# A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations

Brotzakis, Z.F.; Vendruscolo, M.; Bolhuis, P.G.

DOI
10.1073/pnas.2012423118

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the National Academy of Sciences of the United States of America

# A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations

Z. Faidon Brotzakis[a], Michele Vendruscolo[a], and Peter G. Bolhuis[b,1]

[a]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; and [b]van't Hoff Institute for Molecular Sciences, University of Amsterdam, 1090 GD Amsterdam, The Netherlands

Edited by Peter J. Rossky, Rice University, Houston, TX, and approved November 23, 2020 (received for review June 16, 2020)

**From the point of view of statistical mechanics, a full characterization of a molecular system requires an accurate determination of its possible states, their populations, and the respective interconversion rates. Toward this goal, well-established methods increase the accuracy of molecular dynamics simulations by incorporating experimental information about states using structural restraints and about populations using thermodynamics restraints. However, it is still unclear how to include experimental knowledge about interconversion rates. Here, we introduce a method of imposing known rate constants as constraints in molecular dynamics simulations, which is based on a combination of the maximum-entropy and maximum-caliber principles. Starting from an existing ensemble of trajectories, obtained from either molecular dynamics or enhanced trajectory sampling, this method provides a minimally perturbed path distribution consistent with the kinetic constraints, as well as modified free energy and committor landscapes. We illustrate the application of the method to a series of model systems, including all-atom molecular simulations of protein folding. Our results show that by combining experimental rate constants and molecular dynamics simulations, this approach enables the determination of transition states, reaction mechanisms, and free energies. We anticipate that this method will extend the applicability of molecular simulations to kinetic studies in structural biology and that it will assist the development of force fields to reproduce kinetic and thermodynamic observables. Furthermore, this approach is generally applicable to a wide range of systems in biology, physics, chemistry, and material science.**

biomolecular simulation | kinetics | MaxCal | transition path sampling

The first step in the study of a molecular system typically consists of the determination of its conformation, as, for example, most commonly done by using X-ray crystallography (X-ray), cryo-electron microscopy (cryo-EM), or nuclear magnetic resonance spectroscopy (NMR), for obtaining the structures of proteins and of nucleic acids (1). By revealing a wide range of structure–function relationships, this approach has enabled major advances in molecular biology (1). From a procedural point of view, experimental measurements, such as electron densities in X-ray and cryo-EM or interproton distances in NMR, combined with well-established theoretical chemistry rules, facilitate the building of molecular structures using computational methods (2).

As at the molecular level under physiological conditions, thermal fluctuations are relevant, it is becoming increasingly common to perform a second step, which involves the determination of the structures of the thermally excited states of a molecular system, together with their populations (3).* This goal is typically achieved by incorporating experimental measurements as structural restraints in molecular dynamics (MD) simulations to sample the free energy landscape (4). The maximum entropy principle (MaxEnt) provides a rigorous framework to implement this strategy. To carry out this step, a range of methods are now available (5, 6), resulting in the determination of a "thermodynamic ensemble" of structures (7).

One may not, however, stop at this level if kinetic properties are to be characterized. As a third step in the determination of a molecular system, one would like to obtain a "kinetic ensemble," comprising the structures of the different states of a molecular system, their populations, and their interconversion rates (7, 8). Approaches for determining experimentally kinetic ensembles are, however, not yet readily available, as there is no well-established method of incorporating experimental information about kinetic rates in molecular modeling procedures. Our aim here is to make a first step in this direction.

To achieve this goal, we start from a MaxEnt approach, where one maximizes a configurational entropy subject to constraints given by experimental data in order to predict a new configurational probability distribution. MaxEnt can also model uncertainties in the data, effectively turning constraints into restraints (5, 6). Addressing the problem in various ways as a Bayesian or a maximum-likelihood problem (6, 9–13) leads to numerous applications, for example, in cases where force fields are less accurate, such as for intrinsically disordered proteins and RNA (9, 14–17). Apart from ensemble refinement, application of MaxEnt yields perturbative correction terms to the potential energy along order parameters or CVs relating to the experimental data (4, 9, 18).

To enforce experimental information about rate constants, the MaxEnt method can be combined with the maximum caliber principle (MaxCal) (19). This approach seems, again, quite natural, as MaxCal is a general variational framework of nonequilibrium and equilibrium statistical mechanics with a wide scope, from flux–fluctuation relationships to pathway distributions and slow mode identification (19–21). In MaxCal, one maximizes a

## Significance

**Molecular dynamics simulations are often combined with experimental data with the aim to improve structural models. Here, we introduce a method of imposing known rate constants as constraints in such simulations, based on a combination of the maximum-entropy and maximum-caliber principles. The method corrects an initial path ensemble by reweighting it in order to match the calculated and experimental interconversion rates of a molecular transition of interest, yielding improved structure, kinetics, and thermodynamics, as well as mechanistic insights that may not be readily evident directly from the experiments.**

Author contributions: Z.F.B. and P.G.B. designed research; Z.F.B. and P.G.B. performed research; and Z.F.B., M.V., and P.G.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: p.g.bolhuis@uva.nl.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2012423118/-/DCSupplemental.

Published December 29, 2020.

*"Excited" refers to the dynamical structures that the system visits rather than electronic excitation.

path entropy over all possible pathways, subject to dynamical constraints such as average fluxes, in order to predict relative path weights (22). Rigorous and general MaxCal implementations have found, so far, fewer applications compared to MaxEnt approaches due to the difficulty both in sampling path distributions of complex systems and in acquiring experimental data about rate constants. For example, MaxCal enabled reweighting of the equilibrium distribution of macrostates given experimental rate constants for Markov state models or time-discrete pathways (23, 24). A recent implementation of the MaxCal for time-resolved data imposed time-dependent constraints along a few degrees of freedom of the system, or collective variables (CVs), to agree with time-resolved experimental data (25). MaxCal methods have also been extended to nonequilibrium dynamics (22, 26). An important aspect, however, is that such methods rely on spatially discrete models, on limited time-resolved data, or on biased dynamics, while, in practice, one usually only has access to experimental rate or diffusion constants. As starkly captured by Jaynes (27), "reconstructing MaxCal path ensembles containing the microscopic space and time dynamics is difficult." While theoretically rigorous, the MaxCal formalism has not been implemented to date for reweighting time- and space-continuous unbiased trajectories.

Here, we present a method of determining kinetic ensembles using the MaxCal strategy, by reweighting path ensemble distributions a posteriori, according to both kinetic and thermodynamic experimental data. The methodology yields experimentally corrected free energy landscapes and provides structural ensembles that exhibit accurate configurations, including in the barrier regions between states. In addition, it allows statistical mechanistic interpretation in the form of experimentally corrected committor profiles or landscapes, where the committor is the probability for configurations to reach the product state (28), and can be considered the perfect reaction coordinate (RC) for the process at hand (29). The committor landscape is, thus, effectively a statistical representation of the mechanism and gives information on its transition states or dynamical bottlenecks.

Given experimental forward and backward rate constant constraints, we combine MaxEnt with MaxCal to find a biasing function that simultaneously acts on equilibrium properties and on rate constants. This bias function gives correcting weights to the pathways of the equilibrium path ensemble distribution, without altering the dynamical trajectories themselves. The equilibrium path ensemble distribution is generated from computing reweighted path ensembles (RPEs) (30) based on either long MD trajectories or on enhanced sampling of trajectories, e.g., using Transition Interface Sampling (TIS) simulations (31), or in one step using the Virtual Interface Exchange Transition Path Sampling (VIE-TPS) method (32) for pathways sampled by Transition Path Sampling (TPS) (28, 33). Such TPS-based methods focus on reactive or partially reactive pathways, thereby bypassing the computationally expensive sampling of the stable states. In the remainder of the text, RPE will refer to the reconstructed equilibrium (reweighted) path ensemble distribution from simulation, while "kinetic ensemble" will refer to the equilibrium path ensemble distribution after imposing the experimental kinetic constraints.

While our method applies constraints to the distributions, imposing strict equality with experimental data, MaxEnt and MaxCal allow us to model uncertainties in the experimental and simulation data by turning constraints into restraints that impose equality within errors in the data (12).

In this work, we focus on biological problems without losing the generality of our statements. Thus, our approach can be applied to all MD simulations where trajectory reweighting to match target kinetics is possible and helpful.

