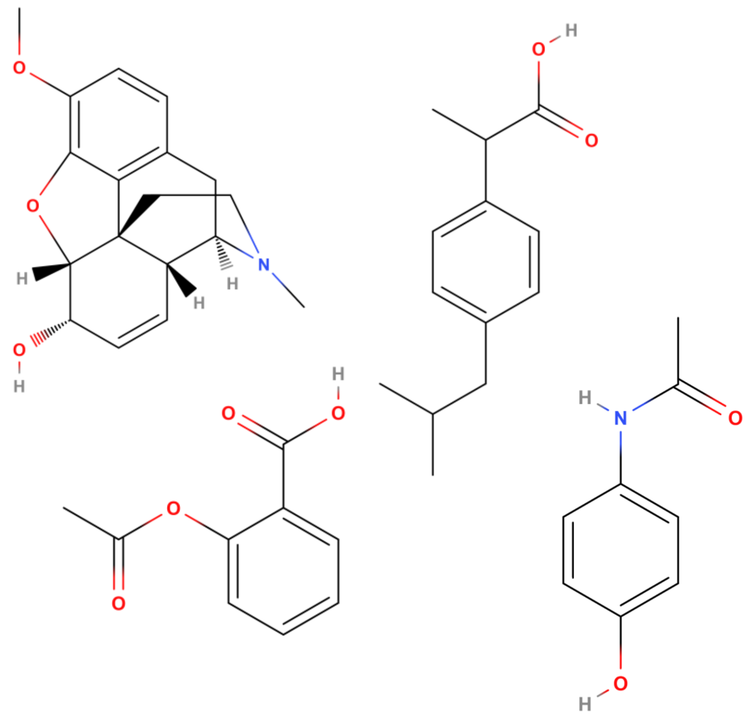
# Design in Compound Space with Machine Learning and Quantum Alchemy

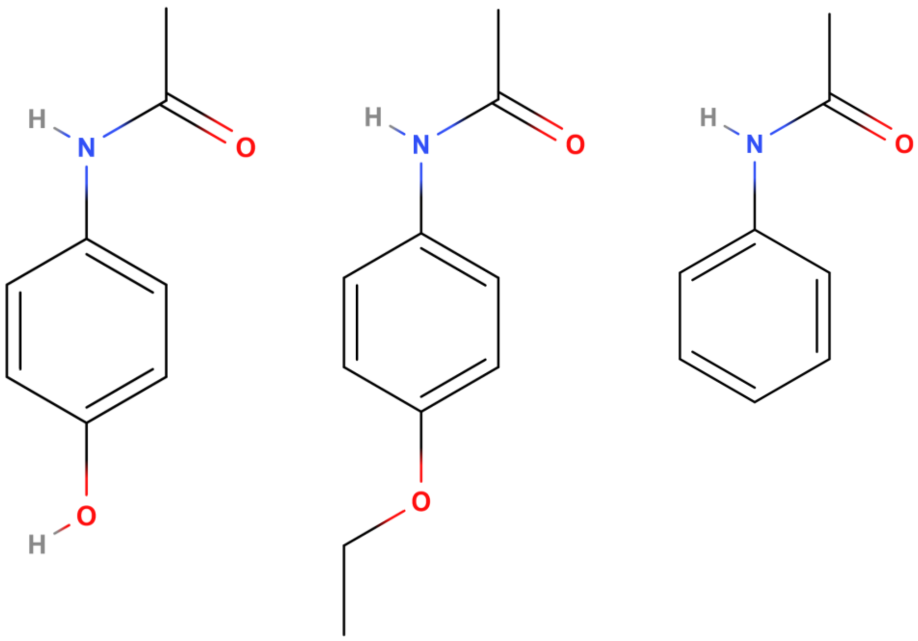
Guido Falk von Rudorff, University of Kassel

ferchault    @ferchault    nablachem.org

# Introduction

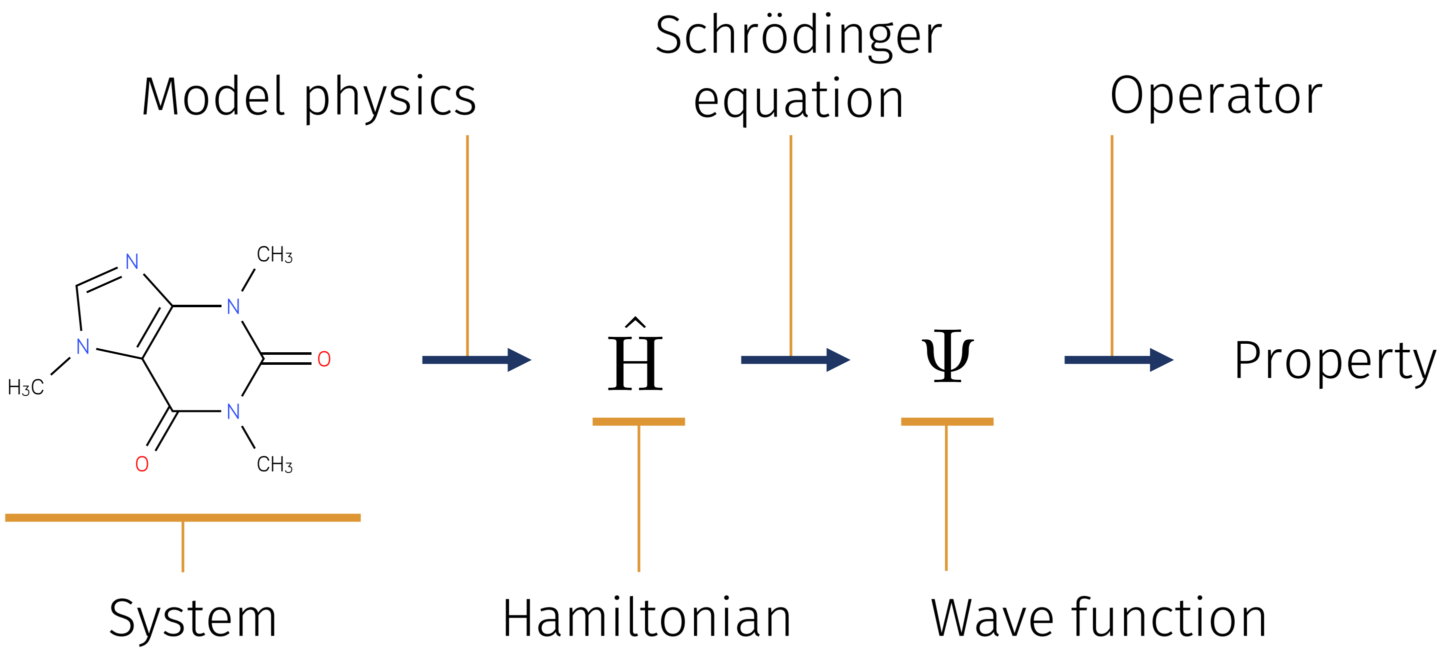


**Global Search Problem**
Which class of compounds?



**Local Search Problem**
Which particular species within that class?

# Solved?



Model physics

Schrödinger equation

Operator

System → $\hat{H}$ → $\Psi$ → Property

Hamiltonian

Wave function

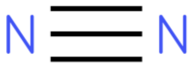
# Wave function

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

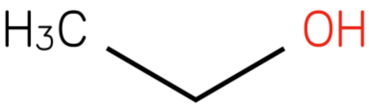
Methane

$CH_4$

$N_2$



Ethanol



Solved by approximations in computational chemistry?

