Basic notions: modelling material properties using machine learning

Bingqing Cheng

University of Cambridge
ICTP Virtual School, Jan 2021

Ab initio, first-principle: from the beginning

The Schrödinger equation (1926)

Erwin Schrödinger

1887 – 1961

$$E|\Phi\rangle = \hat{H}|\Phi\rangle$$

The holy grail of computational physics

"...the rest, is chemistry."

Paul Dirac, 1929

The holy grail of computational physics

"The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus **completely known**, and the difficulty lies only in the fact that application of these laws leads to equations that are **too complex to be solved**.
"

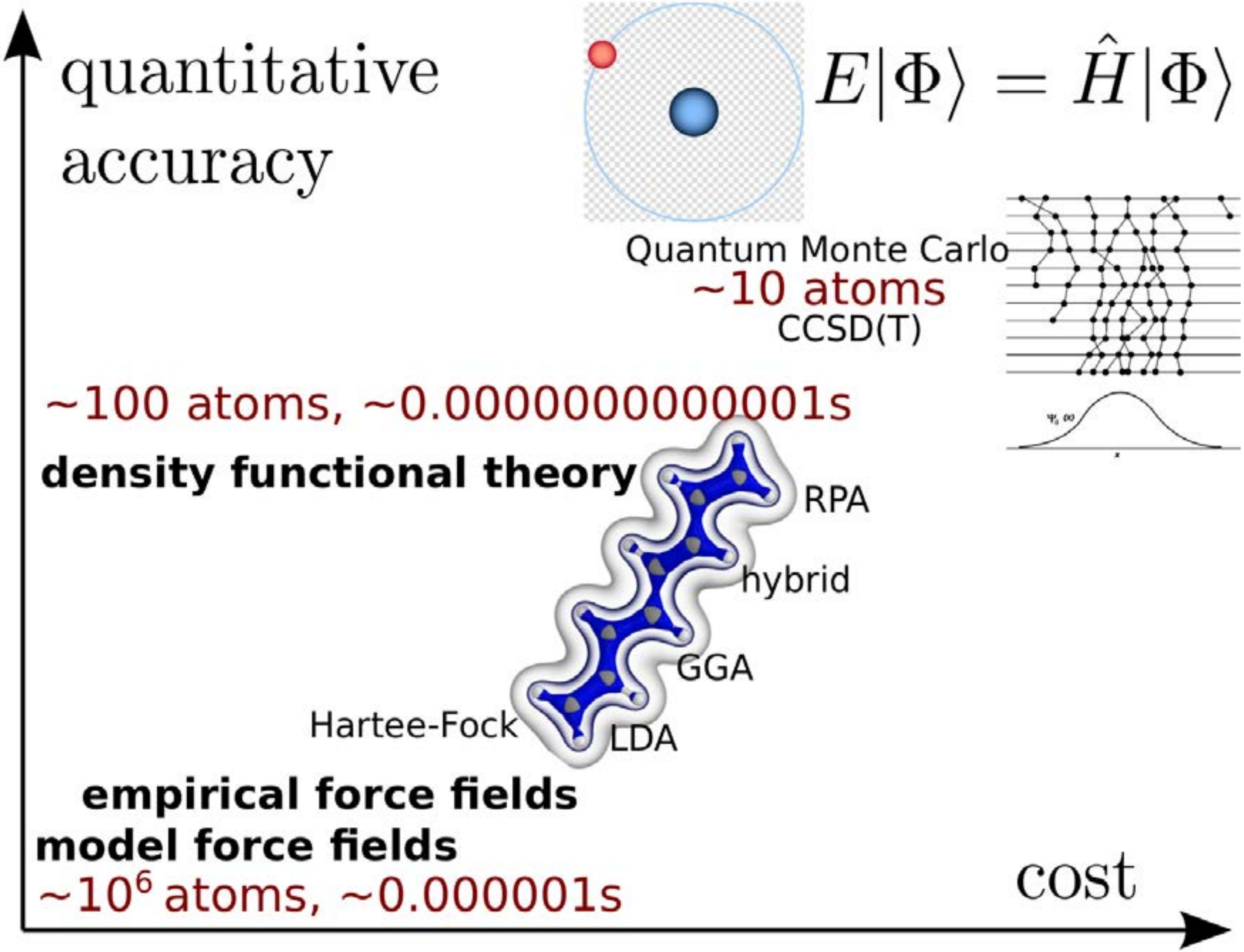
Paul Dirac, 1929

The holy grail of computational physics

"...**approximate practical methods** of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems **without too much computation**."

Paul Dirac, 1929

Trade-off between cost and accuracy



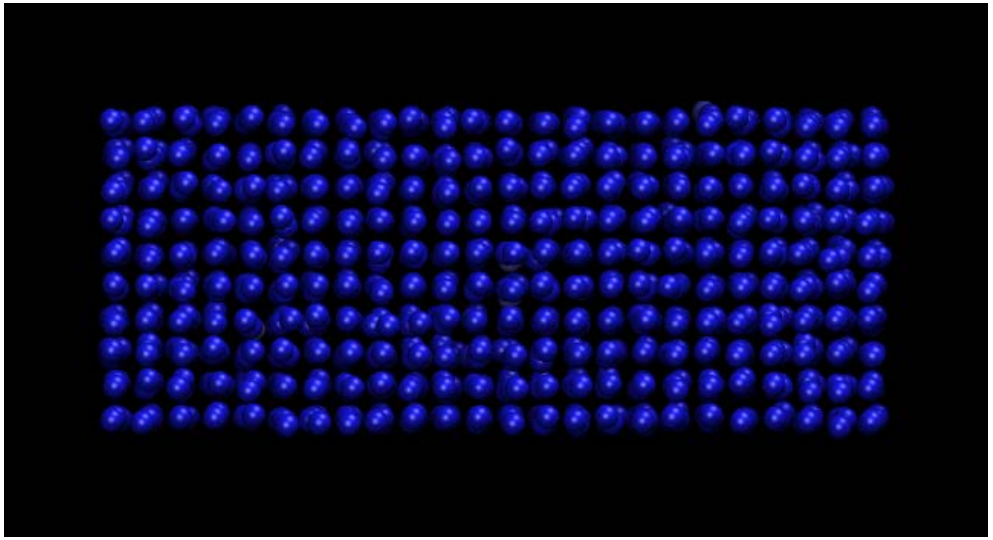
Outline

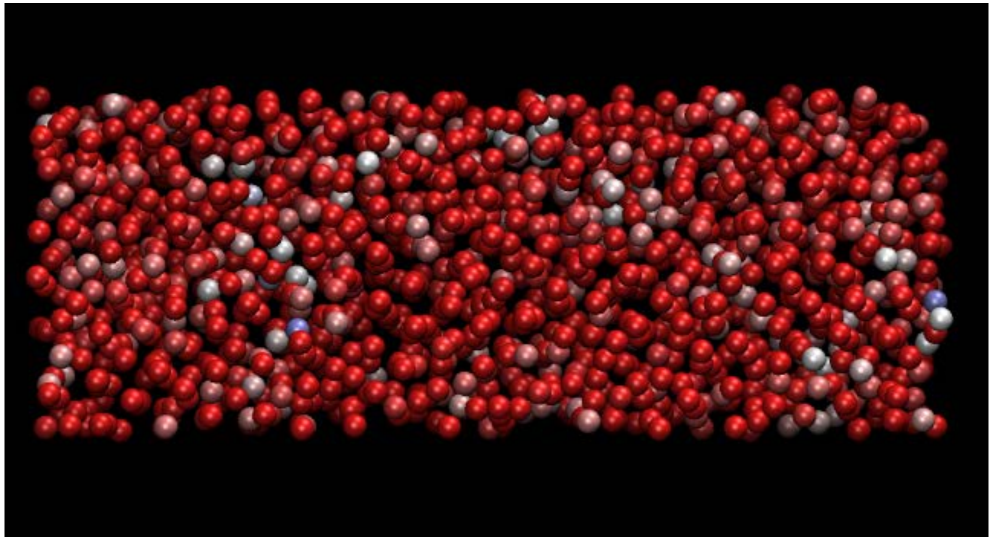
What we will talk about:

- Statistical mechanics & molecular dynamics 101.
  - Metadynamics
  - Thermodynamic integration
  - Nuclear quantum effects (NQEs)
- Translating materials and molecules into matrices.
  - Representations
  - Dimensionality reduction
- Introduction to machine learning potentials.

Thermodynamics

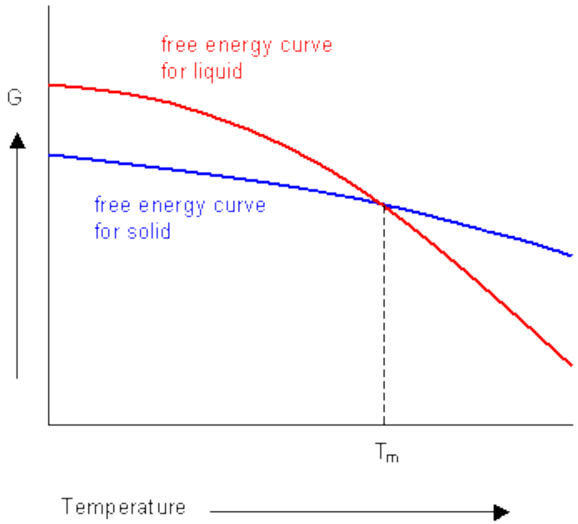
From thermodynamic point of view





$G_s(P,T) = H_s - TS_s$

$G_l(P,T) = H_l - TS_l$



From statistical mechanics point of view ...

Free energy is a measure of *probability*!

$$G_l(P,T) - G_s(P,T) = -(1/kT) ln(\frac{P_l}{P_s})$$

But you have to sum over all the microstates.

$$P_l = \sum_{\Omega \in liquid} P(\Omega)$$

From statistical mechanics point of view ...

Free energy is a measure of *probability*!

$$G_l(P,T) - G_s(P,T) = -(1/kT)ln(\frac{P_l}{P_s})$$

But you have to sum over all the microstates.

$$P_l = \sum_{\Omega \in liquid} P(\Omega)$$

Thermodynamics & statistical mechanics

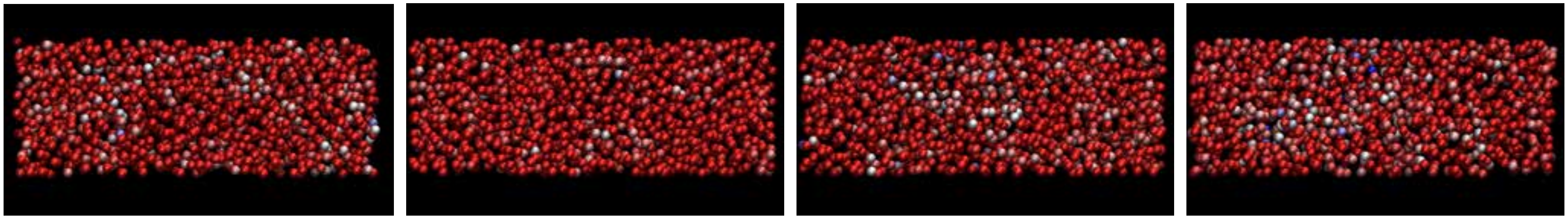
From statistical mechanics point of view ...

Free energy is a measure of *probability*!

$$G_l(P,T) - G_s(P,T) = -(1/kT)ln(\frac{P_l}{P_s})$$

But you have to sum over all the microstates.

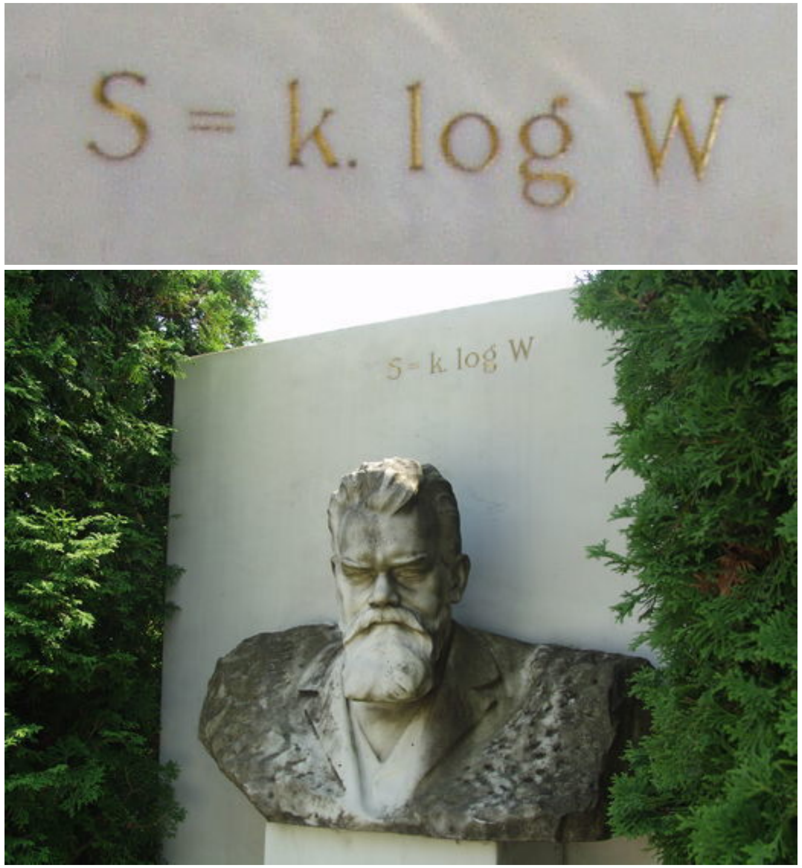
$$P_l = \sum_{\Omega \in liquid} P(\Omega)$$



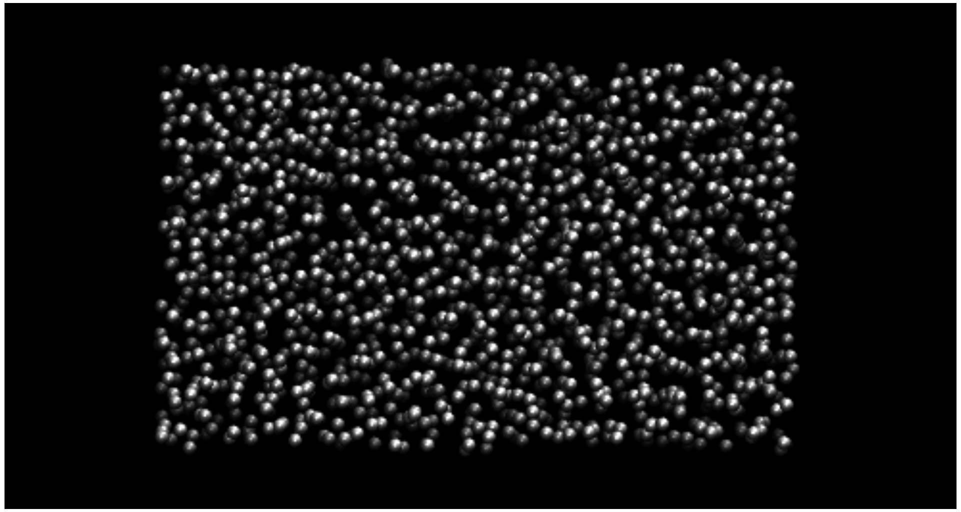
Statistical mechanics

$$P(\Omega) = e^{-\frac{H(\Omega)}{kT}}$$

$$S_I \approx k \sum_{\Omega} P(\Omega) \log(P(\Omega))$$



Ludwig Boltzmann



A microstate is a specific realization of the coordinates and velocities of all atoms in the system.

Monte Carlo sampling

The goal is to sample from: $P(\Omega) = e^{-\frac{H(\Omega)}{kT}}$

- A move is generated from Ω to Ω' with probability $P(\Omega \to \Omega')$.
- The probability distribution is consistent with $P(\Omega)$, if $P(\Omega)$ is invariant under the move, i.e.

$$\int dx P(x \to x') P(\Omega) = P(\Omega')$$

- A stronger condition: detailed balance

$$P(\Omega \to \Omega') P(\Omega) = P(\Omega' \to \Omega) P(\Omega')$$

- One possible option: Metropolis sampling

$$P(\Omega \to \Omega') = \alpha_{\Omega\Omega'} P(\Omega')/P(\Omega), \; \textit{if} \;\; P(\Omega') < P(\Omega)$$

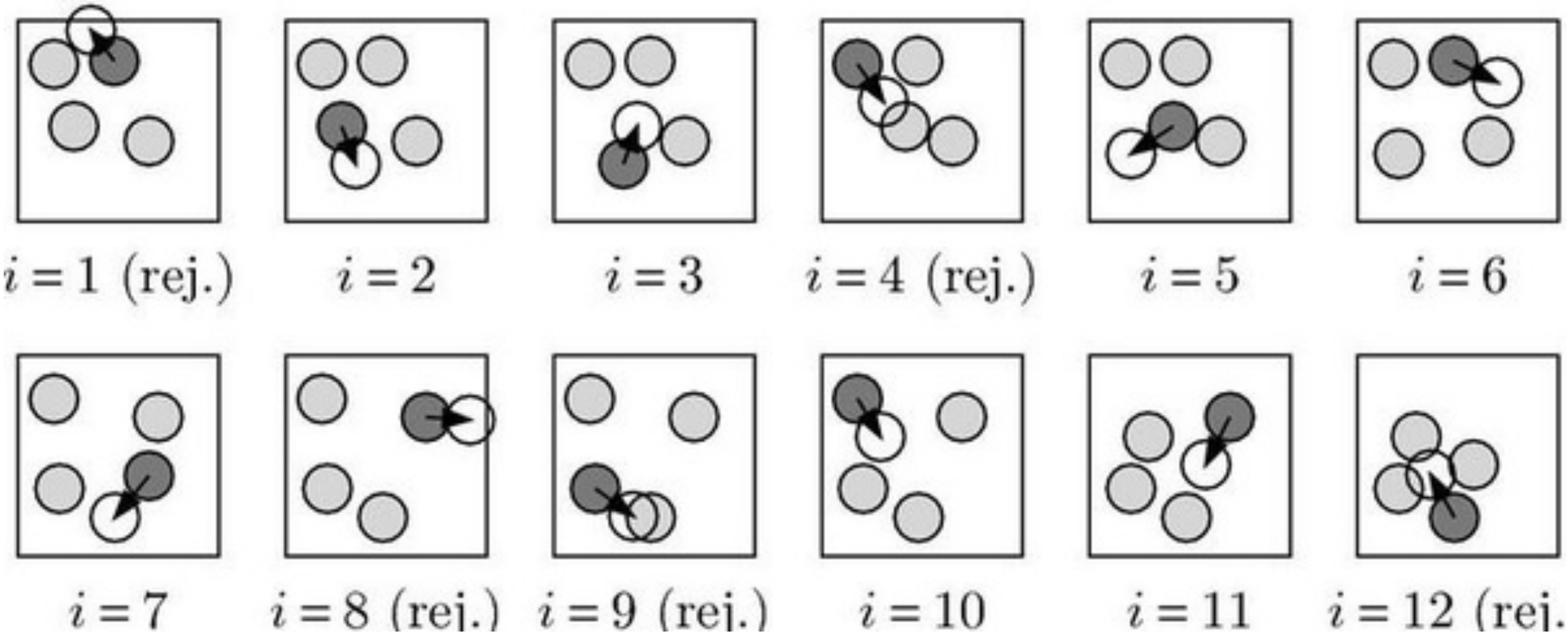
$$P(\Omega \to \Omega') = \alpha_{\Omega\Omega'}, \textit{otherwise}$$

Monte Carlo sampling

- One possible option: Metropolis sampling

$$P(\Omega \to \Omega') = \alpha_{\Omega\Omega'} P(\Omega')/P(\Omega), \ \textit{if} \ \ P(\Omega') < P(\Omega)$$

$$P(\Omega \to \Omega') = \alpha_{\Omega\Omega'}, \textit{otherwise}$$



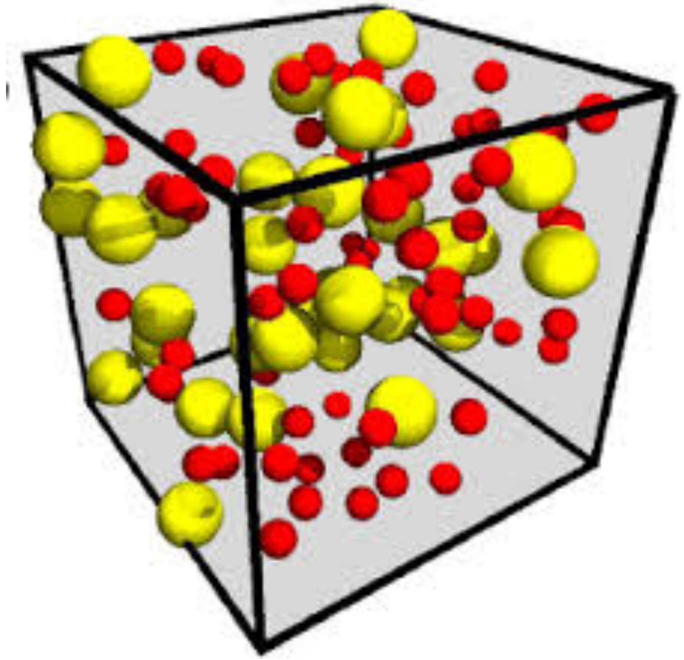
Molecular dynamics

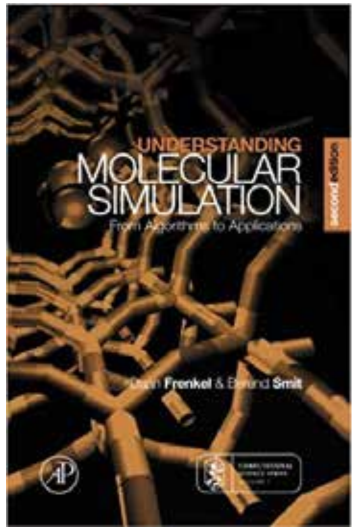
- Microstates can be sampled using molecular dynamics (MD).
- In classical MD, Atoms follow Newton's equation of motion.

$$F = ma$$

$$\Delta x = vt$$

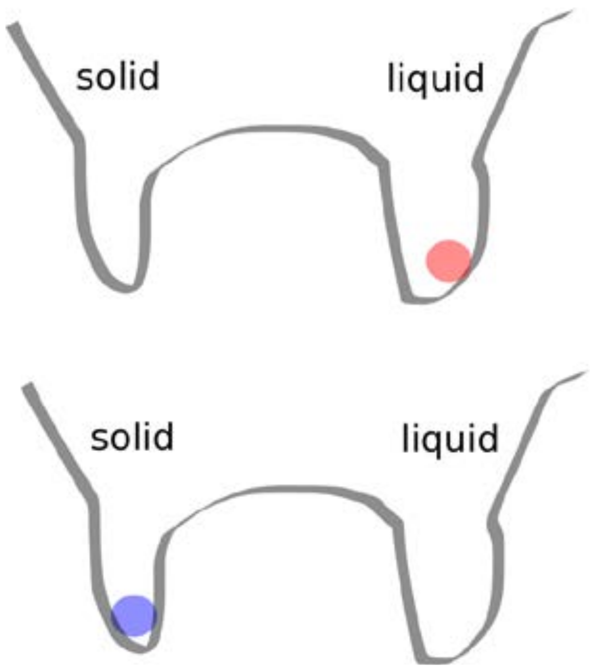
- Proper thermostat and barostat.

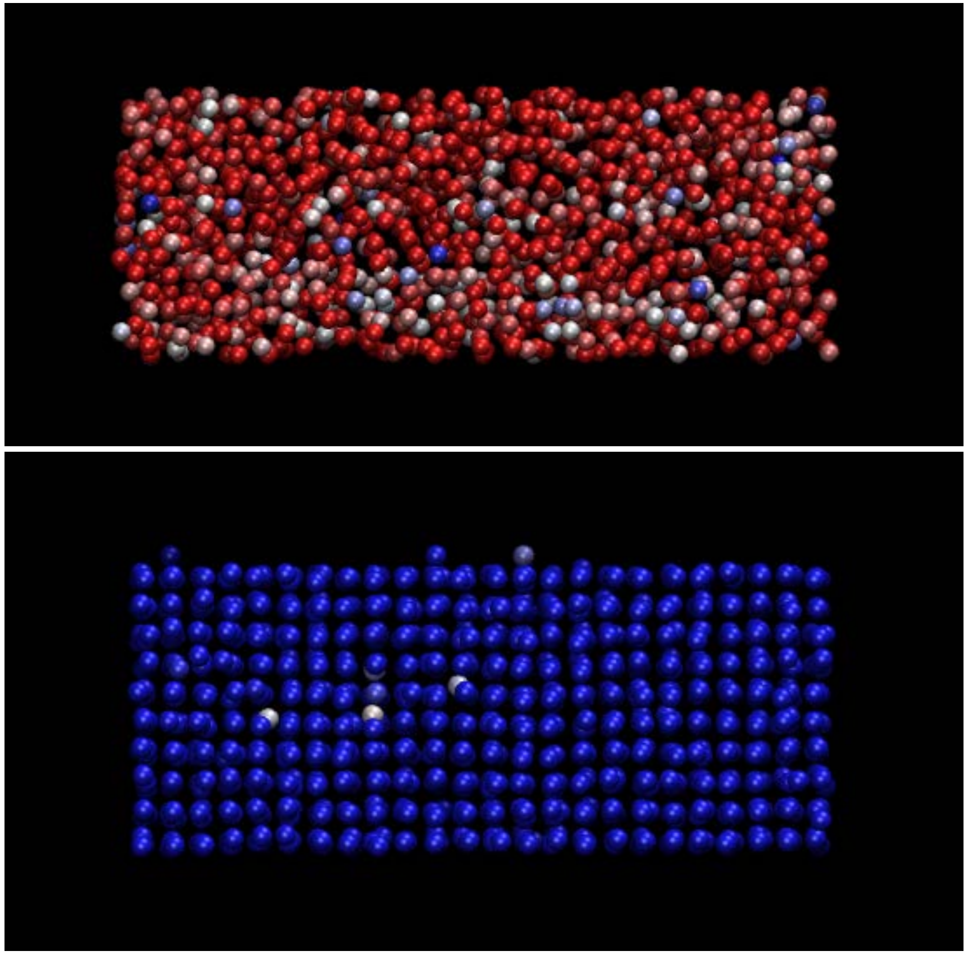




The shortcoming of molecular dynamics

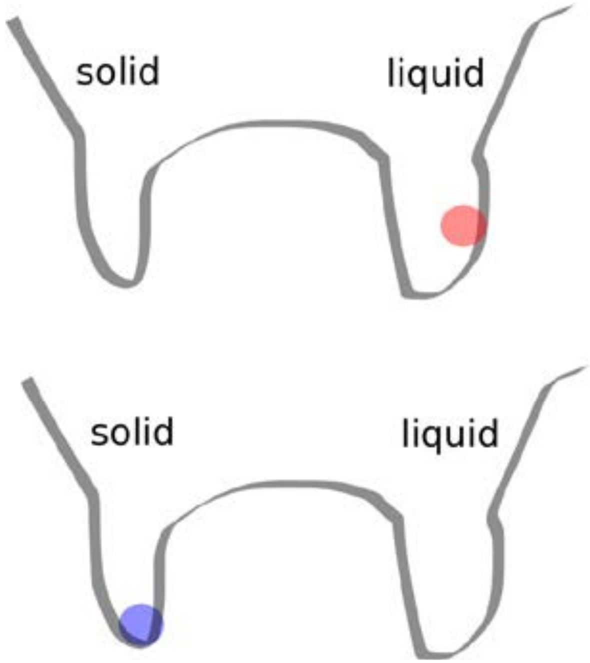
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

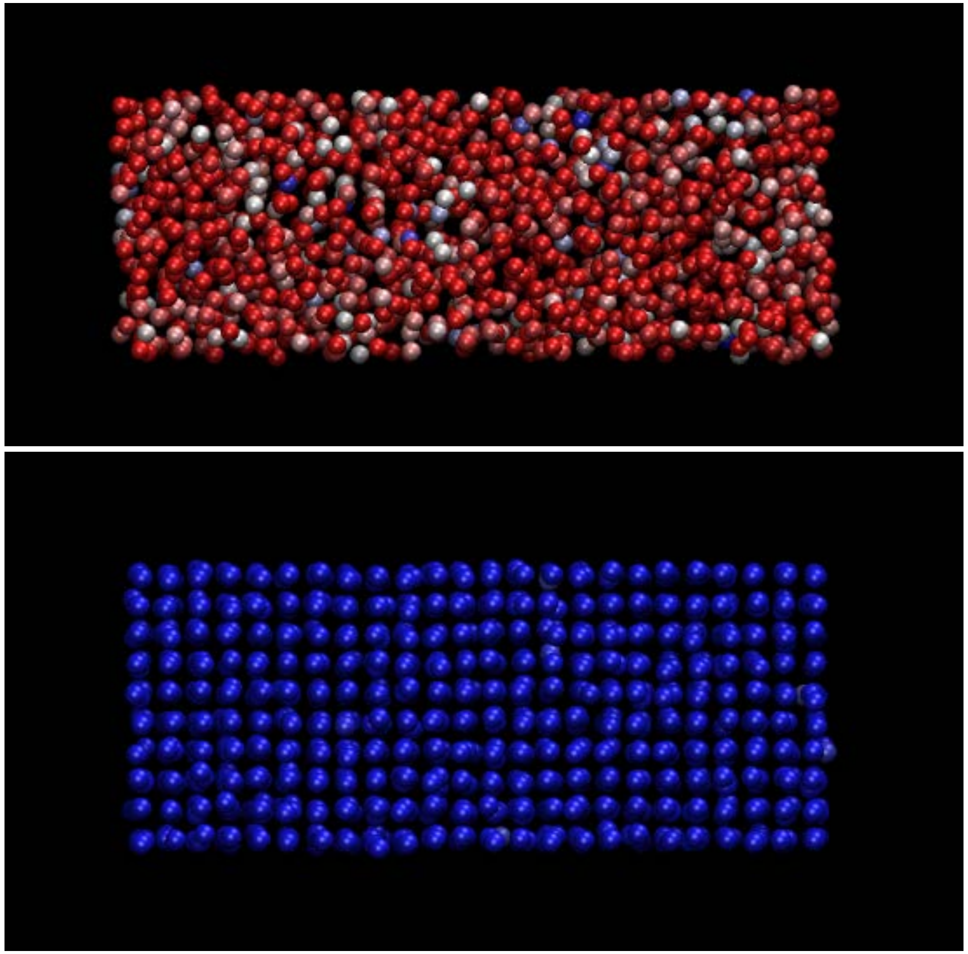




The shortcoming of molecular dynamics

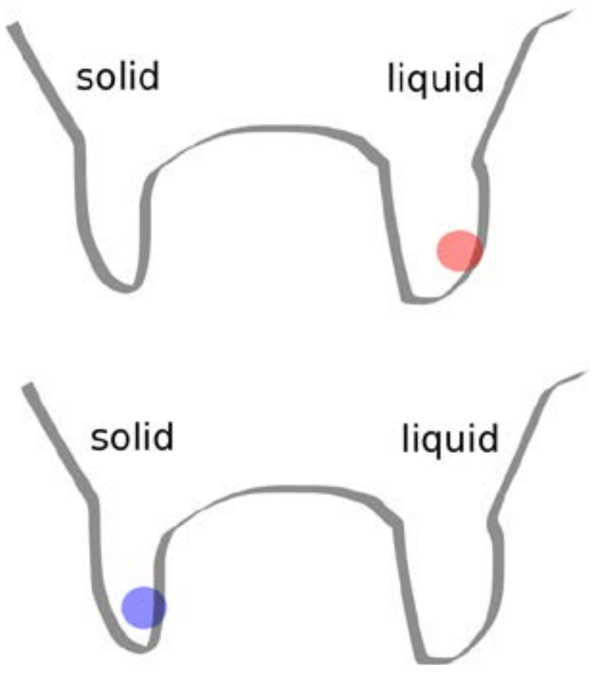
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

