Analyzing complex reaction mechanisms using path sampling

Titus S. van Erp, Mahmoud Moqadam, Enrico Riccardi, and Anders Lervik
Department of Chemistry, Faculty of Natural Sciences and Technology, NTNU,
Norwegian University of Science and Technology, 7941 Trondheim, Norway

We introduce an approach to analyze collective variables regarding their predictive power for a reaction. The method is based on already available path sampling data produced by for instance transition interface sampling or forward flux sampling which are path sampling methods used for efficient computation of reaction rates. By a search in collective variable space a measure of predictiveness can be optimized and, in addition, the number of collective variables can be reduced using projection operations which keep this measure invariant. The approach allows testing hypotheses on the reaction mechanism, but could in principle also be used to construct the phase space committor surfaces without the need of additional trajectory sampling. The procedure is illustrated for a one-dimensional double well potential, a theoretical model for an ion-transfer reaction in which the solvent structure can lower the barrier, and an Ab Initio molecular dynamics study of water auto-ionization. The analysis technique enhances the quantitative interpretation of path sampling data which can provide clues on how chemical reactions can be steered in desired directions.

PACS numbers:

I. INTRODUCTION

Systematic approaches to analyze reaction mechanisms in terms of descriptive reaction coordinates have been focused on committor analysis[1–7]. The committor function tells for each phase point or configuration point what the probability is that a dynamical trajectory launched from that point will end up in the product state rather than the reactant state. Points having the same committor value form iso-committor surfaces. A reaction can then be described by a Markov process in which the system moves from one iso-committor surface to another one. The committor can, hence, be interpreted as a progress coordinate and a growing number of researchers in the field believe it should be viewed as the true reaction coordinate. Literature is, however, not always consistent whether the phase space or configuration space committor should be considered. In principle only the phase space committor gives the full mechanistic information[8,9], but the configurational committor fits better into the original concept of reaction coordinate which is traditionally a purely geometric function[10]. Moreover, the determination of the committor surfaces is computationally intensive since it requires the release of many trajectories from each individual phase- or configuration-point. Although there are systematic approaches based on genetic neural networks[11] and Bayesian techniques[4,5] which can reduce the computational burden, accurate determination of the committor values is difficult especially for low values. Hence, computational studies investigating the committor generally tend to focus on the surface with committor value 1/2, the separatrix. Therefore, a systematic analysis on the required conditions, of how and when the system can reach the separatrix, has received much less attention.

On one hand, the beauty of the phase-space and configurational-space committors is that they are mathematically well-defined and do not require any preassumptions or chemical intuition. On the other hand, however, this also implies a disadvantage. If a divine power would give us the full phase space committor (from which the configurational committor can be obtained by velocity averaging) as an exact nonlinear function of all atom positions and velocities, it will not directly give us a lot of insight. We would probably not be able to make any sense out of this multi-dimensional nonlinear function unless we could simplify it, if necessary through approximations, and rewrite it in a human understandable function of just a few parameters to which we can relate to; i.e parameters based on well-known concepts which are intuitive. An example of such a concept is that of the hydrogen bond (or any chemical bond in general) and order parameters based on it such as the number of hydrogen bonds that a specific molecule donates or accepts. Although there is not a single unique microscopic definition for something like a hydrogen bond[12,13], it provides a tool which helps our understanding of solvent dynamics and the functioning of bio-molecules such as DNA. The knowledge that some well-ordered hydrogen-bonded water networks are essential for reactions[14,15] might eventually lead to rationalized approaches to steer chemical reactions, produce new materials, or design more efficient catalysts.

So in one way or another, we need to translate our findings into intuitively understandable parameters, in other words, gaining an understanding of the reaction mechanism. In this article, we introduce an analysis method to test hypotheses about the reaction mechanism and to identify the essential circumstances which make a reaction proceed or not. The analysis method uses the path sampling data which are produced by rare event sampling methods such as transition interface sampling (TIS)[16], replica exchange TIS (RETIS)[17], and forward flux sampling[18]. These methods employ a Monte Carlo (MC) sampling of trajectories within a series of simulations, each evaluating a different path ensemble.

A trajectory belongs to an ensemble whenever it starts at the boundary of the reactant state, moves along the barrier up to a certain minimum progression, and then either ends by re-entering the reactant state or entering the product state. The minimum progression requirement is enforced by an interface crossing condition: associated with each path ensemble there is an unique interface defined by the value of the progress coordinate ("reaction coordinate") which needs to be crossed. The results from the different path ensembles can be combined and allow one to obtain the reaction rate without additional approximations, but orders of magnitude faster than straight-forward molecular dynamics (MD). Although the progress coordinate in the TIS methods[16–18] is not related to the committor, some have argued that the committor as progress coordinate[19] would give the best possible efficiency for these methods. However, it should be noted that there is a crucial difference between TIS and RETIS on the one hand and FFS on the other hand regarding the efficiency scaling. Whereas the efficiency of TIS and RETIS is relatively insensitive to the choice of progress coordinate and outperforms standard free energy based methods to compute rates whenever both are based on a poor reaction coordinate, FFS is doing worse than the standard methods in that case[20,21]. It is interesting to note that Ref. 19 actually refers to the phase space committor implying that the reaction coordinate providing the most efficient sampling should be momentum-dependent, something which is very unusual. However, as shown in Ref. 21, FFS indeed requires such momentum-dependent reaction coordinate in an underdamped one-dimensional system while it is neither needed nor more efficient for TIS and RETIS.

The aim of our analysis method is, therefore, not finding the committor or a single coordinate per se. Rather, it tries to identify which additional coordinates (possibly momentum-dependent) other than the chosen reaction coordinate determine the progress of the reaction. Though, as we will show below, in principle our method can also be used to determine the full phase-space committor just using the data of the TIS methods. Our article is organized as follows. In Sec. II we give the theoretical definitions that are being used in our methodology. In Sec. III we show how these theoretical measures of predictiveness can be computed using path sampling data from TIS, RETIS or FFS. In Sec. IV, we show numerical results for a one-dimensional double well potential, a theoretical model for an ion-transfer reaction, and an Ab Initio molecular dynamics study of water auto-ionization. In Sec. V we elaborate further on the possibilities of our methodology, in particular we discuss how the number of CVs can be reduced while maintaining the same predictive power and how that ultimately can be used to determine the phase space committor. We end with concluding remarks in Sec. VI.

II. DEFINITIONS

TIS, FFS, and RETIS are based on a partitioning of the phase space using interfaces (from here on called TIS interfaces). Let $\lambda(x)$ be a progress coordinate which is in principle a function of phase space point x. In many cases it can be taken as a geometric function like the length of a bond that needs to be broken, the largest solid cluster in a nucleation study, the radius of gyration for protein folding, etc. Then, the collection of phase points x having a specific value λ_i form the interfaces. This implies that $\{x|\lambda(x) = \lambda_i\}$ comprises interface number i. In the case of $M+1$ interfaces, $\lambda_0 = \lambda^A$ is placed within the reactant well, $\lambda_M = \lambda^B$ is placed in the product well, and the interfaces in between, λ_i for $0 < i < M$, are placed in the barrier region. Here, we use the subscript notation to indicate the integer index for the TIS interfaces and a superscript to indicate a specific value of the λ-parameter. The system is then considered belonging to the *overall state* $\mathcal{A}$ if it crossed λ^A more recently than λ^B. If there is a clear separation of time scales, the overall states will be insensitive to the exact positioning of λ^A and λ^B as long as they are reasonable; it is assumed that once λ^A or λ^B is crossed from the barrier region side the system will relax to that respective state (commit). The other interfaces are placed in order to maximize efficiency.

Within the TIS theory, the rate constant k_{AB} is defined as the number of transitions from *overall state* $\mathcal{A}$ to *overall state* $\mathcal{B}$ per time unit which can be expressed as the flux through λ^A times the overall crossing probability $\mathcal{P}_A(\lambda_M|\lambda_0)$. The last term equals the chance that the system will cross λ_M before λ_0 provided that it just crossed λ_0 in the positive direction (as convention we assume that the reactant state and the product state are situated at the left and right side of the barrier, respectively). This probability is generally too low to be determined directly, but it can be computed by a series of path simulations using the following relation[16]

$$\mathcal{P}_A(\lambda_M|\lambda_0) = \prod_{i=1}^{M} \mathcal{P}_A(\lambda_i|\lambda_{i-1}) \tag{1}$$

Here, $\mathcal{P}_A(\lambda_i|\lambda_{i-1})$ is the conditional probability that the system coming from λ_0 and then crossing λ_{i-1} for the first time will cross λ_i as well before recrossing λ_0 again. Naturally, $\mathcal{P}_A(\lambda_i|\lambda_{i-1})$ will be much larger than the overall crossing probability whenever λ_i is sufficiently close to λ_{i-1} and this property can be computed by Monte Carlo walk in path space.

We denote with X a path of $L+1$ time slices

$$X = \{x_0, x_1, \ldots, x_L\} \tag{2}$$

where L is the path length and x_k is the k-th phase point of the path, also called time slice. We will further refer to the nomenclature of RETIS where the path ensemble $[i^+]$ comprises the collection of trajectories with the following properties:

$$\begin{aligned} &X \in [i^+] \text{ if:} \\ &\quad \lambda(x_0) < \lambda_0, \\ &\quad \lambda(x_L) < \lambda_0 \text{ or } \lambda(x_L) > \lambda_M, \\ &\quad \lambda_0 < \lambda(x_k) < \lambda_M \text{ for } k = 1, 2, \ldots, L-1, \\ &\quad \lambda_{\max} \equiv \max[\lambda(x_1), \lambda(x_2), \ldots, \lambda(x_L)] > \lambda_i \end{aligned} \tag{3}$$

We define characteristic binary functions which relate to whether X is within $[i^+]$ or not

$$h_i(X) = 1 \text{ if } X \in [i^+], \quad 0 \text{ otherwise} \tag{4}$$

and the weight, $\varrho_i(X)$, of a path in ensemble $[i^+]$ is given by

$$\varrho_i(X) = h_i(X)\rho(x_0) \prod_{k=0}^{L-1} p(x_k \to x_{k+1}) \tag{5}$$

Here, ρ is the phase space density and $p(x_k \to x_{k+1})$ are the hopping probability densities; the chance that the system moves to phase point x_{k+1} in a single Δt time step given that it is in x_k. An ensemble average of an arbitrary path function $a(X)$ in the $[i^+]$ ensemble equals

$$\langle a(X) \rangle_{\varrho_i} = \frac{\int a(X)\varrho_i(X)\mathcal{D}X}{\int \varrho_i(X)\mathcal{D}X} \tag{6}$$

where the integral is formally equal to $\int \ldots \mathcal{D}X = \sum_{L=1,\infty} \int \cdots \prod_{k=0,L} \mathrm{d}x_k$. In practice, however, we only compute ratios of two path space integrals like the one of Eq. 6 using MC in trajectory space. In this method we collect a Markov chain of trajectories for specific path ensembles using MC moves (like e.g. shooting[22]) obeying detailed balance $\varrho_i(X^{(\mathrm{o})})P_{\mathrm{gen}}(X^{(\mathrm{o})} \to X^{(\mathrm{n})})P_{\mathrm{acc}}(X^{(\mathrm{o})} \to X^{(\mathrm{n})}) = \varrho_i(X^{(\mathrm{n})})P_{\mathrm{gen}}(X^{(\mathrm{n})} \to X^{(\mathrm{o})})P_{\mathrm{acc}}(X^{(\mathrm{n})} \to X^{(\mathrm{o})})$ where $X^{(\mathrm{o})}$ and $X^{(\mathrm{n})}$ are the old and new paths, respectively, and P_{gen} and P_{acc} are the generation and acceptance probabilities of the MC algorithm. Eq. 6 is then a simple average of the simulation $\langle a(X) \rangle_{\varrho_i} \approx \frac{1}{N_{\mathrm{sim}}} \sum_{n=1,N_{\mathrm{sim}}} a(X_n)$ where X_n is the n-th path sampled in the simulation and N_{sim} is the total number of paths.