## Theory

**MaxEnt in Configuration Space.** In this section, we briefly recapitulate how MaxEnt can be used to combine molecular simulations with experimental data (5, 6). In its original formulation, MaxEnt states that the probability distribution of the states of a system maximally compatible with a set of observed data are the one maximizing the associated Shannon entropy. This principle has been extended to a maximum relative entropy principle, which has the advantage of being invariant with respect to changes of coordinates and coarse-graining and has been shown to play an important role in multiscale problems (17). The entropy is computed here relative to a given prior distribution $P^0(x)$ and, for a system described by a set of continuous variables $x$—e.g., the positions and velocities of all atoms in a molecular system—is defined as

$$S[P\|P^0] = -\int dx P(x) \ln \frac{P(x)}{P^0(x)}. \quad [1]$$

This entropy can be maximized as

$$P^{ME}(x) = \arg\max_{P(x)} S[P\|P^0], \quad [2]$$

$$\text{subject to: } \begin{cases} \int dx P(x) s_i(x) = \langle s_i(x) \rangle = s_i^{exp} \\ \int dx P(x) = 1 \end{cases},$$

where the experimental observations $s_i^{exp}$ ($i \in \{1, 2 \dots M\}$) constrain the ensemble average of $M$ observables $s_i(x)$, computed over the distribution $P(x)$, to be equal to $s_i^{exp}$, and an additional constraint ensures that the distribution $P(x)$ is normalized. $P^0$ $(x)$ is called the "prior" probability distribution, encoding the knowledge available before the experimental measurement. $P^{ME}(x)$ instead represents the best estimate for the probability distribution after the experimental constraints have been enforced and is thus called the "posterior" probability distribution. 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 (KL) divergence $D_{KL}[P\|P^0]$, 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 KL divergence measures how much information is lost when prior knowledge $P^0$ $(x)$ is replaced with $P(x)$. Always nonnegative, the KL divergence is a measure of the difference between the distributions and vanishes only if the two distributions are identical.

A powerful approach to solve the maximization problem in Eq. **1** is based on the method of Lagrange multipliers, namely, searching for the stationary point of the following Lagrange function:

$$\mathcal{L} = S[P\|P^0] - \sum_{i=1}^{M} \mu_i \left( \int dx s_i(x) P(x) - s_i^{exp} \right) - \nu \left( \int dx P(x) - 1 \right), \quad [3]$$

where $\mu_i$ and $\nu$ are suitable Lagrange multipliers taking care of the experimental observations and the probability normalization, respectively. The functional derivative of $\mathcal{L}$ with respect to $P(x)$ is

$$\frac{\delta \mathcal{L}}{\delta P(x)} = -\ln \frac{P(x)}{P^0(x)} - 1 - \sum_{i=1}^{M} \mu_i s_i(x) - \nu. \quad [4]$$

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

$$P^{ME}(x) \propto e^{-\sum_{i=1}^{M} \mu_i s_i(x)} P^0(x). \quad [5]$$

Solving Eq. **4** turns out to be equivalent to minimizing the function

$$\Gamma(\boldsymbol{\mu}) = \ln \left[ \int dx e^{-\sum_{i=1}^{M} \mu_i s_i(x)} P^0(x) \right] + \boldsymbol{\mu} \cdot \boldsymbol{s}^{exp}, \quad [6]$$

with respect to $\mu_i$, leading to the equation(s) $\langle s_i \rangle = \int dx s_i(x) P(x) = s_i^{exp}$, and thus giving for each observable the Lagrange multiplier $\mu_i$.

We also note that MaxEnt can model uncertainties in the data—i.e., the experimental errors (9, 12). This is done by adding the expected error due to the perturbed distribution $\langle e_i \rangle$ to the constraint average—i.e., $\langle s_i \rangle = s_i^{exp} + \langle e_i \rangle$. For a Gaussian-distributed error with a standard deviation $\sigma_i$, the average error is $\langle e_i \rangle = -\mu_i \sigma_i^2$, with $\sigma_i$ the level of confidence in the data—e.g., experimental measurements. Adding this to Eq. **6** yields

$$\Gamma(\boldsymbol{\mu}) = \ln \left[ \int dx P^{ME}(x) \right] + \boldsymbol{\mu} \cdot \boldsymbol{s}^{exp} + \frac{1}{2} \sum_{i=1}^{M} \mu_i^2 \sigma_i^2. \quad [7]$$

Minimizing this function leads to a solution of the Lagrange multipliers $\mu_i$ that account for the error. If $\sigma = 0$, the situation is identical to Eq. **6**, while if $\sigma$ is large, the Lagrange multiplier will be close to zero, almost not perturbing the original distribution. In this way, the constraint on the distribution is turned into a restraint, depending on the level of confidence in the data. In most of our presentation, we will discuss imposing constraints, although one should keep in mind that it is always possible to extend the results to imposing restraints, using the above procedure.

**MaxCal in Path Space.** The MaxEnt principle can be extended to trajectory space (22). Consider a prior path probability distribution $\mathcal{P}^0[\mathbf{x}]$ of trajectories $\mathbf{x}$, each consisting of $L+1$ frames, $\mathbf{x} = \{x_0, x_1, \ldots x_L\}$, where subsequent frames are separated by a time interval $\Delta t$, such that the total duration of a path is $\mathcal{T} = L\Delta t$. The path probability is

$$\mathcal{P}^0[\mathbf{x}] = \rho(x_0) \prod_{i=0}^{L-1} p(x_i \to x_{i+1}), \quad [8]$$

where $\rho(x_0)$ denotes the initial condition, usually the Boltzmann distribution $\rho(x) = \exp(-\beta H(x))$ with $H(x)$ the Hamiltonian of the system, and $p(x_i \to x_{i+1})$ is a short-time Markovian probability that represents dynamical evolution according to the equations of motion, as given, e.g., by an MD simulation, and contains reliable dynamic information, of course up to the extent of the resolution and faithfulness of the force field.

The (relative) path entropy, or caliber, for any path distribution $\mathcal{P}[\mathbf{x}]$ is

$$S[\mathcal{P}||\mathcal{P}^0] = -\int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] \ln \frac{\mathcal{P}[\mathbf{x}]}{\mathcal{P}^0[\mathbf{x}]}, \quad [9]$$

where $\mathcal{D}\mathbf{x}$ indicates an integral over all trajectories or paths $\mathbf{x}$. MaxCal states that the optimal distribution $\mathcal{P}^{MC}[\mathbf{x}]$ is given by

$$\mathcal{P}^{MC}[\mathbf{x}] = \arg\max_{\mathcal{P}[\mathbf{x}]} S[\mathcal{P}||\mathcal{P}^0], \quad [10]$$

$$\text{subject to: } \begin{cases} \int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] s_i[\mathbf{x}] = \langle s_i[\mathbf{x}] \rangle = s_i^{exp} \\ \int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] = 1. \end{cases}$$

That is, $\mathcal{P}^{MC}[\mathbf{x}]$ maximizes the path entropy or caliber, while obeying the constraints given by external constraint $s_i^{exp}$. The observable ensemble average $\langle s_i[\mathbf{x}] \rangle$ can relate to any measurement either giving rise to static/thermodynamic or dynamic/kinetic information. Note that $s[\mathbf{x}]$ is now a path function, which includes (auto)correlation functions. We refer the interested reader to *SI Appendix* for further elaboration.

Following a procedure similar to MaxEnt yields the Lagrange function

$$\mathcal{L} = -\int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] \ln \frac{\mathcal{P}[\mathbf{x}]}{\mathcal{P}^0[\mathbf{x}]} - \nu \left( \int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] - 1 \right) - \sum_i \mu_i \left( \int \mathcal{D}\mathbf{x} \mathcal{P}[\mathbf{x}] s_i[\mathbf{x}] - s_i^{exp} \right), \quad [11]$$

with a derivative

$$\frac{\delta \mathcal{L}}{\delta \mathcal{P}[\mathbf{x}]} = -\ln \frac{\mathcal{P}[\mathbf{x}]}{\mathcal{P}^0[\mathbf{x}]} - 1 - \sum_i \mu_i s_i[\mathbf{x}] - \nu, \quad [12]$$

giving rise to the posterior

$$\mathcal{P}^{MC}[\mathbf{x}] \propto e^{-\sum_i \mu_i s_i[\mathbf{x}]} \mathcal{P}^0[\mathbf{x}]. \quad [13]$$

Thus, given a prior ensemble of paths $\mathcal{P}^0[\mathbf{x}]$, one can reweight each path, while leaving the actual trajectories intact. As such, the reweighting can be interpreted to only affect the distribution of initial conditions in Eq. **8**. We will come back to this later.

