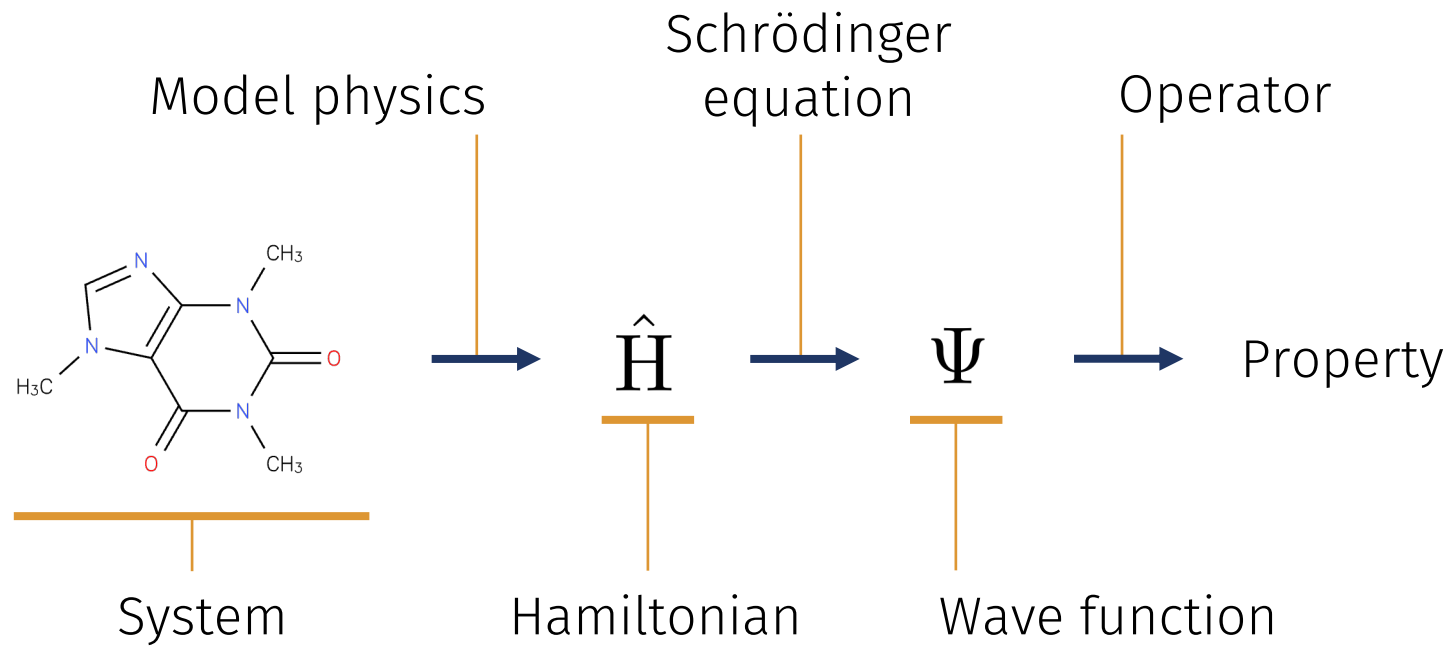


Alchemical Perturbations Characterizing Chemical Space

Guido von Rudorff, University of Kassel

 vonrudorff@uni-kassel.de

 nablachem.org/talks

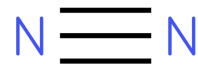


$$\Psi = \Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_n)$$

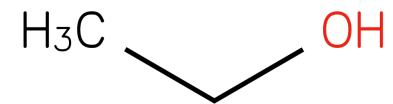
Methane



N2



Ethanol



Solved by approximations in computational chemistry?



Scaling of Molecules

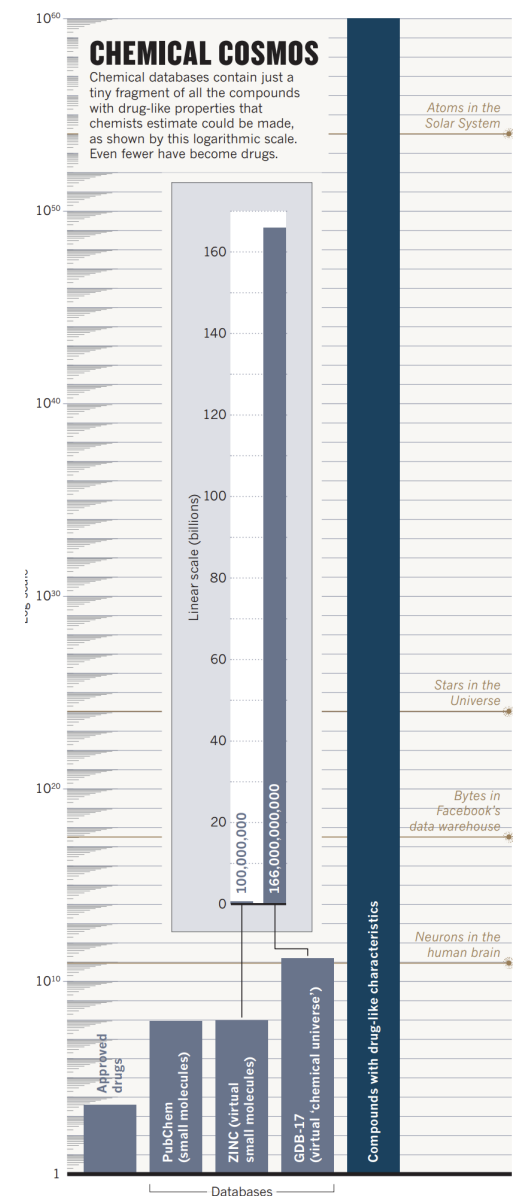
4

Commercial databases

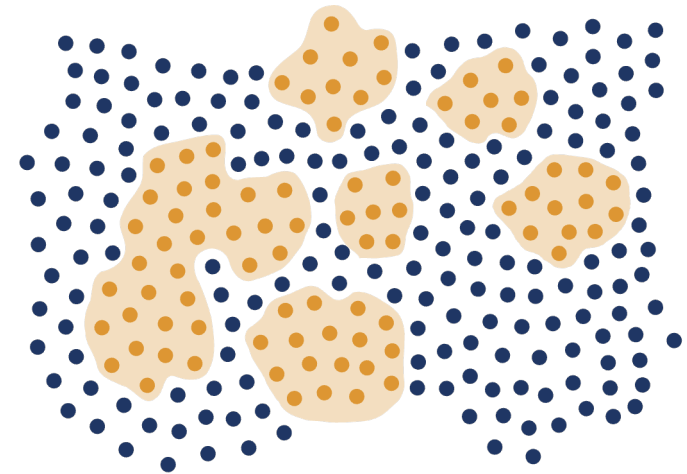
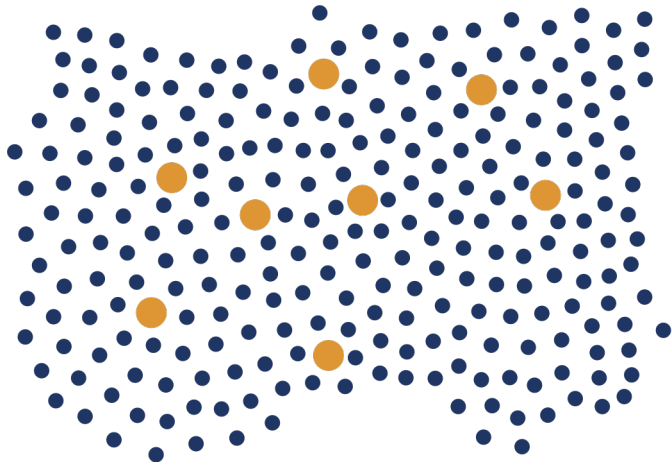
- 164 million molecules
- 15k added daily

Scale

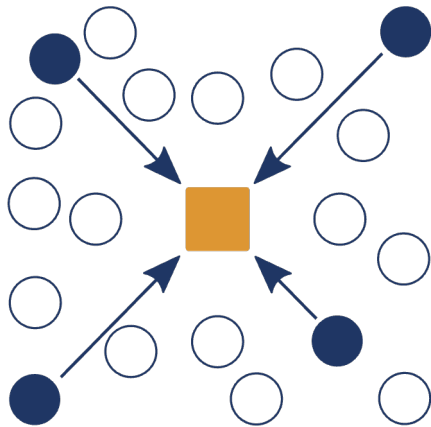
- One person: 1 million compounds/second
- 10 billion people on earth
- 10^{26} universe ages to go through



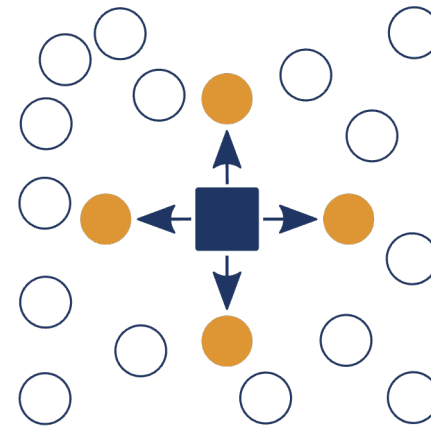
Speed does not matter:
even enumeration is impossible.



