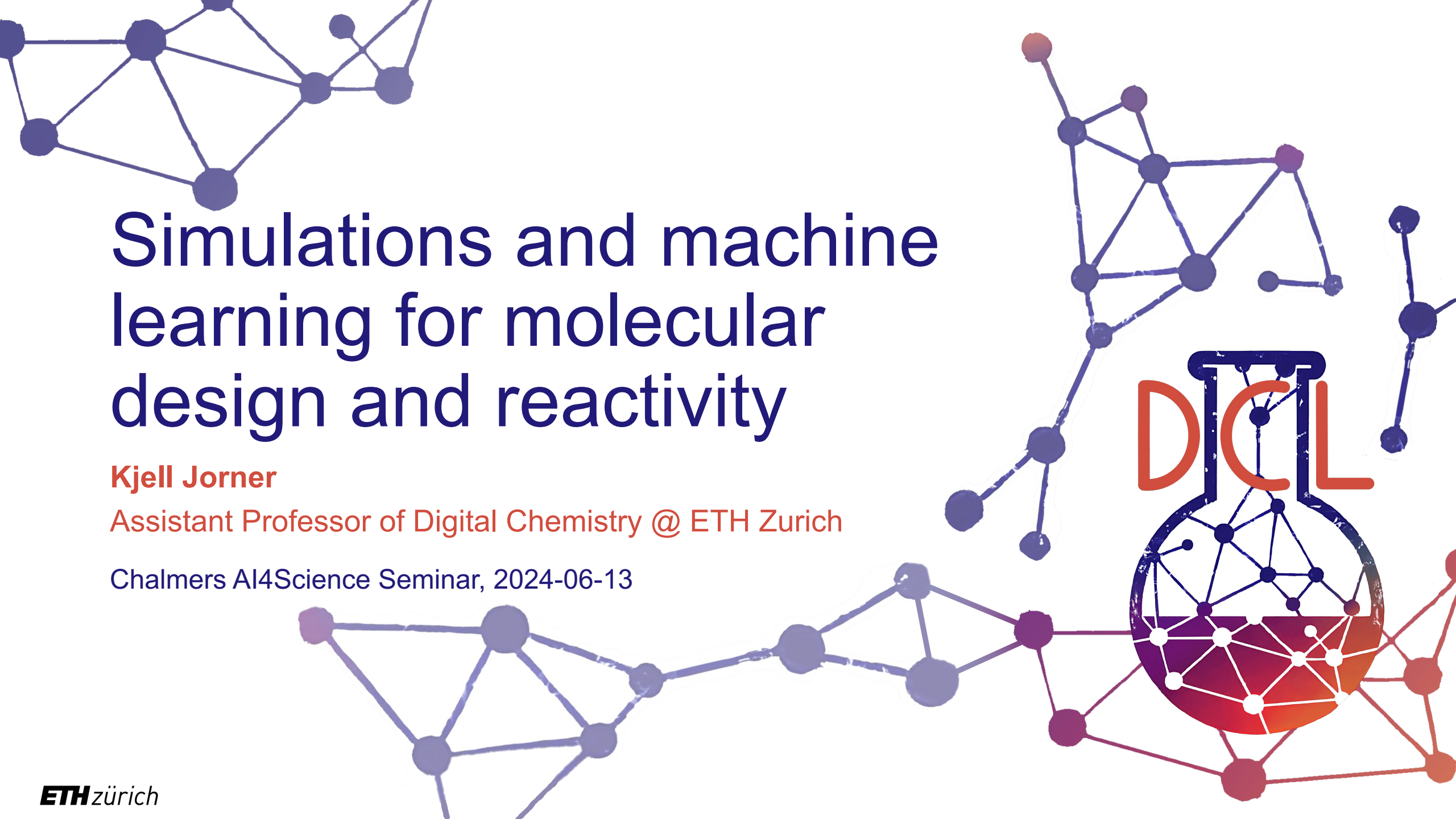
Simulations and machine learning for molecular design and reactivity

