \Delta U)$, Eq. (1). One can justifiably deviate from this proposal to set θ according to available prior knowledge. For example, if larger values of the variance are less likely than implied by the exponential form of the prior with $\theta \cong 2/\mathrm{var}(\beta \Delta U)$, then one should increase θ accordingly. In any case, the knowledge of the expected variance $\mathrm{var}(\beta \Delta U)$ can be applied to set and interpret the scaling parameter θ. Conversely, a particular choice of θ by other reasoning is also an expression of the expected errors in state populations of the reference ensemble. Whereas functions other than an exponential function could be used to define the prior, the simple relation to the force-field error would most likely be lost.

We build up experience about the force-field uncertainties quite naturally. Experienced researchers performing simulations will often be able to state expectations about the energy uncertainties of their favorite force fields for a class of observables. However, the mean-squared force-field error can also be learned from repeated ensemble

refinements against different experimental data across a variety of systems.

VI. CONCLUDING REMARKS

The results presented here can be generalized to non-Boltzmann distributions and are thus valid for the general class of Bayesian/MaxEnt approaches for ill-defined inverse problems. In such cases, we take advantage of the fact that the energy entering Boltzmann's distribution corresponds to Shannon's information content $h(x) = -\ln p(x)$ for arbitrary probability distributions $p(x)$.[39] Consequently, the difference in energy corresponds to a difference in information, $\Delta h(x) = \ln[p(x)/q(x)]$, Eq. (9). Like in gentle ensemble refinement, the KL divergence can be approximated by the variance of this information difference, i.e., $S_{\mathrm{KL}} \approx \mathrm{var}(\Delta h)/2$, if these differences are small. In general, we express our confidence in the 'natural unit of information' or 'nat', which is numerically equal to $k_{\mathrm{B}}T$ for Boltzmann distributions.

Gentle ensemble refinement is a powerful tool for molecular simulations and modeling. Empirical force fields rely on approximations in their functional form to trade off efficiency and accuracy. Therefore, not all errors in the force field can be resolved just by reparameterization. Also, ensembles generated from other methods than molecular simulations suffer from approximations and benefit from ensemble refinement.[20–22] The inaccuracies introduced by such approximations can be alleviated using gentle ensemble refinement to integrate system-specific information, most often in the form of experimental data.

ACKNOWLEDGMENTS

We thank Dr. Jakob T. Bullerjahn, Lisa Pietrek, and Prof. Carlo Camilloni for helpful discussions and the Max Planck Society for financial support.

Appendix A: Kullback-Leibler divergence for finite ensembles

For the sake of completeness, we show here how we estimate the KL divergence, defined in Eq. (8) and in Eq. (15) for Boltzmann distributions, from finite ensembles. Consequently, the results presented in the main text for continuous probability densities also apply to finite ensembles. As we show below, the discrete weights w_α and $w_\alpha^{(0)}$ generally correspond to ratios of probability densities. Note that in the following we use $p_\alpha = w_\alpha$ and $q_\alpha = w_\alpha^{(0)}$ to reduce visual clutter and emphasize the general validity of our results.

In general, we can calculate the information difference expectation in Eq. (8) from an arbitrary normalized probability distribution $\tilde{q}(x)$ by reweighting,

$$S_{\mathrm{KL}}(p||q) = \left\langle \ln \frac{p(x)}{q(x)} \right\rangle_p = \left\langle \frac{p(x)}{\tilde{q}(x)} \ln \frac{p(x)}{q(x)} \right\rangle_{\tilde{q}} \tag{A1}$$

On the right-hand side, we form the expectation with respect to $\tilde{q}(x)$. Equivalently, we obtain

$$S_{\mathrm{KL}}(p||q) = \left\langle \frac{p(x)}{\tilde{q}(x)} \ln \left(\frac{p(x)\tilde{q}(x)}{\tilde{q}(x)q(x)} \right) \right\rangle_{\tilde{q}} \tag{A2}$$

by inserting $1 = \tilde{q}(x)/\tilde{q}(x)$ into the logarithm of Eq. (A1). We use this expression to estimate the KL divergence from finite samples and properly interpret discrete weights.

