Computing Surface Reaction Rates by Adaptive Multilevel Splitting Combined with Machine Learning and Ab Initio Molecular Dynamics

Supplementary information

Thomas Pigeon,[*,†,‡,¶] Gabriel Stoltz,[‡,†] Manuel Corral-Valero,[¶] Ani Anciaux-Sedrakian,[§] Maxime Moreaud,[¶] Tony Lelièvre,[*,‡,†] and Pascal Raybaud[*,¶]

†*MATHERIALS team-project, Inria Paris, 2 Rue Simone Iff, 75012 Paris, France*

‡*CERMICS, École des Ponts ParisTech, 6-8 Avenue Blaise Pascal, 77455,Marne-la-Vallée, France*

¶*IFP Energies Nouvelles, Rond-Point de l'Echangeur de Solaize, BP 3, 69360 Solaize, France*

§*IFP Energies Nouvelles, 1 et 4 avenue de Bois-Préau, F-92852 Rueil-Malmaison Cedex, France*

E-mail: thomas.pigeon@inria.fr; tony.lelievre@enpc.fr; raybaud@ifpen.fr

SI 1: Multilevel splitting estimator and AMS pseudo code

The various approaches using the Hill relation to compute reaction rate constants require a method to estimate the probability $p_{R\to P}(\partial R)$ to reach P before R, starting from a given distribution on the boundary ∂R of the reactant state R. This probability is often estimated using a splitting estimator such as FFS or AMS. Let us first explain why a naive Monte Carlo estimator is plagued by a large variance, before presenting the AMS estimator, and a pseudo code of the AMS algorithm. We also refer to the main text for a detailed explanation of the main steps of the algorithm.

As observing a reaction is a rare event, the probability $p_{R\to P}(\partial R)$ is typically very small which is why resorting to a simple Monte-Carlo estimator is in general not efficient. A naive Monte Carlo estimator consists in running n trajectories starting on the boundary ∂R of R and stopping them once they reach either the state R or the state P. Counting the number n_{success} of trajectories which reach P before R ($R \to P$ transitions) yields the Monte-Carlo estimator:

$$\widehat{p}_{R\to P}(\partial R) = \frac{n_{\text{success}}}{n}. \tag{1}$$

The normalised variance asscoiated with this estimator writes:

$$\text{Var}\left(\frac{\widehat{p}_{R\to P}(\partial R)}{p_{R\to P}(\partial R)}\right) = \frac{(1-p_{R\to P}(\partial R))p_{R\to P}(\partial R)}{n(p_{R\to P}(\partial R))^2} \approx \frac{1}{np_{R\to P}(\partial R)}, \tag{2}$$

as $p_{R\to P}(\partial R)$ is negligible compared to 1. From Equation (2), it is clear that the lower the transition probability is, the larger the number of trials is needed to obtain a sensible relative error.

To alleviate such a difficulty, a splitting estimator uses a product of conditional probabilities to reformulate the problem. The idea is to include the event of interest into an increasing sequence of more likely events. The target probability is then written as a product of conditional probabilities. More precisely, by introducing M surfaces $(\Sigma_j)_{1\leq j\leq M}$ between R and P

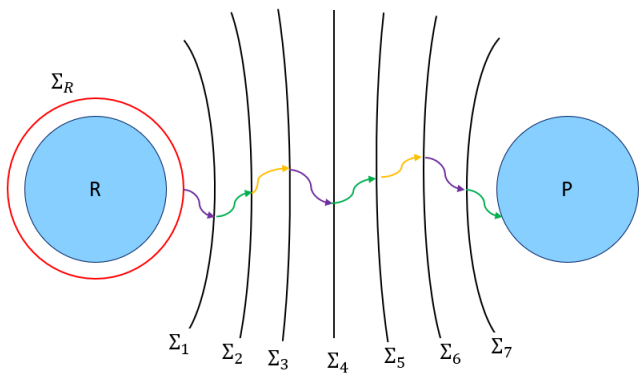
such that any transition path from R to P has to cross each of these surfaces, the probability $p_{R\to P}(\partial R)$ for the trajectory to reach P before R, starting from the boundary of R can be estimated as:

$$\widehat{p}_{R\to P}(\partial R) = \widehat{p}_{R\to\Sigma_1}(\partial R)\left(\prod_{j=1}^{M-1}\widehat{p}_{R\to\Sigma_{j+1}}(\Sigma_j)\right)\widehat{p}_{R\to P}(\Sigma_M) \tag{3}$$

where $\widehat{p}_{R\to\Sigma_1}(\partial R)$ is an estimator of the probability for the path to reach Σ_1 before going back to R, $\widehat{p}_{R\to\Sigma_{j+1}}(\Sigma_j)$ is an estimator of the probability to reach Σ_{j+1} before going back to R conditionally on the fact that Σ_j was reached before going back to R, and finally $\widehat{p}_{R\to P}(\Sigma_M)$ is an estimator of the probability to reach P before going back to R conditionally on the fact that Σ_M was reached before going back to R. It can be shown that such an estimator has a smaller variance than the Monte Carlo estimator. Moreover, for a fixed number of surfaces M, it can be shown that the variance is minimal if the surfaces are chosen such that all the conditional probabilities $p_{R\to\Sigma_{j+1}}(\Sigma_j)$ are equal. This leads to the adaptive multilevel splitting (AMS) algorithm, where the surfaces are placed adaptively on a given simulation so that the estimator of these conditional probabilities are all equal, using empirical quantiles.[1]

As an illustration, imagine than the probability to be estimated is $(1/2)^{M+1}$ (left-hand side of (3)), and that the surfaces are positioned so that all the probabilities to be estimated in the product in the right-hand side are $1/2$ (there is 50% chance to reach the next surface Σ_{j+1} before R, knowing that the path has reached Σ_j before R). In such a situation, a naive Monte Carlo estimator is plagued by a large variance since the probability $(1/2)^{M+1}$ to be estimated is very small. On the other hand, estimating $(M+1)$ times a probability of $1/2$ is much easier.

An illustration of this estimator with 7 surfaces is presented in Figure 1. To use such an estimator in practice, one needs to define the number of surfaces, their positions in phase space and to devise a way to estimate all the conditional probabilities $p_{R\to\Sigma_{j+1}}(\Sigma_j)$. The AMS algorithm is designed to solve these problems all at once.



$$p_{R\to P}(\Sigma_R) = p_{R\to\Sigma_1}(\Sigma_R)p_{R\to\Sigma_2}(\Sigma_1)p_{R\to\Sigma_3}(\Sigma_2)p_{R\to\Sigma_4}(\Sigma_3)p_{R\to\Sigma_5}(\Sigma_4)p_{R\to\Sigma_6}(\Sigma_5)p_{R\to\Sigma_7}(\Sigma_6)p_{R\to P}(\Sigma_7)$$

Figure 1: Schematic representation of a splitting estimator.

The complete pseudo code of the AMS algorithm is presented in Algorithm 1, see the main text for the explanations of each steps.

Algorithm 1: Simplified AMS pseudo algorithm

Requires;
N_{rep}, $k_{\min}$, numerical definition of the states R and P, Reaction coordinate ξ, N_{rep} initial conditions $\{\mathbf{q}^j_{\text{ini}}\}_{1\leq j\leq N_{\text{rep}}}$ on Σ , Molecular dynamics engine **MD_step**, **argsort** function returning the permutation of indices to sort an array of scalars, **randpick** functions that randomly picks an element of an array;

Output:
p: the estimated probability of reaching P before R starting from Σ.