Machine Learning



Quantum Alchemy



Idea

Treat system changes perturbatively^[1,2]

Build a Taylor/Padé approximant^[3]: often 100.000 times faster

Steps

Choose system



Alter system, calculate property response functions



Predict many modified systems



■ Forwards
■ Backwards

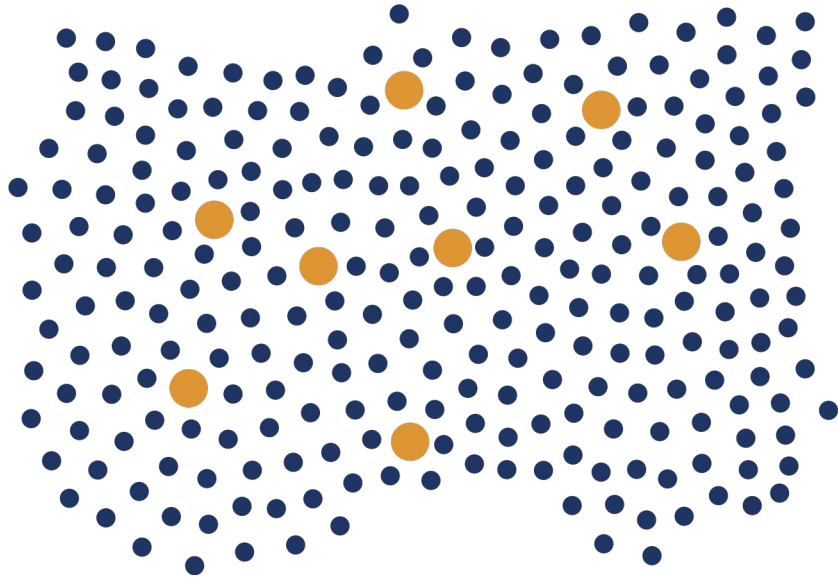
1 | E. B. Wilson, *J. Chem. Phys.* 1962. 2 | GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, 2020.

3 | GFvR, *J. Chem. Phys.*, 2021.

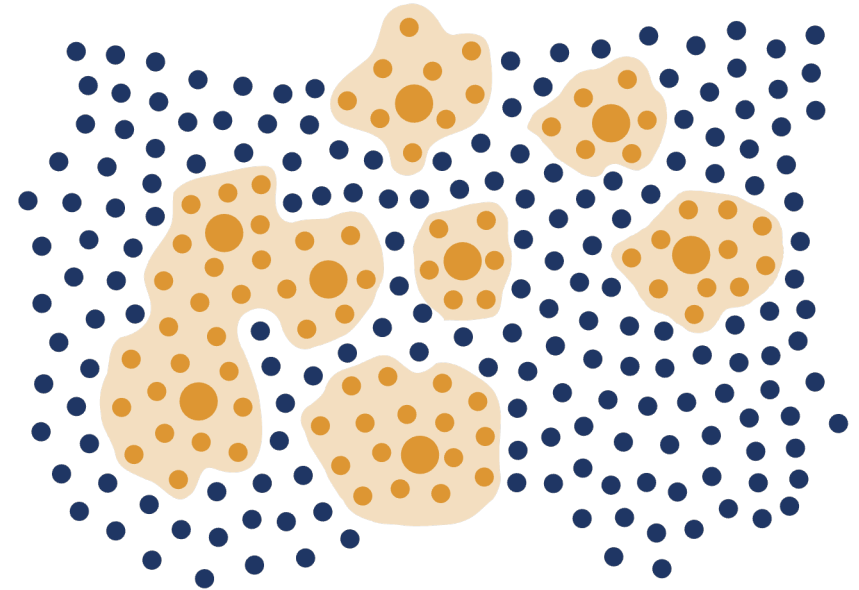
Few highly accurate calculations
instead of many intermediate ones

$$\hat{H} = \hat{H}(\underbrace{Z_i}_{4N}, \underbrace{\mathbf{R}_i}_{1\text{D, close to } \sum_i Z_i}, \underbrace{N_e}_{1\text{D}}, \sigma)$$

Without Perturbation



With Perturbation



Systems

- Any
- Known
- Approximated

Differentiable / Analytic + Converge quickly

- ✓ Total Energy ^[1,2]
- ✓ Dipole moments ^[2]
- ✓ Deprotonation energies ^[3]
- ✓ Photoelectron circular dichroism ^[5]
- ✓ Electron density ^[1,2]
- ✓ Non-covalent interactions ^[1]
- ✓ Ionisation Energy ^[4]
- ✓ Orbital eigenvalues ^[2]
- ✓ Binding energies ^[1,2]
- ✓ Electron Affinity ^[4]