Kjell Jorner

Assistant Professor of Digital Chemistry @ ETH Zurich

Chalmers AI4Science Seminar, 2024-06-13



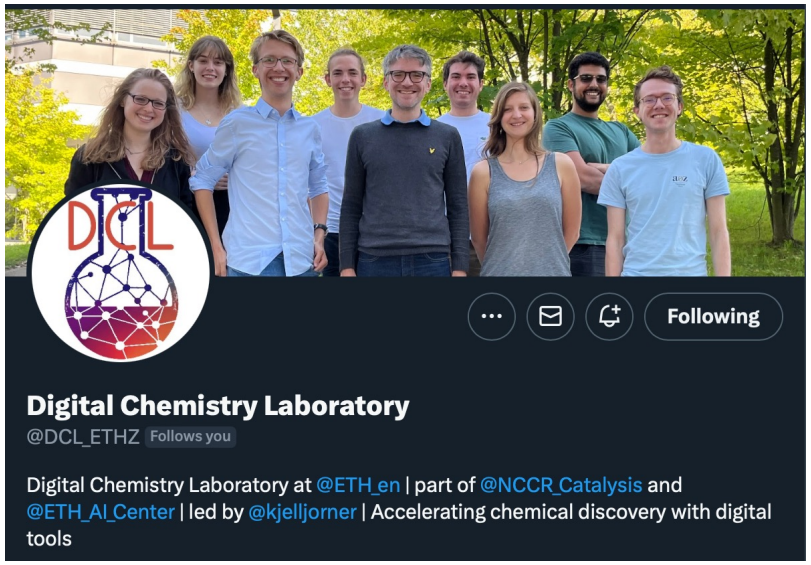
The Digital Chemistry Laboratory

Department of Chemistry and Applied Biosciences (D-CHAB)
Institute of Chemical and Bioengineering (ICB)

Principal investigator
Prof. Dr. Kjell Jorner

PhD students
Lauriane Jacot-Descombes
Giustino Sulpizio
Stefan Schmid
Luca Schaufelberger
Franziska Weißbach

Co-supervisor
Vignesh Ram Somnath (w. Krause)
Riccardo De Santi (w. Krause, He)



@DCL_ETHZ

Open position in Computer-aided Catalyst Design

- Closes **15.06** (on Saturday)
- Ideally computational science (chemistry) & machine learning profile