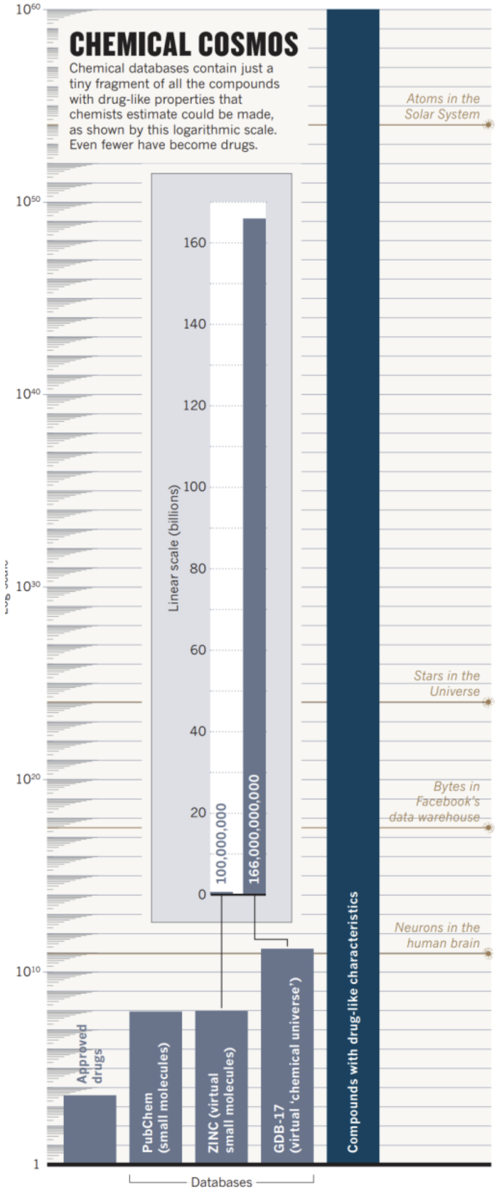
# Chemistry at scale

**Commercial databases**

- 164 million molecules
- 15k added daily

**Scale**

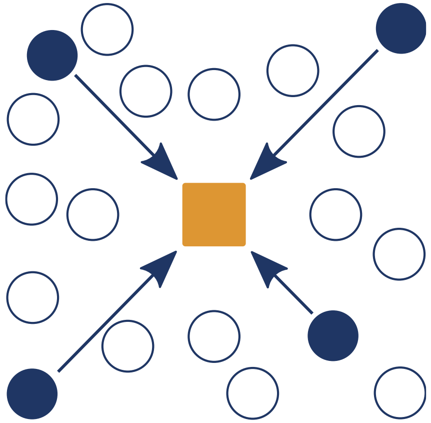
- One person: 1 million compounds/second
- 10 billion people on earth
- $10^{26}$ universe ages to go through



A. Mullard, *Nature*, 2017.

# Introduction

## Machine Learning



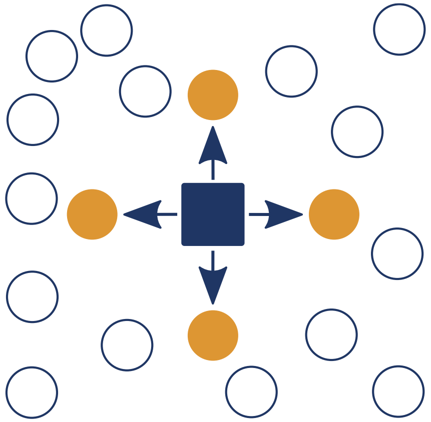
**Foundations** | Statistical modelling

**Accuracy** | Systematically improvable through data and training

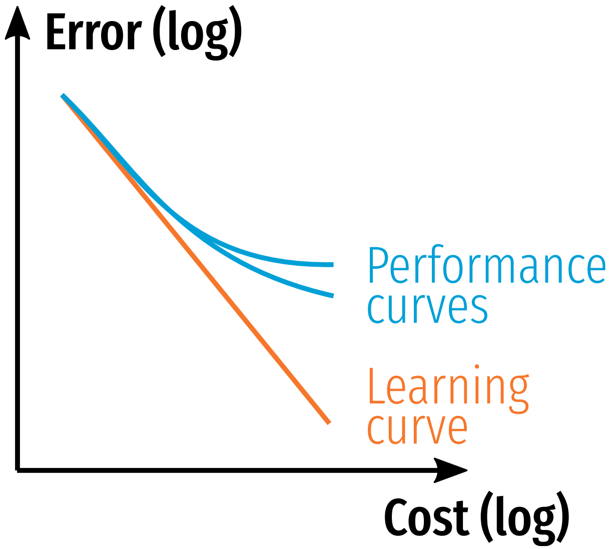
**Specialty** | Universal, scale-bridging, data-driven approach

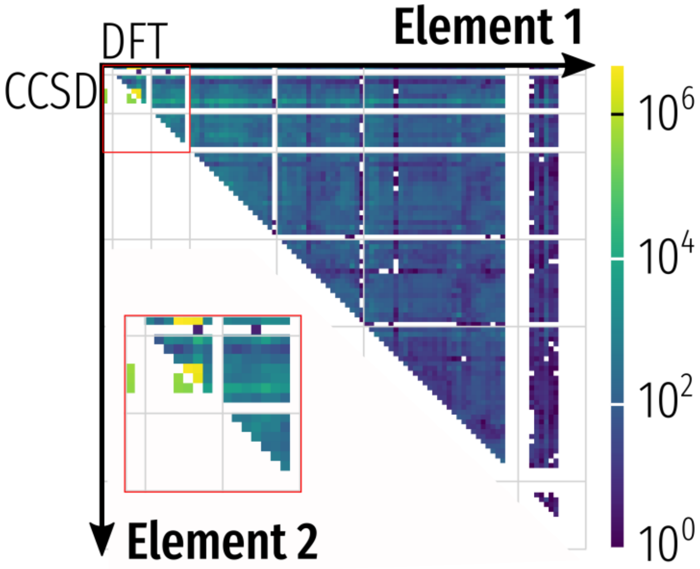
**Limitation** | Requires training data, no black box

## Quantum Alchemy



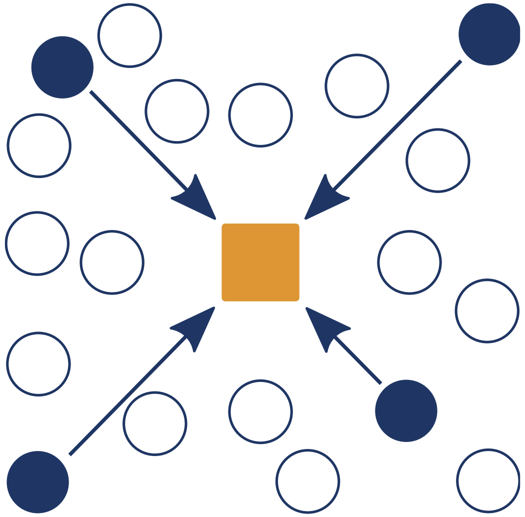
# Data availability





# Introduction: Machine Learning

## Machine Learning

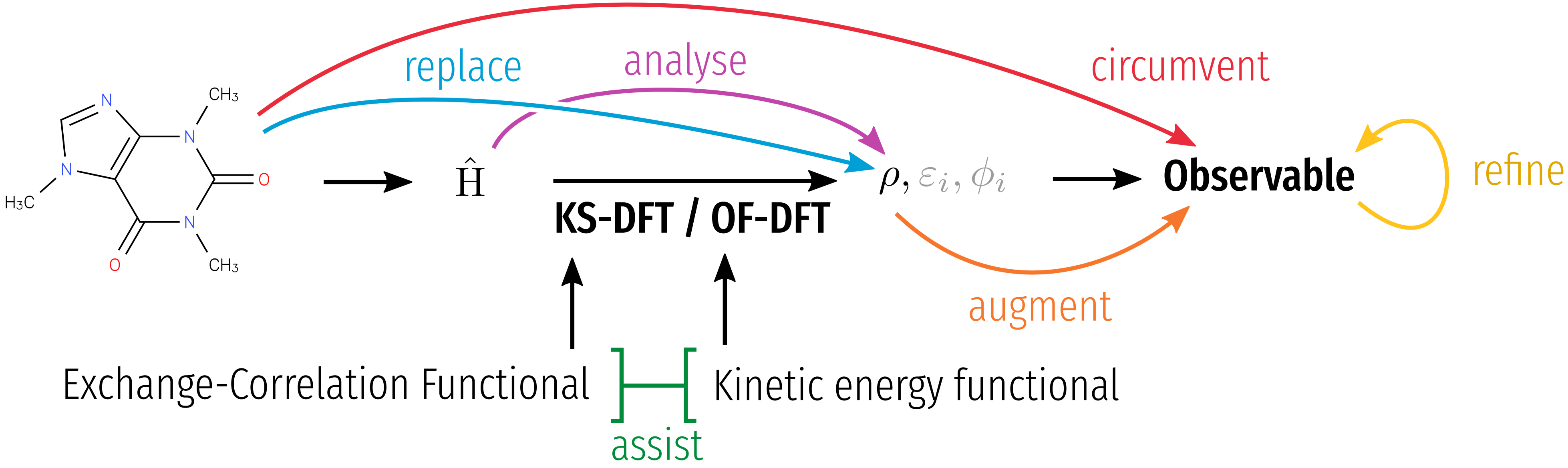


## Kernel-Ridge-Regression

- 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.