**Independence of Partial Path Distributions.** Up to now, we did not specify what the path ensemble distribution refers to. In what follows, we focus on systems that show two-state kinetics between two stable states, A and B. We assume that there is a separation between the molecular timescale and the reaction time (34), to guarantee that well-defined rate constants exist for the interconversions between A and B. The total distribution $\mathcal{P}[\mathbf{x}] = \mathcal{P}_A[\mathbf{x}] + \mathcal{P}_B[\mathbf{x}]$ is the sum of the (unnormalized) partial path distributions $\mathcal{P}_A[\mathbf{x}] \equiv \mathcal{P}[\mathbf{x}] h_A(x_0)$ and $\mathcal{P}_B[\mathbf{x}] \equiv \mathcal{P}[\mathbf{x}] h_B(x_0)$, consisting, respectively, of all paths that start in A and paths that start in B. Here, $h_{A,B}(x)$ are the indicator functions, which are unity when the configuration $x$ is in state $A(B)$, and zero otherwise. Note that we restrict all paths to start and end in one of the stable states.

In what follows, we will focus on applying kinetic constraints on each partial path ensemble separately, as they can be treated independently from each other, as demonstrated in *SI Appendix*:

$$\mathcal{P}_A^{MC}[\mathbf{x}] \propto e^{-\mu_A s_A[\mathbf{x}]} \mathcal{P}_A^0[\mathbf{x}], \quad [14]$$

$$\mathcal{P}_B^{MC}[\mathbf{x}] \propto e^{-\mu_B s_B[\mathbf{x}]} \mathcal{P}_B^0[\mathbf{x}], \quad [15]$$

where $s_{A,B}[\mathbf{x}]$ are now functions of the dynamical paths measuring the observable that is constrained for the partial ensembles A and B, respectively.

**Constraining Rate Constants Using MaxCal.** We now turn to constraining kinetic observables, including, in particular, rate constants. Suppose that we have unbiased simulations that we want to correct in order to match an experimental rate constant, $s_A^{exp} \equiv k_{AB}^{exp}$. First, we need to look at how the rate constant is defined in the path space as the time derivative of the correlation function $C(t) = \langle h_A(x_0) h_B(x_L) \rangle / \langle h_A(x_0) \rangle$

$$k_{AB} = \frac{dC(t)}{dt} = \frac{\langle h_A(x_0) \dot{h}_B(x_L) \rangle}{\langle h_A(x_0) \rangle}, \quad [16]$$

where the indicator functions $h_{A,B}(x)$ are unity when the frame is in state A and B, respectively. This expression thus computes the flux through entering the state B, provided that the trajectories started in A.

To link the flux-correlation function to the path ensembles and the MaxCal approach, we adopt the formalism of TIS (31, 35, 36), which, in turn, is based on the framework of TPS (28, 37, 38). Introducing a CV $\lambda(x)$ that can parameterize a hypersurface, or *interface*, in the configuration space, TIS defines a set of $n+1$ nonintersecting such interfaces, denoted by the parameters $\lambda_0 < \lambda_1 < \ldots < \lambda_n$. In this way, the rate constant can be written as (31)

$$k_{AB} = \phi_0 P_A(\lambda_B|\lambda_0), \quad [17]$$

where the first factor is the effective positive flux through the first interface $\lambda_0 = \lambda_A$, and the second factor is the crossing probability of interface $\lambda_B = \lambda_n$ for all trajectories shot from the first interface that came directly from state A in their backward integration. When evaluating the rate constant using the TIS framework, the first factor is accessible through a regular MD simulation, and the second factor through performing sampling of the interface path ensembles using the TIS algorithm (31) or as an approximation by the VIE-TPS algorithm (32). Both approaches yield an RPE, which is a way to reweight the interface ensembles into effectively the unbiased equilibrium path ensemble (30). $P_A(\lambda_B|\lambda_0)$ can, in principle, also be evaluated by using a very long MD simulation, although this is not very efficient for rare events. The crossing probability connected to each interface ensemble is expressed as a function of $\lambda$

$$P_A(\lambda|\lambda_0) = \int \mathcal{D}\mathbf{x}\mathcal{P}_A[\mathbf{x}]\theta(\lambda_{max}[\mathbf{x}] - \lambda), \quad [18]$$

where $\mathcal{P}_A[\mathbf{x}] = h_A(x_0)\mathcal{P}[\mathbf{x}]$ is the now-normalized (unbiased or reweighted) path ensemble distribution for paths starting in A, $\theta(x)$ is the Heaviside step function, and $\lambda_{max}[\mathbf{x}]$ returns the maximum value of $\lambda$ along the path. Here, we assumed that $\lambda$ is monotonically increasing with $i$.

Imposing the constraint $k_{AB} = k_{AB}^{exp}$ now leads to the Lagrange function

$$\mathcal{L} = -\int \mathcal{D}\mathbf{x}\mathcal{P}_A[\mathbf{x}] \ln \frac{\mathcal{P}_A[\mathbf{x}]}{\mathcal{P}_A^0[\mathbf{x}]} - \nu\left(\int \mathcal{D}\mathbf{x}\mathcal{P}_A[\mathbf{x}] - 1\right) - \mu_A\left(\int \mathcal{D}\mathbf{x}\mathcal{P}_A[\mathbf{x}]\theta(\lambda_{max}[\mathbf{x}] - \lambda_B) - k_{AB}^{exp}\right), \quad [19]$$

where we have left out the flux $\phi_0$ from the rate constant contribution for notational reasons. Following the same reasoning as before, we can optimize the Lagrange function giving rise to the posterior

$$\mathcal{P}_A^{MC}[\mathbf{x}] \propto e^{\mu_A\theta(\lambda_{max}[\mathbf{x}]-\lambda_B)}\mathcal{P}_A^0[\mathbf{x}], \quad [20]$$

and from the analog of Eq. **6**

$$k_{AB}^0 e^{\mu_A} = k_{AB}^{exp}, \quad [21]$$

we obtain the value of the Lagrange multiplier $\mu_A = \ln(k_{AB}^{exp}/k_{AB}^0)$. Note that this equation can easily be extended to the analog of Eq. **7**

$$k_{AB}^0 e^{\mu_A} = k_{AB}^{exp} + \mu_A\sigma_{k_{AB}}^2, \quad [22]$$

where $\sigma_{k_{AB}}$ signifies the level of confidence in the rate constant data. Just as for MaxEnt, one can turn the constraint condition into a restraint condition.

The reweighting procedure can be interpreted as a bias on only the reactive AB paths that make it to the final interface $\lambda_B$, such that the total flux of paths is obeying the kinetic rate constraint. However, this means that this reweighting will introduce a discontinuity in the path distribution, as a path that is nearly reaching B, but is recrossing back to A, is not reactive, and thus is not reweighted. Even though these recrossing paths themselves might be rare, such a discontinuity is undesirable. For an illustration, see *SI Appendix*, Fig S2.

We can make progress by realizing that Eq. **17** can be written as $k_{AB} = \phi_0 P_A(\lambda_B|\lambda_i)P_A(\lambda_i|\lambda_0)$ and that the rate constraint should apply to all values of $\lambda_i$. Choosing interfaces arbitrary close (i.e., large $n$) and enforcing the kinetics at all interfaces simultaneously eliminates the undesired discontinuities. In *SI Appendix*, we show that this yields the general solution

$$\mathcal{P}_A^{MC}[\mathbf{x}] \propto e^{f_A(\lambda_{max}[\mathbf{x}])}\mathcal{P}_A^0[\mathbf{x}], \quad [23]$$

with

$$f_A(\lambda_{max}[\mathbf{x}]) \equiv -\sum_{i=1}^{n} \mu_i\theta(\lambda_{max}[\mathbf{x}] - \lambda_i)P_A(\lambda_n|\lambda_i), \quad [24]$$

where the $P_A(\lambda_n|\lambda_i)$ is the MaxCal-corrected crossing probability. The interpretation is that the weight of each path in the posterior path ensemble is solely determined by $\lambda_{max}[\mathbf{x}]$, which, in turn, means that each path with the same $\lambda_{max}[\mathbf{x}]$ is weighed in the same way (*SI Appendix*). Indeed, for a specific $\lambda_j$

$$f_A(\lambda_j) = -\sum_{i=0}^{j} \mu_i P_A(\lambda_n|\lambda_i). \quad [25]$$

A similar function $f_B(\lambda_{min}[\mathbf{x}])$ follows for the path ensemble from B, based on the minimum value of $\lambda$ along the path.

