Supporting Information:

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@cnrs.fr; fabio.pietrucci@sorbonne-universite.fr

Contents

1 Solving the backward Kolmogorov equation using finite elements

The committor $p(\mathrm{B}|(x,y))$ can be obtained as the solution of the following equation[S1]

$$\begin{cases} \mathcal{L}p(\mathrm{B}|(x,y)) = 0 & (x,y) \in (A \cup B)^c \\ p(\mathrm{B}|(x,y)) = 0 & (x,y) \in A \\ p(\mathrm{B}|(x,y)) = 1 & (x,y) \in B \end{cases} \tag{1}$$

where $\mathcal{L}$ is the infinitesimal generator of the Langevin overdamped dynamics. For the two-dimensional rugged Müller-Brown potential $V(x,y)$, it can be expressed as[S2]

$$\begin{aligned} \mathcal{L}f(x,y) = &- \frac{\partial V(x,y)}{\partial x}\frac{\partial f(x,y)}{\partial x} - \frac{\partial V(x,y)}{\partial y}\frac{\partial f(x,y)}{\partial y} \\ &+ \frac{1}{\beta}\left(\frac{\partial^2 f(x,y)}{\partial x^2} + 2\frac{\partial^2 f(x,y)}{\partial x \partial y} + \frac{\partial^2 f(x,y)}{\partial y^2}\right) \end{aligned} \tag{2}$$

for any two-dimensional function $f(x,y)$. Equation (1) can be solved using finite elements methods.[S3] For a finite element basis $\{f_i(x,y)\}_1^N$, the backward Kolmogorov equation can be expressed as a matrix equation,

$$\begin{cases} \boldsymbol{L}q = 0 & (x,y) \in (A \cup B)^c \\ q = 0 & (x,y) \in A \\ q = 1 & (x,y) \in B \end{cases} \tag{3}$$

where the matrix elements are given by

$$\boldsymbol{L}_{i,j} = \int_{\mathbb{R}^2} \mathrm{d}x\,\mathrm{d}y \left[f_i(x,y) \left(-\frac{\partial V(x,y)}{\partial x}\frac{\partial f_j(x,y)}{\partial x} - \frac{\partial V(x,y)}{\partial y}\frac{\partial f_j(x,y)}{\partial y} \right) - \right.$$
$$\left. \frac{1}{\beta} \left(\frac{\partial f_i(x,y)}{\partial x}\frac{\partial f_j(x,y)}{\partial x} + \frac{\partial f_i(x,y)}{\partial x}\frac{\partial f_j(x,y)}{\partial y} + \frac{\partial f_i(x,y)}{\partial y}\frac{\partial f_j(x,y)}{\partial x} + \frac{\partial f_i(x,y)}{\partial y}\frac{\partial f_j(x,y)}{\partial y} \right) \right]. \quad (4)$$

The committor is then obtained as

$$p(\mathrm{B}|(x,y)) = \sum_i q_i f_i(x,y) \quad (5)$$

The result of Figure 2 of the main text was obtained using a regular triangular mesh of 79202 triangles defined on the rectangular domain $[-1.5,1] \times [-0.5,2]$ and linear element ($\mathbb{P}1$ Lagrange element). This leads to a finite element basis with 39534 degrees of freedom once nodes corresponding to region A and B were removed. Computations were performed using the scikit-fem finite element library.[S4]

2 KRR model optimization procedure

All optimizations (regarding kernel ridge regression expansion coefficients as well as hyperparameters) are performed using the python/C++ package falkon.[S5,S6] A model python script is included in the Supplemental Material. For a given set of hyperparameters (regularization coefficient and bandwidths), optimal expansion coefficients are obtained by solving Equation (3) of the main text. Simultaneous optimization of hyperparameters is achieved by iteratively minimizing the MAE on a training set (distinct from the reference set), using the Adam optimizer,[S7] as implemented in PyTorch,[S8] for 100 steps. For a dataset of N entries, the MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |p(\mathrm{B}|\mathbf{X}_i) - \mathrm{KRRCV}(\xi_i)| \quad (6)$$

We have selected the MAE over the mean squared error (MSE) as in preliminary tests using

various datasets, models optimized with MAE- or MSE-based loss functions were found to be virtually indistinguishable. We opted for the MAE as its interpretation, in terms of a quality metric, is more straightforward than the MSE.