Using Eq. (A2), we can numerically estimate $S_{\mathrm{KL}}(p||q)$ from a sample of conformations x_α sampled from $\tilde{q}(x)$, i.e., $x_\alpha \sim \tilde{q}(x)$, as

$$S_{\mathrm{KL}}(p||q) \;\approx\; \frac{1}{N} \sum_{\alpha=1}^{N} \frac{p(x_\alpha)}{\tilde{q}(x_\alpha)} \ln \frac{p(x)\tilde{q}(x)}{\tilde{q}(x)q(x)} \tag{A3}$$

Exploiting that the probability distributions $p(x)$ and $q(x)$ only show up as ratios with $\tilde{q}(x)$, we can introduce normalized weights or probabilities W_α for the finite sample as

$$W_\alpha = \frac{W(x_\alpha)}{\tilde{q}(x_\alpha)} \left[\sum_{\gamma=1}^{N} \frac{W(x_\gamma)}{\tilde{q}(x_\gamma)} \right]^{-1} \tag{A4}$$

such that $\sum_{\alpha=1}^{N} W_\alpha = 1$ for the finite ensemble. For the reference ensemble, $W(x_\alpha) = q(x_\alpha)$ with discrete weights $W_\alpha = q_\alpha$. For the refined ensemble, $W(x_\alpha) = p(x_\alpha)$ with discrete weights $W_\alpha = p_\alpha$. Importantly, $W(x_\alpha)$ and $\tilde{q}(x_\alpha)$ do not have to be normalized to calculate these weights.

For normalized probability distributions $p(x)$ and $q(x)$, the normalization term in the square bracket above is equal to N,

$$W_\alpha = \frac{1}{N} \frac{W(x_\alpha)}{\tilde{q}(x_\alpha)} \tag{A5}$$

because $x_\alpha \sim \tilde{q}(x)$ and

$$\begin{aligned} \frac{1}{N} \left[\sum_{\gamma=1}^{N} \frac{W(x_\gamma)}{\tilde{q}(x_\gamma)} \right] \;&\approx\; \left\langle \frac{W(x)}{\tilde{q}(x)} \right\rangle_{\tilde{q}} \\ &= \int \mathrm{d}x\, W(x) = 1 \end{aligned} \tag{A6}$$

Using Eq. (A5) for the corresponding probability ratios in Eq. (A3), we obtain for the numerical estimate of the KL divergence

$$S_{\mathrm{KL}}(p||q) \approx \sum_{\alpha=1}^{N} p_\alpha \ln \frac{p_\alpha}{q_\alpha} = -\sum_{\alpha=1}^{N} p_\alpha \Delta u_\alpha \tag{A7}$$

Importantly, we can calculate the discrete weights p_α and q_α using Eq. (A4) without the need for calculating partition functions. If we sample form the reference distribution, $\tilde{q}(x) = q(x)$, then $q_\alpha = 1/N$ and $p_\alpha = \exp(-\Delta u_\alpha)/\sum_{\gamma=1}^{N}\exp(-\Delta u_\gamma)$. If we sample from the refined distribution, $\tilde{q}(x) = p(x)$, then $q_\alpha = \exp(\Delta u_\alpha)/\sum_{\gamma=1}^{N}\exp(\Delta u_\gamma)$ and $p_\alpha = 1/N$.

In BioEn ensemble refinement, we calculate the probabilities q_α for the reference ensembles as $1/N$ for unbiased simulations in agreement with Eq. (A4). We obtain reference weights from biased simulations by reweighting, e.g., using MBAR[40,41] or (binless) WHAM.[42–45] We obtain the optimal probabilities p_α by maximizing the posterior, Eq. (2), or equivalently by minimizing the negative log-likelihood, Eq. (7).

Appendix B: Cumulative ranked weights

We derive derive an analytical expression for the cumulative ranked weights for a Gaussian distribution of the energy change Δu. With mean $\mu = \langle \Delta u \rangle$ and variance $\sigma^2 = \mathrm{var}(\Delta u)$, the weights w are distributed according to a log-normal distribution for uniform reference weights,

$$f(w|\tilde{\mu},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}w}\exp\left[-\frac{(\ln w - \tilde{\mu})^2}{2\sigma^2}\right] \tag{B1}$$

where $\tilde{\mu} = -\mu - \ln N$. The cumulative distribution of the weights then is

$$F(w|\tilde{\mu},\sigma^2) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln w - \tilde{\mu}}{\sqrt{2}\sigma}\right)\right] \tag{B2}$$

where $\mathrm{erf}(\cdot)$ is the error function. We define the cumulative average of weights w as

$$\mathrm{c}(w) = \int_0^w w' f(w'|\tilde{\mu},\sigma^2)\mathrm{d}w' \tag{B3}$$

We normalize this function by the average weight $c(\infty)$ such that

$$\frac{c(w)}{c(\infty)} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\tilde{\mu} + \sigma^2 - \ln 2}{\sqrt{2}\sigma}\right)\right] \tag{B4}$$

The cumulative value of the weights, sorted in descending order, as a function of the rank now corresponds to $1 - F(w|\tilde{\mu},\sigma^2)$ as a function of $1 - \mathrm{c}(w)/\mathrm{c}(\infty)$. Introducing $r = 1 - \mathrm{c}(w)/\mathrm{c}(\infty)$, with $0 \le r \le 1$, we can rewrite $1 - F(w|\tilde{\mu},\sigma^2)$ as a function of r. We obtain the cumulative ranked weights function

$$\mathrm{crw}(r) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\mathrm{erf}^{-1}(1-2r) - \frac{\sigma}{\sqrt{2}}\right)\right] \tag{B5}$$

where $\mathrm{erf}^{-1}(\cdot)$ is the inverse error function. To compare this expression with the cumulative weights of a finite ensemble of size N with discrete rank, we plot $\mathrm{crw}(n/N)$ as a function of the continuous variable n with $0 \le n \le N-1$.

