RARE EVENT SAMPLING METHODS AND MACHINE LEARNING TO STUDY CATALYTIC REACTION MECHANISMS

THOMAS PIGEON

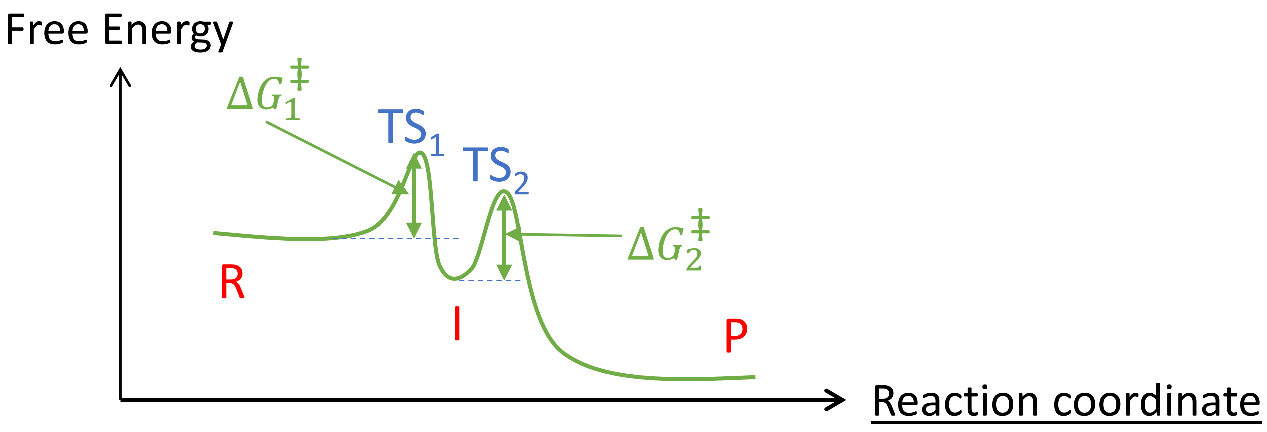
Pascal Raybaud[1], Manuel Corral Valero[1],
Ani Anciaux-Sedrakian[2], Maxime Moreaud[2]
Gabriel Stoltz[3], Tony Lelièvre[3]

[1] IFPEN R06: Catalysis, Biocatalysis and Separation
[2] IFPEN R11: Digital Science and Technology
[3] CERMICS, Ecole des Ponts ParisTech, and Equipe-projet MATHERIALS, Inria Paris,

INTRODUCTION: REACTION MECHANISMS

Targets : compute reaction rates

identify reaction mechanism



Different methods exist:

- Transition State Theory (TST): for instance, Eyring-Polanyi equation[1] $\boldsymbol{k}^{hTST} = \frac{k_B T}{h} e^{-\frac{\Delta G^\ddagger}{k_B T}}$

  Using free energy computed by static approach within harmonic approximation or Molecular Dynamics (MD)
- **Alternatively: MD and Rare events simulation methods to access directly the reaction time**

  Hill relation[2]: $\boldsymbol{k}^{Hill} = p_{R\rightarrow P}\phi_R$

[1] Eyring, H. (1935). The activated complex in chemical reactions. *The Journal of Chemical Physics*, *3*(2), 107-115.

[2] Hill, T. (2012) Free energy transduction in biology: The steady-state kinetic and thermodynamic formalism. *Elsevier Science and Technology Books*

INTRODUCTION: STANDARD MOLECULAR DYNAMICS

Simulates the dynamic of the system by adding a thermostat to newton equations of motion

ex. Langevin formalism[1]

NVE ensemble

NVT ensemble

$$\begin{cases} dq_t = M^{-1} p_t dt \\ dp_t = -\nabla V(q_t) dt - \gamma p_t dt + \sqrt{2\gamma M k_B T} dW_t \end{cases}$$

Preserves energy
Newton equation

Dissipate energy

Provides energy

Langevin part

Not efficient for the simulation of rare events due to high energy barriers and entropic bottlenecks

Time scales: integration time step : $\sim 10^{-15} s$ rare event rate $\sim 10^{-9} s^{-1}$ to $10^{3} s^{-1}$

MD based approaches to overcome barriers:

- TST → biased MD such as Metadynamics[2], Blue-Moon sampling[3] …
  Dynamics is lost but rates are estimated from free energy
- Hill → rare events sampling methods such as Adaptive multi-level splitting[4]
  Dynamics preserved thus the rates can directly be computed

[1] Langevin P. (1908), *Comptes-Rendus de l'Académie des Sciences*, *146*, 530-532

[2] Laio, A., & Parrinello, M. (2002) *Proceedings of the National Academy of Sciences*, *99*(20), 12562-12566.

[3] Carter, E. A., Ciccotti, G., Hynes, J. T., & Kapral, R. (1989). *Chemical Physics Letters*, *156*(5), 472-477.

[4] Cérou, F., & Guyader, A. (2007) *Stochastic Analysis and Applications*, *25*(2), 417-443.

OUTLINE

I. Adaptive Multi-level Splitting (AMS) for reaction rates

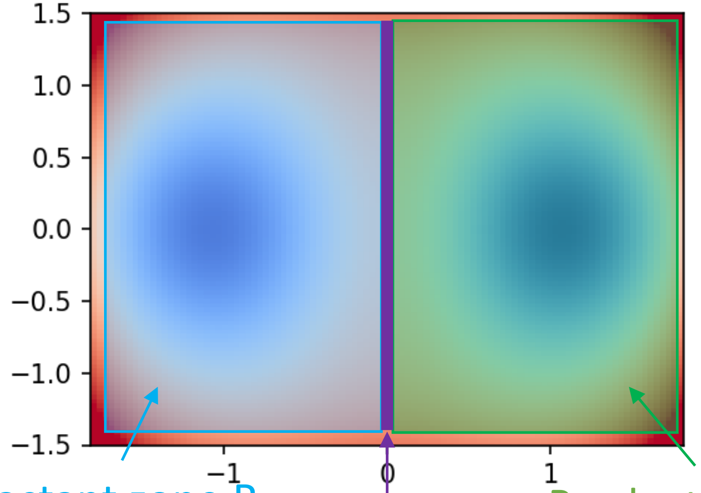
II. Identifying reaction coordinate with Machine Learning (ML)

III. Results of the AMS + ML method.

CASE STUDY: Kinetics of dissociation of H_2O on γ-Al_2O_3 (100) surface

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

Transition State Theory

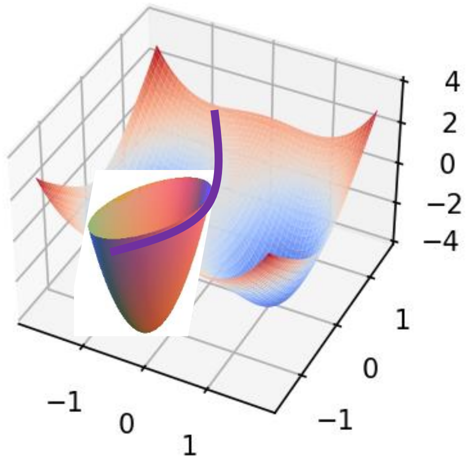


Reactant zone R

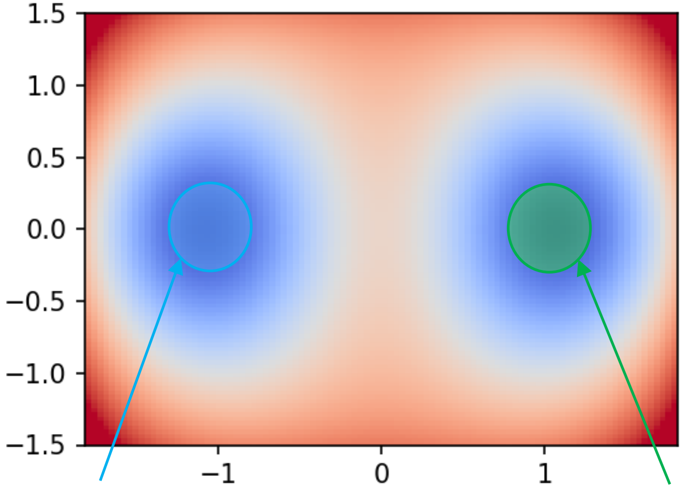
TS surface

Product zone P

2-dimensional potential



Hill relation



Reactant zone R

Product zone P

Rate = probability of being in TS with respect to R × frequency of decomposition to P

$$\boldsymbol{k}^{TST} = p(TS \mid R)\ \phi_{TS \to P} \qquad \boldsymbol{k}^{hTST} = e^{-\frac{\Delta G^{\ddagger}}{k_{\mathrm{B}}T}} \frac{k_{\mathrm{B}}T}{h}$$

Sensitive to the TS definition

TST overestimates rates (κ)

hTST poorly captures entropy

Rate = probability of reaching P before R starting from ∂R × frequency of exits of R