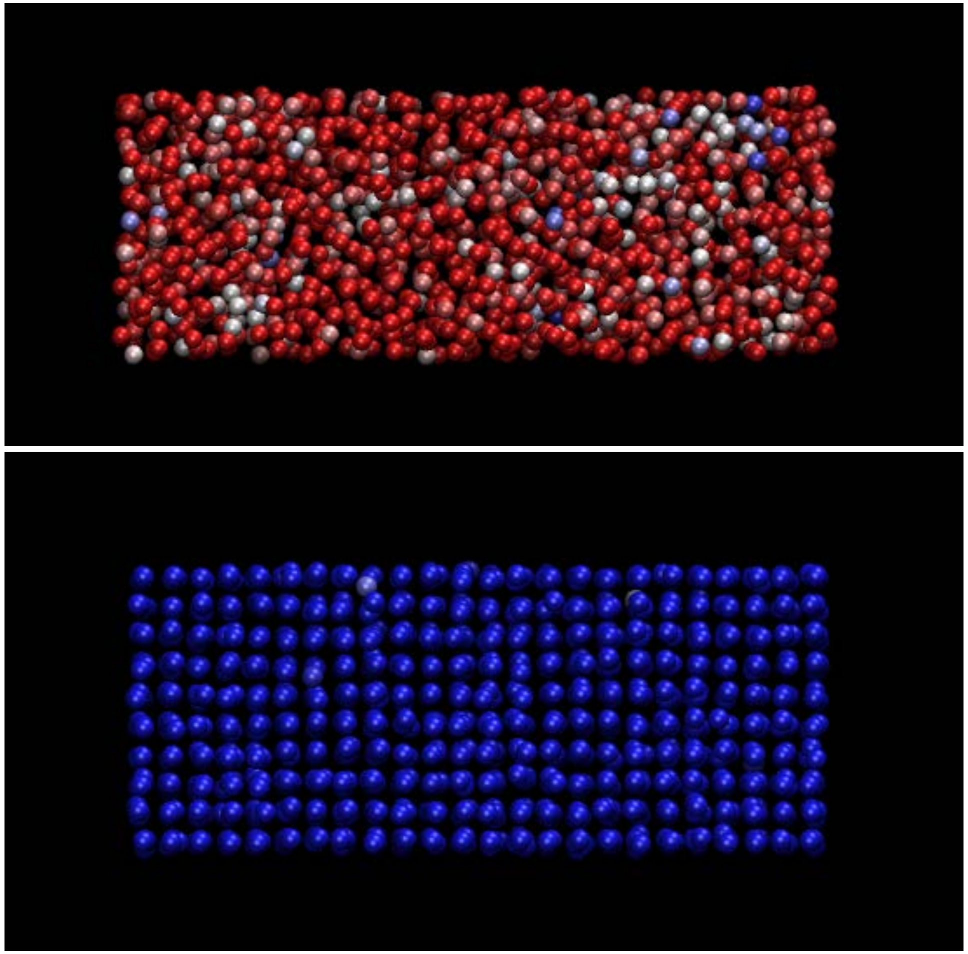




The shortcoming of molecular dynamics

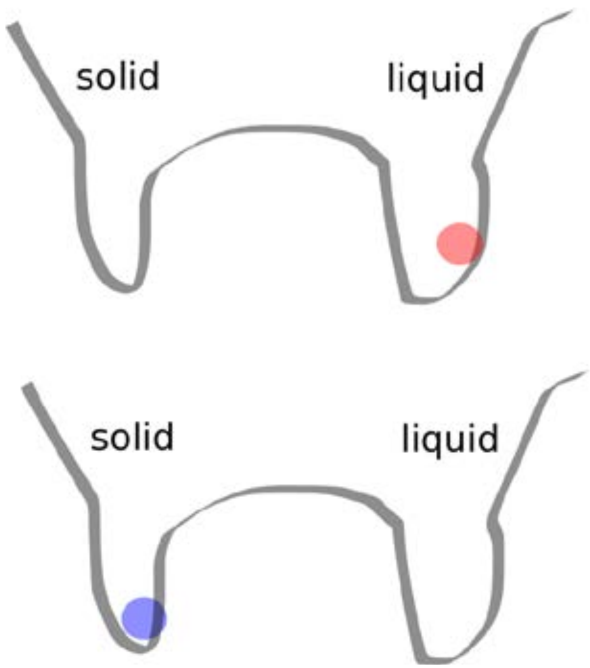
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

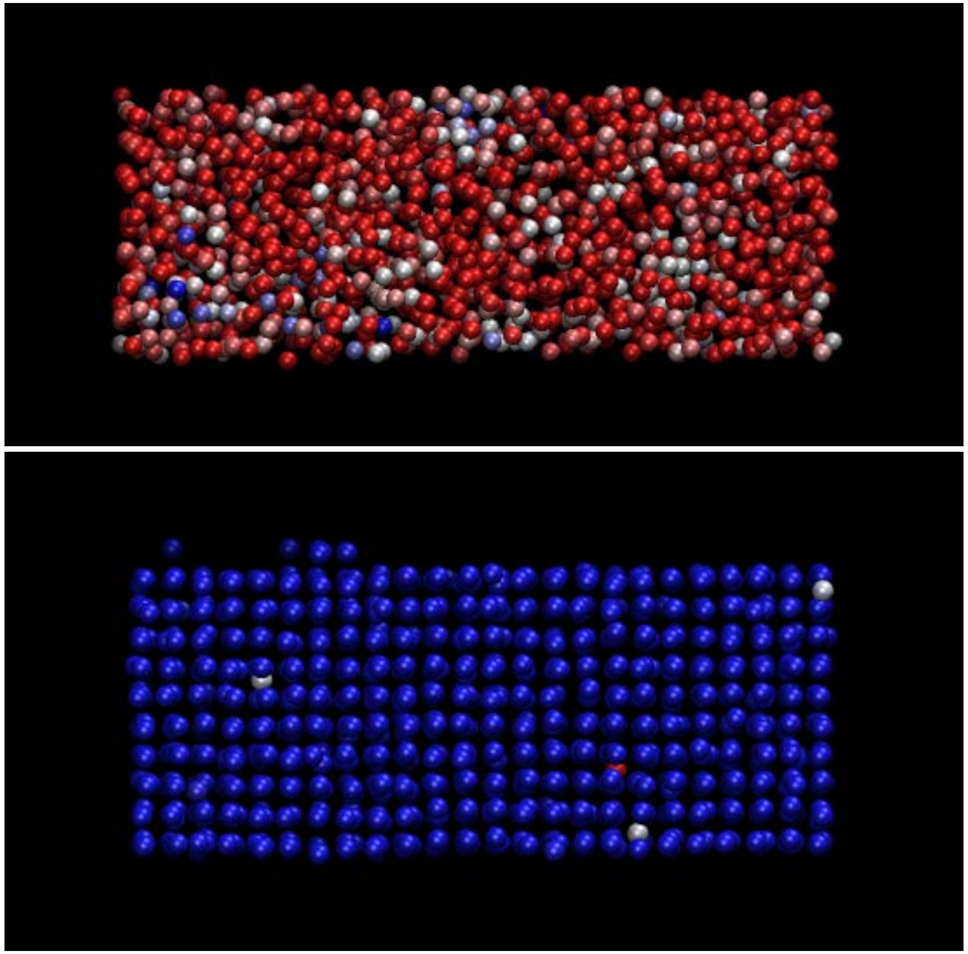




The shortcoming of molecular dynamics

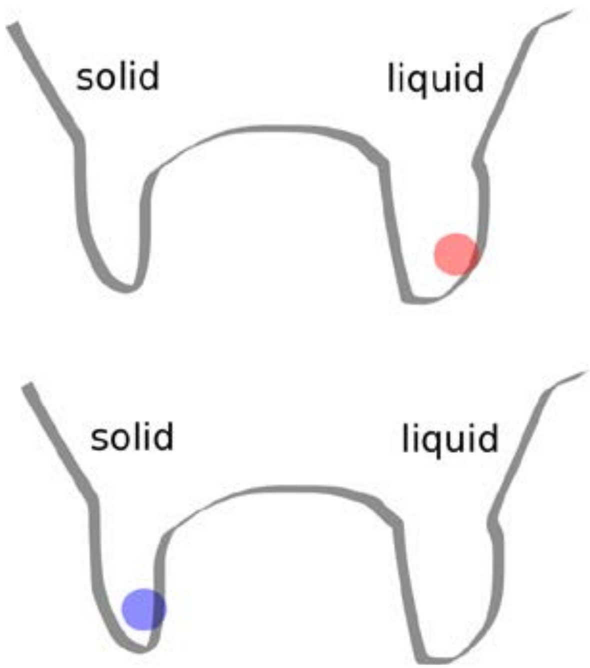
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

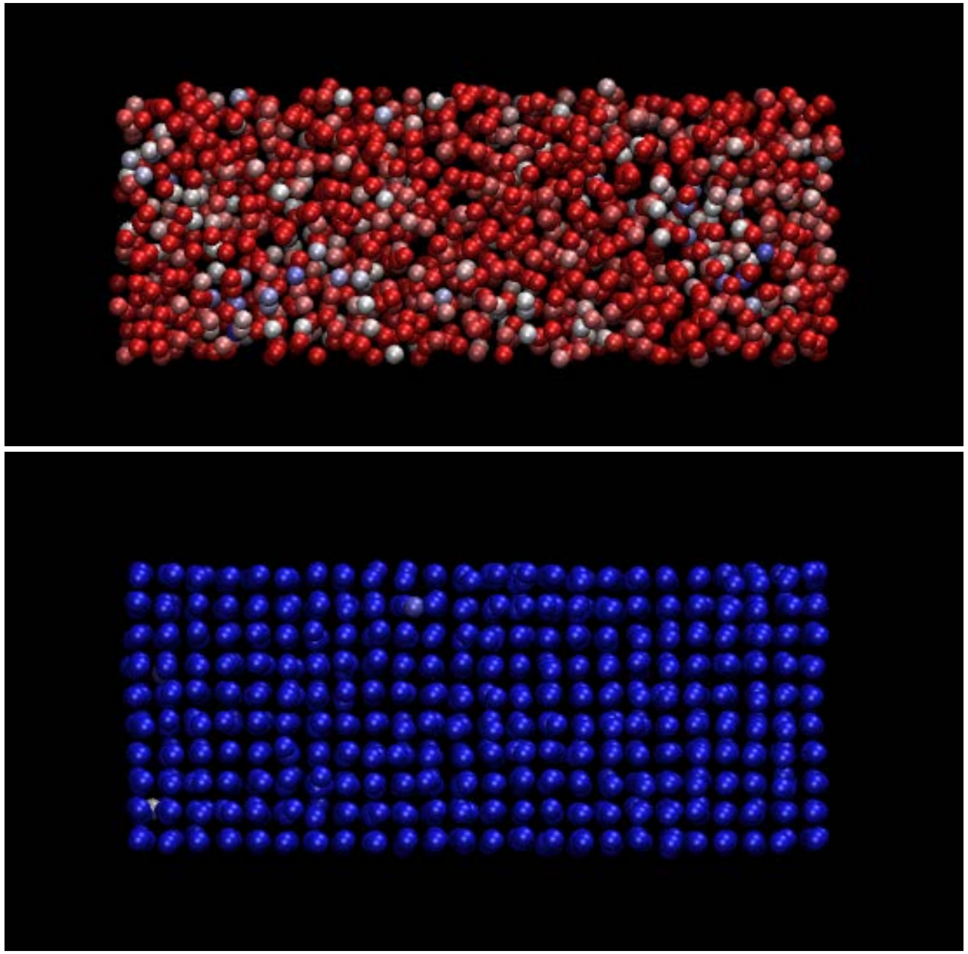




The shortcoming of molecular dynamics

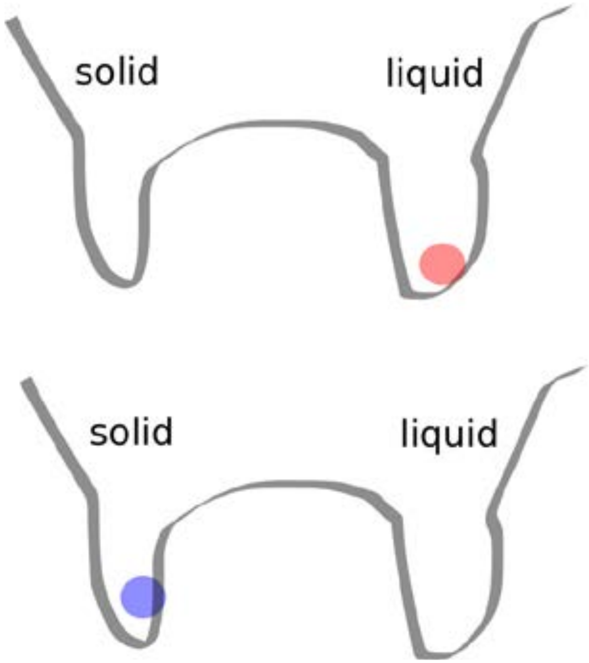
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

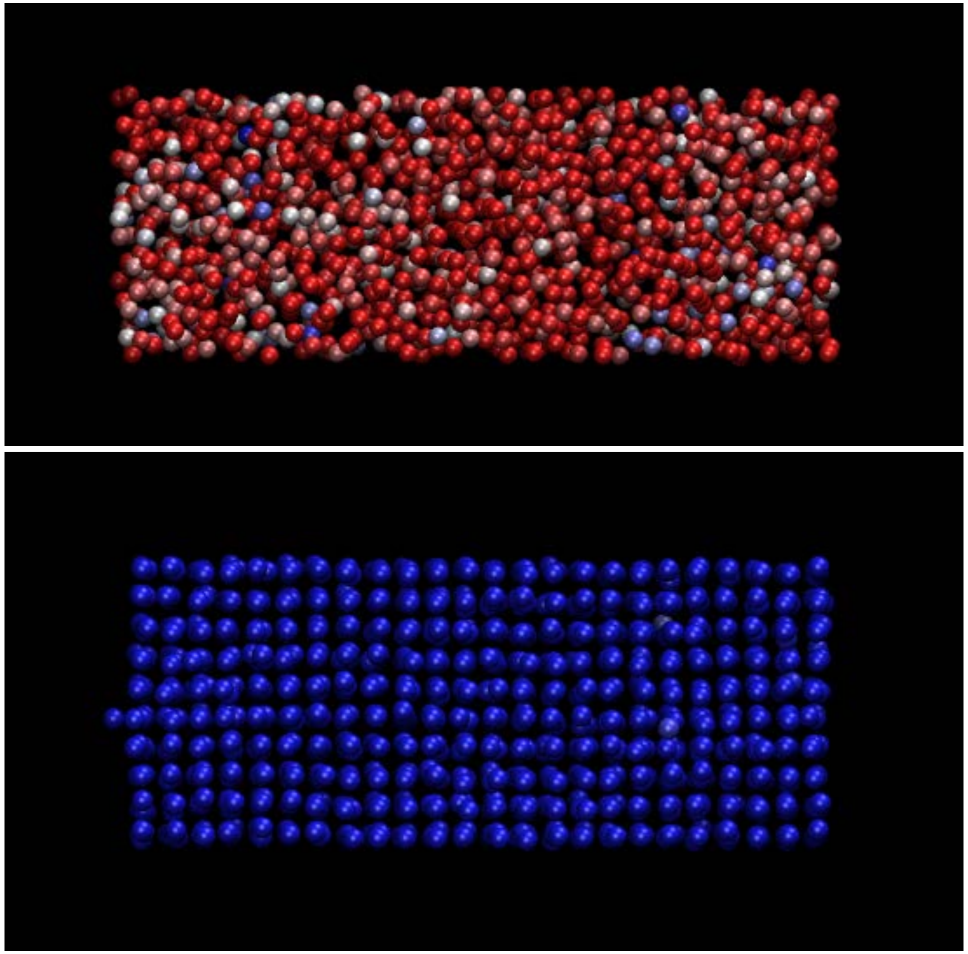




The shortcoming of molecular dynamics

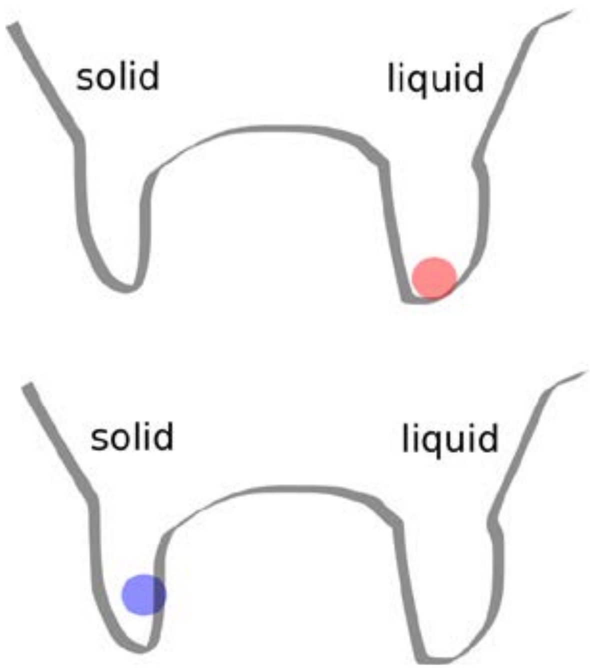
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

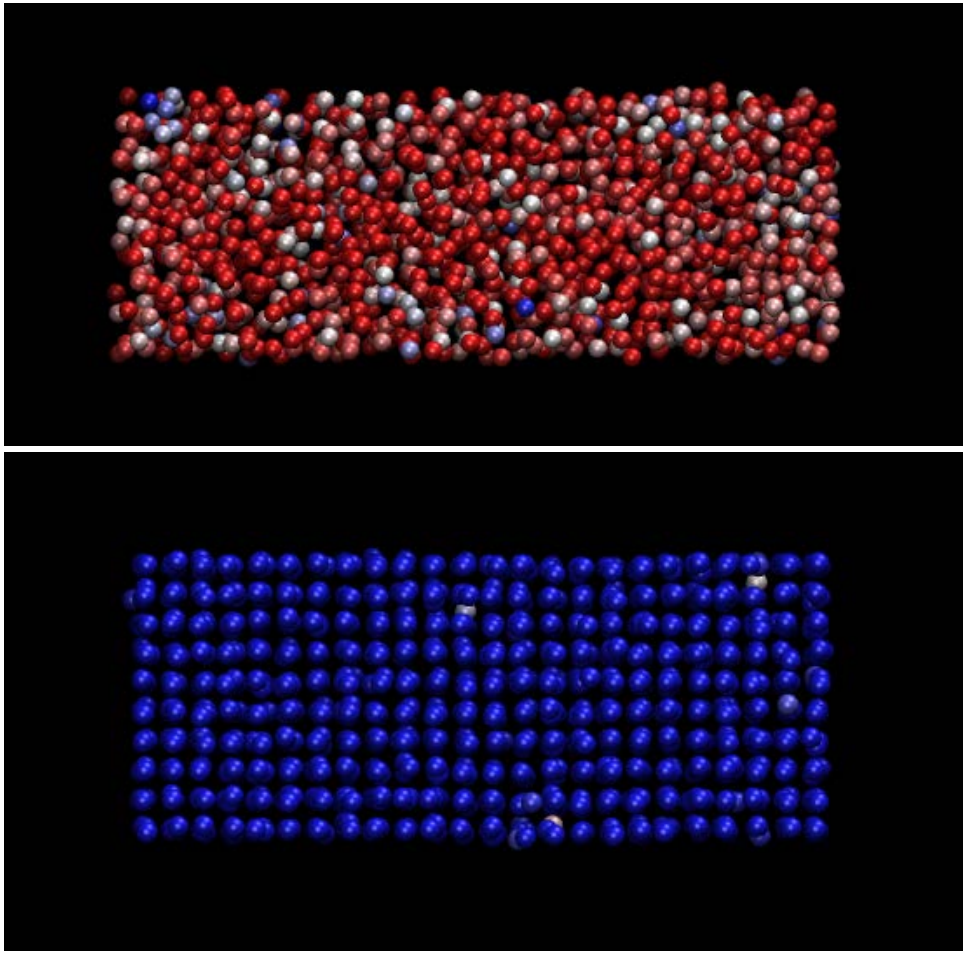




The shortcoming of molecular dynamics

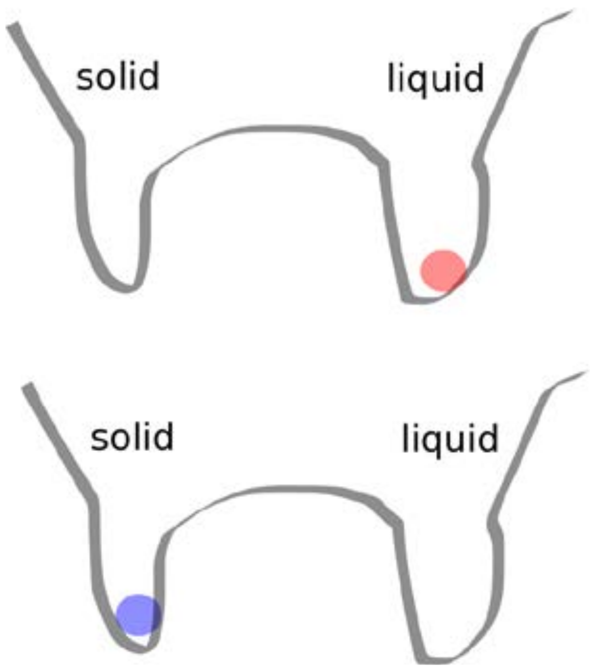
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

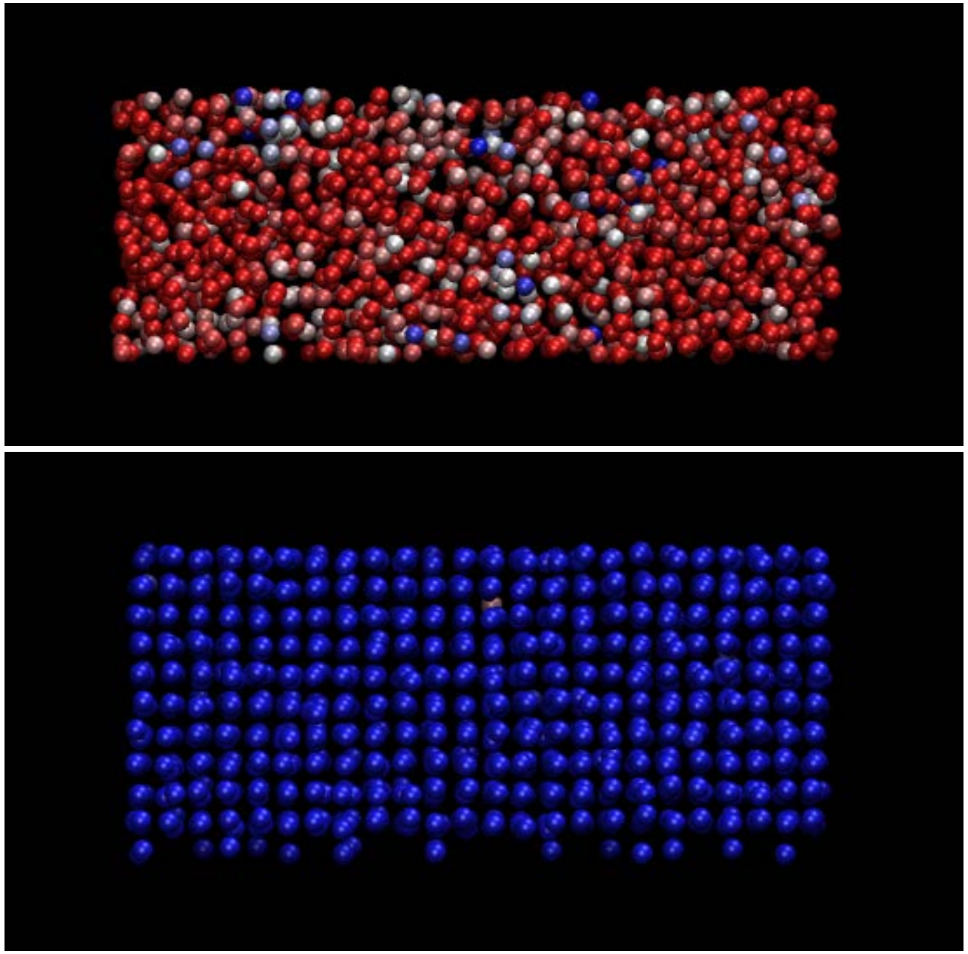




The shortcoming of molecular dynamics

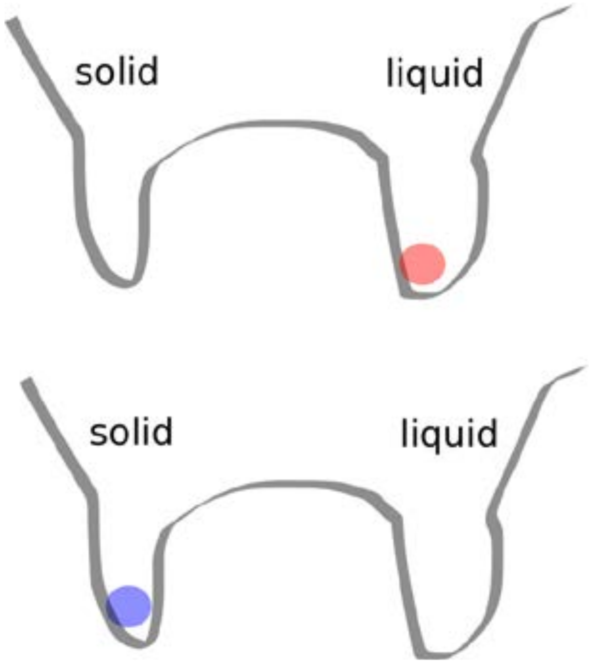
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

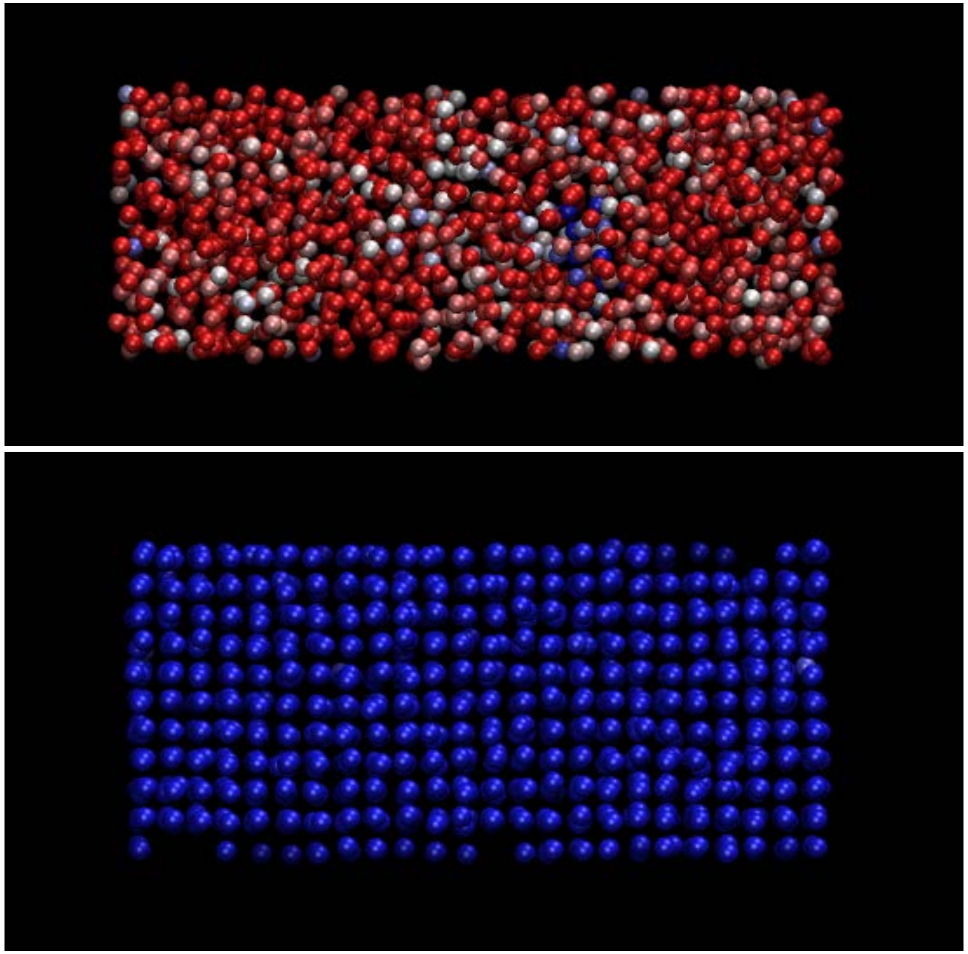




The shortcoming of molecular dynamics

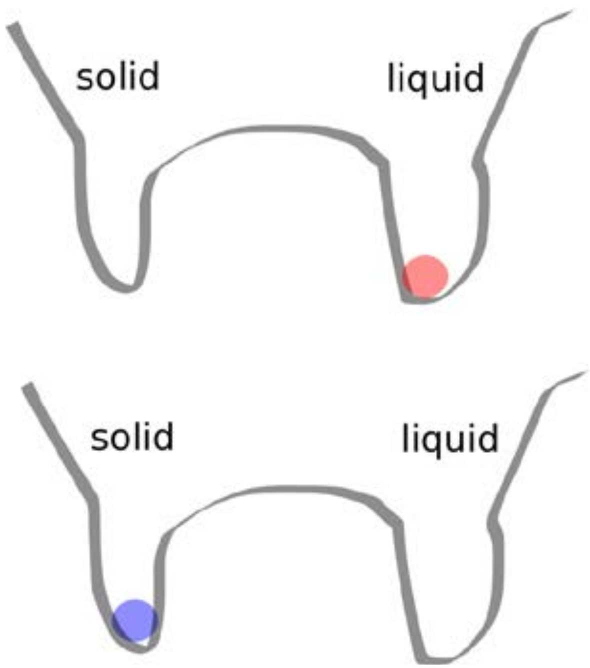
- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.

