Computational Alchemy

Guido Falk von Rudorff, University of Vienna

ferchault @ferchault guido.vonrudorff.de

Introduction

Guido Falk von Rudorff

BSc, Msc	Freie Universität Berlin	Atom clusters, force fields
PhD	London	Electronic structure calculations
PostDoc	Basel, Vienna	Alchemy and machine learning

What about you?

- Background
- Why interest in computational alchemy?
- Aims for this lecture
- Experience with classical or electronic structure calculations?
- How comfortable with Python?

General Information

Each Thursday:

- 10:15-11:00 Lecture
- 11:15-12:00 Lecture
- 12:15-13:00 Exercise (starting next week)

Exercises:

- 10 sets, two exercises each
- 40 points in total, 20 points needed

Setting:

- Interactive and informal: less lecture, more conversation

Joseph Wright, 1771

Outline

$E \simeq -Z^{7/3}$

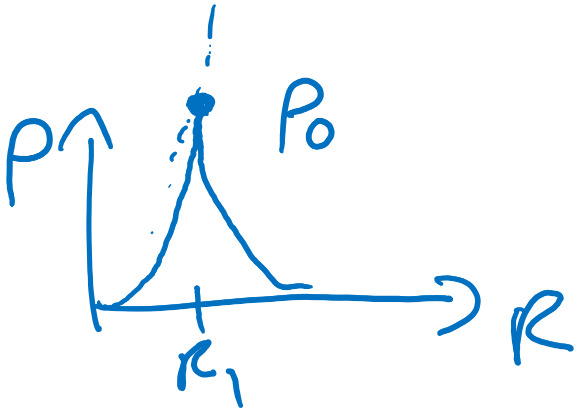
- Why do we care?
- Classical alchemy: free energy calculations
- Quantum alchemy
- High-level introduction to quantum mechanical calculations
- High-level introduction to numerical derivatives and automatic differentiation
- Alchemical Perturbation Density Functional Theory
- Fundamental consequences:
  - Atoms in molecules
  - Alchemical Enantiomers
  - Alchemical normal modes
- Applications
  - Electrostatic potentials
  - Non-covalent interactions

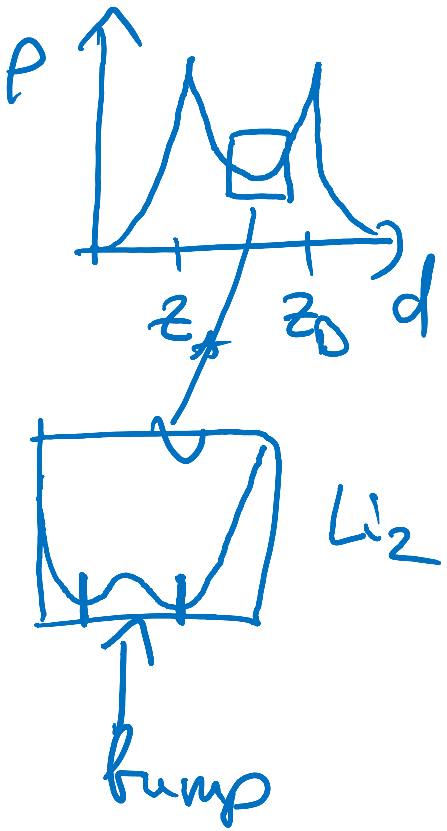
Practical Knowledge

- New tool: interpolation in chemical space
    - Machine learning
- Numerical differentiation/integration
    - Computer Algebra Systems
- Automatic differentiation
    - Machine learning
    - Optimization techniques
- Better understanding of electronic structure calculations

Questions?

- V representability $\rho \to V$
- Kato's cusp

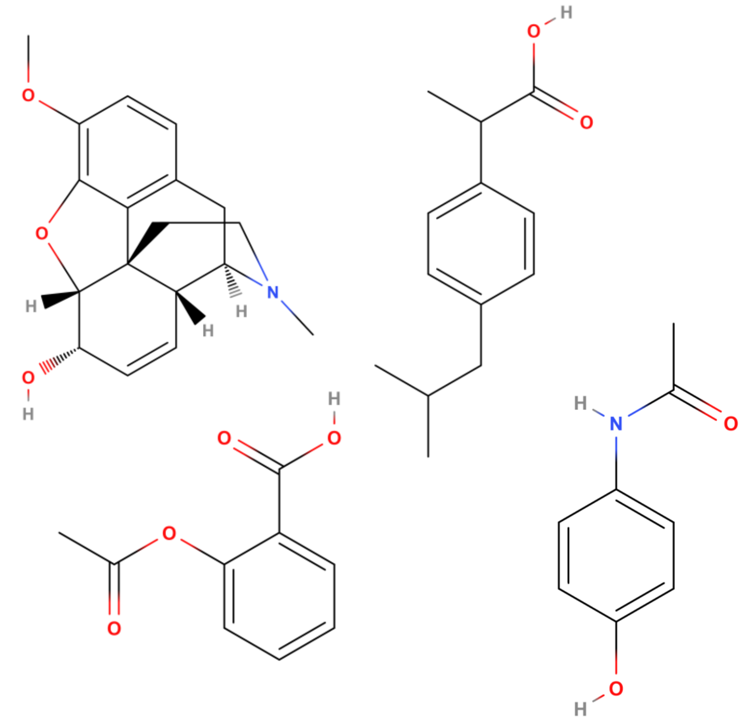


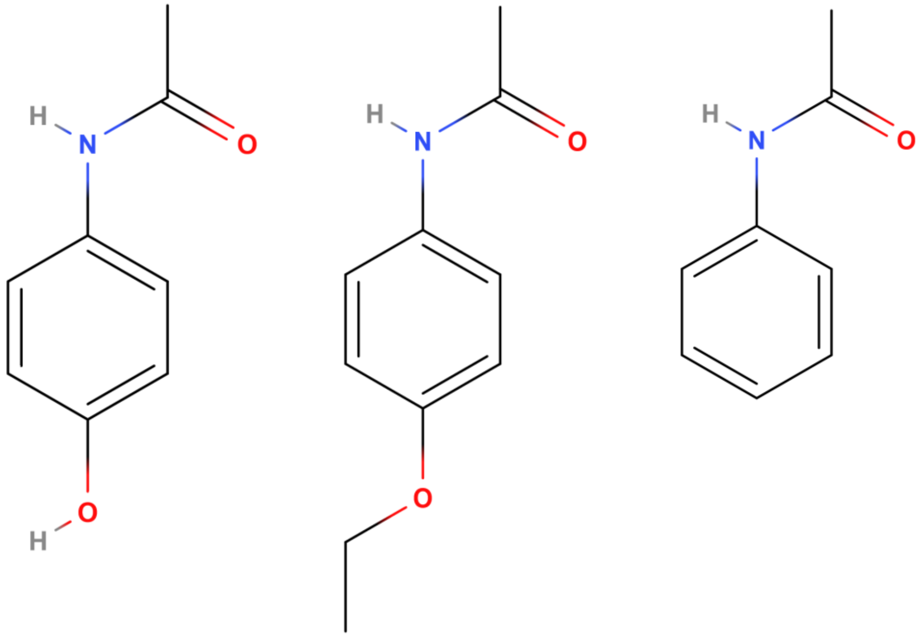


Motivation

Introduction

Materials / compound design efforts face a vast search space

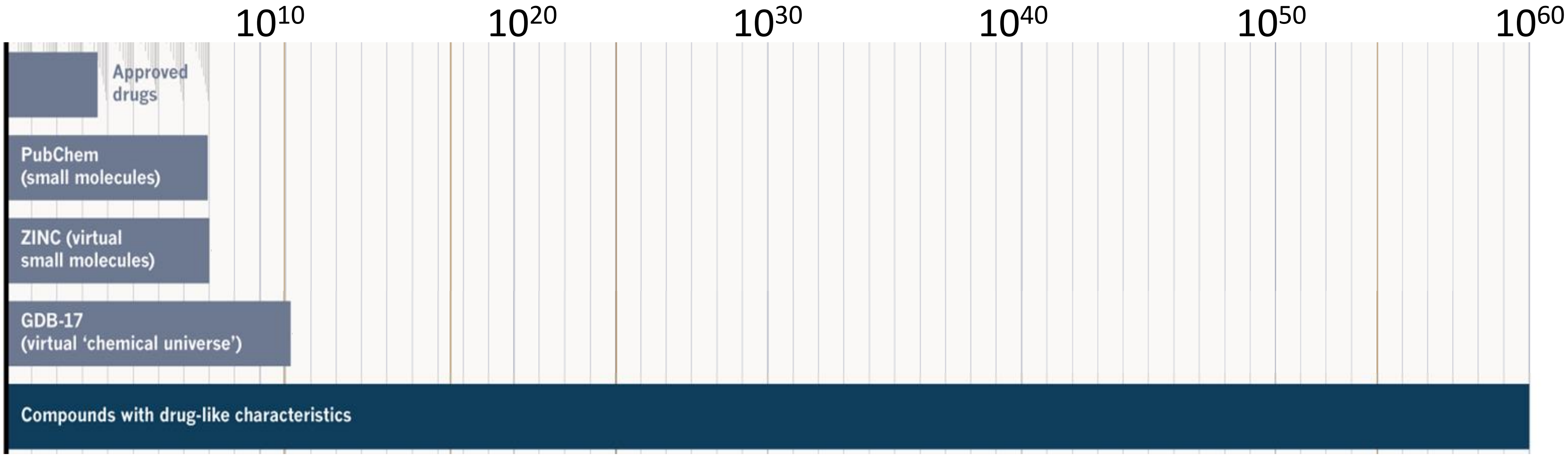




Global Search Problem
Which class of compounds?

Local Search Problem
Which particular species within that class?

Chemical Space: Size



Commercial databases

- 164 million molecules
- 15k added daily

Scale

- One person: 1 million compounds/second
- 10 billion people on earth
- 10^{26} universe ages to go through

Mullard, *Nature,* 2017.

Necessity vs Convenience

Necessity

- Only way to cover problem size
- Still open to systematic evaluation
- Often used as prefiltering step

- Complicated chemistry
- Tricky / error-prone reference calculations

Convenience

- Can be done more accurately
- Uneconomical/cumbersome reference method
- Often used as direct but optional substitute

- Standard energy calculations of well-behaved systems
- Semi-emipirical level sufficient

Chemical Space: Approaches

Restriction

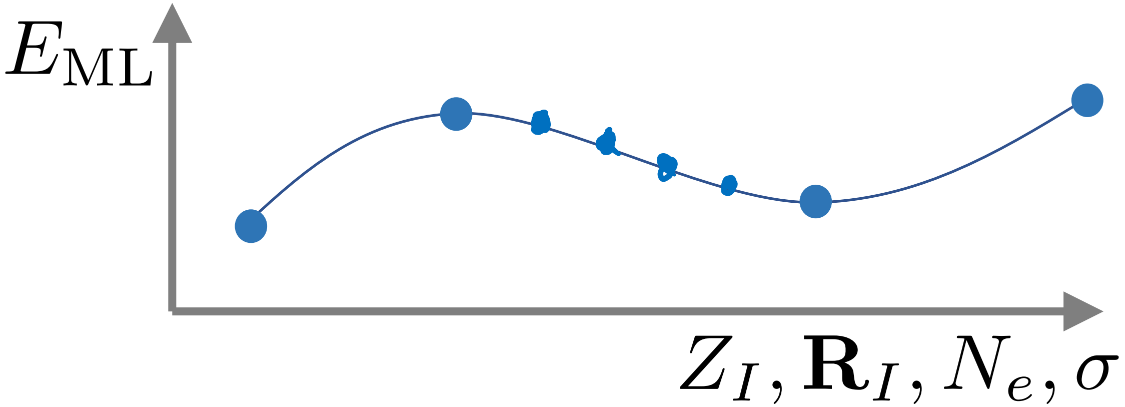
- Subspaces: conformers, constitutional isomers
- Configurations: minima, transition states

Interpolation

- Set of molecules as fixed reference
- Define interpolant
- Small data sets: e.g. KRR
- Large data sets: e.g. NN