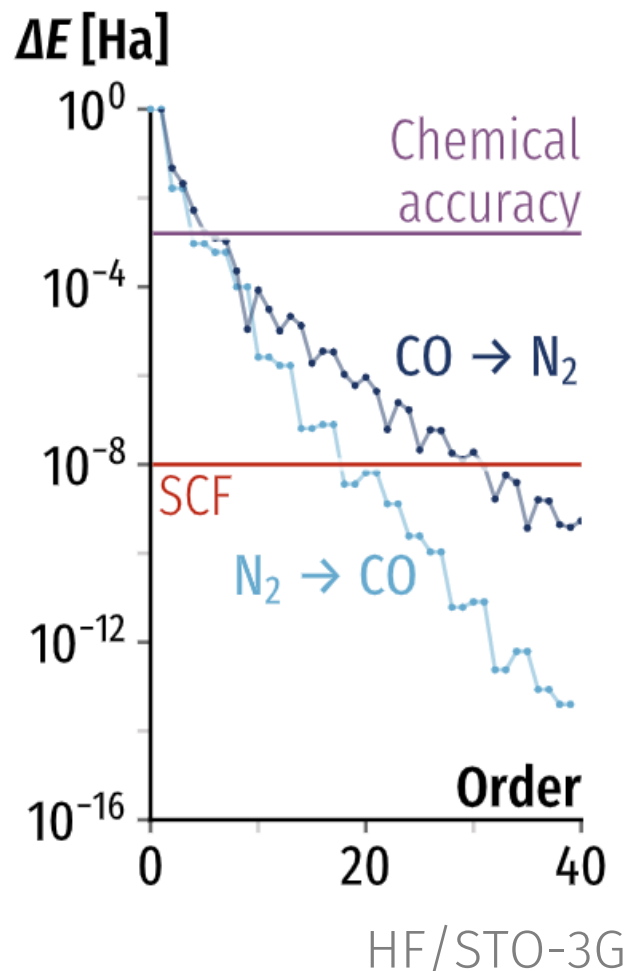
In progress

NMR spectra

Tested in: Gaussian, Psi4, PySCF, MRCC, ORCA, cp2k, CPMD, Quax, dqc, DiffiQult

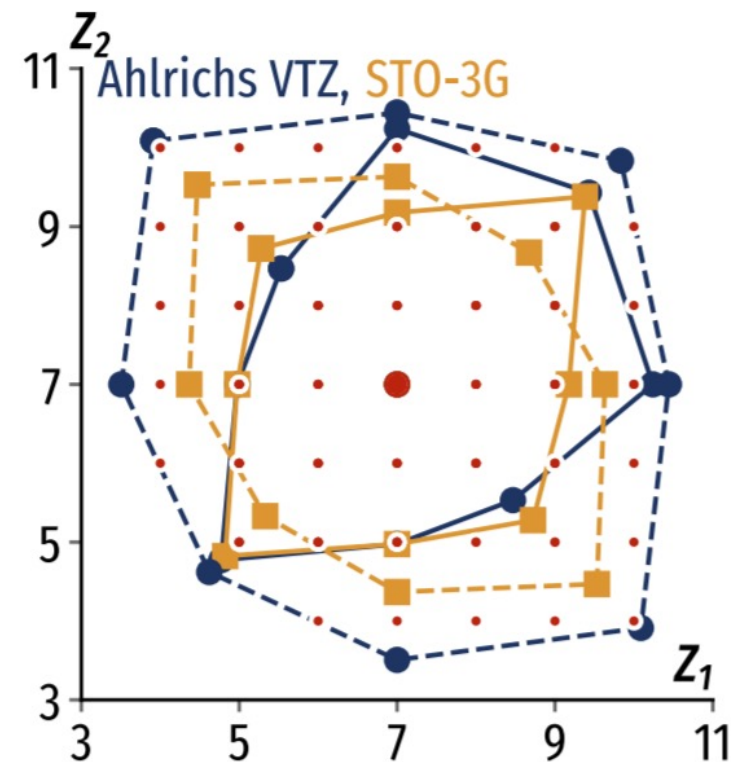
Tested with: HF, DFT, CCSD

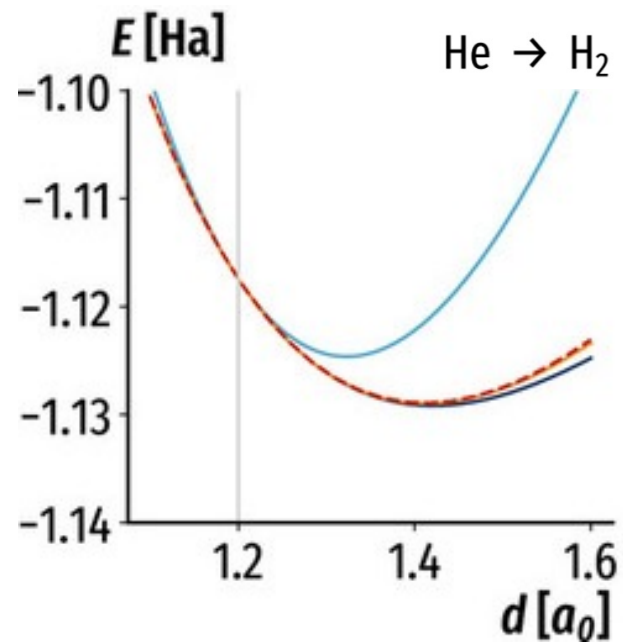
1 | GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, 2020. **2** | GFvR, *J. Chem. Phys.*, 2021. **3** | GFvR, O. A. von Lilienfeld, *Phys. Chem. Chem. Phys.*, 2020. **4** | E Eikey, A Maldonado, C Griego, GFvR, J Keith, *J. Chem. Phys.*, 2022. **5** | GFvR, A Artemyev, B Lagutin, P Demekhin, *J. Chem. Phys.*, 2024.



Taylor expansion

- First terms accurate enough
 - Truncate early
- Converges to the right value
- Large convergence radius
- Scales with chemical space



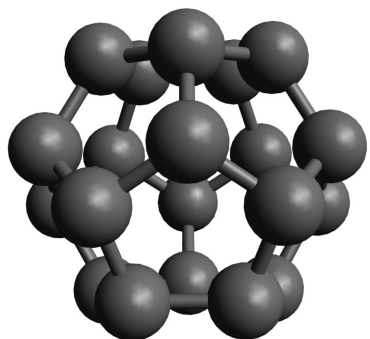


Taylor expansion

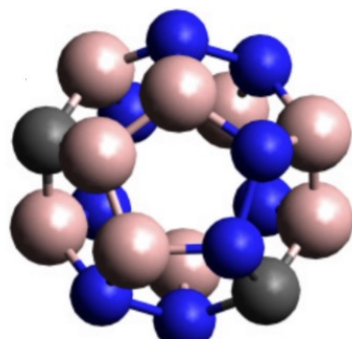
- Large changes still converge (more slowly)
- Geometric response can be recovered

Scaling with chemical space

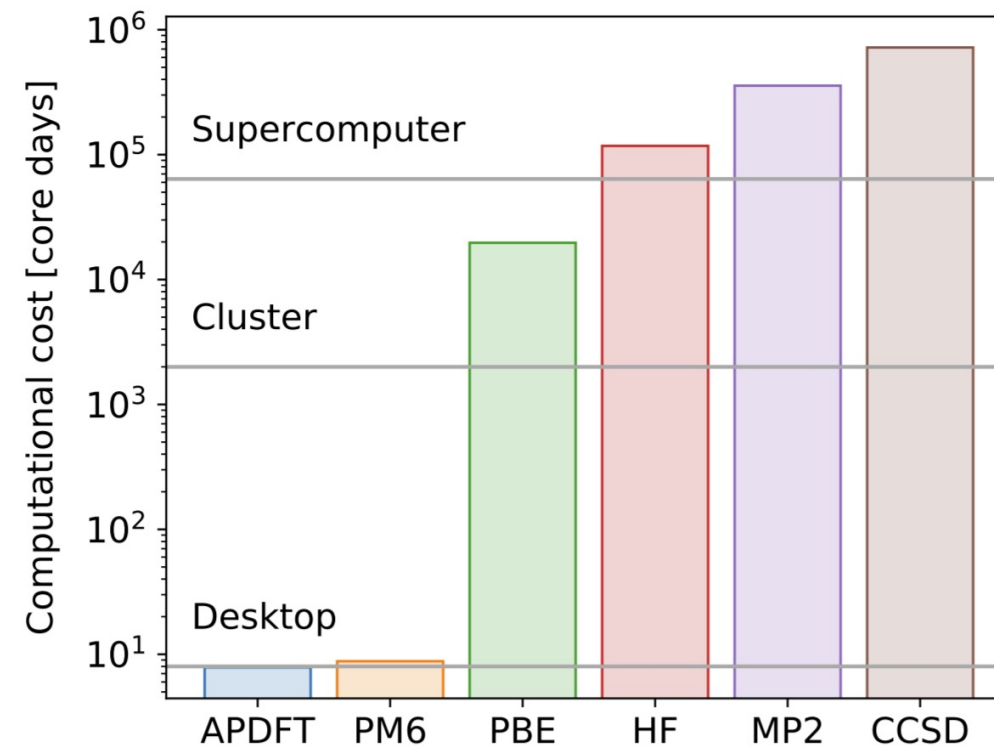
- 1 derivative for second order
- 5 derivatives for third order



C_{20}



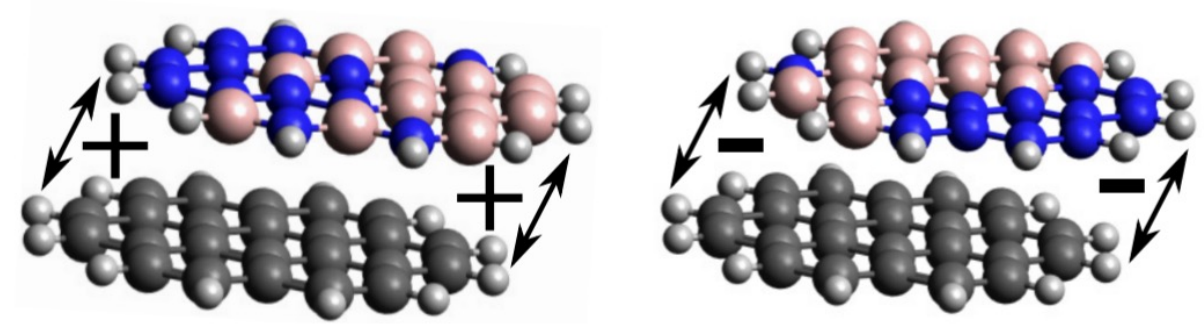
$3.1 \cdot 10^6$
targets



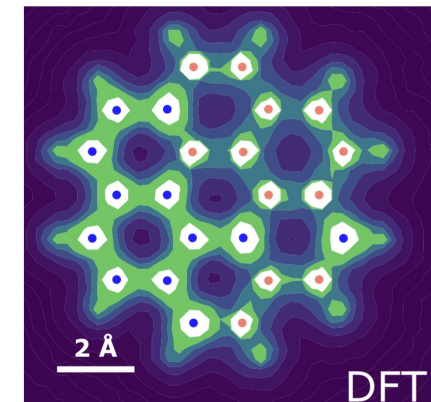
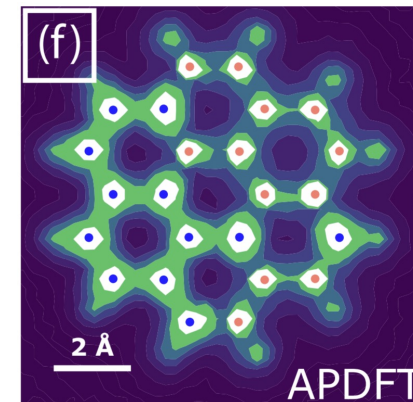
QA: 80.000x faster

