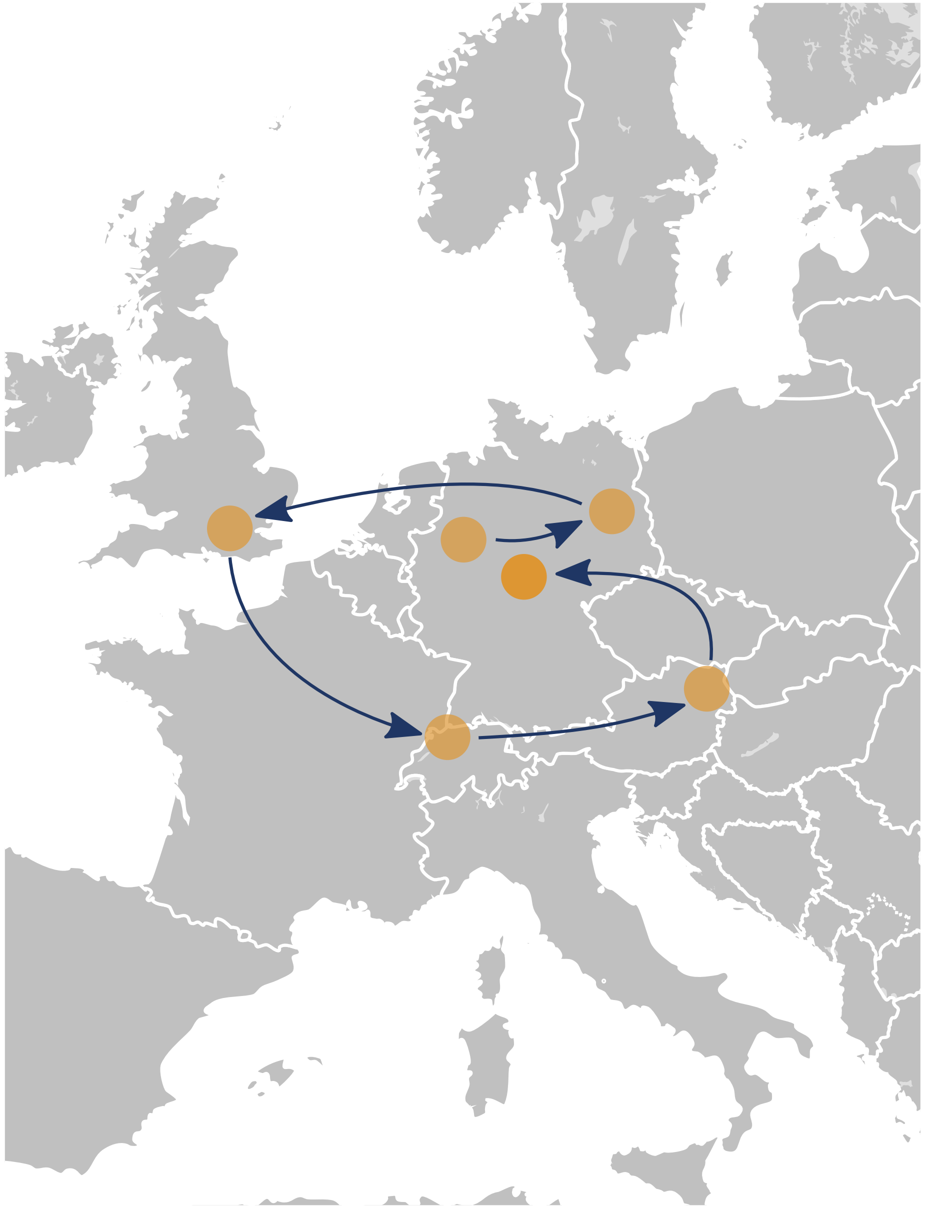
Computational Material Design and the Curse of Dimensionality

Guido Falk von Rudorff, University of Kassel

ferchault @ferchault nablachem.org

About me



Freie Universität Berlin

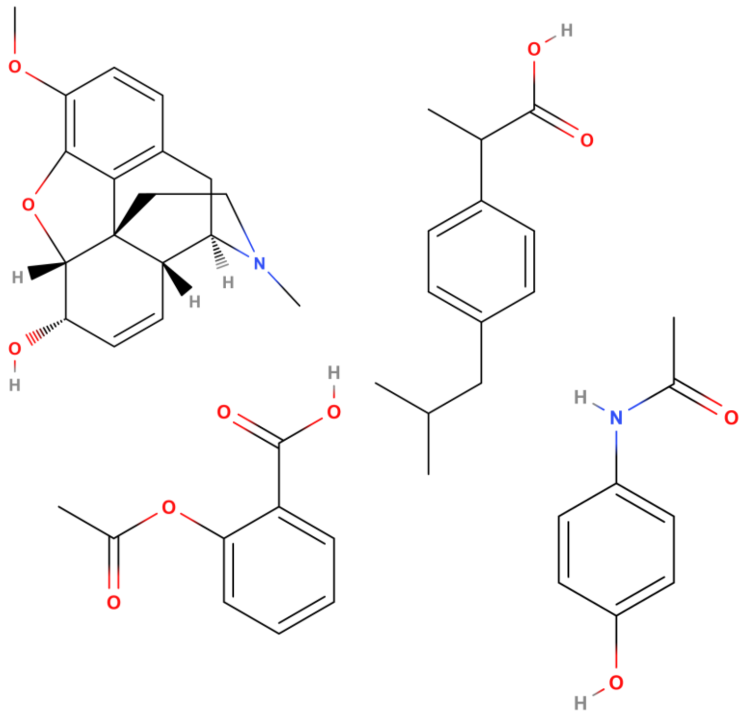
UCL

Universität Basel

universität wien

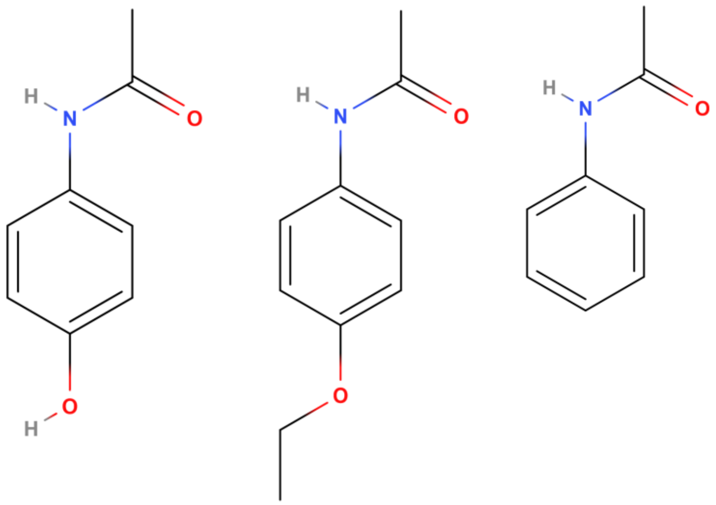
UNI KASSEL VERSITÄT

Introduction



Global Search Problem

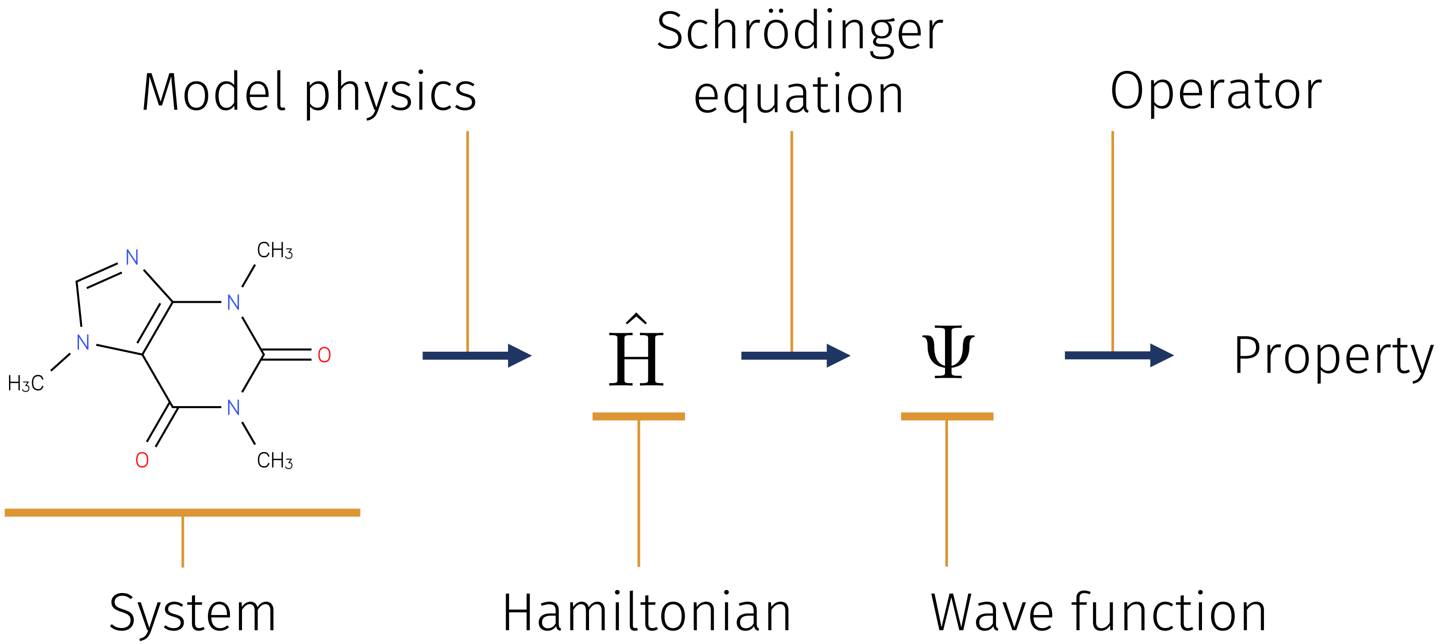
Which class of compounds?



Local Search Problem

Which particular species within that class?

Solved?


