



HAL
open science

Data-Driven Path Collective Variables

Arthur France-Lanord, Hadrien Vroylandt, Mathieu Salanne, Benjamin Rotenberg, A. Marco Saitta, Fabio Pietrucci

► **To cite this version:**

Arthur France-Lanord, Hadrien Vroylandt, Mathieu Salanne, Benjamin Rotenberg, A. Marco Saitta, et al.. Data-Driven Path Collective Variables. *Journal of Chemical Theory and Computation*, 2024, 10.1021/acs.jctc.4c00123 . hal-04549936

HAL Id: hal-04549936

<https://hal.sorbonne-universite.fr/hal-04549936>

Submitted on 17 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Data-driven path collective variables

Arthur France-Lanord,^{*,†,‡} Hadrien Vroylandt,[†] Mathieu Salanne,^{¶,§} Benjamin Rotenberg,[¶] A. Marco Saitta,[‡] and Fabio Pietrucci^{*,‡}

[†]*Sorbonne Université, Institut des Sciences du Calcul et des Données, ISCD, F-75005 Paris, France*

[‡]*Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, F-75005 Paris, France*

[¶]*Physicochimie des Électrolytes et Nanosystèmes Interfaciaux, Sorbonne Université, CNRS, 4 Place Jussieu F-75005 Paris, France*

[§]*Institut Universitaire de France (IUF), 75231 Paris, France*

E-mail: arthur.france-lanord@cnr.fr; fabio.pietrucci@sorbonne-universite.fr

Abstract

Identifying optimal collective variables to model transformations, using atomic-scale simulations, is a long-standing challenge. We propose a new method for the generation, optimization, and comparison of collective variables, which can be thought of as a data-driven generalization of the path collective variable concept. It consists in a kernel ridge regression of the committor probability, which encodes a transformation's progress. The resulting collective variable is one-dimensional, interpretable, and differentiable, making it appropriate for enhanced sampling simulations requiring biasing. We demonstrate the validity of the method on two different applications: a precipitation model, and the association of Li^+ and F^- in water. For the former, we show that global descriptors such as the permutation invariant vector allow to reach an accuracy far from the one achieved *via* simpler, more intuitive variables. For the latter, we show that information correlated with the transformation mechanism is contained in

the first solvation shell only, and that inertial effects prevent the derivation of optimal collective variables from the atomic positions only.

1 Introduction

Accurately modeling how matter transforms at the atomic scale is key in understanding many fundamental physical phenomena, ranging from phase transitions to chemical reactions. An ideal approach to studying such phenomena *in silico* would consist in simply observing the time evolution of an atomistic system, at a given set of thermodynamic conditions; *i.e.* an equilibrium molecular dynamics simulation. However, the free energy barriers associated with transformations are usually several times larger than the thermal energy $k_B T$, so that naturally observing and sufficiently sampling these rare phenomena is simply unachievable. Researchers have therefore developed advanced methods to bias the dynamics of a given system, allowing to effectively lower selected barriers of interest, and to enhance the sampling of rare events over the course of a molecular dynamics simulation. This set of methods¹ has been applied to many different physical problems, including the phase diagram of water² and its putative liquid-liquid phase transition,^{3,4} ice nucleation,⁵ protein folding dynamics,⁶ and chemical reactions governing prebiotic chemistry.⁷

The difficulty and subtlety associated with enhanced sampling methods comes from the high dimensionality ($3N$, where N is the number of particles) of configuration space, *i.e.* the space spanning all possible atomic configurations. Since it is virtually impossible to directly identify relevant barriers in such a high-dimensional space, the solution is to select *collective variables* (CVs), expressed as a function of the system's degrees of freedom. By doing so, the configuration space is projected onto one of much lower dimension. Good CVs should allow to distinguish basins of metastability, as well as transition state ensembles. The way they are chosen affects many crucial elements: i) the observed transformation mechanism in the context of enhanced sampling; ii) CVs are often used as a means of analyzing a reaction mechanism, be it observed from biased or unbiased dynamics. A poorly selected CV can therefore lead to a biased or incomplete rationalization

of the transformation mechanism, even in the context of unbiased simulations; iii) free energies of activation, ΔF^\ddagger , are sometimes extracted from gauge-invariant free energy profiles^{8–10} along collective variables, and used to interpret the kinetics of the system of interest. Using a sub-optimal CV will lead, in this context, to an underestimated ΔF^\ddagger for the transformation under scrutiny;⁹ iv) while in the context of the Eyring-Polanyi equation kinetic rates are independent of the choice of CV, thus a poorly selected one will have an associated transmission coefficient far from unity; v) sub-optimal CVs hamper the efficiency of enhanced sampling methods based on bias forces, since the latter are partly wasted along degrees-of-freedom not related to the transformation, leading to the unphysical overestimation of the free energy barrier height;¹¹ vi) on the contrary, sub-optimal CVs result also in an under-estimation of the free-energy barrier from the viewpoint of the ideal, exact calculation based on marginalizing the canonical probability density (see, *e.g.*, Equation 6 here below), due to overlap between metastable basins.¹² The latter effect has been exploited in a CV optimization approach based on rate minimization,¹³ and concerns, more generally, Langevin models describing the high-dimensional dynamics projected on CVs, constructed from statistical inference on unbiased MD trajectories.^{14,15}

Selecting a CV on solid grounds is therefore an active field of research, with many methods developed in the last decades, following different approaches. Identifying slow modes is one: this has been achieved using diffusion maps,^{16–19} variational approaches based on the overdamped Langevin equation^{13,20} or Markov state models²¹ – combined with independent component analysis^{22,23} and deep learning²⁴ –, or maximizing the spectral gap of a transition matrix.^{25,26} Another set of methods relies on (auto)encoders^{27,28} and the information bottleneck principle.^{29,30} A few techniques^{31–33} analyze fluctuations in the metastable states to extract optimal CVs.

Finally, a large body of work has been dedicated to deriving optimal collective variables based on the committor probability $p(\text{B}|\mathbf{X})$, *i.e.* in a system showing two metastable states A and B, the probability of reaching basin B first, conditioned on an initial point \mathbf{X} in configuration space. The committor is, by construction, the reaction coordinate of the transformation under scrutiny, as it encodes the entire dynamical process of transforming from one metastable state into the other.³⁴

Depending on how optimality is defined, it may or may not be an optimal collective variable: for instance, in the presence of large free energy barriers, the committor’s strong non-linearity near basins³⁵ is impractical to handle. Nevertheless, an intuitive approach is therefore to perform a regression of the committor using a flexible basis of collective variables; this has been pioneered by Ma and Dinner³⁴ using neural networks and a genetic algorithm. Later, Peters and Trout³⁶ devised a fitting approach based on transition path sampling³⁷ data and the transition path probability $p(\text{TP}|\mathbf{X})$, which is analytically related to the committor under the assumption of diffusive dynamics. This was later expanded to obtain CVs with maximal transmission coefficients,³⁸ or using cross-entropy minimization and regularization.³⁹ Recently, this approach has been recast in a reinforcement learning framework to allow the identification of CVs over the course of transition path sampling simulations,^{40,41} including an extension to extract free energies and rates.⁴² Committor-based methods have also been combined with slow process identification.⁴³

In this article, we propose a committor regression technique for CV selection based on kernel ridge regression, which as we show can be thought of as a data-driven generalization of path collective variables.⁴⁴ The resulting optimized collective variable is one-dimensional and differentiable, making it appropriate for biased simulations. It is interpretable, provided that the CV subspace it is based on is interpretable as well.

This paper is structured as follows: in Section 2, we introduce the method and discuss its connection to path collective variables. In Section 3, we apply the method to a two-dimensional toy potential for which we can compute the committor exactly. We then turn to more realistic examples, with a precipitation phenomenon using Lennard-Jones particles (Section 4), and the association of small, monoatomic ions in water (Section 5). Finally, we conclude and provide an outlook regarding possible extensions to the method, in Section 6.

2 From standard to data-driven path collective variables

Path collective variables^{44,45} describe the progress along a reaction pathway connecting metastable states. The main concept, summarized in the top part of Figure 1, is to define reference states, and to compute, for a given atomic configuration, the similarity to these reference states using a kernel function $K(\xi_i, \xi_j)$. The similarity is not computed on the raw atomic coordinates, but rather on a vector of collective variables of arbitrary dimension, ξ . For instance, coordination numbers of a specific set of atoms are often used for chemical reactions.^{45,46} Path CVs therefore involve two projections: from the atomic positions \mathbf{X} to intermediate CVs ξ (which live in what we call the *CV subspace*), and from ξ to the one-dimensional path CV denoted $s(\xi)$:

$$s(\xi) = \frac{\sum_{i=1}^N iK(\xi_i, \xi)}{\sum_{j=1}^N K(\xi_j, \xi)}. \quad (1)$$

A typical choice for the kernel is $K = \exp(-\lambda \|\xi_i - \xi\|_2^2)$, *i.e.* the radial basis function kernel, where λ is a bandwidth parameter, and $\|\xi_i - \xi\|_2^2$ the squared l^2 -norm. Conceptually, s describes the progress along the polygonal chain in ξ -space connecting all references. An additional quantity, z , is usually defined to express the distance to this pathway.⁴⁴ Traditionally, $\mathcal{O}(10^1)$ equidistant references in CV subspace are selected along the pathway using nudged-elastic-band-like algorithms;^{44,47} they can also be optimized during enhanced sampling simulations with an adaptive approach.⁴⁸ Later, a simpler and more explorative approach proposed to set

the number of reference states to the number of identified metastable states of interest.^{2,45} For instance, in the association/dissociation problem considered in Figure 1, the number of states is set to two. In any case, it is then assumed that the path CV will be able to interpolate correctly for configurations halfway between both states. A more important assumption is that the selected CV subspace ξ is able to resolve the transformation. At this stage, both are uncontrolled approximations that should be thoroughly verified, as a good CV should capture the transition state ensemble, whose configurations do not belong to either states.

It is therefore natural to consider including many reference states in the path CV definition.

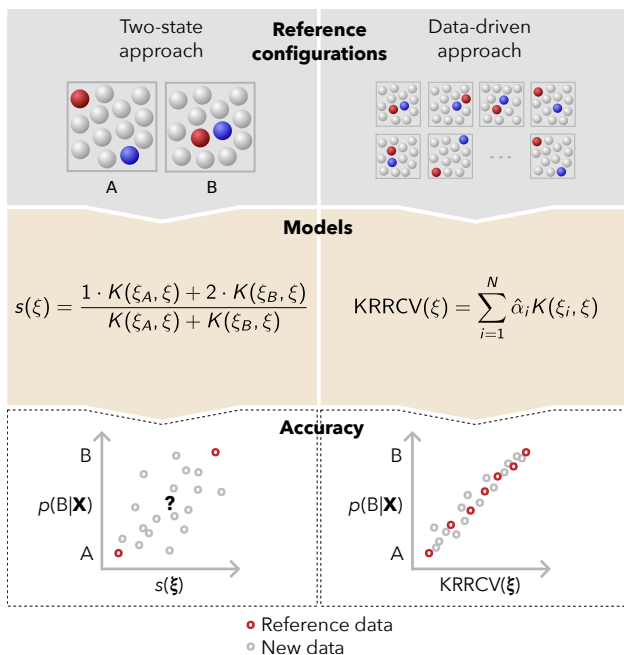


Figure 1: Standard, or two-state (left) and data-driven (right) path collective variables. Consider a system made of a red particle, a blue particle, and several gray particles. The red and blue particles can be in two metastable states: dissociated (A) or associated (B). Left: two reference states describing metastable states A and B are fed to the path CV. This will lead to a good description of configurations close to or in either metastable states, but the behavior outside of the basins is difficult to predict. This can be seen when comparing the committor $p(B|\mathbf{X})$ and $s(\xi)$: configurations far from the basins do not have close by reference states. Right: in a data-driven approach, a much greater number of references is considered, describing the full reaction process, from state A to state B. The committor has to be computed for each reference, and is fed to the path collective variable through kernel ridge regression. Agreement with the true committor function is improved as reference states are distributed over the whole reaction pathway.

