Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments

Andrea Cesari [iD], Sabine Reißer and Giovanni Bussi *

Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy; acesari@sissa.it (A.C.); sreisser@sissa.it (S.R.)
*   Correspondence: bussi@sissa.it; Tel.: +39-040-3787407

Abstract: Molecular dynamics (MD) simulations allow the investigation of the structural dynamics of biomolecular systems with unrivaled time and space resolution. However, in order to compensate for the inaccuracies of the utilized empirical force fields, it is becoming common to integrate MD simulations with experimental data obtained from ensemble measurements. We review here the approaches that can be used to combine MD and experiment under the guidance of the maximum entropy principle. We mostly focus on methods based on Lagrangian multipliers, either implemented as reweighting of existing simulations or through an on-the-fly optimization. We discuss how errors in the experimental data can be modeled and accounted for. Finally, we use simple model systems to illustrate the typical difficulties arising when applying these methods.

Keywords: molecular dynamics; maximum entropy principle; ensemble averages; experimental constraints

1. Introduction

Molecular dynamics (MD) simulations are nowadays a fundamental tool used to complement experimental investigations in biomolecular modeling [1]. Although the accessible processes are usually limited to the microsecond timescale for classical MD with empirical force fields, with the help of enhanced sampling methods [2–4] it is possible to effectively sample events that would require a much longer time in order to spontaneously happen. However, the quality of the results is still limited by the accuracy of the employed force fields, making experimental validations a necessary step. The usual procedure consists in performing a simulation and computing some observable for which an experimental value has been already measured. If the calculated and experimental values are compatible, the simulation can be trusted and other observables can be estimated in order to make genuine predictions. If the discrepancy between calculated and experimental values is significant, one is forced to make a step back and perform a new simulation with a refined force field. For instance, current force fields still exhibit visible limitations in the study of protein-protein interactions [5], in the structural characterization of protein unfolded states [6], in the simulation of the conformational dynamics of unstructured RNAs [7–9], and in the blind prediction of RNA structural motifs [9–11]. However, improving force fields is a far-from-trivial task because many correlated parameters should be adjusted. Furthermore, the employed functional forms might have an intrinsically limited capability to describe the real energy function of the system. Largely due to these reasons, it is becoming more and more common to restrain the simulations in order to enforce agreement with experimental data. Whereas this approach might appear not satisfactory, one should keep in mind that often experimental knowledge is anyway implicitly encoded in the simulation of complex systems (e.g., if the initial structure of a short simulation is taken from experiment, then the simulation will be biased toward it). In addition, one should consider that validation can still be made against independent experiments or against some of the data suitably removed from the set of restraints. From another point of view,

the pragmatic approach of combining experiments with imperfect potential energy models allows one to extract the maximum amount of information from sparse experimental data. Particular care should be taken when interpreting bulk experiments that measure averages over a large number of copies of the same molecule. These experiments are valuable in the characterization of dynamical molecules, where heterogeneous structures might be mixed and contribute with different weights to the experimental observation. If properly combined with MD simulations, these experiments can be used to construct a high resolution picture of molecular structure and dynamics [12–15].

In this review we discuss some recent methodological developments related to the application of the maximum entropy principle to combine MD simulations with ensemble averages obtained from experiments (see, e.g., Refs. [16,17] for an introduction on this topic). We briefly review the maximum entropy principle and show how it can be cast into a minimization problem. We then discuss the equivalent formulation based on averaging between multiple simultaneous simulations. Special explanations are dedicated to the incorporation of experimental errors in the maximum entropy principle and to the protocols that can be used to enforce the experimental constraints. Simple model systems are used to illustrate the typical difficulties encountered in real applications. Source code for the model systems is available at https://github.com/bussilab/review-maxent.

2. The Maximum Entropy Principle

The maximum entropy principle dates back to 1957 when Jaynes [18,19] proposed it as a link between thermodynamic entropy and information-theory entropy. Previously, the definition of entropy was considered as an arrival point in the construction of new theories, and only used as a validation against laws of thermodynamics [18]. In Jaynes formulation, maximum entropy was for the first time seen as the starting point to be used in building new theories. In particular, distributions that maximize the entropy subject to some physical constraints were postulated to be useful in order to make inference on the system under study. In its original formulation, the maximum entropy principle states that, given a system described by a number of states, the best probability distribution for these states compatible with a set of observed data is the one maximizing the associated Shannon's entropy. This principle has been later extended to a maximum relative entropy principle [20] which has the advantage of being invariant with respect to changes of coordinates and coarse-graining [21] and has been shown to play an important role in multiscale problems [22]. The entropy is computed here relative to a given prior distribution $P_0(\boldsymbol{q})$ and, for a system described by a set of continuous variables $\boldsymbol{q}$, is defined as

$$S[P||P_0] = -\int d\boldsymbol{q}\, P(\boldsymbol{q}) \ln \frac{P(\boldsymbol{q})}{P_0(\boldsymbol{q})} \,. \tag{1}$$

This quantity should be maximized subject to constraints in order to be compatible with observations:

$$\begin{cases} P_{ME}(\boldsymbol{q}) = \underset{P(\boldsymbol{q})}{\arg\max}\, S[P||P_0] \\ \int d\boldsymbol{q}\, s_i(\boldsymbol{q}) P(\boldsymbol{q}) = \langle s_i(\boldsymbol{q}) \rangle = s_i^{exp}; \quad i = 1, \dots, M \\ \int d\boldsymbol{q}\, P(\boldsymbol{q}) = 1 \end{cases} \tag{2}$$

Here M experimental observations constrain the ensemble average of M observables $s_i(\boldsymbol{q})$ computed over the distribution $P(\boldsymbol{q})$ to be equal to s_i^{exp}, and an additional constraint ensures that the distribution $P(\boldsymbol{q})$ is normalized. $P_0(\boldsymbol{q})$ encodes the knowledge available before the experimental measurement and is thus called *prior* probability distribution. $P_{ME}(\boldsymbol{q})$ instead represents the best estimate for the probability distribution after the experimental constraints have been enforced and is thus called *posterior* probability distribution. Here, the subscript *ME* denotes the fact that this is the distribution that maximizes the entropy.

Since the relative entropy $S[P||P_0]$ is the negative of the Kullback-Leibler divergence $D_{KL}[P||P_0]$ [23], the procedure described above can be interpreted as a search for the posterior

distribution that is as close as possible to the prior knowledge and agrees with the given experimental observations. In terms of information theory, the Kullback-Leibler divergence measures how much information is gained when prior knowledge $P_0(q)$ is replaced with $P(q)$.

The solution of the maximization problem in Equation (2) can be obtained using the method of Lagrangian multipliers, namely searching for the stationary points of the Lagrange function

$$\mathcal{L} = S[P||P_0] - \sum_{i=1}^{M} \lambda_i \left(\int dq \, s_i(q)P(q) - s_i^{exp} \right) - \mu \left(\int dq \, P(q) - 1 \right) , \tag{3}$$

where λ_i and μ are suitable Lagrangian multipliers. The functional derivative of $\mathcal{L}$ with respect to $P(q)$ is

$$\frac{\delta \mathcal{L}}{\delta P(q)} = - \ln \frac{P(q)}{P_0(q)} - 1 - \sum_{i=1}^{M} \lambda_i s_i(q) - \mu . \tag{4}$$

By setting $\frac{\delta \mathcal{L}}{\delta P(q)} = 0$ and neglecting the normalization factor, the posterior reads

$$P_{ME}(q) \propto e^{-\sum_{i=1}^{M} \lambda_i s_i(q)} P_0(q) . \tag{5}$$

Here the value of the Lagrangian multipliers λ_i should be found by enforcing the agreement with the experimental data. In the following, in order to have a more compact notation, we will drop the subscript from the Lagrangian multipliers and write them as a vector whenever possible. Equation (5) could thus be equivalently written as

$$P_{ME}(q) \propto e^{-\lambda \cdot s(q)} P_0(q) . \tag{6}$$

Notice that the vectors s and λ have dimensionality M, whereas the vector q has dimensionality equal to the number of degrees of freedom of the analyzed system.

In short, the maximum relative entropy principle gives a recipe to obtain the posterior distribution that is as close as possible to the prior distribution and agrees with some experimental observation. In the following, we will drop the word "relative" and we will refer to this principle as the maximum entropy principle.

2.1. Combining Maximum Entropy Principle and Molecular Dynamics

When combining the maximum entropy principle with MD simulations the prior knowledge is represented by the probability distribution resulting from the employed potential energy, that is typically an empirical force field in classical MD. In particular, given a potential energy $V_0(q)$, the associated probability distribution $P_0(q)$ at thermal equilibrium is the Boltzmann distribution $P_0(q) \propto e^{-\beta V_0(q)}$, where $\beta = \frac{1}{k_B T}$, T is the system temperature, and k_B is the Boltzmann constant. According to Equation (5), the posterior will be $P_{ME}(q) \propto e^{-\lambda \cdot s(q)} e^{-\beta V_0(q)}$. The posterior distribution can thus be generated by a modified potential energy in the form

$$V_{ME}(q) = V_0(q) + k_B T \lambda \cdot s(q) . \tag{7}$$

In other words, the effect of the constraint on the ensemble average is that of adding a term to the energy that is linear in the function $s(q)$ with prefactors chosen in order to enforce the correct averages. Such a linear term should be compared with constrained MD simulations, where the value of some function of the coordinates is fixed at every step (e.g., using the SHAKE algorithm [24]), or harmonic restraints, where a quadratic function of the observable is added to the potential energy function. Notice that the words constraint and restraint are usually employed when a quantity is exactly or softly enforced, respectively. Strictly speaking, in the maximum entropy context, ensemble averages $\langle s(q) \rangle$ are constrained whereas the corresponding functions $s(q)$ are (linearly) restrained.

If one considers the free energy as a function of the experimental observables (also known as potential of mean force), which is defined as

$$F_0(s') = -k_B T \ln \int dq \, \delta(s(q) - s') P_0(q) \,, \tag{8}$$

the effect of the corrective potential in Equation (7) is just to tilt the free-energy landscape

$$F_{ME}(s) = F_0(s) + k_B T \boldsymbol{\lambda} \cdot s + C \,, \tag{9}$$

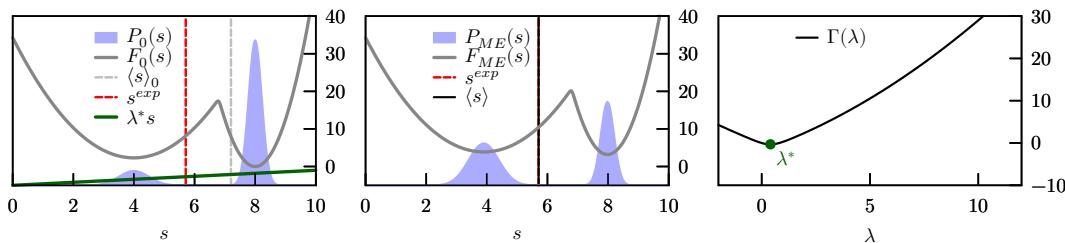where C is an arbitrary constant. A schematic representation of this tilting is reported in Figure 1.

Figure 1. The effect of a linear correcting potential on a given reference potential. $P_0(s)$ is the marginal probability distribution of some observable $s(q)$ according to the reference potential $V_0(q)$ and $F_0(s)$ is the corresponding free-energy profile (**left panel**). Energy scale is reported in the vertical axis and is given in units of $k_B T$. Probability scales are not reported. Vertical lines represent the average value of the observable s in the prior ($\langle s \rangle_0$) and in the experiment (s^{exp}). A correcting potential linear in s (green line) shifts the relative depths of the two free-energy minima, leading to a new free energy profile $F_{ME}(s) = F_0(s) + k_B T \lambda^* s$ that corresponds to a probability distribution $P_{ME}(s)$ (**central panel**). Choosing λ^* equal to the value that minimizes $\Gamma(\lambda)$ (**right panel**) leads to an average $\langle s \rangle = s^{exp}$.