In the following we will focus on first crossing points with interfaces as defined by our progress coordinate. We define x^{λ^c} as the first crossing point with interface λ^c:

$$\begin{aligned} &x^{\lambda^c}(X) = x_k \in X \text{ if } \lambda(x_k) \geq \lambda^c \\ &\text{while } \lambda(x_l) < \lambda^c \text{ for all } l < k \end{aligned} \tag{7}$$

Naturally $\lambda(x^{\lambda^c}(X)) \gtrsim \lambda^c$, but there are many other collective variables (CVs) which can characterize this crossing point. Let us call these coordinates $\Psi_1, \Psi_2, \ldots, \Psi_N$. For instance, if $\lambda(x)$ is the bond length between two atoms, which needs to be broken to establish the reaction, $\Psi_1(x)$ could be related to the relative position of a catalyst, $\Psi_2(x)$ a coordinate describing the solvent structure etc. For a set of N collective variables in addition to $\lambda(x)$ we denote $\Psi^N(x) = \{\Psi_1(x), \Psi_2(x), \ldots, \Psi_N(x)\}$ as the vector describing the additional collective variables. Hence, whereas $\lambda(x^{\lambda^c}(X))$ describes the interface that is being crossed, $\Psi^N(x^{\lambda^c}(X))$ describes the position within this surface where the crossing takes place. We will therefore call Ψ^N the orthogonal coordinates, though we should stress that this does not imply any strict orthogonality as in a Euclidean sense. In fact, λ, Ψ_1, Ψ_2, …, or Ψ_N do not even have to have the same dimensionality; these can be a mix of distances, angles, integer values like the number of hydrogen bonds, Boolean functions, etc. Also, $\lambda(x)$ and $\Psi^N(x)$ do not necessarily have to be mutually independent. For instance, $\Psi_1(x) = (\lambda(x))^2$ would be a valid option. Of course, this Ψ_1 does not add any information about the system which we could not have already known from λ. However, this is a conclusion which should come out of our analysis method. Therefore, there is not a strict need to think very carefully about possible dependencies at this stage. We can choose $\Psi^N(x)$ based on our intuition; the collective variables which we think are important for the reaction.

We will consider three interfaces, the reactant interface λ^A, the crossing interface $\lambda^c > \lambda^A$, and the (partial) reaction interface $\lambda^r > \lambda^c$. Considering all trajectories coming from λ^A which cross λ^c, we can characterize "reactive" and "unreactive" trajectories up to λ^r. The reactive ones cross λ^r, the unreactive ones recross λ^A without crossing λ^r. Naturally, if $\lambda^r = \lambda^B$ the "reactive" trajectories are then fully reactive, but for $\lambda^r < \lambda^B$ we get useful information about the reaction mechanism at intermediate stages of the reaction, and probably better statistics since crossing $\lambda^r < \lambda^B$ is less rare than crossing λ^B.

In the following methodology both λ^c and λ^r can be shifted to the desired region at the reaction barrier. As explained below, we can use the output of standard TIS, RETIS, or FFS simulations to extract a statistically representative subset of trajectories that cross λ^c. This subset can then be used to analyze the first crossing points in CV space: $\Psi_1(x^{\lambda^c}), \Psi_2(x^{\lambda^c}), \ldots$.

By constructing a grid in the collective variable space Ψ^N, we can define bins covering the full accessible surface of the λ^c interface. Let q be the index of these bins. Then, of all trajectories crossing λ^c let t_q be the fraction of trajectories passing through bin q in the λ^c surface, r_q the fraction of trajectories passing through bin q *and* cross λ^r, and u_q the fraction of trajectories passing through bin q but do *not* reach λ^r (See Fig. 1).

We can write down following relations

$$\begin{aligned} t_q = u_q + r_q, \quad \sum_q t_q = 1, \\ \sum_q r_q = \mathcal{P}_A(\lambda^r|\lambda^c), \quad \sum_q u_q = 1 - \mathcal{P}_A(\lambda^r|\lambda^c) \end{aligned} \tag{8}$$

Depending on the CVs and grid spacing we will get different fractions of reactive and unreactive paths in each

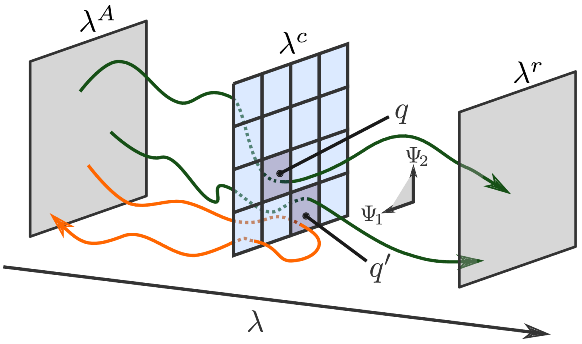


FIG. 1: (color online). Visualization of reactive and unreactive trajectories passing through bins as defined by two orthogonal CVs Ψ_1 and Ψ_2. The green trajectories are reactive up to λ^r while the orange trajectory is unreactive. If we would only base our analysis on these three trajectories we would have $t_q = 1/3, t_{q'} = 2/3$ and $t_{q''} = 0$ for any other bin q''. In addition, the reactive and unreactive distributions for bins q, q' would be $r_q = 1/3$, $u_q = 0$ and $r_{q'} = u_{q'} = 1/3$.

bin. If we would be able to partition the first crossing points such that $r_q/t_q = 1$ or $r_q = 0$ for each bin, the predictive ability is optimal; each time that λ^c is crossed for the first time, we check through which bin it passes and, then, we would be able to say whether it will cross λ^r or not (assuming that there is no problem with the accuracy of our beforehand estimated distributions r and u). In practice this might not be possible, either because the dynamics is stochastic or because it turns out to be too difficult to find the right CVs. In that case each bin q can have any fractional value between zero and one for the reactive ratio r_q/t_q. The overall measure of predictive power, that can be obtained from the orthogonal coordinates, must then be a weighted average of r_q/t_q over q. This measure should be high if there are many bins with $r_q/t_q = 1$. However, if only a very small fraction of the reactive trajectories move through bin q, this will not have a large impact on the overall predictive power. Therefore, we introduce a measure $\mathcal{T}$ for the CVs regarding their predictive ability which is a weighted average of r_q/t_q where each bin is weighted with the fraction of reactive trajectories passing through q:

$$\begin{aligned}\mathcal{T} &\equiv \sum_q \left(\frac{r_q}{\sum_v r_v}\right)\frac{r_q}{t_q} = \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\sum_q \frac{r_q^2}{t_q} \\ &= \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\sum_q \frac{r_q(t_q-u_q)}{t_q} \\ &= \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\sum_q r_q - \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\sum_q \frac{r_q u_q}{t_q} \\ &= 1 - \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\sum_q \frac{r_q u_q}{t_q} \equiv 1 - \mathcal{S} \qquad (9)\end{aligned}$$

In continuous space, S is the overlap integral of the reactive and unreactive distributions.

$$\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N] = \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\int\left(\frac{r^{\lambda^c,\lambda^r}(\Psi^N)u^{\lambda^c,\lambda^r}(\Psi^N)}{t^{\lambda^c}(\Psi^N)}\right)\mathrm{d}\Psi^N \qquad (10)$$

The overlap $\mathcal{S}_A^{\lambda^c,\lambda^r}$ will depend on the selection of CVs which are functions of phase space x. Hence, $\mathcal{S}_A^{\lambda^c,\lambda^r}$ is a functional of $\Psi^N(x)$. The highest possible predictive ability is obtained by finding the collective variables that minimize the overlap

$$\mathcal{S}_{A,0}^{\lambda^c,\lambda^r} = \frac{1}{\mathcal{P}_A(\lambda^r|\lambda^c)}\times \min_{\Psi^N}\left[\int\left(\frac{r^{\lambda^c,\lambda^r}(\Psi^N)u^{\lambda^c,\lambda^r}(\Psi^N)}{t^{\lambda^c}(\Psi^N)}\right)\mathrm{d}\Psi^N\right] \qquad (11)$$

and we call the corresponding collective variables Ψ^N_{min}

$$\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N_{\mathrm{min}}] = \mathcal{S}_{A,0}^{\lambda^c,\lambda^r} \qquad (12)$$

The Ψ^N_{min} variables are in general not unique. For instance, if $\Psi_1(x)$ is a distance between two atoms, we could as well have taken the squared distance. Similarly, we could add or remove CVs to total set of CVs which have no correlation with reactivity. These operations will not change the overlap value. However, since our goal is to gain insight and to provide inspiration how to steer chemical reactions, the ideal set of orthogonal coordinates are those that minimize $\mathcal{S}_A^{\lambda^c,\lambda^r}$ *and* are also intuitive; e.g. based on known concepts such as number of hydrogen bonds, radii of gyration, nucleus size etc.