As illustrated in Figure 1, this can only improve CV quality along the whole reaction path, *i.e.* the pathway, in CV subspace, that connects metastable states. However, a measure of the reaction progress is now needed. Indeed, s can be seen as a weighted average; for two reference states, i is simply set to 1 and 2. A configuration close to basin A will therefore have a s close to 1. If one wants to include reference states outside of metastable basins, it is therefore mandatory to use a measure of the progress along the reaction path, with some methods implemented to evenly space references in the CV subspace.^{49,50}

One way to evaluate the quality of a collective variable is to compare it to the committor probability, which is widely considered the optimal CV³⁶ for a system of two metastable states. For

a configuration already in basin A $p(B|\mathbf{X} \in A) = 0$, and $p(B|\mathbf{X} \in B) = 1$ for one in state B . At the transition state, $p(B|\mathbf{X} \in \text{TS}) = 0.5$. Unfortunately, for realistic – *i.e.* high-dimensional – systems, computing the committor for a given configuration is computationally expensive: one has to run a series of short independent simulations by randomly drawing initial velocities from the Maxwell-Boltzmann distribution. These simulations end when the system reaches either basins, and to reach a satisfying accuracy on the committor, it is necessary to generate hundreds-to-thousands of short trajectories when the committor value is not too close to zero or one.⁵¹ In the presence of high barriers the transition mechanism largely involves configurations with very tiny deviations of the committor from zero or one, requiring more sophisticated estimation approaches.³⁵

In addition, the committor, estimated numerically in this way, is not differentiable, which is problematic for biased simulations. While it is therefore nearly impossible to use the committor directly as a CV, one can compute values for an ensemble of reference configurations, and use these values as weights in the path CV. This is the main idea behind the present data-driven path collective variables.

Interestingly, the expression of s , which is shown in Figure 1, is that of a kernel regression (or kernel smoothing) estimator. Kernel regression is a class of non-parametric methods for non-linear regression. Since kernel regression improves with increasing amounts of data, it is even more natural to consider using many reference states. Surprisingly, the connection to kernel regression was not made in the original article on path CVs,⁴⁴ or in any other subsequent work. In greater detail, s is the Nadaraya-Watson estimator,^{52,53} which is the zeroth-order approximation to the local polynomial estimator, also called the "local constant" approximation. We switch to a global method in the form of kernel ridge regression (KRR), *i.e.* the combination of linear least squares with l^2 -norm regularization and the kernel trick. The KRR estimator is:

$$\hat{f}(\xi) = \sum_{i=1}^N \alpha_i K(\xi_i, \xi), \quad (2)$$

optimal parameters $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1, \dots, \hat{\alpha}_n]^T$ are obtained by solving the following linear system:

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \|y - K\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} = (K_{NN} + \lambda \mathbb{I}_n)^{-1} y, \quad (3)$$

where we introduced the kernel matrix of all pairs of data K_{NN} , the target vector $y = [p(\mathbf{B}|\mathbf{X}_1), \dots, p(\mathbf{B}|\mathbf{X}_N)]^T$, and the regularization parameter λ . In addition, we introduce a bandwidth matrix Σ such that our kernel becomes:

$$K(\xi_i, \xi) = \exp(-(\xi_i - \xi)^T \Sigma^{-1} (\xi_i - \xi)), \quad (4)$$

allowing more flexibility in reproducing the target data. In the following, Σ is always diagonal and reduces to a bandwidth vector $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_d]^T$, for $\xi \in \mathbb{R}^d$. In this way, each component in ξ has a corresponding weight which encodes its importance: a low σ means that the corresponding CV component is highly correlated to the committor. ξ can be composed of simple, intuitive collective variables, but can also include high-dimensional, abstract representations such as commonly used local descriptors for machine-learned interatomic potentials (*e.g.* atom-centered symmetry functions⁵⁴ and the smooth overlap of atomic positions (SOAP)⁵⁵), or global variants such as the permutation-invariant vector (PIV).^{2,56} Optimal values for $\boldsymbol{\sigma}$ and the regularization parameter λ are obtained through optimization, by minimizing a loss function based on a training set, distinct from the set of configurations used in KRR (the reference set).

3 The rugged Müller-Brown potential

As a starting point, we select the two-dimensional rugged Müller-Brown^{57,58} (rMB) analytical potential. Assuming overdamped Langevin dynamics, we can calculate the committor probability over the whole configuration space, as explained in Section 1 of the Supporting Information. This allows us to assess the quality of various regression models without projection errors, by considering the configuration space as the CV subspace. The rMB potential takes the form of the traditional Müller-Brown potential⁵⁹ with added ruggedness:

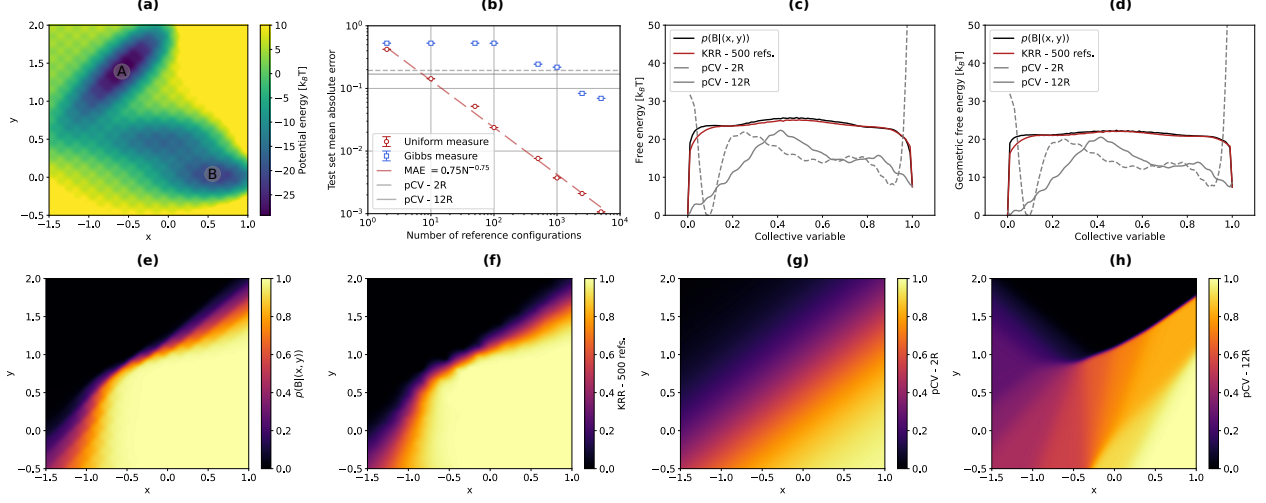


Figure 2: Path collective variables for the rugged Müller-Brown potential. (a) The potential energy surface. Shaded circles correspond to the metastable states definition used for the evaluation of the committor, (b) test set mean absolute error for different path collective variables as a function of the number of reference configurations included. Red circles and blue squares respectively correspond to models built on data sampled from the uniform and Gibbs measures. Error bars correspond to 95% confidence intervals of a distribution of predictions spanning ten different reference and training sets. The red broken line is a power law fit. The full and broken gray lines respectively correspond to the "12R" and "2R" path collective variables, (c) canonical and (d) geometric free energy profiles along various path collective variables, compared to the committor (small wiggles are due to the numerical integration and marginalization), (e-h) collective variables plotted in configuration space: (e) the committor as obtained from the backward Kolmogorov equation, (f) the kernel ridge regression based on 500 references, (g) the "2R" and (h) "12R" path collective variables.

$$\begin{aligned}
 V(x, y) = & \sum_{i=1}^4 D_i \exp [a_i(x - X_i)^2 + b_i(x - X_i)(y - Y_i) \\
 & + c_i(y - Y_i)^2] + \gamma \sin(2k\pi x) \sin(2k\pi y),
 \end{aligned} \tag{5}$$

where $D_{1...4} = [-400, -200, -340, 30]$, $a_{1...4} = [-1, -1, -6.5, 0.7]$, $b_{1...4} = [0, 0, 11, 0.6]$, $c_{1...4} = [-10, -10, -6.5, 0.7]$, $X_{1...4} = [1, 0, -0.5, -1]$, $Y_{1...4} = [0, 0.5, 1.5, 1]$, $\gamma = 9$, and $k = 5$. We define metastable states A and B, shown in Figure 2(a), as circles of radius 0.1 and centered respectively at $(-0.58, 1.39)$ and $(0.55, 0.05)$.

To construct datasets for KRR CVs (*i.e.* a reference set and a training set), we sample $p(B|(x, y))$ values from the uniform measure or from the Gibbs measure on the configuration space. Sampling

configurations in high-dimensional space – especially in the presence of potential energy barriers – is a difficult task, which is why we consider two radically different sampling strategies. As we will see in the next sections, for realistic systems, enhanced sampling strategies are leveraged. The test set is always the same for all models, consisting of 4000 data points sampled randomly. The detailed procedure for model optimization is reported in Section 2 of the Supporting Information. We assess the accuracy of the various path collective variables by comparing their mean absolute error (MAE) on the test set (Figure 2(b)), their value over configuration space 2(d,e)), and free energy profiles 2(f)). Free energy profiles are obtained by evaluating the collective variable’s marginal probability density over configuration space:

$$F(\xi) = -\beta^{-1} \ln \iint e^{-\beta V(x,y)} \delta(\xi(x,y) - \xi) dx dy, \quad (6)$$

where $\beta = 0.1$. We also report a "geometric" free energy profile, $F_g(\xi)$, distinct from the canonical definition of equation 6. The canonical free energy profile is not left invariant by a gauge transformation of a CV, which makes it impossible to compare different choices of ξ , even if they contain the same amount of information with respect to the reaction.^{9,10} The geometric definition incorporates this gauge invariance, which, as we will see later on, will be useful to compare quantities such as free energies of activation for different CVs. The geometric free energy profile can be derived through the application of the coarea formula, as has been discussed many times before:^{8,60}

$$F_g(\xi) = -\beta^{-1} \ln \iint e^{-\beta V(x,y)} |\nabla \xi(x,y)| \delta(\xi(x,y) - \xi) dx dy. \quad (7)$$

Both profiles are related in the following way:

$$e^{-\beta F_g(\xi)} = \langle |\nabla \xi(x,y)| \rangle_{\xi} e^{-\beta F(\xi)}, \quad (8)$$

where $\langle \cdot \rangle$ is an ensemble average. We also construct two variants of traditional path CVs: the first one ("2R") containing two references – one in each basin –, and a more advanced version

("12R"), containing twelve references equally spaced along the minimum energy path, including the exact same two as for the "2R" version. The latter one is assumed to be more advanced than the former since the minimum energy path is usually not available *a priori* for realistic, high-dimensional settings. References are respectively labeled 1,2 and 1,...,12; to compare with the committor, the path CVs are scaled to the $[0, 1]$ interval. The scalar bandwidth parameter λ is set so that the distance between the first and second references times λ is equal to 2.3, a common practice leading to smooth free-energy landscapes, typically with a good separation of reactants, products and transition states.^{44,45}

Results are shown on Figure 2, with the global ability in reproducing the committor of each method reported in Figure 2(b). The largest MAE attainable is roughly 0.5, since the test set is composed of values in the $[0, 1]$ interval with mean roughly equal to 0.5. As one would expect, the test set MAE decreases with increasing number of references. However, sampling from the Gibbs measure is demonstrated to be largely inefficient, which obviously shows the importance of enhanced sampling for realistic systems. Sampling from the Gibbs measure leads to selecting points mostly in the metastable states, where the committor is constant, while it rapidly varies near the separatrix. This is even more critical with a low temperature or large energy barriers.

