barbatti.org

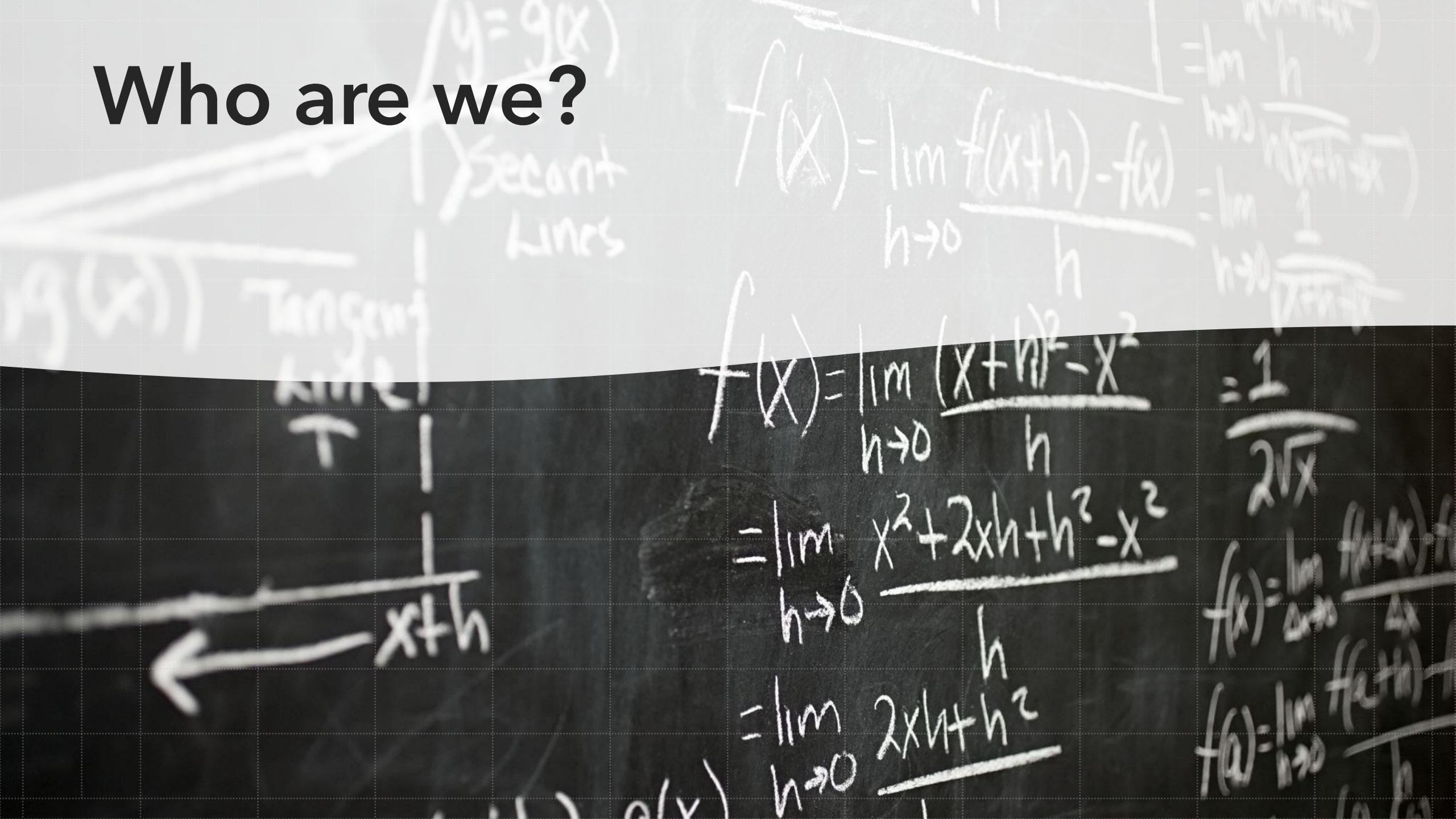
Machine learning for theoretical chemistry

Mario Barbatti

Aix Marseille Université, CNRS, Institut de Chimie Radicalaire

Institut Universitaire de France

Who are we?



The Light & Molecules group

The Light & Molecules Group

Methods

Nonadiabatic dynamics
Nuclear ensembles

Software

NEWTON-X platform

Applications

Photoprocesses in

- Fundamental PhysChem
- Molecular biology
- Organic devices
- Environment

The Light & Molecules Group

Current members

Mario Barbatti (PI)

Baptiste Demoulin (IT researcher)

Josene M Toldo (postdoc)
Saikat Mukherjee (postdoc)
Bidhan Garain (postdoc) *

Rafael Mattos (PhD candidate)
Matheus Bispo (PhD candidate) *

Recent past members

Mariana T do Casal
Ritam Mansour

Shuming Bai
Lijljana Stojanovic
Carlos E de Moura
Fabris Kossoski
Prateek Goel
Max Pinheiro Jr *
Moumita Kar

* Dedicated to ML projects

LIGHT AND MOLECULES

Marseille-Xiamen consortium

Marseille (ICR)
Mario Barbatti

NX
newtonx.org

Nuclear ensembles
Nonadiabatic dynamics
Unsupervised ML

Xiamen Univ
Pavlo Dral

MLatom.com

Atomistic supervised ML

Pinheiro Jr *et al*. *Sci Data* **2023**, *10*, 95
Zhang *et al.*, *In Quantum Chemistry in the Age of ML*, **2023**
Pinheiro Jr and Dral., *In Quantum Chemistry in the Age of ML*, **2023**
Barbatti *et al*. *JCTC* **2022**, *18*, 6851
Dral; Barbatti. *Nat Rev Chem* **2021,** *5*, 388
Pinheiro Jr et al. *Chem Sci* **2021**, *12*, 14396
Dral et al. *Top Curr Chem* **2021**, *379*, 27
Xue; Barbatti; Dral. *J Phys Chem A* **2020**, *124*, 7199
Dral; Barbatti; Thiel. *J Phys Chem Lett* **2018**, *9*, 5660

Marseille (ICR)
Mario Barbatti

Nuclear ensembles
Nonadiabatic dynamics
Unsupervised ML

Xiamen Univ
Pavlo Dral

Atomistic supervised ML

Marseille (LIS)
Thierry Artières
Hachem Kadri

Data science expertise

The Newton-X platform

Newtonian Dynamics Close to the X-Seam

- Surface hopping & Nuclear ensemble simulations
- Freeware
- Open source

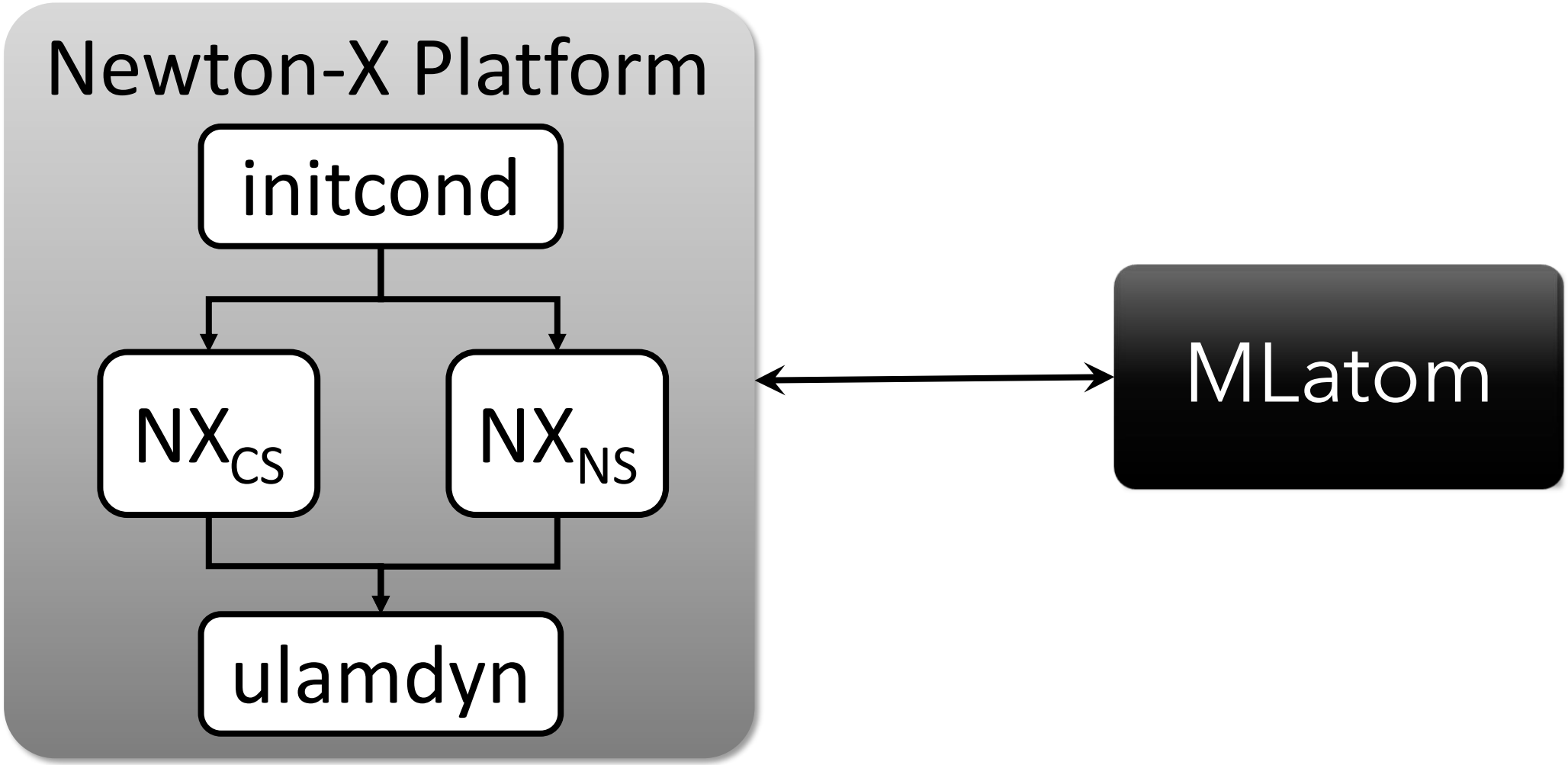
Baptiste Demoulin

Barbatti *et al*. *JCTC* **2022,** *18*, 6851

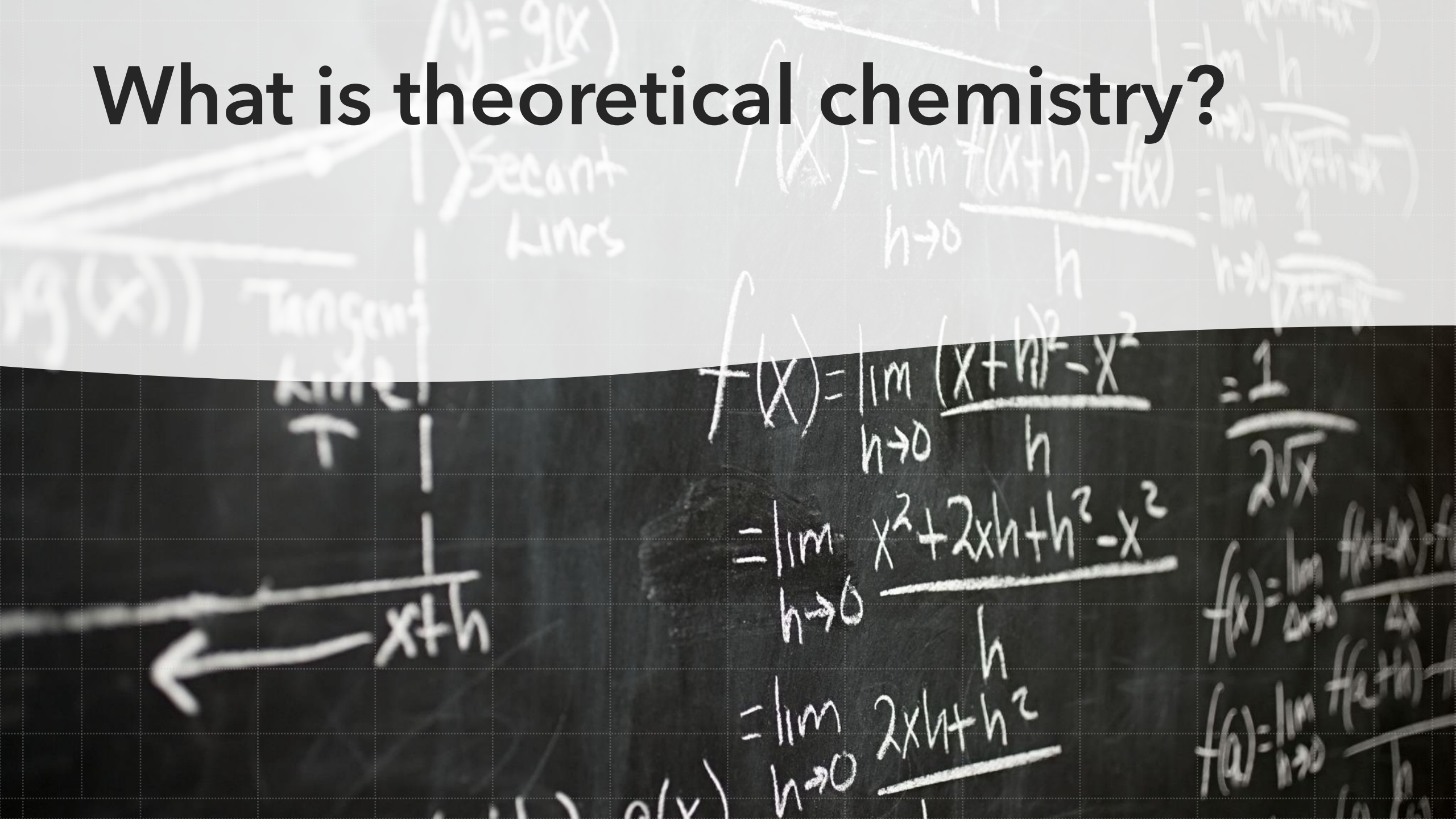
Newtonian Dynamics Close to the X-Seam

- Simulations with MRCI, MCSCF, CASPT2, ADC(2), TDDFT, TD-DFTB, Semiempirical/CI, Analytical models, ML potentials
- Interfaces to Columbus, Turbomole, Gaussian, Bagel, Gamess, CP2K, DFTB+, Mopac (Pisa), ORCA, Open Molcas, MNDO, MLatom

Barbatti *et al*. *JCTC* **2022,** *18*, 6851


Newton-X Platform
initcond
NX$_{CS}$
NX$_{NS}$
ulamdyn
MLatom

What is theoretical chemistry?



Quantum mechanics of molecules

Schrödinger equation for the molecule (including electrons and nuclei)

$$i\hbar \frac{\partial \Psi}{\partial t} = \hat{H}\Psi$$

Following Born and Oppenheimer's approach, this problem simplifies to

Electrons (r)

Electronic Schrödinger equation
(adiabatic approximation)

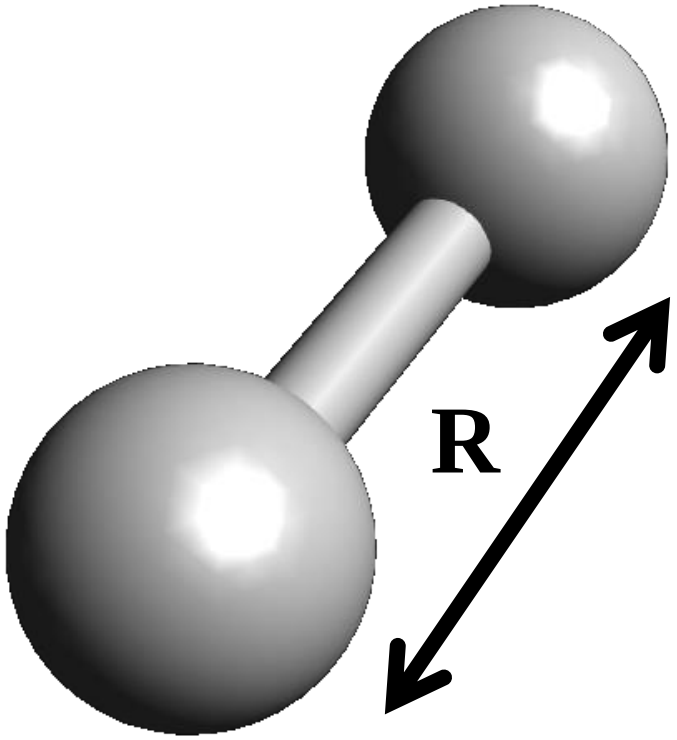
$$\left(T_{elec}(\mathbf{r}) + V(\mathbf{r},\mathbf{R})\right)\varphi(\mathbf{r};\mathbf{R}) = E(\mathbf{R})\varphi(\mathbf{r};\mathbf{R})$$

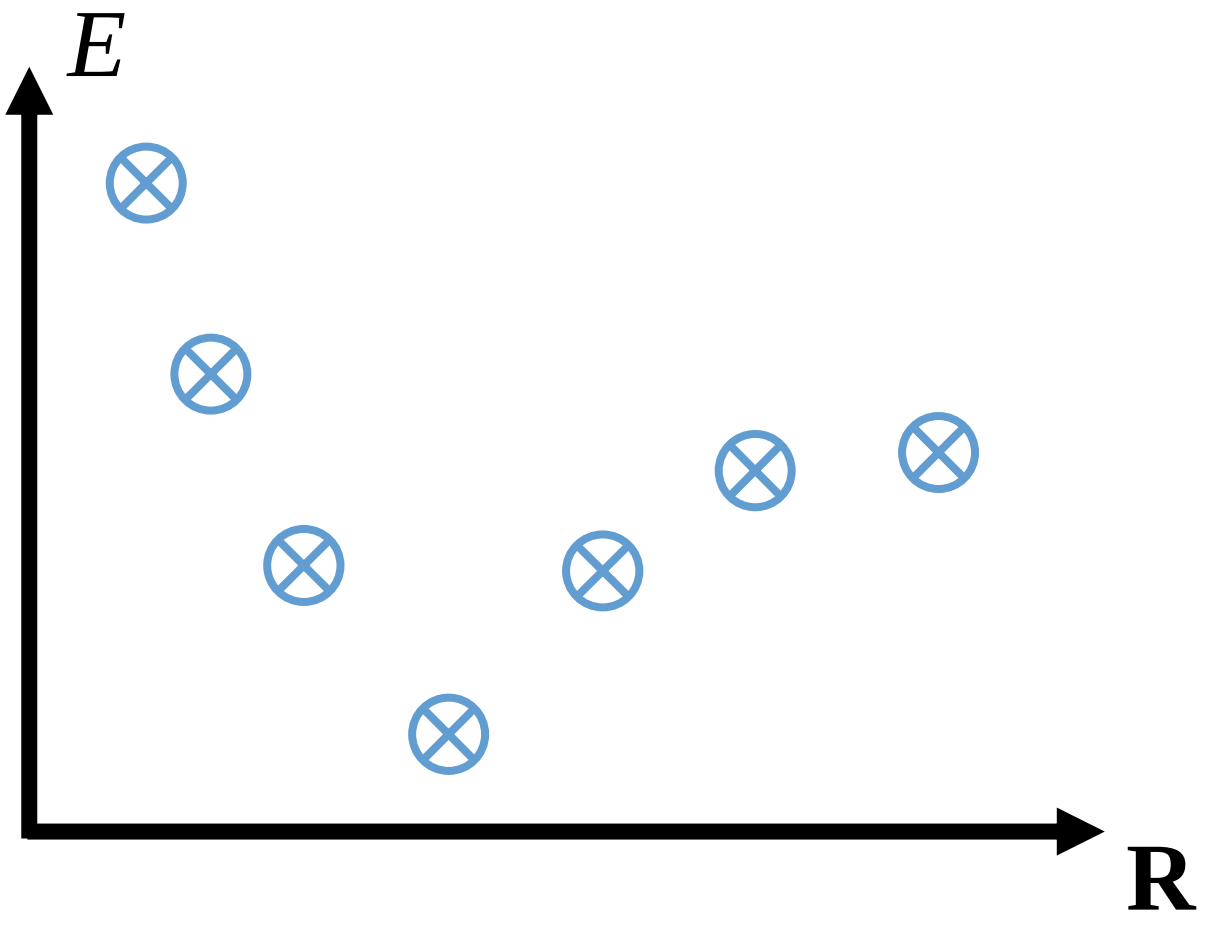
Nuclei (R)

Nuclear Newton's equation
(Classical approximation)

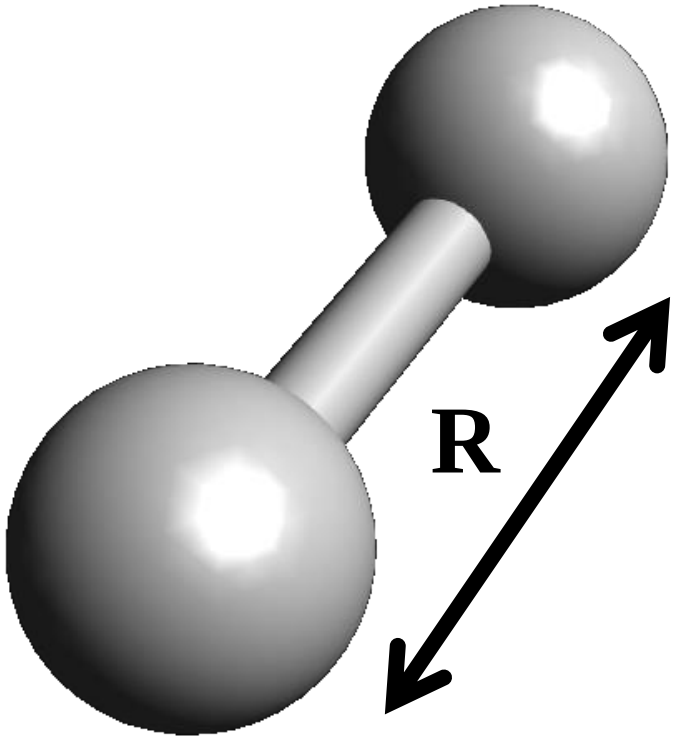
$$M_\alpha \frac{d^2\mathbf{R}_\alpha}{dt^2} = -\nabla_\alpha E(\mathbf{R})$$

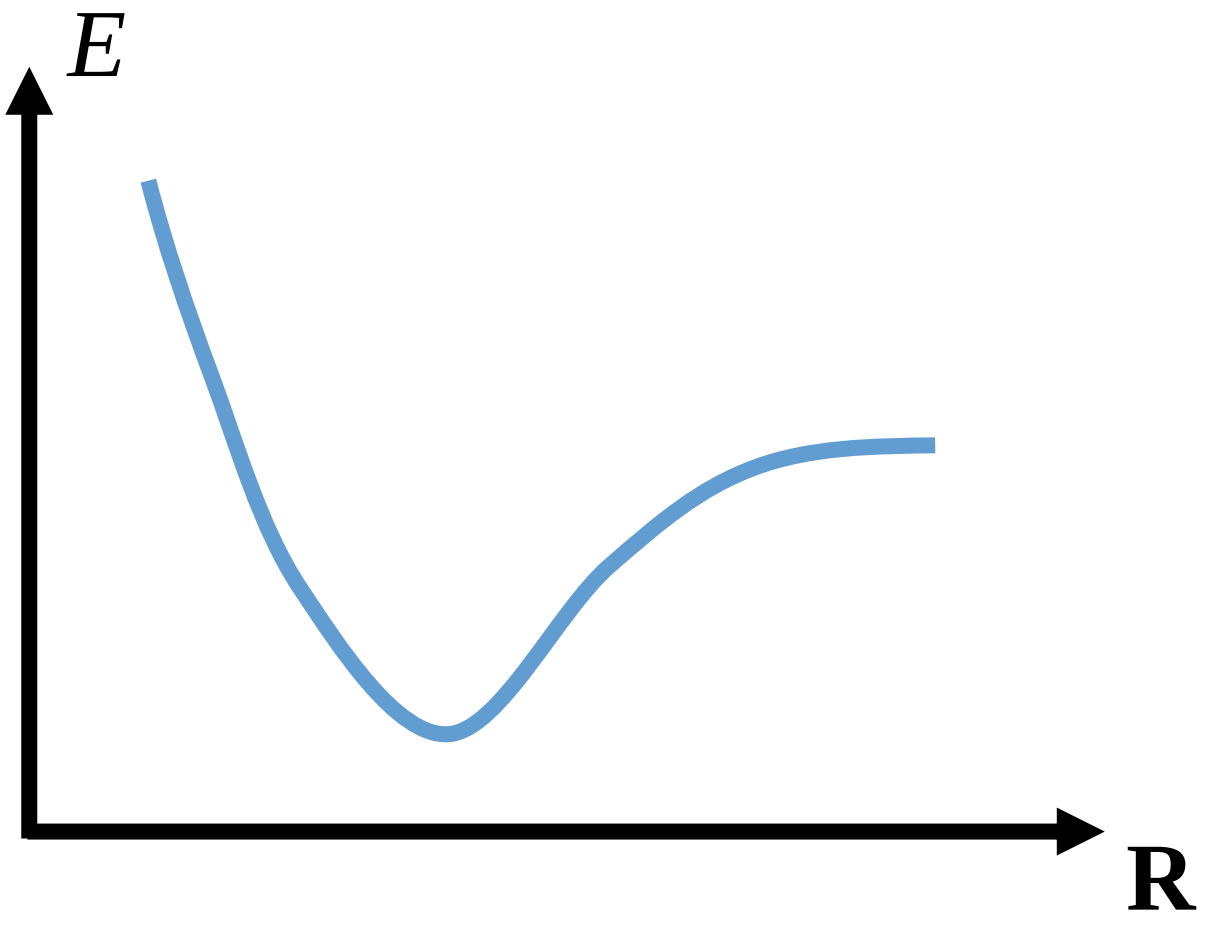
The core quantity is the potential energy surface $E(\mathbf{R})$



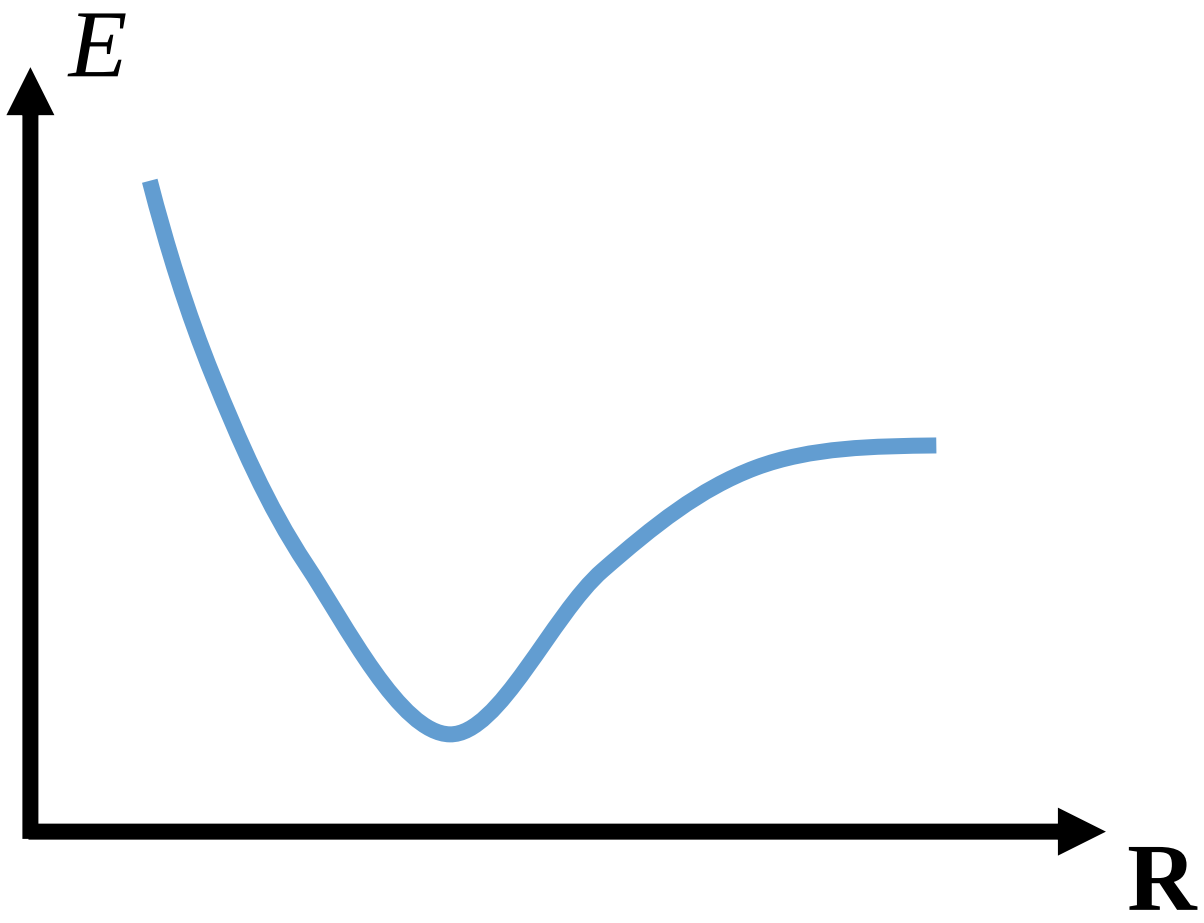


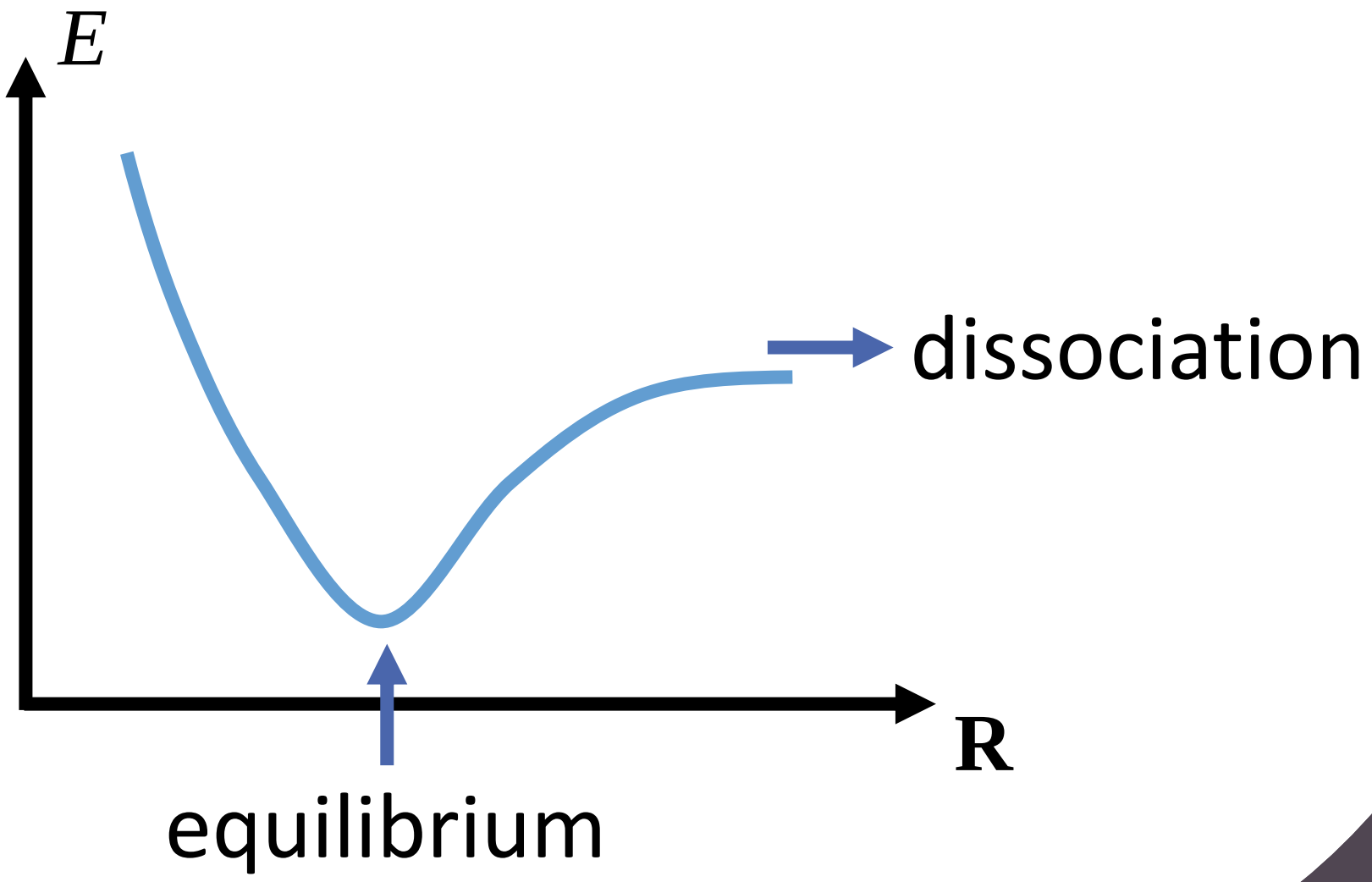
$$H_{elec}\varphi\left(\mathbf{r}_1,\mathbf{r}_2;\mathbf{R}\right)=E\left(\mathbf{R}\right)\varphi\left(\mathbf{r}_1,\mathbf{r}_2;\mathbf{R}\right)$$

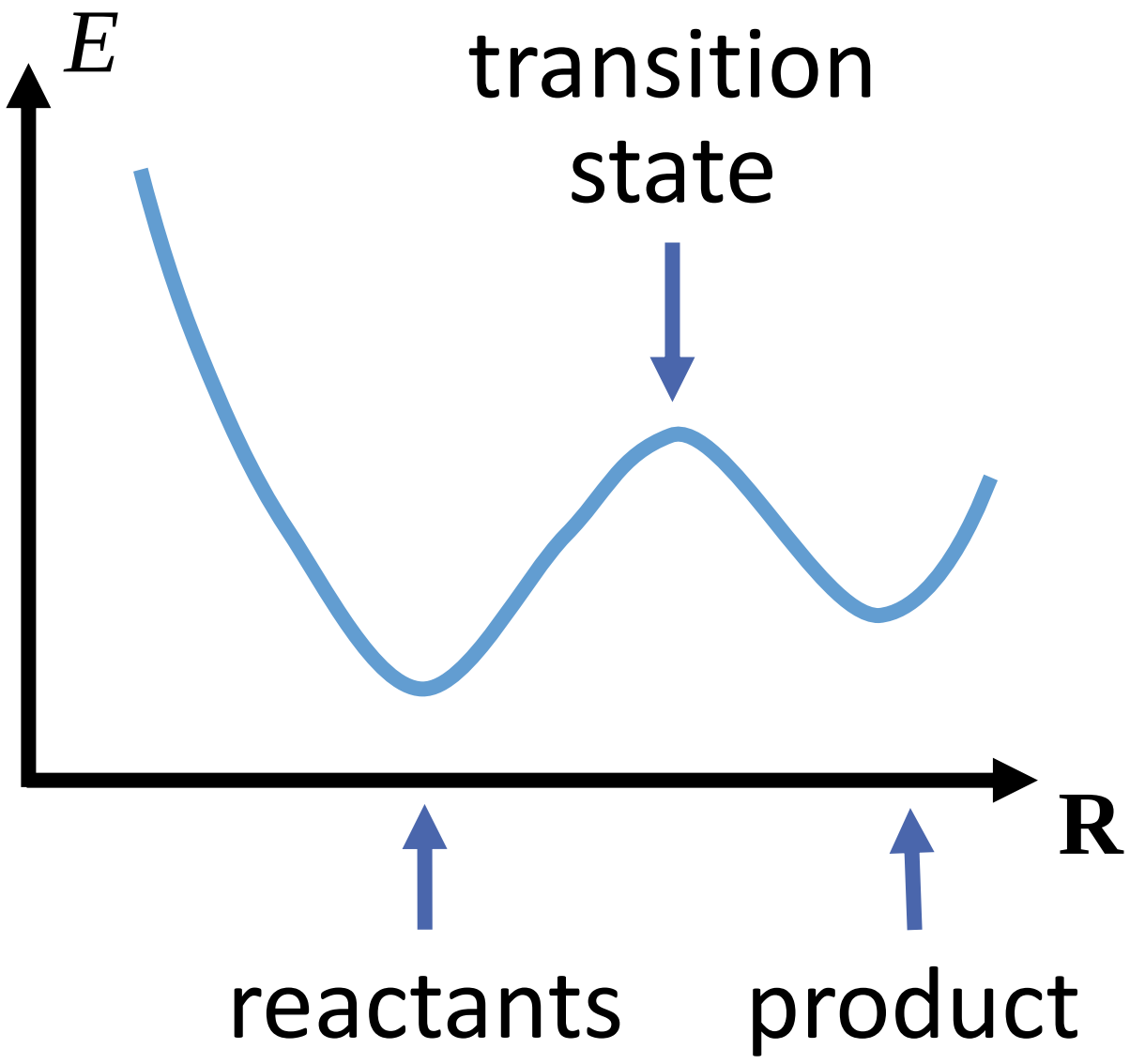


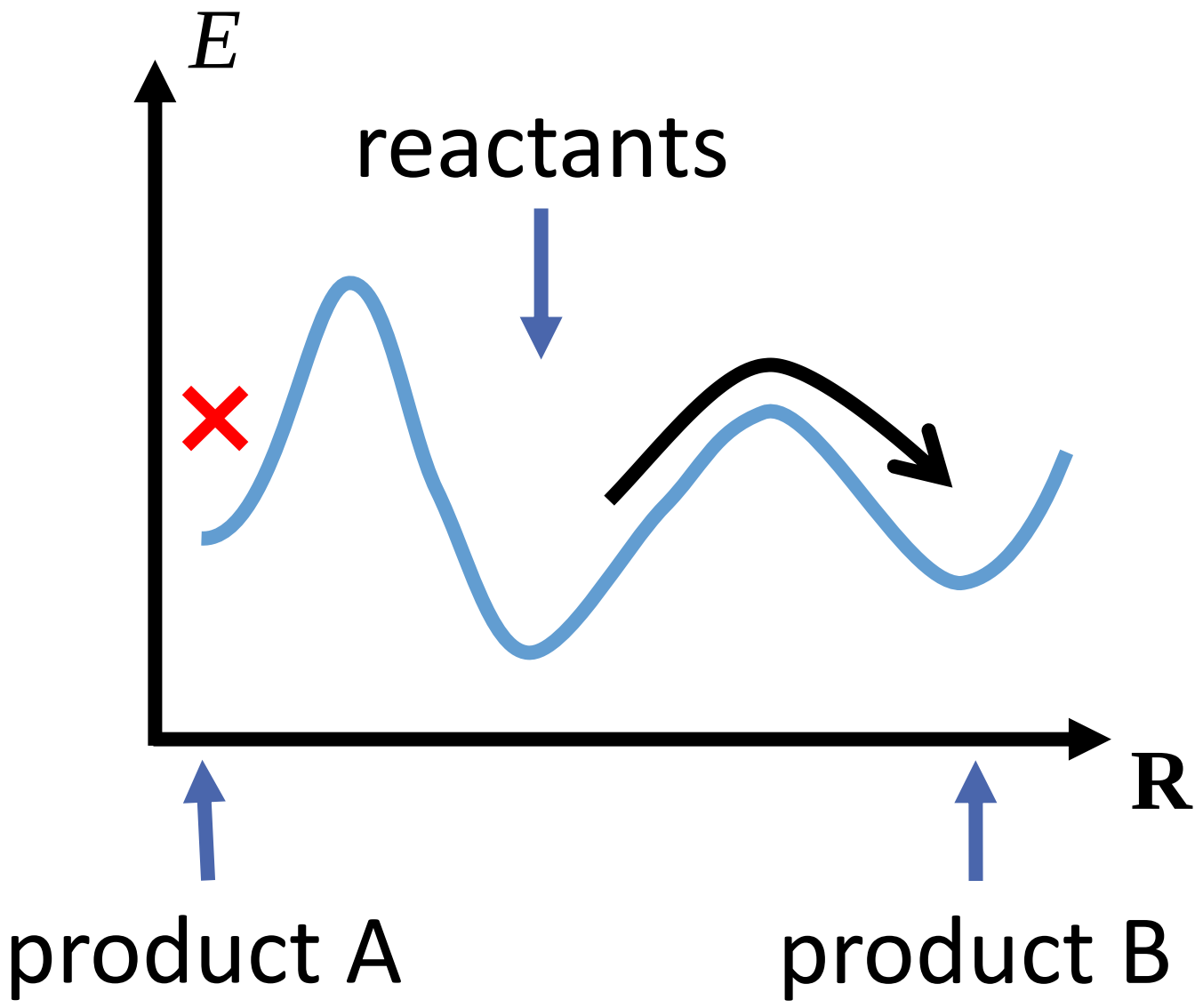


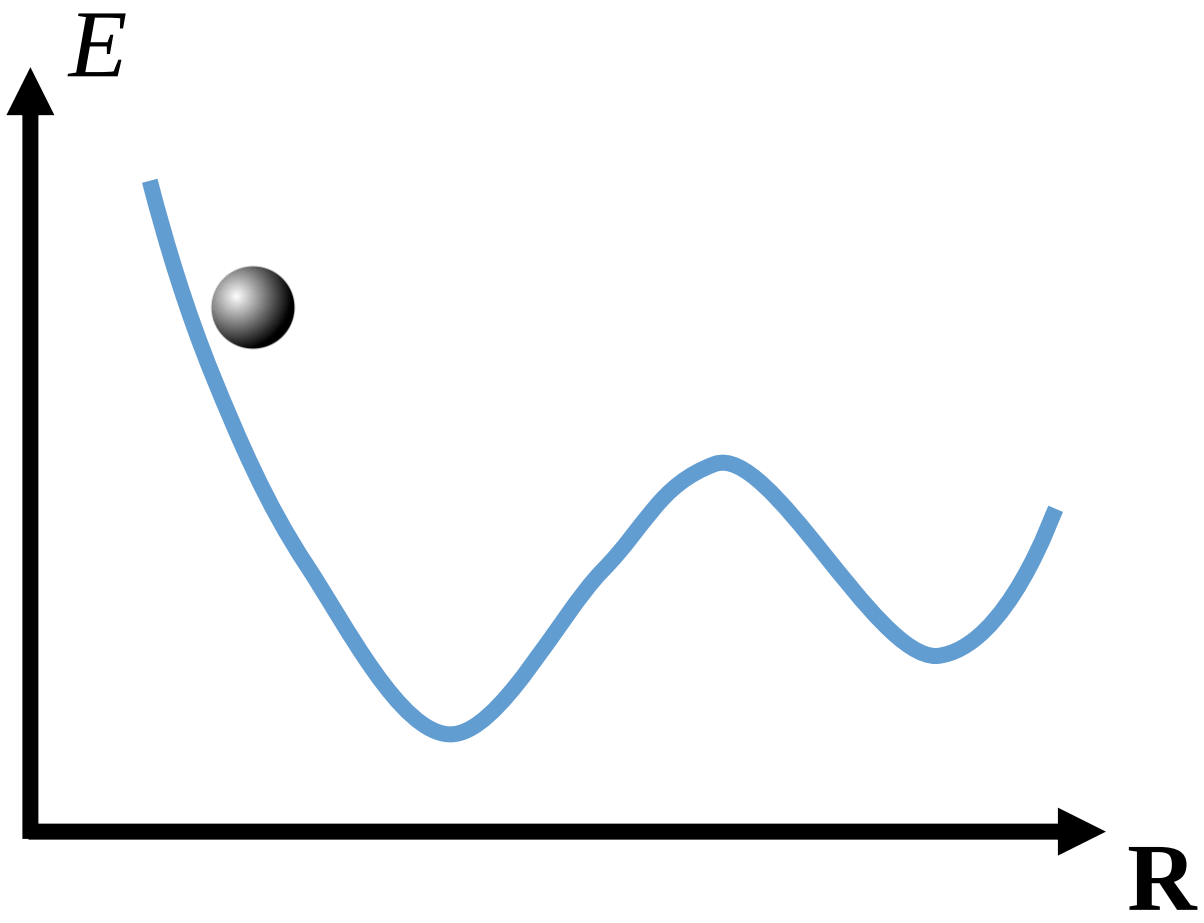
$$H_{elec}\varphi\left(\mathbf{r}_1,\mathbf{r}_2;\mathbf{R}\right)=E\left(\mathbf{R}\right)\varphi\left(\mathbf{r}_1,\mathbf{r}_2;\mathbf{R}\right)$$


E
R


E
dissociation
R
equilibrium

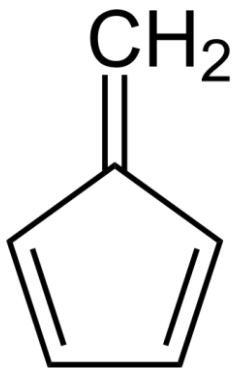

E
transition
state
R
reactants
product


E
reactants
R
product A
product B


E
R

Potential energy surfaces $E(\mathbf{R})$ have $3N_{at}$-6 dimensions

For fulvene, N_{at} = 12, $E(\mathbf{R})$ has 30 dimensions

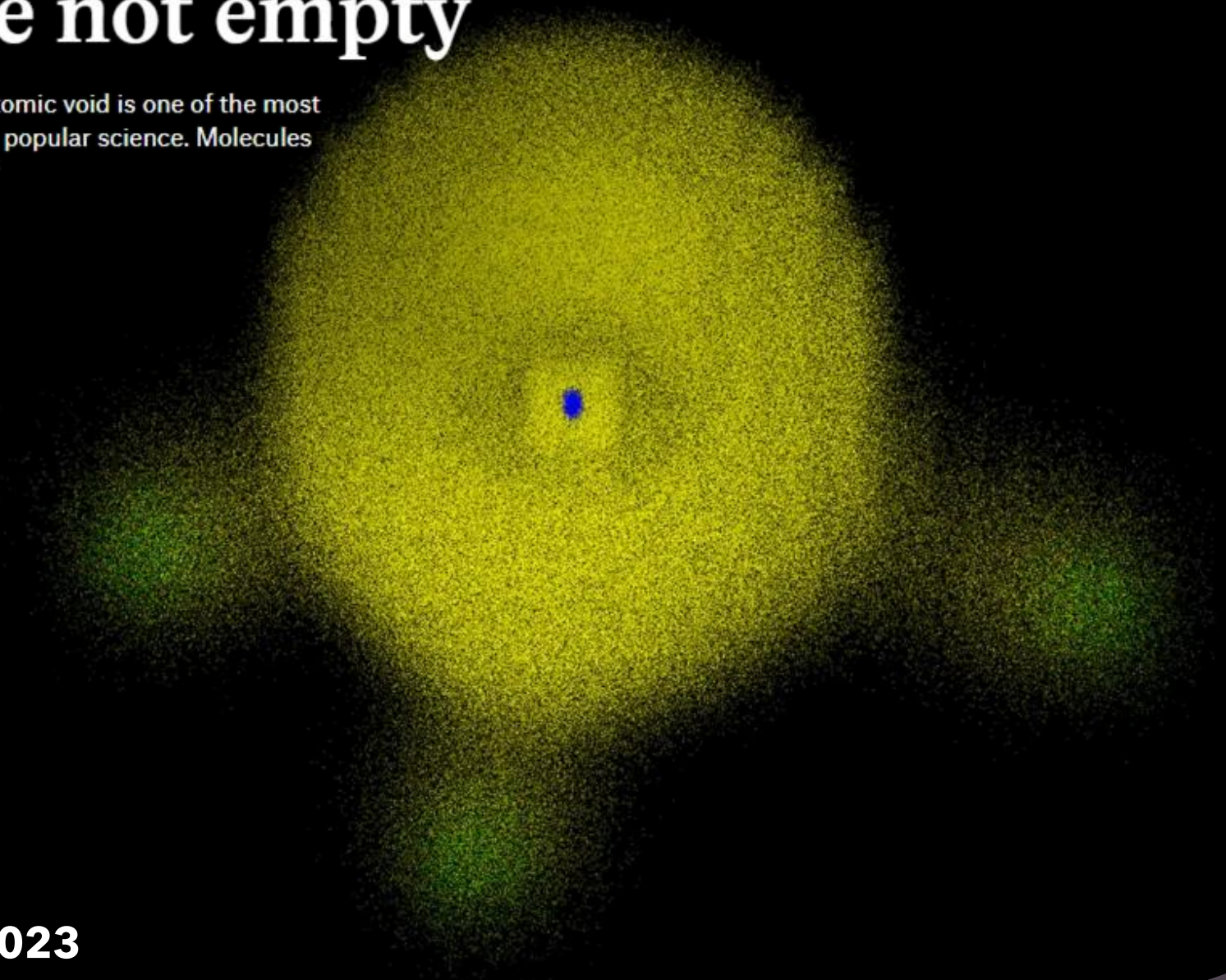


fulvene

MENU / / aeon DONATE / NEWSLETTER / SIGN IN / PSYCHE

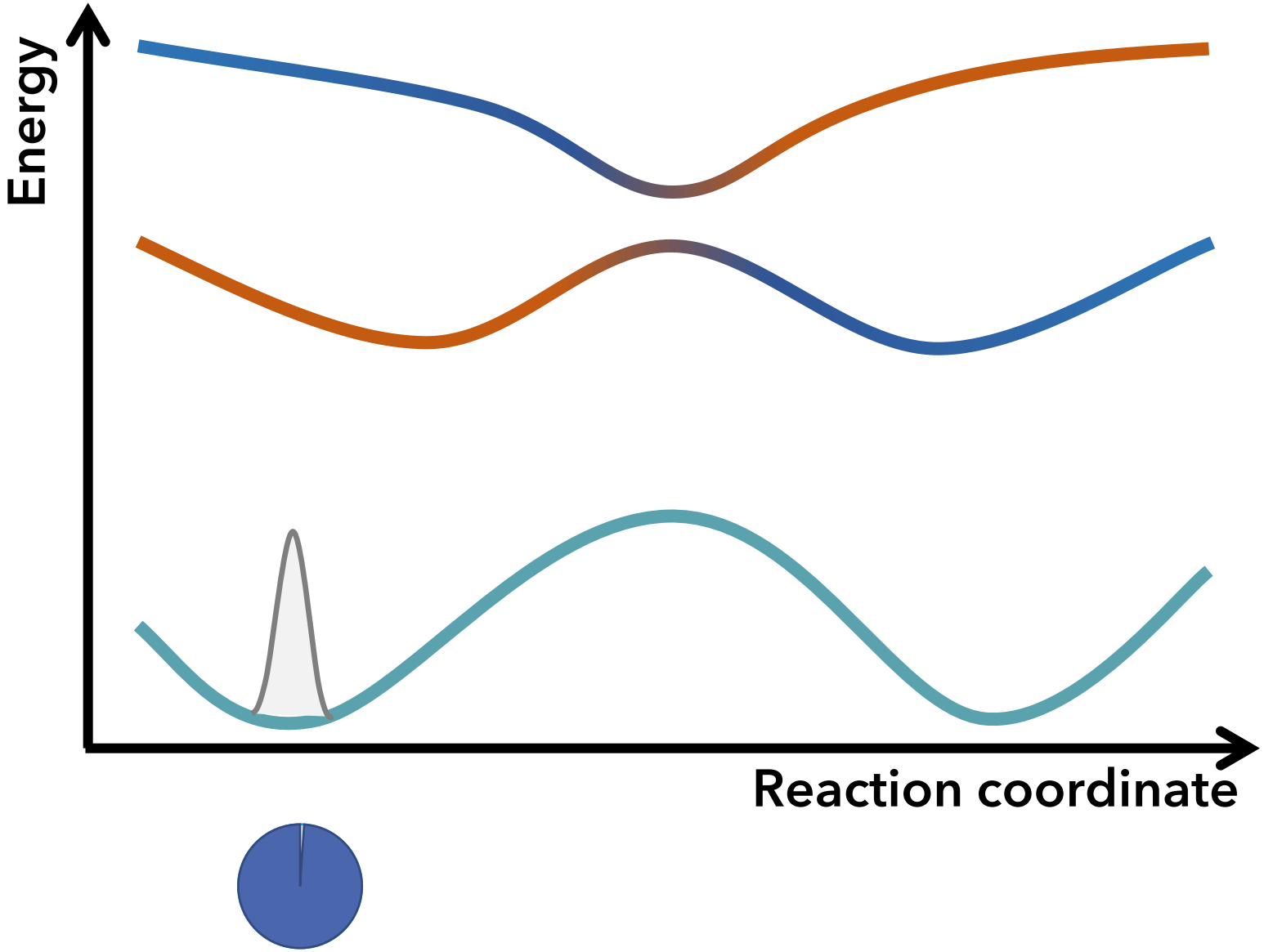
We are not empty

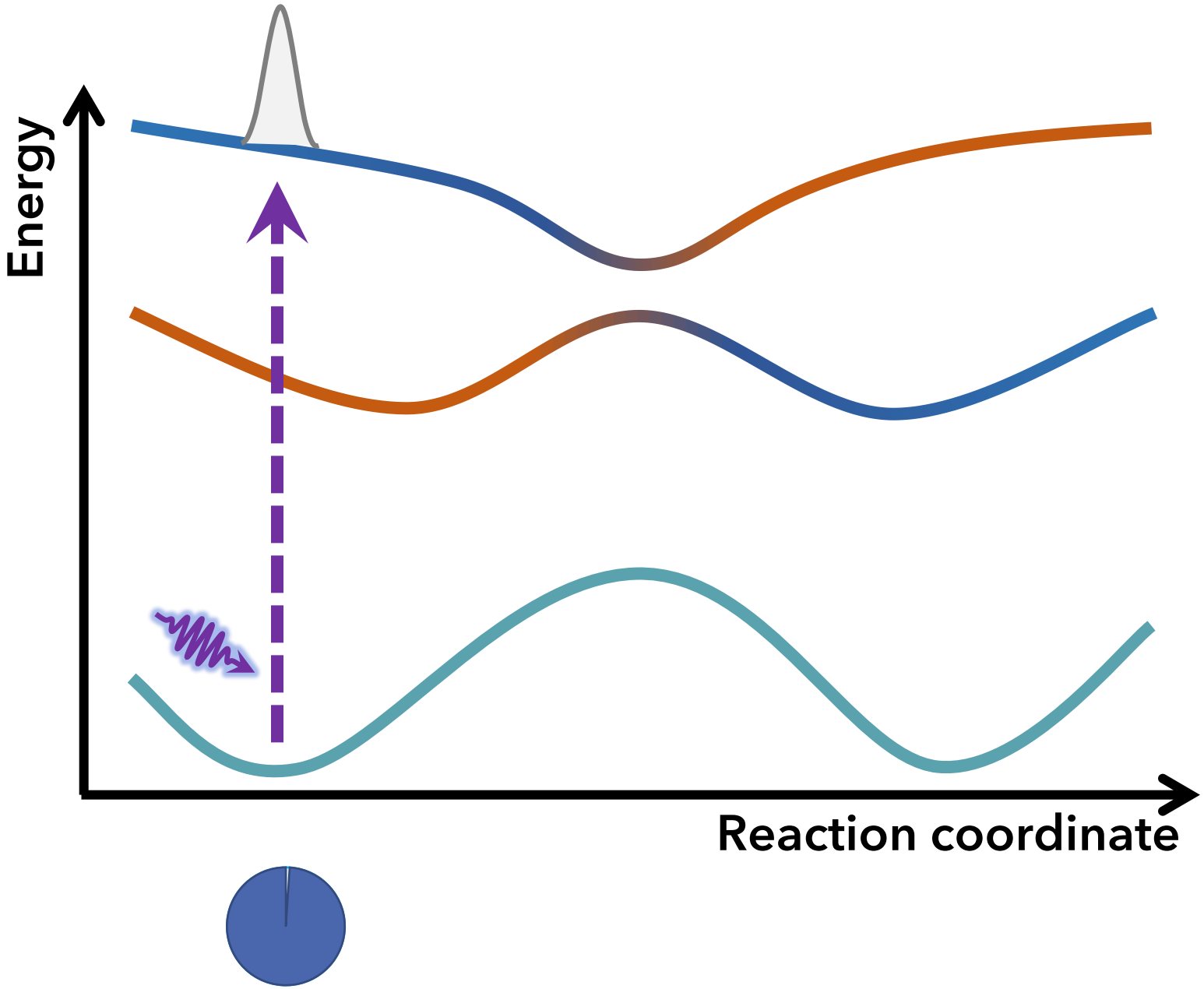
The concept of the atomic void is one of the most repeated mistakes in popular science. Molecules are packed with stuff

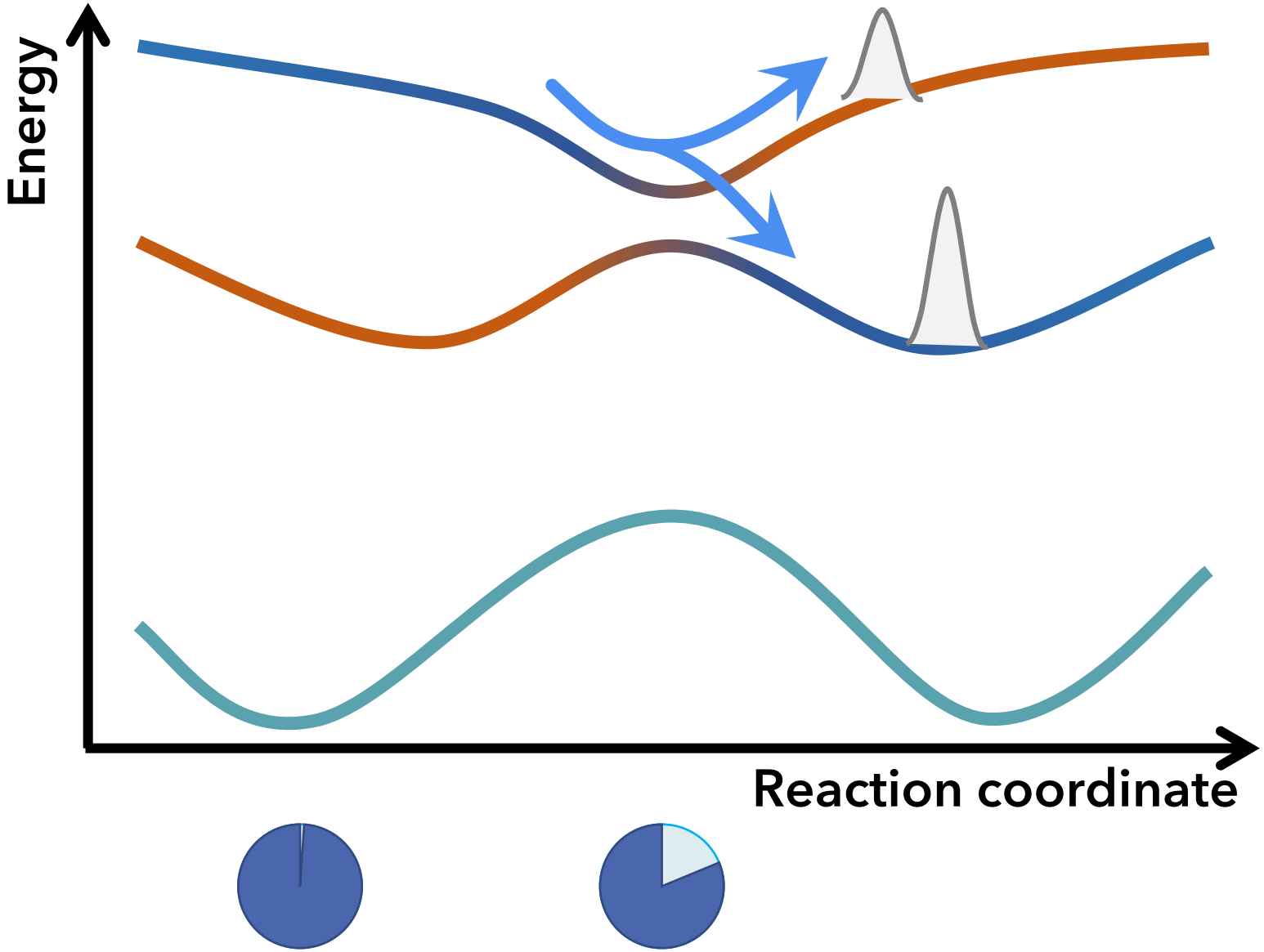


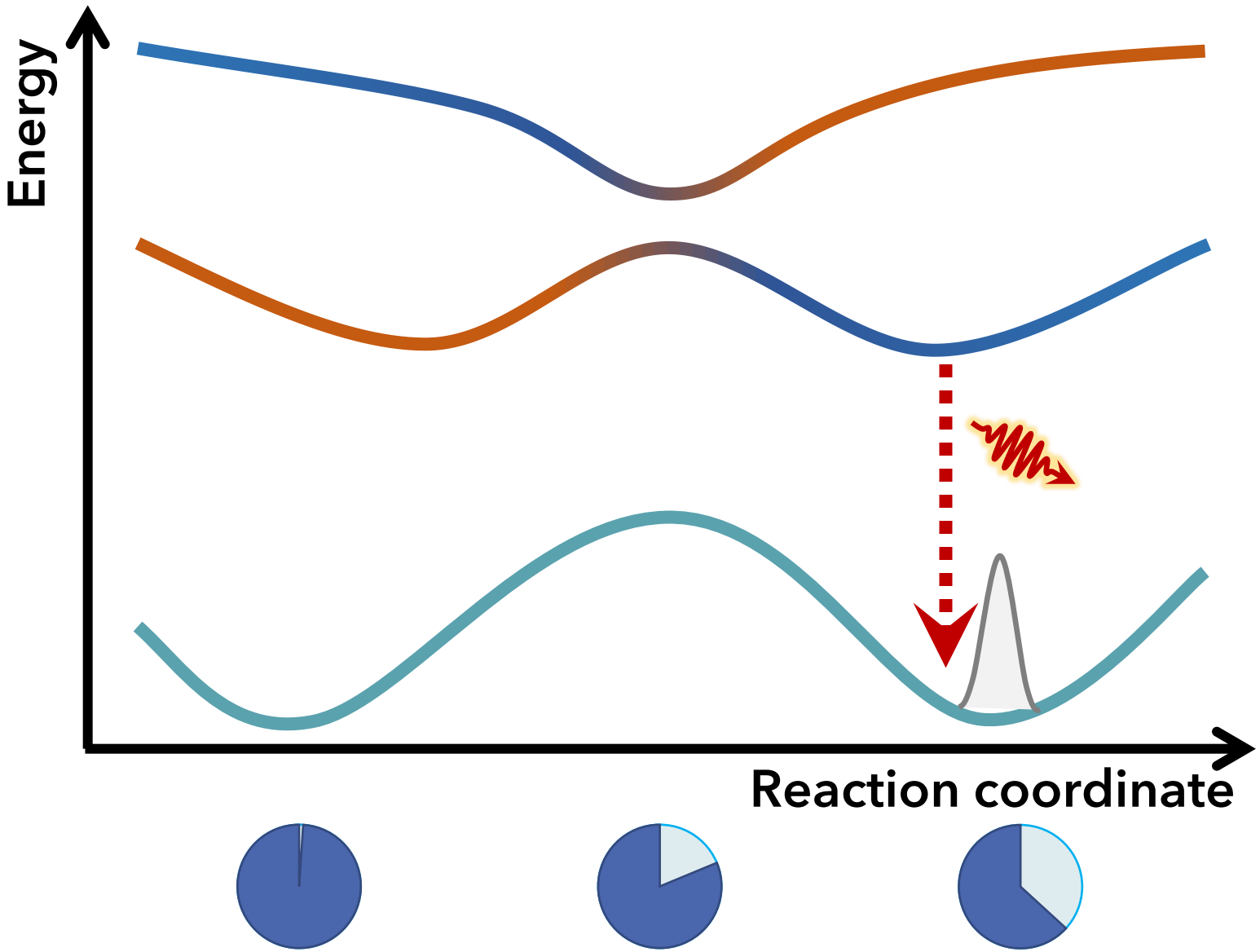
Barbatti. *Aeon Mag* **2023**

Nonadiabatic molecular dynamics (NAMD)


Energy
Reaction coordinate


Energy
Reaction coordinate


Energy
Reaction coordinate

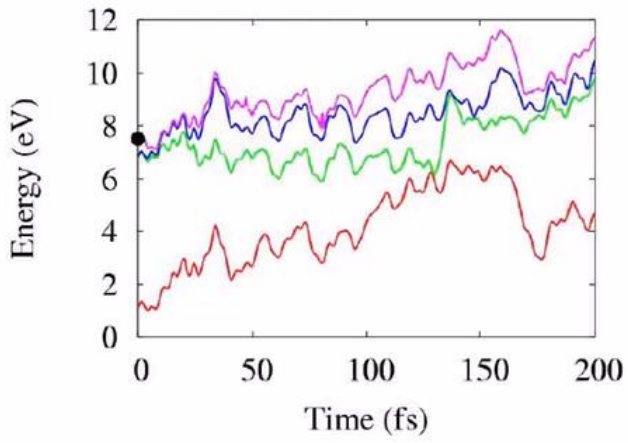

Energy
Reaction coordinate

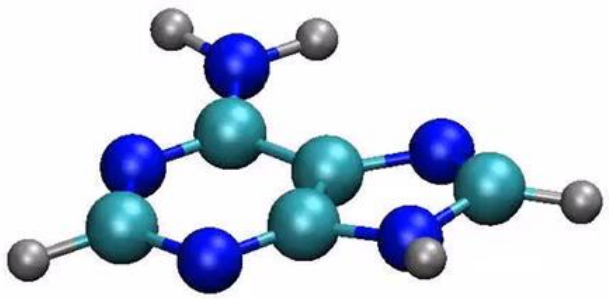
How can we simulate
nonadiabatic molecular dynamics?

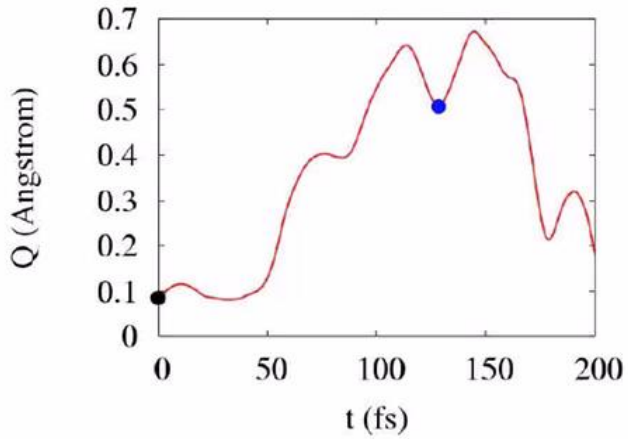
Mixed quantum-classical methods

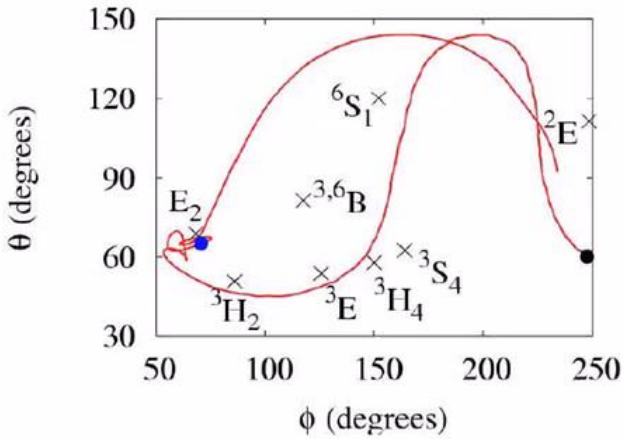
1. Nuclei are treated via *classical trajectories*
2. Electrons are treated *quantum mechanically*
3. A nonadiabatic algorithm introduces *post Born-Oppenheimer effects*

Crespo-Otero; Barbatti. *Chem Rev* **2018,** *118*, 7026



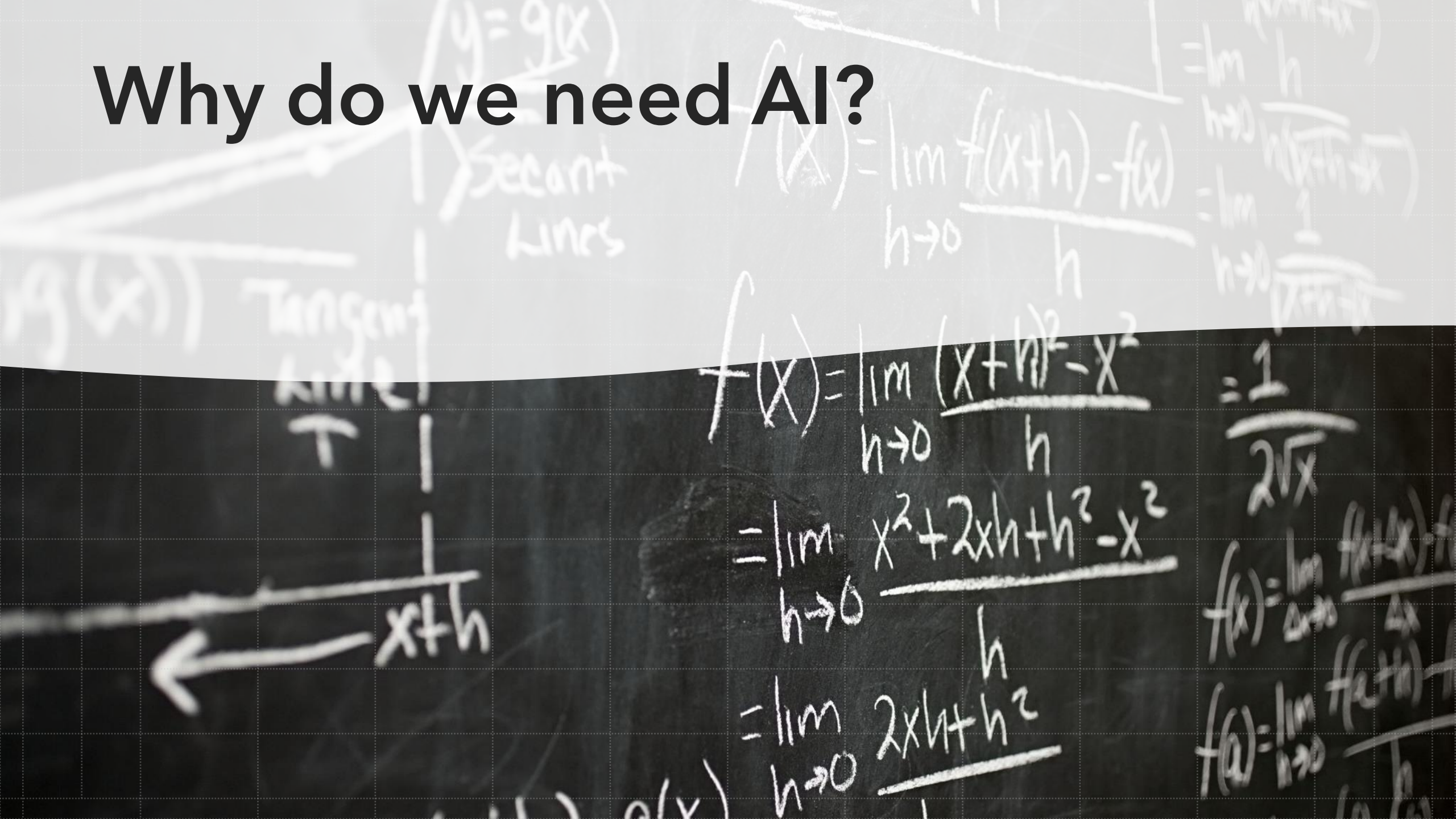






youtube.com/user/mbarbatti

Why do we need AI?



AI for theoretical chemistry has been used to

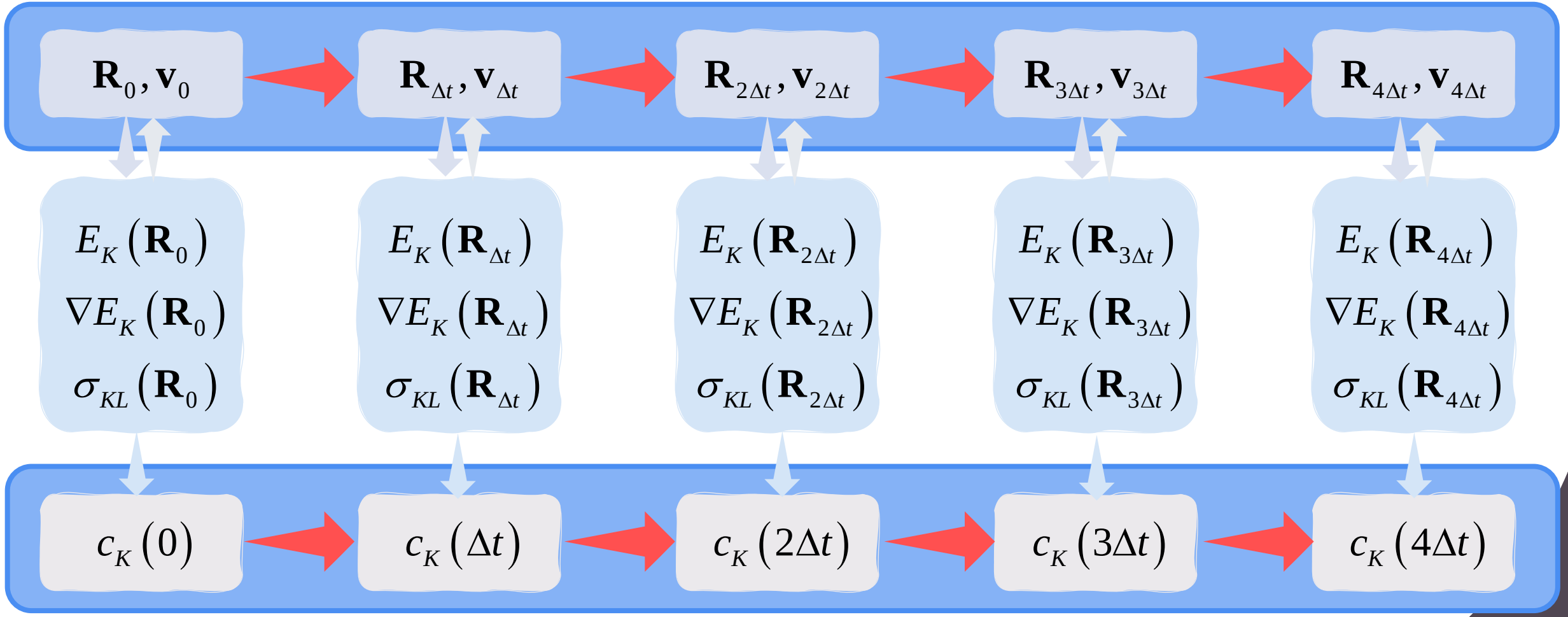
- Search the chemical space of compounds
- Perform dimensionality reduction, clustering, and pattern recognition
- Improve or accelerate quantum chemical methods
- Predict properties as a surrogate approach

Dral (Ed.). *Quantum Chemistry in the Age of ML*, Elsevier, **2023**

Costs of dynamics

Nonadiabatic dynamics

Classical EOM



Quantum EOM

Dynamics may be expensive

$$T_{total} \approx N_{\text{Trajectories}} \times N_{\text{Single Points}} \times T_{\text{Single Point}}$$

How much does dynamics cost? tinyurl.com/dyncost
How many trajectories should we run? tinyurl.com/trajs

Dynamics may be expensive

$$T_{total} \approx N_{\text{Trajectories}} \times \left(\frac{\tau_{\text{chem process}}}{\Delta\tau} \right) \times T_{\text{Single Point}}$$

$N_{\text{Trajectories}}$	= 100 trajectories
$T_{\text{Single Point}}$	= 6 min = 0.1 CPUh
$\tau_{\text{chem process}}$	= 500,000 fs = 0.5 ns
$\Delta\tau$	= 0.5 fs

$$T_{total} \approx 10 \text{ MCPUh}$$

Price 1 CPUh = 0.02 € (France)

Price 10 MCPUh = 200 k€

How much does dynamics cost? tinyurl.com/dyncost
How many trajectories should we run? tinyurl.com/trajs

Dynamics leaves a huge carbon footprint

1 CPUh @ 32 GB = 1.3 g CO_2e

10 MCPUh = 13 tCO_2e

11.5 tCO_2e/year

 → × 2

 → × 7

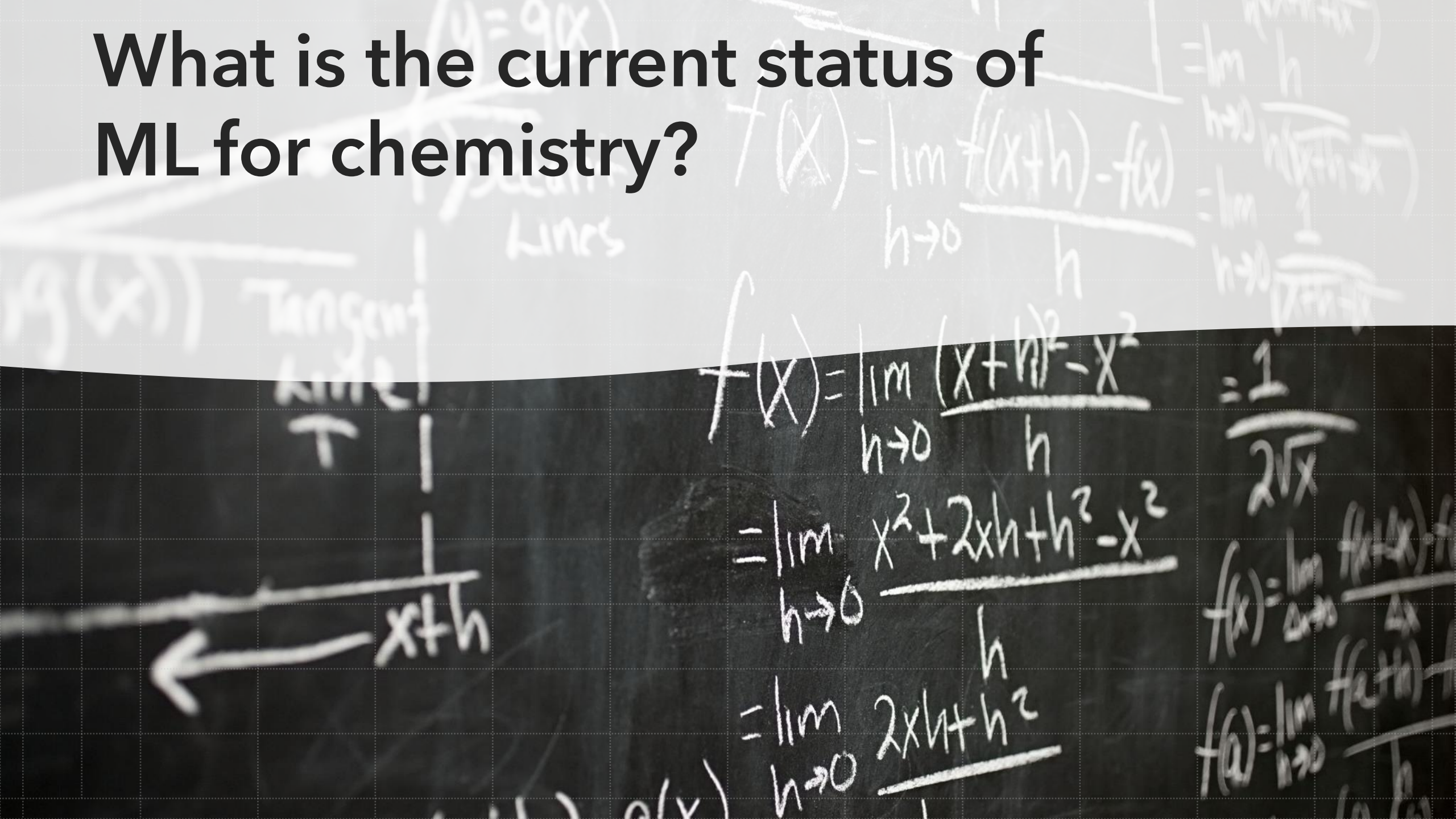
 → × 10

 → × 12

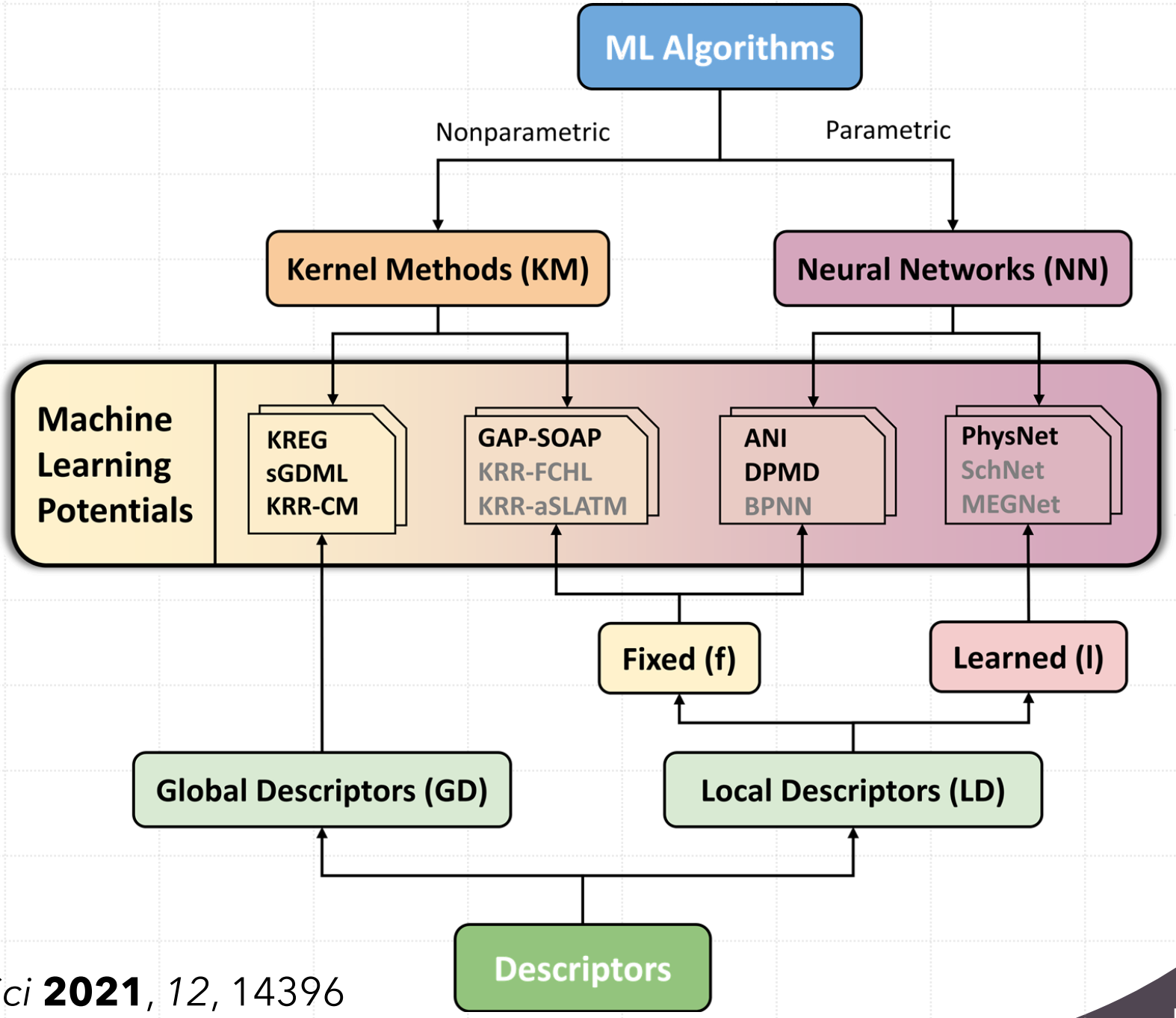
 → × 14

www.green-algorithms.org

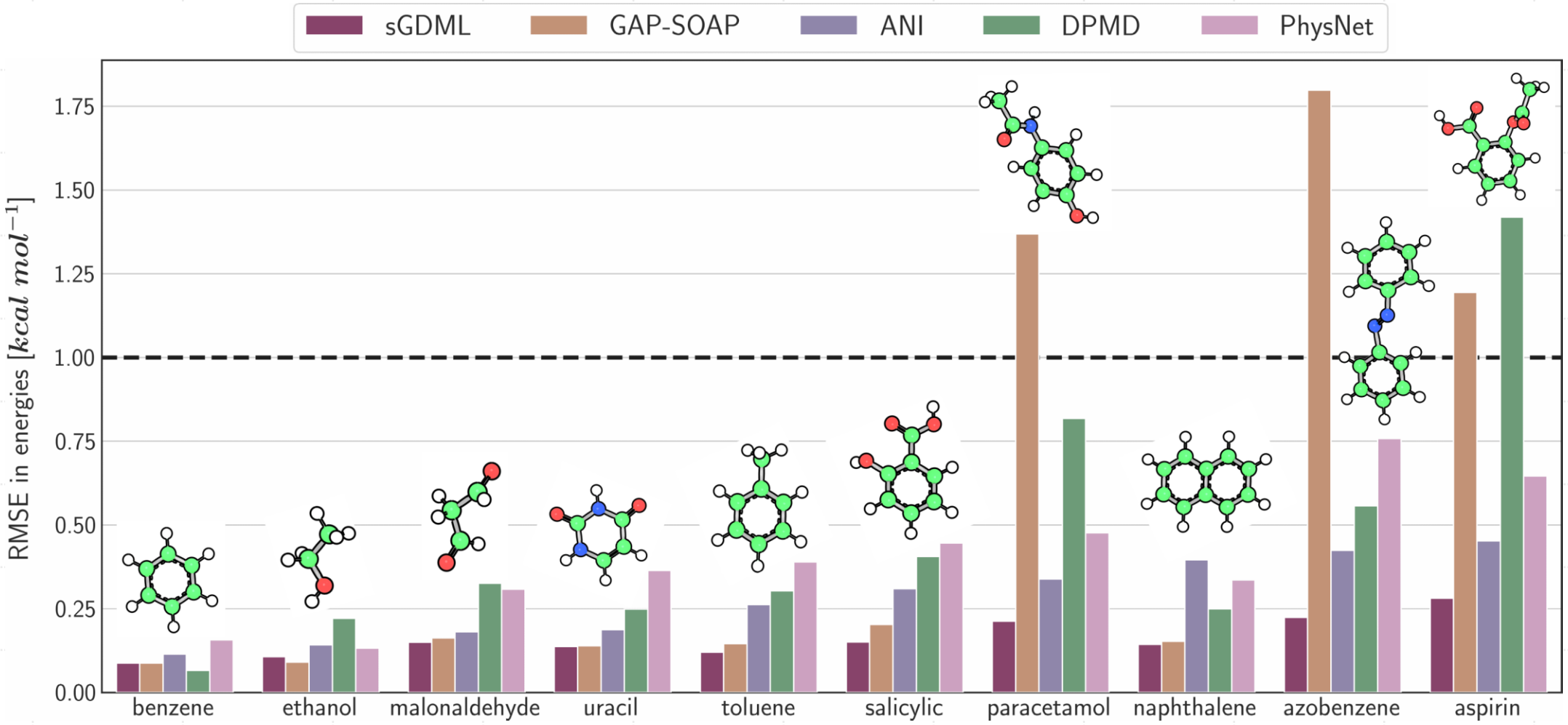
What is the current status of ML for chemistry?



The ground state

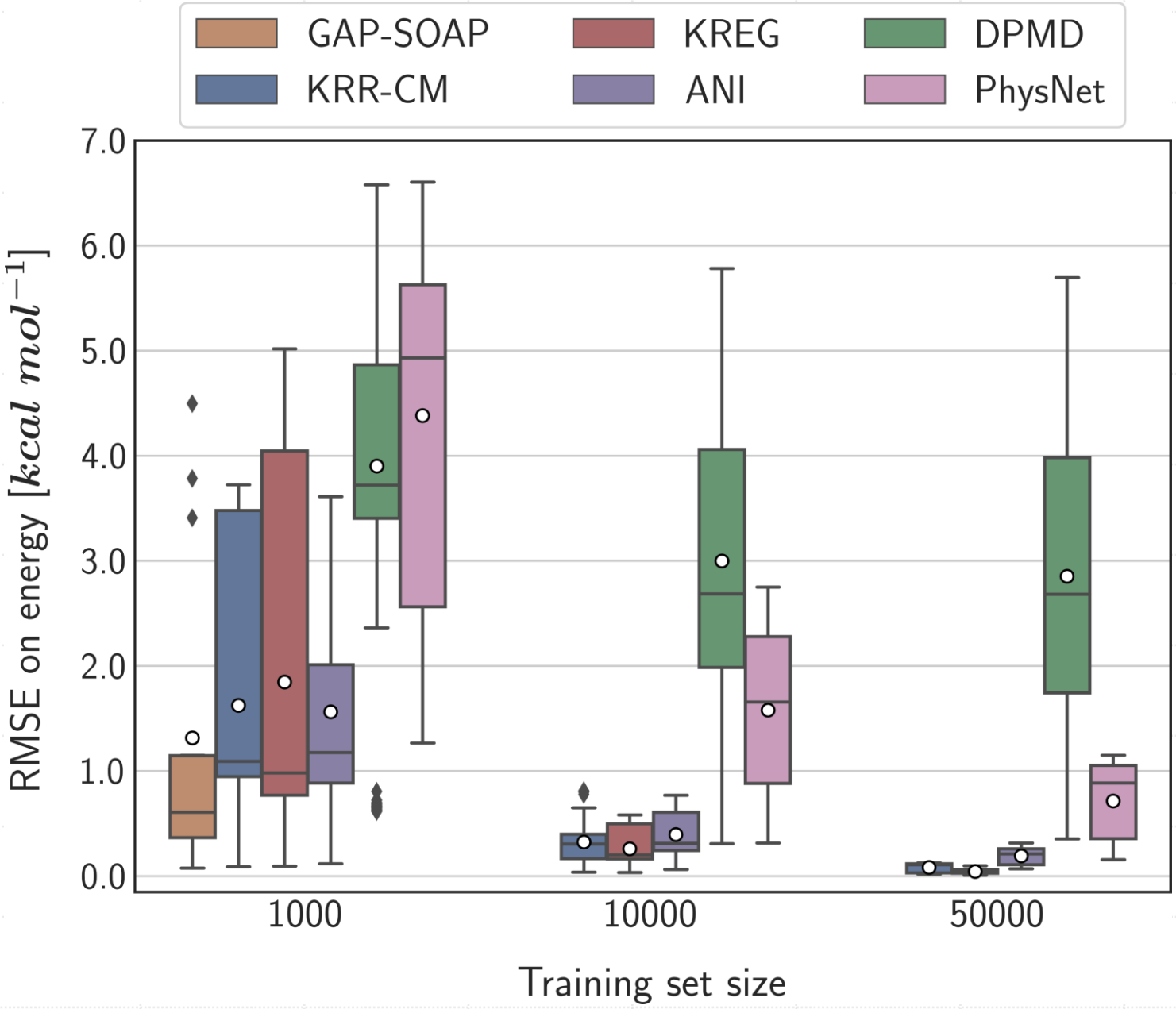


Pinheiro Jr *et al*. *Chem Sci* **2021**, *12*, 14396



- MD17 Database
- Energy + Force
- N_{train} = 1k; N_{model} = 20; N_{test} = 20k

Pinheiro Jr *et al*. *Chem Sci* **2021**, *12*, 14396



- MD17 Database
- Energy only
- N_{test} = 20k

Pinheiro Jr *et al*. *Chem Sci* **2021**, *12*, 14396

The excited states

Simulating excited states is much more challenging:

1. They usually correspond to electronic densities that are difficult to compute
2. They are strongly anharmonic
3. They cluster in state bundles, mixing with each other

ML can simulate excited states for the ground state equilibrium

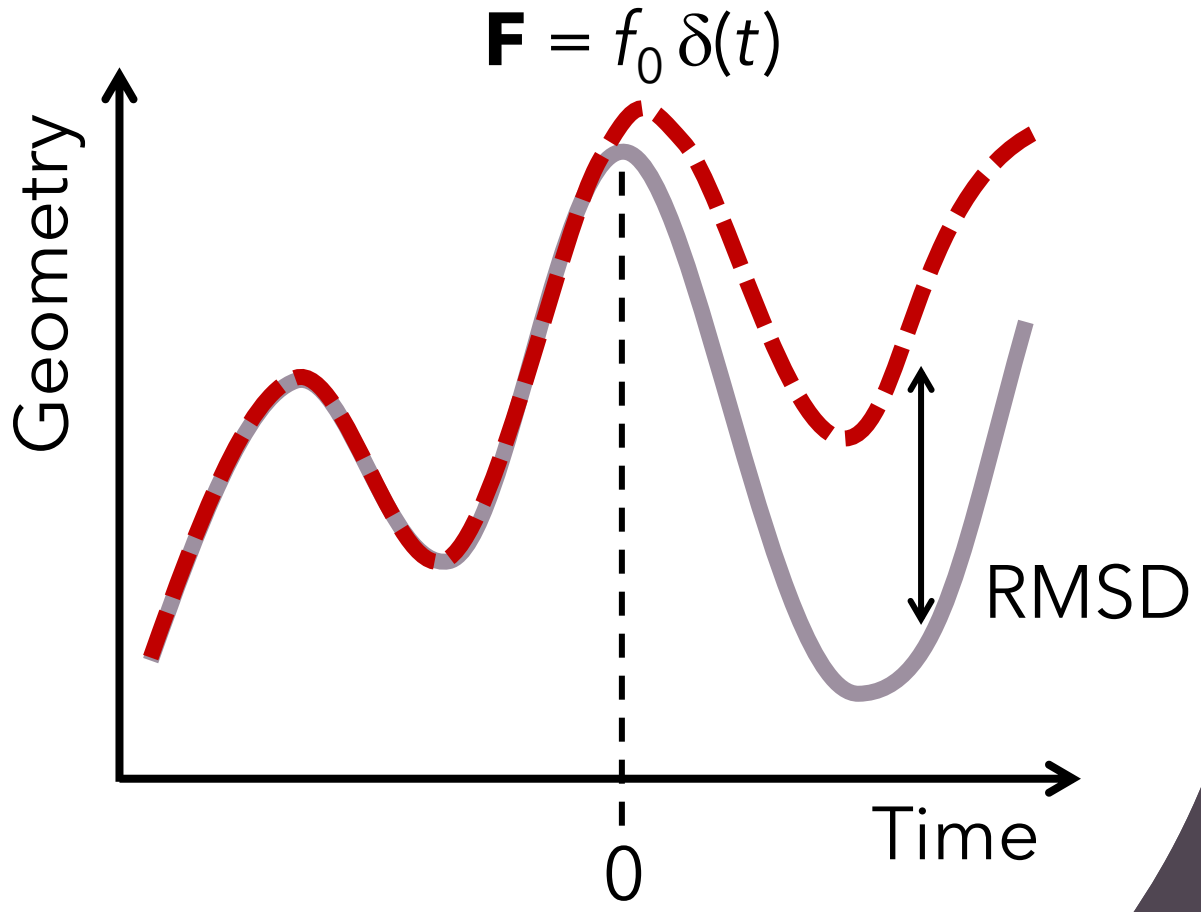
If a dataset spawning the $3N_{at}$-6 dimensions is available, ML can deliver excellent fittings

However, for sparse datasets, robust ML protocols are still missing

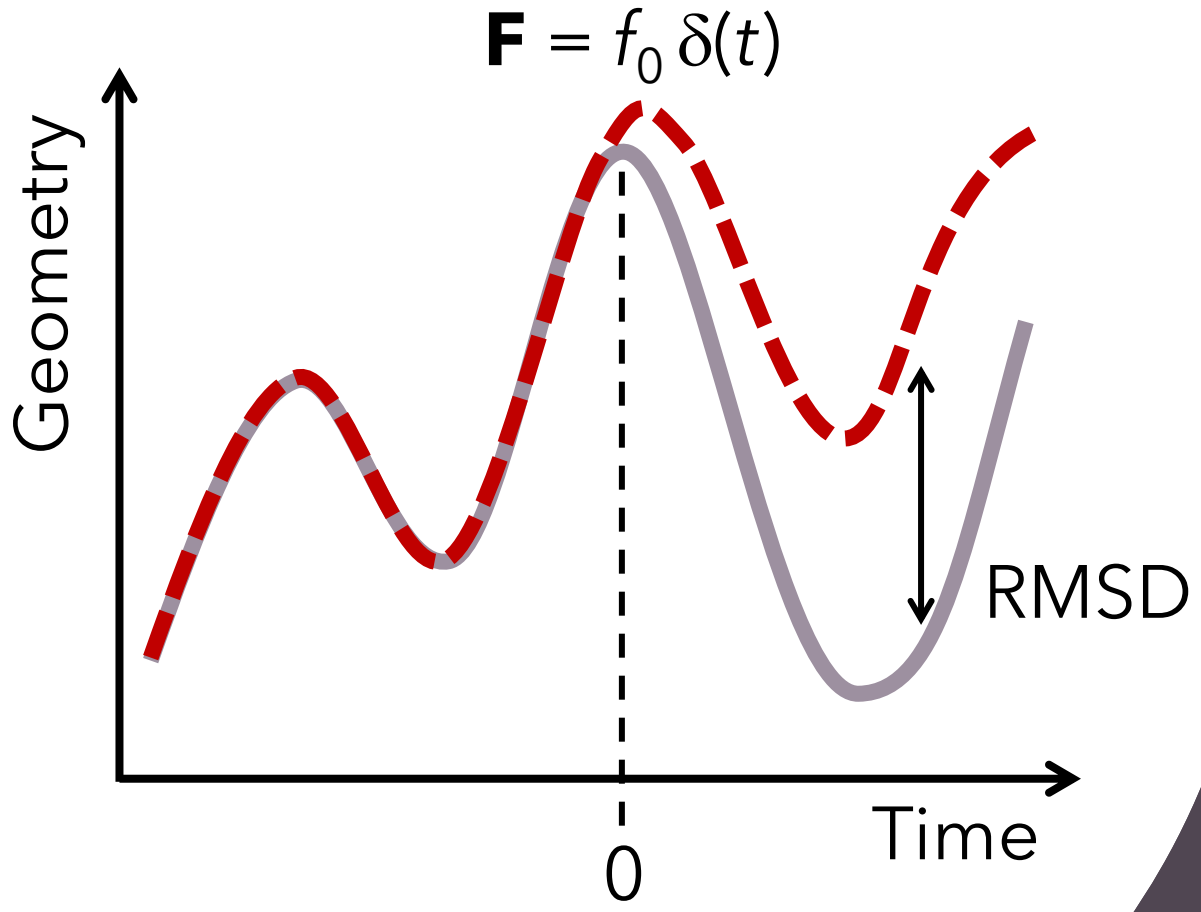
Matheus Bispo

The challenge

Effect of force uncertainty



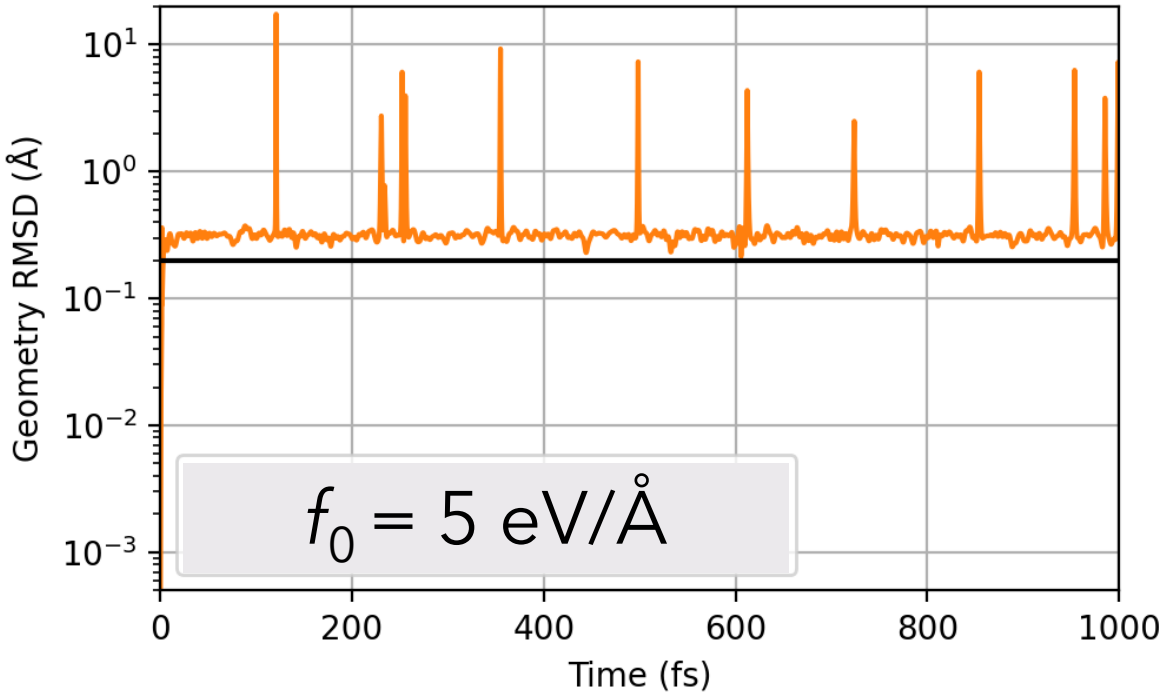
Effect of force uncertainty

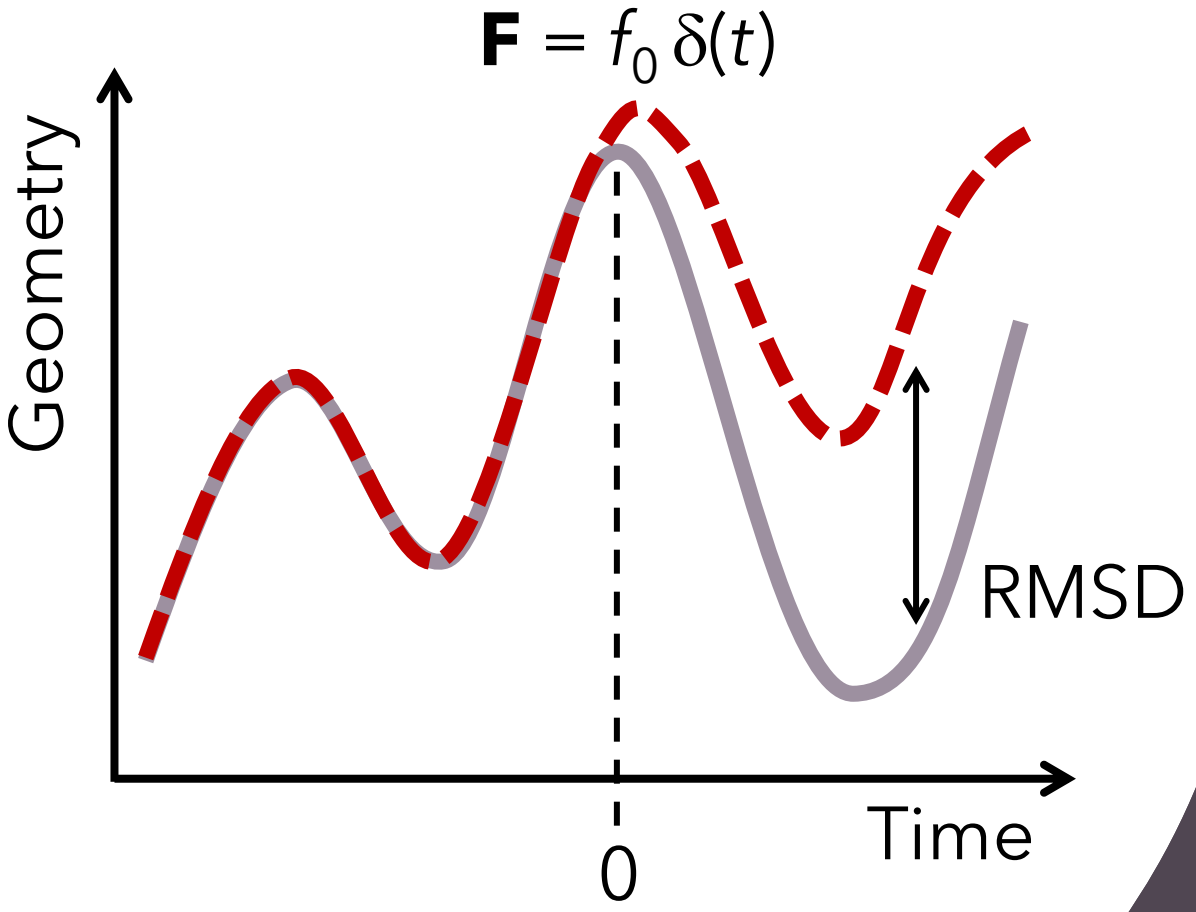


Geometrical accuracy
We want results better than 0.2 Å

- A-SBH 33D
- dynamics on E_1

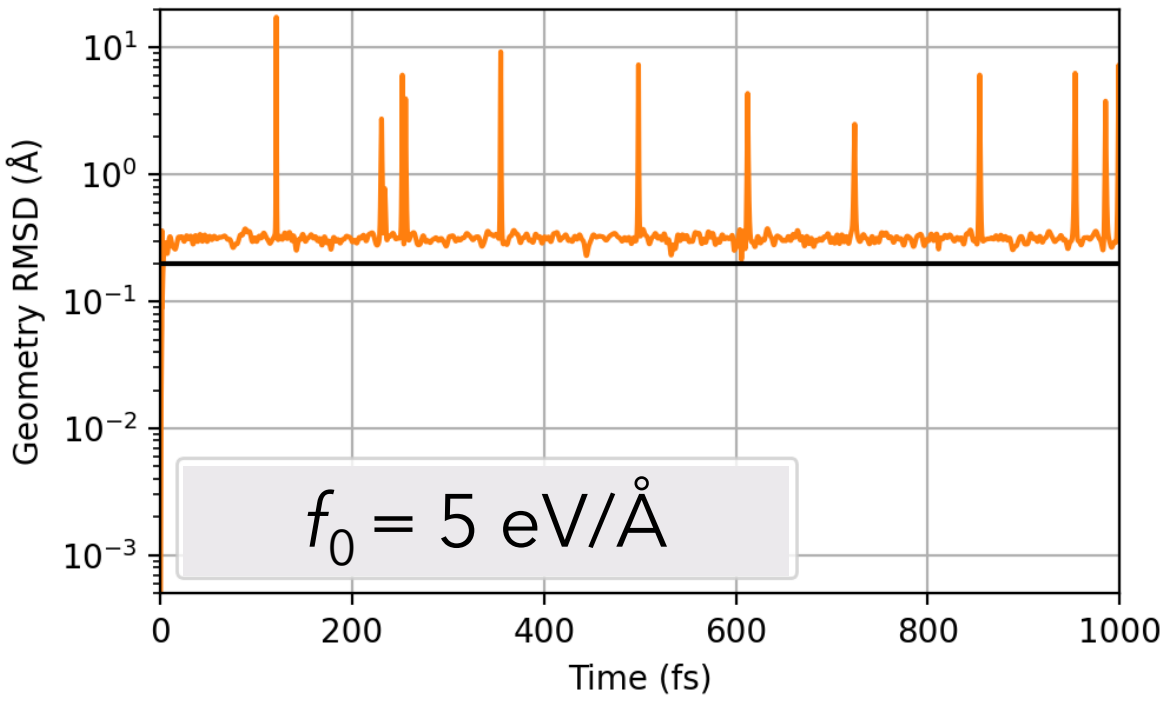
Effect of force uncertainty

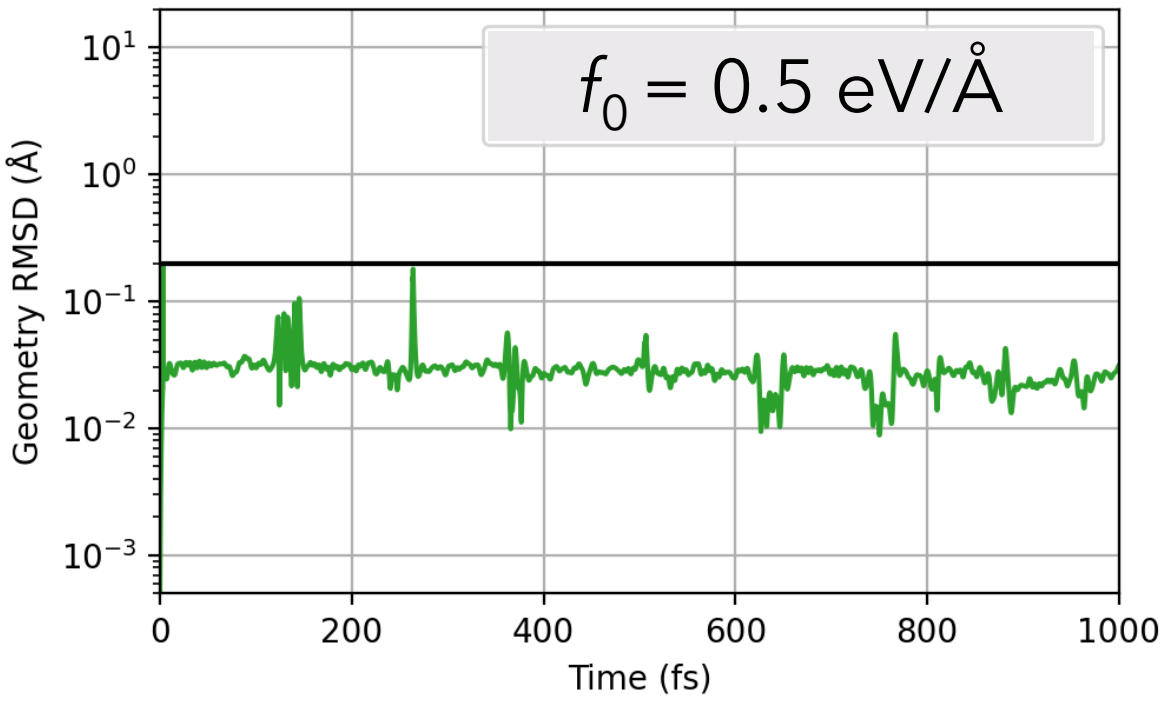


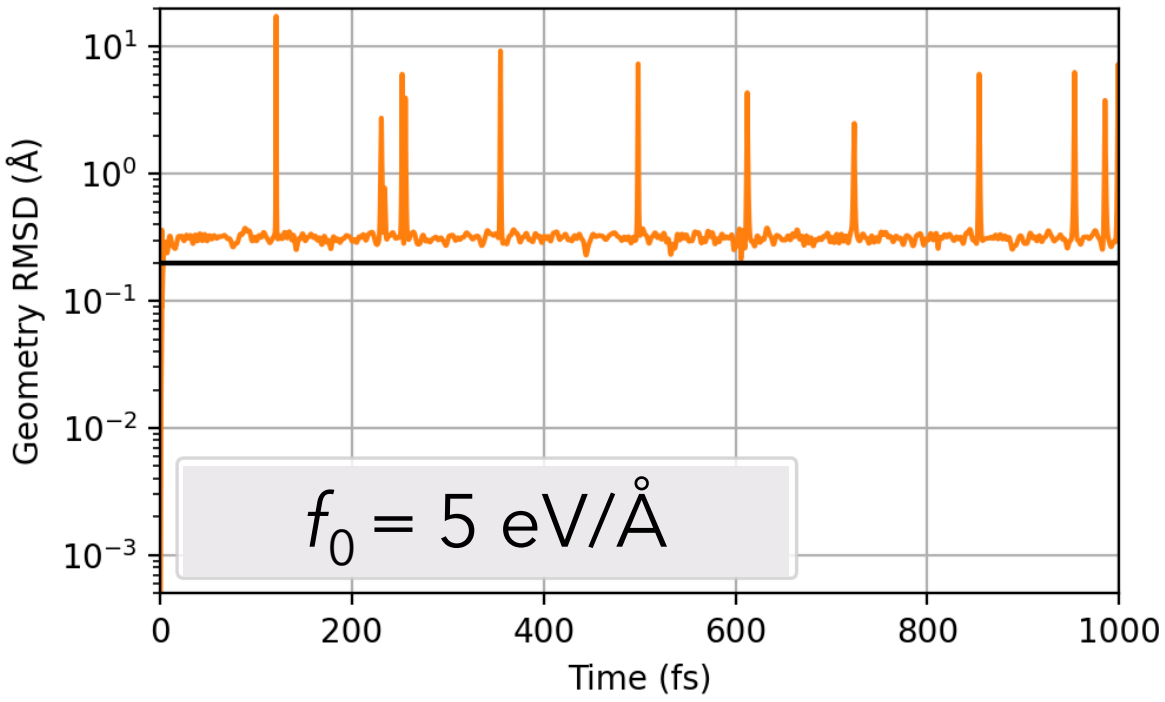


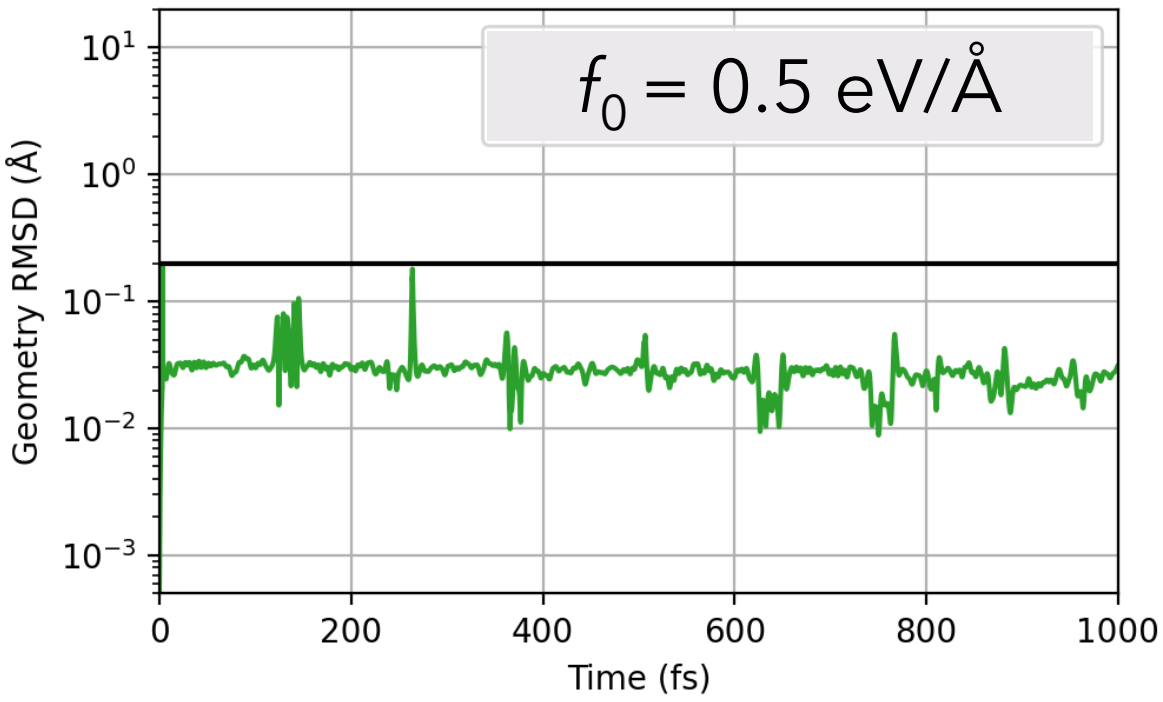
Geometrical accuracy
We want results better than 0.2 Å

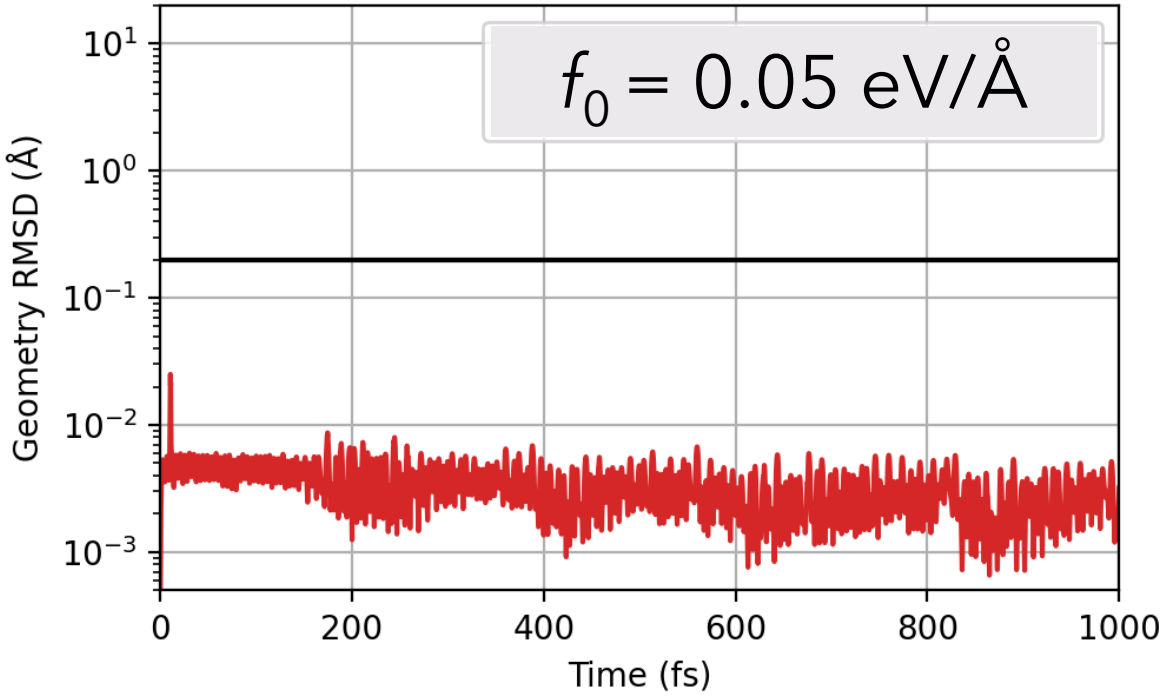
- A-SBH 33D
- dynamics on E_1

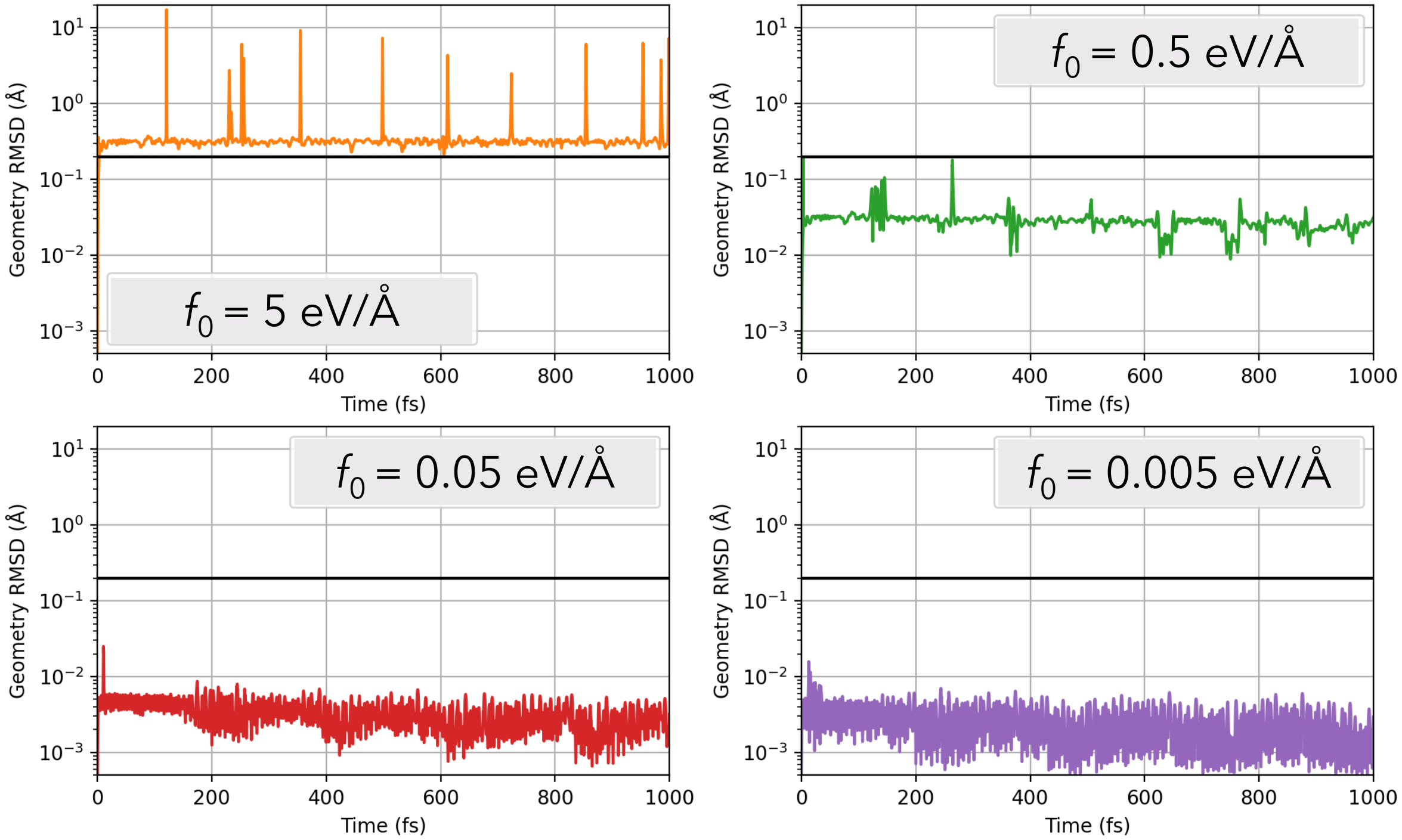

$f_0 = 5$ eV/Å
Geometry RMSD (Å)
Time (fs)
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
0
200
400
600
800
1000


$f_0 = 0.5$ eV/Å
Geometry RMSD (Å)
Time (fs)
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
0
200
400
600
800
1000


$f_0 = 5$ eV/Å
Geometry RMSD (Å)
Time (fs)


$f_0 = 0.5$ eV/Å
Geometry RMSD (Å)
Time (fs)


$f_0 = 0.05$ eV/Å
Geometry RMSD (Å)
Time (fs)

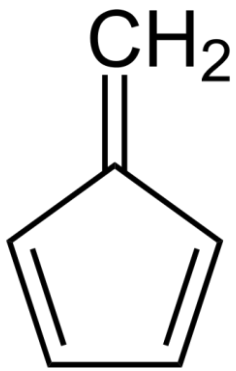

f_0 = 5 eV/Å
f_0 = 0.5 eV/Å
f_0 = 0.05 eV/Å
f_0 = 0.005 eV/Å
Geometry RMSD (Å)
Time (fs)
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
0
200
400
600
800
1000

We must predict forces better than
0.5 eV/Å (0.001 Hartree/Bohr)

(Maximum absolute error)

Max Pinheiro Jr

ML-NAMD test cases



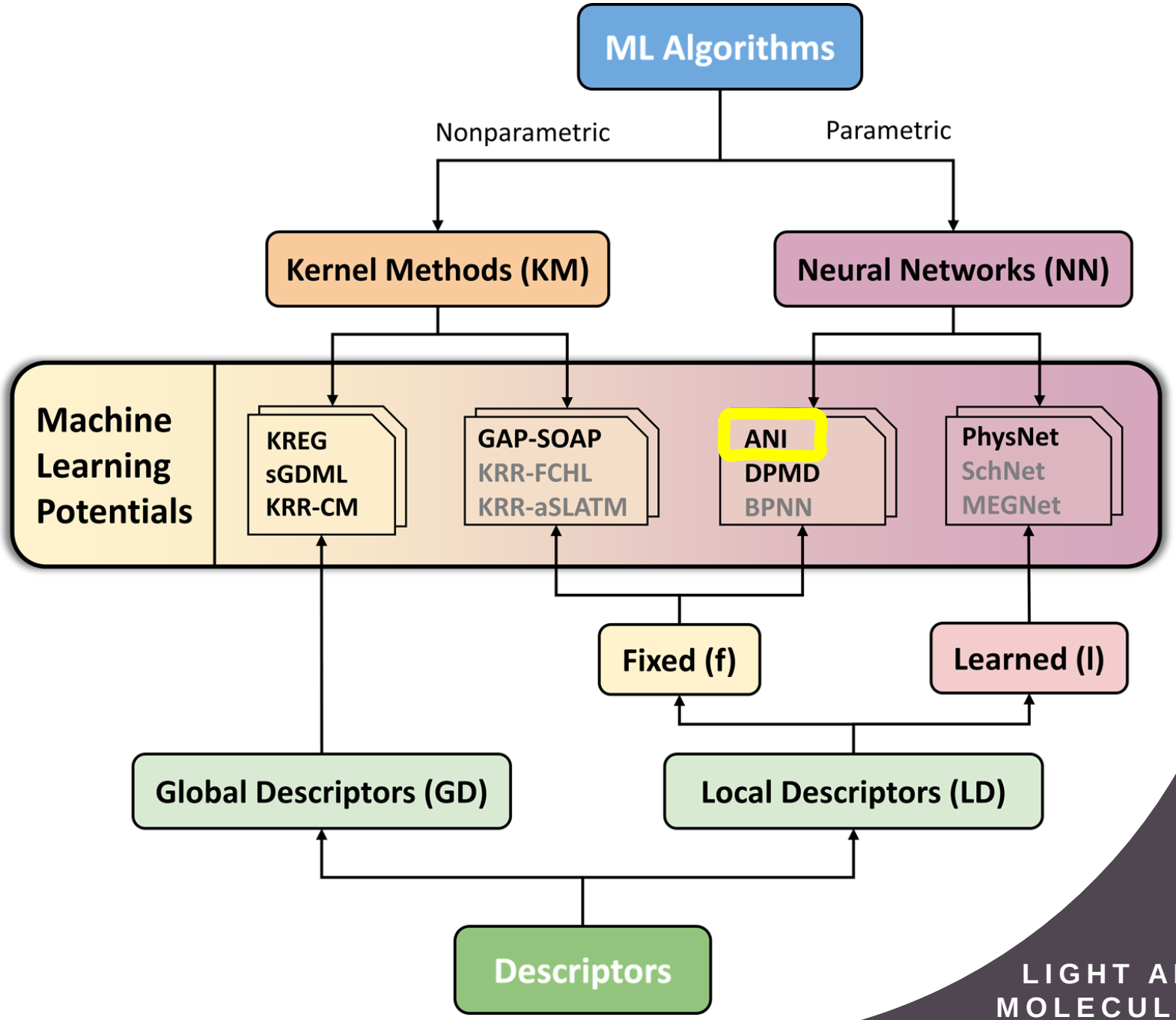
Fulvene test

Dataset:

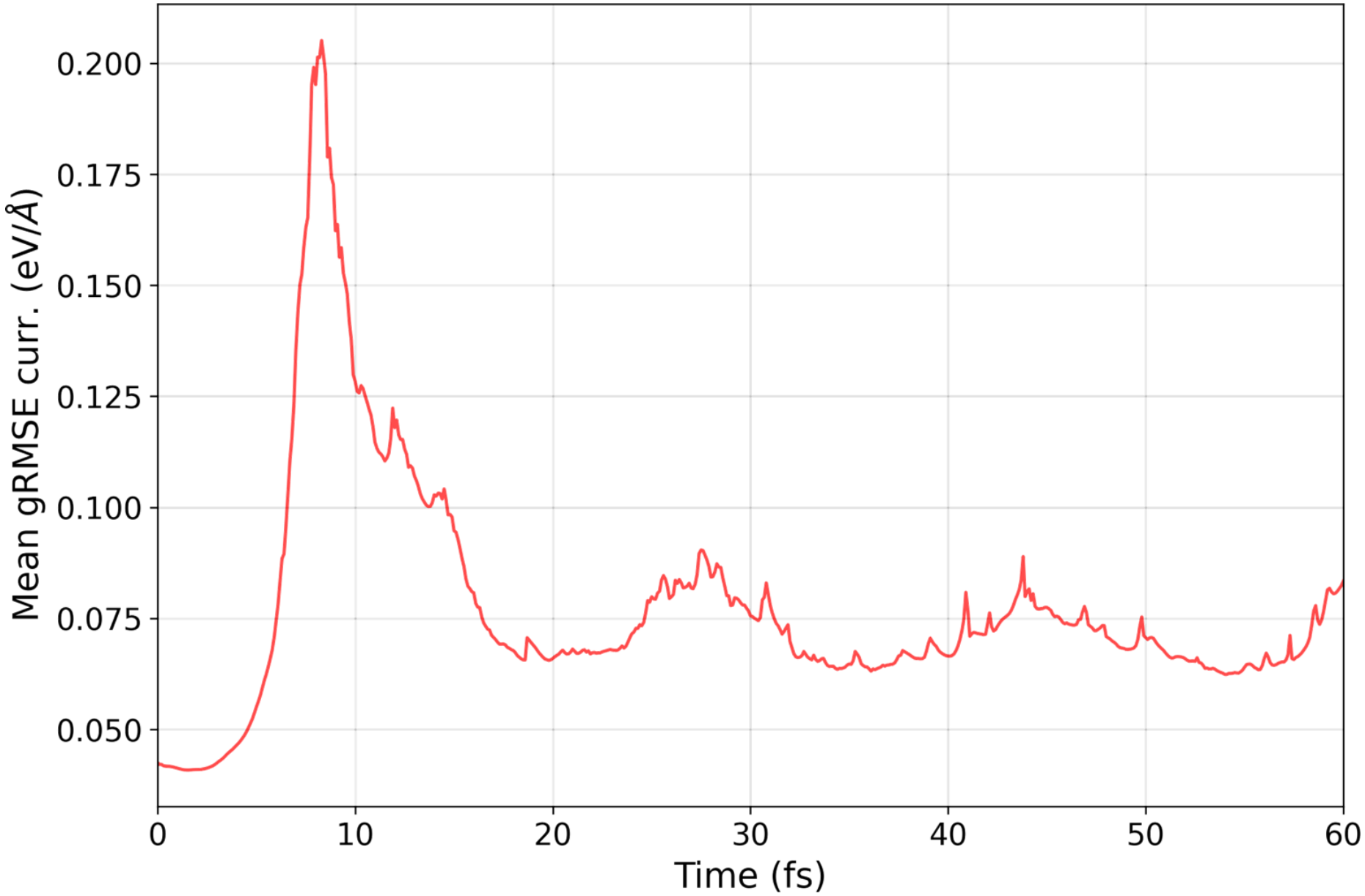
- DC-FSSH / CASSCF
- 200 trajs; 2 states
- t_{max} = 60 fs; Δt = 0.1 fs

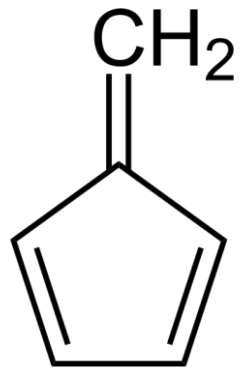
ML potential:

- ANI on 40k
- energy + forces



Gradient (current state)

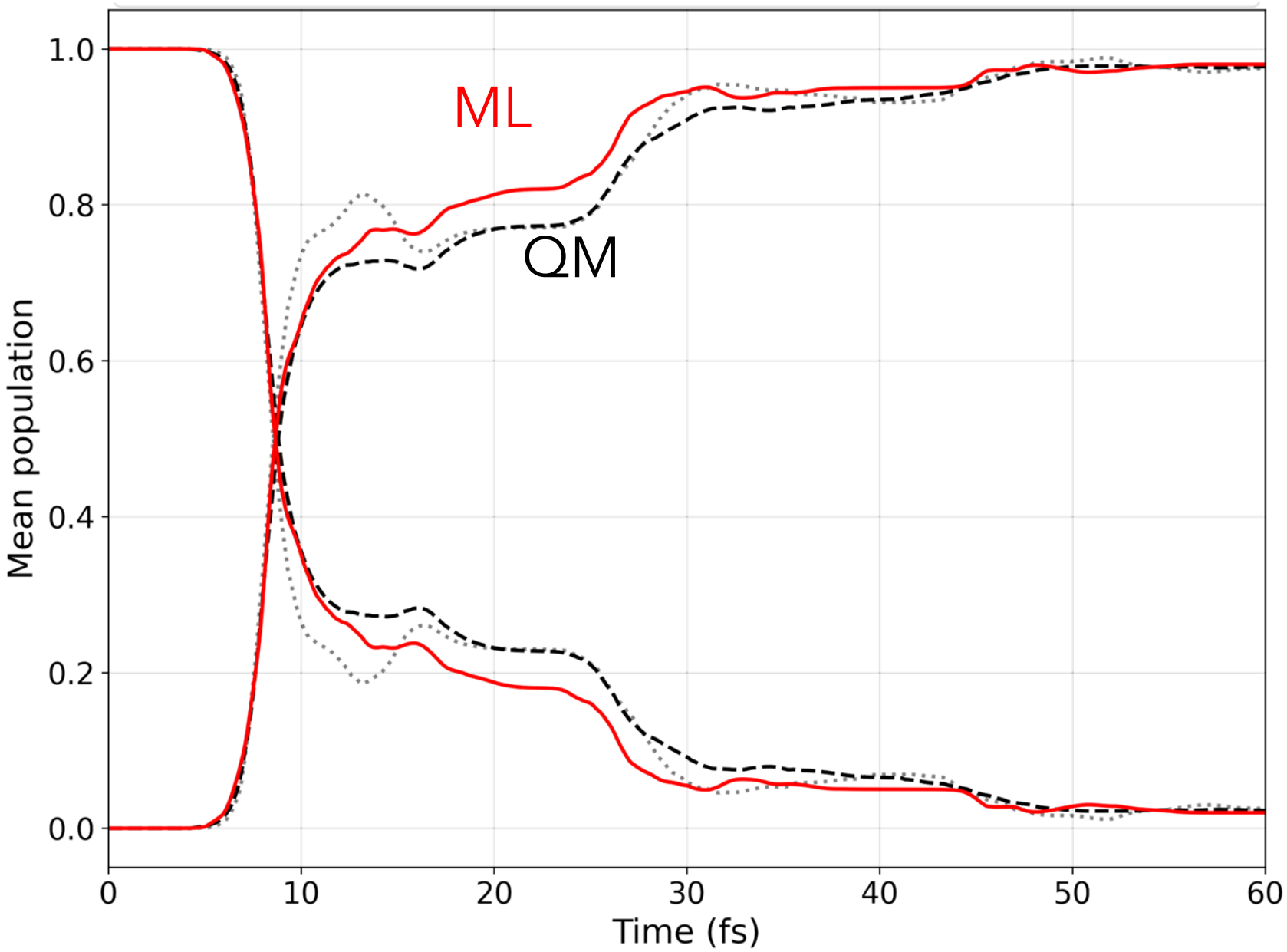


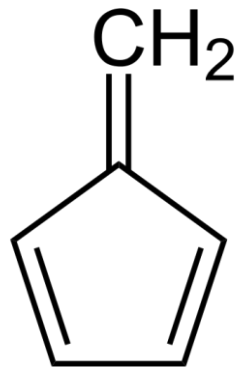


fulvene

Error for the gradient norm (current state)