In the case that the CVs do not correlate with reactivity: $P_A(\lambda^r|\lambda^c,\Psi^N) = P_A(\lambda^r|\lambda^c)$. In other words, the chance to cross λ^r after crossing λ^c is independent of where the first crossing with λ^c takes place in the Ψ^N space. This might either indicate that the CVs were badly chosen or because the λ^c surface is an isocommittor surface with respect to λ^r. The former implies that these specific CVs do not improve predictivity, while the latter implies that there simply are no CVs which potentially could improve the predictive power. For both cases, the absence of correlation implies that $r^{\lambda^c,\lambda^r}(\Psi^N) = \mathcal{P}_A(\lambda^r|\lambda^c)t^{\lambda^c}(\Psi^N)$, $u^{\lambda^c,\lambda^r}(\Psi^N) = [1-\mathcal{P}_A(\lambda^r|\lambda^c)]\,t^{\lambda^c}(\Psi^N)$. Substitution in Eq. 10 gives $\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N] = 1 - \mathcal{P}_A(\lambda^r|\lambda^c)$ and $\mathcal{T}_A^{\lambda^c,\lambda^r} = \mathcal{P}_A(\lambda^r|\lambda^c)$. In other words, Ψ^N does not provide more information which can help us to tell whether λ^r will be crossed or not. Based on the fact that we observe an effective positive crossing with λ^c, we know already that the chance of a (partial) reaction is $\mathcal{P}_A(\lambda^r|\lambda^c)$. Any knowledge about the orthogonal space expressed in the CVs Ψ^N does not increase the quality of our predictions. Of course, if λ^c is sufficiently beyond the transition state $\mathcal{P}_A(\lambda^r|\lambda^c) = 1$. Hence, we can still have a high predictive power. However, Ψ^N does not improve it and $\mathcal{T}_A^{\lambda^c,\lambda^r}/\mathcal{P}_A(\lambda^r|\lambda^c)$ will be equal to one. Therefore, $\mathcal{T}_A^{\lambda^c,\lambda^r}$ is a useful measure of predictive capacity using all information (both λ^c and Ψ^N) while $\mathcal{T}_A^{\lambda^c,\lambda^r}/\mathcal{P}_A(\lambda^r|\lambda^c)$ is a measure of the enhancement of predictive capacity due to the information of the selected orthogonal coordinates. Note that $\mathcal{T}_A^{\lambda^c,\lambda^r} \leq 1$ and $\mathcal{T}_A^{\lambda^c,\lambda^r}/\mathcal{P}_A(\lambda^r|\lambda^c) \geq 1$, which basically means that predictions can never be more than 100% correct and additional information on Ψ^N can never be harmful for the predictive power.

Since, the path sampling data allow computing the overlap for different values of λ^c and λ^r, these functions can be plotted for the full range, $\lambda^A \leq \lambda^c < \lambda^B$ and $\lambda^c \leq \lambda^r < \lambda^B$, in order to provide information about the predictive power of the CVs Ψ^N at each stage of the reaction. Numerical examples showing such plots are given in Sec. IV.

III. PATH REWEIGHING

In this section we will show how the results from TIS, FFS, or RETIS can be used to compute $\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N]$ for a predefined set of CVs. As mentioned above, these path sampling methods for computing reaction rates consist of a series of simulations. Each simulation samples a so-called path ensemble. The $[i^+]$ path ensemble consists of trajectories that start at λ^A, cross λ_i at least once, and then might end at either λ^A or λ^B. In TIS, RETIS, and FFS different path simulations sample the different ensembles $[0^+], [1^+], \ldots [(M-1)^+]$. In addition, TIS and FFS also require a short MD simulation initiated from the reactant state while RETIS employs an extra path ensemble $[0^-]$. These will not be part of our analysis and in the following, when referring to the i-th simulation, we mean the simulation exploring the $[i^+]$ path ensemble.

Hence, from the trajectories generated in the i-th path ensemble, we can in principle straightforwardly determine $\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N]$ for $\lambda^c = \lambda_i$. We simple gather the effective crossing points with λ_i and from these we can construct histograms for t^{λ^c}, r^{λ^c,λ^r}, and u^{λ^c,λ^r} in the Ψ^N space choosing appropriate bin widths, and by checking whether the trajectories cross λ^r or not. Once the histograms are constructed, integrations of Eq. 10 can be carried out to obtain $\mathcal{S}_A^{\lambda^c,\lambda^r}$ and $\mathcal{T}_A^{\lambda^c,\lambda^r}$.

However, we would like to determine $\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N]$ or $\mathcal{T}_A^{\lambda^c,\lambda^r}[\Psi^N]$ on the full range and not restrict λ^c to any of the TIS interfaces. In addition, we would also like to combine all data of the different path simulations to reduce statistical errors, especially if crossing λ^r from λ^c is a rare event. We can achieve this by path reweighting[23] based on the weighted histogram analysis method (WHAM)[24–26].

For convenience, we introduce following notation for the multidimensional Dirac delta function in CV space

$$\delta^{\lambda^c}(\Psi'^N, X) \equiv \prod_{m=1}^{N} \delta(\Psi_m(x^{\lambda^c}(X)) - \Psi'_m) \tag{13}$$

Now, suppose that $\lambda^c = \lambda_i$ and $\lambda^r = \lambda_j$ are both identical to one of TIS interfaces. Then we can write for r^{λ_i,λ_j}:

$$\begin{aligned} r^{\lambda_i,\lambda_j}(\Psi^N) &= \left\langle h_j(X)\delta^{\lambda_i}(\Psi^N, X)\right\rangle_{\varrho_i} \\ &= \frac{\int \varrho_i(X) h_j(X) \delta^{\lambda_i}(\Psi^N, X)\mathcal{D}X}{\int \varrho_i(X)\mathcal{D}X} \end{aligned} \tag{14}$$

Then, using that for $j > i$: $h_i(X)h_j(X) = h_j(X)$ or $\varrho_i(X)h_j(X) = \varrho_j(X)$, we can rewrite the above expression to get an ensemble average in $[j^+]$ ensemble.

$$\begin{aligned} &r^{\lambda_i,\lambda_j}(\Psi^N) = \\ &\left(\frac{\int \varrho_j(X)\delta^{\lambda_i}(\Psi^N, X)\mathcal{D}X}{\int \varrho_j(X)\mathcal{D}X}\right)\left(\frac{\int \varrho_i(X)h_j(X)\mathcal{D}X}{\int \varrho_i(X)\mathcal{D}X}\right) \\ &= \left\langle \delta^{\lambda_i}(\Psi^N, X)\right\rangle_{\varrho_j} \mathcal{P}_A(\lambda_j|\lambda_i) \end{aligned} \tag{15}$$

Here, $\mathcal{P}_A(\lambda_j|\lambda_i)$ is a known result from the interface path sampling simulation since the computation of the full crossing probability $\mathcal{P}_A(\lambda|\lambda_0)$ is a central output to the TIS, FFS, and RETIS and $\mathcal{P}_A(\lambda_j|\lambda_i) = \mathcal{P}_A(\lambda_j|\lambda_0)/\mathcal{P}_A(\lambda_i|\lambda_0)$. In principle, also the data of the other path ensembles can be used to obtain r^{λ_i,λ_j} since for any $k < j$:

$$r^{\lambda_i,\lambda_j}(\Psi^N) = \left\langle \delta^{\lambda_i}(\Psi^N, X)h_j(X)\right\rangle_{\varrho_k} \mathcal{P}_A(\lambda_k|\lambda_i) \tag{16}$$

The ensembles $[k^+]$ with $k > j$ can by itself not be used to fully compute the r^{λ_i,λ_j} distribution, but still these data can be used to reduce its statistical errors. WHAM[24–26] provides a way to take a weighted average of the distributions which have been obtained using different bias functions (also called windows). That is, for an arbitrary parameter $\xi(x)$ the most accurate distribution that can be obtained from the different biased simulations is

$$\rho(\xi) = \frac{\sum_{i=1}^{N_w} \omega_i(\xi)\rho_i^{\text{unb.}}(\xi)}{\sum_{j=1}^{N_w} \omega_j(\xi)} \tag{17}$$

Here N_w is the number of windows and $\rho(\xi)_i^{unb}$ is the unbiased distribution of simulation i. This is the distribution after proper rescaling to remove the effect bias. Further, ω_i are weights depending on ξ, chosen to be proportional to inverse square of the estimated error in each simulation.

As shown in the appendix, also the crossing probability itself can then be expressed using WHAM:

$$\mathcal{P}_A(\lambda|\lambda_0) = \frac{\sum_{i=0}^{K(\lambda)} n_i[\lambda]}{\sum_{j=0}^{K(\lambda)} n_j[\mathcal{P}_A(\lambda_j|\lambda_0)]^{-1}} \tag{18}$$

Here $n_i[\lambda]$ is the number of trajectories in simulation i having a $\lambda_{\max} > \lambda$, n_j is the total number of trajectories in simulation j, and $K(\lambda)$ is the integer which fulfills

$$K(\lambda) = \begin{cases} k & \text{if } \lambda_k < \lambda \leq \lambda_{k+1} \text{ and } \lambda < \lambda^B \\ M-1 & \text{if } \lambda > \lambda^B \end{cases} \tag{19}$$

The maximum of $M-1$ is due to the fact that there is generally not a simulation that just considers the $[M^+]$ ensemble since it gives the trivial unit contribution in the product expression, Eq. 1. Eq. 18 can be solved iteratively and is presumably somewhat more accurate than Eq. 1 since it is based on more data. If n_j is equal for all simulations, then Eq. 18 is identical to the crossing probability derived by Rogal et al.[23] in a different manner. In our derivation (see appendix) we also provide two refinements of the above expression. One is standard and appears also in e.g Roux[26] and is related to the effect of correlated trajectories which might be more severe in some of the simulation than in the others. Another refinement, that is non-standard, is related to the non-negligible size of the bins in the determination of crossing probability. The use of more refined expressions is not always preferred since they rely on the facts that errors can be obtained accurately while, in practice, simulations which tend to get trapped for a long time can provide artificially low standard deviations. In the refined expression these simulations could get the highest weights and overwhelm the more converged results of the other simulations. We have therefore used the simpler expression, Eq. 18, in the remainder of this article.

In order to obtain our distributions $t_q^{\lambda^c}$, $r_q^{\lambda^c,\lambda^r}$, and $u_q^{\lambda^c,\lambda^r}$ we first compute the following ensemble averages $\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0}$ in which $H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}(X)$ is 1 (otherwise 0) if and only if trajectory X passes through bin q on the λ^c surface while $\lambda_{\max}(X)$ is inside the interval $[\lambda^a:\lambda^b]$. Moreover, we restrict ourselves to intervals which do not overlap with any of the TIS interfaces. In other words $\lambda^b \leq \lambda_{K(\lambda^a)+1}$. Then for any $i \leq K(\lambda^a)$:

$$\begin{aligned}\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0} &= \frac{\int H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\varrho_0(X)\mathrm{d}X}{\int \varrho_0(X)\mathrm{d}X} \\ &= \frac{\int H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\varrho_i(X)\mathrm{d}X}{\int \varrho_i(X)\mathrm{d}X}\frac{\int \varrho_i(X)\mathrm{d}X}{\int \varrho_0(X)\mathrm{d}X} \\ &= \left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_i}\mathcal{P}_A(\lambda_i|\lambda_0) \end{aligned} \tag{20}$$

Here we used the relation $H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\varrho_0 = H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\varrho_i$ valid for $\lambda_i < \lambda^a$. Hence, this property can be determined using different interface ensemble simulations. As shown in the appendix, the WHAM weights which are assumed to minimize the error equals

$$\omega_i = \frac{[\mathcal{P}_A(\lambda_i|\lambda_0)]^{-1}}{\sum_{j=0}^{K(\lambda^a)}[\mathcal{P}_A(\lambda_j|\lambda_0)]^{-1}} \tag{21}$$

This implies that the WHAM expression equals

$$\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0} = \frac{\sum_{i=0}^{K(\lambda^a)} n_i(q,\lambda^c;[\lambda^a:\lambda^b])}{\sum_{j=0}^{K(\lambda^a)} n_j[\mathcal{P}_A(\lambda_j|\lambda_0)]^{-1}} \tag{22}$$

where $n_i(q,\lambda^c;[\lambda^a:\lambda^b])$ is the number of trajectories in simulation i moving through bin q at its first crossing with λ^c and having $\lambda_{\max}$ in the interval $[\lambda^a:\lambda^b]$.