The projection of the (normalized) partial path ensemble $\mathcal{P}_A^0[\mathbf{x}]$ yields the crossing probability $P_A^0(\lambda|\lambda_0)$ and the configurational density $\rho_A^0(\lambda)$, respectively,

$$P_A^0(\lambda|\lambda_0) = \int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}]\theta(\lambda_{max}[\mathbf{x}] - \lambda), \quad [26]$$

$$\rho_A^0(\lambda) \propto \int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}] \sum_{k=0}^{L[\mathbf{x}]} \delta(\lambda(x_k) - \lambda), \quad [27]$$

where the sum is over all frames of the trajectory $\mathbf{x}$. Using the MaxCal path reweighting for the configurational density yields

$$\rho_A^{MC}(\lambda) \propto \int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}]e^{f_A(\lambda_{max}[\mathbf{x}])} \sum_{k=0}^{L[\mathbf{x}]} \delta(\lambda(x_k) - \lambda). \quad [28]$$

For the crossing probability, the reweighting is a bit more subtle. In *SI Appendix*, we show that

$$P_A^{MC}(\lambda|\lambda_0) = \int_{\lambda_n}^{\lambda} R_A^0(\lambda|\lambda_0)e^{f_A(\lambda)}d\lambda, \quad [29]$$

where $R_A^0(\lambda|\lambda_0)$ is the "reaching" histogram of paths that just reach $\lambda$

$$R_A^0(\lambda|\lambda_0) = \int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}]\delta(\lambda_{max}[\mathbf{x}] - \lambda). \quad [30]$$

The configurational density and crossing probability for the partial path ensembles from B are done likewise.

**The MaxCal Bias Function $f_A(\lambda)$ Follows from MaxEnt for the Density.** MaxCal does not give an explicit solution for the function $f_A(\lambda)$, as it only concerns the final rate value, which is satisfied

as long as $f_A(\lambda_n)$ is set to the proper value. Indeed, a solution to the constraint equation will be correct for all functions $f_A$ under the condition that $f_A(\lambda_B)$ gives the correct kinetic constraint. This can also be deduced from Eq. **25**, where the solution to the Lagrange multipliers $\mu_j$ allow virtually all reasonably shaped functions $f_A(\lambda)$.

Therefore, it seems that we have not made progress, since $f_A(\lambda)$ is unknown. Here is where the configurational density $\rho^0(x)$ comes in. Applying the regular MaxEnt approach by setting $P^0(x) = \rho^0(x)$ and $s_1(x) = g(\lambda(x))$ ($M = 1$) in Eq. **5** yields

$$\rho^{ME}(x) \propto e^{-\mu g(x)} \rho^0(x), \qquad [31]$$

where $g(\lambda(x))$ is an a priori unknown function that imposes the constraint (taking the role of the $s$ in Eq. **2**). When projecting onto the CV $\lambda$, this expression reduces to

$$\rho^{ME}(\lambda) \propto e^{-\mu g(\lambda)} \rho^0(\lambda), \qquad [32]$$

where the constraint imposed is

$$\frac{\int d\lambda g(\lambda)\rho(\lambda)}{\int d\lambda \rho(\lambda)} = g^{exp}. \qquad [33]$$

Now, what is $g^{exp}$ if we constrain the rate constants $k_{AB}$ and $k_{BA}$? The obvious candidate is the ratio $k_{AB}/k_{BA} \equiv K_{eq}$, which is equal to the equilibrium constant $K_{eq} = \pi_B/\pi_A$. In fact, it turns out better to consider the equilibrium fraction $K = \pi_B/(\pi_A + \pi_B) = K_{eq}/(1 + K_{eq})$. Here, we use $\pi_{A,B}$ to denote the total equilibrium population in A and B, to avoid confusion with $\rho_{A,B}(\lambda)$. Thus, the question is which function $g(\lambda)$ would obey

$$\frac{\int d\lambda g(\lambda)\rho(\lambda)}{\int d\lambda \rho(\lambda)} = K. \qquad [34]$$

In *SI Appendix*, we show that a natural solution for $g(\lambda)$ is the (projected) committor $p_B(\lambda)$, as the points that commit to B are both determining the committor (39) and the equilibrium fraction $K$. We remind the reader that the committor $p_B(x)$ is the probability for a MD trajectory initialized from configuration $x$ with randomized velocities to reach state B before A (28, 29). In the field of protein folding, this is also sometimes referred to as p-fold (40), while in the chemical literature, this is known as Onsager's splitting probability (41). Just as free energies, committors can be projected on CVs, leading to the concept of committor distributions, profiles, or landscapes $p_B(\lambda)$ (39, 42).

The reweighted MaxEnt densities, given in Eq. **32**, then become

$$\rho_A(\lambda) = \rho_A^0(\lambda) e^{\mu_A p_B(\lambda)} \qquad [35]$$

$$\rho_B(\lambda) = \rho_B^0(\lambda) e^{-\mu_B p_B(\lambda)} e^{\mu_A}, \qquad [36]$$

where the latter equation has a negative exponent and a shift, and we considered two different Lagrange multipliers, one for each direction AB and BA. To solve for $p_B$, we note that $p_B(\lambda) = \rho_B(\lambda)/(\rho_A(\lambda) + \rho_B(\lambda))$ and substituting the ME densities gives

$$p_B(\lambda) = \frac{\rho_B^0(\lambda)}{\rho_A^0(\lambda) e^{-\mu_A} e^{(\mu_A+\mu_B)p_B(\lambda)} + \rho_B^0(\lambda)}. \qquad [37]$$

This self-consistent equation can be solved numerically, given $\rho_B^0(\lambda), \rho_B^0(\lambda)$, and the values of $\mu_A$ and $\mu_B$. These last quantities follow from the MaxCal constraint that the rate constants need to be correct. That is,

$$e^{\mu_A} = \frac{k_{AB}^{exp}}{k_{AB}^0} \qquad e^{\mu_B} = \frac{k_{BA}^{exp}}{k_{BA}^0}, \qquad [38]$$

so that the ratio of these equations is

$$e^{\mu_A - \mu_B} = K_{eq}^{exp}/K_{eq}^0. \qquad [39]$$

Note that these last two equations can be extended to account for the experimental error (Eq. **22**). While we use MaxEnt here for clarifying purposes, we note that, in principle, we can also add static constraints in the MaxCal formalism.

We illustrate this approach for a toy example density. By taking simple exponential forms for the density (Fig. 1*A*), we plot the initial committor in Fig. 1*B*. We can then apply the self-consistent solution to the committor using $\mu_A = 1$ and $\mu_B = 1.5$ (Fig. 1*B*) and reweight the densities (Fig. 1*C*).

To obtain $f_A(\lambda)$ from $g(\lambda)$, we use the fact that the MaxCal-corrected RPE configurational density and the MaxEnt-corrected configurational density should be identical, i.e.,

$$\rho_A^{MC}(\lambda) = e^{-\mu g(\lambda)} \rho_A^0(\lambda), \qquad [40]$$

or

$$\int \mathcal{D}\mathbf{x} \mathcal{P}_A^0[\mathbf{x}] e^{f_A(\lambda_{max}[\mathbf{x}])} \sum_{k=0}^{L} \delta(\lambda(x_k) - \lambda) = \\ = e^{-\mu g(\lambda)} \int \mathcal{D}\mathbf{x} \mathcal{P}_A^0[\mathbf{x}] \sum_{k=0}^{L} \delta(\lambda(x_k) - \lambda). \qquad [41]$$

In practice, this Volterra equation of the first kind should be solved numerically (*SI Appendix*). A similar equation needs to be solved for the partial ensemble of paths starting in B, in order to obtain $f_B$.

**Optimal Path Distributions by Varying the CV.** The final perturbed distributions will be dependent on the choice of the CV $\lambda$. In principle, it is possible to vary the CV and maximize the entropy

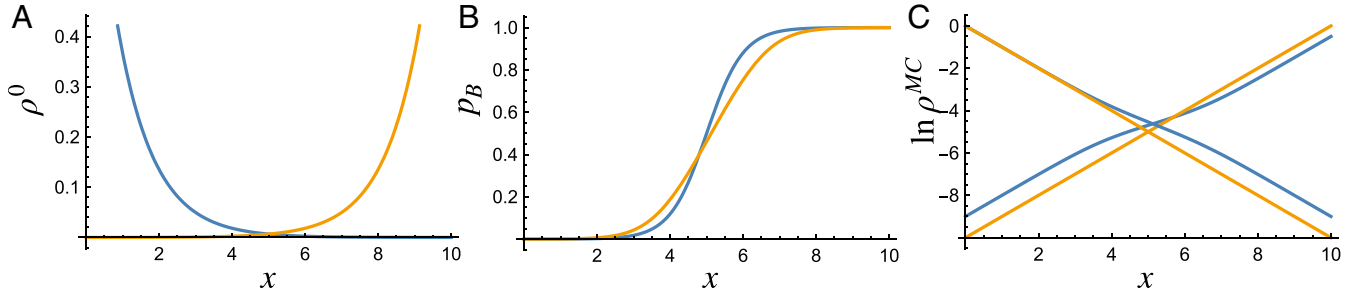


**Fig. 1.** A simple example of the approach of reweighting densities described in this work. (*A*) Initial densities $\rho_A^0(\lambda)$ (blue) and $\rho_B^0(\lambda)$ (orange). (*B*) Initial committor $p_B^0(\lambda)$ (blue). Solution of the self-consistent equation Eq. **37** (orange). (*C*) Reweighted densities (blue) compared to initial densities (orange).