Any experimental data that is the result of an ensemble measurement can be used as a constraint. Typical examples for biomolecular systems are nuclear-magnetic-resonance (NMR) experiments such as measures of chemical shifts [25], scalar couplings [26], or residual dipolar couplings [27], and other techniques such as small-angle X-ray scattering (SAXS) [28], double electron-electron resonance (DEER) [29], and Förster resonance energy transfer [30]. The only requirement is the availability of a so-called *forward model* for such experiments. The forward model is a function mapping the atomic coordinates of the system to the measured quantity and thus allows the experimental data to be back-calculated from the simulated structures. For instance, in the case of 3J scalar couplings, the forward model is given by the so-called Karplus relations [26], that are trigonometric functions of the dihedral angles. It must be noted that the formulas used in standard forward models are often parameterized empirically, and one should take into account errors in these parameters on par with experimental errors (see Section 3). Without entering in the complexity of the methods mentioned above, we will only consider cases where experimental data can be trusted to be ensemble averages.

In short, the maximum entropy principle can be used to derive corrective potentials for MD simulations that constrain the value of some ensemble average. The choice to generate an ensemble that is as close as possible to the prior knowledge implies that the correcting potential has a specific functional form, namely that it is linear in the observables that have been measured.

2.2. A Minimization Problem

In order to chose the values of λ that satisfy Equation (2), it is possible to recast the problem into a minimization problem. In particular, consider the function [16,31]

$$\Gamma(\boldsymbol{\lambda}) = \ln \left[\int dq \, P_0(q) e^{-\boldsymbol{\lambda} \cdot s(q)} \right] + \boldsymbol{\lambda} \cdot s^{exp} \,. \tag{10}$$

Notice that the first term is the logarithm of the ratio between the two partition functions associated to the potential energy functions $V(\boldsymbol{q})$ and $V_0(\boldsymbol{q})$, that is proportional to the free-energy difference between these two potentials. The gradient of $\Gamma(\boldsymbol{\lambda})$ is

$$\frac{\partial \Gamma}{\partial \lambda_i} = s_i^{exp} - \frac{\int d\boldsymbol{q} \, P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})} s_i(\boldsymbol{q})}{\int d\boldsymbol{q} \, P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})}} = s_i^{exp} - \langle s_i(\boldsymbol{q}) \rangle \tag{11}$$

and is thus equal to zero when the average in the posterior distribution is identical to the enforced experimental value. This means that the constraints in Equation (2) can be enforced by searching for a stationary point $\boldsymbol{\lambda}^*$ of $\Gamma(\boldsymbol{\lambda})$ (see Figure 1). The Hessian of $\Gamma(\boldsymbol{\lambda})$ is

$$\frac{\partial \Gamma}{\partial \lambda_i \partial \lambda_j} = \langle s_i(\boldsymbol{q}) s_j(\boldsymbol{q}) \rangle - \langle s_i(\boldsymbol{q}) \rangle \langle s_j(\boldsymbol{q}) \rangle \tag{12}$$

and is thus equal to the covariance matrix of the forward models in the posterior distribution. Unless the enforced observables are dependent on each other, the Hessian will be positive definite [16]. The solution of Equation (2) will thus correspond to a minimum of $\Gamma(\boldsymbol{\lambda})$ that can be searched for instance by a steepest descent procedure. However there are cases where such minimum might not exist. In particular, one should pay attention to the following cases:

- When data are incompatible with the prior distribution.
- When data are mutually incompatible. As an extreme case, one can imagine two different experiments that measure the same observable and report different values.

In both cases $\Gamma(\boldsymbol{\lambda})$ will have no stationary point. Clearly, there is a continuum of possible intermediate situations where data are almost incompatible. In Section 4 we will see what happens when the maximum entropy principle is applied to model systems designed in order to highlight these difficult situations.

2.3. Connection with Maximum Likelihood Principle

The function $\Gamma(\boldsymbol{\lambda})$ allows to easily highlight a connection between maximum entropy and maximum likelihood principles. Given an arbitrary set of N_s molecular structures $\boldsymbol{q}_t$ chosen such that $\frac{1}{N_s} \sum_{t=1}^{N_s} \boldsymbol{s}(\boldsymbol{q}_t) = \boldsymbol{s}^{exp}$, it is possible to rewrite $e^{-N_s \Gamma(\boldsymbol{\lambda})}$ as

$$e^{-N_s \Gamma(\boldsymbol{\lambda})} = \frac{e^{-N_s \boldsymbol{\lambda} \cdot \boldsymbol{s}^{exp}}}{\left[\int d\boldsymbol{q} P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})} \right]^{N_s}} = \frac{e^{-\boldsymbol{\lambda} \cdot \sum_t \boldsymbol{s}(\boldsymbol{q}_t)}}{\left[\int d\boldsymbol{q} P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})} \right]^{N_s}} = \prod_{t=1}^{N_s} \frac{e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q}_t)}}{\int d\boldsymbol{q} P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})}} = \prod_{t=1}^{N_s} \frac{P(\boldsymbol{q}_t)}{P_0(\boldsymbol{q}_t)} \tag{13}$$

The last term is the ratio between the probability of drawing the structures $\boldsymbol{q}_t$ from the posterior distribution and that of drawing the same structures from the prior distribution. Since the minimum of $\Gamma(\boldsymbol{\lambda})$ corresponds to the maximum of $e^{-N_s \Gamma(\boldsymbol{\lambda})}$, the distribution that maximizes the entropy under experimental constraints is identical to the one that, among an exponential family of distributions, maximizes the likelihood of a set of structures with average value of the observables $\boldsymbol{s}$ equal to the experimental value [32,33]. This equivalence can be considered as an added justification for the maximum entropy principle [32]: if the notion of selecting a posterior $P(\boldsymbol{q})$ that maximizes the entropy is not compelling enough, one can consider that this same posterior is, among the distributions with the exponential form of Equation (5), the one that maximizes the likelihood of being compatible with the experimental sample.

Equation (13) can also be rearranged to $\Gamma(\boldsymbol{\lambda}) = -\frac{1}{N_s} \sum_{t=1}^{N_s} \ln P(\boldsymbol{q}_t) + \frac{1}{N_s} \sum_{t=1}^{N_s} \ln P_0(\boldsymbol{q}_t)$ and, after proper manipulation, it can be shown that

$$\Gamma(\boldsymbol{\lambda}) = D_{KL}[P^{exp} || P] - D_{KL}[P^{exp} || P_0] \tag{14}$$

where P^{exp} is an arbitrary distribution with averages equal to the experimental ones. Thus, minimizing $\Gamma(\boldsymbol{\lambda})$ is equivalent to choosing the distribution with the exponential form of Equation (5) that is as close

as possible to the experimental one. Since at its minimum, by construction, $\Gamma(\lambda^*) \leq \Gamma(0)$, it follows that $D_{KL}[P^{exp}||P_{ME}] \leq D_{KL}[P^{exp}||P_0]$. In other words, the maximum entropy restraint is guaranteed to make the posterior distribution closer to the experimental one than the prior distribution [34].

2.4. Enforcing Distributions

We so far considered the possibility of enforcing ensemble averages. However, one might be interested in enforcing the full distribution of an observable. This can be done by noticing that the marginal probability distribution $\rho(s)$ of a quantity s can be computed as the expectation value of a Dirac-delta function:

$$\rho(s') = \langle \delta(s(q) - s') \rangle \, . \tag{15}$$

An example of experimental technique that can report distance distributions is the already mentioned DEER [29]. If the form of $\rho(s)$ has been measured experimentally, the maximum entropy principle can be used to enforce it in a MD simulation. Notice that this corresponds to constraining an infinite number of data points (that is, the occupation of each bin in the observable s). In this case, λ will be a function of s and Equation (5) will take the following form

$$P_{ME}(q) \propto e^{-\lambda(s(q))} P_0(q) \, . \tag{16}$$

Thus, the correction to the potential should be a function of the observable s chosen in order to enforce the experimental distribution $\rho^{exp}(s)$. Different approaches can be used to construct the function $\lambda(s)$ with such property. For instance, one might take advantage of iterative Boltzmann inversion procedures originally developed to derive coarse-grained models from atomistic simulations [35]. As an alternative, one might use a time-dependent adaptive potential. In target metadynamics [36,37] such potential is constructed as a sum of Gaussians centered on the previously visited values of s. It can be shown that by properly choosing the prefactors of those Gaussians an arbitrary target distribution can be enforced.

Alternatively, it is possible to directly minimize the function $\Gamma(\lambda)$ as mentioned in Section 2.2. In this context, Γ would be a functional of $\lambda(s)$ with the form

$$\Gamma[\lambda] = \ln \int dq \, e^{-\lambda(s(q))} P_0(q) + \int ds \lambda(s) \rho^{exp}(s) \, . \tag{17}$$

Interestingly, this functional is identical to the one introduced in the variationally enhanced sampling (VES) method of Ref. [38]. In its original formulation, VES was used to enforce a flat distribution in order to sample rare events. However, the method can also be used to enforce an arbitrary a priori chosen distribution [39,40]. The analogy with maximum entropy methods, together with the relationship in Equation (14), was already noticed in Ref. [40] and is interesting for a twofold reason: (a) It provides a maximum-entropy interpretation of VES, and (b) the numerical techniques used for VES might be used to enforce experimental averages in a maximum-entropy context. We will further comment about this second point in Section 5.4.

2.5. Equivalence to the Replica Approach

A well established method to enforce ensemble averages in molecular simulations is represented by restrained ensembles [41–43]. The rationale behind this method is to mimic an ensemble of structures by simulating in parallel N_{rep} multiple identical copies (replicas) of the system each of which having its own atomic coordinates. The agreement with the M experimental data is then enforced by adding a harmonic restraint for each observable, centered on the experimental reference and acting on the average over all the simulated replicas. This results in a restraining potential with the following form:

$$V_{RE}\left(q_1, q_2, \ldots, q_{N_{rep}}\right) = \sum_{i=1}^{N_{rep}} V_0\left(q_i\right) + \frac{k}{2} \sum_{j=1}^{M} \left(\frac{1}{N_{rep}} \sum_{i=1}^{N_{rep}} s_j(q_i) - s_j^{exp}\right)^2, \tag{18}$$

where k is a suitably chosen force constant. It has been shown [16,44,45] that this method produces the same ensemble as the maximum entropy approach in the limit of large number of replicas $\left(N_{rep} \to \infty\right)$. Indeed, the potential in Equation (18) results in the same force $-\frac{k}{N_{rep}} \left(\frac{1}{N_{rep}} \sum_{i=1}^{N_{rep}} s_j(q_i) - s_j^{exp}\right)$ applied to the observable $s_j(q)$ in each replica. As the number of replicas grows, the fluctuations of the average decrease and the applied force becomes constant in time, so that the explored distribution will have the same form as Equation (5) with $\lambda = \frac{k}{N_{rep}k_BT} \left(\frac{1}{N_{rep}} \sum_{i=1}^{N_{rep}} s(q_i) - s^{exp}\right)$. If k is chosen large enough, the average between the replicas will be forced to be equal to the experimental one. It is possible to show that, in order to enforce the desired average, k should grow faster than N_{rep} [45]. In practical implementations, k should be finite in order to avoid infinite forces. A direct calculation of the entropy-loss due to the choice of a finite N_{rep} has been proposed to be an useful tool in the search for the correct number of replicas [46]. An approach based on a posteriori reweighting (Section 5.1) of replica-based simulations and named Bayesian inference of ensembles has been also proposed in order to eliminate the effect of choosing a finite number of replicas [47].

