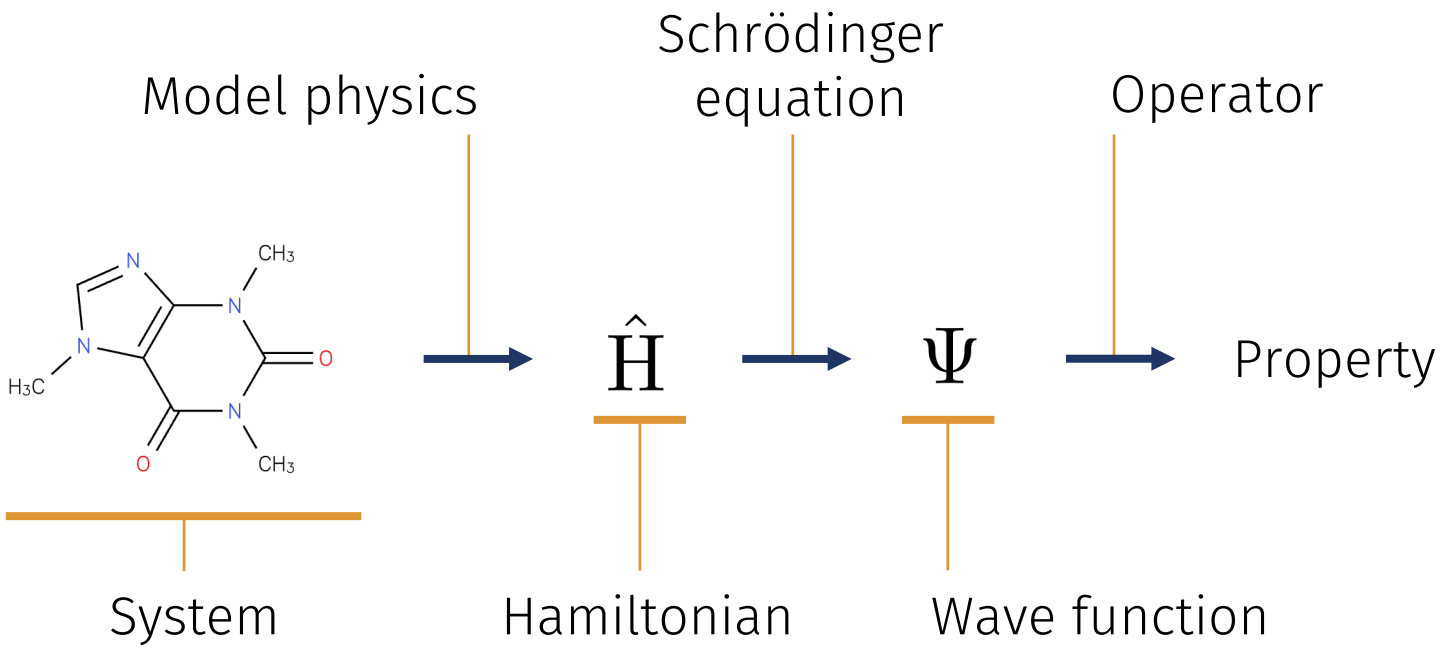
Alchemical Perturbations and Machine Learning

Guido von Rudorff, University of Kassel

vonrudorff@uni-kassel.de

nablachem.org/talks

Solved?


Model physics
Schrödinger equation
Operator
$\hat{H}$
Ψ
Property
System
Hamiltonian
Wave function

Wave function

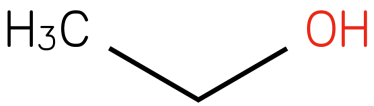
$$\Psi = \Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_n)$$

Methane

CH_4

N_2

Ethanol



Solved by approximations in computational chemistry?

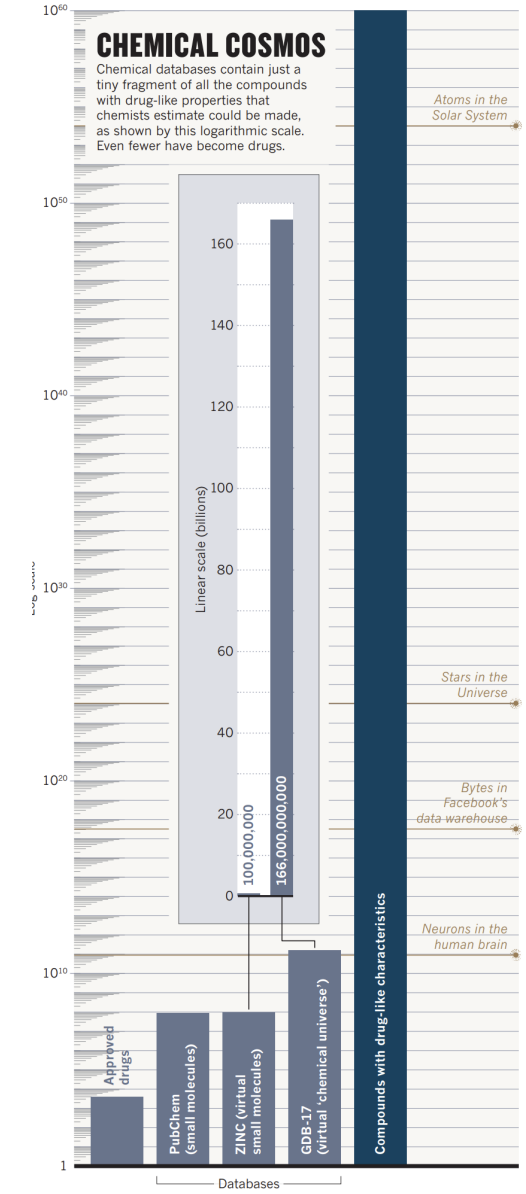
Scaling of Molecules

Commercial databases

- 164 million molecules
- 15k added daily

Scale

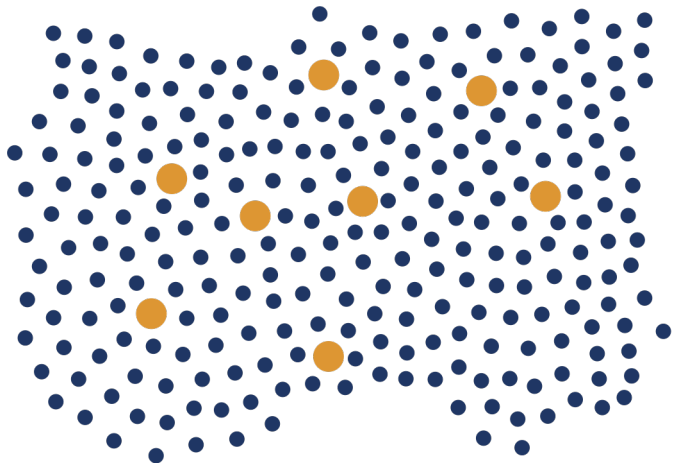
- One person: 1 million compounds/second
- 10 billion people on earth
- 10^{26} universe ages to go through

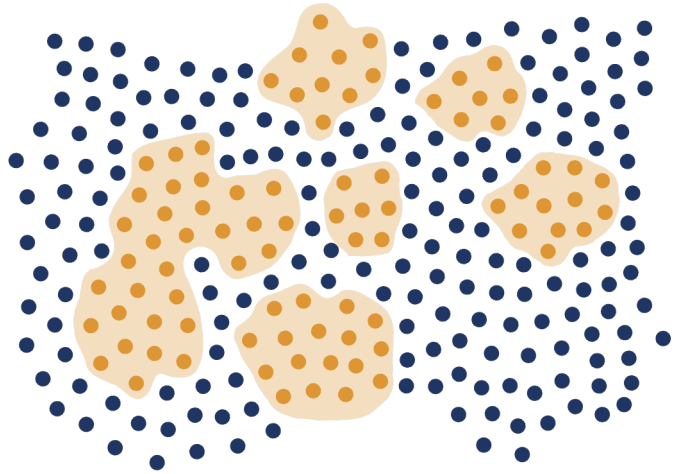


A. Mullard, *Nature*, 2017.

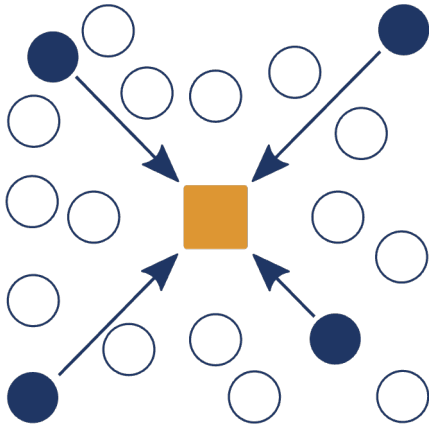
Properties of domains?

Speed does not matter:
even enumeration is impossible.