PhD position in computer-aided catalyst design

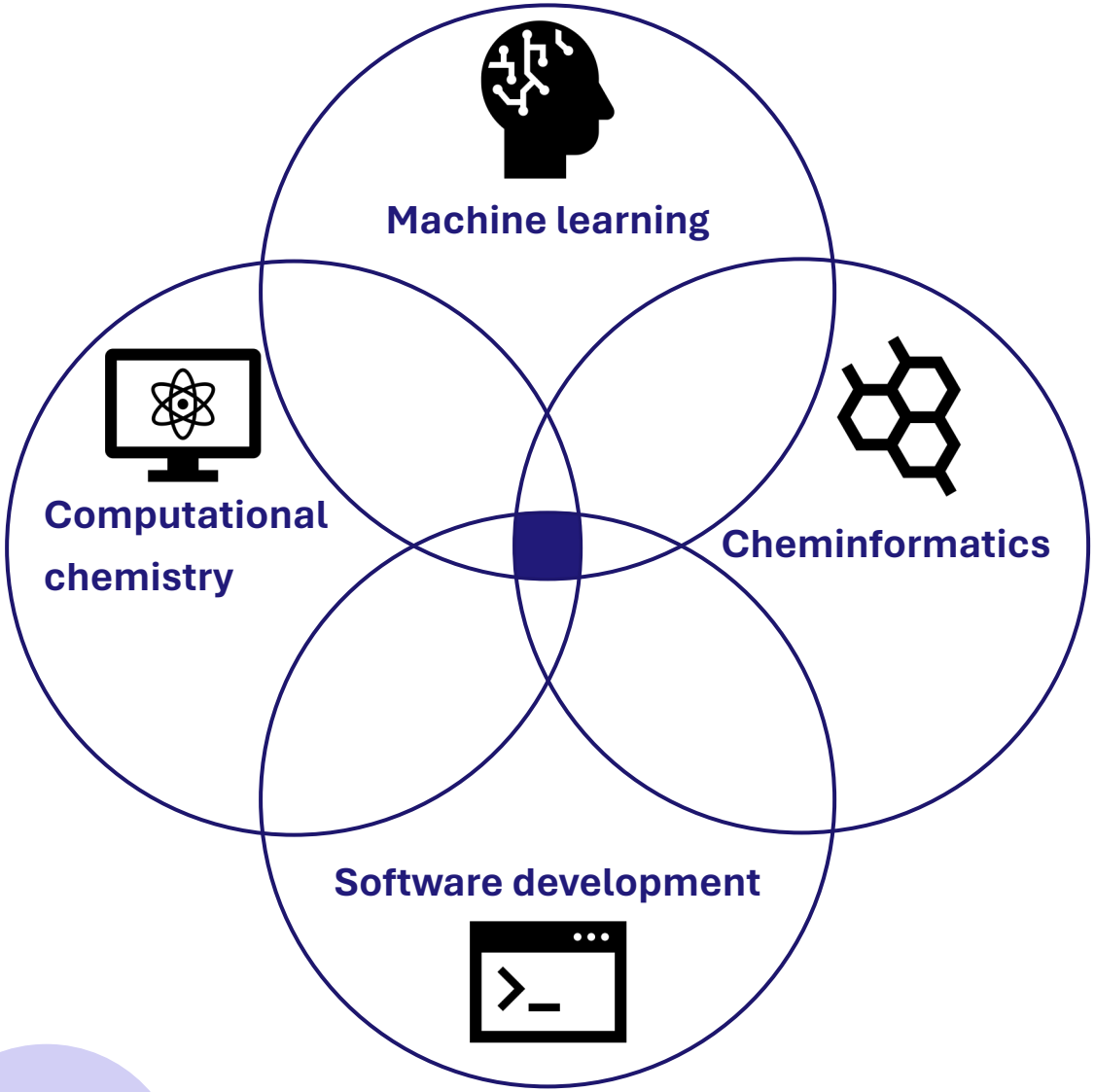
100%, Zurich, fixed-term

Print

The Digital Chemistry Laboratory is led by Prof. Dr. Kjell Jorner at the Institute of Chemical and Bioengineering, within the Department of Chemistry and Applied Biosciences at ETH Zurich. Our mission is to accelerate chemical discovery using digital tools. We predict chemical reactivity and molecular properties using the tools of machine learning, computational chemistry, and cheminformatics. Our ultimate goal is the computer-aided design of molecules and especially catalysts.

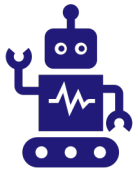
https://jobs.ethz.ch/job/view/JOPG_ethz_9XnHx97EHSgGA7ypih