We perform distinct optimizations with different values of the learning rate parameter (α): for the rugged Müller-Brown case, $\alpha \in \left[10^{-1}, 10^{0}, 10^{1}\right]$, for the Lennard-Jones case, $\alpha \in \left[10^{-1}, 10^{0}, 10^{1}, 10^{2}\right]$, and for the LiF association case, $\alpha \in \left[10^{-1}, 5\cdot10^{-1}, 10^{0}, 5\cdot10^{0}, 10^{1}, 5\cdot10^{1}, 10^{2}, 5\cdot10^{2}\right]$. We partition reference and training sets randomly in 10 different ways, to obtain uncertainty estimates. The test set is always the same. We perform 100 optimizations starting from randomly selected initial hyperparameter values. For each γ value, this amounts to 1000 optimizations. Finally, out of all optimized models, we select the one that minimizes the training set error. We observe a positive correlation between the training set and test set error metrics, as shown on Figure S1.

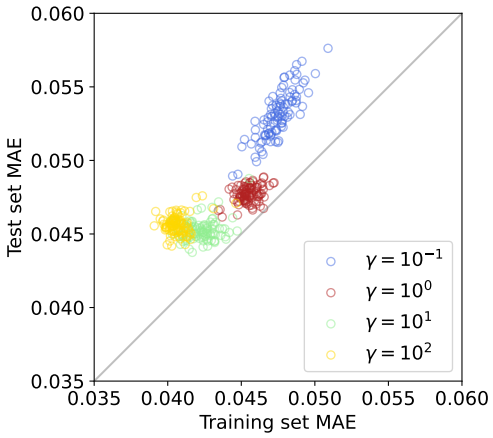

Figure S1: Correlation between the training set and test set mean absolute errors, for the Lennard-Jones precipitation case, with the PIV collective variable, for the dataset split 3. Points are colored according to the learning rate α; each point corresponds to the final metrics optained for one optimization, starting from randomly selected parameters.

We note that for high-dimensional representations, the optimal models are "ridgeless", meaning that the optimal regularization parameter λ approaches zero. This has been discussed before[S9] for non-linear kernels, and does not prevent generalization.