3. Modelling Experimental Errors

The maximum entropy method can be modified in order to account for uncertainties in experimental data. This step is fundamental in order to reduce over-fitting. In this section we will briefly consider how the error can be modeled according to Ref. [48]. Here errors are modeled modifying the experimental constraints introduced in Equation (2) by introducing an auxiliary variables ϵ_i for each data point representing the discrepancy or residual between the experimental and the simulated value. The new constraints are hence defined as follows:

$$\langle(s(q) + \epsilon)\rangle = s^{exp}. \tag{19}$$

The auxiliary variable ϵ is a vector with dimensionality equal to the number of constraints and models all the possible sources of error, including inaccuracies of the forward models (Section 2.1) as well as experimental uncertainties. Errors can be modeled by choosing a proper prior distribution function for the variable ϵ. A common choice is represented by a Gaussian prior with a fixed standard deviation σ_i for the ith observable

$$P_0\left(\epsilon\right) \propto \prod_{i=1}^{M} \exp\left(-\frac{\epsilon_i^2}{2\sigma_i^2}\right). \tag{20}$$

The value of σ_i corresponds to the level of confidence in the ith data point, where $\sigma_i = \infty$ implies to completely discard the data in the optimization process while $\sigma_i = 0$ means having complete confidence in the data, that will be fitted as best as possible. Notice that, for additive errors, q and ϵ are independent variables and Equation (19) can be written as:

$$\langle s(q)\rangle = s^{exp} - \langle\epsilon\rangle \tag{21}$$

where $\langle\epsilon\rangle$ is computed in the posterior distribution $P(\epsilon) \propto P_0(\epsilon)e^{-\lambda\cdot\epsilon}$. Incorporating the experimental error in the maximum entropy approach is then as easy as enforcing a different experimental value, corresponding to the one in Equation (21). Notice that the value of $\langle\epsilon\rangle$ only depends on its prior distribution $P_0(\epsilon)$ and on λ. For a Gaussian prior with standard deviation σ_i Equation (20) we have:

$$\langle\epsilon_i\rangle = -\lambda_i\sigma_i^2. \tag{22}$$

Thus, as λ grows in magnitude, a larger discrepancy between simulation and experiment will be accepted. In addition, it can be seen that applying the same constraint *twice* is exactly equivalent to applying a constraint with a σ_i^2 reduced by a factor two. This is consistent with the fact that the confidence in the repeated data point is increased.

Other priors are also possible in order to better account for outliers and to deal with cases where the standard deviation of the residual is not known a priori. One might consider the variance of the *i*th residual $\sigma_{0,i}^2$ as a variable sampled from a given prior distribution $P_0(\sigma_{0,i}^2)$:

$$P_0\left(\boldsymbol{\epsilon}\right) = \prod_{i=1}^{M} \int_0^{\infty} d\sigma_{0,i}^2 P_0(\sigma_{0,i}^2) \frac{1}{\sqrt{2\pi\sigma_{0,i}^2}} \exp\left(-\frac{\epsilon_i^2}{2\sigma_{0,i}^2}\right) . \tag{23}$$

A flexible functional form for $P_0(\sigma_{0,i}^2)$ can be obtained using the following Gamma distribution

$$P_0(\sigma_{0,i}^2) \propto (\sigma_{0,i}^2)^{\kappa-1} \exp\left(-\frac{\kappa\sigma_{0,i}^2}{\sigma_i^2}\right) . \tag{24}$$

In the above equation σ_i^2 is the mean parameter of the Gamma function and must be interpreted as the typical expected variance of the error on the *i*th data point. κ, which must satisfy $\kappa > 0$, is the shape parameter of the Gamma distribution and expresses how much the distribution is peaked around σ_i^2. In practice, it controls how much the optimization is tolerant to large discrepancies between the experimental data and the enforced average. Notice that in Ref. [48] a different convention was used with a parameter $\alpha = 2\kappa - 1$. By setting $\kappa = \infty$ a Gaussian prior on $\boldsymbol{\epsilon}$ will be recovered. Smaller values of κ will lead to a prior distribution on $\boldsymbol{\epsilon}$ with "fatter" tails and thus able to accommodate larger differences between experiment and simulation. For instance, the case $\kappa = 1$ leads to a Laplace prior $P_0(\boldsymbol{\epsilon}) \propto \prod_i \exp\left(-\frac{\sqrt{2}|\epsilon|}{\sigma_i}\right)$. After proper manipulation, the resulting expectation value $\langle \boldsymbol{\epsilon} \rangle$ can be shown to be

$$\langle \epsilon_i \rangle = -\frac{\lambda_i \sigma_i^2}{1 - \frac{\lambda_i^2 \sigma_i^2}{2\kappa}} . \tag{25}$$

In this case, it can be seen that applying the same constraint twice is exactly equivalent to applying a constraint with a σ_i^2 reduced by a factor two and a κ multiplied by a factor two.

In terms of the minimization problem of Section 2.2, modeling experimental errors as discussed here is equivalent to adding a contribution Γ_{err} to Equation (10):

$$\Gamma(\boldsymbol{\lambda}) = \ln \int d\boldsymbol{q}\, P_0(\boldsymbol{q}) e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})} + \boldsymbol{\lambda} \cdot \boldsymbol{s}^{exp} + \Gamma_{err}(\boldsymbol{\lambda}) . \tag{26}$$

For a Gaussian noise with preassigned variance (Equation (20)) the additional term is

$$\Gamma_{err}(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^{M} \lambda_i^2 \sigma_i^2 . \tag{27}$$

For a prior on the error in the form of Equations (23) and (24) one obtains

$$\Gamma_{err}(\boldsymbol{\lambda}) = -\kappa \sum_{i=1}^{M} \ln\left(1 - \frac{\lambda_i^2 \sigma_i^2}{2\kappa}\right) . \tag{28}$$

In the limit of large κ, Equation (28) is equivalent to Equation (27). If the data points are expected to all have the same error σ_0, unknown but with a typical value σ (see Ref. [48]), Equation (28) should be modified to $\Gamma_{err}(\boldsymbol{\lambda}) = -\kappa \ln\left(1 - \frac{|\boldsymbol{\lambda}|^2 \sigma^2}{2\kappa}\right)$.

Equation (28) shows that by construction the Lagrangian multiplier λ_i will be limited in the range $\left(-\frac{\sqrt{2\kappa}}{\sigma_i}, +\frac{\sqrt{2\kappa}}{\sigma_i} \right)$. The effect of using a prior with $\kappa < \infty$ is thus that of restricting the range of allowed λ in order to avoid too large modifications of the prior distribution. In practice, values of λ chosen outside these boundaries would lead to a posterior distribution $P(\epsilon) \propto P_0(\epsilon) e^{-\lambda \cdot \epsilon}$ that cannot be normalized.

Except for trivial cases (e.g., for Gaussian noise with $\sigma = 0$), the contribution originating from error modeling has positive definite Hessian and as such it makes $\Gamma(\lambda)$ a strongly convex function. Thus, a suitable error treatment can make the minimization process numerically easier.

It is worth mentioning that a very similar formalism can be used to include not only errors but more generally any quantity that influences the experimental measurement but cannot be directly obtained from the simulated structures. For instance, in the case of residual dipolar couplings [27], the orientation of the considered molecule with the respect to the external field is often unknown. The orientation of the field can then be used as an additional vector variable to be sampled with a Monte Carlo procedure, and suitable Lagrangian multipliers can be obtained in order to enforce the agreement with experiments [49]. Notice that in this case the orientation contributes to the ensemble average in a non-additive manner so that Equation (21) cannot be used. Interestingly, thanks to the equivalence between multi-replica simulations and maximum entropy restraints (Section 2.5), equivalent results can be obtained using the tensor-free method of Ref. [50].

Finally, we note that several works introduced error treatment using a Bayesian framework [47,51–53]. Interestingly, Bayesian ensemble refinement [47] introduces an additional parameter (θ) that takes into account the confidence in the prior distribution. In case of Gaussian error, this parameter enters as a global scaling factor in the errors σ_i for each data point. Thus, the errors σ_i discussed above can be used to modulate both our confidence in experimental data and our confidence in the original force field. The equivalence between the error treatment of Ref. [47] and the one reported here is further discussed in Ref. [48], in particular for what concerns non-Gaussian error priors.

4. Exact Results on Model Systems

In this section we illustrate the effects of adding restraints using the maximum entropy principle on simple model systems. In order to do so we first derive some simple relationship valid when the prior has a particular functional form, namely a sum of N_G Gaussians with center s_α and covariance matrix A_α, where $\alpha = 1, \ldots, N_G$:

$$P_0(s) = \sum_{\alpha=1}^{N_G} \frac{w_\alpha}{\sqrt{2\pi \det A_\alpha}} e^{-\frac{(s-s_\alpha)A_\alpha^{-1}(s-s_\alpha)}{2}} . \tag{29}$$

The coefficients w_α provide the weights of each Gaussian and are normalized ($\sum_\alpha w_\alpha = 1$). We assume here that the restraints are applied on the variable s. For a general system, one should first perform a dimensional reduction in order to obtain the marginal prior probability $P_0(s)$. By constraining the ensemble averages of the variable s to an experimental value s^{exp} the posterior becomes:

$$P_{ME}(s) = \frac{e^{-\lambda \cdot s}}{Z(\lambda)} \sum_\alpha \frac{w_\alpha}{\sqrt{2\pi \det A_\alpha}} e^{-\frac{(s-s_\alpha)A_\alpha^{-1}(s-s_\alpha)}{2}} . \tag{30}$$

With proper algebra it is possible to compute explicitly the normalization factor $Z(\lambda) = \sum_\alpha w_\alpha e^{\frac{\lambda A_\alpha \lambda}{2} - \lambda \cdot s_\alpha}$. The function $\Gamma(\lambda)$ to be minimized is thus equal to:

$$\Gamma(\lambda) = \ln\left(\sum_\alpha w_\alpha e^{\frac{\lambda A_\alpha \lambda}{2} - \lambda \cdot s_\alpha} \right) + \lambda \cdot s^{exp} + \Gamma_{err}(\lambda) \tag{31}$$

and the average value of s in the posterior is

$$\langle s \rangle = \frac{\sum_\alpha w_\alpha e^{\frac{\lambda A_\alpha \lambda}{2} - \lambda \cdot s_\alpha} (s_\alpha - A_\alpha \lambda)}{\sum_\alpha w_\alpha e^{\frac{\lambda A_\alpha \lambda}{2} - \lambda \cdot s_\alpha}} . \tag{32}$$

We could not find a close formula for λ^* given s^{exp} and Γ_{err}. However, the solution can be found numerically with the gradient descent procedure discussed in Section 5 (see Equation (33)).