Digital Chemistry



Lab chemist

High-throughput experimentation



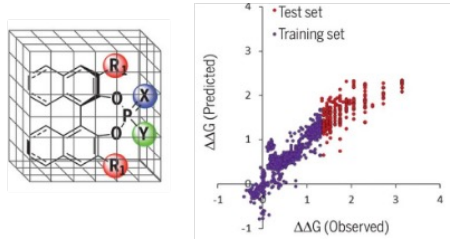
Self-driving labs

Mission: Accelerate chemical discovery with digital tools

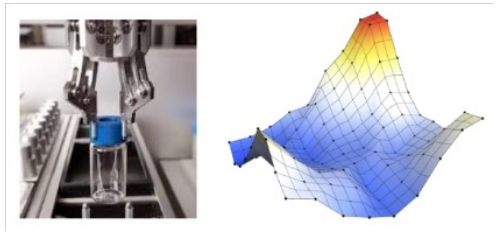
Strategic core research

Reactivity and catalysis

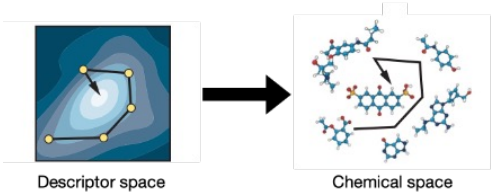
Reaction outcome prediction



Reaction optimization



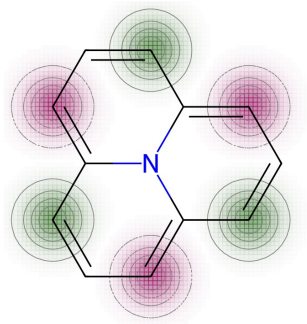
Catalyst & reaction design



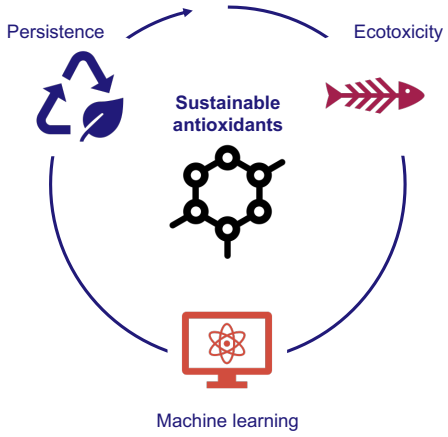
Opportunity-driven

Computer-aided molecular design

Organic electronic materials

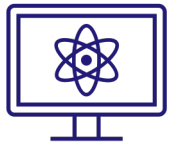


Safe and sustainable by design

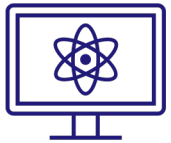


NCCR Catalysis

Our niche: Models in the low-data regime



Simulations



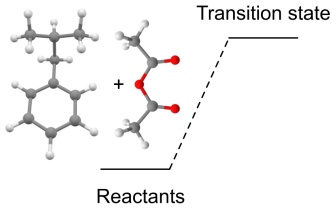
Chemistry-informed machine learning

Machine learning

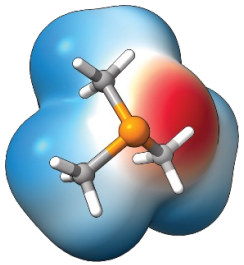
No data

Lots of data

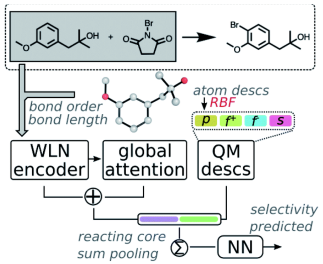
Mechanistic DFT



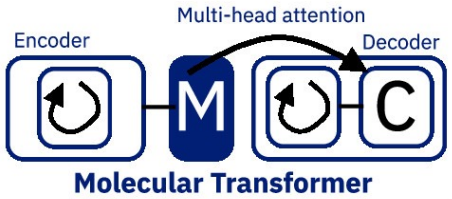
Chemical descriptors



Chemistry-augmented deep learning



Deep learning

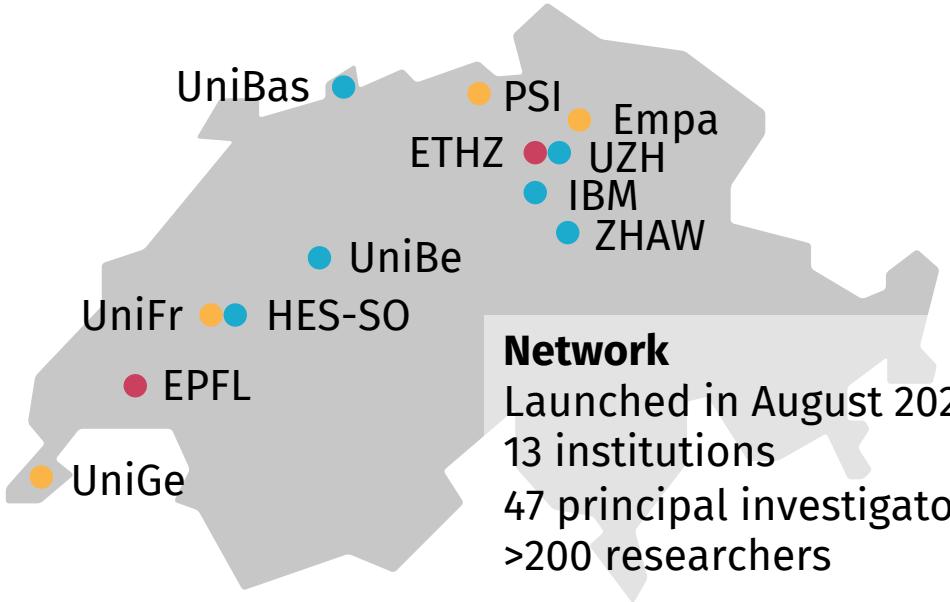


NCCR Catalysis

Swiss National Science Foundation

The National Centres of Competence in Research (NCCRs) are a funding scheme of the Swiss National Science Foundation

Research themes
Small molecule activation
Fine chemicals
from renewable platforms
Advanced tools
Digitalization
Implementation



Network
Launched in August 2020
13 institutions
47 principal investigators
>200 researchers

