JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

On the Use of Experimental Observations to Bias Simulated Ensembles

Jed W. Pitera*,[†] and John D. Chodera[‡]

[†]IBM Almaden Research Center, San Jose, California

[‡]California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, California

ABSTRACT: Historically, experimental measurements have been used to bias biomolecular simulations toward structures compatible with those observations via the addition of *ad hoc* restraint terms. We show how the maximum entropy formalism provides a principled approach to enforce concordance with these measurements in a minimally biased way, yielding restraints that are linear functions of the target observables and specifying a straightforward scheme to determine the biasing weights. These restraints are compared with instantaneous and ensemble-averaged harmonic restraint schemes, illustrating their similarities and limitations.

I. INTRODUCTION

Many biophysical experiments (e.g., X-ray diffraction, NMR, FRET) measure properties sensitive to molecular geometry (interatomic distances, angles, etc.) and therefore provide information about molecular conformations populated at equilibrium. With a sufficient number of observations, detailed structural models of highly populated conformations can be constructed.[1] One of the earliest applications of biomolecular simulation was structure refinement.[2,3] In typical refinement applications, biasing potentials or restraints are used to drive the system toward conformations consistent with experimental observations. The physical information encoded in the molecular topology and interatomic potential complements the restraints and allows the efficient generation of a structure or set of structures consistent with experimental observations and with chemical knowledge. However, most experimental observations correspond to ensemble averages, providing information about the mean of a distribution which may possess significant variance or multiple modes. As a result, it can be impossible for a single structure to simultaneously satisfy the full set of experimental data available.[4] These limitations have long been recognized, and methods like time-averaged restrained molecular dynamics[5,6] and ensemble dynamics/dynamic ensemble refinement[7,8] explicitly attempt to generate the ensemble of structures whose ensemble-averaged observables agree with the data. However, in underdetermined cases, there may be many distributions whose ensemble averages satisfy experimental constraints, but which impose varying degrees of bias on the resulting distribution.[9] In the present work, we derive the unique *minimally biased* choice of restraint potential that can be used to bias a simulated ensemble toward assuming a given set of experimental measurements on average without introducing unphysical artifacts, and compare it against other approaches. In particular, we show that ensemble-averaged refinement schemes can achieve this minimally biased result under certain conditions.

It should be noted that there are examples where maximum entropy approaches were used *post hoc* to reweight clustered simulation data to form an ensemble reproducing experimental measurements. These post hoc approaches have been used for satisfying experimental NMR constraints[10,11] or SAXS observations.[12] In the present work, however, we extend the use of maximum entropy methods to derive restraints that can be applied as biasing potentials during a molecular simulation to satisfy an experimental restraint in a straightforward manner.

II. DERIVATION

We illustrate the issue with a general form of the problem. Consider some experimental observable f and corresponding observation (experimental measurement) f_{exp}. For simplicity, we assume f is a function of the vector of atomic coordinates $\vec{r}$ only and consider only the configurational partition function of the system. This excludes from consideration spectroscopic observables that depend on time correlation functions, but our formalism is readily extended into the time domain through the use of path sampling Monte Carlo schemes.[13] We will also make the simplifying assumption that f_{exp} is known exactly, with no uncertainty or error. Bayesian approaches, like those advocated by Nilges and co-workers,[14] offer a promising route for the inclusion of error, potentially of unknown magnitude, in these measurements.

For a given potential $U(\vec{r})$, typically a molecular mechanics force field, an unbiased simulation in the canonical ensemble yields the unbiased value f_0:

$$f_0 = \langle f(\vec{r}) \rangle = \frac{\int \mathrm{d}\vec{r}\, f(\vec{r})\, \mathrm{e}^{-\beta U(\vec{r})}}{\int \mathrm{d}\vec{r}\, \mathrm{e}^{-\beta U(\vec{r})}} = \int \mathrm{d}\vec{r}\, f(\vec{r})\, p_0(\vec{r}) \tag{1}$$

where p_0 is the unbiased probability density function of the coordinate vector $\vec{r}$, β is the inverse temperature $1/k_{\mathrm{B}}T$ of the simulation, and k_{B} is the Boltzmann constant. If our computational model of the system is perfect and sampling is complete, f_0 will equal f_{exp}. However, imperfect models and incomplete sampling generally result in a disagreement between the unbiased simulation result f_0 and f_{exp}. A maximum entropy approach allows

Special Issue: Wilfred F. van Gunsteren Festschrift

Published: August 21, 2012

us to derive the unique way to bias our simulation to ensure agreement with the experimental result without adding additional unwarranted distortions to the ensemble, or assuming additional information about the distribution of configurations that is not carried by the experimental measurement. Following the maximum entropy approach to thermodynamics introduced by Jaynes,[15] we consider a biased simulation which yields the result f_{exp} from the biased probability density function p_1:

$$f_1 = f_{\text{exp}} = \int \mathrm{d}\vec{r}\, f(\vec{r})\, p_1(\vec{r}) \tag{2}$$