Expansion

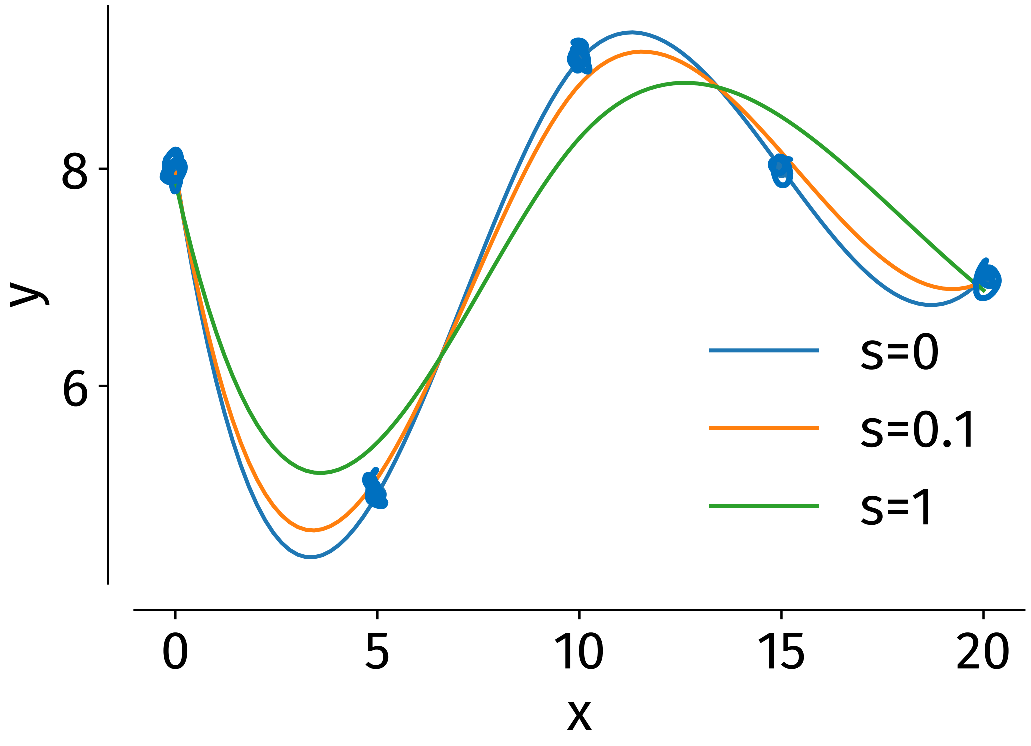
- Molecules get perturbed
- Quantum Alchemy e.g. APDFT
- Taylor expansion w.r.t. elements



Interpolation != Interpolation

- Computational alchemy
  True functional value is always defined / obtainable

- Mathematical interpolation
  True value is a model construct and depends on the model employed (s)

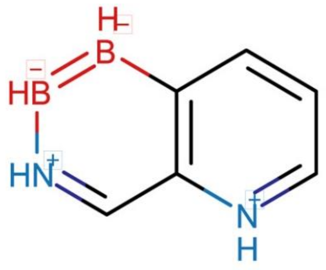


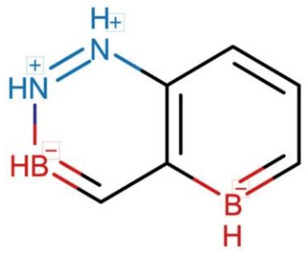
Chemical Space

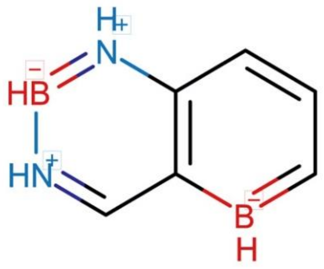
Smoothly connecting discrete states

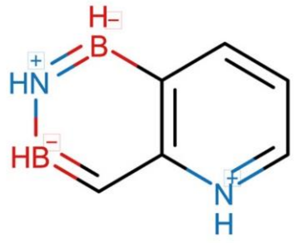
- Classical thermodynamic integration
  interpolate parameters representing elements

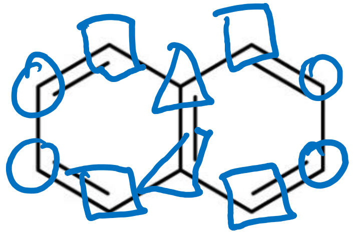
- Quantum no interpolant needed in many directions (Schrödinger’s equation)
  interpolate coordinates, nuclear charges (not: number of electrons)





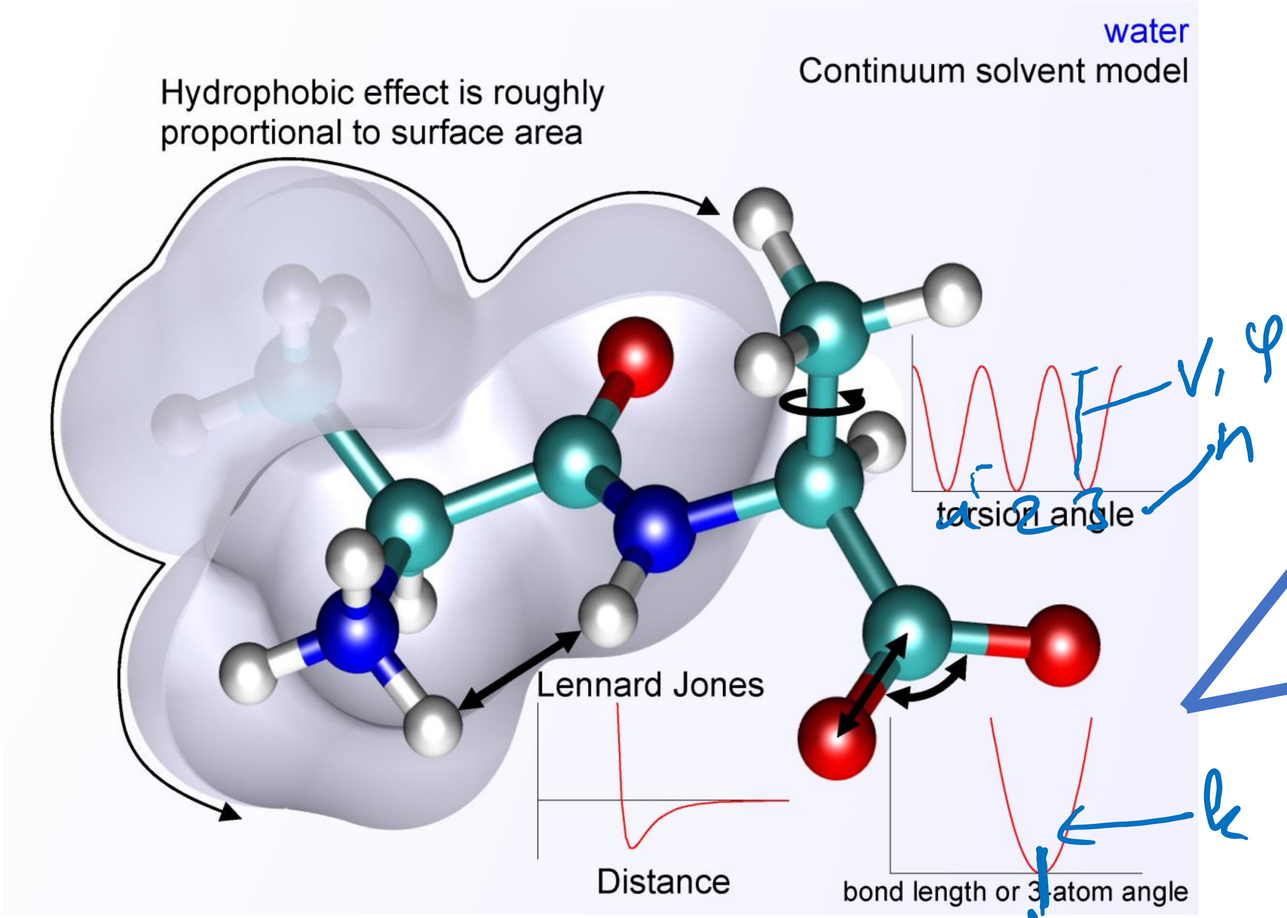




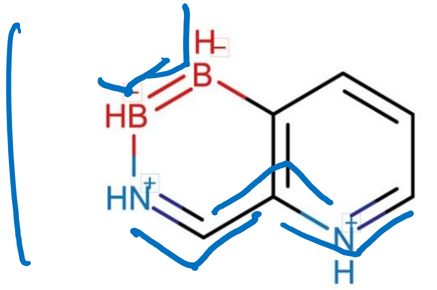


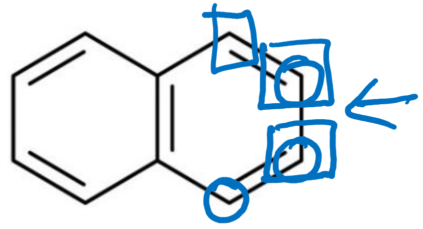
Classical picture

- Force field describes material/molecule
- No electronic structure, but physical effects


water
Continuum solvent model
Hydrophobic effect is roughly proportional to surface area
V, φ
n
1 2 3
torsion angle
Lennard Jones
Distance
k
bond length or 3-atom angle
d0, α0

https://commons.wikimedia.org/wiki/File:MM_PEF.png

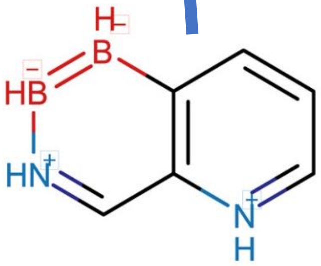


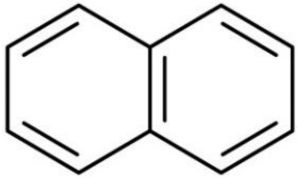


Quantum picture

- Hamiltonian describes electronic structure
- Mixing Hamiltonians creates paths

$$\hat{H}(\lambda) \equiv \lambda \hat{H}_{\mathrm{t}} + (1-\lambda)\hat{H}_{\mathrm{r}} \qquad \lambda \in [0,1]$$





$$-\sum_i \frac{\hbar^2}{2M_i}\nabla^2_{\mathbf{R}_i}$$

$$-\sum_i \frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}_i}$$

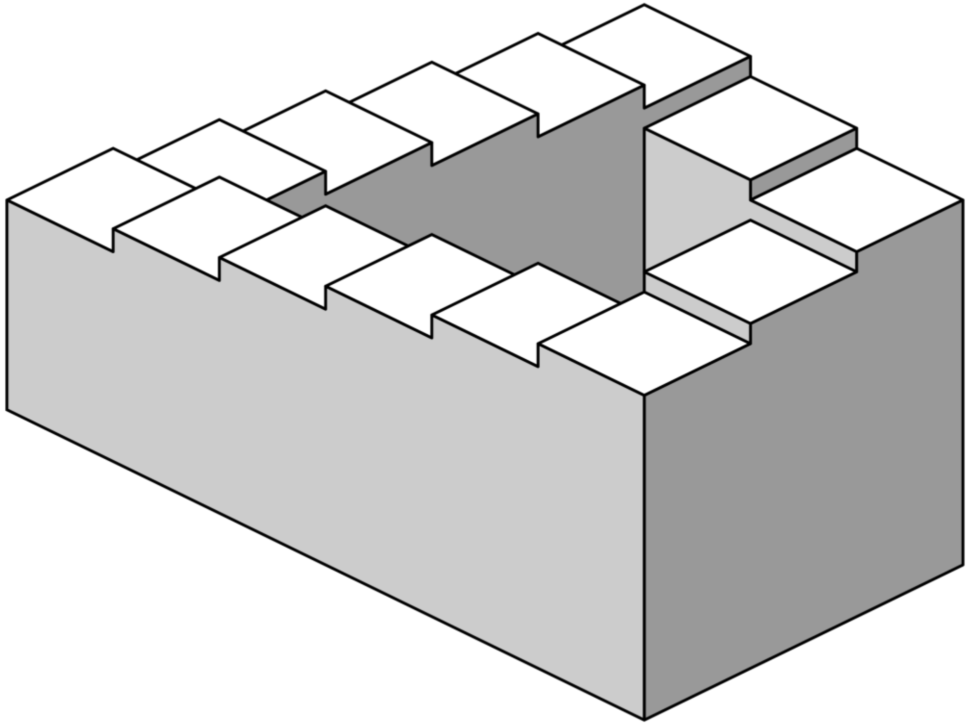
$$-\sum_i\sum_j \frac{Z_i e^2}{4\pi\epsilon_0 \left|\mathbf{R}_i - \mathbf{r}_j\right|}$$

$$\frac{1}{2}\sum_i\sum_{j\neq i} \frac{e^2}{4\pi\epsilon_0 \left|\mathbf{r}_i - \mathbf{r}_j\right|} = \sum_i\sum_{j>i} \frac{e^2}{4\pi\epsilon_0 \left|\mathbf{r}_i - \mathbf{r}_j\right|}$$

$$\frac{1}{2}\sum_i\sum_{j\neq i} \frac{Z_i Z_j e^2}{4\pi\epsilon_0 \left|\mathbf{R}_i - \mathbf{R}_j\right|} = \sum_i\sum_{j>i} \frac{Z_i Z_j e^2}{4\pi\epsilon_0 \left|\mathbf{R}_i - \mathbf{R}_j\right|}.$$

Physical meaning?