$$\boldsymbol{k}^{Hill} = p_{R \to P}(\partial R)\ \phi_R$$

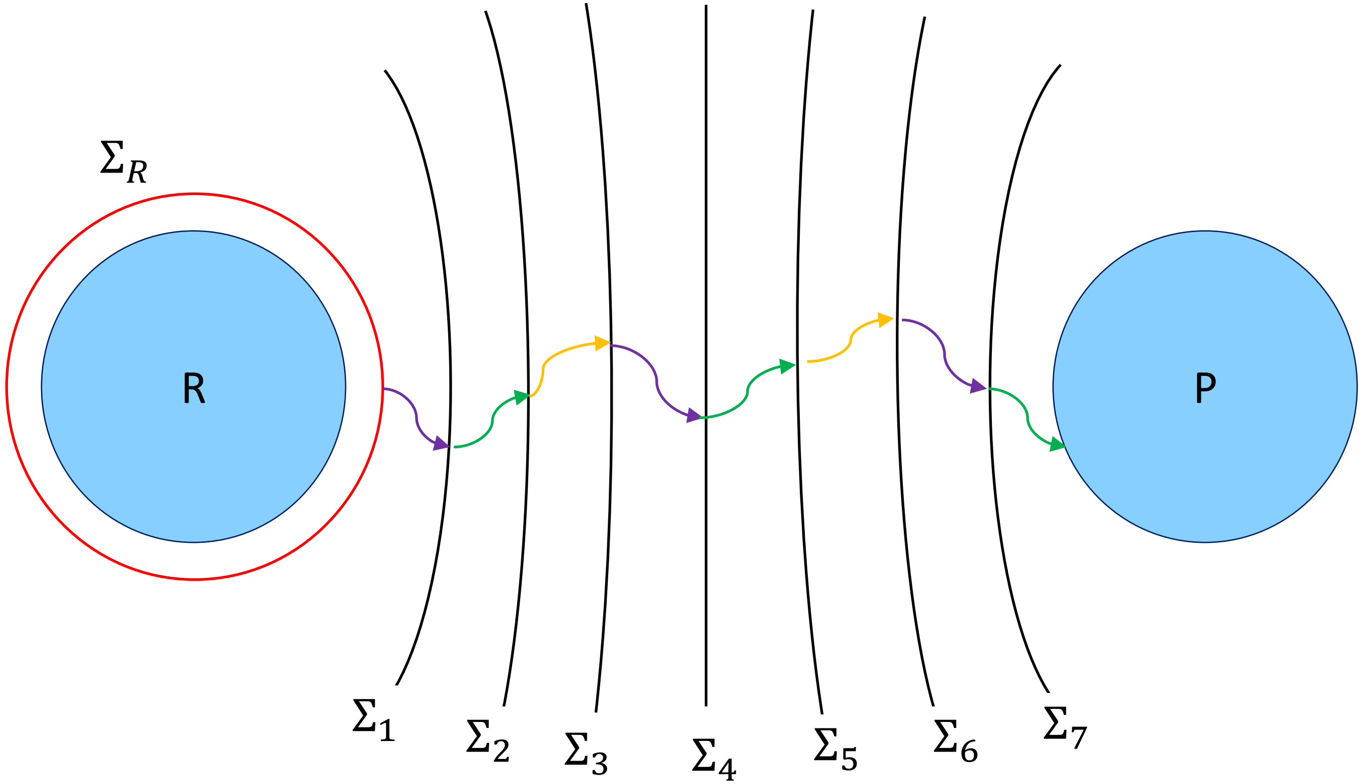
Not extremely sensitive to the definition of R and P

[1] Hänggi, P. Talkner, P. Borkovec, M. (1990) Reaction-rate theory: fifty years after Kramers *Reviews of Modern Physics* , Vol. 62, No. 2 American Physical Society (APS) p. 251-341

[2] Hill, T. (2012) Free energy transduction in biology: The steady-state kinetic and thermodynamic formalism. *Elsevier Science and Technology Books*

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

What is a Multilevel Splitting estimator:



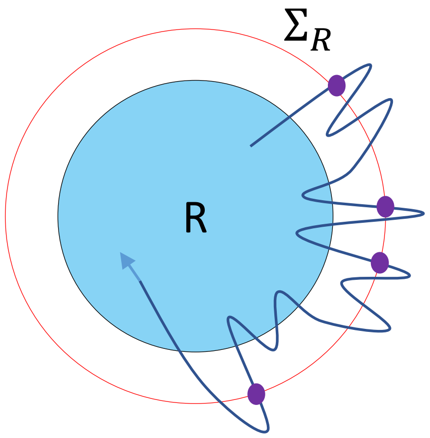
$$p_{R\to\Sigma_1}(\Sigma_R)p_{R\to\Sigma_2}(\Sigma_1)p_{R\to\Sigma_3}(\Sigma_2)p_{R\to\Sigma_4}(\Sigma_3)p_{R\to\Sigma_5}(\Sigma_4)p_{R\to\Sigma_6}(\Sigma_5)p_{R\to\Sigma_7}(\Sigma_6)p_{R\to P}(\Sigma_7)$$
$$= p_{R\to P}\ (\Sigma_R)$$

How to place Σ_i and compute $p_{\mathrm{R}\to\Sigma_{i+1}}(\Sigma_i)$?

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

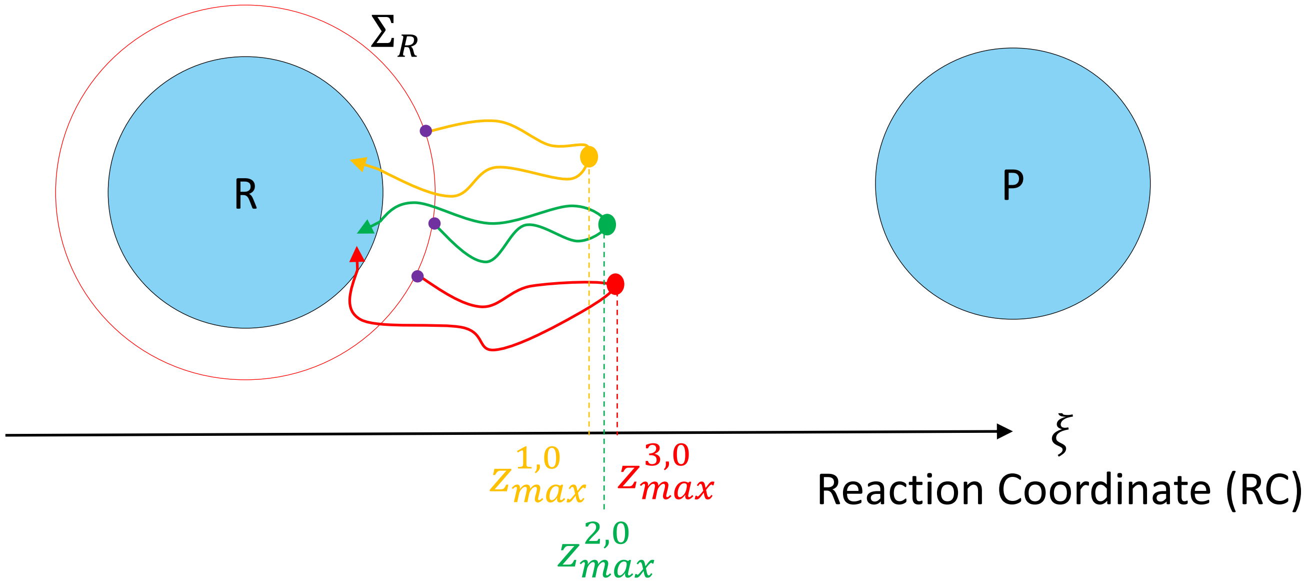
- AMS aims at estimating $p_{\Sigma \to P}$[1,2]. It can be split in 3 steps:
  1. Generating initial conditions on Σ and estimate $t_{R-\Sigma-R} = \frac{1}{\phi_R}$
  2. Initialize N replicas by running an unbiased dynamics until it reaches R or P. Set p = 1. Classify all the replicas by increasing ξ_{max}.
  3. Apply the AMS loop until all replicas have reached P.

1. Initial conditions and flux



MD for Initial conditions.

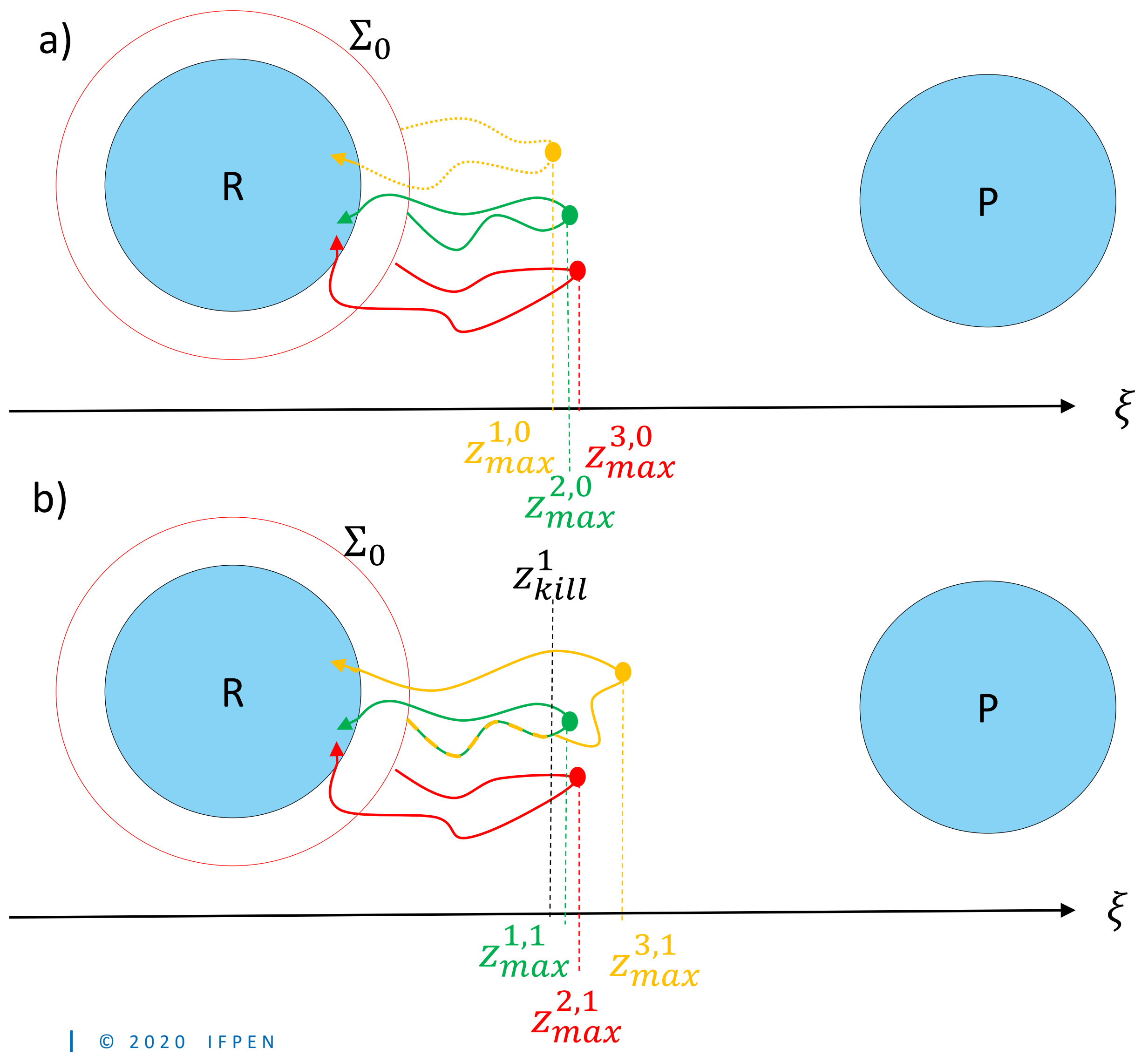
2. Initialization



[1] F. Cérou, A. Guyader, ***Stochastic Analysis and Applications* 25**, 417-443 (2007).
[2] L. J. S. Lopes, T. Lelièvre, ***Journal of computational chemistry* 40**, 1198-1208 (2019).