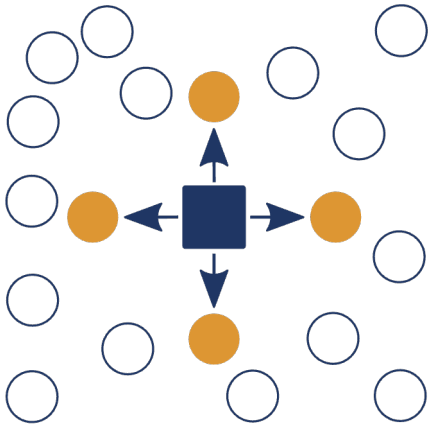




Machine Learning



Quantum Alchemy



Quantum Alchemy

Idea

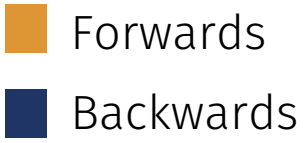
Treat system changes perturbatively[1,2]
Build a Taylor/Padé approximant[3] : **often 100.000 times faster**

Steps

Choose system

Alter system, calculate property response functions

Predict many modified systems



1 | E. B. Wilson, *J. Chem. Phys.* 1962. 2 | GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, 2020.
3 | GFvR, *J. Chem. Phys.*, 2021.

Perspective shift

Few highly accurate calculations
instead of many intermediate ones

$$\hat{\mathrm{H}} = \hat{\mathrm{H}}(Z_i, \mathbf{R}_i, N_e, \sigma)$$

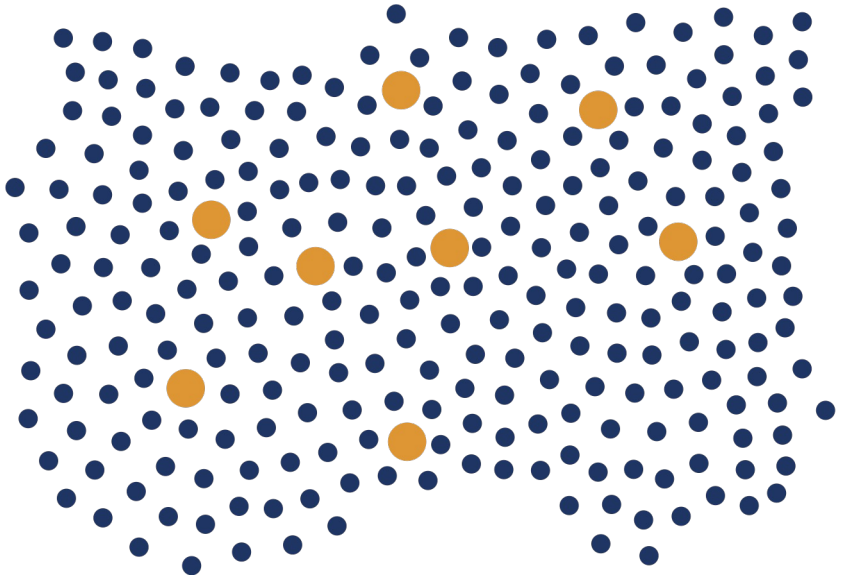
4N

1D, close to $\sum_i Z_i$

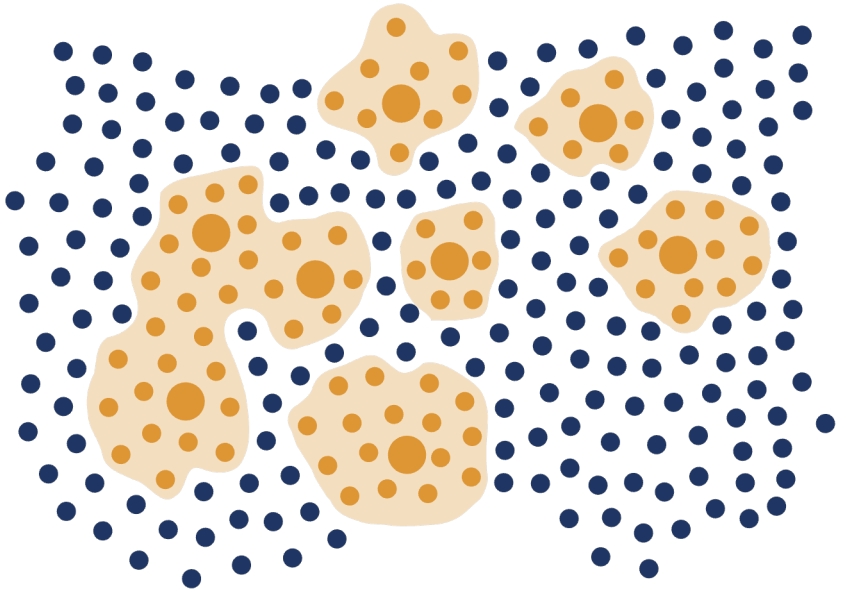
1D

Motivation

Without Perturbation



With Perturbation



Systems
- Any
- Known
- Approximated

Requirements

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]

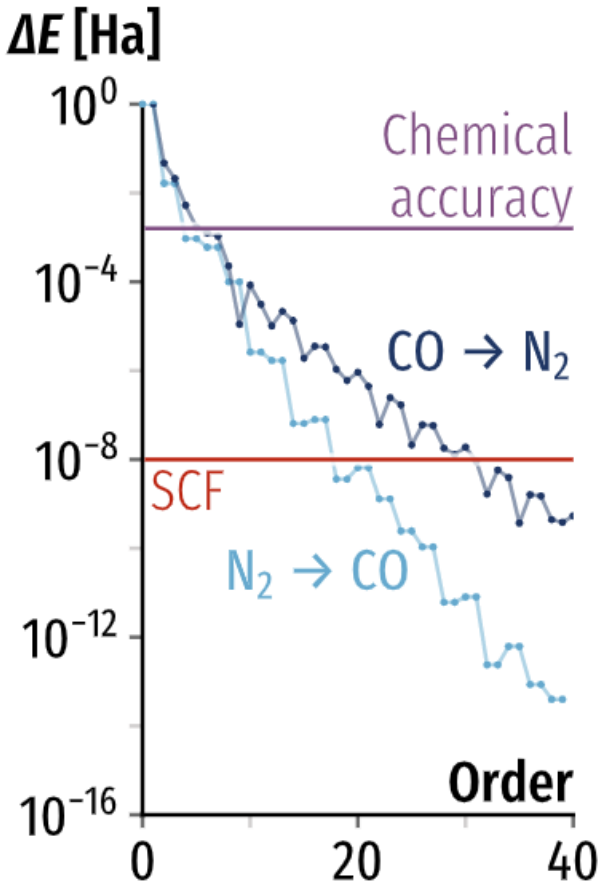
In progress

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.

