Agnostic machine learning description of chemical reactions in solution

Timothée Devergne

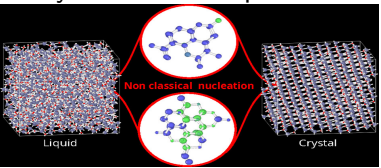
03/07/2023

Supervisors: A. Marco Saitta-Fabio Pietrucci (IMPMC)

Transformations in physics

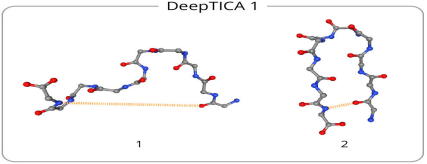
Transformations are at the heart of physics:

Study of nucleation processes:



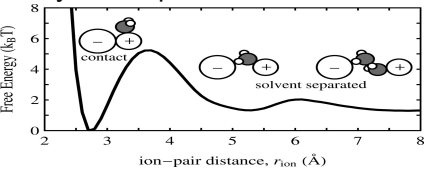
Goniakowski et al *J Phys Chem C* 2022 126 (40)

Study of protein folding:



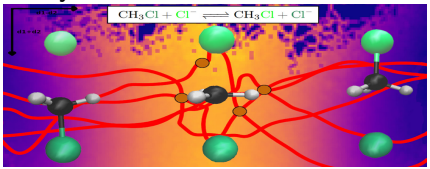
Novelli et al J. Chem. Theory Comput. 2022

Study of ion pair dissociation:



Mullen et al J. Chem. Theory Comput. 2014, 10, 2

Study of chemical reactions:

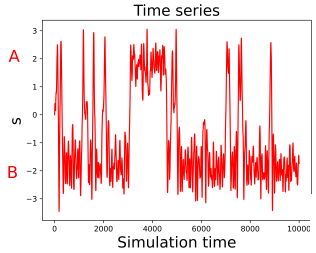


Magrino et al J. Phys. Chem. A 2022, 126, 47

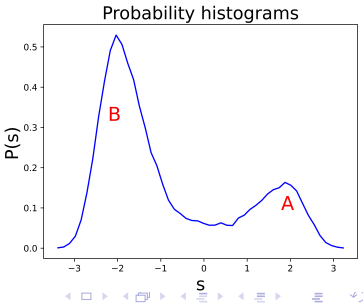
Molecular dynamics simulations

A chemical reaction is a transition between two metastable states A and B

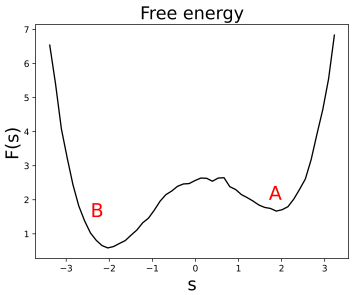
Molecular dynamics(MD): Track the evolution of a system by numerically solving Newton's equations of motion



$$P(s) = \frac{1}{T}\int_0^T \delta(s(x(t) - s)dt$$

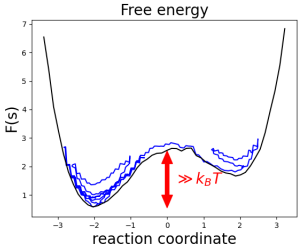


$$F(s) = -k_B T ln(P(s))$$



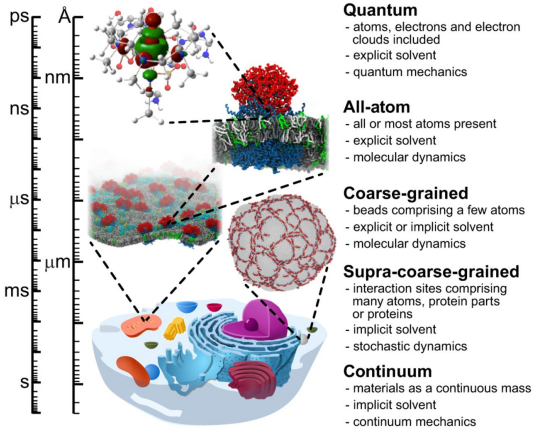
The timescale problem

The Bottleneck: need to observe many transitions to have enough statistics to recover the thermodynamics and the kinetics



Transition rates at 300K for different barrier heights:

20 kcal/mol: 1/minute
30 kcal/mol: 1/human life
40 kcal/mol: 1/billion years

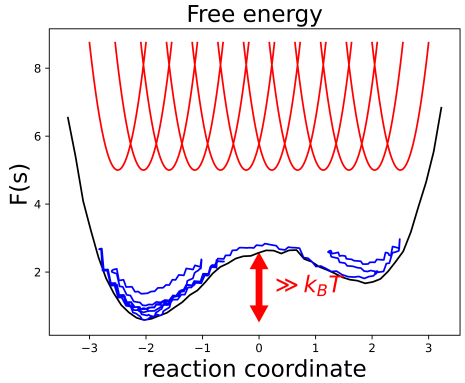


One solution: Umbrella sampling

- Split the chemical space into bins
- In each bin perform a simulation in a potential of the form:

$$V_{biased}(r) = V(r) + k(r - r_0)^2$$

- Assemble the statistics of all the bins and get the free energy
- Main problem: it depends on the reaction coordinate



Reaction coordinate

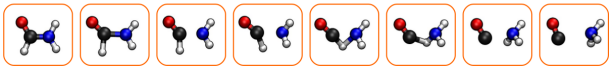
Projection of the 3N dimensional position space onto a 1D space that contains all the information about the reaction

The reaction coordinate

Use of a reaction coordinate based on a putative reactive pathway : path collective variables

Branduardi, Gervasio, Parrinello, JCP 126, 054103 (2007)

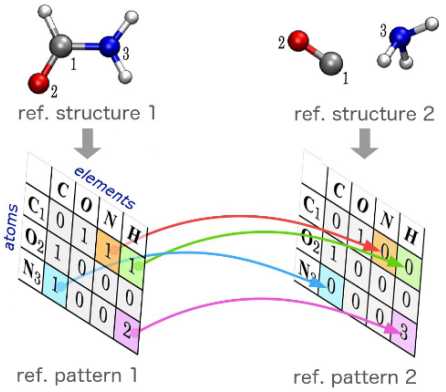
initial guess of path



$$\begin{cases} s(t) = \dfrac{1}{N-1}\left(\dfrac{\sum_{\alpha=1}^{N} \alpha \exp(-\lambda D[x(t), X_\alpha])}{\sum_{\alpha=1}^{N} \exp(-\lambda D[x(t), X_\alpha])} - 1\right) \\ z(t) = \dfrac{-1}{\lambda} \log \sum_{\alpha=1}^{N} \exp(-\lambda D[x(t), X_\alpha]) \end{cases}$$

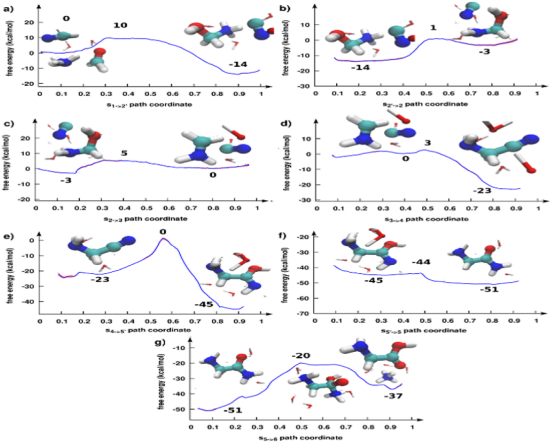
Metric that includes the solvent:

Pietrucci, Saitta, PNAS, 112(49) (2015)

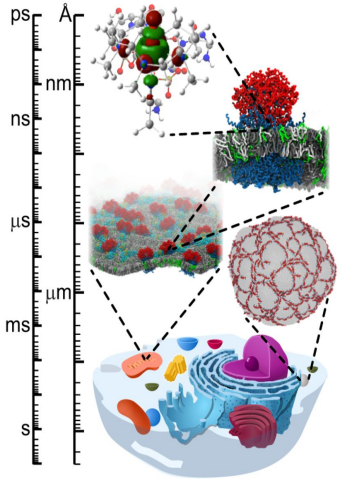


Recent results: Strecker synthesis of glycine

Thorough study of every step of the prebiotic synthesis of glycine :



Umbrella sampling was used, but each step needed at least 500k CPU.h



Quantum
- atoms, electrons and electron clouds included
- explicit solvent
- quantum mechanics

All-atom
- all or most atoms present
- explicit solvent
- molecular dynamics

Coarse-grained
- beads comprising a few atoms
- explicit or implicit solvent
- molecular dynamics

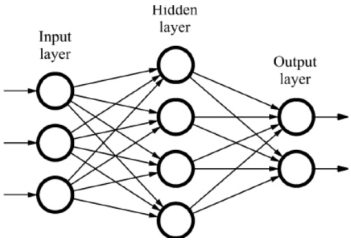
Supra-coarse-grained
- interaction sites comprising many atoms, protein parts or proteins
- implicit solvent
- stochastic dynamics

Continuum
- materials as a continuous mass
- implicit solvent
- continuum mechanics