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

3. AMS interations



3. AMS iterations: $i \geq 0$

a) Save the smallest ($z_{max}^{1,i}$) as z_{kill}^{i+1} and delete all the trajectories that did not "go above" z_{kill}^{i+1}

b) Randomly select one trajectory within the remaining ones. Copy it until it reaches z_{kill}^{i+1} and continue it until it reaches R or P.

c) Classify all the replicas by increasing z_{max}.

$$\tilde{p} = \prod_{i=0}^{i_{max}} \tilde{p}_{\Sigma_{z_{kill}^{i}} \to \Sigma_{z_{kill}^{i+1}}} = \left(1 - \frac{1}{N}\right)^{i_{max}}$$

Unbiased estimator:	Variance depends on RC:
$\mathbb{E}[\tilde{p}] = p_{R-P}(\Sigma_R)$	$\mathrm{Var}[\tilde{p}] = f(\xi)$

OUTLINE

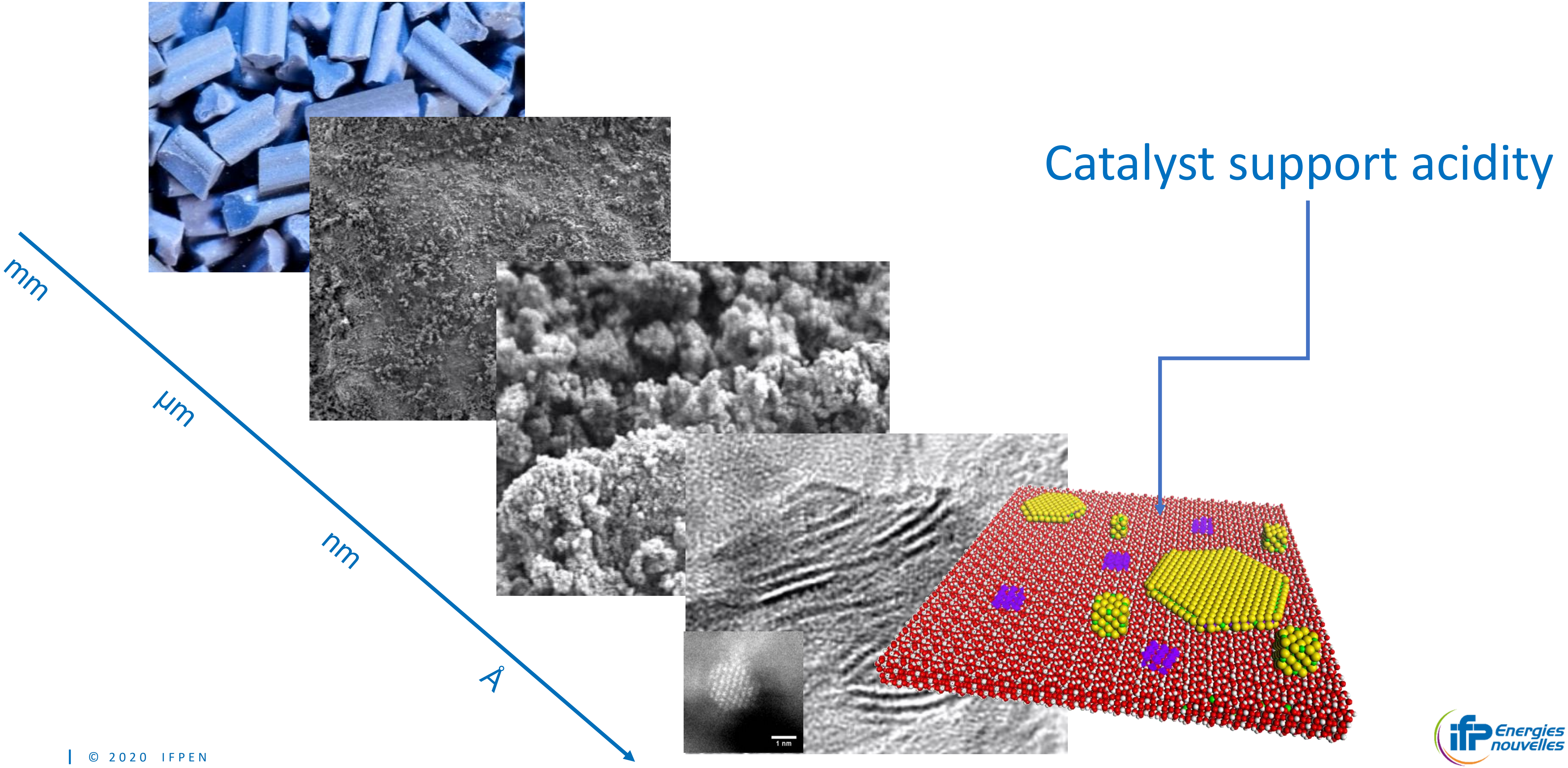
I. Adaptive Multi-level Splitting (AMS) for reaction rates

II. Identifying reaction coordinate with Machine Learning (ML)

III. Results of the AMS + ML method.

CASE STUDY: Dissociation of H_2O on γ-Al_2O_3 (100) surface

SYSTEM PRESENTATION



Catalyst support acidity

II. IDENTIFYING COLLECTIVE VARIABLES WITH MACHINE LEARNING

- Each structure of N atoms is a point in $\mathbb{R}^{3N}$
- Collective Variables(CV) are synthetic variables in lower dimensions.

$$\xi : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{n}, n = 1, 2, 3...$$

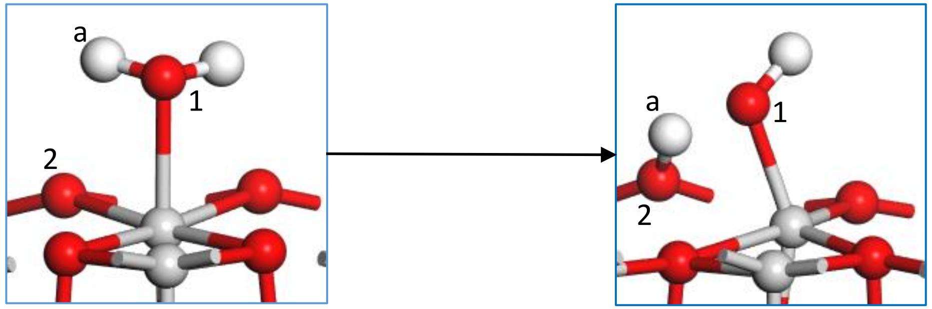
- A reaction coordinate is one, or a set of collective variables able to discriminate the important states of the system.
- An ideal reaction coordinate answers : **how committed is the dynamic in the process of going from Reactants to Products ?**

Potentially bad RC:

dist(Al-O_1)

Reactant

Product



Potentially good RC:

dist(O_1-H_a)

dist(O_2-H_a)

● H

● O

● Al

II. IDENTIFYING COLLECTIVE VARIABLES WITH MACHINE LEARNING

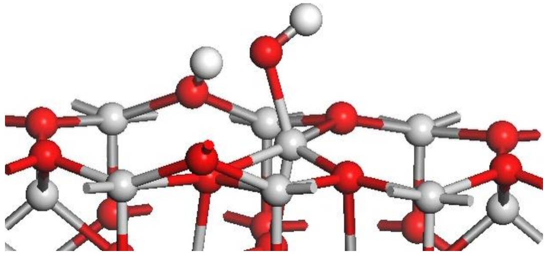
Method:

1. Identify the various metastable states (intermediates)

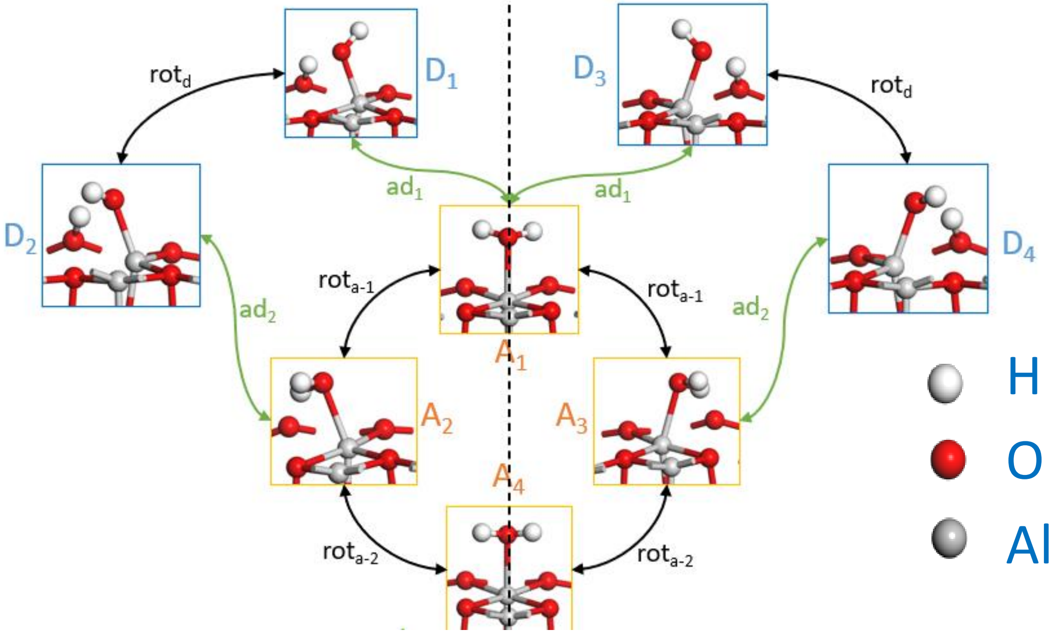
   → dissociated (D_i) or associated (A_i)

2. Run short dynamics in these states to sample Potential Energy Surface (PES) around the minima

D_1 dissociated structure



VASP software
43 atoms
dt = 1 fs
Total time = 1 ps
3 – 4 wall clock hours on 1 node with 36 CPUs



Identified structures and **intuitively** plausible transitions

3. Train supervised machine learning model

   (with the proper labelling)

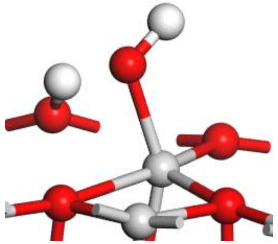
II. IDENTIFYING COLLECTIVE VARIABLES WITH MACHINE LEARNING

SOAP[1] atom centered descriptors to numerically encode the structure for training the ML algorithm.

