Learning Reaction Barriers And Transition State Geometries

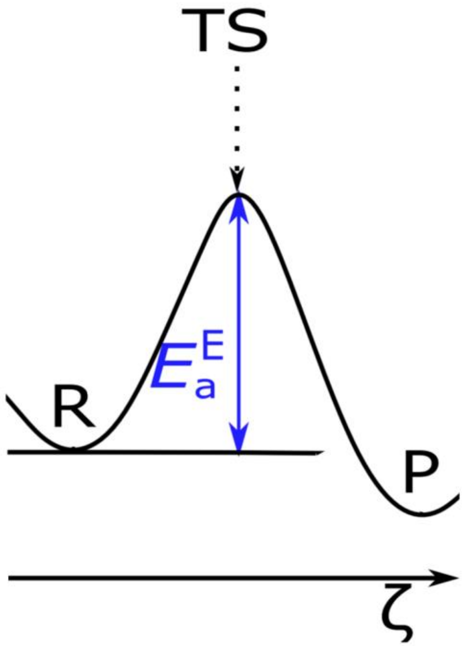
Guido Falk von Rudorff, University of Vienna

ferchault @ferchault guido.vonrudorff.de

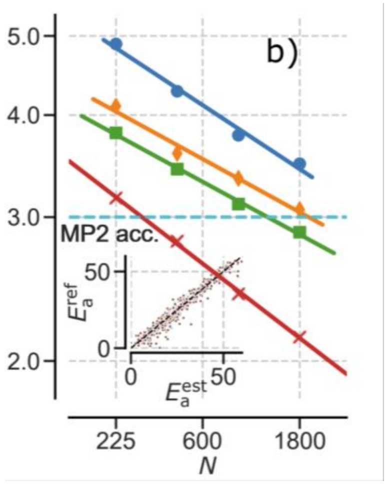
Introduction

- Reactions: complicated landscape
- Not only expensive but also hard problem
- Even if the reaction mechanism is known:
  - Find reactant (R)/product (P) complexes
  - Find transition state (TS) geometries
  - Describe energy near TS
  - Low level of automation available

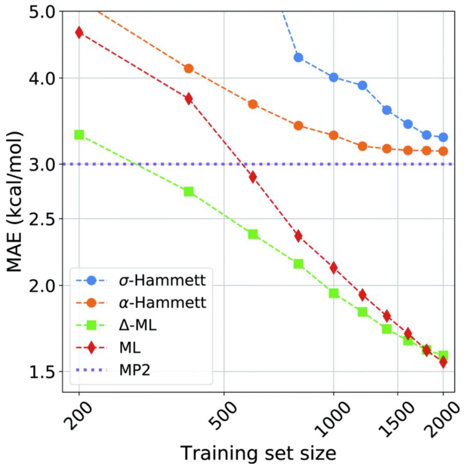
- Machine learning: accelerate & less supervised



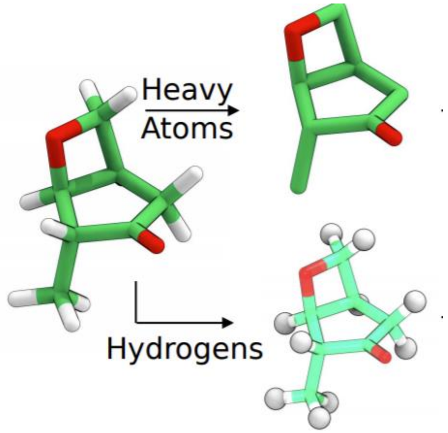
Overview



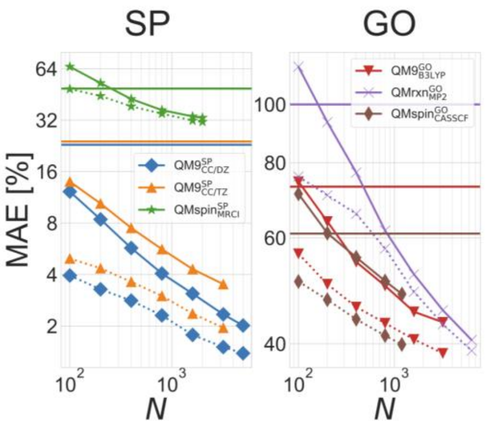
Energies with KRR



Detrending with Hammett’s equation



Geometries with Graph2Structure



Estimate computational cost

qmlcode/qml

chemspacelab/ Enhanced-Hammett

qmlcode/qml

ferchault/mlscheduling

Kernel Ridge Regression

Idea

- Molecular representation for each molecule *i*
  - CM, BoB, FCHL, SLATM, …
- Distance metric
  - Typically L1 or L2 norm
- Kernel function
  - Laplacian, Gaussian

$$\mathbf{M}_i$$

$$d_{ij} \equiv d(\mathbf{M}_i, \mathbf{M}_j)$$

$$k_{ij} \equiv k(d_{ij})$$

Kernel Ridge Regression

Procedure

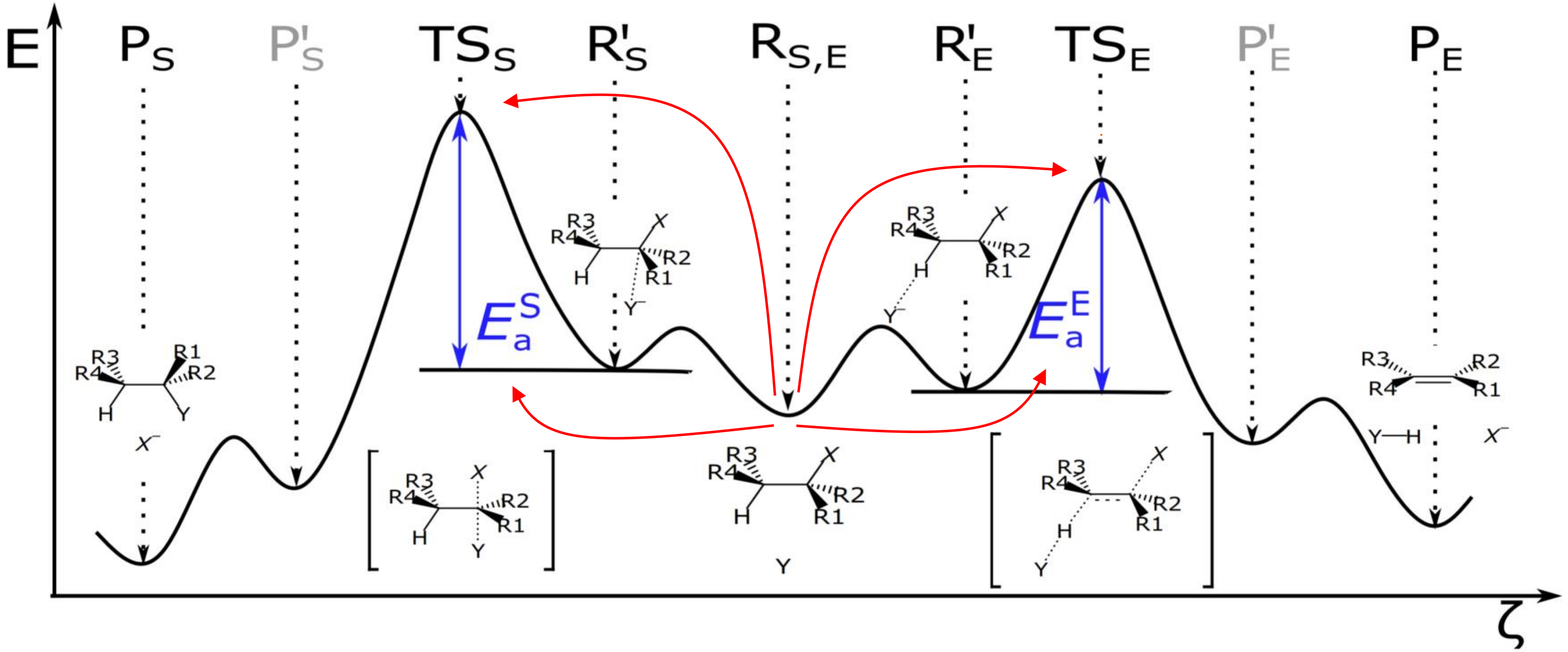
- Get *i* data points with scalar property (label) $\{q_i\}$
  - E.g. atomisation energy
- Calculate all representations $\{\mathbf{M}_i\}$
  - typically ~1k
- Find distance and kernel matrices $\mathbf{D}, \mathbf{K}$
  - Symmetric
- Train model for predictions $\{\tilde{q}_i\}$
- Find best hyperparameters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