BN-doped coronene dimer

- Identify most/least attractive doping pattern
- Design case

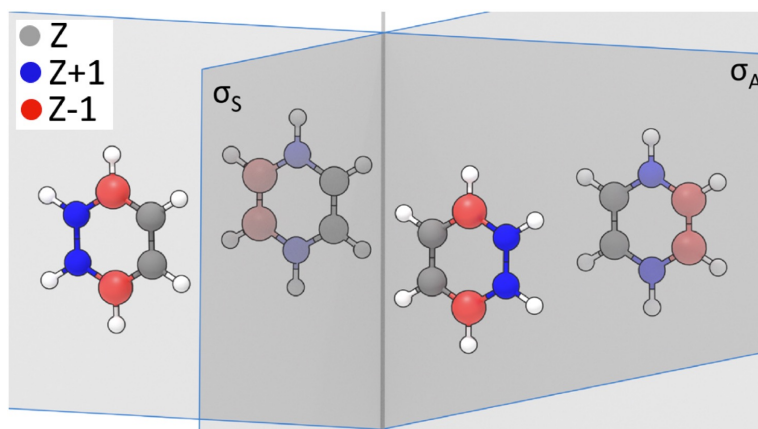


QA: 20.000x faster

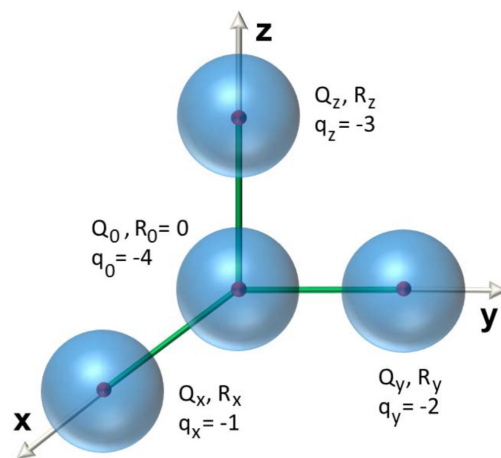


$2.8 \cdot 10^{10}$ targets

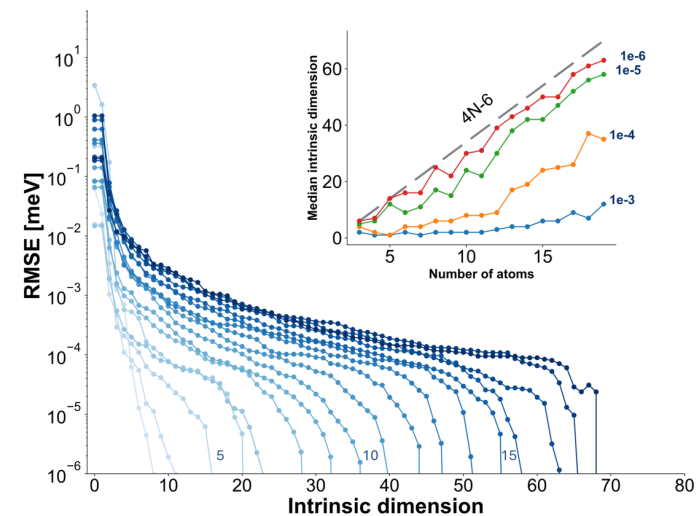
Ok, it works
What can I learn?



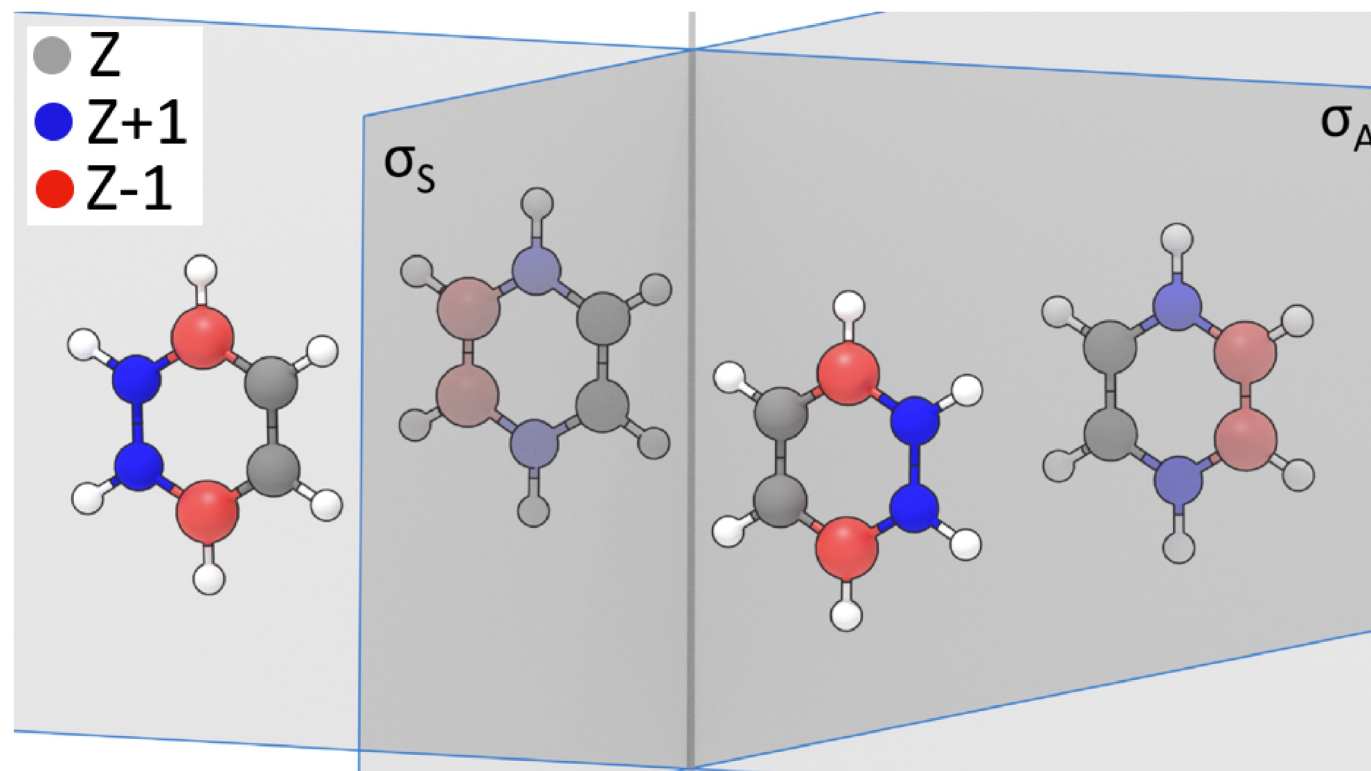
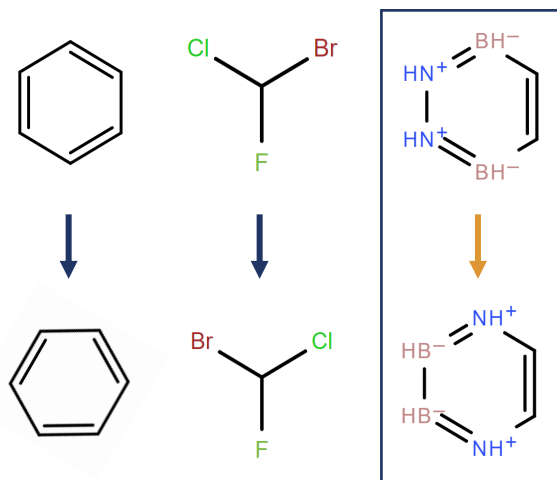
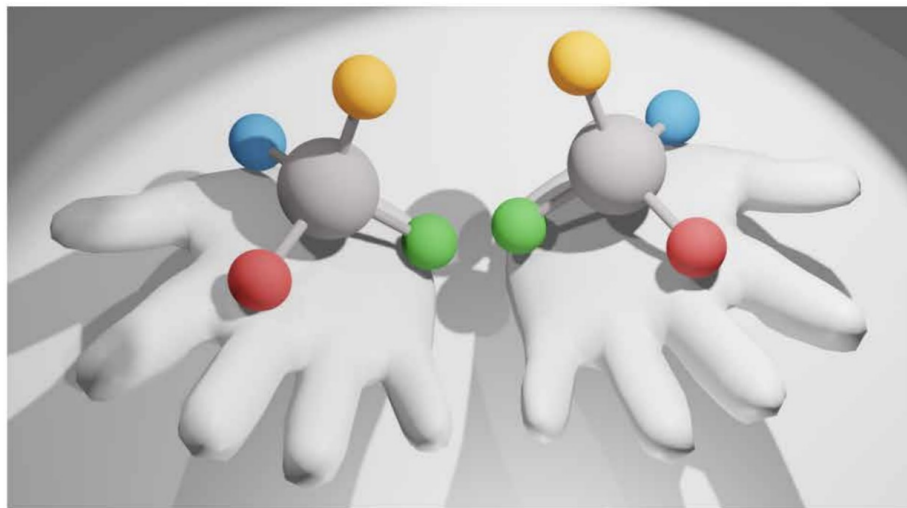
Symmetries