Model physics
Schrödinger
equation
Operator
N
CH3
N
N
O
H3C
N
O
CH3
Ĥ
Ψ
Property
System
Hamiltonian
Wave function

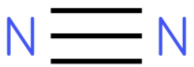
Wave Function

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

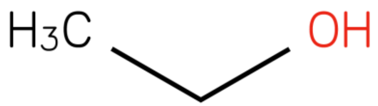
Methane

CH_4

N_2



Ethanol



Solved by approximations in computational chemistry?

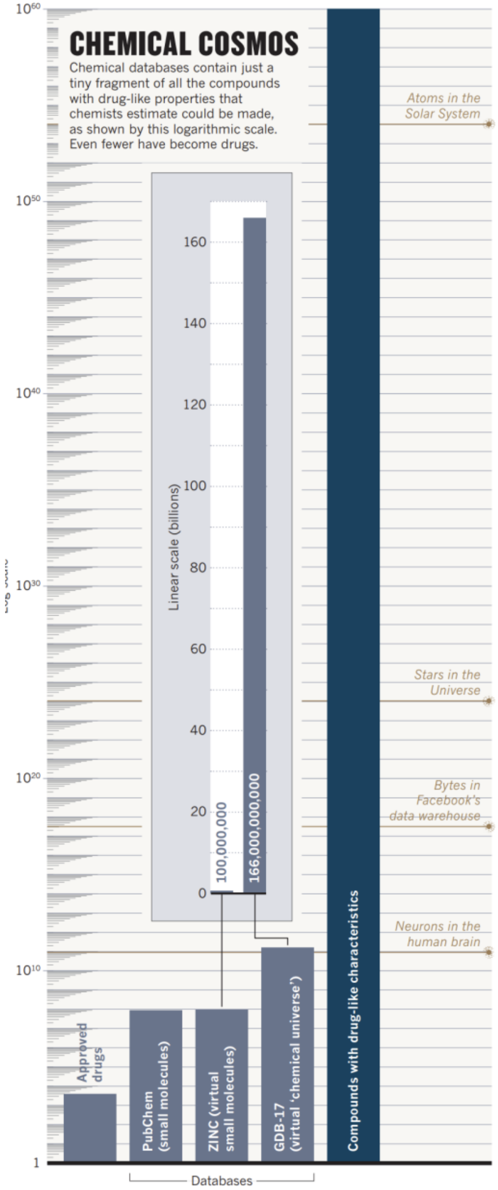
Scaling of Molecules

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



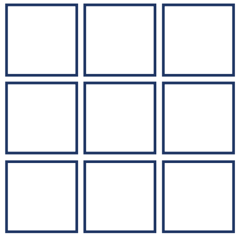
A. Mullard, *Nature*, 2017.

Scaling of Materials

Face centered cubic and 70 elements only

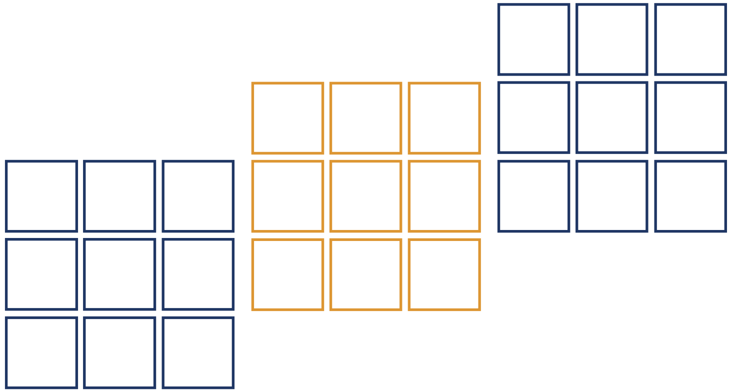
9 primitive cells

- Binary 10^7
- Ternary: 10^{13}
- Quaternary: 10^{15}



27 primitive cells

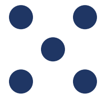
- Binary: ~ 10^{17}
- Ternary: ~ 10^{29}
- Quaternary: ~ 10^{36}



K. Shinohara, A. Seko, T. Horiyama, M. Ishihata, J. Honda, I. Tanaka, *J Chem Phys*, **2020**.

3D Geometries

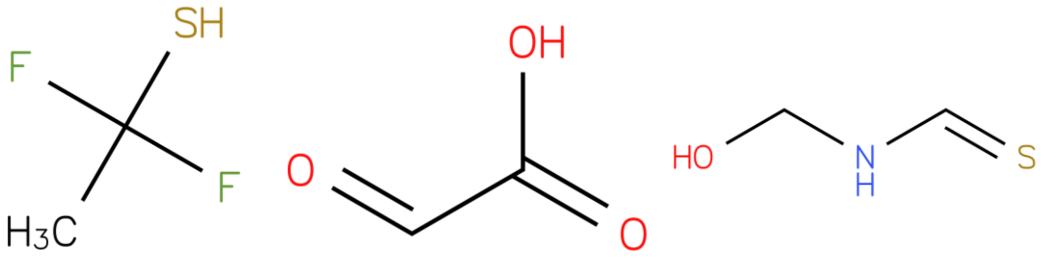
Take some atoms from {C, O, N, F, S}, H-saturated



349:

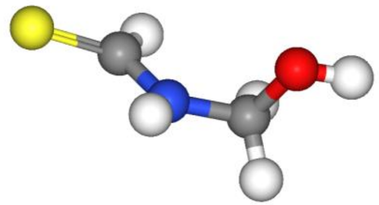
$C_2H_4F_2S$
$C_2H_2O_3$
C_2H_5NOS
...

9,917:

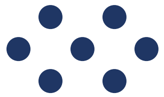


...

~52k:



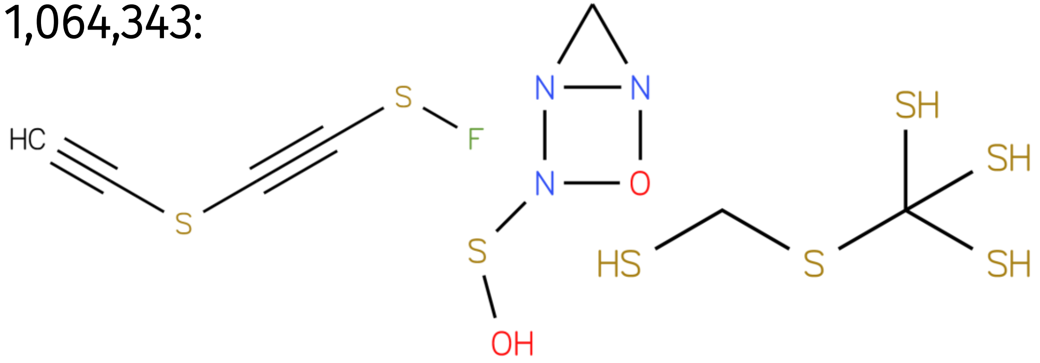
...



1,050:

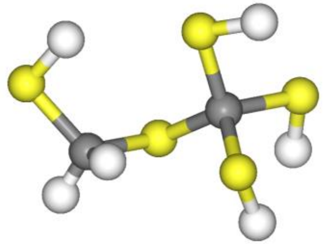
C_4HFS_2
$CH_3N_3O_2S$
$C_2H_6S_5$
...

1,064,343:



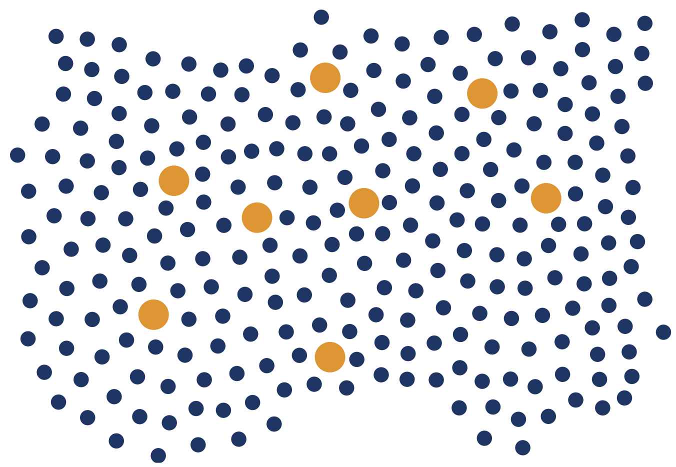
...

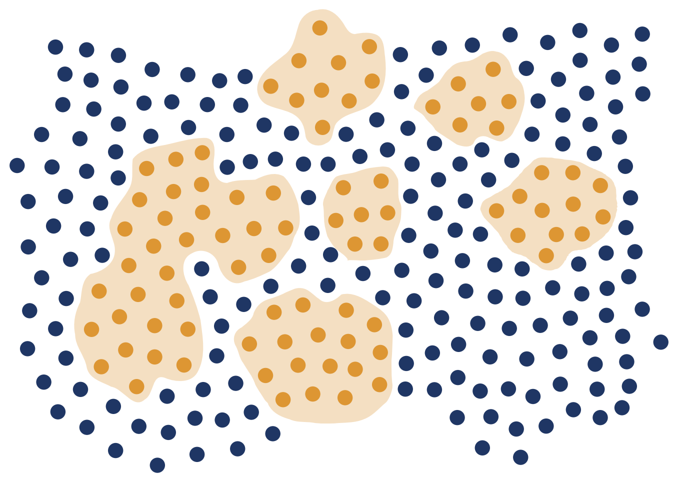
~23M:



...

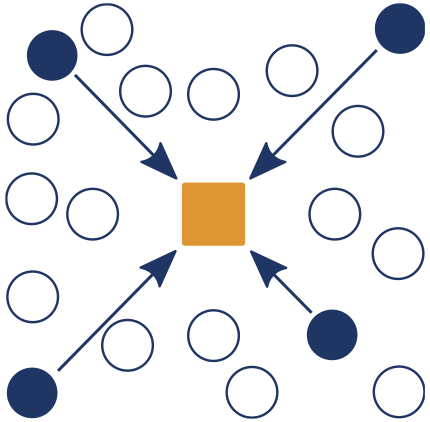
Speed does not matter:
even enumeration is impossible.