Key Questions of Any Research Area



https://commons.wikimedia.org/wiki/File:Beaulieu_National_Motor_Museum,_Hampshire_(460903)_(9454847609).jpg

What could possibly go wrong?

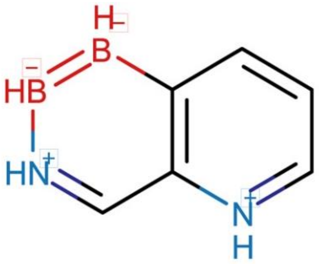
How hard can it be?

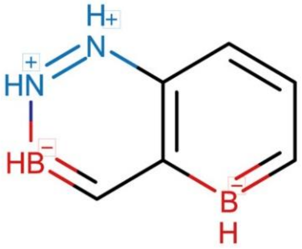
Pre-Requisites

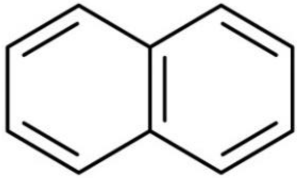
- State function: uniquely defined for a system = $Z_I\,, \mathbf{R}_I\,, N_e\,, \sigma$

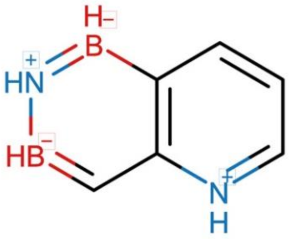
Energy, and any other observable (!)

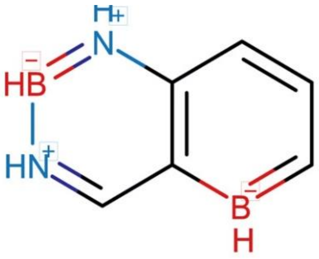
Other examples?









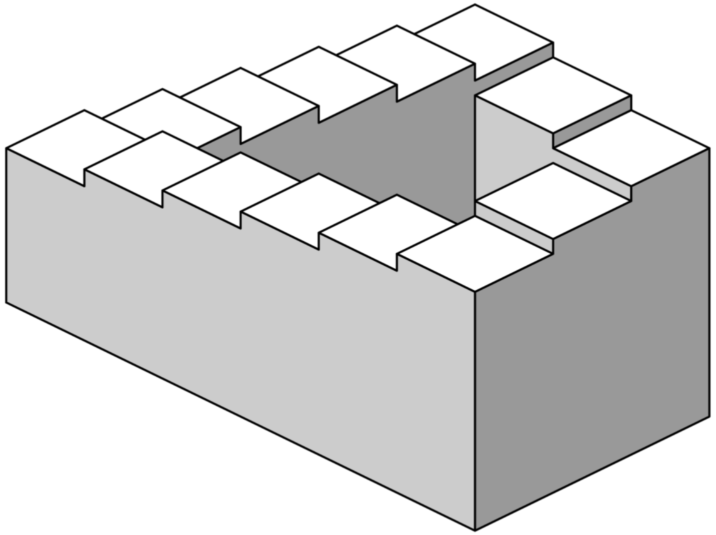


Pre-Requisites

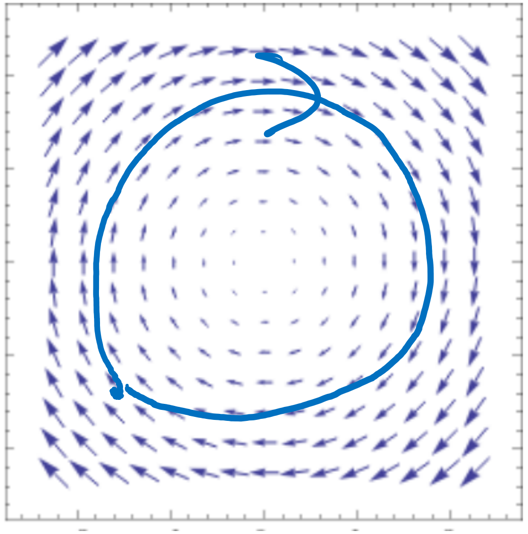
- State function: uniquely defined for a system
- Integrals being path-independent*

$\leftarrow \quad < 0$

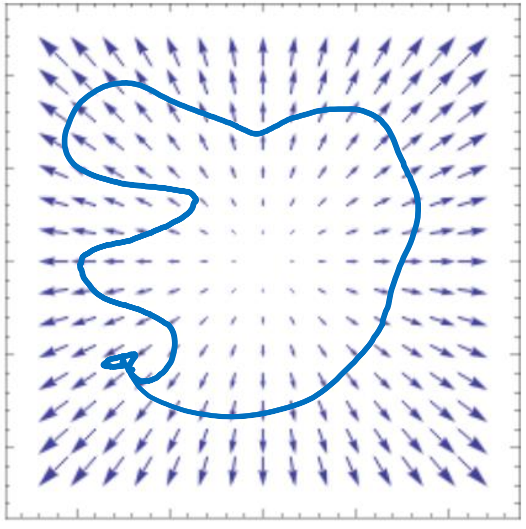
$\rightarrow \quad > 0$



$$\int_{C_1} \mathbf{F} \cdot d\mathbf{s} \neq \int_{C_2} \mathbf{F} \cdot d\mathbf{s}$$



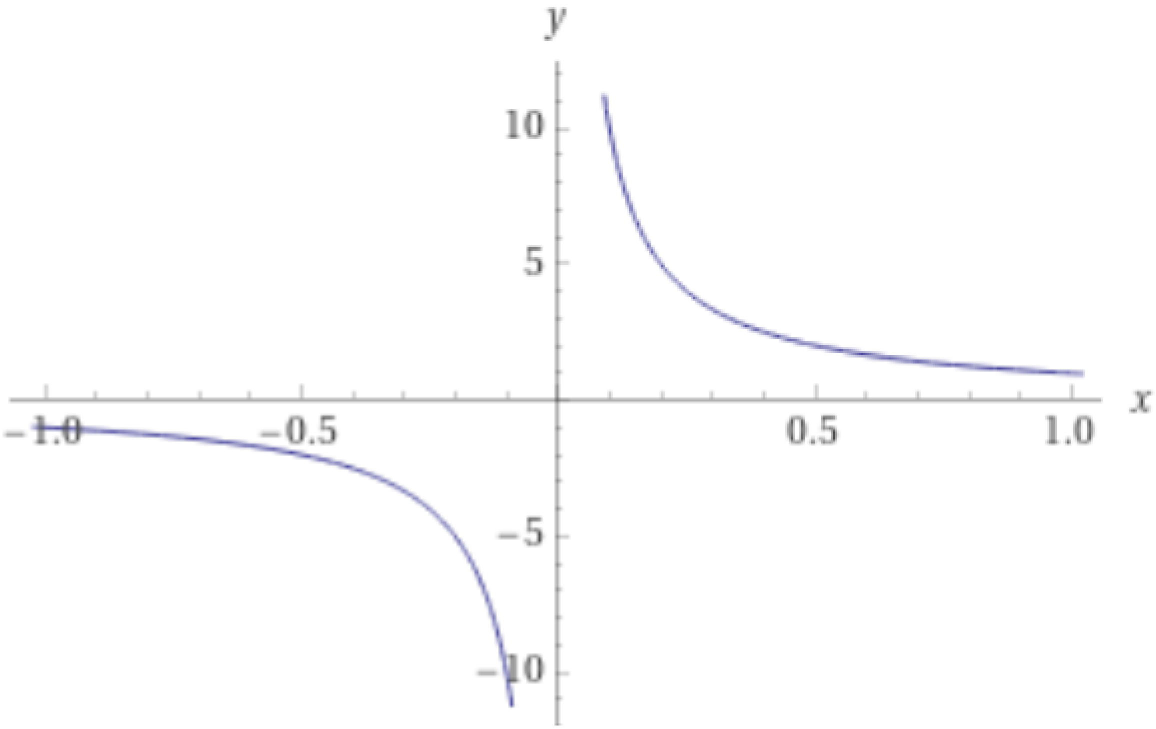
non-conservative



conservative

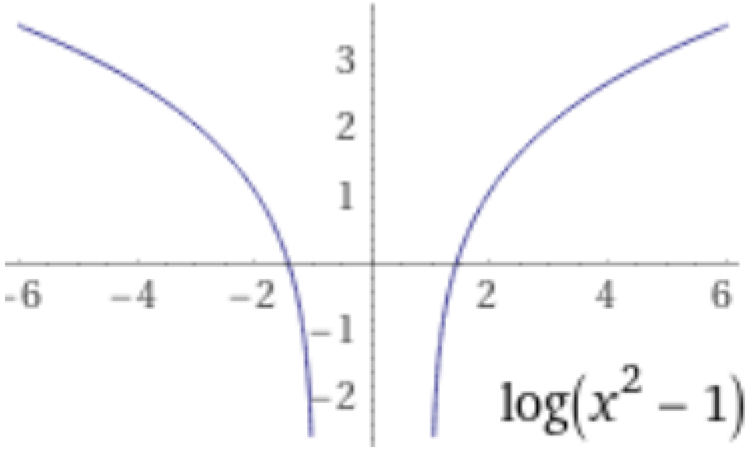
Pre-Requisites

- State function: uniquely defined for a system
- Integrals being path-independent*
- Smoothness



Pre-Requisites

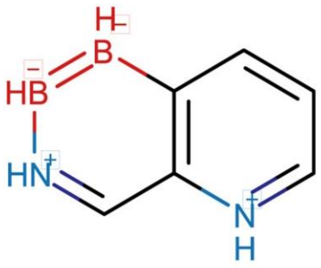
- State function: uniquely defined for a system
- Integrals being path-independent*
- Smoothness
- Defined for the relevant domain



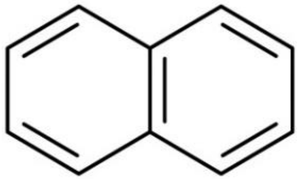
Non-integer values are no problem

Pre-Requisites

Fundamental theorem of calculus



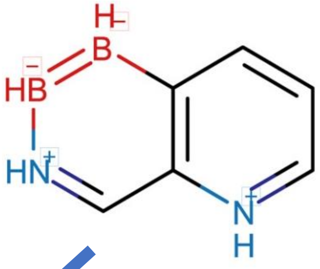
$$F(x) = \int_a^x f(t)\,dt.$$



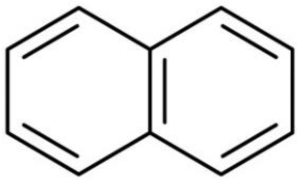
$$\hat{H}(\lambda) \equiv \lambda \hat{H}_\mathrm{t} + (1 - \lambda)\hat{H}_\mathrm{r}$$

Pre-Requisites

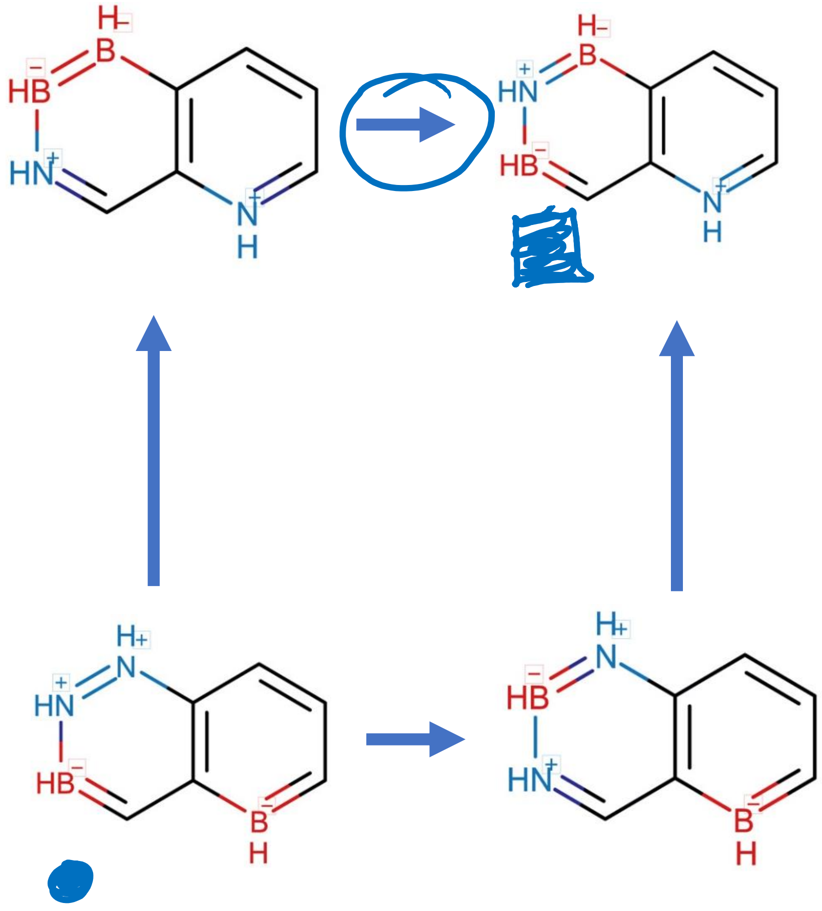
Fundamental theorem of calculus



$$F(x) = \int_a^x f(t)\,dt.$$



$$\hat{H}(\lambda) \equiv \lambda \hat{H}_{\mathrm{t}} + (1-\lambda)\hat{H}_{\mathrm{r}}$$