Interpretability

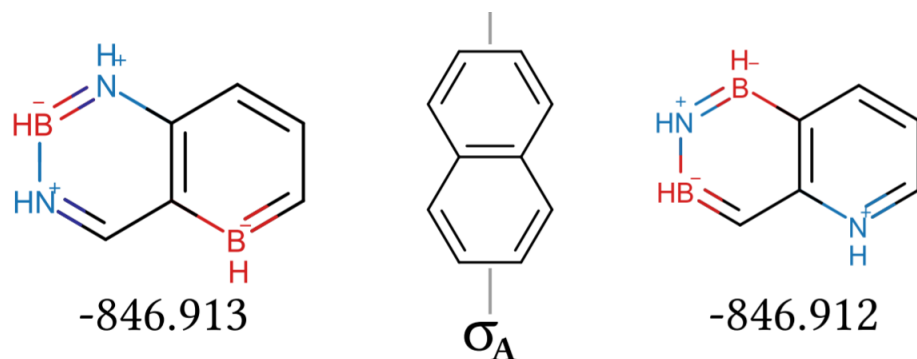


Intrinsic Dimension

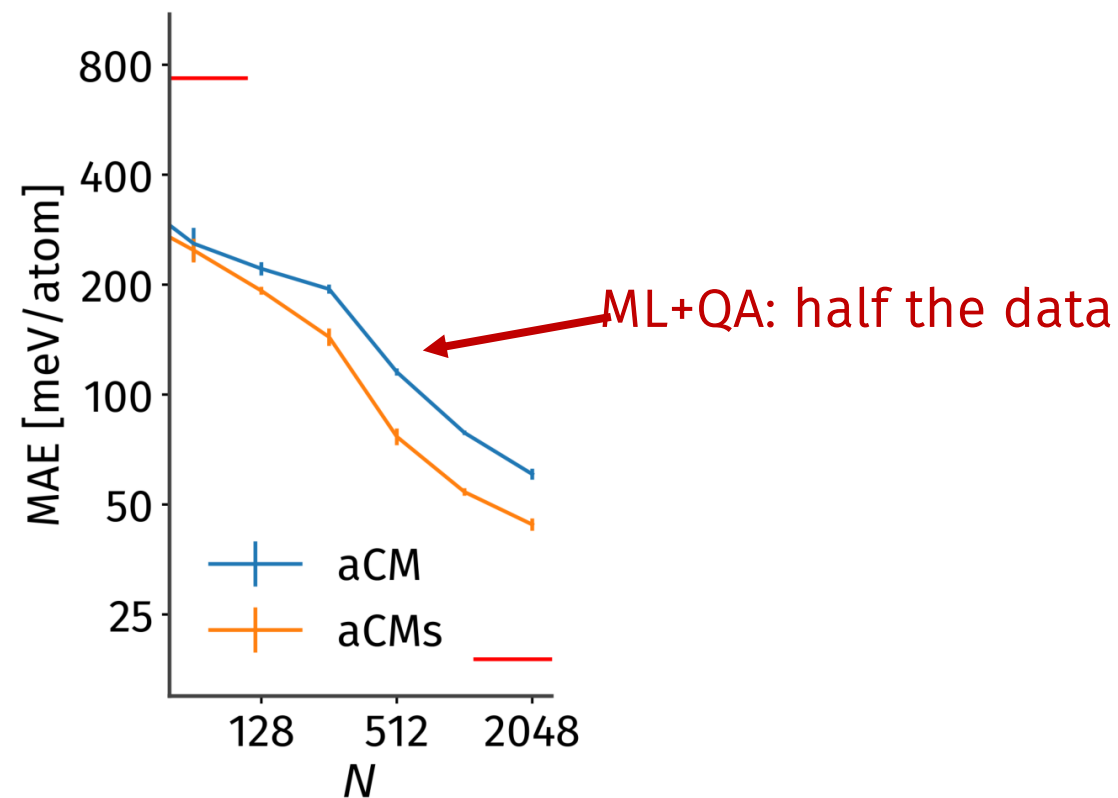


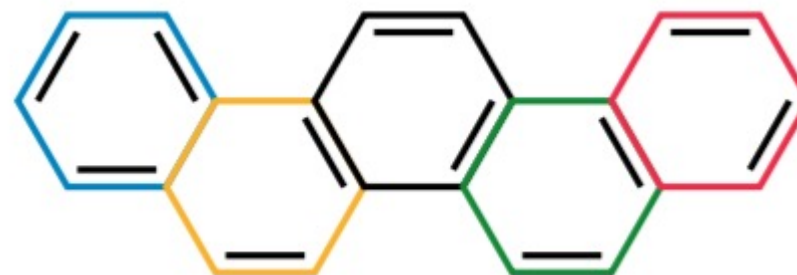
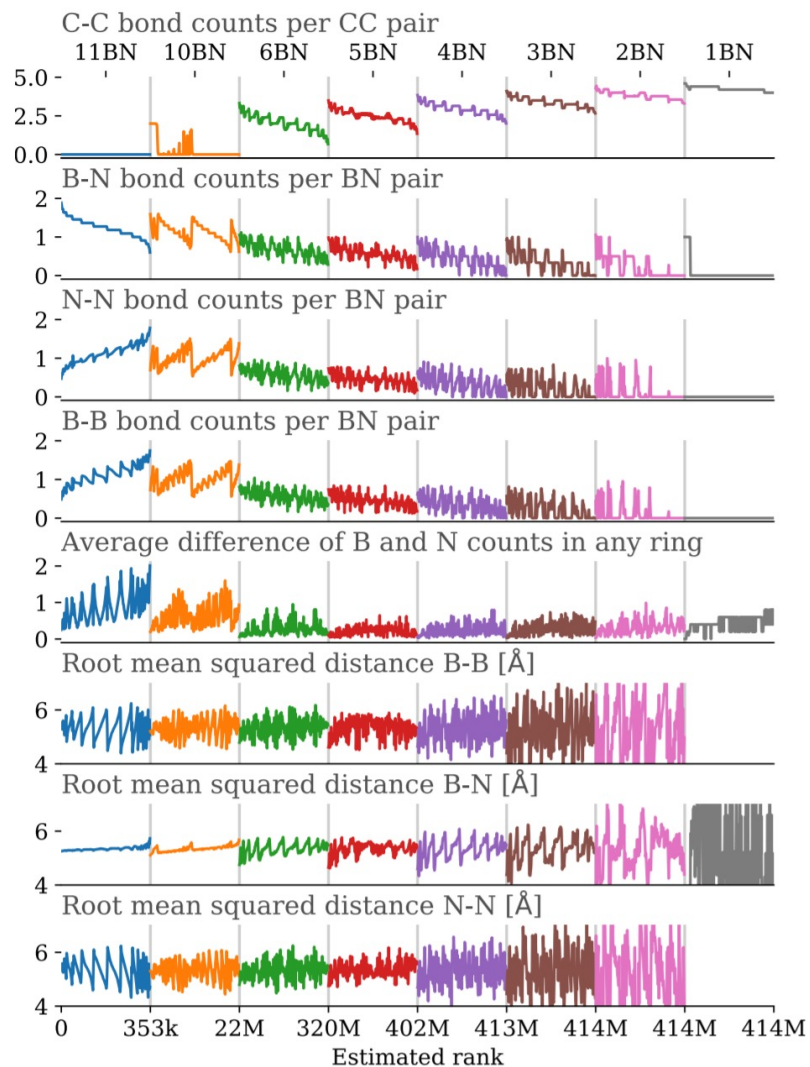
Fundamentally new symmetry

Electronic energy only



Speed up machine learning





x 414 M

Design rules in order of decreasing strength

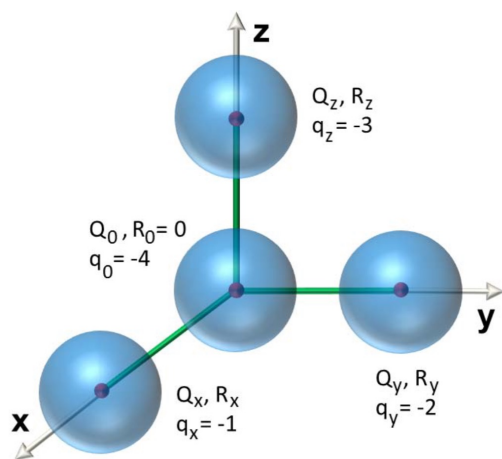
- Add BN pairs
- Maximize CC bonds
- Substitute sites shared between rings
- Maximize BN bonds
- Avoid N substitutions on rings sharing a larger amount of bonds with other rings
- Balance BN substitutions in each ring

QA: Millions at once!