Approaches

Machine Learning



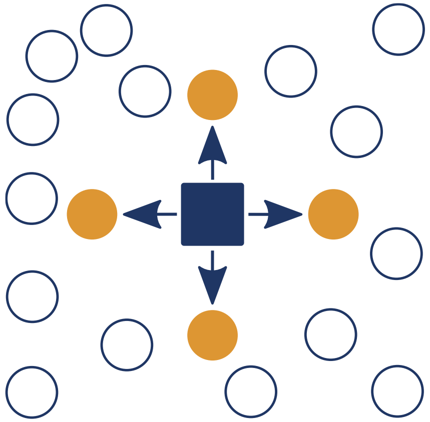
Foundations | Statistical modelling

Accuracy | Systematically improvable through data and training

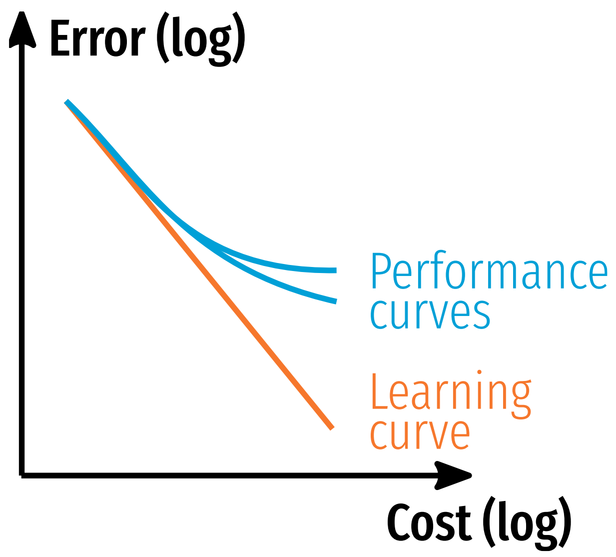
Specialty | Universal, scale-bridging, data-driven approach

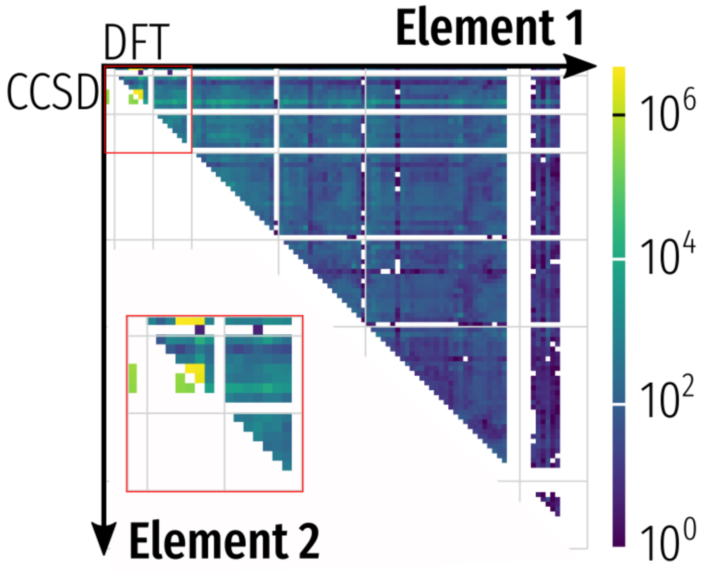
Limitation | Requires training data, no black box

Quantum Alchemy



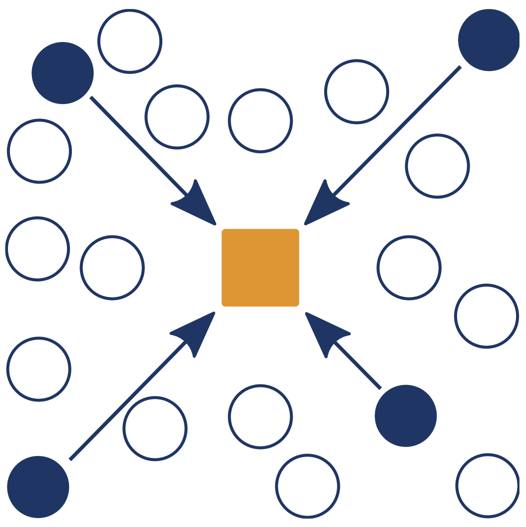
Data Availability





Low-Data Regime

Machine Learning

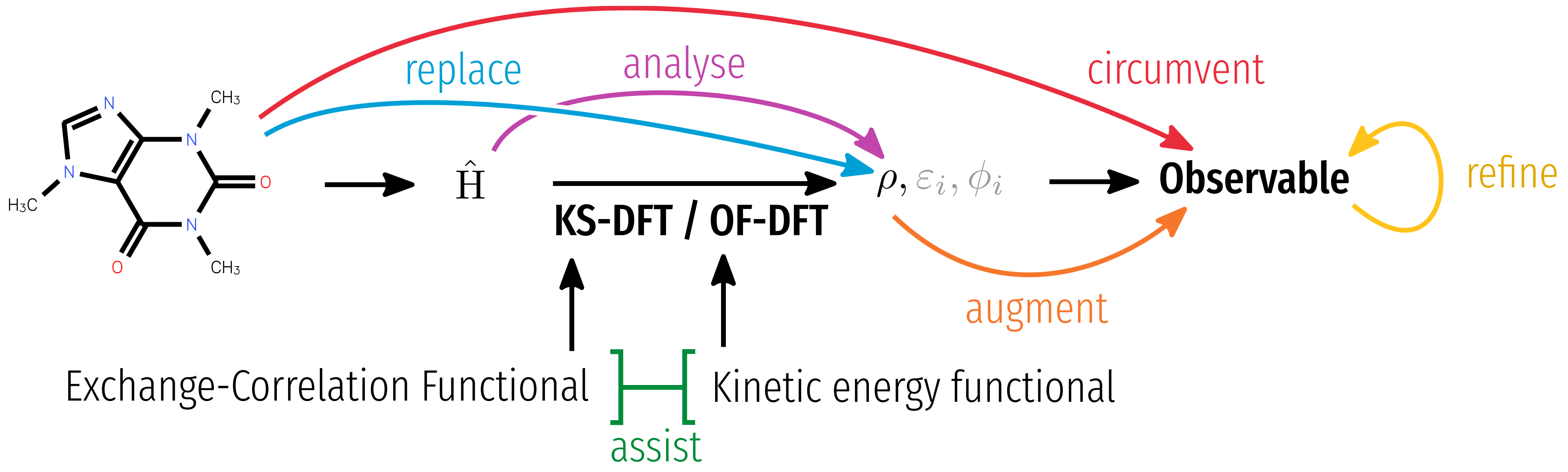


Kernel-Ridge-Regression

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

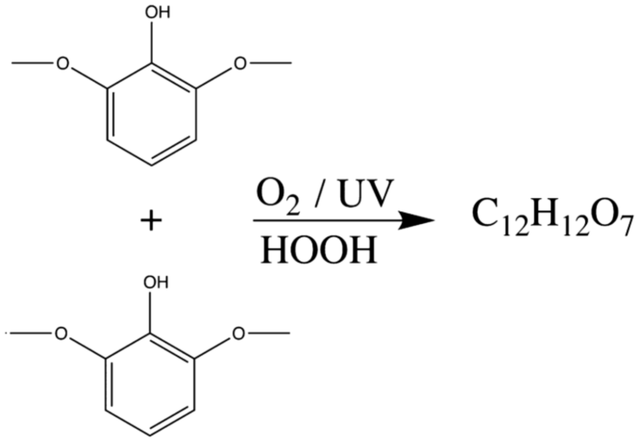
V. N. Vapnik: The Nature of Statistical Learning Theory, *Springer* **2000**.
B. Schölkopf, A. J. Smola: Learning with Kernels, *MIT Press*, **2001**.

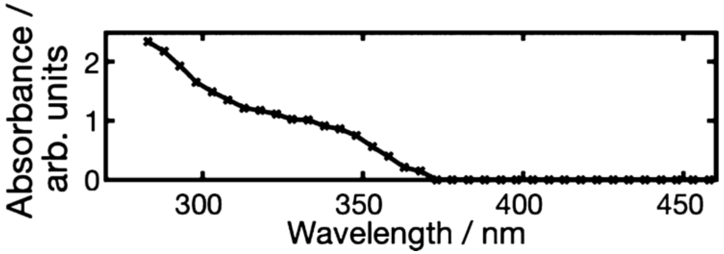
Strategies


CH3
N
N
N
O
N
H3C
O
CH3
Ĥ
KS-DFT / OF-DFT
ρ, εi, φi
Observable
replace
analyse
circumvent
refine
augment
Exchange-Correlation Functional
Kinetic energy functional
assist

Filter by Spectrum

Experiment

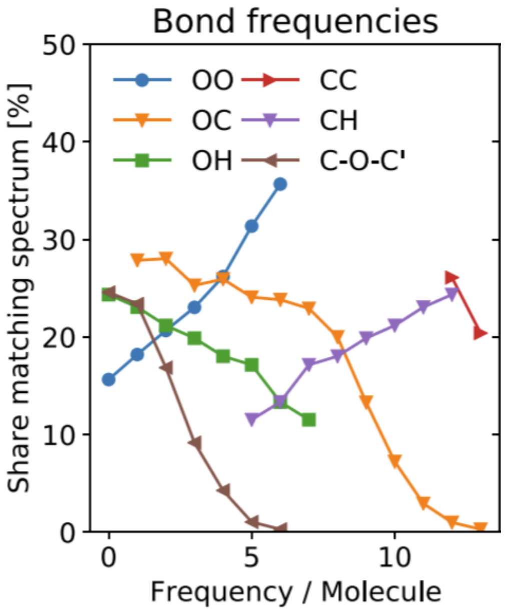




Search space

Molecular graphs: 264 M

Stable molecules: 123 M



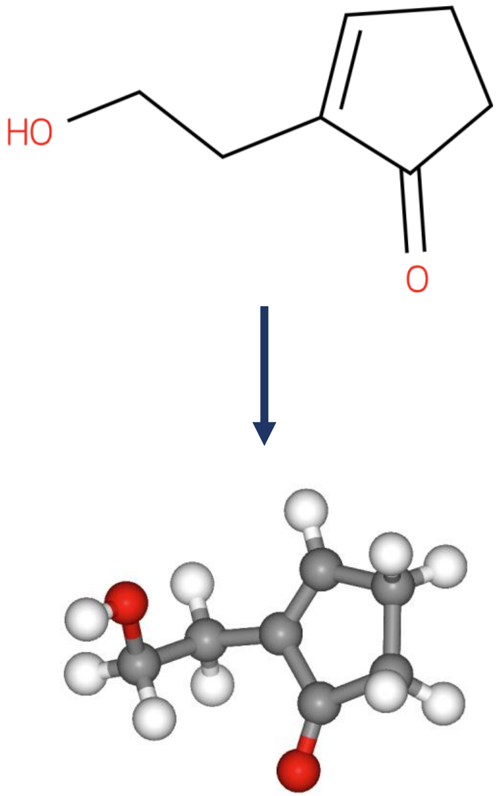
Millions at once!