○ H

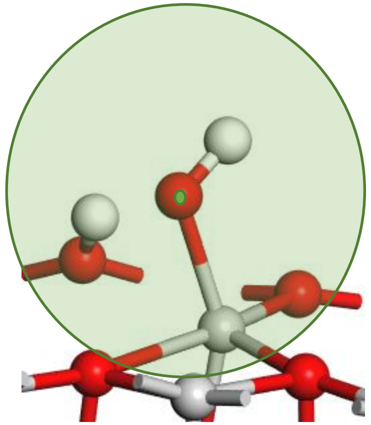
● O

● Al



Periodic structures
(vector of 3N lines)

Select central atom



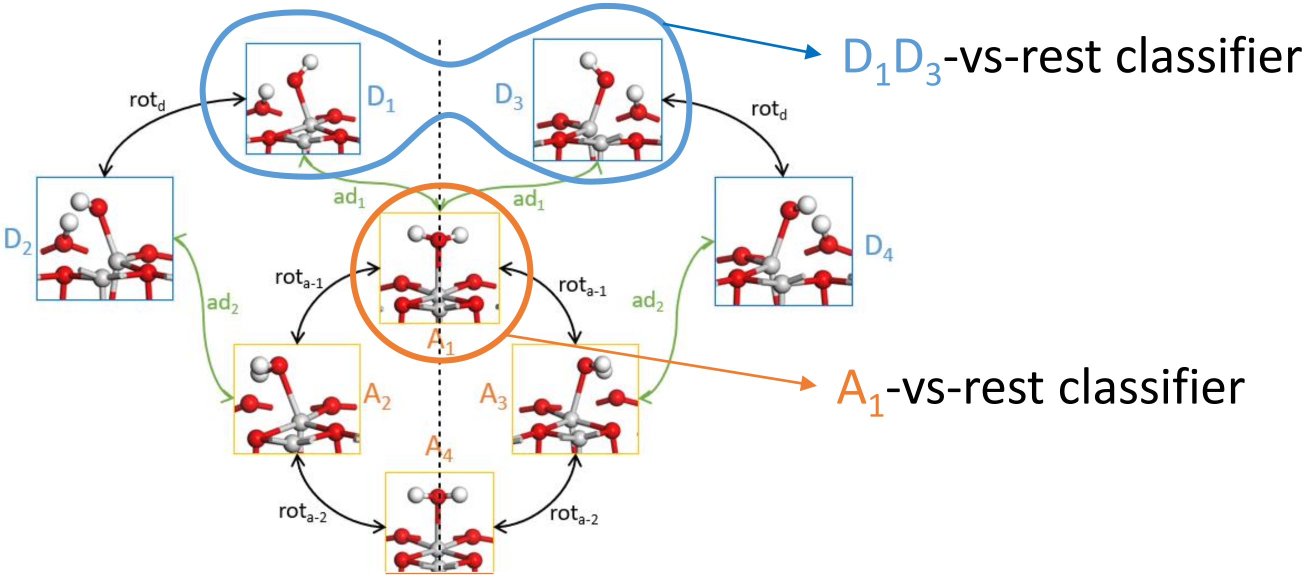
Atom centered description of the structure
(vector of ~ 10^3-10^4 lines)

[1] Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. *Physical Review B*, *87*(18), 184115.

II. IDENTIFYING COLLECTIVE VARIABLES WITH MACHINE LEARNING

SVM classifiers separate two sets of points by the highest margin plane.

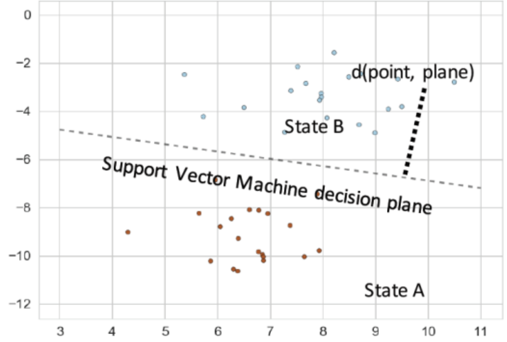
SOAP-SVM CV : classifier decision function (f_X): algebraic distance to the plane.

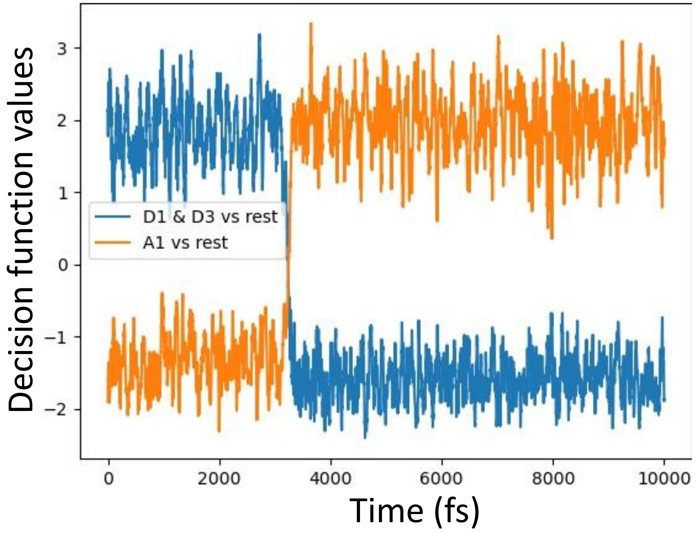


D_1D_3-vs-rest classifier

A_1-vs-rest classifier

[1]





Classifier decision function interpretation:

$$f_X(\boldsymbol{q}) \in (-\infty, -1] \iff \boldsymbol{q} \in X$$

[1] Sultan, M. M.; Pande, V. S. (2018) Automated design of collective variables using supervised machine learning *The Journal of Chemical Physics*, 149, 094106.

OUTLINE

I. Adaptive Multi-level Splitting (AMS) for reaction rates

II. Identifying reaction coordinate with Machine Learning (ML)

III. Results of the AMS + ML method.

CASE STUDY: Dissociation of H_2O on γ-Al_2O_3 (100) surface

III. RESULTS OF THE AMS + ML METHOD.

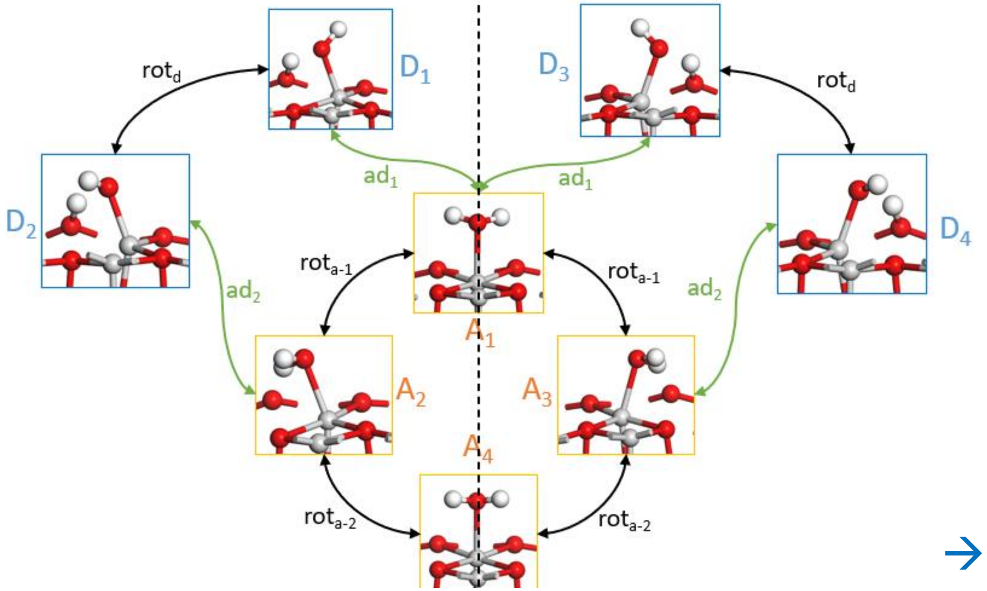
With
$R = A_1$
$\Sigma_R = \Sigma_{A_1}$
$P = A_2A_3 \cup A_4 \cup D_1D_3 \cup D_2D_4$

AMS can sample :

$A_1 \rightarrow A_2A_3$
$A_1 \rightarrow A_4$
$A_1 \rightarrow D_1D_3$
$A_1 \rightarrow D_2D_4$

→ Answers how A_1 can decompose ?

