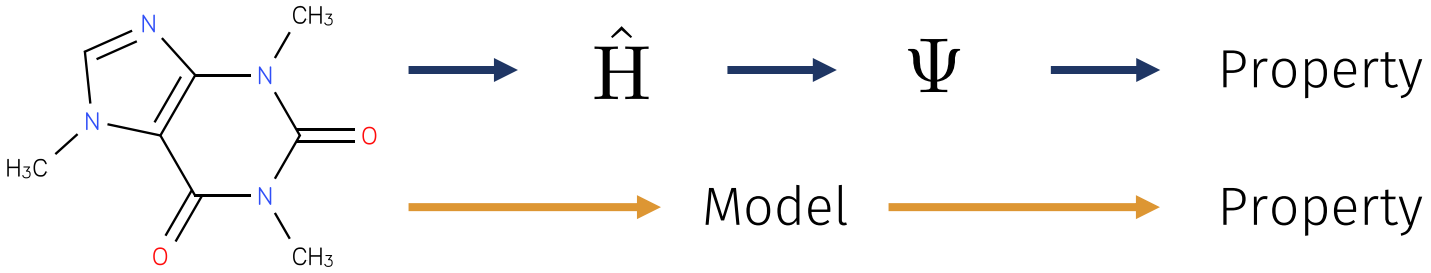
Symmetries and the Intrinsic Dimensionality of Chemical Space

Guido Falk von Rudorff, University of Kassel

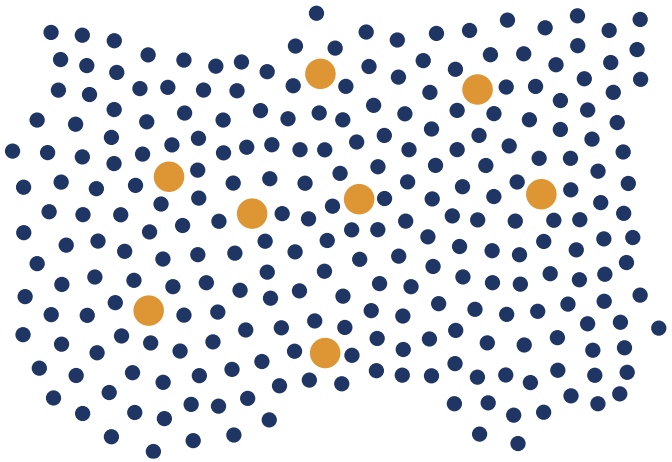
vonrudorff@uni-kassel.de nablachem.org/talks ferchault @ferchault

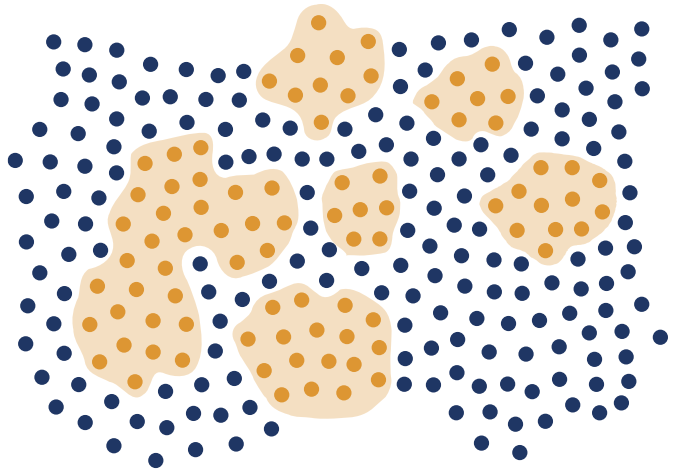
Properties of a single system



Properties of domains?

Speed does not matter:
even enumeration is impossible.





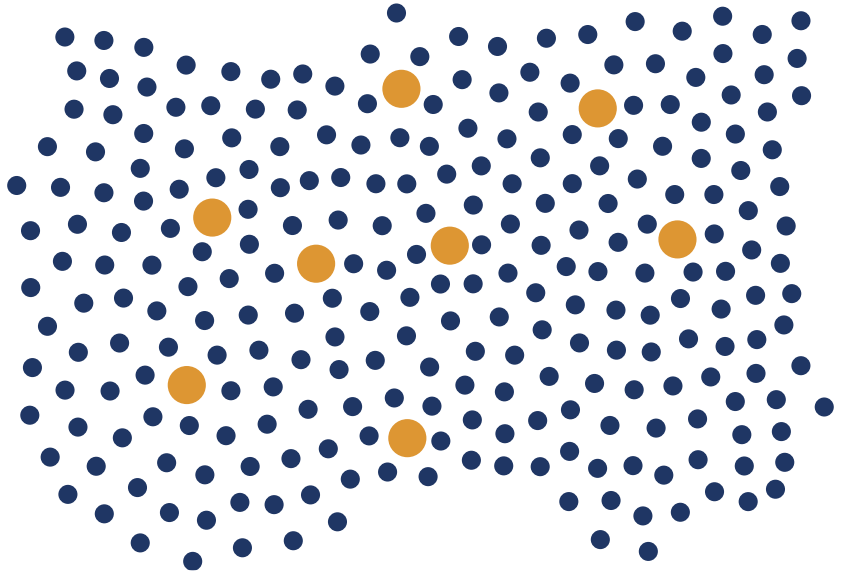
Domain definition

Chemical Space
All thermally accessible configurations for some constraints on Z_i.

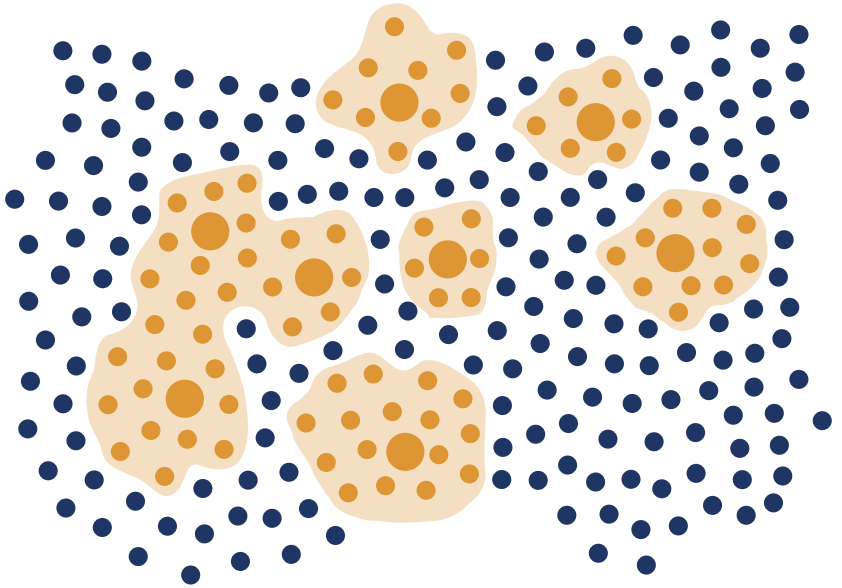
$$\hat{\mathrm{H}} = \hat{\mathrm{H}}(\underbrace{Z_i, \mathbf{R}_i}_{\text{4N}}, \underbrace{N_e}_{\text{1D, close to } \sum_i Z_i}, \underbrace{\sigma}_{\text{1D}})$$

Perturbations

Without Perturbation



With Perturbation

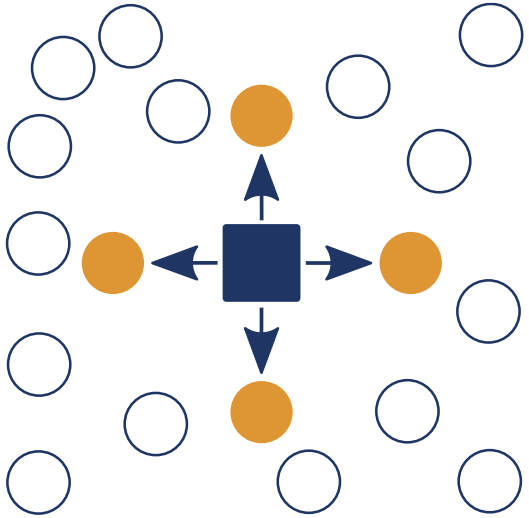


Systems/Molecules

- Any
- Known
- Approximated

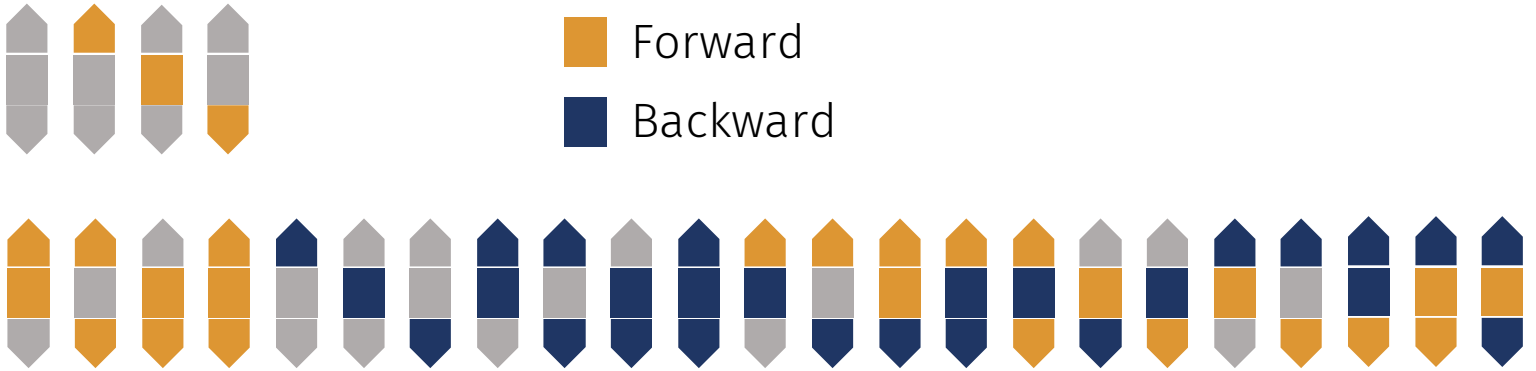
Quantum Alchemy

Quantum Alchemy



Taylor/Padé approximant

- Energy function of
  - Geometry — Forces, Vibrations
  - Nuclear charges — Alchemical changes
- Idea: obtain dominant leading derivatives, predict many systems



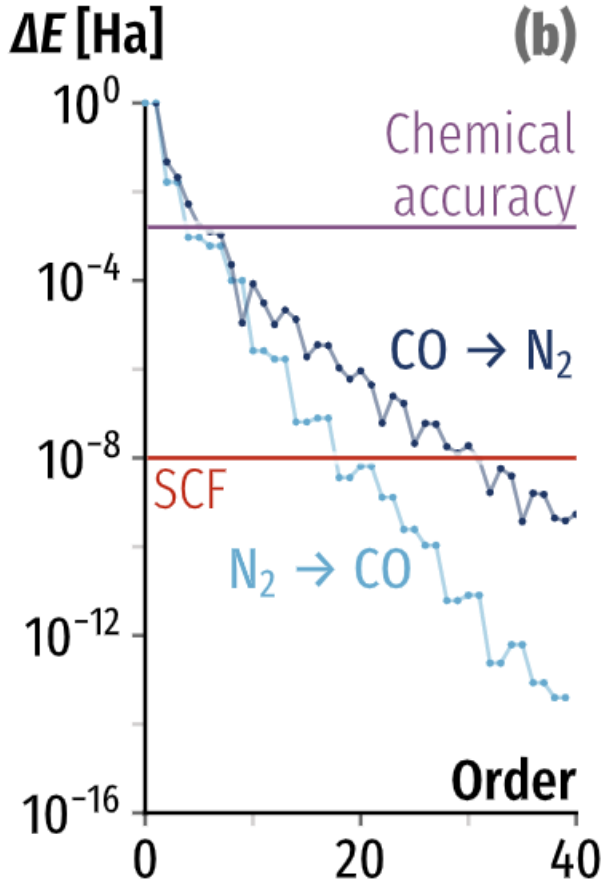
$$Q(\boldsymbol{x}) \simeq \sum_{|\alpha| \leq k} \frac{\partial^{|\alpha|} Q(\mathbf{a})}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \frac{(\mathbf{x}-\mathbf{a})^{\alpha}}{\alpha!}$$

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

Case Studies

Differentiable / Analytic / Converge Quickly

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



1 | GFvR, OA von Lilienfeld, *Phys. Rev. Res.*, 2020. 2 | GFvR, *J. Chem. Phys.*, 2021. 3 | GFvR, OA 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.*, accepted.

Physical Insights from Closed Form

Angular emission (Photoelectron circular dichroism)

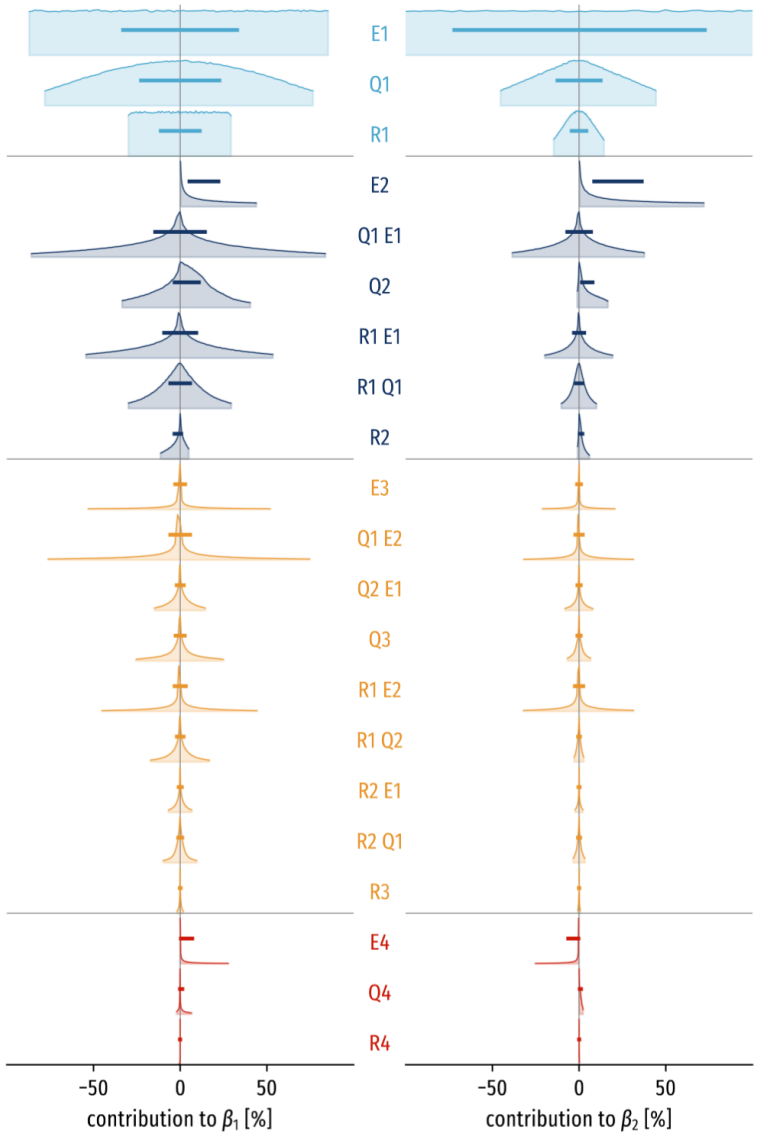
- Expensive to calculate
- Highly coupled degrees of freedom: multidimensional expansion

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

dichroic parameter

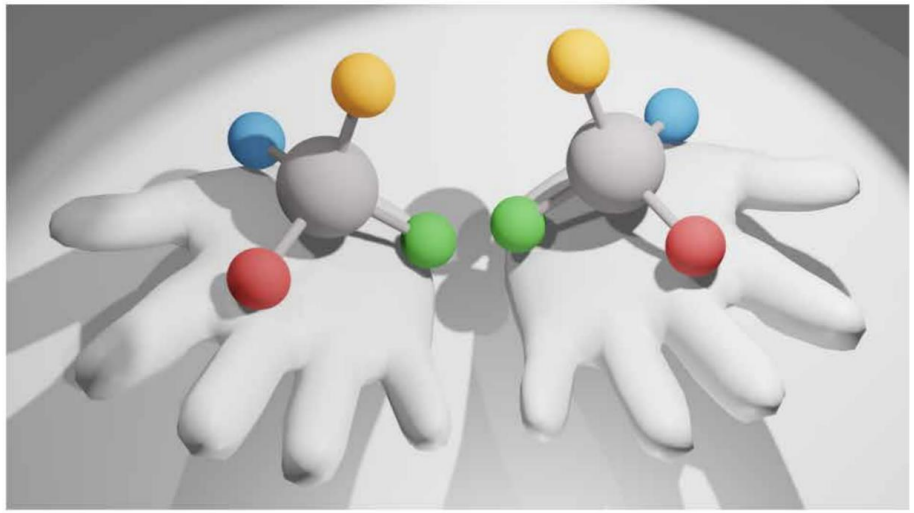
anisotropy parameter

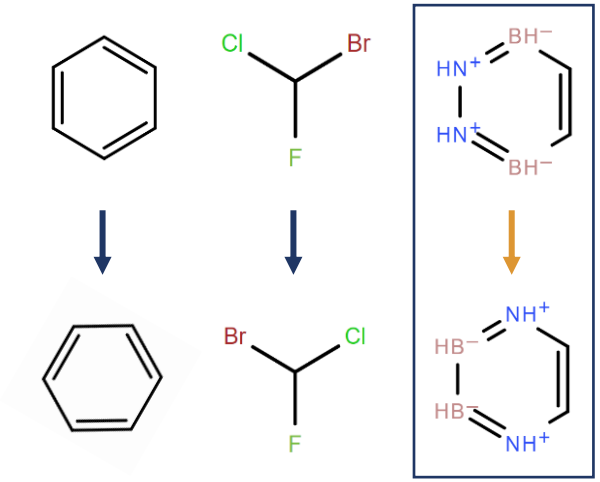
$$\frac{\partial \beta_i}{\partial R_x} \in \text{R1} \qquad \frac{\partial^2 \beta_i}{\partial Q_y^2} \in \text{Q2} \qquad \frac{\partial^2 \beta_i}{\partial R_x Q_y} \in \text{R1 Q1}$$

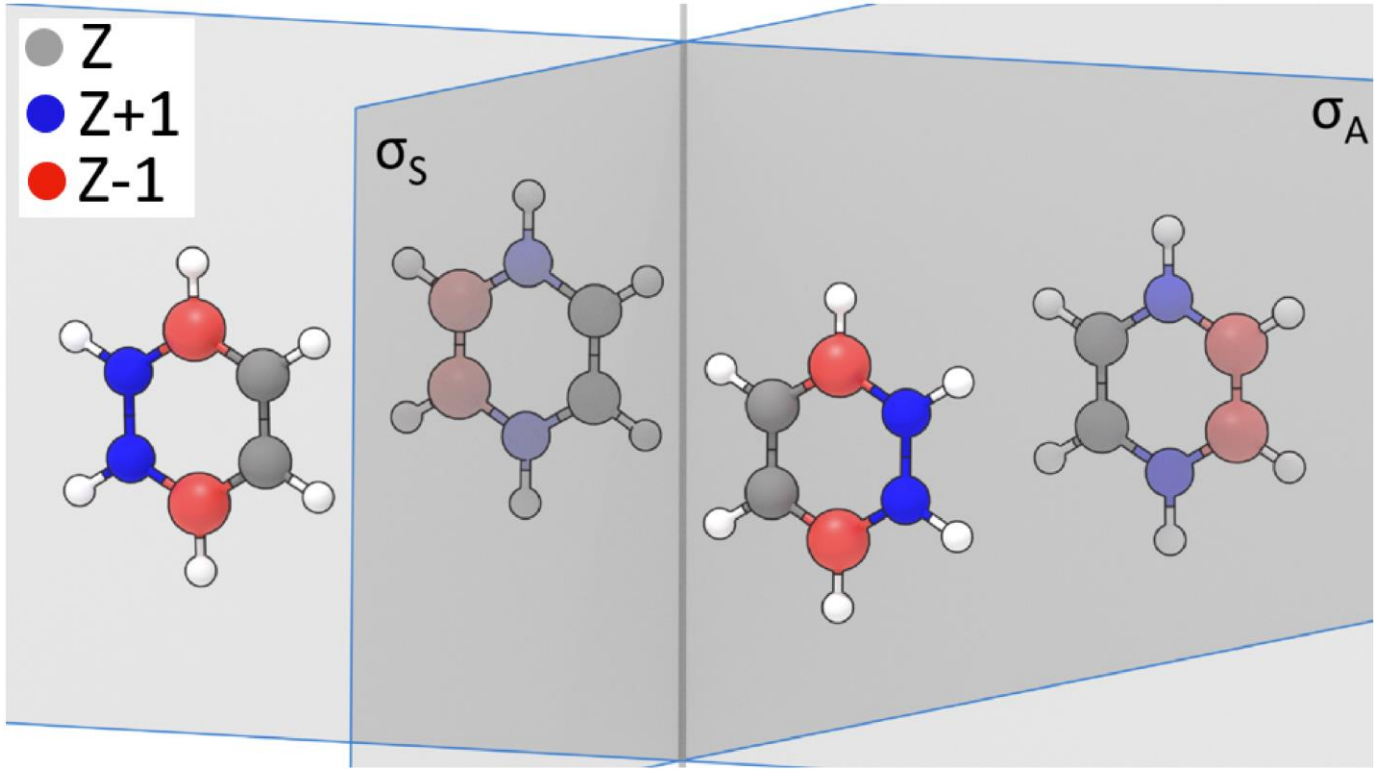


GFvR, A.A. Artemyev, B.M. Lagutin, P.V. Demekhin, *accepted.*

Alchemical Enantiomers





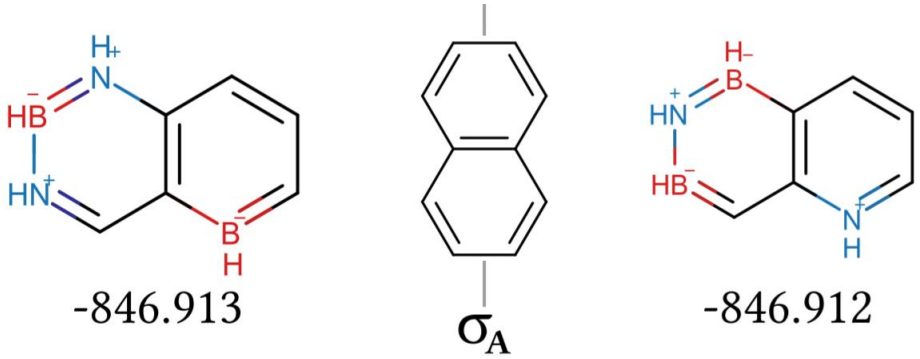


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

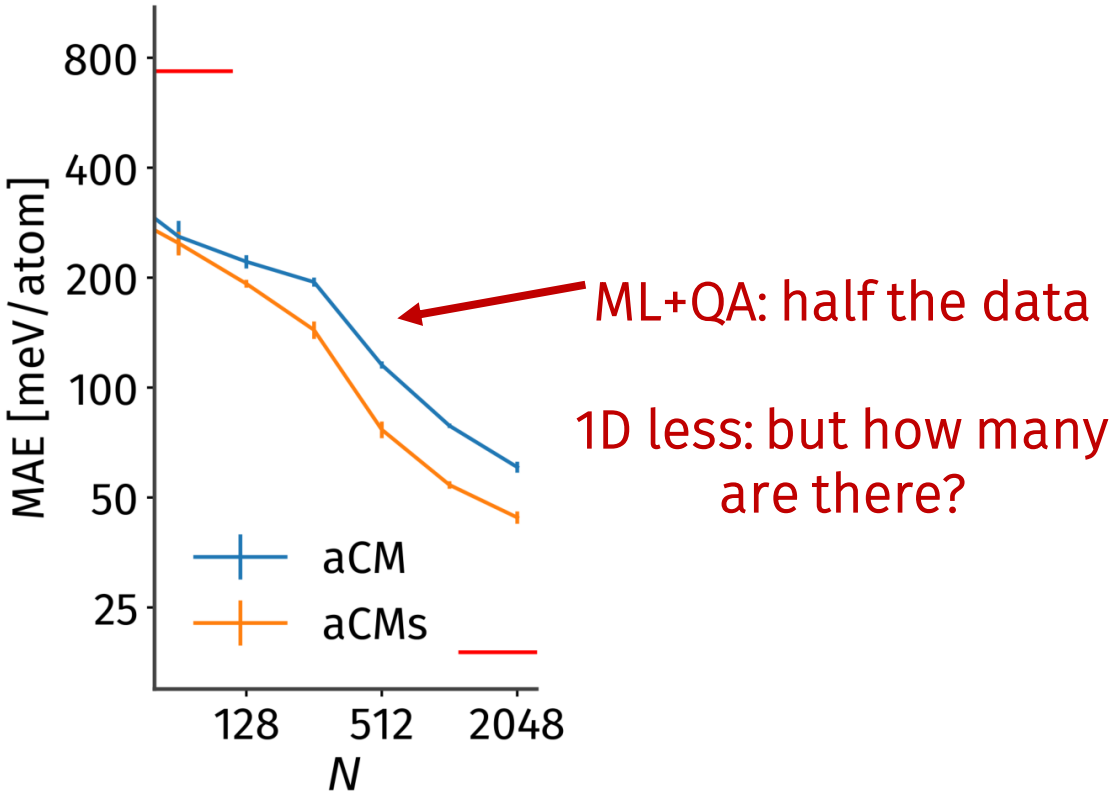
Alchemical Enantiomers

Fundamentally new symmetry

Electronic energy only



Speed up machine learning



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

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

A Banjafar, **GFvR**, *in preparation.*

Quantifying Dimensions

How to measure?

Dynamics in 4N embedding using quantum alchemy, keeping the property Q constant

$$U(\Delta\mathbf{R}, \Delta\mathbf{Z}) = k_1 \left[Q(\Delta\mathbf{R}, \Delta\mathbf{Z}) - Q_{\mathrm{ref}}\right]^2 + k_2 \|\Delta\mathbf{R} \oplus \Delta\mathbf{Z}\|$$

Trajectories form a level set / hyperplane of the null space.

(Local) intrinsic dimension D_Q

Apply point cloud estimators to level set: ID of null-space $D_{NS,Q}$

$$D_Q = 4N - D_{\mathrm{NS,Q}}$$

Limitations

- Sampling (long enough and bias-free)
- Point cloud ID estimators

Small molecules (up to six atoms)

- Equilibrium and non-equilibrium ID identical
- Scaling linearly, prefactor << 4

A Banjafar, **GFvR**, *in preparation.*

Random Sampling

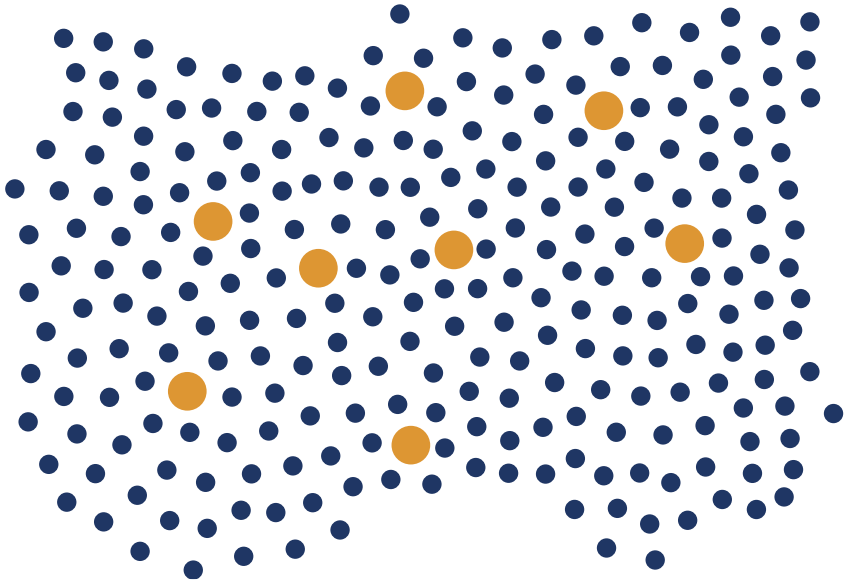
Allow for data-driven fundamental statements
“Most molecules do X”, “High X means low Y”

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



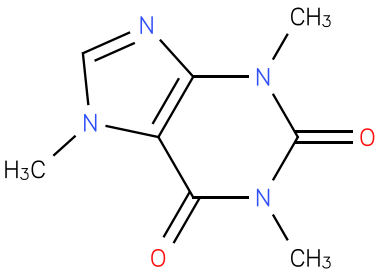
Random Sampling

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

Counting without enumeration

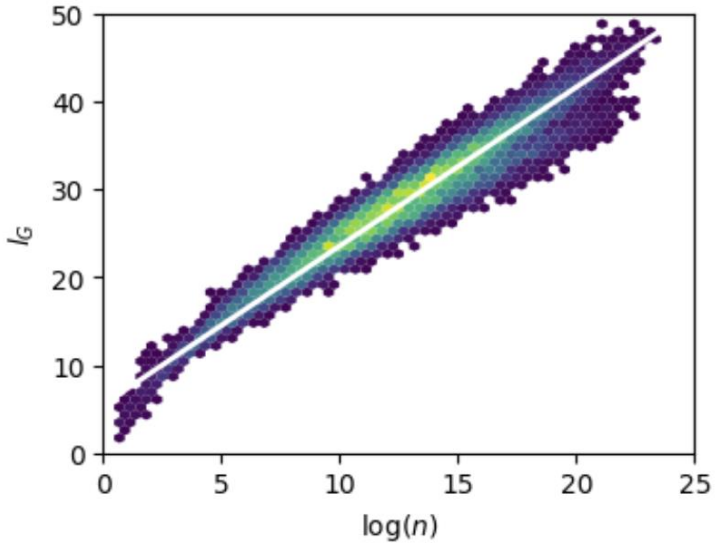
Goal

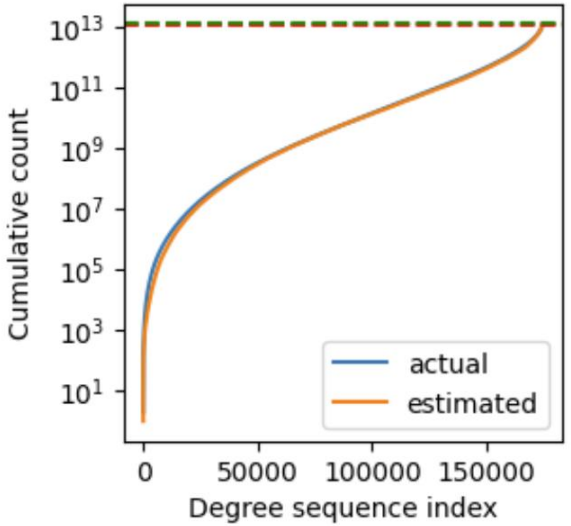
Estimate number n of loop-free multigraphs with given degree sequence.

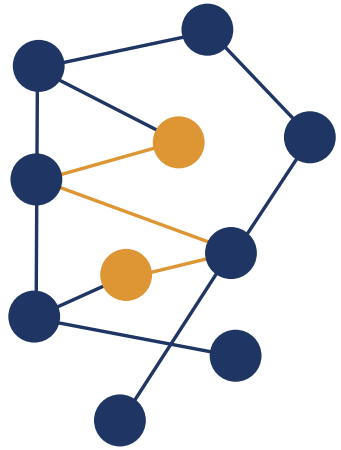
Average Path Length l_G

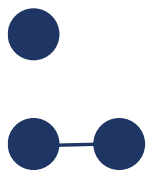
Sample from random molecule pairs

$$l_G \sim \log n$$









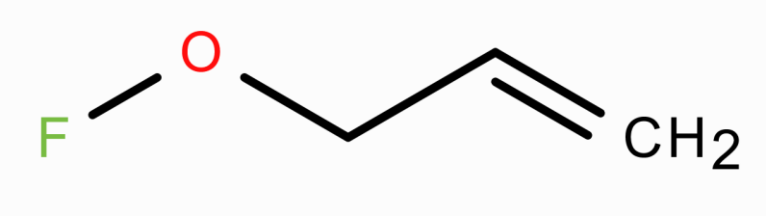
Molecule

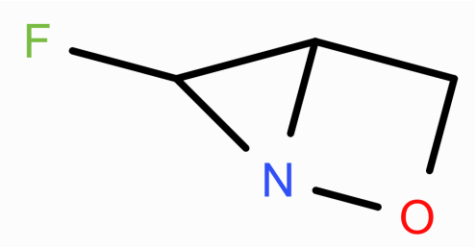
Identical up to one bond

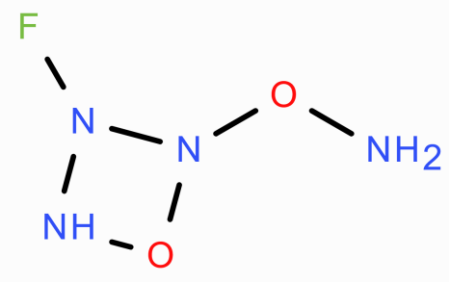
Minimum path

Examples

10 atoms, CHONF, at least 3 hydrogens and one fluorine



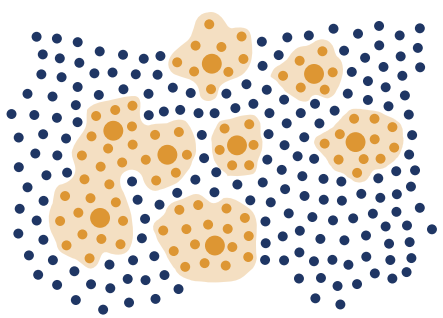


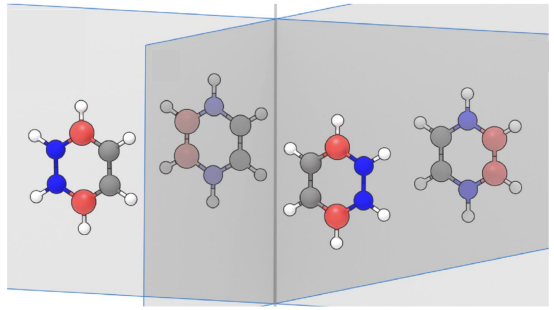


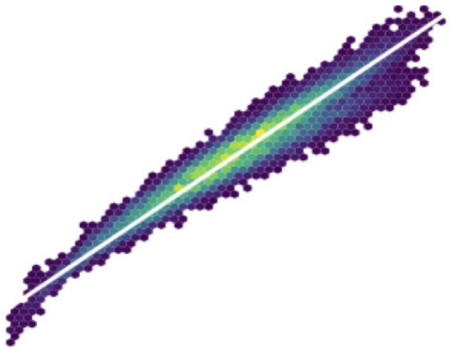
Available conditions in this approach

- Number of atoms
- Atoms per element (and combinations)
- Valences
- Via rejection sampling:
  - Bonds (count and bond orders)
  - Ring presence / membership
  - Substructures
  - …

Summary







Symmetries | Reducing (“folding”) search space

Dimensionality | Limit data efficiency

Random Sampling | Allows for collective statements

Thanks

Anton Artemyev
Ali Banjafar
Emily Eikey
Philipp Demekhin
Chasz Griego
John Keith
Boris Lagutin
Anatole von Lilienfeld
Alex Maldonado
Sana Qureshi

vonrudorff@uni-kassel.de

We’re hiring!

nablachem.org/talks

ferchault

@ferchault