Guide experiment

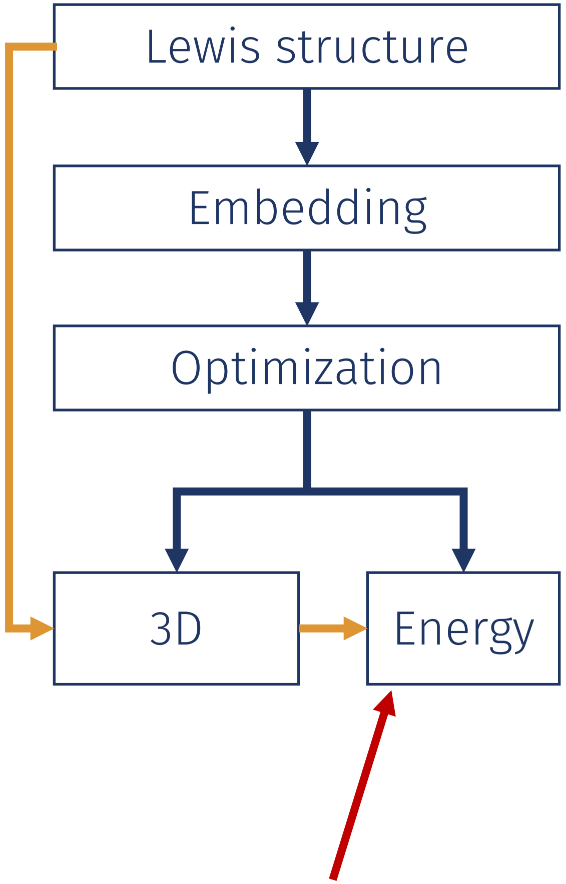
How many molecules are left and which feature to measure next?

E. Tapavicza, **GFvR**, D. O. De Haan, M. Contin, C. George, M. Riva, O. A. von Lilienfeld, *Environ. Sci. Technol.* **2021**.

Filter by Geometry



Traditional / G2S



ML: 100.000x faster,
can filter graphs

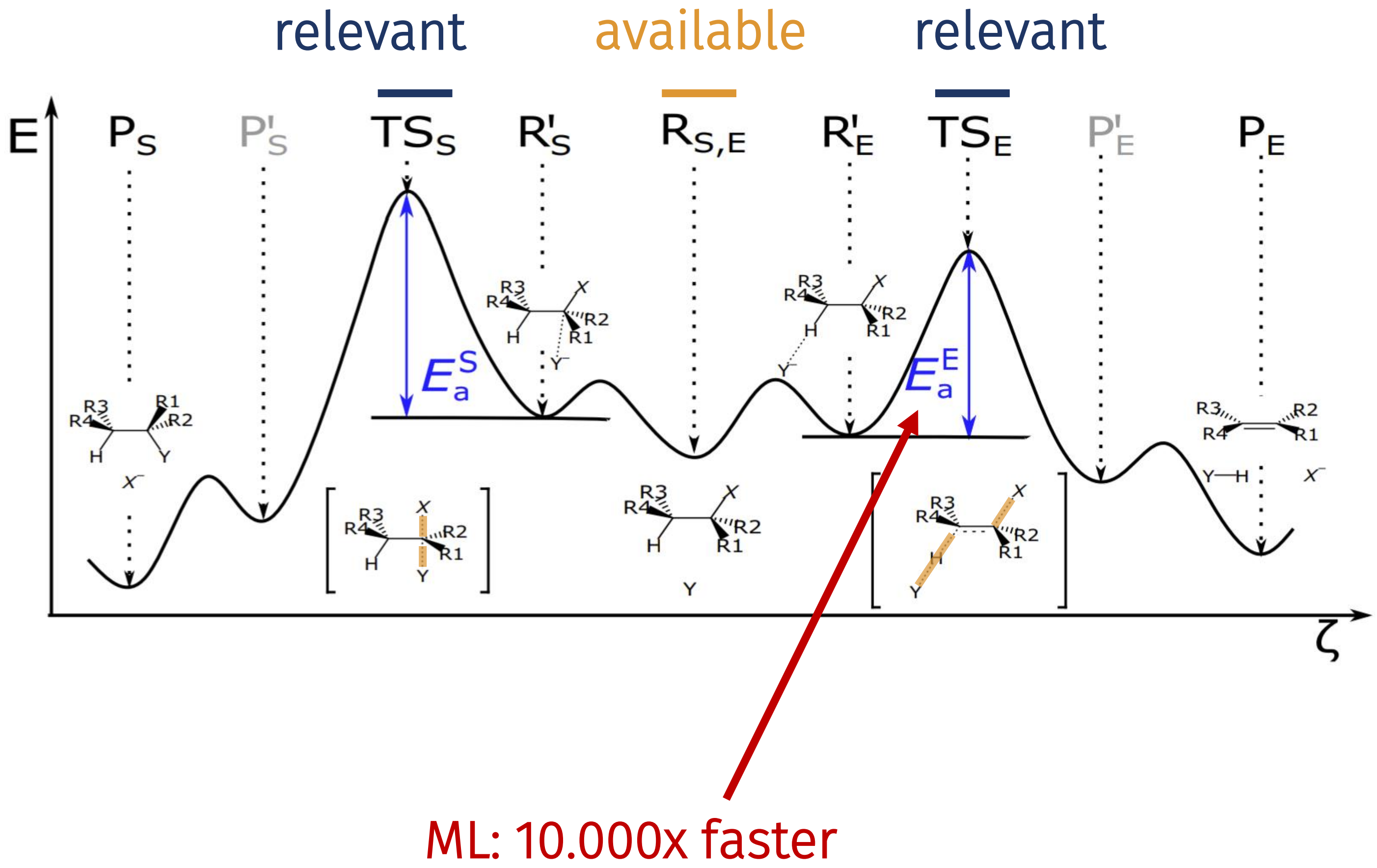
G2S

- Closer to DFT than common methods
  - Small molecules

- Applicable to complex chemical spaces
  - Transition state geometries
  - Carbenes
  - Elpasolite crystals

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

Filter by Barriers

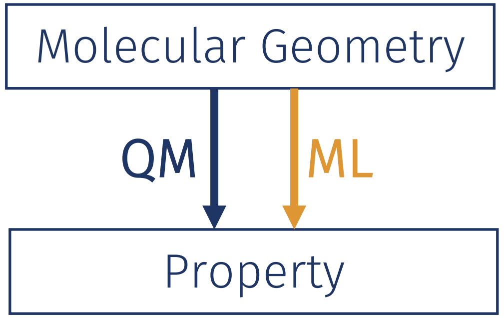


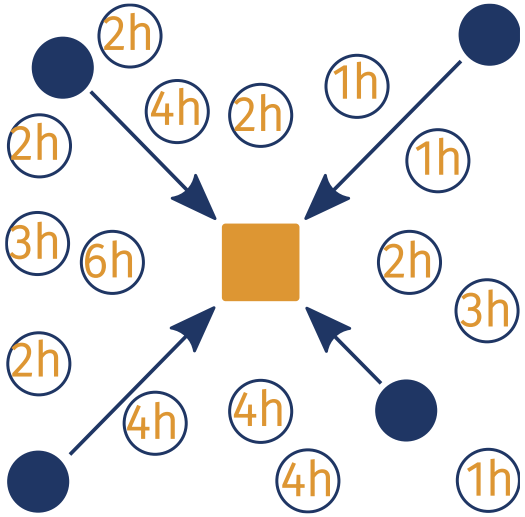
Competing reactions: E2, S_N2

- 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

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.

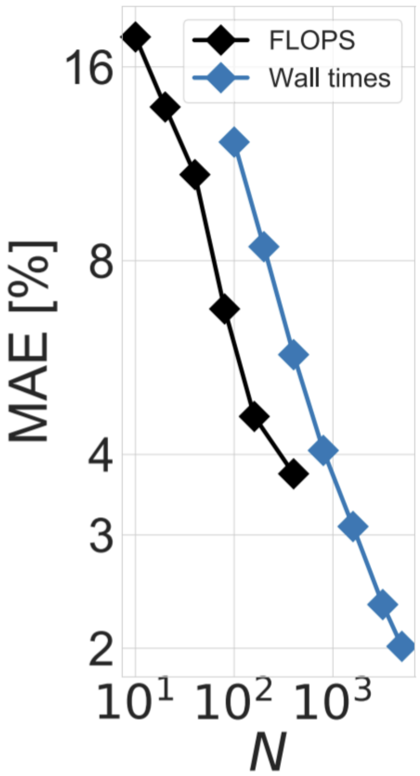
Filter by Cost





Computational effort as molecular property

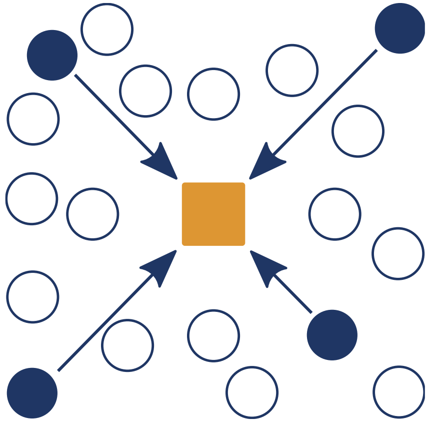
- Improves models
- Accuracy depends on problem
  - Single points: 2%
  - Transition state search: 25%
  - Geometry optimisations: 40%



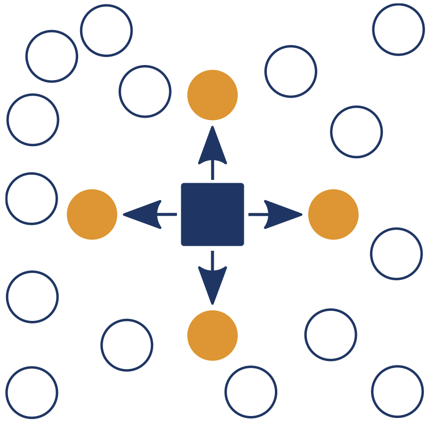
S.N. Heinen, M. Schwilk, **GFvR**, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* **2020**.