Canonical path CVs, while they do not aim at reproducing the committor, perform relatively decently, with test set MAEs respectively equal to 0.194 and 0.169 for the "2R" and "12R" variants (see also Figure 2(g,h)). Using only two references, or aligned references in the CV subspace, leads to constant path CVs along the direction normal to the line connecting the references. Here, as can be seen in 2(e), there are non-negligible variations of the committor in directions normal to the line connecting both metastable states. In addition, with no measure of the reaction progress, one can only label references in a linear fashion.

As mentioned before, the true committor varies non-linearly, depending on the underlying potential energy surface and the temperature. This highlights the importance of extracting such information, possibly by computing committor values. We note that other methods are used for labeling references, for instance based on transition path sampling data.⁵⁰ With a "reasonable" number of

500 references (reasonable because it can be rather easily achieved with moderate computational cost for the realistic examples reported in the next sections), one can achieve excellent, yet still improvable, agreement with the true committor, with a test set MAE smaller than 10^{-2} . This also translates into a qualitative agreement across configuration space (Figure 2(e,f)), and quantitative agreement of free energy profiles (Figure 2(c,d)).

Features of the free energy profiles along path collective variables such as the free energy of activation are in good agreement with the committor. Depending on the system, this might not always be the case: in Section 3 of the Supporting Information, we show results for a three wells model potential for which simple path collective variables do not perform as well. Realistic, high-dimensional systems typically show many shallow intermediate states. Nevertheless, these results demonstrate the strength of path collective variables. Of course, this is valid only in the limit where the CV subspace is the configuration space itself, *i.e.* when there is no loss of information due to a projection.

To assess the robustness of the KRR approach, we add parasitic dimensions to the potential, uncorrelated to the committor:

$$V(x, y, \mathbf{z}) = V(x, y) + \sum_{j=1}^{d_+} z_j^2, \tag{9}$$

in which we vary d_+ from 10 to 1000. The test set mean absolute error as a function of d_+ and the number of references is presented in Figure 3. As one can see, variations along d_+ are relatively negligible. This mostly demonstrates the robustness of the bandwidth optimization approach, since the problem here is finding directions correlated with the committor in a high-dimensional space. Finally, as shown in Section 4 of the Supporting Information, we can also embed the two-dimensional potential in a five-dimensional space using non-linear transformations leading to large degeneracies, and still recover a low test set MAE. This is perhaps expected, since kernel ridge regression models can learn non-linear features.

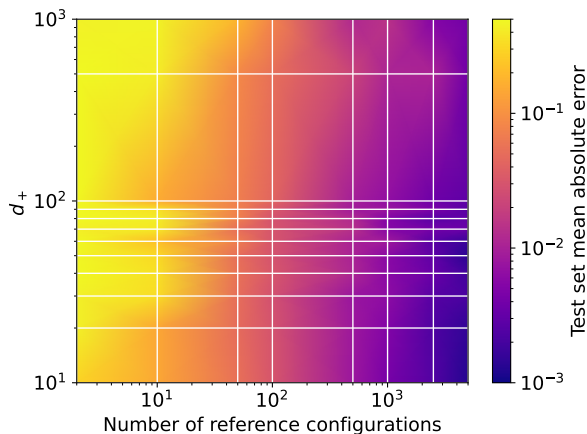


Figure 3: Test set mean absolute error for KRR models trained using different number of references and parasitic dimensions. The error is interpolated over data points represented by the intersections of the white lines (including $N = 2$ and $d_+ = 10$).

4 Precipitation of Lennard-Jones particles

We now turn to a more realistic system involving Lennard-Jones interactions, aimed at capturing the onset of a precipitation process. Precipitation is a critical phenomenon in inorganic or organic chemistry, as well as in biochemistry and metallurgy. For instance, in Li-ion batteries, decomposition products precipitate to form a solid interphase between a solid electrode and the liquid electrolyte. Here 20 particles represent the precipitating species, and 4086 smaller particles represent the solvent. The solute species interact strongly with one another ($\epsilon_{11} = 3.1$, $\sigma_{11} = 1.2$), while the solvent-solvent and solute-solvent interactions are weaker ($\epsilon_{12} = \epsilon_{22} = 1.0$, $\sigma_{12} = \sigma_{22} = 1.0$). The interaction cutoff is set to 6σ . All masses are set to 1. These parameters are chosen such that two metastable states separated by a free energy barrier exist: an associated state, where solute particles clump together to form an aggregate, and a dissociated state, where solvent particles coordinate solute particles. The integration time step (δt) is set to $2 \cdot 10^{-3}$, and all simulations are run at $T = 1$, $p = 1$, using the LAMMPS program⁶¹ (version 7 Aug 2019), with PLUMED^{62,63} (version 2.5.3) as an add-on to compute collective variables and perform enhanced sampling.

Our sampling strategy is the following: 1) converging free energy profiles from unbiased molecular dynamics, 2) sampling the putative transition state ensemble using a simple collective

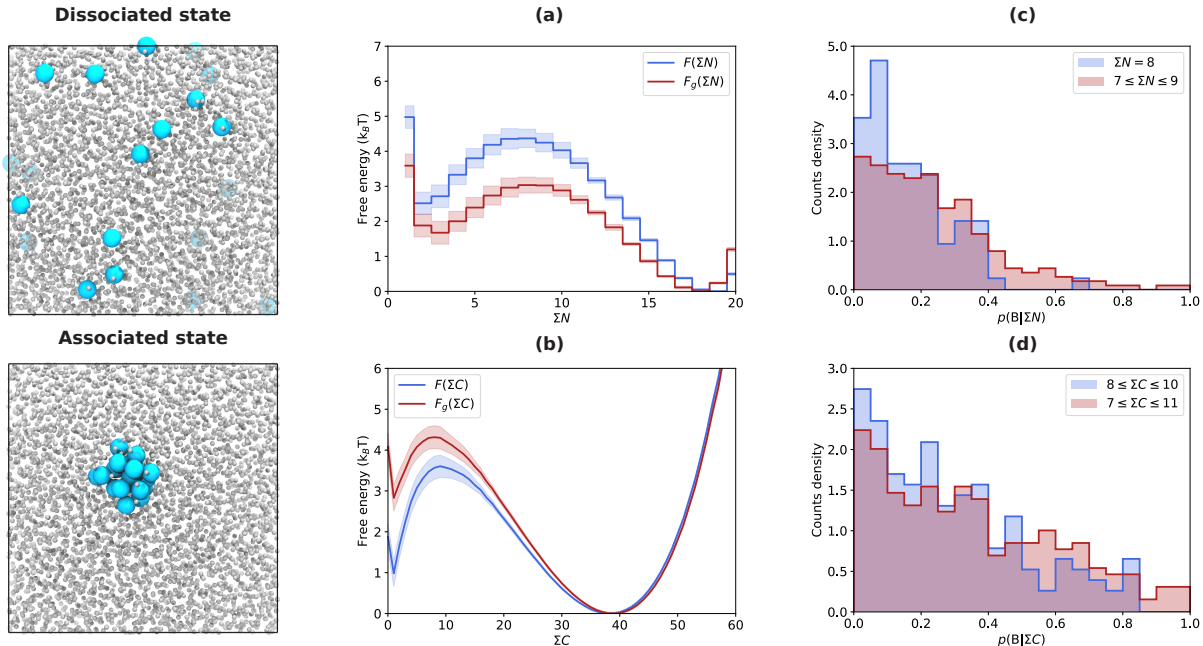


Figure 4: Lennard-Jones model of a precipitation phenomenon. Left panels: system snapshots in the dissociated and associated states. Large cyan and small gray spheres respectively represent solute and solvent particles. (a,b) Free energy profiles along collective variables ΣN and ΣC . (c,d) distributions of committor values for configurations at the putative transition state ensemble of ΣN and ΣC .

variable, 3) sampling close to the true transition state ensemble using transition path sampling. The committor is numerically estimated for configurations extracted from steps 2) and 3); these pairs of structure and committor constitute the datasets we then use to optimize and validate KRRCV models. All computational details regarding these steps are presented in Section 5 of the Supporting Information.

We start by obtaining free energy profiles along two simple one-dimensional collective variables, which we select to be properties of the largest aggregate. We use a depth-first search (DFS) clustering algorithm to identify the largest aggregate in a given configuration,⁶⁴ using a distance-based criterion to assess if particles are aggregated or not. Our two collective variables correspond to the number of particles in the largest cluster (ΣN), and of the sum of solute coordination numbers over solute particles in the largest cluster (ΣC). Profiles are obtained from long unbiased molecular dynamics in the canonical (nVT) ensemble.

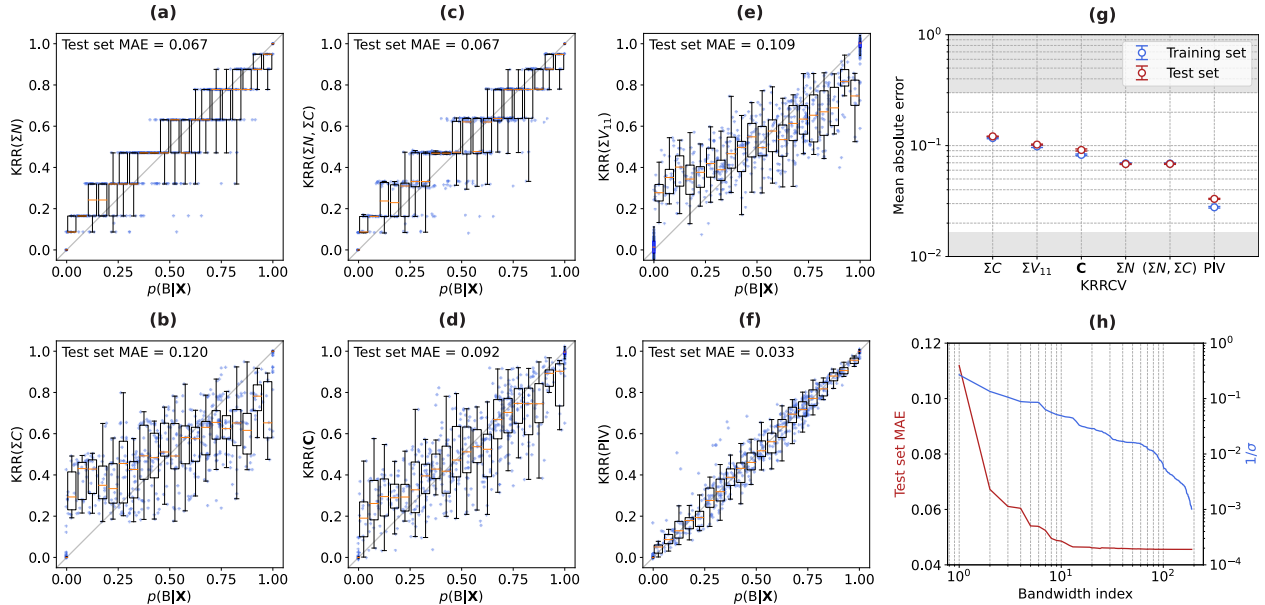


Figure 5: KRRCV models for the Lennard-Jones precipitation example. (a-f) Performance on the test set of models based on various collective variables. Small blue circles represent individual data points. The data is also shown using boxplots (the orange line is the median, the box extends from the first to the third quartile of the distribution, whiskers extend from the box edges to the last data point smaller or larger than ± 1.5 times the interquartile range). Data is partitioned in 0.05 wide bins, with two additional blue bins for the basins. (g) Training and test set mean absolute error, averaged over ten dataset splits, plotted as a function of the selected collective variable. Gray boxes show bounds as described in the main text. (h) Sorted inverse bandwidths (blue line), and corresponding test set mean absolute error of reduced models (red line) as a function of the number of included components.

Free energy profiles are reported in Figure 4(a,b). As can be seen, with both ΣN and ΣC , the associated state is considered more thermodynamically stable than the dissociated one; in addition, both states are separated by a barrier of a few $k_B T$.