Population

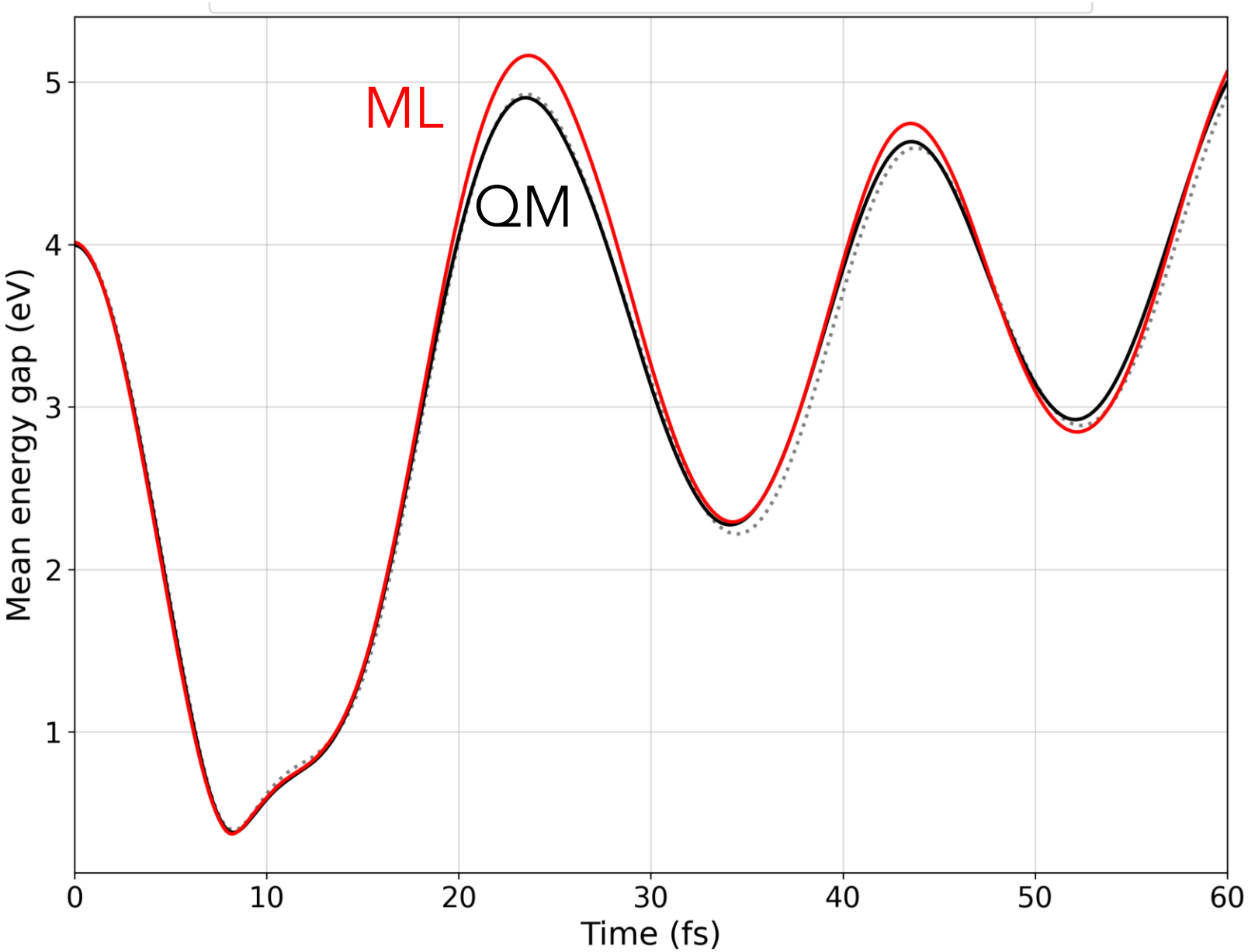


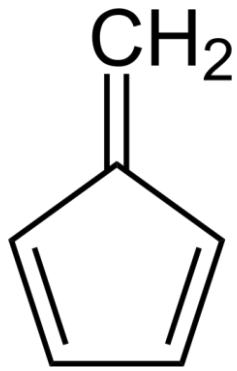


fulvene

Population averaged over 100 new trajectories

Energy gap

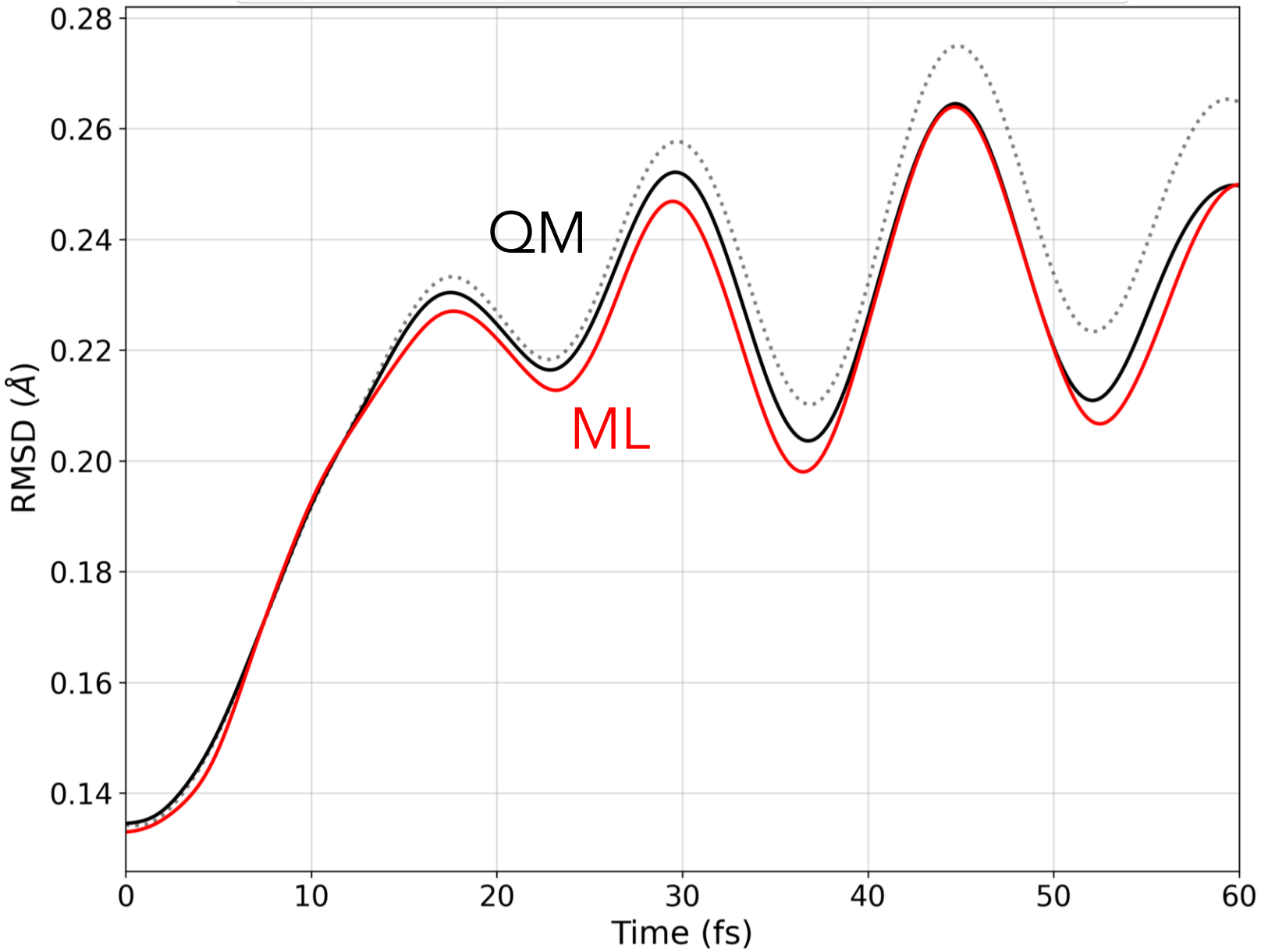


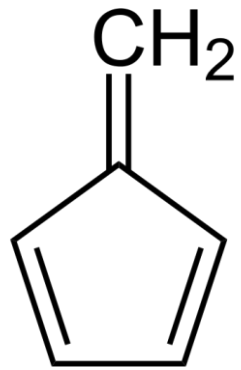


fulvene

ΔE_{10} evolution averaged over 100 new trajectories

Geometry

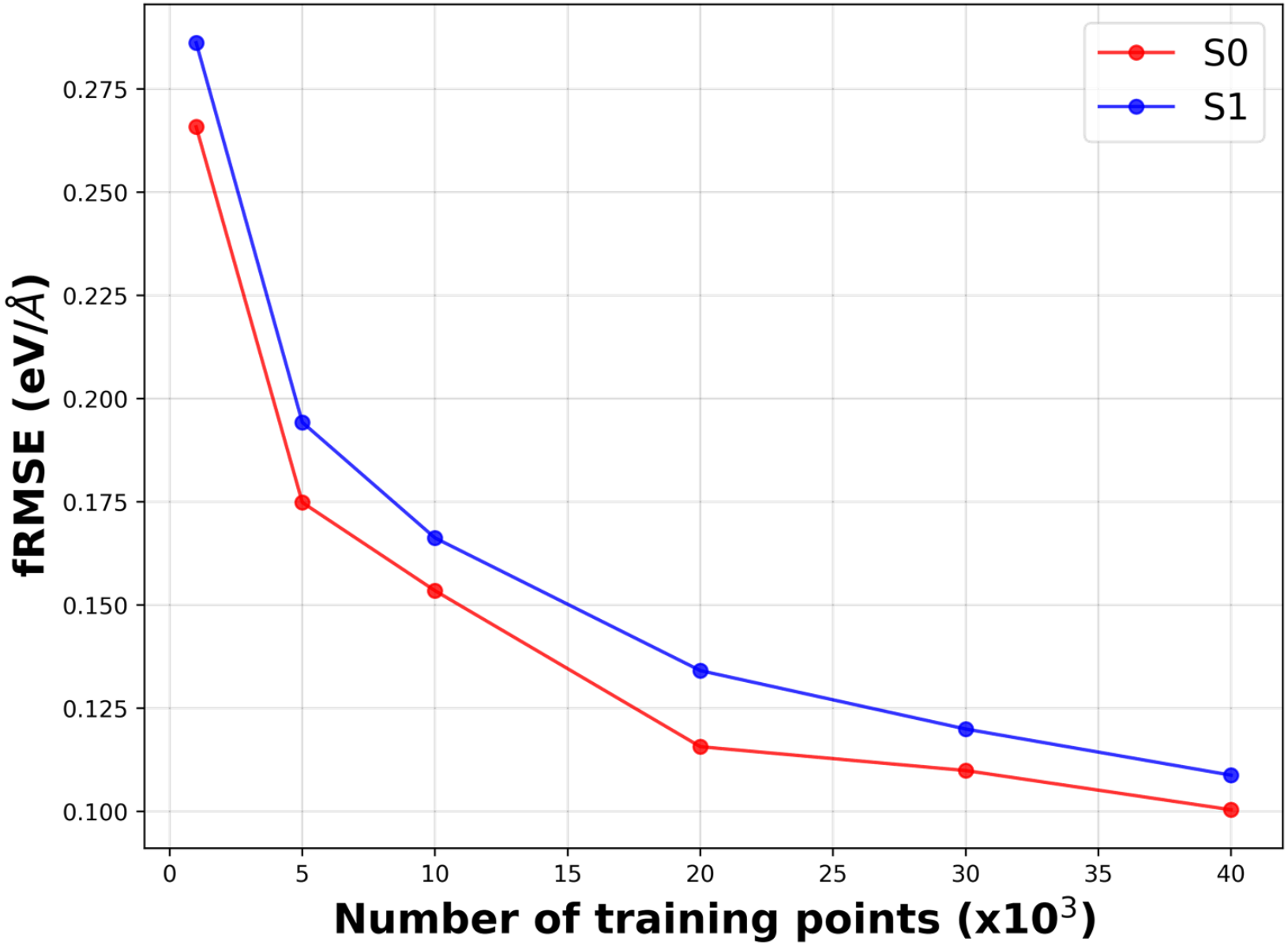




fulvene

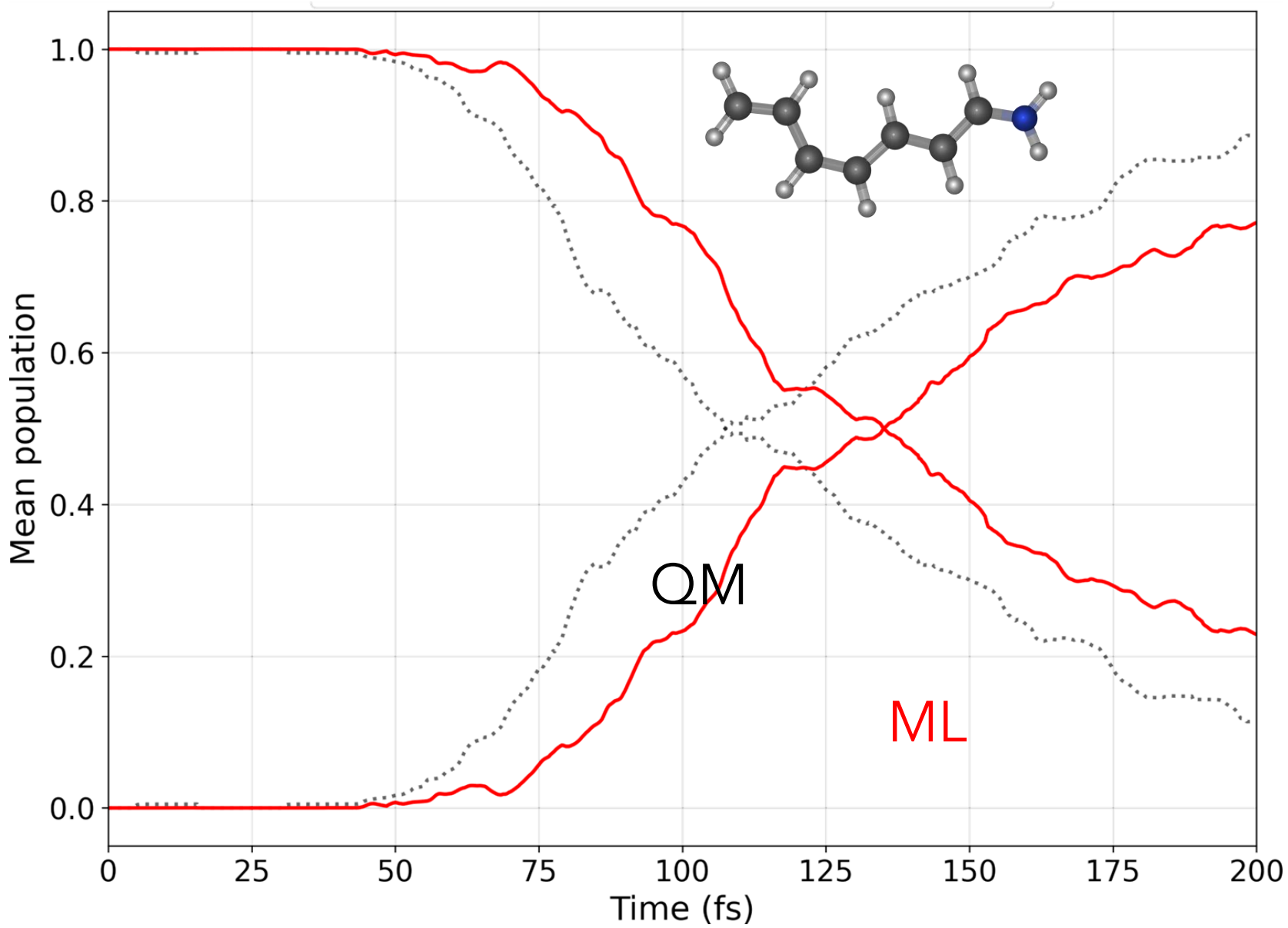
Geometry evolution averaged over 100 new trajectories

Problem 1: Training set size



40k points are too expensive to adopt as routine protocol

Problem 2: Lack of robustness



Using the same ML model with the same training set size for another molecule may not give adequate results

To address both problems, we are testing **active learning**:

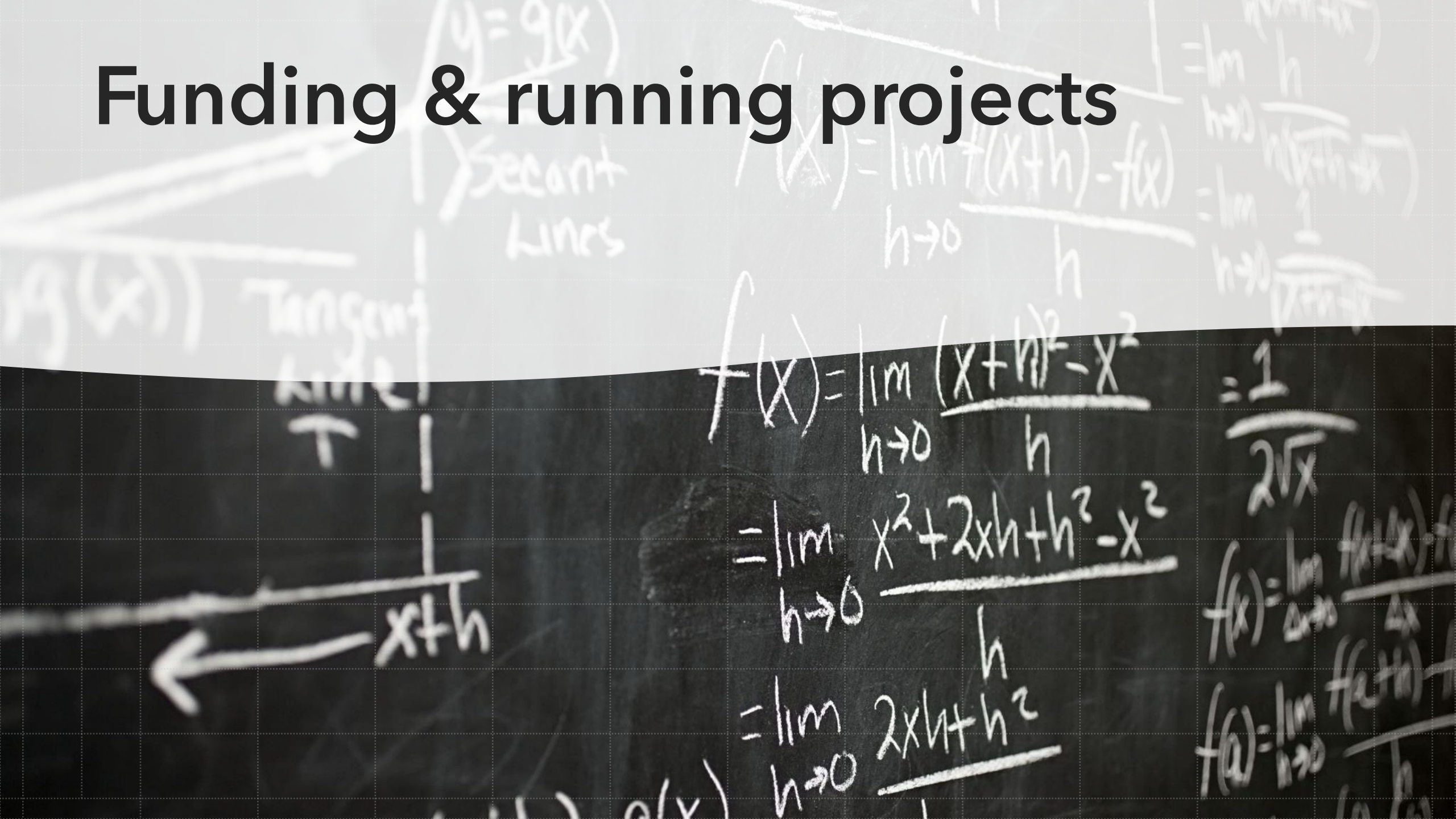
1. We train multiple machines with an initially small number of points

2. Then, we run dynamics until the prediction between machines diverges

3. The geometries showing divergence are included in the training set, and the machine is retrained

4. Repeat the procedure until no divergence is observed

Funding & running projects



Sep 19 — Aug 24

Subnano (ERC AdG)

Goal: ML potentials for dynamics in long timescale

- 72 months postdoc
- 1 PhD

Apr 24 — Mar 25

Sony (Award)

Goal: ML potentials NAMD of large chromophores

- 12 months postdoc

Jan 24 — Dec 27

MLChem (A*Midex 2022 Res. & Training)

Goal: Establish an international AI center for chemistry at AMU

- 3x36 months postdoc
- 1 PhD

A*Midex MLChem

Aix*Marseille université
Socialement engagée

ICR
Institut Chimie Radicalaire

CHIMIE Théorique

- Barbatti
- Ferré

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES
UMR 7020

QARMA

- Artières
- Kadri

Center of AI for Chemistry

厦門大學化學化工學院
College of Chemistry and Chemical Engineering, Xiamen University

- Dral

XACS

MLChem aims to establish an **international AI center for chemistry** at AMU

MLChem mission 1: ML showcase

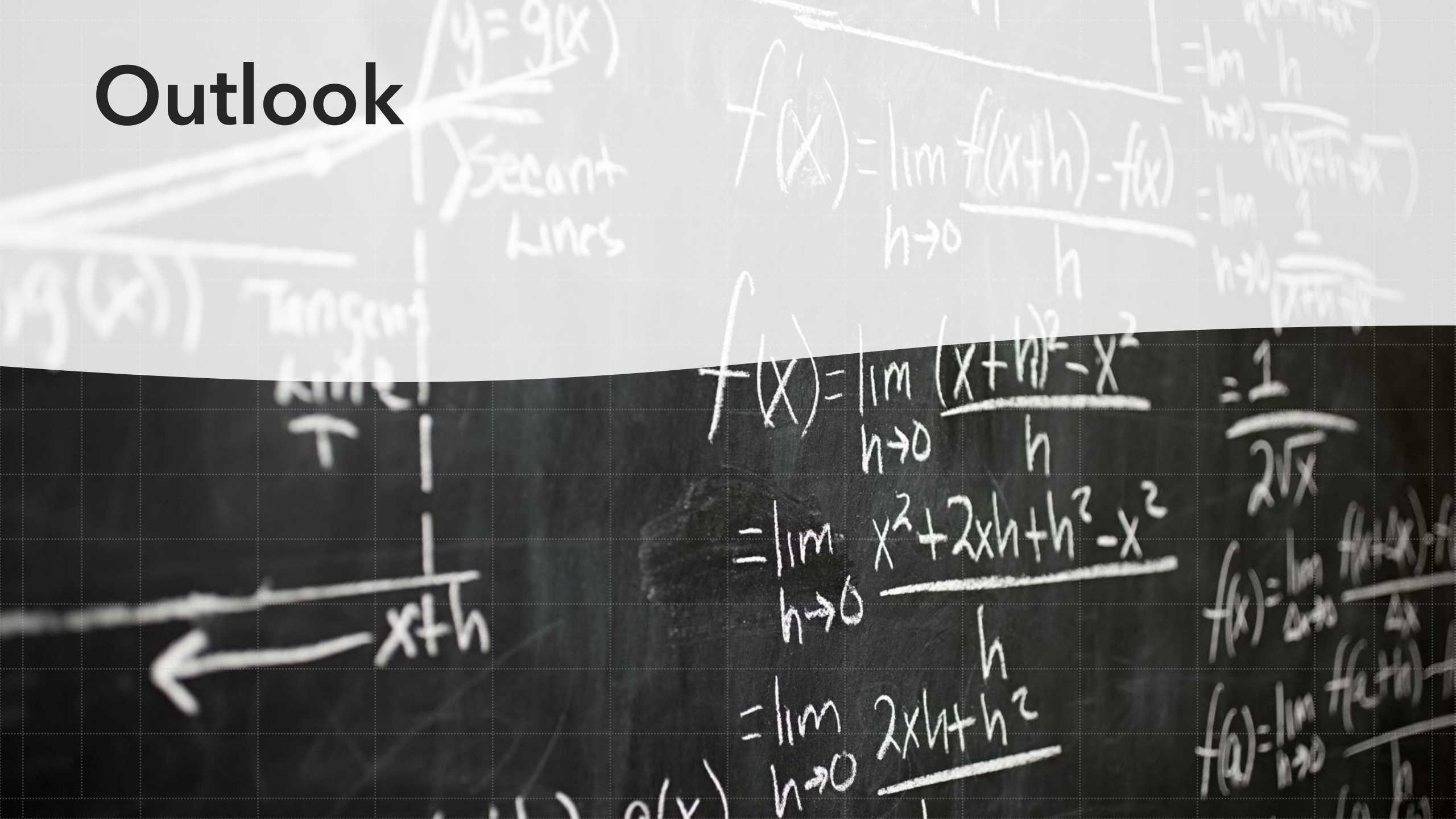
- Developing new methodologies for computational chemistry using ML for energy and charge propagation in organic crystals
- Implementing unsupervised-learning approaches for analyzing computational chemistry data
- Creating open data protocols for sharing ML results in a dedicated system server

MLChem mission 2: AI culture

- Offering project-tailored consulting on using AI solutions for chemistry groups at AMU
  1. help to identify preexistent software and hardware available
  2. help writing grant proposals to secure funds to implement those solutions.

- Offering courses and tutorials on ML for chemistry at several levels
  1. Two virtual ML schools
  2. Courses for ED250
  3. ML at master programs

Outlook



Supervised ML has great potential for accelerating excited-state potential energy predictions.

However, no robust protocol has been published
(If the model and training procedure worked for a molecule, it does not mean they will also work for another one)

Problems boil down to the high accuracy required in the predictions in highly multidimensional spaces

Our experience developing methods, benchmarks, and programs for ML for theoretical chemistry can be a seed for creating a culture of AI for chemistry at AMU

Our main challenge may not be scientific.

It is our need for career attractiveness for young researchers.

SubNanO

erc

A*Midex
Initiative d'excellence Aix-Marseille

Aix*Marseille université
Initiative d'excellence

institut universitaire de France

LIGHT AND MOLECULES

www.barbatti.org

mario.barbatti@univ-amu.fr

@MarioBarbatti