Approaches

Machine Learning



Quantum Alchemy



Foundations | Perturbation theory

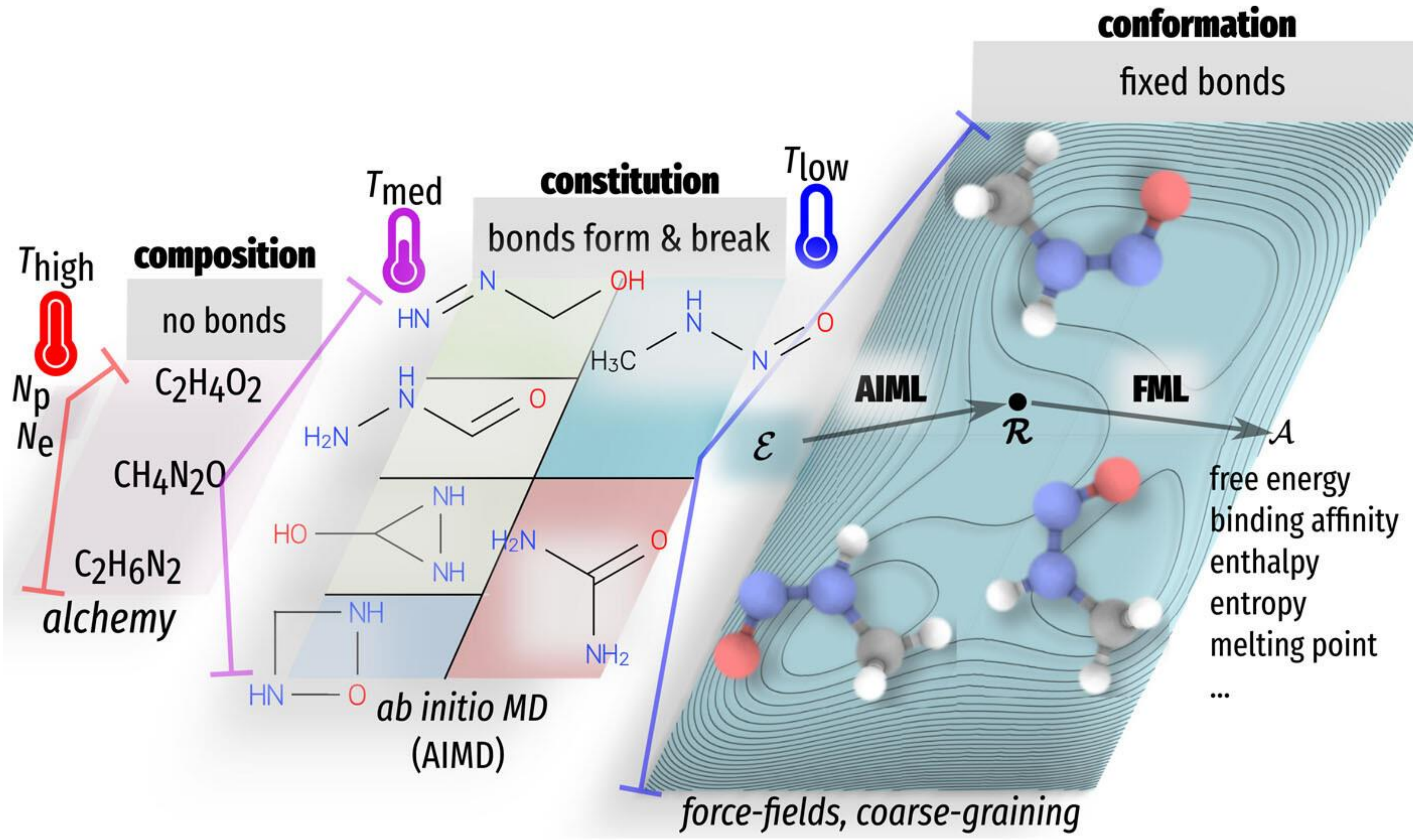
Accuracy | Systematically improvable through higher orders terms

Specialty | Combinatorial scaling with chemical diversity

Limitation | Finite range in chemical space

Joseph Wright, 1771

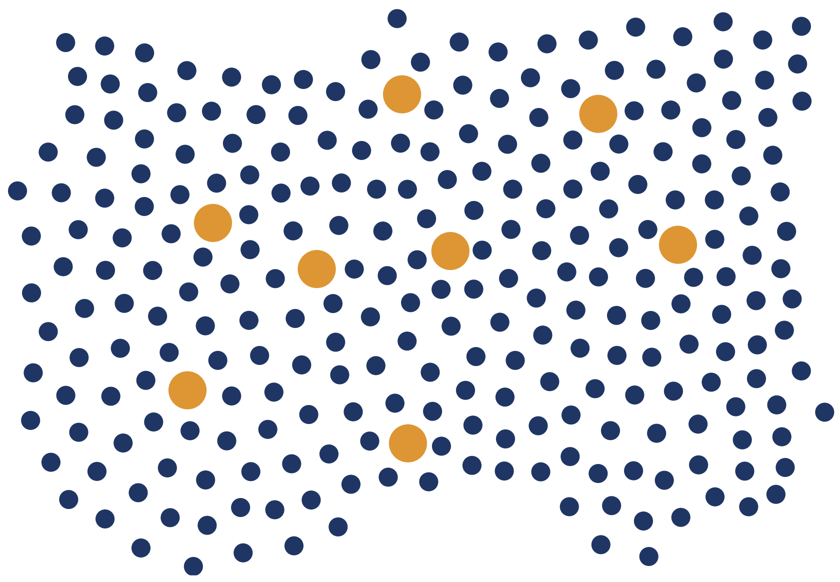
Discrete points?



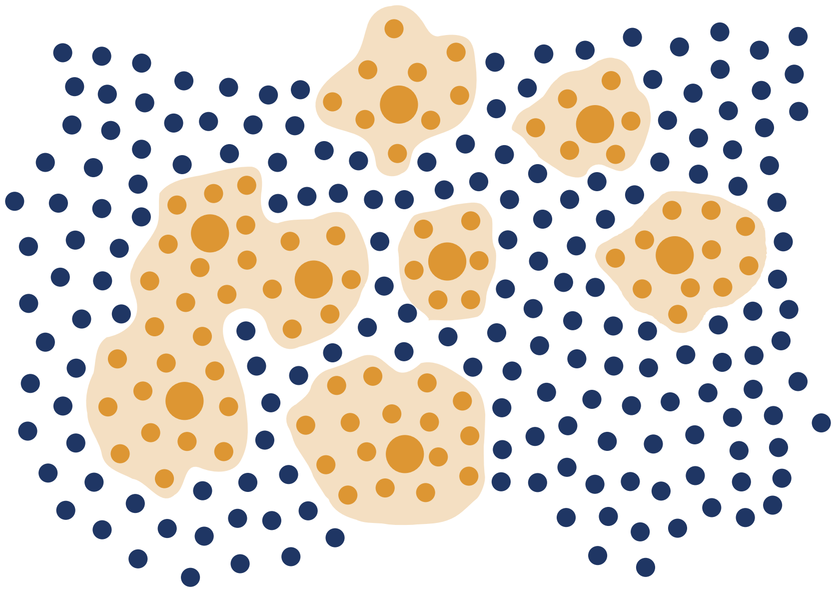
J. Weinreich, D. Lemm, GFvR, O.A. von Lilienfeld, *J. Chem. Phys.*, 2022.

Motivation

Without Perturbation



With Perturbation



Systems/Molecules

- Any
- Known
- Approximated

Motivation

Perspective shift

Few highly accurate calculations instead of many intermediate ones

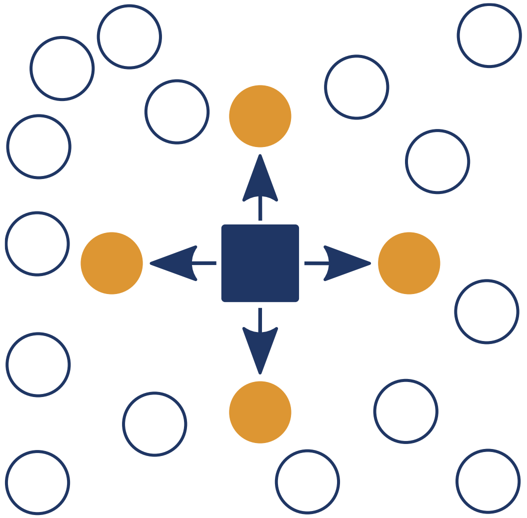
$$\hat{\mathrm{H}} = \hat{\mathrm{H}}(\underbrace{Z_i, \mathbf{R}_i}_{4N}, \underbrace{N_e}_{1D,\ \text{close to } \sum_i Z_i}, \underbrace{\sigma}_{1D})$$

