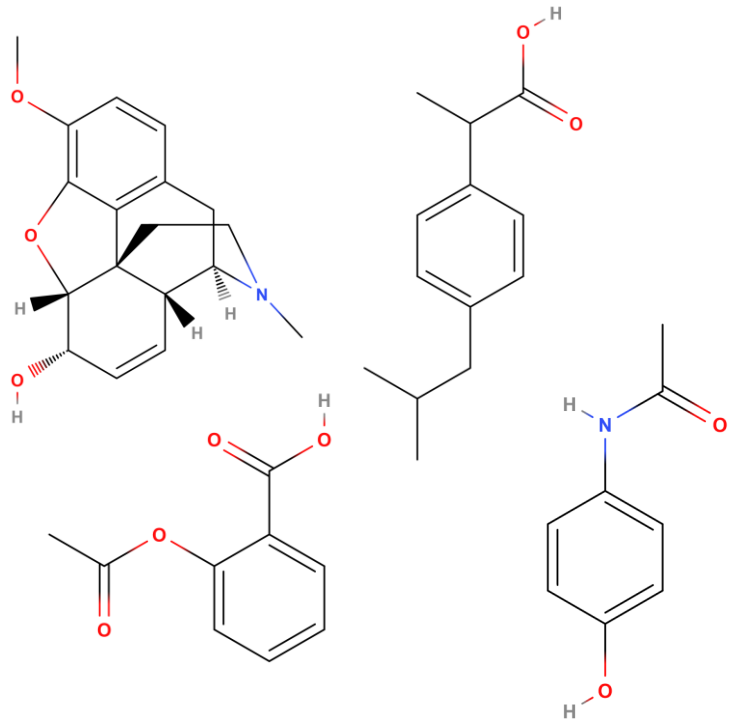
Machine Learning and Quantum Alchemy

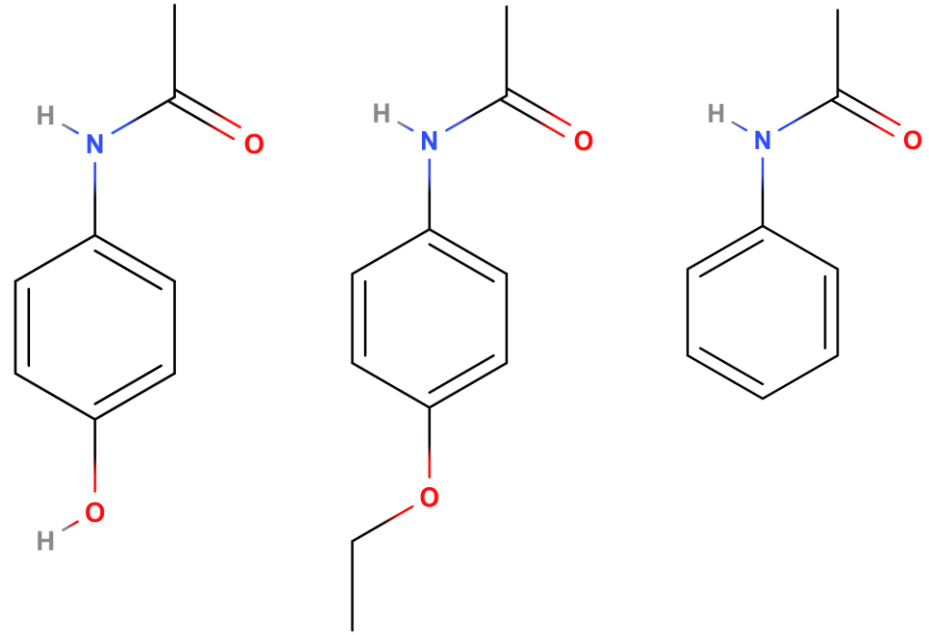
Guido Falk von Rudorff, University of Vienna

 ferchault  @ferchault  guido.vonrudorff.de

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?

Introduction

Global Search Problem
Which class of compounds?

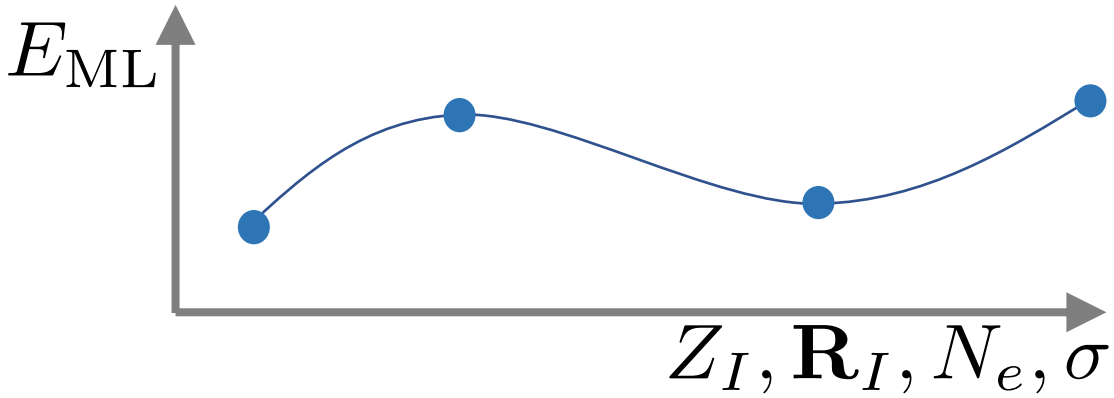
- Human expertise
- Bioinformatics
- Machine learning

Local Search Problem
Which particular species within that class?

- Human expertise
- High-throughput experiments
- Quantum alchemy

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
- Alchemical perturbation density functional theory
- Taylor expansion w.r.t. elements

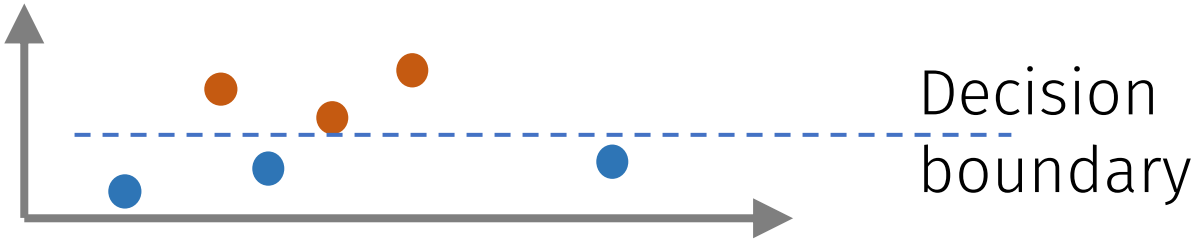


Machine Learning

Kernel Ridge Regression

Idea

- Molecular representation for each molecule *i* $\mathbf{M}_i$
  - CM, BoB, FCHL, SLATM, ...
- Distance metric $d_{ij} \equiv d(\mathbf{M}_i, \mathbf{M}_j)$
  - Typically L1 or L2 norm
- Kernel function $k_{ij} \equiv k(d_{ij})$
  - Laplacian, Gaussian
- Linearly separable in higher dimensions
  - ...without explicit higher dimensions



Kernel Ridge Regression

Procedure

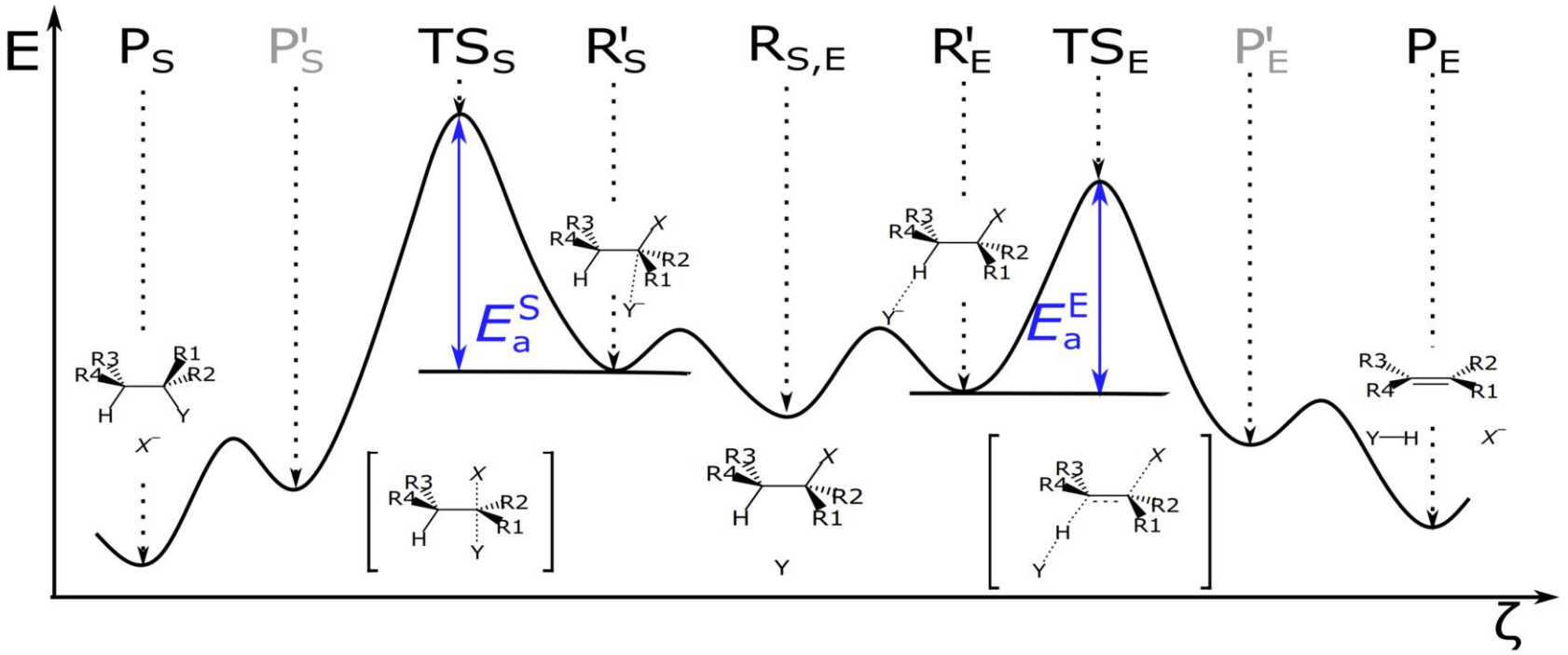
- Get i data points with scalar property (label) $\{q_i\}$
  - E.g. atomisation energy
- Calculate all representations $\{\mathbf{M}_i\}$
  - typically ~1k
- Find distance and kernel matrices $\mathbf{D}, \mathbf{K}$
  - Symmetric
- Train model for predictions $\{\tilde{q}_i\}$
- Find best hyperparameters (cross-validation)

$$\arg\min_{\alpha} \sum_i (q_i - \tilde{q}_i)^2 + \lambda \sum_{ij} \alpha_i \alpha_j k_{ij}$$

$$\Rightarrow \alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1}$$

$$\tilde{q}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$

Competing Reactions: E2 and S_N2

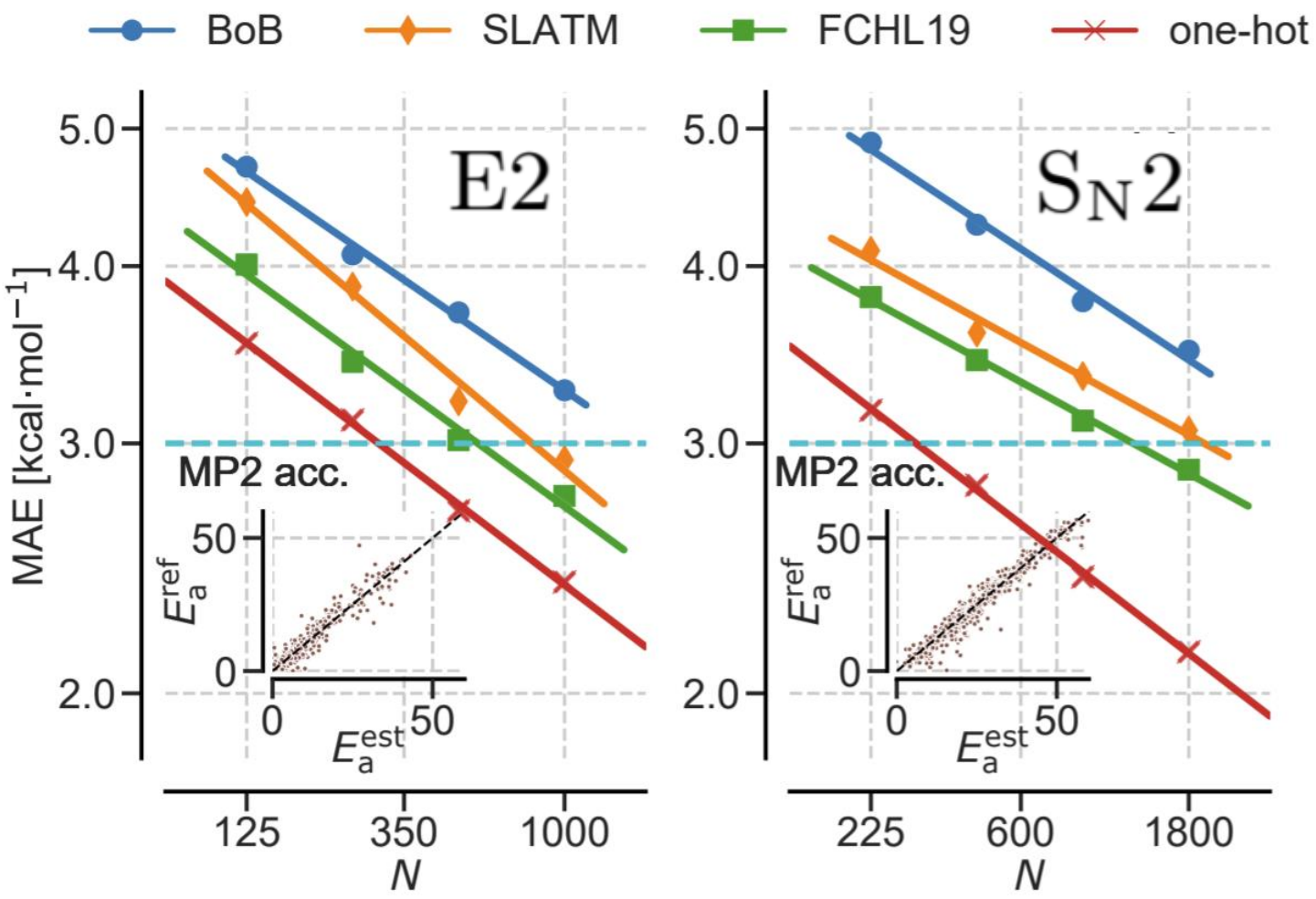


R	X	Y
H	F	H
NO_2	Cl	F
CN	Br	Cl
CH_3		Br
NH_2		

- Activation energies E_a
- Transition state geometries
- Dataset of 4.5k transition states, 143k reactant geometries, part MP2, part DF-LCCSD

GFvR, S. Heinen, M. Bragato, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* **2020** (*arXiv* 2006.00504).

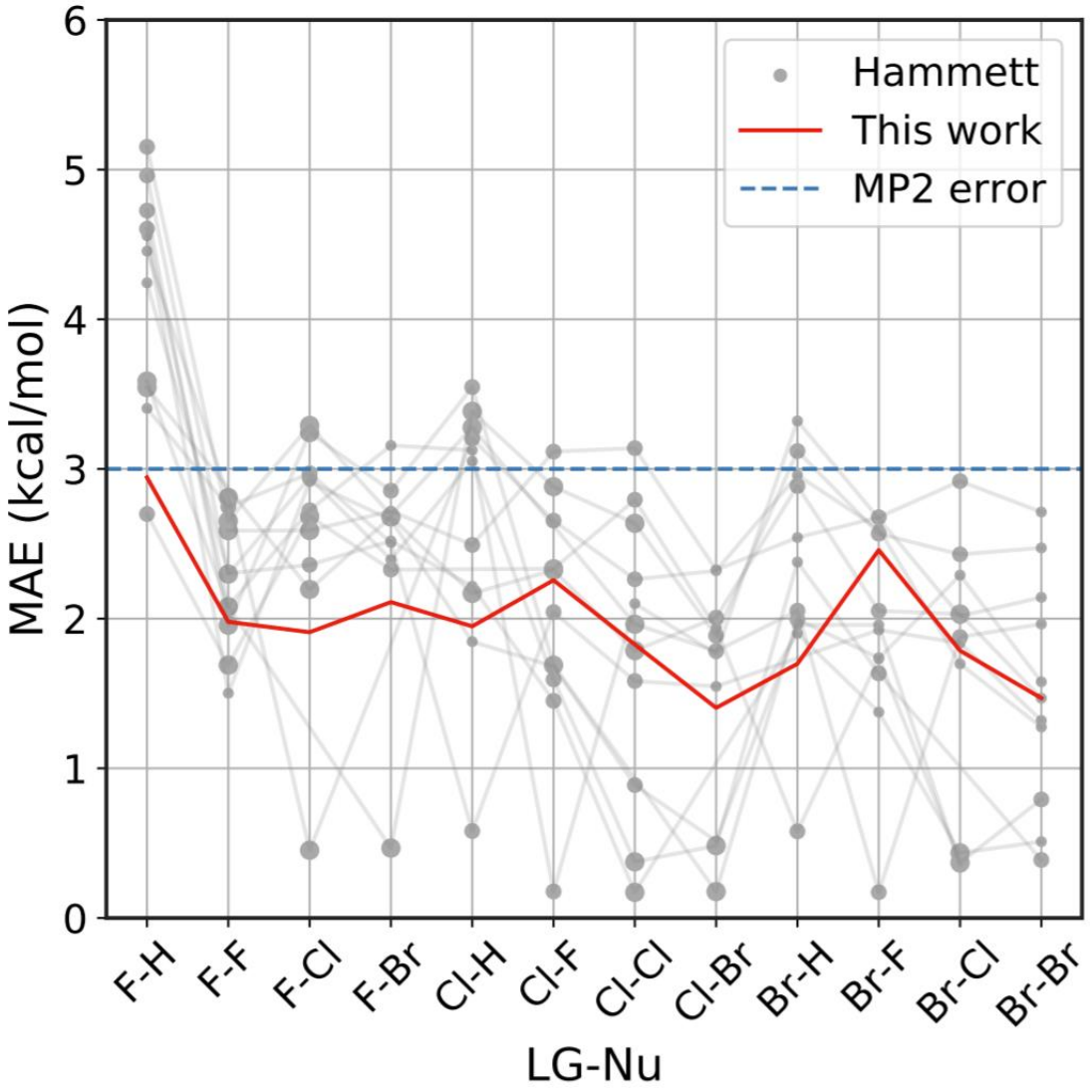
Learning Activation Energies



- Geometry-based representations
  - BoB
  - SLATM
  - FCHL19
- Graph-based representations
  - One-hot

GFvR, S. Heinen, O. A. von Lilienfeld, *arXiv* 2009.13429.

Learning Activation Energies



Hammett's equation (1935):

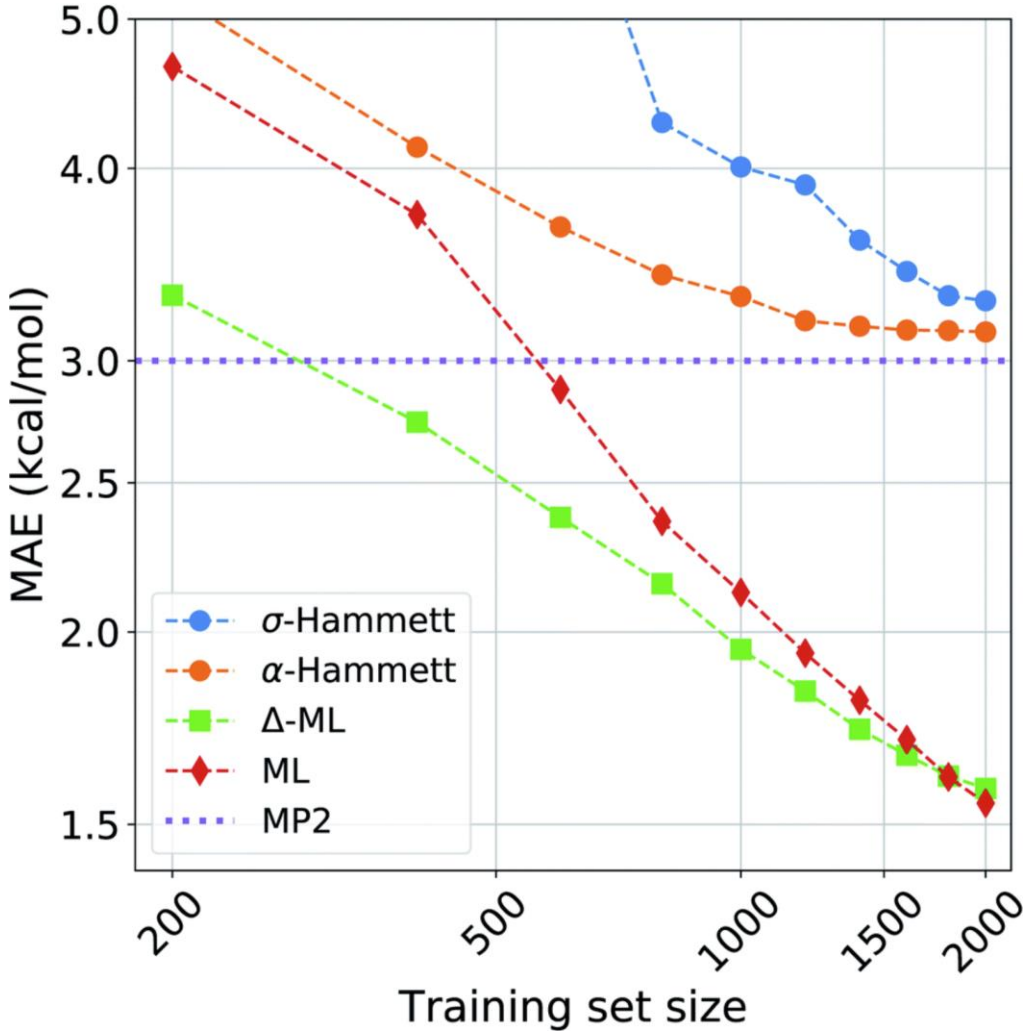
$$\log\left(\frac{K}{K_0}\right) \simeq \rho\sigma$$

Can be used to remove linear trends in the data

1. Find two aspects (e.g. solute/solvent) that are orthogonal and approximately balanced in the data set
2. Fit rho, sigma in a robust manner

M. Bragato, GFVR, O. A. von Lilienfeld, *Chem. Sci. 2020 (arXiv 2004.14946).*

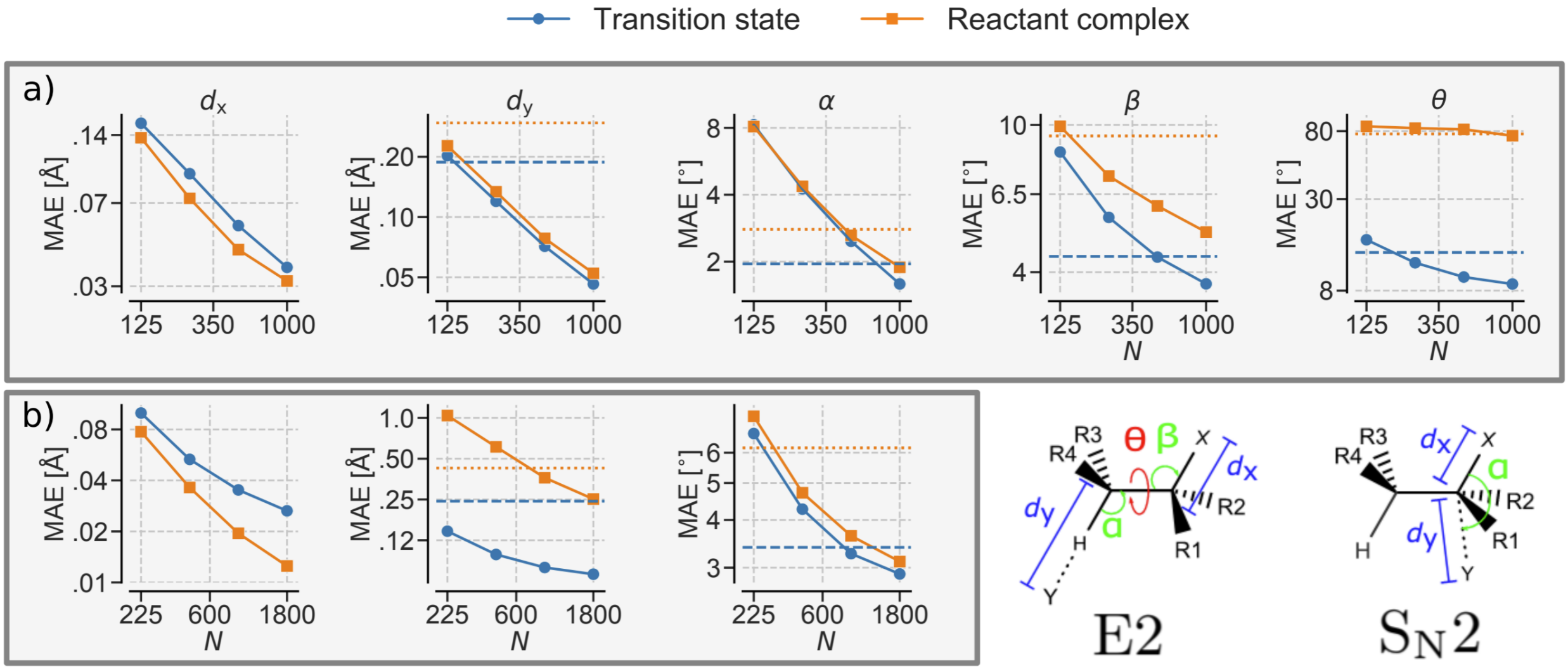
Learning Activation Energies



- Hammett nearly reaches MP2 accuracy
- Residuals are easier to learn
- Preprocessing of datasets most helpful for small training sets

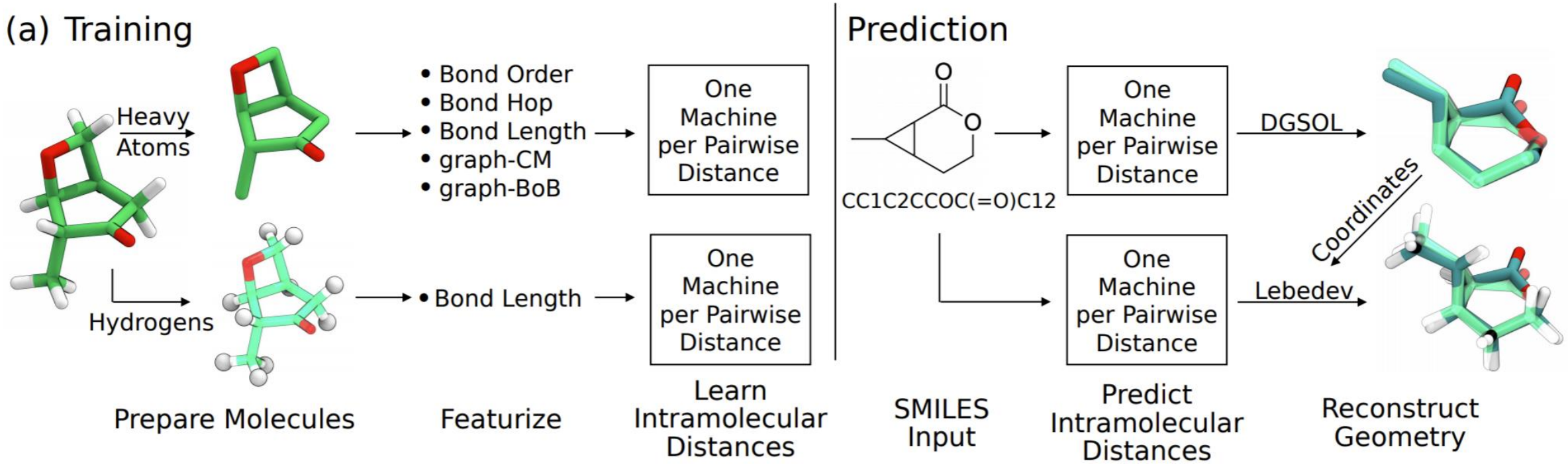
M. Bragato, GFVR, O. A. von Lilienfeld, *Chem. Sci. 2020 (arXiv 2004.14946).*

Learning Transition State Geometries



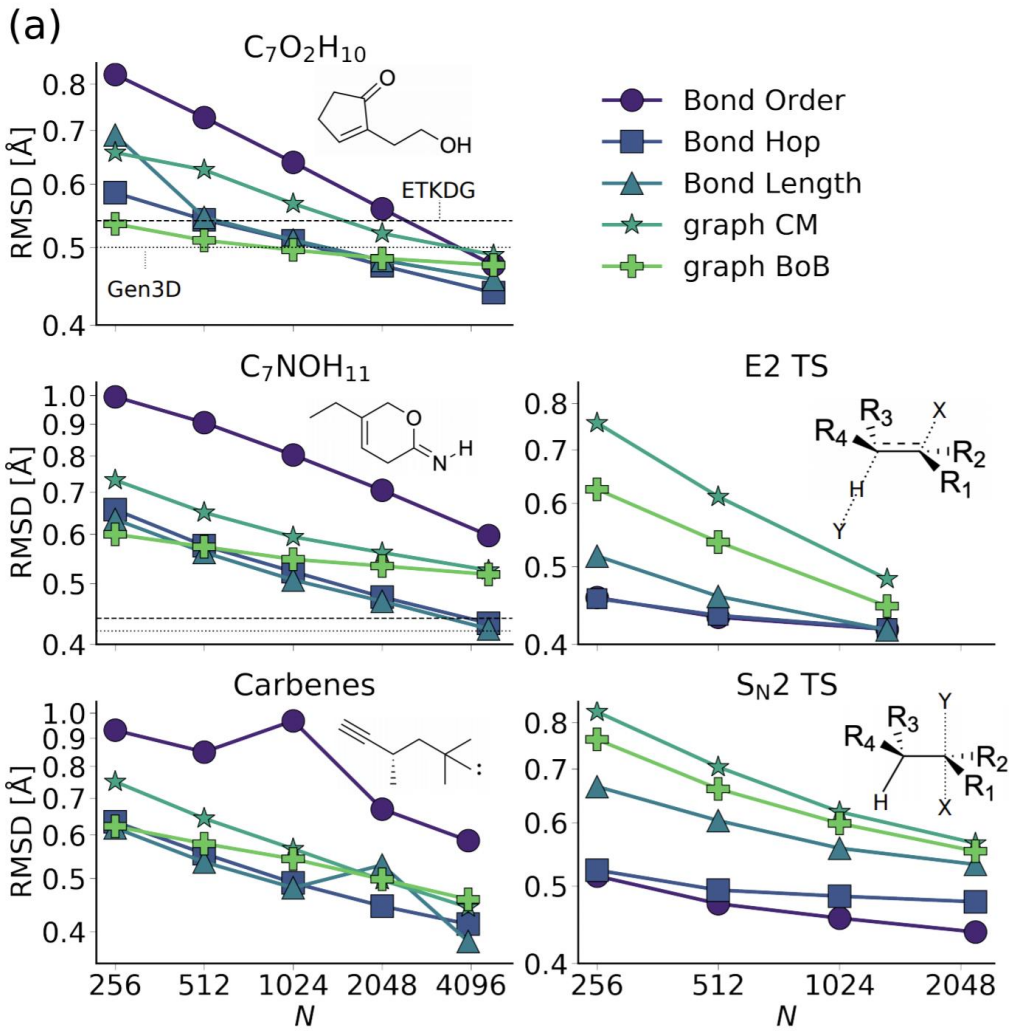
GFvR, S. Heinen, O. A. von Lilienfeld, *arXiv* 2009.13429.

Graph2Structure



D. Lemm, GFVR, O. A. von Lilienfeld, *arXiv* 2102.02806.

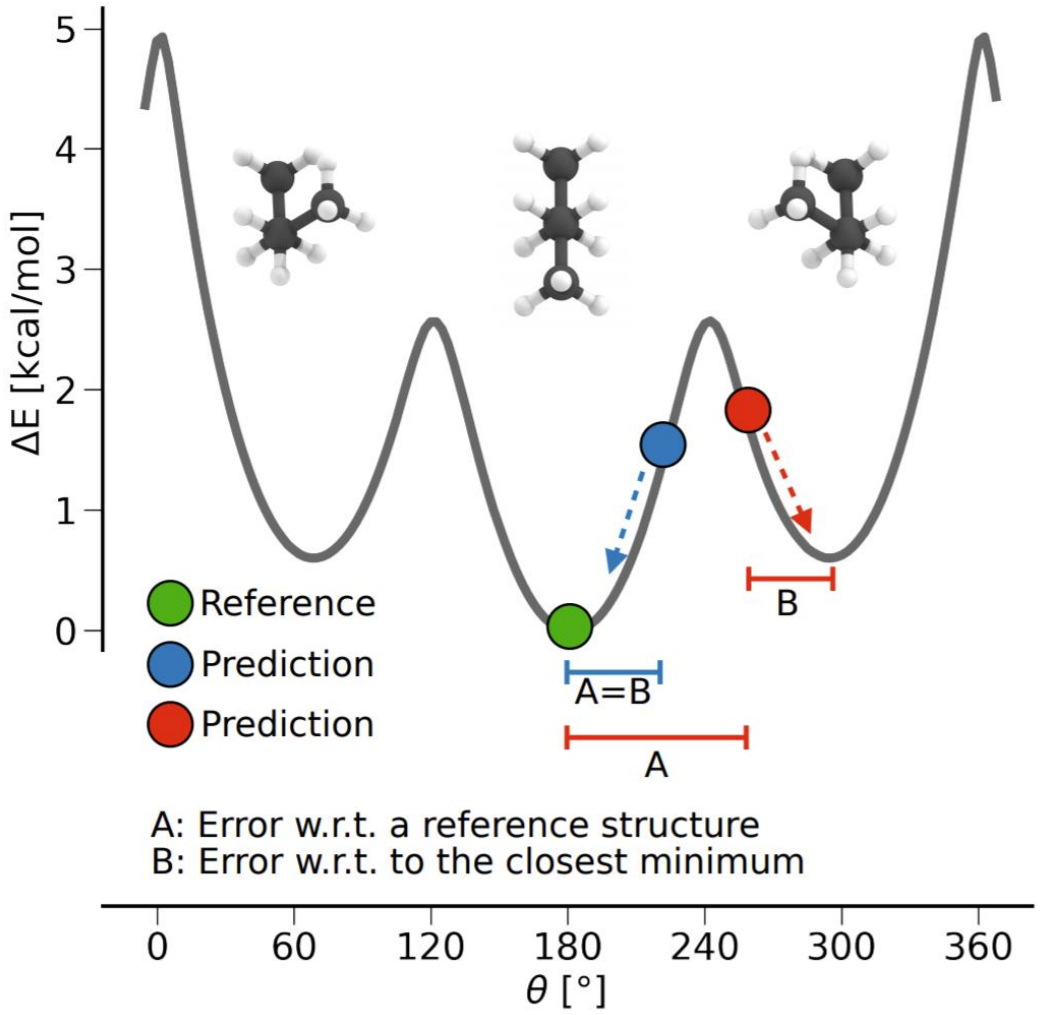
Graph2Structure



- Learns standard chemistry, but also carbenes, transition state geometries
- More accurate w.r.t. to QM calculations than state-of-the-art embedding methods (which only do standard chemistry)
- Can produce initial guesses for e.g. transition state searches

D. Lemm, GFVR, O. A. von Lilienfeld, *arXiv* 2102.02806.

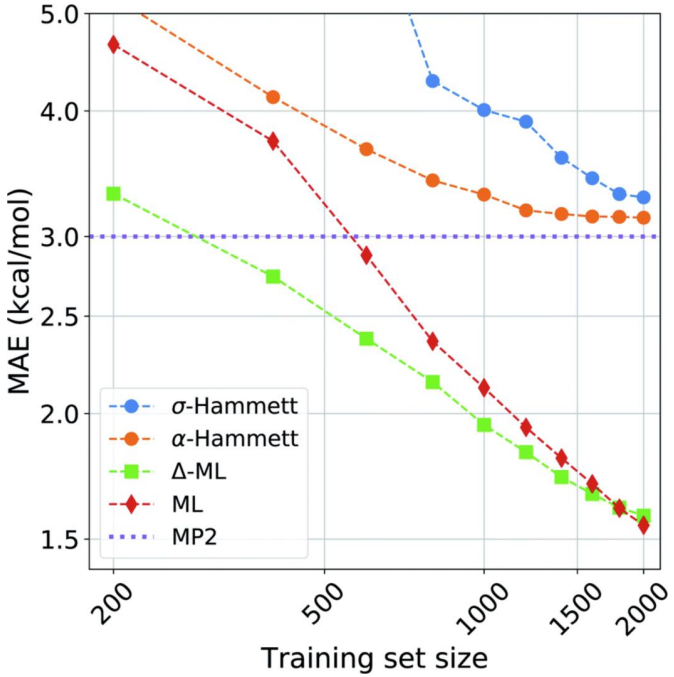
Graph2Structure



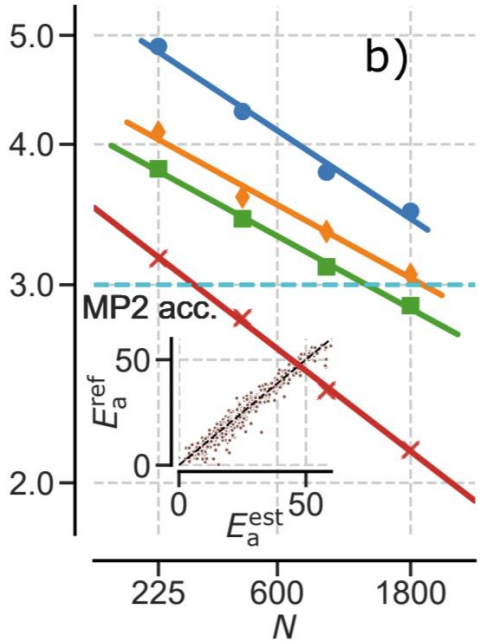
- Most efficient for cases with wide minima
- Multiple small minima e.g. stereoisomers account for most of the error

D. Lemm, GFVR, O. A. von Lilienfeld, *arXiv* 2102.02806.

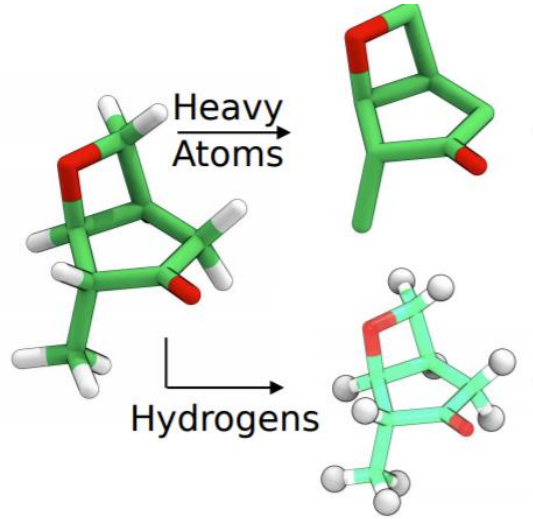
Summary Machine Learning



Detrending with Hammett's equation



Energies with Kernel-Ridge-Regression



Geometries with Graph2Structure

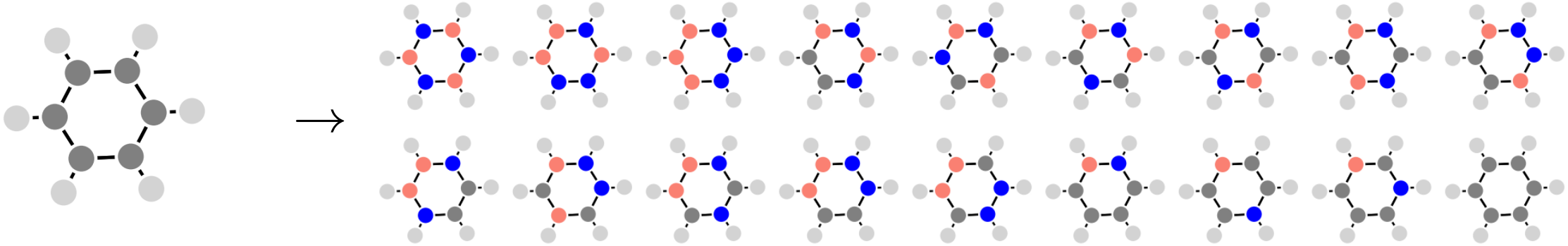
chemspacelab/Enhanced-Hamme

qmlcode/qml

Quantum Alchemy

Example

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



$$E, \rho, \{\partial^i_\lambda \rho\} \rightarrow \{E_i\}, \{\rho_i\}, \{F_i\}, \{\mu_i\}, \{Q_i\}, \ldots$$

1 system $\rightarrow$ Millions of systems

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

Theory

Interpolate between molecular Hamiltonians

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

Taylor expansion around reference molecule

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

Hellmann-Feynman theorem

$$\partial_\lambda E = \left\langle \psi_\lambda \middle| \hat{H}_\mathrm{t} - \hat{H}_\mathrm{r} \middle| \psi_\lambda \right\rangle = \Delta E^\mathrm{NN} + \int_\Omega d\mathbf{r} \underbrace{\left(v_\mathrm{t}(\mathbf{r}) - v_\mathrm{r}(\mathbf{r})\right)}_{\equiv \Delta v} \rho_\lambda(\mathbf{r})$$

O. A. von Lilienfeld, *J. Chem. Phys.* **2009**.

Alchemical Perturbation Density Functional Theory (APDFT)

$$E_\mathrm{t} = E_\mathrm{r} + \Delta E^\mathrm{NN} + \int_\Omega d\mathbf{r} \sum_{n=0}^{\infty} \frac{1}{(n+1)!} \Delta v \frac{\partial^n \rho_\lambda(\mathbf{r})}{\partial \lambda^n}\bigg|_{\lambda=0}$$

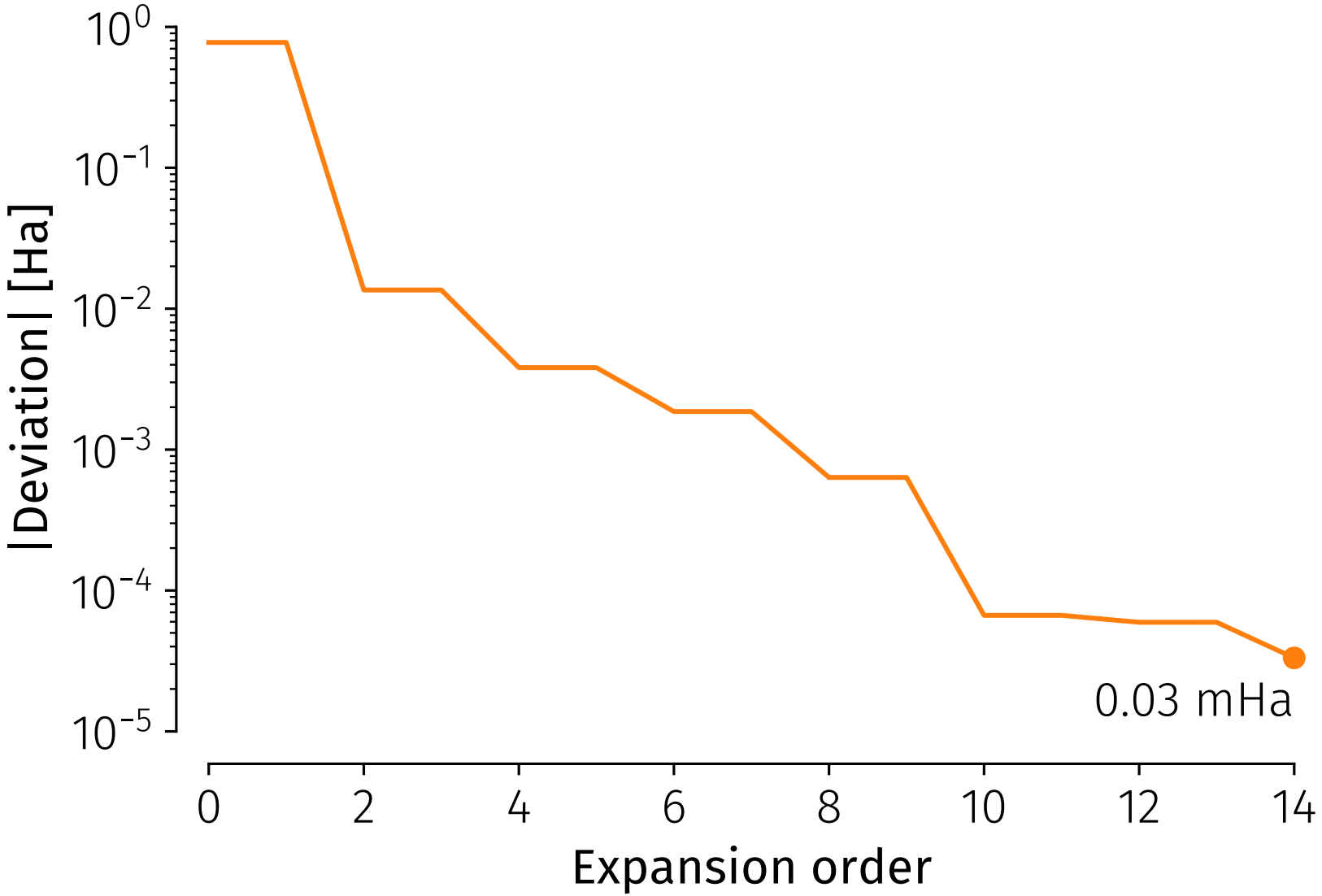
$$\rho_t = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n \rho}{\partial \lambda^n}\bigg|_{\lambda=0}$$

- Gives consistent energies, densities, forces, …
- Uses the same derivatives for all predictions
- In practice: truncate after some order *n*

ferchault/APDFT

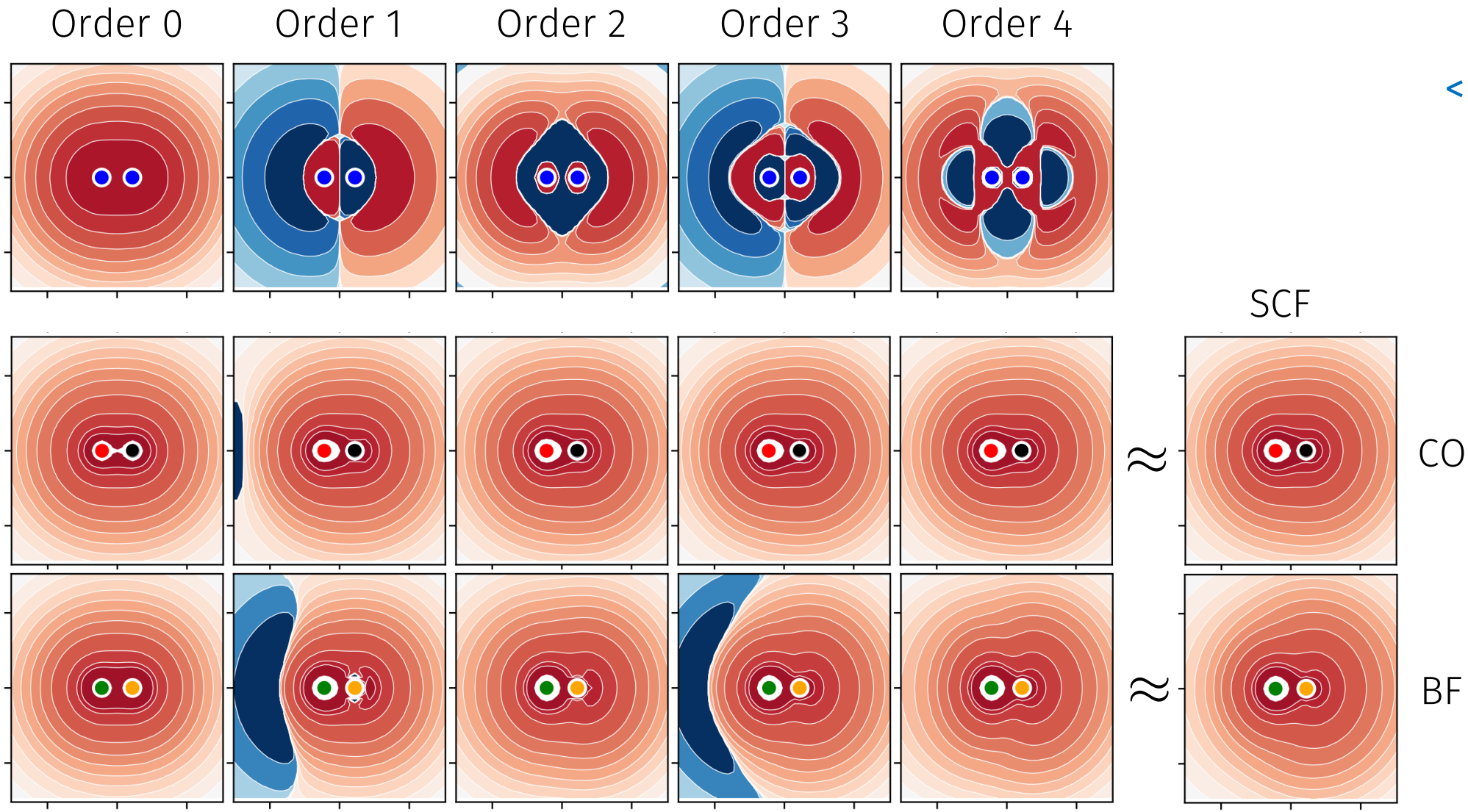
GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.* **2020** (*arXiv* 1809.01647).

Analytical convergence: Energy



< H_2 to He
HF / def2-TZVP
DiffiQult, algopy

Numerical convergence: Density

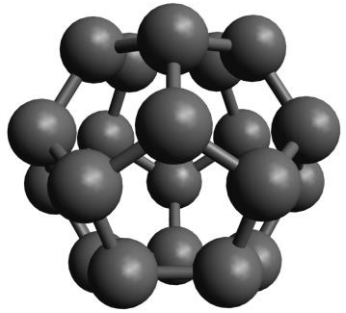


< N_2 to CO, BF
def2-TZVP
Horton

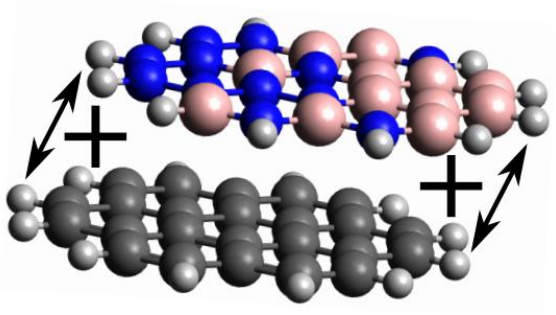
GFvR, O. A. von Lilienfeld, *Phys. Rev. Res.* **2020**, accepted (*arXiv* 1809.01647).

Applications

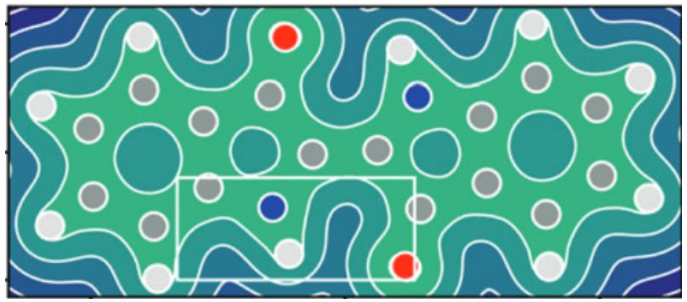
Covalent Energies



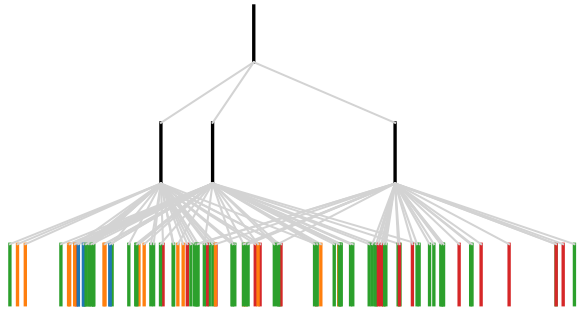
Non-covalent Interactions



Deprotonation Energies

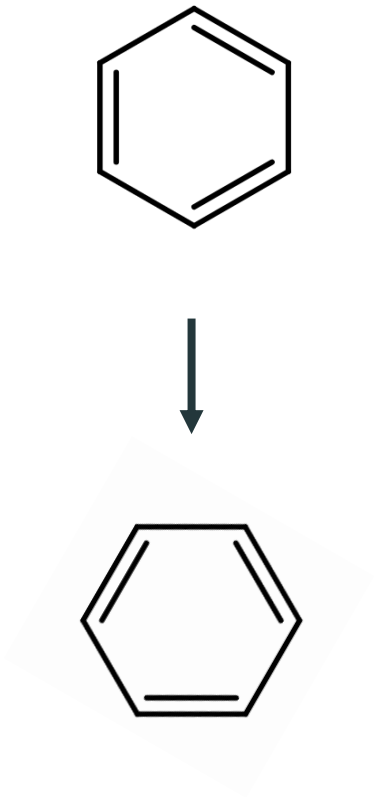


Energy Decomposition

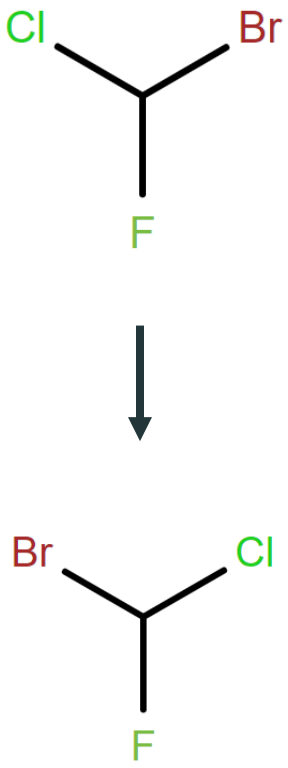


Symmetries to exploit

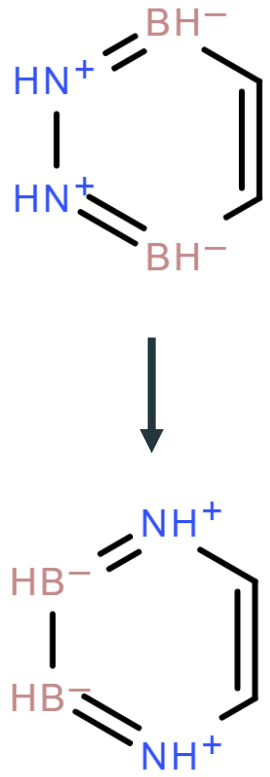
Translation
Exact



Rotation
Exact



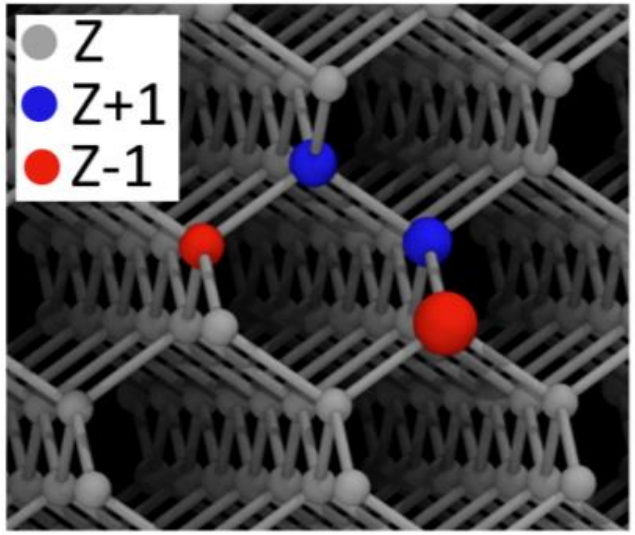
Spatial reflection
Approximate

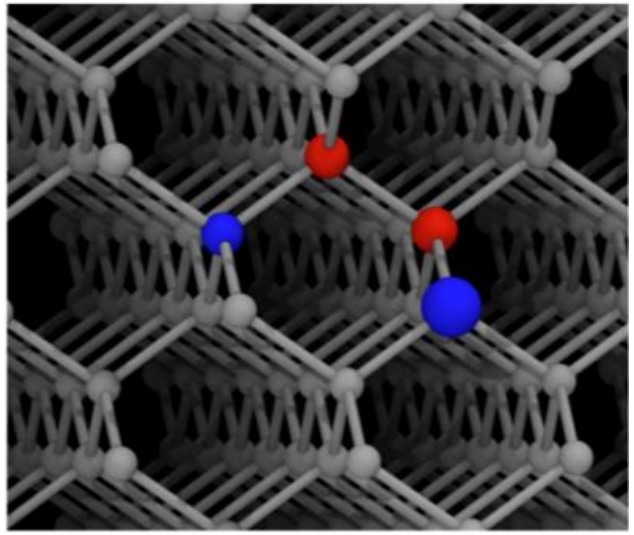


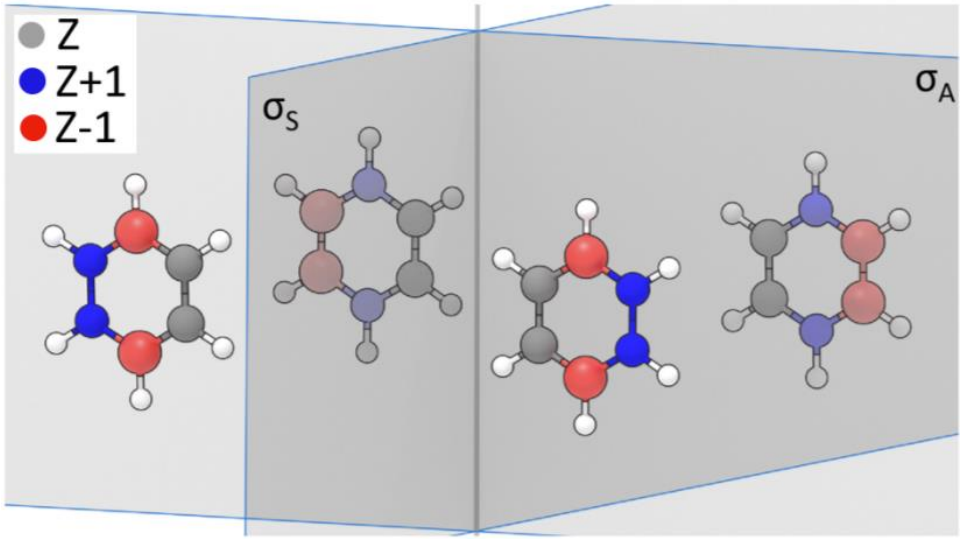
Alchemical reflection
More approximate

GFvR, O. A. von Lilienfeld, *Science Adv.* 2021 (arXiv 2008.02784).

Symmetries to exploit







$$E_\mathrm{t} = E_\mathrm{r} + \Delta E^\mathrm{NN} + \int_\Omega d\mathbf{r} \sum_{n=0}^{\infty} \frac{1}{(n+1)!}\, \Delta v \frac{\partial^n \rho_\lambda(\mathbf{r})}{\partial \lambda^n}\bigg|_{\lambda=0}$$

GFvR, O. A. von Lilienfeld, *Science Adv.* 2021 (arXiv 2008.02784).

Theory

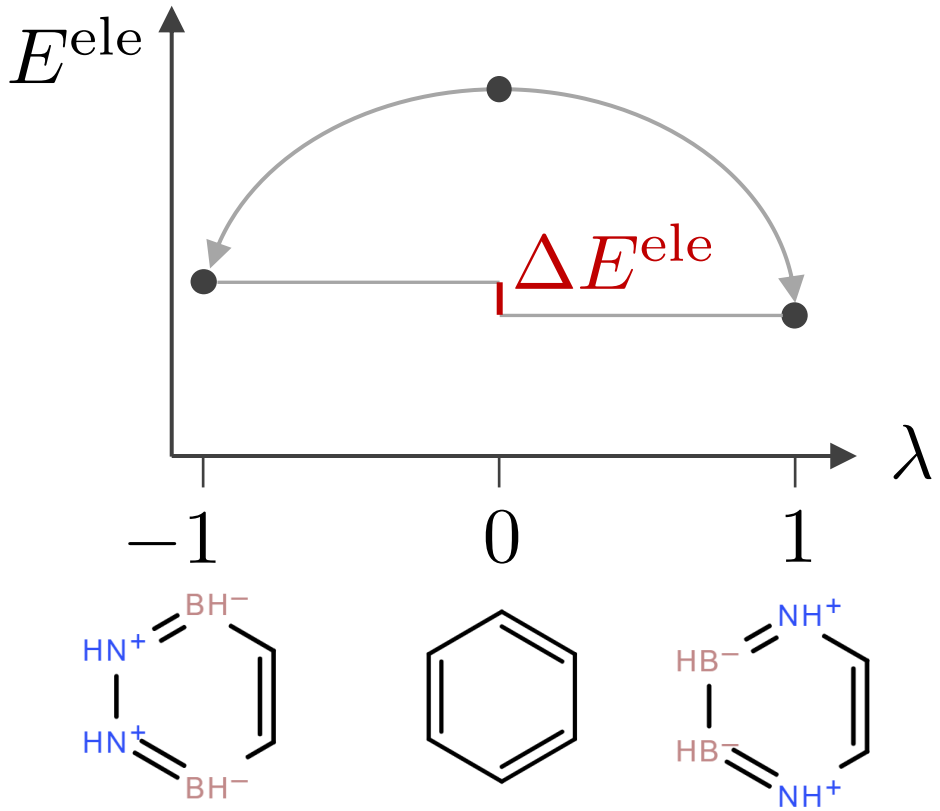
$$\Delta E_{ij}^{\mathrm{ele}} = E_0 - E_0 + \int_\Omega d\mathbf{r} \sum_{n=0}^{\infty} \frac{\Delta v_i}{(n+1)!} \left[\frac{\partial^n \rho}{\partial \lambda_i^n} + \frac{\partial^n \rho}{\partial \lambda_j^n} \right]$$

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

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

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

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

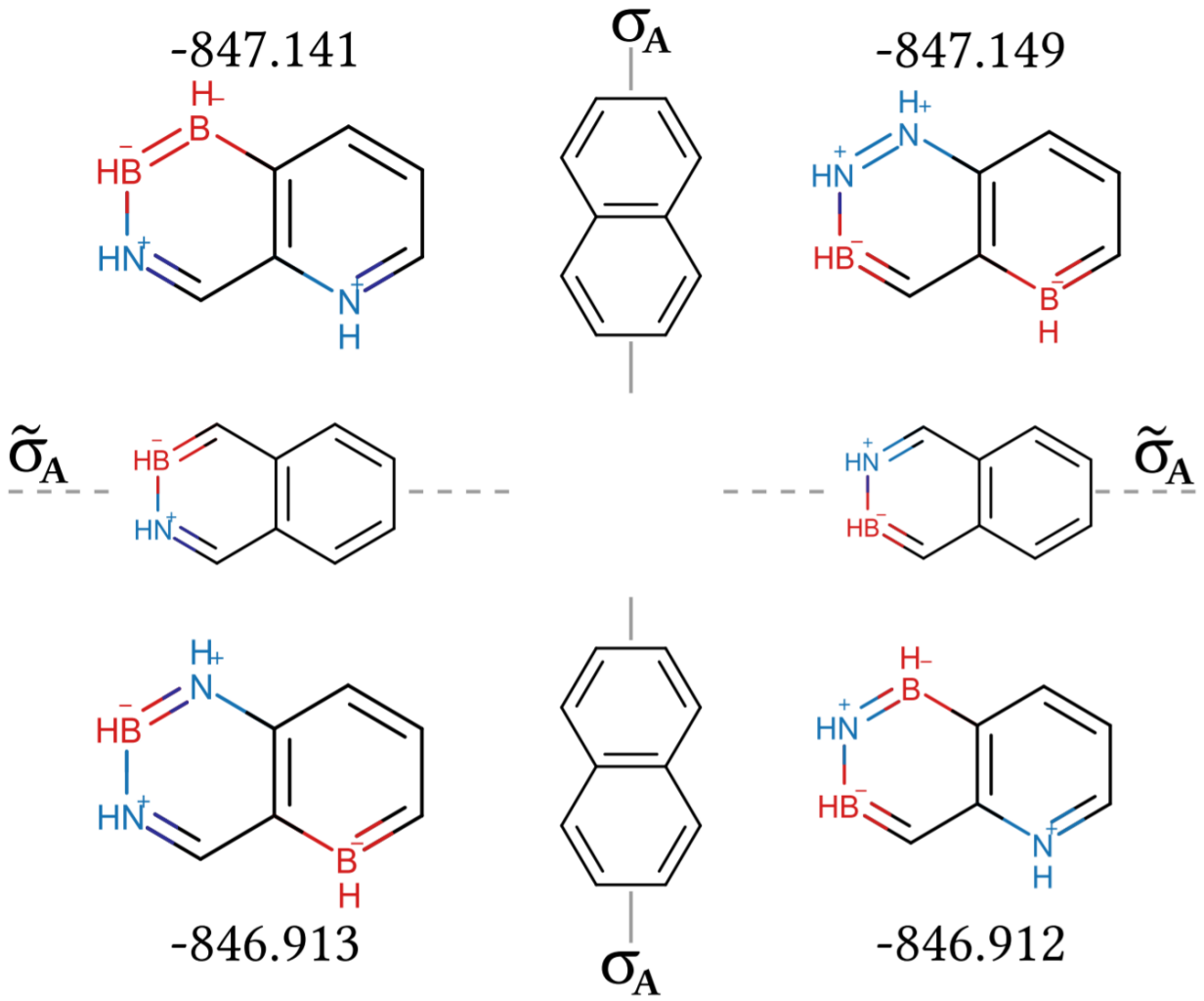


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021 (arXiv 2008.02784).

Alchemical symmetry

Approximate alchemical enantiomers are
- two spatially non-superimposable,
- alchemically coupled,
- and iso-electronic compounds with the same formal charge,
- where each transmutating atom is assigned to exactly one subset within each of which averaging of nuclear charges results in *nearly identical* chemical environments.

BN-doped naphthalene >
CCSD/cc-pVDZ
Molpro



GFvR, O. A. von Lilienfeld, *Science Adv.* 2021 (arXiv 2008.02784).

Summary

Global Search Problem
Which class of compounds?

Machine Learning as global exploration method

Local Search Problem
Which particular species within that class?

Quantum Alchemy to reduce the search space

Prof. von Lilienfeld

Dominik Lemm

Marco Bragato

Stefan Heinen

ferchault @ferchault guido.vonrudorff.de