Our distributions can then be constructed from these since

$$\begin{aligned} R_q^{\lambda^c,\lambda^r} &= \left\langle H_{q,\lambda^c}^{[\lambda^r:\lambda_{K(\lambda^r)+1}]}\right\rangle_{\varrho_0} + \sum_{k=K(\lambda^r)+1}^{M}\left\langle H_{q,\lambda^c}^{[\lambda_k:\lambda_{k+1}]}\right\rangle_{\varrho_0} \\ U_q^{\lambda^c,\lambda^r} &= \left\langle H_{q,\lambda^c}^{[\lambda^c:\lambda^r]}\right\rangle_{\varrho_0} \text{ if } \lambda^r < \lambda_{K(\lambda^c)+1} \text{ or} \\ &= \left\langle H_{q,\lambda^c}^{[\lambda^c:\lambda_{K(\lambda^c)+1}]}\right\rangle_{\varrho_0} + \sum_{k=K(\lambda^c)+1}^{K(\lambda^r)}\left\langle H_{q,\lambda^c}^{[\lambda_k:\lambda_{k+1}]}\right\rangle_{\varrho_0} \\ &+ \left\langle H_{q,\lambda^c}^{[\lambda_{K(\lambda^r)}:\lambda_r]}\right\rangle_{\varrho_0} \text{ if } \lambda^r > \lambda_{K(\lambda^c)+1} \end{aligned} \tag{23}$$

where $\lambda_{M+1} = \infty$. Now, by rescaling we obtain the actual distributions

$$r_q^{\lambda^c,\lambda^r} = \frac{R_q^{\lambda^c,\lambda^r}}{\mathcal{P}_A(\lambda^c|\lambda_0)}, \quad u_q^{\lambda^c,\lambda^r} = \frac{U_q^{\lambda^c,\lambda^r}}{\mathcal{P}_A(\lambda^c|\lambda_0)} \tag{24}$$

IV. NUMERICAL RESULTS

In this section we will give a detailed description of the implementation for calculating the crossing probability and distribution functions. We will also exemplify the method by applying it to three systems we have studied with RETIS simulations.

A. Implementation

1. *The Crossing Probability*

In order to obtain the distribution functions using WHAM we need, first of all, to obtain the crossing probability. This can be done using the product expression, Eq. 1, or the more accurate expression based on WHAM, Eq. 18. We will discuss the last one. The first step is to obtain the values of the crossing probability at the TIS interfaces. Setting $\mathcal{P}_A(\lambda_0|\lambda_0) = 1$ gives directly

$$\mathcal{P}_A(\lambda_1|\lambda_0) = \frac{n_0(\lambda_1)}{n_0} \tag{25}$$

or simply the number of trajectories in simulation 0 crossing λ_1 divided by the total number of trajectories in simulation 0. The next interface gives

$$\mathcal{P}_A(\lambda_2|\lambda_0) = \frac{n_0(\lambda_2)+n_1(\lambda_2)}{n_0 + n_1\left[\mathcal{P}_A(\lambda_1|\lambda_0)\right]^{-1}} \tag{26}$$

and so we can continue determining $\mathcal{P}_A(\lambda_3|\lambda_0)$, $\mathcal{P}_A(\lambda_4|\lambda_0)$, ..., $\mathcal{P}_A(\lambda_M|\lambda_0)$. For convenience we define

$$Q_k \equiv \frac{1}{\sum_{j=0}^{k} n_j\left[\mathcal{P}_A(\lambda_j|\lambda_0)\right]^{-1}} \tag{27}$$

and the crossing probability for any continuous value λ is then obtained by a simple summation which includes all trajectories X in all path sampling data of ensembles $i = 0, 1, \ldots, M-1$

$$\begin{aligned}\mathcal{P}_A(\lambda|\lambda_0) &= Q_{K(\lambda)} \sum_{i=0}^{K(\lambda)} n_i[\lambda] \\ &= Q_{K(\lambda)} \sum_{i=0}^{M-1} n_i[\lambda]\theta(\lambda-\lambda_i) \\ &= Q_{K(\lambda)} \sum_{i=0}^{M-1} \sum_{X\in[i^+]} \theta(\lambda_{\max}(X)-\lambda)\theta(\lambda-\lambda_i)\end{aligned} \tag{28}$$

with $\theta(\cdot)$ being the Heaviside step function.

In an actual computer algorithm, $\mathcal{P}_A(\lambda|\lambda_0)$ is computed using a small step-size along the λ-parameter which define a fine grid of sub-intervals. Let $v(\alpha)$ be the vector for determining $\mathcal{P}_A(\lambda|\lambda_0)$ on this fine grid such that α is an index of the sub-interface λ^α and $v(\alpha) = \mathcal{P}_A(\lambda^\alpha|\lambda_0)$ after completion of the algorithm. We can then determine the full vector v as follows:

1. Set all entries of $v(\alpha)$ equal to 0: $v(\alpha) = 0$.
2. Loop over all data sets corresponding to the path ensembles $i = 0, 1, \ldots, M-1$, and for each trajectory X in data set i:
   - 2.1. Determine $\lambda_{\max}(X)$.
   - 2.2. For each α where $\lambda_i \leq \lambda^\alpha < \lambda_{\max}(X)$, increment $v(\alpha)$: $v(\alpha) = v(\alpha) + 1$
3. For each α determine $K(\alpha)$ and $Q_{K(\alpha)}$ and multiply this with the vector entry: $v(\alpha) = v(\alpha) \times Q_{K(\alpha)}$.

2. Probability Distribution Functions

We can also determine the probability distribution function $u_q^{\lambda^c,\lambda^r}, r_q^{\lambda^c,\lambda^r}$ based on Eqs. 22-24 using a single loop over all trajectories. Substituting Eq. 27 into Eq. 22 yields

$$\begin{aligned}\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0} &= Q_{K(\lambda^a)} \sum_{i=0}^{K(\lambda^a)} n_i(q,\lambda^c;[\lambda^a:\lambda^b]) \\ &= Q_{K(\lambda^a)} \sum_{i=0}^{M-1} n_i(q,\lambda^c;[\lambda^a:\lambda^b]) \\ &= \sum_{i=0}^{M-1} \sum_{X\in[i^+]} Q_{K(\lambda_{\max})}\delta_{q,\lambda_c} \\ &\quad \times\ \theta(\lambda^a-\lambda_{\max})\theta(\lambda_{\max}-\lambda^b)\end{aligned} \tag{29}$$

where δ_{q,λ_c} is 1 (otherwise zero) whenever X has a first crossing through bin q at the λ^c surface. In the second equation we used the fact that we only consider intervals $[\lambda^a : \lambda^b]$ that are not overlapping with the TIS interfaces. This implies that $\lambda^b < \lambda_{K(\lambda^a)+1}$ and, hence, $n_i(q,\lambda^c;[\lambda^a:\lambda^b]) = 0$ for any $i \geq K(\lambda^a)+1$. In the third expression we use that whenever $\lambda_{\max}(X)$ is within this interval $[\lambda^a : \lambda^b]$, we must have $K(\lambda^a) = K(\lambda_{\max})$. Eq. 29 shows that $\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0}$ can be expressed as a sum over all trajectories without considering the actual simulation i it was taken from. The values to be summed are either zero or $Q_{K(\lambda_{\max})}$. A non-zero contribution can only occur whenever $\lambda_{\max}$ is within the interval $[\lambda^a : \lambda^b]$. Now let us consider Eqs. 23. To calculate $R_q^{\lambda^c,\lambda^r}$ and $U_q^{\lambda^c,\lambda^r}$ we need to add $\left\langle H_{q,\lambda^c}^{[\lambda^a:\lambda^b]}\right\rangle_{\varrho_0}$ for different intervals and since these intervals are not overlapping, each X can only give a contribution 0 or $Q_{K(\lambda_{\max})}$ to the total sum as well. The non-zero contribution occurs whenever there is a first crossing through bin q and $\lambda_{\max}$ is within $[\lambda^c : \lambda^r]$ for $U_q^{\lambda^c,\lambda^r}$ and whenever $\lambda_{\max}$ is larger than λ^r for $R_q^{\lambda^c,\lambda^r}$. Now, let $M_u(q,\alpha,\beta)$ and $M_r(q,\alpha,\beta)$ be the matrices used to construct the $u_q^{\lambda^\alpha,\lambda^\beta}$ and $r_q^{\lambda^\alpha,\lambda^\beta}$ distributions, respectively, where α, β are indices of the fine grid along λ. The computational algorithm is then as follows:

1. Set all entries of matrices $M_u(q,\alpha,\beta)$ and $M_r(q,\alpha,\beta)$ equal to 0.
2. Loop over all data sets corresponding to path ensembles $i = 0, 1, \ldots M-1$, and for each trajectory X in data set i:
   - 2.1. Determine $\lambda_{\max}$ and $Q_{K(\lambda_{\max})}$.
   - 2.2. For each α such that $\lambda^\alpha < \lambda_{\max}$:
     - Determine x^{λ^α} and the corresponding bin q.
     - For each β such that $\lambda^\beta > \lambda_{\max}$, add $Q_{K(\lambda_{\max})}$ to the entries of $M_u(q,\alpha,\beta)$: $M_u(q,\alpha,\beta) = M_u(q,\alpha,\beta) + Q_{K(\lambda_{\max})}$.
     - For each β such that $\lambda^\beta \leq \lambda_{\max}$, add $Q_{K(\lambda_{\max})}$ to all entries of $M_r(q,\alpha.\beta)$: $M_r(q,\alpha,\beta) = M_r(q,\alpha,\beta) + Q_{K(\lambda_{\max})}$.
3. For each α, apply for all β and q the normalizations $M_r(q,\alpha,\beta) = M_r(q,\alpha,\beta)/v(\alpha)$ and $M_u(q,\alpha,\beta) = M_u(q,\alpha,\beta)/v(\alpha)$. The normalized matrices M_r and M_u are now estimates of the distributions $r_q^{\lambda^c,\lambda^r}$ and $u_q^{\lambda^c,\lambda^r}$, respectively.