T. Magrino, F. Pietrucci, A.M. Saitta, J. Phys. Chem. Lett. 2021, 12

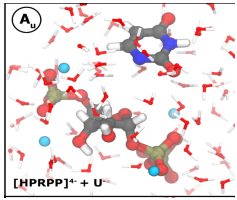
Machine learning

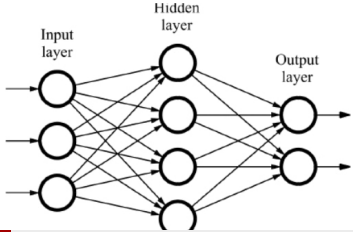
Machine learning in every day life:



ML answer:
It is a cat

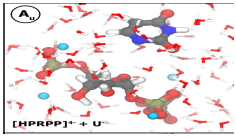
Machine learning in molecular dynamics simulations:

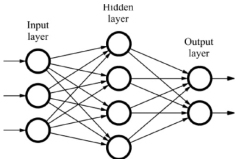




ML answer:
quantum energies and forces

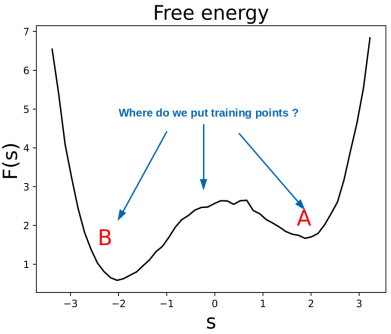
Using ML to recover thermodynamics and kinetics of chemical reactions





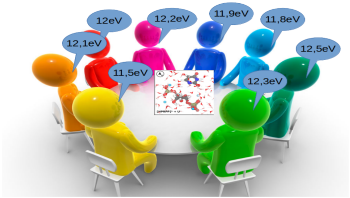
ML answer: quantum energies and forces along the chemical space

- Can we use machine learning interatomic potentials (MLIP) in such complex environments?
- Between 40-60 US windows: need of a MLIP capable of producing *ab initio* quality simulations in each window

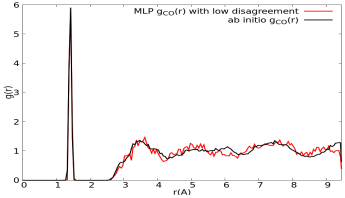


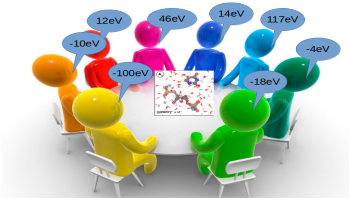
Assessing the quality of a simulation

Train several ML models to form a "committee", and compute the deviation of the prediction between them

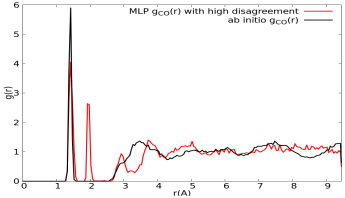


Agreement between all the members gives good physical results





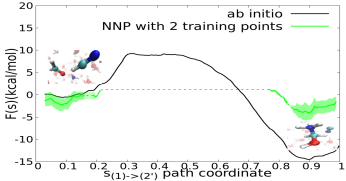
Disagreement between all the members gives bad physical results



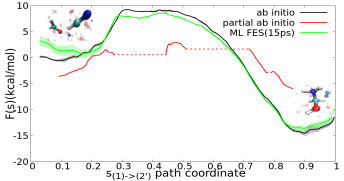
First method: Pick training points along a pre defined CV

With this method :

- Umbrella sampling is the most expensive part of the method
- Simulations scale with the number of atoms(N) instead of N^3
- Bigger systems can be studied



add training points at the transition state

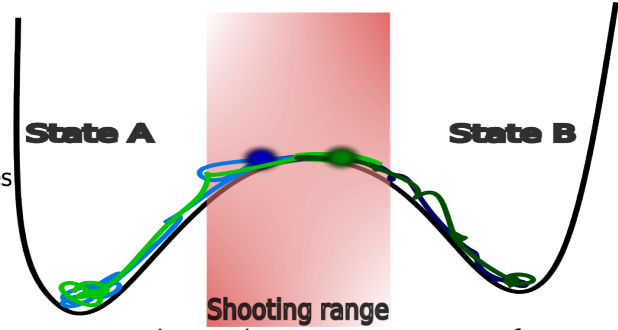


Devergne, Magrino, Pietrucci, Saitta, *J. Chem. Theory Comput.*, 2022, 18, 5410-5421)