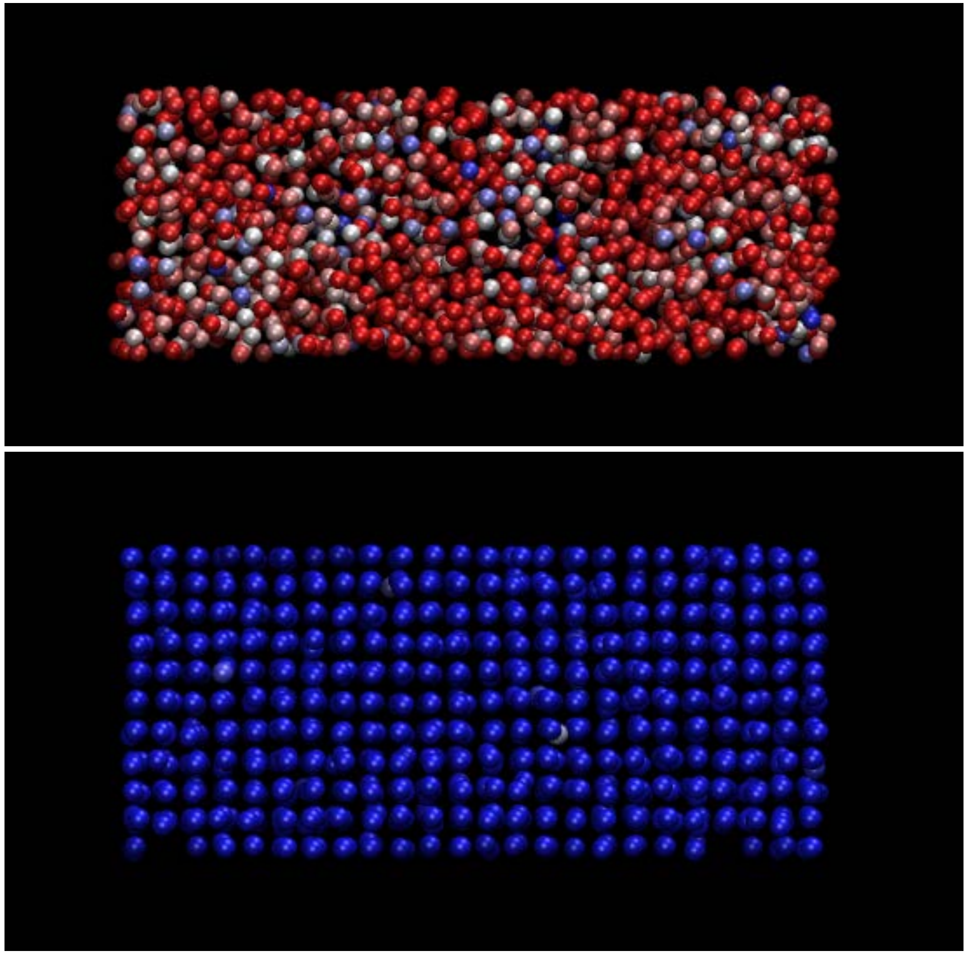




The shortcoming of molecular dynamics

- The low probability states are under-represented in a finite run.
- It is difficult to cross the energy barrier between two equilibrium states.





Overcome activation barrier

- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

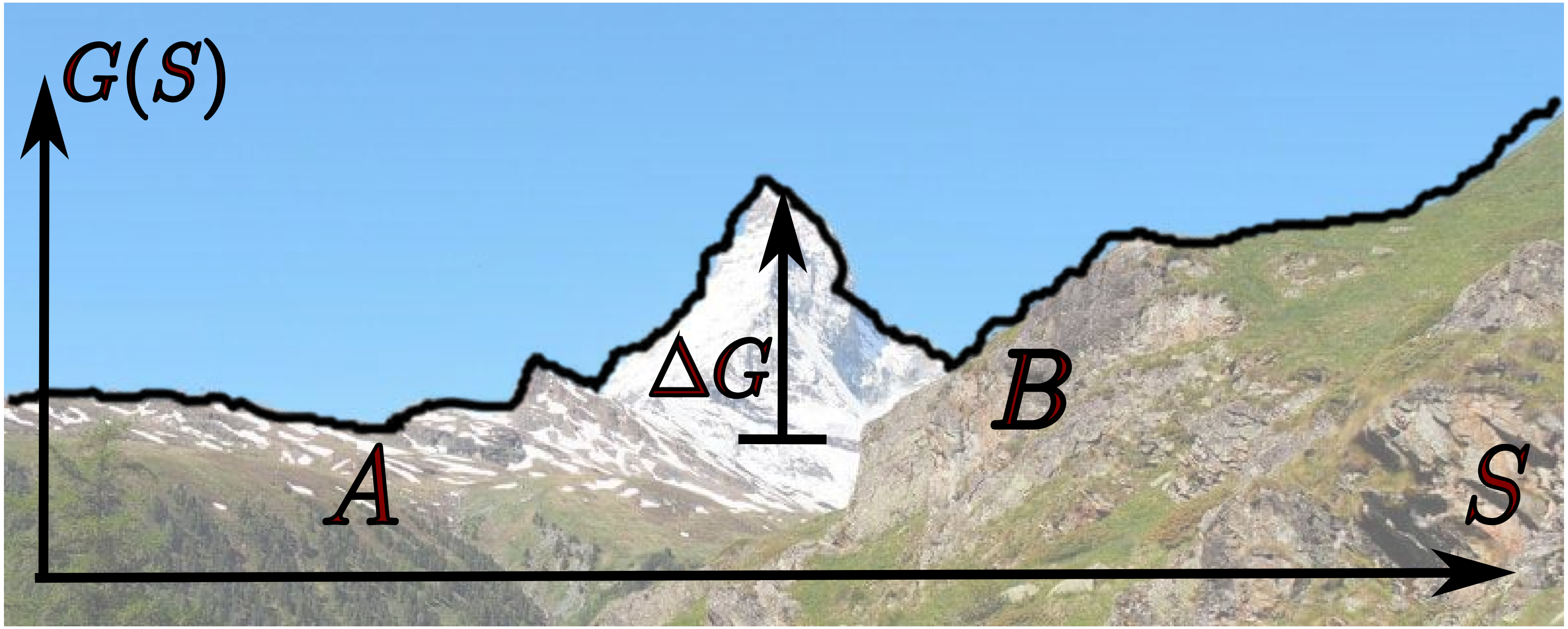
$$H_{biased}(q) \leftarrow H(q) + V(S)$$

And $V_t(S) = \sum_{t'<t} g(S(t'))$

- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



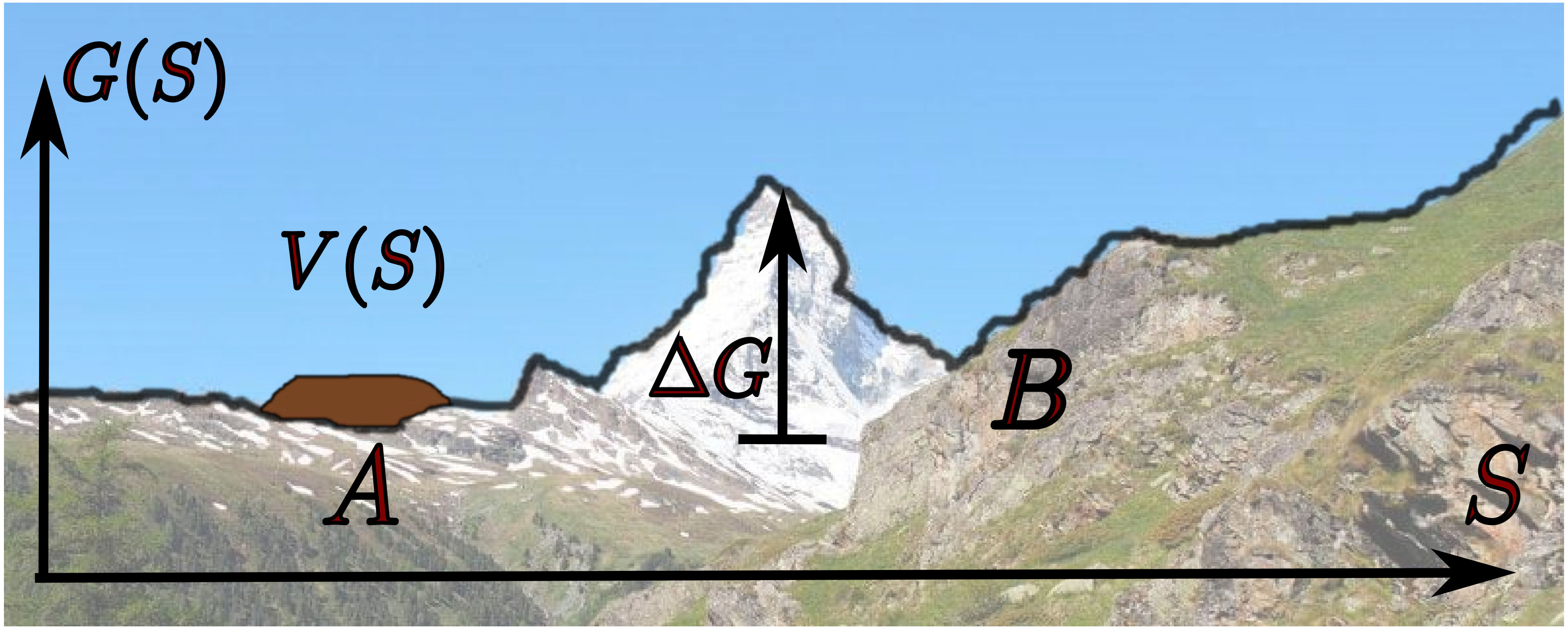
- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

  And $V_t(S) = \sum_{t'<t} g(S(t'))$
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

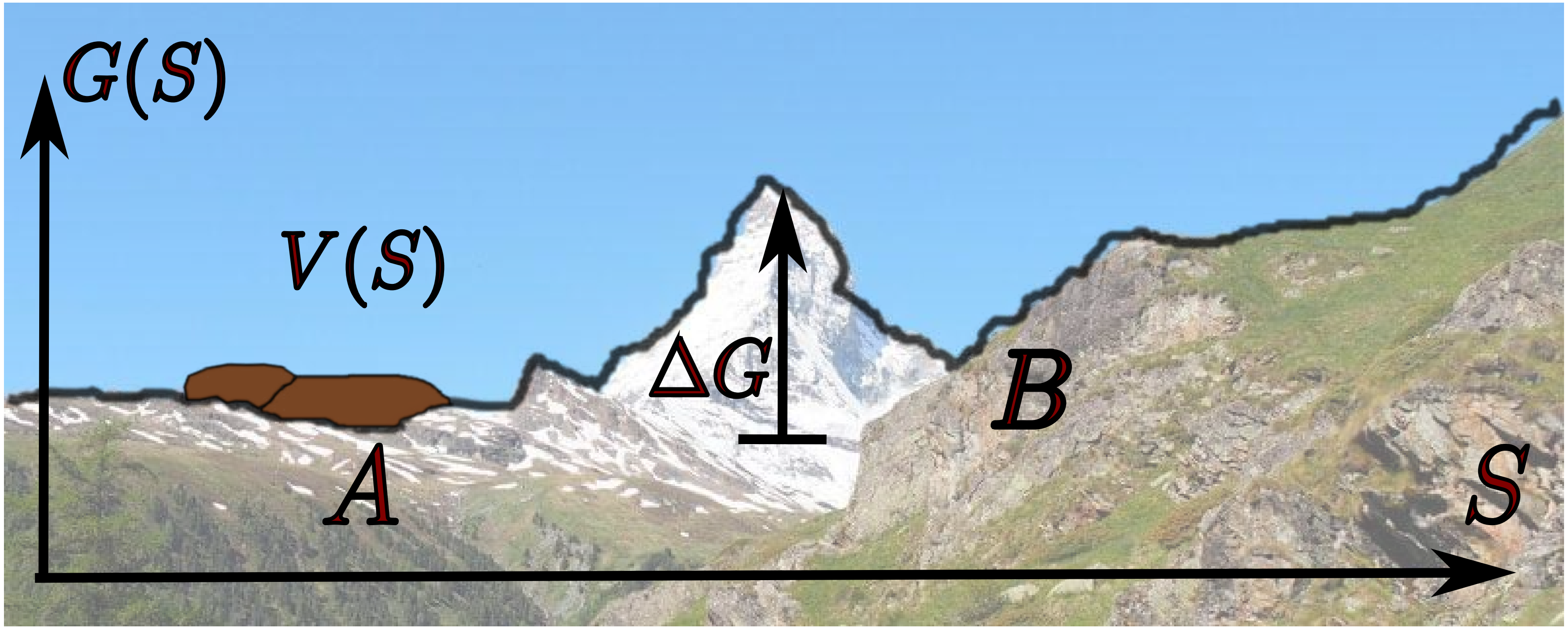
$$H_{biased}(q) \leftarrow H(q) + V(S)$$

And $V_t(S) = \sum_{t'<t} g(S(t'))$

- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



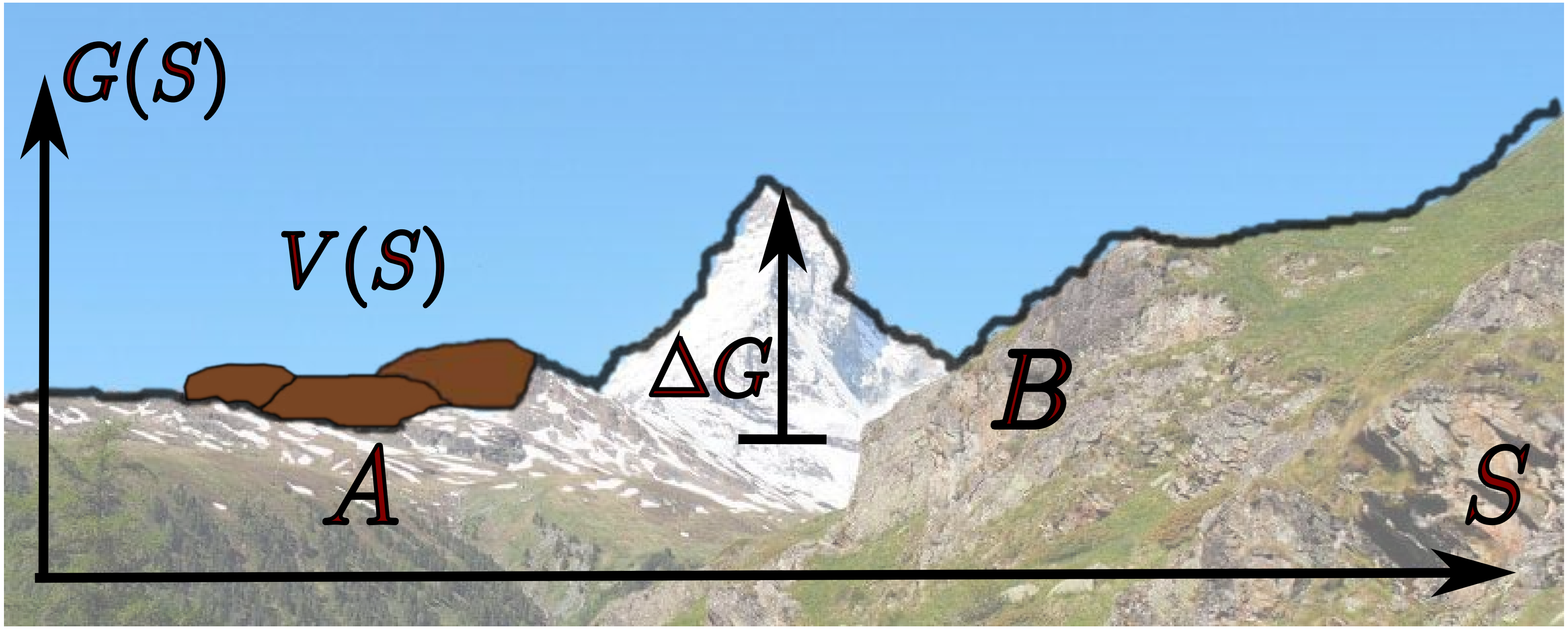
- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

  And $V_t(S) = \sum_{t'<t} g(S(t'))$
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



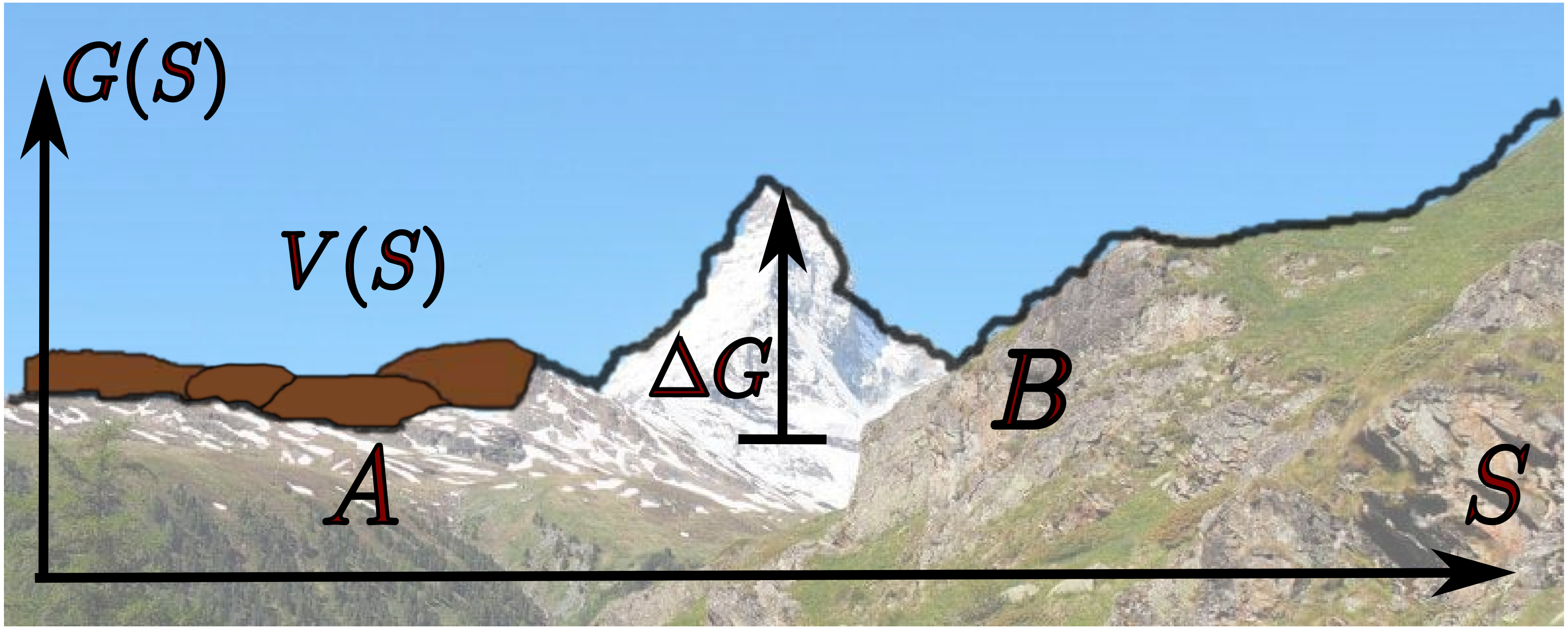
- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

  And $V_t(S) = \sum_{t'<t} g(S(t'))$
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

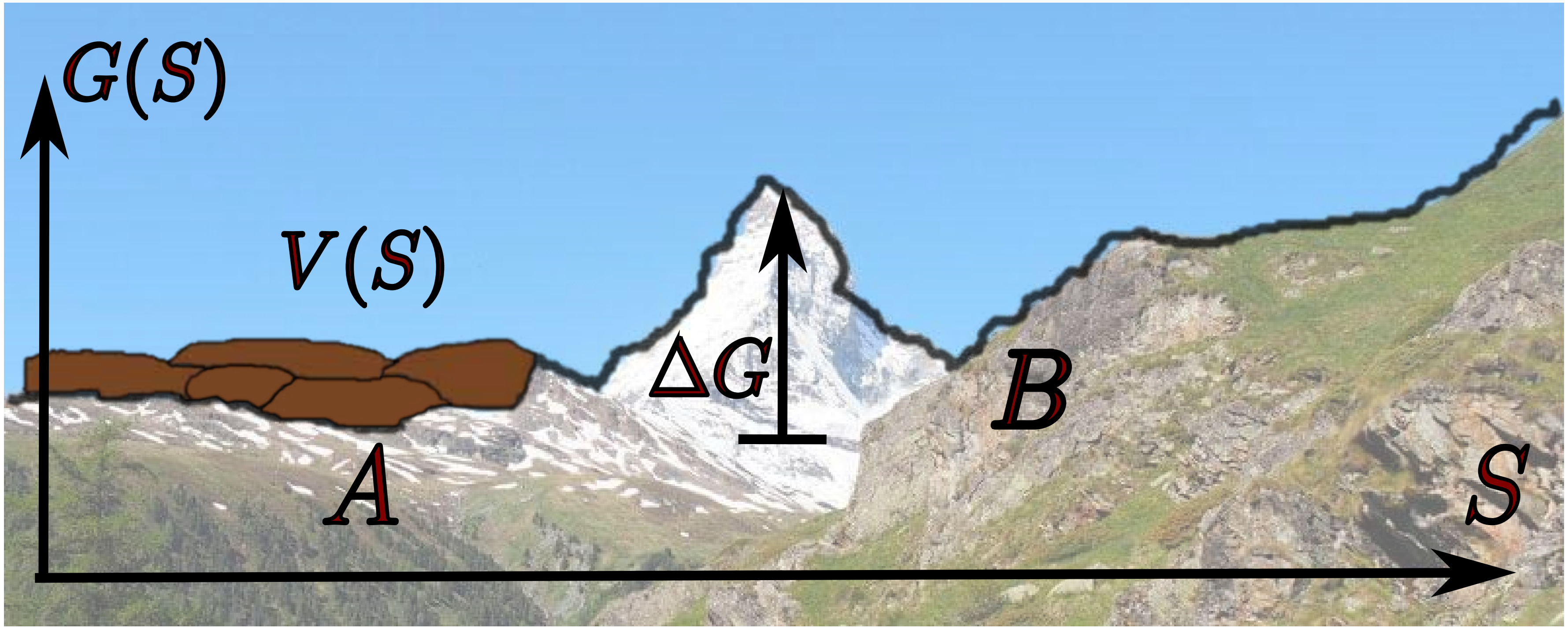
$$H_{biased}(q) \leftarrow H(q) + V(S)$$

And $V_t(S) = \sum_{t'<t} g(S(t'))$

- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



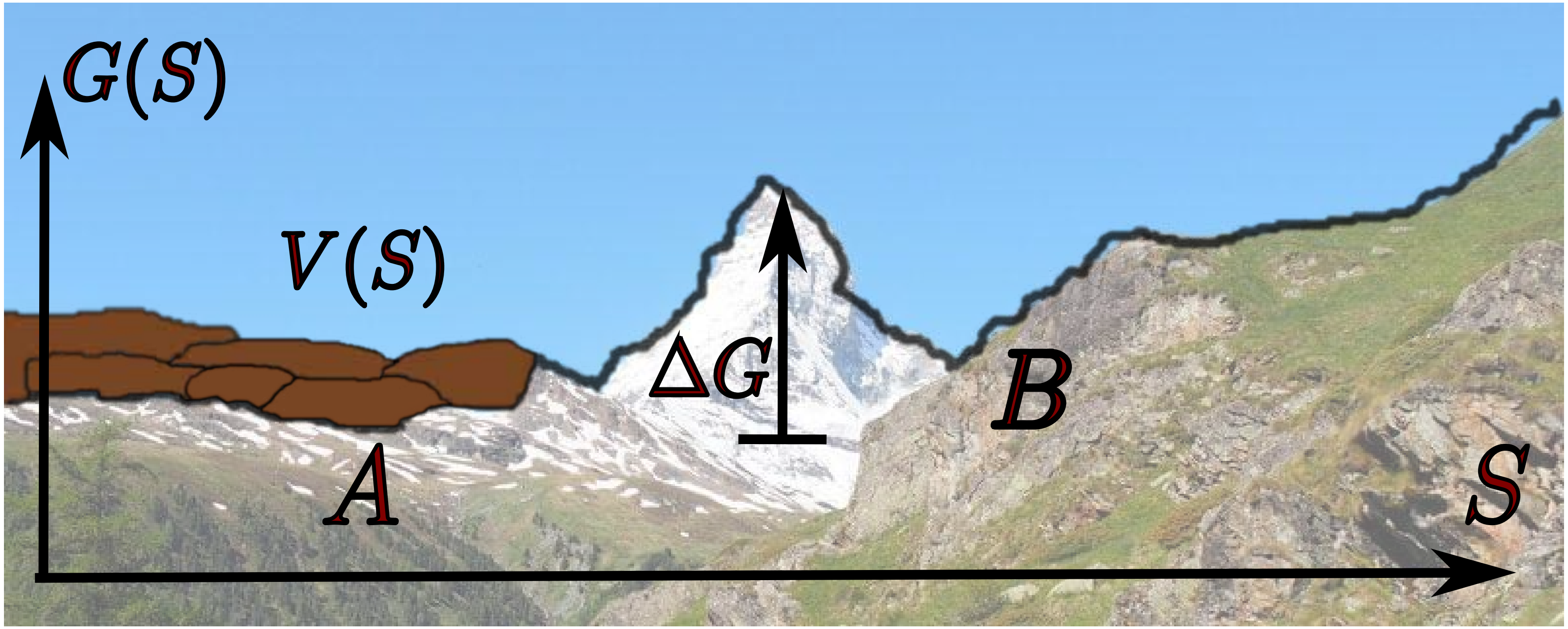
- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

  And $V_t(S) = \sum_{t'<t} g(S(t'))$
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



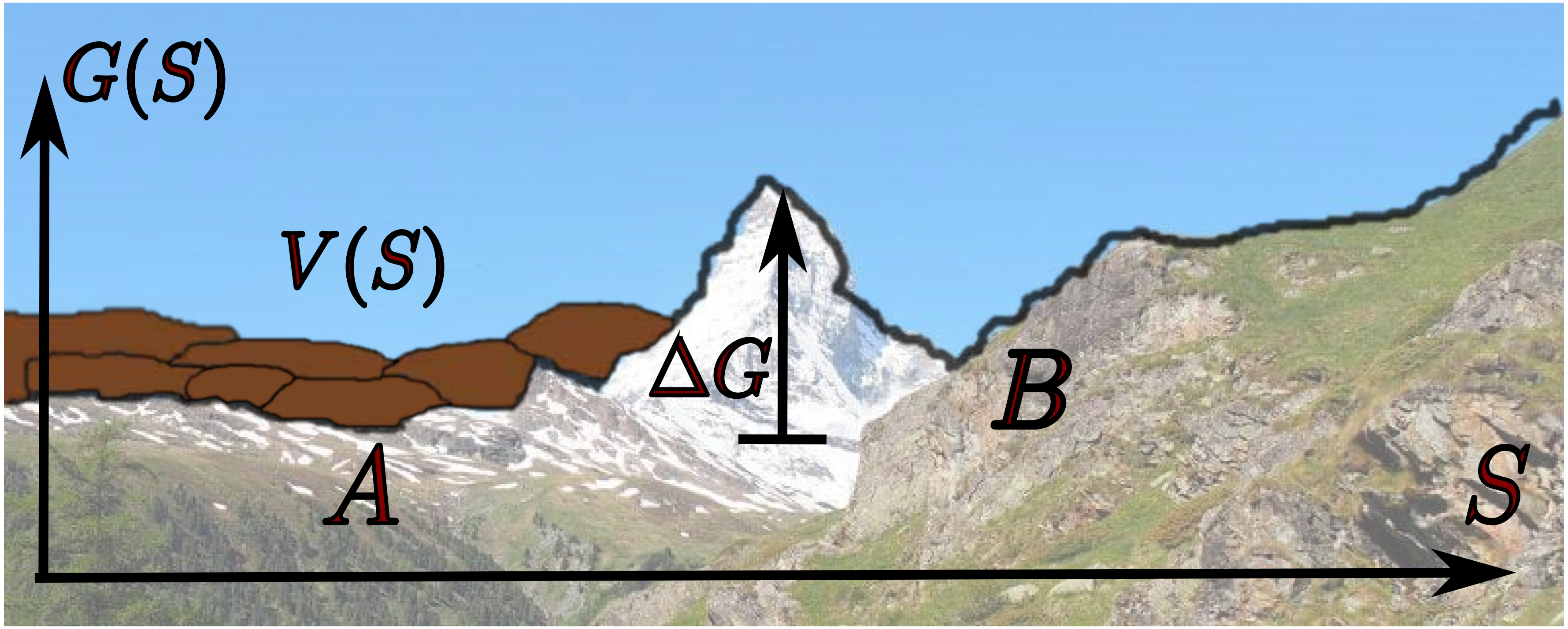
- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