Algorithm;
$p \leftarrow 1$;
for $i = 1$ **to** $i = N_{\text{rep}}$ **do**
  $\mathbf{q}^i_0 \leftarrow \mathbf{q}^i_{\text{ini}}$;
  $t \leftarrow 0$;
  while $\boldsymbol{q}^i_t \notin R \cup P$ **do**
    $\mathbf{q}^i_{t+\delta t} \leftarrow$ **MD_step**$(\mathbf{q}^i_t)$;
    $t \leftarrow t + \delta t$;
  $z^i_{\max} \leftarrow \sup_t(\xi(\mathbf{q}^i_t))$;
while $\exists\, i \in [\![1, N_{\text{rep}}]\!] \,/\, z^i_{\max} \neq +\infty$ **do**
  sorted $\leftarrow$ **argsort**$(z_{\max})$;
  $z_{\text{kill}} \leftarrow z_{\max}^{\text{sorted}[\text{k}_{\min}]}$;
  killed $\leftarrow \{i \in [\![1, N_{\text{rep}}]\!] \mid z^i_{\max} \leq z_{\text{kill}}\}$;
  alive $\leftarrow \{i \in [\![1, N_{\text{rep}}]\!] \mid z^i_{\max} > z_{\text{kill}}\}$;
  $p \leftarrow p\left(1 - \frac{\textbf{length}(\text{killed})}{N_{\text{rep}}}\right)$;
  for $k \in$ killed **do**
    delete $\left(\mathbf{q}^k_t\right)\ \forall t$;
    j $\leftarrow$ **randpick**(alive);
    $t \leftarrow 0$;
    while $\xi\left(\mathbf{q}^k_t\right) \leq z_{\text{kill}}$ **do**
      $\mathbf{q}^k_{t+\delta t} \leftarrow \mathbf{q}^j_{t+\delta t}$;
      $t \leftarrow t + \delta t$;
    while $\mathbf{q}^k_t \notin R \cup P$ **do**
      $\mathbf{q}^k_{t+\delta t} \leftarrow$ **MD_step**$(\mathbf{q}^k_t)$;
      $t \leftarrow t + \delta t$;
    $z^k_{\max} \leftarrow \sup_t(\xi(\mathbf{q}^k_t))$;
return p

SI 2: Rate constant error estimation

We provide in this section, an expression for the confidence interval for the reaction rate using the Delta method, which is a standard technique in statistics.[2] The Hill relation writes:

$$k_{\text{Hill}} = \Phi_R p_{R\to P}(\partial R), \tag{4}$$

where Φ_R is the flux of trajectories leaving the state R and $p_{R\to P}$ is the probability of reaching P before R starting on the boundary of R. The flux in (4) is estimated via:

$$\Phi_R = \frac{n_{\text{loop}-R\Sigma_R R}}{t_{\text{tot}}} = \frac{1}{t_{\text{loop}-R\Sigma_R R}}. \tag{5}$$

To obtain uncorrelated loop times $t_{\text{loop}-R\Sigma_R R}$, the results presented were computed considering one loop every five loops. From (4) and (5), the reaction rate writes:

$$k_{R\to P} = \frac{p_{R\to P}(\Sigma_R)}{t_{\text{loop}-R\Sigma_R R}}. \tag{6}$$

Assuming M_{real} realizations of AMS were done, let us consider the two estimators of the term of the quotient:

$$\begin{aligned} \widehat{p}_{R\to P}(\Sigma_R) &= \frac{1}{M_{\text{real}}} \sum_{i=1}^{M_{\text{real}}} p_i, \\ \widehat{t}_{loop-R\Sigma_R R} &= \frac{1}{n_{\text{loop}-R\Sigma_R R}} \sum_{j=1}^{n_{\text{loop}-R\Sigma_R R}} t_i, \end{aligned} \tag{7}$$

where p_i are independent results of AMS with the same N_{rep} and $k_{\min}$, the times t_i are the times of different loops going from R to Σ and then back to R.

The AMS estimator satisfies the central limit theorem in the limit of an infinitely large number of replicas, see Ref. 3. Concerning the estimator of this flux, the central limit theorem can be invoked only if the times t_i are not correlated. It is clear that two successive times might be correlated but due to the fact that the Langevin dynamics is stochastic, it

is possible to assume that t_i and t_{i+n} are not correlated if n is large enough. The number of times n one should skip to ensure that depends on the friction parameter of the dynamics and was assumed to be 5 in this work. Assuming that the central limit theorem holds, one obtains

$$\begin{aligned} \widehat{t}_{\text{loop}-R\Sigma_R R} &= t_{\text{loop}-R\Sigma_R R} + \sqrt{\frac{\text{Var}(\widehat{t}_{\text{loop}-R\Sigma_R R})}{n_{\text{loop}-R\Sigma_R R}}} G_t, \\ \widehat{p}_{R\to P}(\Sigma_R) &= p_{R\to P}(\Sigma_R) + \sqrt{\frac{\text{Var}(\widehat{p}_{R\to P}(\Sigma_R))}{M_{\text{real}}}} G_p \end{aligned} \tag{8}$$