Total investment
32 MCHF, 4 years

Prof. Javier Pérez-Ramírez
Director
ETH zürich

Prof. Jérôme Waser
Co-Director
EPFL

nccr-catalysis.ch NCCR Catalysis @NCCR_Catalysis @nccr_catalysis

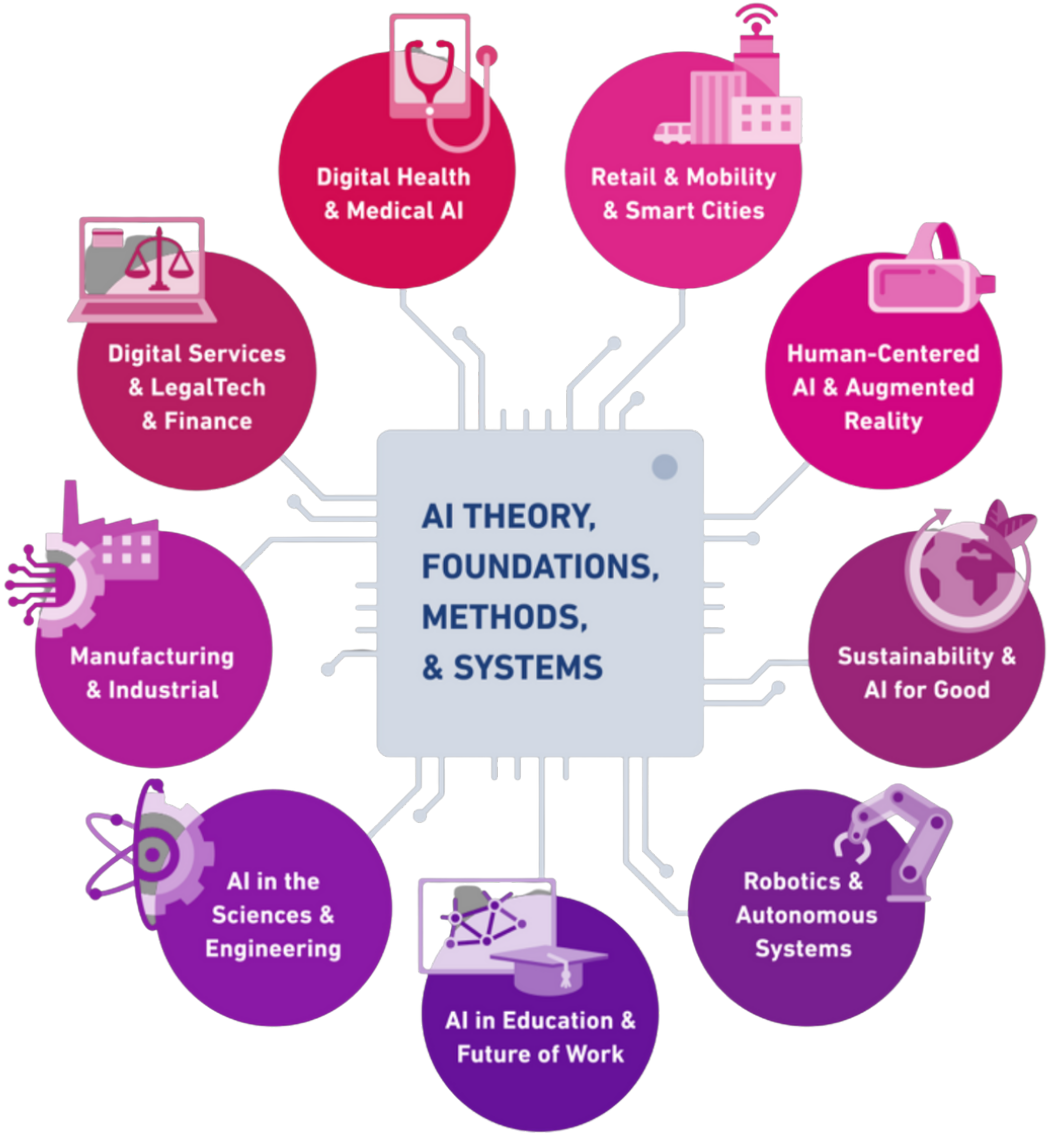
ETH AI CENTER

114
Faculty members

41
Fellows

PD Dr. Alexander Ilic
Executive Director

Prof. Andreas Krause
Chair of Steering Committee
and Core Faculty Member



*As **ETH's central hub for artificial intelligence**, we bring together researchers of AI foundations, applications, and implications across all departments. We foster research excellence, industry innovation, and AI entrepreneurship to **promote trustworthy, accessible, and inclusive AI systems**.*

ai.ethz.ch

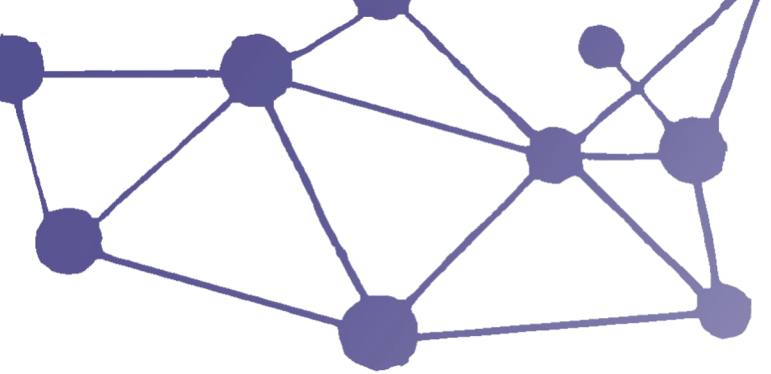
eth_ai_center

@ETH_AI_Center

ETH AI Center

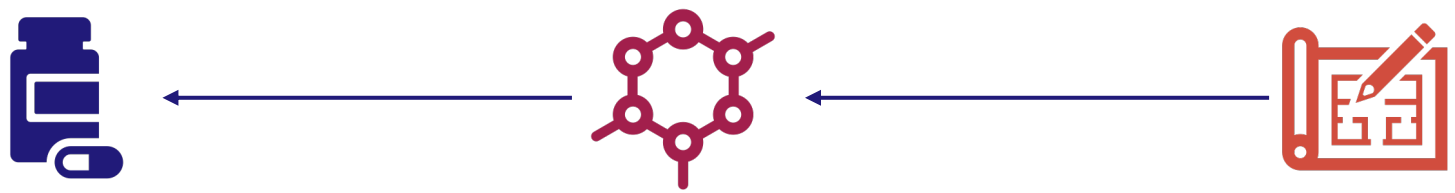
Outline

- Computer-aided molecular design
- Molecular design with Hückel theory
- Designing singlet fission materials with uncertainty-aware genetic algorithms
- Reaction design with reactive force field methods