  And $V_t(S) = \sum_{t'<t} g(S(t'))$
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

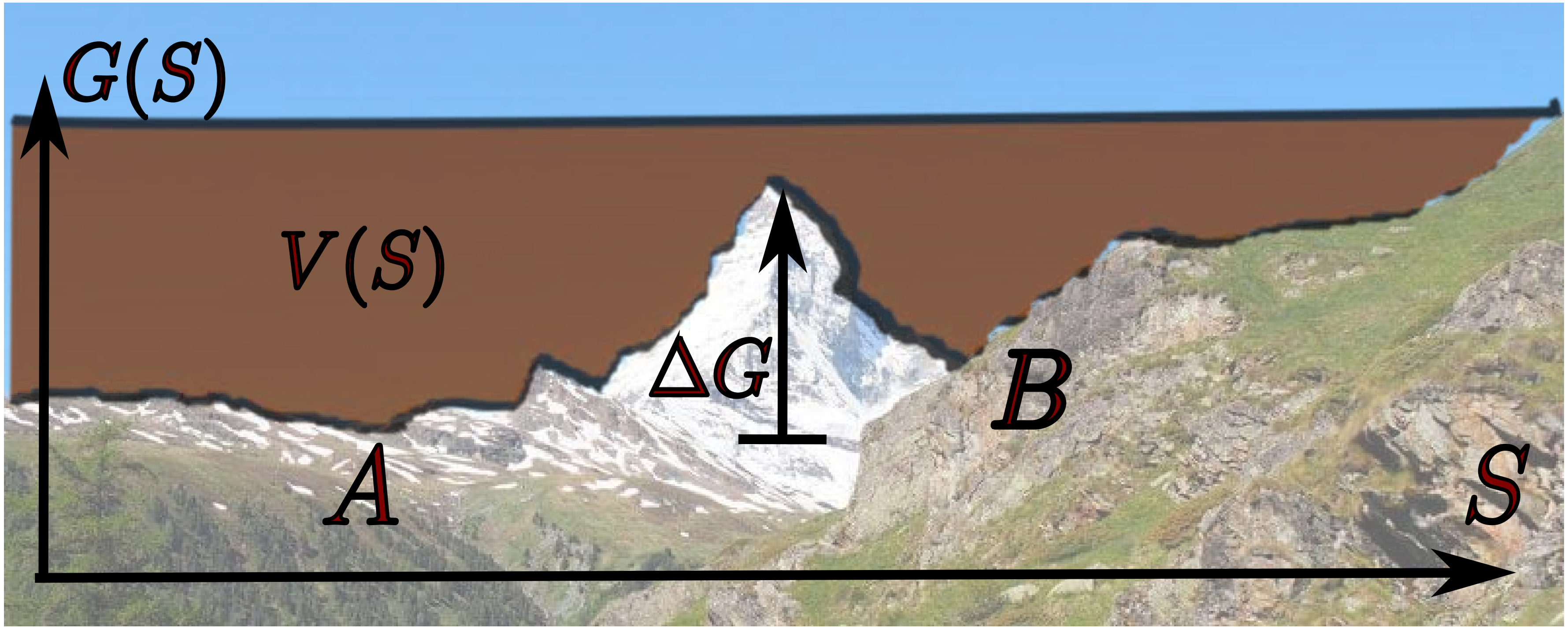
$$H_{biased}(q) \leftarrow H(q) + V(S)$$

And $V_t(S) = \sum_{t'<t} g(S(t'))$

- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

Overcome activation barrier



- The free energy surface $G(S)$ as a function of the order parameters S.
- Add bias to the system by altering the system Hamiltonian

$$H_{biased}(q) \leftarrow H(q) + V(S)$$

And $V_t(S) = \sum_{t'<t} g(S(t'))$

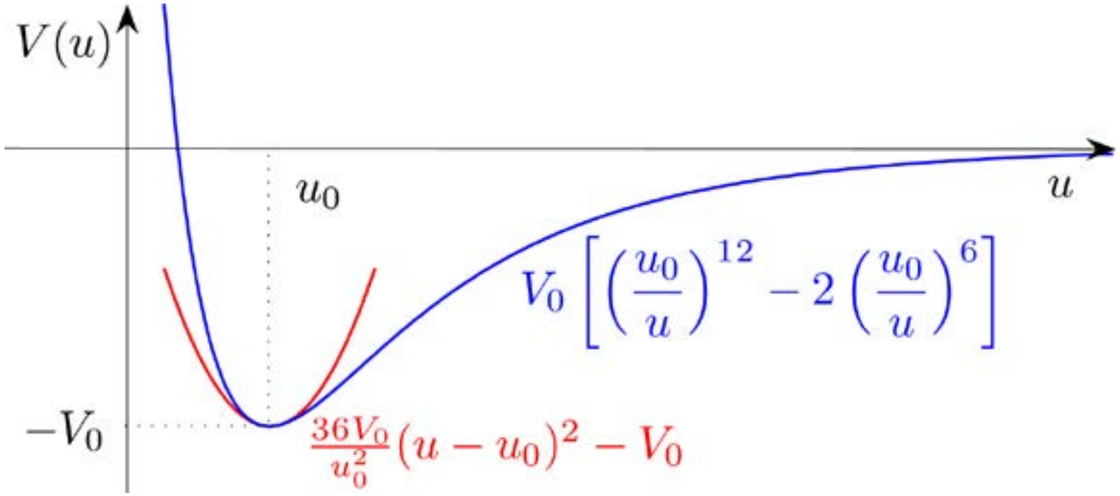
- Upon convergence, $-V(S) = G(S)$

[Laio and Parrinello. PNAS (2002)]

The classical Gibbs free energy

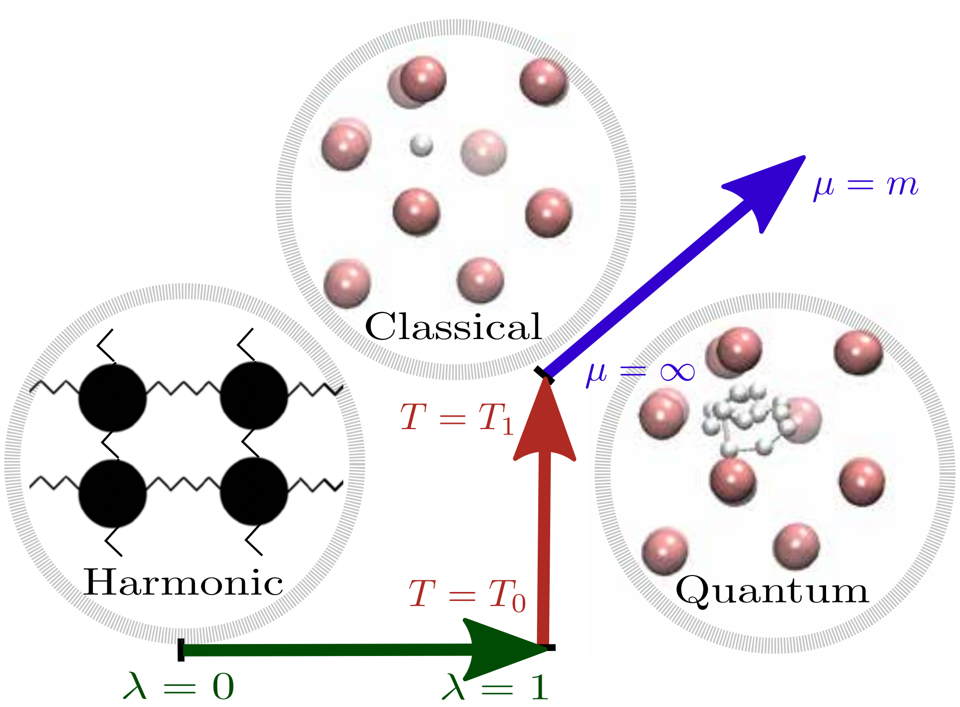
In thermodynamics, the Gibbs free energy is $G(P,T) = U + PV - TS$.
In classical statistical mechanics, the Gibbs free energy is

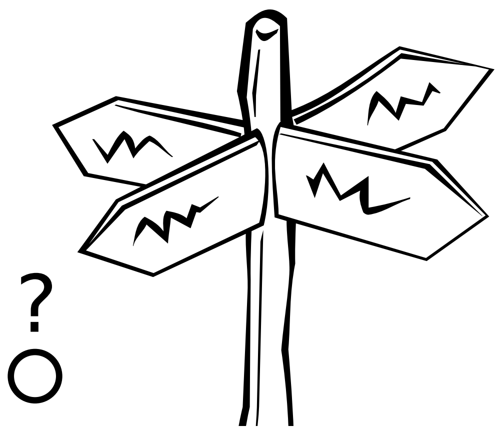
$$G(P,T) = -k_B T \ln \int dV \exp\left[-\frac{PV}{k_B T}\right] \int_{D(V)} d\mathbf{q} \exp\left[-\frac{U(\mathbf{q})}{k_B T}\right]$$



- Minimum potential energy at 0 K.
- Harmonic approximation $G = k_B T \sum_{i=1}^{3N-3} \ln \frac{\hbar\omega_i}{k_B T}$.
- Self-consistent phonons [Monserrat & Needs]
- **Thermodynamic integration. [Polson & Frenkel JCP 1998; Li, Totton & Frenkel JCP 2018]**
- Rare event sampling methods (e.g. umbrella sampling, metadynamics, transition path sampling).

Thermodynamic integration





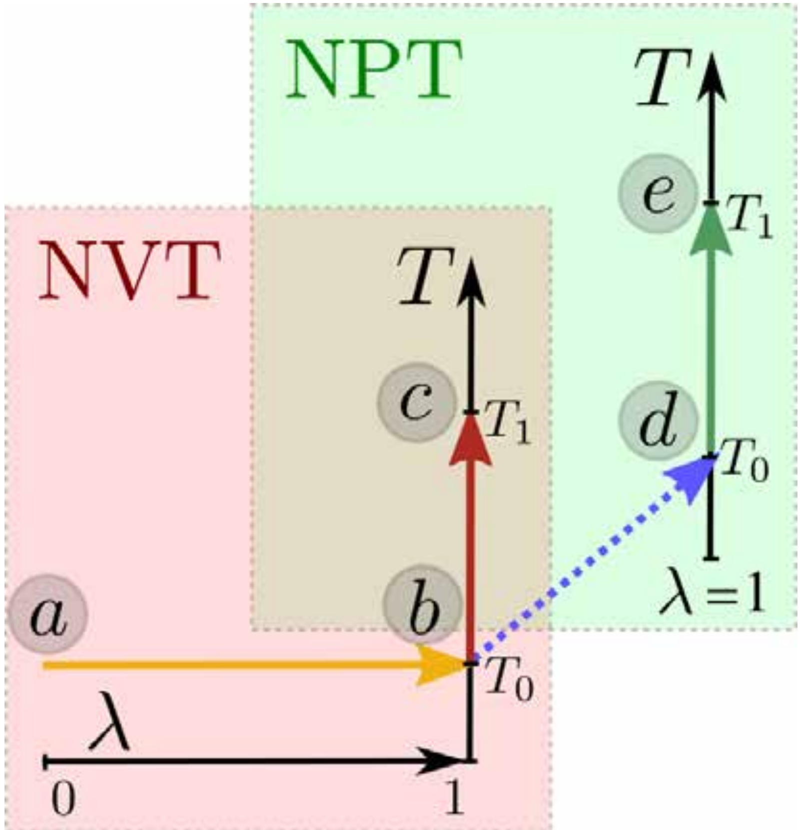
Consider two systems, A and B, which can be transformed continuously between each other via a parameter λ,

$$F_A - F_B = \int_{\lambda_A}^{\lambda_B} \frac{dF(\lambda)}{d\lambda} d\lambda$$

This parameter can be

- Thermodynamic variables (temperature, volume, concentration, etc.)
- Switching parameter between different Hamiltonians
- Order parameters (reaction coordinates)

Thermodynamic integration routes



→ Between harmonic and real crystal.
$\Delta A = \int_0^1 d\lambda \, \langle U - U_{\mathrm{h}} \rangle_{V,T_0,\lambda}$

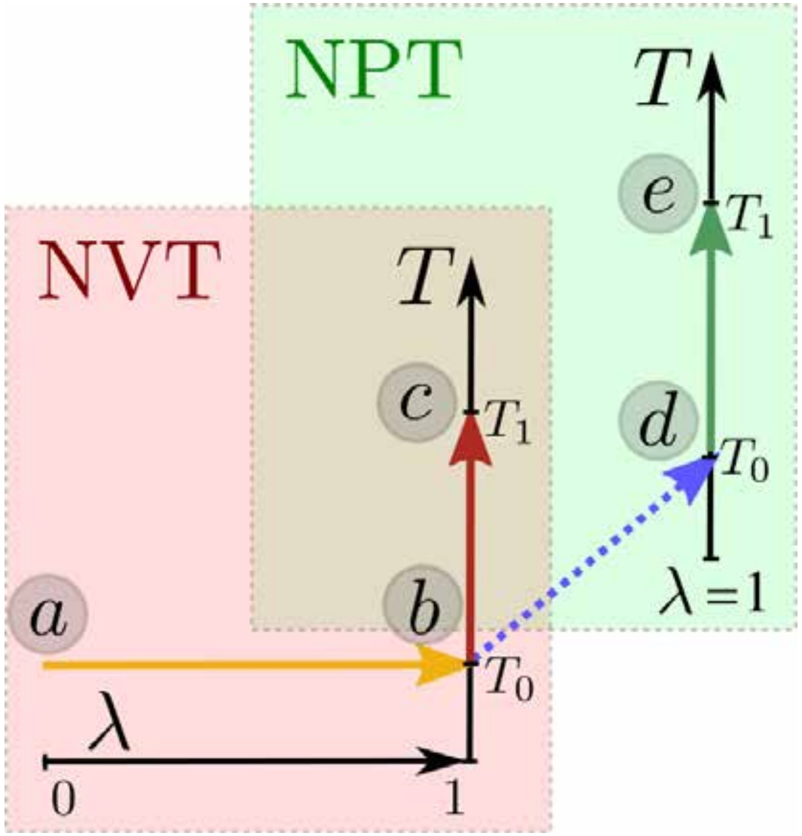
→ Integrate with respect to temperature.
$$\Delta A = -\int_{T_0}^{T_1} \frac{\langle K + U \rangle_{V,T}}{T^2} dT$$

→ From NVT to NPT ensemble; from A to G.

→ Integrate with respect to temperature.
$$\Delta G = -\int_{T_0}^{T_1} \frac{\langle H \rangle_{P,T}}{T^2} dT$$

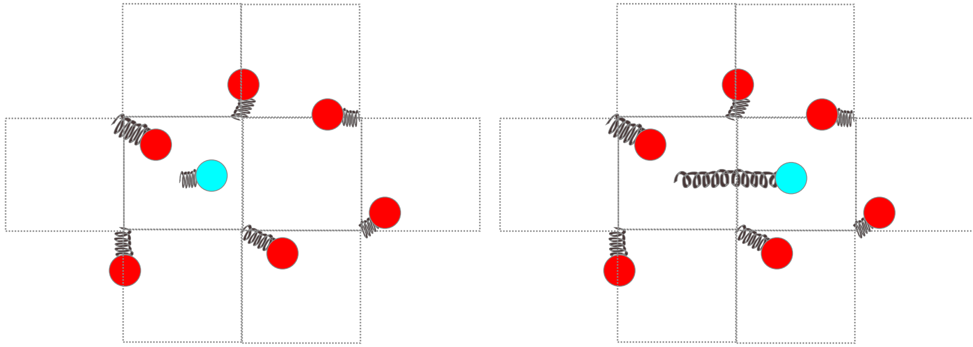
- To get the Helmholtz free energy A : → →
- To get the Gibbs free energy G: → → →

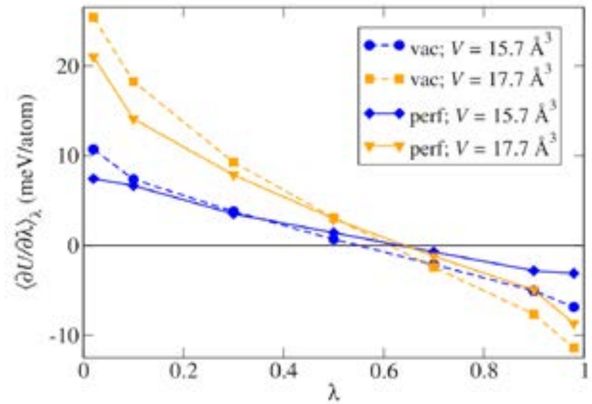
Some justifications



Why integrate from a harmonic to a real crystal at a low temperature?

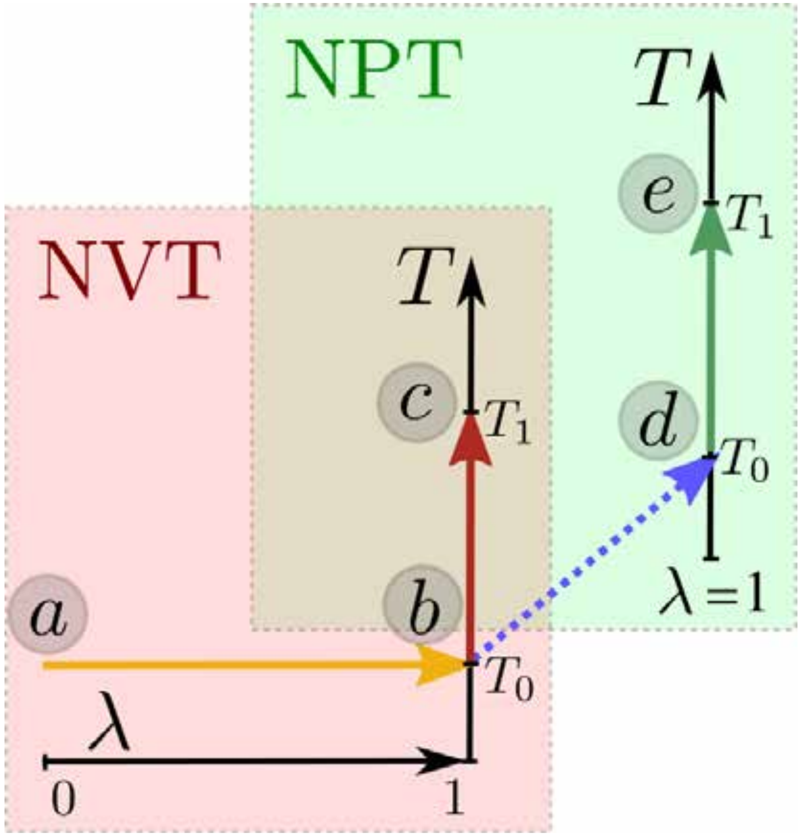
$$\Delta A = \int_0^1 d\lambda \, \langle U - U_\mathrm{h} \rangle_{V,T_0,\lambda}$$



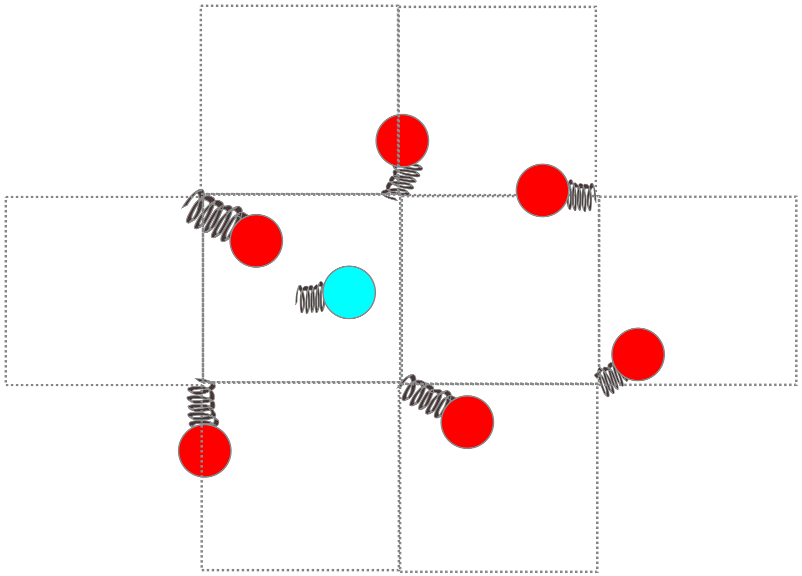


[Grabowski, Ismer, Hickel & Neugebauer PRB 2009]

Some justifications

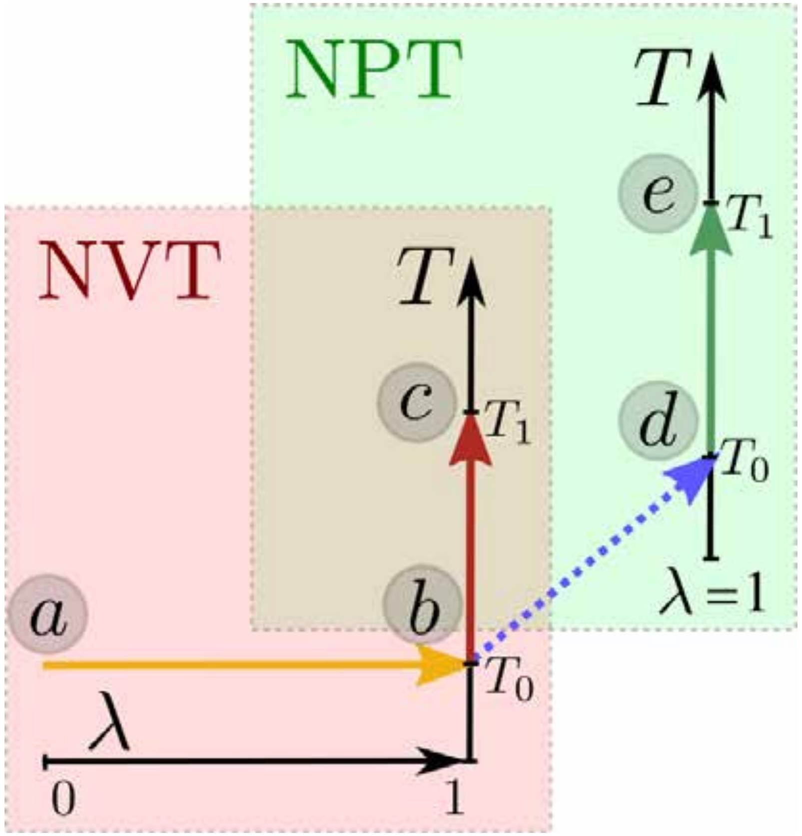


Why switch between the NVT and the NPT ensemble?



Because pressure may not be well-defined for the reference harmonic system.

Some justifications



- Choose a reference harmonic crystal that has the same frequency modes and equilibrium configuration as the real crystal.
- Separate the harmonic and the anharmonic part of the potential energy.
- Apply the viral theorem.
- Change the variable in the integration.
- Perform parallel tempering.

Getting started

A detailed yet simple description of the methodology

PHYSICAL REVIEW B **97**, 054102 (2018)

Computing the absolute Gibbs free energy in atomistic simulations: Applications to defects in solids

Bingqing Cheng* and Michele Ceriotti

Python notebooks and scripts

Init

In [7]:
```python
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [8]:
```python
kb = 3.1668105e-06 # in hartree
ha2ev = 27.211386
kbev = kb*ha2ev # in eV
T0=100
natombcc=250
natomvacancy=250-1
nfreebcc=(natombcc-1)*3
nfreevacancy=(natomvacancy-1)*3
```

The energy units in this notebook is in eV, unless specified otherwise

0K results

In [9]:
```python
U0bcc=-1030.68877518368
U0vacancy=-1024.77431283251
```

The free energy of the reference harmonic system with fixed center of mass at 100K

$A_h(T_0) = k_B T_0 \sum_{i=1}^{3N-3} \ln(\frac{\hbar\omega_i}{k_B T_0})$

In [10]:
```python
eva = np.loadtxt("ipi-100K/bcc-phonon/perfect-fd.eigval")
# The square root of the eigenvalues of the phonon modes is the frequency in the unit of hartree
AharmbccT0 = ha2ev*T0*kb*np.log(np.sqrt(eva[3:])/(T0*kb)).sum()
print AharmbccT0
```

6.94539173078

In [11]:
```python
eva = np.loadtxt("ipi-100K/vacancy-phonon/vacancy-fd.eigval")
# The square root of the eigenvalues of the phonon modes is the frequency in the unit of hartree
AharmvacancyT0 = ha2ev*T0*kb*np.log(np.sqrt(eva[3:])/(T0*kb)).sum()
print AharmvacancyT0
```

6.89595727377

Anharmonic correction of A at 100K

$$A(T_0) = A_h(T_0) + U(0) - k_B T_0 \ln \left\langle \exp\left[\frac{-(U - U_h - U(0))}{k_B T_0}\right]\right\rangle_{V,T_0,\lambda=0}$$

now we first compute $A_{anh} = -k_B T_0 \ln \langle \exp[\frac{-(U - U_h - U(0))}{\ldots}]\rangle$

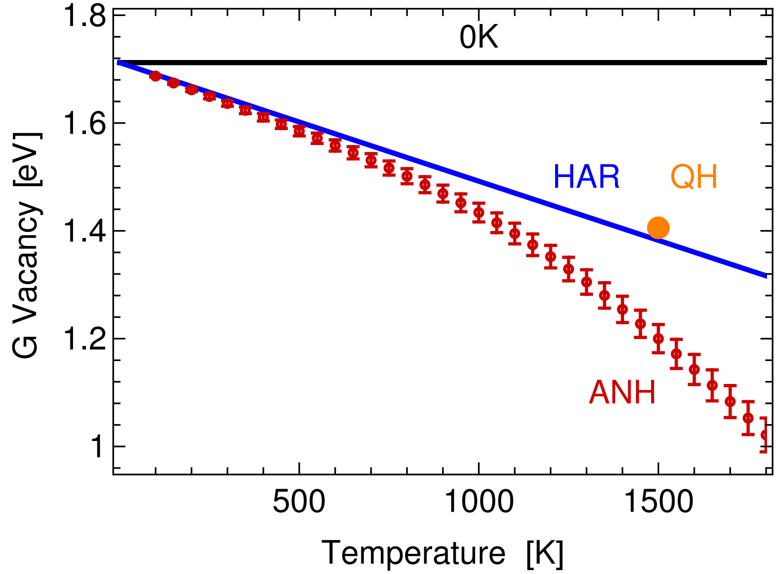
Sample input files

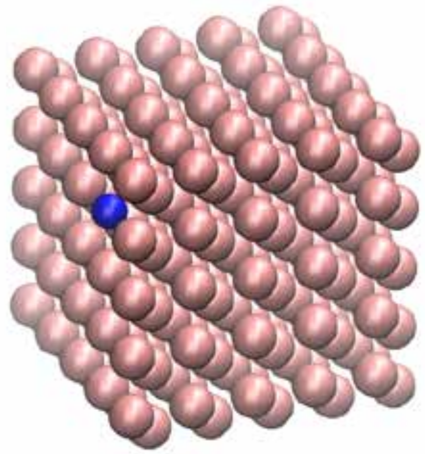
```
# Input file for Stack Fault Energy surface of Nickel
# Richard Glare, 2014
# ------------------- INITIALIZATION -------------------
clear
units           metal
dimension       3
boundary        p p p
atom_style      atomic
variable latparam1 equal 2.85532463
variable atomradius equal sqrt(3)/4
variable xdim equal 5
variable ydim equal 5
variable zdim equal 5

variable systemp equal 0

# ------------------- ATOM DEFINITION -------------------
lattice         bcc ${latparam1}
region          simbox block 0 ${xdim} 0 ${ydim} 0 ${zdim}
create_box      1 simbox
lattice bcc ${latparam1} orient x 1 0 0 orient y 0 1 0 orient z 0 0 1
create_atoms    1 region simbox

# ------------------- To create a vacancy -------------------
region vacancy sphere 1 1 1 ${atomradius} units lattice
delete_atoms region vacancy compress yes

#-------------------To create an interstitial -------------------
#create_atoms 1 single 3.0 3.3 3.5 units lattice

change_box all x final 0.0 14.37 y final 0.0 14.37 z final 0.0 14.37 remap units box

# ------------------- FORCE FIELDS -------------------
# ---------- Define Interatomic Potential -------------------
pair_style      eam/alloy
pair_coeff * * ../FeCuNi.eam.alloy Fe

neighbor     0.3 bin              ## define parameters for neighbor list
neigh_modify    delay 0 every 1 check yes

thermo 1
thermo_style custom step temp pe enthalpy lx ly lz atoms press pxx pyy pzz pxy pyz pxz