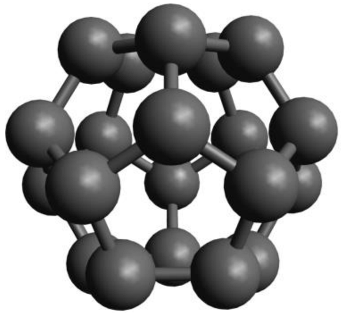


Non Pre-Requisites

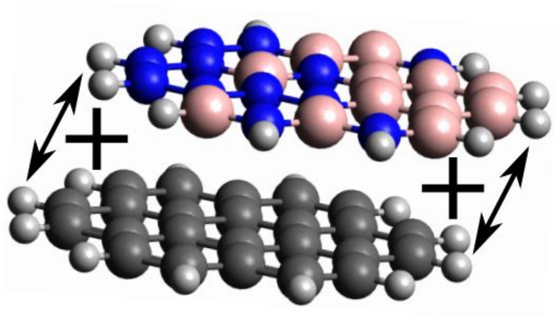
- Experimentally observable
- Energetically stable
- Simple relationships / easy paths

Applications

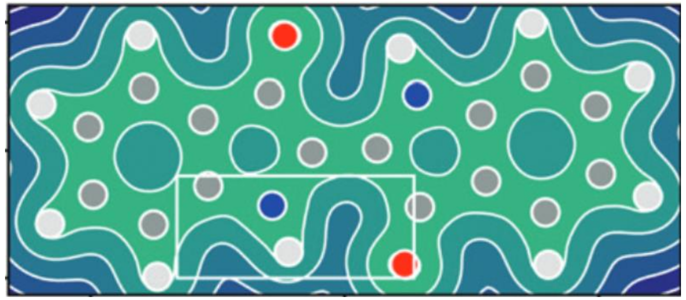
Covalent Energies



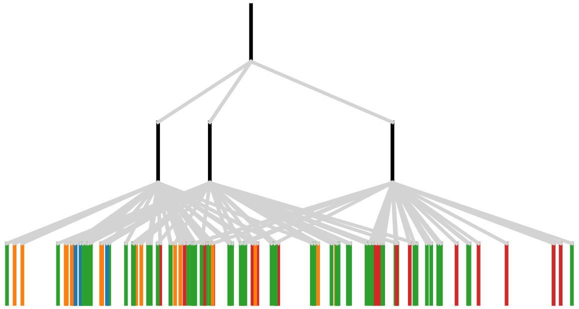
Non-covalent Interactions



Deprotonation Energies



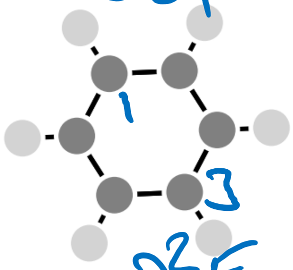
Energy Decomposition



Example

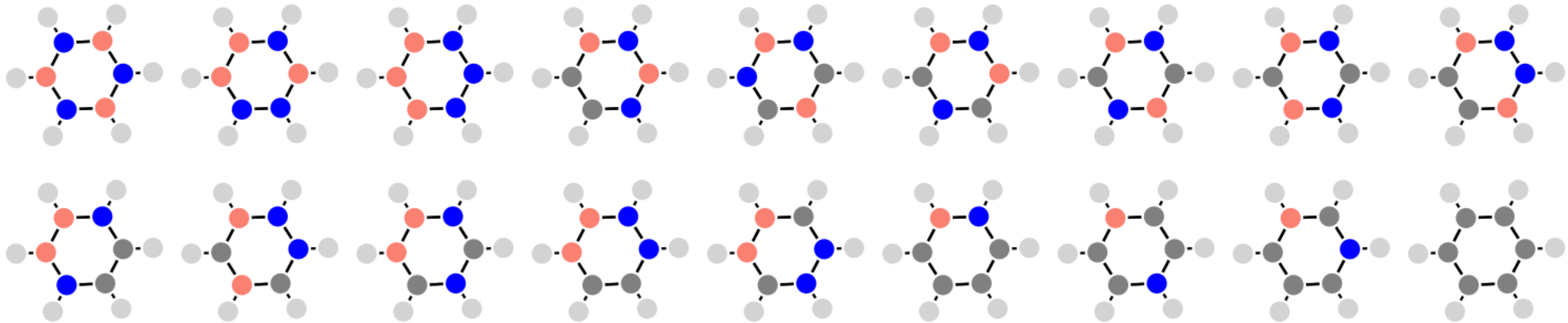
Alchemical Perturbation Density Functional Theory (APDFT)
Uses calculations of *one* molecule to estimate *many* molecules

$$\frac{\partial E}{\partial Z_1} = \frac{\partial E}{\partial Z_3}$$

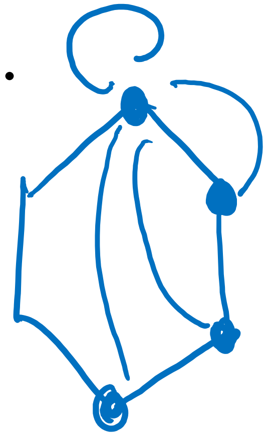


$$\frac{\partial^2 E}{\partial Z_1 \partial Z_3}$$

→



$$\underline{E}, \underline{\rho}, \underline{\{\partial_\lambda^i \rho\}} \rightarrow \{E_i\}, \{\rho_i\}, \{F_i\}, \{\mu_i\}, \{Q_i\}, \ldots$$



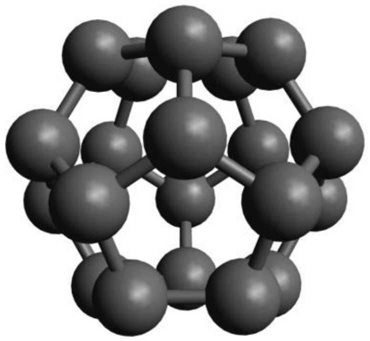
1 system → Millions of systems

GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.* **2020** (*arXiv* 1809.01647).

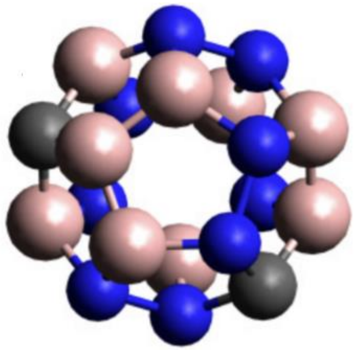
Covalent Energies

- BN-doped C_{20}
- APDFT2: one SCF, one derivative
- Key trick to scale with chemical space:

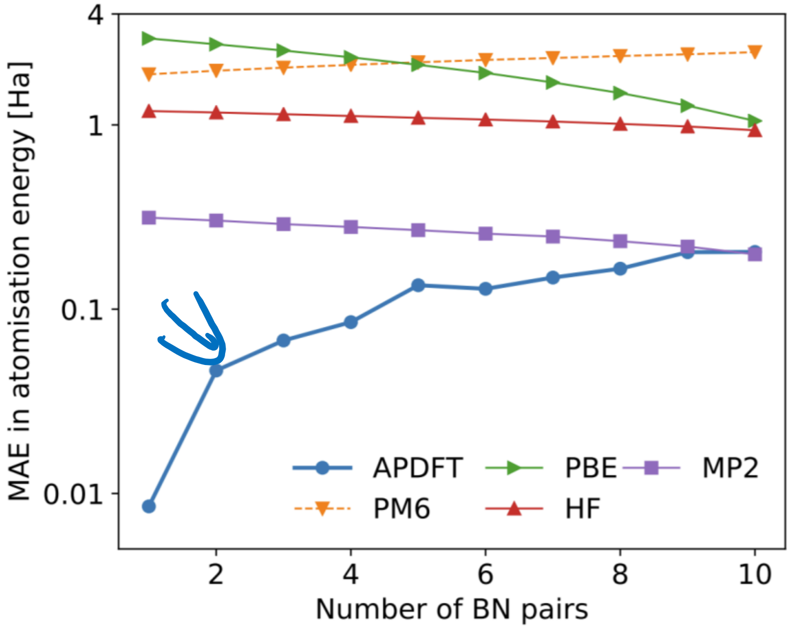
$$\frac{\partial \rho}{\partial \lambda} = \sum_I \frac{\partial \rho}{\partial Z_I} \frac{\partial Z_I}{\partial \lambda}$$

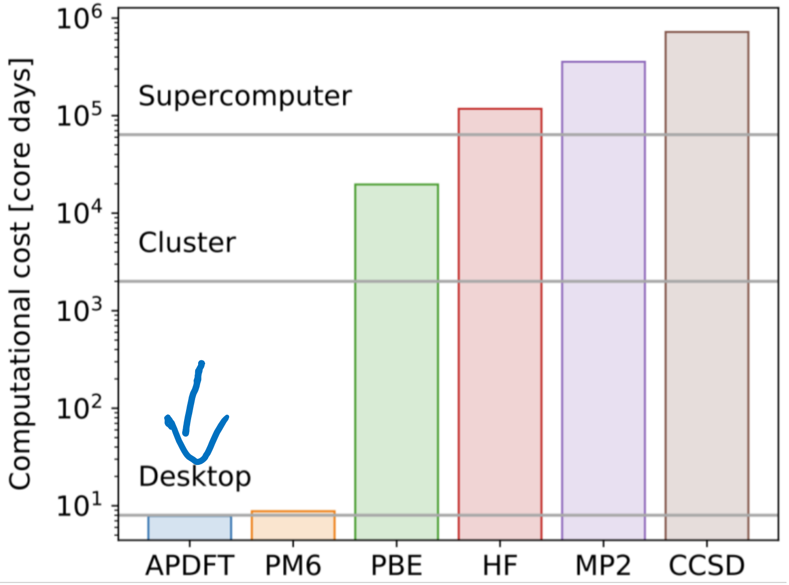


C_{20}



$3.1 \cdot 10^6$
targets



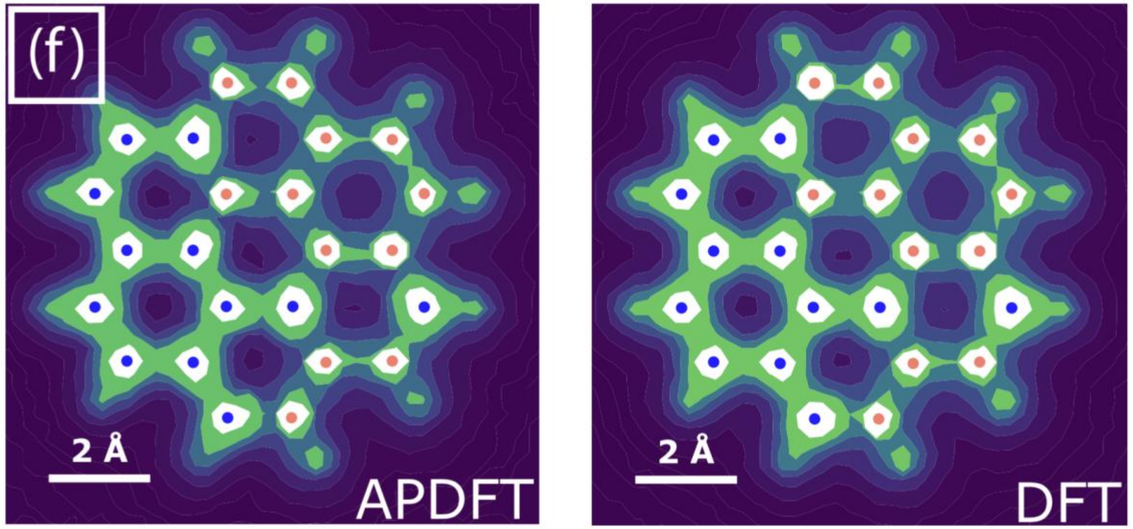


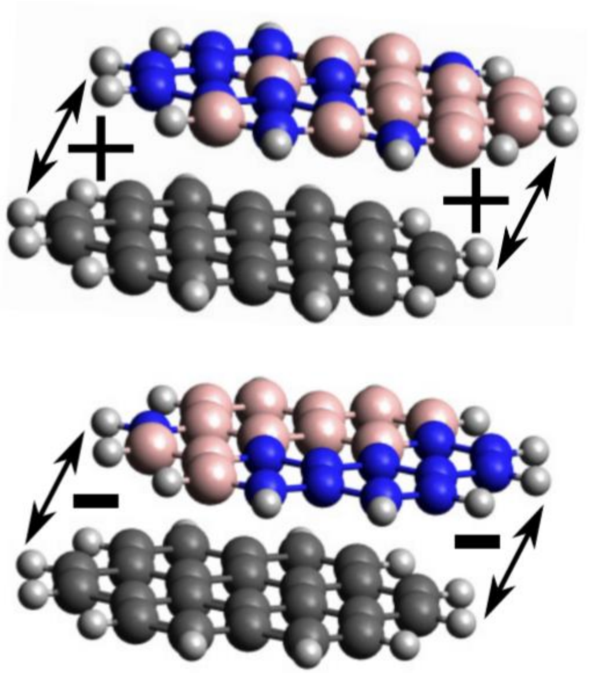
GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.* **2020** (*arXiv* 1809.01647).