Enhanced sampling involving biasing along ΣN is not achievable as it is a discontinuous quantity; we therefore perform umbrella sampling along ΣC ,⁶⁴ from which we extract 500 configurations. We also perform transition path sampling simulations, using the flexible-length aimless shooting algorithm.^{36,65} We initialize such simulations by generating 100 reactive trajectories connecting both basins. The basin definition is the following: $A : \Sigma N < 5$, $B : \Sigma N > 14$. From these initial reactive paths, we perform aimless shooting simulations, exploring configurations close to the separatrix. We select 500 configurations from umbrella sampling simulations, and 1000 con-

configurations from transition path sampling – 500 accepted, 500 rejected – for which we compute the committor. The evaluation of the committor is performed by repeated, independent simulations for which initial velocities are drawn from the Maxwell-Boltzmann distribution, and the system is evolved in the nVT ensemble, until a basin is reached. For each configuration, we perform 200 simulations. In the end, we obtain a dataset of 1500 configurations and corresponding committors.

We are also interested in seeing how well ΣN and ΣC capture the true transition state ensemble. For that purpose, from the umbrella sampling dataset, we select configurations at the putative transition state for both simple collective variables using the following conditions: $\Sigma N = 8$ (85 configurations), $7 \leq \Sigma N \leq 9$ (227 configurations), $8 \leq \Sigma C \leq 10$ (153 configurations), and $7 \leq \Sigma C \leq 11$ (259 configurations). These conditions are selected according to the free energy profiles, and correspond to configurations at or close to the transition state ensembles according to both collective variables. Results, in the form of distributions of committor values for these sets of configurations, are reported in Figure 4(c,d). A good CV would lead to a sharp, unimodal distribution centered at $p(B|\mathbf{X}) = 0.5$, meaning that its putative transition state ensemble overlaps well with the actual one. Here, the distributions largely lean towards the dissociated state: that most configurations supposed to be transition states according to these CVs are in fact close to one of the metastable states. This means that ΣN and ΣC lack important information mandatory to appropriately describe the transition between both states.

From our configuration-committor dataset, we optimize various KRR models, based on different CV subspaces: ΣN , ΣC , the pairwise interaction energy between all solute particles (ΣV_{11}), a combination of ΣC and individual solute coordination numbers for all solute particles (\mathbf{C} , $d = 21$), and the permutation invariant vector⁵⁶ of all solute particles (PIV, $d = 190$). The PIV consists of all pairwise distances of the group of atoms, sorted to enforce permutation invariance. While a cutoff function is usually used on the distances to allow focusing on a certain range, here we simply use the inverse distance, which smoothly converges to zero with increasing distance. The detailed procedure for model optimization is reported in Section 2 of the Supporting Information.

Performances of the various KRR models are reported in Figure 5(a-f), comparing test set

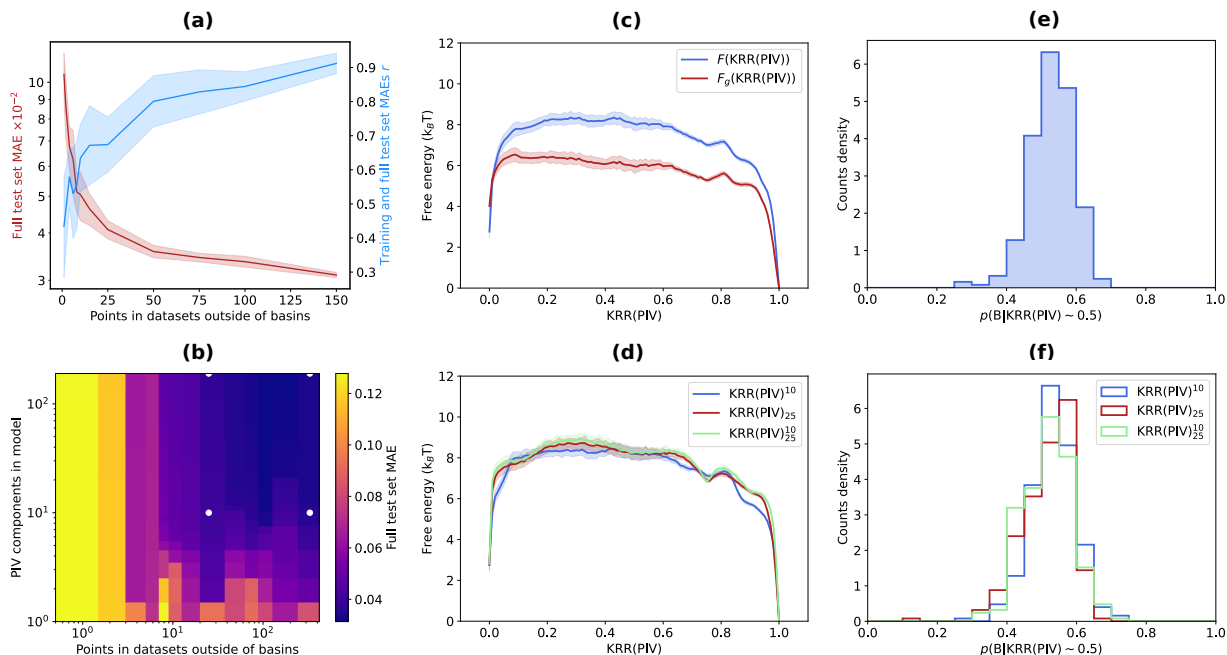


Figure 6: Performance of various KRR(PIV) models. (a) Evolution of the full test set MAE as a function of the number of points in the reference and training sets not in metastable basins (red curve). Also shown is the corresponding evolution of the Pearson correlation coefficient between the training set and full test set MAEs (blue curve). Results are averaged for 20 dataset splits. (b) Full test set MAE as a function of the number of PIV components included in the model (sorted by bandwidth importance), and number of points in datasets outside of basins. White dots correspond to selected models for further analyses. (c) Free energy and geometric free energy profiles along KRR(PIV). (d) Free energy profiles along reduced KRR(PIV) models. (e) Distribution of committor values for configurations at the putative transition state ensemble of KRR(PIV). (f) Same as (e), for reduced KRR(PIV) models.

committor values with model predictions. Training and test set mean absolute errors are reported in Figure 5(g). Two bounds on the MAE values are derived, and represented on Figure 5(g) as gray boxes. The upper bound corresponds to the MAE of a naïve model always returning the average of its training set. The lower bound, which represents a limit on the meaningful accuracy a model can reach, is due to the uncertainty on the committor estimation, due to the finite number of trials. Here, as demonstrated in Section 6 of the Supporting Information, the MAE on the test set distribution of committor values calculated with 200 trials is about 0.017. First, from Figure 5(g), it is apparent that CVs rank differently in correlating with the committor. The worst performing ones are ΣC and ΣV_{11} ; looking at Figure 5(b,e), they seem weakly positively correlated

with the committor, with larger deviations closer to the basins ($p(B|\mathbf{X}) = 0$ and 1). ΣN (Figure 5(a)) performs better, in particular close to the basins. Combining ΣN and ΣC (Figure 5(c)) as a two-dimensional CV does not seem to improve the performances, which means that all the relevant information relative to the transition in ΣC is already captured by ΣN . The more complicated list of solute coordination numbers \mathbf{C} (Figure 5(d)) performs as well as ΣN . Contrarily to the latter, it is however a continuous quantity, which makes it appropriate for biasing. Finally, the PIV based KRR model (Figure 5(f)) performs significantly better than all other collective variables, with a test set MAE close to the lower bound. A key difference between \mathbf{C} and the PIV collective variables, is that they are composed respectively of single-particle (an atom’s coordination number) and two-particle (an interatomic distance) descriptors. While this leads to having to handle a descriptor of much larger dimensionality, the difference in correlation to the committor hints to the importance of two-particle features in properly describing a model precipitation transformation.

An analysis of the bandwidth components of the PIV-based model (Figure 5(h)) reveals that most of the relevant information is contained in about 10 pairwise distances, out of 190 in total. The analysis proceeds in the following way: inverse bandwidths are sorted in decreasing order; the largest valued components are the most important. Then, models containing only the n ($1 \leq n \leq 190$) most important components are trained. Finally, the test set MAE of each model is plotted as a function of the number of components included. There is a good correlation between inverse bandwidth values and the performance of the resulting reduced model. This shows that the descriptor can be significantly compressed, allowing to bypass the costly evaluation of all pairwise distances.

In addition, it is beneficial to minimize the amount of reference configurations in the kernel ridge regression model, most importantly to reduce the amount of committor evaluations, but also to decrease training and inference times. To this end, we trained PIV models with varying amounts of data points outside of basins (both for reference and training sets), keeping the number of points in basins constant, where the committor can be evaluated at no cost. Results are presented in Figure 6(a,b). As expected, predictions are poor for datasets only including basin configurations. When

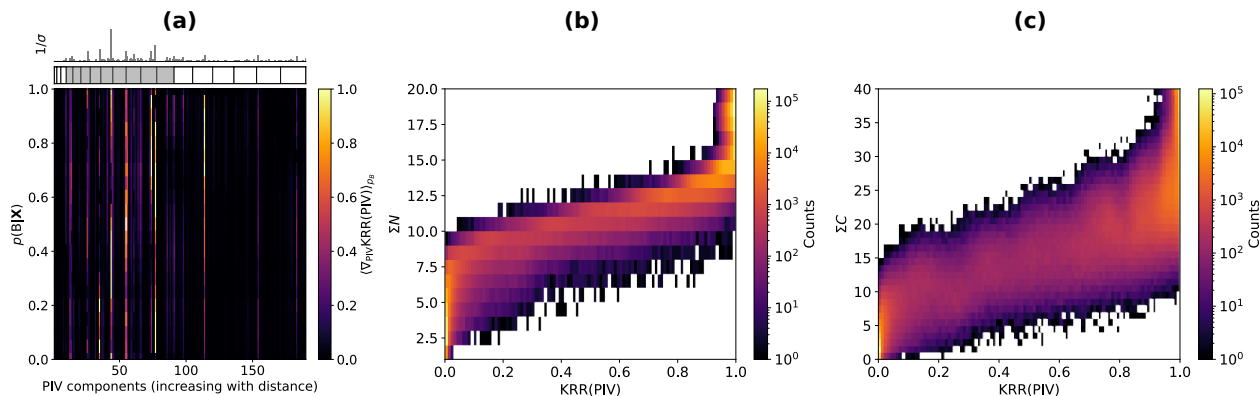


Figure 7: (a) Absolute gradient in PIV space of the KRR(PIV) model, averaged and normalized along $p(\mathbf{B}|\mathbf{X})$ bins computed over the test set. The greater the PIV component, the greater the distance out of all pair distances of solute particles. Upper panels: inverse bandwidth values for each PIV components, and corresponding number of particles (white and gray boxes), ranging from 2 (left) to 20 (right) particles. The gray area, from 5 to 14, corresponds to the range between basins definition. (b,c) 2d histograms of configurations sampled from unbiased molecular dynamics, along KRR(PIV) and either ΣN (b) or ΣC (c).

increasing the amount of points outside of basins, the accuracy of the model rapidly increases. In addition, the correlation between the reduced training set and full test set MAEs steadily increases. From Figure 6(b), it is clear that a range of models with reduced dimensionality or data points can achieve similar performance on the full test set. Eventually, we selected four models, represented as white points on Figure 6(b), for which we performed subsequent tests: the original one (190 PIV components and 330 data points), one with only 10 PIV components preserved ($\text{KRR}(\text{PIV})^{10}$), one with only 25 data points outside of basins ($\text{KRR}(\text{PIV})_{25}$), and one fully reduced model ($\text{KRR}(\text{PIV})_{25}^{10}$). Ideally, selecting an appropriate number of reference points would not require obtaining a large test set. We therefore propose, in Section 5.8 of the Supporting Information, an empirical approach for model selection based on training set error and noise level. Finally, modifying the loss function for bandwidth optimization by increasing the weight on data points outside of basins could allow to obtain satisfying models with even less reference configurations.