Note that if the approximation of the KL divergence in terms of the variance alone holds, then Eq. (B5) is parameterized by the KL divergence, $\sigma/\sqrt{2} \approx \sqrt{S_{\mathrm{KL}}}$. For example, the weight fraction of the top-half of the weights is $\mathrm{crw}(1/2) \approx \frac{1}{2}\left[1 + \mathrm{erf}\left(\sqrt{S_{\mathrm{KL}}}\right)\right]$. These approximations are exact in the Gaussian case.

Appendix C: Equivalence of Kullback-Leibler divergence in the space of conformations and observables

As we have shown in Ref. 9, the BioEn refinement can be performed equivalently in the space of conformations x and in the space of observables $\mathbf{y}$. Here, we show that the KL divergences calculated in the respective spaces have identical numerical values.

Following the Appendix in Ref. 9, let $q(\mathbf{y})$ and $p(\mathbf{y})$ be the distributions in the space of observables according to Eq. (42). As shown in the lead-up to Eq. (A5) of Ref. 9, these distributions can be written in terms of a vector $\mathbf{z}$ of constants z_i as

$$p(x) = \frac{q(x)\exp\left(\sum_{i=1}^{M} y_i(x)\, z_i\right)}{\int \mathrm{d}x'\, q(x')\exp\left(\sum_{i=1}^{M} y_i(x')\, z_i\right)}\,, \tag{C1}$$

$$p(\mathbf{y}) = \frac{q(\mathbf{y})\exp\left(\sum_{i=1}^{M} y_i\, z_i\right)}{\int \mathrm{d}\mathbf{y}'\, q(\mathbf{y}')\exp\left(\sum_{i=1}^{M} y_i'\, z_i\right)}\,. \tag{C2}$$

where we use $\mathrm{d}\mathbf{y}' = \prod_{i=1}^{M}\mathrm{d}y_i'$. Importantly, the coefficients z_i are the same in both configuration and observable space.[9]

The KL divergence in observable space $\mathbf{y}$ is by definition,

$$\begin{aligned} S_{\mathrm{KL}}^{(y)} &= \int \mathrm{d}\mathbf{y}\, p(\mathbf{y}) \ln\frac{p(\mathbf{y})}{q(\mathbf{y})} \\ &= \int \mathrm{d}\mathbf{y}\, p(\mathbf{y}) \sum_{i=1}^{M} y_i z_i - \ln \int \mathrm{d}\mathbf{y} q(\mathbf{y}) e^{\sum_{i=1}^{M} y_i z_i} \end{aligned}$$

We now rewrite the KL divergence in the space of conformations x in terms of observables $\mathbf{y}$ by inserting Dirac

https://doi.org/10.26434/chemrxiv-2023-71rr6-v2 **ORCID:** https://orcid.org/0000-0001-8367-1077

delta functions,

$$\begin{aligned} S_{\mathrm{KL}}^{(x)} &= \int dx\, p(x) \ln \frac{p(x)}{q(x)} \\ &= \int \mathrm{d}x\, p(x) \sum_{i=1}^{M} y_i(x) z_i - \ln \int \mathrm{d}x\, p(x) e^{\sum_{i=1}^{M} y_i(x) z_i} \\ &= \int \mathrm{d}\mathbf{y} \int \mathrm{d}x\, p(x) \prod_{i=1}^{M} \delta[y_i(x) - y_i] \sum_{i=1}^{M} y_i(x) z_i \\ &\quad - \ln \int \mathrm{d}\mathbf{y} \int \mathrm{d}x\, p(x) \prod_{i=1}^{M} \delta[y_i(x) - y_i] e^{\sum_{i=1}^{M} y_i(x) z_i} \\ &= \int \mathrm{d}\mathbf{y}\, p(\mathbf{y}) \sum_{i=1}^{M} y_i z_i - \ln \int \mathrm{d}\mathbf{y} q(\mathbf{y}) e^{\sum_{i=1}^{M} y_i z_i} \end{aligned} \quad \text{(C3)}$$

We find that the KL divergences in the x and $\mathbf{y}$ spaces are identical,

$$S_{\mathrm{KL}}^{(x)} = S_{\mathrm{KL}}^{(\mathbf{y})} \equiv S_{\mathrm{KL}} \quad \text{(C4)}$$

We can thus estimate S_{KL} in either space.