We can think of this relationship as one of three constraint equations defining p_1:

$$c_0 = \int \mathrm{d}\vec{r}\, p_1(\vec{r}) - 1 \tag{3}$$

$$c_1 = \int \mathrm{d}\vec{r}\, U(\vec{r})\, p_1(\vec{r}) - \frac{3}{2}Nk_{\text{B}}T \tag{4}$$

$$c_2 = \int \mathrm{d}\vec{r}\, f(\vec{r})\, p_1(\vec{r}) - f_{\text{exp}} \tag{5}$$

where choosing p_1 such that each c_i is zero ensures that p_1 is normalized (eq 3), has average potential energy given by the thermal energy $3/2Nk_{\text{B}}T$ (eq 4), and yields the experimental measurement in expectation (eq 5). To infer the new biased probability density function p_1, we maximize its entropy S

$$S = -\int \mathrm{d}\vec{r}\, p_1(\vec{r}) \ln p_1(\vec{r}) \tag{6}$$

subject to these three constraints by introducing Lagrange multipliers. This process is equivalent to minimizing the additional information content of p_1 relative to p_0 and yields the functional extremum condition

$$\frac{\partial S}{\partial p} - \lambda_0\frac{\partial c_0}{\partial p} - \lambda_1\frac{\partial c_1}{\partial p} - \lambda_2\frac{\partial c_2}{\partial p} = 0 \tag{7}$$

Substituting the corresponding derivatives, rearranging, and factoring out constants yields the minimally biased solution for p_1 of

$$p_1(\vec{r}) = \frac{\mathrm{e}^{-\lambda_1 U(\vec{r}) - \lambda_2 f(\vec{r})}}{\int \mathrm{d}\vec{r}\, \mathrm{e}^{-\lambda_1 U(\vec{r}) - \lambda_2 f(\vec{r})}} \tag{8}$$

Recognizing λ_1 as the inverse temperature β, we can think of p_1 as arising from sampling a system with a modified potential energy:

$$U_{\text{ME}}(\vec{r};\, \alpha) = U(\vec{r}) + \alpha f(\vec{r}) \tag{9}$$

where

$$\alpha = \frac{\lambda_2}{\beta} \tag{10}$$

The second term in eq 9, which is linear in $f(\vec{r})$, is the only form of restraint or biasing potential that can be used to impose the experimental observation f_{exp} on the system without introducing additional uncontrolled bias not contained in the measurement. Any other functional form will reduce the entropy of the system and alter the observed distribution by a greater degree than is warranted by the single observation, effectively assuming additional information not communicated by the measurement. The linear restraint coefficient α (which depends on the unknown Lagrange multiplier λ_2 and has units of energy divided by the units of the observable, f) is uniquely determined by ensuring satisfaction of the corresponding constraint ($c_2 = 0$ in eq 5).

A straightforward approach to determine α (following ref 16) is to define a convex objective function $\Gamma(\alpha)$

$$\Gamma(\alpha) = \ln Z(\alpha) + \alpha\beta f_{\text{exp}} \tag{11}$$

where $Z(\alpha)$is the α-dependent partition function,

$$Z(\alpha) = \int \mathrm{d}\vec{r}\, \mathrm{e}^{-\beta U_{\text{ME}}(\vec{r};\alpha)} \tag{12}$$

The partial derivative of $\Gamma(\alpha)$ can be written as

$$\frac{\partial\Gamma}{\partial\alpha} = -\beta\langle f\rangle + \beta f_{\text{exp}} \tag{13}$$

showing that at α^*, a particular value of α such that the partial derivative (eq 13) is equal to zero, and assuming β (the inverse temperature of the measurement and the simulation) is finite, the constraint equation (eq 5) is exactly satisfied. The convexity of $\Gamma(\alpha)$, and hence the existence of a unique solution, can be guaranteed by noting that $\partial^2\Gamma/\partial\alpha^2 = \beta^2(\langle f\rangle^2) - \langle f\rangle^2)$, which is positive everywhere.

Determination of α^* can be done in several ways. The objective $\Gamma(\alpha)$ could be minimized by the use of iterative simulation to compute the gradient (eq 16, below) and potentially the Hessian (eq 17) for use in an optimization scheme. Simulation at every trial value of α could be avoided by the use of reweighting methods from one or more sets of simulation data (such as the multistate Bennett acceptance ratio[17]). Alternatively, weak-coupling methods could propagate α in a manner that attempts to minimize $\Gamma(\alpha)$.[18] For any scheme, simulating at the final estimated α^* can verify that the constraint equation (eq 5) is satisfied.