# Strategies



replace

analyse

circumvent

$\hat{H}$

**KS-DFT / OF-DFT**

$\rho, \varepsilon_i, \phi_i$

**Observable**

refine

augment

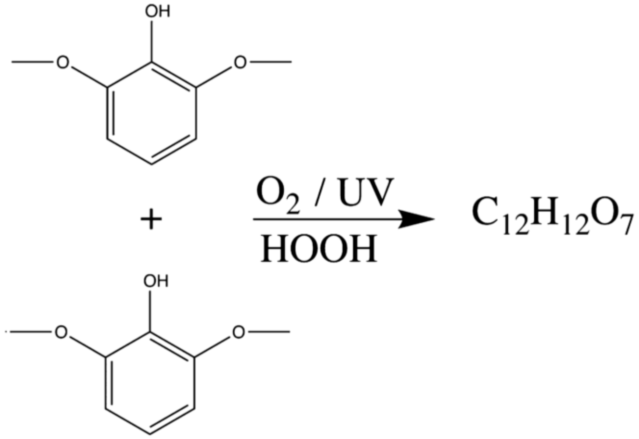
Exchange-Correlation Functional

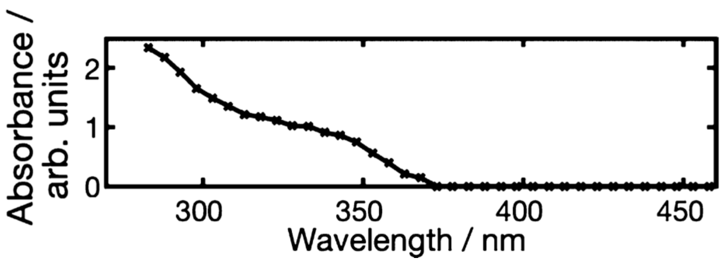
assist

Kinetic energy functional

# Identification of spectra

## Experiment

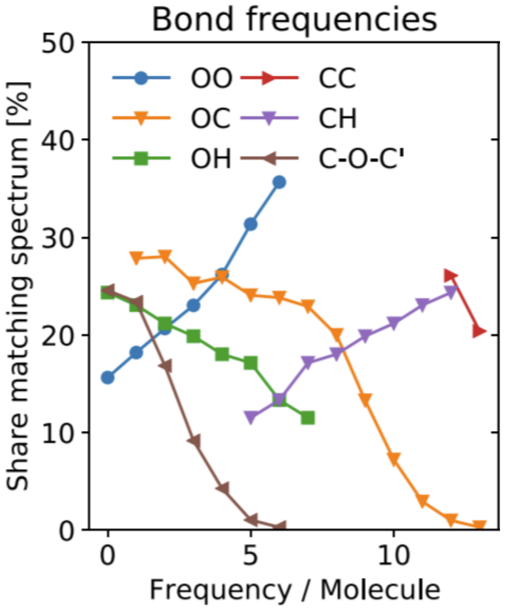




## Search space

| | |
|---|---|
| Molecular graphs: | 264 M |
| Stable molecules: | 123 M |

## Identified features

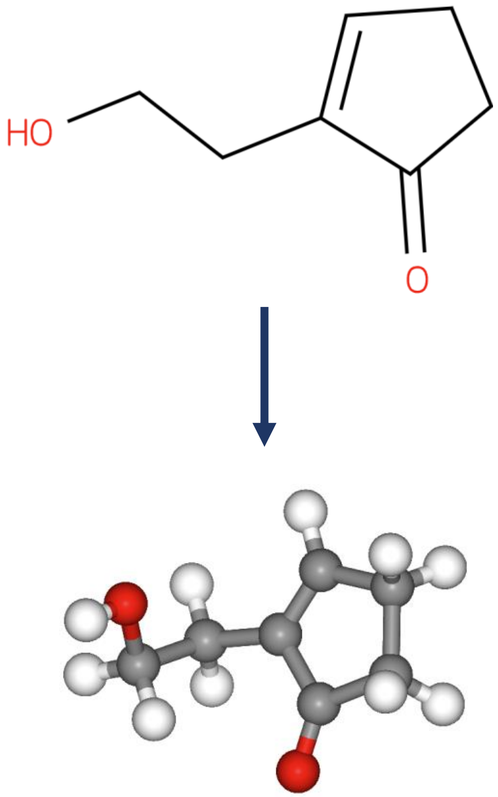


## Guide experiment

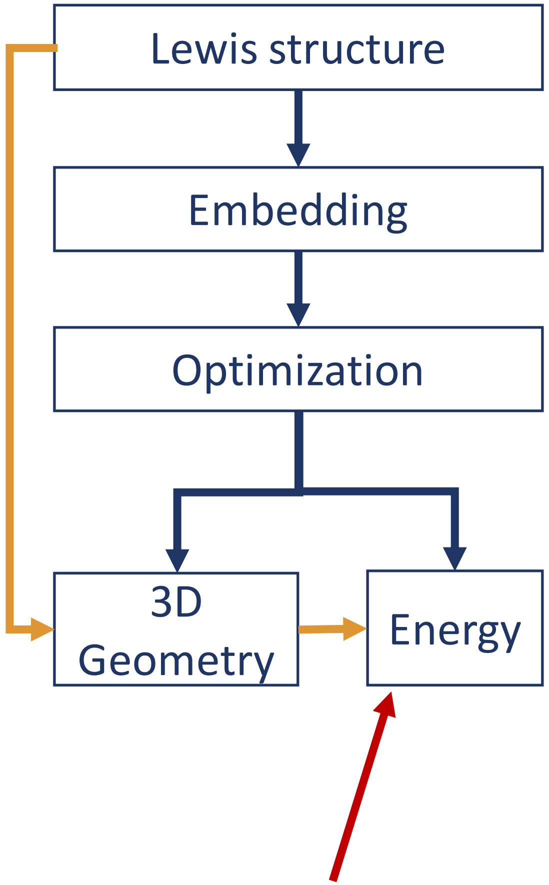
How many molecules are left and which feature to measure next?