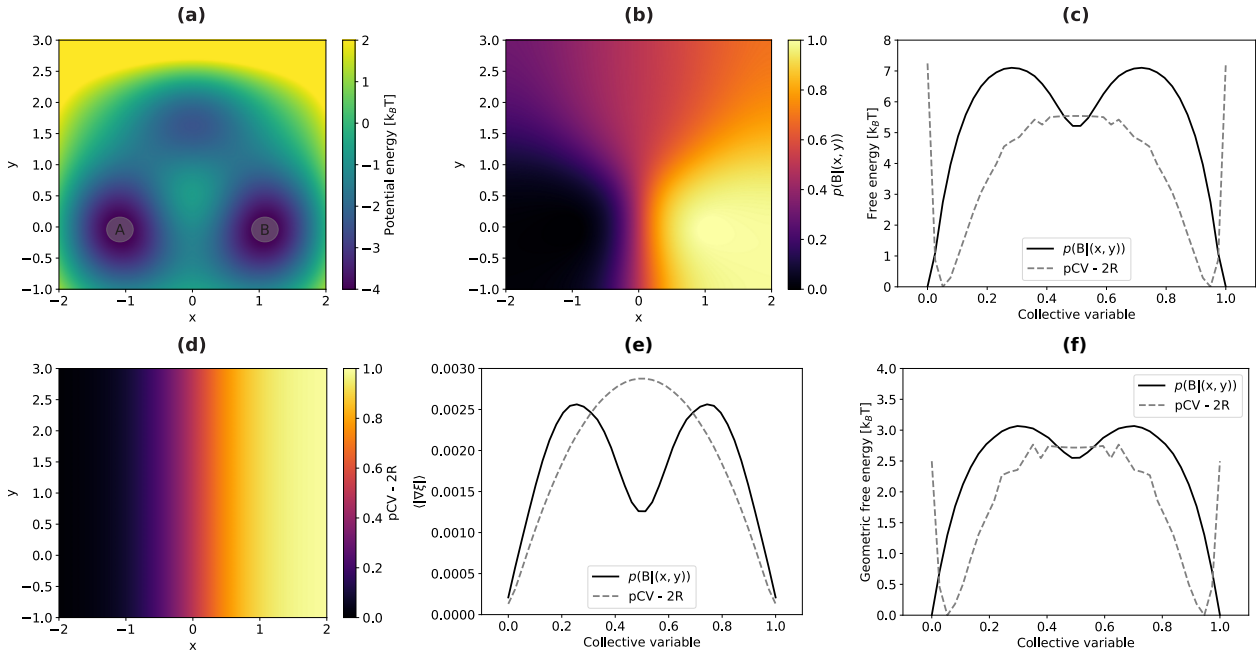

Figure S2: (a) The potential energy surface. Shaded circles correspond to the metastable states definition used for the evaluation of the committor, (b) the committor as obtained from the backward Kolmogorov equation, (c) Canonical free energy profiles, (d) the "2R" path collective variable in configuration space, (e) Ensemble-averaged collective variable gradients, (f) Geometric free energy profiles.

3 Three wells model potential

In Figure S2, we compare a "2R" path collective variable with the true committor for a two-dimensional potential showing three metastable states. The two deeper ones are labeled as A and B, and the shallowest one is an intermediate state. The free energy profiles differ quite significantly, especially in the vicinity of the intermediate state.

4 Müller-Brown potential embedded in a five-dimensional space

To complicate the learning process of the committor in Section 3 of the main text, we use non-linear transformation to embed the two-dimensional Müller-Brown potential in a five-dimensional space, in a similar way as in Ref.:[S10]

$$V(x_1,x_2,x_3,x_4,x_5) = V(x,y), \tag{7}$$

$$\begin{cases} x_1 = & x + 0.1y^2 \\ x_2 = & y - 2x + 3 \\ x_3 = & \sqrt{4|xy|} \\ x_4 = & x^3 - y^2 \\ x_5 = & xy^4 \end{cases} . \tag{8}$$

We train a KRR model using all five dimensions, with 500 reference and training data points. The resulting test set MAE, $\approx 7 \cdot 10^{-3}$, is on par with the one of the model trained on the native two-dimensional representation.

5 Precipitation of Lennard-Jones particles: computational details and additional information

5.1 System generation, initial relaxation

We begin with an initial configuration composed of 4096 particles arranged on a simple cubic lattice of spacing set to $l = \sigma$, in a cubic box with 16 σ-long edges. 20 particles, selected randomly, are set to being of type 2 (the larger, precipitating species). The atomic velocities are initialized by drawing from the Maxwell-Boltzmann distribution at $T = 1$. A $10^6 \delta t$ simulation in the npT ensemble at $T = 1$, $p = 1$ is then performed to relax the system and estimate the equilibrium box size. The box size is subsequently fixed at 17.20 σ.

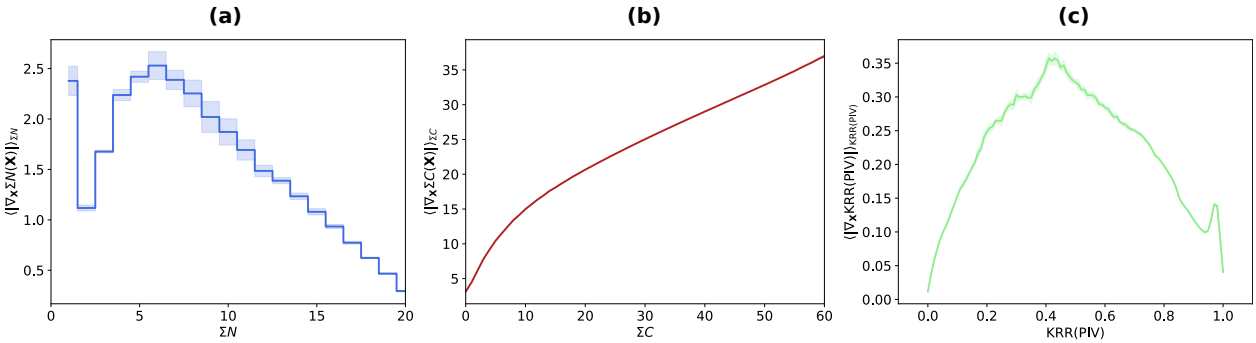

Figure S3: Norm of the gradient of collective variables along Cartesian coordinates: (a) ΣN, (b) ΣC, (c) KRR(PIV)