To further confirm the quality of the newly obtained CVs, we perform committor analyses. To sample new configurations at the putative transition state ensembles of KRRCVs, we implement the approach in the `hack-the-tree` branch of PLUMED.^{63,66} Enhanced sampling simulations are automatically enabled as long as the input collective variables forming the subspace are dif-

ferentiable; we therefore performed umbrella sampling on all four KRRCV models based on the PIV, and obtained datasets constrained at $0.45 \leq \xi \leq 0.55$. We then compute the committor of the sampled configurations, for both collective variables; committor distributions are reported in Figure 6(e,f). These are in stark contrast with those presented in Figure 4(c,d). For all four models, the distribution is unimodal and sharply peaked at 0.5, which demonstrates that the projection of the configuration space onto this collective variable preserves the features of the transition state ensemble. Furthermore, we evaluate free energy profiles along these CV models using the previously obtained unbiased molecular dynamics simulations; results are shown, for the original PIV model, in Figure 6(c). The profile, mostly flat, and rapidly varying close to the basins, is typical of committor approximations.⁶⁷ Profiles obtained for the reduced models (Figure 6(d)) are very similar. It should be noted that such free energy profiles – showing strongly-varying free energy gradients close to the basins – are inconvenient to handle using enhanced sampling methods such as metadynamics or umbrella sampling. For committor models, it is common practice to alleviate this problem using sigmoid functions such as the error function³⁸ or the logistic function;⁴⁰ the same can certainly be done for KRRCVs. The ensemble-averaged gradients along Cartesian coordinates of ΣN , ΣC , and KRR(PIV) are reported in Figure S3 of the Supporting Information. These gradients are used to compute the geometric free energy profile, and intuitively, should be zero in metastable basins, and for monotonous CVs like the committor, maximal at the transition state ensemble. Obviously, this is here only achieved by KRR(CV). Finally, the free energy of activation values for various collective variables (reported in Table 1) are significantly different, even when enforcing gauge-invariance in the free energy profile definition. Additional information is therefore encoded in the KRR(PIV) model, for which ΔF^\ddagger is maximized. Since the reaction rate is independent of the collective variable, using the Eyring-Polanyi equation, the transmission coefficients of ΣN and ΣC should be smaller than the one of KRR(PIV). Recrossings due to a poor choice of collective variable (and therefore separatrix) are therefore reduced.

A kernel ridge regression approach also ensures the interpretability of the model, since a single parameter (the bandwidth) rates the importance of each component of the collective variable space.

Table 1: Free energy of activation, for both free energy profile definitions, and for three collective variables. The error estimates are 95% confidence intervals computed through uncertainty propagation, assuming zero covariance.

| ξ | $\Delta F^\ddagger (k_B T)$ | $\Delta F_g^\ddagger (k_B T)$ |
|------------|-----------------------------|-------------------------------|
| ΣN | 1.86 ± 0.21 | 1.36 ± 0.21 |
| ΣC | 2.61 ± 0.21 | 1.48 ± 0.22 |
| KRR(PIV) | 5.59 ± 0.21 | 2.52 ± 0.23 |

For more involved estimators, such as neural networks, this is far from being the case. Obviously, in the end, interpretability will be dictated by the abstractness of the selected collective variable: it is easier to extract knowledge from a small list of simple, intuitive descriptors than from broad, general representations such as a PIV. In the top panel of Figure 7(a), we display the inverse bandwidth values of the KRR(PIV) model, along with the corresponding number of particles in the PIV displayed as white or gray boxes (the number of distance components grows as $N(N-1)/2$, where N is the number of particles in the system). Most of the important bandwidths are in a range corresponding to 5 to 14 particles in the largest cluster (gray boxes), which is precisely the condition on ΣN used for basin definition. In addition, we also evaluate the gradient in PIV space of the KRR(PIV) model along $p(\mathbf{B}|\mathbf{X})$, by binning the test set configurations. The absolute gradient, normalized for each committor bin, is displayed in the bottom panel of Figure 7(a). This analysis allows to identify the most important PIV components along the reaction process. Although there is some dispersion, along the transition from the dissociated state to the associated state, important components seem to concern larger distances with increasing transformation progress.

Finally, we can use the KRR(PIV) as a proxy to the committor, to analyze configurations along the reaction process. To this end, we perform an analysis over configurations sampled from unbiased molecular dynamics, here mostly to assess the quality of simple collective variables (ΣN and ΣC) in describing the transformation. We emphasize that a much more thorough analysis using other collective variables can be performed easily, if the interest is in understanding the transformation process, since an inexpensive-to-compute proxy to the committor is now available. 2d histograms along KRR(PIV) and either ΣN (b) or ΣC (c) are presented in Figure 7(b,c), computed

over the $6 \cdot 10^6$ configurations sampled. Looking at the distribution of ΣN at $\text{KRR}(\text{PIV}) = 0.5$, it is interesting to notice that for most configurations, $\Sigma N \approx 10 - 11$, while the top of the barrier shown in Figure 4(a) corresponds to $\Sigma N = 8$. This also explains the committor distribution in Figure 4(c), skewed towards the dissociated state. Overall, for constant $\text{KRR}(\text{PIV})$ values, the ΣN distributions are rather unimodal, although skewed towards large values. For ΣC , the distributions are very broad over the whole transformation. The basins also overlap with the transition zone. This translates to poor correlation with the committor, and inappropriateness in describing the transformation under scrutiny.

5 Ion association in solution: LiF in water

Finally, we investigate the association of Li^+ and F^- ions in solution; this mechanism is an important precursor of the formation of the solid electrolyte interphase (SEI) in Li-ion batteries,^{68,69} its understanding is therefore crucial to design devices with greater lifetimes. We select water as a solvent, both for simplicity, and because an LiF-rich SEI forms at the negative electrode in Li-ion batteries based on aqueous electrolytes.⁷⁰ LiF, being composed of the smallest monovalent inorganic ions, has a very high lattice energy, and is therefore only very moderately soluble in water (about 1 g/L at room temperature⁷¹). From the fundamental point of view, the association of atomic, monovalent ions in water is a well-known illustration of the counter-intuitive failure of simple collective variables – namely the interionic distance – in describing transformations.^{67,72,73} Arguments put forward include solvent degrees of freedom correlated with the transition as well as inertial effects, *i.e.* when the time-derivative of degrees of freedom has to be accounted for.

We follow a similar sampling strategy as in section 4: we first converge the free energy profile along the interionic distance (r) by running unbiased molecular dynamics simulations. Then, we sample configurations with $r = r^* \approx 2.63 \text{ \AA}$, the putative transition state along this collective variable, using umbrella sampling. Finally, we sample configurations near the true separatrix using aimless shooting. Committors are then computed using repeated molecular dynamics for con-

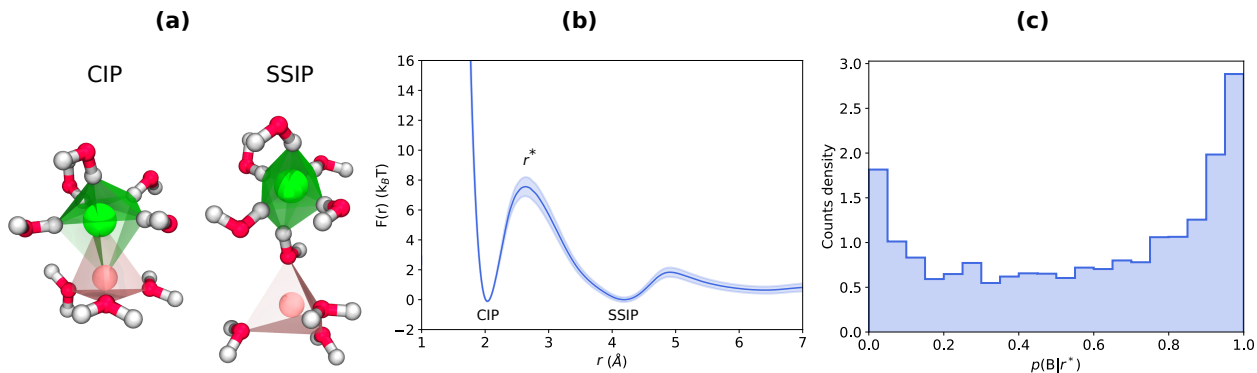


Figure 8: LiF association in water. (a) Contact ion pair (CIP) and solvent-separated ion pair (SSIP). Pink, green, red, and white spheres respectively represent Li, F, O, and H atoms. (b) Free energy profile along the interionic distance (r). (c) Distribution of committor values for configurations at the putative transition state ensemble of r sampled using umbrella sampling.

figurations sampled using both methods. Details regarding all computational steps are included in Section 7 of the Supporting Information. The ions are modeled using the Joung-Cheatham⁷⁴ potential, and water using the SPC/E rigid model.⁷⁵

The free energy along the interionic distance is reported in Figure 8(b). At short distance, ions form a first metastable state, the contact ion pair (CIP). Separated from the CIP by a free energy barrier of more than $7k_B T$ is the solvent-separated ion pair (SSIP) metastable state, where one water molecule coordinates both ions. Further increasing the distance leads to completely separating the ions. We define both states using the following criteria: CIP, state A: $r \leq 2.036 \text{ \AA}$, SSIP, state B: $r \geq 4.190 \text{ \AA}$. These distances correspond to the bottom of the free energy wells for both states. Other state definitions are possible, such as the interionic distances corresponding to the minima of the wells plus $k_B T$, but a stringent state definition allows to clearly separate the transition regime and the metastable states. The distribution of committor values for configurations sampled at r^* using umbrella sampling is reported in Figure 8(c). The distribution is bimodal, with sharp peaks at $p(B|r^*) = 0$ and 1, and mostly flat otherwise. Similar distributions have been recovered for other monovalent ions in water, using different interatomic potentials.^{67,72,73}

Interestingly, configurations sampled close to the true transition state ensemble correlate fairly well with the interionic distance. This is demonstrated in Figure 9(a), showing results from aimless shooting simulations. At r^* , configurations sampled from aimless shooting are distributed

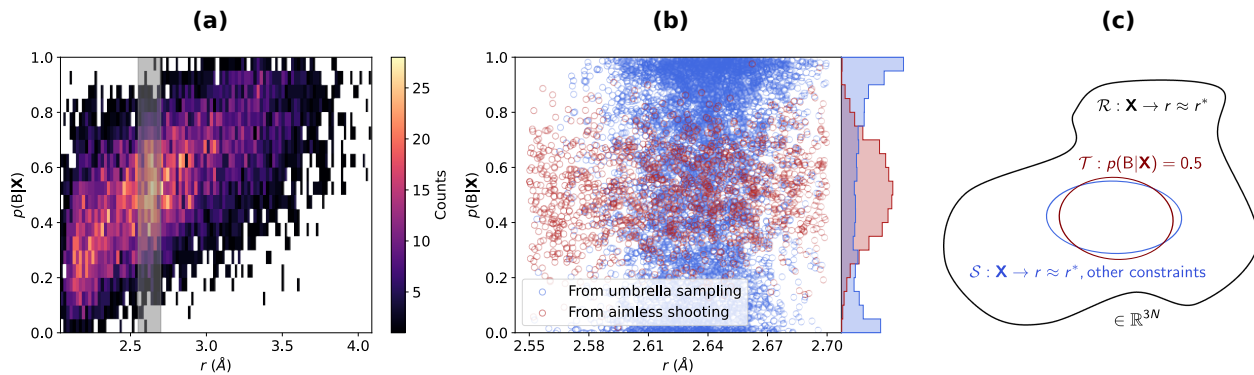


Figure 9: (a) Correlation between r and $p(\mathbf{B}|\mathbf{X})$ for configurations sampled from aimless shooting simulations. The gray zone corresponds to $r \in [2.55, 2.70] \approx r^*$. (b) Correlation between $r \approx r^*$ and $p(\mathbf{B}|\mathbf{X})$ for configurations sampled from aimless shooting (red circles) and umbrella sampling (blue circles) simulations. Distributions are also displayed on the right side, with matching colors. (c) Hypothesis regarding the hierarchy of ensembles in configuration space.