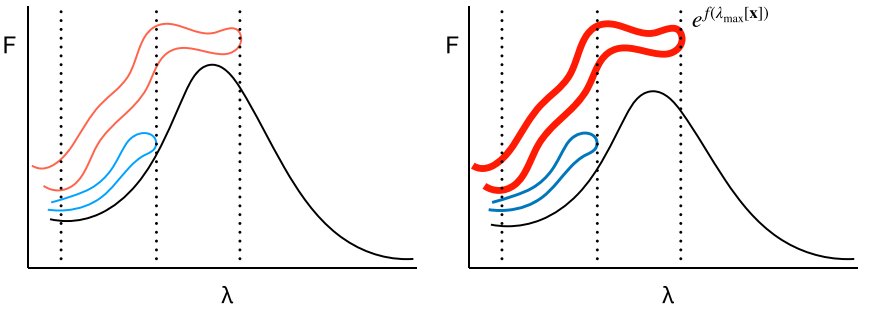


**Fig. 2.** Illustration of the method of reweighting paths described in this work. (*Left*) A free energy barrier $F(\lambda)$ is shown in black. The blue path has a high weight, while the red path has a low one, as it has to travel further up the barrier. The maximum $\lambda$-values are indicated by dotted vertical lines. (*Right*) After the reweighting, the red path is relatively more abundant in the ensemble, as indicated by the thicker curve. The resulting free energy barrier is lowered in line with the kinetic constraints.

and caliber as function of the CV. The most optimal CV is then the one that leads to the least perturbed path distribution.

Inserting the optimized MaxCal distributions $\mathcal{P}_A^{MC}[\mathbf{x}] = C_A^{-1}\mathcal{P}_A^0[\mathbf{x}]\exp[f_A(\lambda_{max}[\mathbf{x}])]$ and $\mathcal{P}_B^{MC}[\mathbf{x}] = C_B^{-1}\mathcal{P}_B^0[\mathbf{x}]\exp[f_B(\lambda_{min}[\mathbf{x}])]$, with $C_A, C_B$ appropriate normalization constants, into the expression for the caliber of the distributions and using the definition of the "reaching histograms" $R_A^0(\lambda|\lambda_0), R_B^0(\lambda|\lambda_n)$, yields

$$S[\mathcal{P}_A\|\mathcal{P}_A^0] = -\frac{1}{C_A}\int d\lambda R_A^0(\lambda|\lambda_0)e^{f_A(\lambda)}(f_A(\lambda) - \ln C_A)$$
$$S[\mathcal{P}_B\|\mathcal{P}_B^0] = -\frac{1}{C_B}\int d\lambda R_B^0(\lambda|\lambda_n)e^{f_B(\lambda)}(f_B(\lambda) - \ln C_B), \quad \textbf{[42]}$$

where the normalization $C_A = \int d\lambda R_A^0(\lambda|\lambda_0)e^{f_A(\lambda)}$ is now also expressed using the reaching histograms. Note that we have assumed all subdistributions $\mathcal{P}_A^0, \mathcal{P}_A, \mathcal{P}_B, \mathcal{P}_B^0$ to be normalized. However, when computing the total entropy, we need to use the normalized total path distributions $\mathcal{P}$ and $\mathcal{P}^0$. It is possible to express the caliber for the full distributions in terms of $S[\mathcal{P}_A\|\mathcal{P}_A^0]$ and $S[\mathcal{P}_B\|\mathcal{P}_B^0]$ as

$$S[\mathcal{P}\|\mathcal{P}^0] = \alpha S[\mathcal{P}_A\|\mathcal{P}_A^0] + (1-\alpha)S[\mathcal{P}_B\|\mathcal{P}_B^0] + +\alpha\ln\frac{\alpha}{\alpha_0} + (1-\alpha)\ln\frac{1-\alpha}{1-\alpha_0}, \quad \textbf{[43]}$$

with $\alpha = C_A/(C_A + C_B)$, and $\alpha_0 = C_A^0/(C_A^0 + C_B^0)$. The last two terms provide a kind of penalty for how much the partial ensembles differ in their respective weight. For a symmetric potential, identical sampling, and a symmetric bias, $\alpha = 1/2$ and these terms vanish.

**Generalizing the Approach.** When deriving the $g(\lambda)$ function, we use $\lambda$ as a CV. We can generalize the approach and look for the $g(x)$ as a function of any configuration $x$. In analogy with Eq. **35** and Eq. **36**, the reweighted MaxEnt densities are given by:

$$\rho_A(x) = \rho_A^0(x)e^{\mu_A p_B(x)}$$
$$\rho_B(x) = \rho_B^0(x)e^{-\mu_B p_B(x)}e^{\mu_A}. \quad \textbf{[44]}$$

To solve for $p_B$, we use again the definition $p_B(x) = \rho_B(x)/(\rho_A(x) + \rho_B(x))$ and substitute the MaxEnt densities,

$$p_B(x) = \frac{\rho_B^0(x)}{\rho_A^0(x)e^{-\mu_A}e^{(\mu_A+\mu_B)p_B(x)} + \rho_B^0(x)}. \quad \textbf{[45]}$$

Again, this self-consistent equation needs to be solved numerically, given $\rho_A^0(x), \rho_B^0(x)$, and the values of $\mu_A$ and $\mu_B$.

The $f_A(x)$ function then follows from identifying the MaxCal-corrected RPE configurational density with the MaxEnt-corrected configurational density

$$\rho_A^{MC}(x) = e^{-\mu g(x)}\rho_A^0(x), \quad \textbf{[46]}$$

or, setting $g(x) = p_B(x)$,

$$\int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}]e^{f_A(p_{B,max}[\mathbf{x}])}\sum_{k=0}^{L}\delta(x_k - x) = = e^{-\mu p_B(x)}\int \mathcal{D}\mathbf{x}\mathcal{P}_A^0[\mathbf{x}]\sum_{k=0}^{L}\delta(x_k - x), \quad \textbf{[47]}$$

where $p_{B,max}[\mathbf{x}]$ is the maximum value of the committor along the path $\mathbf{x}$. In practice, this equation should again be solved numerically.

This approach is consistent with the idea that $p_B(x)$ is the most optimal RC (29, 42).

**Interpretation of the Method.** The method that we have described takes as input an unbiased ensemble of paths and reweights each

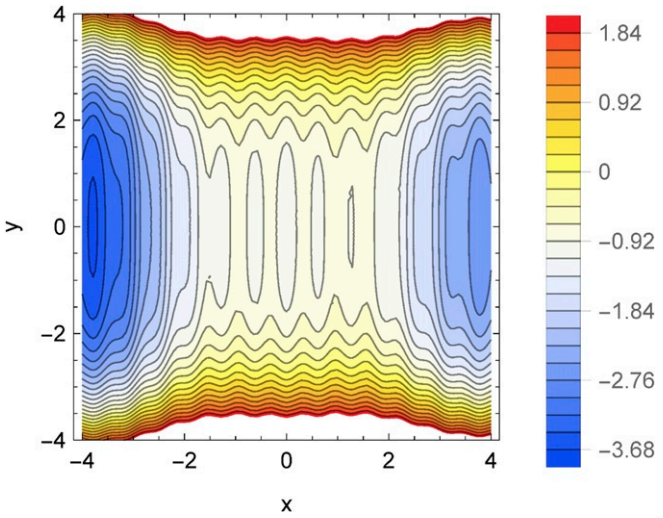


**Fig. 3.** Representation of a 2D toy potential (32) used to illustrate the application of the reweighting method described in this work. Energies are in units of $k_BT$. The two states, A and B, are separated by an energy barrier along the *x* axis. Oscillations are added to show better resolution of the projections.