If we have multiple independent experimental observations $\{f_{\text{exp},i}\}$ to satisfy, the functional form of eq 9 is easily generalized to

$$U_{\text{ME}}(\vec{r};\, \{\alpha_i\}) = U(\vec{r}) + \sum_{i=1}^{M} \alpha_i f_i(\vec{r}) \tag{14}$$

where each of the M experimental observations generates its own maximum entropy restraint term $\alpha_i\, f_i(\vec{r})$ with corresponding unknown linear restraint coefficient α_i. The M values α_i must be simultaneously determined by one of the schemes described above. The corresponding convex objective $\Gamma(\{\alpha_i\})$ is given by

$$\Gamma(\{\alpha_i\}) = \ln Z(\{\alpha_i\}) + \beta\sum_{i=1}^{M} \alpha_i f_{\text{exp},i} \tag{15}$$

with its gradient and Hessian now given by

$$\frac{\partial\Gamma}{\partial\alpha_i} = -\beta\langle f_i\rangle + \beta f_{\text{exp},i} \tag{16}$$

$$\frac{\partial^2\Gamma}{\partial\alpha_i\partial\alpha_j} = \beta^2(\langle f_i f_j\rangle - \langle f_i\rangle\langle f_j\rangle) \tag{17}$$

From eq 17, a complication of this multiple observation case is clear. The convexity of $\Gamma(\{\alpha_i\})$ depends on the covariance of all pairs of observables f_i and f_j. Uncorrelated pairs of observables correspond to null off-diagonal elements of the Hessian. If all the observables are uncorrelated with one another, the Hessian matrix will be positive definite and $\Gamma(\{\alpha_i\})$ therefore convex. However, the presence of one or more pairs of correlated (for

which eq 17 will be positive) or anticorrelated (for which eq 17 will be negative) observables may alter the character of the Hessian to be positive semidefinite (i.e., having one or more zero eigenvalues). If this occurs, $\Gamma(\{\alpha_i\})$ is no longer convex. In such a case, there is no guarantee that a set of linear restraint coefficients $\{\alpha_i\}$ can be found that will ensure the simulated ensemble simultaneously satisfies the full set of observations. This is expected to be a particular problem for observables that are linear functions of the particle coordinates. For example, consider a single particle uniformly distributed on the one-dimensional interval $[0,1]$. If we have two linear observations $f_1 = x = 0.25$ and $f_2 = 1-x = 0.25$, there are no possible choices of α_1 and α_2 that ensure the simulated ensemble simultaneously satisfies both conditions. For observables which are nonlinear functions of the particle coordinates, this problem is expected to be less severe as the ensemble average of each observable will tend to be strongly influenced by a different portion of configuration space.

III. COMPARISON TO OTHER RESTRAINT SCHEMES

As noted earlier, there are three restraint schemes commonly used to enforce agreement between a simulated ensemble and an experimental observation: instantaneous, time-averaged, and ensemble-averaged harmonic restraints. In an instantaneous harmonic restraint scheme, quadratic residuals are added to the potential to bias the value of the observable for the instantaneous configuration, $f(\vec{r})$, toward the measurement f_{exp}:

$$U_{\text{HR}}(\vec{r}) = U(\vec{r}) + \frac{k_r}{2}(f(\vec{r}) - f_{\text{exp}})^2 \tag{18}$$

where k_r denotes the restraint force constant. This yields terms both linear and quadratic in $f(\vec{r})$, in contrast to the single linear term prescribed by the maximum entropy approach. While the instantaneous harmonic restraint attempts to ensure that eq 2 is satisfied, it also inappropriately affects the variance of f. For example, if such a restraint is applied to a model initially having a uniformly distributed f, the mean of the resulting f_1 will approximate f_{exp}, but the variance becomes $(\beta k_r)^{-1}$. In effect, a single piece of information (the measurement f_{exp}) is being incorrectly used to constrain two moments of f_1, when the measurement only provides information about one moment. In addition, no absolute prescription for the selection of k_r is generally given, unlike our constraint equation (eq 2) that guides the selection of α. There are many variants of this type of restraint, including half-harmonic and "flat-well" restraints which only act on $f(\vec{r})$ when it moves outside a selected region.[3,19] These are also inconsistent with our derived result, as the biasing potential $\alpha f(\vec{r})$ has to be applied uniformly across all configurations. Otherwise, the choice of "biased" and "unbiased" regions represents additional information added into the sampling which will alter the distribution.

Time-averaged harmonic restraint molecular dynamics schemes[6] use similar quadratic terms but act on a historical time average rather than the instantaneous value of $f(\vec{r})$.

$$U_{\text{TR}}(\vec{r}) = U(\vec{r}) + \frac{k_r}{2}\left(\frac{1}{\tau(1 - e^{-T/\tau})}\int_{t_0-T}^{t_0} dt\, e^{-(t_0-t)/\tau} f(\vec{r}(t)) - f_{\text{exp}}\right)^2 \tag{19}$$

The nonstationary nature of this potential places it outside the scope of the present work, but the idea of biasing the observable averaged over multiple structures toward the experimental observation was further expanded in a stationary form as ensemble-averaged harmonic restraints, discussed below.

Ensemble-averaged harmonic restraints were initially introduced by Fennen et al. to generate an ensemble of representative conformations from NMR data.[20] The method builds on the idea of time-averaged restraints by attempting to generate a finite set of configurations representative of the true experimentally biased equilibrium distribution directly. The method was extended and popularized by Vendruscolo and co-workers as "dynamic ensemble refinement" or "ensemble dynamics".[7,8] In ensemble-averaged harmonic restraint schemes, the ensemble is explicitly simulated as a set of N replicas of the system (each with coordinate vector $\vec{r}_i$, i = 1 to N) subjected to biasing potentials which attempt to bring instantaneous averages over the simulated set in agreement with experimental results

$$U_{\text{ED}}(\vec{r}_i) = U(\vec{r}_i) + N\frac{k_r}{2}\left(\frac{1}{N}\sum_{i=1}^{N} f(\vec{r}_i) - f_{\text{exp}}\right)^2 \tag{20}$$

The biasing potential is multiplied by the number of replicas N to keep the magnitude of the restraint force roughly independent of N. The method is also commonly applied in an annealing approach, rather than as a direct biased simulation. If we expand eq 20 and retain only the $\vec{r}_i$-dependent terms influencing the replica where i = 1, we get

$$U_{\text{ED}}(\vec{r}_i) = U(\vec{r}_i) + N\frac{k_r}{2}\left(\frac{1}{N^2}f(\vec{r}_1)\sum_{i=1}^{N} f(\vec{r}_i) - \frac{2}{N}f(\vec{r}_1)f_{\text{exp}}\right) \tag{21}$$

$$U_{\text{ED}}(\vec{r}_i) = U(\vec{r}_i) + \frac{k_r}{2N}f(\vec{r}_1)^2 + \frac{k_r}{2N}f(\vec{r}_1)\sum_{i=2}^{N} f(\vec{r}_i) - k_r f_{\text{exp}} f(\vec{r}_1) \tag{22}$$

From this rearrangement, we see that the ensemble-averaged biasing includes terms of the correct linear order in $f(\vec{r}_1)$ but has an additional unwarranted quadratic term. In the limit where N becomes very large, only the linear term remains, yielding

$$\lim_{N\to\infty} U_{\text{ED}}(\vec{r}_1) = U(\vec{r}_1) - k_r f_{\text{exp}} f(\vec{r}_1) \tag{23}$$

which recovers the ideal minimally biased maximum entropy form of eq 9. Interestingly, this only occurs because of the additional factor of N multiplying the force constant k_r. Without it, all the ensemble-averaged restraint terms vanish in the large N limit. If there is a single observation, it can be correctly satisfied by choosing the factor $-k_r f_{\text{exp}}$ to be equal to λ_2/β, just like its counterpart α in eq 10. In this way, the ensemble-averaged functional form achieves the minimally unbiased functional form in the large N limit. For multiple observations, a problem arises in that each observation requires its own unique and unknown constant α_i to reach the maximum entropy solution, and there is no guarantee that the constant factor $-k_r f_{\text{exp},i}$ will yield the target value of each observable in a standard ensemble-averaged harmonic restraint refinement simulation. The value of k_r sufficient to make the simulated ensemble reproduce one observation $f_{\text{exp},1}$ might prevent it from simultaneously reproducing a different observation $f_{\text{exp},2}$. Observations that are already nearly reproduced by the unbiased simulated ensemble require little additional restraint, corresponding to small restraint force constants, while observations where the unbiased simulated

dx.doi.org/10.1021/ct300112v | J. Chem. Theory Comput. 2012, 8, 3445−3451

ensemble is far from experimental values can require substantial biasing forces, corresponding to large restraint force constants. At a minimum, it would be necessary to vary the force constant k_r to different values $k_{r,i}$ for each restraint, which could be done using a scheme similar to optimization of the convex function Γ above, with expectations replaced by ensemble means. A variant of ensemble-averaged harmonic restraints which varies the number of replicas used for different observables, designed to address issues of over- and underfitting to experimental data, may effectively be an *ad hoc* solution to the need for different force constants for different restraints.[21] Finally, the consistency between ensemble-averaged harmonic restraints in the large N limit and maximum entropy restraints only holds for experimental measurements that are *linear* expectations of observables of the form of eq 1 above. For nonlinearly averaged observables like Nuclear Overhauser Effect (NOE) derived distances (which have a r^{-6} or r^{-3} dependence on r), ensemble-averaged harmonic restraints are often expressed in terms of the nonlinear average of the function applied to the coordinates, here represented by the averaging function F and its inverse F^{-1} (e.g., $F(x) = x^{-6}$ and $F^{-1}(x) = x^{-1/6}$):

$$U_{\mathrm{ED}}(\vec{r}_i) = U(\vec{r}_i) + N\frac{k_r}{2}\left(F^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}F(f(\vec{r}_i))\right) - f_{\mathrm{exp}}\right)^2 \tag{24}$$

In contrast, the maximum entropy form of a nonlinearly averaged observable consists of the restraint form

$$U_{\mathrm{ME}}(\vec{r};\,\alpha) = U(\vec{r}) + \alpha F(f(\vec{r})) \tag{25}$$

and associated constraint equation

$$c_2 = F^{-1}\left(\int \mathrm{d}\vec{r}\; F(f(\vec{r}))\, p_1(\vec{r})\right) - f_{\mathrm{exp}} \tag{26}$$

The differences between these two forms can result in systematic deviations of nonlinear ensemble-averaged harmonic restrained ensembles from maximum entropy restrained ensembles, even in the large N limit. Numerical results showing these deviations are presented in section IV.

IV. ANALYTICAL AND NUMERICAL COMPARISONS

Instantaneous and ensemble-averaged harmonic restraints can be compared to maximum entropy restraints in simple systems to illustrate the nature of the biases they introduce. In the case of a one-dimensional harmonic oscillator fixed to the origin with spring constant k_s and at an inverse temperature β, the problem is simple enough that analytic solutions are possible for maximum entropy or instantaneous harmonic restraints. For maximum entropy restraint of the value of α needed to shift $\langle x\rangle$ to x_{exp} is

$$\alpha = -k_s x_{\mathrm{exp}} \tag{27}$$

From eq 23 above, ensemble-averaged harmonic restraints in the large N limit yield a similar result so long as the restraint force constant k_r is set equal to k_s.

In contrast, for instantaneous harmonic restraints there is no unique solution for the needed k_r. Instead, the expected value of $\langle x\rangle$ is

$$\langle x\rangle = \frac{k_r x_{\mathrm{exp}}}{k_s + k_r} \tag{28}$$

The target value is recovered exactly only when either k_s is zero or k_r is infinite. Figure 1 shows the observed distributions from

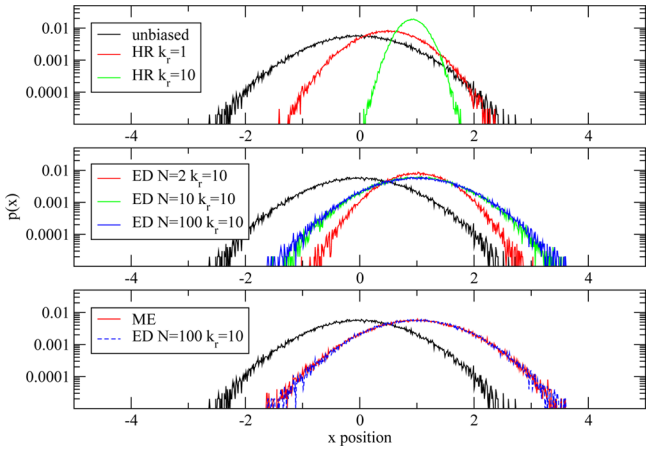

Figure 1. Simulated probability distributions for a one-dimensional harmonic oscillator with spring constant $k_s = 2$ in the presence of different restraint schemes, biased toward the linear average $\langle x\rangle = 1$. The black curve in each panel shows the distribution for the unbiased harmonic oscillator. The panels show: (top) instantaneous harmonic restraint with k_r values of 2 (red) or 20 (green); (center) ensemble-averaged harmonic restraint results for a k_r value of 10 and $N = 2$ (red), 10 (green), or 100 (blue) replicas in the ensemble; (bottom) the maximum entropy restraint result for $\alpha = -k_s$ (red) along with the $N = 100$ ensemble-averaged harmonic restraints result (dashed blue).

Metropolis Monte Carlo simulations (10^8 trial moves of unit displacements) using each of these restraint schemes. The bias introduced by the instantaneous restraint is evident by the large effect on the variance of the distribution, and the ensemble-averaged scheme shows biased instantaneous restraint-like behavior at small N that recovers the maximum entropy result for large N. Unlike the instantaneous restraints, the maximum entropy restraint shows almost no effect on the variance, while shifting the mean to exactly satisfy the experimental measurement.

