COMPUTING REACTION TIMES AND PATHS WITH MACHINE LEARNING AND RARE EVENT SAMPLING METHODS

THOMAS PIGEON

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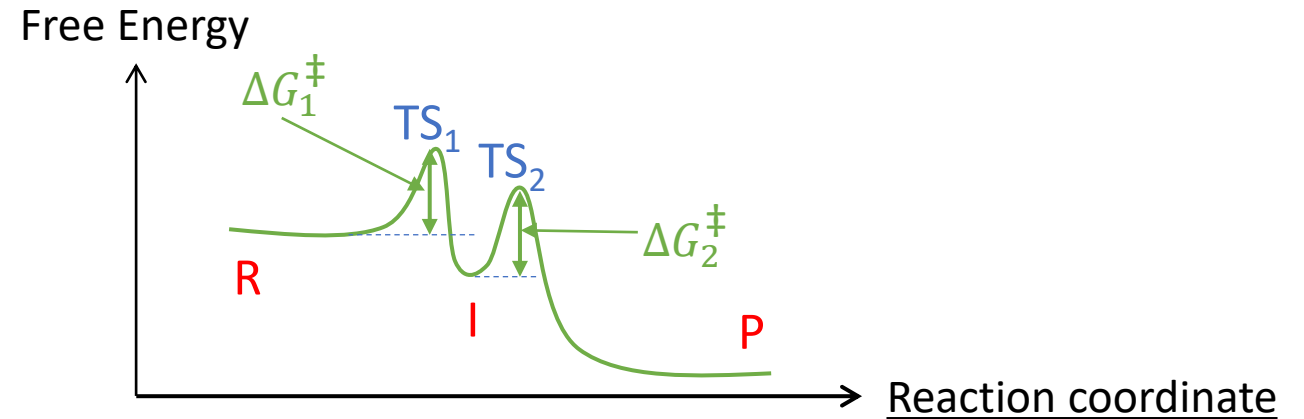Gabriel Stoltz[3], Tony Lelièvre[3]

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

Targets : compute reaction rates
identify reaction mechanism

Free Energy

$\Delta G_1^{\ddagger}$

TS$_1$
TS$_2$

$\Delta G_2^{\ddagger}$

R

I

P

Reaction coordinate

Different methods exist:

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

  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*

ifp Energies nouvelles

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
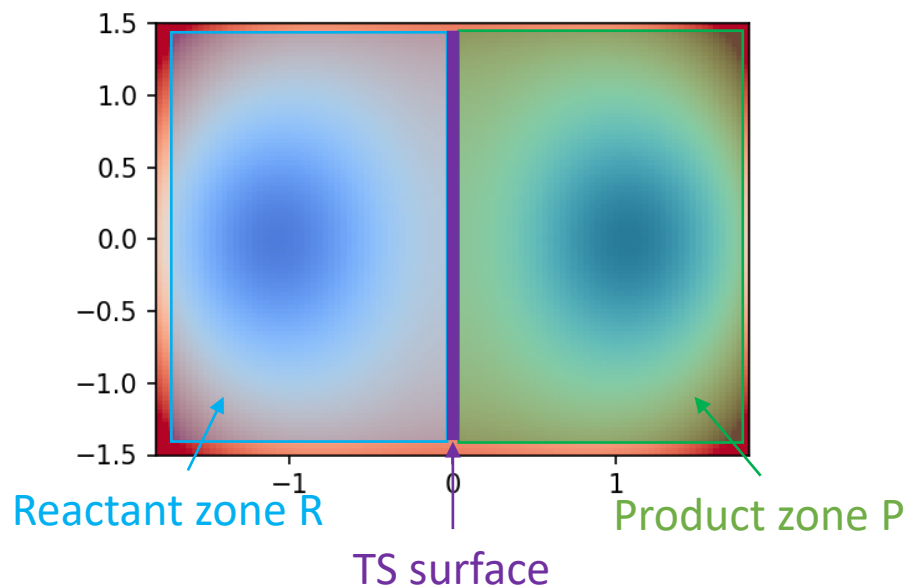[2] Laio, A., & Parrinello, M. (2002) *Proceedings of the National Academy of Sciences, 99*(20), 12562-12566.
[3] Carter, E. A., Ciccotti, G., Hynes, J. T., & Kapral, R. (1989). *Chemical Physics Letters, 156*(5), 472-477.
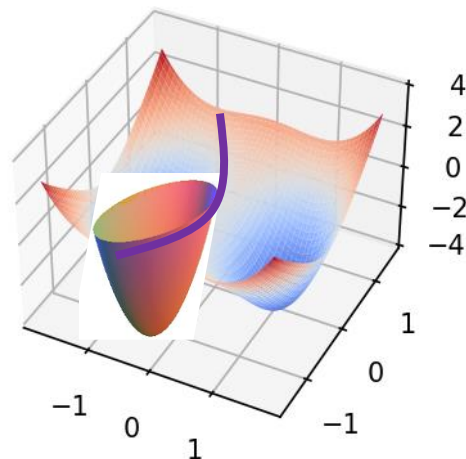[4] Cérou, F., & Guyader, A. (2007) *Stochastic Analysis and Applications, 25*(2), 417-443.

Transition State Theory

Reactant zone R

TS surface

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

2-dimensional potential

Hill relation

Reactant zone R

Product zone P

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*

What is a Multilevel Splitting estimator:
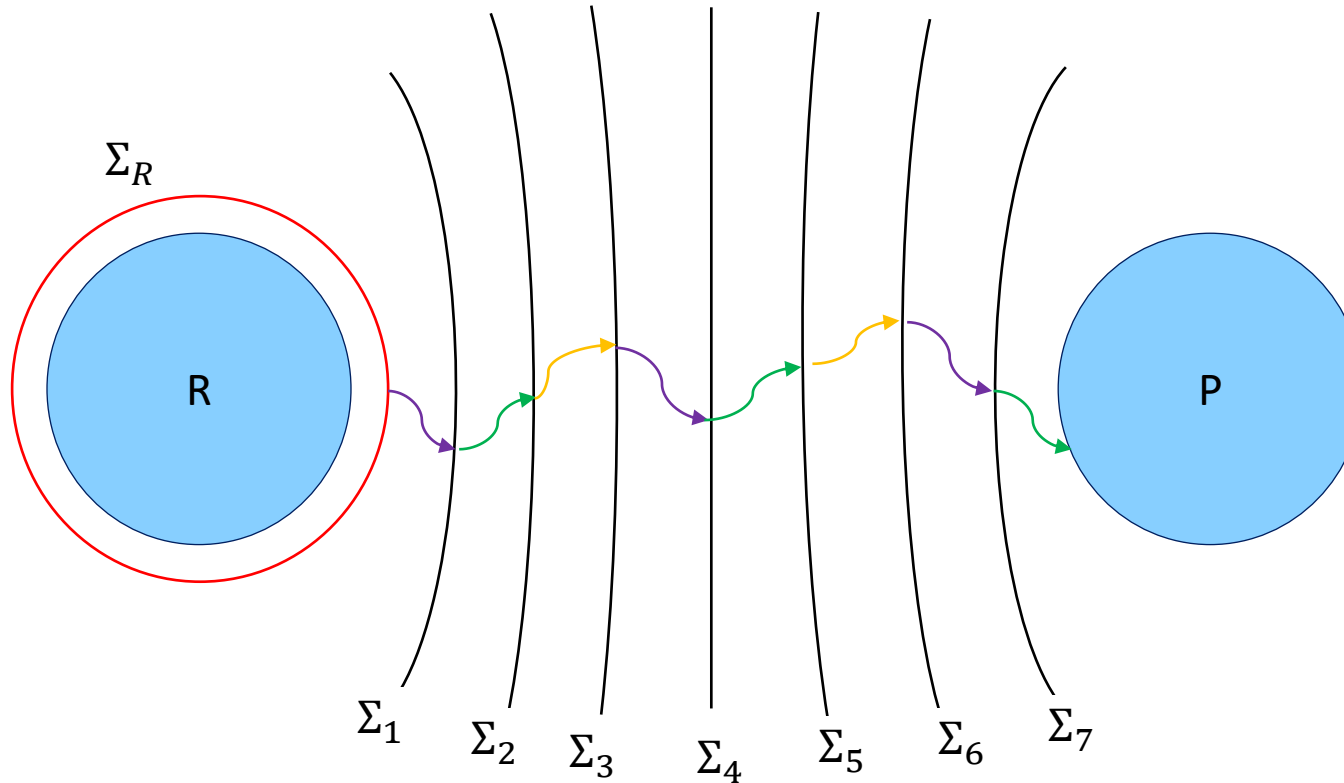
$$p_{R\to\Sigma_1}(\Sigma_R)p_{R\to\Sigma_2}(\Sigma_1)p_{R\to\Sigma_3}(\Sigma_2)p_{R\to\Sigma_4}(\Sigma_3)p_{R\to\Sigma_5}(\Sigma_4)p_{R\to\Sigma_6}(\Sigma_5)p_{R\to\Sigma_7}(\Sigma_6)p_{R\to P}(\Sigma_7)$$
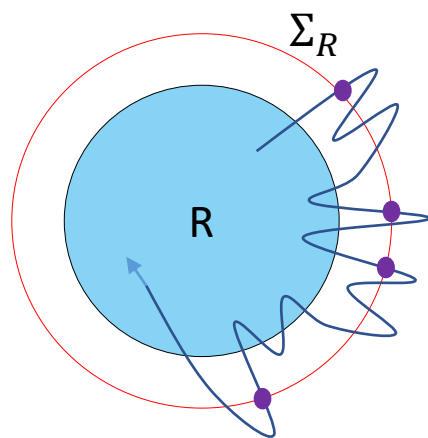$$= p_{R\to P}(\Sigma_R)$$