[1] S. Gull and G. Daniell, Nature **272**, 686 (1978).
[2] M. Nilges, M. Habeck, and W. Rieping, Comptes Rendus Chimie **11**, 356 (2008).
[3] A. B. Ward, A. Sali, and I. A. Wilson, Science **339**, 913 (2013).
[4] S. Bottaro and K. Lindorff-Larsen, Science **361**, 355 (2018).
[5] E. T. Jaynes, in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice (Cambridge University Press, 1986) pp. 27–58.
[6] S. Kullback and R. A. Leibler, Ann. Math. Statist. **22**, 79 (1951).
[7] P. C. Hansen and D. P. O'Leary, SIAM J. Sci. Comput. **14**, 1487 (1993).
[8] B. Różycki, Y. C. Kim, and G. Hummer, Structure **19**, 109 (2011).
[9] G. Hummer and J. Köfinger, J. Chem. Phys. **143**, 243150 (2015).
[10] J. Köfinger, L. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer, J. Chem. Theory Comput. **15**, 3390 (2019).
[11] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, in *2011 31st International Conference on Distributed Computing Systems Workshops* (2011) pp. 166–171, iSSN: 2332-5666.
[12] A. Cultrera and L. Callegaro, IOPSciNotes **1**, 025004 (2020), publisher: IOP Publishing.
[13] S. F. Gull, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Springer Netherlands, Dordrecht, 1989) pp. 53–71.
[14] J. Skilling, in *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*, edited by J. Skilling (Springer Netherlands, Dordrecht, 1989) pp. 45–52.
[15] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[16] J. W. Pitera and J. D. Chodera, J. Chem. Theory Comput. **8**, 3445 (2012).
[17] A. Cesari, A. Gil-Ley, and G. Bussi, J. Chem. Theory Comput. **12**, 6192 (2016).
[18] S. Bottaro, G. Bussi, S. D. Kennedy, D. H. Turner, and K. Lindorff-Larsen, Sci. Adv. **4**, eaar8521 (2018).
[19] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by P. F. Fougère (Springer Netherlands, Dordrecht, 1990) pp. 341–350.
[20] K. Reichel, L. Stelzl, J. Köfinger, and G. Hummer, J. Phys. Chem. Lett. **9**, 5748 (2018).
[21] L. S. Stelzl, L. M. Pietrek, A. Holla, J. Oroz, M. Sikora, J. Köfinger, B. Schuler, M. Zweckstetter, and G. Hummer, JACS Au **2**, 673 (2022).
[22] L. M. Pietrek, L. S. Stelzl, and G. Hummer, J. Chem. Theory Comput. (2024), https://doi.org/10.1021/acs.jctc.3c01049, (accepted).
[23] S. Bottaro, T. Bengtsen, and K. Lindorff-Larsen, in *Methods in Molecular Biology*, Vol. 2112, edited by Z. Gáspári (Humana, New York, NY, 2020) pp. 219–240.
[24] R. B. Best and M. Vendruscolo, J. Am. Chem. Soc. **126**, 8090 (2004).
[25] M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo, Sci. Adv. **2**, 1 (2016).
[26] E. T. Jaynes, in *Statistical Physics* (K. Ford (ed.), Benjamin, New York, 1963) p. p. 181, section: Information Theory and Statistical Mechanics.
[27] T. Dannenhoffer-Lafage, A. D. White, and G. A. Voth, J. Chem. Theory Comput. **12**, 2144 (2016).
[28] G. Hummer, J. Chem. Phys. **114**, 7330 (2001).
[29] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **100**, 12564 (2003).
[30] G. Hummer, L. R. Pratt, and A. E. Garcia, J. Phys. Chem. **99**, 14188 (1995).
[31] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[32] W. G. Noid, J. Phys. Chem. B **127**, 4174 (2023).
[33] J. Köfinger and G. Hummer, Eur. Phys. J. B **94**, 245 (2021).
[34] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe, J. Am. Chem. Soc. **129**, 1179 (2007).
[35] M. Karplus, J. Chem. Phys. **30**, 11 (1959).
[36] D. A. Case, C. Scheurer, and R. Brüschweiler, J. Am. Chem. Soc. **122**, 10390 (2000).
[37] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM Rev. **59**, 65 (2017).
[38] P. K. Mogensen and A. N. Riseth, J. Open Source Softw. **3**, 615 (2018).
[39] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[40] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[41] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[42] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[43] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[44] M. Souaille and B. Roux, Comput. Phys. Commun. **135**, 40 (2001).
[45] E. Rosta, M. Nowotny, W. Yang, and G. Hummer, J. Am. Chem. Soc. **133**, 8934 (2011).