4.1. Consistency between Prior Distribution and Experimental Data

We consider a one dimensional model with a prior expressed as a sum of two Gaussians, one centered in $s_A = 4$ with standard deviation $\sigma_A = 0.5$ and one centered in $s_B = 8$ with standard deviation $\sigma_B = 0.2$. The weights of the two Gaussians are $w_A = 0.2$ and $w_B = 0.8$, respectively. The prior distribution is thus $P_0(s) \propto \frac{w_A}{\sigma_A} e^{-(s-s_A)^2/2\sigma_A^2} + \frac{w_B}{\sigma_B} e^{-(s-s_B)^2/2\sigma_B^2}$, has an average value $\langle s \rangle_0 = 7.2$, and is represented in Figure 2, left column top panel.
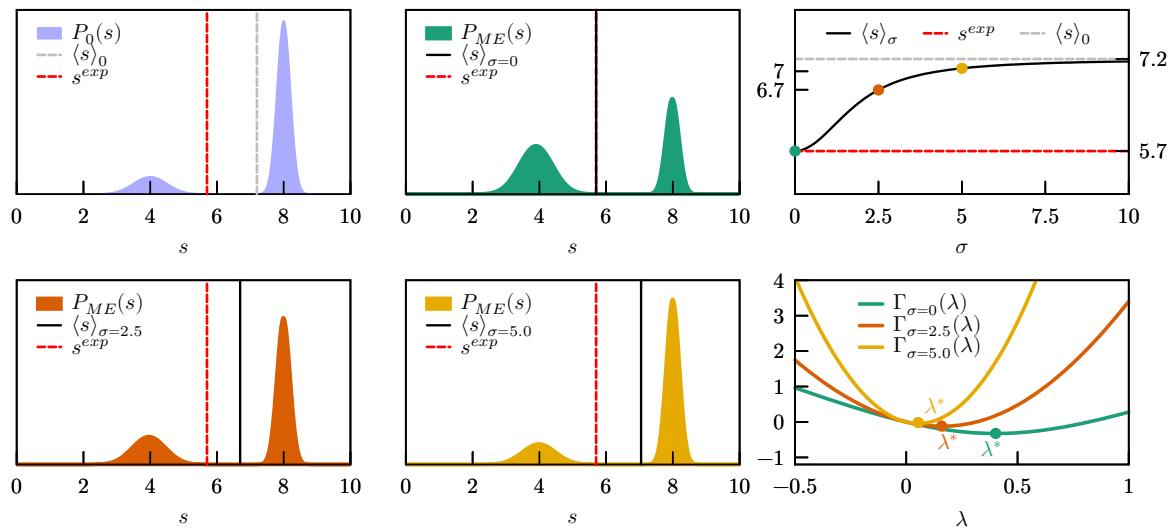

Figure 2. Effect of modeling error with a Gaussian probability distribution with different standard deviations σ on the posterior distribution $P_{ME}(s)$. The experimental value is here set to $s^{exp} = 5.7$, which is compatible with the prior distribution. **Left** and **middle** column: prior $P_0(s)$ and posterior $P_{ME}(s)$ with $\sigma = 0$, 2.5, 5.0. **Right** column: ensemble average $\langle s \rangle$ plotted as a function of σ and $\Gamma(\lambda)$ plotted for different values of σ. λ^* denotes that value of λ that minimizes $\Gamma(\lambda)$.

We first enforce a value $s^{exp} = 5.7$, which is compatible with the prior probability. If we are absolutely sure about our experimental value and set $\sigma = 0$, the λ^* which minimizes $\Gamma(\lambda)$ is $\lambda^* \approx 0.4$ (Figure 2 right column, bottom panel). In case values of $\sigma \neq 0$ are used, the $\Gamma(\lambda)$ function becomes more convex and the optimal value λ^* is decreased. As a result, the average s in the posterior distribution is approaching its value in the prior. The evolution of the ensemble average $\langle s \rangle_\sigma$ with σ values between zero and ten, with respect to the initial $\langle s \rangle_0$ and the experimental s^{exp}, is shown in Figure 2, right column top panel. In all these cases the posterior distributions remain bimodal and the main effect of the restraint is to change the relative population of the two peaks (Figure 2, left and middle columns).

We then enforce an average value $s^{exp} = 2$, which is far outside the original probability distribution (see Figure 3). If we are absolutely sure about our experimental value and set $\sigma = 0$, the λ^* which minimizes $\Gamma(\lambda)$ is very large, $\lambda^* \approx 8$ (Figure 3 right column, bottom panel). Assuming zero error on the experimental value is equivalent to having poor confidence in the probability distribution sampled by the force field, and leads in fact to a $P_{ME}(s)$ completely different from $P_0(s)$. The two peaks in $P_0(s)$ are replaced by a single peak centered around the experimental value, which is exactly met by the ensemble average ($\langle s \rangle_{\sigma=0} = s^{exp} = 2$; Figure 3 middle column top panel). Note that this is possible

only because the experimental value is not entirely incompatible with the prior distribution, i.e., it has a small, non-zero probability also in the prior. If the probability had been zero, $\Gamma(\lambda)$ would have had no minimum and no optimal λ^* would have been found. If we have more confidence in the distribution sampled by the force field, assume that there might be an error in our experimental value, and set $\sigma = 2.5$, λ^* is more than one order of magnitude lower ($\lambda^* \approx 0.52$). The two peaks in $P_0(s)$ are only slightly shifted towards lower s, while their relative populations are shifted in favor of the peak centered around 4 (Figure 3, left column bottom panel). According to our estimate of the probability distribution of the error, the ensemble average $\langle s \rangle_{\sigma=2.5} \approx 5.2$ is more probably the true value than the experimentally measured one. In case we have very high confidence in the force field and very low confidence in the experimental value and set $\sigma = 5.0$, the correction becomes very small ($\lambda^* \approx 0.18$) and the new ensemble average $\langle s \rangle_{\sigma=5.0} \approx 6.6$, very close to the initial $\langle s \rangle_0 = 7.2$ (Figure 3, middle column bottom panel). The evolution of the ensemble average $\langle s \rangle_\sigma$ with σ values between zero and ten, with respect to the initial $\langle s \rangle_0$ and the experimental s^{exp}, is shown in Figure 3, right column top panel.
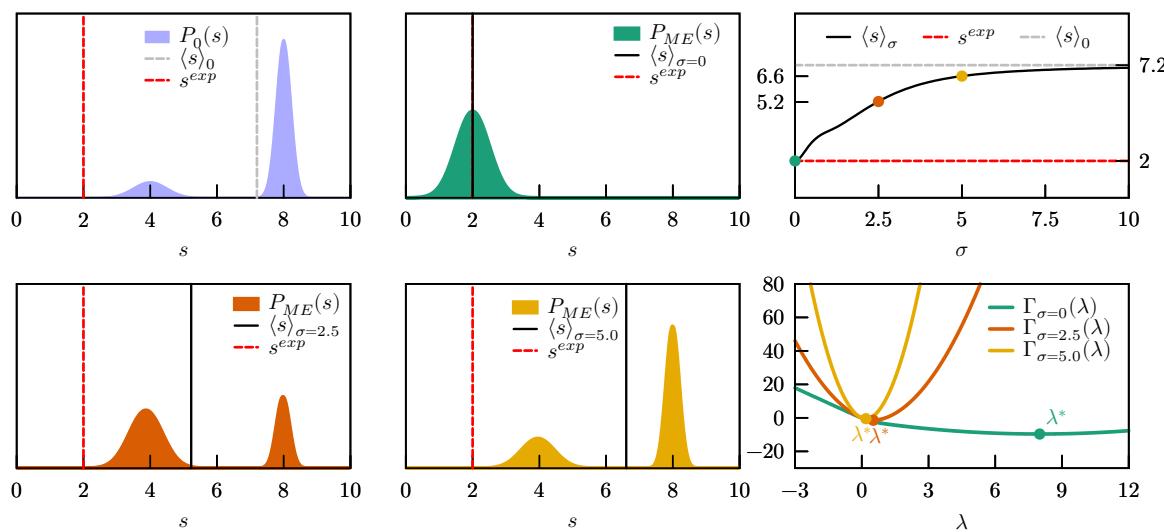

Figure 3. Same as Figure 2, but the experimental value is here set to $s^{exp} = 2$, which is almost incompatible with the prior distribution.

In conclusion, when data that are not consistent with the prior distribution are enforced, the posterior distribution could be severely distorted. Clearly, this could happen either because the prior is completely wrong or because the experimental values are affected by errors. By including a suitable error model in the maximum entropy procedure it is possible to easily interpolate between the two extremes in which we completely trust the force field or the experimental data.

4.2. Consistency between Data Points

We then consider a two dimensional model with a prior expressed as a sum of two Gaussians centered in $\boldsymbol{s}_A = (0,0)$ and $\boldsymbol{s}_B = (3,3)$ with identical standard deviations $\sigma_A = \sigma_B = 0.2$ and weights $w_A = w_B = 0.5$. The prior distribution is represented in Figure 4.

This model is particularly instructive since, by construction, the two components of $\boldsymbol{s}$ are highly correlated and is hence possible to see what happens when inconsistent data are enforced. To this aim we study the two scenarios (i.e., consistent and inconsistent data) using different error models (no error model, Gaussian prior with $\sigma = 1$, and Laplace prior with $\sigma = 1$), for a total of six combinations. In the *consistent* case we enforce $\boldsymbol{s}^{exp} = (1,1)$, whereas in the *inconsistent* one we enforce $\boldsymbol{s}^{exp} = (1,0)$. Figure 4 reports the posterior distributions obtained in all these cases.

When consistent data are enforced the posterior distribution is very similar to the prior distribution, the only difference being a modulation in the weights of the two peaks. The optimal

value λ^*, marked with a $\star$ in Figure 4, does not depend significantly on the adopted error model. The main difference between including or not including error models can be seen in the form of the $\Gamma(\lambda)$ function. When errors are not included, $\Gamma(\lambda)$ is almost flat in a given direction, indicating that one of the eigenvalues of its Hessian is very small. On the contrary, when error modeling is included, the $\Gamma(\lambda)$ function becomes clearly convex in all directions. In practical applications, the numerical minimization of $\Gamma(\lambda)$ would be more efficient.
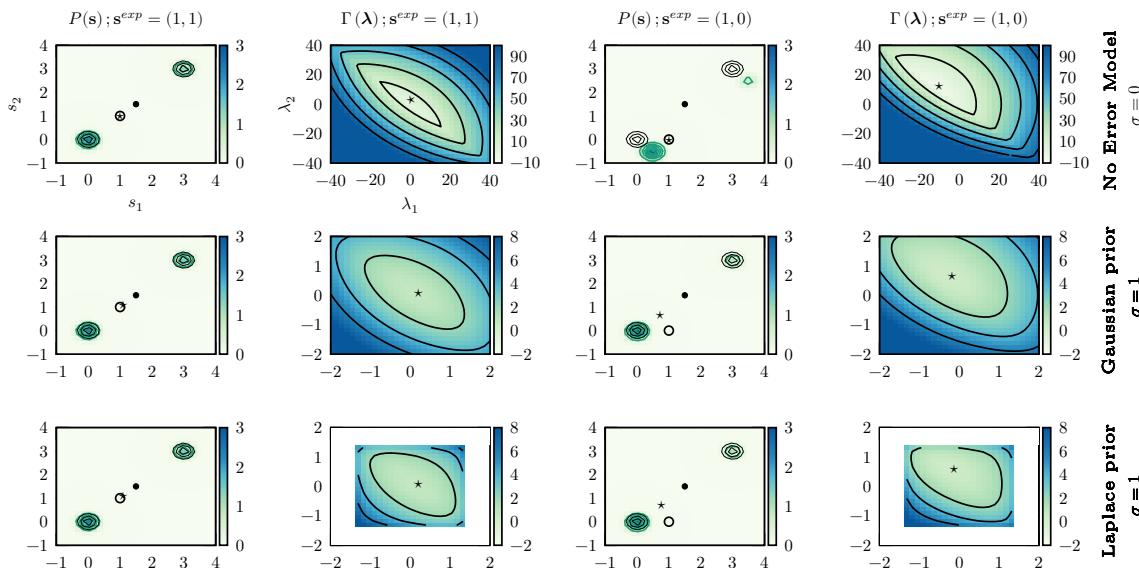

Figure 4. Effect of different prior distributions for the error model in a two-dimensional system. In the first (last) two columns, compatible (incompatible) data are enforced. In the first and the third column, prior distributions are represented as black contour lines and posterior distributions are shown in color scale. A black dot and a $\star$ are used to indicate the average values of s in the prior and posterior distributions respectively, while an empty circle is used to indicate the target s^{exp}. In the second and the fourth column, the function $\Gamma(\lambda)$ is shown, and its minimum λ^* is indicated with a $\star$. The first row reports results where errors are not modeled, whereas the second and the third row report results obtained using Gaussian and Laplace prior for the error model respectively. Notice that a different scale is used to represent $\Gamma(\lambda)$ in the first row. For the Laplace prior, the region of λ where $\Gamma(\lambda)$ is undefined is marked as white.