A simulation of the interparticle distance r that would be sampled by two Lennard-Jones particles ($\sigma = 1$, $\varepsilon = 1$) in three dimensions illustrates the effects of nonlinear averaging. The experimental observation is taken to be either linear in r or a nonlinear r^{-6} (as in FRET and some NOE experiments) average of the interparticle separation biased toward either short (4σ) or long (20σ) distances. Without an applied restraint, the average value of $\langle r\rangle$ is 18.7σ. Figure 2 shows Metropolis Monte Carlo simulation results (5×10^7 trials of new interparticle distances r drawn with probability proportional to r^2 from the interval $[0, 25\sigma]$) for each restraint scheme. For $N = 100$, linear ensemble-averaged harmonic restraints (panel 2B) yield a distribution very similar to the proper maximum entropy restraint result (panel 2C). In the nonlinear case, however, ensemble-averaged harmonic restraints with $N = 100$ incorrectly yield no population in Lennard-Jones contact when biasing to ensure $\langle r^{-6}\rangle^{-1/6} = 4$ (panel 2D).

For a more biomolecular example, we consider a system of two particles representing the two ends of a random coil polymer of contour length 25σ. As above, the two particles interact with a Lennard-Jones potential ($\sigma = 1$, $\varepsilon = 1$), but this potential is now combined with a Gaussian distribution representing the random coil entropy. The Gaussian distribution was introduced by adding a harmonic potential in r with a force constant of 0.002. Both linear and nonlinear averages of the interparticle distance r were considered. Without an applied restraint, the average value of $\langle r\rangle$ is 5.4σ. The panels of Figure 3 show Metropolis Monte

dx.doi.org/10.1021/ct300112v | J. Chem. Theory Comput. 2012, 8, 3445−3451

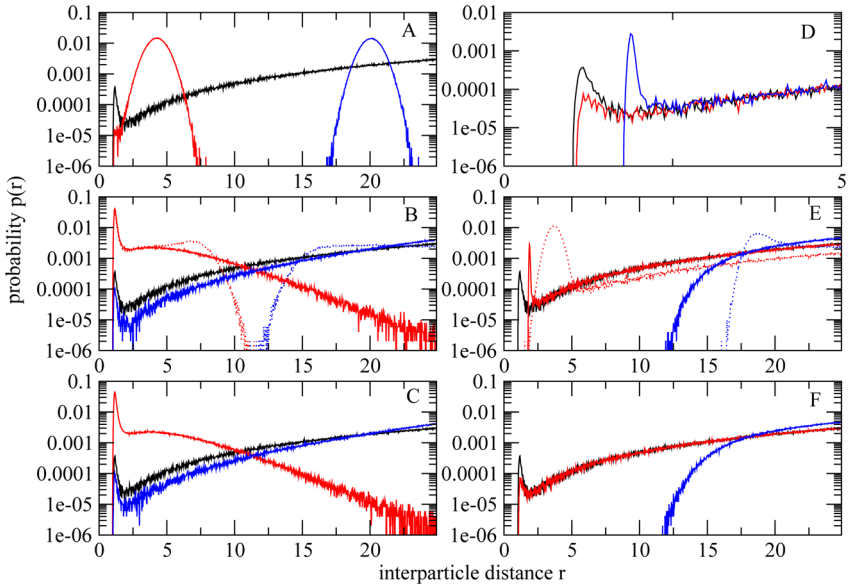

Figure 2. Simulated probability distributions for the separation r between a pair of Lennard-Jones particles ($\sigma = 1$, $\varepsilon = 1$). The black curve in each panel shows the unbiased probability distribution. Red curves correspond to restraints targeting an ensemble averaged value of $\langle r \rangle = 4$ or $\langle r^{-6} \rangle^{-1/6} = 4$, while blue curves correspond to $\langle r \rangle = 20$ or $\langle r^{-6} \rangle^{-1/6} = 20$. The panels show (A) instantaneous harmonic restraint in r with $k_r = 2$; (B) ensemble-averaged harmonic restraint in $\langle r \rangle$ with $k_r = 2$ and $N = 2$ (dashed lines) or 100 (solid lines); (C) maximum entropy restraint in $\langle r \rangle$; (E) ensemble-averaged harmonic restraint in $\langle r^{-6} \rangle^{-1/6}$ with $k_r = 2$ and $N = 2$ (dashed lines) or 100 (solid lines); (F) maximum entropy restraint in $\langle r^{-6} \rangle^{-1/6}$. Panel D shows a detailed view of the short-ranged probability distributions for the unbiased case (black), the $N = 100$ ensemble-averaged harmonic restraint distributions for $\langle r^{-6} \rangle^{-1/6} = 4$ from panel E (blue), and the corresponding maximum entropy restraint distribution yielding $\langle r^{-6} \rangle^{-1/6} = 4$ from panel F (red).