The fine grid along λ does not have to be commensurate with the TIS interfaces. The accuracy is not affected by the spacing between the sub-interfaces unlike the binning in the orthogonal directions; The bins q have to be sufficiently large in order to determine the path density going through it, while still be sufficiently small enough to get enough resolution in order to discriminate between r_q and u_q.

B. Numerical Example 1: 1D Double Well Potential

The 1D double well potential, $V(r) = r^4 - 2r^2$, models the transition for a single particle between two stable states (located at $r \pm 1$) separated by a barrier (at $r = 0$)[21]. The progress coordinate is in this case given by the position of the particle in the potential: $\lambda = r$. We have investigated the transition between the two stable states using RETIS simulations under Langevin dynamics with a friction coefficient $\gamma = 0.3$ and a reduced temperature of 0.07 as described in Ref. 21.

Two additional collective variables have been considered for our analysis: (1) the velocity of the order parameter $v = \mathrm{d}\lambda/\mathrm{d}t$ and (2) the random Langevin force averaged over 9 steps (labeled “9rf”) after the crossing. The use of the second parameter might be viewed as a not completely ”fair” way to improve predictions since it assumes that, after crossing λ^c, one already knows which random numbers will be generated for the stochastic force. Still, it is a CV that we can use in a computer experiment in order to measure the balance of initial conditions at the crossing before the reaction takes place and the stochastic contributions during the coarse of the reaction. The use of velocity is also not a common parameter in reaction coordinate analysis studies. Notably exceptions are Ref. 27,28 in which not the configurational committor but the transmission coefficient selected to choose the reaction coordinates.

The first collective variable, v, is expected to improve the predictive capacity for small values of γ when the dynamics is largely deterministic, while the second, “9rf”, may improve the predictive capacity for stochastic dynamics resulting from a large friction coefficient γ. In Fig. 2, we show the crossing probability and the predictive capacities for the three combinations of collective variables ($\{v\}, \{9\mathrm{rf}\}, \{v, 9\mathrm{rf}\}$). These results show that the predictive capacity is largely improved by inclusion of the collective variable v, but only minimally with the inclusion of 9rf. This shows that the chance for a barrier crossing is much more determined by the right initial conditions than by a rare sequence of random kicks during the barrier crossing process. This is also the reason that FFS is not able to predict the crossing rate in an adequate manner[21].

A closer inspection (see Fig. 3) shows that the predictive capacities can be increased by several orders of magnitude due to the knowledge of v. The use of 9rf as CV only improves the predictions by 4% as is shown in the inset of the top panel.

C. Numerical Example 2: Ion Transfer Model in a Solvent

This example models the ion transfer reaction Ax + A → A + Ax where an ion, x is transferred between two molecules of the same type, A. The reaction takes place in a solvent (molecules of type B) which may reduce the barrier due to a cooperative effect. For the detailed description of the potential and the interactions, please see Ref. 29.

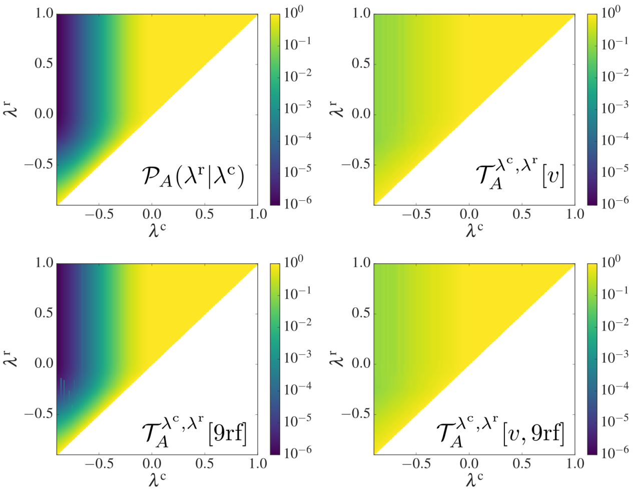


FIG. 2: (Color online.) The crossing probability, $\mathcal{P}_A(\lambda^r|\lambda^c)$, and the predictive capacities, $\mathcal{T}_A^{\lambda^c,\lambda^r}$, for the 1D double well potential: (top left) the crossing probability, (top right) the predictive capacity using the velocity (v) of the progress coordinate as the collective variable, (bottom left) the predictive capacity using the random Langevin force averaged over 9 steps (“9rf”) as the collective variable, (bottom right) the predictive capacity using both v and 9rf as collective variables. We used 200 sub-interfaces both for λ^r and λ^c. The histograms in the Ψ^N space were constructed using 20 bins for $-2 < v < 2$ and 20 bins for $-3 < 9\mathrm{rf} < 3$.

The ion is initially bound to one of the A molecules (labeled “A_1“) and it is transferred to the other (labeled A_2) over the course of the reaction. The progress coordinate, λ, is defined using the distance, $r_{\mathrm{x},\mathrm{A}_2}$, between x and A_2:

$$\lambda = -r_{\mathrm{x},\mathrm{A}_2} \tag{30}$$

where the minus sign ensures that the progress coordinate changes from a low value to a high value while the ion transfer advances. The reactant state is defined by $\lambda^A = -0.7$ and the product state by $\lambda^B - 0.4$.

For our analysis we have defined two additional collective variables: (1) the velocity, v, of the progress coordinate ($v = \mathrm{d}\lambda/\mathrm{d}t$), and (2) the coordination number, CN, for solvent molecules surrounding the ion, defined by

$$CN = \sum_{j \in \{\text{type B}\}} \frac{1}{1 + \exp\left[N_\mathrm{d}\left(r_{\mathrm{x},j} - R_\mathrm{coop}\right)\right]} \tag{31}$$

where $r_{\mathrm{x},j}$ is the distance between the ion and solvent molecule j and R_coop and N_d are parameters of the potential[29]. Like in the previous example, we expect the velocity v to be important if the crossing process is largely

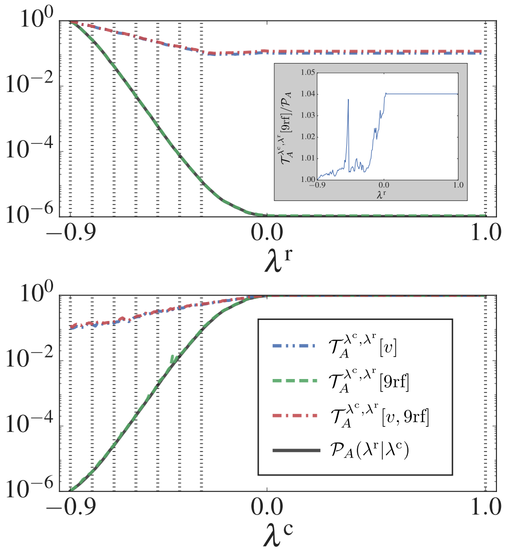


FIG. 3: (Color online.) The crossing probability, $\mathcal{P}_A(\lambda^r|\lambda^c)$, and the predictive capacities, $\mathcal{T}_A^{\lambda^c,\lambda^r}$, for the double well potential at (top) $\lambda^c = -0.9$ and (bottom) $\lambda^r = 1$. The position of the interfaces used in the RETIS simulations are indicated with dotted vertical lines. In both cases we find that $\mathcal{T}_A^{\lambda^c,\lambda^r}[v, 9\text{rf}] > \mathcal{T}_A^{\lambda^c,\lambda^r}[v] >> \mathcal{P}_A(\lambda^r|\lambda^c)$ and $\mathcal{T}_A^{\lambda^c,\lambda^r}[9\text{rf}] > \mathcal{P}_A(\lambda^r|\lambda^c)$ and in the inset in the top figure we show the enhancement of the predictive capacity, $\mathcal{T}_A^{\lambda^c,\lambda^r}[9\text{rf}]/\mathcal{P}_A(\lambda^r|\lambda^c)$, when using 9rf alone as the collective variable.

non-stochastic. This would be the case if the typical collision time with solvent molecules is larger than the time required to cross the reaction barrier. For dense systems, we expect that the second collective variable will be more important as it's directly linked to the height of the barrier[29].

In Fig. 4 we show the crossing probability and the predictive capacity for the three combinations of collective variables ($\{v\}, \{CN\}, \{v, CN\}$) using results from a RETIS simulation carried out as described in Ref. 29. In this case, we see that both variables improve the predictive capacity. However, the coordination number improves the predictive capacity more than the velocity of the progress coordinate. The distributions $t^{\lambda^c,\lambda^r}(CN)$, $r^{\lambda^c,\lambda^r}(CN)$, and $u^{\lambda^c,\lambda^r}(CN)$ for some of the λ^c, λ^r values are shown in Figs. 5 and 6. Fig. 5 shows the distributions with $\lambda^c = \lambda^A$ and different values for λ^r, while Fig 6 shows the distributions for λ^r fixed at λ^B and different values for λ^c. The values correspond to the black dots in the left-bottom panel of Fig. 4. Fig. 5 shows a clear cross-over for increasing λ^r. For λ^r values close to $\lambda^c = \lambda^A$, the reactive distribution $r^{\lambda^c,\lambda^r}(CN)$ is almost identical to the total distribution $t^{\lambda^c,\lambda^r}(CN)$. However, when λ^r is moved towards λ^B the unreactive distribution increases at the expense of the reactive distribution. Despite, as shown by the insets, the reactive distribution is always higher at the large coordination numbers.

Fig. 6 shows an opposite trend regarding the height of the distributions. The unreactive distribution is initially the largest but for increasing λ^c the reactive distribution rises at the expense of $u^{\lambda^c,\lambda^r}(CN)$. At large coordination numbers, the reactive distribution is always the largest just as in Fig. 5.