# ---------- EQUILIBRATION -------------------
fix 1 all box/relax aniso 0.0 vmax 0.0001
min_style cg
minimize 1e-20 1e-20 10000 10000
min_style sd
minimize 1e-40 1e-40 10000 10000
unfix 1

reset_timestep 0
run 1

variable PE equal pe

######################################
## SIMULATION DONE
```

```xml
- <simulation verbosity="medium">
  - <output prefix="perfect">
    - <properties stride="1" filename="out">
        [ step, time{picosecond}, conserved, temperature{kelvin}, potential, pressure_md, volume, cell_h ]
      </properties>
      <trajectory filename="pos" stride="1" bead="0" format="xyz" cell_units="angstrom"> positions{angstrom} </trajectory>
    </output>
    <total_steps>1000000</total_steps>
  - <prng>
```

Example: Vacancy formation free energy in BCC iron

$$G_v = G_{vacancy} - \frac{N_{vacancy}}{N_{perfect}} G_{perfect}$$

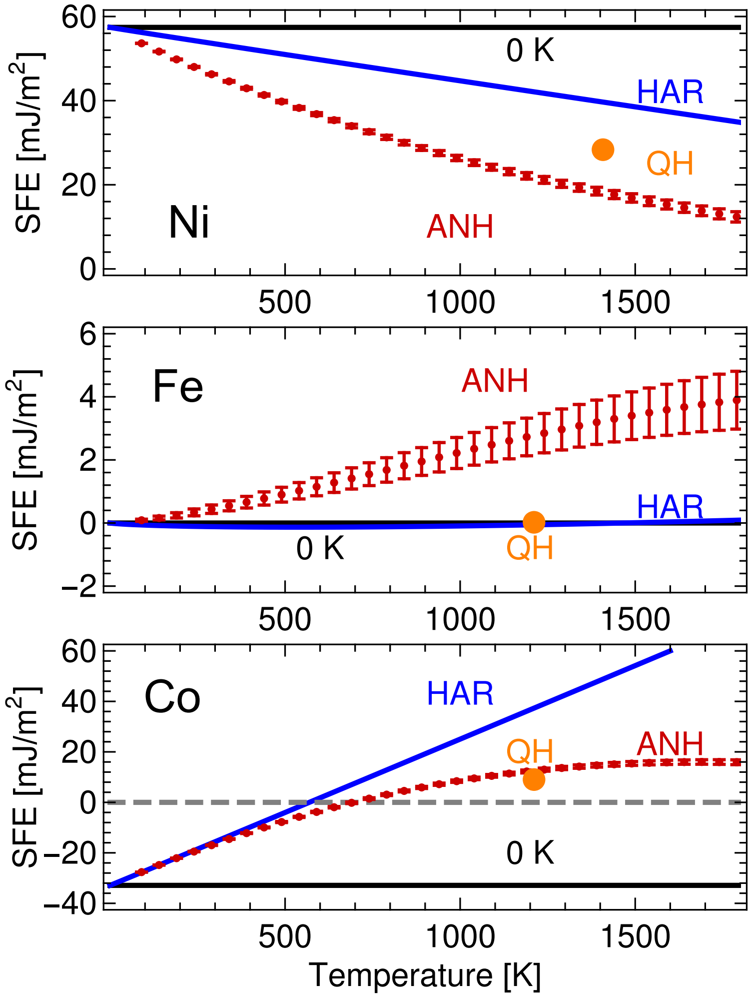


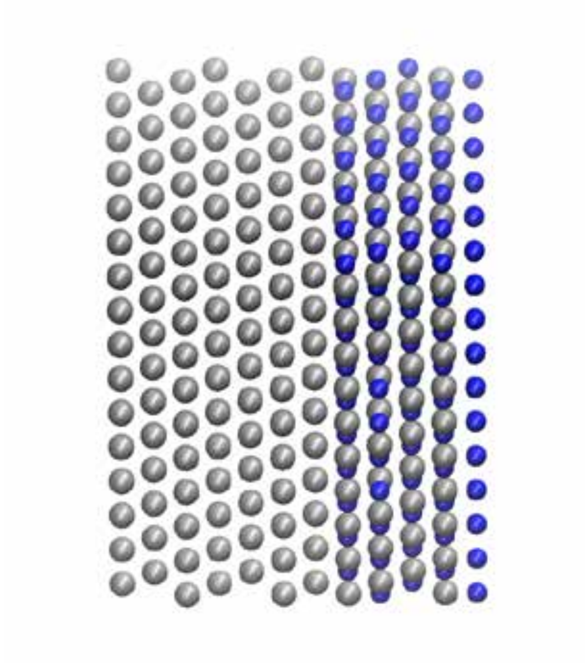


- BCC iron system using a widely used EAM potential. [M. Mendelev, S. Han, D. Srolovitz, G. Ackland, D. Sun & M. Asta 2003]
- NPT, 250 atoms for the bulk system, 249 atoms for the system with a vacancy

Example: Stacking fault free energies in FCC Ni, Fe and Co

$$\gamma_{\text{sf}} \times Area = G_{\text{sf}} - \frac{N_{\text{sf}}}{N_{\text{perfect}}} G_{\text{perfect}}$$

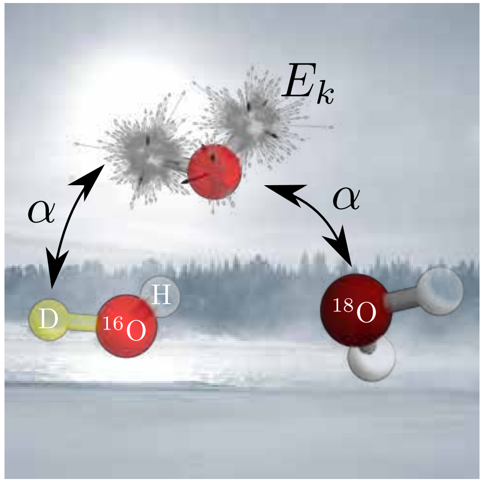


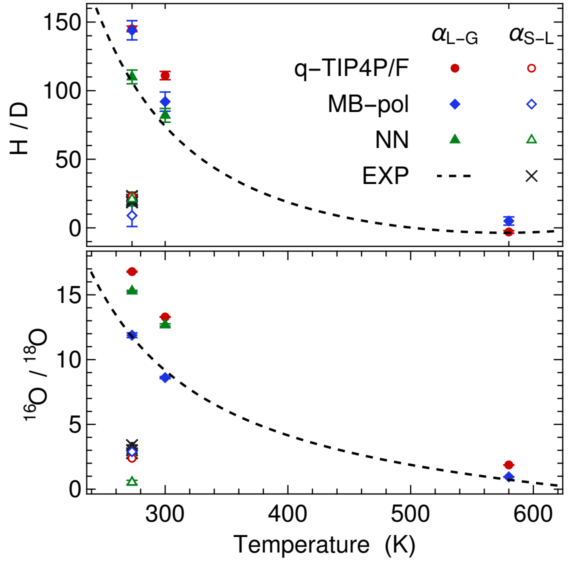


- EAM Ni: [J. A. Zimmerman, H. Gao, and F. F. Abraham 2000]
- EAM Fe: [G. Ackland, D. Bacon, A. Calder, and T. Harry 1997]
- EAM Co: [G. P. Pun and Y. Mishin 2012]

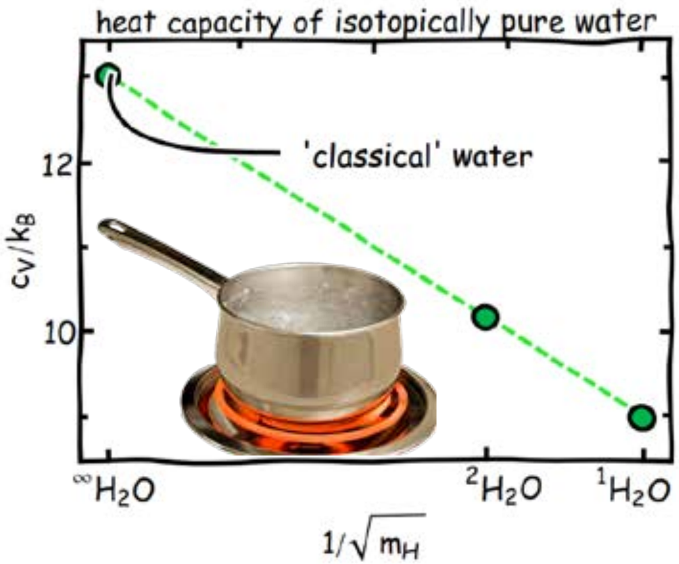
Nuclear Quantum effects

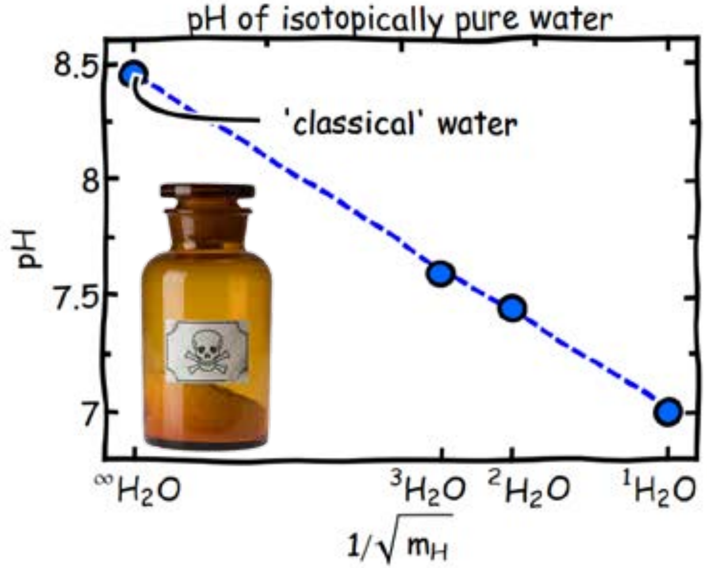
$H_2O^g + HDO^l \rightleftharpoons$
$HDO^g + H_2O^l$





- Particle momentum distribution
- Isotope fractionation
- Heat capacity
- Hydrogen bond strength
- Diffusivity...





Path-integral formalism

The density of states: $\rho(q) = \langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle$

$$\hat{K}\hat{U} \neq \hat{U}\hat{K} \rightarrow e^{-\frac{\hat{K}}{k_B T}} e^{-\frac{\hat{U}}{k_B T}}) \neq e^{-\frac{\hat{U}}{k_B T}} e^{-\frac{\hat{K}}{k_B T}}$$

Trotter expansion:

$$e^{-\frac{\hat{K}+\hat{U}}{k_B T}} = \lim_{p\rightarrow\infty} \left[e^{-\frac{\hat{U}}{2Pk_B T}} e^{-\frac{\hat{K}}{Pk_B T}} e^{-\frac{\hat{U}}{2Pk_B T}} \right]^P = \lim_{p\rightarrow\infty} \hat{\Omega}^P$$

Insert identity: $1 = \int dq \langle q | q \rangle$

$$\langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle = \lim_{p\rightarrow\infty} \int dq^{(1)} q^{(2)} \dots q^{(P-1)}$$

$$\times \langle q | \hat{\Omega} | q^{(1)} \rangle \langle q^{(1)} | \hat{\Omega} | q^{(2)} \rangle \langle q^{(2)} | \dots | q^{(P-1)} \rangle \langle q^{(P-1)} | \hat{\Omega} | q \rangle$$

$$\langle q^{(j)} | \hat{\Omega} | q^{(j+1)} \rangle = \sqrt{\frac{mPk_B T}{2\pi\hbar^2}} e^{-\frac{\hat{U}(q^{(j)})+\hat{U}(q^{(j+1)})}{2Pk_B T}} e^{-\frac{mPk_B T}{2\hbar^2}(q^{(j)}-q^{(j+1)})^2}$$

Path-integral formalism

The density of states: $\rho(q) = \langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle$

$$\hat{K}\hat{U} \neq \hat{U}\hat{K} \rightarrow e^{-\frac{\hat{K}}{k_B T}} e^{-\frac{\hat{U}}{k_B T}}) \neq e^{-\frac{\hat{U}}{k_B T}} e^{-\frac{\hat{K}}{k_B T}}$$

Trotter expansion:

$$e^{-\frac{\hat{K}+\hat{U}}{k_B T}} = \lim_{p\rightarrow\infty} \left[e^{-\frac{\hat{U}}{2Pk_B T}} e^{-\frac{\hat{K}}{Pk_B T}} e^{-\frac{\hat{U}}{2Pk_B T}} \right]^P = \lim_{p\rightarrow\infty} \hat{\Omega}^P$$

Insert identity: $1 = \int dq \langle q | q \rangle$

$\langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle = \lim_{p\rightarrow\infty} \int dq^{(1)} q^{(2)} \ldots q^{(P-1)}$

$$\times \langle q | \hat{\Omega} | q^{(1)} \rangle \langle q^{(1)} | \hat{\Omega} | q^{(2)} \rangle \langle q^{(2)} | \ldots | q^{(P-1)} \rangle \langle q^{(P-1)} | \hat{\Omega} | q \rangle$$

$$\langle q^{(j)} | \hat{\Omega} | q^{(j+1)} \rangle = \sqrt{\frac{mPk_B T}{2\pi\hbar^2}} e^{-\frac{\hat{U}(q^{(j)})+\hat{U}(q^{(j+1)})}{2Pk_B T}} e^{-\frac{mPk_B T}{2\hbar^2}(q^{(j)}-q^{(j+1)})^2}$$

Path-integral formalism

The density of states: $\rho(q) = \langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle$

$$\hat{K}\hat{U} \neq \hat{U}\hat{K} \rightarrow e^{-\frac{\hat{K}}{k_B T}} e^{-\frac{\hat{U}}{k_B T}}) \neq e^{-\frac{\hat{U}}{k_B T}} e^{-\frac{\hat{K}}{k_B T}}$$

Trotter expansion:

$$e^{-\frac{\hat{K}+\hat{U}}{k_B T}} = \lim_{p\rightarrow\infty} \left[e^{-\frac{\hat{U}}{2Pk_B T}} e^{-\frac{\hat{K}}{Pk_B T}} e^{-\frac{\hat{U}}{2Pk_B T}} \right]^P = \lim_{p\rightarrow\infty} \hat{\Omega}^P$$

Insert identity: $1 = \int dq \langle q | q \rangle$

$$\langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle = \lim_{p\rightarrow\infty} \int dq^{(1)} q^{(2)} \dots q^{(P-1)}$$

$$\times \langle q | \hat{\Omega} | q^{(1)} \rangle \langle q^{(1)} | \hat{\Omega} | q^{(2)} \rangle \langle q^{(2)} | \dots | q^{(P-1)} \rangle \langle q^{(P-1)} | \hat{\Omega} | q \rangle$$

$$\langle q^{(j)} | \hat{\Omega} | q^{(j+1)} \rangle = \sqrt{\frac{mPk_B T}{2\pi\hbar^2}} e^{-\frac{\hat{U}(q^{(j)})+\hat{U}(q^{(j+1)})}{2Pk_B T}} e^{-\frac{mPk_B T}{2\hbar^2}(q^{(j)}-q^{(j+1)})^2}$$

Path-integral formalism

The density of states: $\rho(q) = \langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle$

$$\hat{K}\hat{U} \neq \hat{U}\hat{K} \rightarrow e^{-\frac{\hat{K}}{k_B T}} e^{-\frac{\hat{U}}{k_B T}}) \neq e^{-\frac{\hat{U}}{k_B T}} e^{-\frac{\hat{K}}{k_B T}}$$

Trotter expansion:

$$e^{-\frac{\hat{K}+\hat{U}}{k_B T}} = \lim_{p\to\infty} \left[e^{-\frac{\hat{U}}{2Pk_B T}} e^{-\frac{\hat{K}}{Pk_B T}} e^{-\frac{\hat{U}}{2Pk_B T}} \right]^P = \lim_{p\to\infty} \hat{\Omega}^P$$

Insert identity: $1 = \int dq \langle q|q \rangle$

$$\langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle = \lim_{p\to\infty} \int dq^{(1)} q^{(2)} \ldots q^{(P-1)}$$

$$\times \langle q|\hat{\Omega}|q^{(1)}\rangle \langle q^{(1)}|\hat{\Omega}|q^{(2)}\rangle \langle q^{(2)}| \ldots |q^{(P-1)}\rangle \langle q^{(P-1)}|\hat{\Omega}|q\rangle$$

$$\langle q^{(j)}|\hat{\Omega}|q^{(j+1)}\rangle = \sqrt{\frac{mPk_B T}{2\pi\hbar^2}} e^{-\frac{\hat{U}(q^{(j)})+\hat{U}(q^{(j+1)})}{2Pk_B T}} e^{-\frac{mPk_B T}{2\hbar^2}(q^{(j)}-q^{(j+1)})^2}$$

Path-integral formalism

The density of states: $\rho(q) = \langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle$

$$\hat{K}\hat{U} \neq \hat{U}\hat{K} \rightarrow e^{-\frac{\hat{K}}{k_B T}} e^{-\frac{\hat{U}}{k_B T}}) \neq e^{-\frac{\hat{U}}{k_B T}} e^{-\frac{\hat{K}}{k_B T}}$$

Trotter expansion:

$$e^{-\frac{\hat{K}+\hat{U}}{k_B T}} = \lim_{p\rightarrow\infty} \left[e^{-\frac{\hat{U}}{2Pk_B T}} e^{-\frac{\hat{K}}{Pk_B T}} e^{-\frac{\hat{U}}{2Pk_B T}} \right]^P = \lim_{p\rightarrow\infty} \hat{\Omega}^P$$

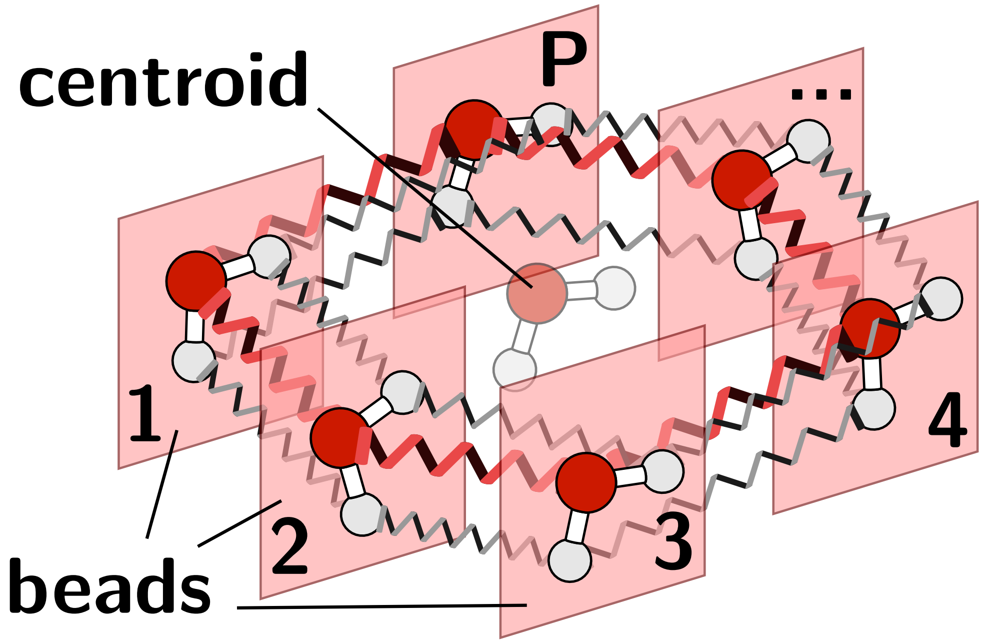
Insert identity: $1 = \int dq \langle q | q \rangle$

$$\langle q | e^{-\frac{\hat{K}+\hat{U}}{k_B T}} | q \rangle = \lim_{p\rightarrow\infty} \int dq^{(1)} q^{(2)} \ldots q^{(P-1)}$$

$$\times \langle q | \hat{\Omega} | q^{(1)} \rangle \langle q^{(1)} | \hat{\Omega} | q^{(2)} \rangle \langle q^{(2)} | \ldots | q^{(P-1)} \rangle \langle q^{(P-1)} | \hat{\Omega} | q \rangle$$

$$\langle q^{(j)} | \hat{\Omega} | q^{(j+1)} \rangle = \sqrt{\frac{mPk_B T}{2\pi\hbar^2}} e^{-\frac{\hat{U}(q^{(j)})+\hat{U}(q^{(j+1)})}{2Pk_B T}} e^{-\frac{mPk_B T}{2\hbar^2}(q^{(j)}-q^{(j+1)})^2}$$

Ring polymer molecular dynamics



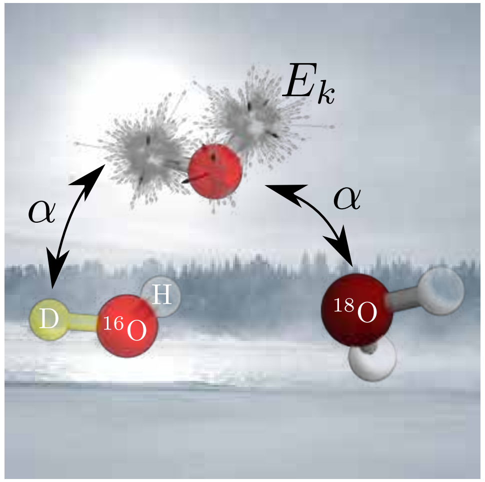
Isomorphism between a quantum mechanical particle and a ring polymer connected by harmonic springs. The Hamiltonian can be expressed as

$$H(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{P} \frac{[\mathbf{p}^{(j)}]^2}{2m} + V(\mathbf{q}^{(j)}) + \frac{1}{2}m(\frac{Pk_BT}{\hbar})^2[\mathbf{q}^{(\mathbf{j})} - \mathbf{q}^{(\mathbf{j-1})}]^2$$

$$\mathbf{q}^{(\mathbf{0})} = \mathbf{q}^{(\mathbf{P})}$$

[Tuckerman, Statistical Mechanics]

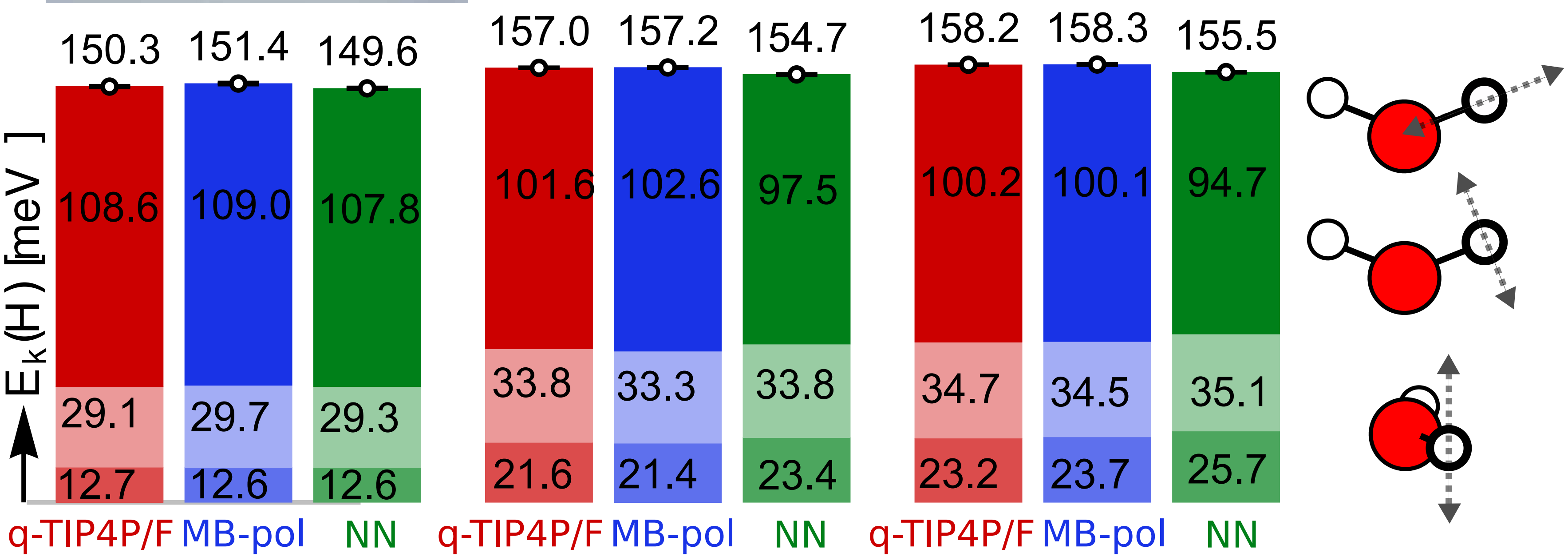
Quantum kinetic energy



[Cheng & Ceriotti JCP 2014;
Cheng, Behler & Ceriotti JPCL 2016]

$$\frac{dG}{du} = -\frac{\langle E_k(u)\rangle}{u}$$

$$G_{qm} - G_{cl} = \int_m^\infty du \frac{\langle E_k(u)\rangle}{u}$$

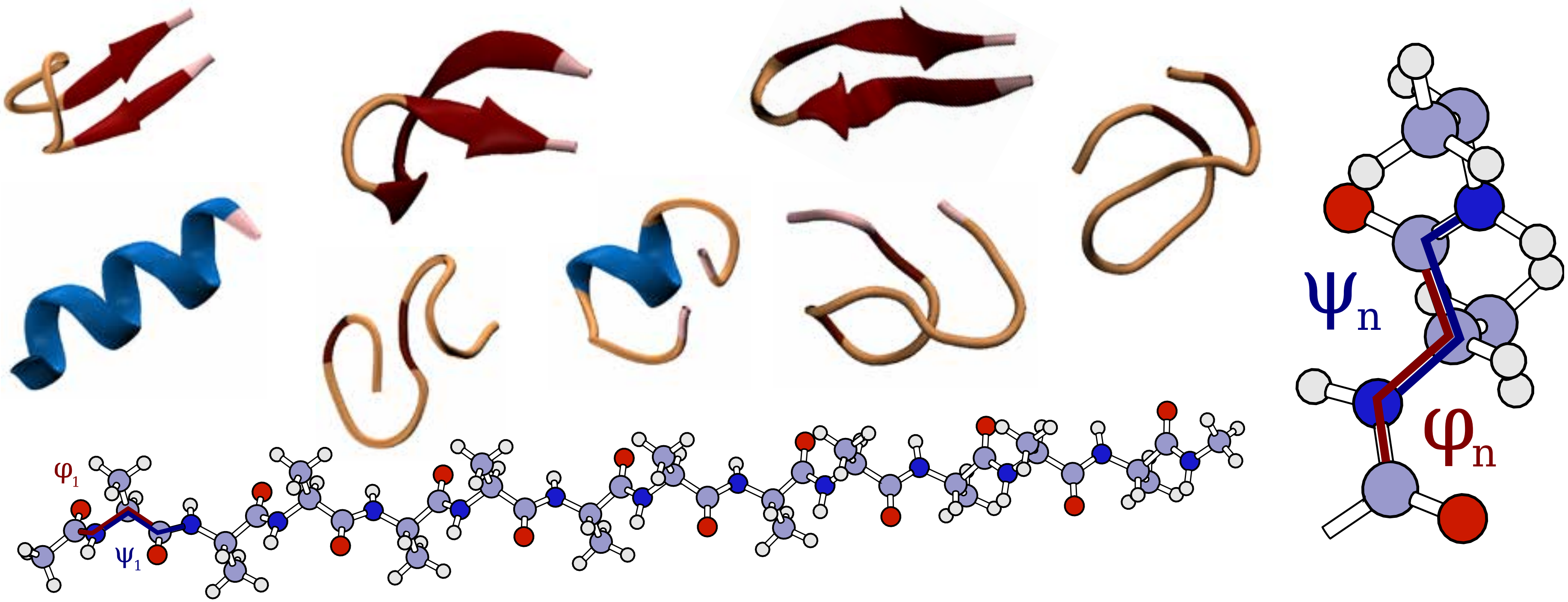


Outline

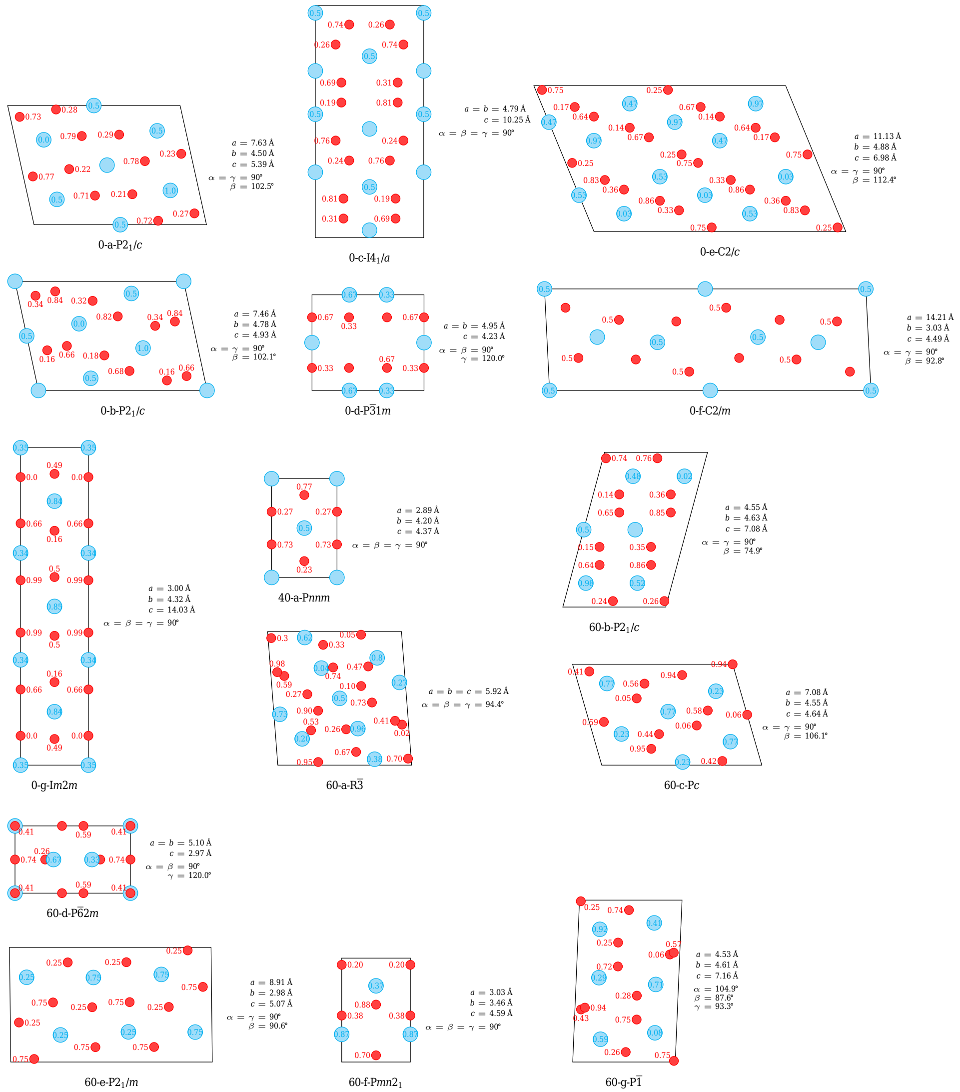
What we will talk about:

- Statistical mechanics & molecular dynamics 101.
  - Metadynamics
  - Thermodynamic integration
  - Nuclear quantum effects (NQEs)
- Translating materials and molecules into matrices.
  - Representations
  - Dimensionality reduction
- Introduction to machine learning potentials.

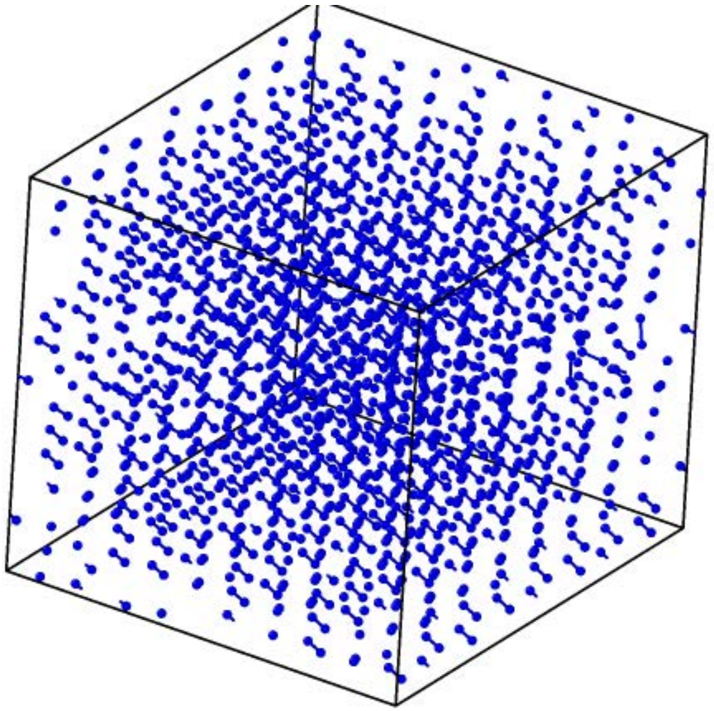
Molecules and materials live in high-D space



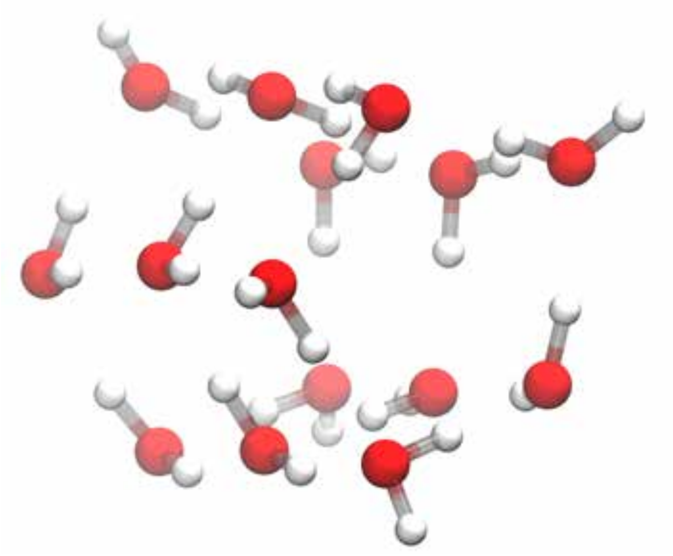
Molecules and materials live in high-D space



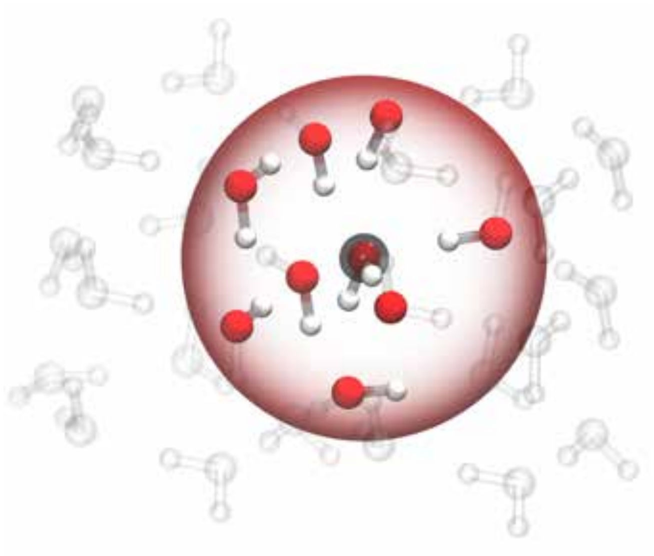
Molecules and materials live in high-D space



Divide system into atomistic environments



A



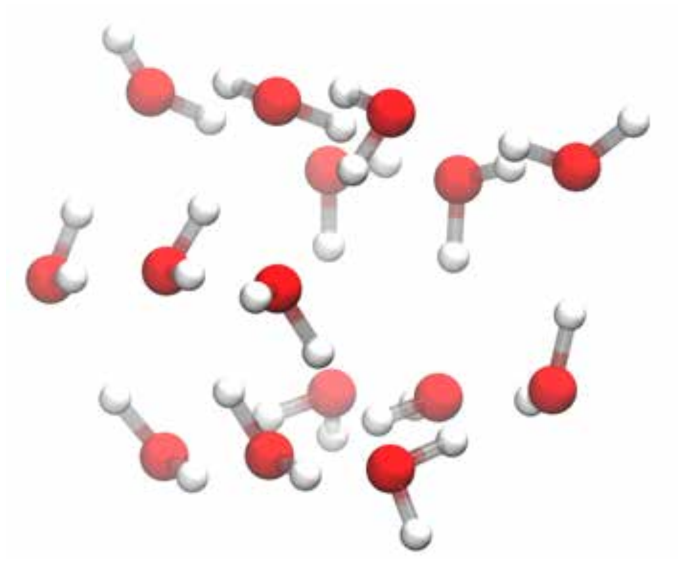
$\mathcal{X}_i$

The first step is to divide the system into a set of atomic environments.
So the task becomes representing atomic environments.

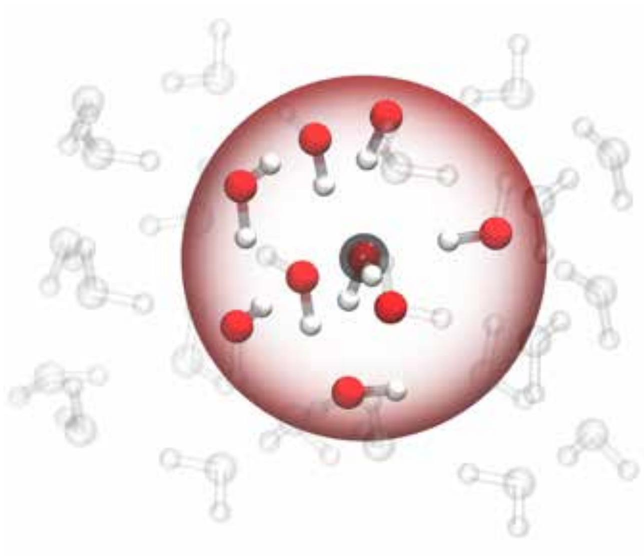
Popular representations (Invariant with respect to translation, rotation and permutation.):

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]

Divide system into atomistic environments



$\Phi \leftarrow \{\Psi(\mathcal{X}_i)\}$



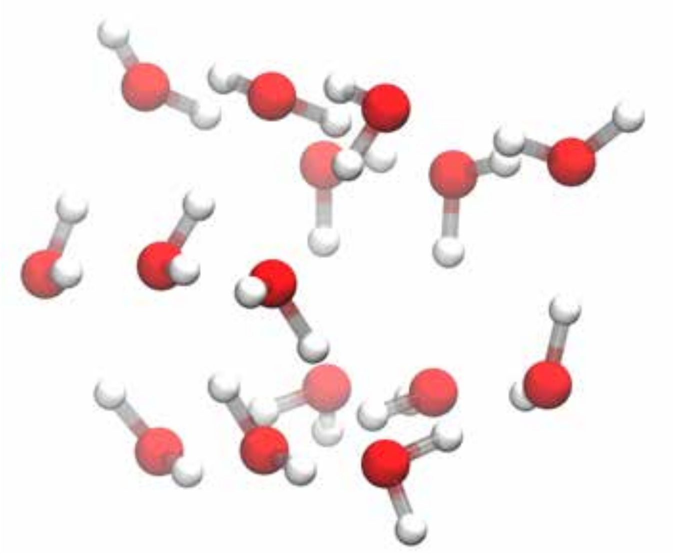
$\Psi(\mathcal{X}_i)$

The first step is to divide the system into a set of atomic environments.
So the task becomes representing atomic environments.

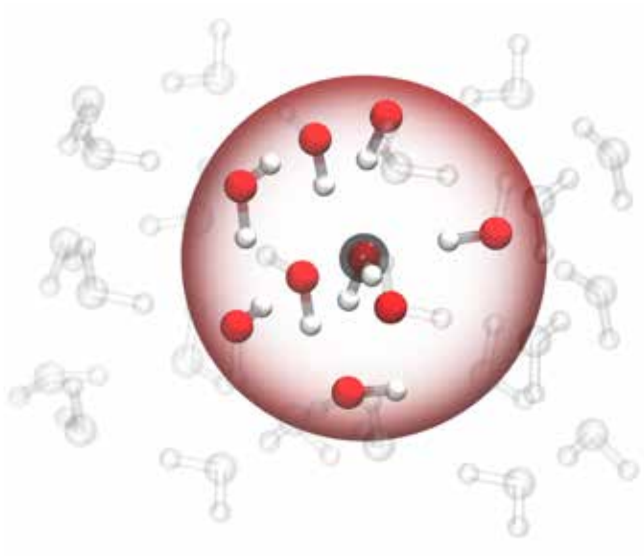
Popular representations (Invariant with respect to translation, rotation and permutation.):

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]

Divide system into atomistic environments



$$O(A) = F(\Phi(A))$$



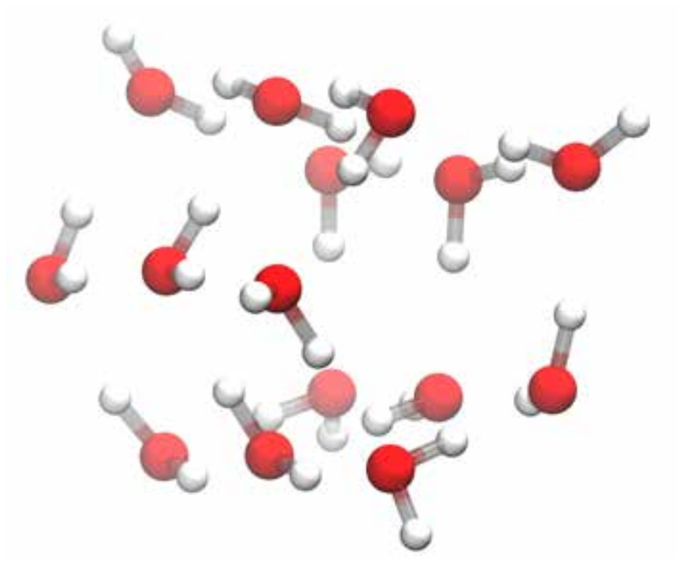
$$O_i = f(\Psi(\mathcal{X}_i))$$

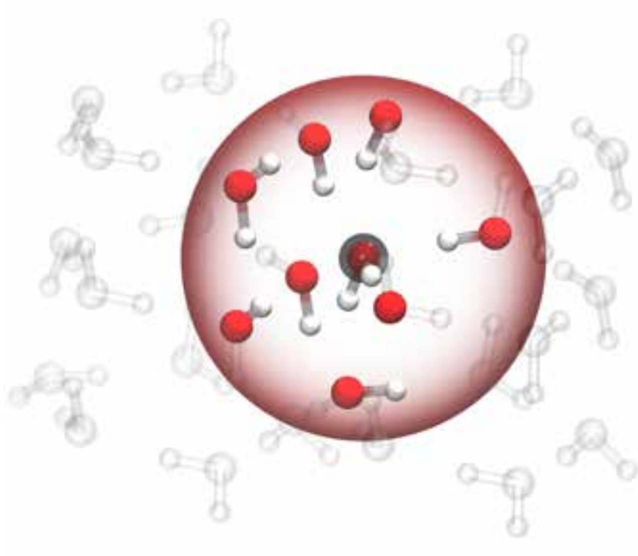
The first step is to divide the system into a set of atomic environments.
So the task becomes representing atomic environments.

Popular representations (Invariant with respect to translation, rotation and permutation.):

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]

Divide system into atomistic environments





$$E(A) = \sum E_i \qquad E_i = e(\Psi(\mathcal{X}_i))$$

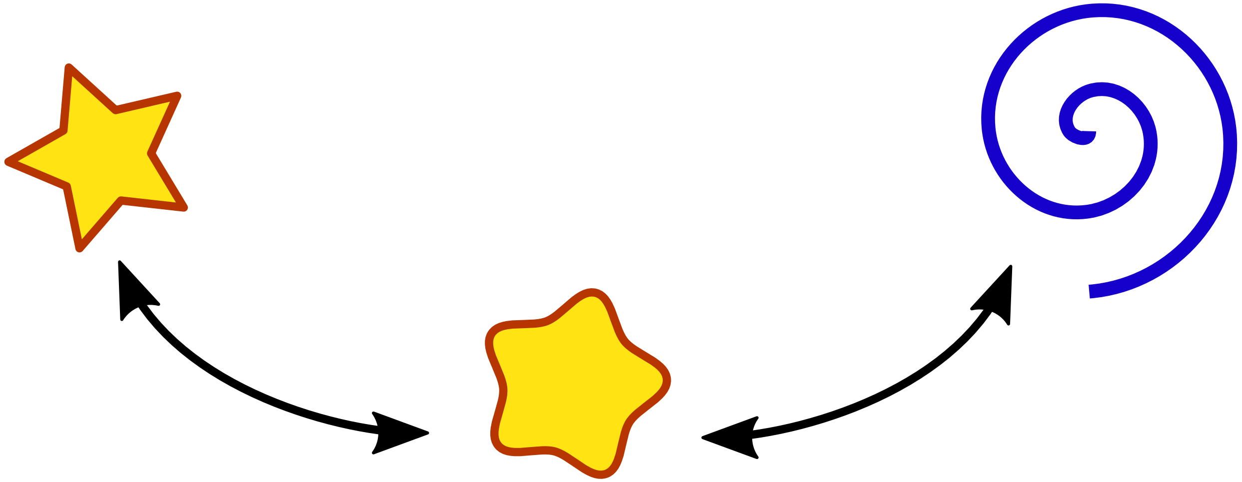
The first step is to divide the system into a set of atomic environments.
So the task becomes representing atomic environments.

Popular representations (Invariant with respect to translation, rotation and permutation.):

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

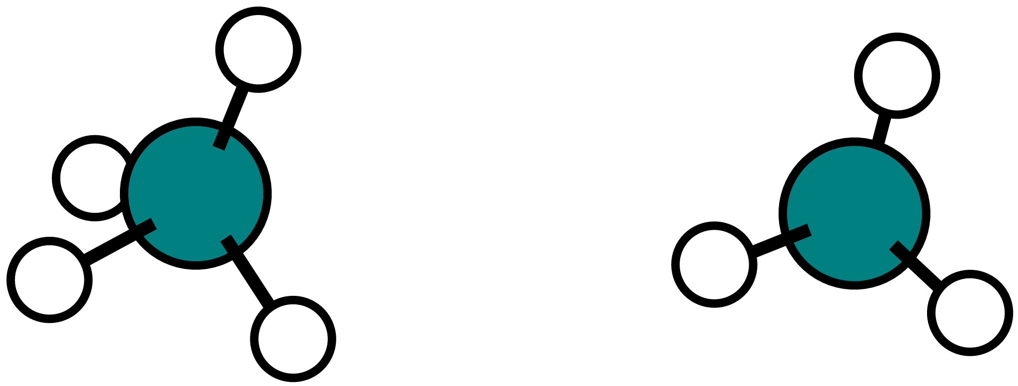


$$K(\mathcal{X}, \mathcal{X}') \approx 1 \qquad K(\mathcal{X}, \mathcal{X}') \approx 0$$

$$D(\mathcal{X}, \mathcal{X}') \approx 0 \qquad D(\mathcal{X}, \mathcal{X}') \gg 0$$

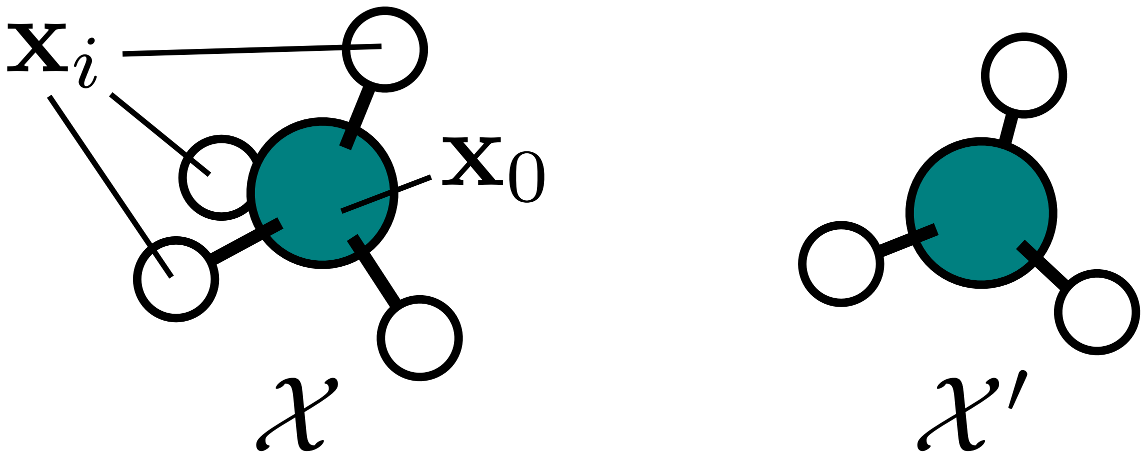
Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]



Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]



$$\{\mathbf{x}_i - \mathbf{x}_0\} \leftrightarrow \mathcal{X}$$