Not a single QM calculation required!

Angular emission

- Expensive to calculate
- Highly coupled degrees of freedom: multidimensional expansion



Data

- 85 pairs of β_i
- Center: $Q_{xyz}=2.5$, $R_{xyz} = 3$, photoelectron = 6 eV

$$\frac{d\sigma^\pm}{d\Omega} = \frac{\sigma}{4\pi} \left[1 \pm \underbrace{\beta_1 P_1(\cos \theta)}_{\text{dichroic parameter}} - \frac{1}{2} \underbrace{\beta_2 P_2(\cos \theta)}_{\text{anisotropy parameter}} \right]$$

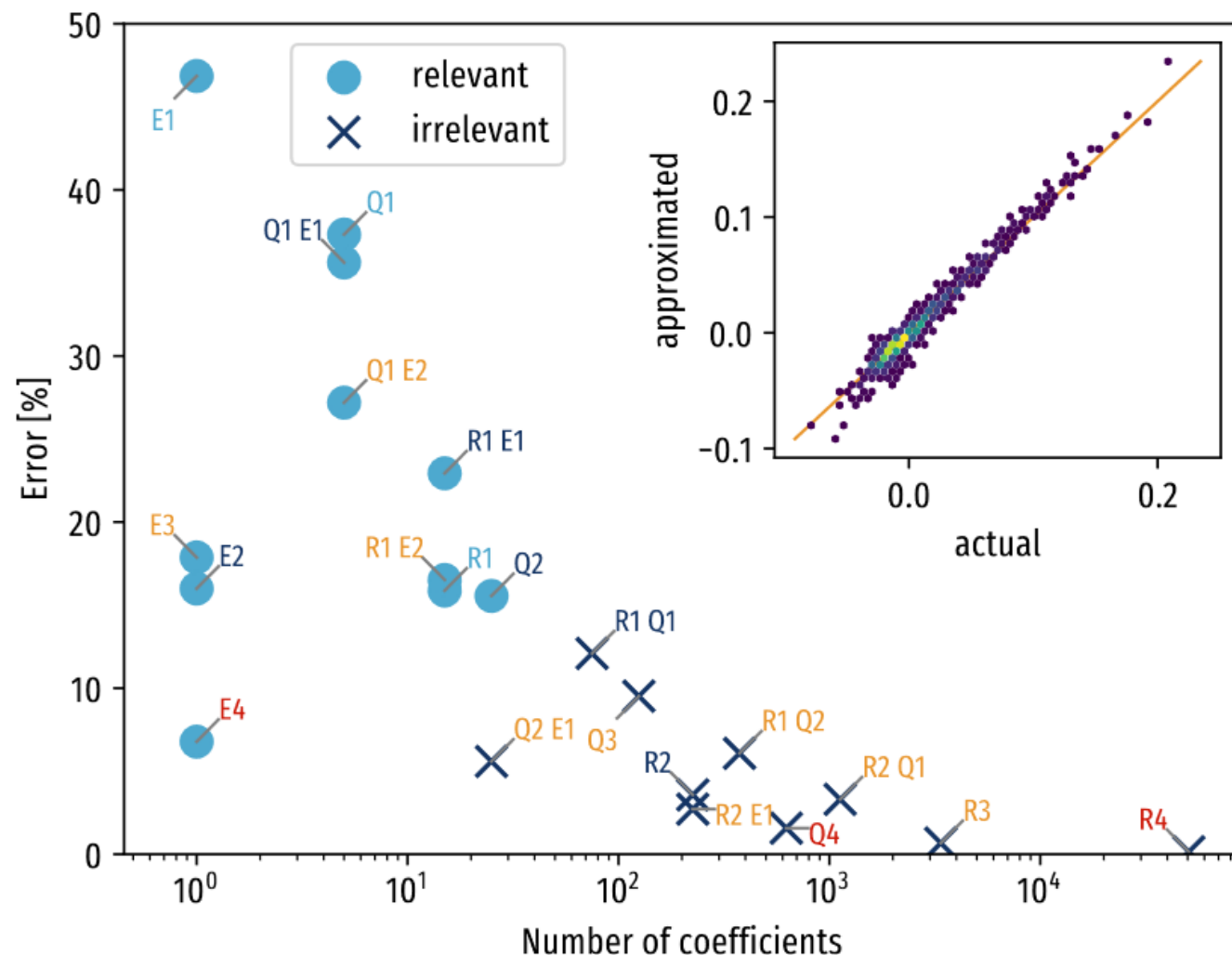
$$\beta_i(\mathbf{x}) \simeq \sum_{|\alpha| \leq k} \frac{\partial^{|\alpha|} \beta_i(\mathbf{a})}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \frac{(\mathbf{x} - \mathbf{a})^\alpha}{\alpha!}$$

Relate to cost

- Expensive terms also irrelevant
- For molecules, subset only
- Allows for efficient stencil design

Estimated speedup

$$\frac{3^N}{N^2 + 12N - 1}$$



Definition Intrinsic Dimension

Minimal number of degrees of freedom to describe a property.

≠ intrinsic dimension of a point cloud!

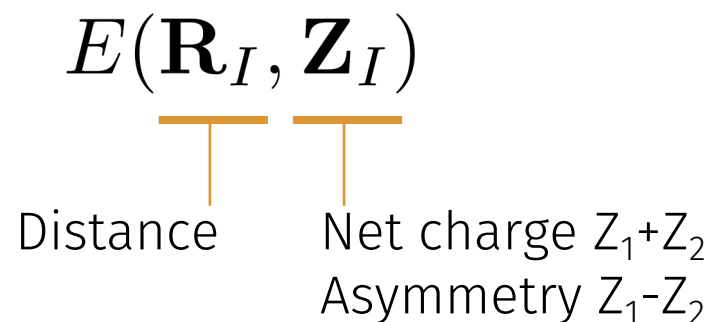
$$f(x, y) = x + y$$

Example: Dimers

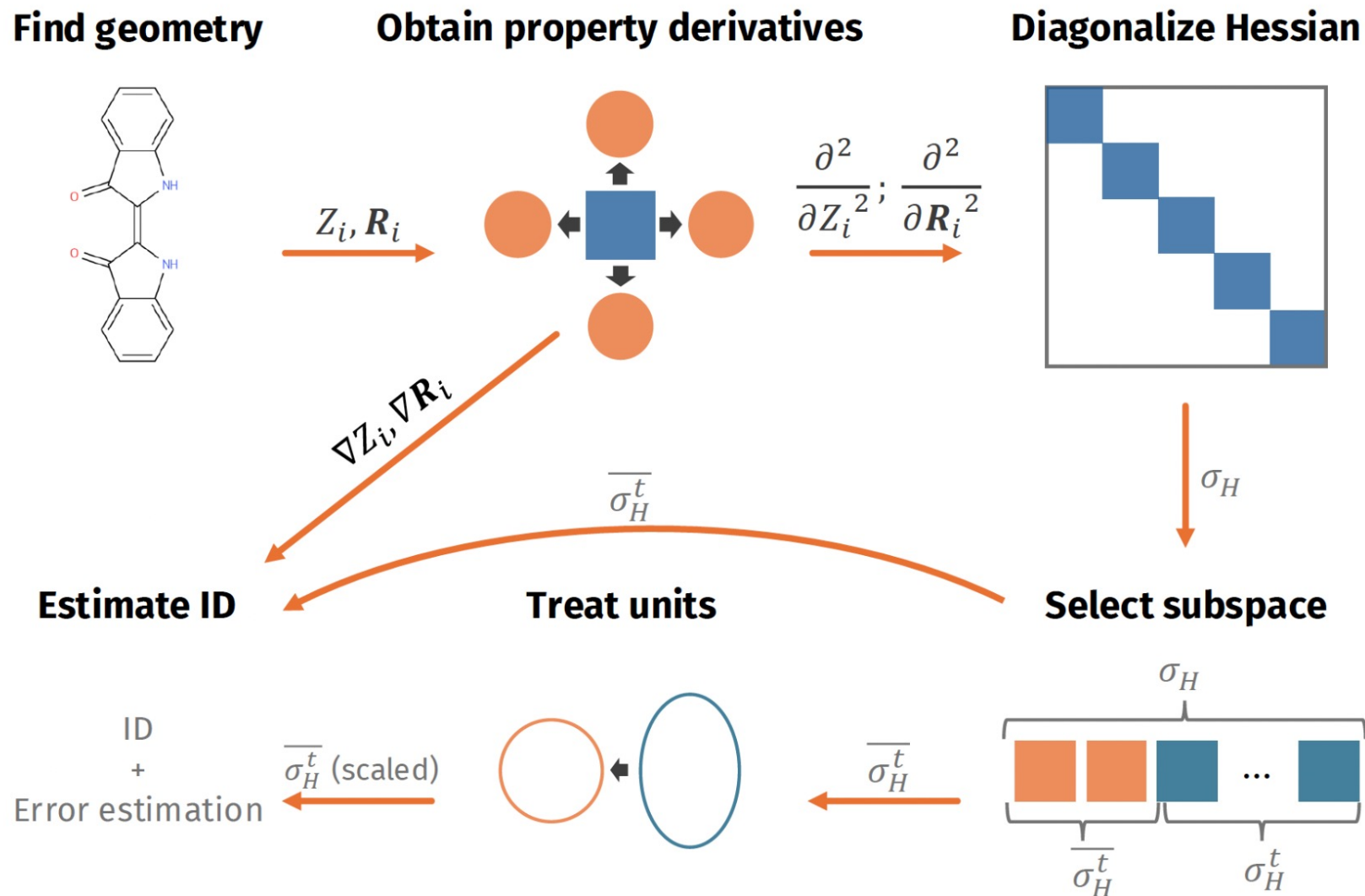
- Energy: 3 dimensions
- Net charge: 1 dimension