Computer-aided molecular design

Principle of computer-aided molecular design



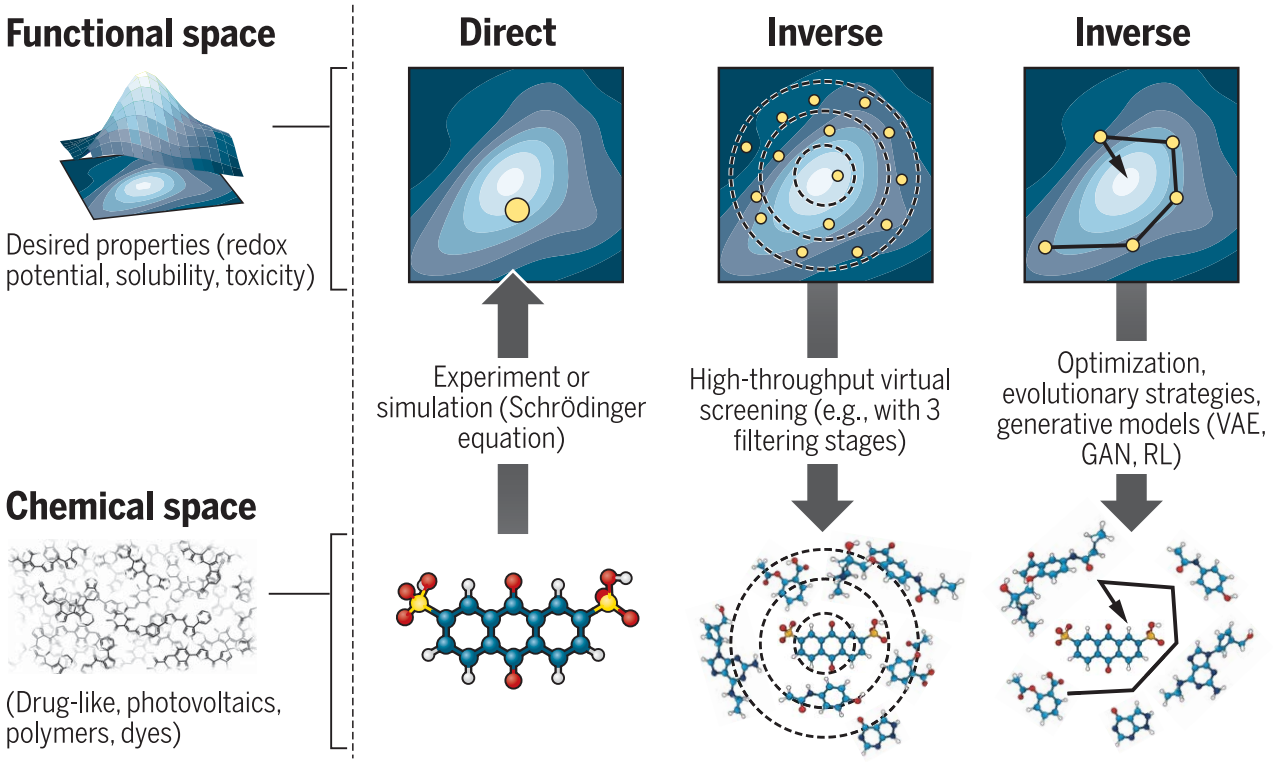
Drug against disease

Active pharmaceutical ingredient

Human expertise
Experimental iteration
Computational tools
“AI” design methods

Inverse design

- Promises more effective searching of very large chemical design spaces (10^{60})
- Complementary to human design and older techniques such as virtual screening