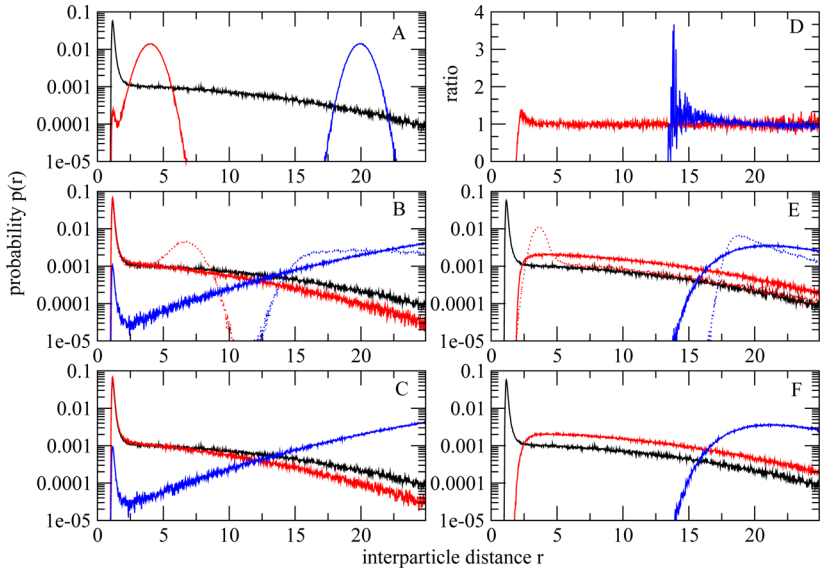

Figure 3. Simulated probability distributions for the separation r between a pair of Lennard-Jones particles ($\sigma = 1$, $\varepsilon = 1$) representing the ends of a random coil polymer in the presence of different restraint schemes. The black curve in each panel shows the unbiased probability distribution. Red curves correspond to restraints targeting an ensemble averaged value of $\langle r \rangle = 4$ or $\langle r^{-6} \rangle^{-1/6} = 4$ while blue curves correspond to $\langle r \rangle = 20$ or $\langle r^{-6} \rangle^{-1/6} = 20$. The panels show (A) instantaneous harmonic restraint in r with $k_r = 2$; (B) ensemble-averaged harmonic restraint in $\langle r \rangle$ with $k_r = 2$ and $N = 2$ (dashed lines) or 100 (solid lines); (C) maximum entropy restraint in $\langle r \rangle$; (E) ensemble-averaged harmonic restraint in $\langle r^{-6} \rangle^{-1/6}$ with $k_r = 2$ and $N = 2$ (dashed lines) or 100 (solid lines); (F) maximum entropy restraint in $\langle r^{-6} \rangle^{-1/6}$. Panel D shows the ratio between the $N = 100$ ensemble-averaged harmonic restraint distributions from panel E and the corresponding maximum entropy restraint distribution from panel F.

Carlo simulation results (5×10^7 trials of new interparticle distances drawn uniformly from the interval $[0, 25\sigma]$) for each restraint scheme when the experimental observable corresponds to either a short (4σ) or long (20σ) interparticle distance. Panel 3D shows the ratio of the ensemble-averaged harmonic restraint probabilities with $N = 100$ to the corresponding maximum entropy result for the nonlinear case. In this case, the ensemble-averaged harmonic restraint distributions systematically exaggerate the probability of short interparticle separations.

For nonlinear averages, the error introduced by the ensemble-averaged harmonic restraint method depends greatly on both the nonlinear averaging function and the unbiased distribution of the restrained observable. The inhomogenous weighting of configuration space by the nonlinear averaging means that there are

many possible distributions which can satisfy a nonlinearly averaged observation.[9] For the r^{-6} averaging we have considered, any average will be dominated by short interparticle separations. In the hypothetical case of an ensemble of structures, one of which is at a short interparticle separation and the remainder of which are at long interparticle separations, the restraint term of eq 24 effectively becomes a harmonic restraint acting only on the single short-range structure. In the Lennard-Jones case of Figure 2, the unbiased population is dominated by large interparticle separations. Agreement with the experimental observation is achieved in the ensemble-averaged case by substantial distortions to the small population of short-range structures, leaving the long-range populations largely untouched. In contrast, in the polymer case of Figure 3, the unbiased population predominantly consists of short interparticle distances already, so the distortion is manifest as smaller changes across a range of distances.

V. CONCLUSIONS

The maximum entropy approach to statistical mechanics provides a simple, unambiguous recipe for the use of experimental observations to bias molecular simulations. These "maximum entropy restraints" are linearly proportional to the instantaneous value of the observable, with constants of proportionality uniquely determined by the constraint that the observable expectations satisfy the corresponding experimental observations. By contrast, traditional harmonic restraints introduce strong biases in the distribution. However, the ensemble-averaged harmonic restraints approach is shown to approach the maximum entropy result for linearly averaged observables in the limit that the ensemble contains many replicas. For nonlinearly averaged observables, this result does not hold, and ensemble-averaged harmonic restraints still introduce some bias even in that limit. Specifically, ensemble-averaged harmonic restraints for a NOE-like distance observable can either over- or underestimate the population at short range, yielding a distorted picture of the correct ensemble. This can be corrected by enforcing constraints in the linearly averaged form of the observable. As shown by the present results, maximum entropy restraints are free of such biases and obviate the need to simulate many coupled replicas of the system since the ensemble-averaged restraint distributions converge toward the maximum entropy restraint distributions in the large N limit.