where G_t and G_p are two real valued random variables distributed according to a standard Gaussian distribution. By truncation of the Taylor expansion at the first order in $\frac{1}{\sqrt{M_{\text{real}}}}$ and $\frac{1}{\sqrt{n_{\text{loop}-R\Sigma_R R}}}$, one gets:

$$\widehat{k}_{R\to P} \approx \frac{p_{R\to P}(\Sigma_R)}{t_{\text{loop}-R\Sigma_R R}} + \frac{1}{t_{\text{loop}-R\Sigma_R R}} \sqrt{\frac{\text{Var}(\widehat{p}_{R\to P}(\Sigma_R))}{M_{\text{real}}}} G_p - \frac{p_{R\to P}(\Sigma_R)}{t^2_{\text{loop}-R\Sigma_R R}} \sqrt{\frac{\text{Var}(\widehat{t}_{\text{loop}-R\Sigma_R R})}{n_{\text{loop}-R\Sigma_R R}}} G_t.$$

As the sum of two zero mean Gaussian random variables is also a zero mean Gaussian random variable, it therefore holds:

$$\widehat{k}_{R\to P} \approx k_{R\to P} + \sqrt{\frac{\text{Var}(\widehat{p}_{R\to P}(\Sigma_R))}{t^2_{\text{loop}-R\Sigma_R R} M_{\text{real}}} + \frac{p_{R\to P}(\Sigma_R)^2 \text{Var}(\widehat{t}_{\text{loop}-R\Sigma_R R})}{t^4_{\text{loop}-R\Sigma_R R} n_{\text{loop}-R\Sigma_R R}}} G_k. \tag{9}$$

Using the unbiased variance estimators

$$\begin{aligned} \text{Var}(\widehat{p}_{R\to P}(\Sigma_R)) &= \frac{1}{M_{\text{real}} - 1} \sum_{i=1}^{M_{\text{real}}} \left(p_i - \widehat{p}_{R\to P}(\Sigma_R)\right)^2, \\ \text{Var}(\widehat{t}_{\text{loop}-R\Sigma_R R}) &= \frac{1}{n_{\text{loop}-R\Sigma_R R} - 1} \sum_{i=1}^{n_{\text{loop}}} \left(t_i - \widehat{t}_{loop-R\Sigma_R R}\right)^2, \end{aligned} \tag{10}$$

and replacing $t_{loop-R\Sigma_R R}$ and $p_{\Sigma_R\to P}$ by their estimators in (9), the following confidence

interval are finally deduced:

$$
\begin{aligned}
k_{R\to P} &\in \left[\frac{\widehat{p}_{R\to P}(\Sigma_R)}{\widehat{t}_{\text{loop}-R\Sigma_R R}} - \theta_{\frac{\alpha}{2}}\sigma, \frac{\widehat{p}_{R\to P}(\Sigma_R)}{\widehat{t}_{\text{loop}-R\Sigma_R R}} + \theta_{\frac{\alpha}{2}}\sigma\right], \\
\sigma &= \sqrt{\frac{\text{Var}(\widehat{p}_{R\to P}(\Sigma_R))}{\widehat{t}^2_{\text{loop}-R\Sigma_R R} M_{\text{real}}} + \frac{\widehat{p}_{R\to P}(\Sigma_R)^2 \text{Var}(\widehat{t}_{\text{loop}-R\Sigma_R R})}{\widehat{t}^4_{loop-R\Sigma_R R} n_{\text{loop}-R\Sigma_R R}}},
\end{aligned}
\tag{11}
$$

where $\theta_{\frac{\alpha}{2}}$ stand for the quantile $\frac{\alpha}{2}$ of the Gaussian law to obtain a $1 - alpha$ precision confidence interval.

SI 3: State to state probability estimation in a multi-state case

Only two states R and P are necessary for AMS while multiple states are generally present in catalysis and the reaction rate constants between all the states $\{E_1, .. E_i, .. E_N\}$ are of interest. To address this issue with AMS, two approaches can be proposed.