5.2 Unbiased free energy estimates

Starting from the previously equilibrated geometry, we perform 300 independent simulations in the nVT ensemble with randomized initial velocities, ran for $10^6\delta t$ for equilibration and $2\cdot 10^7\delta t$ for sampling, which amounts to a total of $6\cdot 10^9\delta t$. Collective variables are computed every 10^3 time steps. Free energy profiles are then calculated through binning, with error estimates obtained by splitting the whole dataset into six distinct subsets, and computing 95% confidence intervals over the distribution of estimates. We performed identical simulations in the npT ensemble, to verify that box size fluctuations do not significantly influence free energy profiles.

5.3 Computing collective variable gradients

We compute the derivatives of the collective variable with respect to the Cartesian coordinates using a second-order central difference scheme, with a displacement of atomic positions set to 0.05σ:

$$\frac{\mathrm{d}\xi(x)}{\mathrm{d}x} \approx \frac{\xi(x+0.05\sigma) - 2\xi(x) + \xi(x-0.05\sigma)}{(0.05\sigma)^2} \tag{9}$$

5.4 Umbrella sampling simulations

To sample configurations biased along ΣC, we perform simulations with five different harmonic biasing potentials of the form $0.5k\left(\Sigma C - \Sigma C_0\right)^2$ centered at $\Sigma C_0 = 10, 11, 12, 13, 14$, and with a force constant $k = 10 k_B T$. Trajectories last $10^7 \delta t$ and configurations are sampled every 10^5 steps; the first configuration is discarded to allow for equilibration. We therefore sample a total of 500 configurations.

To sample configurations at the putative transition state ensembles of KRR($\mathbf{C}$) and KRR(PIV), we perform simulations with a harmonic biasing potential centered at $\mathrm{KRR}(\xi)_0 = 0.5$, and with a force constant $k = 10^4 k_B T$. Trajectories last $2.5 \cdot 10^6 \delta t$ and configurations are sampled every 10^4 steps; the first configuration is discarded to allow for equilibration. We therefore sample a total of 250 configurations for each collective variable.

5.5 Transition path sampling: brute force

In order to generate initial transition paths to be used as starting points for aimless shooting simulations, we select the 100 configurations from the $\Sigma C_0 = 12$ umbrella sampling window and propagate them forward and backward in time with randomized initial velocities. If both forward and backward dynamics reach the same metastable basin, we perform dynamics again with new random initial velocities. In this setting, a transition path is typically achieved after less than 10 tries. Eventually, all initial contributions lead to transition pathways. The largest number of tries was 28. We report in Figure S4 a histogram of the number of tries, for all 100 starting configurations (which we call the $t = 0$ configurations).

5.6 Transition path sampling: aimless shooting

We used the transition paths generated using brute force to initialize 100 independent aimless shooting simulations.[S11,S12] The approach is the following: starting from the $t = 0$ configuration in the initial path, the system is propagated both forward and backward in time, with randomized

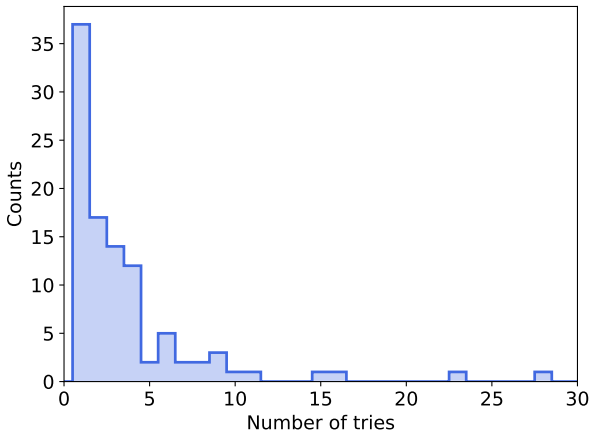

Figure S4: Number of tries needed to connect both basins in brute force transition path sampling, for 100 starting configurations sampled from an umbrella sampling simulation.