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

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 $t_{R-\Sigma-R} = \dfrac{1}{\phi_R}$

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.

2. Initialization

1. Initial conditions and flux

MD for Initial conditions.

Σ_R

R

$z_{max}^{1,0}$ $z_{max}^{3,0}$

$z_{max}^{2,0}$

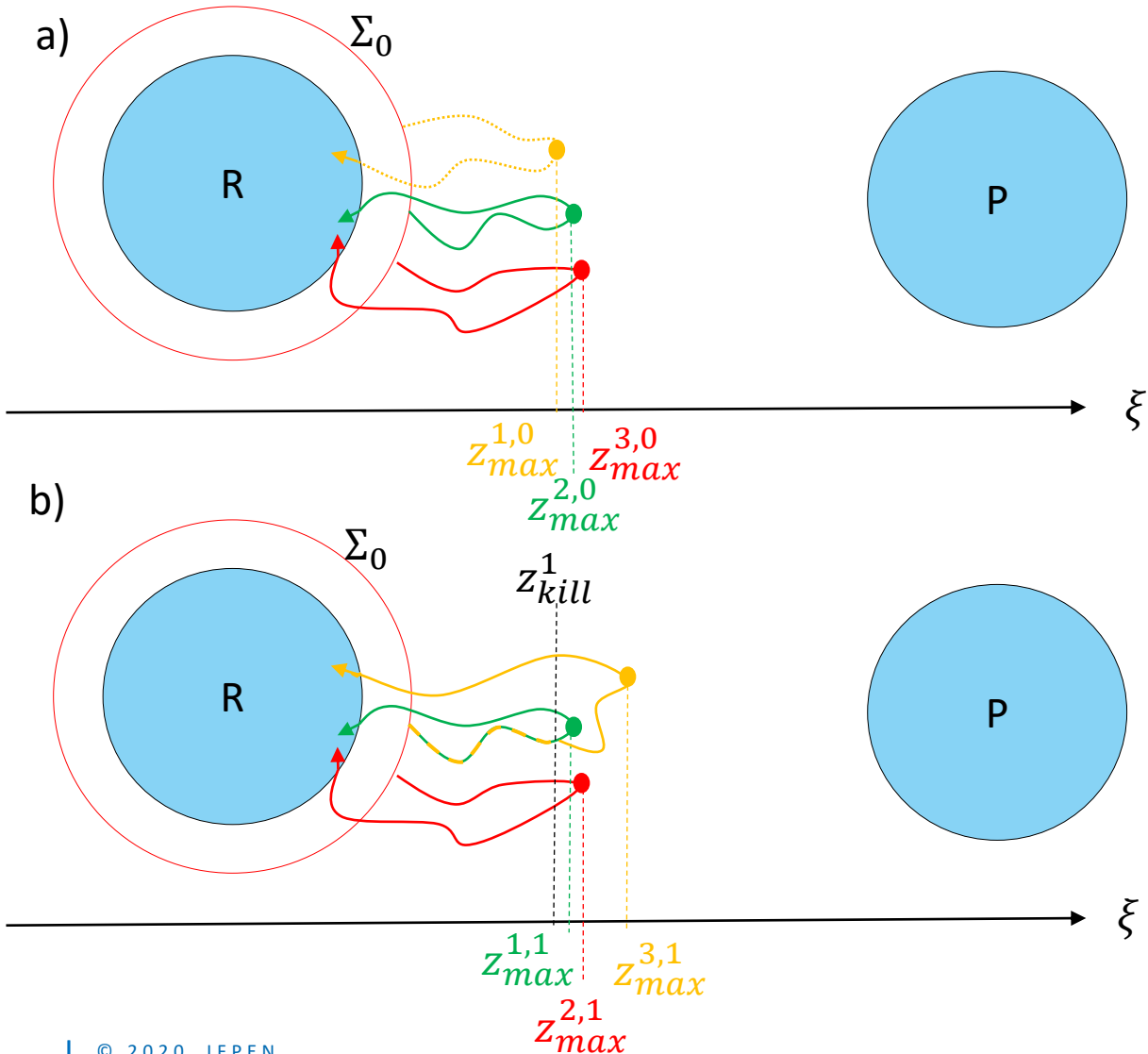ξ

Reaction Coordinate (RC)