Direct design

$f(\boldsymbol{x}) = \boldsymbol{y}$

Inverse design

$\boldsymbol{x} = f^{-1}(\boldsymbol{y})$

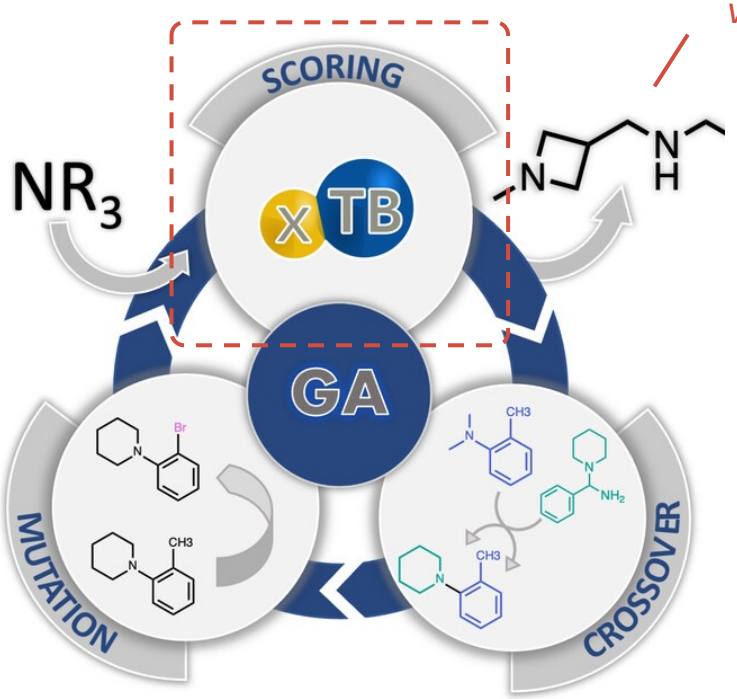
Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**. 10.1126/science.aat2663

Inverse design methods

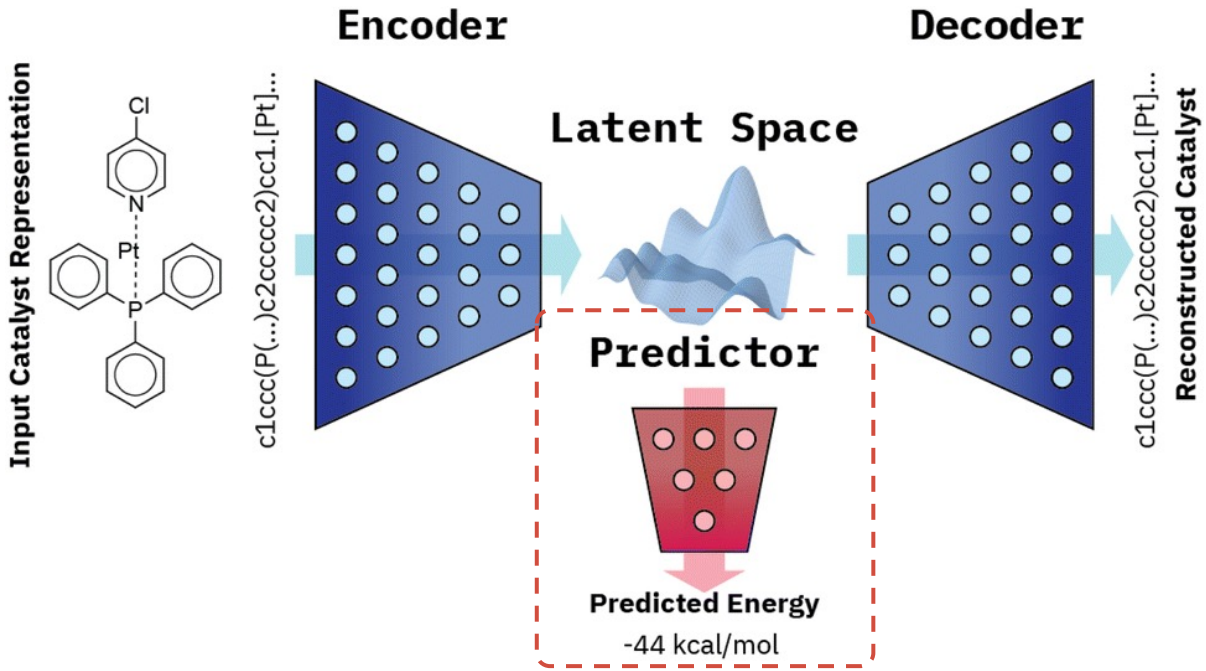
- Models that suggest new molecules with targeted properties
- Avoid exhaustive enumeration of chemical/materials space (>10^{60})
- Rely on property predictors to assess molecular function

Genetic algorithms

Experimentally verified



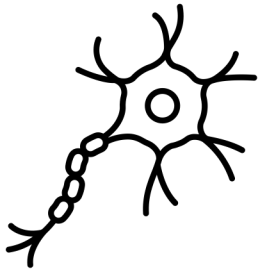
VAEs



(a) Seumer, J.; Kirschner Solberg Hansen, J.; Brøndsted Nielsen, M.; Jensen, J. H. *Angew. Chem. Int. Ed.* **2023**, *62* (18), e202218565.
(b) Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. *Dig. Disc.* **2023**, 10.1039.D2DD00125J.