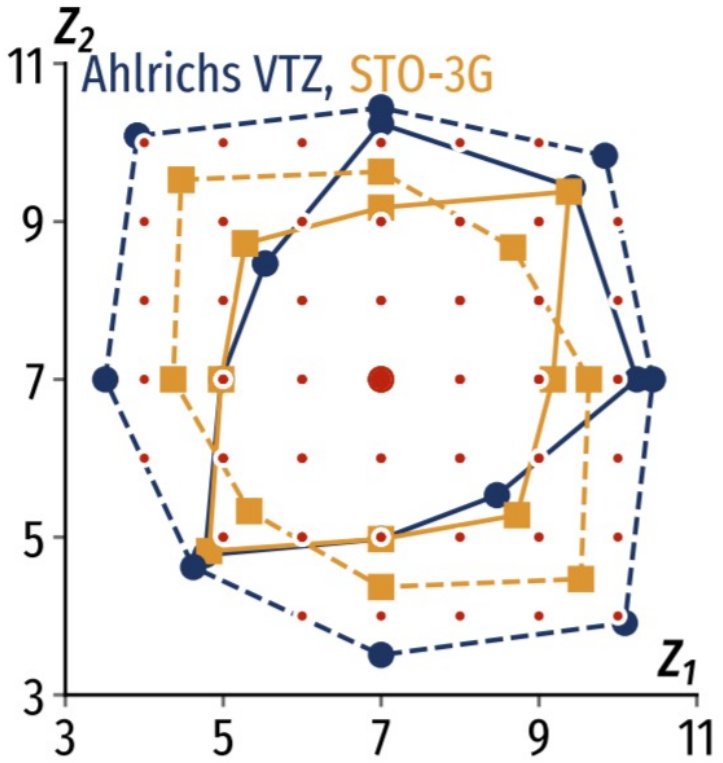
Convergence



HF/STO-3G

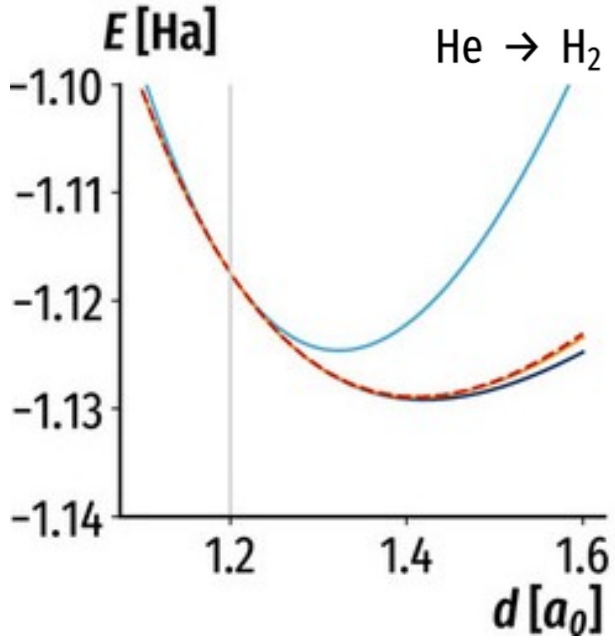
Taylor expansion

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



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

Geometry Relaxation



Taylor expansion

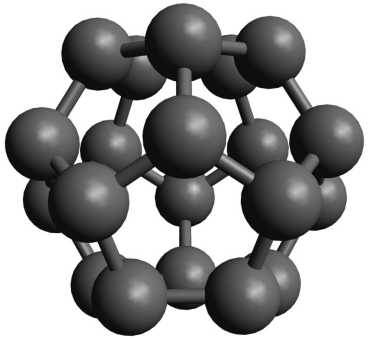
- Large changes still converge (more slowly)
- Geometric response can be recovered

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

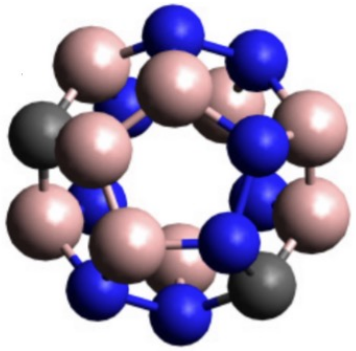
Covalent Interactions

Scaling with chemical space

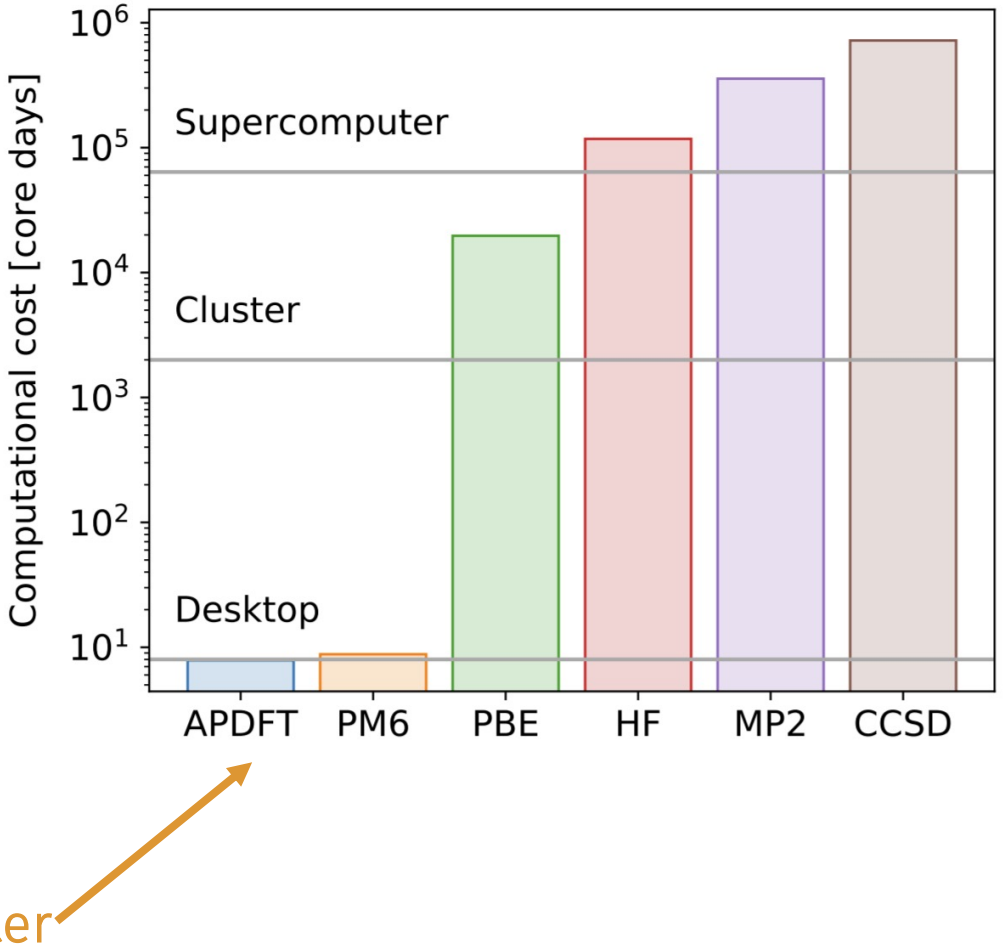
- 1 derivative for second order
- 5 derivatives for third order



C_{20}



$3.1 \cdot 10^6$ targets



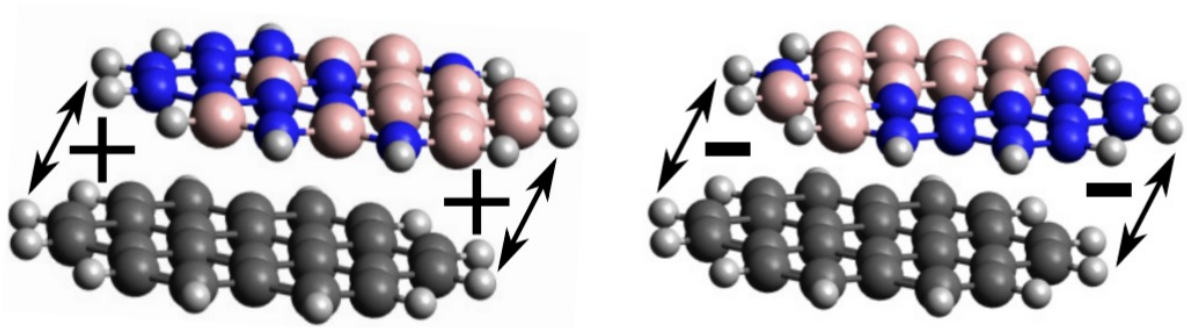
QA: 80.000x faster

GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, **2020**.

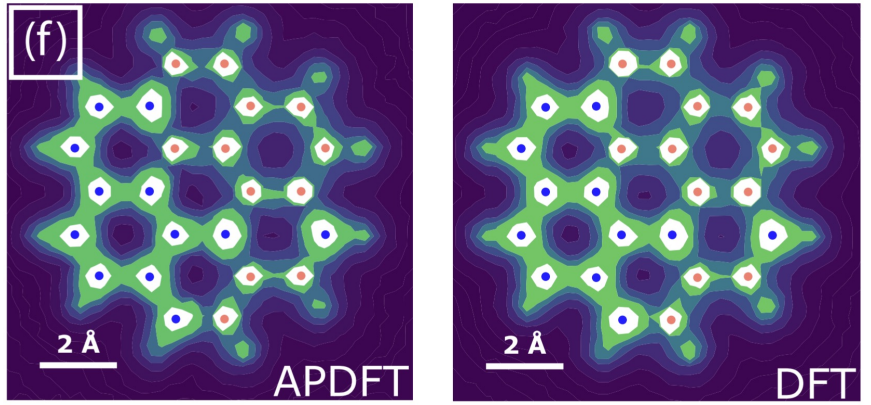
Non-covalent interactions

BN-doped coronene dimer

- Identify most/least attractive doping pattern
- Design case



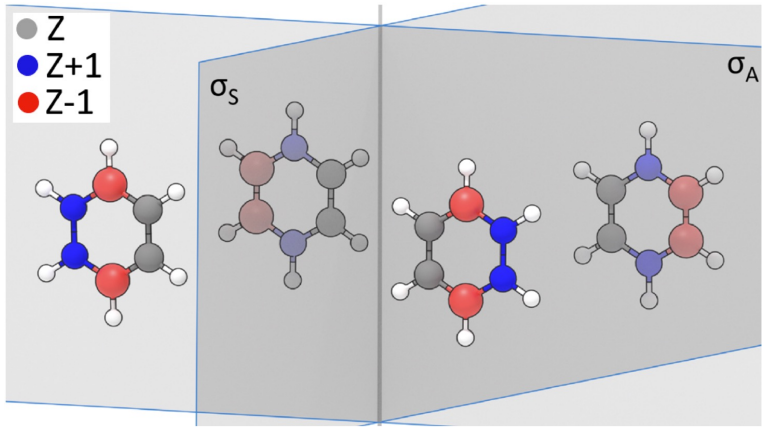
QA: 20.000x faster



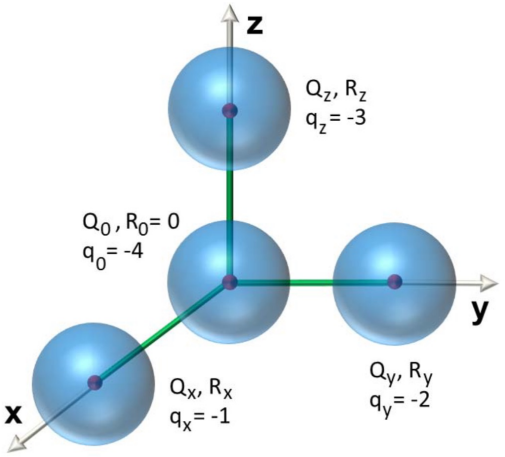
$2.8 \cdot 10^{10}$ targets

GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, **2020**.