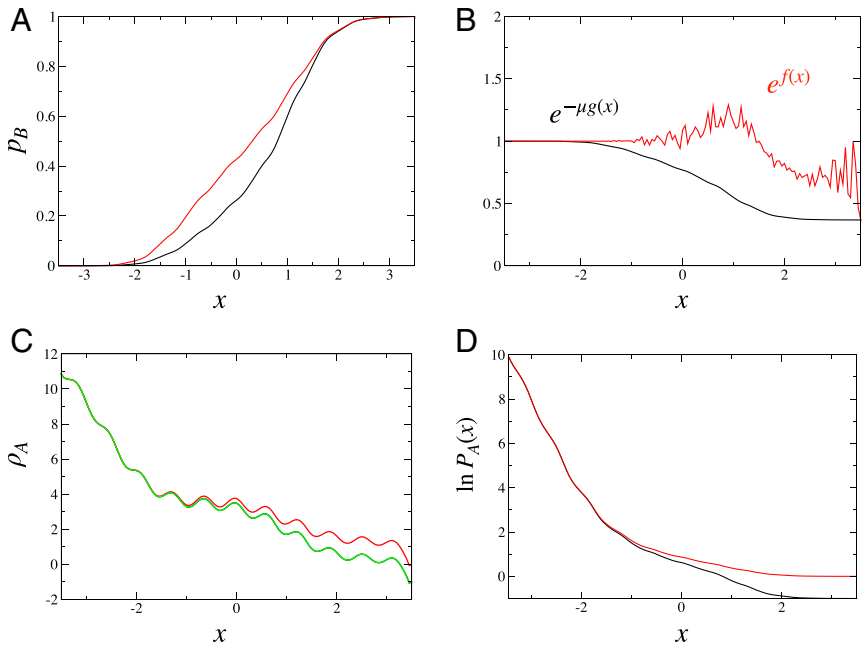


**Fig. 4.** Application of the reweighting method to the 2D potential of Fig. 3*A* committor $p_B^0(\lambda)$ function (red) and solution of the self-consistent Eq. **37** (black) for the explicit simulation. (*B*) Original weight function $e^{-\mu g(\lambda)}$ (black) and back-iterated function $e^{f_A(\lambda)}$ (red). (*C*) Logarithm of the configurational densities with the original in red, the reweighted with the *g* function (green), and the RPE-corrected with $e^{f_A(\lambda)}$ in black (not visible, behind green). (*D*) Logarithm of the crossing probability with the original in red and the RPE-corrected with $e^{f_A(\lambda)}$ in black.



trajectory in the ensemble according to how far it progresses along a predefined CV (Fig. 2). This includes the paths that cross the barrier and reach the other state, so the rate constants are automatically constrained to the correct value (via the functions $f_{A,B}(\lambda)$). The more involved part of the framework is to also ensure that the thermodynamic properties are correct, in particular, the equilibrium constant. This requires a specific bias function $g(\lambda)$ based on the committor function, which produces the least perturbed path ensemble, while still obeying the constraints. So imposing $g(\lambda)$ can be viewed as responsible for constraining equilibrium conditions, whereas $f_{A,B}(\lambda)$ takes care of the dynamical corrections. The interpretation of the reweighting procedure is that trajectories are artificially made more (or less) probable in the path ensembles, analogous of changing the weight of each conformation in the Boltzmann distribution, using the MaxEnt approach. Note that the method is enslaved to the original dynamics: The distribution of initial conditions for the trajectories is altered via the reweighting procedure, but the trajectories themselves do not change. This is analogous (but not identical) to how microcanonical trajectories can be reweighted to give a canonicaly distributed path ensemble (see, e.g., ref. 43). The altered path ensemble can lead to improved mechanistic insight by putting more emphasis to certain (transition) states and routes.

Of course, ideally, one would like to know how to alter the underlying force field, in order to achieve the same corrections. This could be the subject of future research.

## Results and Discussion

In this section, we illustrate the approach on a toy model as well as all-atom MD simulations of a folding/unfolding transition. In *SI Appendix*, we present further illustrations using toy models and a peptide unbinding transition.

**Two-Dimensional Toy Models.** We first investigated a two-dimensional (2D) potential (Fig. 3), which was recently studied using the VIE-TPS method (32). For details about the potential and the method, we refer to ref. 32. Setting the (reciprocal) temperature $\beta = 1/k_B T = 3$, we performed $10^7$ trial shots, where paths of flexible length were generated by using Metropolis

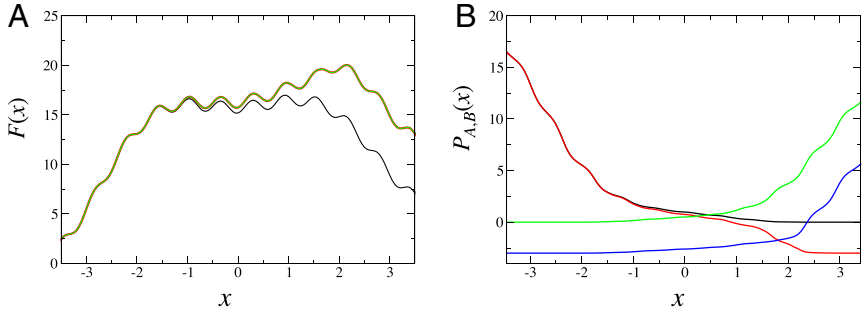


**Fig. 5.** Analysis of the 2D potential of Fig. 3 by increasing the bias. (*A*) The free energy for a tilt with $\mu_A = -3$, $\mu_B = 3$: original (black), reweighted with *g* function (green), and RPE-corrected with $e^{f(\lambda)}$ (red) (behind green). (*B*) Logarithm of the crossing probability: original in black/green, RPE-corrected with $e^{f(\lambda)}$ in red/blue.

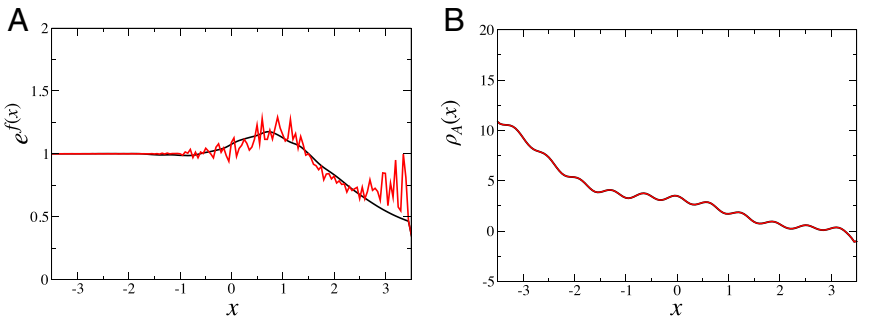


**Fig. 6.** Analysis of the 2D potential of Fig. 3 by parametrizing $e^{f_A(\lambda)}$ with a functional form. (*A*) Back-iterated function $f_A(\lambda)$ (red) and the fit (black curve). (*B*) The corresponding densities are identical.

Monte Carlo dynamics, on average roughly 1,000 frames long. Applying the VIE-TPS method on this potential gave the two partial path ensembles $\mathcal{P}_A^0$ and $\mathcal{P}_B^0$. We then applied our MaxCal approach, reweighting with $\mu_A = -1$ and $\mu_B = 0$, which corresponds to the lowering of the rate $k_{AB}$ by a factor $e$. The results are shown in Fig. 4. Fig. 4*A* shows the committor based on the original data (red curve), as well as the self-consistent solution to the committor (black curve). Fig. 4*B* gives the solution to Eq. **41** using back substitution (*SI Appendix*). The original weight $e^{-\mu g(\lambda)} = e^{-\mu p_B(\lambda)}$ is shown in black and the back iteration in red. Note that the red curve oscillates, due to numerical inaccuracies.

Next, we show the reweighted densities $\rho_A$ in Fig. 4*C*. The original density is shown in red, and the reweighted with the $e^{-\mu g(\lambda)}$ is shown in green. The RPE-corrected density should be identical and is shown in black (not visible, as it is, indeed, exactly the same as the green curve). Finally, we show the logarithm of the crossing probability in Fig. 4*D*, with the original curve in red and the RPE-corrected one in black. Indeed, the final log-rate was lowered with one, as imposed.

For positive bias $\mu_A > 0$, this treatment is also possible, but can result in some negative weights $e^{f_A(\lambda)}$ for $\lambda$ just below $\lambda_n$. We ameliorated this by putting the weights to zero for these cases, which precludes a precise solution. Still, the reweighted densities are almost correct. In any case, the values of $f_A(\lambda)$ do not affect the densities strongly at these values.

In Fig. 5, the bias is increased to $\mu_A = -3$ and $\mu_B = 3$, and the crossing probabilities now show a dramatic change. Both forward (AB) and backward (BA) curves are shown in Fig. 5*B*. The crossing-probability curves are shifted to match the minimum values. Note that the BA curve (blue) is thus shifted by $6k_BT$, as required by detailed balance. The free energy (Fig. 5*A*) shows a strong shift of the transition state toward the final state.