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

# Geometry learning



Traditional / G2S



ML: 100.000x faster

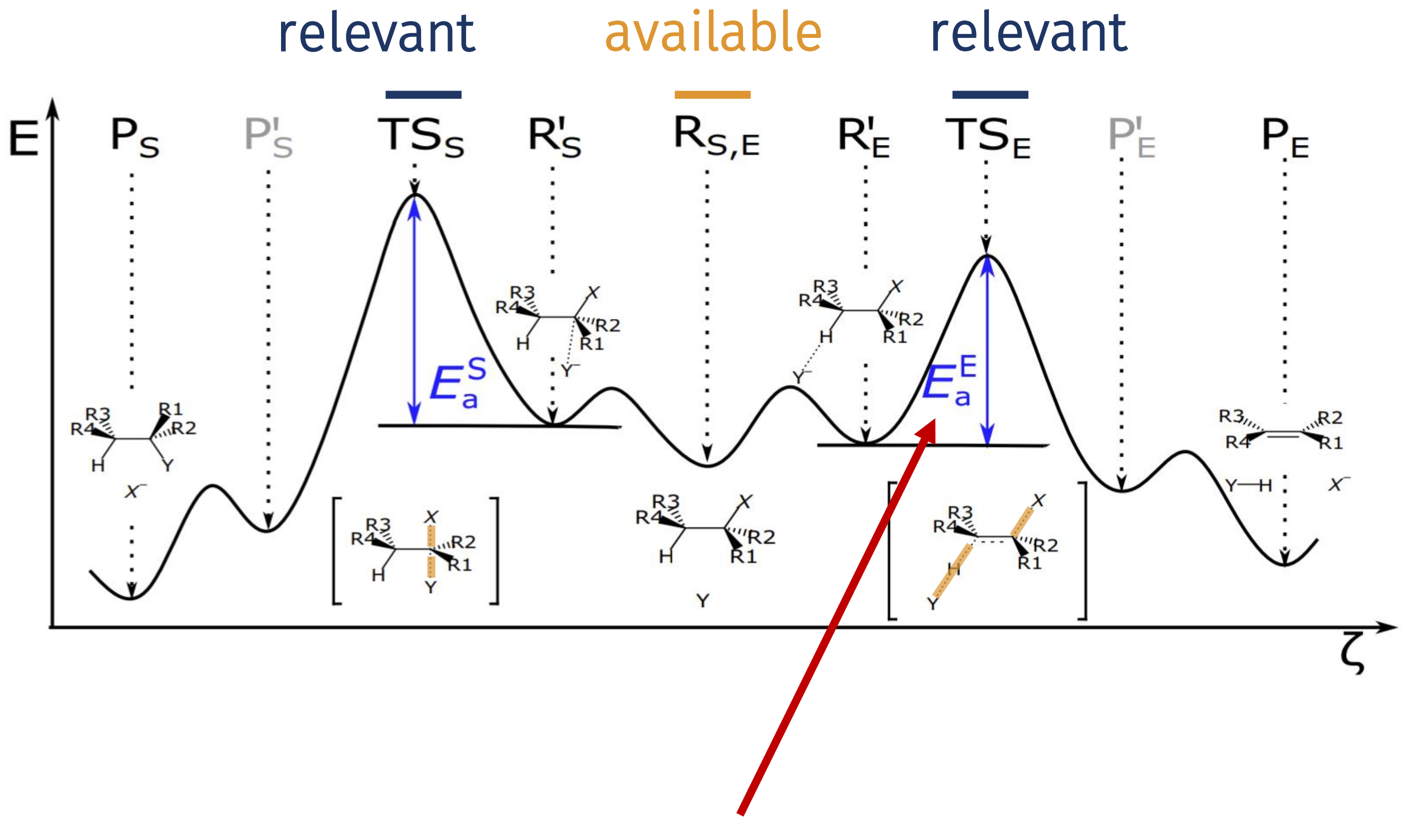
G2S

- Closer to DFT than common methods
  - Small molecules

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

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

# Reaction barrier learning



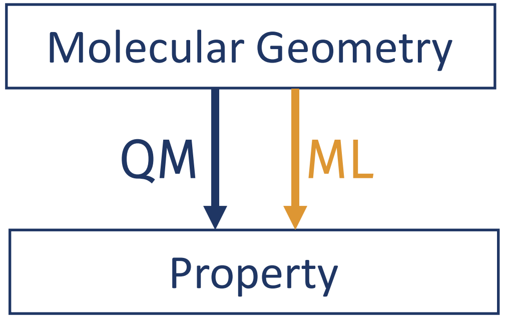
ML: 10.000x faster

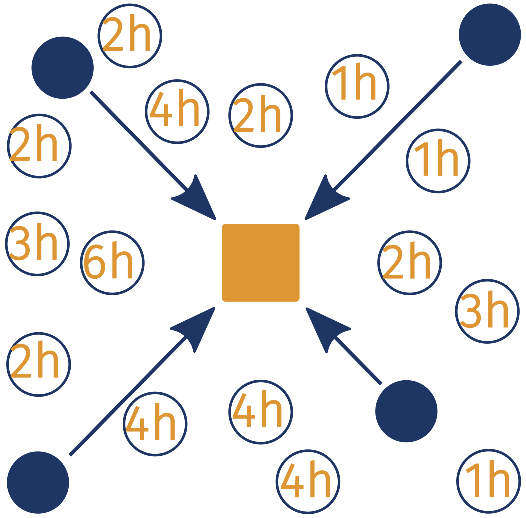
Competing reactions: E2, $S_N2$

- 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

**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**.

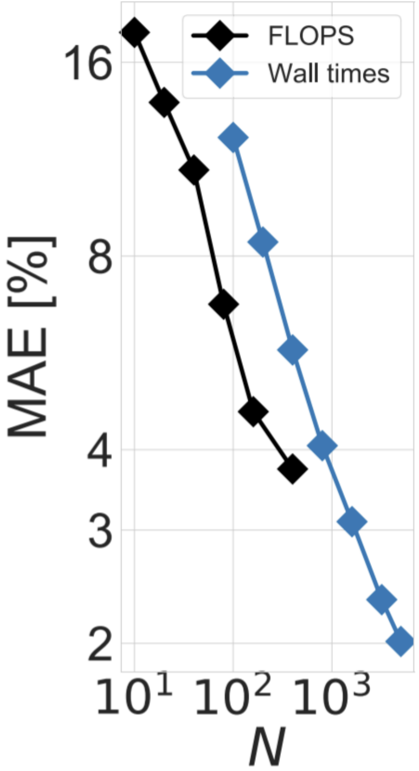
# Computational cost





## Computational effort as molecular property

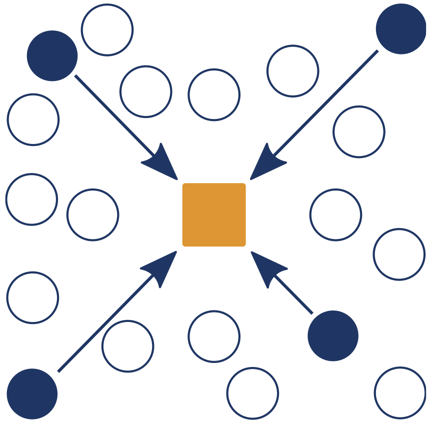
- Improves models
- Accuracy depends on problem
  - Single points: 2%
  - Transition state search: 25%
  - Geometry optimisations: 40%



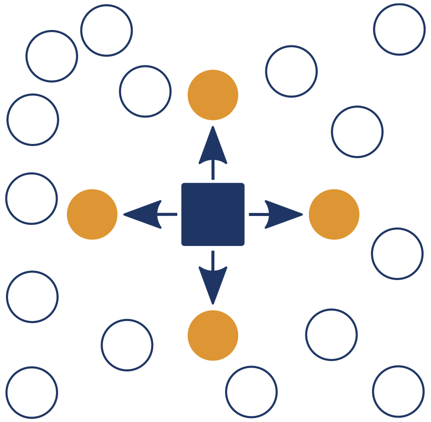
S.N. Heinen, M. Schwilk, **GFvR**, O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.* 2020.

# Introduction

## Machine Learning



## Quantum Alchemy



**Foundations** | Perturbation theory

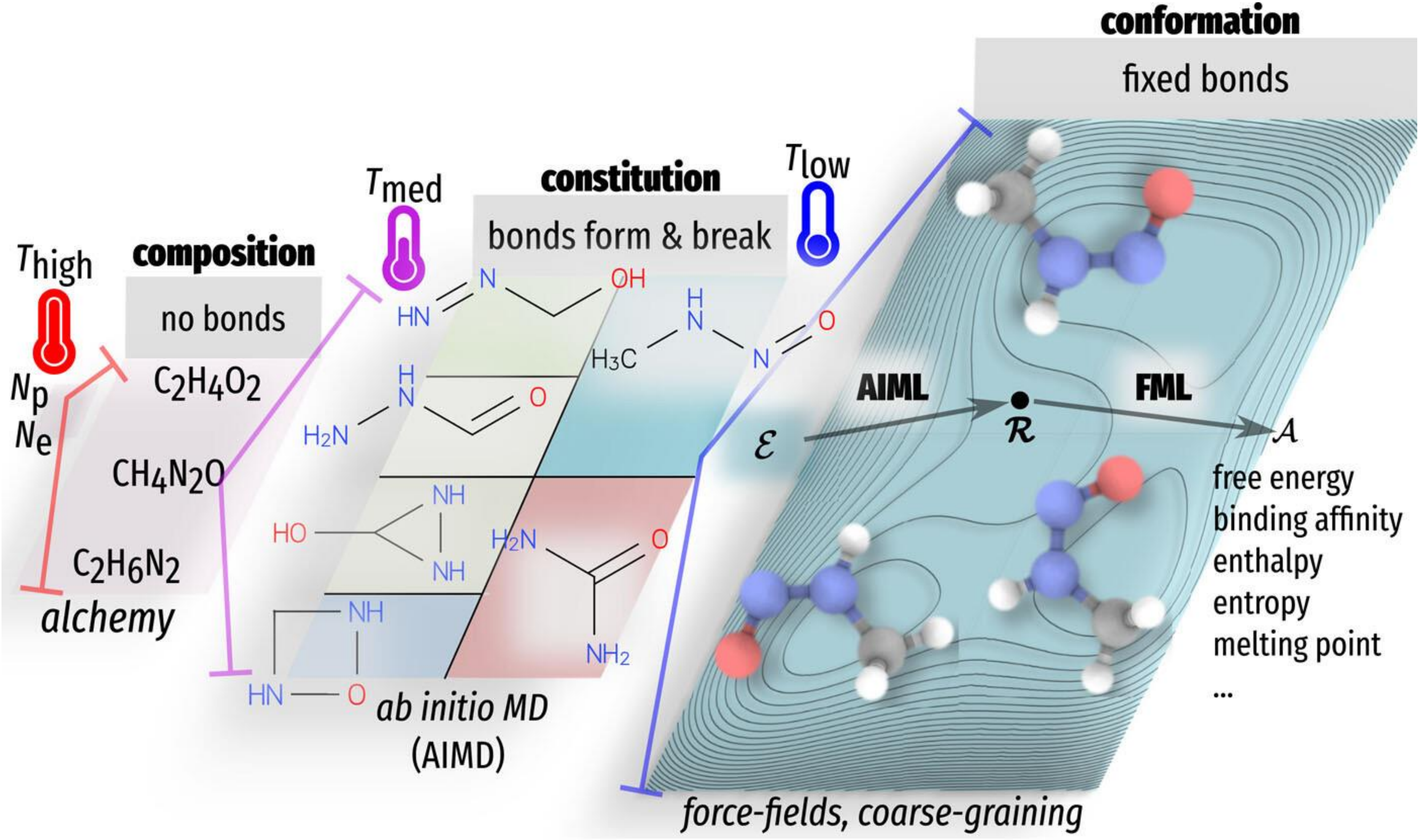
**Accuracy** | Systematically improvable through higher orders terms