First approach. We first define $R = E_j$ and $P = \bigcup_{i \neq j} E_i$. The initial conditions are sampled on the surface Σ_{E_j} surrounding the state E_j, then the transition probability $\Sigma_{E_j} \to P$ can be estimated. Finally the probabilities $\Sigma_{E_j} \to E_i$ can be estimated by counting the number of trajectories $n^{\text{in}}_{E_i}$ that indeed finished in the E_i state.

$$\widehat{p}_{E_j \to E_i}(\Sigma_{E_j}) = \frac{n^{\text{in}}_{E_i}}{N_{\text{rep}}} \widehat{p}_{E_j \to P}(\Sigma_{E_j}). \tag{12}$$

This formula is motivated later on by considering the order of the first hitting time of each state. This approach allows to observe various types of transitions using a single AMS run. However if one transition is less likely to occur compared to another, the sampling of the less probable transition might not be satisfactory as most trajectories would sample the most probable one.

Second approach. To circumvent this issue, one can decide to change the definition of the reactant and product states. Using initial conditions sampled on Σ_{E_j} and setting $R = \bigcup_{i \neq k} E_i$ and $P = E_k$, the AMS is compelled to sample the $E_j \to E_k$ trajectories. The state R contains states E_i with $i \neq k$ as it allows to consider a $E_j \to E_i$ as a non reactive trajectory. This matter is discussed in the result section.

Justification of equation (12)**.** Let us consider a 3 states case. The reactant state is $R = E_1$ and the product state is $P = E_2 \cup E_3$. Then we define the time τ_{E_i} as the first that the dynamics starting on Σ_R reaches the state E_i. As three states are considered, six possibilities for the ranking of these three time are possible:

1. $\tau_{E_1} < \tau_{E_2} < \tau_{E_3}$,

2. $\tau_{E_1} < \tau_{E_3} < \tau_{E_2}$,

3. $\tau_{E_2} < \tau_{E_1} < \tau_{E_3}$,

4. $\tau_{E_2} < \tau_{E_3} < \tau_{E_1}$,

5. $\tau_{E_3} < \tau_{E_1} < \tau_{E_2}$,

6. $\tau_{E_3} < \tau_{E_2} < \tau_{E_1}$.

Hence, the sum of the probability of these 6 events in equals to 1 and, given that the states do not overlap, these events are independent. It is possible to identify that the event corresponding to reaching P before R correspond the events 3 to 6 while the events 1 and 2 correspond to reaching R first. Reaching E_2 before R corresponds to the events 3 and 4 and finally reaching E_3 before R corresponds to events 5 and 6. Then, to estimate the probability of the last two events one has just to identify the fraction of events 3 and 4 (or 5 and 6) that occurred during the realization of the events 3 to 6. This is exactly what the factor $\frac{n^{\text{in}}_{E_i}}{N_{\text{rep}}}$ represents in expression (12).

SI 4: Calculation parameters

SI 4.1: DFT parameters

The DFT functional was PBE[4] with the D3 dispersion correction.[5] A Gaussian smearing was used with $\sigma = 0.05$ eV. The γ-alumina bulk structure taken from Ref. 6. The K point grid was set to $2 \times 2 \times 4$ and centered at the Γ point. The bulk structure was first fully relaxed (allowing box volume to change) with a 800 eV kinetic energy cutoff to ensure low Pulay stress. All other DFT calculations on slabs representing the γ-alumina (100) surface were achieved by keeping the cell's volume constant with a kinetic energy cutoff of 450 eV. The (100) surface model is composed of a four layer slab structure with 15 Å of vacuum inserted in the direction perpendicular to the surface plane (that is, the x direction in this case). Calculations on this system were carried out with a K point grid set to $1 \times 2 \times 4$. Geometry optimizations were done using the conjugate gradient algorithm as implemented in VASP with a convergence criterion of 0.01 eV/Å.

SI 4.2: AIMD parameters.

All the molecular dynamics runs were generated using the Brünger Brooks Karplus integrator of the Langevin dynamics implemented in VASP. A 1.0 fs time step was used for all of them. The length of the dynamics runs and the used friction parameter varied upon the purpose of the molecular dynamics run as detailed in the results section of the main text.