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

$$\rho_\alpha(\mathbf{x}) = \sum_{i\in\alpha} g(\mathbf{x} - \mathbf{x}_i)$$

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

$$k(\mathcal{X}, \mathcal{X}') = \int \rho(\mathbf{x}) \rho'(\mathbf{x})$$

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

$$k(\mathcal{X}, \mathcal{X}') = \int \mathrm{d}\hat{R} \left| \int \rho(\mathbf{x}) \rho'(\hat{R}\mathbf{x}) \right|^2$$

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

$$k(\mathcal{X},\mathcal{X}') = \int \mathrm{d}\hat{R} \left| \int \rho(\mathbf{x})\rho'(\hat{R}\mathbf{x}) \right|^2$$

Representing atomic environments

[Bartók, Kondor & Csányi PRB 2013]

$$\kappa(\mathcal{X},\mathcal{X}') = \int \mathrm{d}\hat{R} |\rho(\mathbf{x})\rho'(\hat{R}\mathbf{x})|^2$$

$$\rho(\mathbf{x}) = \sum_{nlm} c_{nlm} g_n(|r|) Y_{lm}(\hat{r})$$

$$k_{nn'l}(\mathcal{X}) = \pi\sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^* c_{n'lm}$$

The list of vector $\{k_{nn'l}(\mathcal{X})\}$ is the descriptor of the atomic environment $\mathcal{X}$.

Similarity measurement between structures

- The kernel matrix $\{K\}$ records the similarity measurement for each pair of structures in the data set.
- The kernel function $K(A, B)$ for structure A and B is

$$K(A,B) = \Phi(A)^{\mathsf{T}}\Phi(B) = \sum_{i=1}^{M} \phi_i(A)\phi_i(B)$$

- Global features are constructed from local features by taking the average:

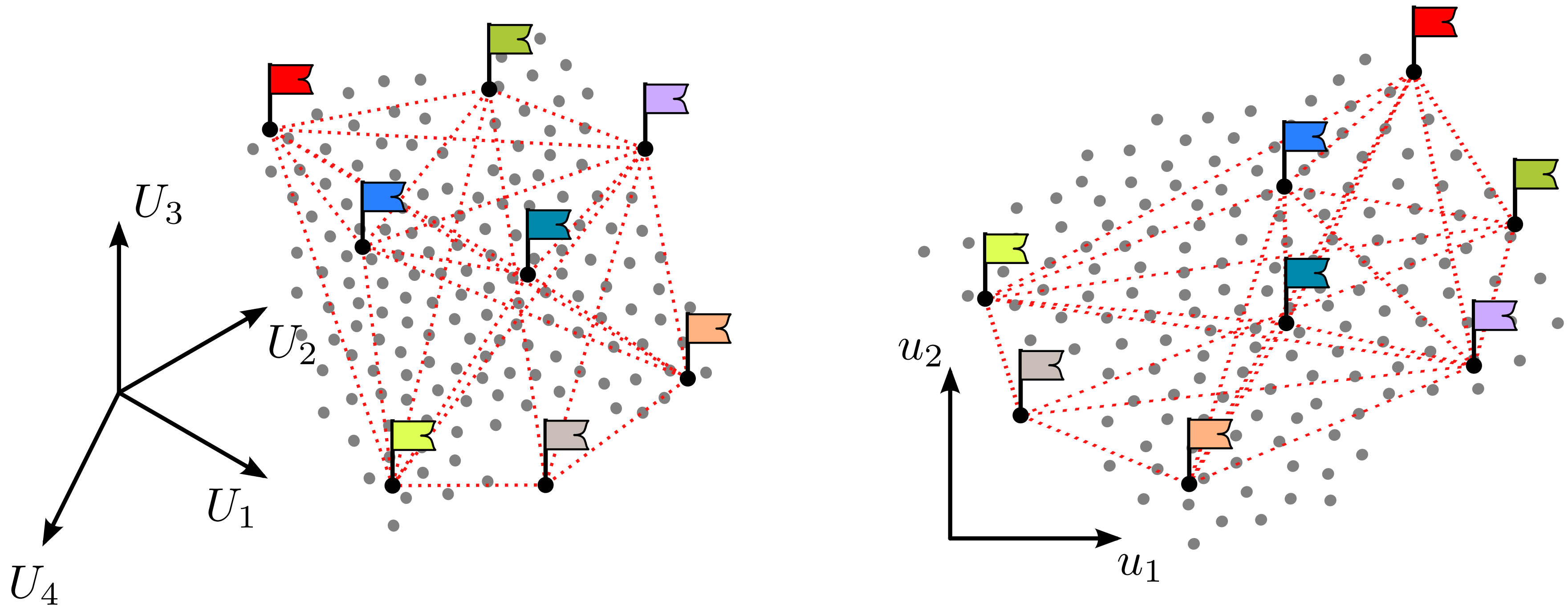
$$\Phi(A) = \frac{1}{N_A}\sum_{n=1}^{N_A} \Psi(\mathcal{X}_n^A)$$

- Other choices available...

ML methods to apply to the design matrix

- Build low-dimensional map using dimensionality reduction (e.g. PCA)
- Sparsity the train set using farthest point sampling, CUR or uniform sampling
- Clustering
- Regression (Kernel ridge regression (KRR), neural networks)

Dimensionality reduction



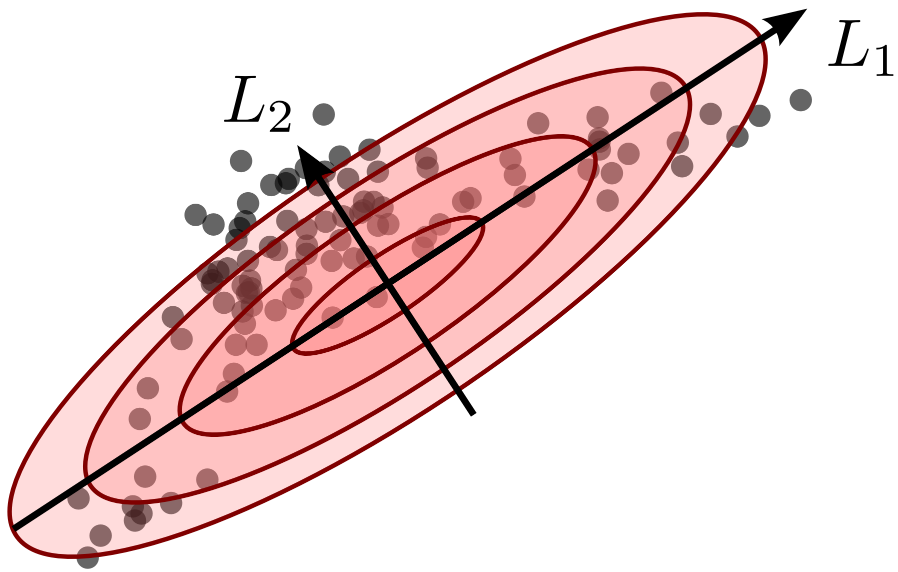
$\{\mathbf{X}_i\}^D$ $\{\mathbf{x}_i\}^d$

(k)PCA, MDS, t-SNE/UMAP

Principal component analysis

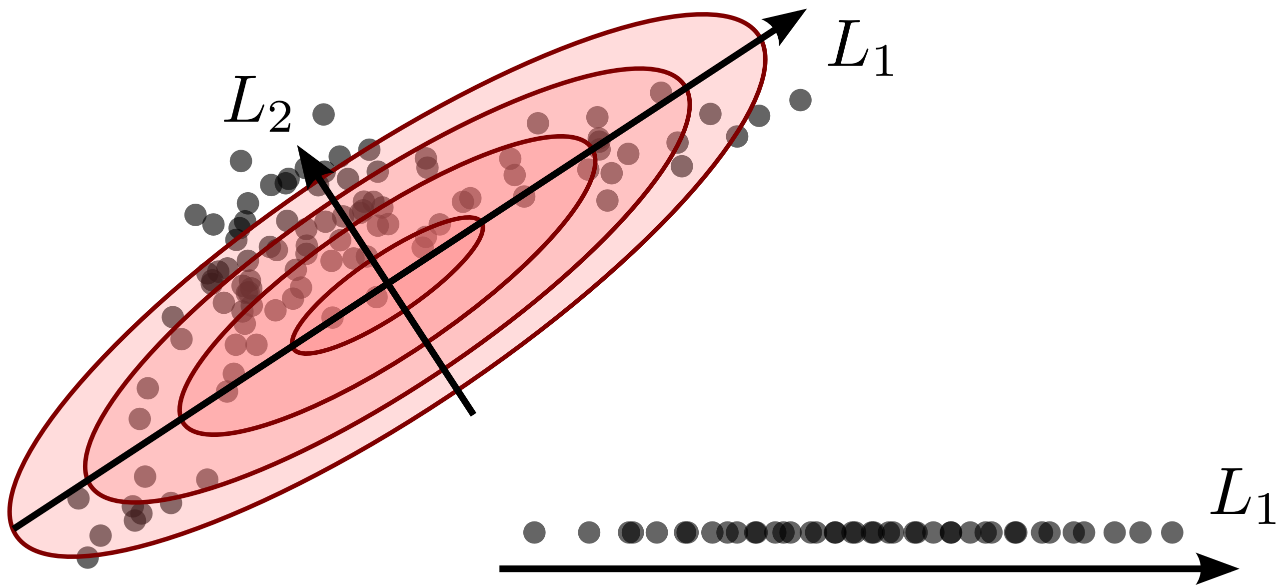
Question:
What is preserved during PCA?

Principal component analysis



Question:
What is preserved during PCA?

Principal component analysis



Question:
What is preserved during PCA?

Principal component analysis

PCA identifies the axis that accounts for the largest amount of variance in the data set.

- $\{\mathbf{X}_i\}^D$: data in the D-dimensional space
- $\{\mathbf{x}_i\}^d$: linear projection in the low d dimensional space
- $\mathbf{c}$: normalized projection matrix

$$\mathbf{x}_i = \mathbf{X}_i \mathbf{c}$$

- Covariance of the data: $\mathbf{C} = \mathbf{X}^T \mathbf{X}$
- Covariance of the projected data: $\mathbf{x}^T \mathbf{x}$

Principal component analysis

PCA identifies the axis that accounts for the largest amount of variance in the data set.

- $\{\mathbf{X}_i\}^D$: data in the D-dimensional space
- $\{\mathbf{x}_i\}^d$: linear projection in the low d dimensional space
- $\mathbf{c}$: normalized projection matrix

$$\mathbf{x}_i = \mathbf{X}_i \mathbf{c}$$

- Covariance of the data: $\mathbf{C} = \mathbf{X}^T \mathbf{X}$
- Covariance of the projected data: $\mathbf{x}^T \mathbf{x}$

Given d , how to reserve the largest amount of variance?

Principal component analysis

PCA identifies the axis that accounts for the largest amount of variance in the data set.

- $\{\mathbf{X}_i\}^D$: data in the D-dimensional space
- $\{\mathbf{x}_i\}^d$: linear projection in the low d dimensional space
- $\mathbf{c}$: normalized projection matrix

$$\mathbf{x}_i = \mathbf{X}_i\mathbf{c}$$

- Covariance of the data: $\mathbf{C} = \mathbf{X}^T\mathbf{X}$
- Covariance of the projected data: $\mathbf{x}^T\mathbf{x}$

Keep the first d eigenvectors of the covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$

Principal component analysis

- The covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$: $D \times D$ form.
- Eigenvalues $\{\lambda^j\}$
- Corresponding eigenvectors $\{\mathbf{v}^j\}$ of the matrix

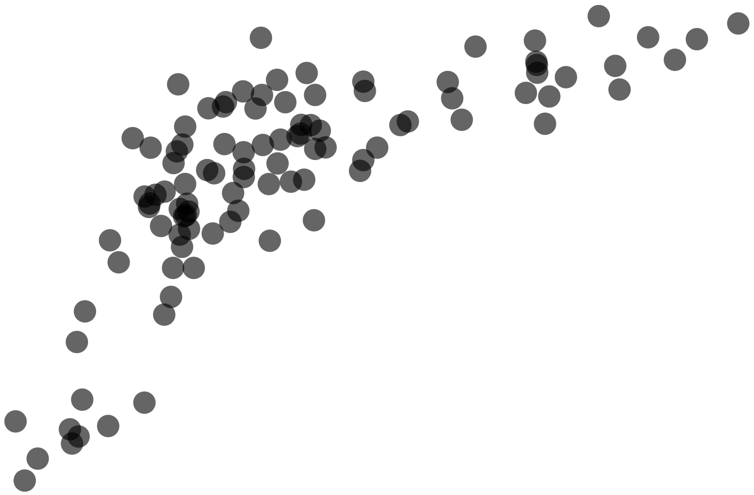
Eigenvalues and eigenvectors fulfills

$$\mathbf{C}\mathbf{v}^j = \lambda^j \mathbf{v}^j$$

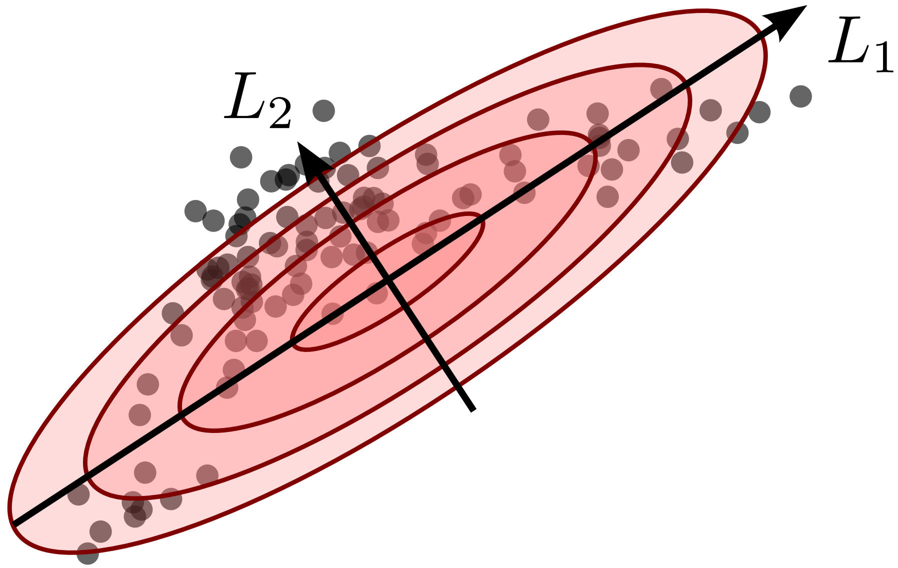
for $j = 1 \ldots D$.
One can find the eigenvalues $\{\lambda^j\}$ by solving

$$det(\mathbf{C} - \lambda \mathbf{I}) = 0$$

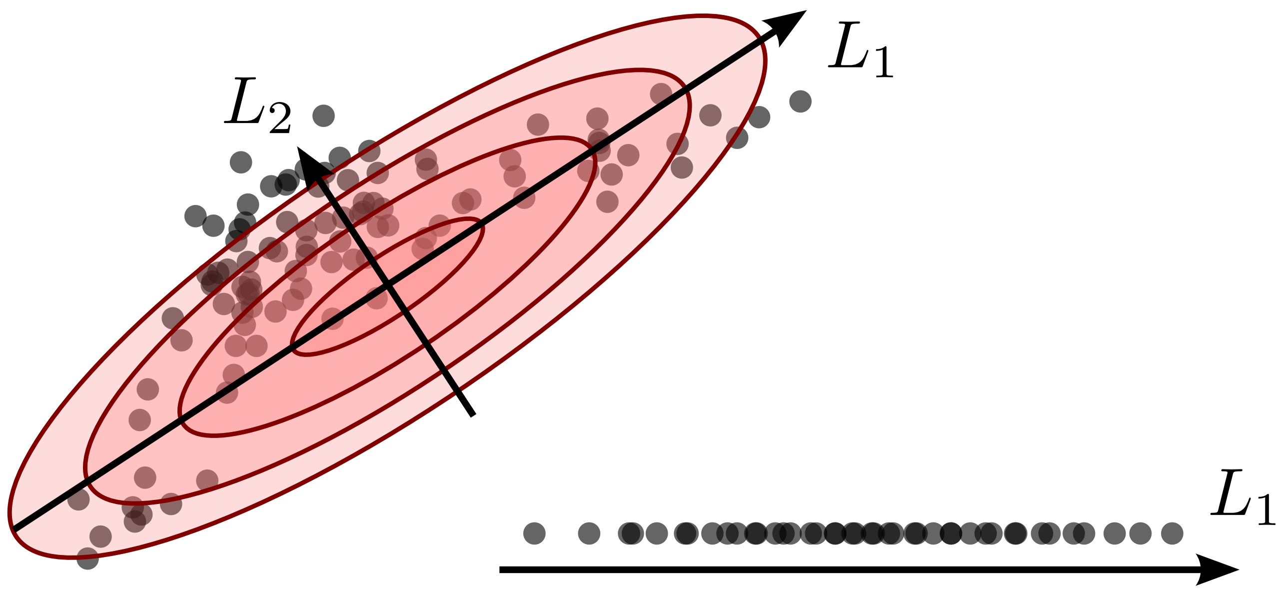
Principal component analysis



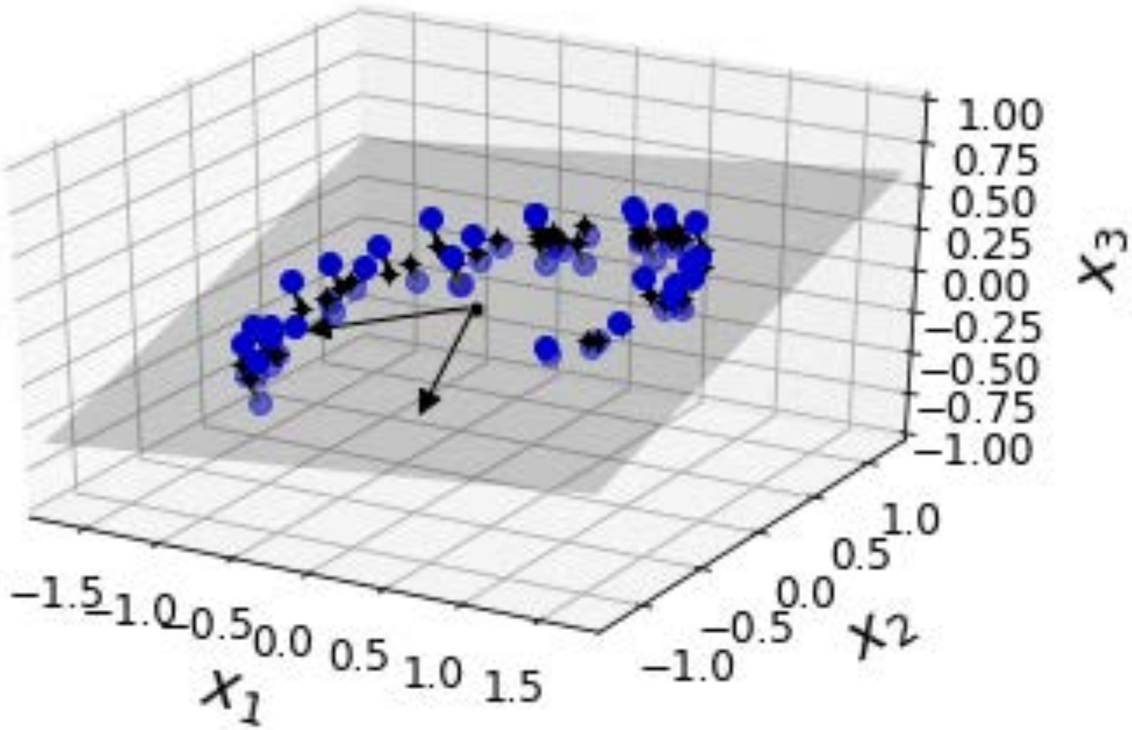
Principal component analysis

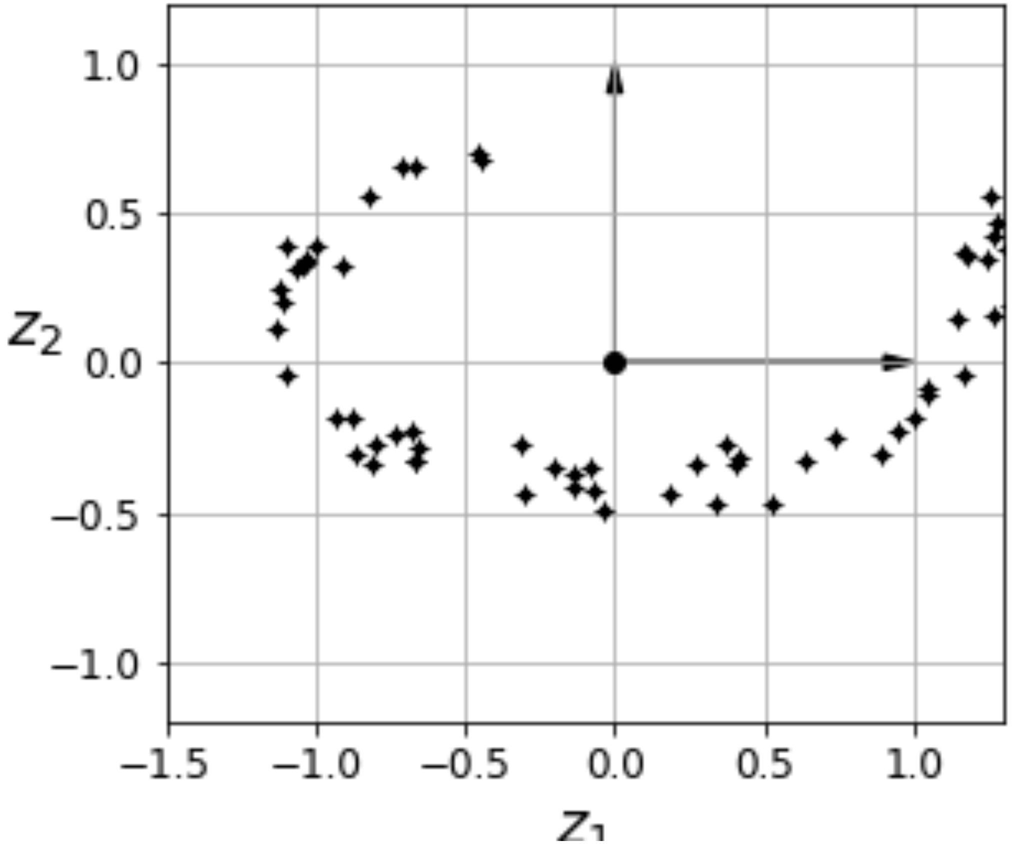


Principal component analysis



Principal component analysis




1.0
0.5
0.0
−0.5
−1.0
z_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
z_1

https://github.com/BingqingCheng/ASAP

Contributors: Ryan-Rhys Griffiths, Tamas Stenczel, Bonan Zhu, Felix Faber, Noam Bernstein

ASAP

Automatic Selection And Prediction tools for materials and molecules

Basic usage

Type `asap` and use the sub-commands for various tasks.

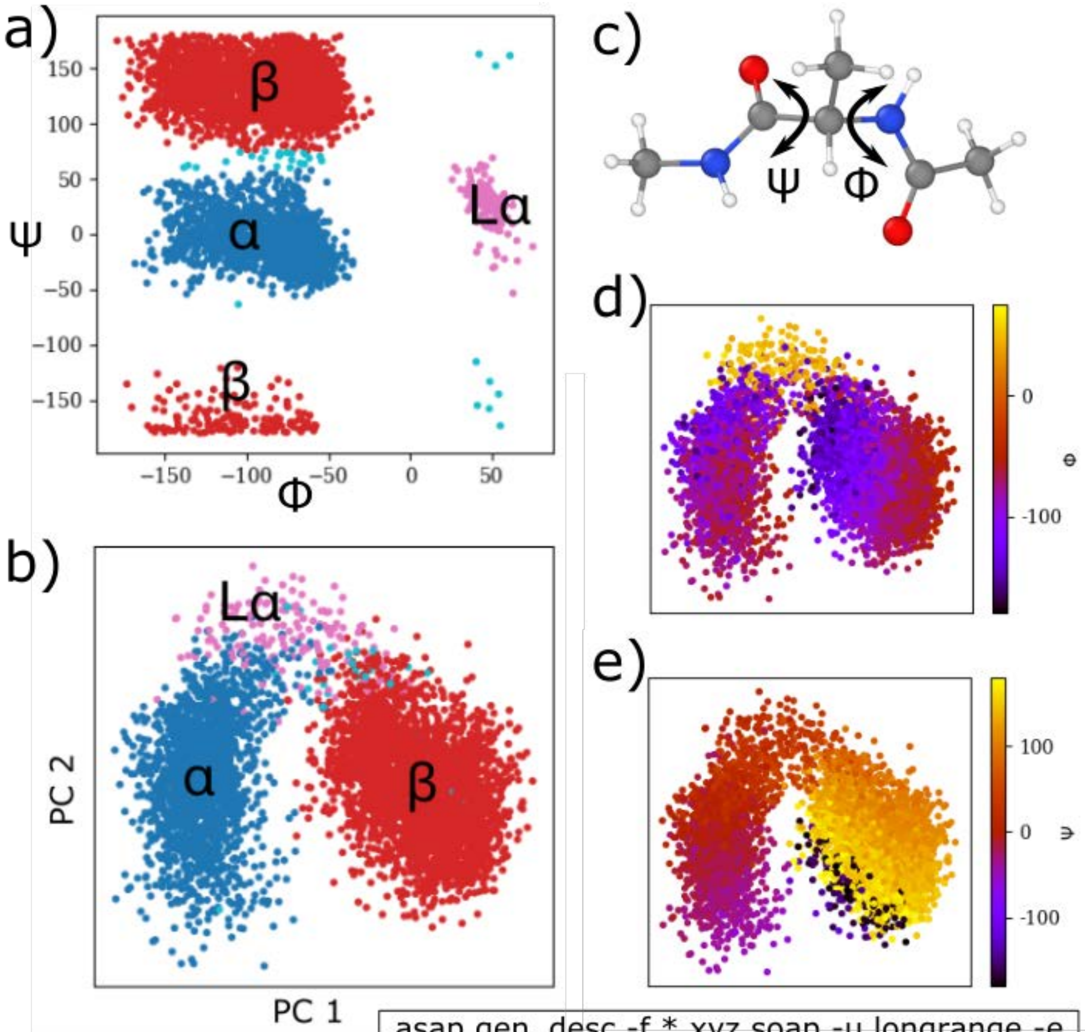
- Low-dimensional embedding, regression
- Sparsification
- Clustering, kernel density estimation

https://github.com/BingqingCheng/ASAP

1. Generate descriptor matrix
   ‘asap gen_desc -f *.xyz soap‘
2. Make map
   ‘asap map –f *.xyz -dm [*] pca‘
3. Other tasks: regression, clustering, sparsification, kernel density estimation, e.g.
   ‘asap cluster‘, ‘asap fit‘, ‘asap kde‘, ‘asap select‘

KPCA map for alanine dipeptide

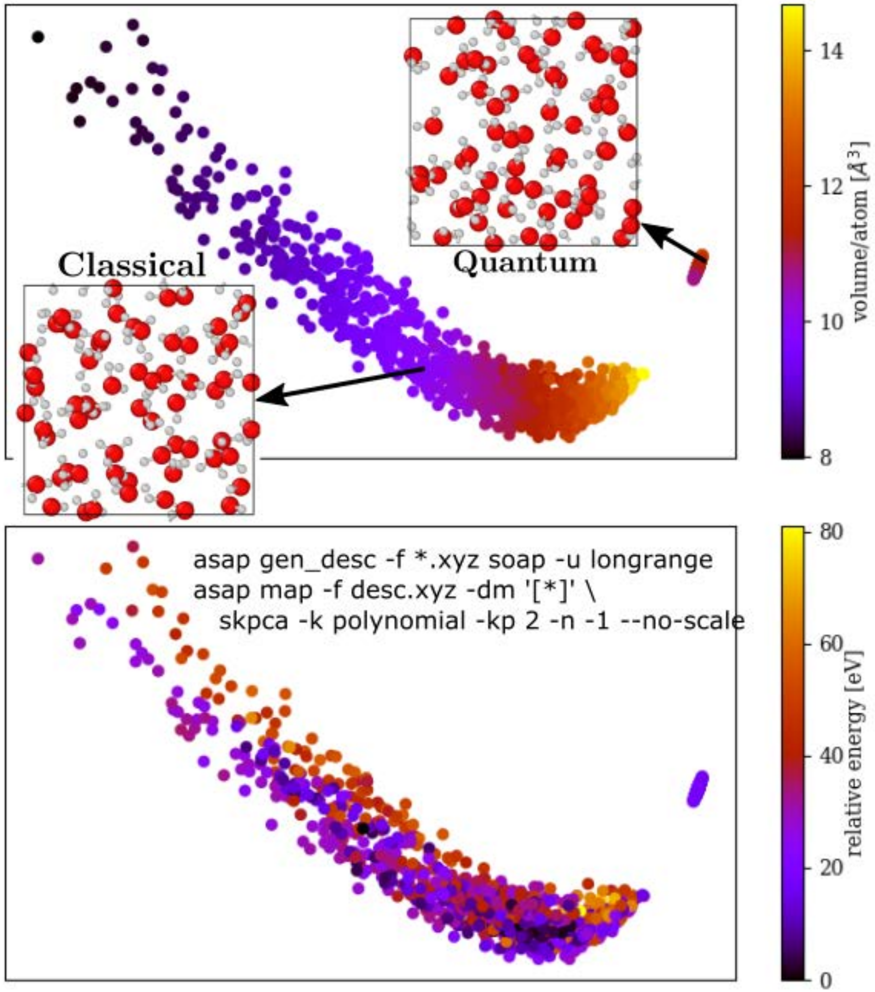
[Cheng et al. Accounts of Chemical Research 2020]