in a unimodal fashion centered at $p(\mathbf{B}|\mathbf{X}) = 0.5$, in stark contrast with configurations extracted from umbrella sampling, as shown in Figure 9(b). We also verify that the umbrella sampling distribution corresponds to the the distribution under the same conditions obtained from unbiased equilibrium simulations, in Section 7.7 of the Supporting Information. This could explain why collective variable optimization techniques based exclusively on transition path sampling data show fairly good correlation between the committor and the interionic distance⁴⁰ – not forgetting that the ions and interatomic potentials are not the same. From these observations, we formulate the following hypothesis regarding what is lacking for the proper description of the dissociation process, schematized in Figure 9(c): the subset of configuration space \mathcal{R} matching the $r \approx r^*$ criterion includes the transition state ensemble \mathcal{T} . A subset $\mathcal{S} \subset \mathcal{R}$ approximating \mathcal{T} would match the $r \approx r^*$ condition, with additional unknown constraints on degrees of freedom related to water. Identifying these constraints has been the purpose of many studies.^{40,67,72,73}

Following these observations, we train six KRRCV models based on different collective variables: i) r only ($d = 1$), ii) r and f_p , the interionic force projected on the vector connecting both ions ($d = 2$), iii) a list of scalar collective variables (r , Li^+ and F^- 's hydrogen and oxygen coordination numbers, the number of water molecules coordinating both ions, and the solvent-contributed Madelung potential⁷⁶ on Li^+ and F^- ($d = 8$), iv) the PIV of the subsystem composed of both ions

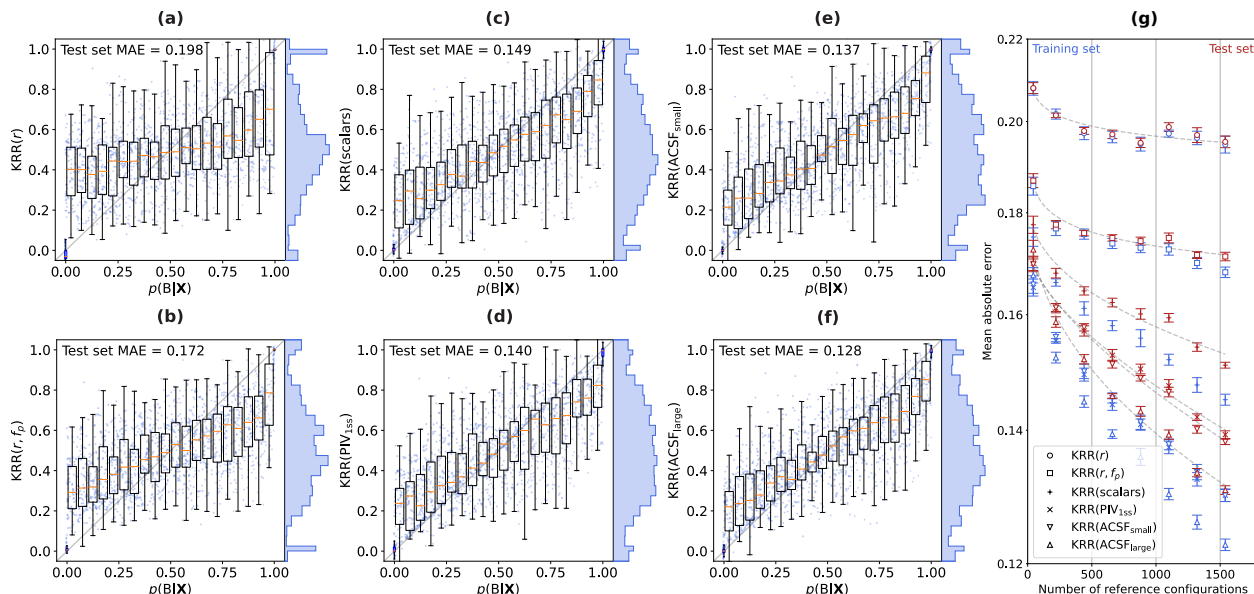


Figure 10: KRRCV models for LiF association in water. (a-f) Performance on the test set of models based on various collective variables. Small blue circles represent individual data points. The data is also shown using boxplots; data is partitioned in 0.05 wide bins, with two additional blue bins for the basins. These are the bins used to generate homogeneous datasets. Distributions of estimated committor values are also provided: an optimal estimator would return a constant distribution, since the test set is homogeneous. (g) Training and test set mean absolute errors, averaged over ten dataset splits, plotted as a function of the selected collective variable. Gray dashed lines are power law fits on the test set MAEs.

and their first coordination sphere, the four closest oxygens to Li^+ and the six closest hydrogens to F^- ($d = 66$), v) a compact set of atom-centered symmetry functions⁵⁴ centered on both ions (ACSFs) designed for aqueous systems⁷⁷ ($d = 90$), and vi) a larger set of ACSFs automatically designed for organic matter^{78,79} ($d = 595$). Reference, training and test sets are homogeneously distributed in terms of committor values (using 22 bins, with half of the configurations from umbrella sampling, and the other half from aimless shooting). We include basin configurations sampled from equilibrium simulations. The models are trained with a variable number of reference configurations, from 44 to 1540 (2 to 70 per bin). Training and test sets have a fixed number of configurations (1760, or 80 per bin). Results are presented in Figure 10.

The first thing to notice is that contrary to the precipitation case, here no single model achieves satisfying performance – although going from r to more involved collective variables certainly leads to significant improvements. Unsurprisingly, the interionic distance alone leads to a model

with high variance along the whole range of committor values. The performances near the basins are particularly poor. Adding f_p slightly improves the situation; in particular, the variance is reduced for configurations close to the transition state. The list of six scalar collective variables, as well as the PIV of the first solvation shell, lead to an improved description of the onset of the reversible transformation mechanism. Estimated committor distributions become noticeably flatter. With the more involved ACSF collective variables, in particular with the high-dimensional ($d = 595$) ACSF_{large}, the performance close to the transition state becomes acceptable, although configurations close to the basins remain poorly described. Often the committor is first modeled using a switching function,^{36,40} with parameters controlling the shape of the actual function. This allows to obtain a quantity which varies more linearly in the collective variable space, which is therefore easier to fit. Here, this did not improve the situation, with performances of transformed models using a sigmoid function (excluding basins in the training) almost identical to non-transformed models. Kernel ridge regression already allows to learn non-linear patterns, although enforcing these patterns, when known *a priori*, usually makes the learning easier. From Figure 10(g), it is clear that complex collective variables (i) lead to larger overfitting (*i.e.* the inability of a model to perform satisfyingly far from the training data), and ii) require more reference configurations to converge. Increasing the number of reference configurations beyond the range considered here would certainly improve the performances of the high-dimensional collective variable models. The inability to find an accurate model for the association of LiF in water could be explained by two causes: i) structure beyond the first solvation shell is important, since almost all collective variables considered here focus at most on the ions' first solvation shells (apart from the Madelung potentials), and/or ii) structure alone cannot predict the average evolution of the system, *i.e.* inertial effects are important. In the following, we test both hypotheses.

To assess the importance of structure beyond the first solvation shell, we considered PIV-based KRRCV models centered on either the cation or the anion. These always include both ions, and a variable number of water molecules (from 1 to 16), ranked by their distance to the central ion (O-Li⁺, or H-F⁻). The dimension of the collective variable ranges from 10 to 1225. Performances

of the models, using 1760 reference configurations, are reported in Figure 11.

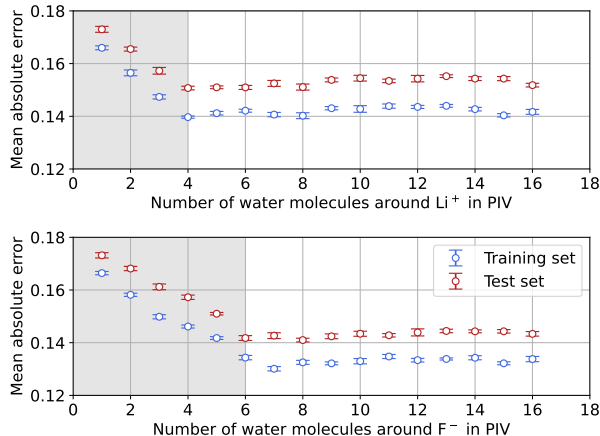


Figure 11: Performances of KRRCV models based on the ion-centered PIV with varying number of surrounding water molecules. They gray zones correspond to the first solvation shell of both ions.

While increasing the number of water molecules initially leads to reducing both training and test set MAEs, which corresponds to including more information correlated to the transformation mechanism, a plateau is rapidly reached. Interestingly, this plateau is exactly reached at a number of water molecules corresponding to the first solvation shell of both ions (4 for the cation, 6 for the anion). Increasing the number of water molecules, and therefore the dimension of the collective variable, does not lead to KRRCV models with degraded performances, as can be seen from the plateau in the MAEs. This means that the optimizer still performs well, for high-dimensional data. Within a PIV representation, there is therefore no information correlated to the LiF association mechanism beyond the first solvation shell. Assuming that the mechanism is similar for NaCl, this finding is both in agreement⁸⁰ and disagreement⁷³ with previous studies.

To investigate inertial effects, we start by evaluating their importance in the present setting. Hummer showed⁸¹ that in the diffusive limit, the committor and the transition path probability $p(\text{TP}|\mathbf{X})$ (the probability that a trajectory passing through configuration \mathbf{X} connects both basins) are related in the following way:

$$p(\text{TP}|\mathbf{X}) = 2p(\text{B}|\mathbf{X})(1 - p(\text{B}|\mathbf{X})). \tag{10}$$

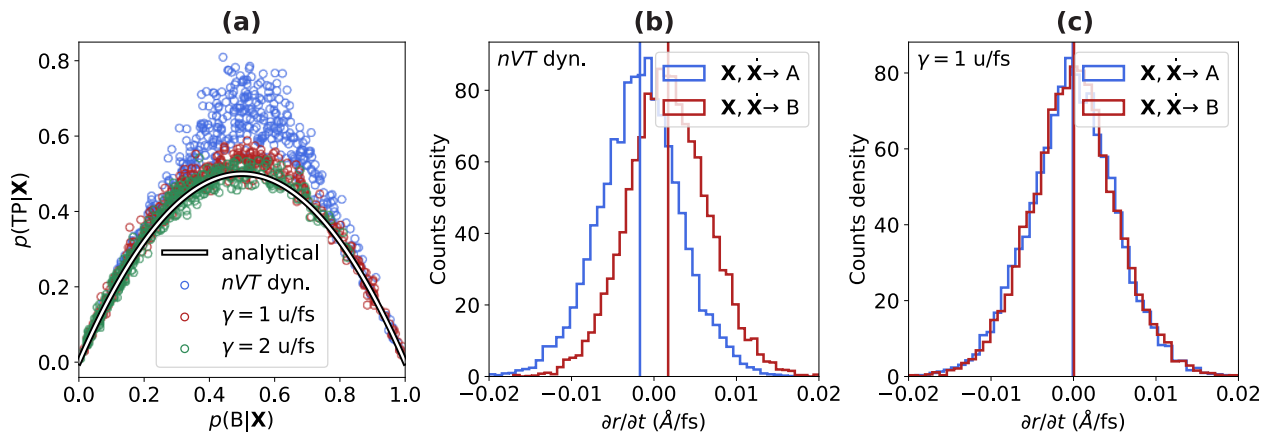


Figure 12: Inertial effects: (a) correlation between $p(B|\mathbf{X})$ and $p(TP|\mathbf{X})$ for different cases: nVT dynamics using a chain of Nosé-Hoover thermostats (blue circles), two overdamped Langevin dynamics with different friction values (red and green circles), and the limit of infinite friction (black and white line), (b,c) outcome histograms over initial interionic velocity, for nVT dynamics (b), and overdamped Langevin dynamics (c). The red and blue vertical lines correspond to the average of each distribution.