SI 4.3: NEB and saddle points

Saddle points on the potential energy surface were identified by nudged elastic band (NEB) methods using the VASP TST tools.[7,8] The spring force was taken to 5.0 eV.Å^{-2} and nudging was turned on. The number of images was 10 including the reactant and product. The optimizer used was FIRE with the default parameters that can be found in the VTST documentation.[8] The initial path was created by an interpolation between the z-matrix represen-

tation of the reactant and product structures with the Opt'n path code.[9] The identification of the relevant saddle points on the potential energy surface was done starting from the NEB results and refined using the quasi-Newton method. The vibrational frequencies of the minima and the saddle points were evaluated using a finite difference method as implemented in the VASP package based on displacements of 0.01 Å. Using these frequencies, the free energies of the meta-stable basins and the transition states were computed within the harmonic approximation. The rotational components of the entropy were not explicitly computed and assumed to cancel out between minima and transition states. Detailed expressions used can be found in Ref. 10. Finally the hTST rates were computed using Eyring-Polanyi equation assuming that the transmission coefficient is equal to 1.

SI 4.4: SOAP descriptor parameters

The SOAP descriptors were computed using the dscribe python package.[11] The cutoff radius was set to 6 Å as the main structural changes in the example system are within a sphere of this radius around a central atoms. The atomic density in the neighborhood of an atom is approximated as a sum of Gaussians centered on the nearby atomic nuclei and their width σ was 0.05 Å. The width of these Gaussians is chosen so that there is not too much overlap between two different structures. The parameters $n_{\max}$ and $l_{\max}$ controlling the size of the basis on which the atomic density is projected were respectively set to 8 and 6.

SI 5: Implementation with VASP software

The Fleming-Viot particle process to sample initial conditions and the AMS were implemented in python scripts calling the execution of the VASP software for the integration of the unbiased Langevin dynamics. Both these algorithms require to stop the dynamics when it enters a certain state. This means that at every time-step, one has to evaluate the criterion used to define this state and then decide whether the dynamics is to be continued or not. This kind of stopping condition cannot be enforced with the current implementation of VASP and had to be implemented. The collective variable to define the states is computed using a python script `CV.py` that is used as input. The choice of the stopping conditions is monitored by INCAR tags.

SI 6: Detailed numerical results

The most precise results for each observed transitions during this study are presented in Table 1. From these results, some reaction heats were computed and presented in Table 2.

Table 1: Transition rate constants computed with AMS for all the transition observed in this study with 90% precision.

Transition	$t_{\text{loop}-R\Sigma_R R}$ (fs)	$p_{\text{Transition}}$	$k_{\text{Transition}}(\text{s}^{-1})$
Water rotations			
$A_1 \rightarrow A_2A_3$ [a]	110 ± 5	$(3.38 \pm 1.56)\ 10^{-3}$	$(3.08 \pm 1.43)\ 10^{10}$
$A_2A_3 \rightarrow A_1$ [b]	85 ± 9	$(1.34 \pm 0.41)\ 10^{-2}$	$(1.49 \pm 0.46)\ 10^{11}$
$A_2A_3 \rightarrow A_4$ [b]	85 ± 9	$(3.90 \pm 1.97)\ 10^{-3}$	$(4.33 \pm 2.20)\ 10^{10}$
$A_4 \rightarrow A_2A_3$ [c]	100 ± 5	$(2.24 \pm 0.83)\ 10^{-2}$	$(2.35 \pm 0.87)\ 10^{11}$
$A_1 \rightarrow A_4$ [a]	110 ± 3	$(3.66 \pm 7.18)\ 10^{-7}$	$(3.34 \pm 6.56)\ 10^{6}$
$A_4 \rightarrow A_1$ [c]	100 ± 5	$(1.28 \pm 0.65)\ 10^{-3}$	$(1.34 \pm 0.68)\ 10^{10}$
Hydroxyl rotation			
$D_1D_3 \rightarrow D_2D_4$ [d]	$\varnothing$	$\varnothing$	$\varnothing$
$D_2D_4 \rightarrow D_1D_3$ [e]	63 ± 4	$(1.72 \pm 2.84)\ 10^{-5}$	$(2.86 \pm 4.71)\ 10^{8}$
Formation and dissociation of water			
$A_1 \rightarrow D_1D_3$ [f]	109 ± 3	$(1.78 \pm 1.64)\ 10^{-4}$	$(1.64 \pm 1.59)\ 10^{9}$
$D_1D_3 \rightarrow A_1$ [d]	88 ± 4	$(2.05 \pm 1.40)\ 10^{-3}$	$(2.32 \pm 1.59)\ 10^{10}$
$A_2A_3 \rightarrow D_2D_4$ [b]	85 ± 9	$(7.07 \pm 6.77)\ 10^{-4}$	$(7.86 \pm 7.53)\ 10^{9}$
$D_2D_4 \rightarrow A_2A_3$ [e]	63 ± 4	$(7.71 \pm 3.24)\ 10^{-3}$	$(1.28 \pm 0.54)\ 10^{11}$
$A_2A_3 \rightarrow D_1D_3$ [b]	$\varnothing$	$\varnothing$	$\varnothing$
$D_1D_3 \rightarrow A_2A_3$ [d]	88 ± 4	$(2.05 \pm 2.77)\ 10^{-5}$	$(2.33 \pm 3.14)\ 10^{8}$