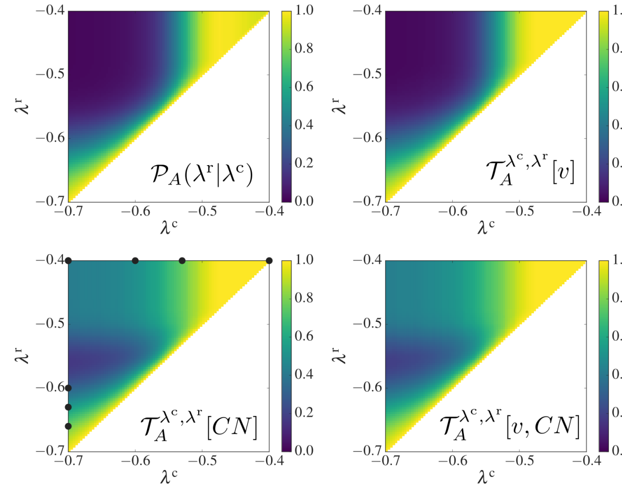


FIG. 4: (Color online.) The crossing probability, $\mathcal{P}_A(\lambda^r|\lambda^c)$, and the predictive capacities, $\mathcal{T}_A^{\lambda^c,\lambda^r}$, for the ion transfer potential: the crossing probability (top left), the predictive capacity using the velocity (v) of the progress coordinate as the collective variable (top right), the predictive capacity using the coordination number (CN) as the collective variable (bottom left), the predictive capacity using both v and CN as collective variables (bottom right). We used 85 sub-interfaces both for λ^r and λ^c. The histograms in the Ψ^N space were constructed using 20 bins for $0 < CN < 3$ and 20 bins for $-45 < v < 25$. The circles placed on the λ^r and λ^c axes in the bottom left figure indicate points where we have obtained the distributions in the CN space, shown in Fig. 5 and 6

Fig. 7 shows the intersections of Fig. 4 corresponding to a fixed $\lambda^c = -0.7$ and a fixed $\lambda^r = -0.4$. The results clearly show that the coordination number is a much better indicator for the ion-transfer reaction than the velocity along the reaction coordinate λ. Still, having knowledge of both parameters will improve the predictive capacity slightly more compared to the situation in which one only knows CN.

D. Numerical Example 3: Ab Initio MD of Water Dissociation

For this example, we have performed RETIS simulations of dissociation of water at a low density. Water was modeled with the BLYP functional[30,31] and a DZVP-MOLOPT basis set[32]. A plane-wave cut-off of 300 Ry was used and the simulations were performed at a low density with 8 water molecules in a cubic simulation box of 9.85 × 9.85 × 9.85 Å^3. Periodic boundary conditions

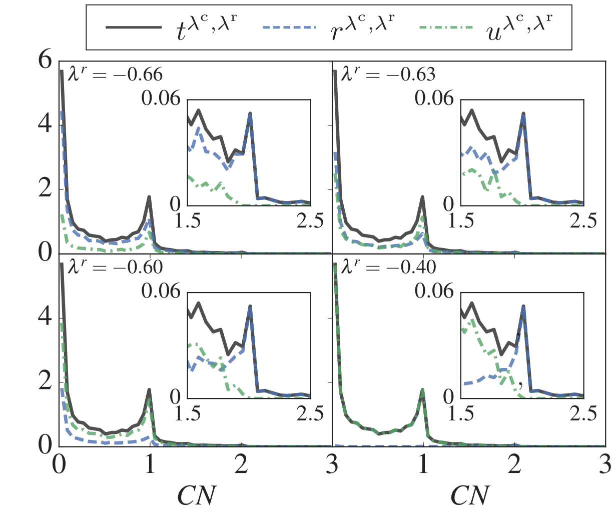


FIG. 5: (Color online.) Distribution of t^{λ^c,λ^r}, r^{λ^c,λ^r} and u^{λ^c,λ^r} using the coordination number (CN) as the collective variable for $\lambda^c = -0.7$ and $\lambda^r = -0.66$ (top-left), $\lambda^r = -0.63$ (top-right), $\lambda^r = -0.60$ (bottom-left) and $\lambda^r = -0.4$ (bottom-right). For the analysis, we used 22 sub-interfaces both for λ^r and λ^c, and the distributions shown here were obtained using 50 bins for $0 < CN < 3$.

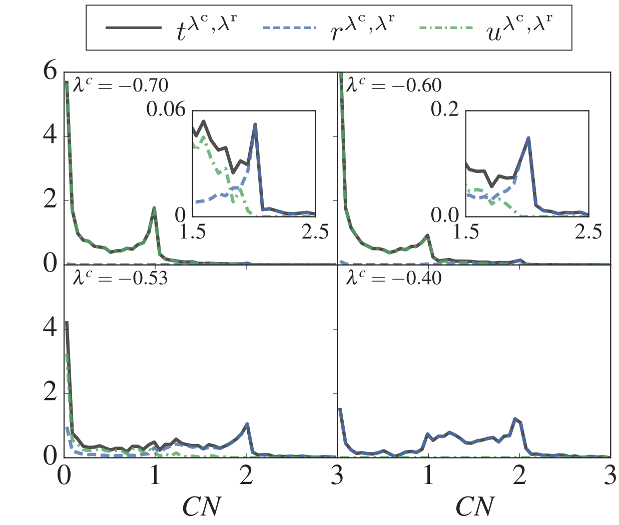


FIG. 6: (Color online.) Distribution of t^{λ^c,λ^r}, r^{λ^c,λ^r} and u^{λ^c,λ^r} using the coordination number (CN) as the collective variable for $\lambda^r = -0.4$ and $\lambda^c = -0.70$ (top-left), $\lambda^c = -0.60$ (top-right), $\lambda^c = -0.53$ (bottom-left) and $\lambda^c = -0.40$ (bottom-right). For the analysis, we used 22 sub-interfaces both for λ^r and λ^c, and the distributions shown here were obtained using 50 bins for $0 < CN < 3$.

were employed in all directions. The DFT-based MD simulations were carried out using the CP2K program package[33] with a time step of 0.5 fs and NVE dynamics. Shooting moves were performed by randomly reselecting the velocities at the shooting point from a Maxwellian distribution corresponding to a temperature of 600 K.

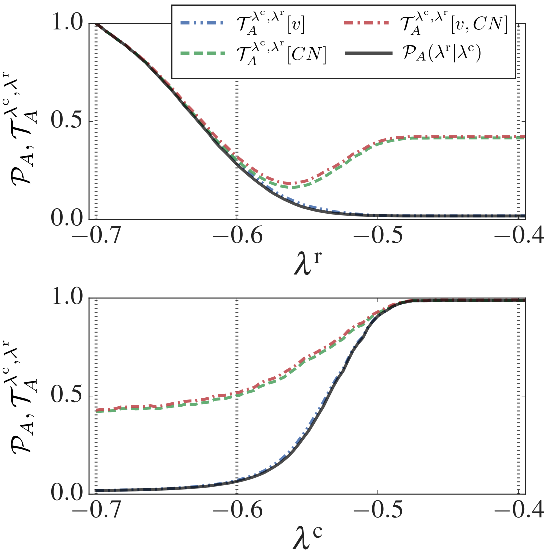


FIG. 7: (Color online.) The crossing probability, $\mathcal{P}_A(\lambda^r|\lambda^c)$, and the predictive capacities, $\mathcal{T}_A^{\lambda^c,\lambda^r}$, for the ion transfer potential at (top) $\lambda^c = -0.7$ and (bottom) $\lambda^r = -0.4$. The position of the interfaces used in the RETIS simulations are indicated with dotted vertical lines.

The progress coordinate was defined using the distances between oxygen and hydrogen atoms. We first assign each hydrogen to the oxygen atom it is closest to. This allows us to classify molecules, e.g. as H_2O, H_3O^+ or OH^-, and calculate bond lengths between hydrogen and oxygen. If the system only contains H_2O molecules, the progress coordinate is defined as the longest hydrogen-oxygen bond length. If the system contains H_3O^+ and OH^- species, the order parameter is taken as the shortest distance from the oxygen in OH^- to a hydrogen in H_3O^+.

For this example, we have considered one additional collective variable defined as the length, w, of the shortest hydrogen bond wire connecting 4 water molecules/species where one of the species contain the oxygen atom used for the progress coordinate. Hassanali et al.[34] highlighted the importance of compression of such wires for the recombination reaction and hypothesized that a similar phenomena is likely to be the rate-limiting step for autoionization. In order to obtain the hydrogen bond wire length, we first obtained all hydrogen bonds (defined as in Ref. 35) in the system and used this to create a graph of hydrogen-bond connected water molecules. The relevant hydrogen bond wire was obtained using the following criteria: (i) The wire should contain the oxygen atom used for the order parameter (identified as explained above) when the order parameter first crossed the 1.15 interface, (ii) the wire should contain 4 water molecules, (iii) the wire should be the shortest of the wires where criterion (i) and (ii) is met. The length of the wire was defined as the sum of the oxygen-oxygen distances of consecutive molecules in the wire.

The results of including this additional collective variable are shown in Fig. 8. The values of $\mathcal{T}_A^{\lambda^c,\lambda^r}[w]$, show that the dissociation reaction involves rare fluctuations in the hydrogen bonded network and as shown in the bottom-left and bottom-right figures, including the hydrogen bond wire length improves the predictive capacity compared by several orders of magnitude.

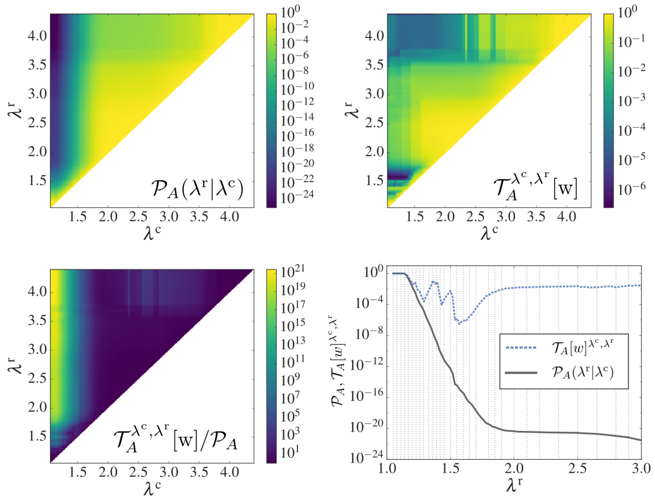


FIG. 8: (Color online.) The crossing probability, $\mathcal{P}_A(\lambda^r|\lambda^c)$, and the predictive capacity, $\mathcal{T}_A^{\lambda^c,\lambda^r}$, for the ab initio water simulations: (top left) the crossing probability, (top right) the predictive capacity using the hydrogen bond wire length (w) the collective variable (see the main text for the definition of this quantity), (bottom left) the enhancement of the predictive capacity, $\mathcal{T}_A^{\lambda^c,\lambda^r}[w]/\mathcal{P}_A(\lambda^r|\lambda^c)$, (bottom right) the predictive capacity and crossing probability as a function of λ^r for $\lambda^c = 1.05$ (positioned at the leftmost interface; the position of the interfaces used in the RETIS simulations are indicated with dotted vertical lines). We used 200 sub-interfaces both for λ^r and λ^c. The histograms in the Ψ^N space were constructed using 50 bins for $7 < w < 50$.