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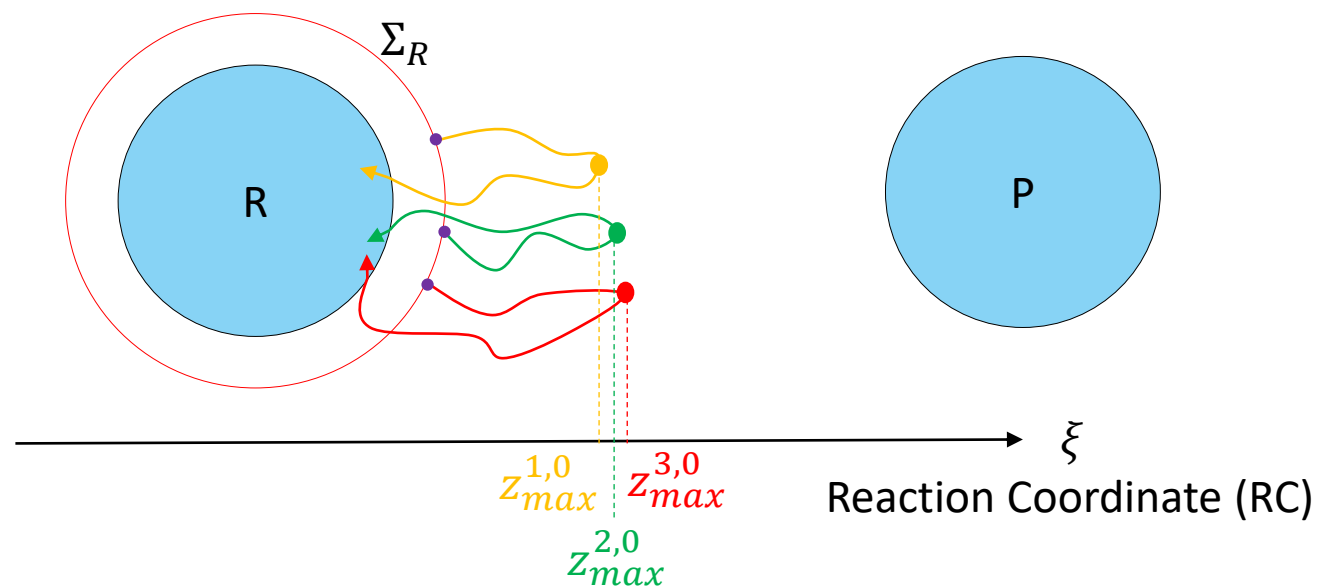
[1] F. Cérou, A. Guyader, *Stochastic Analysis and Applications* **25**, 417-443 (2007).
[2] L. J. S. Lopes, T. Lelièvre, *Journal of computational chemistry* **40**, 1198-1208 (2019).

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

3. AMS interations

a)

Σ_0

R

P

ξ

$z_{max}^{1,0}$ $z_{max}^{3,0}$

$z_{max}^{2,0}$

b)

Σ_0

z_{kill}^1

R

P

ξ

$z_{max}^{1,1}$ $z_{max}^{3,1}$

$z_{max}^{2,1}$

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}.

$$\tilde{p} = \prod_{i=0}^{i_{max}} \tilde{p}_{\Sigma_{z_{kill}^i} \to \Sigma_{z_{kill}^{i+1}}} = \left(1 - \frac{1}{N}\right)^{i_{max}}$$
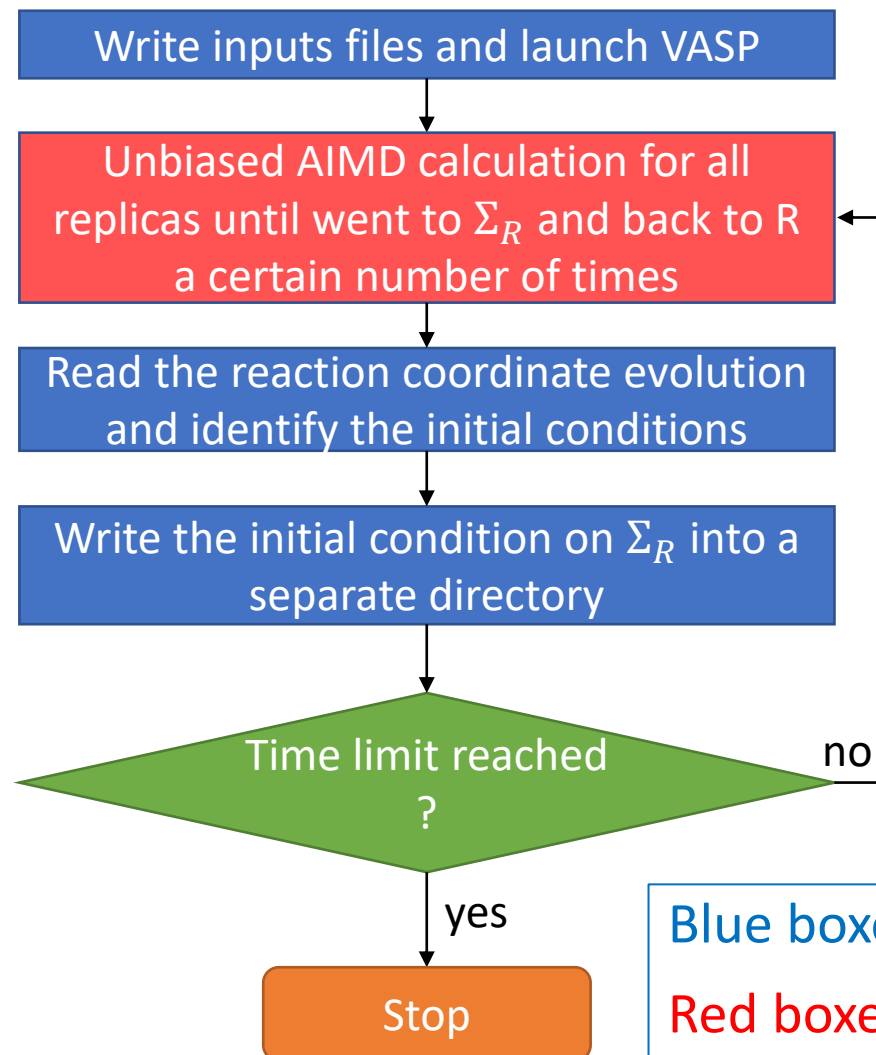
Unbiased estimator: Variance depends on RC:

$$\mathbb{E}[\tilde{p}] = p_{R-P}(\Sigma_R) \qquad \text{Var}[\tilde{p}] = f(\xi)$$

I. ADAPTIVE MULTI-LEVEL SPLITTING METHOD FOR REACTION RATES

Sampling of initial conditions

- Write inputs files and launch VASP
- Unbiased AIMD calculation for all replicas until went to Σ_R 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_R 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[1,2]

[1]Kresse, G.; Hafner, J. *Phys. Rev. B* **1993**, 47, 558–561.
[2]Kresse, G.; Joubert, D *Phys. Rev. B* **1999**, 59, 1758–1775.
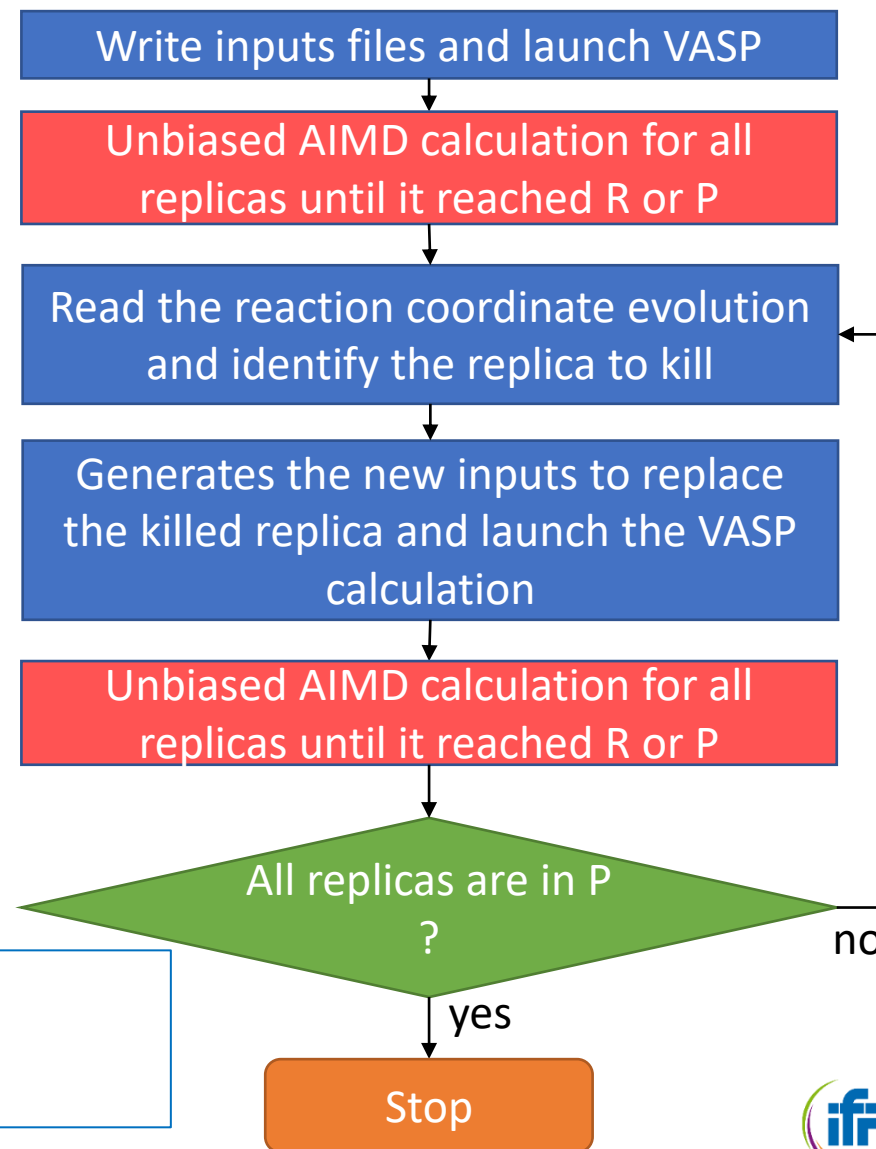
Multistate problem