[a] Rate sampled using $N_{\text{rep}} = 200$, $M_{\text{real}} = 10$, $R = A_1$, $P = A_2A_3 \cup A_4 \cup D_1D_3 \cup D_2D_4$ and $\xi = A_1$-vs-all SOAP-SVM RC.
[b] Rate sampled using $N_{\text{rep}} = 100$, $M_{\text{real}} = 20$, $R = A_2A_3$, $P = A_1 \cup A_4 \cup D_1D_3 \cup D_2D_4$ and $\xi = A_2A_3$-vs-all SOAP-SVM RC.
[c] Rate sampled using $N_{\text{rep}} = 200$, $M_{\text{real}} = 10$, $R = A_4$, $P = A_1 \cup A_2A_3 \cup D_1D_3 \cup D_2D_4$ and $\xi = A_4$-vs-all SOAP-SVM RC.
[d] Rate sampled using $N_{\text{rep}} = 200$, $M_{\text{real}} = 10$, $R = D_1D_3$, $P = A_1 \cup A_2A_3 \cup A_4 \cup D_2D_4$ and $\xi = D_1D_3$-vs-all SOAP-SVM RC.
[e] Rate sampled using $N_{\text{rep}} = 200$, $M_{\text{real}} = 10$, $R = D_2D_4$, $P = A_1 \cup A_2A_3 \cup A_4 \cup D_1D_3$ and $\xi = D_2D_4$-vs-all SOAP-SVM RC.
[f] Rate sampled using $N_{\text{rep}} = 200$, $M_{\text{real}} = 10$, $R = A_1 \cup A_2A_3 \cup A_4 \cup D_2D_4$, $P = D_1D_3$, $\Sigma_R = \Sigma_{A_1}$ and $\xi =$ interpolated SOAP-PCV

Similar results can be obtained via the hTST approach and are presented in Tables 3 and 4.

Table 2: Reaction heats at 200 K computed AMS reaction rates of Table 1

	Value (kJ.mol^{-1})
Water rotations	
$\Delta G_{A_1 \to A_2 A_3}$	2.62 ± 2.66
$\Delta G_{A_2 A_3 \to A_4}$	2.81 ± 2.83
$\Delta G_{A_1 \to A_4}$	13.8 ± 4.45
Water dissociations	
$\Delta G_{A_2 A_3 \to D_2 D_4}$	4.64 ± 3.54
$\Delta G_{A_1 \to D_1 D_3}$	4.41 ± 3.88

Table 3: Activation energies and rate constant computed with the harmonic approximation at 200K.

Transition	$\Delta G^{\ddagger}_{\text{Transition}}$ (kJ.mol^{-1})	$k_{\text{Transition}}$(s^{-1})
Water rotations		
$A_1 \to A_2 A_3$	6.67	$7.55\ 10^{10}$
$A_2 A_3 \to A_1$	1.17	$2.06\ 10^{12}$
$A_2 A_3 \to A_4$	7.88	$3.64\ 10^{10}$
$A_4 \to A_2 A_3$	3.31	$5.66\ 10^{11}$
$A_1 \to A_4$	16.5	$2.04\ 10^{8}$
$A_4 \to A_1$	6.44	$8.65\ 10^{10}$
Hydroxyl rotation		
$D_1 D_3 \to D_2 D_4$	12.4	$2.38\ 10^{9}$
$D_2 D_4 \to D_1 D_3$	11.5	$4.15\ 10^{9}$
Formation and dissociation of water		
$A_1 \to D_1 D_3$	4.18	$3.37\ 10^{11}$
$D_1 D_3 \to A_1$	2.17	$1.13\ 10^{12}$
$A_2 A_3 \to D_2 D_4$	-4.28	$5.45\ 10^{13}$
$D_2 D_4 \to A_2 A_3$	-1.71	$1.17\ 10^{13}$