4N

1D, close to $\sum_i Z_i$

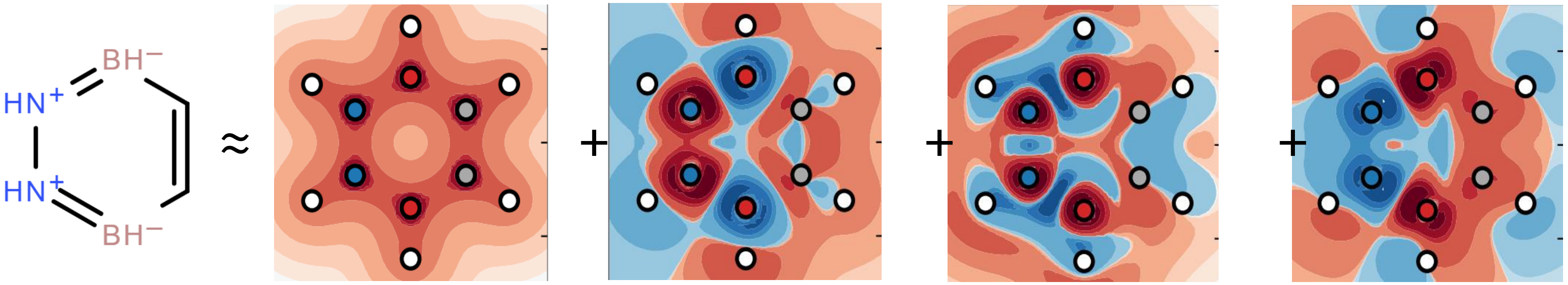
1D

Quantum Alchemy



Taylor expansion

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



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

Quantum Alchemy

Interpolate between molecular isoelectronic Hamiltonians

$$\hat{H}(\lambda) \equiv \lambda \hat{H}_{\mathrm{t}} + (1-\lambda)\hat{H}_{\mathrm{r}}$$

Taylor expansion around reference molecule

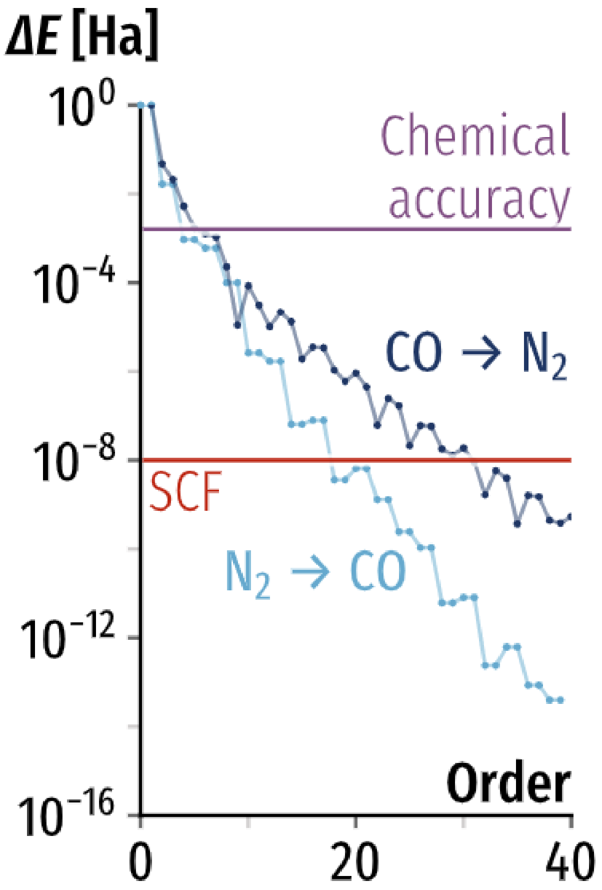
$$E_{\mathrm{t}} = E_{\mathrm{r}} + \Delta E^{\mathrm{NN}} + \int_{\Omega} d\mathbf{r} \sum_{n=0}^{\infty} \frac{1}{(n+1)!} \, \Delta v \frac{\partial^n \rho_\lambda(\mathbf{r})}{\partial \lambda^n}\bigg|_{\lambda=0}$$

- Gives consistent energies, densities, forces, …
- Uses the same derivatives for all predictions

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

xkcd.com/2605

Convergence



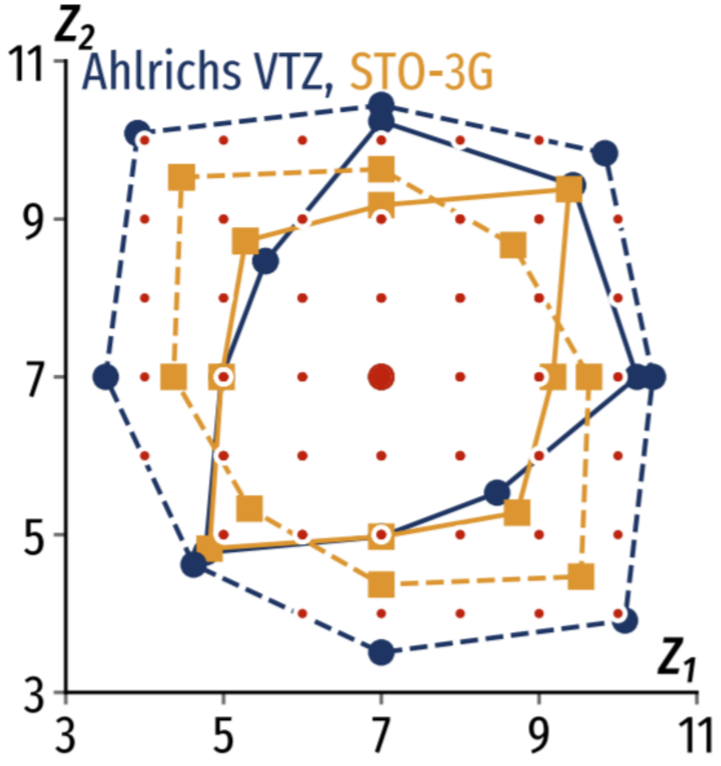
HF/STO-3G

Taylor expansion

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

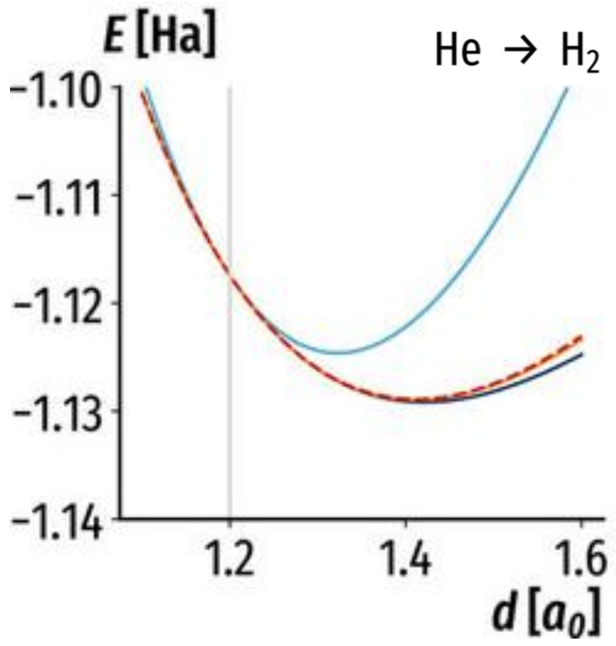
Limits

- Only few orders feasible
- Iterative in chemical space



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

Geometry Relaxation



Taylor expansion

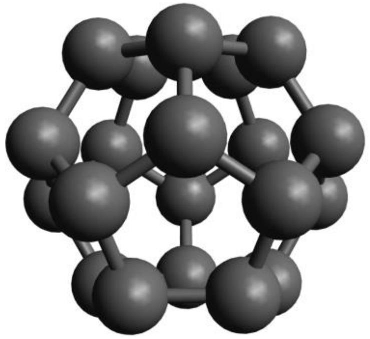
- Large changes still converge (more slowly)
- Geometric response can be recovered

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

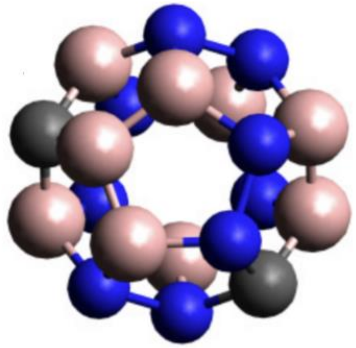
Covalent Interactions

Scaling with chemical space

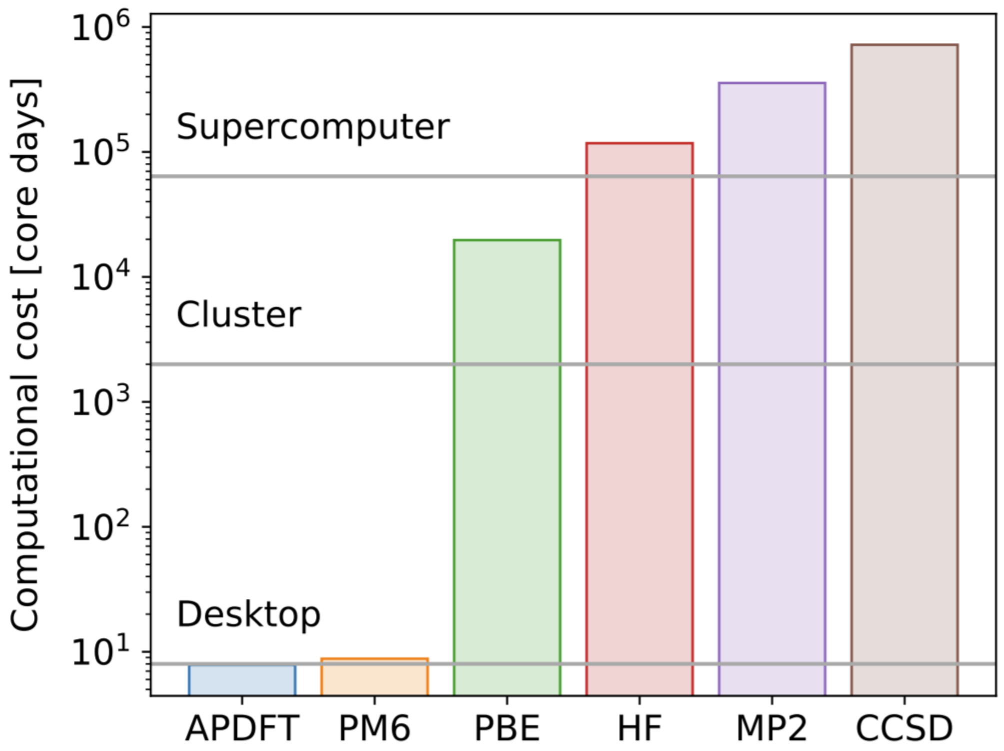
- 1 derivative for second order
- 5 derivatives for third order