When enforcing inconsistent data without taking into account experimental error, the behavior is significantly different. Indeed, the only manner to enforce data where the value of the two components of s are different is to significantly displace the two peaks. On the contrary, the distortion is significantly alleviated when taking into account experimental errors. Obviously, in this case the experimental value is not exactly enforced and, with both Gaussian and Laplace prior, we obtain $\langle s \rangle \approx (0.7, 0.7)$.

By observing $\Gamma(\lambda)$ it can be seen that the main effect of using a Laplace prior instead of a Gaussian prior for the error is that the range of suitable values for λ is limited. This allows one to decrease the effect of particularly wrong data points on the posterior distribution.

In conclusion, when data that are not consistent among themselves are enforced, the posterior distribution could be severely distorted. Inconsistency between data could either be explicit (as in the case where constraints with different reference values are enforced on the same observable) or more subtle. In the reported example, the only way to know that the two components of s should have similar values is to observe their distribution according to the original force field. In the case of complex molecular systems and of observables that depend non-linearly on the atomic coordinates, it is very difficult to detect inconsistencies between data points a priori. By properly modeling experimental error it is possible to greatly alleviate the effect of these inconsistencies on the resulting

posterior. Clearly, if the quality of the prior is very poor, correct data points might artificially appear as inconsistent.

5. Strategies for the Optimization of Lagrangian Multipliers

In order to find the optimal values of the Lagrangian multipliers, one has to minimize the function $\Gamma(\boldsymbol{\lambda})$. The simplest possible strategy is gradient descent (GD), that is an iterative algorithm in which function arguments are adjusted by following the opposite direction of the function gradient. By using the gradient in Equation (11) the value of λ at the iteration $k+1$ can be obtained from the value of λ at the iteration k as:

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \eta_i \frac{\partial \Gamma}{\partial \lambda_i} = \lambda_i^{(k)} - \eta_i \left(s_i^{exp} - \langle s_i(\boldsymbol{q}) \rangle_{\boldsymbol{\lambda}^{(k)}} - \langle \epsilon_i \rangle_{\boldsymbol{\lambda}^{(k)}} \right) , \tag{33}$$

where η represents the step size at each iteration and might be different for different observables. Here we explicitly indicated that the average $\langle s_i(\boldsymbol{q}) \rangle$ should be computed using the Lagrangian multipliers at the kth iteration $\boldsymbol{\lambda}^{(k)}$. In order to compute this average it is in principle necessary to sum over all the possible values of $\boldsymbol{q}$. This is possible for the simple model systems discussed in Section 4, where integrals can be done analytically. However, for a real molecular system, summing over all the conformations would be virtually impossible. Below we discuss some possible alternatives.

Notice that this whole review is centered on constraints in the form of Equation (2). The methods discussed here can be applied to inequality restraints as well, as discussed in Ref. [48].

5.1. Ensemble Reweighting

If a trajectory has been already produced using the prior force field $V_0(\boldsymbol{q})$, samples from this trajectory might be used to compute the function $\Gamma(\boldsymbol{\lambda})$. In particular, the integral in Equation (10) can be replaced by an average over N_s snapshots $\boldsymbol{q}_t$ sampled from $P_0(\boldsymbol{q})$:

$$\tilde{\Gamma}(\boldsymbol{\lambda}) = \ln \left(\frac{1}{N_s} \sum_{t=1}^{N_s} e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q}_t)} \right) + \boldsymbol{\lambda} \cdot \boldsymbol{s}^{exp} + \Gamma_{err}(\boldsymbol{\lambda}) . \tag{34}$$

A gradient descent on $\tilde{\Gamma}$ results in a procedure equivalent to Equation (33) where the ensemble average $\langle \boldsymbol{s}(\boldsymbol{q}) \rangle_{\boldsymbol{\lambda}^{(k)}}$ is computed as a weighted average on the available frames:

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \eta_i \frac{\partial \tilde{\Gamma}}{\partial \lambda_i} = \lambda_i^{(k)} - \eta_i \left(s_i^{exp} - \frac{\sum_{t=1}^{N_s} s_i(\boldsymbol{q}_t) e^{-\boldsymbol{\lambda}^{(k)} \cdot \boldsymbol{s}(\boldsymbol{q}_t)}}{\sum_{t=1}^{N_s} e^{-\boldsymbol{\lambda}^{(k)} \cdot \boldsymbol{s}(\boldsymbol{q}_t)}} - \langle \epsilon_i \rangle_{\boldsymbol{\lambda}^{(k)}} \right) . \tag{35}$$

It is also possible to use conjugated gradient or more advanced minimization methods. Once the multipliers $\boldsymbol{\lambda}^*$ have been found one can compute any other expectation value by just assigning a normalized weight $w_t = e^{-\boldsymbol{\lambda}^* \cdot \boldsymbol{s}(\boldsymbol{q}_t)} / \sum_{t'=1}^{N_s} e^{-\boldsymbol{\lambda}^* \cdot \boldsymbol{s}(\boldsymbol{q}_{t'})}$ to the snapshot $\boldsymbol{q}_t$.

A reweighting procedure related to this one is at the core of the ensemble-reweighting-of-SAXS method [54], that has been used to construct structural ensembles of proteins compatible with SAXS data [54,55]. Similar reweighting procedures were used to enforce average data on a variety of systems [47,51,53,56–60]. These procedures are very practical since they allow incorporating experimental constraints a posteriori without the need to repeat the MD simulation. For instance, in Ref. [59] it was possible to test different combinations of experimental restraints in order to evaluate their consistency. However, reweighting approaches must be used with care since they are effective only when the posterior and the prior distributions are similar enough [61]. In case this is not true, the reweighted ensembles will be dominated by a few snapshots with very high weight, leading to a large statistical error. The effective number of snapshots with a significant weight can be estimated using the Kish's effective sample size [62], defined as $1/\left(\sum_{t=1}^{N_s} w_t^2 \right)$ where w_t are the normalized

weights, or similar measures [63], and is related to the increase of the statistical error of the averages upon reweighting.

5.2. Iterative Simulations

In order to decrease the statistical error, it is convenient to use the modified potential $V(\boldsymbol{q}) = V_0(\boldsymbol{q}) + k_B T \boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})$ to run a new simulation, in an iterative manner. For instance, in the iterative Boltzmann method, pairwise potentials are modified and new simulations are performed until the radial distribution function of the simulated particles does match the desired one [35].

It is also possible to make a full optimization of $\Gamma(\boldsymbol{\lambda})$ using a reweighting procedure like the one illustrated in Section 5.1 at each iteration. One would first perform a simulation using the original force field and, based on samples taken from that simulation, find the optimal $\boldsymbol{\lambda}$ with a gradient descent procedure. Only at that point a new simulation would be required using a modified potential that includes the extra $k_B T \boldsymbol{\lambda} \cdot \boldsymbol{s}(\boldsymbol{q})$ contribution. This whole procedure should be then repeated until the value of $\boldsymbol{\lambda}$ stops changing. This approach was used in Ref. [64] in order to adjust a force field to reproduce ensembles of disordered proteins. The same scheme was later used in a maximum entropy context to enforce average contact maps in the simulation of chromosomes [65,66]. A similar iterative approach was used in Refs. [67,68].

In principle, iterative procedures are supposed to converge to the correct values of $\boldsymbol{\lambda}$. However, this happens only if the simulations used at each iteration are statistically converged. For systems that exhibit multiple metastable states and are thus difficult to sample it might be difficult to tune the length of each iteration so as to obtain good estimators of the $\Gamma(\boldsymbol{\lambda})$ gradients.

5.3. On-the-Fly Optimization with Stochastic Gradient Descent

Instead of trying to converge the calculation of the gradient at each individual iteration and, only at that point, modify the potential in order to run a new simulation, one might try to change the potential on-the-fly so as to force the system to sample the posterior distribution:

$$V(\boldsymbol{q}, t) = V_0(\boldsymbol{q}) + k_B T \boldsymbol{\lambda}(t) \cdot \boldsymbol{s}(\boldsymbol{q}) \,. \tag{36}$$

An earlier approach aimed at enforcing time-averaged constraints was reported in Ref. [69]. However, here we will focus on methods based on the maximum-entropy formalism.

The simplest choice in order to minimize the $\Gamma(\boldsymbol{\lambda})$ function is to use a stochastic gradient descent (SGD) procedure, where an unbiased estimator of the gradient is used to update $\boldsymbol{\lambda}$. In particular, the instantaneous value of the forward model computed at time t, that is $\boldsymbol{s}(\boldsymbol{q}(t))$, can be used to this aim. The update rule for $\boldsymbol{\lambda}$ can thus be rewritten as a differential equation:

$$\dot{\lambda}_i(t) = -\eta_i(t) \left(s_i^{exp} - s_i(\boldsymbol{q}(t)) - \langle \epsilon_i \rangle_{\lambda_i(t)} \right) \tag{37}$$

with initial condition $\boldsymbol{\lambda}(0) = \boldsymbol{0}$.

Notice that now η plays the role of a *learning rate* and depends on the simulation time t. This choice is motivated by the fact that approximating the true gradient with its unbiased estimator introduces a noise into its estimate. In order to decrease the effect of such noise, a common choice when using SGD is to reduce the learning rate as the minimization (learning) process progresses with a typical schedule $\eta(t) \propto 1/t$ for large times. In our previous work [48] we adopted a learning rate from the class *search then converge* [70], which prescribes to choose $\eta_i(t) = k_i / \left(1 + \frac{t}{\tau_i} \right)$. Here k_i represents the initial learning rate and τ_i represents its damping time. In this manner, the learning rate is large at the beginning of the simulation and decreases proportionally to $1/t$ for large simulation times. The parameters k_i and τ_i are application specific and must be tuned by a trial and error procedure. In particular, a very small value of τ will cause the learning rate to decrease very fast, increasing the probability to get stuck in a suboptimal minimum. On the other hand, a very large value of τ will

prevent step-size shrinking and thus will hinder convergence. Analogous reasoning also applies to k (see Section 6.1 for numerical examples). Also notice that the k_i's are measured in units of the inverse of the observable squared multiplied by an inverse time and could thus in principle be assigned to different values in case of heterogeneous observables. It appears reasonable to choose them inversely proportional to the observable variance in the prior, in order to make the result invariant with respect to a linear transformation of the observables. On the other hand, the τ_i parameter should probably be independent of i in order to avoid different λ_i's to converge on different timescales.