**Specialty** | Combinatorial scaling with chemical diversity

**Limitation** | Finite range in chemical space

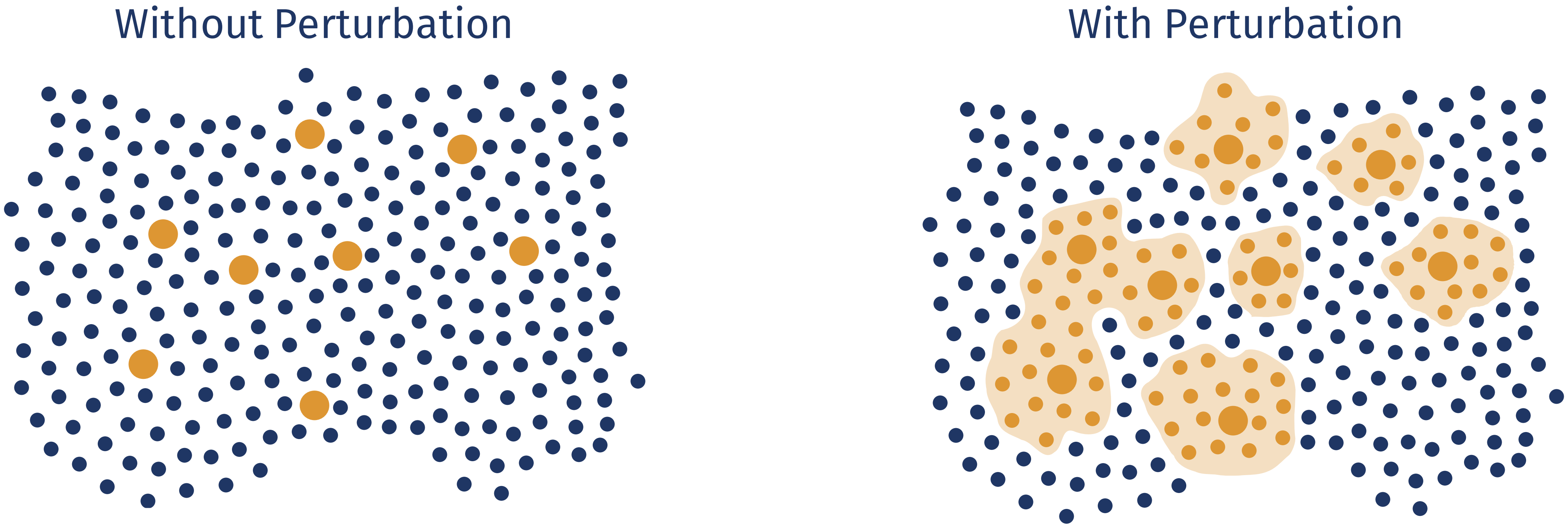
Joseph Wright, 1771

# Discrete points?



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

# Introduction



Without Perturbation

With Perturbation

Systems/Molecules

- • Any
- ● Known
- ● Approximated

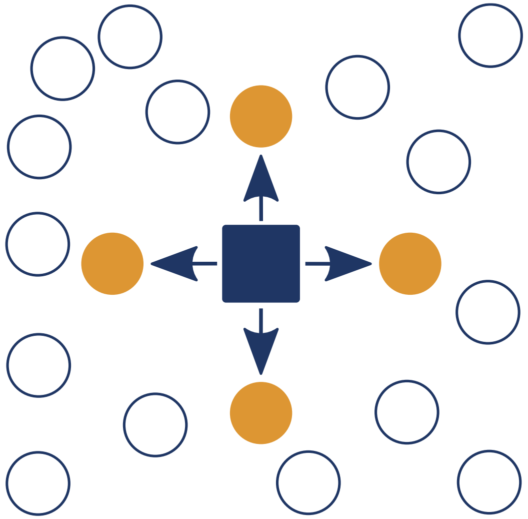
# Introduction

**Perspective shift**
Few highly accurate calculations
instead of many intermediate ones

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

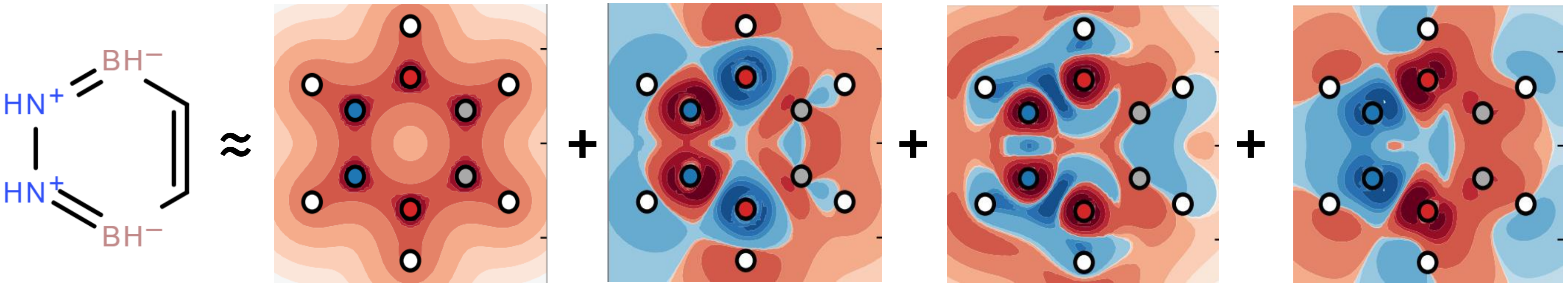
# Quantum Alchemy

## 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

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

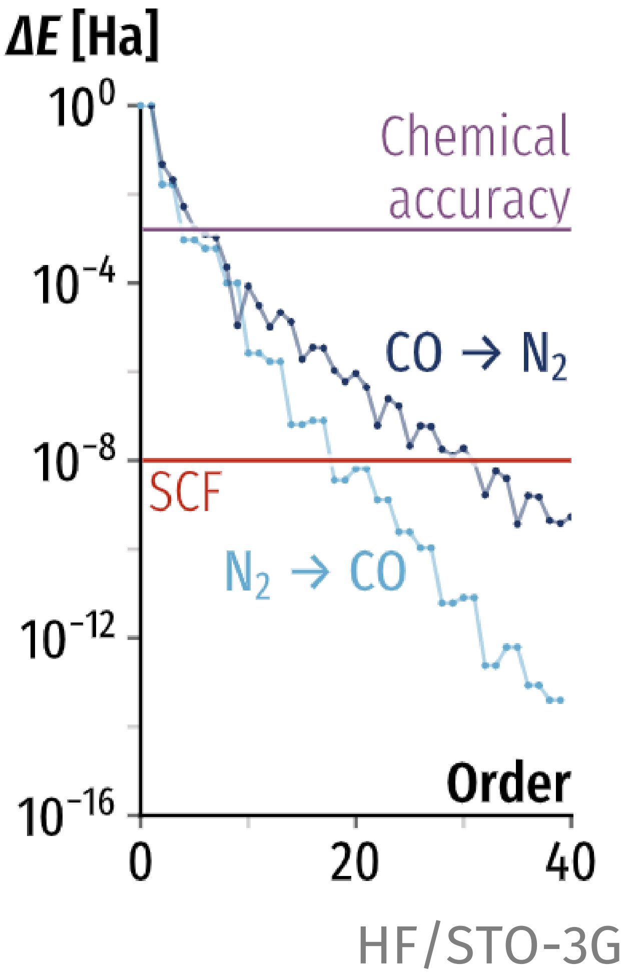
- Gives consistent energies, densities, forces, ...
- Uses the same derivatives for all predictions

AT THIS POINT, YOU'RE PROBABLY THINKING, "I LOVE THIS EQUATION AND WISH IT WOULD NEVER END!"
WELL, GOOD NEWS!