V. REDUCTION OF CVS AND RELATION TO THE ISOCOMMITTOR

Whenever a predictive set of CVs is obtained with a relatively low overlap value $\mathcal{S}_A^{\lambda^c,\lambda^r}[\Psi^N]$ for a certain set of interfaces λ^c, λ^r, one can attempt to reduce the number of CVs without raising the overlap value. The idea is graphically illustrated in Fig. 9 where we show how a two-dimensional CV-space can be projected on a single coordinate. In the figure the best possible one-dimensional coordinate is given as a linear combination of the previously examined coordinates Ψ_1 and Ψ_2. This approach can also be used to construct coordinates that are non-linear functions of the original CVs. However, in some cases it might be preferred to project on a more simple coordinate even if it is not the best one in terms of minimizing the overlap since the more complex functional form might be less intuitive.

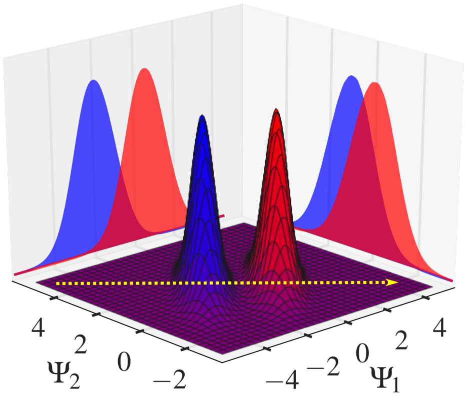


FIG. 9: (color online). Reduction of a set of CVs having a low overlap. Let red represent the distribution $r^{\lambda^c,\lambda^r}(\Psi_1,\Psi_2)$ and blue $u^{\lambda^c,\lambda^r}(\Psi_1,\Psi_2)$ where Ψ_1, Ψ_2 are two CVs. By selecting Ψ_1 as a single one-dimensional coordinate the projected distributions still show little overlap while the coordinate Ψ_2 is a much poorer choice since the projected distributions $r^{\lambda^c,\lambda^r}(\Psi_2)$ and $u^{\lambda^c,\lambda^r}(\Psi_2)$ will almost fully overlap. The optimal coordinate is shown by a yellow arrow and corresponds to the linear combination $c_1\Psi_1 + c_2\Psi_2$ where c_1 and c_2 are two constants.

The projection procedure can, in principle, also be used to find the committor, at least if sufficient path data is available. Suppose that we use the full phase space as orthogonal coordinates ($\Psi^N = x$) and take $\lambda^r = \lambda^B$. In that case $r^{\lambda^c,\lambda^B}(x)/t^{\lambda_i}(x) = P_B(x)$ where $P_B(x)$ is the phase space committor. For the overlap integral we get

$$\mathcal{S}_A^{\lambda^c,\lambda^B}[x] = \frac{1}{\mathcal{P}_A(\lambda^B|\lambda^c)} \int_{\lambda^c} \mathrm{d}x\, t^{\lambda^c}(x) P_B(x)(1 - P_B(x)) \tag{32}$$

Suppose we bin the full phase space such that the integral can be solved numerically as

$$\mathcal{S}_A^{\lambda^c,\lambda^B}[x] = \frac{\mathrm{d}x}{\mathcal{P}_A(\lambda^B|\lambda^c)} \sum_q s(x_q) \tag{33}$$

where

$$\begin{aligned} s(x_q) &= \frac{r^{\lambda^c,\lambda^B}(x_q) u^{\lambda^c,\lambda^B}(x_q)}{t^{\lambda^c}(x_q)} \\ &= t^{\lambda^c}(x_q) P_B(x_q)(1 - P_B(x_q)) \end{aligned} \tag{34}$$

is the unnormalized contribution of bin q belonging to phase point x_q. Naturally, the contribution to Eq. 33 of two bins corresponding to phase points x_1 and x_2 is given as

$$s(x_1) + s(x_2) = (1 - P_B(x_1))P_B(x_1)t(x_1) + (1 - P_B(x_2))P_B(x_2)t(x_2) \tag{35}$$

Now, our projection operations can basically be viewed as a process in which two or more bins are merged into a single bin in the reduced coordinate space. After merging these two bins we can write for the collective bin

$$\begin{aligned} s(x_1 + x_2) &= \frac{(r(x_1) + r(x_2))\,(u(x_1) + u(x_2))}{t(x_1) + t(x_2)} \qquad (36) \\ &= \left(\frac{(1 - P_B(x_1))t(x_1) + (1 - P_B(x_2))t(x_2)}{t(x_1) + t(x_2)}\right) \\ &\times \left(\frac{P_B(x_1)t(x_1) + P_B(x_2)t(x_2)}{t(x_1) + t(x_2)}\right)(t(x_1) + t(x_2)) \end{aligned}$$

The difference between Eq. 36 and Eq. 35 is

$$\begin{aligned} s(x_1 + x_2) - s(x_1) - s(x_2) = \\ \left(\frac{(P_B(x_1) - P_B(x_2))^2 t(x_1)t(x_2)}{t(x_1) + t(x_2)}\right) \end{aligned} \tag{37}$$

Naturally, this difference is always positive except if $P_B(x_1) = P_B(x_2)$, then it is zero. Therefore, any projection will increase the overlap unless it is done such that phase points having the same committor end up in the same bin after the projection. In other words, if after the projection only a single orthogonal coordinate is left while the overlap has not increased, then all points having the same value for this orthogonal coordinate must have the same committor value. As such, we basically obtain an intersection of the committor surfaces with the λ^c plane and the final one-dimensional Ψ is a descriptor of this committor.

VI. CONCLUSIONS

We devised a quantitative analysis method for identifying reaction mechanisms and initiation conditions for reactive events. The analysis is performed on the path sampling data that are already produced by path sampling simulations for computing reaction rates such as TIS[16], RETIS[17], and FFS[18]. Hence, a big advantage of our technique is that it does not require additional simulations which is generally needed for other analysis methods such a committor analysis. Also, our method does not require intensive iterations such as in the FFS-least-square estimation for determining the committor on-the-fly as linear or a polynomial function of predefined CVS during a FFS simulation. In contrast, our approach is a pure *a posteriori* method that can be applied after the simulation is finished, which allows for testing any possible set of CVs, which could either come from intuition after analyzing molecular trajectories or even from machine learning techniques.

Another advantage is that it is very flexible and also allows identifying momenta dependent variables which might be crucial steps in the reaction mechanism. The main idea is to determine probability distributions of first crossing points along order parameters orthogonal to the chosen reaction coordinate. Each plane with points having the same value of the reaction coordinate, also called interface, can be used to collect the first crossing points. Another plane further towards the product state can be used to set a condition of partial reactivity. Trajectories from the first crossing points with the first plane might or might not cross the plane defining partial reactivity. Based on this, the first crossing points are categorized and define “reactive” and “unreactive” distributions. Then, a simple overlap integral defines how well the orthogonal coordinates can help in the prediction of reactivity or not. Since crossing the full barrier can be a rare event, we showed how reweighting techniques can be used based on WHAM[24–26] to improve statistics. Moreover, the number of orthogonal coordinates can be reduced by applying projection operations which keep the overlap to its minimum. The latter approach, in principle, can also be applied to determine the phase space committor. We are aware that the analysis method described here, possibly with some adaptations, could be used in a wide range of different scientific fields such as economics and social sciences. Certainly, this is not the first method that tries, based on available data, to early identify events or parameters which possibly could predict whether something happens or not. One of such techniques is determination of receiver operator characteristic curves[36,37] which is a common method in signal detection theory. These methods, although having similar aims, are based on a rather different mathematical formulations to measure the quality of predictiveness of some parameters. In addition, it would be very instructive to compare the kind of information that can be subtracted from the predictive power method, described in this article, and the information obtained from likelihood maximization[4,5]. We plan to analyze possible analogies of these approaches in a future study.

The approach presented here allows one to get more valuable data from path sampling simulations and provides a mean to analyze reaction mechanism in a quantitative way. This output is likely to unravel hidden initiation events. Knowledge of these can then be exploited for designing new synthesis routes in which either new products are made or existing chemicals are generated with a lower energy cost or impact on the environment.

Appendix A: WHAM Approach for Path Sampling

The WHAM methodology is well explained in previous publications[24–26] and also the WHAM approach applied on path sampling simulations have been reported before. The derivation that we give here is, however, slightly different than reported elsewhere. We give it here for completeness and to show that the WHAM weights can be optimized using non-standard terms. These terms come in addition to the terms depending on the correlation number, which are standard but often omitted. Whether these more refined weights should be applied or not will mainly depend on the accuracy of the path simulations.

The WHAM approach is based on the idea that whenever different simulations produce the same output, the best numerical result should be a weighted average of these outputs in which the weights have to chosen in order to minimize the overall error. One way to derive these weights is to write an general expression of the overall error for arbitrary weights. The optimizing set of weights can then be found by minimizing this expression with respect to the weights under the condition that the sum of weights must be equal to one. Intuitively we can, however, also use the following argument. Suppose there are two simulations with different simulation lengths computing the same average. If the two simulations are equally efficient, it is obvious that the best overall result is obtained by taking a weighted average in which the weights are taken to be proportional to the simulation length. Reversely, since the error scales as the inverse square root of the simulation length, it makes sense to weight different types of simulations, possibly using different algorithmic approaches or biases, with the inverse square of their error: $\omega_i \propto \epsilon_i^{-2}$.

Now, consider a certain probability $p(\xi)$ which is for instance the probability that the system is within a bin as defined by the order parameter ξ. To improve the statistics we can apply biases in the sampling and unbias the results using a proper rescaling.

$$p_i^{\text{unb.}}(\xi) = Y_i(\xi) p_i^{\text{b.}}(\xi) \tag{A1}$$

where $p_i^{\text{b.}}$ is the biased distribution of simulation i and Y_i is the ξ-dependent scaling factor to obtain the i-th realization of the unbiased distribution $p_i^{\text{unb.}}$. If we assume that $Y_i(\xi)$ can be viewed as a constant not bearing any error, then the error in $p_i^{\text{unb.}}$ is simply $\epsilon(p_i^{\text{unb.}}) = Y_i \epsilon(p_i^{\text{b.}})$. Moreover, since the calculation of $p_i^{\text{unb.}}(\xi)$ is generally related to the average of a binary function (being 1 if the system visits the bin at ξ and zero otherwise), we can use the well known expression for its error (see e.g.[20,38])

$$\begin{aligned}
\epsilon(p_i^{\text{b.}}(\xi)) &= \sqrt{\frac{p_i^{\text{b.}}(\xi)\left((1 - p_i^{\text{b.}}(\xi)\right)}{n_i/\mathcal{N}_i}} \\
&= \sqrt{\frac{p_i^{\text{unb.}}(\xi)}{n_i' Y_i(\xi)}}
\end{aligned} \tag{A2}$$

Here, $\mathcal{N}_i$ is the effective correlation (also called statistical inefficiency) and $n_i' = n_1/[\mathcal{N}_i\left((1 - p_i^{\text{b.}}(\xi)\right)]$. Then, by taking the weights proportional to $\epsilon_i^{-2}\left(p_i^{\text{unb.}}(\xi)\right)$ with the condition $\sum_i \omega(\xi) = 1$ we get

$$\begin{aligned}
\omega_i(\xi) &= \frac{\left[Y_i(\xi)\,\epsilon\left(p_i^{\text{b.}}(\xi)\right)\right]^{-2}}{\sum_j \left[Y_j(\xi)\,\epsilon\left(p_j^{\text{b.}}(\xi)\right)\right]^{-2}} \\
&= \frac{\frac{1}{Y_i^2(\xi)} \frac{n_i' Y_i(\xi)}{p_i^{\text{unb.}}(\xi)}}{\sum_j \frac{1}{Y_j^2(\xi)} \frac{n_j' Y_j(\xi)}{p_j^{\text{unb.}}(\xi)}} \\
&\approx \frac{n_i' Y_i^{-1}(\xi)}{\sum_j n_j' Y_j^{-1}(\xi)}
\end{aligned} \tag{A3}$$

where we used Eqs. A1 and A2 and the fact that $p_i^{\text{unb.}}(\xi)$ should be similar for all i since these values should converge for each simulation to the true unbiased distribution $\rho(\xi)$.

Hence, the weighted average, Eq. 17, equals

$$\rho(\xi) = \frac{\sum_{i=1} n_i' \rho_i^{\text{b.}}(\xi)}{\sum_j n_j' Y_j^{-1}(\xi)} = \frac{\sum_{i=1} n_i'[\xi]}{\sum_j n_j' Y_j^{-1}(\xi)} \tag{A4}$$

where $n_i'[\xi]$ is the effective number of cycles in simulation i that visit bin ξ. Here, effective means that one counts all the cycles visiting bin ξ but finally divides this number by $\mathcal{N}_i\left((1 - p_i^{\text{b.}}(\xi)\right)$. The reduction of n_i and $n_i[\xi]$ with the correlation is standard though often omitted. Although $\mathcal{N}_i$ can be large, it cancels out in Eq A4 if the correlation number is similar for all i. The other factor $\left((1 - p_i^{\text{b.}}(\xi)\right)$ can generally be omitted whenever the bin-width is small enough such that $p_i^{\text{b.}}(\xi) \ll 1$.

Now let us come back to the crossing probability. For $\mathcal{P}_A(\lambda|\lambda_0)$ we can write:

$$\begin{aligned}
\mathcal{P}_A(\lambda|\lambda_0) &= \langle \theta(\lambda_{\max}(X) - \lambda) \rangle_{\varrho_0} \\
&= \langle \theta(\lambda_{\max}(X) - \lambda) \rangle_{\varrho_i} \mathcal{P}_A(\lambda_i|\lambda_0) \text{ for any } i < K(\lambda)
\end{aligned} \tag{A5}$$

where, in the second step, we applied a similar mathematical operation as the one in Eq. 20. Now, we can use Eq. A4 in which we replace Y_i with $\mathcal{P}_A(\lambda_i|\lambda_0)$ and replace the bin around ξ with an extended interval $[\lambda : \infty]$ for $\lambda_{\max}$:

$$\mathcal{P}_A(\lambda|\lambda_0) = \frac{\sum_{i=0}^{K(\lambda)} n_i'[\lambda]}{\sum_{j=0}^{K(\lambda)} n_j' [\mathcal{P}_A(\lambda_j|\lambda_0)]^{-1}} \tag{A6}$$

where

$$\begin{aligned}
n_i' &= \frac{n_i}{\mathcal{N}_i[1 - \langle \theta(\lambda_{\max}(X) - \lambda) \rangle_{\varrho_i}]} \\
&= \frac{n_i}{\mathcal{N}_i \langle \theta(\lambda - \lambda_{\max}(X)) \rangle_{\varrho_i}}
\end{aligned} \tag{A7}$$

is the number of generated path in simulation i reduced by a factor proportional to the statistical inefficiency and the fraction of trajectories with $\lambda_{\max} < \lambda$. Similarly,

$$n_i'[\lambda] = n_i' \langle \theta(\lambda_{\max}(X) - \lambda) \rangle_{\varrho_i} \tag{A8}$$

is the number of trajectories in simulation i with $\lambda_{\max} > \lambda$ scaled by the same factor. The bin at position ξ is here replaced by the region $\lambda_{\max} > \lambda$. The small bin-width assumption can, hence, not be made. In other words, we can't assume that $[1 - \langle \theta(\lambda_{\max}(X) - \lambda) \rangle_{\varrho i}] = \langle \theta(\lambda - \lambda_{\max}(X)) \rangle_{\varrho i} \approx 1$ for all i. In fact, it can even be zero which, unless it is based on bad statistics, implies that the crossing probability has reached a plateau. However, if it is based on bad statistics the infinite weight of $1/ \langle \theta(\lambda - \lambda_{\max}(X)) \rangle_{\varrho i}$ will influence the results in a negative way. We, therefore, used the simpler expression of Eq. 18 which involves n_i and $n[\lambda]$ instead of n'_i and $n'[\lambda]$.

Acknowledgments

The authors thank the Research Council of Norway project no. 237423 and the Faculty of Natural Sciences and Technology (NTNU) for support. Computational resources were granted by The Norwegian Metacenter for Computational Science (NOTUR), project NN9254K.

[1] Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.

[2] Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

[3] Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, *123*, 1–78.

[4] Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.

[5] Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 034109.

[6] Best, R.; Hummer, G. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732–6737.

[7] Weinan, E.; Ren, W.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2005**, *413*, 242–247.

[8] E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503–523.

[9] E, W.; Vanden-Eijnden, E. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420.

[10] Peters, B. *Annu. Rev. Phys. Chem.* **2016**, *67*, 669–690.

[11] Ma, A.; Dinner, A. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.

[12] Ferrario, M.; Haughney, M.; McDonald, I.; Klein, M. *J. Chem. Phys.* **1990**, *93*, 5156–5166.

[13] Luzar, A.; Chandler, D. *Phys. Rev. Lett.* **1996**, *76*, 928–931.

[14] Geissler, P.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, *291*, 2121–2124.

[15] van Erp, T.; Meijer, E. *Angew. Chem.-Int. Edit.* **2004**, *43*, 1659–1662.

[16] van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762–7774.

[17] van Erp, T. S. *Phys. Rev. Lett.* **2007**, *98*, 268301.

[18] Allen, R. J.; Warren, P. B.; ten Wolde, P. R. *Phys. Rev. Lett.* **2005**, *94*, 018104.

[19] Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.

[20] van Erp, T. S. *J. Chem. Phys.* **2006**, *125*, 174106.

[21] van Erp, T. S. In *Kinetics and Thermodynamics of Multistep Nucleation and Self-Assembly in Nanoscale Materials: Advances in Chemical Physics, Vol 151*; Nicolis, G and Maes, D,, Ed.; Advances in Chemical Physics; John Wiley and Sons, Inc: Hoboken, NJ, USA, 2012; Vol. 151; pp 27–60.

[22] Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 9236.

[23] Rogal, J.; Lechner, W.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. *J. Chem. Phys.* **2010**, *133*, 174109.

[24] Ferrenberg, A.; Swendsen, R. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

[25] Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[26] Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

[27] Peters, B. *Chem. Phys. Lett.* **2012**, *554*, 248–253.

[28] Mullen, R. G.; Shea, J.-E.; Peters, B. *J. Chem. Theory Comput.* **2014**, *10*, 659–667.

[29] Lervik, A.; van Erp, T. S. *J. Chem. Theory Comput.* **2015**, *11*, 2440–2450.

[30] Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[31] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

[32] VandeVondele, J.; Hutter, J. *J. Chem. Phys* **2007**, *127*, 114105.

[33] Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. *Wiley Interdiscip. Rev: Comput. Mol. Sci.* **2014**, *4*, 15–25.

[34] Hassanali, A.; Prakash, M. K.; Eshet, H.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20410–20415.

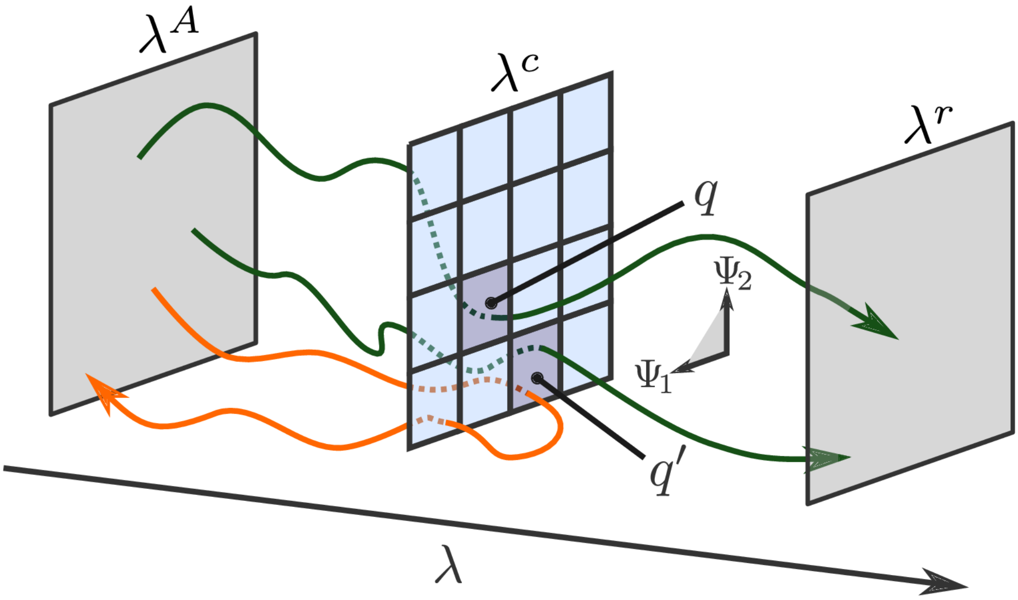
[35] Luzar, A.; Chandler, D. *Phys. Rev. Lett.* **1996**, *76*, 928–931.

[36] Egan, J. P. *Signal detection theory and ROC analysis*; Series in Cognition and Perception; Academic Press: New York, NY, 1975.

[37] Hallerberg, S.; de Wijn, A. S. *Phys. Rev. E* **2014**, *90*, 062901.

[38] Ruiz-Montero, M. J.; Frenkel, D.; Brey, J. J. *Mol. Phys.* **1996**, *90*, 925–941.

TOC-graphic


λ^A
λ^c
λ^r
q
q'
Ψ_2
Ψ_1
λ