The most probable transition will be sampled, with precision conditioned by ξ



With
$R = A_1 \cup A_2A_3 \cup A_4 \cup D_2D_4$
$\Sigma_R = \Sigma_{A_1}$
$P = D_1D_3$

AMS can sample :

$A_1 \rightarrow D_1D_3$

→ Focus specifically on one dissociation

Quality of the sampling depends on ξ

III. RESULTS OF THE AMS + ML METHOD.

Multiple type of trajectories:

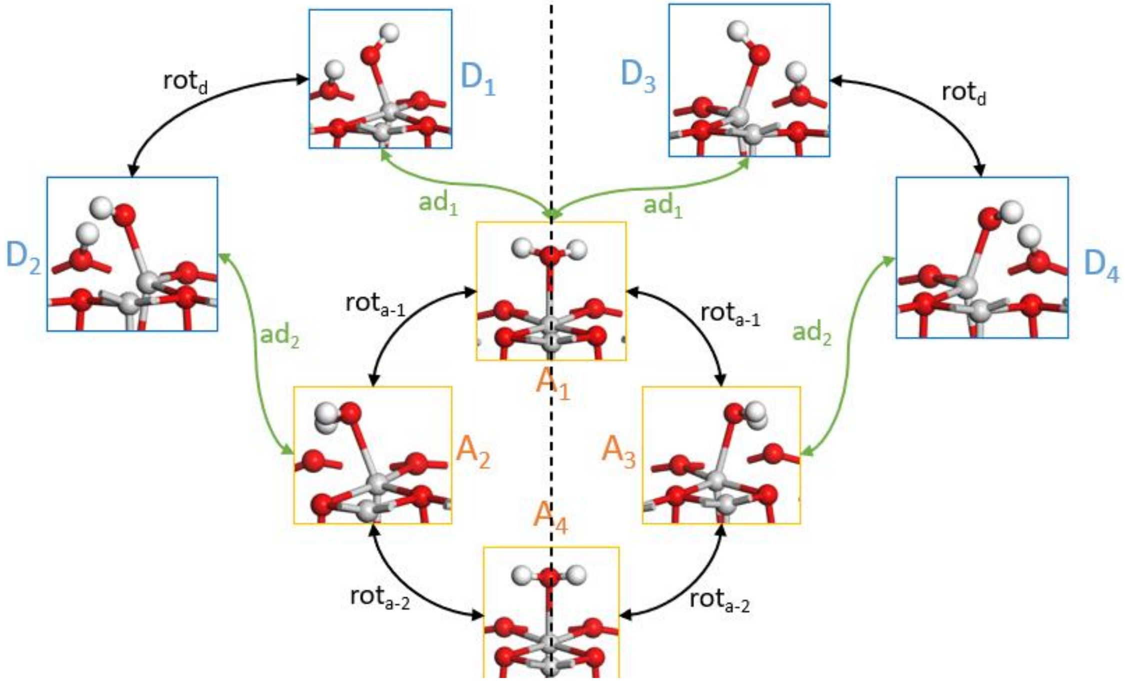
$\tilde{p}_{\mathrm{A_1 \to any}}$, $\tilde{p}_{\mathrm{A_2 A_3 \to any}}$, $\tilde{p}_{\mathrm{A_4 \to any}}$, $\tilde{p}_{\mathrm{D_1} D_3 \mathrm{\to any}}$ and $\tilde{p}_{\mathrm{D_2 D_4 \to any}}$

Final state can be identified using all CVs

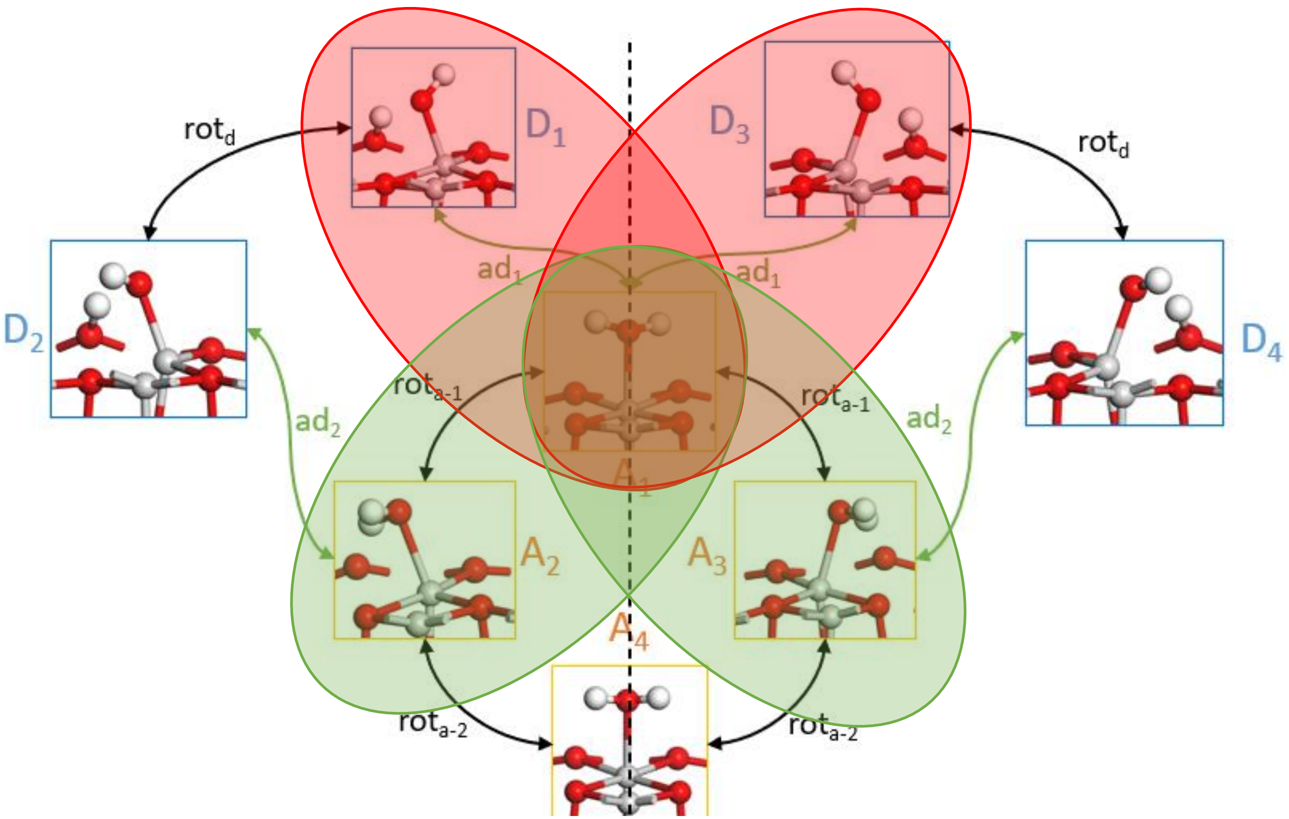
Count the number of replicas $n_{\mathrm{X}}^{\mathrm{in}}$ finishing in the state X

$$\tilde{p}_{\mathrm{A_1 \to X}} = \frac{n_{\mathrm{X}}^{\mathrm{in}}}{N} \tilde{p}_{\mathrm{A_1 \to any}}$$

$$\tilde{k}_{\mathrm{A_1 \to X}} \approx \frac{\tilde{p}_{\mathrm{A_1 \to X}}}{t_{loop-A_1}}$$



III. RESULTS OF THE AMS + ML METHOD.



Dissociation	Hill	hTST
$k_{A_1 \to D_1 D_3} =$	$1.6\ 10^{9}\ s^{-1}$	$3.4\ 10^{11}\ s^{-1}$
$k_{D_1 D_3 \to A_1} =$	$2.3\ 10^{10}\ s^{-1}$	$1.1\ 10^{12}\ s^{-1}$

Rotation	Hill	hTST
$k_{A_1 \to A_2 A_3} =$	$3.8\ 10^{10}\ s^{-1}$	$7.6\ 10^{10}\ s^{-1}$
$k_{A_2 A_3 \to A_1} =$	$1.5\ 10^{11}\ s^{-1}$	$2.1\ 10^{12}\ s^{-1}$

hTST rates are larger

Might come from entropy estimation.

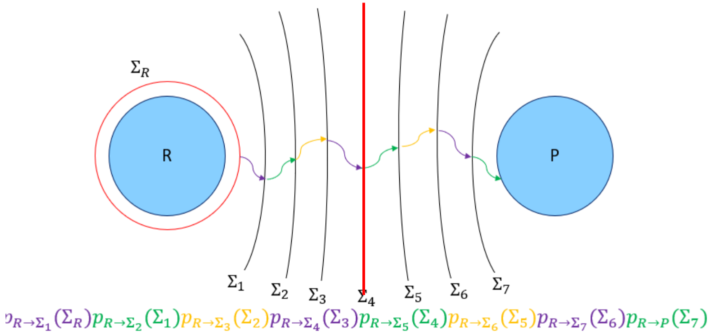
III. RESULTS OF THE AMS + ML METHOD.

Identify TS structures

TS in the sense of committor function $p_{R\to P}$ (probability of reaching P before R)[1]

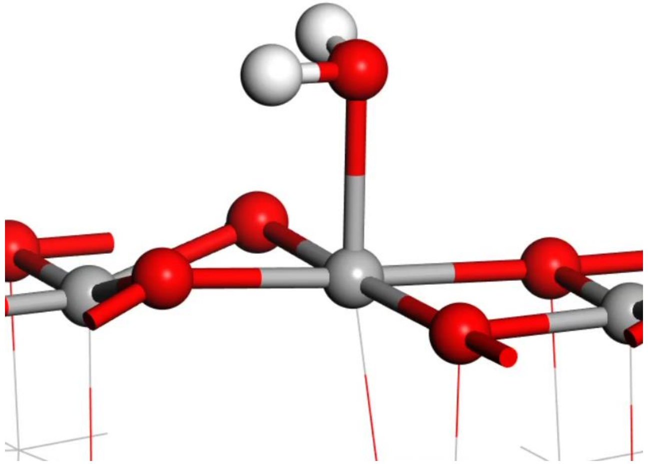
Find the level of the RC z_{kill}^{n} such that $p_{R\to P} = 0.5$

$$\prod_{i=n}^{i_{max}} \tilde{p}_{R\to\Sigma_{z_{kill}^{i+1}}}\left(\Sigma_{z_{kill}^{i}}\right) = 0.5$$

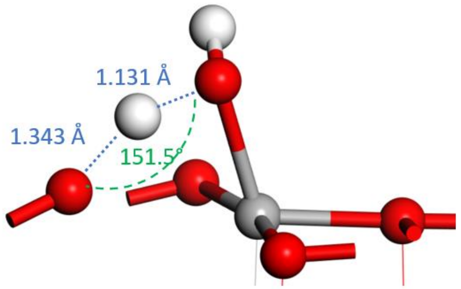


Σ_R
R
P
Σ_1
Σ_2
Σ_3
Σ_4
Σ_5
Σ_6
Σ_7
$p_{R\to\Sigma_1}(\Sigma_R)p_{R\to\Sigma_2}(\Sigma_1)p_{R\to\Sigma_3}(\Sigma_2)p_{R\to\Sigma_4}(\Sigma_3)p_{R\to\Sigma_5}(\Sigma_4)p_{R\to\Sigma_6}(\Sigma_5)p_{R\to\Sigma_7}(\Sigma_6)p_{R\to P}(\Sigma_7)$