Second method: Use of transition path sampling

Even though we got good results with the previous method, it depends on the collective variable. Use of transition path sampling(TPS) to obtain many transition trajectories

- Based on a metropolis algorithm
- Used to explore the transition state ensemble, and generate transition trajectories
- Drawback and bottleneck: computational time needed



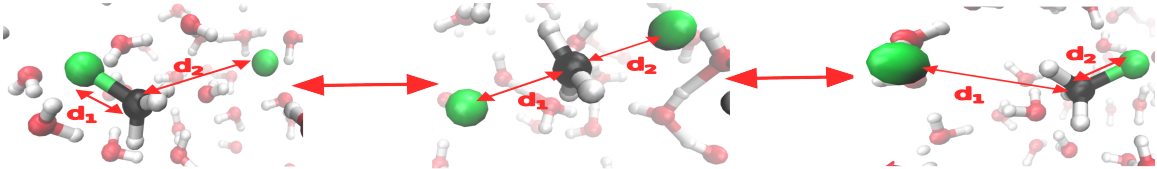
Can we train a machine learning potential on transition path sampling trajectories to perform more TPS and recover the free energy ?

A well known toy reaction:

$CH_3Cl + Cl^- \rightleftharpoons CH_3Cl + Cl^-$

- Simple, symmetric, S_N2 reaction
- Extensively studied in the team

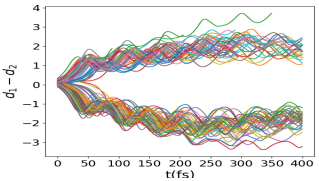
  Magrino, Huet, Saitta, Pietrucci J. Phys. Chem. A 2022, 126, 47, 8887–8900
- Many TPS data available
- Use of this data to train a MLIP

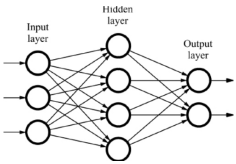


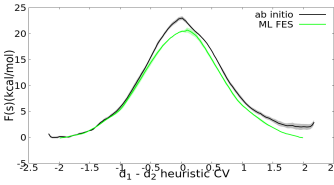
Reactants: d_1-d_2<0 **Transition state: d_1-d_2≈0** **Reactants: d_1-d_2>0**

First results

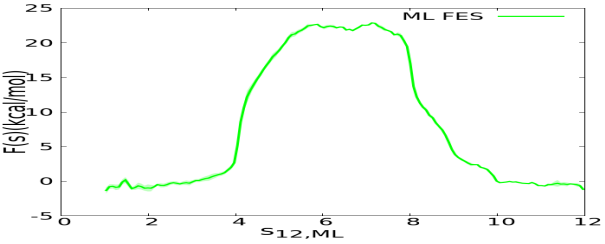
Train a MLIP on 54 reactive AIMD trajectories, perform US simulations on a simple CV

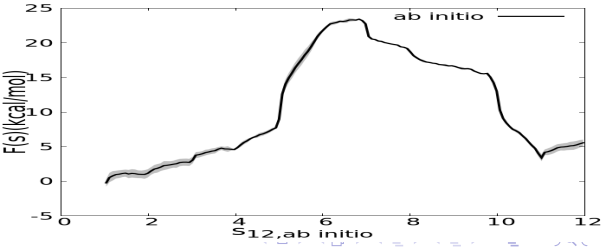






With the same MLP, perform US on a different, more generic CV



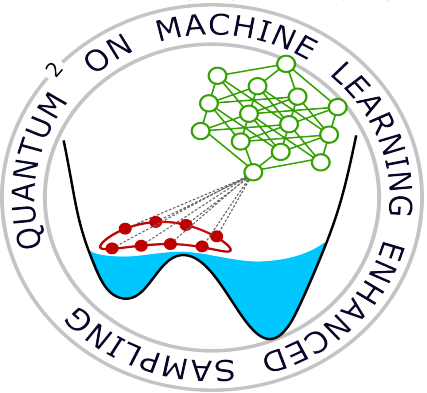


Summary

- We devised a method to have an efficient machine learning potential but this depends on the collective variable
- Luckily, TPS exists
- We now have a machine learning potential accurate on the whole chemical space without the prior knowledge of the transition mechanism
- We can perform more TPS simulations with this MLP

Advertisement

If you want to know more about ML, enhanced sampling and nuclear quantum effects come to our CECAM workshop:
Quantum^2 on enhanced sampling and machine learning
Lausanne, 29/11/2023-01/12/2023, https://www.cecam.org/workshop-details/1255



Among invited speakers:

- Jörg Behler
- Dominik Marx
- Alexandre Tchtchenko
- and more

Thank you for your attention

Chouchen the Machine learning cat: