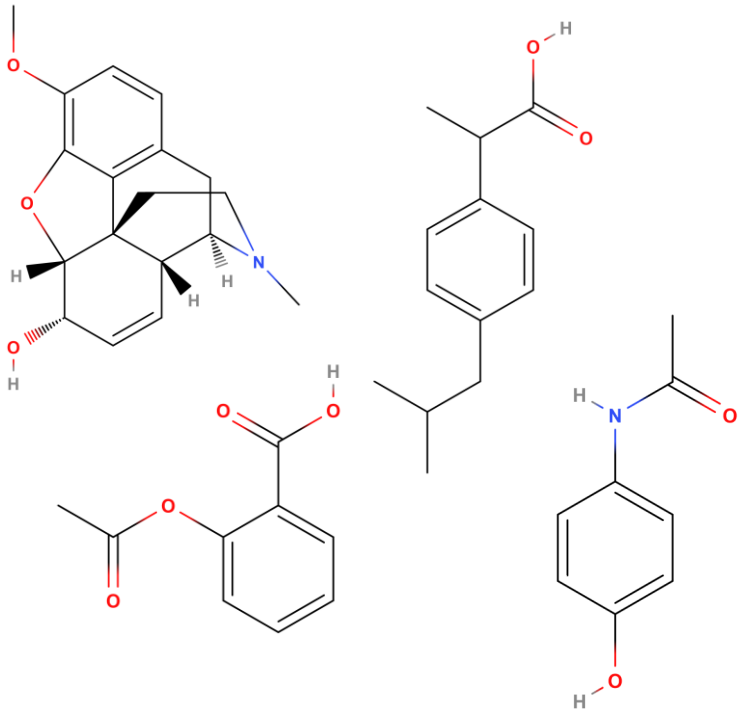
Design in Compound Space With Machine Learning and Quantum Alchemy

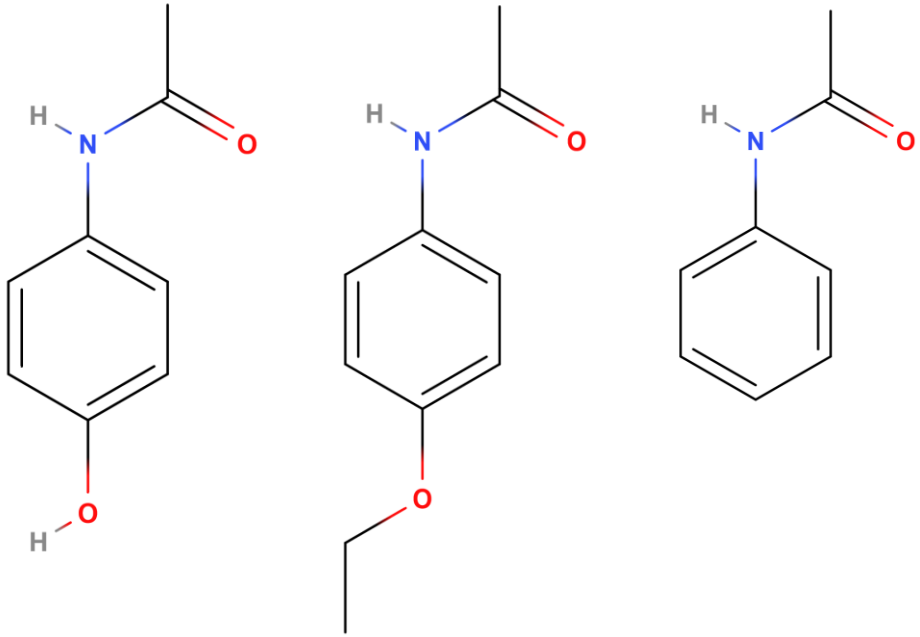
Guido Falk von Rudorff, University of Vienna

ferchault @ferchault guido.vonrudorff.de

Introduction

Design: sample by guided trial-and-error.





Global Search Problem
Which class of compounds?

Drug-like: 10^{60}

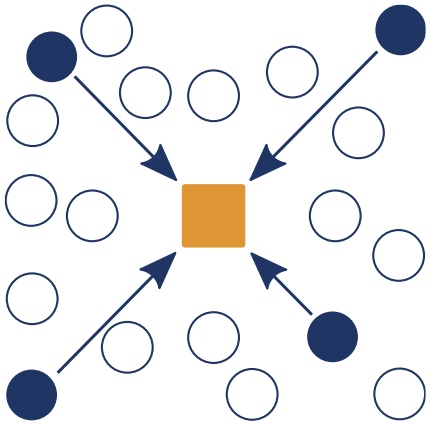
Local Search Problem
Which particular species within that class?

BN-doped 8x8 graphene: 10^{50}

G. Ceder, *Science* **1998**. A. Franceschetti, A. Zunger, *Nature*, **1999**. A. Mullard, *Nature*, **2017**.

Introduction

Machine Learning



Foundations | Statistical modelling

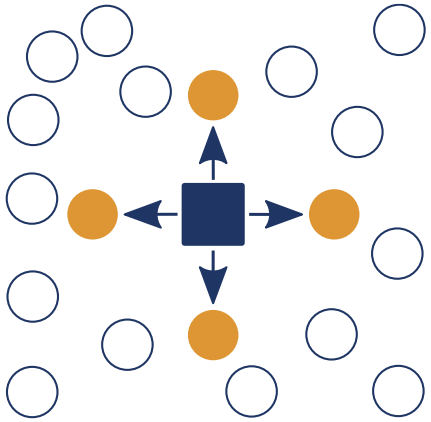
Accuracy | Systematically improvable through data and training

Specialty | Universal, scale-bridging, data-driven approach

Limitation | Requires training data, no black box

Chemistry: exciting new domains become accessible!

Quantum Alchemy



Foundations | Perturbation theory

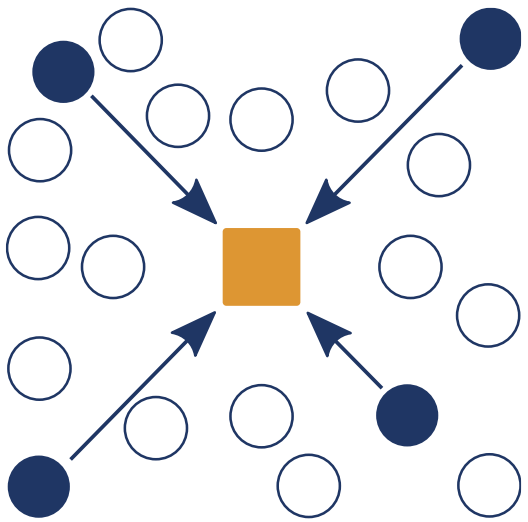
Accuracy | Systematically improvable through higher orders terms

Specialty | Combinatorial scaling with chemical diversity

Limitation | Finite range in chemical space

Introduction: Machine Learning

Machine Learning



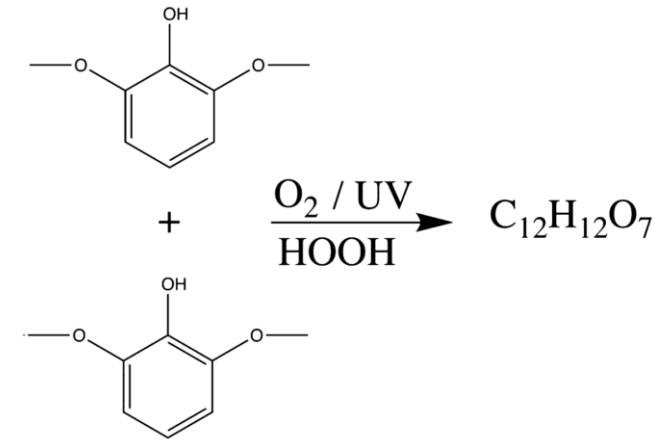
Kernel-Ridge-Regression

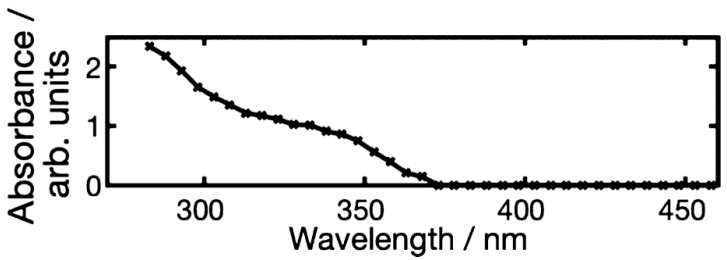
- Efficient in the low-data regime (around 1k points)
- Ingredients
  - Representation $\mathbf{M}$
  - Similarity measure $k(\mathbf{M}_i, \mathbf{M}_j)$
  - Observed properties $\mathbf{y}$
- Training
  - Pairwise similarities $\mathbf{K}$
  - Model coefficients $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}$
- Predictions
  - Compare to training $\tilde{q}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i)$

V. N. Vapnik: The Nature of Statistical Learning Theory, *Springer* 2000.
B. Schölkopf, A. J. Smola: Learning with Kernels, *MIT Press*, 2001.

Identification of atmospheric brown carbon

Experiment



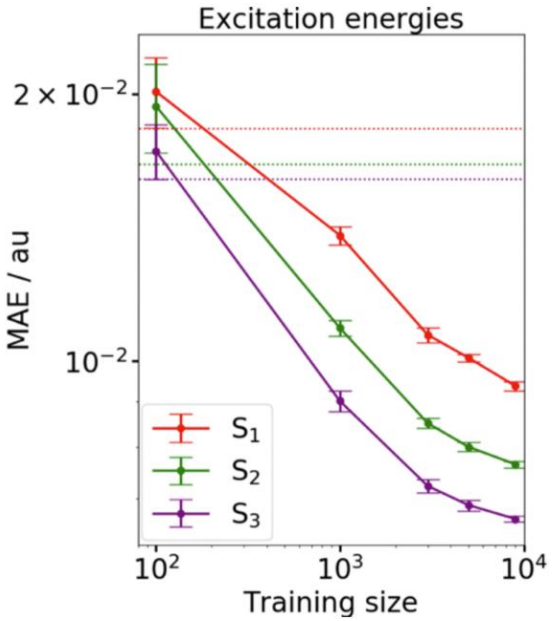


Search space

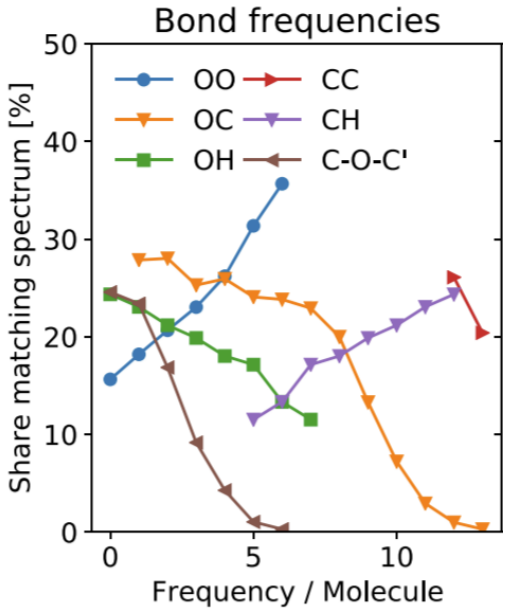
Molecular graphs: 264 M
Stable molecules: 123 M

Model

Excitation energies
Oscillator strengths



Identified features

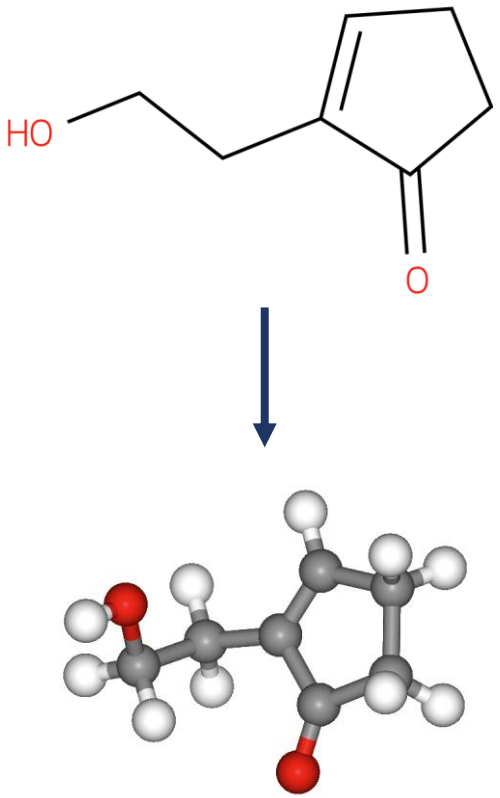


Guide experiment

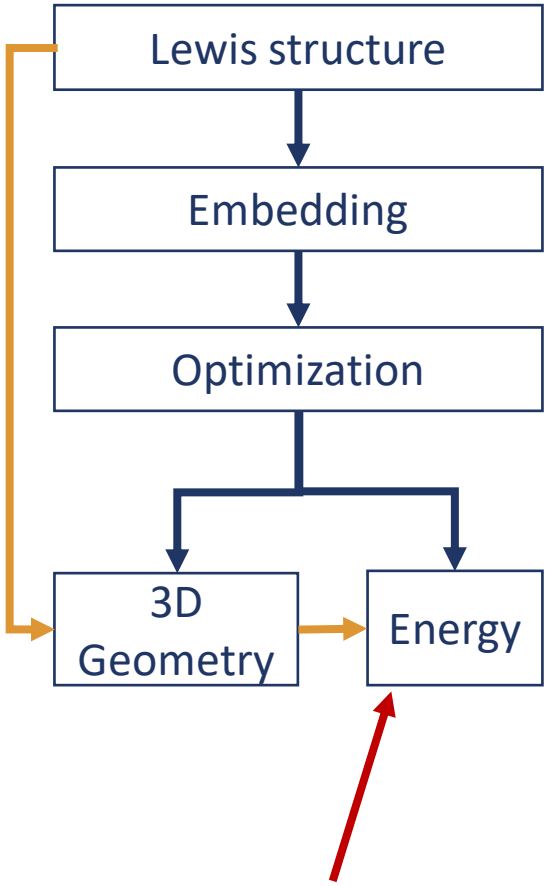
How many molecules are left and which feature to measure next?