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

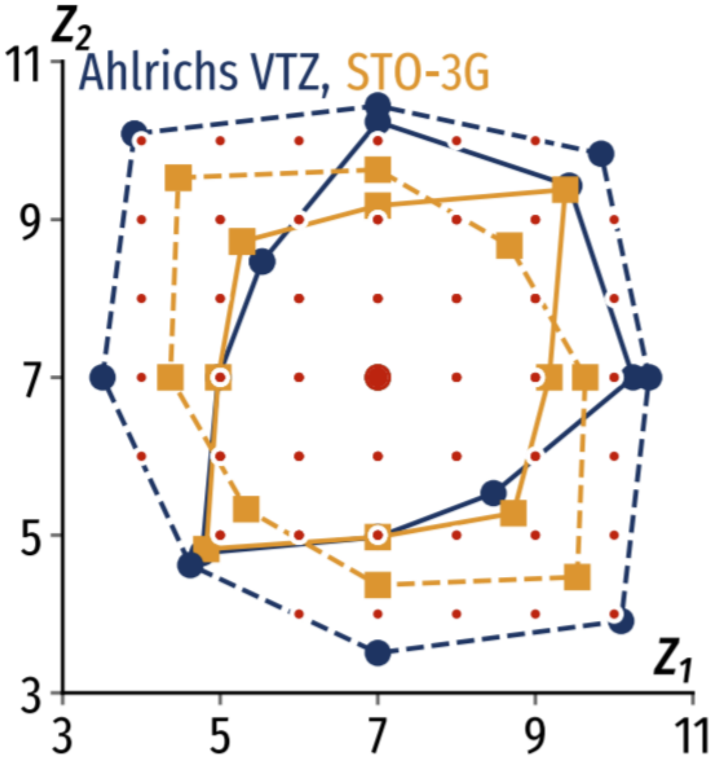
xkcd.com/2605

# Convergence



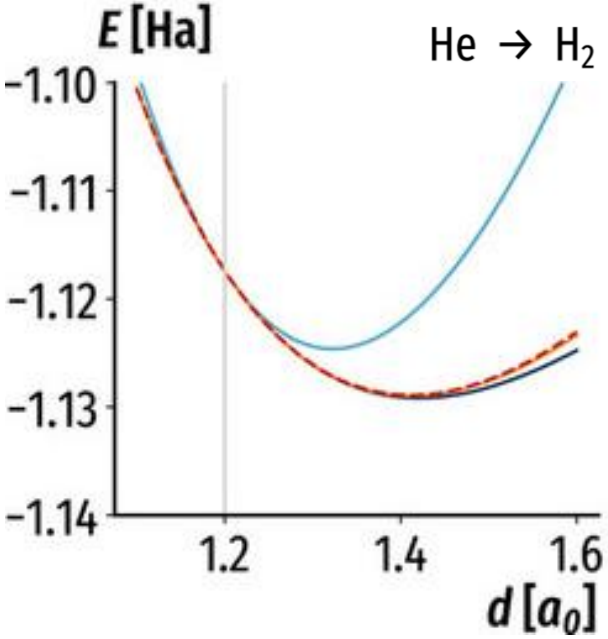
### Taylor expansion

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



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

# Geometry relaxation



**Taylor expansion**

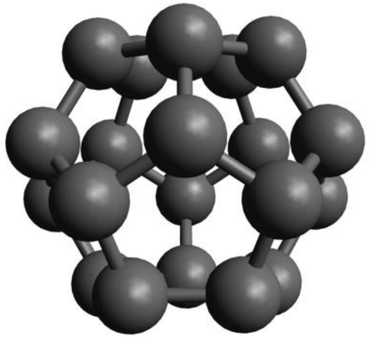
- Large changes still converge (more slowly)
- Geometric response can be recovered

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

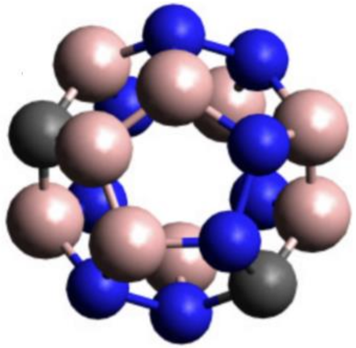
# Covalent interactions

**Scaling with chemical space**

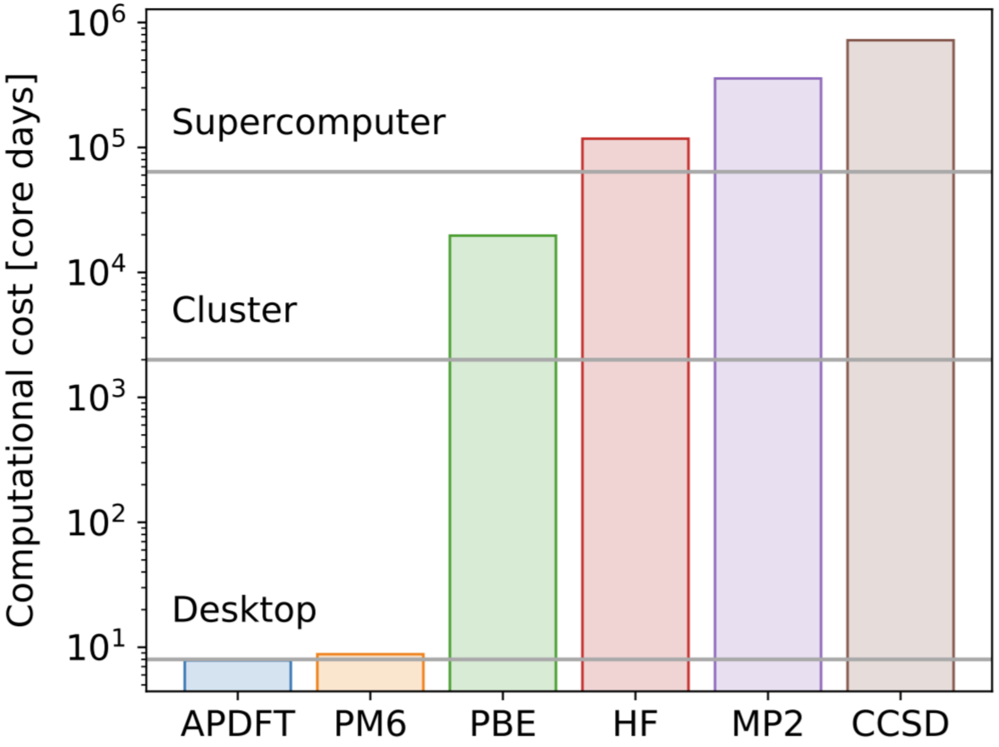
- 1 derivative for second order
- 5 derivatives for third order