Non-covalent Interactions

- BN-doped coronene dimer
- APDFT2: one SCF, 24 derivatives
- SCAN+VV10

$$\rho_t = \sum_{n=0}^{\infty} \frac{1}{n!} \left. \frac{\partial^n \rho}{\partial \lambda^n} \right|_{\lambda=0}$$





$2.8 \cdot 10^{10}$ targets

GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.* **2020** (*arXiv* 1809.01647).

Exercises

1 A tweet used to be up to 140 characters. Summarize the key idea and relevance of computational alchemy in 140 characters or less.

2 a) The fundamental theorem of calculus allows us to obtain one function value by integrating the derivative of that function along a path. In practise, this is done by evaluating the function derivative at a finite number of equidistant points *n* along the path.

Why can this method yield wrong results?

Can you sketch a function where the integral as obtained from *n=3* and *n=5* differs in sign?

Which features in an integrand would be problematic for this approach and why?

b) In classical calculations, a common application of computational alchemy is to grow a molecule (e.g. benzene) inside an existing environment (e.g. water or a membrane). This is done by turning on the interactions between a static benzene and environment molecules, while keeping the interactions between the environment molecules unchanged.

Why does this have to be done slowly?

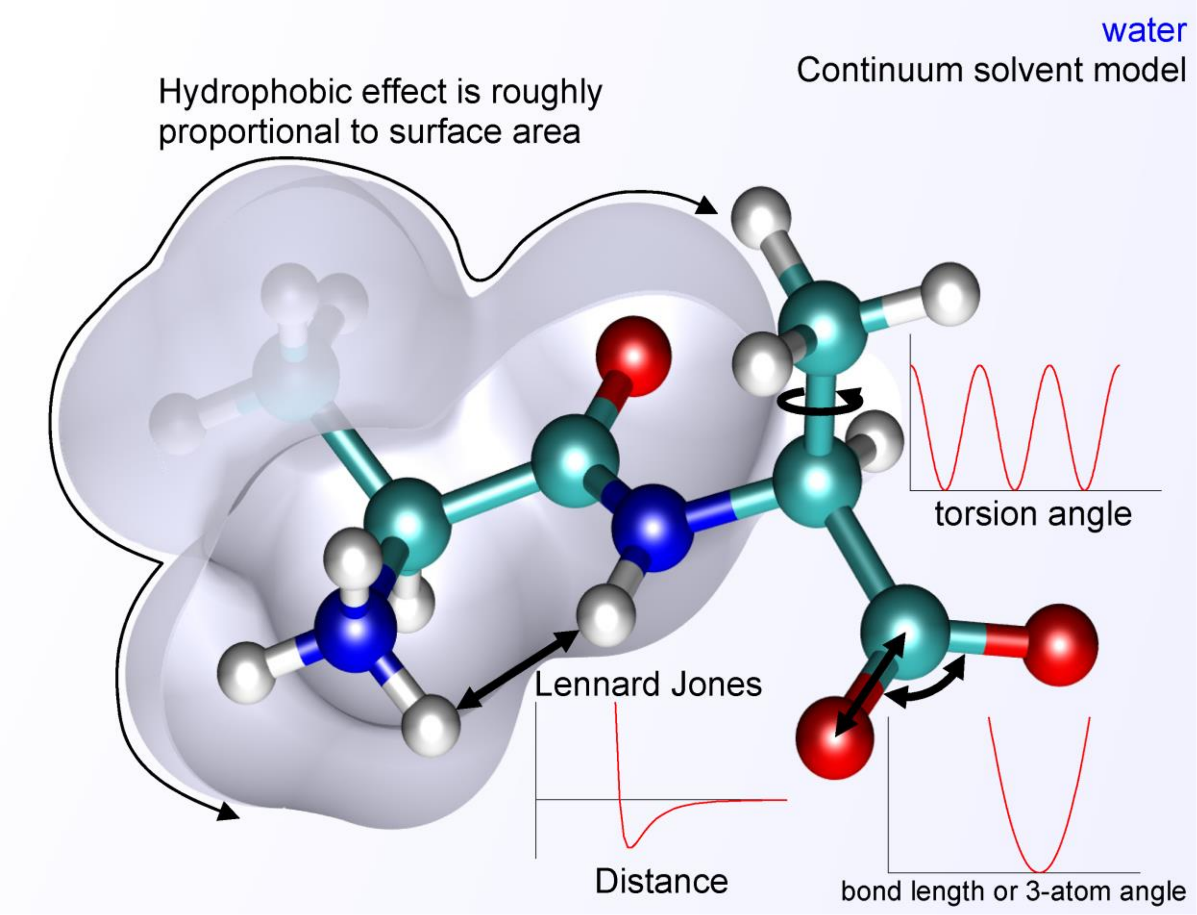
Classical Alchemy: Theory

Core idea

- Free energy is a state function
- Free energy differences
  … drive action
  … are hard to calculate (if at all)
- Following alchemical paths yields such differences

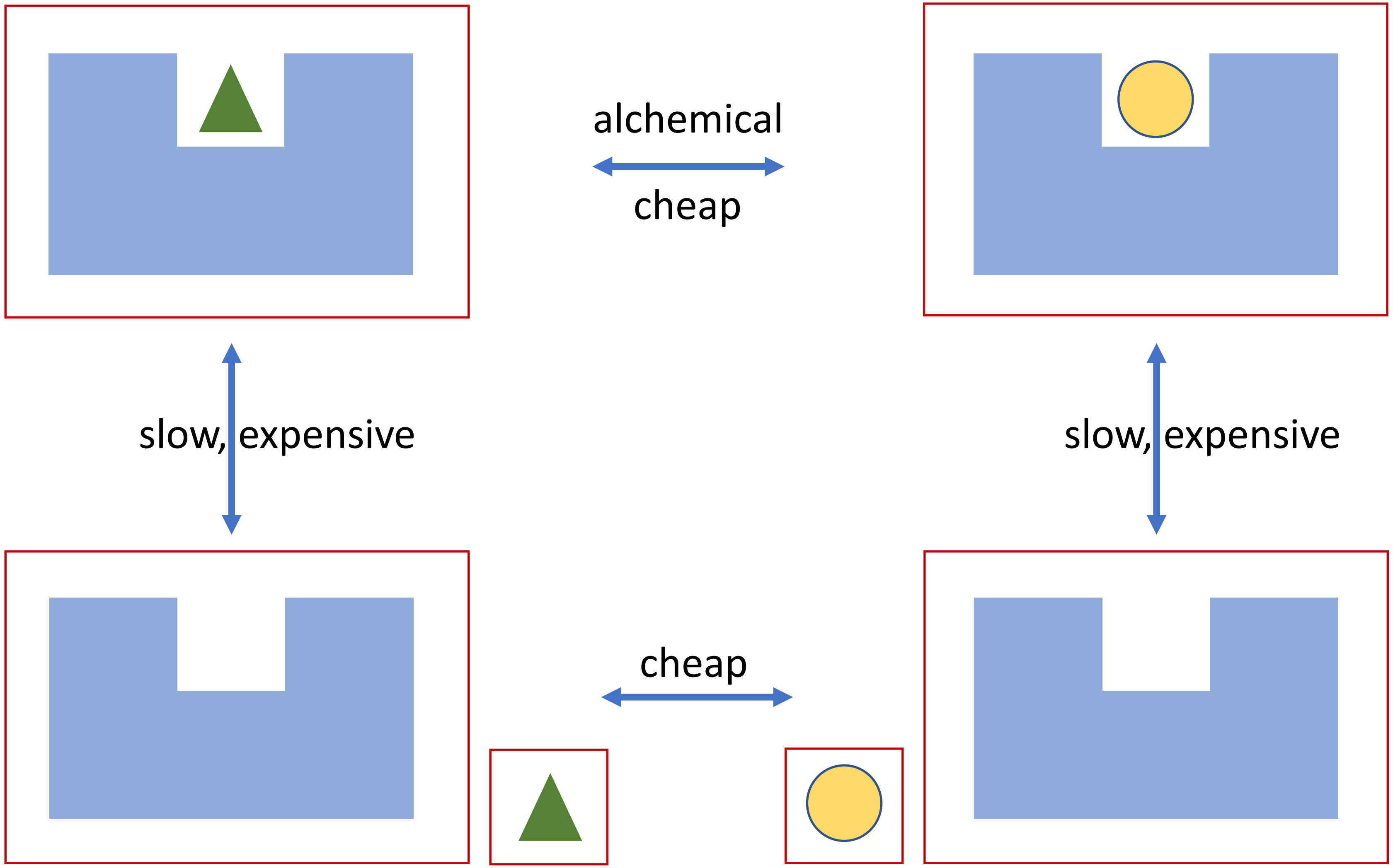
Context

- Classical force fields (all-classical description)
  - Bond, angle, dihedral
  - Non-bonded: Lennard Jones
- Interactions get scaled
  - Masses stay constant
  - Charges are scaled
- No need to be linear in interpolation
- Often seen within a molecular dynamics trajectory
- No charge conservation required!


water
Continuum solvent model
Hydrophobic effect is roughly proportional to surface area
torsion angle
Lennard Jones
Distance
bond length or 3-atom angle

https://commons.wikimedia.org/wiki/File:MM_PEF.png

Typical application: Thermodynamic Cycles



Theory

- Two states: i, j
- NVT ensemble, equilibrated

$$\Delta A_{ij} \equiv -k_B T \ln \frac{Q_j}{Q_i}$$

$$Q_i \equiv \int_{\Gamma_i} \exp\left[-\frac{U_i(\vec{q})}{k_B T}\right] d\vec{q}$$

Helmholtz free energy

Temperature

Partition function

Phase space

Potential energy

Theory: Molecular dynamics

- Need: ensemble average
- Ergodicity: Time average = Ensemble average
- Propagate in small steps (~fs) explicit positions from repeated force evaluations

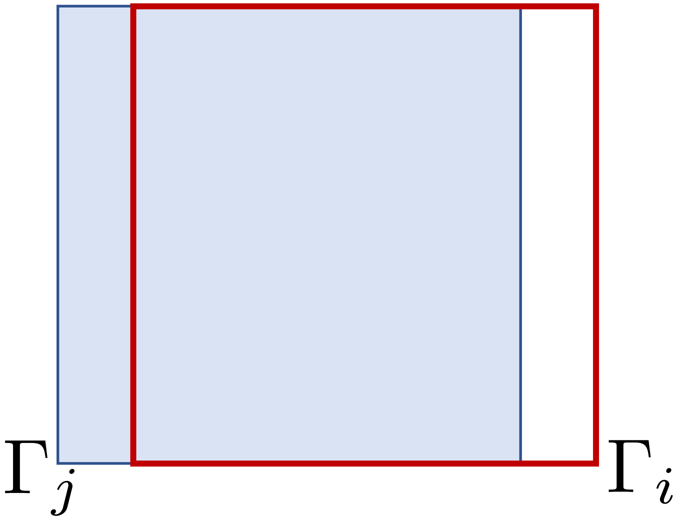
- Challenges
  - Numerically stable
  - Time reversibility
  - Thermostats / Barostats
  - Equilibration

Questions:

- Why are classical calculations typically time-reversible but quantum mechanical calculations are not?
- What does it mean if a setup is lacking time-reversibility?