With
$$R = A_1$$
$$\Sigma_R = \Sigma_{A_1}$$
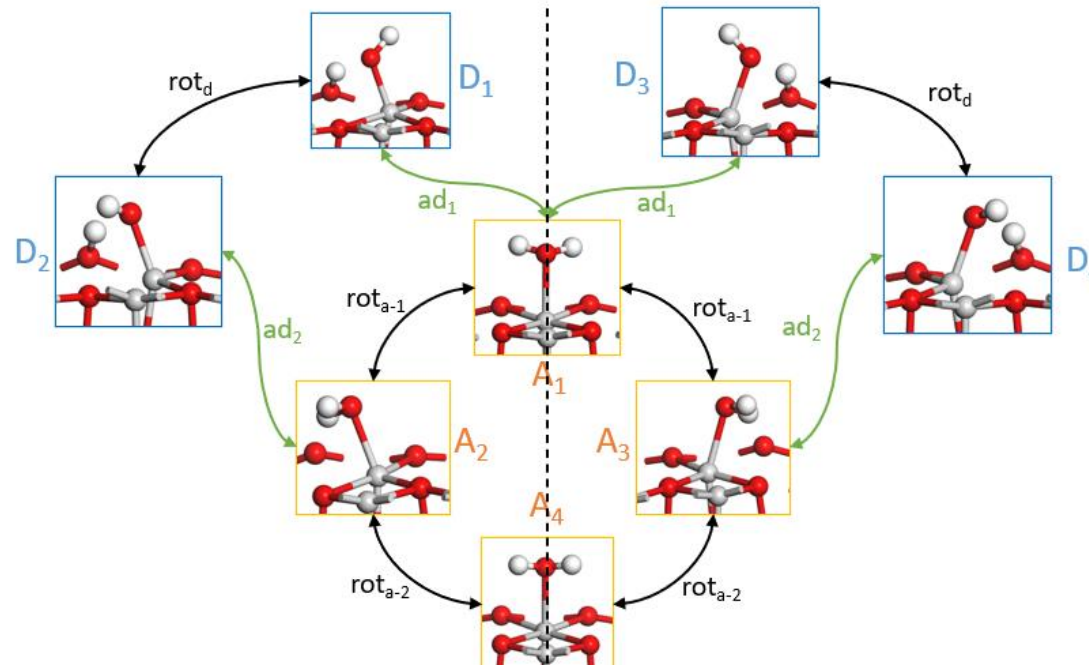$$P = A_2 A_3 \cup A_4 \cup D_1 D_3 \cup D_2 D_4$$

AMS can sample :

$$A_1 \rightarrow A_2 A_3$$
$$A_1 \rightarrow A_4$$
$$A_1 \rightarrow D_1 D_3$$
$$A_1 \rightarrow D_2 D_4$$

→ Answers how A_1 can decompose ?

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

Metastable states of H$_2$O on the (100) surface of γ-alumina

With
$$R = A_1 \cup A_2 A_3 \cup A_4 \cup D_2 D_4$$
$$\Sigma_R = \Sigma_{A_1}$$
$$P = D_1 D_3$$

AMS can sample :

$$A_1 \rightarrow D_1 D_3$$

→ Focus specifically on one event

Quality of the sampling depends on ξ

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

3. SOAP[1] atom centered descriptors to numerically encode the structure for training the MLCV.

H
O
Al

Select central atom
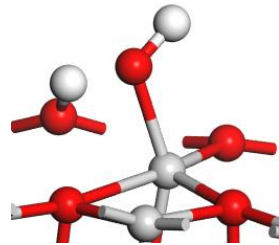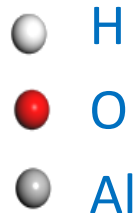
Cartesian coordinates
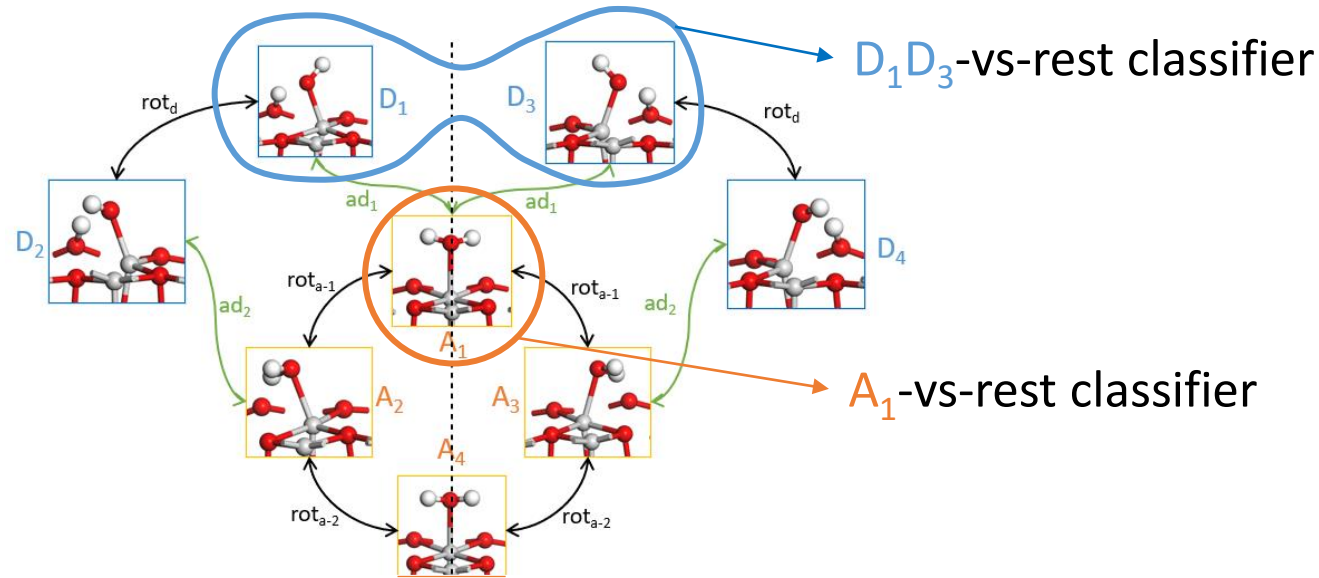(vector of 3N lines)
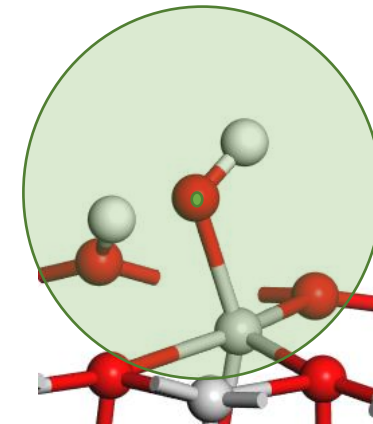
Atom centered descriptor of the structure
(vector of ~ 10^3-10^4 lines)

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

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.