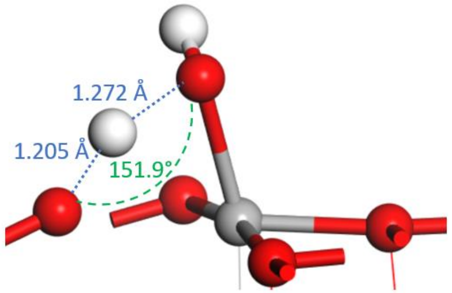
Along each trajectory, take the structure right after the level $\Sigma_{z_{kill}^{n}}$ is crossed, then find the average structure



Example for the $A_1 \to D_1D_3$ reaction



Saddle point



AMS estimated $p_{A_1\to D_1D_3} = 0.5$

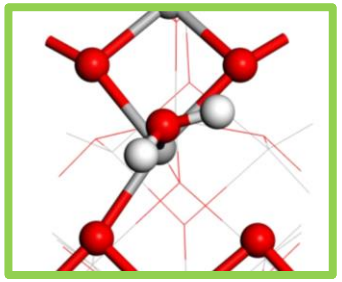
[1] Vanden-Eijnden, E. Transition Path Theory (2006) in Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1 Springer Berlin Heidelberg: Berlin, Heidelberg p. 453-493

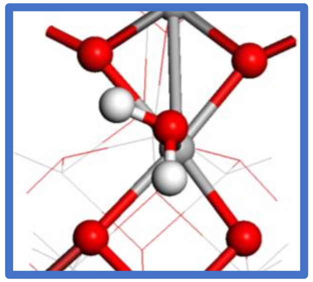
III. RESULTS OF THE AMS + ML METHOD.

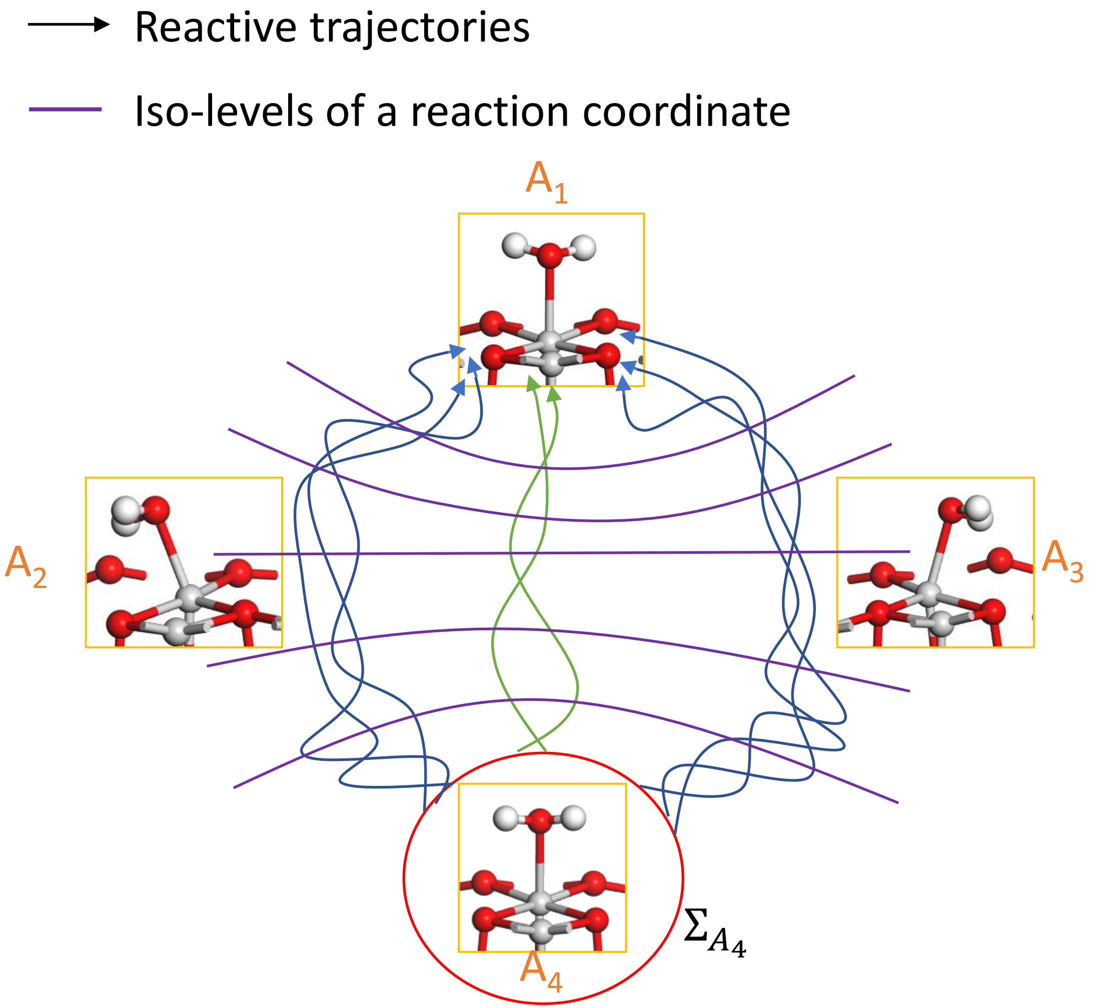
Use K-means clustering method to identify groups of trajectories.

Based on SOAP descriptor + PCA to describe 5 structures per trajectory.

5 Structures = First time trajectory cross RC iso-levels

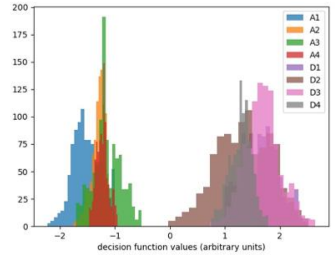






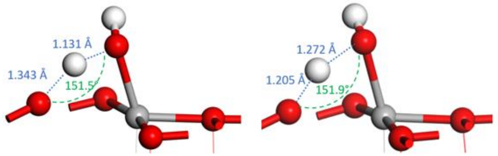
CONCLUSION

- Collective Variables :
  - ✓SVM Allows to define RCs discriminating the metastable states



- AMS :
  - ✓SOAP – SVM RCs allow to sample transitions

- Analysis of reactive trajectories:
  - ✓Some key structures for the transition can be identified
  - ✓Clustering allows to differentiate types of paths



PERSPECTIVES

- Application:
  - ✓ Apply this method to a more challenging reaction on alumina (such as Alcohol dehydration)

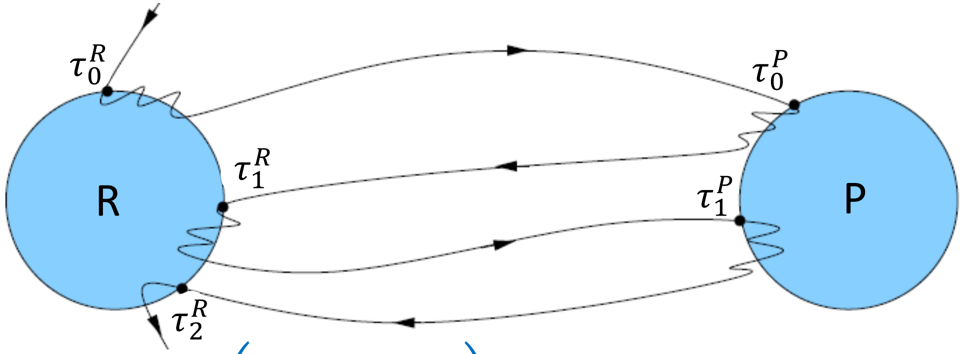
- Theoretical aspects:
  - ✓ Auto-encoders models can be used to define RCs.

[1] Jinnouchi, R., Miwa, K., Karsai, F., Kresse, G., & Asahi, R. (2020). *The Journal of Physical Chemistry Letters*, *11*(17), 6946-6955.

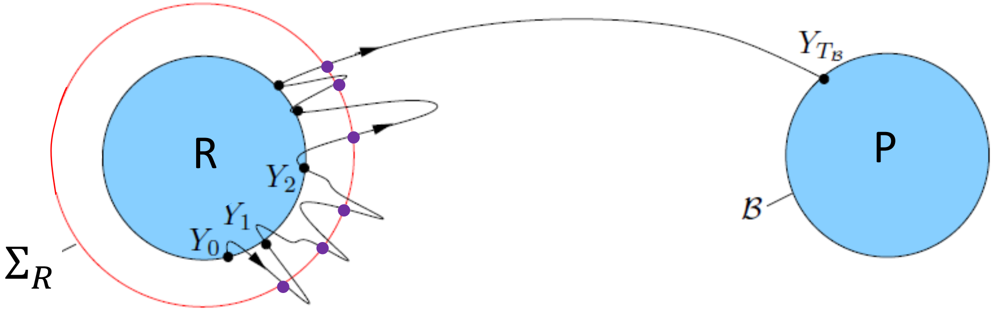
Thank you for you attention

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

Transition time: $\frac{1}{k_{RP}} = t_{RP} = \; mean(\tau_i^R - \tau_i^P)$



We model the reaction time as: $\frac{1}{k_{RP}} = t_{RP} = \left(\frac{1}{p_{\Sigma_R - P}} - 1\right)(t_{R-\Sigma} + t_{\Sigma-R}) + t_{R\ -\Sigma}^{\dagger} + t_{\Sigma-P} \approx \frac{t_{R\to\Sigma\to R}}{p_{\Sigma_R \to P}} = \frac{1}{p_{\Sigma_R\to P}\,\phi_R}$

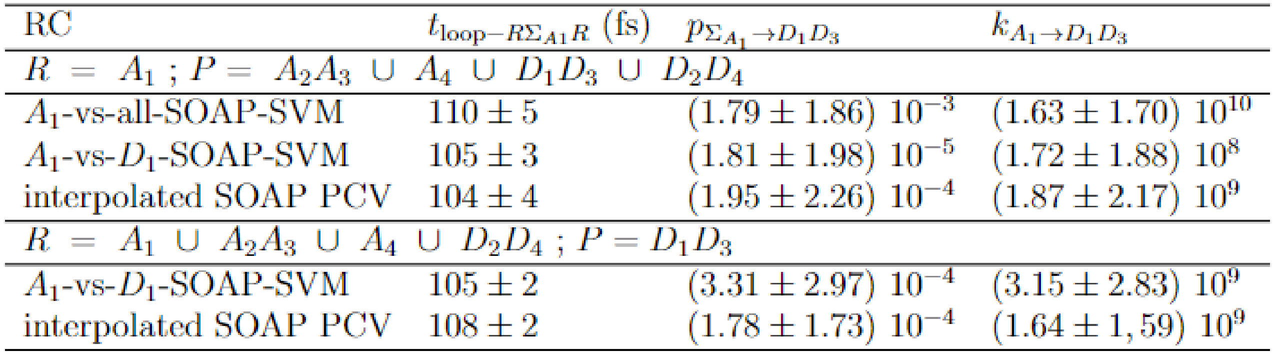


$p_{\Sigma_R - P}$: probability of reaching P before R when starting from Σ_R.