```
asap gen_desc -f *.xyz soap -u longrange -e
asap map -f desc.xyz -dm '[*]' skpca --no-scale -k polynomial -kp 2
```

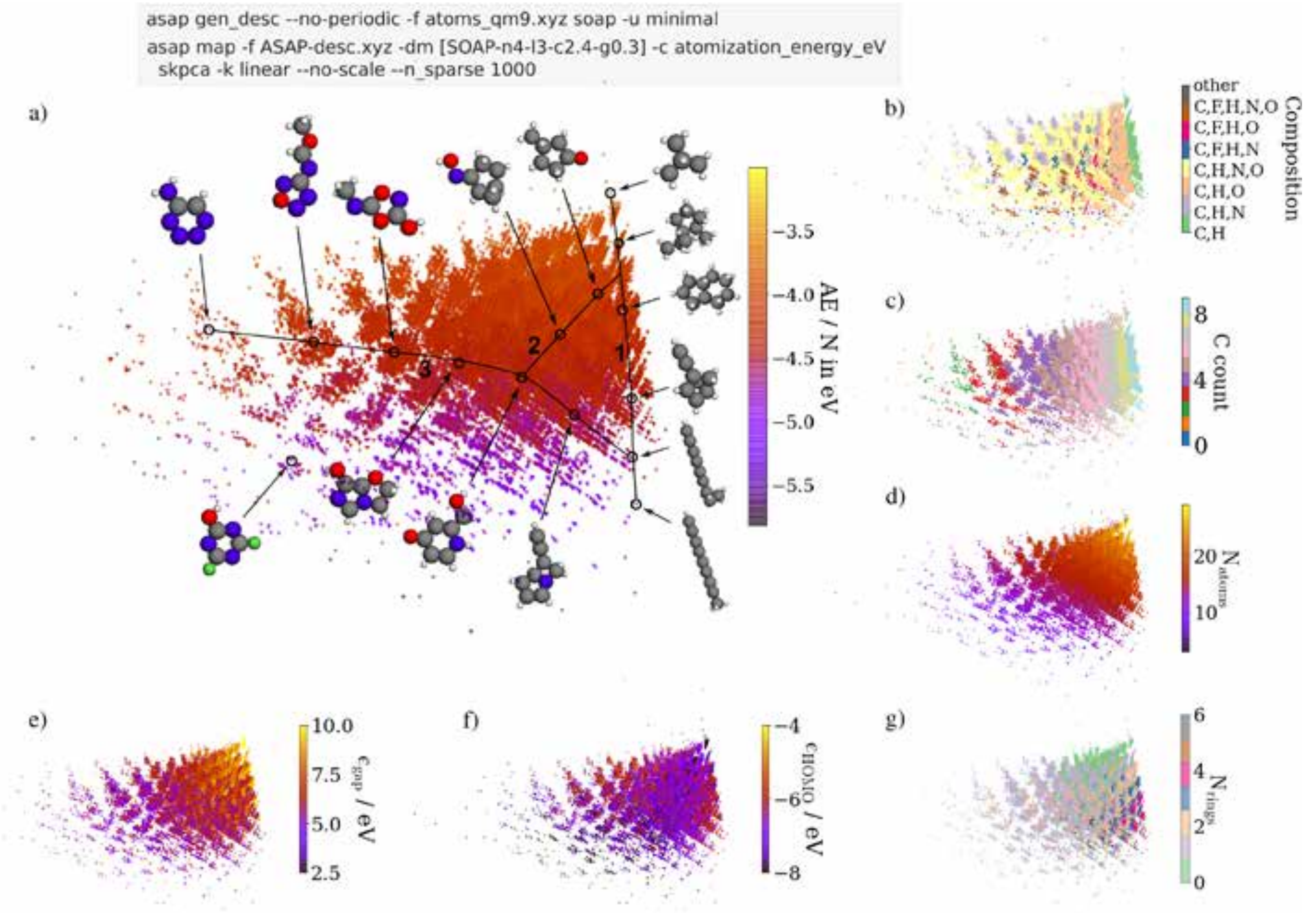
KPCA map for water configurations

[Cheng et al. Accounts of Chemical Research 2020]



KPCA map for QM9 data set

[Cheng et al. Accounts of Chemical Research 2020]



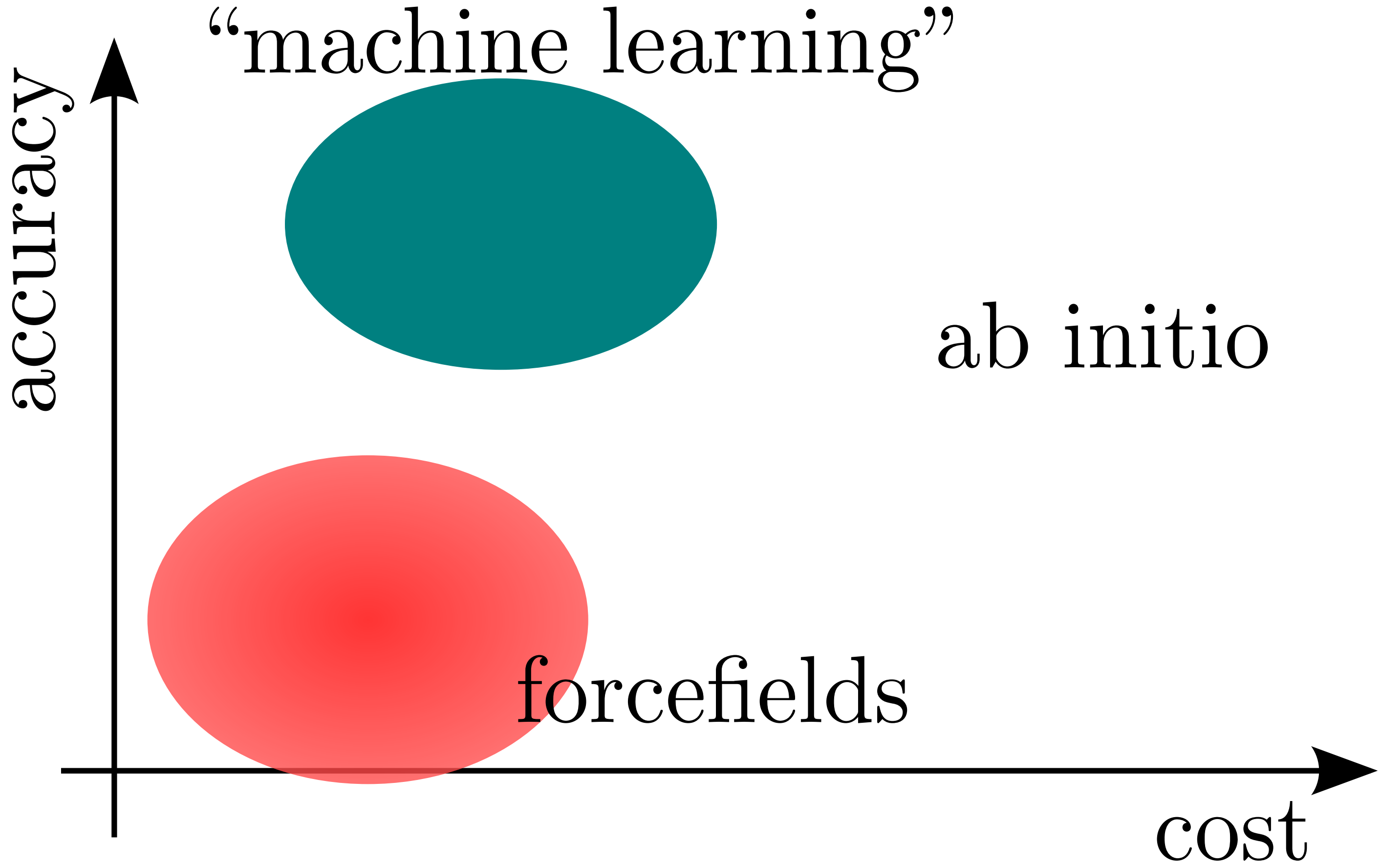
[Figure made by Simon Wengert, Christian Kunkel, Johannes Margarf]

Outline

What we will talk about:

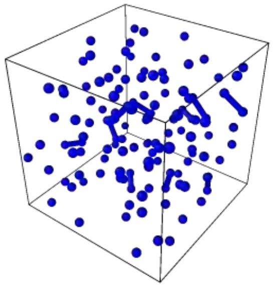
- Statistical mechanics & molecular dynamics 101.
  - Metadynamics
  - Thermodynamic integration
  - Nuclear quantum effects (NQEs)
- Translating materials and molecules into matrices.
  - Representations
  - Dimensionality reduction
- Introduction to machine learning potentials.

ML potentials



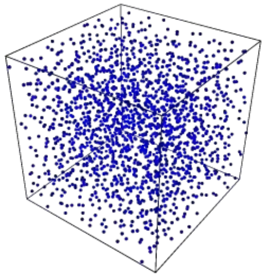
ML potentials

Density functional theory
(1/6 of all supercomputer usage!)



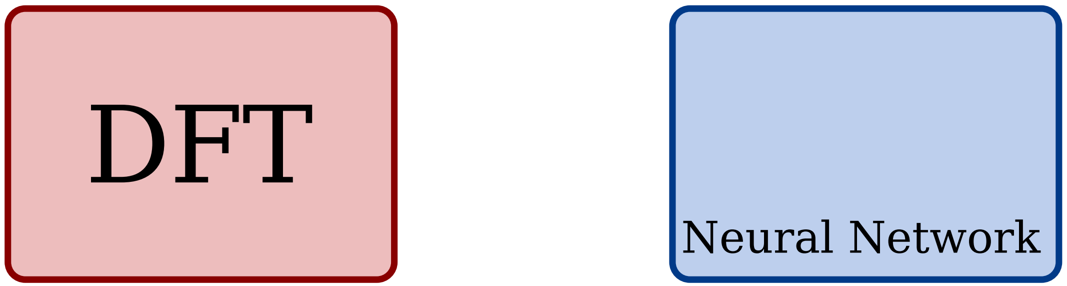
Size: ~100 atoms
Time: picoseconds (10^{-12} S)
Scaling: cubic ($\mathcal{O}N^3$)
Cost: up to millions of CPU hours

ML potentials
[Behler & Parrinello PRL 2008;
Bartók et al PRL 2010]

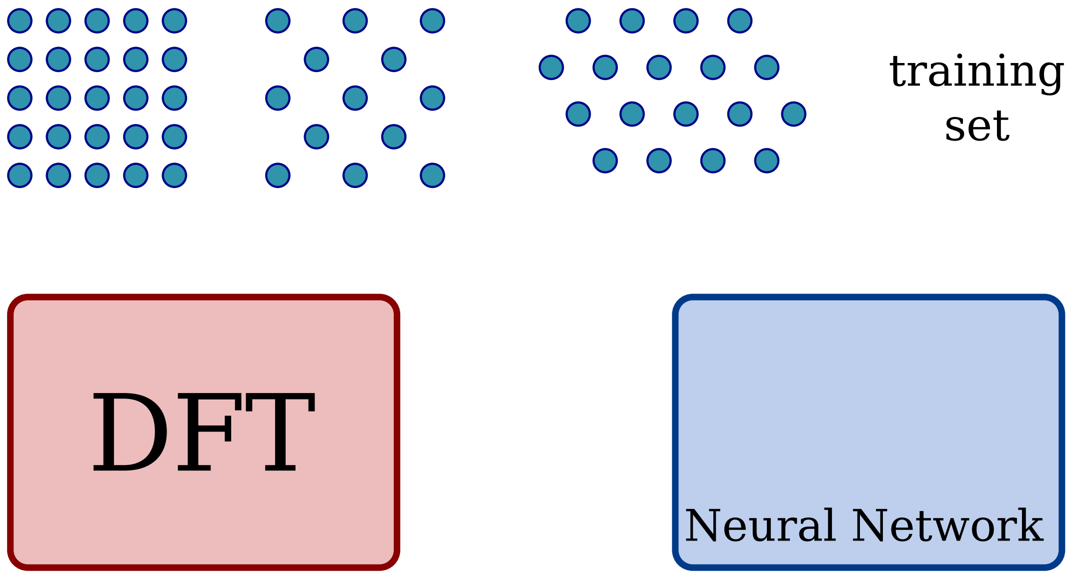


Size: >10,000 atoms
Time: nanoseconds (10^{-9} S)
Scaling: linear ($\mathcal{O}N$)
Cost: laptop friendly

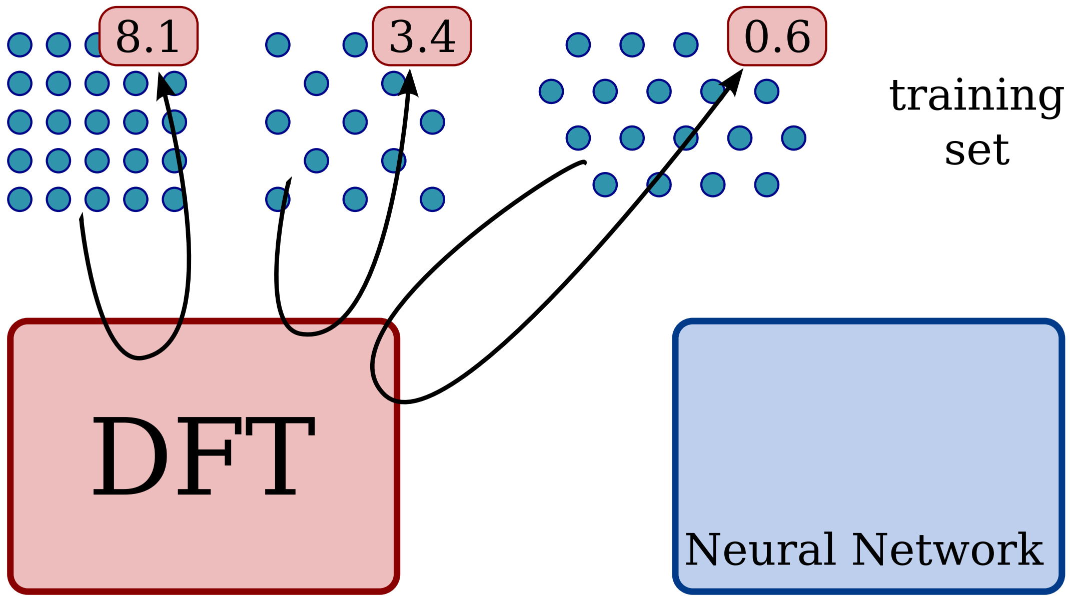
ML potential, a black-box view



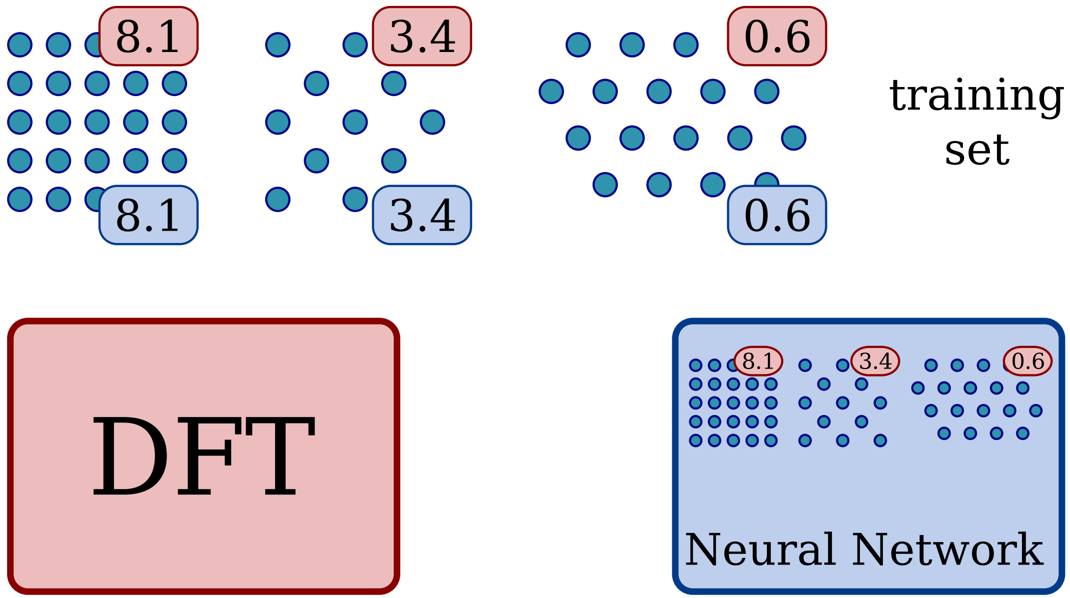
ML potential, a black-box view



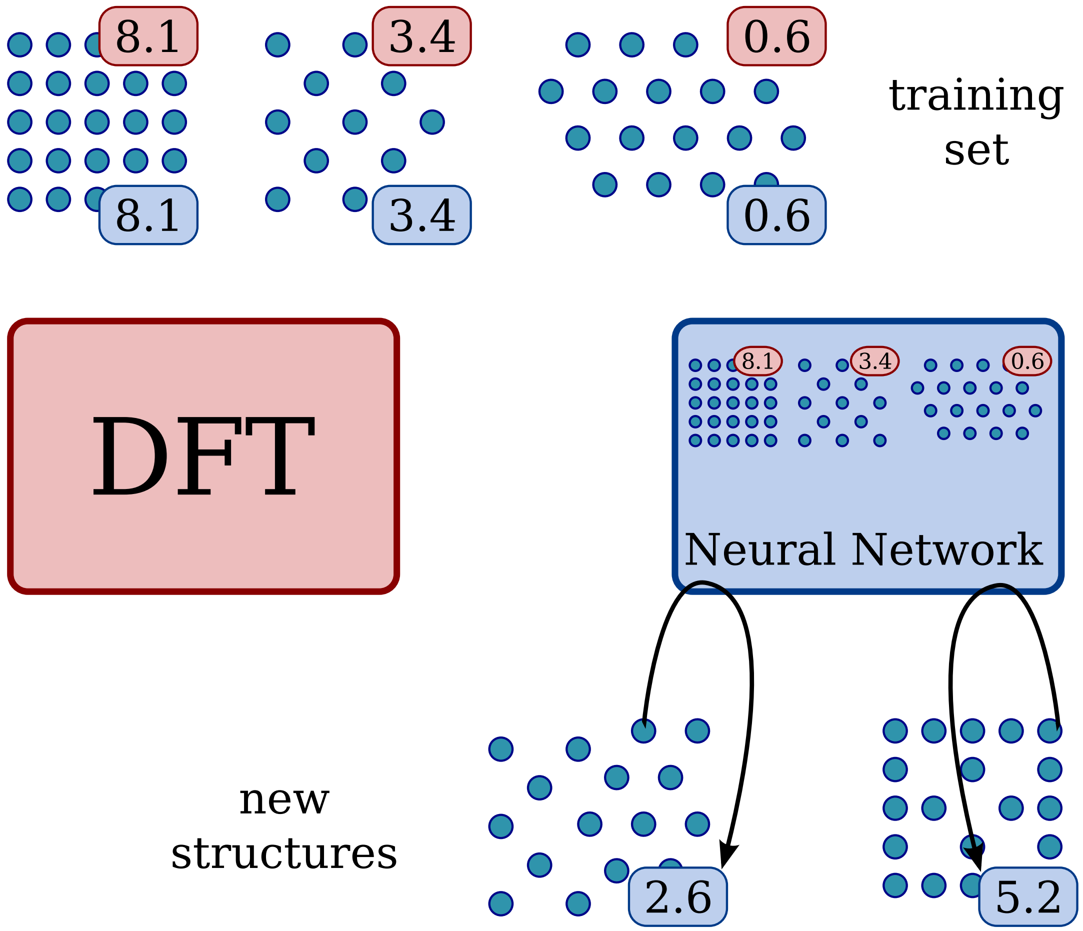
ML potential, a black-box view



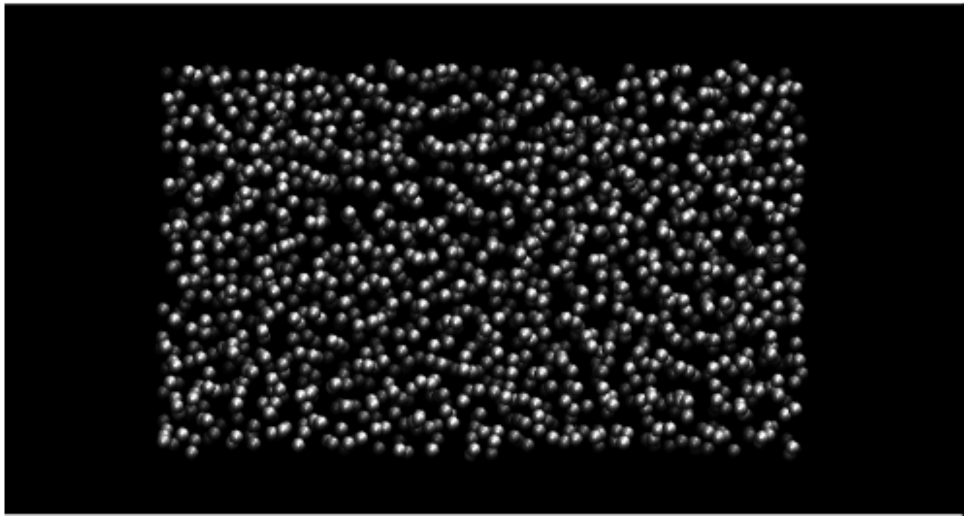
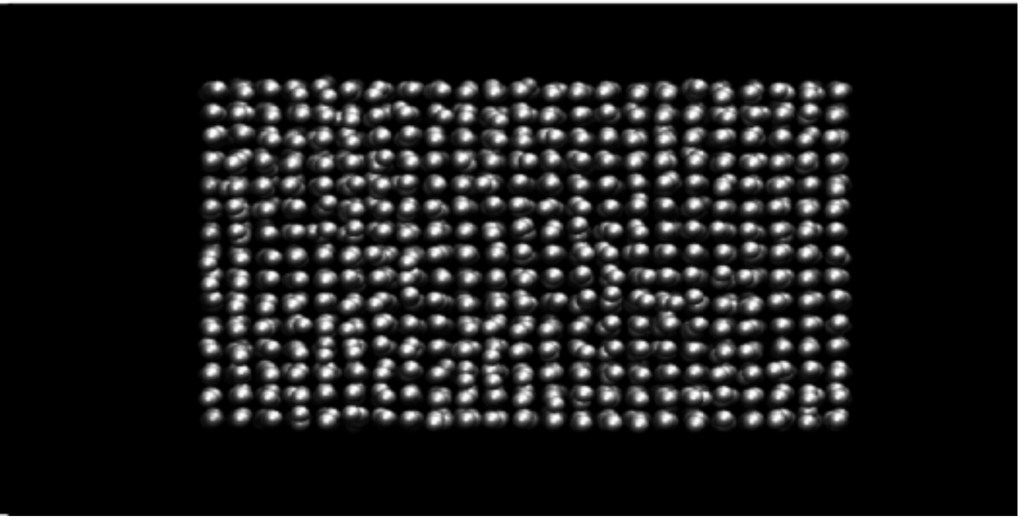
ML potential, a black-box view



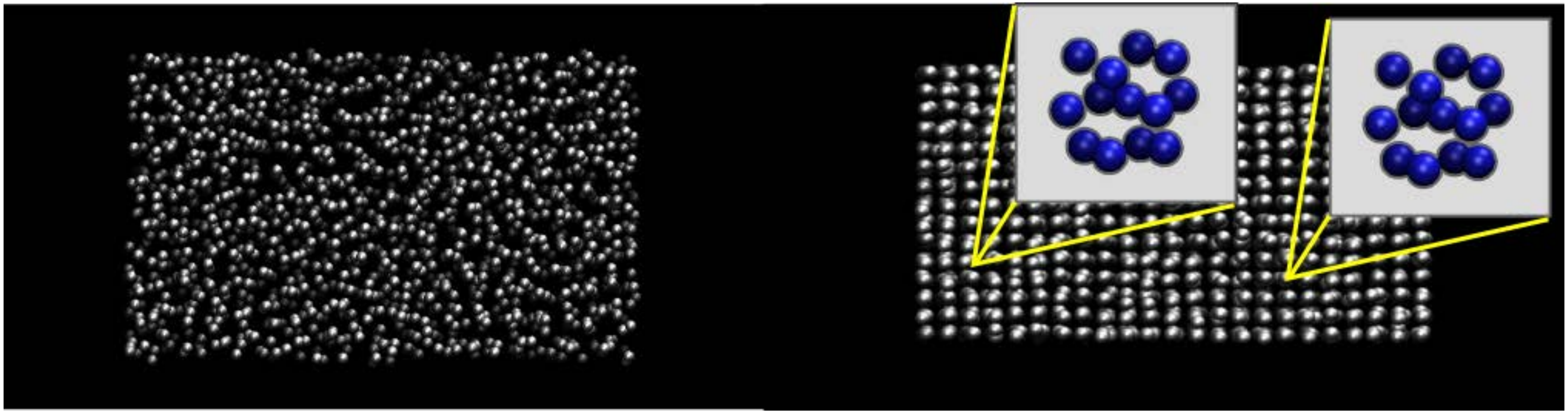
ML potential, a black-box view



Local atomic environments

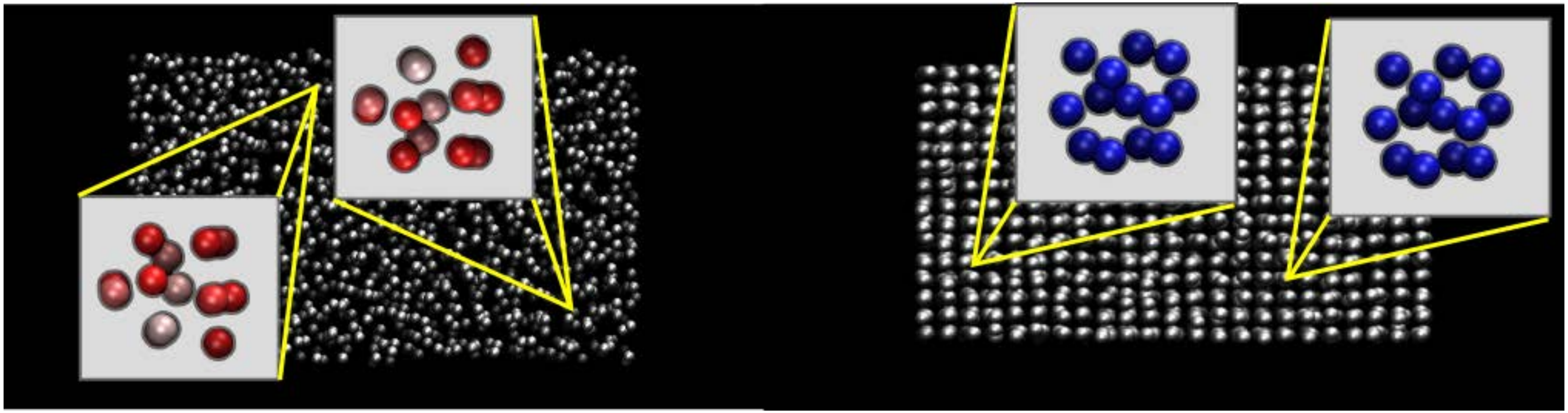
Local atomic environments



Popular representations for comparing atomic environments

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]
- Permutation invariant polynomials [Braam & Bowman 2008]

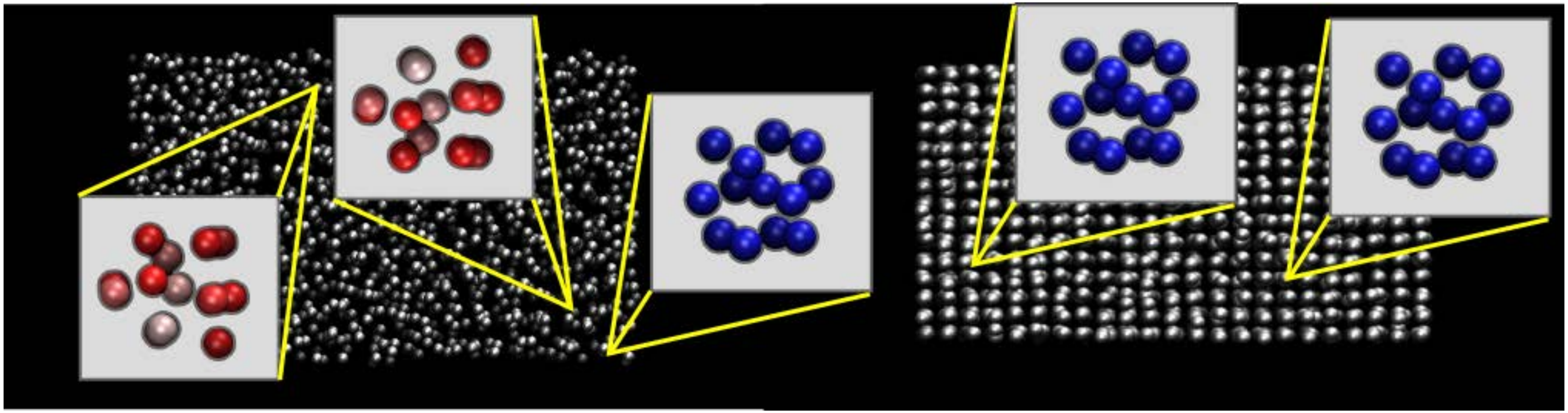
Local atomic environments



Popular representations for comparing atomic environments

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]
- Permutation invariant polynomials [Braam & Bowman 2008]

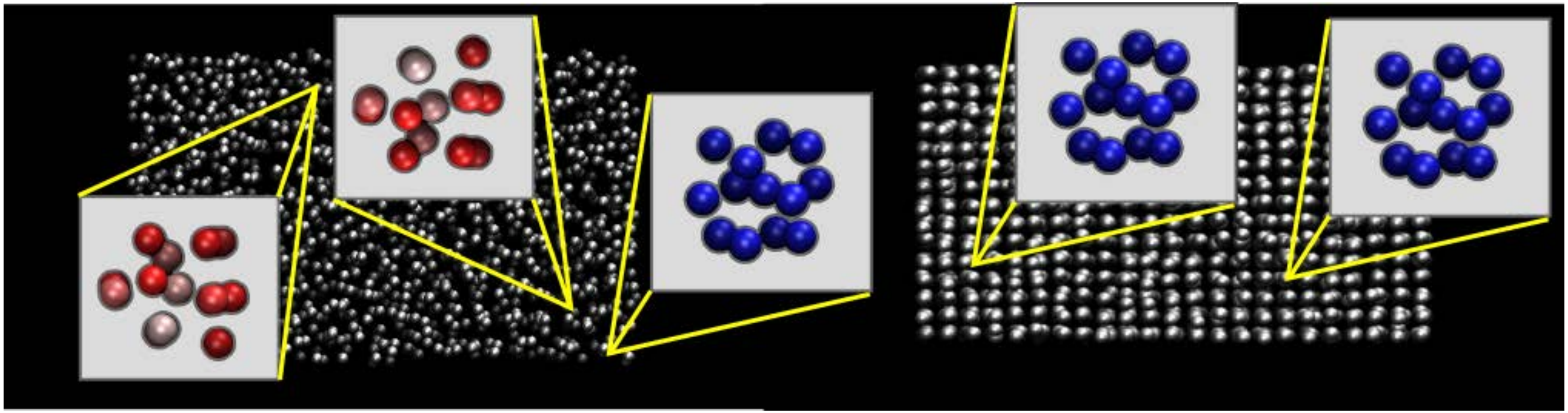
Local atomic environments



Popular representations for comparing atomic environments

- Smooth overlap of atomic positions (SOAP) [Bartók, Kondor & Csányi PRB 2013]
- Behler-Parrinello symmetry functions [Behler & Parrinello PRL 2008]
- Permutation invariant polynomials [Braam & Bowman 2008]

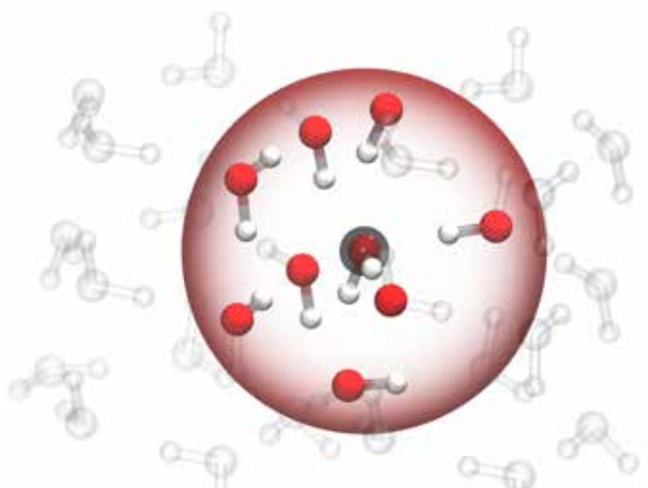
Local atomic environments



- Similar atomic environments are encountered over and over again.
- If you compute all configurations using quantum mechanics, you lose!
- Near-sightedness of energy and forces of each environment.

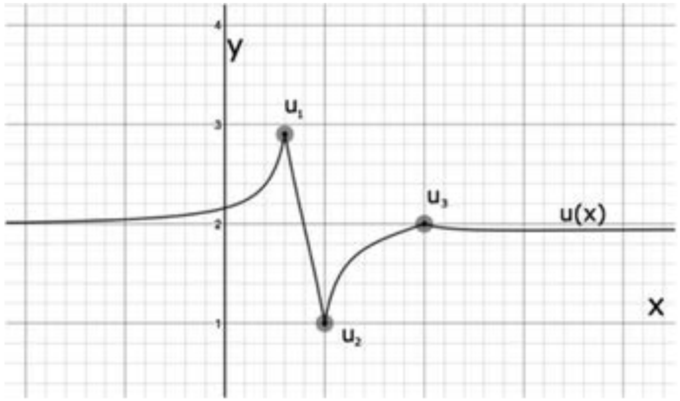
Construct ML potentials

Step 1: Collect environments.



$$E = \sum E_i$$

Step 2: Interpolate.

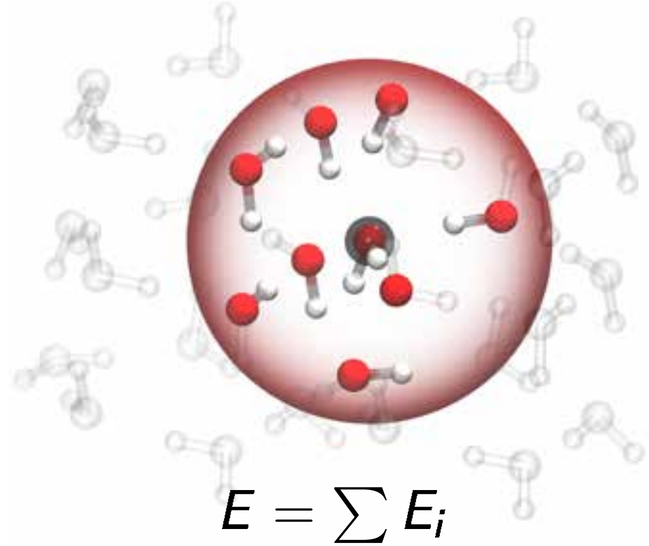


Construct ML potentials

Ways to collect atomic environments:

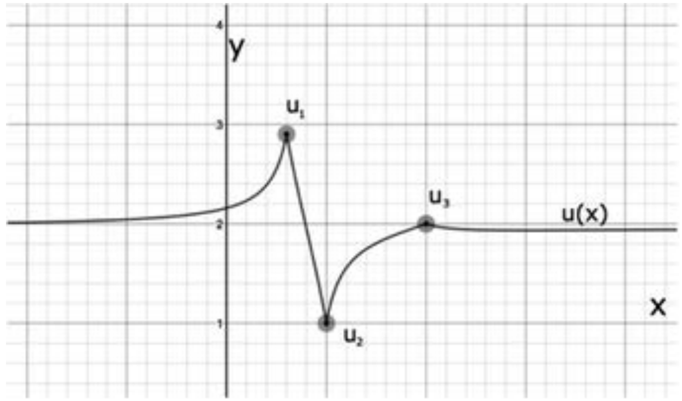
- Enumerating possible structures
- Random displacement
- Stretching and compression
- Molecular dynamics (MD) and PIMD
- On-the-fly learning [Li, Kermode & Vita PRL 2015]
- Random searches [Deringer, Pickard & Csányi PRL 2018]
- Active learning [Podryabinkin & Shapeev Com. Mat. Sci. 2017]

Step 1: Collect environments.

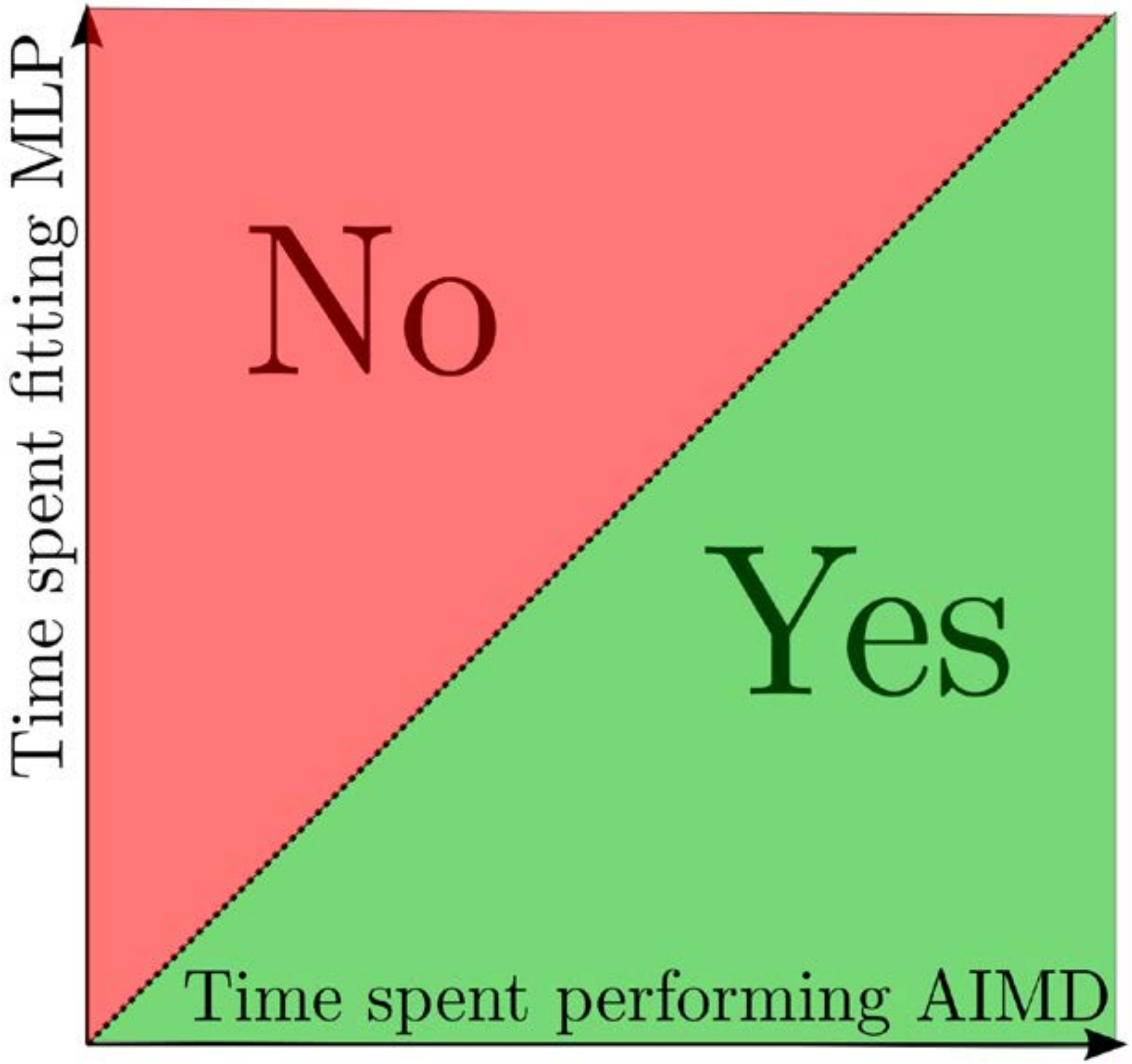


$$E = \sum E_i$$

Step 2: The rest is interpolation.



Making a decision



https://github.com/BingqingCheng/ASAP

Contributors: Ryan-Rhys Griffiths, Tamas Stenczel, Bonan Zhu, Felix Faber, Noam Bernstein

ASAP

Automatic Selection And Prediction tools for materials and molecules

Basic usage

Type `asap` and use the sub-commands for various tasks.

- Low-dimensional embedding, regression
- Sparsification
- Clustering, kernel density estimation