initial velocities. Both simulations end when a metastable basin is reached. If the new path connects both basins, the configuration is stored as an "accepted" configuration, and a new starting point is obtained from the new path by selecting the configuration separated by $\pm 500\delta t$ from the $t = 0$ configuration. If the new path does not connect both basins, the configuration is stored as "rejected", and the selection strategy is applied again to the former path. The aimless shooting selection step ($500\delta t$) has been adjusted to roughly match a 35% acceptance ratio, which represents a good balance between sampling quality (a small selection step leads to highly correlated configurations), and efficiency (there is a large enough number of accepted paths). For each initial path, we perform 2200 aimless shooting iterations. Finally, 5 "accepted" and 5 "rejected" configurations per aimless shooting simulation are selected, evenly spaced across both datasets. This leads to a final dataset of 500 "accepted" and 500 "rejected" configurations.

5.7 Numerical estimation of the committor

To compute the committor of each configuration from the sampled configurations, we launch 200 dynamics with randomized initial velocities, which end once a basin is reached, or once the trajectory reaches $10^6\delta t$. In this case, which represents about 0.35% of all trajectories, it is discarded. Histograms of trajectory lengths are displayed on Figure S5. The committor is then evaluated

based on the number of outcomes. Overall, this required about $5.2 \cdot 10^{10}$ molecular dynamics time steps, which highlights the cost of evaluating the committor for systems showing slow committment kinetics.

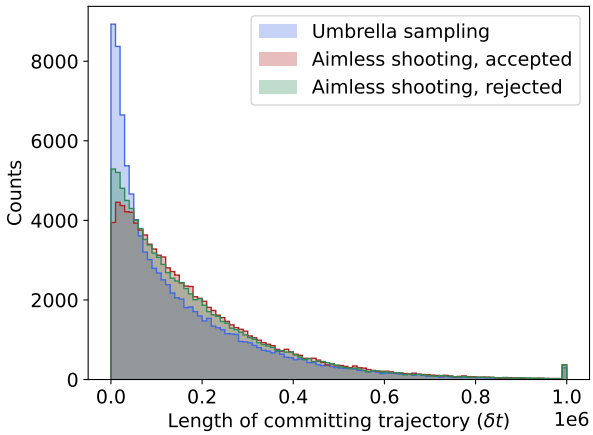

Figure S5: Committing trajectory lengths, in δt. There is a peak at $10^6 \delta t$ since this is the maximum trajectory length we allow.

5.8 A model selection strategy for data points and dimensionality

When trying to minimize the amount of committor evaluations, the map reported in Figure 6(b) of the main text is generally not available, since the metric used for discrimination is the performance on the full test set. What is directly available is the performance on a small test set, or on the small training set (Figure S6(a)). The latter quantity does not allow to select the appropriate minimal number of data points outside of basins, since it will be minimal for the smallest datasets, resulting in significant overfitting. However, one can also estimate the noise MAE as a function of the dataset distribution (Figure S6(b)), following Appendix 6, which decreases as a function of the number of points in the datasets. This quantity can be used as a baseline to prevent overfitting: in Figure S6(c), we plot the training set MAE divided by the noise MAE. When this quantity is smaller or close to one, overfitting is significant. When it is large (*e.g.* at low dimensionality), the model performs poorly even on the training set. When it reaches an intermediate value ($\approx 2-3$), it seems to lead to appropriate models, balancing accuracy and overfitting. As a strategy, we therefore recommend