While experimental measurements are not free of error, there has been rapid development in recent years on sophisticated Bayesian approaches to treating the uncertainty in the true measurements given observations.[14] While these methods typically enforce that a *single* structure match all experimental observables simultaneously, rather than an ensemble, this Bayesian machinery for treating the uncertainty in true observations is expected to be useful in extending the maximum-entropy framework presented here to create simulation ensembles representative of experimental measurements contaminated with error.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pitera@us.ibm.com. Phone: +1(408)927-2084. Fax: +1(408)927-2100.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Prof. Gerhard Stock (University of Freiburg) for stimulating discussions on the topic of maximum entropy. J.D.C. acknowledges funding from a California Institute for Quantitative Biosciences (QB3) Distinguished Postdoctoral Fellowship.

DEDICATION

The authors dedicate this work to Prof. Wilfred F. van Gunsteren to celebrate the happy occasion of his 65th birthday.

REFERENCES

(1) Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993; p 507.
(2) Brunger, A. T.; Nilges, M. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* **1993**, *26* (1), 49−125.
(3) Kaptein, R.; Zuiderweg, E. R.; Scheek, R. M.; Boelens, R.; van Gunsteren, W. F. A protein structure from nuclear magnetic resonance data. lac repressor headpiece. *J. Mol. Biol.* **1985**, *182* (1), 179−182.
(4) Kessler, H.; Griesinger, C.; Lautz, J.; Mueller, A.; van Gunsteren, W. F.; Berendsen, H. J. C. Conformational Dynamics Detected by Nuclear Magnetic Resonance NOE Values and J Coupling Constants. *J. Am. Chem. Soc.* **1988**, *110* (11), 3393−3396.
(5) Torda, A. E.; Brunne, R. M.; Huber, T.; Kessler, H.; van Gunsteren, W. F. Structure refinement using time-averaged J-coupling constant restraints. *J Biomol. NMR* **1993**, *3* (1), 55−66.
(6) Torda, A. E.; Scheek, R. M.; van Gunsteren, W. F. Time-averaged nuclear Overhauser effect distance restraints applied to Tendamistat. *J. Mol. Biol.* **1990**, *214* (1), 223−235.
(7) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433* (7022), 128−132.
(8) Best, R. B.; Vendruscolo, M. Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.* **2004**, *126* (26), 8090−8091.
(9) Burgi, R.; Pitera, J.; van Gunsteren, W. F. Assessing the effect of conformational averaging on the measured values of observables. *J. Biomol. NMR* **2001**, *19* (4), 305−320.
(10) Groth, M.; Malicka, J.; Czaplewski, C.; Oldziej, S.; Lankiewicz, L.; Wiczk, W.; Liwo, A. Maximum entropy approach to the determination of solution conformation of flexible polypeptides by global conformational analysis and NMR spectroscopy–application to DNS1-c-[D-A2,bu2,Trp4,Leu5]enkephalin and DNS1-c-[D-A2bu2,Trp4,D-Leu5]-enkephalin. *J. Biomol. NMR* **1999**, *15* (4), 315−330.
(11) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *J. Am. Chem. Soc.* **2007**, *129* (5), 1179−1189.
(12) Rozycki, B.; Kim, Y. C.; Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011**, *19* (1), 109−116.
(13) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(14) Rieping, W.; Habeck, M.; Nilges, M. Inferential structure determination. *Science* **2005**, *309* (5732), 303−306.
(15) Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106* (4), 620−630.
(16) Mead, L. R.; Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **1984**, *25* (8), 2404−2417.
(17) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129* (12), 124105.
(18) Njo, S. L.; van Gunsteren, W. F.; Mueller-Plathe, F. Determination of force field parameters for molecular simulation by molecular simulation: An application of the weak-coupling method. *J. Chem. Phys.* **1995**, *102* (15), 6199−6207.

(19) Pearlman, D. A.; Kollman, P. A. Are time-averaged restraints necessary for nuclear magnetic resonance refinement? A model study for DNA. *J. Mol. Biol.* **1991**, *220* (2), 457−479.
(20) Fennen, J.; Torda, A. E.; van Gunsteren, W. F. Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. *J. Biomol. NMR* **1995**, *6* (2), 163−170.
(21) Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR* **2007**, *37* (2), 117−135.