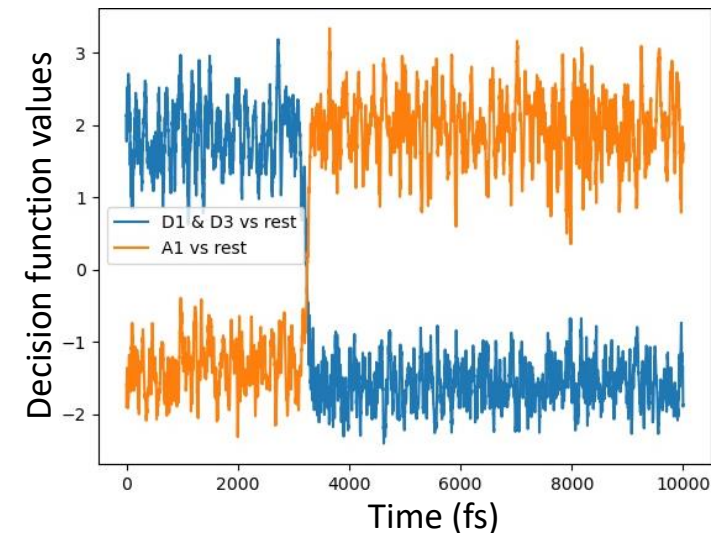D_1D_3-vs-rest classifier

A_1-vs-rest classifier

Histogram of SVM classifier decision function values
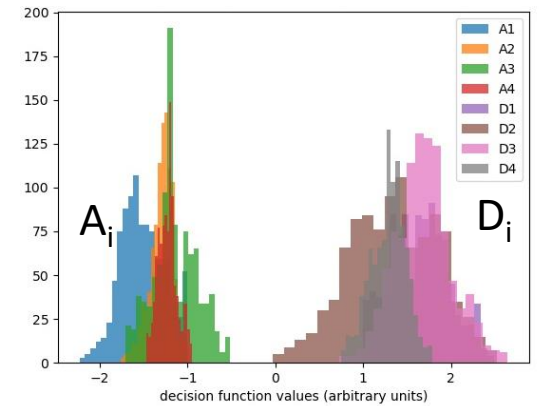
Training set : A1 = 0, D1 = 1

Classifier decision function interpretation:
$$f_X(\boldsymbol{q}) \in (-\infty, -1] \iff \boldsymbol{q} \in X$$
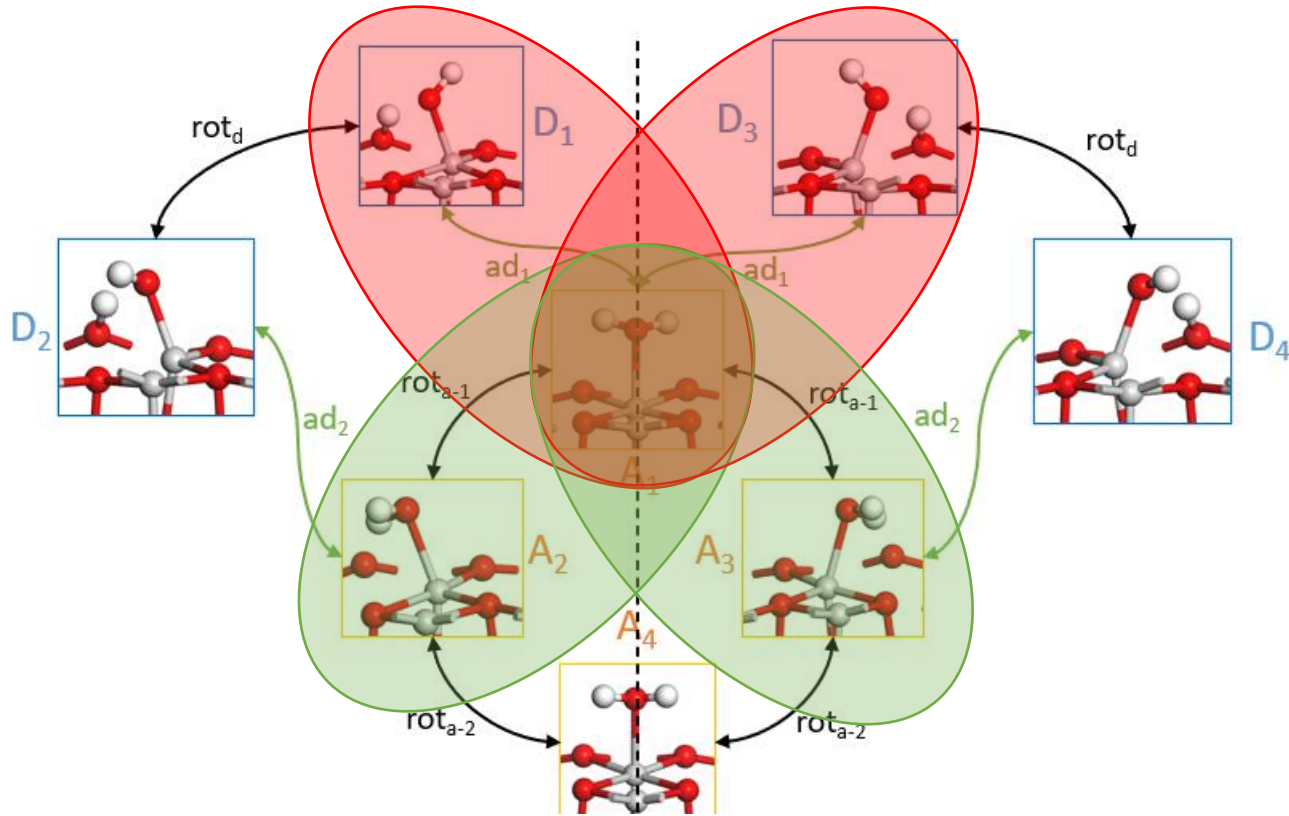
Dissociation   Hill   hTST

$$k_{A_1 \to D_1 D_3} =$$ $1.6 \; 10^9 \; s^{-1}$ $3.4 \; 10^{11} \; s^{-1}$

$$k_{D_1 D_3 \to A_1} =$$ $2.3 \; 10^{10} \; s^{-1}$ $1.1 \; 10^{12} \; s^{-1}$

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.
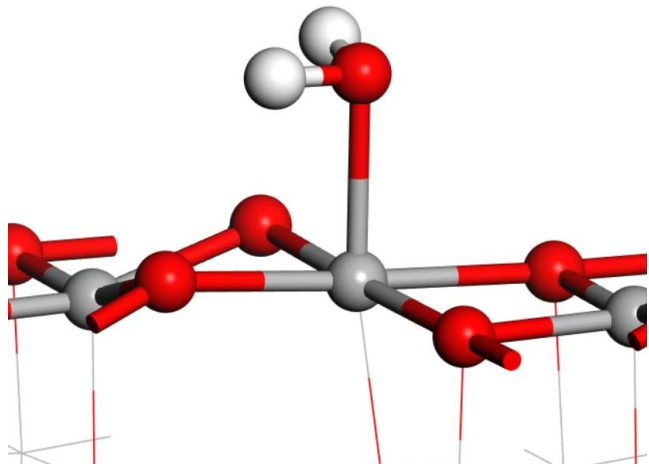
~ 2 10^6 CPU Hours

Published

Identify TS structures

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

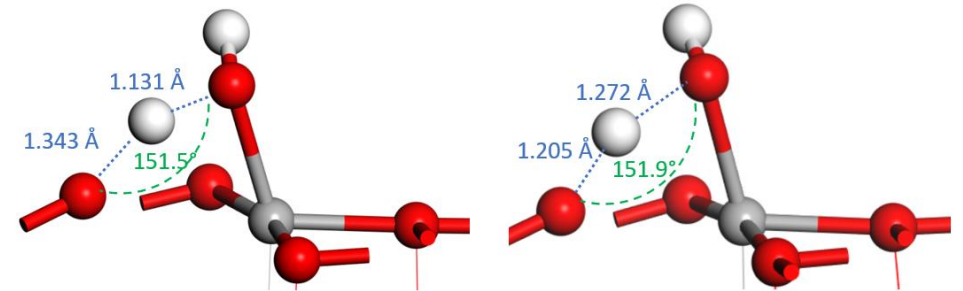Find the level of the RC z_{kill}^n such that $p_{R \to P}\left(\Sigma_{z_{kill}^n}\right) = 0.5$

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

$p_{R \to \Sigma_1}(\Sigma_R) p_{R \to \Sigma_2}(\Sigma_1) p_{R \to \Sigma_3}(\Sigma_2) p_{R \to \Sigma_4}(\Sigma_3) p_{R \to \Sigma_5}(\Sigma_4) p_{R \to \Sigma_6}(\Sigma_5) p_{R \to \Sigma_7}(\Sigma_6) p_{R \to P}(\Sigma_7)$
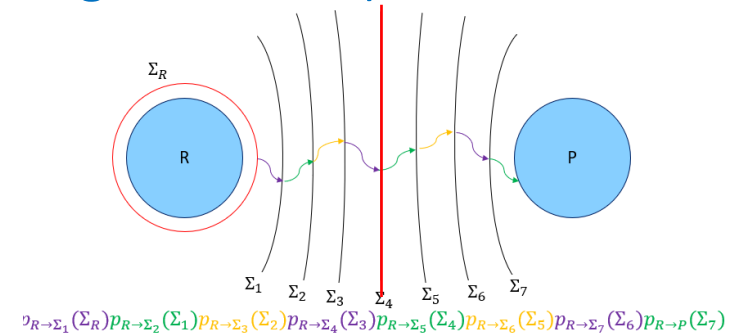
Along each trajectory, take the structure right after the level $\Sigma_{z_{kill}^n}$ is crossed, then find the average structure

Example for the $A_1 \to D_1 D_3$ reaction

Saddle point

AMS estimated
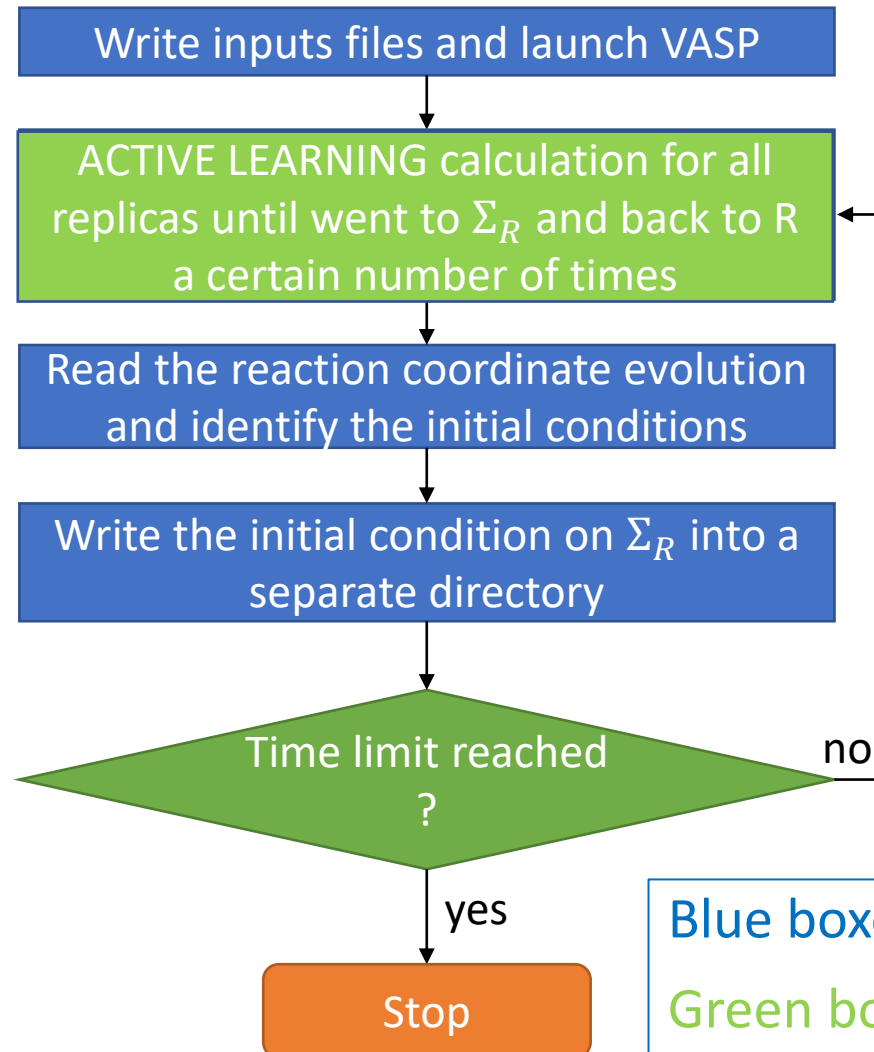$p_{A_1 \to D_1 D_3} = 0.5$