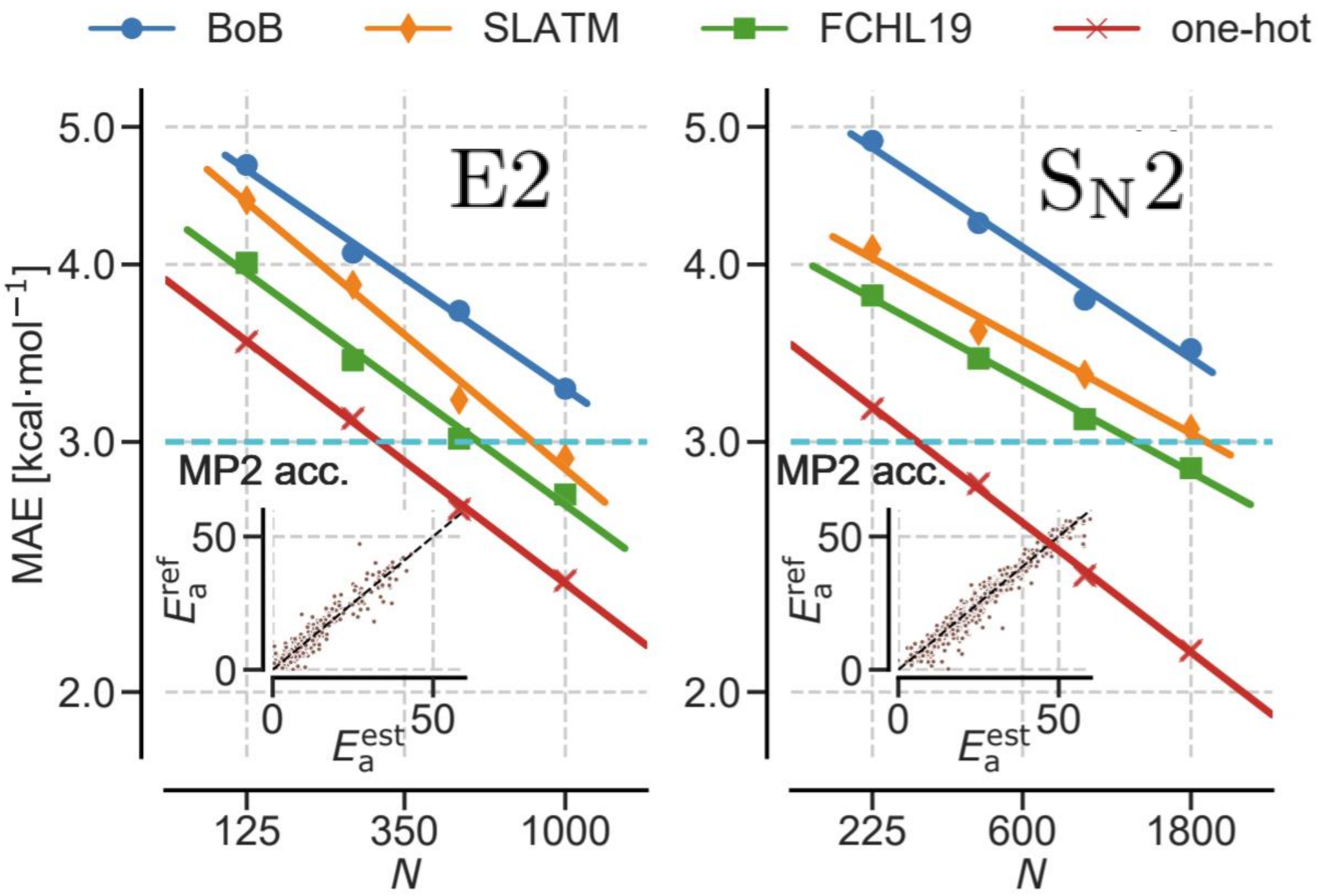


R	X	Y
H	F	H
NO_2	Cl	F
CN	Br	Cl
CH_3		Br
NH_2		

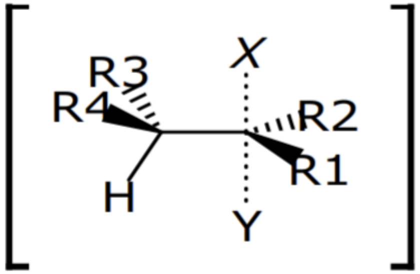
- Activation energies E_a
- Transition state geometries
- Dataset of 4.5k transition states, 143k reactant geometries, part MP2, part DF-LCCSD

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

Learning Activation Energies



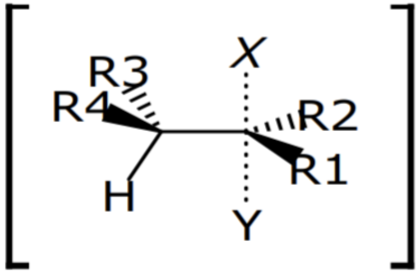
- Geometry-based representations on lowest conformer
  - BoB
  - SLATM
  - FCHL19
- Graph-based representations
  - One-hot