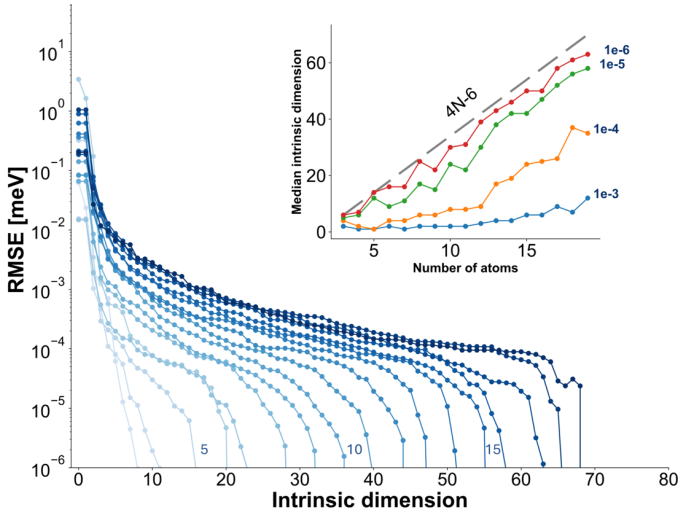
Ok, it works
What can we learn?



Symmetries

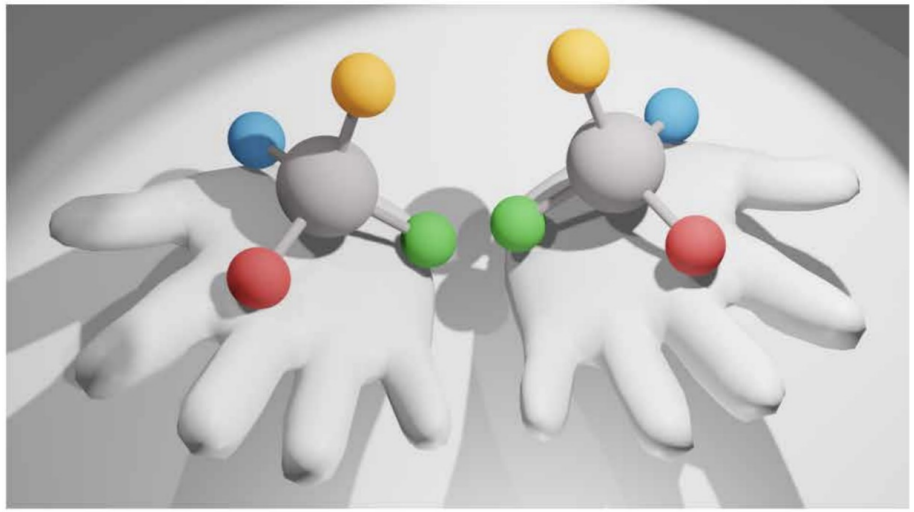


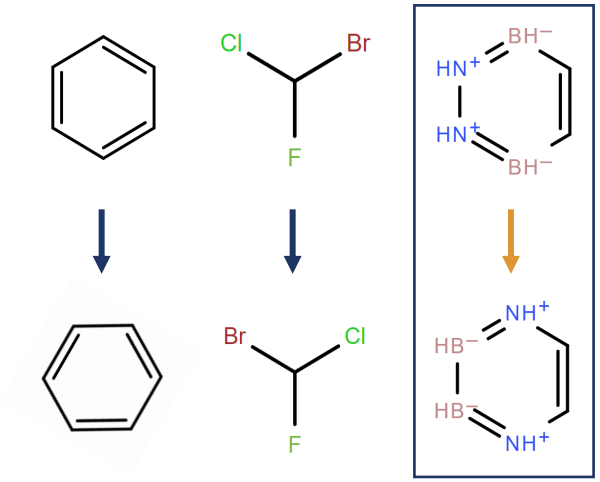
Interpretability

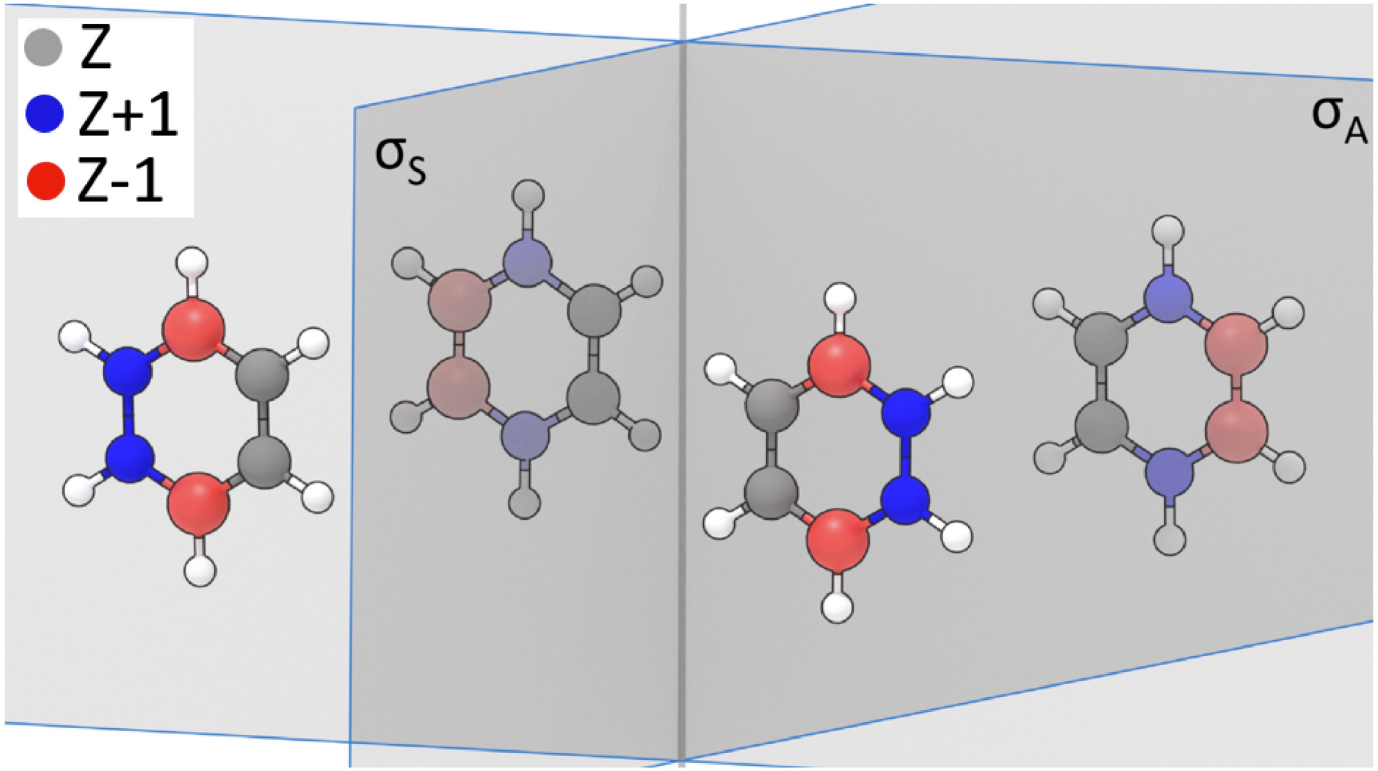


Intrinsic Dimension

Alchemical Enantiomers





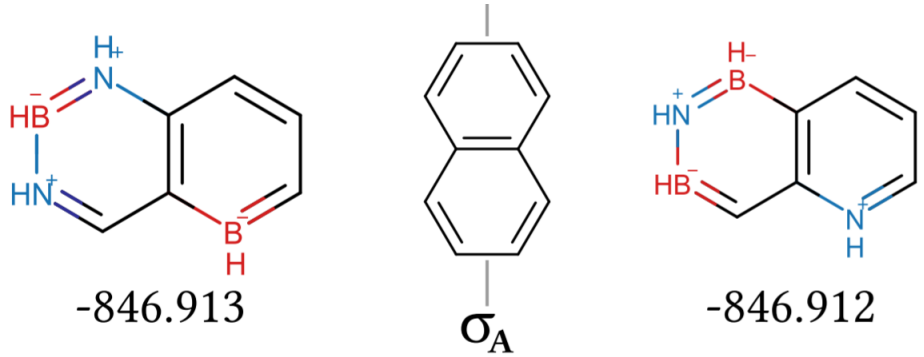


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

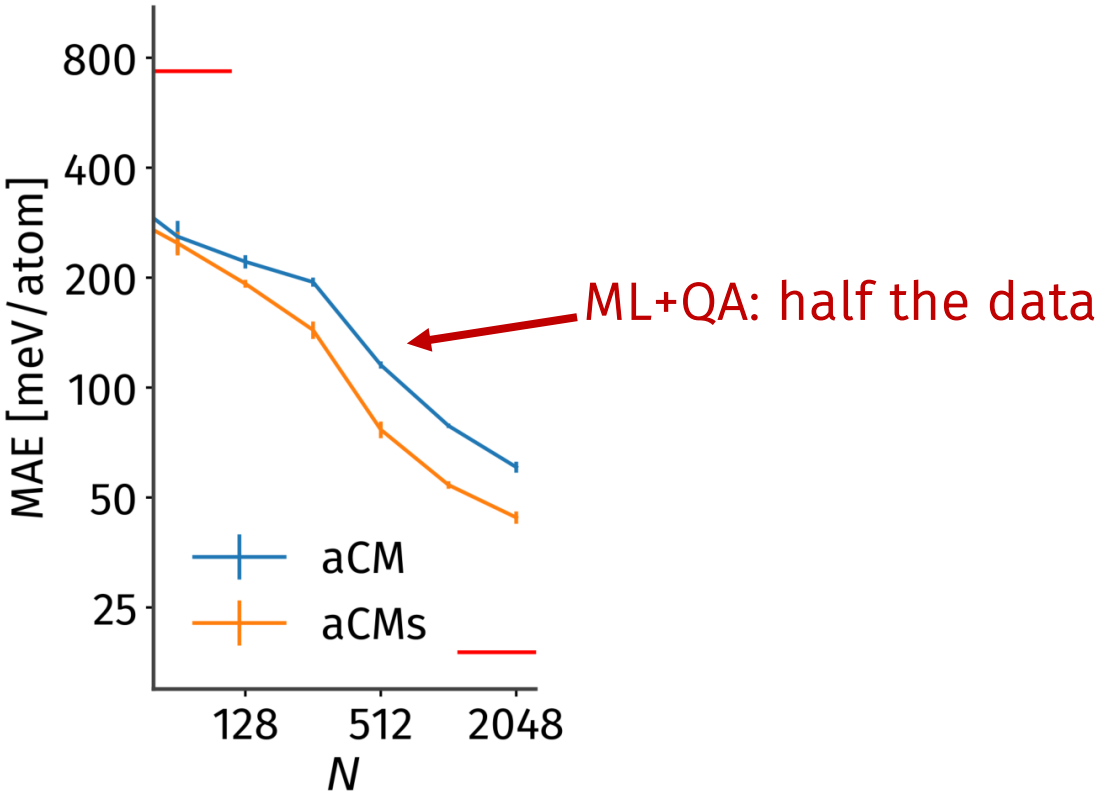
Alchemical Enantiomers

Fundamentally new symmetry

Electronic energy only

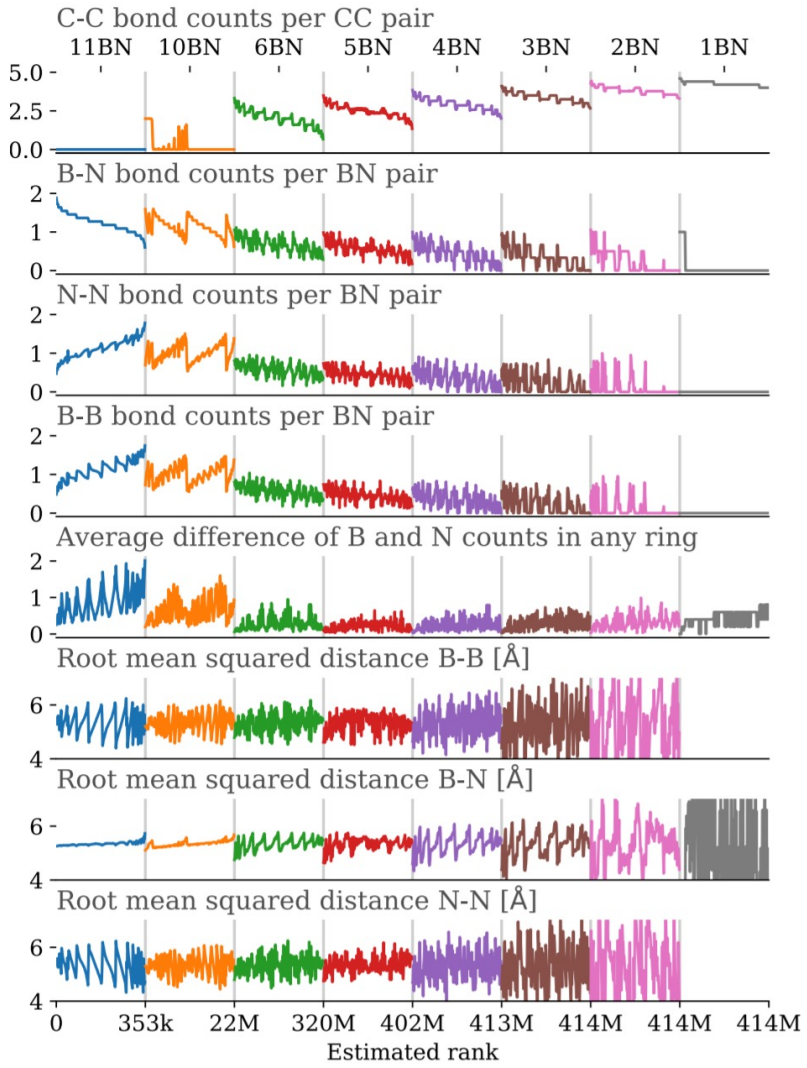


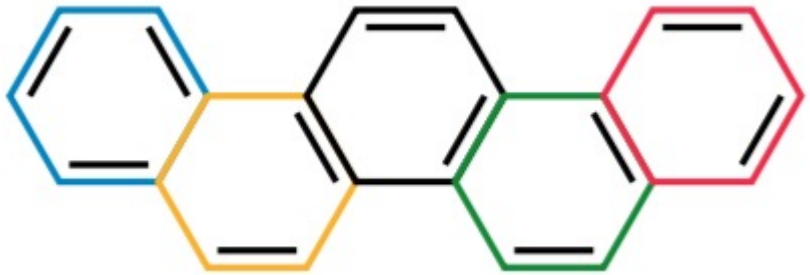
Speed up machine learning



GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

Alchemical Enantiomers





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

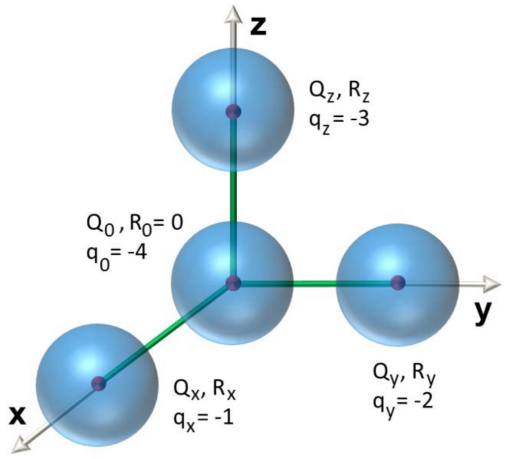
Not a single QM calculation required!

GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

Interpretability: PECD

Angular emission

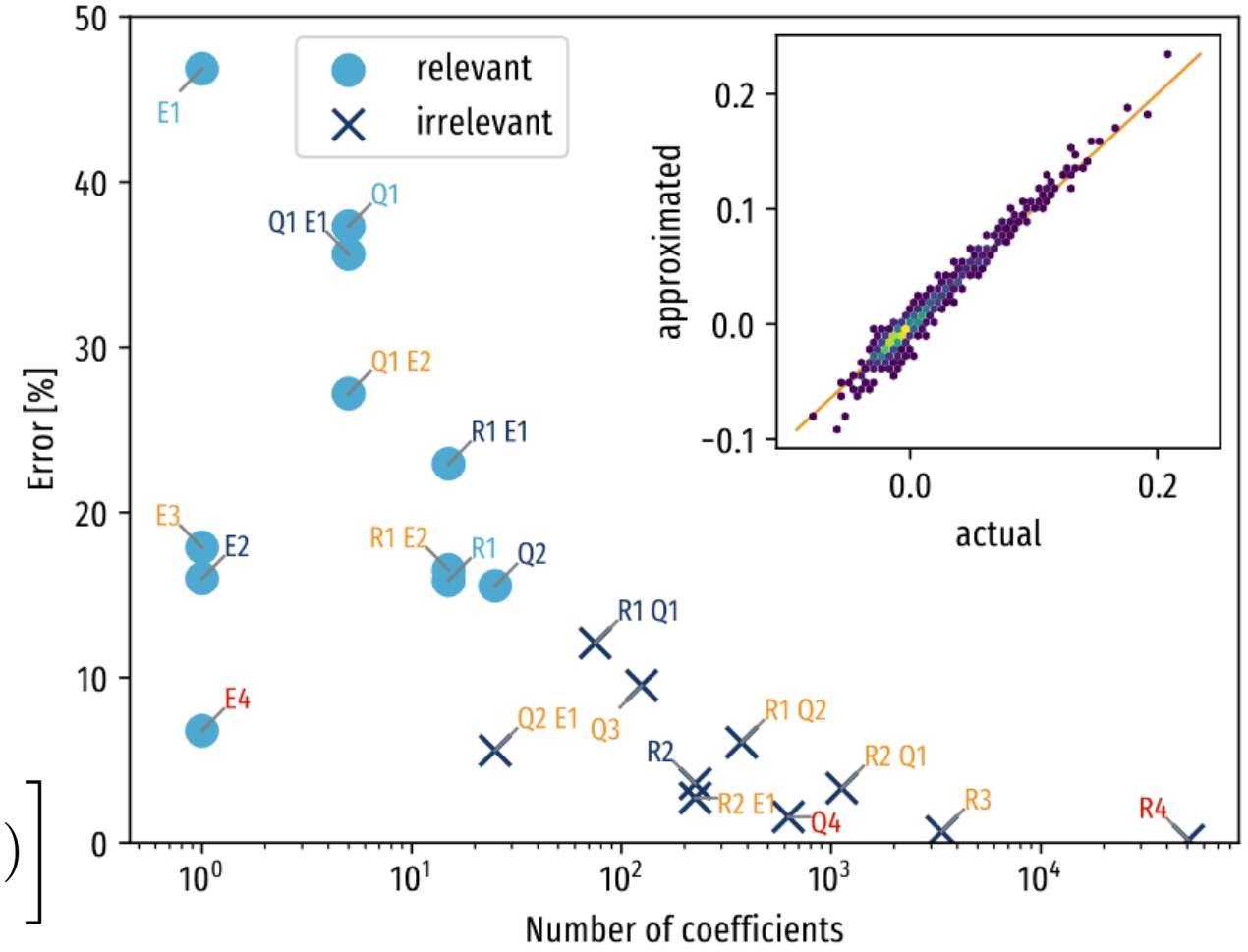
- Expensive to calculate
- Highly coupled degrees of freedom: multidimensional expansion



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

dichroic parameter

anisotropy parameter



GFvR, A Artemyev, B Lagutin, P Demekhin, *J. Chem. Phys.*, 2024.

Quantifying Dimensions

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

$$E(\mathbf{R}_I, \mathbf{Z}_I)$$

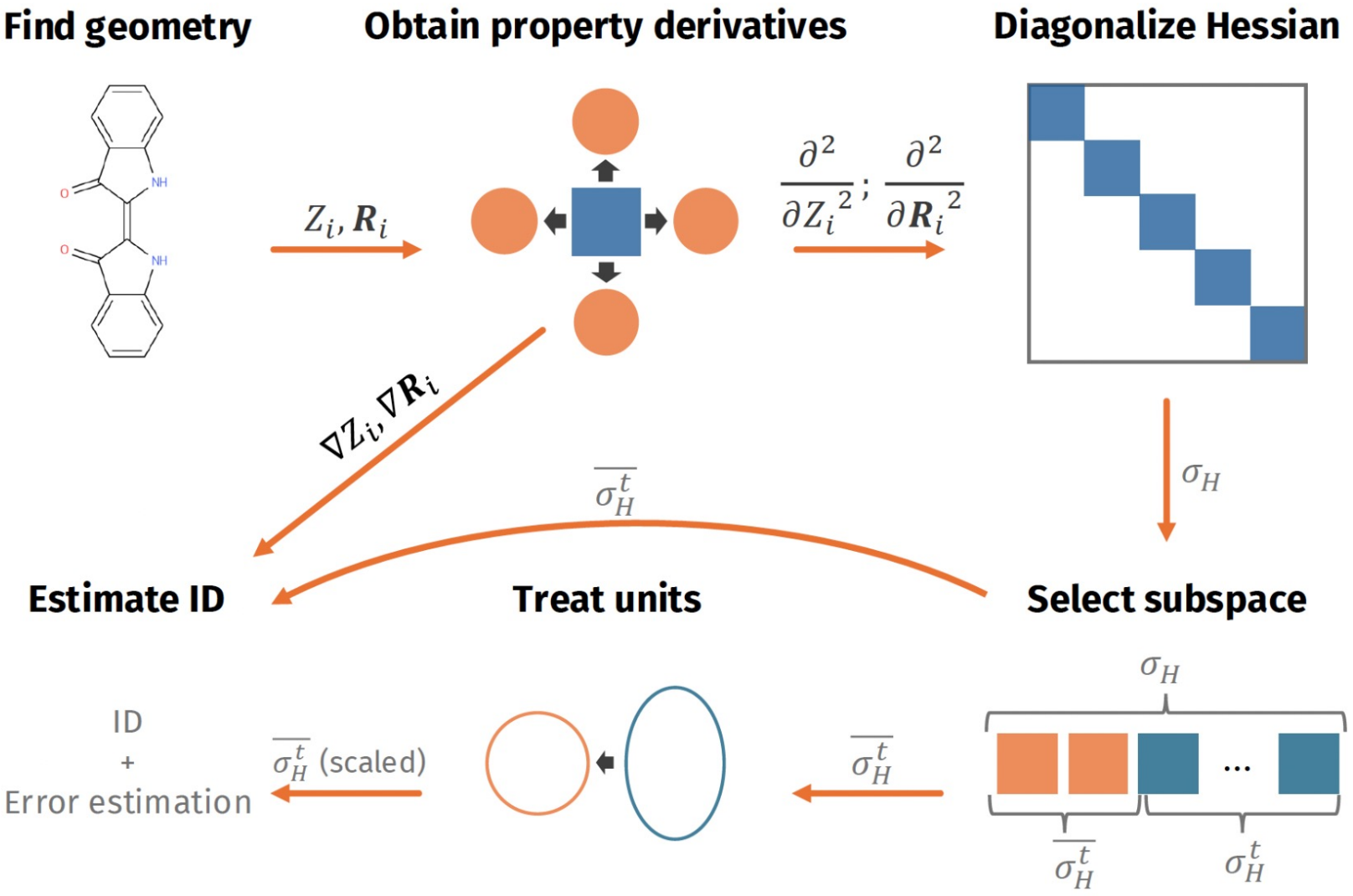
Distance

Net charge Z_1+Z_2

Asymmetry Z_1-Z_2

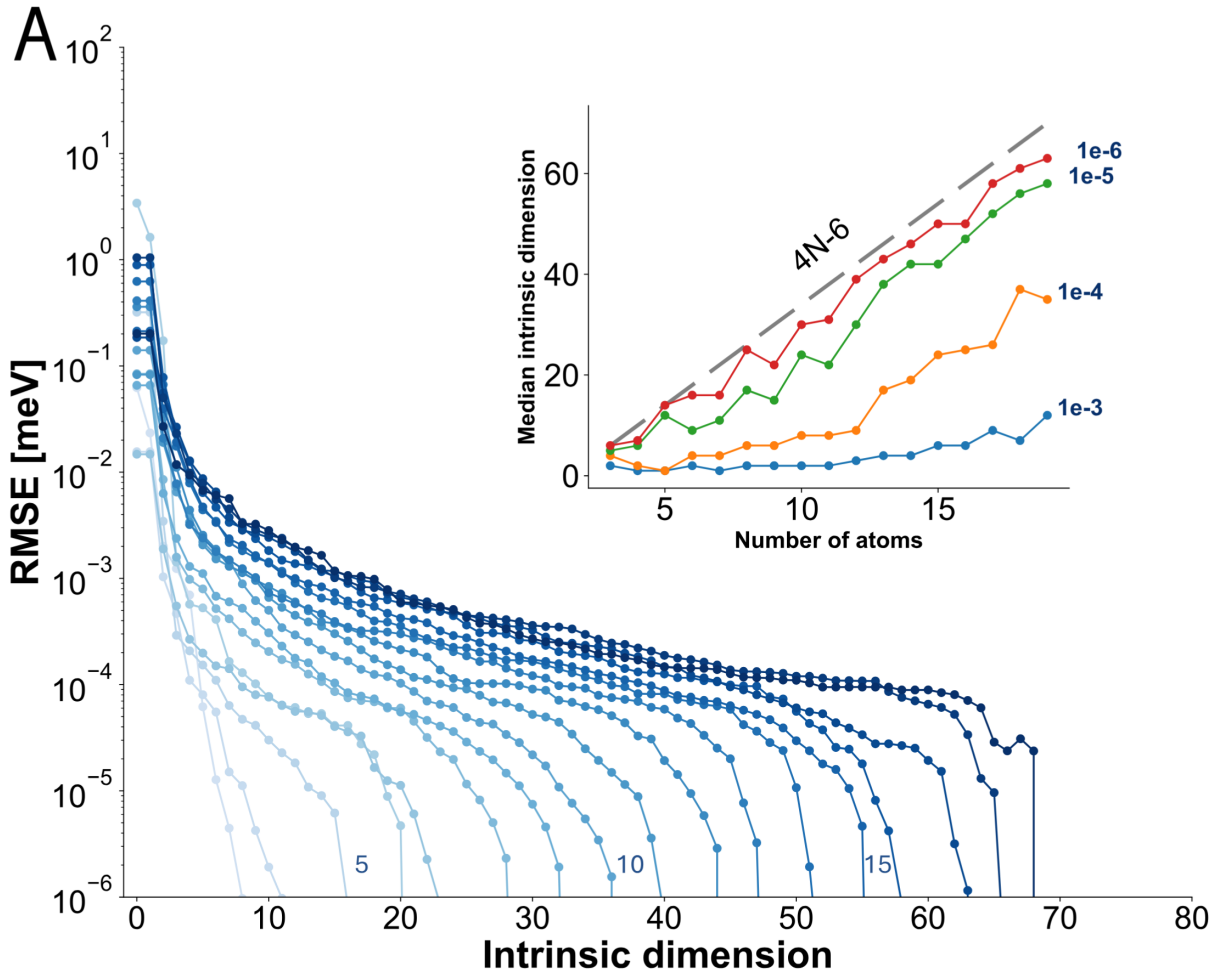
Ali Banjafar

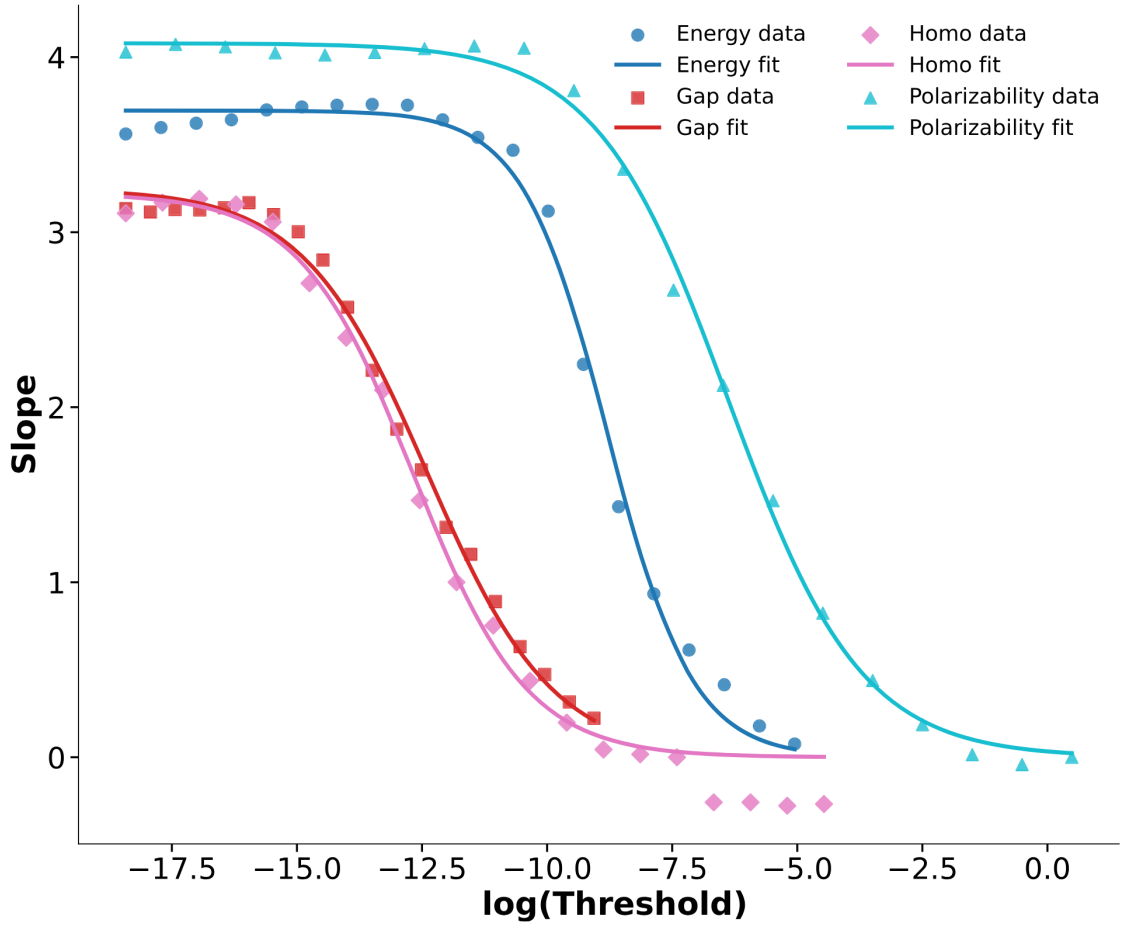
Quantifying Dimensions



A Banjafar, **GFvR**, *J Chem Phys, 2025.*

Quantifying Dimensions

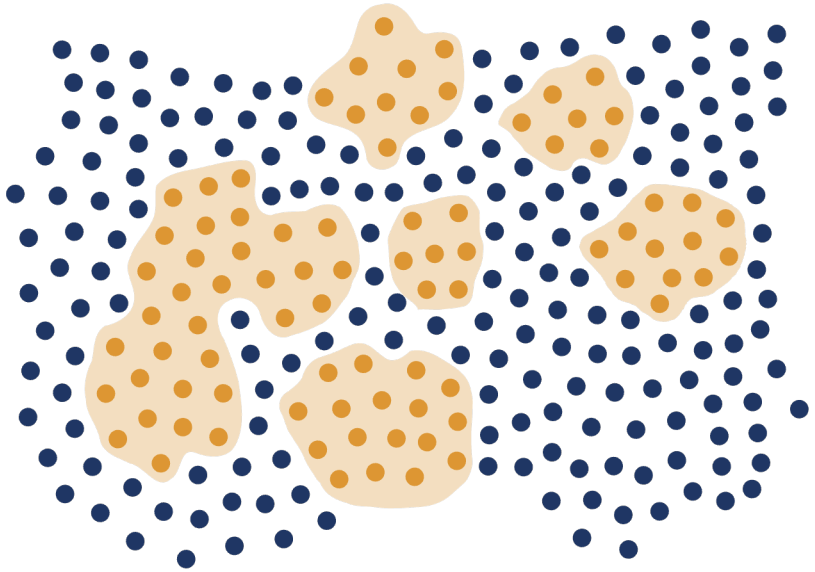




A Banjafar, **GFvR**, *J Chem Phys, 2025.*

Properties of domains?