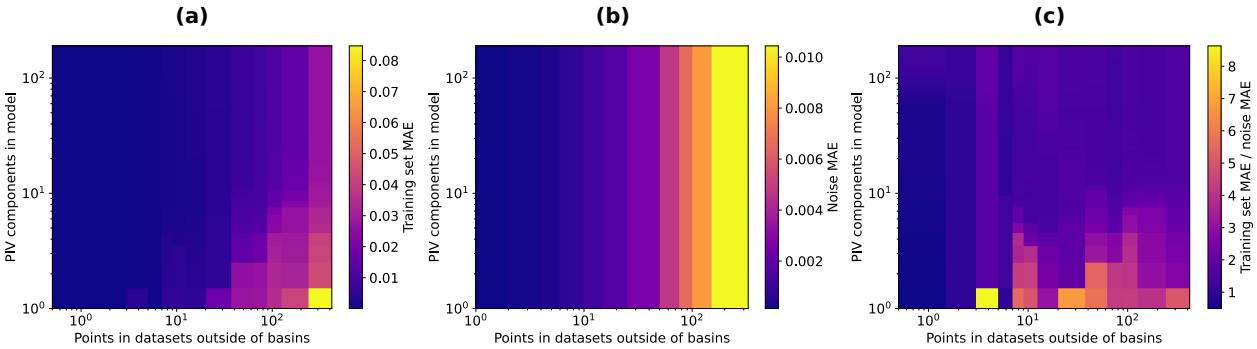

Figure S6: Model selection strategy. (a) Training set MAE, (b) noise MAE, and (c) ratio of the two quantities as a function of the number of reference points included outside of basins, and of the number of PIV components included in the KRRCV mode.

to optimize models with a small number of data points, progressively adding more points until this quantity reaches $\approx 2-3$.

5.9 List of collective variables investigated

Table S1: The list of CVs, or CV combinations, employed in the Lennard-Jones precipitation application, with corresponding dimension.

CV	Designation	d
ΣN	Number of solute particles in the largest cluster	1
ΣC	Sum of solute coordination numbers over solute particles in the largest cluster	1
$\Sigma N, \Sigma C$	Combination of the CVs above	2
ΣV_{11}	Pairwise interation energy between all solute particles	1
$\mathbf{C}$	ΣC and individual solute coordination numbers for all solute particles	21
PIV	Sorted vector of all solute-solute inverse distances	190

6 Uncertainty on numerical estimates of the committor probability: mean absolute error

Since the committor must be estimated numerically, there is a numerical uncertainty associated that leads to a lower bound on the MAE, *i.e.* even if the KRR would correlate perfect with the comittor, we would get a finite value of the MAE that can be estimated as follows. The numerical

estimation of the committor probability p_B through N repeated and independent trial molecular dynamics simulations is a Bernoulli process. The number of successes $k = N \cdot p_B$, *i.e.* the number of trajectories committing to basin B for N trials, therefore follows a binomial distribution. The mean absolute error (MAE) of a binomially-distributed random variable has a closed-form expression identified by de Moivre:[S13,S14]

$$\mathrm{MAE}(p_B, N) = \mathbb{E}\,|p_B - \mathbb{E}p_B| = \frac{1}{N} 2k\,(1 - p_B) \binom{N}{k} b\,(k, N, p_B)\,, \tag{10}$$

where $b\,(k, N, p_B)$ is the probability mass function of the binomial distribution:

$$b\,(k, N, p_B) = p_B^k\,(1 - p_B)^{N-k}\,. \tag{11}$$

We can therefore compute the MAE as a function of the committor probability, and of the number of trials. For the LiF association in water, since datasets are uniform in committor values, we can estimate the MAE on data as:

$$\mathrm{MAE}(N) = \int_0^1 \mathrm{MAE}(p_B, N)\mathrm{d}p_B. \tag{12}$$

The MAE on data for homogeneous datasets is reported in Figure S7, with the MAE dependence on both p_B and N. For the LiF association in water, $N = 1000$ and MAE ≈ 0.010. The Lennard-Jones datasets being heterogeneously distributed along p_B, we use the actual test set distribution to evaluate MAE(N), *i.e.*, for a dataset with M data points:

$$\mathrm{MAE}(N) = \frac{1}{M} \sum_{i=1}^{M} \mathrm{MAE}(p_B^i, N) \tag{13}$$

For $N = 200$, MAE ≈ 0.017. A homogeneously-distributed dataset would lead to MAE $\approx$ 0.021; the – small – reduction is due to the larger amount of basin configurations, for which the error is minimal.