The oscillations occurring in Fig. 4*B* are related to numerical inaccuracies during the backward-substitution solution for $f_A(\lambda)$. These oscillations, indeed, decrease with the amount of path ensemble data that is available. In the limit of infinite amounts of data, this curve should be smooth. It should therefore be possible to parametrize $f_A(\lambda)$ with a functional form, e.g., with

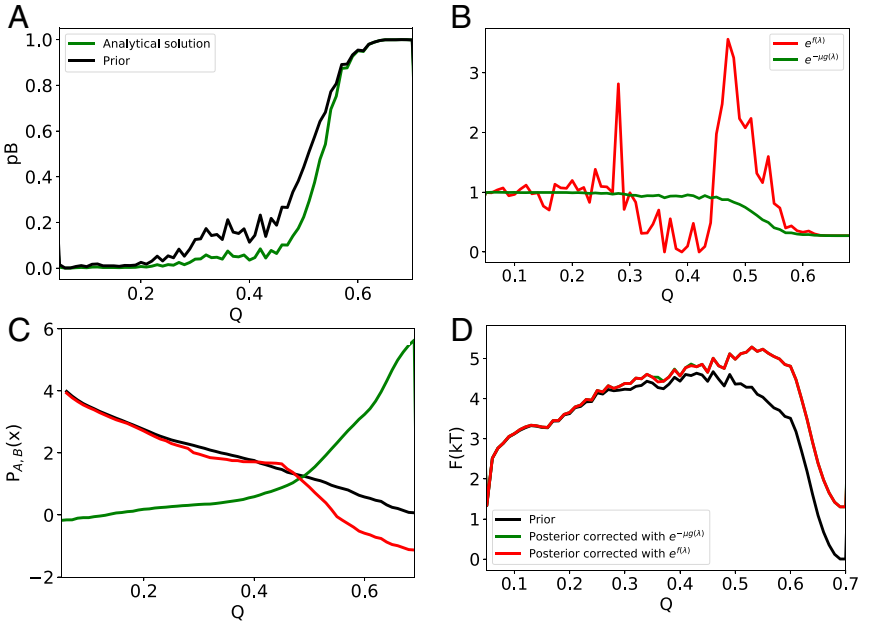


**Fig. 7.** Simulations of folding and unfolding of chignolin. (*A*) Committor $p_B^0(\lambda)$ function (black) and solution of the self-consistent Eq. **37** (green) for the explicit simulation using $\mu_A$ = −1.3. (*B*) Original weight function $e^{-\mu g(\lambda)}$ (green) and back-iterated function $e^{f(\lambda)}$ (red). (*C*) Logarithm of the crossing-probability histogram of the original (black) and RPE-corrected with $e^{f(\lambda)}$ (red). (*D*) Free energies as a function of the fraction of native contacts *Q*, original (black), reweighted with *g* function (green), and RPE-corrected with $e^{f(\lambda)}$ (red) (green not visible, behind red).

$$f_A(\lambda) = g(\lambda) + \sum_i^{n_p} a_{i,0} \exp(-a_{i,1}(x - a_{i,2})^2),$$

with $n_p$ as the number of Gaussian functions, and optimize the $a_{i,j}$ coefficients. The result for $n_p = 1$ is shown in Fig. 6. This opens up the possibility to optimize parameterizations of $f_A(\lambda)$ using advanced regression procedures, and even machine learning.

In *SI Appendix*, we further explore 2D potentials in order to study the influence of the choice of CV.

**Folding and Unfolding of Chignolin.** The kinetics of fast folding $\beta$-hairpins have been studied by temperature-jump spectroscopy (44), reporting microsecond-timescale folding. Chignolin is a two-state, $\beta$-hairpin, mini-protein that folds in the $\mu$s timescale (45). Despite its simple fold (Protein Data Bank ID code 2RVD), molecular simulation fails to capture the experimentally determined melting temperature of 341 K (45–47). Here, we perform our kinetic analysis on an equilibrium MD trajectory of chignolin at 341 K (45). While at this temperature, experiments suggest that the folding and unfolding rates should be the same, simulations report a folding rate of $k_f = 1.667\ \mu s^{-1}$ and an unfolding rate of $k_u = 0.455\ \mu s^{-1}$, respectively. The corresponding enhanced stability of the folded state is likely to arise from inaccuracies in the force field used in the MD simulations. In the absence of an experimental folding rate for chignolin, but in light of 1) knowledge that the barrier heights should not exceed $4.5 k_B T$ (45), and 2) that folding and unfolding rates should be the same at the melting temperature, we perform our kinetic analysis by constraining only the folding rate $k_f^{exp} = 0.455\ \mu s^{-1}$. This leads to posterior kinetic ensemble of (un)folding pathways exhibiting new kinetics and thermodynamics, as shown in Fig. 7. We use the fraction of native contacts $Q$ as the CV for the order parameter $\lambda$. In the remainder of this section, states A and B refer to the unfolded ($Q < 0.05$) and folded state ($Q > 0.7$), respectively.

The posterior committor distribution becomes steeper and gets shifted to higher $Q$ values (Fig. 7*A*). In particular, the isocommittor surfaces $p_B = 0.5$ shift by 10%, from $Q = 0.5$ to $Q = 0.55$. This is in agreement with an effect of lowering the temperature to the increase of the nativeness of the transition state (48), as well as the knowledge of native-like transition states in protein zippers (44). Fig. 7*B* illustrates the solution to the Volterra equation Eq. **41** by back substitution. The original weight $e^{-\mu g(\lambda)} = e^{-\mu p_B(\lambda)}$ is shown in green, and the back iteration gives the MaxCal bias on the path weights $e^{f(\lambda)}$. Applying this bias to reweight the path ensemble results in posterior crossing probabilities (Fig. 7*C*), where the folding conditional probability becomes steeper, indicating slower folding kinetics. At the same time, the final shift in the folding crossing probabilities is $\exp(-1.3)$, giving, indeed, rise to the imposed folding rate of $k_f^{exp} = 0.455\ \mu s^{-1}$. Finally, Fig. 7*D* illustrates the effect of the kinetic constraint on the free energy. The free energy difference between folded and unfolded states becomes zero, as expected by the constraint, and amends the current force field's inaccuracy in predicting the simulated temperature (341 K) as the melting temperature. Moreover, the free energy barrier becomes more asymmetric, shifting toward a more native-like conformations (TS′), as indicated also in Fig. 7*A*. The interactions in structure TS′ can inform mutation experiments in order to regulate the folding kinetics.

In Fig. 8, we assess how the kinetic correction alters the free energy and committor landscapes as a function of fraction of native contacts and solvent-accessible surface. The kinetic constraint increases the population of the misfolded configurations ($0.05 < Q < 0.4$) state (Fig. 8*B*), as well as widens the distribution of the solvent-accessible area (SAS) of the protein. Moreover, Fig. 8 *C* and *D* show that the kinetic constraint

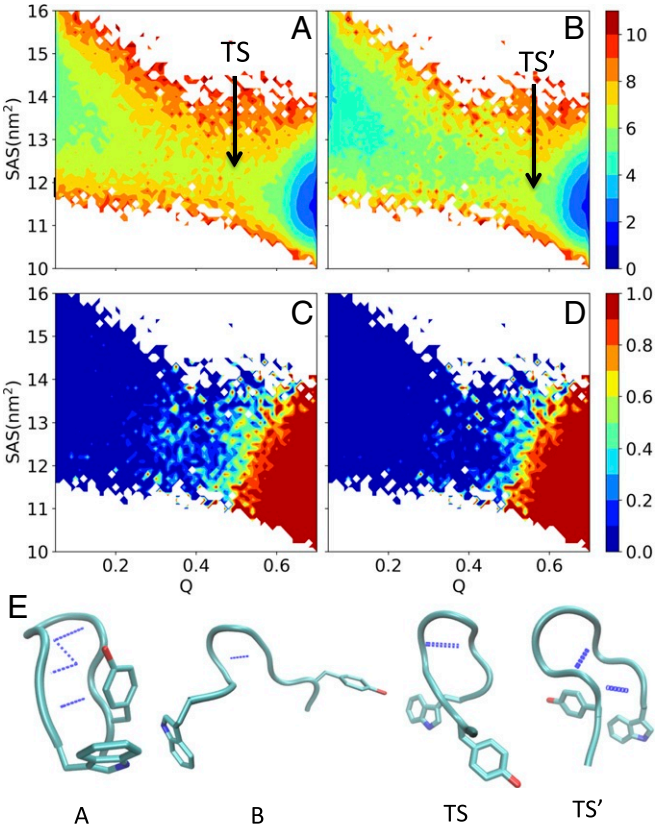


**Fig. 8.** (*A* and *B*) Free energy as a function of the fraction of native contacts, $Q$, and the solvent-accessible surface, SAS, of the prior (*A*) and the posterior (*B*). *C* and *D* highlight the respective committor landscapes of the prior (*C*) and the posterior (*D*). (*E*) The structure panel refers to the folded (A), unfolded (B), prior transition state (TS), and posterior transition state (TS′).

shifts the transition state—i.e., the 0.5 isocommittor surface—to a higher fraction of native contacts $Q \approx 0.55$ and a slightly lower SAS of 12.2 $nm^2$, indicating a more packed structure. As illustrated in Fig. 8*E*, the prior transition-state configuration TS forms one backbone hydrogen bond and has the residues Y2 and W9, crucial for hydrophobic collapse, facing away from each other. On the contrary, the posterior transition-state configuration TS′ is more native-like and shown to form more native backbone hydrogen bonds, while forming contacts at the key hydrophobic collapse residues Y2 and W9.

## Conclusions

Molecular simulations can be used to accurately characterize protein structural ensembles and their corresponding thermodynamics (4). Yet, as the functions of proteins often depend on the transition rates between their different states, there is still a need for developing accurate methods for characterizing the kinetics of these molecules.

To address this challenge, in this work we have developed a framework to determine kinetic ensembles from experimental information. This framework combines MaxCal and MaxEnt concepts in order to match experimentally determined kinetic rate constants with MD simulations. The matching is done by biasing the paths in the unbiased RPE based on how far they are progressing along a chosen CV. In this reweighting, both the rate constant as well as the equilibrium free energy are constrained. In doing so, we are able to ameliorate dynamical profiles, such as conditional probabilities, committor functions, and transition states, as well as the long time kinetics and the equilibrium thermodynamics. In addition, the method can

account for uncertainty in the data by imposing restraints rather than constraints.

To illustrate possible applications of this method, we showed that matching the rate constants of folding of chignolin to a simulated structural ensemble yields accurate melting temperature and a more native-like transition-state ensemble.

We anticipate that this method will extend the applicability of MD simulations as a kinetic tool in structural biology—e.g., for accurate mechanistic and reaction coordinate analysis. Furthermore, the approach can be extended to amend imperfections in current atomistic force fields to reproduce the kinetics and thermodynamic observables. Such a possible method would require computing the derivative of the kinetic rate constant in path ensembles. We leave this for future research.

We finally note that, in principle, the method is general and could be applied to a wide range of problems amenable to molecular simulations.

## Materials and Methods

Simulation data were generated by using home-written code or from previous work (32, 49). Reweighting of path ensembles was done by using home-written analysis scripts.

**Data Availability.** All study data are included in the article and *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank Tristan Bereau, Carlo Camilloni, Christoph Dellago, and Pieter Rein ten Wolde for carefully commenting on the manuscript. Z.F.B. was supported by the Federation of European Biochemical Societies through a Long Term Fellowship.

1. B. Alberts *et al.*, (*Molecular Biology of the Cell* Garland Science, New York, NY, ed. 4, 2002).
2. A. T. Brünger *et al.*, Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 905–921 (1998).
3. A. Mittermaier, L. E. Kay, New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228 (2006).
4. M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
5. A. Cavalli, C. Camilloni, M. Vendruscolo, Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **138**, 094112 (2013).
6. W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Combining experiments and simulations using the maximum entropy principle. *PLoS Comput. Biol.* **10**, e1003406 (2014).
7. M. Bonomi, M. Vendruscolo, Determination of protein structural ensembles using cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **56**, 37–45 (2019).
8. T. Löhr, K. Kohlhoff, G. T. Heller, C. Camilloni, M. Vendruscolo, A kinetic ensemble of the Alzheimer's Aβ peptide. bioRxiv, https://doi.org/10.1101/2020.05.07.082818 (8 May 2020).
9. A. Cesari, S. Reißer, G. Bussi, Using the maximum entropy principle to combine simulations and solution experiments. *Computation* **6**, 15 (2018).
10. J. W. Pitera, J. D. Chodera, On the use of experimental observations to bias simulated ensembles. *J. Chem. Theor. Comput.* **8**, 3445–3451 (2012).
11. G. Hummer, J. Köfinger, Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
12. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2015).
13. S. Olsson, H. Wu, F. Paul, C. Clementi, F. Noé, Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci.U.S.A.* **114**, 8265–8270 (2017).
14. G. T. Heller *et al.*, Sequence specificity in the entropy-driven binding of a small molecule and a disordered peptide. *J. Mol. Biol.* **429**, 2772–2779 (2017).
15. G. T. Heller *et al.*, Small molecule sequestration of amyloid-β as a drug discovery strategy for Alzheimers disease. *Sci. Adv.* **6**, eabb5924 (2020).
16. A. N. Borkar *et al.*, Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7171–7176 (2016).
17. S. Orioli, A. H. Larsen, S. Bottaro, K. Lindorff-Larsen, How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **170**, 123–176 (2020).
18. A. Cesari, A. Gil-Ley, G. Bussi, Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theor. Comput.* **12**, 6192–6200 (2016).
19. E. T. Jaynes, The minimum entropy production principle. *Annu. Rev. Phys. Chem.* **31**, 579–601 (1980).
20. C. Monthus, Non-equilibrium steady states: Maximization of the Shannon entropy associated with the distribution of dynamical trajectories in the presence of constraints. *J. Stat. Mech. Theor. Exp.* **2011**, P03008 (2011).
21. P. Tiwary, B. J. Berne, Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2839–2844 (2016).
22. S. Pressé, K. Ghosh, J. Lee, K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **85**, 1115–1141 (2013).
23. P. D. Dixit, K. A. Dill, Caliber corrected Markov modeling (C2M2): Correcting equilibrium Markov models. *J. Chem. Theor. Comput.* **14**, 1111–1119 (2018).
24. A. A. Filyukov, V. Y. Karpov, Method of the most probable path of evolution in the theory of stationary irreversible processes. *J. Eng. Phys.* **13**, 416–419 (1967).
25. R. Capelli, G. Tiana, C. Camilloni, An implementation of the maximum-caliber principle by replica-averaged time-resolved restrained simulations. *J. Chem. Phys.* **148**, 184114 (2018).
26. M. Bause, T. Wittenstein, K. Kremer, T. Bereau, Microscopic reweighting for nonequilibrium steady-state dynamics. *Phys. Rev. E* **100**, 60103 (2019).
27. E. T. Jaynes, "*Macroscopic prediction*" in *Complex Systems—Operational Approaches in Neurobiology, Physics, and Computers*, H. Haken, Ed. (Springer Series in Synergetics, Springer, Berlin, Germany, 1985), vol. 31, pp. 254–269.
28. P. G. Bolhuis, D. Chandler, C. Dellago, P. Geissler, Transition path sampling: Throwing ropes over mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53**, 291–318 (2002).
29. E. W., W. Ren, E. Vanden-Eijnden, Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **413**, 242–247 (2005).
30. J. Rogal, W. Lechner, J. Juraszek, B. Ensing, P. G. Bolhuis, The reweighted path ensemble. *J. Chem. Phys.* **133**, 174109 (2010).
31. T. S. van Erp, D. Moroni, P. G. Bolhuis, A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **118**, 7762 (2003).
32. Z. F. Brotzakis, P. G. Bolhuis, Approximating free energy and committor landscapes in standard transition path sampling using virtual interface exchange. *J. Chem. Phys.* **151**, 174111 (2019).
33. C. Dellago, P. G. Bolhuis, F. S. Csajka, D. Chandler, Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108**, 1964 (1998).
34. D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, UK, 1987).
35. P. G. Bolhuis, C. Dellago, Trajectory based molecular rare event simulations. *Rev. Comput. Chem.* **27**, 1–105 (2010).
36. T. S. van Erp, Dynamical rare event simulation techniques for equilibrium and nonequilibrium systems. *Adv. Chem. Phys.* **151**, 27–60 (2012).
37. C. Dellago, P. G. Bolhuis, F. S. Csajka, D. Chandler, Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108**, 1964–1977 (1998).
38. C. Dellago, P. G. Bolhuis, Transition path sampling and other advanced simulation techniques for rare events. *Adv. Polym. Sci.* **221**, 167–233 (2009).
39. P. G. Bolhuis, W. Lechner, On the relation between projections of the reweighted path ensemble. *J. Stat. Phys.* **145**, 841–859 (2011).
40. R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, E. S. Shakhnovich, On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350 (1998).
41. L. Onsager, Initial recombination of ions. *Phys. Rev.* **54**, 554–557 (1938).
42. E. W., E. Vanden-Eijnden, Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* **61**, 391–420 (2010).
43. P. G. Bolhuis, G. Csányi, Nested transition path sampling. *Phys. Rev. Lett.* **120**, 250601 (2018).
44. C. D. Snow *et al.*, Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4077–4082 (2004).
45. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
46. S. Honda *et al.*, Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* **130**, 15327–15331 (2008).
47. P. Robustelli, S. Piana, D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4758–E4766 (2018).
48. J. Karanicolas, C. L. Brooks, The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11**, 2351–2361 (2002).
49. Z. F. Brotzakis, P. G. Bolhuis, A one-way shooting algorithm for transition path sampling of asymmetric barriers. *J. Chem. Phys.* **145**, 164112 (2016).