[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. USING AMS WITH ACTIVE LEARNING
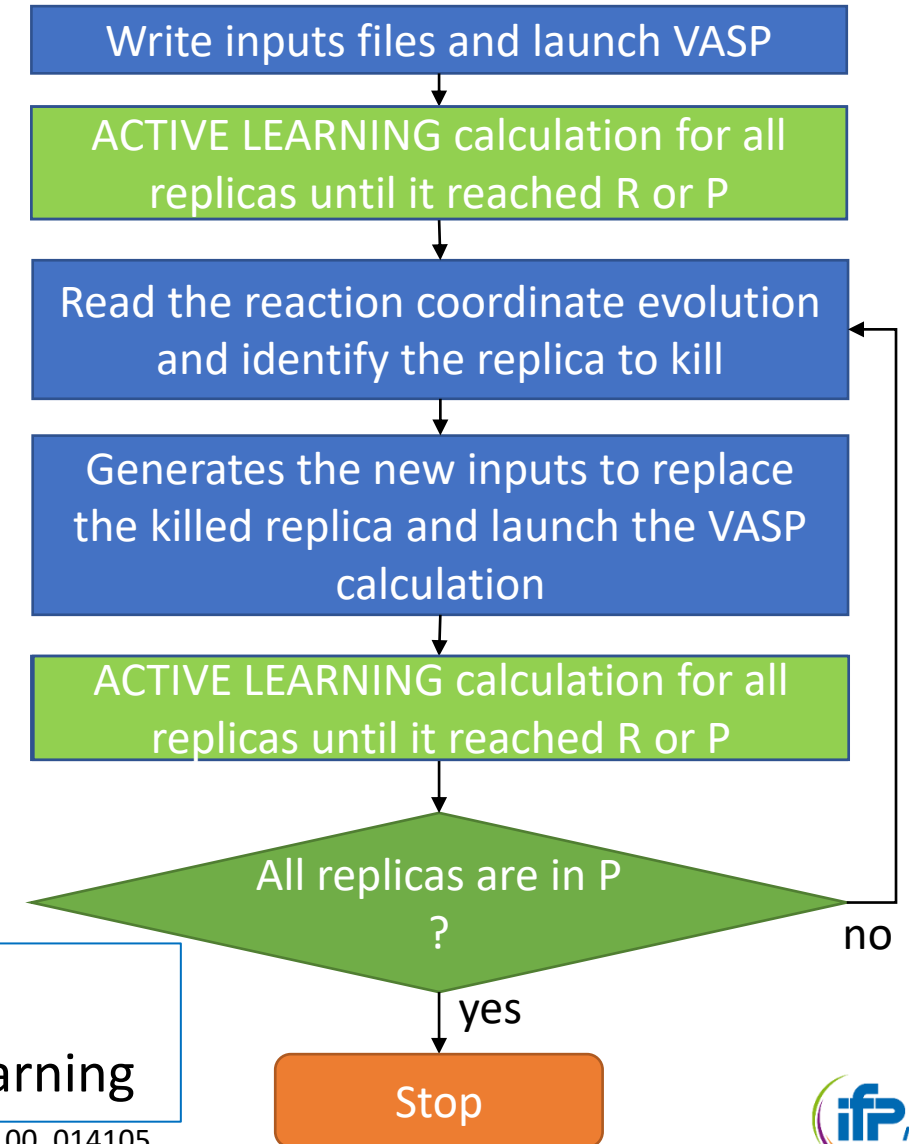## AMS IMPLEMENTATION WITH VASP (PLANE WAVE DFT)

Sampling of initial conditions

Write inputs files and launch VASP

↓

ACTIVE LEARNING calculation for all replicas until went to Σ_R 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_R into a separate directory

↓

Time limit reached ? — no →

yes ↓

Stop

Probability estimation with AMS

Write inputs files and launch VASP

↓

ACTIVE LEARNING 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

↓

ACTIVE LEARNING calculation for all replicas until it reached R or P

↓

All replicas are in P ? — no →

yes ↓

Stop

Blue boxes: python code

Green boxes: VASP active learning

[1] Jinnouchi, R.; Karsai, F.; Kresse, G. *Phys. Rev. B* **2019,** 100, 014105.
[2] Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. *The Journal of Physical Chemistry Letters* **2020**, 11, 6946–6955.

IFP Energies nouvelles
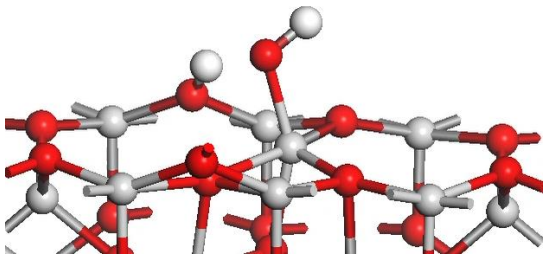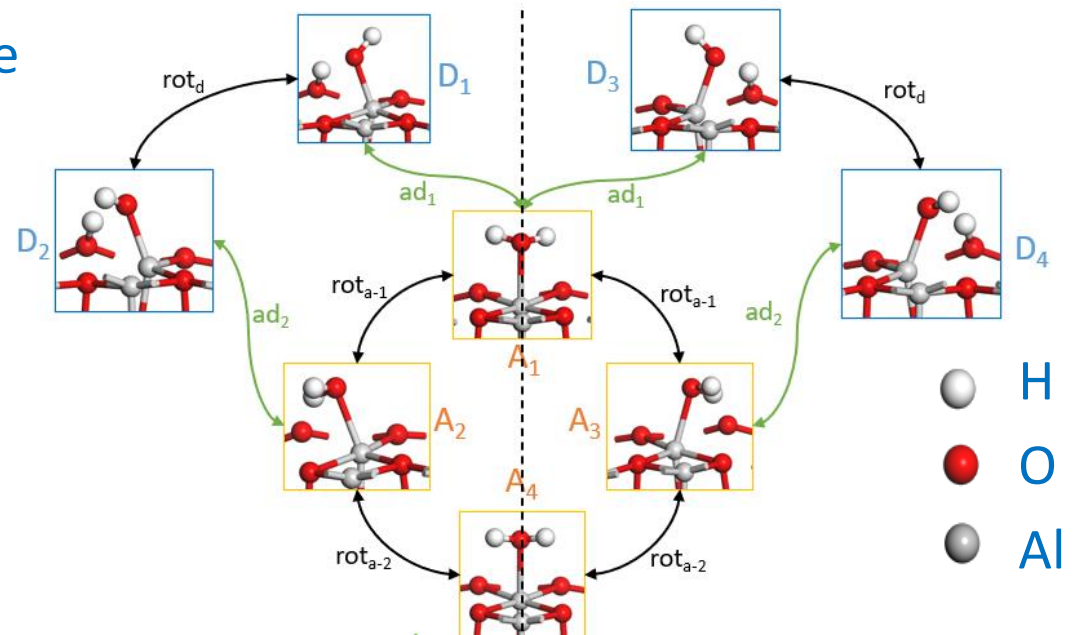
Method:

1. Identify the various metastable states (intermediates)
   → dissociated (D_i) or associated (A_i)

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

D_1 dissociated structure

dt = 1 fs
Total time = 50 ps

3. Concatenate the dataset
   (containing E, Forces, and positions)

Identified structures and **intuitively** plausible transitions

Wall clock time

Dissociation   with N_{rep} = 200 and M_{real} = 10

DFT

~10 days

$k_{A_1 \rightarrow D_1 D_3} =$ $1.64\ 10^9\ s^{-1}$ $\pm\ 1.59\ 10^9\ s^{-1}$

MLFF prediction

~0.5 days

$k_{A_1 \rightarrow D_1 D_3} =$ $2.76\ 10^9\ s^{-1}$ $\pm\ 3.81\ 10^9\ s^{-1}$
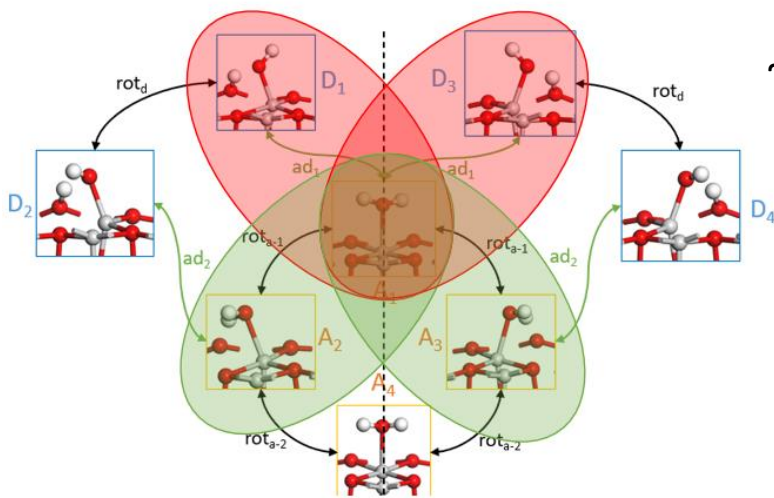
with N_{rep} = 800 and M_{real} = 10

~2 days

$k_{A_1 \rightarrow D_1 D_3} =$ $2.43\ 10^9\ s^{-1}$ $\pm 1.15\ 10^9\ s^{-1}$

CONCLUSION AND PERSPECTIVES

- hTST overestimate the DFT-MD rate estimated using AMS

- MLFF-MD and DFT-MD rates are consistent

- MLFF used in prediction mode drastically reduces de computational cost

- Current implementation of AMS with VASP limits the application of active learning
    - → Restart does have an important cost for the active learning.