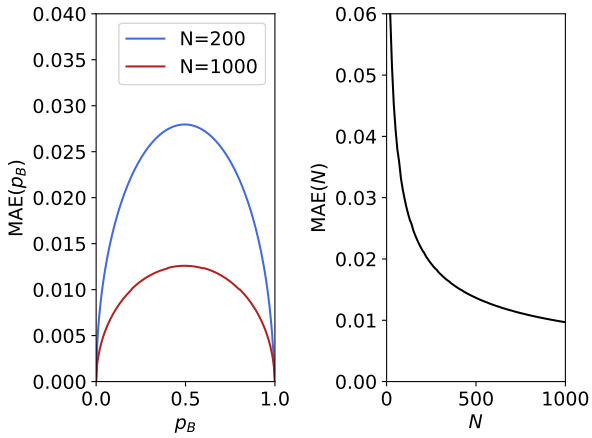

Figure S7: Numerical estimation of the committor probability. Left: MAE as a function of committor value, right: MAE as a function of the number of trials.

7 LiF association in water: computational details and additional information

7.1 System generation, initial relaxation

We generate 100 initial configurations using a Monte Carlo algorithm implemented in the packmol program,[S15] with one cation, one anion, and 160 water molecules, in a cubic box with 16.90 Å-long edges. The atomic velocities are initialized by drawing from the Maxwell-Boltzmann distribution at $T = 300$K. We relax these configurations in the npT ensemble at $T = 300$K, $p = 1$atm, for 5 ns. The box size is subsequently fixed at the average equilibrium value obtained, 16.83 Å.

7.2 Unbiased free energy estimates

From the previous, relaxed geometries, we perform 100 independent simulations in the nVT ensemble with randomized initial velocities, 1 ns of equilibration, and 20 ns for sampling, which amounts to a total of 2 μs of dynamics. The interionic distance (r) is computed every 10 fs; this dataset is subsequently used to compute the free energy profile along r by binning. Uncertainty estimates are obtained by computing 95% confidence intervals over the distribution made of the

100 estimates.

7.3 Umbrella sampling simulations

We perform a single umbrella sampling simulation by constraining the interionic distance at $r_0 =$ 2.636 Å, using a harmonic biasing potential with a force constant set to 5000 kcal/mol/Å, for 50 ns. Configurations are sampled every 10 ps; we therefore obtain a dataset of 5000 configurations matching the constraint on r.

7.4 Transition path sampling: brute force

In order to generate initial transition paths to be used as starting points for aimless shooting simulations, we randomly select 200 configurations with $0.3 \leq p(\mathrm{B}|\mathbf{X}) \leq 0.7$ from the umbrella sampling dataset, and propagate them forward and backward in time with randomized initial velocities. If both forward and backward dynamics reach the same metastable basin, we perform dynamics again with new random initial velocities. Transition paths are achieved with less than 5 tries for all configurations; as shown in Figure 12(a) of the main text, $p(\mathrm{TP}|\mathbf{X})$ is high for transition state configurations.

7.5 Transition path sampling: aimless shooting

Starting from the previously generated transition paths, we perform 200 independent aimless shooting simulations of 10^4 iterations, with a selection step of 10 fs, leading to an average acceptance ratio of 42%. Finally, we downsample the list of sampled structures by a factor of 100, leading to a dataset of 8551 configurations.

7.6 Numerical estimation of the committor and of the transition path probability

We obtain numerical estimates of the committor by launching 1000 independent unbiased dynamics from each configuration. We also perform backward dynamics to estimate $p(\mathrm{TP}|\mathbf{X})$; these backward dynamics statistics are however discarded when estimating $p(\mathrm{B}|\mathbf{X})$.

7.7 Committor distribution at the critical interionic distance from unbiased molecular dynamics

We investigate the shape of the committor distribution for configurations matching $r \approx r^*$ sampled from unbiased molecular dynamics. We perform 100 independent simulations of 10 ns, amounting to a total sampling time of 1 μs. Every 10 fs, configurations for which $r \in [2.55, 2.70]$ Å are saved. We obtain 2706 configurations, out of which 581 are separated by at least 1 ps, for which we compute the committor (using 500 velocity initializations). The distribution, whose shape matches the one of the umbrella sampling distribution, is displayed in Figure S8.