Problem: Phase space overlap

- Two states: i, j
- NVT ensemble



$$Q_i \equiv \int_{\Gamma_i} \exp\left[-\frac{U_i(\vec{q})}{k_B T}\right] d\vec{q}$$

Partition function

Phase space

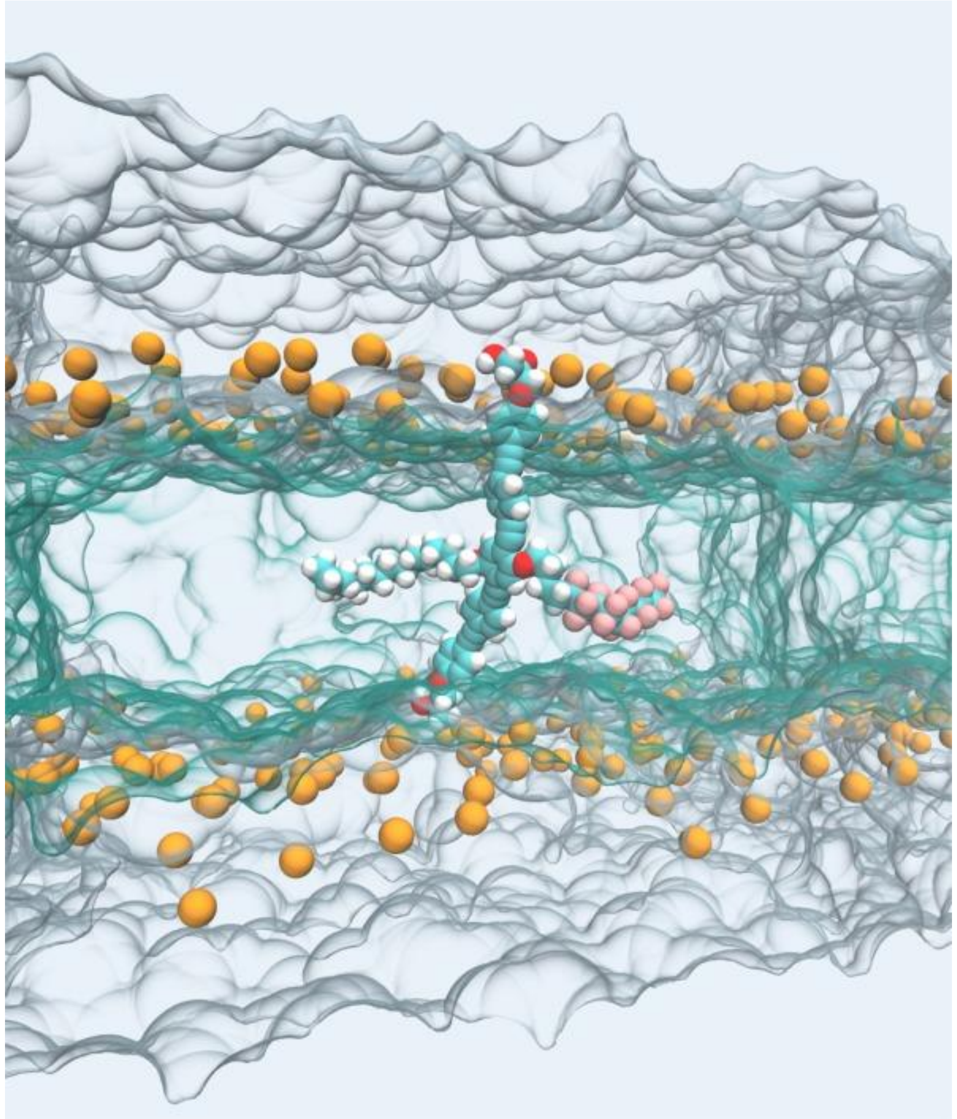
Potential energy

Example: hard spheres with different radii: close interaction never happens

Problem: Singularities

Insert molecule in membrane: turn on interactions

$$V_{\mathrm{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$



Questions:

1. How to “turn on” interactions?
2. Why might that fail?
3. What happens if that fails?

GFvR, T. Watermann, X.-Y. Guo, D. Sebastiani, *J. Comput. Chem.* **2017.**

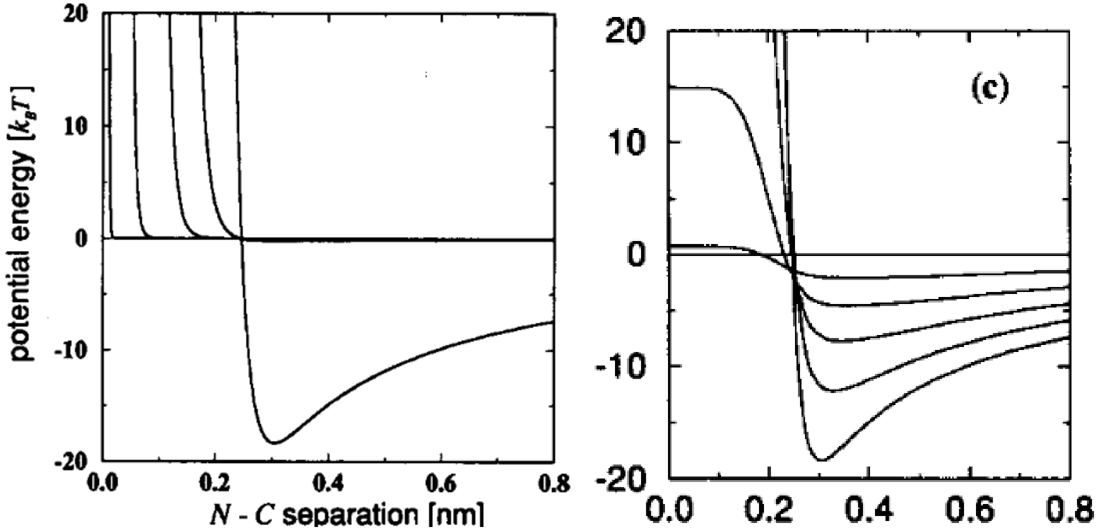
Problem: Hard potentials

- No matter the scaling: unbounded energy

$$V_{\mathrm{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

- Unbounded energy: no reliable derivatives
  - Question: Why is that an issue in molecular dynamics?
- Solution: soft-core potentials

T.C. Beutler et al. / Chemical Physics Letters 222 (1994) 529–539



$$U(\lambda, r) = 4\epsilon\lambda^{n}\left[\left(\alpha(1-\lambda)^{m} + \left(\frac{r}{\sigma}\right)^{6}\right)^{-2} - \left(\alpha(1-\lambda)^{m} + \left(\frac{r}{\sigma}\right)^{6}\right)^{-1}\right]$$

Problem: Charges and Lennard Jones

- Typical molecules: effective charges on each site + Lennard Jones

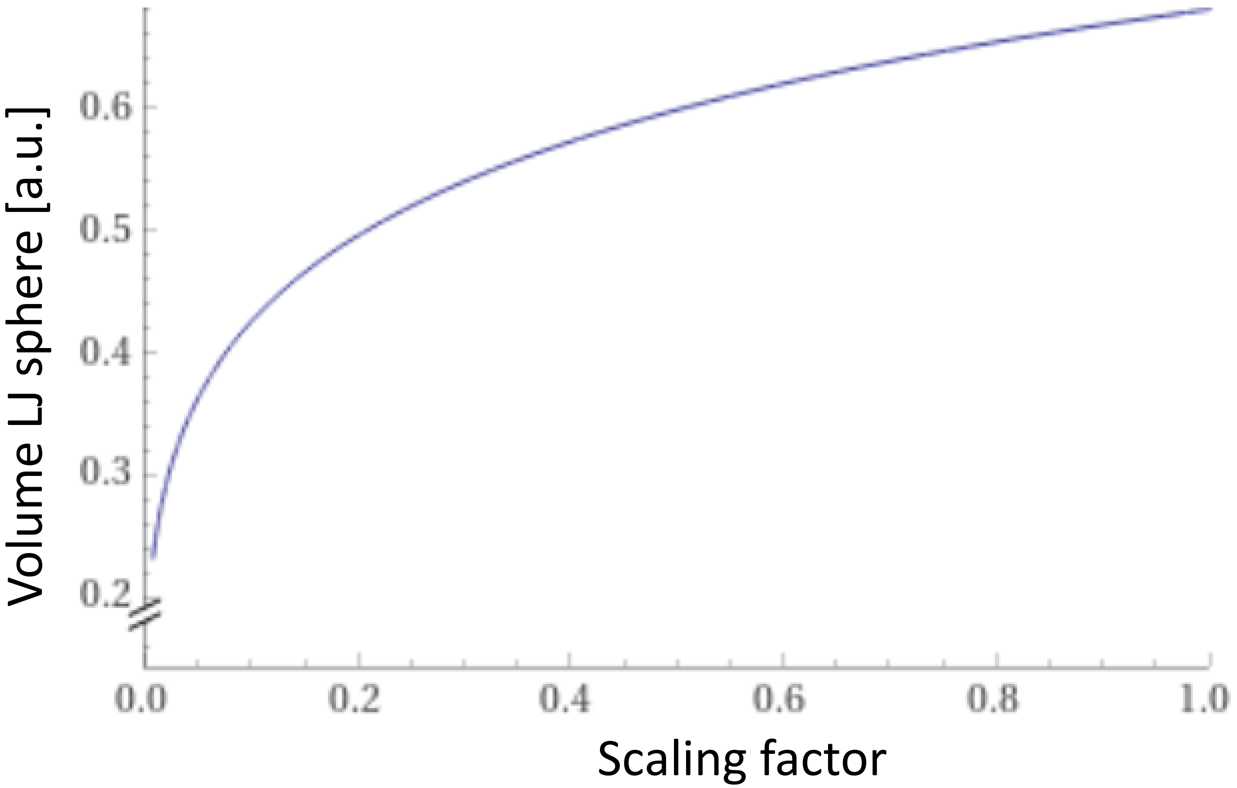
$$V_{\mathrm{LJ}}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right]$$

- Can be scaled independently
  - Energies remain state function of parameters
- Caveat:
  - If LJ is scaled: charges can get closer to each other. If charges are of opposite sign: trapping
  - Therefore: electrostatics first, LJ second
- Question: Would separate paths be acceptable and if so, why?

Problem: Pathways

Linear pathway not necessarily efficient / rarely “effectively linear”

$$V_{\mathrm{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$



Problem: Pathways

- Avoid constrained/restrained configurations

Why?

- Choose low-change path: large changes mean large derivatives

Why is that bad?

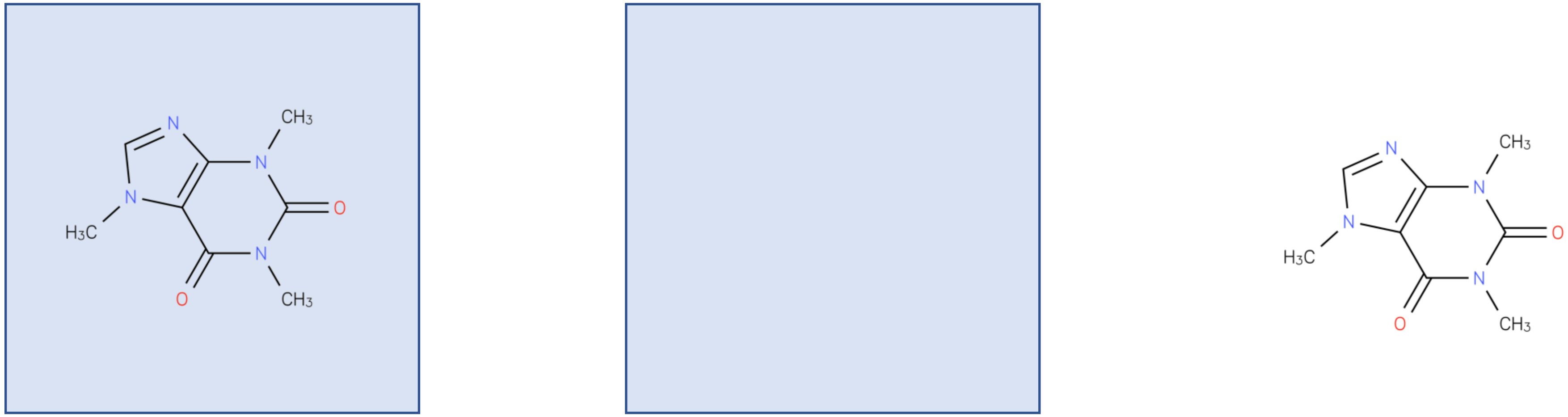
- Change parameters to create effectively linear results
- Restrict number of intermediates (=mixed states)
- Beware electrostatics: keep net charge

Why?