- Using FLARE[1] active learning with VASP as calculator of ab-initio seems a good opportunity as AMS could be implemented more easily

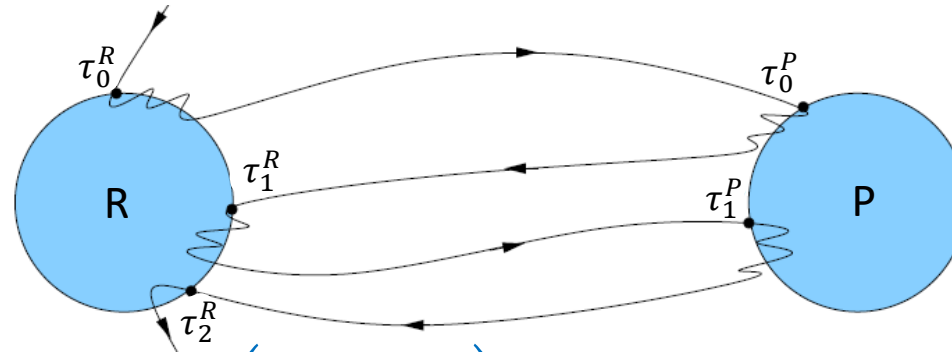- Active learning of RC ξ can be included in the workflow

[1] Vandermause, J., Xie, Y., Lim, J.S., Owen, C.J. and Kozinsky, B., 2021. *Active learning of reactive Bayesian force fields: Application to heterogeneous hydrogen-platinum catalysis dynamics.* arXiv preprint arXiv:2106.01949.
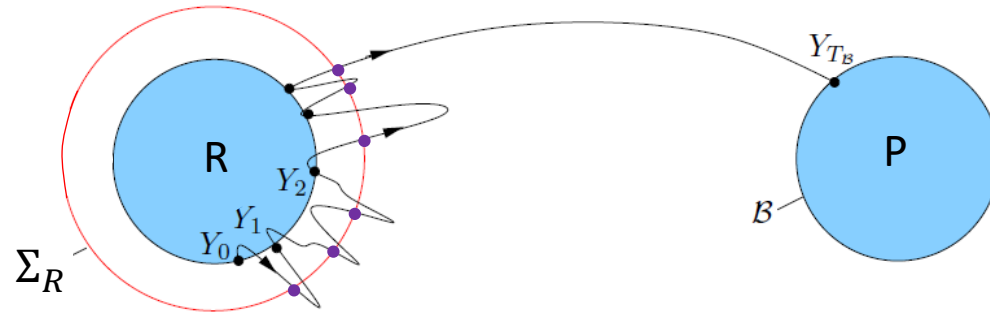
Thank you for you attention
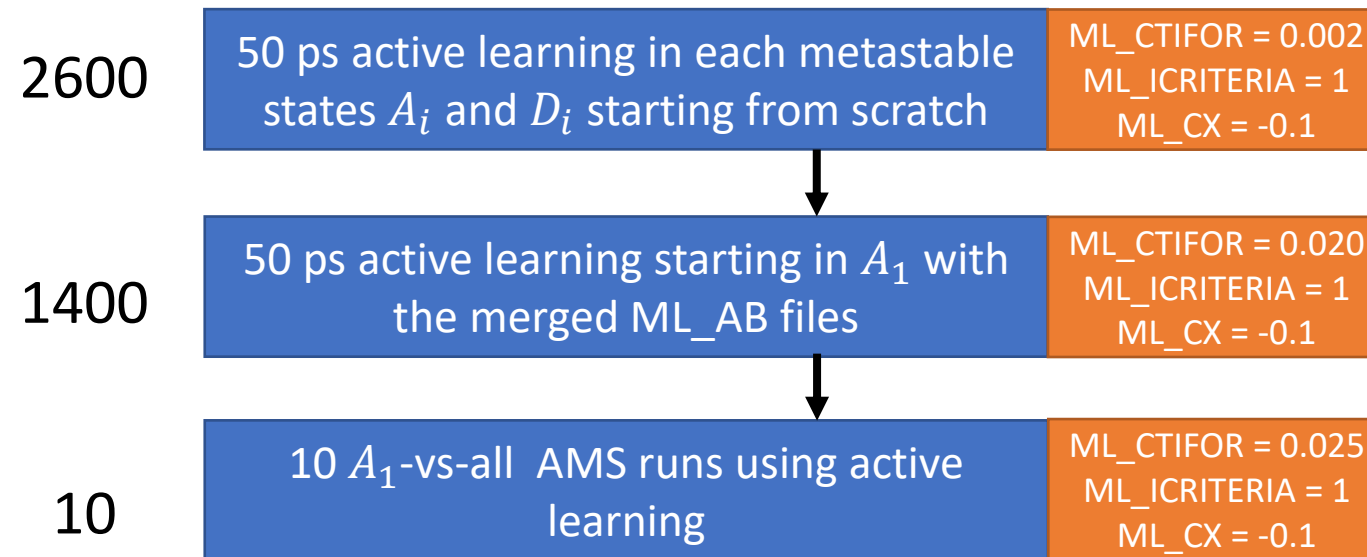
Transition time: $\frac{1}{k_{RP}} = t_{RP} = mean(\tau_i^R - \tau_i^P)$

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*.

Number of structures added to the dataset:

2600	50 ps active learning in each metastable states A_i and D_i starting from scratch	ML_CTIFOR = 0.002 ML_ICRITERIA = 1 ML_CX = -0.1

↓

1400	50 ps active learning starting in A_1 with the merged ML_AB files	ML_CTIFOR = 0.020 ML_ICRITERIA = 1 ML_CX = -0.1

↓

10	10 A_1-vs-all AMS runs using active learning	ML_CTIFOR = 0.025 ML_ICRITERIA = 1 ML_CX = -0.1

Wall clock time	Number of structures		
~0.5 days	2600	50 ps active learning in each metastable states A_i and D_i starting from scratch	Final force mean Bayesian error 10 to 60 meV/Å
~0.5 days	1400	50 ps active learning starting in A_1 with the merged datasets	25 meV/Å
~10 days	10	10 A_1-vs-all AMS runs using active learning	25 meV/Å

Threshold updated using the stored Bayesian errors[1,2]

→ Transitions already sampled during first two steps?

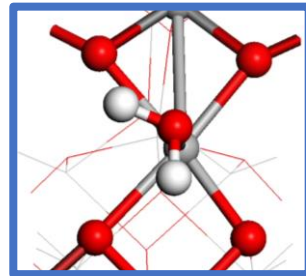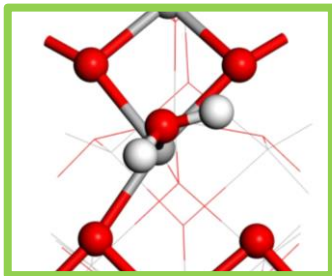[1] Jinnouchi, R.; Karsai, F.; Kresse, G. *Phys. Rev. B* **2019,** 100, 014105.
[2] Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. *The Journal of Physical Chemistry Letters* **2020**, 11, 6946–6955.

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

$\longrightarrow$ Reactive trajectories

— Iso-levels of a reaction coordinate

A_1

A_2

A_3

Σ_{A_4}

A_4