$C_{20}$



$3.1 \cdot 10^6$ targets



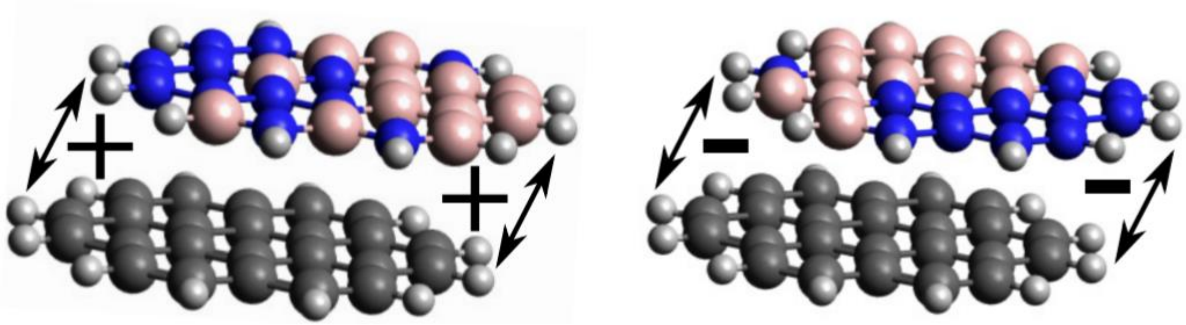
QA: 80.000x faster

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

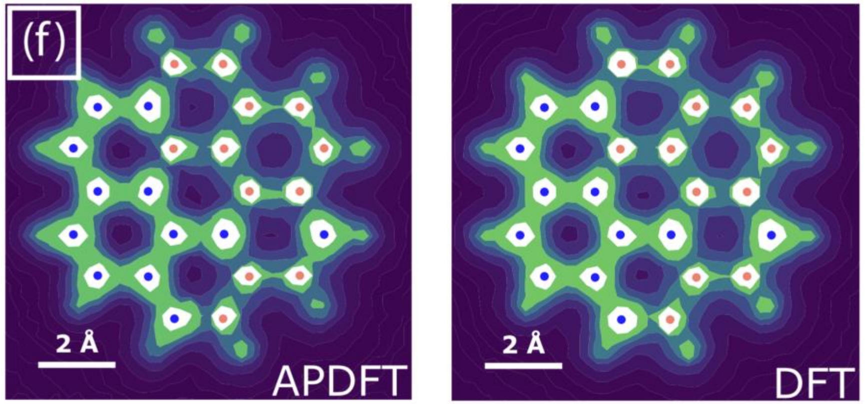
# Non-covalent interactions

**BN-doped coronene dimer**

- Identify most/least attractive doping pattern
- Design case



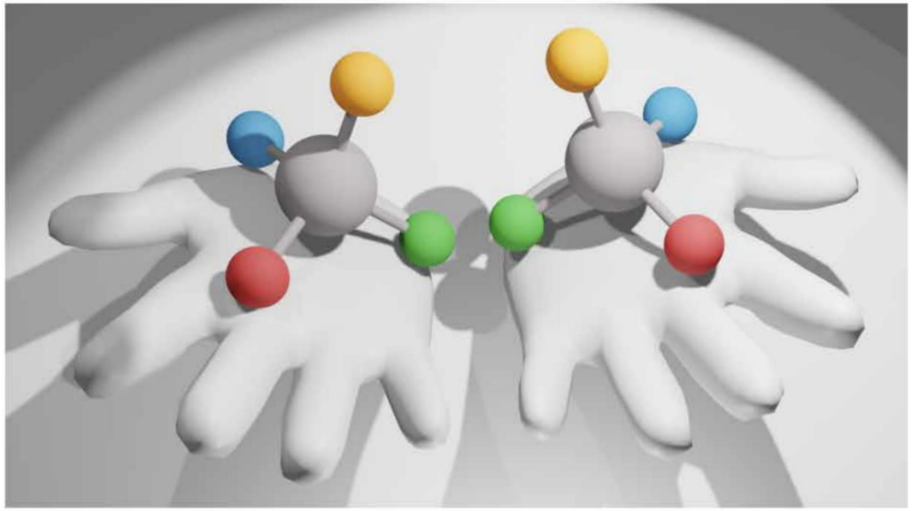
QA: 20.000x faster

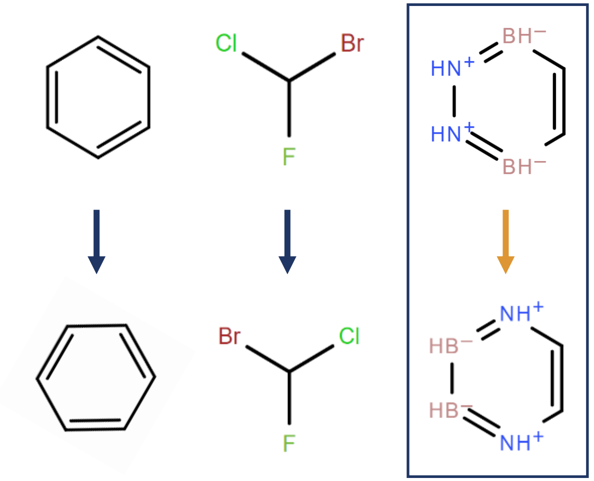


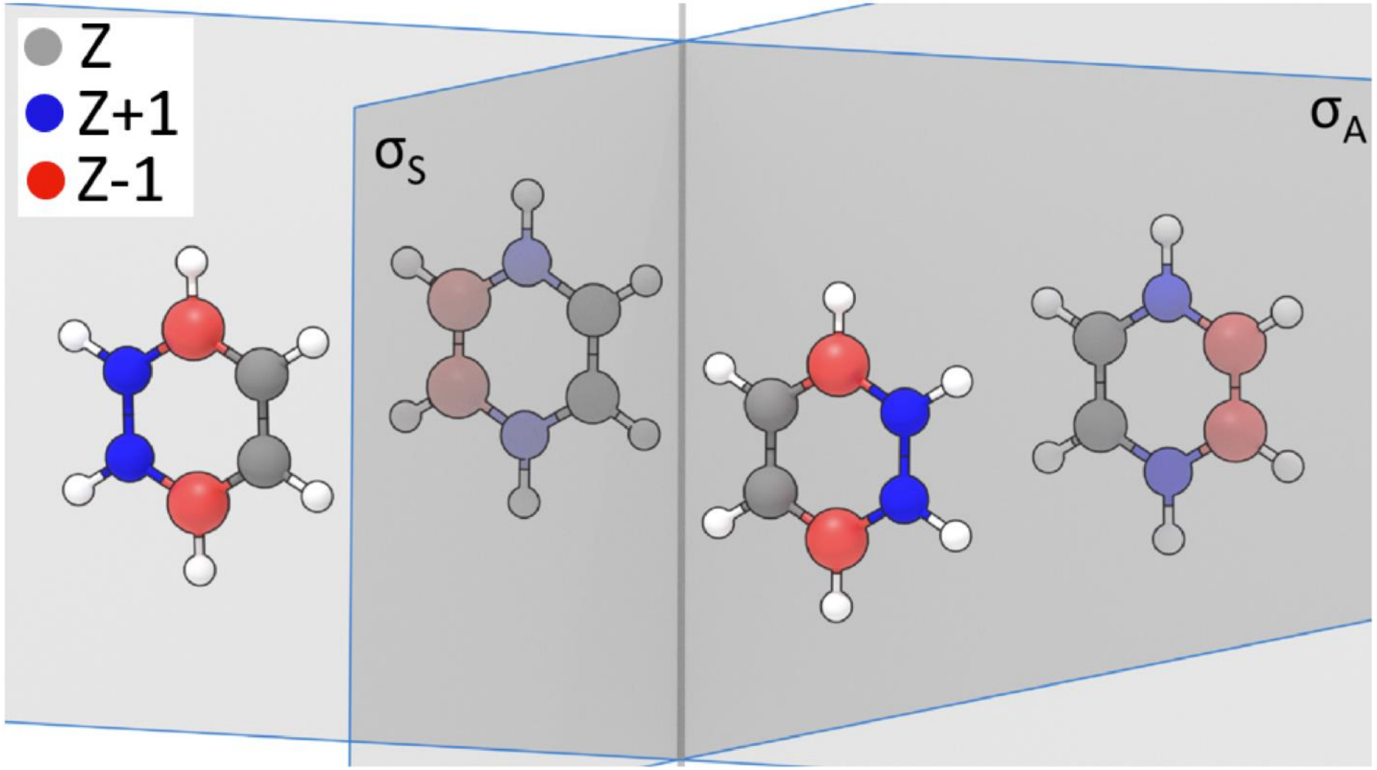
$2.8 \cdot 10^{10}$ targets

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

# Alchemical enantiomers





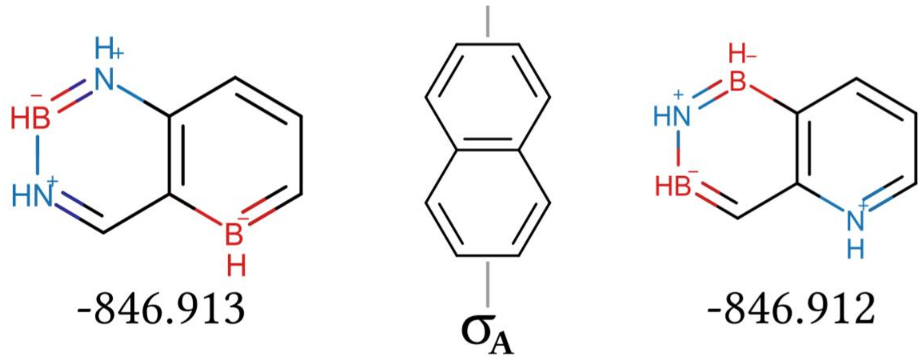


GFvR, O. A. von Lilienfeld, *Science Adv.* 2021.