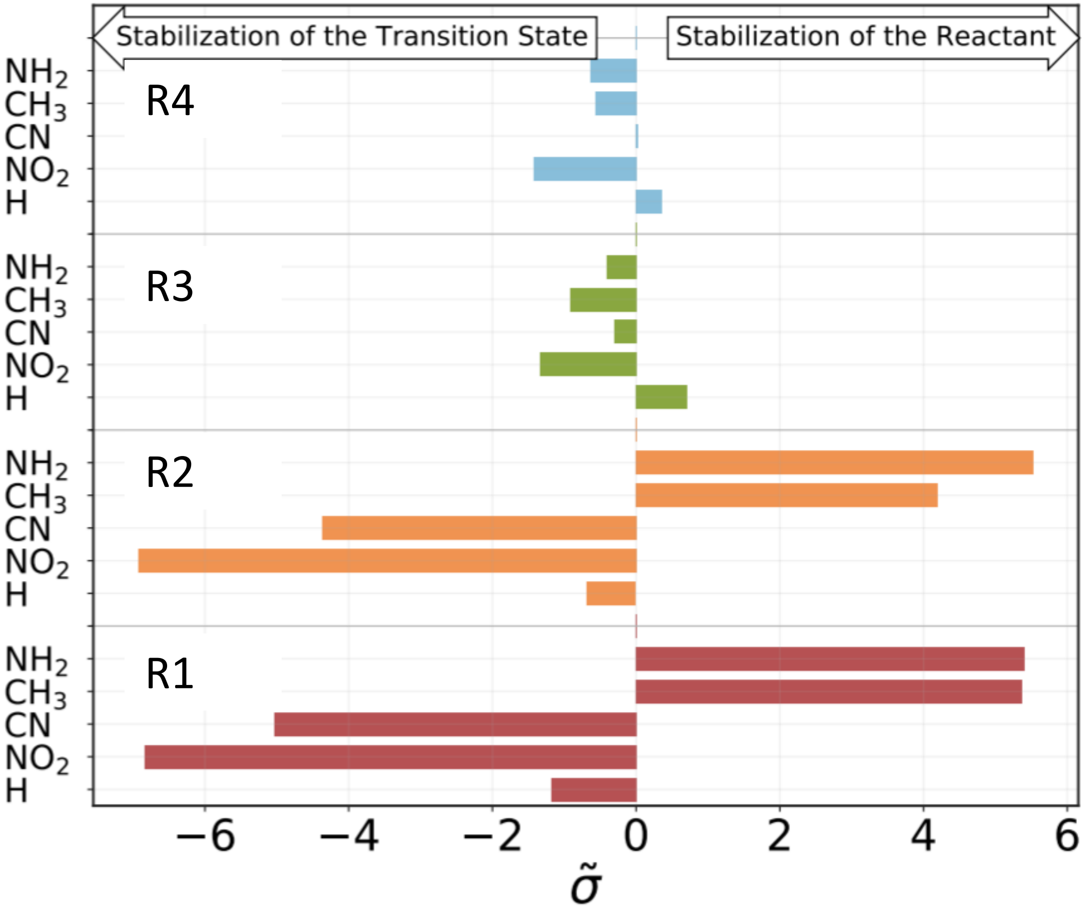


S. Heinen, GFvR, O. A. von Lilienfeld, *arXiv* 2009.13429.

Learning Activation Energies

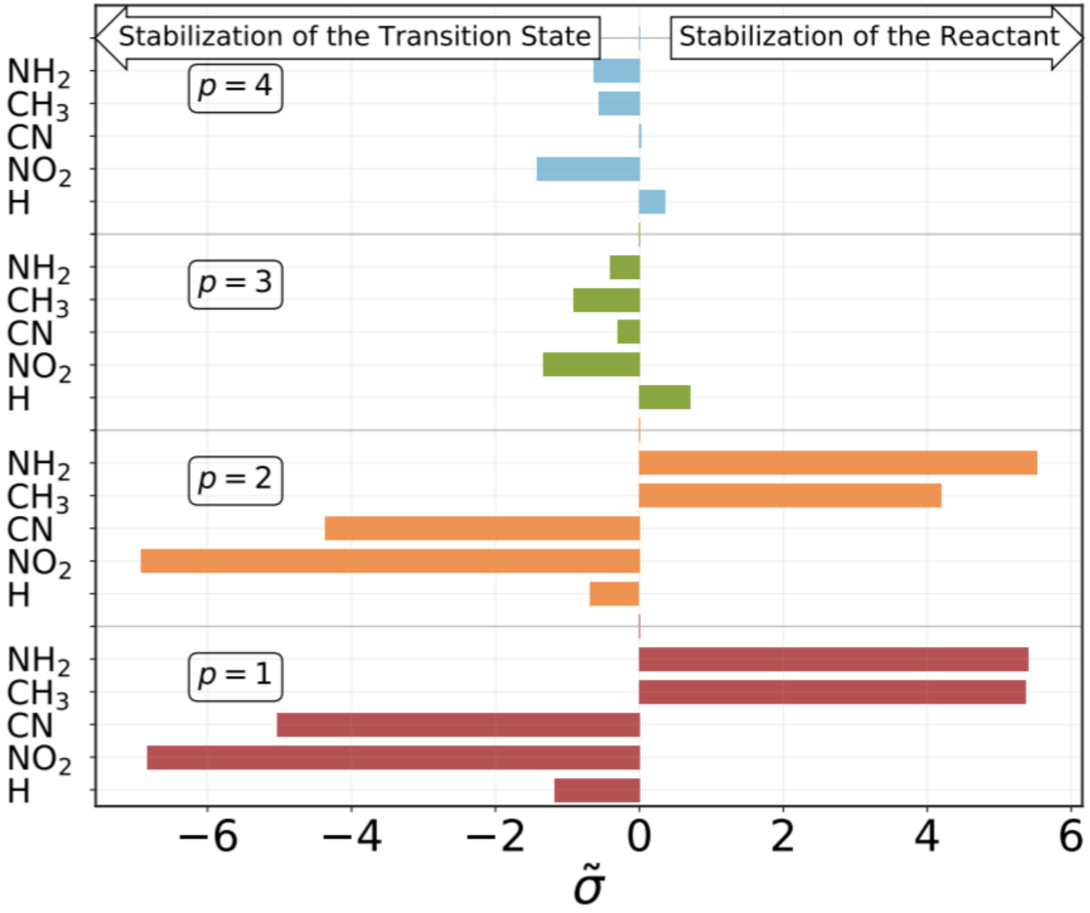
- Often in chemistry: trends obscure relevant detail
  - Electron density dominated by individual atoms
  - Energies dominated by elemental composition
  - Bond energies dominated by element pairs
  - …





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

Learning Activation Energies



Hammett's equation (1935):

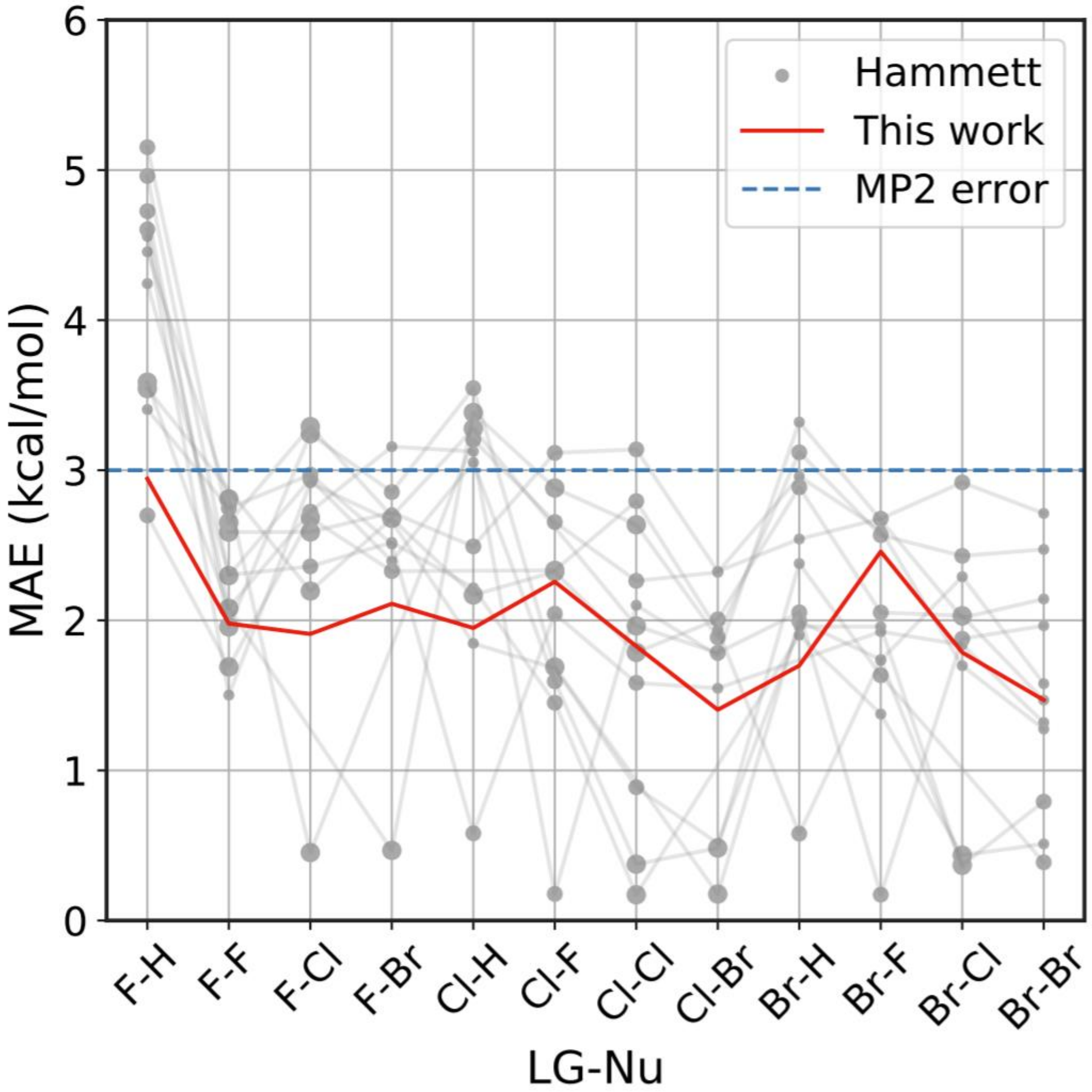
$$\log\left(\frac{K}{K_0}\right) \simeq \rho\sigma$$

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's equation (1935):

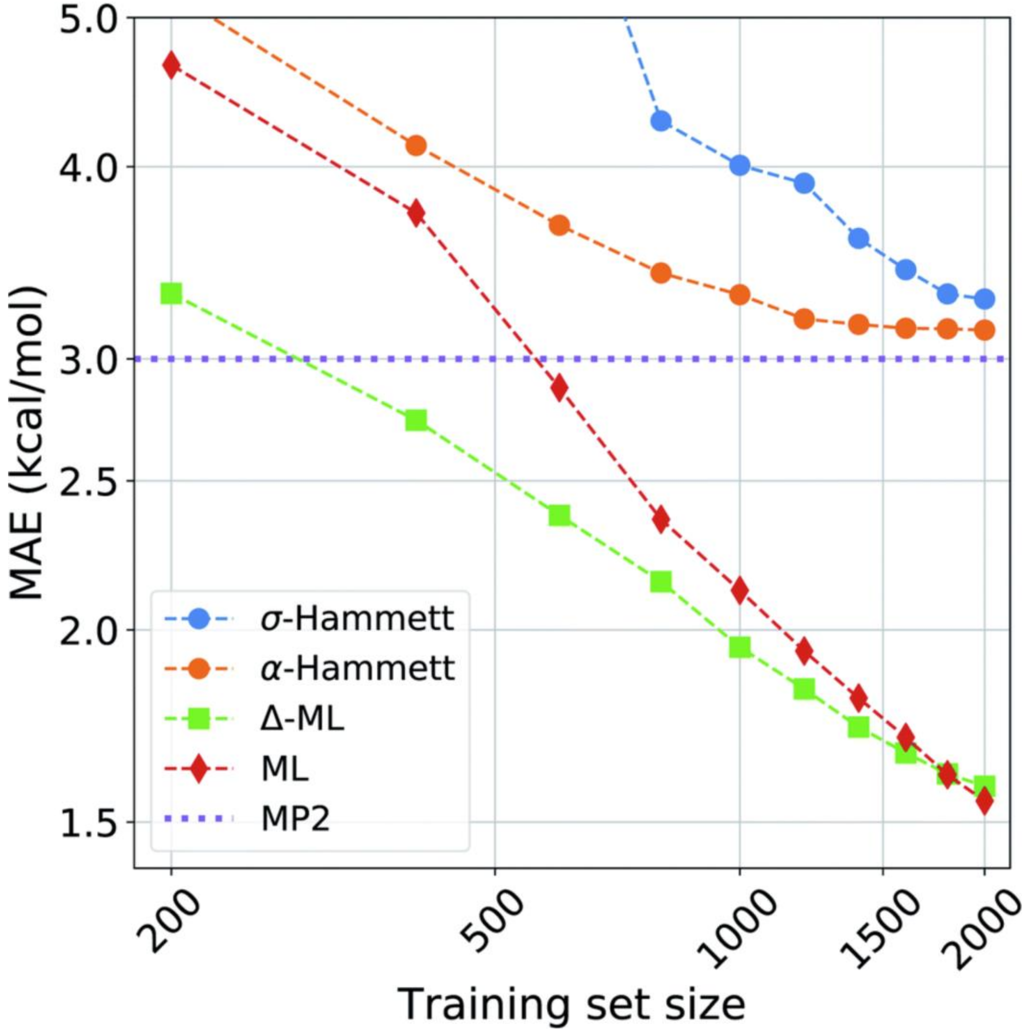
$$\log\left(\frac{K}{K_0}\right) \simeq \rho\sigma$$

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

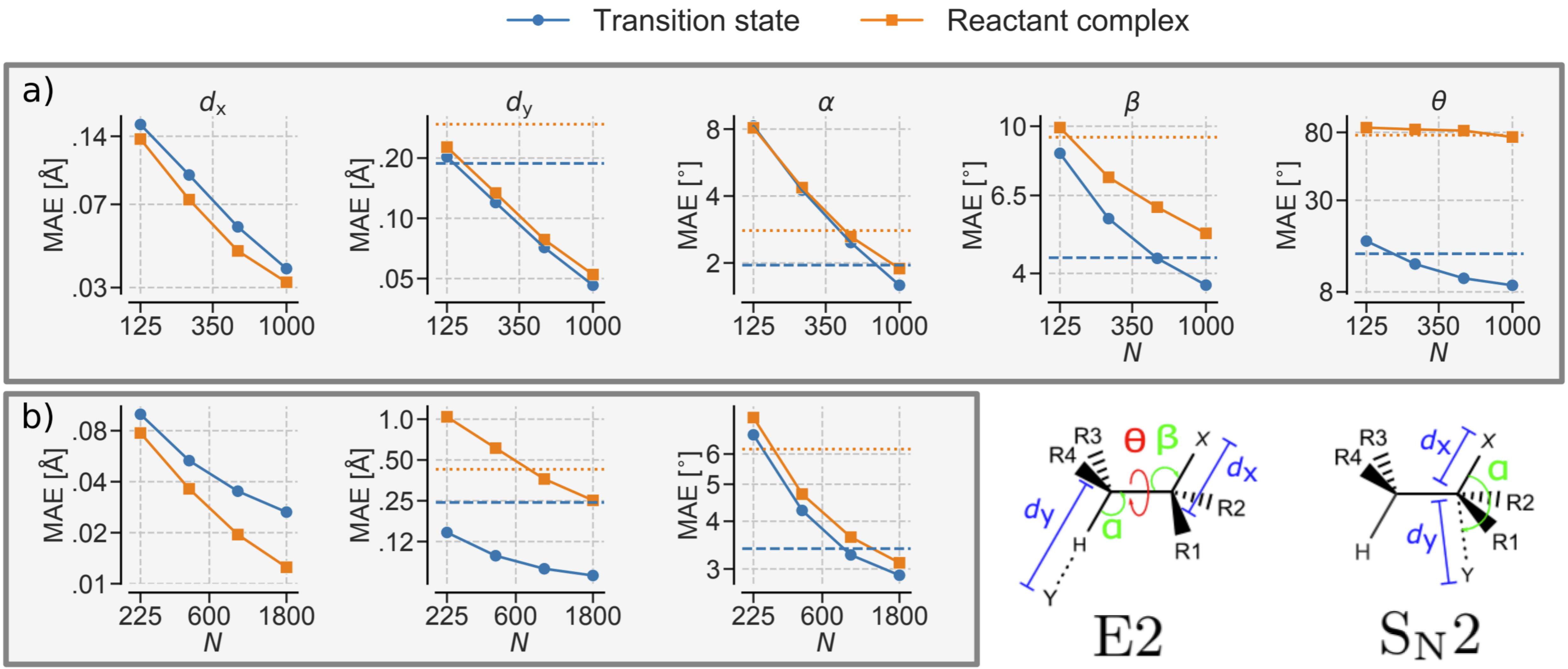


- Hammett nearly reaches MP2 accuracy
- Residuals are easier to learn
- Preprocessing of datasets most helpful for small training sets

chemspacelab/Enhanced-Hammett

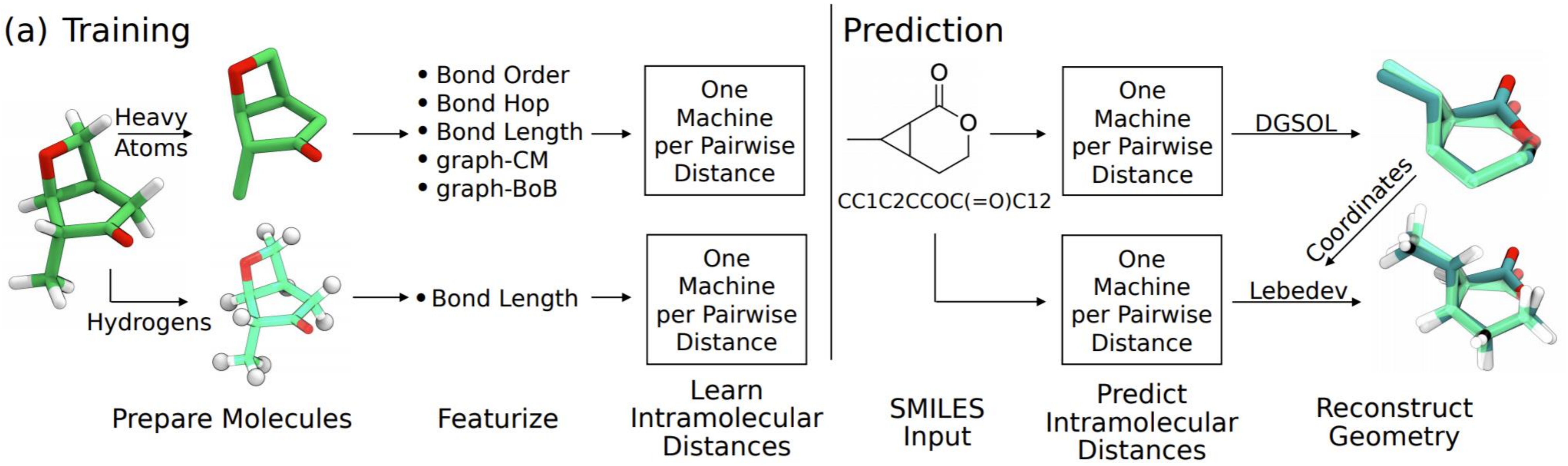
M. Bragato, GFVR, O. A. von Lilienfeld, *Chem. Sci. 2020 (arXiv 2004.14946).*

Learning Transition State Geometries



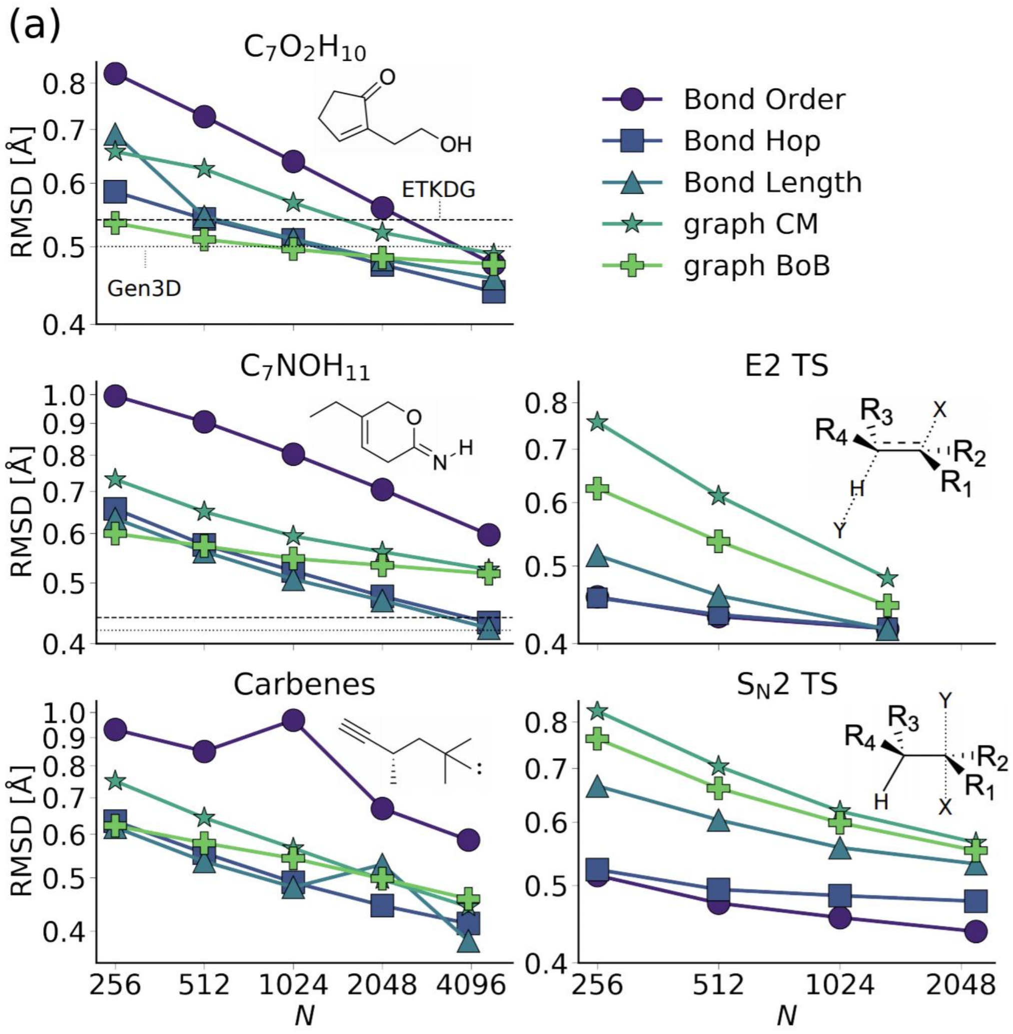
S. Heinen, GFvR, O. A. von Lilienfeld, *arXiv* 2009.13429.

Graph2Structure



D. Lemm, GFVR, O. A. von Lilienfeld, *Nat. Commun.* (accepted) *arXiv* 2102.02806.

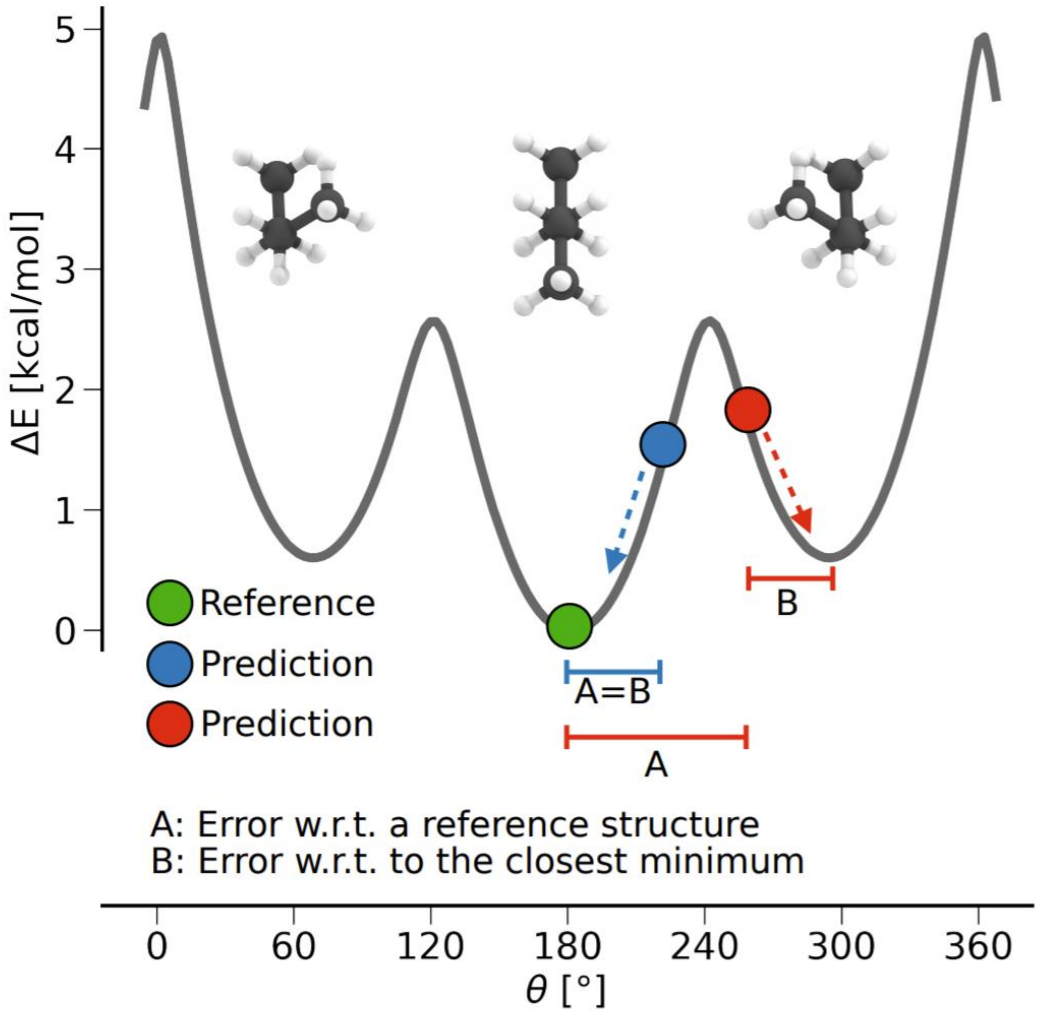
Graph2Structure



- 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, *Nat. Commun.* (accepted) *arXiv* 2102.02806.

Graph2Structure



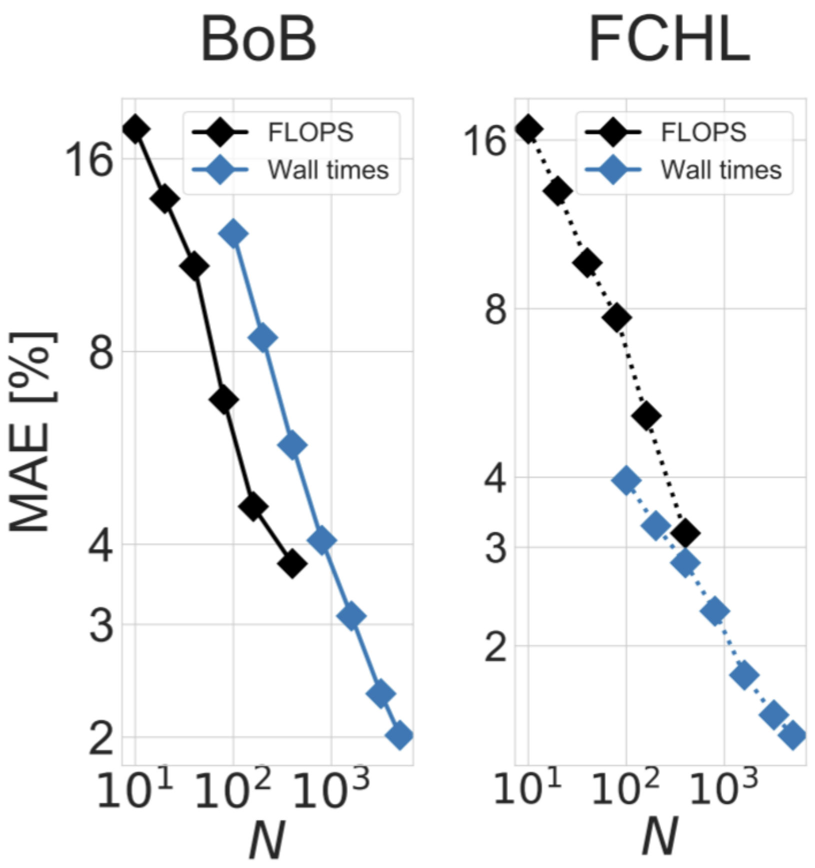
- Most efficient for cases with wide minima
- Multiple small minima e.g. stereoisomers account for most of the error

qmlcode/qml

D. Lemm, GFVR, O. A. von Lilienfeld, *Nat. Commun.* (accepted) *arXiv* 2102.02806.

Learning Computational Cost

- Training data expensive
- Not all training points equally expensive
  - Geometry optimisations may take longer
  - SCF might not converge or converge more slow
- Treating this as “molecular property”

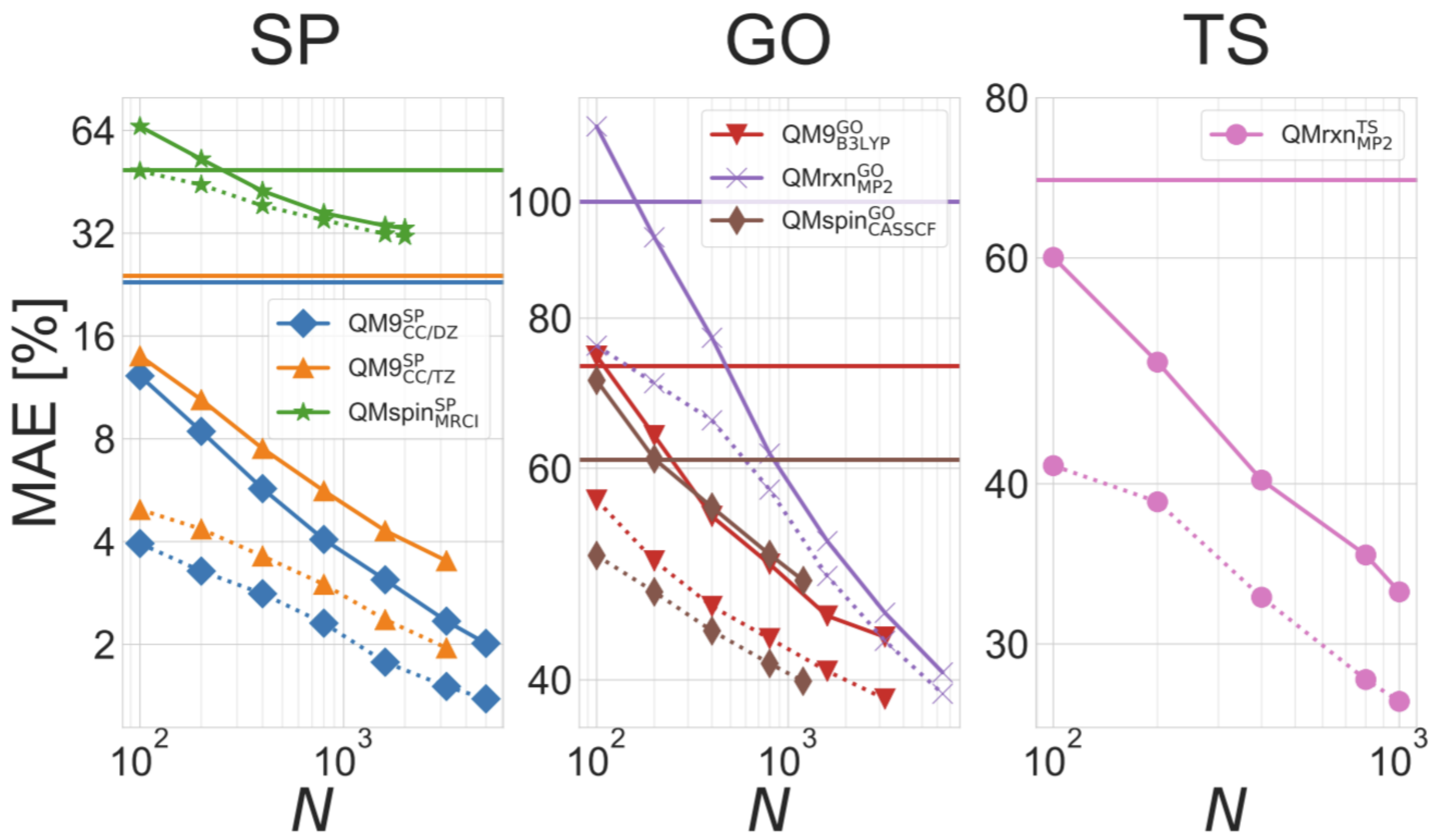


- Best case: controlled environment, single points (QM9)