E. Tapavicza, **GFvR**, D. O. De Haan, M. Contin, C. George, M. Riva, O. A. von Lilienfeld, *Environ. Sci. Technol.*

Geometry learning



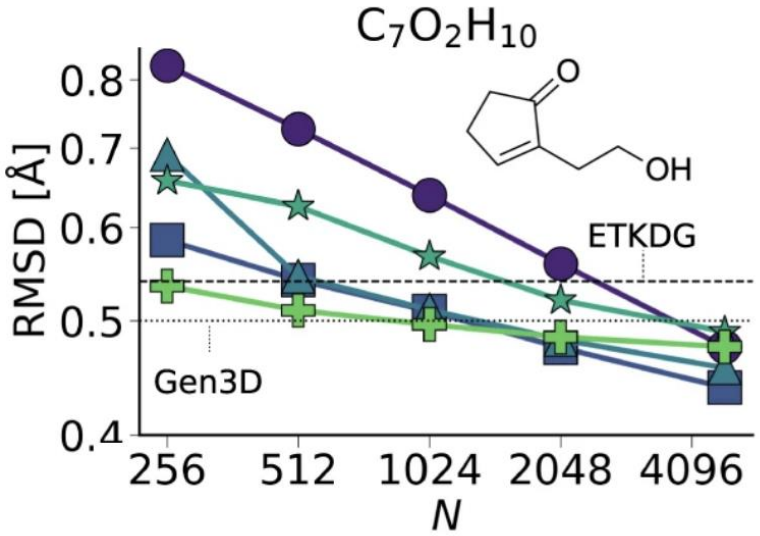
Traditional / G2S



ML: 100.000x faster

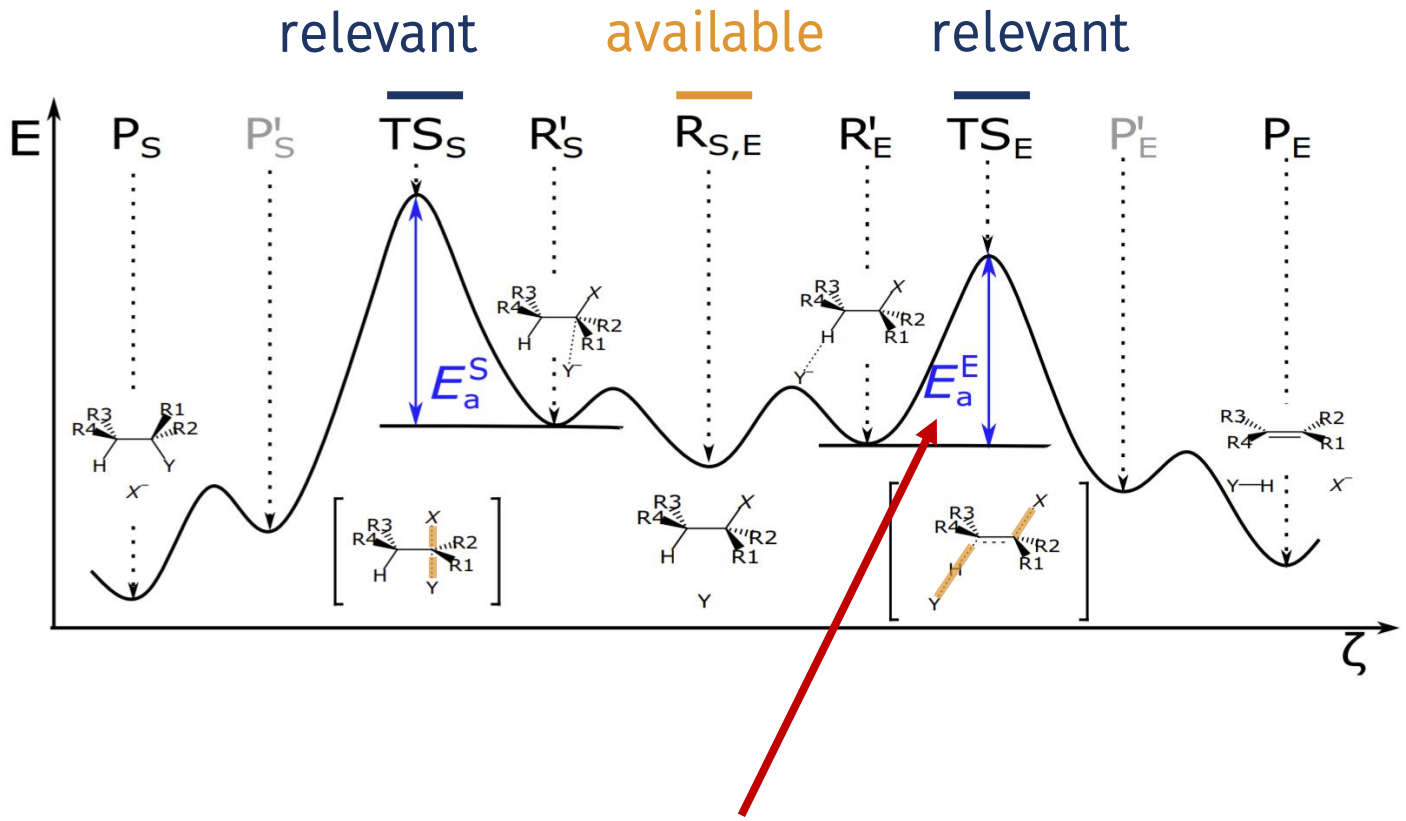
G2S

- Closer to DFT than common methods
  - Small molecules
- Applicable to complex chemical spaces
  - Transition state geometries
  - Carbenes



D. Lemm, **GFvR**, O. A. von Lilienfeld, *Nat. Commun.* **2021**.

Reaction barrier learning



ML: 10.000x faster

Competing reactions: E2, S_N2

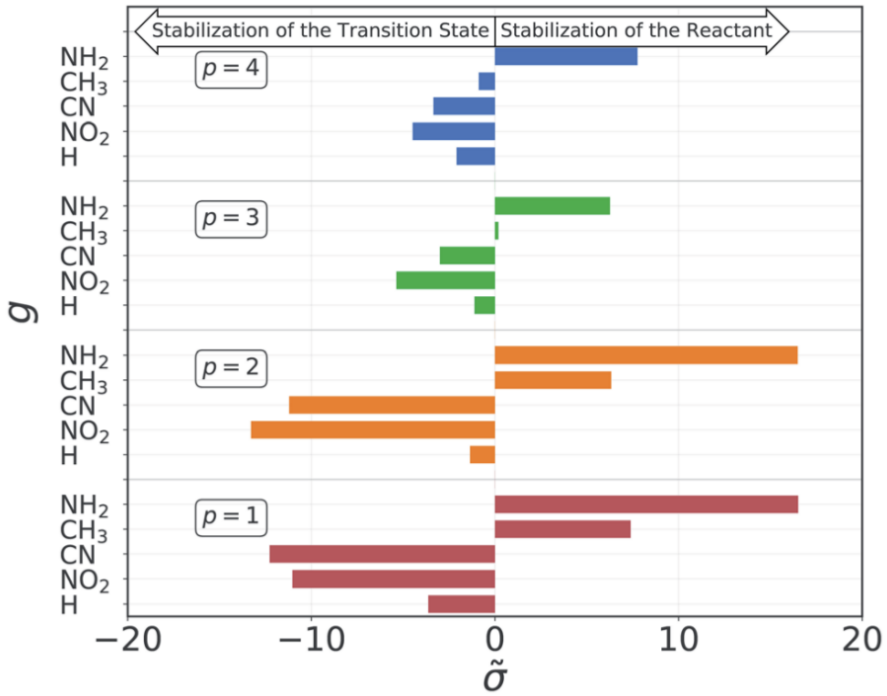
- 4.5k reactions in one new dataset
- Learning activation energies from reactants only reaching 2.5 kcal/mol with 800 data points
- Learning geometries of transition states
  - direct
    0.05 Angstrom for distances
  - G2S
    0.45 Angstrom heavy-atom RMSD

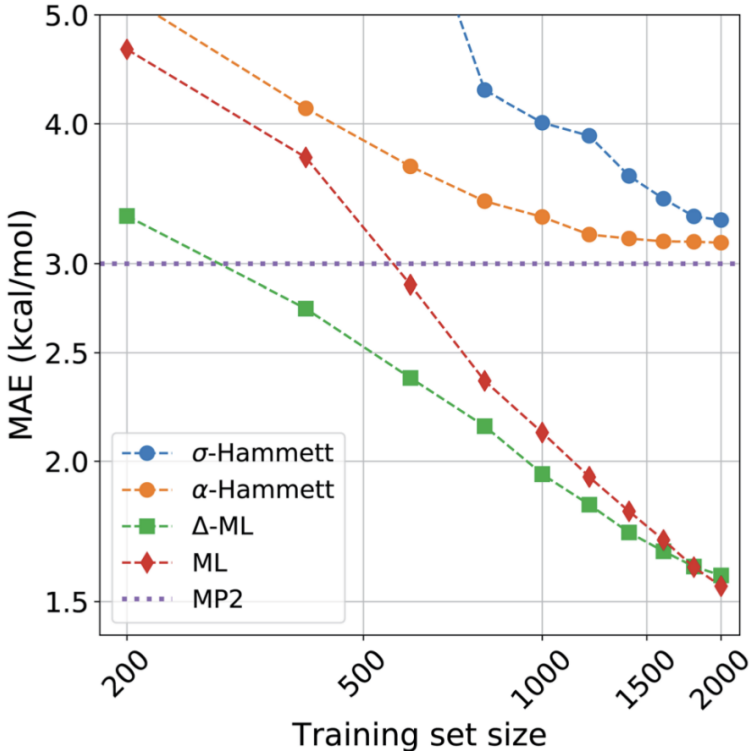
Hammond's postulate

- valid for S_N2, not for E2
- E2 preferred in 75% of the cases: less sensitive to reactant complex geometry

GFvR, S.N. Heinen, M. Bragato, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.*, **2020**.
S.N. Heinen, **GFvR**, O. A. von Lilienfeld, *J. Chem. Phys.*, **2021**.

Reaction barrier learning





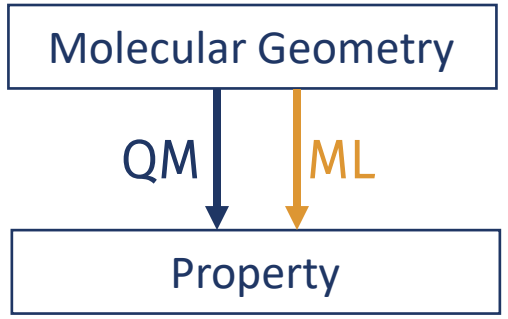
Hammett equation

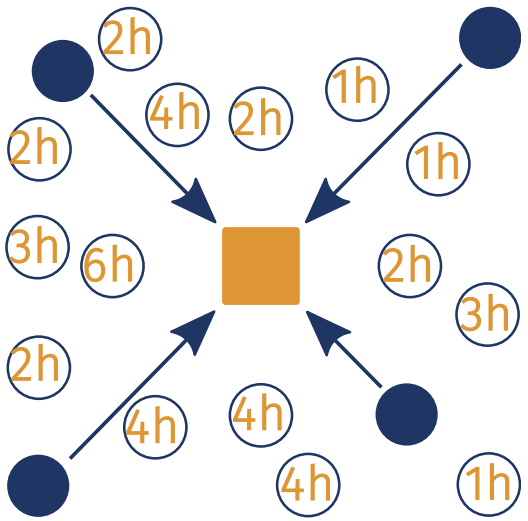
Electron withdrawing/-donating order of substituents recoverable

- Physics-based model
- Use Hammett as baseline

M. Bragato, **GFvR**, O. A. von Lilienfeld, *Chem. Sci.* **2020**.

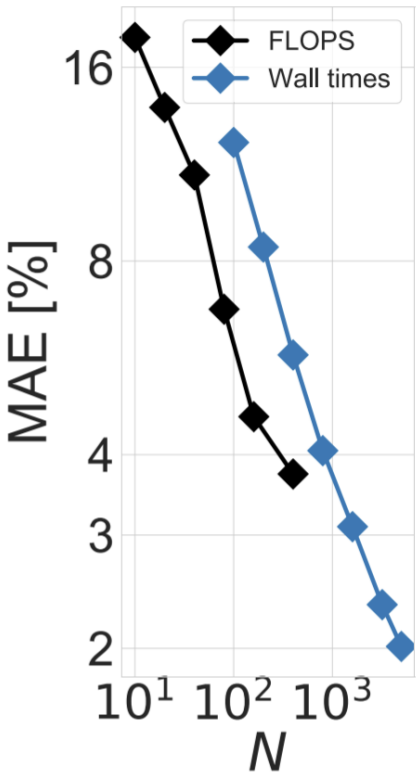
Computational cost





Computational effort as molecular property

- Improves models
- Accuracy depends on problem
  - Single points: 2%
  - Transition state search: 25%
  - Geometry optimisations: 40%



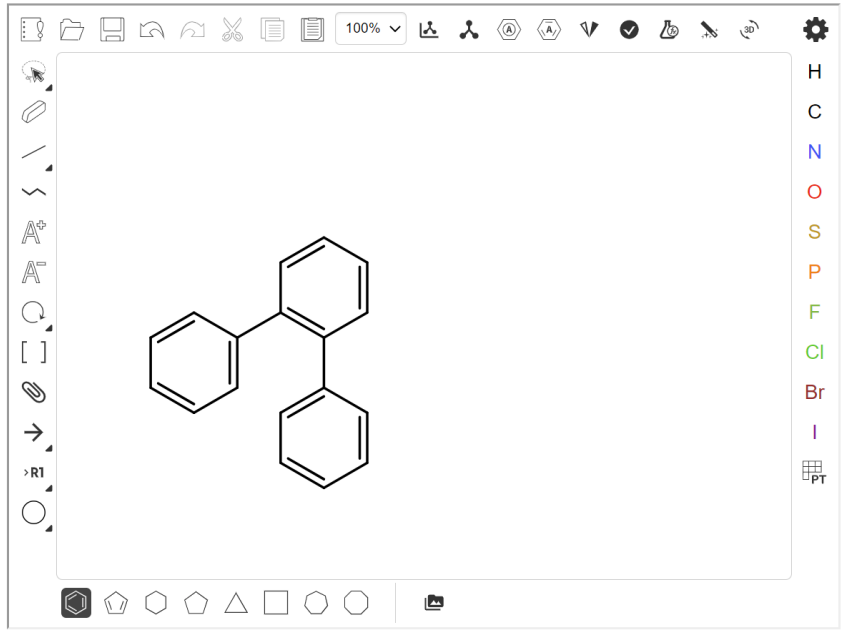
S.N. Heinen, M. Schwilk, **GFvR**, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* **2020**.

Model availability

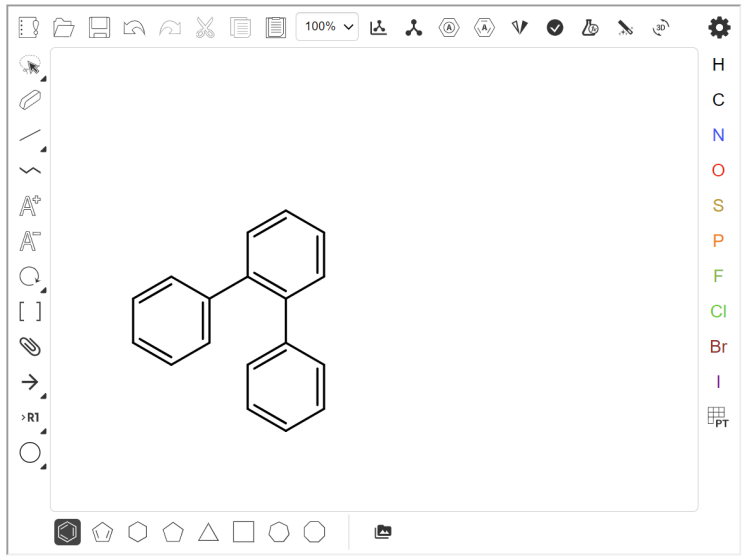
- chemspacelab/Enhanced-Hammett
- qmlcode/qml
- ferchault/mlscheduling
- ...

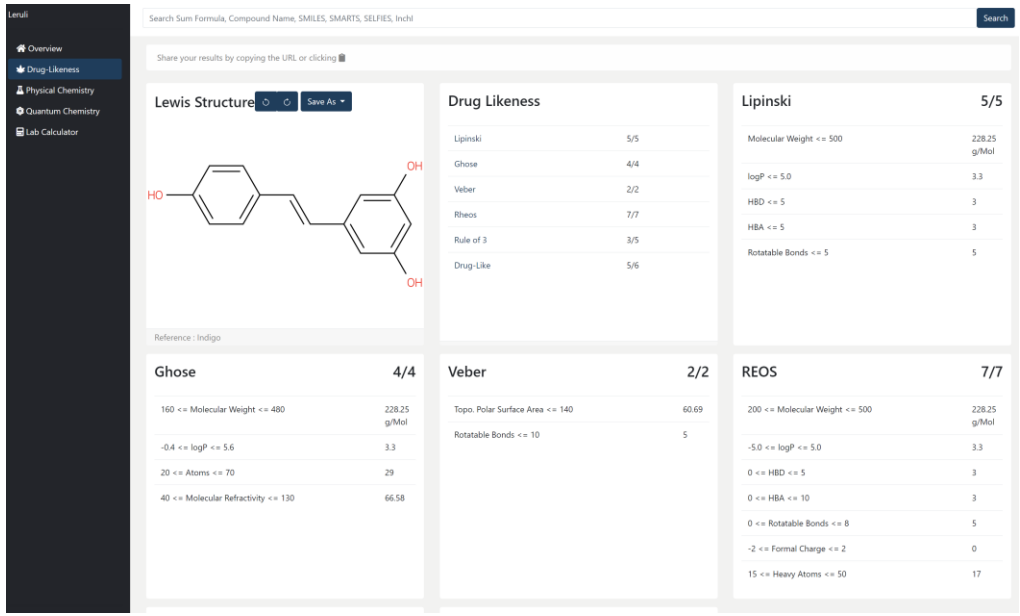
- Dependencies might break
- Might be an old model
- Tedious/risky to get started
- Therefore:

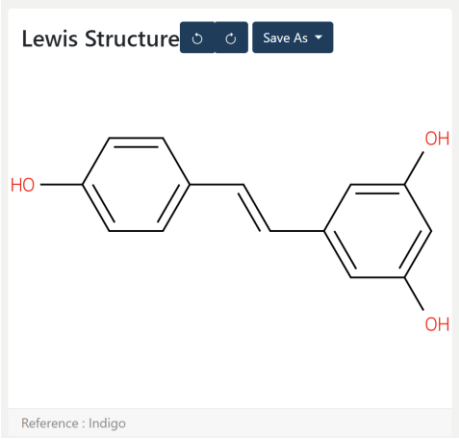
leruli.com

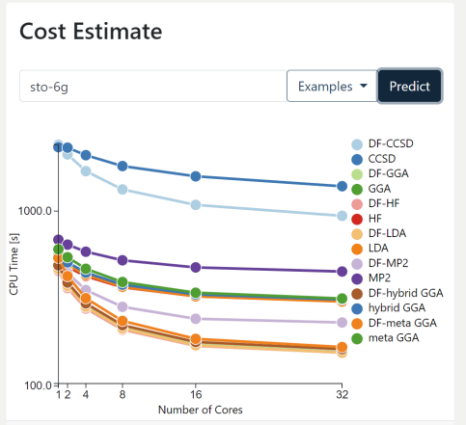


Model availability





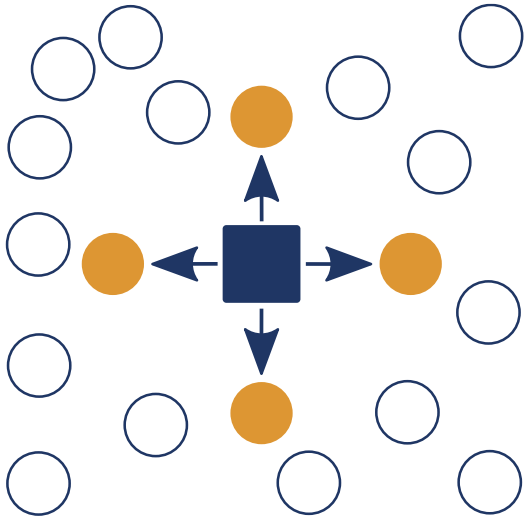




Want to include your model? Let me know

Introduction: Quantum Alchemy

Quantum Alchemy

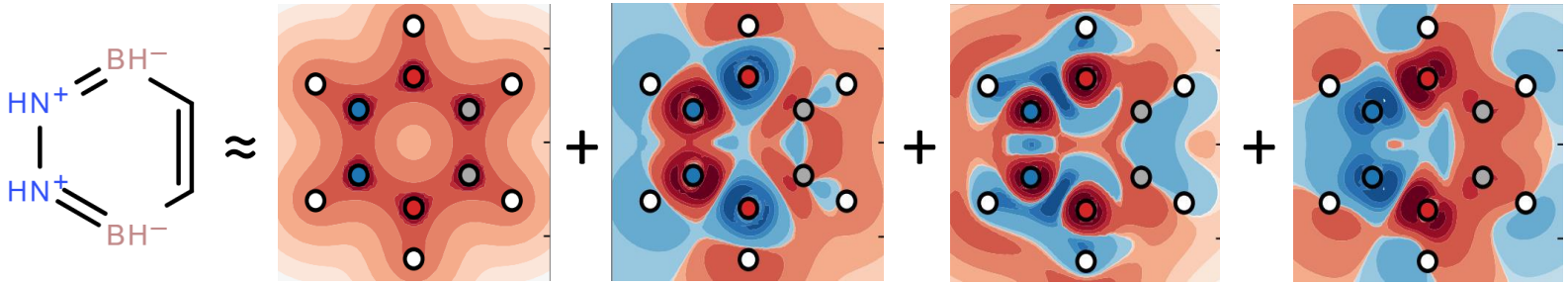


Taylor expansion

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

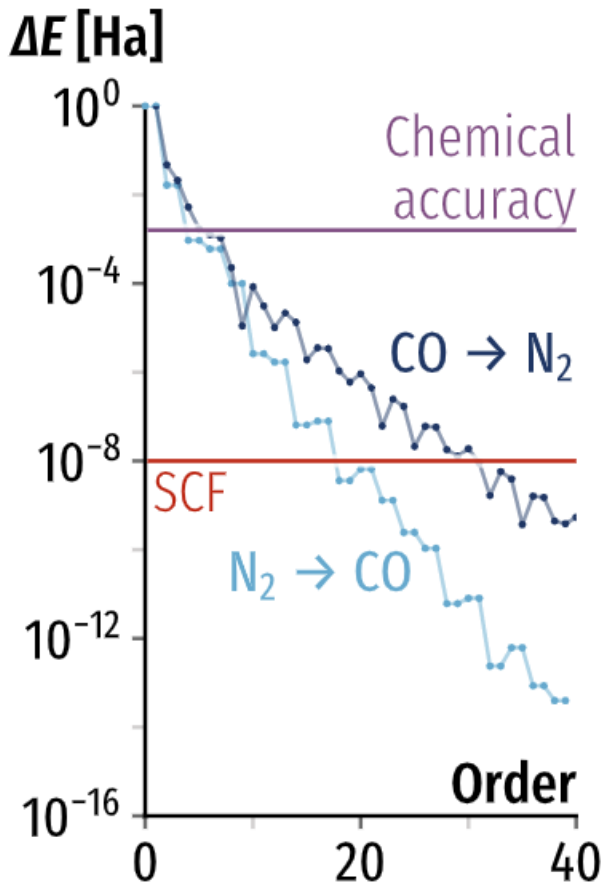
$$\hat{H}(\lambda) \equiv \lambda \hat{H}_\mathrm{t} + (1-\lambda)\hat{H}_\mathrm{r} \qquad \lambda \in [0,1]$$

$$E_\mathrm{t} = E_\mathrm{r} + \sum_{n=1}^{\infty} \frac{1}{n!} \left. \frac{\partial^n E(\lambda)}{\partial \lambda^n} \right|_{\lambda=0}$$



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

Covalent interactions



Taylor expansion

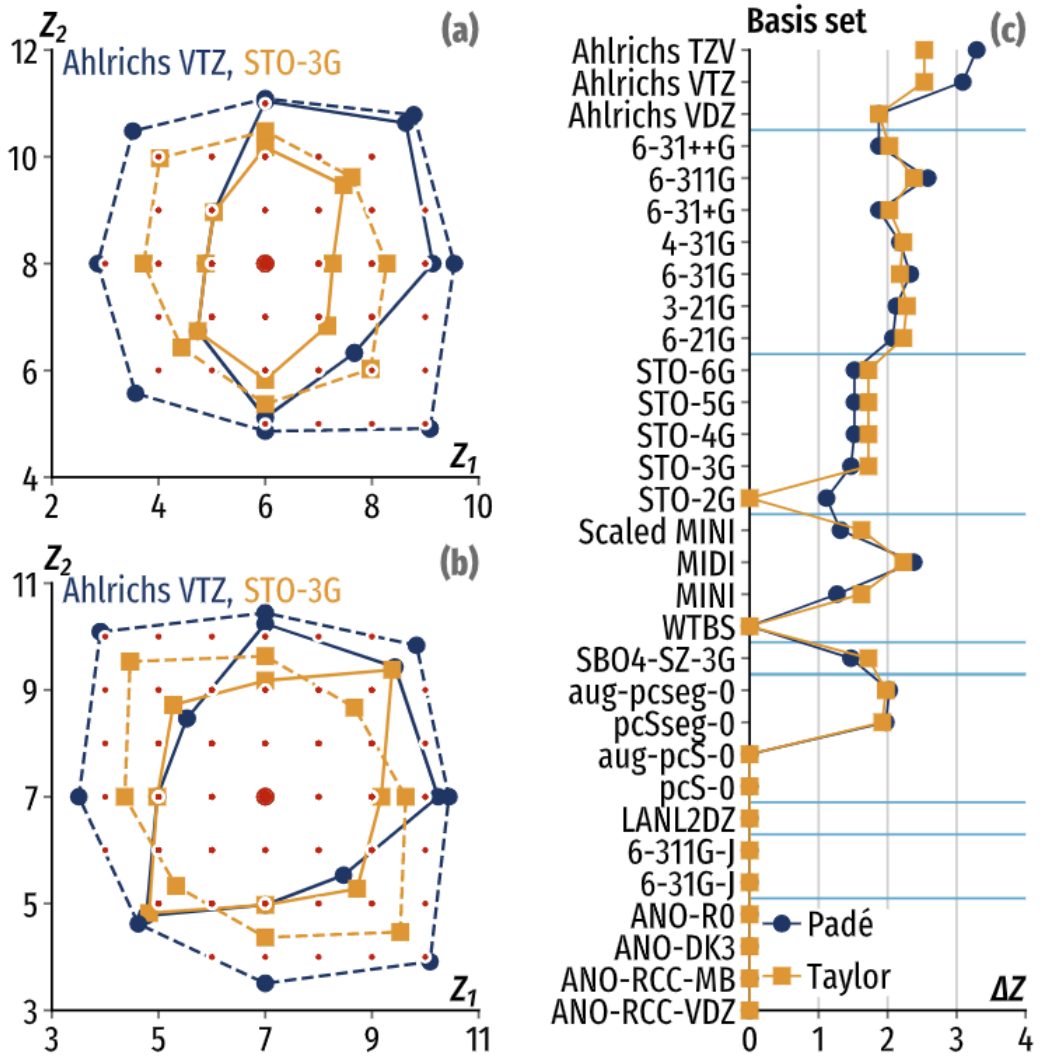
- First terms accurate enough
  - Truncate early
- Scales with chemical space

Paradigm shift

Few highly accurate calculations instead of many intermediate ones

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

Covalent interactions

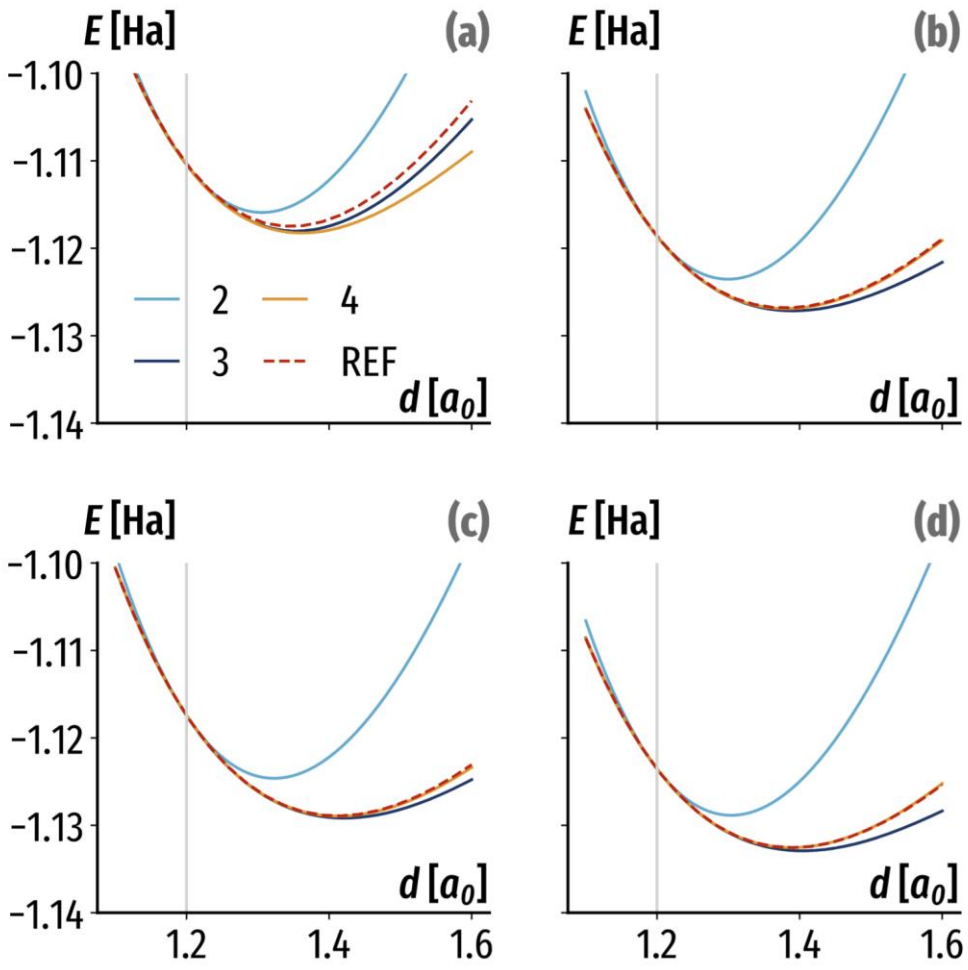


Padé approximant

Larger convergence radius

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

Covalent interactions



Mixed derivatives

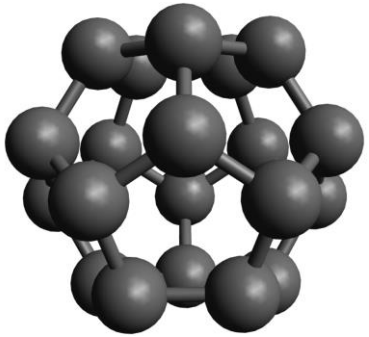
A path to relaxation

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

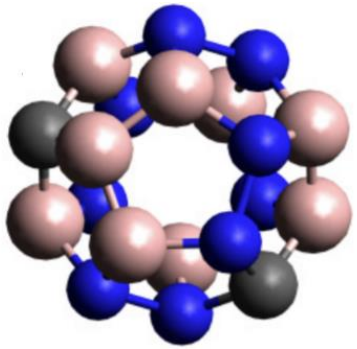
Covalent interactions

Scaling with chemical space

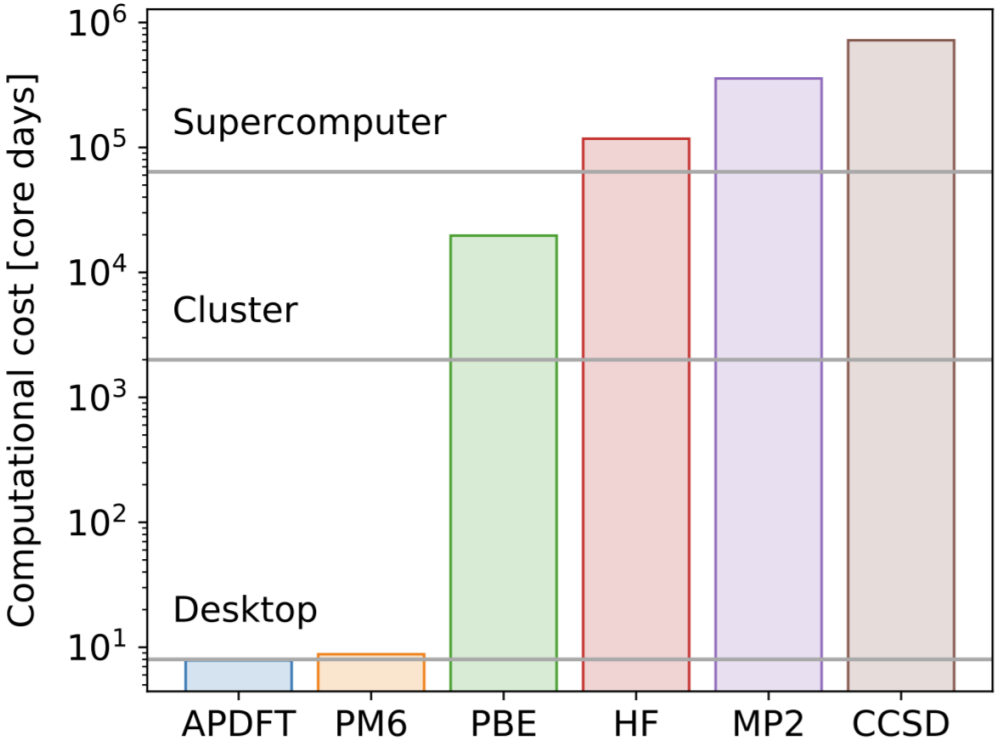
- 1 derivative for second order
- 5 derivatives for third order



C_{20}



$3.1 \cdot 10^6$ targets



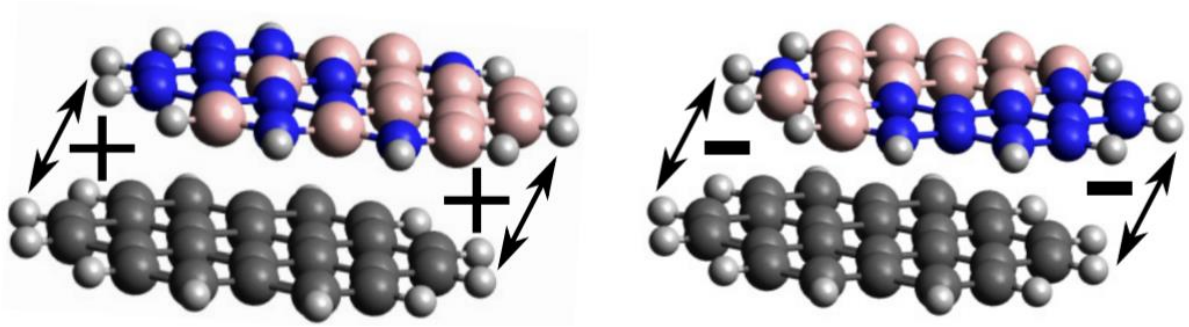
QA: 80.000x faster

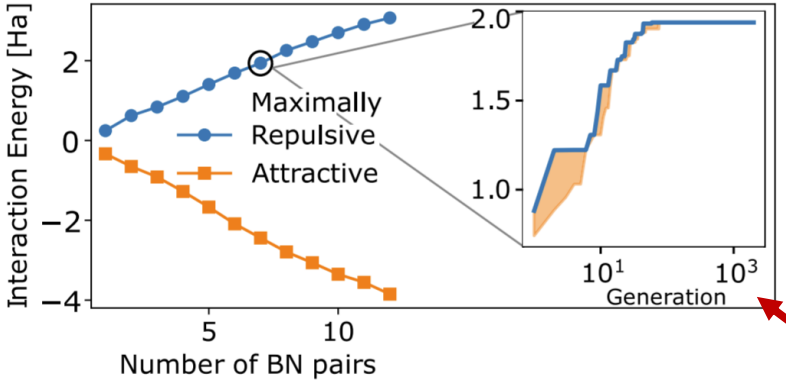
GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, 2020.

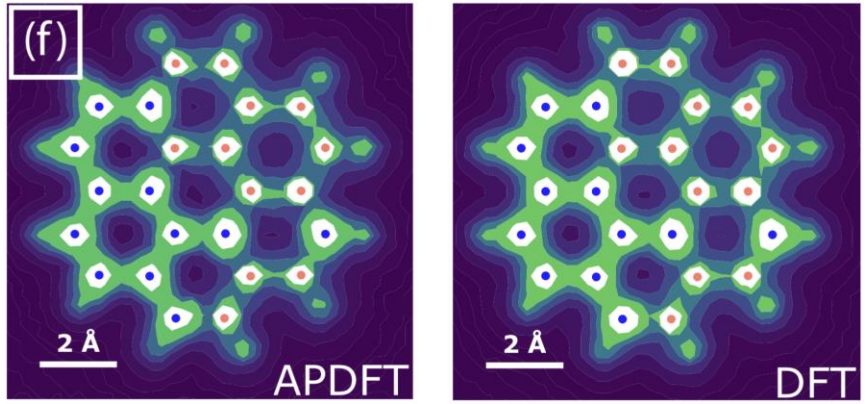
Non-covalent interactions

BN-doped coronene dimer

- Identify most/least attractive doping pattern
- Design example!







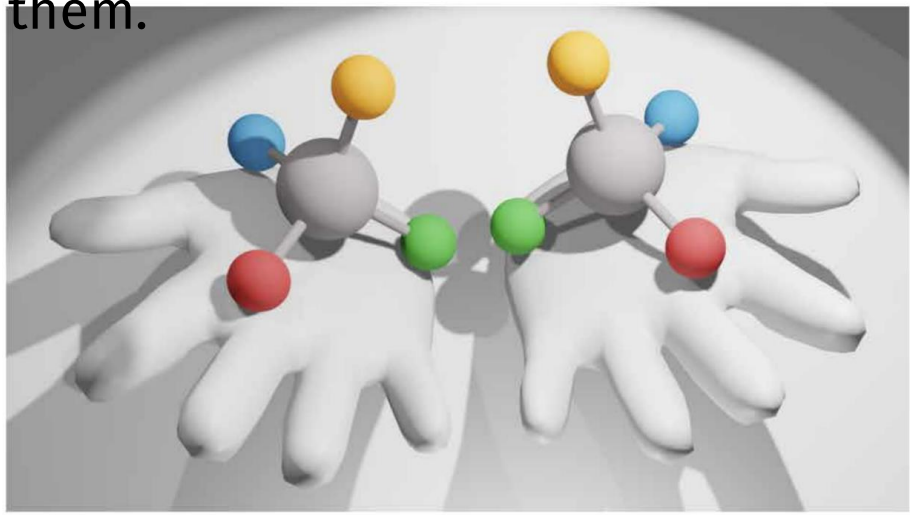
QA: 20.000x faster

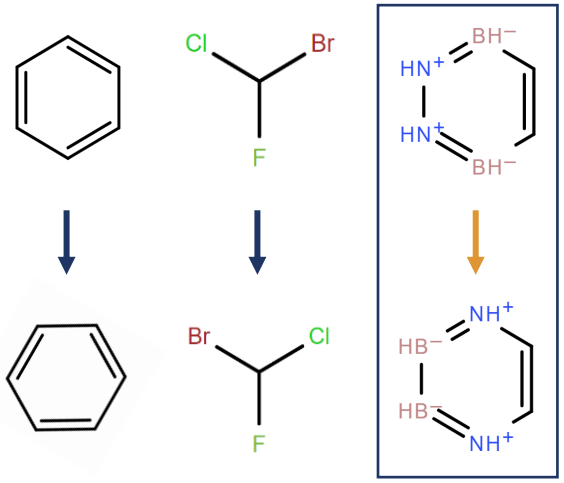
$2.8 \cdot 10^{10}$ targets

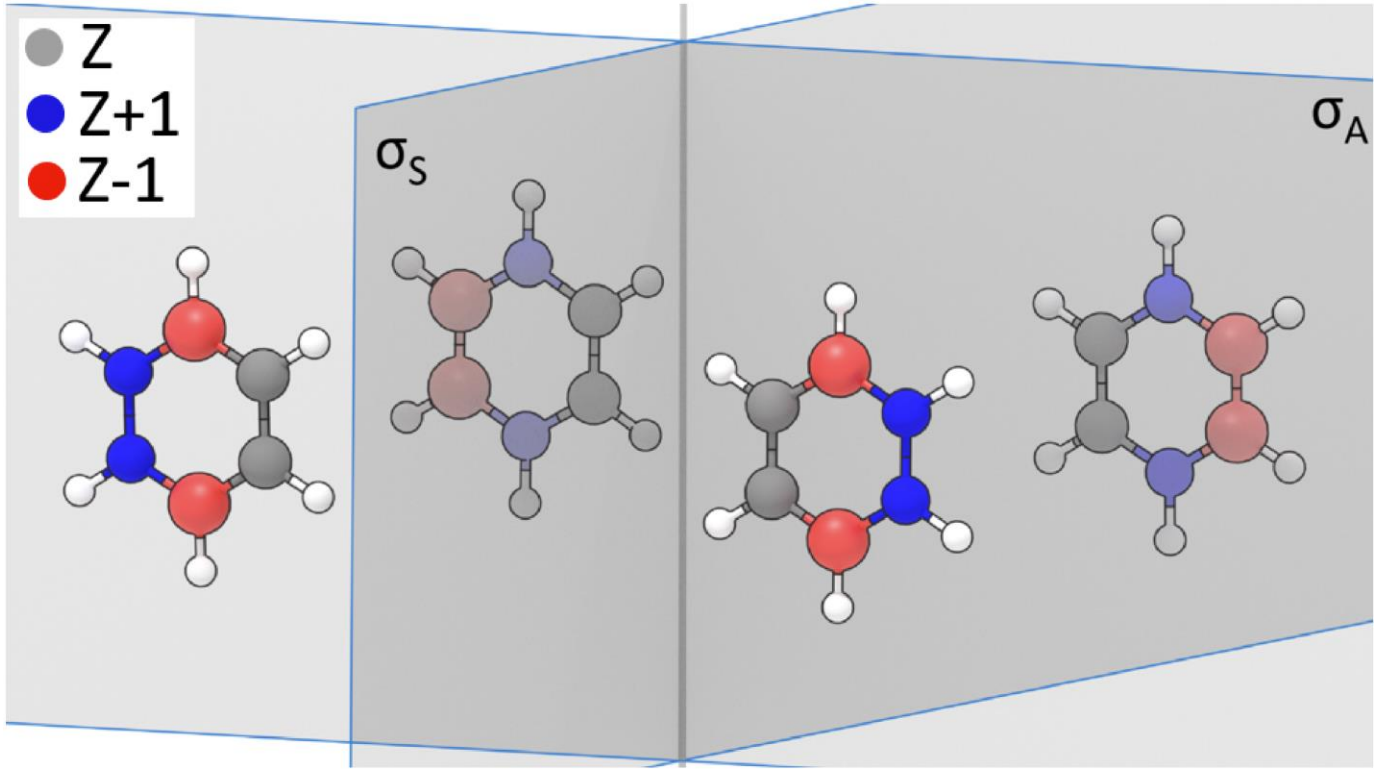
GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.*, 2020.

Alchemical enantiomers

Quasi-degeneracy for systems if this symmetry applies to them.







Theory

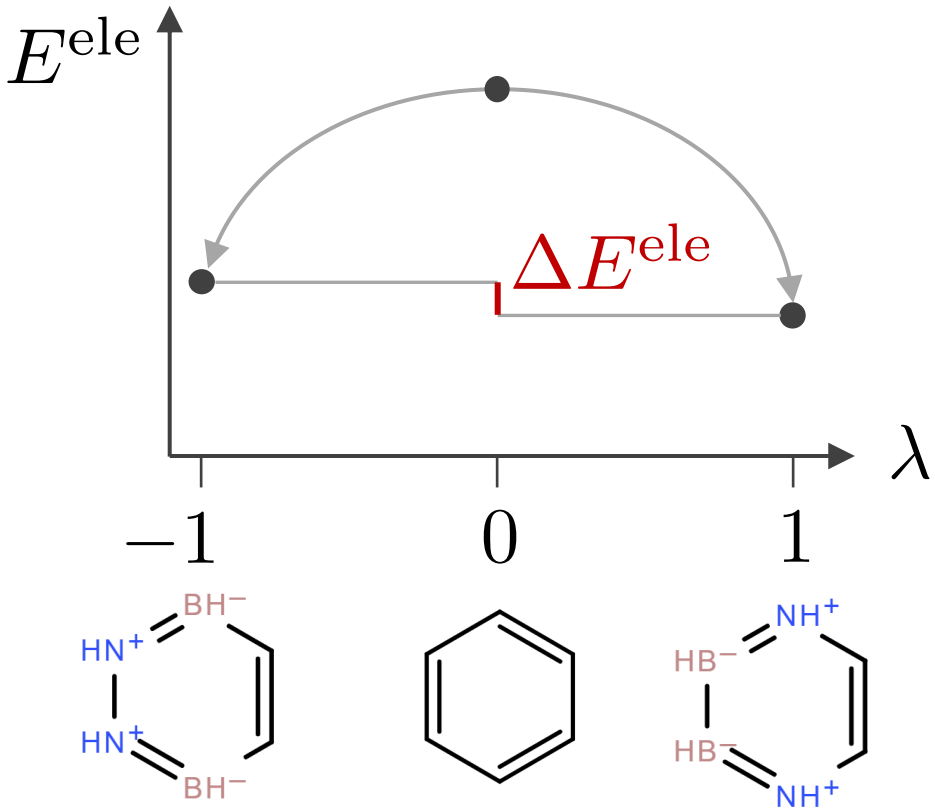
$$\Delta E_{ij}^{\text{ele}} = E_0 - E_0 + \int_\Omega d\mathbf{r} \sum_{n=0}^{\infty} \frac{\Delta v_i}{(n+1)!} \left[\frac{\partial^n \rho}{\partial \lambda_i^n} + \frac{\partial^n \rho}{\partial \lambda_j^n} \right]$$

$$\Delta E_{(0)}^{\text{ele}} = E_0 - E_0 = 0$$

$$\Delta E_{(1)}^{\text{ele}} = 2 \int_\Omega \Delta v \rho = \int_\Omega e \cdot o = 0$$

$$\Delta E_{(2)}^{\text{ele}} = \int_\Omega \Delta v \left[\frac{\partial \rho}{\partial \lambda_i} + \frac{\partial \rho}{\partial \lambda_j} \right]$$

$$= \int_\Omega \Delta v \left[\sum_I \frac{\partial \rho}{\partial Z_I} \frac{\partial Z_I}{\partial \lambda_i} + \frac{\partial \rho}{\partial Z_I} \frac{\partial Z_I}{\partial \lambda_j} \right] = 0$$

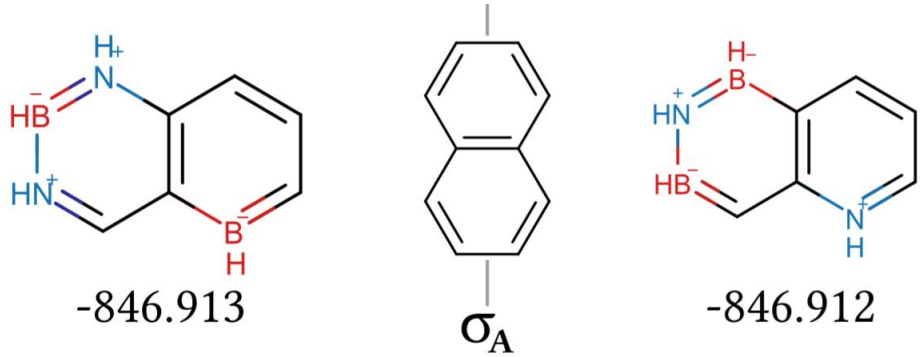


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

Alchemical enantiomers

Fundamentally new symmetry

Electronic energy only



Bond energy rules

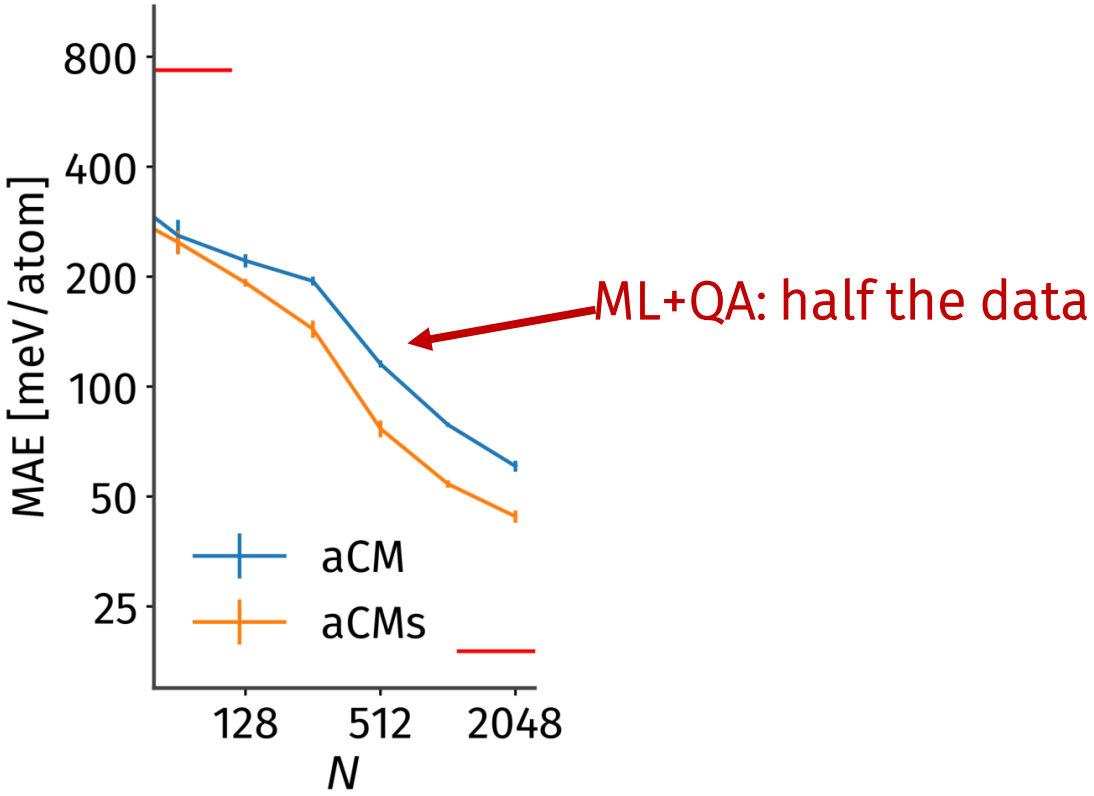
Consecutive Elements

Q R S

B C N

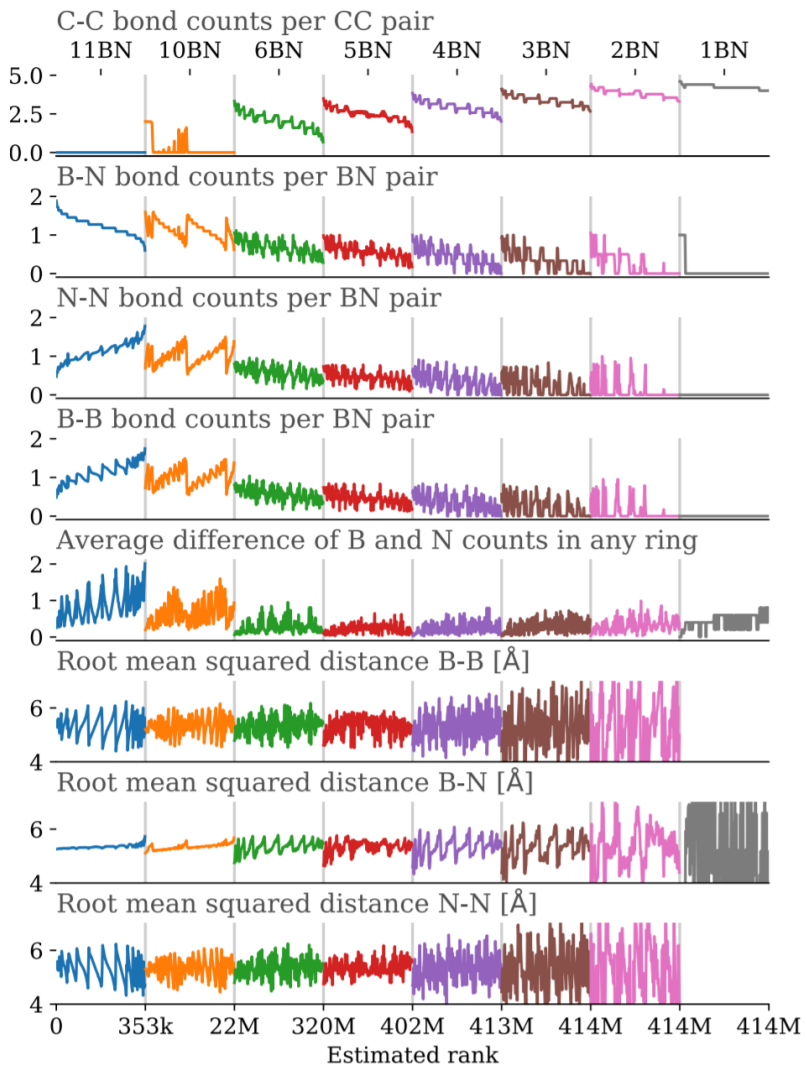
$$E_{\mathrm{QR}} \simeq E_{\mathrm{SR}} + 0.5(E_{\mathrm{QQ}} - E_{\mathrm{SS}})$$

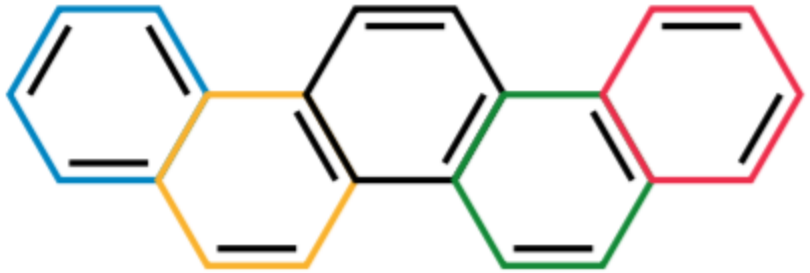
Speed up machine learning



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

Alchemical enantiomers





x 414 M

QA: Millions at once!

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.

Summary

Machine Learning | Global search, information transfer

Quantum Alchemy | Local gradients, structure in chemical space, replaces brute force

Reaction Barriers | S.N. Heinen, **GFvR**, O. A. von Lilienfeld, *J. Chem. Phys.*, 2021.

Geometry Learning | D. Lemm, **GFvR**, O. A. von Lilienfeld, *Nat. Commun.* 2021.

Quantum Alchemy | **GFvR**, *J. Chem. Phys.* 2021.

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

Prof. von Lilienfeld

Dominik Lemm

Marco Bragato

Stefan Heinen

Dr. Max Schwilk

ferchault @ferchault guido.vonrudorff.de leruli.com