Table 4: Reaction heats computed from harmonic approximation of the free energy at 200 K

	Value (kJ.mol^{-1})
Water rotations	
$\Delta G_{A_1 \to A_2 A_3}$	5.50
$\Delta G_{A_2 A_4 \to A_4}$	4.56
$\Delta G_{A_1 \to A_4}$	$10,1$
Water dissociations	
$\Delta G_{D_1 D_3 \to D_2 D_4}$	$0,93$
Water dissociations	
$\Delta G_{A_2 A_3 \to D_2 D_4}$	-2.56
$\Delta G_{A_1 \to D_1 D_3}$	2.01

SI 7: Clustering reactive trajectories

K-means and most clustering algorithms better perform for clustering problems in low dimensions. Performing clustering using the whole reactive trajectories is therefore expected to be inefficient. Moreover, the sampled reactive trajectories do not necessary have the same length and cannot directly be compared. To alleviate these problems, a first preprocessing step is to summarize each trajectory by a small number of structures. These structures correspond to the first point of the trajectory crossing certain reaction coordinate isolevels. In the example presented in the result section of the main text, five levels were used. These levels are equally spaced between the largest minimal values of the RC along the trajectories and the smallest maximal values of the RC along the trajectories. The next preprocessing step is to numerically represent these few structures per trajectory via the SOAP descriptor centered on the oxygen atom of the water molecule. As these SOAP descriptors are also in high dimension, a principal component analysis is performed for all the normalized SOAP descriptors of structures corresponding to the same level of the RC. Finally, the first four principal components are used as descriptors of the structure. The choice of using the first four principal components is motivated by the fact that these four components capture at least 90% of the variance of the SOAP descriptors for a given level. Finally, a full reactive trajectory of is represented as a vector of size four times the number of levels chosen. As the K-means algorithm strongly depends on its (random) initialisation, the algorithm was repeated 20 times and the best set of clusters was kept. Setting the number of clusters to find to 3 allows to find the two types of pathways for the $A_4 \rightarrow A_1$ rotation which were previously discussed. The trajectories are attached as videos in the electronic SI. Three clusters are necessary as both paths are not equally sampled by the AMS simulation. Indeed, one path being less probable, is less sampled. This behavior is typical of K-means which prefers to split one large cluster in two parts rather than identifying a large one and a much smaller one. Of course such issues could be alleviated by resorting to other clustering methods but

such study is beyond the scope of the present work. However, the result obtained at this level paves the way to future more detailed investigations.

References

(1) Bréhier, C.-E.; Lelièvre, T.; Rousset, M. Analysis of adaptive multilevel splitting algorithms in an idealized case. *ESAIM - Probab. Stat.* **2015**, *19*, 361–394.

(2) Hogg, R. V.; McKean, J. W.; Craig, A. T. *Introduction to Mathematical Statistics*; Pearson, p 768.

(3) Cérou, F.; Delyon, B.; Guyader, A.; Rousset, M. On the Asymptotic Normality of Adaptive Multilevel Splitting. *SIAM-ASA J. Uncertain. Quantif.* **2019**, *7*, 1–30.

(4) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(5) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(6) Krokidis, X.; Raybaud, P.; Gobichon, A.-E.; Rebours, B.; Euzen, P.; Toulhoat, H. Theoretical study of the dehydration process of Boehmite to γ-Alumina. *J. Phys. Chem. B.* **2001**, *105*, 5121–5130.

(7) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. Classical and quantum dynamics in condensed phase simulations. 1998; pp 385–404.

(8) Henkelman, G. `https://theory.cm.utexas.edu/vtsttools/index.html`.

(9) Fleurat-Lessard, P. `http://pfleurat.free.fr/ReactionPath.php`.

(10) McDouall, J. J. W. *Computational quantum chemistry*; Royal Society of Chemistry, 2013.

(11) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.