S. Heinen, M. Schwilk, GFvR, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* **2020** (*arXiv* 1908.06714).

Learning Computational Cost

- Realistic case: I/O, noise: different machines



ferchault/mlscheduling

S. Heinen, M. Schwilk, GFvR, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* **2020** (*arXiv* 1908.06714).

Model availability

chemspacelab/Enhanced-Hammett

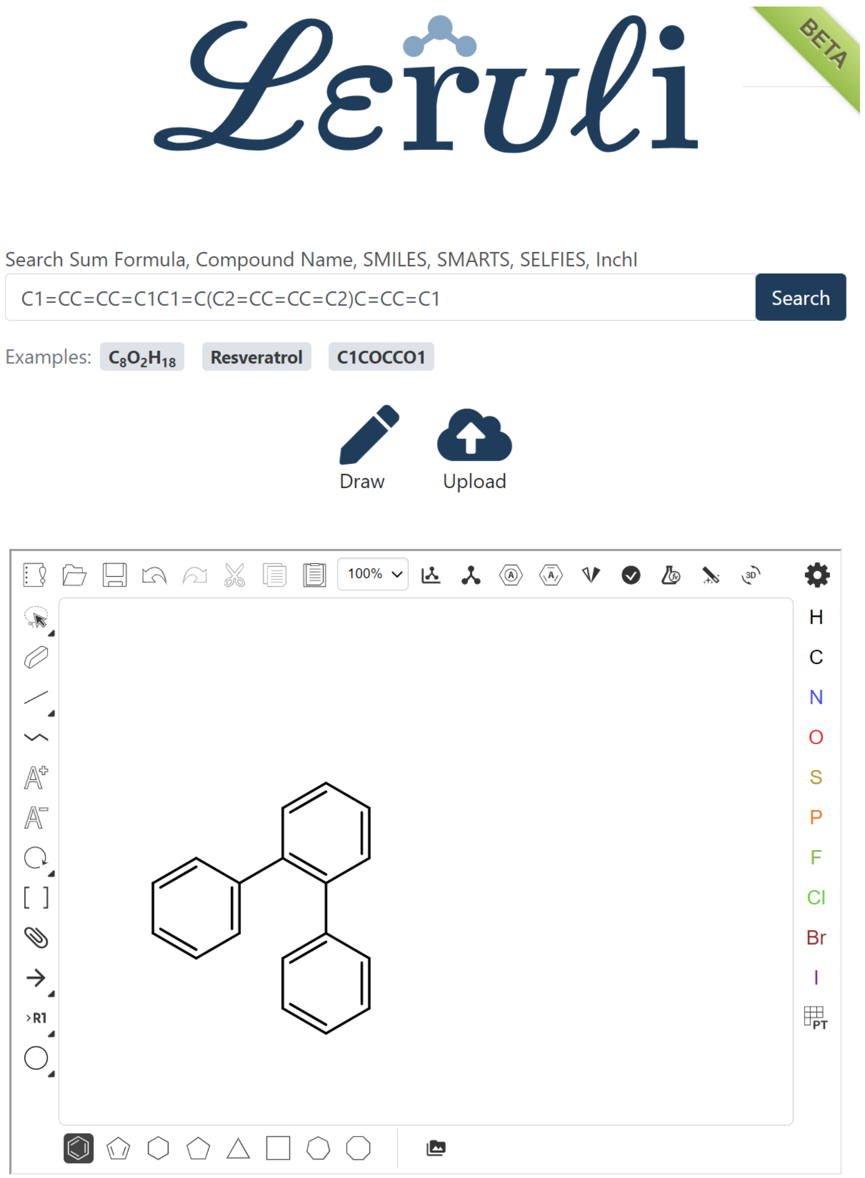
qmlcode/qml

ferchault/mlscheduling

...