C_{20}



$3.1 \cdot 10^6$ targets



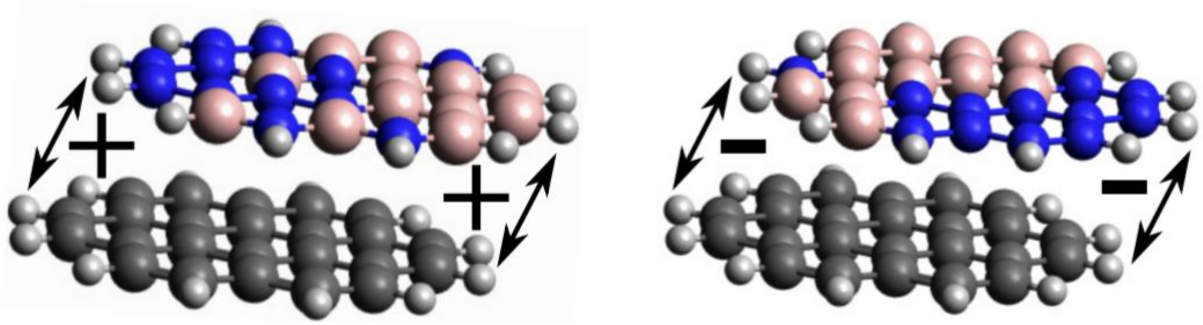
QA: 80.000x faster

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

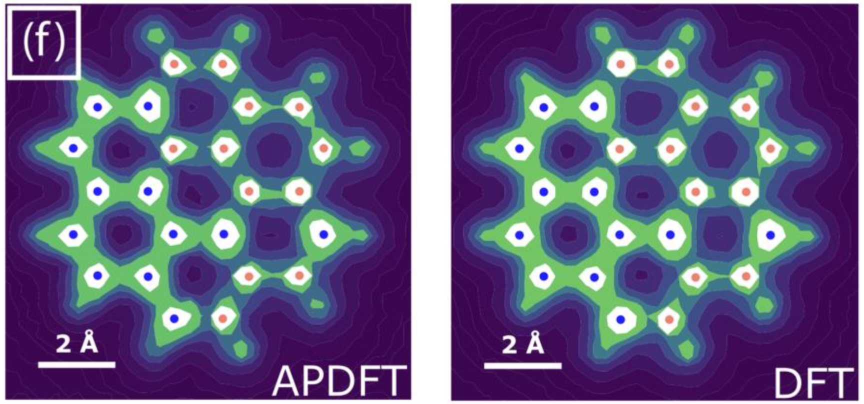
Non-covalent Interactions

BN-doped coronene dimer

- Identify most/least attractive doping pattern
- Design case



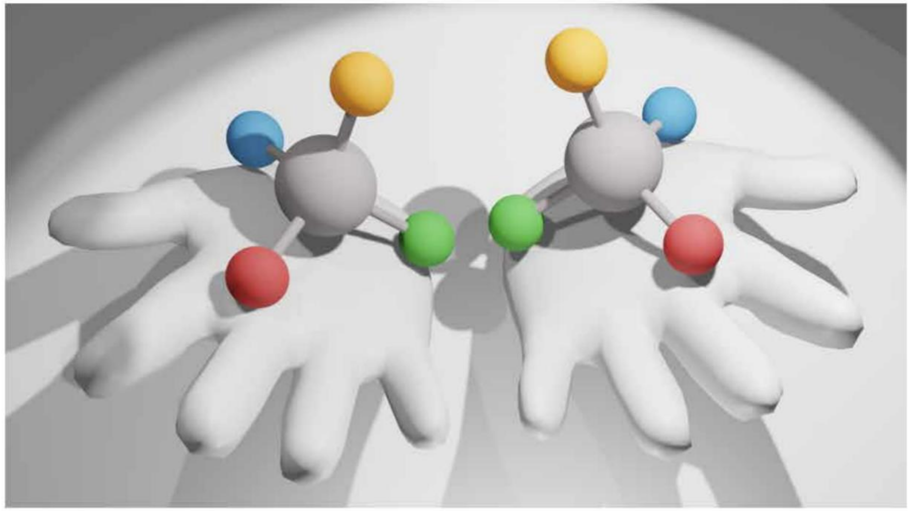
QA: 20.000x faster

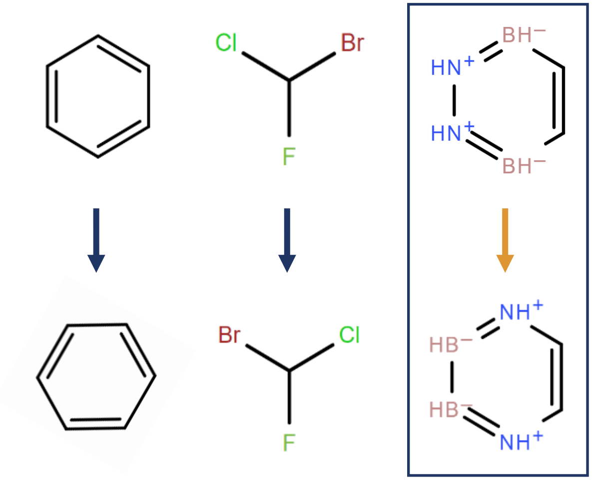


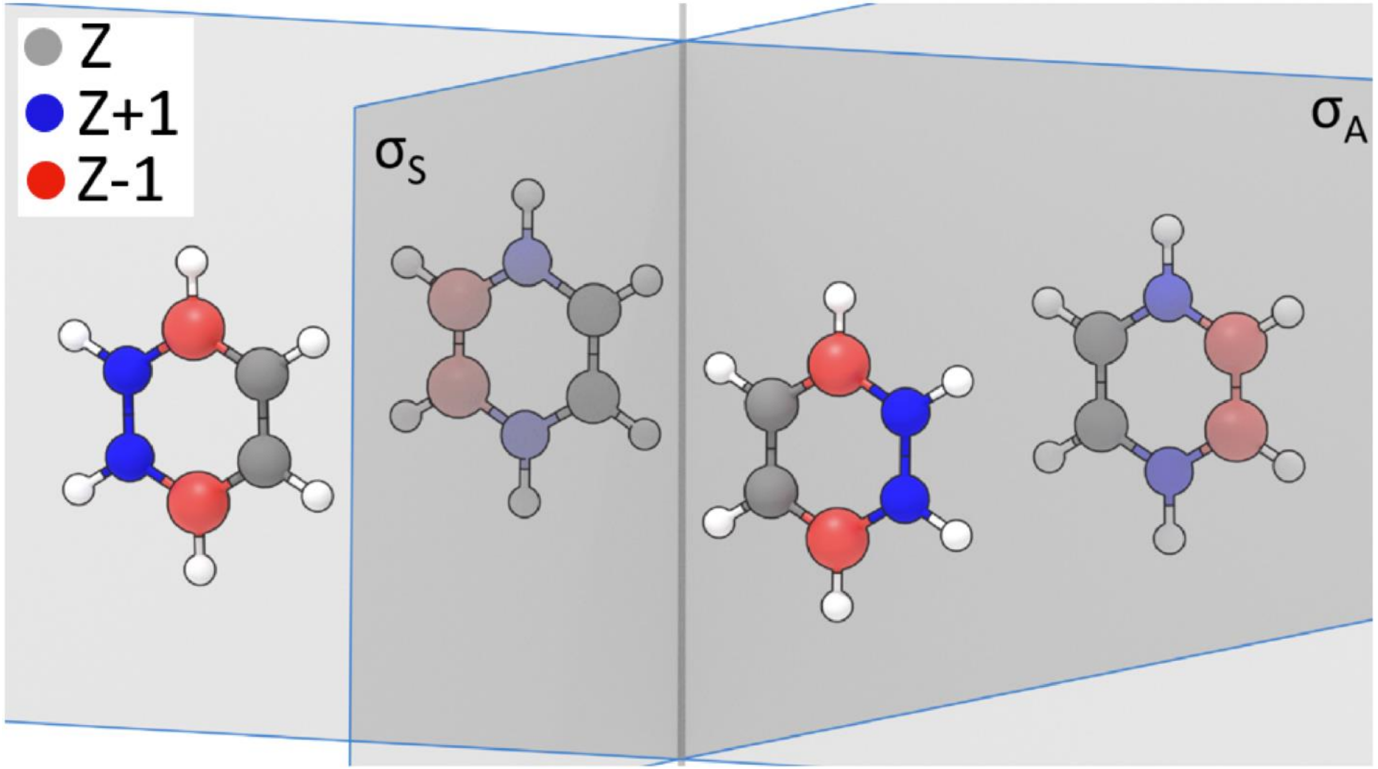
$2.8 \cdot 10^{10}$ targets

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

Alchemical Enantiomers





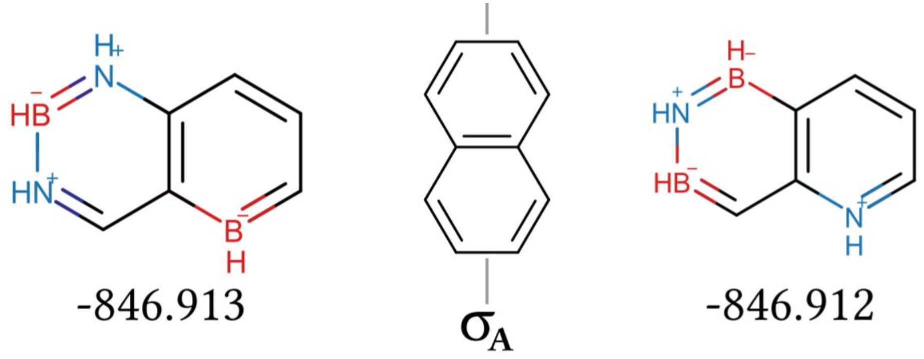


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

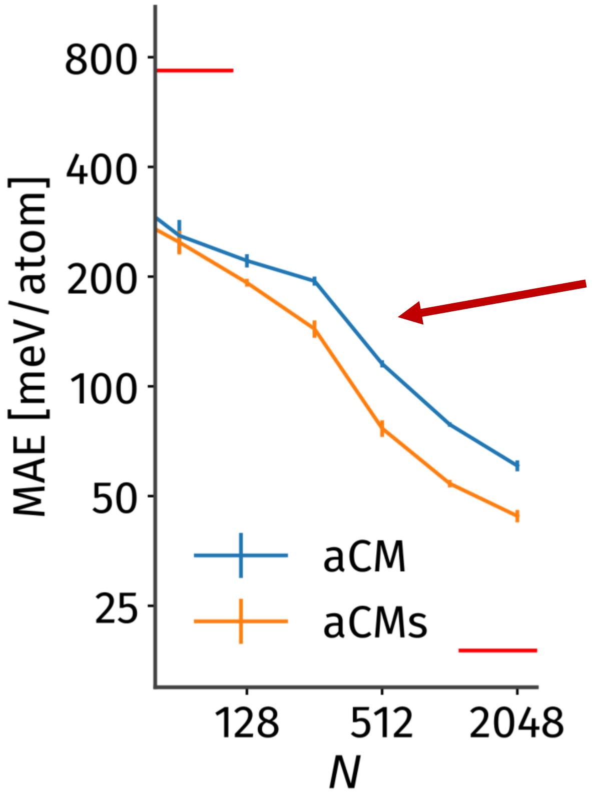
Alchemical Enantiomers

Fundamentally new symmetry

Electronic energy only



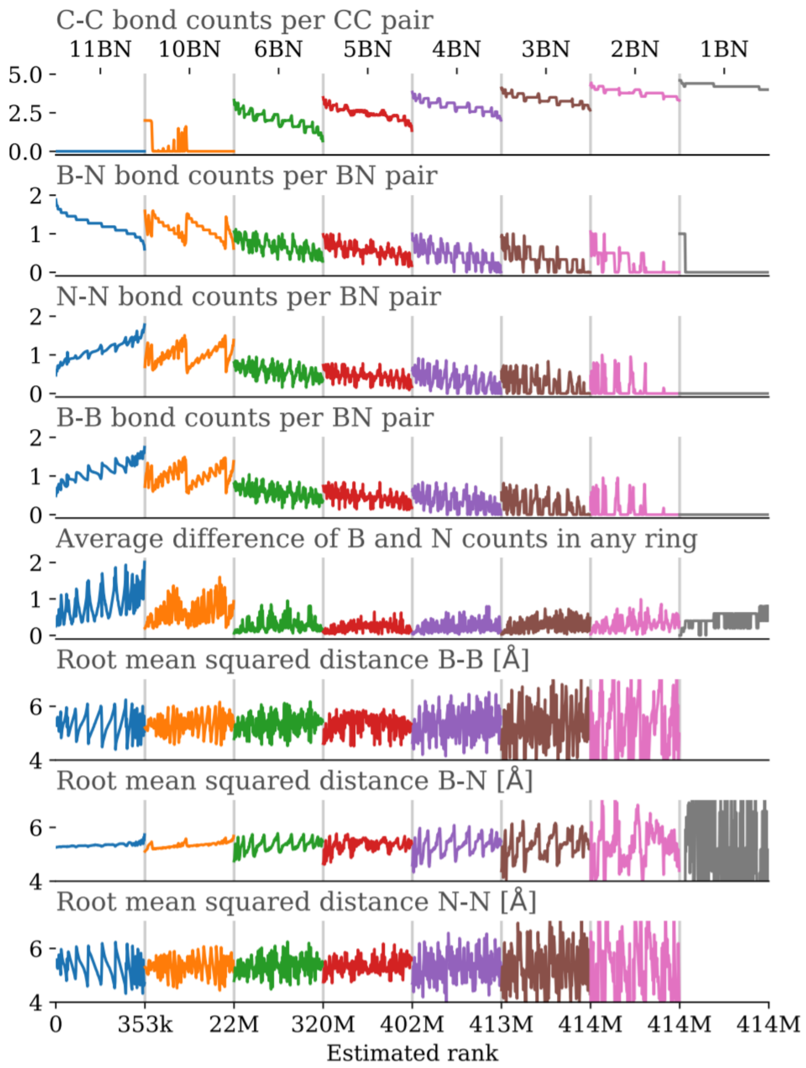
Speed up machine learning

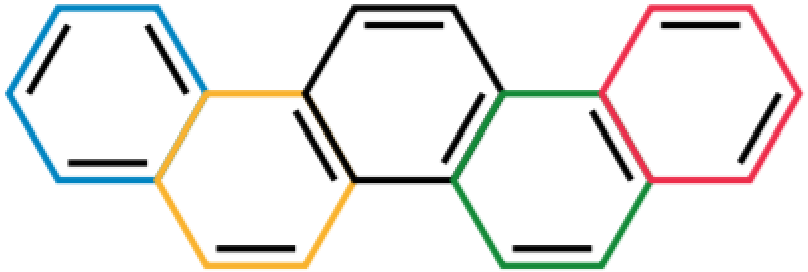


ML+QA: half the data

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

Alchemical Enantiomers





x 414 M

QA: Millions at once!

Design rules in order of decreasing strength

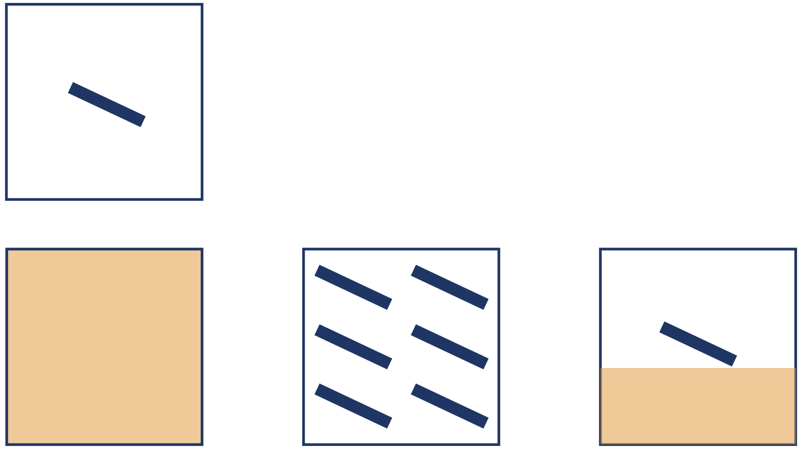
- Add BN pairs
- Maximize CC bonds
- Substitute sites shared between rings
- Maximize BN bonds
- Avoid N substitutions on rings sharing a larger amount of bonds with other rings
- Balance BN substitutions in each ring

Not a single QM calculation required!