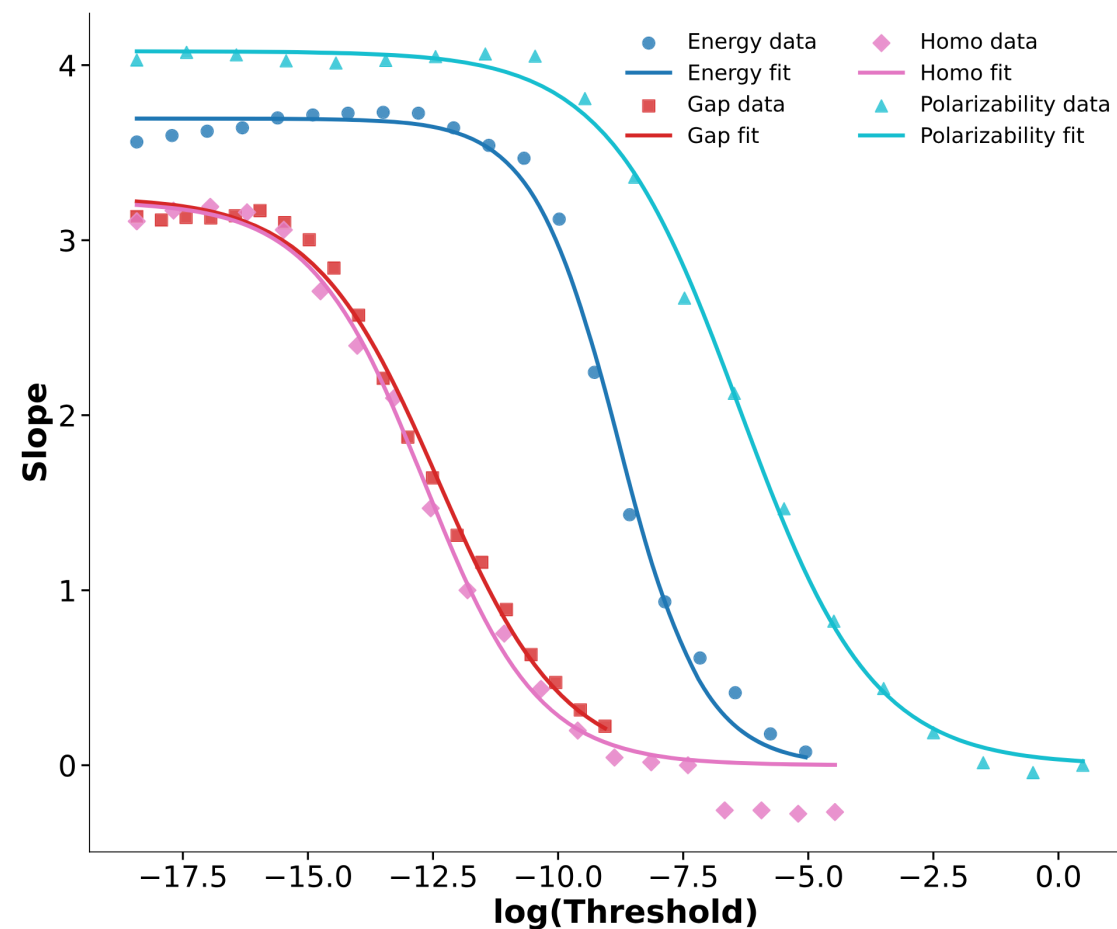
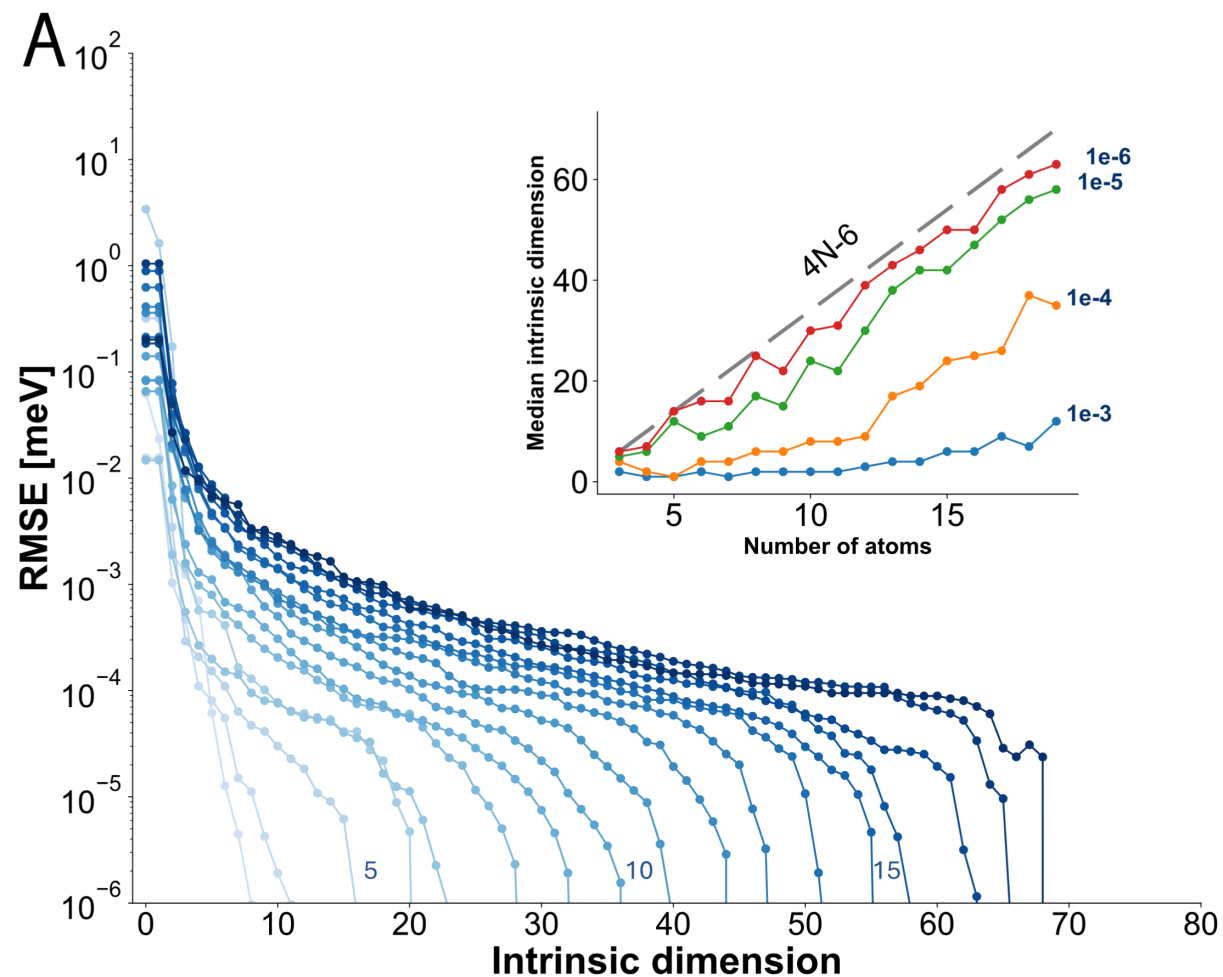
Example: Atom

- 1 dimension
- 1 dimension

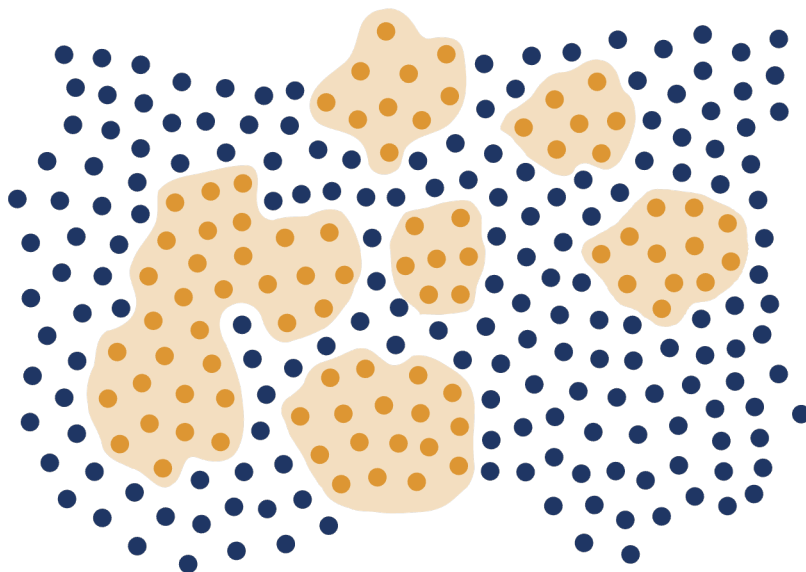


Ali Banjafar





Enumeration is impossible:
what's in this chemical space?