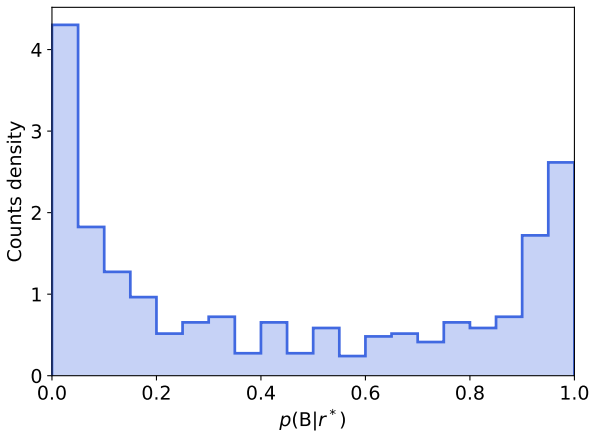

Figure S8: Distribution of committor values for configurations at the putative transition state ensemble of r sampled from unbiased simulations.

Table S2: The list of CVs, or CV combinations, employed in the ion association application, with corresponding dimension.

CV	Designation	d
r	Li-F interionic distance	1
r, f_p	r and the interionic force projected on the vector connecting both ions	2
scalars	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 potential on Li^+ and F^-	8
PIV	Sorted distances between particles of the subsystem composed of both ions and their first coordination sphere, the four closest oxygens to Li^+ and the six closest hydrogens to F^-	66
$ACSF_{small}$	A compact set of atom-centered symmetry functions[S16] centered on both ions, designed for aqueous systems[S17]	90
$ACSF_{large}$	A larger set of ACSFs automatically designed for organic matter[S18,S19]	595
PIV^{F}_{Nw}	PIV including both ions, and the N closest water molecules to F^- ($N = 1-16$)	$10-1225$
PIV^{Li}_{Nw}	PIV including both ions, and the N closest water molecules to Li^+ ($N = 1-16$)	$10-1225$

7.8 List of collective variables investigated

References

(S1) Zhang, W.; Hartmann, C.; Schütte, C. Effective Dynamics along given Reaction Coordinates, and Reaction Rate Theory. *Faraday Discussions* **2017**, *195*, 365–394.

(S2) Pavliotis, G. A. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*; Texts in Applied Mathematics; Springer: New York, NY, 2014; Vol. 60.

(S3) Lapelosa, M.; Abrams, C. F. Transition-Path Theory Calculations on Non-Uniform Meshes in Two and Three Dimensions Using Finite Elements. *Computer Physics Communications* **2013**, *184*, 2310–2315.

(S4) Gustafsson, T.; McBain, G. D. Scikit-Fem: A Python Package for Finite Element Assembly. *Journal of Open Source Software* **2020**, *5*, 2369.

(S5) Meanti, G.; Carratino, L.; Rosasco, L.; Rudi, A. Kernel methods through the roof: handling billions of points efficiently. Advances in Neural Information Processing Systems 32. 2020.

(S6) Meanti, G.; Carratino, L.; De Vito, E.; Rosasco, L. Efficient Hyperparameter Tuning for Large Scale Kernel Ridge Regression. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. 2022.

(S7) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(S8) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*.

(S9) Liang, T.; Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics* **2020**, *48*, 1329–1347.

(S10) Sun, L.; Vandermause, J.; Batzner, S.; Xie, Y.; Clark, D.; Chen, W.; Kozinsky, B. Multitask machine learning of collective variables for enhanced sampling of rare events. *Journal of Chemical Theory and Computation* **2022**, *18*, 2341–2353.

(S11) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *The Journal of chemical physics* **2006**, *125*, 054108.

(S12) 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.

(S13) de Moivre, A. *The doctrine of chances: or, A method of calculating the probabilities of events in play*; Chelsea Publishing Company, Incorporated, 1756; Vol. 200.

(S14) Diaconis, P.; Zabell, S. Closed form summation for classical distributions: variations on a theme of de Moivre. *Statistical Science* **1991**, 284–302.

(S15) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: a Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(S16) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **2007**, *98*, 146401.

(S17) 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.

(S18) 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.

(S19) 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*.