Once Lagrangian multipliers are converged or, at least, stably fluctuating around a given value, the optimal value λ^* can be estimated by taking a time average of λ over a suitable time window. At that point, a new simulation could be performed using a static potential $V^*(q) = V_0(q) + k_B T \lambda^* \cdot s(q)$, either from a different molecular structure or starting from the structure obtained at the end of the learning phase. Such a simulation done with a static potential can be used to rigorously validate the obtained λ^*. Notice that, if errors have been included in the model, such validation should be made by checking that $\langle s \rangle \approx s^{exp} - \langle \epsilon \rangle$. Even if the resulting λ^* are suboptimal, it is plausible that such a simulation could be further reweighted (Section 5.1) more easily than the one performed with the original force field. When modeling errors, if an already restrained trajectory is reweighted one should be aware that restraints will be overcounted resulting in an effectively decreased experimental error (see Section 3).

As an alternative, one can directly analyze the learning simulation. Whereas strictly speaking this simulation is performed out of equilibrium, this approach has the advantage that it allows the learning phase to be prolonged until the agreement with experiment is satisfactory.

The optimization procedure discussed in this Section was used in order to enforce NMR data on RNA nucleosides and dinucleotides in Ref. [48], where it was further extended in order to simultaneously constrain multiple systems by keeping their force fields chemically consistent. This framework represents a promising avenue for the improvement of force fields, although it is intrinsically limited by the fact that the functional form of the correcting potential is by construction related to the type of available experimental data. However, the method in its basic formulation described here can be readily used in order to enforce system-specific experimental constraints.

Finally, notice that Equation (37) is closely related to the on-the-fly procedure proposed in the appendix of Ref. [47], where a term called "generalized force" and proportional to λ is calculated from an integral over the trajectory. Using the notation of this review, considering a Gaussian prior for the error (Section 3), and setting the confidence in the force field $\theta = 1$, the time-evolution of λ proposed in Ref. [47] could be rewritten in differential form as Equation (37) with $\eta(t) = 1/(\sigma_i^2 t)$, however with a different initial condition $\lambda(0) = s(q(0))/\sigma_i^2$.

5.4. Other On-the-Fly Optimization Strategies

Other optimization strategies have been proposed in the literature. The already mentioned target metadynamics (Section 2.4) provides a framework to enforce experimental data, and was applied to enforce reference distributions obtained from more accurate simulation methods [36], from DEER experiments [37], or from conformations collected over structural databases [71]. It is however not clear if it can be extended to enforce individual averages.

Also the VES method [38] (Section 2.4) is designed to enforce full distributions. However, in its practical implementation, the correcting potential is expanded on a basis set and the average values of the basis functions are actually constrained, resulting thus numerically equivalent to the other methods discussed here. In VES, a function equivalent to $\Gamma(\lambda)$ is optimized using the algorithm by Bach and Moulines [72] that is optimally suitable for non-strongly-convex functions. This algorithm requires to estimate not only the gradient but also the Hessian of the function $\Gamma(\lambda)$. We recall that $\Gamma(\lambda)$ can be made strongly convex by suitable treatment of experimental errors (see Section 3). However, there might be situations where the Bach-Moulines algorithm outperforms the SGD.

The experiment-directed simulation (EDS) approach [73] instead does not take advantage of the function $\Gamma(\lambda)$ but rather minimizes with a gradient-based method [74] the square deviation between the experimental values and the time-average of the simulated ones. A later paper tested a number of related minimization strategies [75]. In order to compute the gradient of the ensemble averages $\langle s_i \rangle_\lambda$ with respect to λ it is necessary to compute the variance of the observables s_i in addition to their average. Average and variance are computed on short simulation segments. It is worth observing that obtaining an unbiased estimator for the variance is not trivial if the simulation segment is too short. Errors in the estimate of the variance would anyway only affect the effective learning rate of the Lagrangian multipliers. In the applications performed so far, a few tens of MD time steps were shown to be sufficient to this aim, but the estimates might be system dependent. A comparison of the approaches used in Refs. [73,75] with the SGD proposed in Ref. [48] in practical applications would be useful to better understand the pros and the cons of the two algorithms. EDS was used to enforce the gyration radius of a 16-bead polymer to match the one of a reference system [73]. Interestingly, the restrained polymer was reported to have not only the average gyration radius in agreement with the reference one, but also its distribution. This is a clear case where a maximum entropy (linear) restraint and a harmonic restraint give completely different results. The EDS algorithm was recently applied to a variety of systems (see, e.g., Refs. [34,75,76]).

6. Convergence of Lagrangian Multipliers in Systems Displaying Metastability

Evaluating the Lagrangian multipliers on-the-fly might be nontrivial especially in systems that present multiple metastable states. We here present some example using a model system and provide some recommendation for the usage of enhanced sampling methods.

6.1. Results for a Langevin System

We first illustrate the effect of the choices in the learning schedule on the convergence of the Lagrangian multipliers and on the sampled distribution when using a SGD approach (Section 5.3). We consider a one dimensional system subject to a potential $V_0(s) = -k_B T \ln \left(e^{-(s-s_A)^2/2\sigma^2} + e^{-(s-s_B)^2/2\sigma^2} \right)$, with $s_A = 0$, $s_B = 3$, and $\sigma = 0.4$. The system is evolved according to an overdamped Langevin equation with diffusion coefficient $D = 1$ using a timestep $\Delta t = 0.01$. The average value of s in the prior distribution is $\langle s \rangle_0 = (s_A + s_B)/2 = 1.5$. The potential has been chosen in order to exhibit a free-energy barrier and is thus representative of complex systems where multiple metastable states are available.

We then run an on-the-fly SGD scheme [48] in order to enforce an experimental average $s^{exp} = 1$. For simplicity, experimental error is not modeled. By using the analytical results of Section 4, it can be seen that the exact Lagrangian multipliers required to enforce this average is $\lambda^* = 0.214$. In particular, we test different choices for k and τ which represent respectively the initial value of the learning rate and its damping factor (see Section 5.3 for more details on these parameters). The list of parameters and the results are summarized in Table 1, whereas Figure 5 reports the actual trajectories, their histogram, and the time evolution of the Lagrangian multipliers.

Panels a1 and a3 in Figure 5 report results obtained with a correct choice of the parameters. Panel a1 shows that the Lagrangian multiplier has quite large fluctuations at the beginning of the simulation (as expected from SGD), which are then damped as the simulations proceeds. The resulting sampled posterior distribution (red bars in panel a3) is in close agreement with the analytical solution (continuous blue line). The resulting average $\langle \lambda \rangle \approx 0.207$ reported in Table 1 is in very good agreement with the analytical result ($\lambda^* = 0.214$).

Table 1. Summary of the results obtained with the Langevin model, including learning parameters (k and τ) and average $\langle \lambda \rangle$ and $\langle s \rangle$ computed over the second half of the simulation. In addition, we report the exact Lagrangian multiplier $\lambda^*_{\langle s \rangle}$ required to enforce an average equal to $\langle s \rangle$ and the exact average $\langle s \rangle_{\langle \lambda \rangle}$ corresponding to a Lagrangian multiplier $\langle \lambda \rangle$. The last two columns are obtained by using the analytical solutions described in Section 4. Panel labels match those in Figure 5.
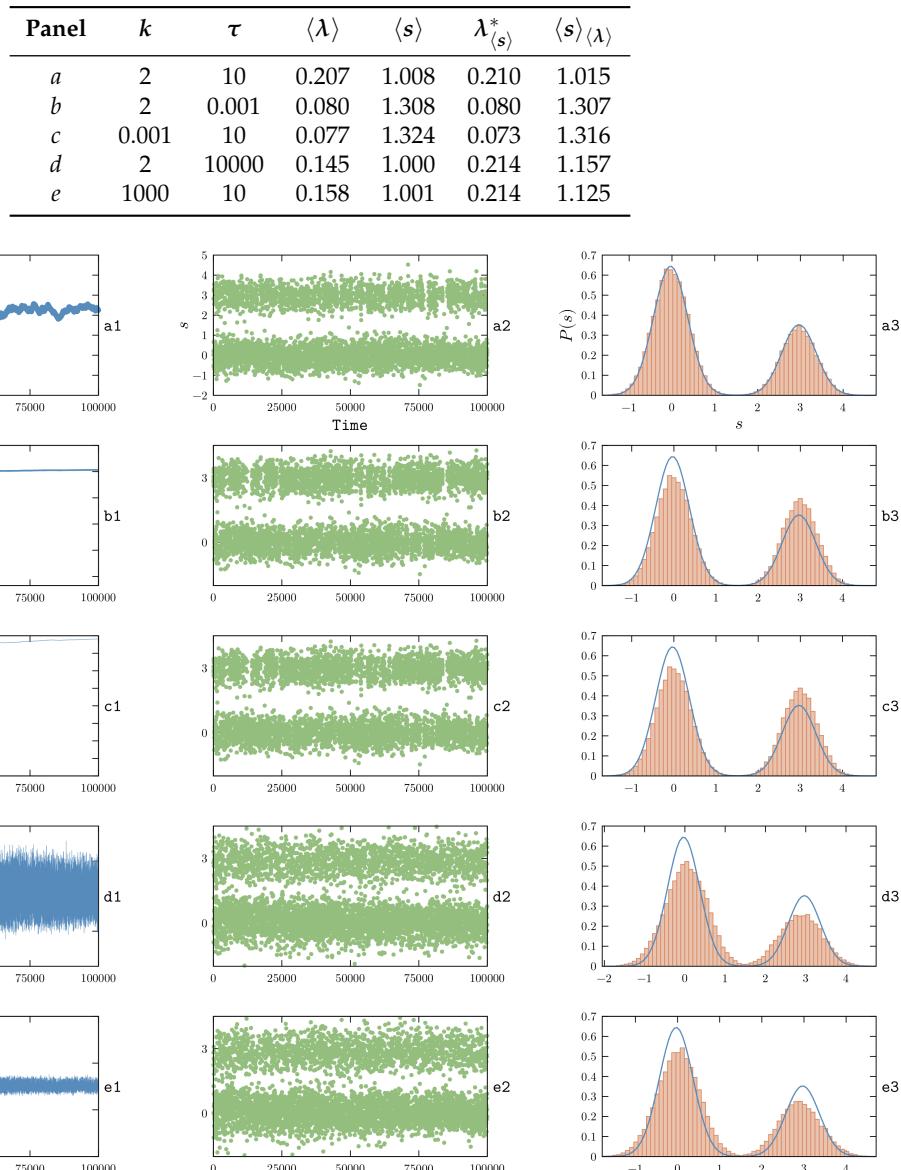
Panel	k	τ	$\langle \lambda \rangle$	$\langle s \rangle$	$\lambda^*_{\langle s \rangle}$	$\langle s \rangle_{\langle \lambda \rangle}$
a	2	10	0.207	1.008	0.210	1.015
b	2	0.001	0.080	1.308	0.080	1.307
c	0.001	10	0.077	1.324	0.073	1.316
d	2	10000	0.145	1.000	0.214	1.157
e	1000	10	0.158	1.001	0.214	1.125

Figure 5. Effect of choosing different values for k and τ when using stochastic gradient descent (SGD) on-the-fly during molecular dynamics (MD) simulations. Panel labels (**a–e**) refer to different sets of k and τ values matching those of Table 1. In particular, for each set of k and τ we show the convergence of the Lagrangian multipliers (number 1 of each letter), the time series of the observable (number 2 of each letter), and the resulting sampled posterior distribution, red bars, together with the analytical result, continuous line (number 3 of each letter).