Allow for data-driven fundamental statements

“Most molecules do X”, “High X means low Y”

Transferability

More reliable understanding of trends

Lower data bias

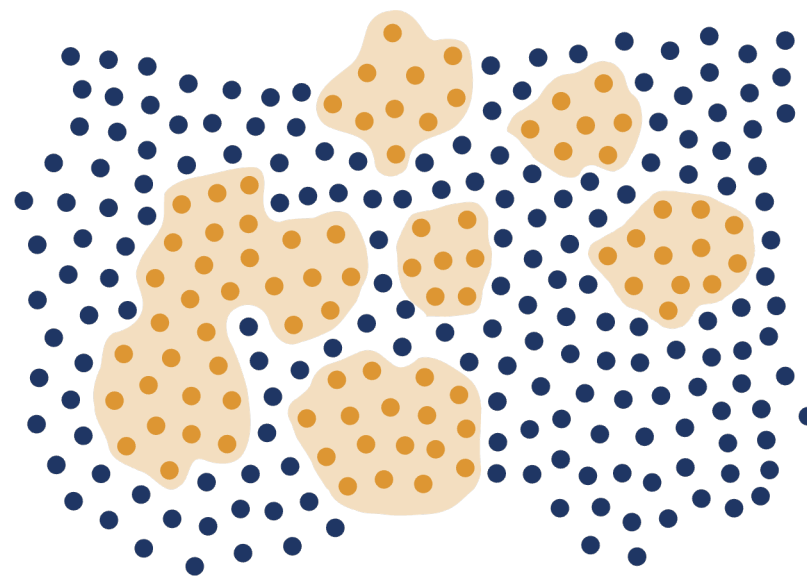
More realistic generalisation error

Formal statements

Often require uniform sampling

Problems

- Total number unknown
- Distribution unknown

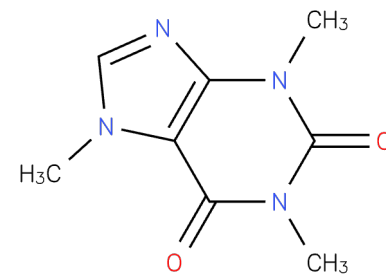
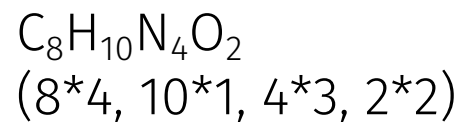


Goal

Sample all molecules (with given constraints) with known probabilities.

Sampling

- Choose **weighted** random sum formula
- Choose **weighted** random degree sequence
- Choose **weighted** random molecular graph



Requirements

- Find all sum formulas and degree sequences
- Sample loop-free multigraphs with given degree sequences uniformly
- Find **weights**

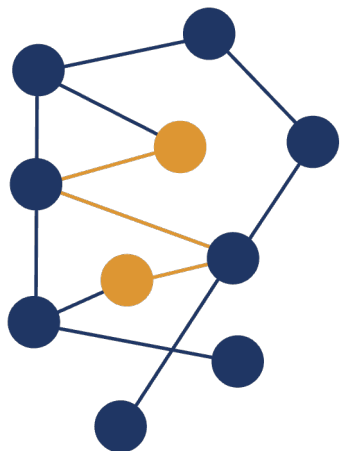
Solved
Solved, Seconds^[1]

Counting without enumeration

27

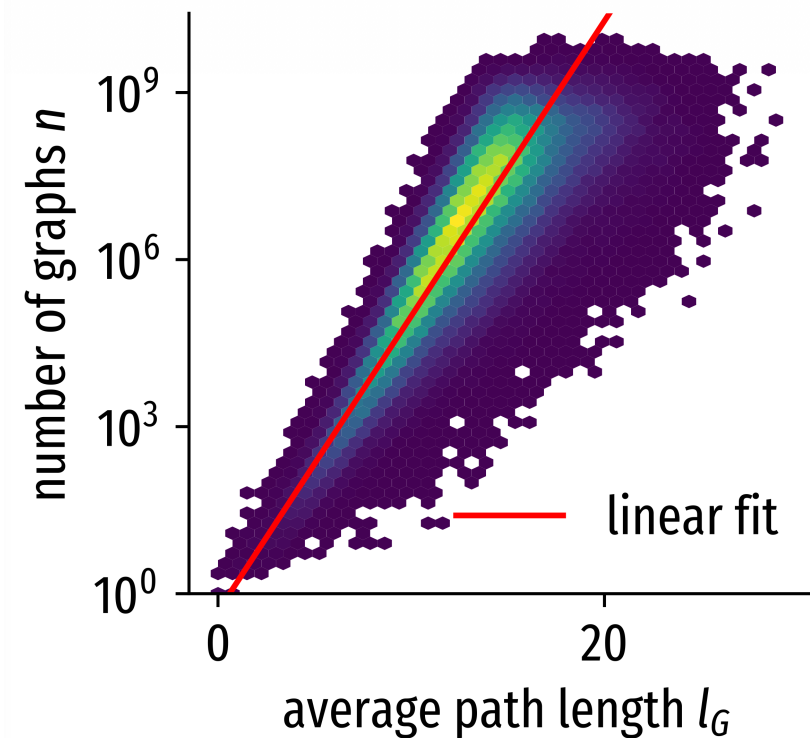
Average Path Length l_G

Sample from random molecule pairs



$$l_G \sim \log n$$

- Molecule
- Identical up to one switch
- Minimum path



285,656 stoichiometries
26,076,359,902,577 molecular graphs

Try it: random-molecule.org

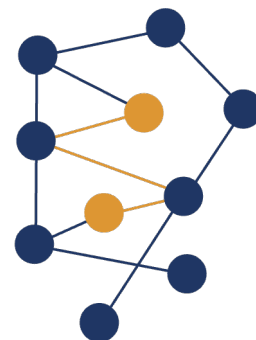
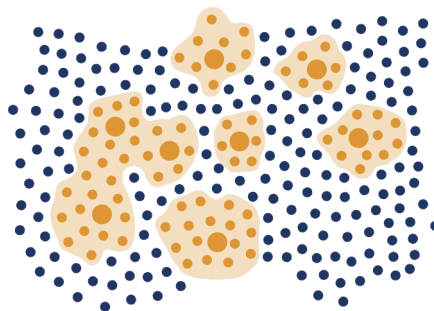
▼ Code to run yourself

```
# pip install --upgrade nablachem
import nablachem.space as ncs
counter = ncs.ApproximateCounter(show_progress=False)
space = ncs.SearchSpace("N:3 H:1 O:2 C:4")
criterion = ncs.Q("C = 8 & H = 10 & N = 4 & O = 2")
natoms = 24

total_molecule_count = counter.count(space, natoms, criterion)
mols = ncs.random_sample(
    counter, space, natoms=natoms, nmols=3, selection=criterion
)
```



I think there are 1,084,249,439,809 molecules with 24 atoms in the search space that satisfy the filter condition.



Thanks

Anton Artemyev

Ali Banjafar

Marco Bragato

Philipp Demekhin

Giorgio Domenichini

Emily Eikey

Chasz Griego

Nicolas Grimblat

John Keith

Simon Krug

Boris Lagutin

Anatole von Lilienfeld

Alex Maldonado

Diego Monterrubio-Chanca

Michael Sahre

- Symmetries** | Reducing (“folding”) search space
- Physics-driven** | Predictive power
- Closed-Form** | Explainable and shows structure
- Sampling** | Can assess whole regions at once