Away from the diffusive limit, inertial effects are important and lead to a deviation from this analytic result. For 400 configurations sampled from aimless shooting, we compute explicitly $p(TP|\mathbf{X})$; results are presented in Figure 12(a) (blue circles, nVT dynamics using a chain of Nosé-Hoover thermostats). A strong deviation from the diffusive limit, maximal at the transition state ensemble, is apparent. This deviation has also been observed for the association of NaCl in water.⁷³ Configurations near the transition state ensemble have an enhanced transition path probability. In the diffusive limit, atomic velocities are randomized at each timestep, while under nVT dynamics, the persistence in time of momenta can influence directly the occurrence of a reactive event. To further confirm the importance of inertial effects, we perform 1000 velocity initializations for 20 configurations from umbrella sampling for which $0.475 \leq p(B|\mathbf{X}) \leq 0.525$, and monitor the initial atomic velocities projected on the interionic distance ($\partial r/\partial t$), as well as the basin the system eventually commits to. Histograms depending on the commitment outcome are presented in Figure 12(b). In the absence of inertial effects, both histograms would overlap, indicating that the outcome is not conditioned on the atomic velocities. However, a clear separation is visible here, with a negative (positive) interionic velocity being more likely to lead to association (dissociation). This is a trivial result if one assumes that the persistence of atomic momenta is non-zero. Consid-

ering the reversibility of the equations of motion, and therefore, in particular, of committor trajectories, this result can also be interpreted in a specular way: initial velocities Boltzmann-distributed in each basin lead to reactive trajectories with transition-state velocities having an asymmetric non-Boltzmann distribution.

Having shown that inertial effects do play a role in the context of LiF association in water, we wish to alter the dynamics in such a way that they are reduced. To this end, we modify the equation of motion of the ions only (*i.e.* that for water is unchanged), following the overdamped Langevin equation:

$$d\mathbf{X} = -\gamma^{-1}\nabla U dt + \sqrt{2k_B T \gamma^{-1/2}} d\mathbf{W}, \quad (11)$$

where \mathbf{W} is a $3n$ -dimensional white noise term (with n the number of ions), and γ is the friction, identical for all ions. We perform simulations with $\gamma = 1$ u/fs and 2 u/fs, corresponding to inertial timescales – *i.e.*, timescales for decorrelation of the CV time derivative – of about 7 fs and 14 fs on Li^+ , and 19 fs and 38 fs on F^- . As demonstrated in Figure 12(a), increasing the friction parameter leads to a better agreement with the diffusive limit (which corresponds to $\gamma \rightarrow \infty$). As shown in Figure 12(c), the outcome histograms now overlap, signifying that inertial effects are removed. We then evaluate the committor on the umbrella sampling and aimless shooting datasets previously obtained, with both friction values, and train new KRRCV models. Results are presented in Figure 13.

The performances of KRRCV models are systematically and significantly improved compared to nVT dynamics. This is valid for each collective variable. Obviously, in the diffusive limit, the interionic distance would be the ideal collective variable, but considering that such an improvement is already obtained at low to reasonable friction values, this demonstrates that inertial effects prevent finding a collective variable only based on the structure that optimally reproduces the committor. In addition, even for $\gamma = 2$ u/fs, the hierarchy of performance between collective variables is preserved.

There are therefore two effects to take into account when searching for an optimal collective

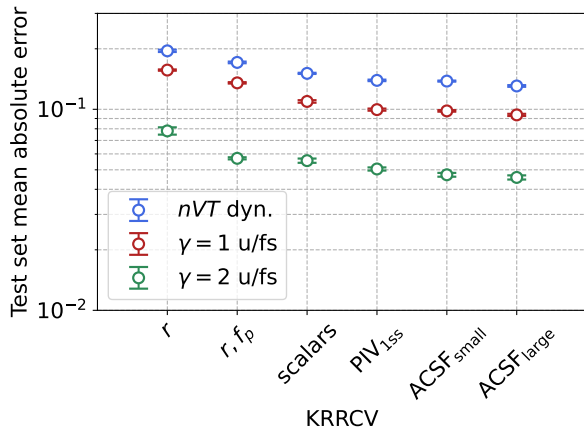


Figure 13: Performances of various KRRCV models, for different dynamics.

variable for the association of small atomic ions in water: structural degrees of freedom related to the solvent, which contain information correlated with the committor, and inertial effects. An approach purely based on the committor in the case where inertial effects are important seems inappropriate, as $p(\mathbf{B}|\mathbf{X})$ does not encode all the necessary information describing the transformation under scrutiny.

Note that the velocity-dependent committor $p(\mathbf{B}|\mathbf{X}, \dot{\mathbf{X}})$ is a deterministic quantity in the micro-canonical ensemble, as it corresponds to a single point in phase space, while it retains a probabilistic nature in presence of a thermostat (see, *e.g.*, Ref.⁸² for a discussion of velocity-dependent committors in transition path theory).

A possible approach suitably accounting for inertial effects could be based on a committor conditioned on the velocity of a collective variable. This quantity is still a probability, and it would be interesting to determine to which extent inertial effects are important: do they only concern the ions, in which case evaluating $p(\mathbf{B}|\mathbf{X}, \dot{r})$ should be enough, or are the momenta of the water molecules also correlated to the transformation mechanism? To efficiently evaluate $p(\mathbf{B}|\mathbf{X}, \dot{r})$, velocity initialization techniques with fixed values of \dot{r} would have to be implemented.

6 Summary and conclusions

We have introduced a new method for the construction of optimized collective variables, based on a data-driven generalization of path collective variables. It consists in performing a regression of the committor – which, in the absence of inertial effects, directly encodes the transformation progress – starting from a given basis of collective variables, through kernel ridge regression. The final collective variable is one dimensional and differentiable, which makes it appropriate for enhanced sampling simulations. Our approach also allows to rank and compare CVs: the performance of a certain CV subspace choice is evaluated by computing a mean absolute error with respect to a test set of committor values. In addition, using a bandwidth vector in the kernel, with values optimized by minimizing a loss score defined on a training set, allows for interpretability of the resulting model. We apply the method to three different test cases: a two-dimensional toy model, a precipitation model of Lennard-Jones particles, and the LiF association in water.

From the two-dimensional toy model test case, we learn that a reasonable number of reference configurations is sufficient to obtain an accurate KRRCV model (Kernel Ridge Regression Collective Variable), although how they are sampled is important, with enhanced sampling being a viable strategy for constructing datasets. In addition, the optimization process of kernel bandwidth parameters is shown to be robust even when including many collective variables uncorrelated with the committor.

For the precipitation model of Lennard-Jones particles, we show that our method allows to obtain an accurate collective variable from a high-dimensional representation of the structure (the PIV). This is promising for further investigation of realistic precipitation, crystallization and nucleation phenomena. The model obtained can be significantly compressed, both in terms of the number of reference configurations and components included, without degrading its performances, evaluated on committor distributions and free energy profiles. In particular, for committor distributions, we were able to perform biased simulations using KRRCVs. The free energy of activation of the best-performing model, extracted from the gauge-invariant free energy profile, is larger than the ones obtained using simple collective variables, hinting at a significantly improved transmis-

sion coefficient too. Finally, we show that the model can be interpreted, and is useful to evaluate issues related to less well-performing collective variables.

For the association of LiF in water, we demonstrate that no information correlated with the transformation mechanism is contained beyond the first solvation shell of the ions. We obtain various KRRCV models performing better than the simple interionic distance. The highest accuracy reached is however moderate, even when resorting to flexible representations such as high-dimensional ACSFs. We show that this is mostly due to inertial effects. In this regime, the committor depends not only on the atomic positions, but also on their momenta.

In situations where inertial effects are important, such as reactions in solution, it therefore seems mandatory to account for the atomic momenta. A solution could be to evaluate committor probabilities conditioned on a collective variable momentum. In this setting, much remains to be understood. For instance, should the momenta be projected onto collective variables in the same way as the structure? Or is there a hierarchy that would allow to use simpler projections for the velocities, compared to the atomic positions; for instance, in the ion association example, could we get away with only conditioning the committor on the interionic distance time derivative?

In situations where inertial effects are negligible (precipitation, nucleation, crystallization), the method introduced here allows to help understand transformation mechanisms, select collective variables on solid grounds, and allow for efficient enhanced sampling simulations. A word of caution should be expressed here: enhanced sampling is a strategy usually employed to gain access to regions of configuration space difficult to reach by directly sampling from the Gibbs measure, such as transition states, or unknown products of a chemical reaction for trajectory-based samplers (*e.g.* molecular dynamics). To construct KRRCVs, as exemplified in the previous Sections, one already needs to significantly sample the transformation under scrutiny. This is a clear example of the "chicken or the egg?" causality dilemma faced in modeling rare events: optimal CVs are required to sample transformations, yet designing optimal CVs requires sampling of said transformations. For KRRCVs, we have sketched an approach aimed at minimizing the sampling (Figure 6(a,b) and Section 5.8 of the Supporting Information), which will require more detailed, future investigations.

An iterative sampling strategy^{40,83} could also be leveraged. Overall, to resolve this dilemma, there is a need for more work on the consequences of using suboptimal – or at least uncontrolled – CVs for enhanced sampling, and on CV selection techniques bypassing the need for significant sampling. Finally, let us recall that the KRRCV model is first thought as a generalization of the path CV concept. In cases where sampling committors is challenging, *e.g.* for slow commitment kinetics, it is always possible to construct simple KRRCV models using only a single or few committor estimates near the transition state, with a scalar bandwidth parameter, and a low-dimensional CV subspace. By construction, this model will always be more accurate than a path CV constructed only with reference configurations from the metastable basins.

Acknowledgement

The authors would like to gratefully acknowledge fruitful discussions with Sara Bonella, Michele Casula, Ludovic Goudenège, Pierre Monmarché, Rocio Semino, Alessandra Serva, Gabriel Stoltz, and Rodolphe Vuilleumier, within the MAterials for Energy through STochastic sampling and high-peRformance cOmputing (MAESTRO) collaboration, hosted at the Institut des Sciences du Calcul et des Données. We also thank Gareth Tribello for his advice in implementing KRRCVs in the `hack-the-tree` branch of PLUMED. This project received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 863473). This work was granted access to the HPC resources of IDRIS under the allocations A0130811069 and A0140901387, and also benefited from the computing resources at Sorbonne Université managed by SACADO.

Data Availability Statement

The collective variable datasets used in the Lennard-Jones precipitation test case, a python script to optimize KRRCV models, as well as examples of optimization results are available at the following GitHub repository: https://github.com/physix-repo/krrcv_data

Supporting Information Available

The Supporting Information is available free of charge at url.

Details of solving the backward Kolmogorov equation with finite elements; details of the kernel ridge regression model optimization procedure; a three-wells toy model; KRRCV performance on embedding the Müller-Brown potential in a five-dimensional space; details of the Lennard-Jones precipitation test case; Bounds on the MAE associated with the numerical estimation of the committor; details of the LiF association in water test case

References

- (1) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living Journal of Computational Molecular Science* **2022**, *4*, 1583.
- (2) Pipolo, S.; Salanne, M.; Ferlat, G.; Klotz, S.; Saitta, A. M.; Pietrucci, F. Navigating at will on the water phase diagram. *Physical Review Letters* **2017**, *119*, 245701.
- (3) Jedrecy, A.; Saitta, A. M.; Pietrucci, F. Free energy calculations and unbiased molecular dynamics targeting the liquid–liquid transition in water no man’s land. *The Journal of Chemical Physics* **2023**, *158*.
- (4) Gartner III, T. E.; Piaggi, P. M.; Car, R.; Panagiotopoulos, A. Z.; Debenedetti, P. G. Liquid–liquid transition in water from first principles. *Physical review letters* **2022**, *129*, 255702.
- (5) Lupi, L.; Hudait, A.; Peters, B.; Grünwald, M.; Gotchy Mullen, R.; Nguyen, A. H.; Molinero, V. Role of stacking disorder in ice nucleation. *Nature* **2017**, *551*, 218–222.
- (6) Shea, J.-E.; Brooks III, C. L. From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annual Review of Physical Chemistry* **2001**, *52*, 499–535.

- (7) Saitta, A. M.; Saija, F. Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13768–13773.
- (8) Hartmann, C.; Latorre, J. C.; Ciccotti, G. On two possible definitions of the free energy for collective variables. *The European Physical Journal Special Topics* **2011**, *200*, 73–89.
- (9) Bal, K. M.; Fukuhara, S.; Shibuta, Y.; Neyts, E. C. Free energy barriers from biased molecular dynamics simulations. *The Journal of Chemical Physics* **2020**, *153*.
- (10) Dietschreit, J. C.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *The Journal of Chemical Physics* **2022**, *157*.
- (11) Palacio-Rodriguez, K.; Vroylandt, H.; Stelzl, L. S.; Pietrucci, F.; Hummer, G.; Cossio, P. Transition rates and efficiency of collective variables from time-dependent biased simulations. *The Journal of Physical Chemistry Letters* **2022**, *13*, 7490–7496.
- (12) Jungblut, S.; Dellago, C. Pathways to self-organization: Crystallization via nucleation and growth. *The European Physical Journal E* **2016**, *39*, 1–38.
- (13) Mouaffac, L.; Palacio-Rodriguez, K.; Pietrucci, F. Optimal Reaction Coordinates and Kinetic Rates from the Projected Dynamics of Transition Paths. *Journal of Chemical Theory and Computation* **2023**, *19*, 5701–5711.
- (14) Vroylandt, H.; Goudenège, L.; Monmarché, P.; Pietrucci, F.; Rotenberg, B. Likelihood-based non-Markovian models from molecular dynamics. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2117586119.
- (15) Girardier, D. D.; Vroylandt, H.; Bonella, S.; Pietrucci, F. Inferring free-energy barriers and kinetic rates from molecular dynamics via underdamped Langevin models. *The Journal of Chemical Physics* **2023**, *159*.

- (16) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis* **2006**, *21*, 113–127.
- (17) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation* **2008**, *7*, 842–864.
- (18) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of chemical physics* **2011**, *134*.
- (19) Rydzewski, J.; Chen, M.; Ghosh, T. K.; Valsson, O. Reweighted Manifold Learning of Collective Variables from Enhanced Sampling Simulations. *Journal of Chemical Theory and Computation* **2022**, *18*, 7179–7192.
- (20) Zhang, W.; Hartmann, C.; Schütte, C. Effective dynamics along given reaction coordinates, and reaction rate theory. *Faraday discussions* **2016**, *195*, 365–394.
- (21) Noé, F.; Nuske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation* **2013**, *11*, 635–655.
- (22) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013**, *139*.
- (23) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters* **1994**, *72*, 3634.
- (24) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **2018**, *9*, 5.
- (25) Tiwary, P.; Berne, B. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences* **2016**, *113*, 2839–2844.

- (26) Rydzewski, J. Spectral Map: Embedding Slow Kinetics in Collective Variables. *The Journal of Physical Chemistry Letters* **2023**, *14*, 5216–5220.
- (27) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Physical Review E* **2018**, *97*, 062412.
- (28) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *The Journal of chemical physics* **2018**, *149*.
- (29) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature communications* **2019**, *10*, 3573.
- (30) Wang, D.; Tiwary, P. State predictive information bottleneck. *The Journal of Chemical Physics* **2021**, *154*.
- (31) Mendels, D.; Piccini, G. M.; Parrinello, M. Collective variables from local fluctuations. *The journal of physical chemistry letters* **2018**, *9*, 2776–2781.
- (32) Sultan, M. M.; Pande, V. S. Automated design of collective variables using supervised machine learning. *The Journal of chemical physics* **2018**, *149*.
- (33) Bonati, L.; Rizzi, V.; Parrinello, M. Data-driven collective variables for enhanced sampling. *The journal of physical chemistry letters* **2020**, *11*, 2998–3004.
- (34) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *The Journal of Physical Chemistry B* **2005**, *109*, 6769–6779.
- (35) Lechner, W.; Rogal, J.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. Nonlinear reaction coordinate analysis in the reweighted path ensemble. *The Journal of chemical physics* **2010**, *133*.
- (36) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *The Journal of chemical physics* **2006**, *125*, 054108.

- (37) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry* **2002**, *53*, 291–318.
- (38) Peters, B. Inertial likelihood maximization for reaction coordinates with high transmission coefficients. *Chemical Physics Letters* **2012**, *554*, 248–253.
- (39) Mori, Y.; Okazaki, K.-i.; Mori, T.; Kim, K.; Matubayasi, N. Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide. *The Journal of Chemical Physics* **2020**, *153*.
- (40) Jung, H.; Covino, R.; Arjun, A.; Leitold, C.; Dellago, C.; Bolhuis, P. G.; Hummer, G. Machine-guided path sampling to discover mechanisms of molecular self-organization. *Nature Computational Science* **2023**, 1–12.
- (41) Jung, H.; Covino, R.; Arjun, A.; Leitold, C.; Bolhuis, P. G.; Dellago, C.; Hummer, G. Artificial intelligence for molecular mechanism discovery. *Biophysical Journal* **2023**, *122*, 281a–282a.
- (42) Lazzeri, G.; Jung, H.; Bolhuis, P. G.; Covino, R. Molecular free energies, rates, and mechanisms from data-efficient path sampling simulations. *Journal of Chemical Theory and Computation* **2023**,
- (43) Chen, H.; Roux, B.; Chipot, C. Discovering Reaction Pathways, Slow Variables, and Committer Probabilities with Machine Learning. *Journal of Chemical Theory and Computation* **2023**,
- (44) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *The Journal of chemical physics* **2007**, *126*, 054103.
- (45) Pietrucci, F.; Saitta, A. M. Formamide reaction network in gas phase and solution via a unified

- theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences* **2015**, *112*, 15030–15035.
- (46) France-Lanord, A.; Pietrucci, F.; Saitta, A. M.; Tarascon, J.-M.; Grimaud, A.; Salanne, M. Chemical Decomposition of the TFSI Anion under Aqueous Basic Conditions. *PRX Energy* **2022**, *1*, 013005.
- (47) Saladino, G.; Gauthier, L.; Bianciotto, M.; Gervasio, F. Assessing the performance of metadynamics and path variables in predicting the binding free energies of p38 inhibitors. *Journal of Chemical Theory and Computation* **2012**, *8*, 1165–1170.
- (48) Pérez de Alba Ortíz, A.; Vreede, J.; Ensing, B. The adaptive path collective variable: a versatile biasing approach to compute the average transition path and free energy of molecular transitions. *Biomolecular Simulations: Methods and Protocols* **2019**, 255–290.
- (49) Hovan, L.; Comitani, F.; Gervasio, F. L. Defining an optimal metric for the path collective variables. *Journal of Chemical Theory and Computation* **2018**, *15*, 25–32.
- (50) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900.
- (51) Peters, B. Using the histogram test to quantify reaction coordinate error. *The Journal of chemical physics* **2006**, *125*.
- (52) Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications* **1964**, *9*, 141–142.
- (53) Watson, G. S. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* **1964**, 359–372.
- (54) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **2007**, *98*, 146401.

- (55) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.
- (56) Gallet, G. A.; Pietrucci, F. Structural cluster analysis of chemical reactions in solution. *The Journal of Chemical Physics* **2013**, *139*, 074101.
- (57) Lai, R.; Lu, J. Point Cloud Discretization of Fokker–Planck Operators for Committor Functions. *Multiscale Modeling & Simulation* **2018**, *16*, 710–726.
- (58) Li, Q.; Lin, B.; Ren, W. Computing committor functions for the study of rare events using deep learning. *The Journal of Chemical Physics* **2019**, *151*, 054112.
- (59) Müller, K.; Brown, L. D. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta* **1979**, *53*, 75–93.
- (60) Weinan, E.; Vanden-Eijnden, E. *Multiscale modelling and simulation*; Springer, 2004; pp 35–68.
- (61) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; others LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*, 108171.
- (62) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; others Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673.
- (63) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer physics communications* **2014**, *185*, 604–613.
- (64) Tribello, G. A.; Giberti, F.; Sosso, G. C.; Salvalaglio, M.; Parrinello, M. Analyzing and driving cluster formation in atomistic simulations. *Journal of chemical theory and computation* **2017**, *13*, 1317–1327.

- (65) Mullen, R. G.; Shea, J.-E.; Peters, B. Easy transition path sampling methods: Flexible-length aimless shooting and permutation shooting. *Journal of Chemical Theory and Computation* **2015**, *11*, 2421–2428.
- (66) PLUMED – hack-the-tree branch. <https://github.com/plumed/plumed2/tree/hack-the-tree>, 2023.
- (67) Mullen, R. G.; Shea, J.-E.; Peters, B. Transmission coefficients, committors, and solvent coordinates in ion-pair dissociation. *Journal of chemical theory and computation* **2014**, *10*, 659–667.
- (68) Peled, E.; Menkin, S. SEI: past, present and future. *Journal of The Electrochemical Society* **2017**, *164*, A1703.
- (69) Alzate-Vargas, L.; Vikrant, K.; Allu, S.; Fattebert, J.-L. Atomistic modeling of LiF microstructure ionic conductivity and its influence on nucleation and plating. *Physical Review Materials* **2022**, *6*, 095402.
- (70) Droguet, L.; Hobold, G. M.; Lagadec, M. F.; Guo, R.; Lethien, C.; Hallot, M.; Fontaine, O.; Tarascon, J.-M.; Gallant, B. M.; Grimaud, A. Can an inorganic coating serve as stable SEI for aqueous superconcentrated electrolytes? *ACS Energy Letters* **2021**, *6*, 2575–2583.
- (71) Jones, J.; Anouti, M.; Caillon-Caravanier, M.; Willmann, P.; Lemordant, D. Thermodynamic of LiF dissolution in alkylcarbonates and some of their mixtures with water. *Fluid phase equilibria* **2009**, *285*, 62–68.
- (72) Geissler, P. L.; Dellago, C.; Chandler, D. Kinetic pathways of ion pair dissociation in water. *The Journal of Physical Chemistry B* **1999**, *103*, 3706–3710.
- (73) Ballard, A. J.; Dellago, C. Toward the mechanism of ionic dissociation in water. *The Journal of Physical Chemistry B* **2012**, *116*, 13490–13497.

- (74) Joung, I. S.; Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B* **2008**, *112*, 9020–9041.
- (75) Berendsen, H. J.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.
- (76) Kattirtzi, J. A.; Limmer, D. T.; Willard, A. P. Microscopic dynamics of charge separation at the aqueous electrochemical interface. *Proceedings of the National Academy of Sciences* **2017**, *114*, 13374–13379.
- (77) Schran, C.; Thiemann, F. L.; Rowe, P.; Müller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2110077118.
- (78) Bircher, M. P.; Singraber, A.; Dellago, C. Improved description of atomic environments using low-cost polynomial functions with compact support. *Machine Learning: Science and Technology* **2021**, *2*, 035026.
- (79) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *The Journal of chemical physics* **2018**, *148*.
- (80) Wang, D.; Zhao, R.; Weeks, J. D.; Tiwary, P. Influence of long-range forces on the transition states and dynamics of NaCl ion-pair dissociation in water. *The Journal of Physical Chemistry B* **2022**, *126*, 545–551.
- (81) Hummer, G. From transition paths to transition states and rate coefficients. *The Journal of chemical physics* **2004**, *120*, 516–523.
- (82) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Illustration of transition path theory on a collection of simple examples. *The Journal of chemical physics* **2006**, *125*.

- (83) Kulichenko, M.; Barros, K.; Lubbers, N.; Li, Y. W.; Messerly, R.; Tretiak, S.; Smith, J. S.; Nebgen, B. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science* **2023**, 3, 230–239.

TOC Graphic