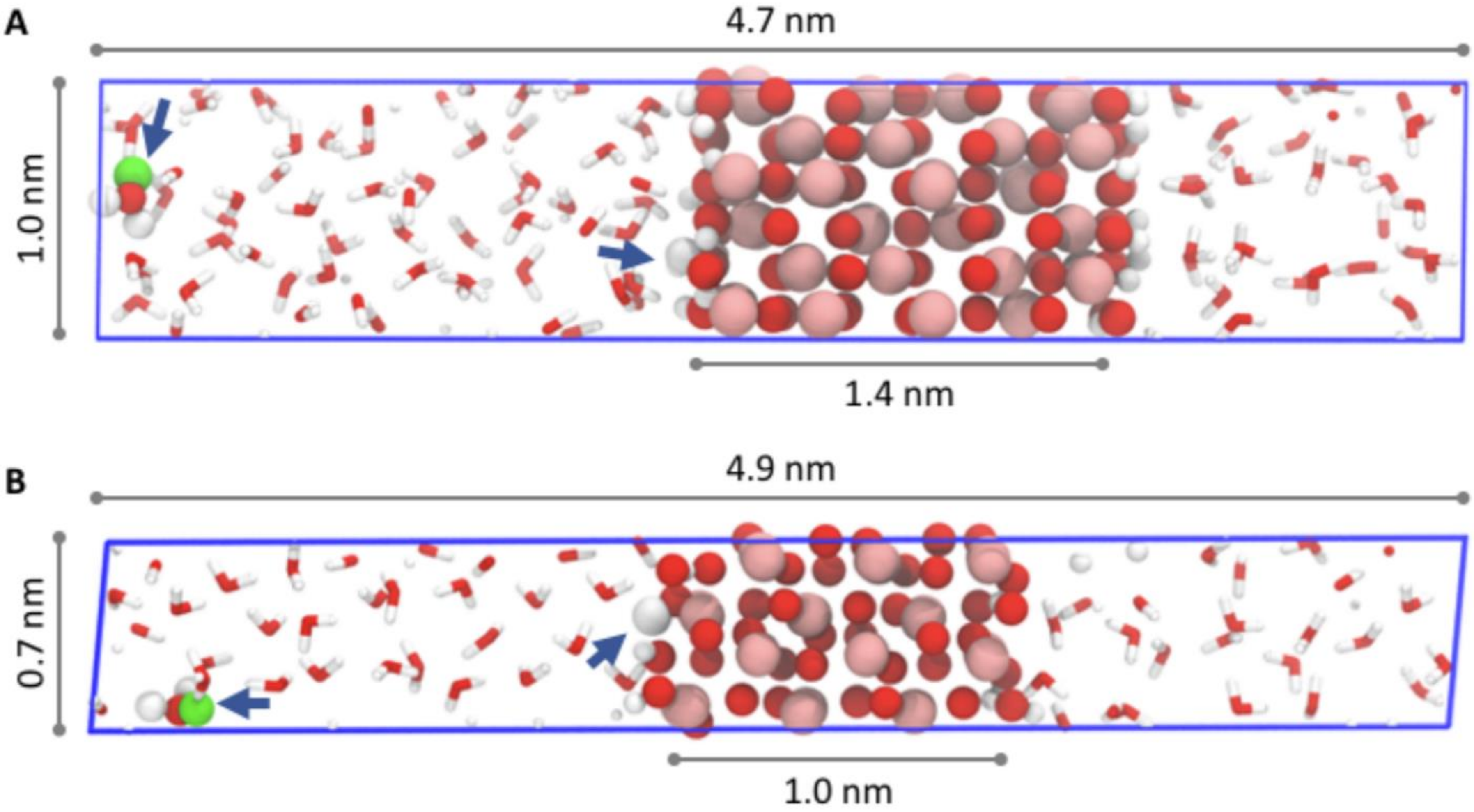
Problem: Pathways

Consider NVT:

Why is the free energy of solvation NOT simply the free energy differences with solute-solvent interactions turned off?



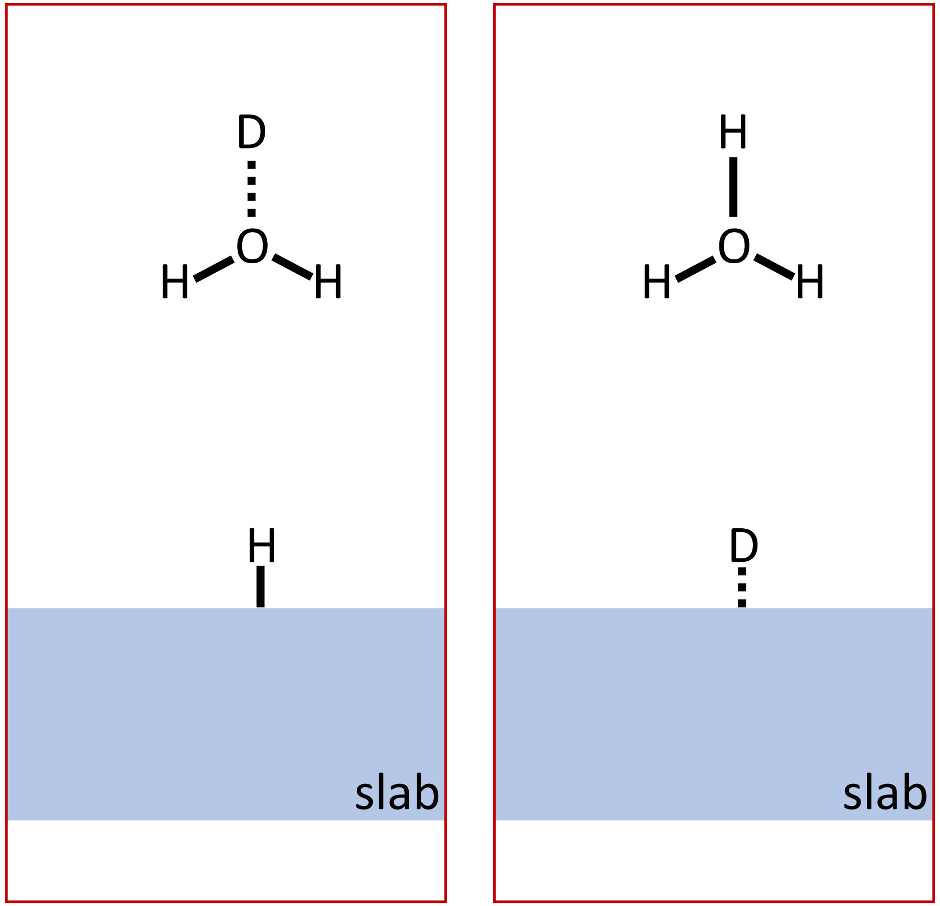
Problem: System mismatch



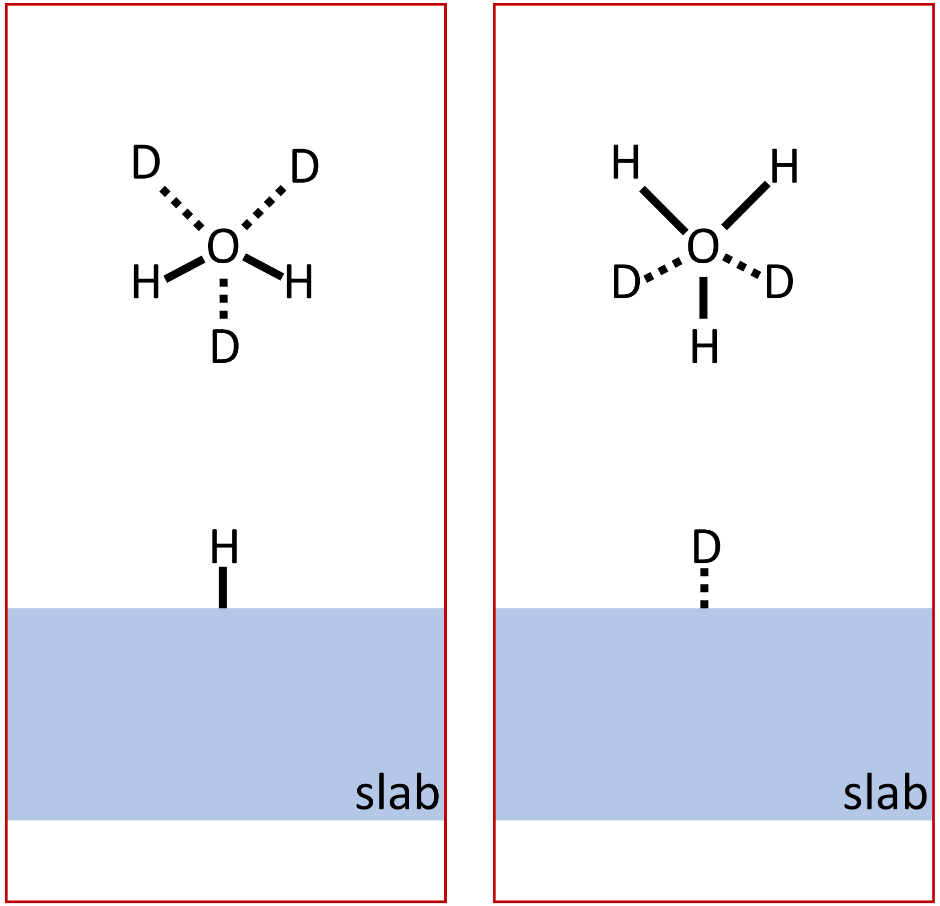
O. Gittus, GFvR, K. Rosso, J. Blumberger, *J. Phys. Chem. Lett.* **2018.**