[1] Baudel, M., Guyader, A., & Lelièvre, T. (2020). On the Hill relation and the mean reaction time for metastable processes. *arXiv preprint, arXiv:2008.09790*.

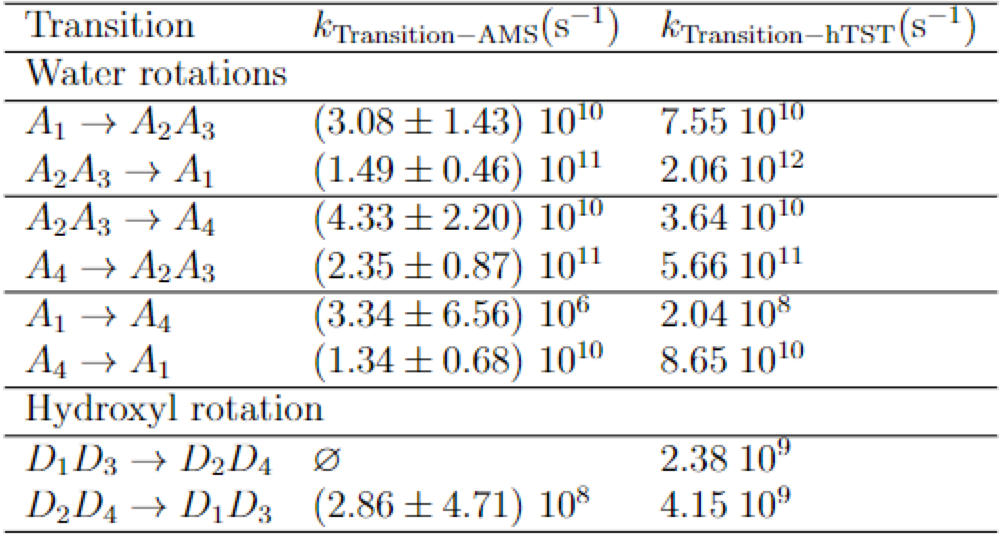
III. RESULTS OF THE AMS + ML METHOD.

Path Collective Variable (PCV) with proper definition of states allows precise estimation

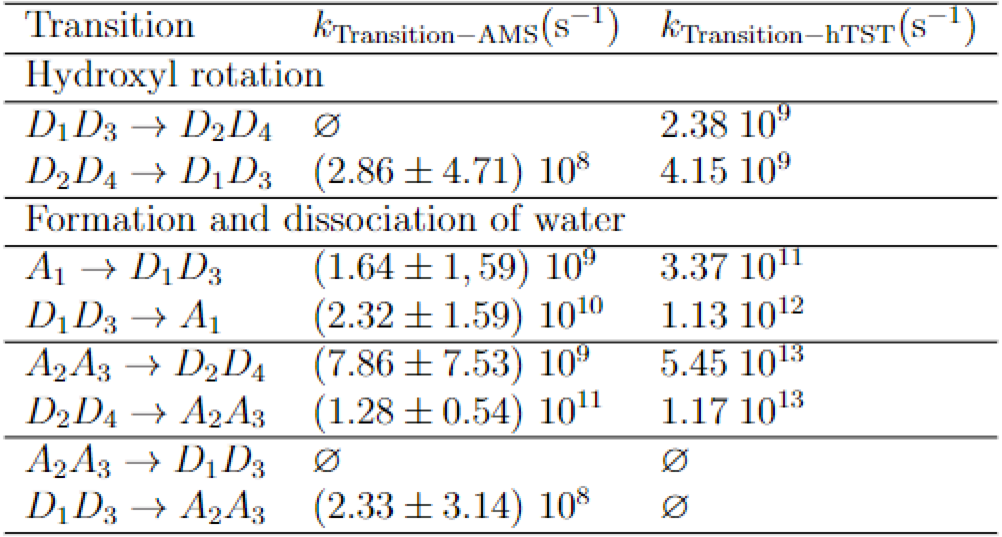


RC	$t_{\text{loop}-R\Sigma_{A1}R}$ (fs)	$p_{\Sigma_{A_1}\to D_1D_3}$	$k_{A_1\to D_1D_3}$
$R = A_1$; $P = A_2A_3 \cup A_4 \cup D_1D_3 \cup D_2D_4$			
A_1-vs-all-SOAP-SVM	110 ± 5	$(1.79 \pm 1.86)\ 10^{-3}$	$(1.63 \pm 1.70)\ 10^{10}$
A_1-vs-D_1-SOAP-SVM	105 ± 3	$(1.81 \pm 1.98)\ 10^{-5}$	$(1.72 \pm 1.88)\ 10^{8}$
interpolated SOAP PCV	104 ± 4	$(1.95 \pm 2.26)\ 10^{-4}$	$(1.87 \pm 2.17)\ 10^{9}$
$R = A_1 \cup A_2A_3 \cup A_4 \cup D_2D_4$; $P = D_1D_3$			
A_1-vs-D_1-SOAP-SVM	105 ± 2	$(3.31 \pm 2.97)\ 10^{-4}$	$(3.15 \pm 2.83)\ 10^{9}$
interpolated SOAP PCV	108 ± 2	$(1.78 \pm 1.73)\ 10^{-4}$	$(1.64 \pm 1,59)\ 10^{9}$

Summary of AMS results vs hTST results



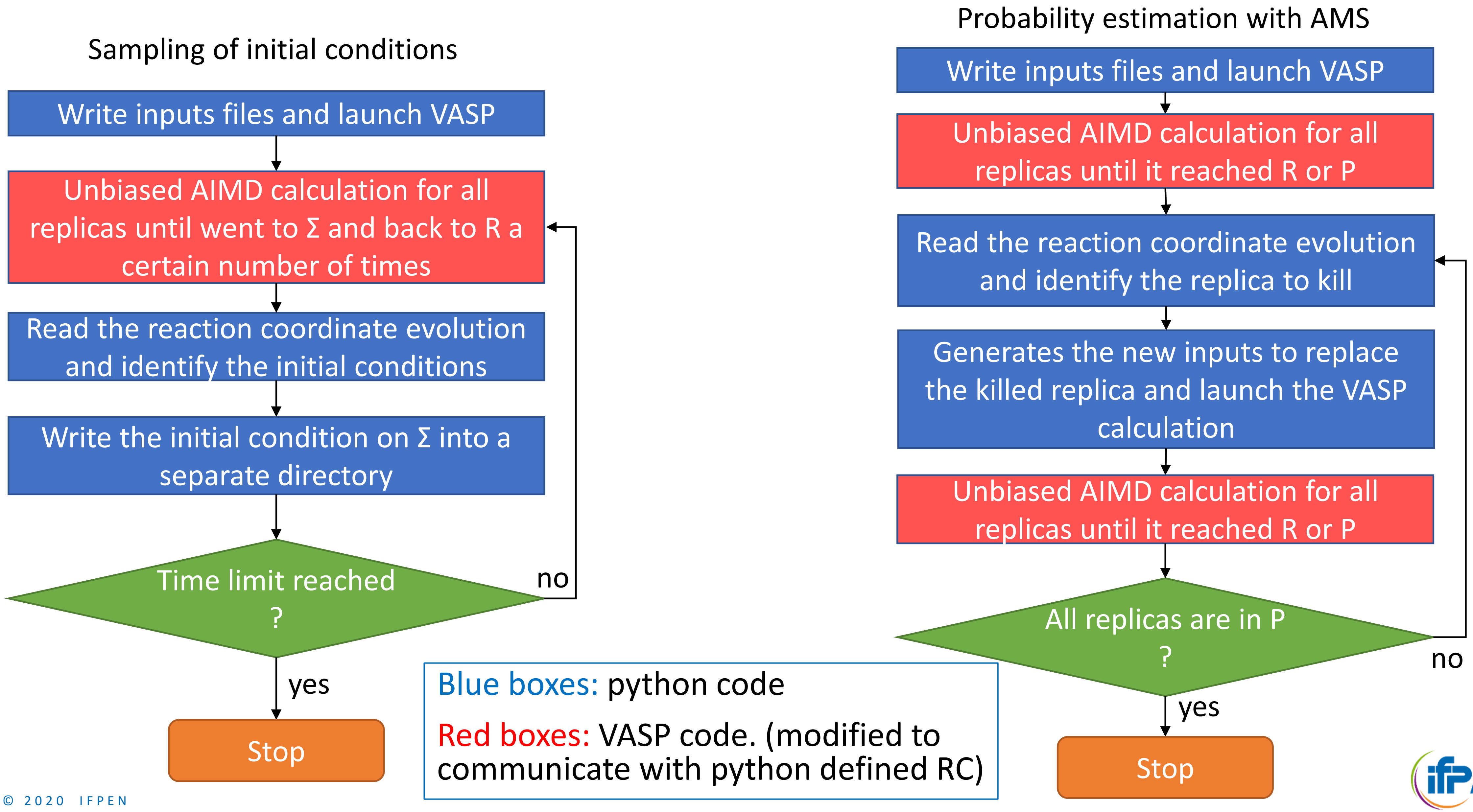
Transition	$k_{\text{Transition}-\text{AMS}}(\text{s}^{-1})$	$k_{\text{Transition}-\text{hTST}}(\text{s}^{-1})$
Water rotations		
$A_1 \to A_2A_3$	$(3.08 \pm 1.43)\ 10^{10}$	$7.55\ 10^{10}$
$A_2A_3 \to A_1$	$(1.49 \pm 0.46)\ 10^{11}$	$2.06\ 10^{12}$
$A_2A_3 \to A_4$	$(4.33 \pm 2.20)\ 10^{10}$	$3.64\ 10^{10}$
$A_4 \to A_2A_3$	$(2.35 \pm 0.87)\ 10^{11}$	$5.66\ 10^{11}$
$A_1 \to A_4$	$(3.34 \pm 6.56)\ 10^{6}$	$2.04\ 10^{8}$
$A_4 \to A_1$	$(1.34 \pm 0.68)\ 10^{10}$	$8.65\ 10^{10}$
Hydroxyl rotation		
$D_1D_3 \to D_2D_4$	$\varnothing$	$2.38\ 10^{9}$
$D_2D_4 \to D_1D_3$	$(2.86 \pm 4.71)\ 10^{8}$	$4.15\ 10^{9}$