Panels b1 and b3 in Figure 5 show the effect of choosing a very small value of τ. This choice not only kills the noise but also hinders the convergence of λ by shrinking too much the step-size during the minimization. The resulting distribution shown in panel b3 is clearly in disagreement with the analytical one having wrong populations for the two peaks. This example shows that apparently converged Lagrangian multipliers (panel b1) are not a sufficient condition for convergence to the

correct result, and it is necessary to check that the correct value was actually enforced. Panels c1 and c3 in Figure 5 show the effect of choosing a too small value of k. This scenario is very similar to the previous one since both cases result in small values of the learning rate η. Thus, what said for b1 and b3 also applies to c1 and c3. As reported in Table 1, in both cases the final average is $\langle s \rangle \approx 1.3$ and is thus visibly different from $s^{exp} = 1$. Thus, in a real application, this type of pathological behavior would be easy to detect. We recall that in case error is explicitly modeled (Section 3) one should compare $\langle s \rangle$ with $s^{exp} - \langle \epsilon \rangle_\lambda$.

Panels d1 and d3 in Figure 5 show the effect of choosing a very large value of τ. The effect of such choice is that the damping rate of the noise in Lagrangian multipliers is much slower than in the ideal case. This is reflected in the larger fluctuations of Lagrangian multipliers (panel d1) but also in an incorrect reconstruction of the posterior. The last example, panels e1 and e3 in Figure 5, shows the effect of choosing a very large value of k. In this case, the fluctuations of the Lagrangian multiplier (panel e1) are even higher than in the previous case. As reported in Table 1, in both cases the final average is equal to $s^{exp} = 1$. So, even though the sampled distribution has the correct average it is not the distribution that maximizes the entropy. This is a suboptimal solution that might be at least qualitatively satisfactory in some case. However, it is clear that there is no way to detect the incorrectness in the resulting distribution by just monitoring the enforced average. The only practical way to detect the problem indeed is to consider the resulting value of $\langle \lambda \rangle \approx 0.15$ and run a new simulation with a static potential. An additional indication of the problematic behavior is the large (several units) fluctuations in the Lagrangian multipliers. Indeed, the problem can be rationalized noting that the timescale at which λ evolves is too fast when compared with the typical time required to see a transition between one state and the other and the restraining force is overpushing the system forcing it to spend too much time in the region between the two peaks. The problem can be solved either slowing down the λ evolution (as in panel a) or by using enhanced sampling methods to increase the number of transitions.

6.2. Comments about Using Enhanced Sampling Methods

The model potential discussed above displays a free-energy barrier separating two metastable states. In order to properly sample both peaks in the distribution it is necessary to wait the time required to cross the barrier. If the transition is forced by very large fluctuations of λ, one can see that the resulting distribution is significantly distorted. For this reason, whenever a system displays metastability, it is highly recommended to use enhanced sampling techniques [2–4]. It is particularly important to employ methods that are capable to induce transitions between the states that contribute to the measured experimental averages. NMR timescales of typically µs-ms, i.e., upper limits for the lifetimes of interconverting conformations that are indistinguishable in the spectra, can be reached better using enhanced sampling techniques, since they result in probability distributions that would be effectively sampled by a much longer un-enhanced continuous simulation.

Replica exchange methods where one replica is unbiased are easy to apply since the learning procedure can be based on the reference replica. Methods such as parallel tempering [77], solute tempering [78], bias-exchange metadynamics with a neutral replica [79], or collective-variable tempering [80] can thus be used straightforwardly. Notice that in this case the higher replicas might feel either the same correcting potential as the reference replica (as it was done in Ref. [48]) or might be independently subject to the experimental restraints, provided the differences in the potential energy functions are properly taken into account in the acceptance rate calculation. Leaving the higher replicas uncorrected (i.e., simulated with the original force field) is suboptimal since they would explore a different portion of the space leading to fewer exchanges with the reference replica. It is also important to consider that, thanks to the coordinate exchanges, the reference replica will be visited by different conformations. These multiple conformations will all effectively contribute to the update of the Lagrangian multipliers. For instance, if an SGD is used, in the limit of very frequent exchanges,

the update will be done according to the average value of the observables over the conformations of all replicas, properly weighted with their probability to visit the reference replica.

Methods based on biased sampling, such as umbrella sampling [81], metadynamics [82], parallel-tempering metadynamics [83], bias-exchange (without a neutral replica) [79] or parallel-bias [84] metadynamics, require instead the implementation of some on-the-fly reweighting procedure in order to properly perform the update of the Lagrangian multipliers. The weighting factors should be proportional to the exponential of the biasing potential and, as such, might lead to a very large variability of the increment of λ (Equation (37)) that could make the choice of the learning parameters more difficult. For instance, this could result in very large λ_i's in the initial transient leading to large forces that make the simulation unstable. When reweighting (see Section 5.1) a trajectory generated using one of these enhanced sampling methods it is sufficient to use the weighting factors in the evaluation of Equation (34). Notice that similar arguments apply to replica-based methods (see Section 2.5), where on-the-fly reweighting is required in order to correctly compute the replica average [85]. If the resulting weights of different replicas are too different, the average value might be dominated by a single or a few replicas. A low number of replicas contributing to the average might in turn lead to large forces, unless the spring constant is suitable reduced [86], and to an entropy decrease.

Even in the absence of any enhanced sampling procedure, we notice that Lagrangian multipliers could be updated in a parallel fashion by multiple equivalent replicas, in a way resembling that used in multiple-walkers metadynamics to update the bias potential [87]. Since $s_i(q)$ enters linearly in Equation (37), this would be totally equivalent to using the arithmetic average between the walkers to update the Lagrangian multipliers (to be compared with the weighted average discussed above for replica-exchange simulations), showing an interesting analogy between Lagrangian multiplier optimization and replica-based methods (see Section 2.5). Such a multiple-walkers approach was used for instance in the well-tempered variant of VES [88], although in the context of enhanced sampling rather than to enforce experimental data.

7. Discussion and Conclusions

In this work, we reviewed a number of recently introduced techniques that are based on the maximum entropy principle and that allow experimental observations to be incorporated in MD simulations preserving the heterogeneity of the ensemble. We discuss here some general features of the reviewed methods.

First, one must keep in mind that, by design, the maximum entropy principle provides a distribution that, among those satisfying the experimental constraints, is as close as possible to the prior distribution. If the prior distribution is reasonable, a minimal correction is expected to be a good choice. However, for systems where the performance of classical force fields is very poor, the maximum entropy principle should be used with care and, if possible, should be based on a large number of experimental data so as to diminish the impact of force-field deficiencies on the final result. As a related issue, different priors are in principle expected to lead to different posteriors, and thus different ensemble averages for non-restrained quantities. There are indications that current force fields restrained by a sufficient number of experimental data points lead to equivalent posterior distributions at least for trialanine [51] and for larger disordered peptides [86,89]. It would be valuable to perform similar tests on other systems where force fields are known to be poorly predictive, such as unstructured RNAs or difficult-to-predict RNA structural motifs.

We here discussed both the possibility of reweighting a posteriori a trajectory and that of performing a simulation where the restraint is iteratively modified. Techniques where the elements of a previously generated ensemble are reweighted have the disadvantage that if the initial ensemble averages are far from the experimental values the weights will be distributed very inhomogeneously (i.e., very large λ_i will be needed), which means that singular conformations with observables close to the experimental values can be heavily overweighted to obtain the correct ensemble average. In the extreme case, it might not even be possible to find weights that satisfy the desired ensemble average,

since important conformations are simply missing in the ensemble. On the other hand, reweighting techniques have the advantage that they can be readily applied to new or different experimental data, without performing new simulations. Additionally, they can be used to reweight a non-converged simulation performed with an on-the-fly optimization.

When simulating systems that exhibit multiple metastable states, it might be crucial to combine the experimental constraints with enhanced sampling methods. This is particularly important if multiple metastable states contribute to the experimental average. As usual in enhanced sampling simulations, one should observe as many as possible transitions between the relevant metastable states. When using replica-based methods (either to enhance sampling or to compute averages), transitions should be observed in the continuous trajectories.

Several methods are based on the idea of simulating a number of replicas of the system with a restraint on the instantaneous average among the replicas and have been extended to treat experimental errors. These methods are expected to reproduce the maximum entropy distribution in the limit of a large number of replicas. However, if the number of replicas is too low, the deviation from the maximum entropy distribution might be significant. Indeed, the number of replicas should be large enough for all the relevant states to be represented with the correct populations. The easiest way to check if the number of replicas is sufficient is to compare simulations done using a different number of replicas. Methods based on Lagrangian multipliers reproduce the experimental averages by means of an average over time rather than an average over replicas. Thus, they can be affected by a similar problem if the simulation is not long enough. This sort of effect is expected to decrease when the simulation length increases and when using enhanced sampling techniques.

The on-the-fly refinement of Lagrangian multipliers typically requires ad hoc parameters for the learning phase that should be chosen in a system-dependent manner. Properly choosing these parameters is not trivial. Several different algorithms have been proposed in the last years and a systematic comparison on realistic applications would be very useful. It might also be beneficial to consider other stochastic optimization algorithms that have been proposed in the machine-learning community. Interestingly, all the methods discussed in this review for on-the-fly optimization (target metadynamics, maximum entropy with SGD, VES, and EDS) are available in the latest release of the software PLUMED [90] (version 2.4), which also implements replica-based methods, forward models to calculate experimental observables [91], and enhanced sampling methods.

Finally, we notice that there are cases where results might be easier to interpret if only a small number of different conformations were contributing to the experimental average. In order to obtain small sets of conformations that represent the ensemble and provide a clearer picture about the different states, several *maximum parsimony* approaches have been developed. Naturally, the selection of a suitable set of structures is done on an existing ensemble and not on-the-fly during a simulation. While some approaches use genetic algorithms to select the structures of a fixed-size set [28,92,93], others use matching pursuit [94] or Bayes-based reweighting techniques to obtain correct ensemble averages [95–98] while minimizing the number of non-zero weights, i.e., structures, in the set. These approaches are not central to this review and so were not discussed in detail.

In conclusion, the maximum (relative) entropy principle provides a consistent framework to combine molecular dynamics simulations and experimental data. On one hand, it allows improving not-satisfactory results sometime obtained when simulating complex systems with classical force fields. On the other hand, it allows the maximum amount of structural information to be extracted from experimental data, especially in cases where heterogeneous structures contribute to a given experimental signal. Moreover, if experimental errors are properly modeled, this framework allows to detect experimental data that are either mutually inconsistent or incompatible with the employed force field. For all these reasons, we expect this class of methods to be increasingly applied for the characterization of the structural dynamics of biomolecular systems in the coming future.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DEER   double electron-electron resonance
EDS   experiment-directed simulation
GD   gradient descent
MD   molecular dynamics
NMR   nuclear magnetic resonance
SAXS   small-angle X-ray scattering
SGD   stochastic gradient descent
VES   variationally enhanced sampling

References

1. Dror, R.O.; Dirks, R.M.; Grossman, J.; Xu, H.; Shaw, D.E. Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.
2. Bernardi, R.C.; Melo, M.C.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta Gen. Subj.* **2015**, *1850*, 872–877.
3. Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67*, 159–184.
4. Mlỳnskỳ, V.; Bussi, G. Exploring RNA structure and dynamics through enhanced sampling simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 63–71.
5. Petrov, D.; Zagrovic, B. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput. Biol.* **2014**, *10*, e1003638.
6. Piana, S.; Lindorff-Larsen, K.; Shaw, D.E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **2011**, *100*, L47–L49.
7. Condon, D.E.; Kennedy, S.D.; Mort, B.C.; Kierzek, R.; Yildirim, I.; Turner, D.H. Stacking in RNA: NMR of Four Tetramers Benchmark Molecular Dynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2729–2742.
8. Bergonzo, C.; Henriksen, N.M.; Roe, D.R.; Cheatham, T.E. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **2015**, *21*, 1578–1590.
9. Šponer, J.; Bussi, G.; Krepl, M.; Banáš, P.; Bottaro, S.; Cunha, R.A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jurečka, P.; et al. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chem. Rev.* **2018**, doi:10.1021/acs.chemrev.7b00427.
10. Kuhrová, P.; Best, R.B.; Bottaro, S.; Bussi, G.; Šponer, J.; Otyepka, M.; Banáš, P. Computer Folding of RNA Tetraloops: Identification of Key Force Field Deficiencies. *J. Chem. Theory Comput.* **2016**, *12*, 4534–4548.
11. Bottaro, S.; Banáš, P.; Šponer, J.; Bussi, G. Free Energy Landscape of GAGA and UUCG RNA Tetraloops. *J. Phys. Chem. Lett.* **2016**, *7*, 4032–4038.
12. Schröder, G.F. Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* **2015**, *31*, 20–27.
13. Ravera, E.; Sgheri, L.; Parigi, G.; Luchinat, C. A critical assessment of methods to recover information from averaged data. *Phys. Chem. Chem. Phys.* **2016**, *18*, 5686–5701.
14. Allison, J.R. Using simulation to interpret experimental data in terms of protein conformational ensembles. *Curr. Opin. Struct. Biol.* **2017**, *43*, 79–87.
15. Bonomi, M.; Heller, G.T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116.
16. Pitera, J.W.; Chodera, J.D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.
17. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **2014**, *10*, e1003406.

18. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620.

19. Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171.

20. Caticha, A. Relative entropy and inductive inference. In *AIP Conference Proceedings*; AIP: College Park, MD, USA, 2004; Volume 707; pp. 75–96.

21. Banavar, J.; Maritan, A. The maximum relative entropy principle. *arXiv* **2007**, arXiv:0703622.

22. Shell, M.S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.

23. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

24. Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

25. Case, D.A. Chemical shifts in biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172–176.

26. Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.

27. Tolman, J.R.; Ruan, K. NMR residual dipolar couplings as probes of biomolecular dynamics. *Chem. Rev.* **2006**, *106*, 1720–1736.

28. Bernadó, P.; Mylonas, E.; Petoukhov, M.V.; Blackledge, M.; Svergun, D.I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664.

29. Jeschke, G. DEER distance measurements on proteins. *Annu. Rev. Phys. Chem.* **2012**, *63*, 419–446.

30. Piston, D.W.; Kremers, G.J. Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem. Sci.* **2007**, *32*, 407–414.

31. Mead, L.R.; Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **1984**, *25*, 2404–2417.

32. Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.

33. Chen, S.F.; Rosenfeld, R. A Gaussian Prior for Smoothing Maximum Entropy Models; Technical Report; 1999. Available online: http://reports-archive.adm.cs.cmu.edu/anon/anon/1999/CMU-CS-99-108.pdf (accessed on 4 February 2018).

34. Dannenhoffer-Lafage, T.; White, A.D.; Voth, G.A. A Direct Method for Incorporating Experimental Data into Multiscale Coarse-Grained Models. *J. Chem. Theory Comput.* **2016**, *12*, 2144–2153.

35. Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624–1636.

36. White, A.D.; Dama, J.F.; Voth, G.A. Designing free energy surfaces that match experimental data with metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2451–2460.

37. Marinelli, F.; Faraldo-Gómez, J.D. Ensemble-biased metadynamics: A molecular simulation method to sample experimental distributions. *Biophys. J.* **2015**, *108*, 2779–2782.

38. Valsson, O.; Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **2014**, *113*, 090601.

39. Shaffer, P.; Valsson, O.; Parrinello, M. Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 1150–1155.

40. Invernizzi, M.; Valsson, O.; Parrinello, M. Coarse graining from variationally enhanced sampling applied to the Ginzburg–Landau model. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3370–3374.

41. Fennen, J.; Torda, A.E.; van Gunsteren, W.F. Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. *J. Biomol. NMR* **1995**, *6*, 163–170.

42. Best, R.B.; Vendruscolo, M. Determination of Protein Structures Consistent with NMR Order Parameters. *J. Am. Chem. Soc.* **2004**, *126*, 8090–8091.

43. Lindorff-Larsen, K.; Best, R.B.; DePristo, M.A.; Dobson, C.M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433*, 128–132.

44. Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, 094112.

45. Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*, 084107.

46. Olsson, S.; Cavalli, A. Quantification of Entropy-Loss in Replica-Averaged Modeling. *J. Chem. Theory Comput.* **2015**, *11*, 3973–3977.

47. Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.

48. Cesari, A.; Gil-Ley, A.; Bussi, G. Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192–6200.

49. Olsson, S.; Ekonomiuk, D.; Sgrignani, J.; Cavalli, A. Molecular dynamics of biomolecules through direct analysis of dipolar couplings. *J. Am. Chem. Soc.* **2015**, *137*, 6270–6278.

50. Camilloni, C.; Vendruscolo, M. A tensor-free method for the structural and dynamical refinement of proteins using residual dipolar couplings. *J. Phys. Chem. B* **2014**, *119*, 653–661.

51. Beauchamp, K.A.; Pande, V.S.; Das, R. Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. *Biophys. J.* **2014**, *106*, 1381–1390.

52. Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, e1501177.

53. Brookes, D.H.; Head-Gordon, T. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **2016**, *138*, 4530–4538.

54. Różycki, B.; Kim, Y.C.; Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011**, *19*, 109–116.

55. Boura, E.; Różycki, B.; Herrick, D.Z.; Chung, H.S.; Vecer, J.; Eaton, W.A.; Cafiso, D.S.; Hummer, G.; Hurley, J.H. Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9437–9442.

56. Sanchez-Martinez, M.; Crehuet, R. Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings. *Phys. Chem. Chem. Phys.* **2014**, *16*, 26030–26039.

57. Leung, H.T.A.; Bignucolo, O.; Aregger, R.; Dames, S.A.; Mazur, A.; Bernéche, S.; Grzesiek, S. A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *J. Chem. Theory Comput.* **2015**, *12*, 383–394.

58. Cunha, R.A.; Bussi, G. Unraveling Mg2+–RNA binding with atomistic molecular dynamics. *RNA* **2017**, *23*, 628–638.

59. Bottaro, S.; Bussi, G.; Kennedy, S.D.; Turner, D.H.; Lindorff-Larsen, K. Conformational Ensemble of RNA Oligonucleotides from Reweighted Molecular Simulations. *bioRxiv* **2017**, 230268, doi:10.1101/230268.

60. Podbevsek, P.; Fasolo, F.; Bon, C.; Cimatti, L.; Reisser, S.; Carninci, P.; Bussi, G.; Zucchelli, S.; Plavec, J.; Gustincich, S. Structural determinants of the SINEB2 element embedded in the long non-coding RNA activator of translation AS Uchl1. *Sci. Rep.* **2018**, accepted.

61. Shen, T.; Hamelberg, D. A statistical analysis of the precision of reweighting-based simulations. *J. Chem. Phys.* **2008**, *129*, 034103.

62. Gray, P.G.; Kish, L. Survey Sampling. *J. R. Stat. Soc. A* **1969**, *132*, 272.

63. Martino, L.; Elvira, V.; Louzada, F. Effective sample size for importance sampling based on discrepancy measures. *Signal Process.* **2017**, *131*, 386–401.

64. Norgaard, A.B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **2008**, *94*, 182–192.

65. Giorgetti, L.; Galupa, R.; Nora, E.P.; Piolot, T.; Lam, F.; Dekker, J.; Tiana, G.; Heard, E. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **2014**, *157*, 950–963.

66. Tiana, G.; Amitai, A.; Pollex, T.; Piolot, T.; Holcman, D.; Heard, E.; Giorgetti, L. Structural fluctuations of the chromatin fiber within topologically associating domains. *Biophys. J.* **2016**, *110*, 1234–1245.

67. Zhang, B.; Wolynes, P.G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6062–6067.

68. Zhang, B.; Wolynes, P.G. Shape transitions and chiral symmetry breaking in the energy landscape of the mitotic chromosome. *Phys. Rev. Lett.* **2016**, *116*, 248101.

69. Torda, A.E.; Scheek, R.M.; van Gunsteren, W.F. Time-dependent distance restraints in molecular dynamics simulations. *Chem. Phys. Lett.* **1989**, *157*, 289–294.

70. Darken, C.; Moody, J. Towards faster stochastic gradient search. In Proceedings of the Neural Information Processing Systems 4 (NIPS 1991), Denver, CO, USA, 2–5 December 1991; pp. 1009–1016.

71. Gil-Ley, A.; Bottaro, S.; Bussi, G. Empirical Corrections to the Amber RNA Force Field with Target Metadynamics. *J. Chem. Theory Comput.* **2016**, *12*, 2790–2798.

72. Bach, F.; Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). In Proceedings of the Neural Information Processing Systems 16 (NIPS 2013), Lake Tahoe, CA, USA, 4–11 December 2013; pp. 773–781.

73. White, A.D.; Voth, G.A. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *J. Chem. Theory Comput.* **2014**, *10*, 3023–3030.

74. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.

75. Hocky, G.M.; Dannenhoffer-Lafage, T.; Voth, G.A. Coarse-grained Directed Simulation.*J. Chem. Theory Comput.* **2017**, *13*, 4593–4603.

76. White, A.D.; Knight, C.; Hocky, G.M.; Voth, G.A. Communication: Improved ab initio molecular dynamics by minimally biasing with experimental data. *J. Chem. Phys* **2017**, *146*, 041102.

77. Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

78. Liu, P.; Kim, B.; Friesner, R.A.; Berne, B. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13749–13754.

79. Piana, S.; Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.

80. Gil-Ley, A.; Bussi, G. Enhanced Conformational Sampling Using Replica Exchange with Collective-Variable Tempering. *J. Chem. Theory Comput.* **2015**, *11*, 1077–1085.

81. Torrie, G.M.; Valleau, J.P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

82. Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

83. Bussi, G.; Gervasio, F.L.; Laio, A.; Parrinello, M. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

84. Pfaendtner, J.; Bonomi, M. Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 5062–5067.

85. Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic metainference: enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* **2016**, *6*, 31232.

86. Löhr, T.; Jussupow, A.; Camilloni, C. Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J. Chem. Phys.* **2017**, *146*, 165102.

87. Raiteri, P.; Laio, A.; Gervasio, F.L.; Micheletti, C.; Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* **2006**, *110*, 3533–3539.

88. Valsson, O.; Parrinello, M. Well-tempered variational approach to enhanced sampling.*J. Chem. Theory Comput.* **2015**, *11*, 1996–2002.

89. Tiberti, M.; Papaleo, E.; Bengtsen, T.; Boomsma, W.; Lindorff-Larsen, K. ENCORE: Software for quantitative ensemble comparison. *PLoS Comput. Biol.* **2015**, *11*, e1004415.

90. Tribello, G.A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

91. Bonomi, M.; Camilloni, C. Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* **2017**, *33*, 3999–4000.

92. Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M.R.; Blackledge, M. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.

93. Pelikan, M.; Hura, G.L.; Hammel, M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **2009**, *28*, 174–189.

94. Berlin, K.; Castañeda, C.A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595–16609.

95. Yang, S.; Blachowicz, L.; Makowski, L.; Roux, B. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 15757–15762.

96. Fisher, C.K.; Huang, A.; Stultz, C.M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919–14927.

97. Cossio, P.; Hummer, G. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.* **2013**, *184*, 427–437.

98. Molnar, K.S.; Bonomi, M.; Pellarin, R.; Clinthorne, G.D.; Gonzalez, G.; Goldberg, S.D.; Goulian, M.; Sali, A.; DeGrado, W.F. Cys-Scanning Disulfide Crosslinking and Bayesian Modeling Probe the Transmembrane Signaling Mechanism of the Histidine Kinase, PhoQ. *Structure* **2014**, *22*, 1239–1251.