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



Random Sampling

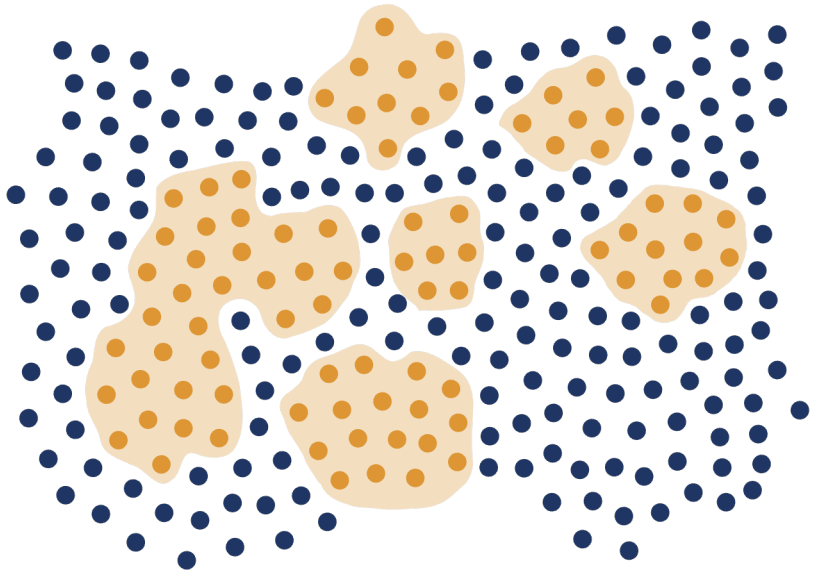
Allow for data-driven fundamental statements
“Most molecules do X”, “High X means low Y”

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



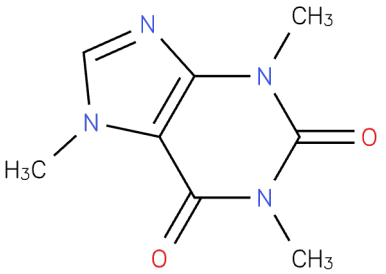
Random Sampling

Goal

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

Sampling

- Choose weighted random sum formula — $C_8H_{10}N_4O_2$
- Choose weighted random degree sequence — (8*4, 10*1, 4*3, 2*2)
- Choose weighted random molecular graph



Requirements

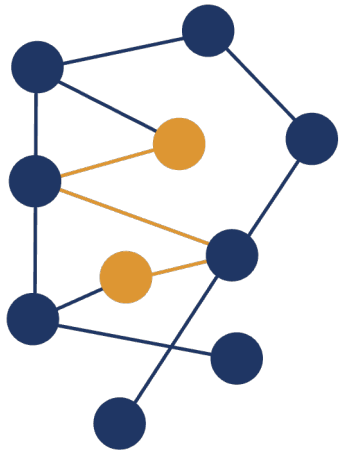
- Find all sum formulas and degree sequences — Solved
- Sample loop-free multigraphs with given degree sequences uniformly — Solved, Seconds[1]
- Find weights

1 | C Greenhill, M Sfragara, *Theo Comput Sci*, **2018**.

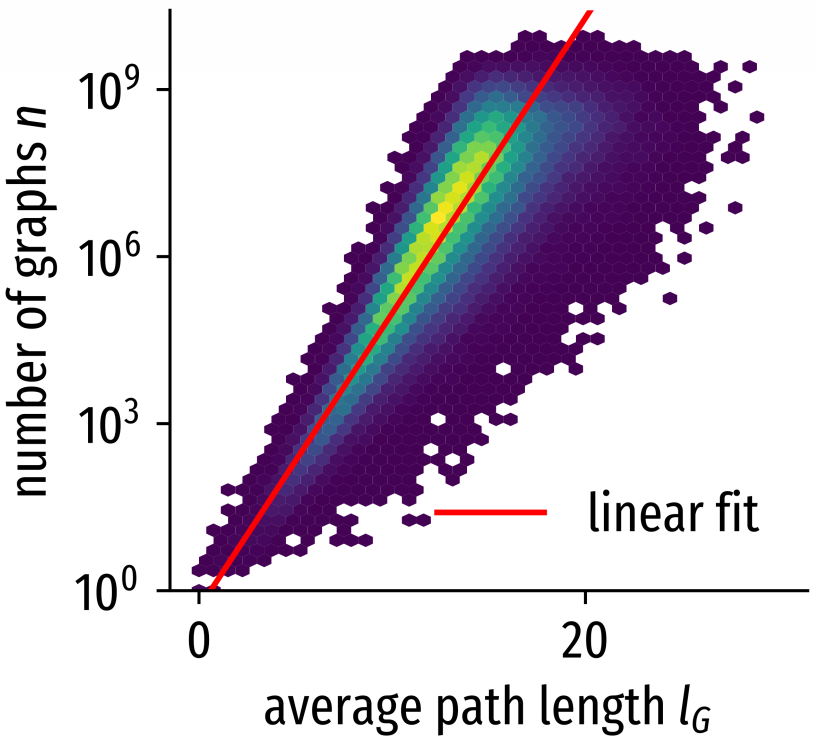
Counting without enumeration

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

Counting without enumeration

Try it: random-molecule.org

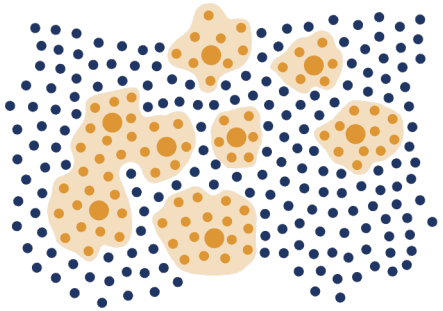
Code to run yourself

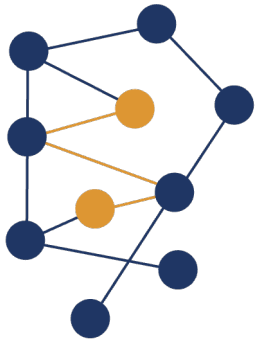
```python
# pip install --upgrade nablachem
import nablachem.space as ncs
counter = ncs.ApproximateCounter(show_progress=False)
space = ncs.SearchSpace("N:3 H:1 O:2 C:4")
criterion = ncs.Q("C = 8 & H = 10 & N = 4 & O = 2")
natoms = 24

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.

Summary



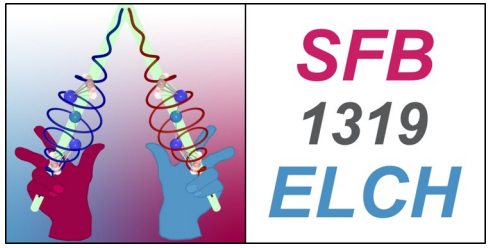


Symmetries | Reducing (“folding”) search space

Physics-driven | Predictive power

Closed-Form | Explainable and shows structure

Sampling | Can assess whole regions at once



We are hiring!

vonrudorff@uni-kassel.de

nablachem.org/talks

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

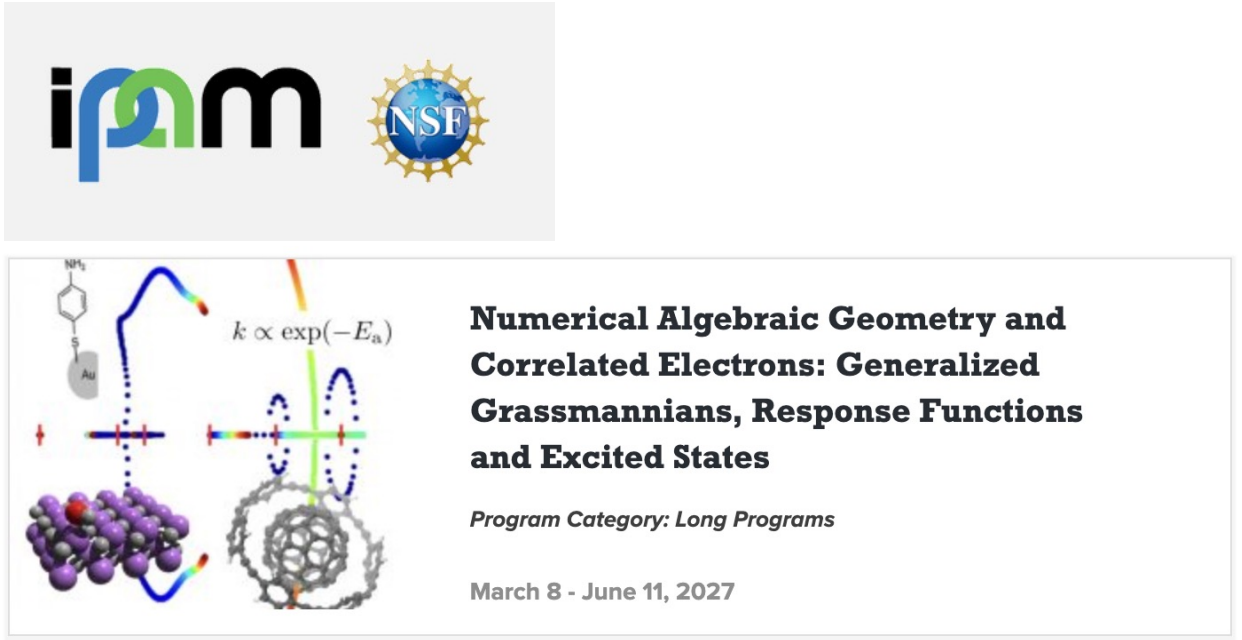
`pip install nablachem`
`random-molecule.org`

Chemical
Compound
Space
Conference

2026

Chemical Compound Space Conference 2026

• Munich • March, 10th to 13th • Germany •



Numerical Algebraic Geometry and Correlated Electrons: Generalized Grassmannians, Response Functions and Excited States

Program Category: Long Programs

March 8 - June 11, 2027