Transition	$k_{\text{Transition}-\text{AMS}}(\text{s}^{-1})$	$k_{\text{Transition}-\text{hTST}}(\text{s}^{-1})$
Hydroxyl rotation		
$D_1D_3 \to D_2D_4$	$\varnothing$	$2.38\ 10^{9}$
$D_2D_4 \to D_1D_3$	$(2.86 \pm 4.71)\ 10^{8}$	$4.15\ 10^{9}$
Formation and dissociation of water		
$A_1 \to D_1D_3$	$(1.64 \pm 1,59)\ 10^{9}$	$3.37\ 10^{11}$
$D_1D_3 \to A_1$	$(2.32 \pm 1.59)\ 10^{10}$	$1.13\ 10^{12}$
$A_2A_3 \to D_2D_4$	$(7.86 \pm 7.53)\ 10^{9}$	$5.45\ 10^{13}$
$D_2D_4 \to A_2A_3$	$(1.28 \pm 0.54)\ 10^{11}$	$1.17\ 10^{13}$
$A_2A_3 \to D_1D_3$	$\varnothing$	$\varnothing$
$D_1D_3 \to A_2A_3$	$(2.33 \pm 3.14)\ 10^{8}$	$\varnothing$

III. RESULTS OF THE AMS + ML METHOD.

AMS IMPLEMENTATION WITH VASP (PLANE WAVE DFT)



Sampling of initial conditions

Write inputs files and launch VASP
Unbiased AIMD calculation for all replicas until went to Σ and back to R a certain number of times
Read the reaction coordinate evolution and identify the initial conditions
Write the initial condition on Σ into a separate directory
Time limit reached ?
no
yes
Stop

Probability estimation with AMS

Write inputs files and launch VASP
Unbiased AIMD calculation for all replicas until it reached R or P
Read the reaction coordinate evolution and identify the replica to kill
Generates the new inputs to replace the killed replica and launch the VASP calculation
Unbiased AIMD calculation for all replicas until it reached R or P
All replicas are in P ?
no
yes
Stop

Blue boxes: python code

Red boxes: VASP code. (modified to communicate with python defined RC)

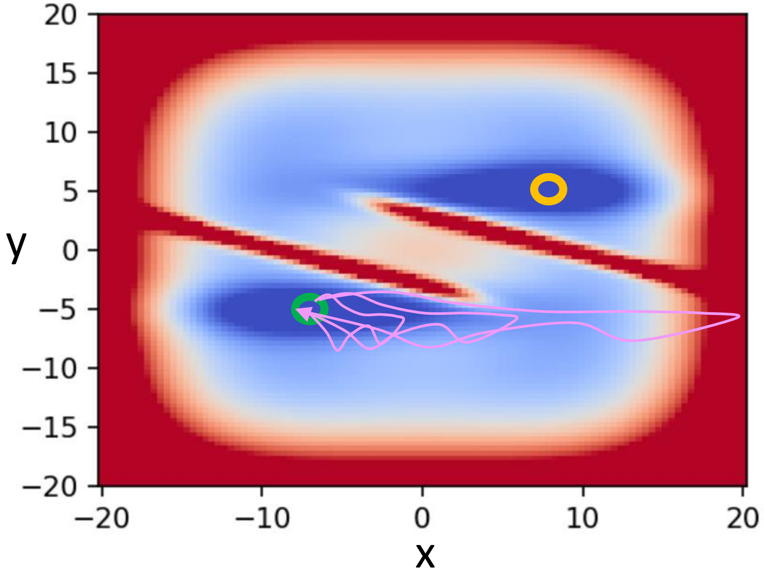
I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

Langevin dynamics with $\mathrm{m} = 1, \beta = 4, \Delta t = 0.1$

AMS extinction case, illustration on the Z-potential[1]:

AMS with $N_{rep} = 100, \quad k_{min} = 1$

$\xi(x, y) = x$



○ P State

○ R State

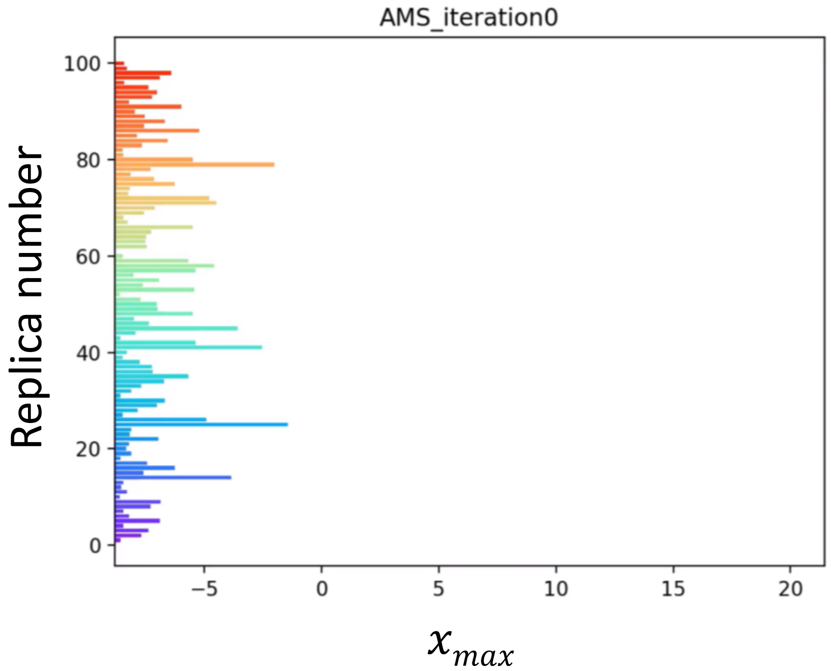
— “initial” “best” trajectory

“final” “best” trajectory

No reactive trajectories !

$\tilde{p} = 0$

Algorithm “extinction”



AMS favors high values of the Reaction coordinate.

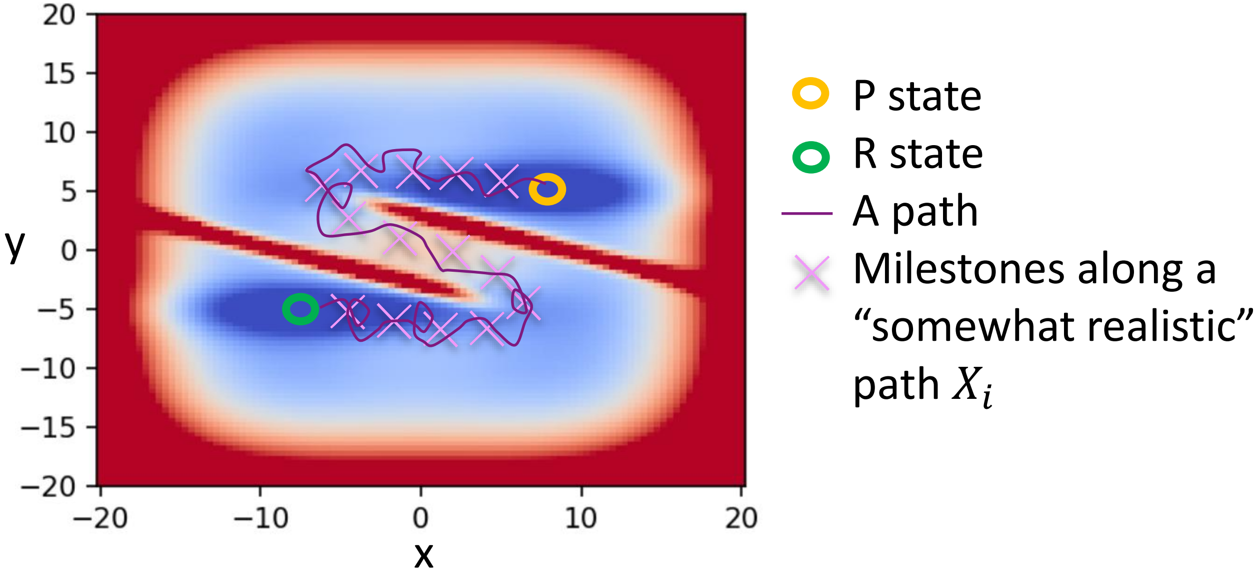
Bad indexation of the reaction path by the RC ξ can lead to frequent “extinction”.

→ Alternative RC needed!

[1] Lechner, W.; Rogal, J.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. (2010) *The Journal of Chemical Physics*, 133, 174110.

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

Alternative approach, Path Collective Variable (PCV):



$$\text{PCV} : \xi(q) = \frac{1}{14}\frac{\sum_{i=0}^{14} i e^{-\lambda|q-X_i|^2}}{\sum_{i=0}^{14} e^{-\lambda|q-X_i|^2}}$$

Langevin dynamics with $\mathrm{m} = 1, \beta = 4, \Delta t = 0.1$

AMS with $N_{rep} = 100,\ \ k_{min} = 1$

Initial condition: $\begin{cases} q = (2.01, -5.75), \\ p = (0.61, -0.15). \end{cases}$

	DNS	AMS $\xi(x,y) = x$	AMS $\xi(x,y) = y$	AMS PCV
$FM(\tilde{p})$	3.14×10^1	1.57×10^0	6.41×10^1	6.45×10^2

Figure of merit: $FM(\tilde{p}) = \dfrac{1}{c\,Var(\tilde{p})}$

c: computational cost per evaluation

A "**somewhat realistic path**" is needed (reverse reaction path, NEB path …)

Decent milestones definition method has to be considered (Intuition based milestones, unsupervised clustering)

* Results presented via a poster during the workshop "Méthodes machine-learning pour la modélisation des matériaux" organized by the GDR MODMAT from the 22th to the 24th Sept 2021