# Alchemical enantiomers

## Fundamentally new symmetry

Electronic energy only



## Bond energy rules

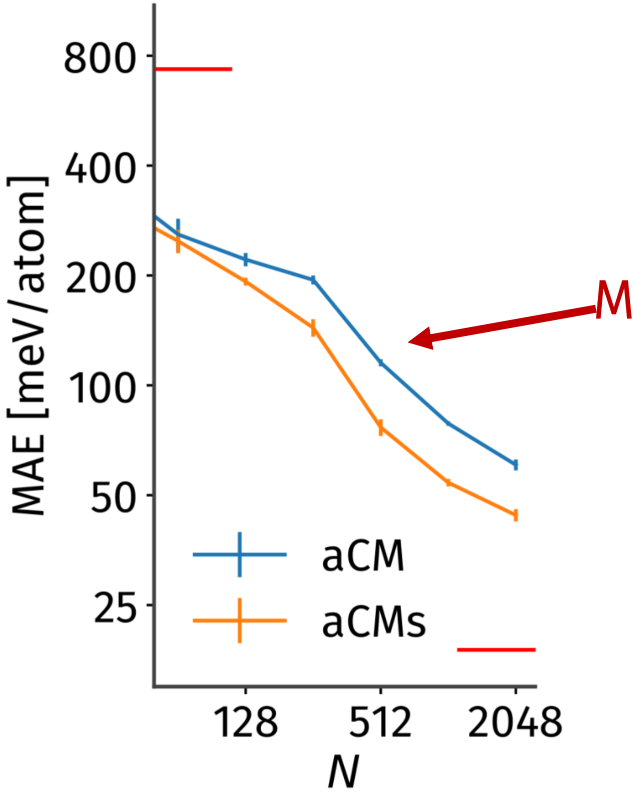
Consecutive Elements

Q R S

B C N

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

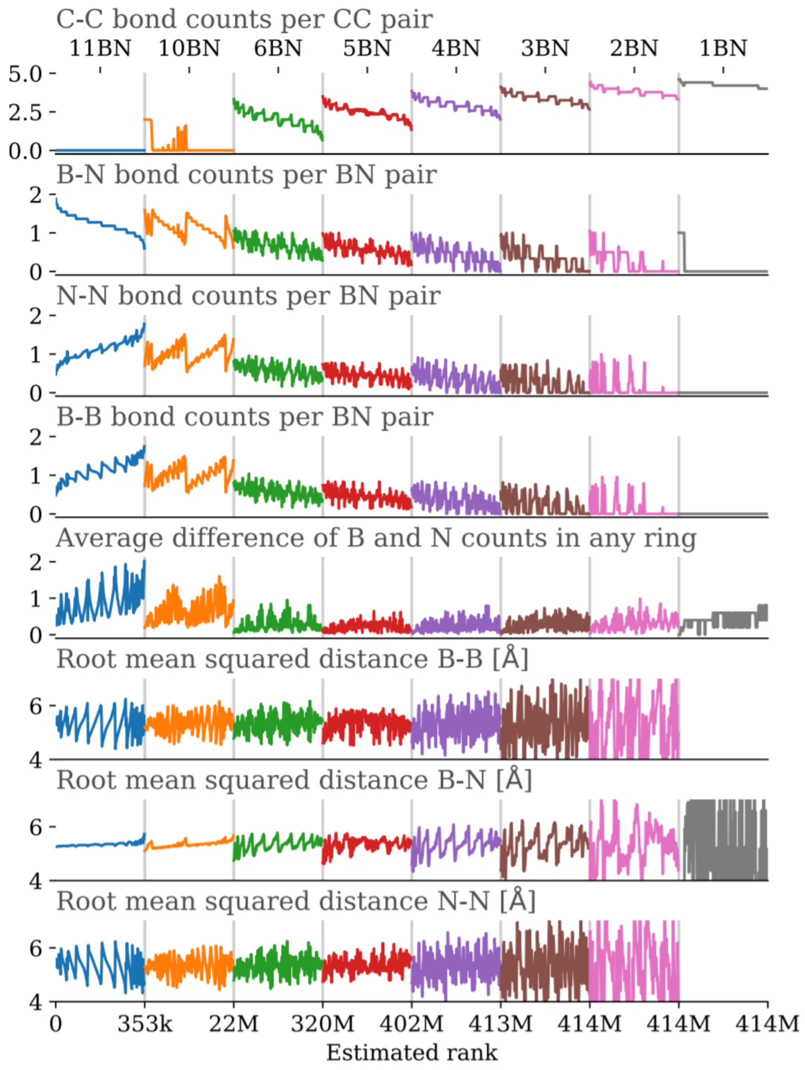
## Speed up machine learning

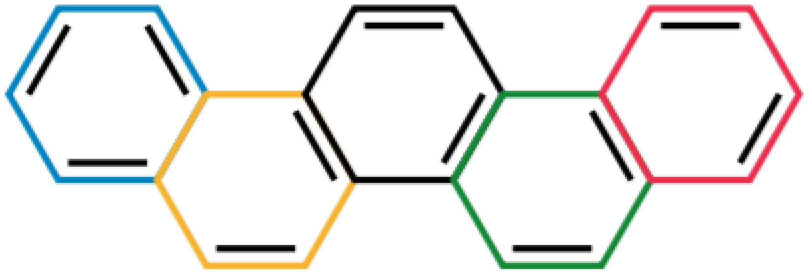


ML+QA: half the data

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

# Alchemical enantiomers





x 414 M

**Design rules** in order of **decreasing strength**

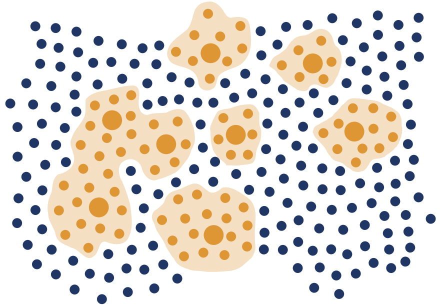
- Add BN pairs
- Maximize CC bonds
- Substitute sites shared between rings
- Maximize BN bonds
- Avoid N substitutions on rings sharing a larger amount of bonds with other rings
- Balance BN substitutions in each ring

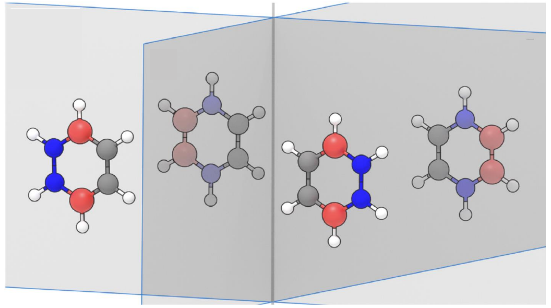
QA: Millions at once!

Not a single QM calculation required!

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

# Summary





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

Symmetries | Reducing (“folding”) search space

Constraints? | Exclude regions of interest

Differentiable Chemistry? | Arbitrary derivatives

Representations? | Better data efficiency

Ensembles? | Derivatives on dynamic observables

**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

# 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

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

Hellmann-Feynman theorem

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

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

# Quantum Alchemy

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**.