Problem: System mismatch

Single topology approach

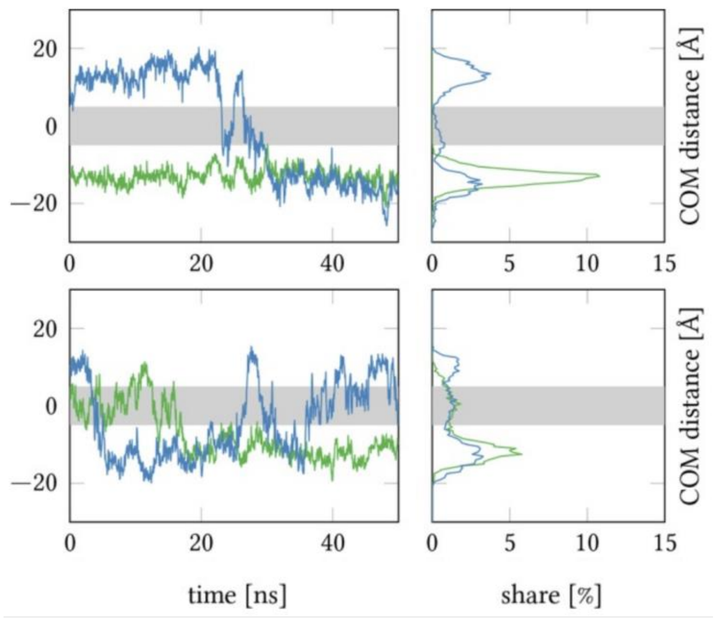


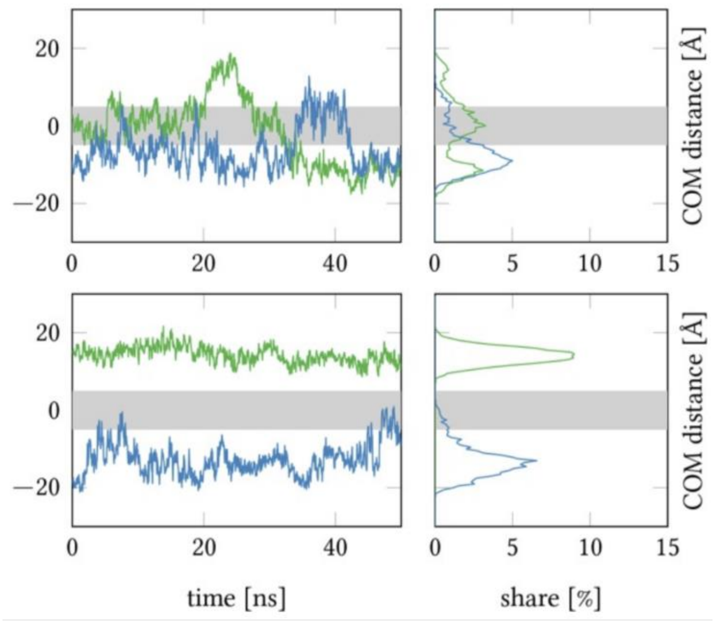
Double topology approach

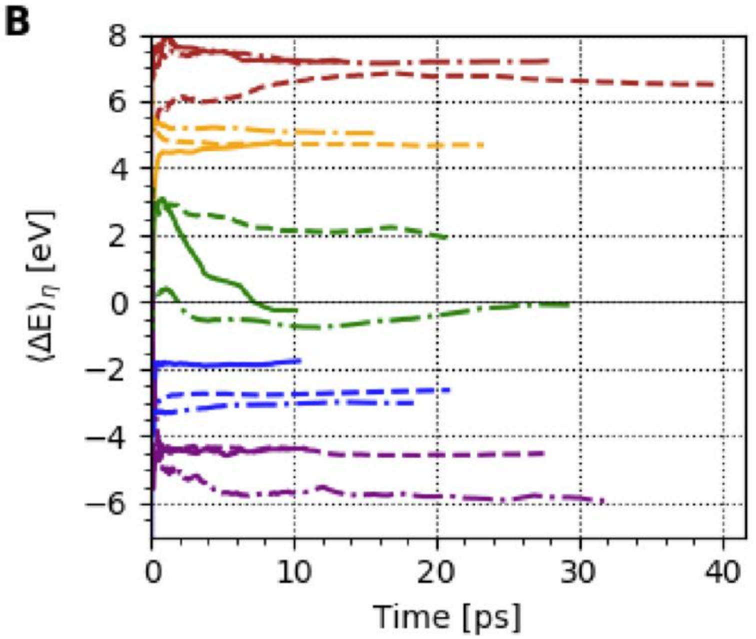


Problem: Equilibration

Conformational sampling slow – four independent trajectories





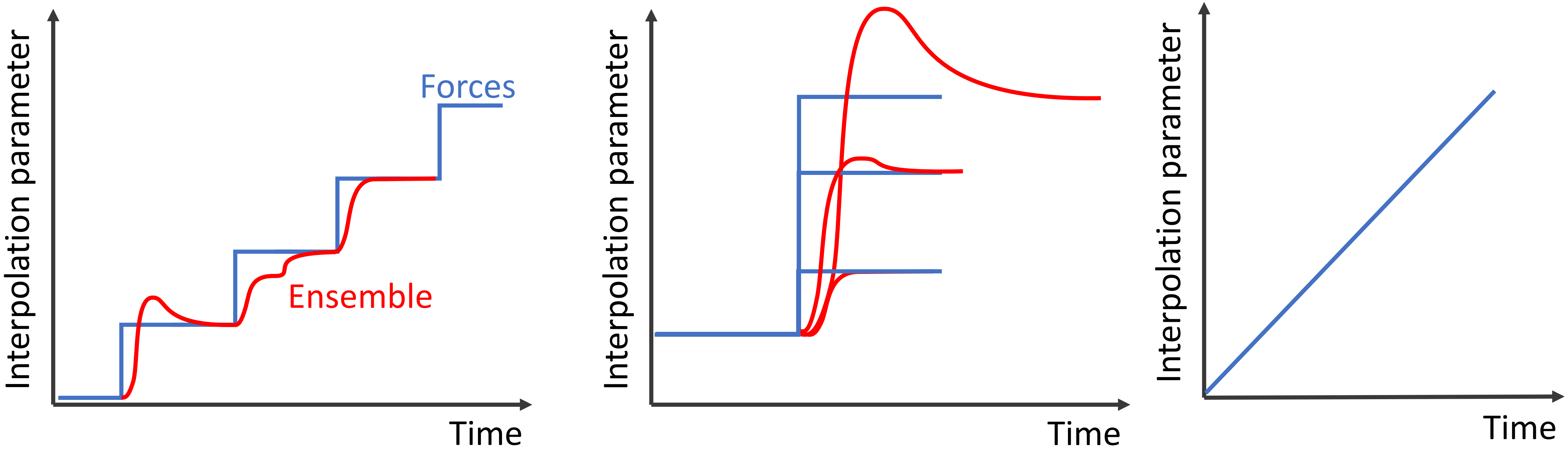


High auto-correlation

O. Gittus, GFvR, K. Rosso, J. Blumberger, *J. Phys. Chem. Lett.* **2018.**
GFvR, T. Watermann, X.-Y. Guo, D. Sebastiani, *J. Comput. Chem.* **2017.**

Problem: Lagged ensemble

Molecular dynamics: propagation of configurations: auto-correlated, delayed

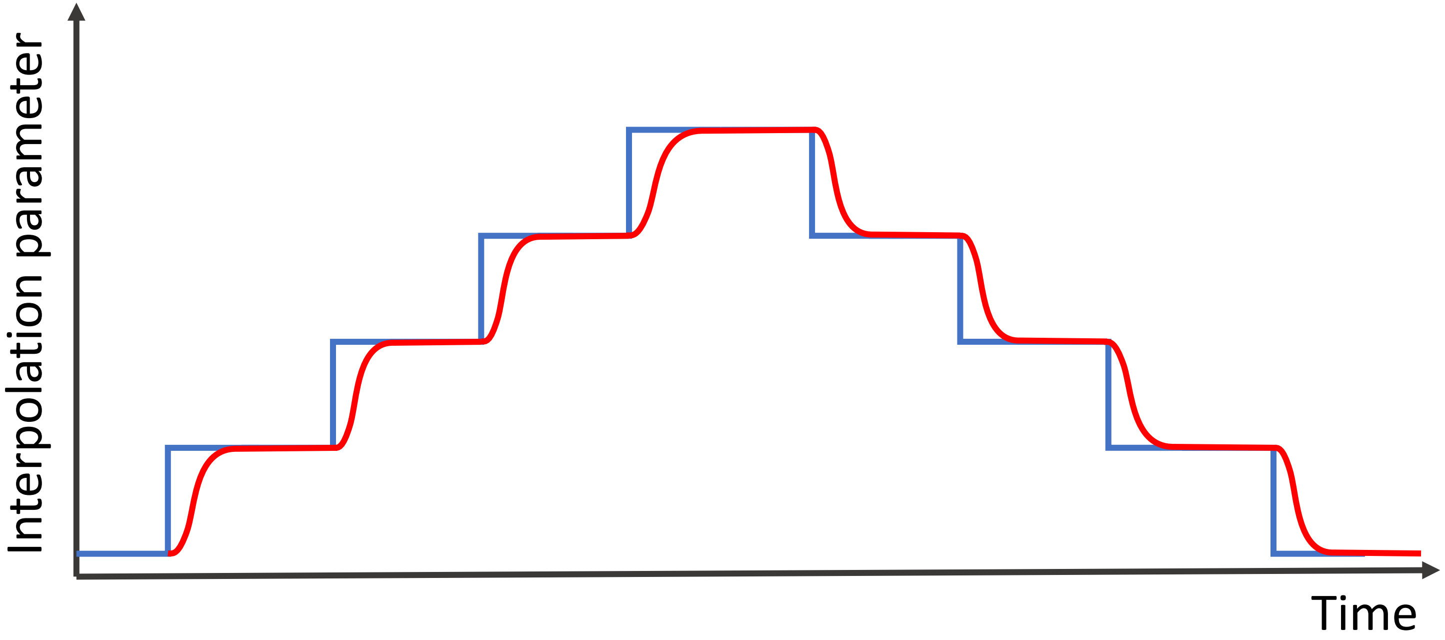


Questions:

1. Which limit is required?
2. Third panel: What does the ensemble do?

Problem: Hysteresis

- Convergence hard to achieve in practise
  - How about being wrong twice?



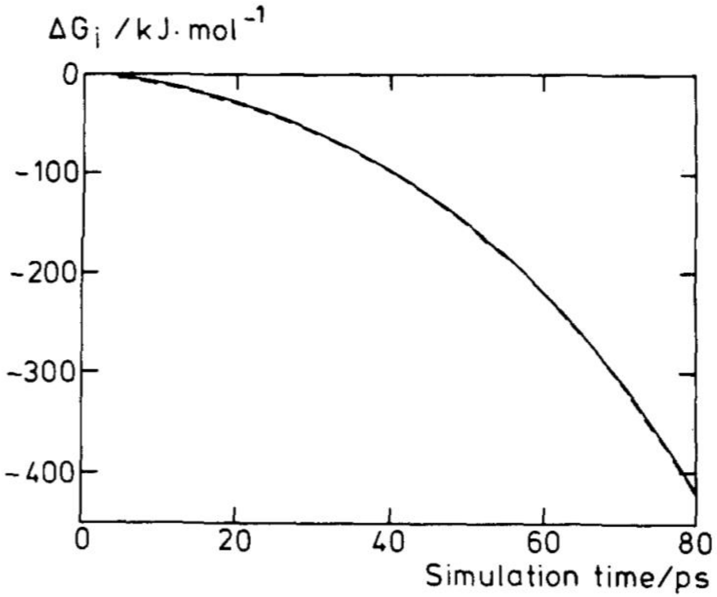


FIG. 1. Free energy of hydration change in a thermodynamic integration from neon to a sodium cation in 80 ps. The dashed curve gives the free energy change of the reverse process.

More accurate: slower change. Low hysteresis does not mean accurate: why?

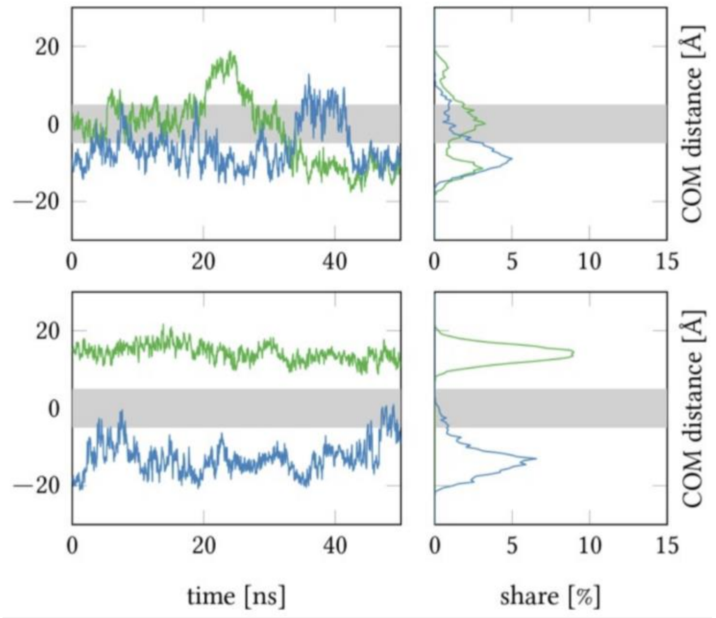
T.P. Straatsma, H.J.C. Berendsen, *J. Chem. Phys.* **1988.**

Problem: Uncertainties

Main sources: Correlation and sampling uncertainty

Method 1: **Bootstrapping** (What if we knew less?)

- Sub-sample observations, estimate uncertainty from variance
- Risks:
  - Rare events
  - Biased sampling



Method 2: **Autocorrelation** (What if we measured independently?)

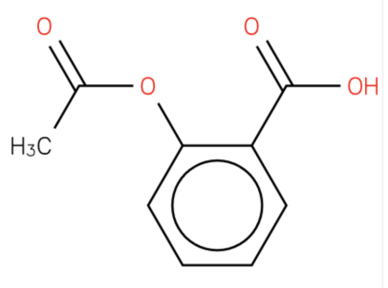
- Observable over time, quantify lagging
- Risks:
  - Not necessarily static: fast/slow process
  - Complex interplay with constraints

Method 3: **Block averages** (What if time domains are representative?)

- Slice time in blocks and treat them as observables
- Risks:
  - Hard to figure the duration out

Exercises

1 Sketch the single topology and double topology approach for one specific alchemical path change from aspirin to acetozone. Try to find the shortest path in terms of sites alchemically changed. Which method is likely to be preferable and why?

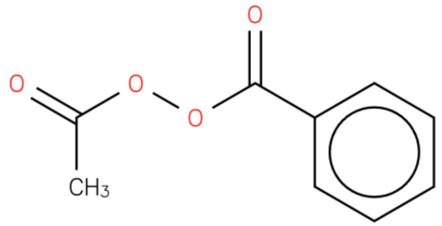


aspirin

2 Consider the surface deprotonation case: One question you might have is why not to vanish a surface proton instead of the complicated way of placing the proton in the liquid.

a) List reasons why vanishing a proton is not the right thing to do.

b) Using the dimensions from the slides and approximating the unit cell to be rectangular: what change in pH is expected if you add/remove one proton? (Pick one slab only)



acetozone