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

leruli.com

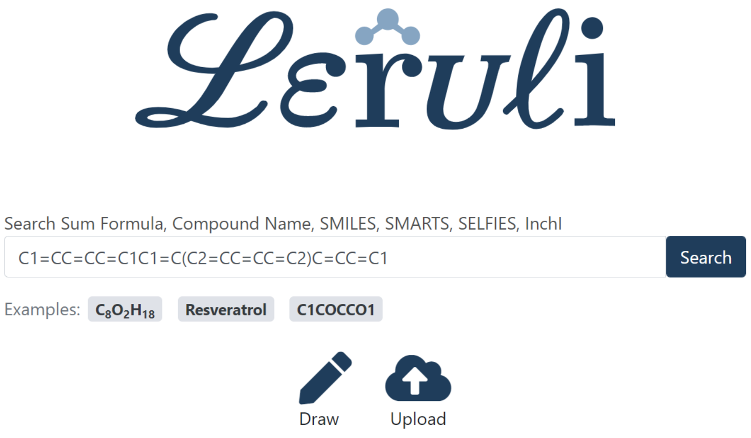


Search Sum Formula, Compound Name, SMILES, SMARTS, SELFIES, InchI

C1=CC=CC=C1C1=C(C2=CC=CC=C2)C=CC=C1

Examples: $C_8O_2H_{18}$ Resveratrol C1COCCO1

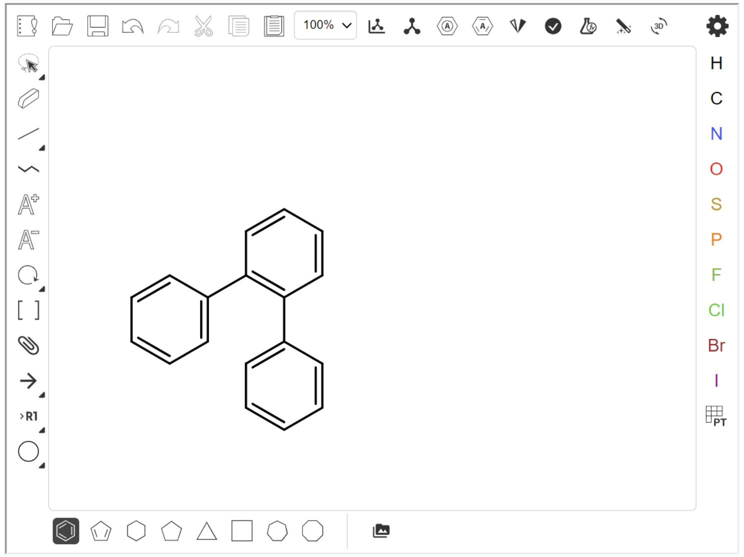
Model availability

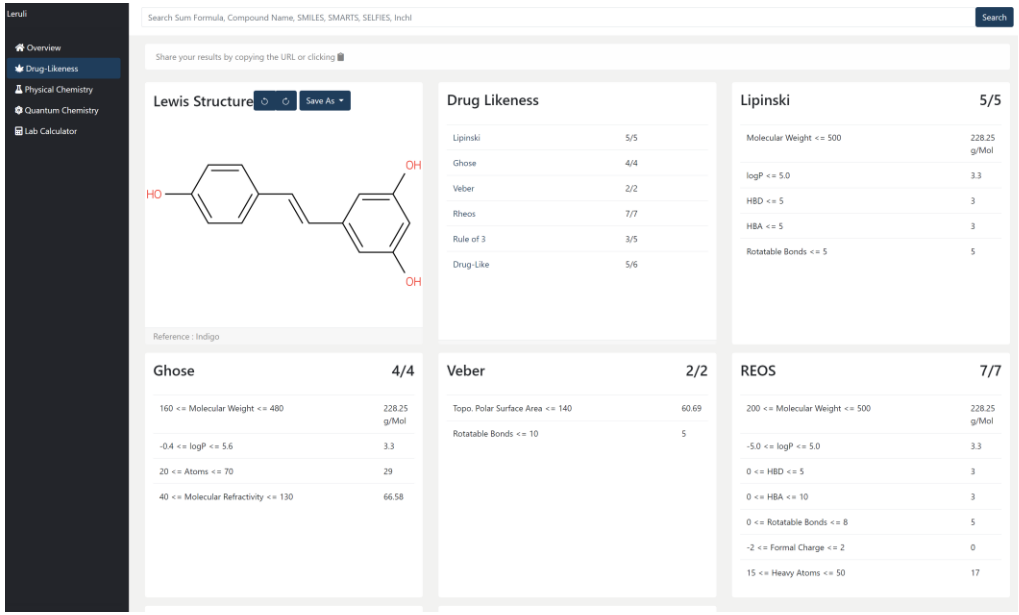


Leruli

Search Sum Formula, Compound Name, SMILES, SMARTS, SELFIES, InchI

C1=CC=CC=C1C1=C(C2=CC=CC=C2)C=CC=C1





Share your results by copying the URL or clicking

Lewis Structure

Reference : Indigo

Drug Likeness

Lipinski	5/5
Ghose	4/4
Veber	2/2
Rheos	7/7
Rule of 3	3/5
Drug-Like	5/6

Lipinski 5/5

Molecular Weight <= 500	228.25 g/Mol
logP <= 5.0	3.3
HBD <= 5	3
HBA <= 5	3
Rotatable Bonds <= 5	5

Ghose 4/4

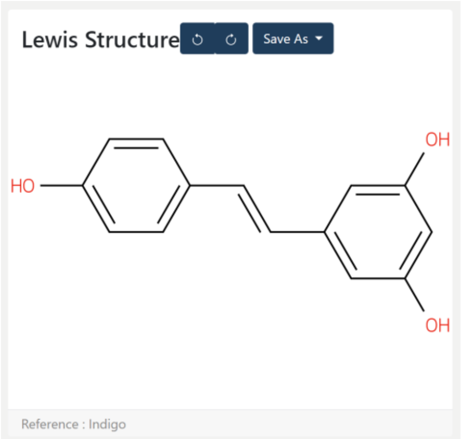
160 <= Molecular Weight <= 480	228.25 g/Mol
-0.4 <= logP <= 5.6	3.3
20 <= Atoms <= 70	29
40 <= Molecular Refractivity <= 130	66.58

Veber 2/2

Topo. Polar Surface Area <= 140	60.69
Rotatable Bonds <= 10	5

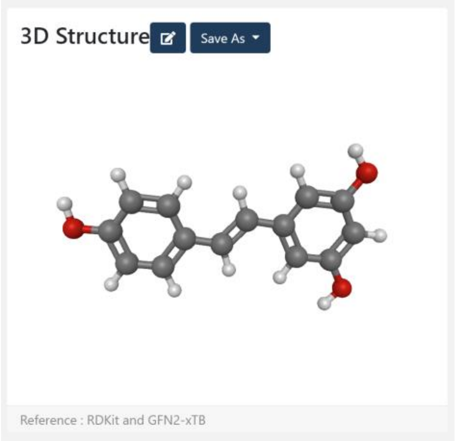
REOS 7/7

200 <= Molecular Weight <= 500	228.25 g/Mol
-5.0 <= logP <= 5.0	3.3
0 <= HBD <= 5	3
0 <= HBA <= 10	3
0 <= Rotatable Bonds <= 8	5
-2 <= Formal Charge <= 2	0
15 <= Heavy Atoms <= 50	17



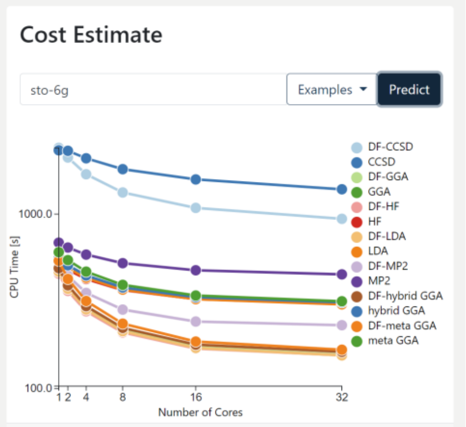
Lewis Structure

Reference : Indigo



3D Structure

Reference : RDKit and GFN2-xTB

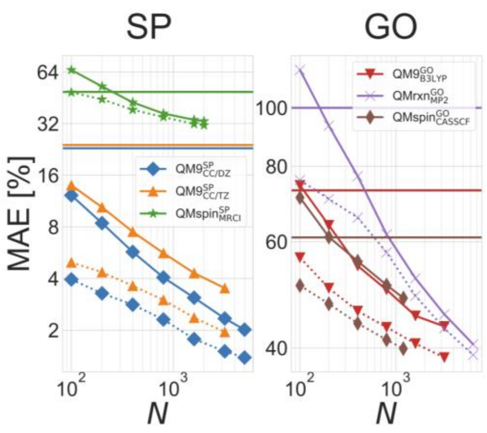


Cost Estimate

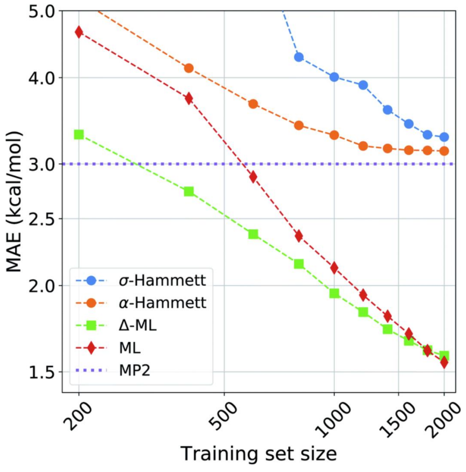
sto-6g

Want to include your model? Let me know

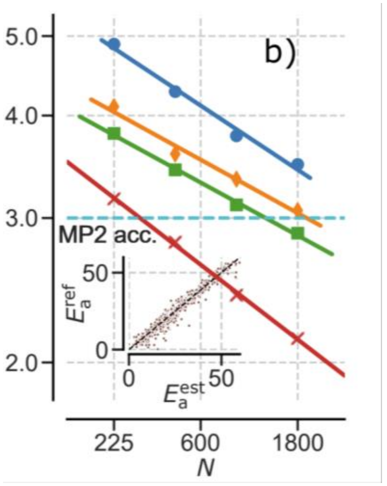
Summary



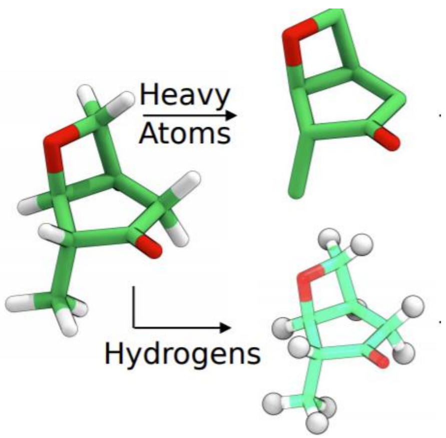
Estimate computational cost



Detrending with Hammett’s equation



Energies with KRR



Geometries with Graph2Structure

ferchault/mlscheduling

chemspacelab/ Enhanced-Hammett

qmlcode/qml

qmlcode/qml

Acknowledgements

Prof. von Lilienfeld

Dominik Lemm

Marco Bragato

Stefan Heinen

Dr. Max Schwilk

ferchault @ferchault guido.vonrudorff.de