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

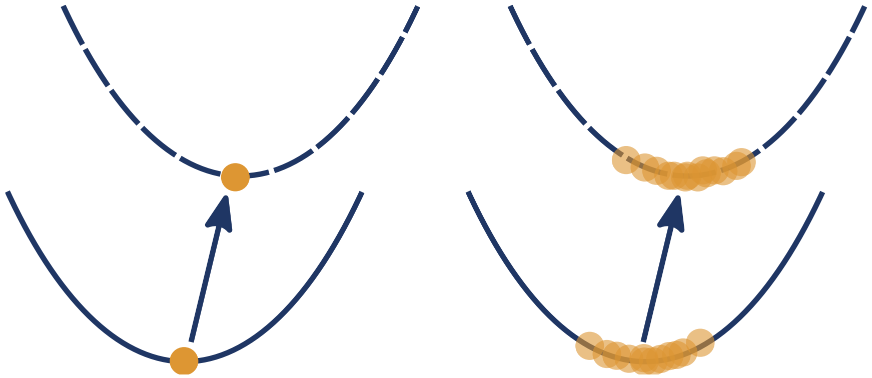
Future Steps

Include materials



also in collaboration with CINSaT members

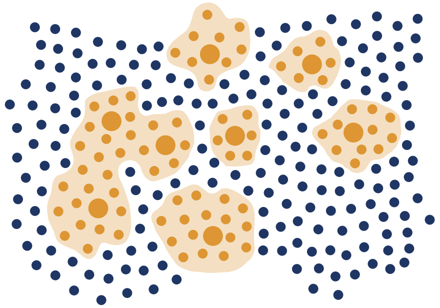
Include ensembles

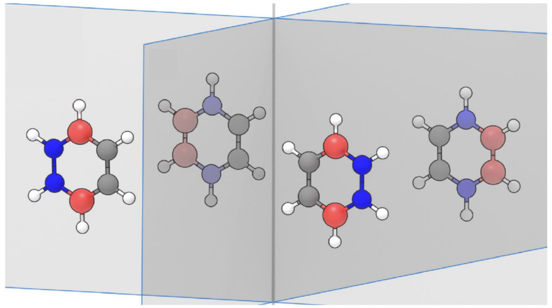


Method

- Basis sets and pseudopotentials need alchemy built-in
- Exploit pen&paper structure for fundamental aspects
- Push differentiable quantum chemistry
- Simplify use

Summary





Efficient | Re-use knowledge, no one-by-one

Symmetries | Reducing (“folding”) search space

Constraints | Exclude whole regions

Differentiable Chemistry | Arbitrary derivatives

Representations | Better data efficiency

Thanks
Marco Bragato
Giorgio Domenichini
Emily Eikey
Stefan Heinen
Chasz Griego
Konstantin Karandashev
John Keith
Mario Krenn
Simon Krug
Dominik Lemm
Anatole von Lilienfeld
Alex Maldonado
Michael Sahre
Max Schwilk
Enrico Tapavicza
Jan Weinreich

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

Quantum Alchemy

Interpolate between molecular isoelectronic 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

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

Hellmann-Feynman theorem

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

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

ferchault/APDFT

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