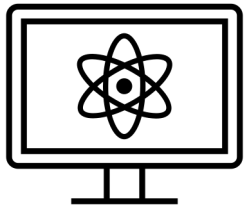
Property predictors

Machine learning



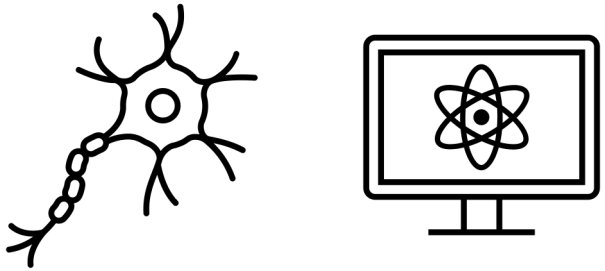
⊕ Fast
⊕ Accurate for “interpolation”
⊖ Limited applicability domain

Physics-based simulation

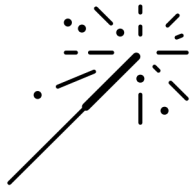


⊕ Wide applicability domain
⊖ Slow
⊖ Inaccurate for some problems

Hybrid models



⊕ Fast
⊕ Wide applicability domain
⊖ Not generally available yet



“Hybrid light”: Semi-empirical methods

Thiel, W. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (2), 145–157.

Goodhart's law

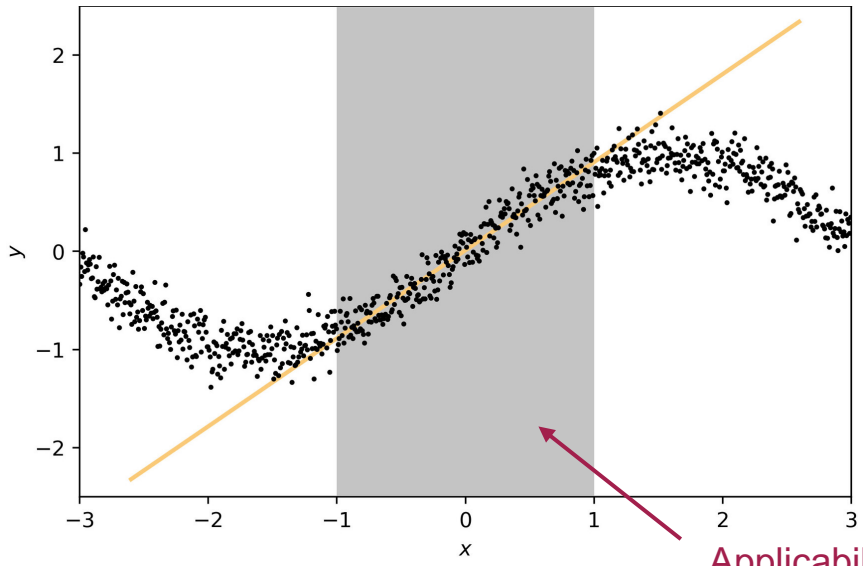
"When a measure becomes a target, it ceases to be a good measure."

https://en.wikipedia.org/wiki/Goodhart%27s_law

Applicability domain

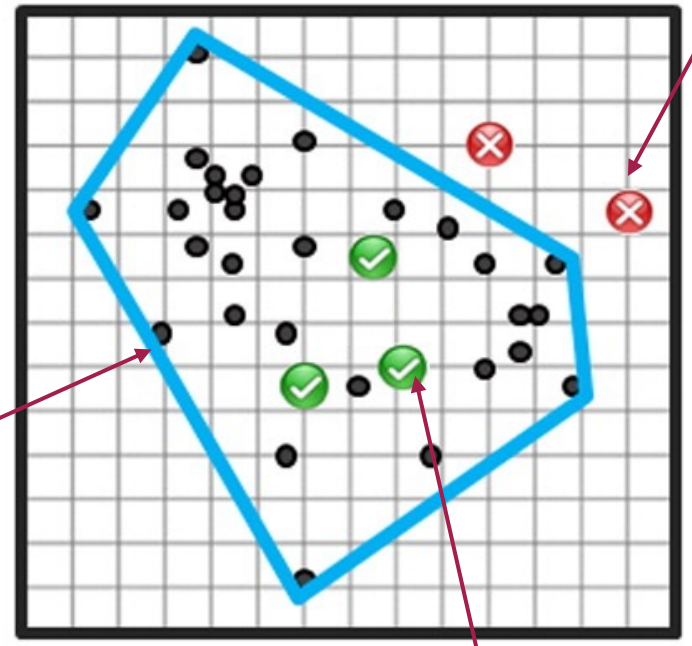
- Abstract concept of when a (chemical) ML model can be used
- Grows with more data
- Physical models have wider domain
- Machine learning models have narrower domain

Main principle in 1D



Applicability domain in 1D

Multidimensional space



Outside applicability domain

Training data

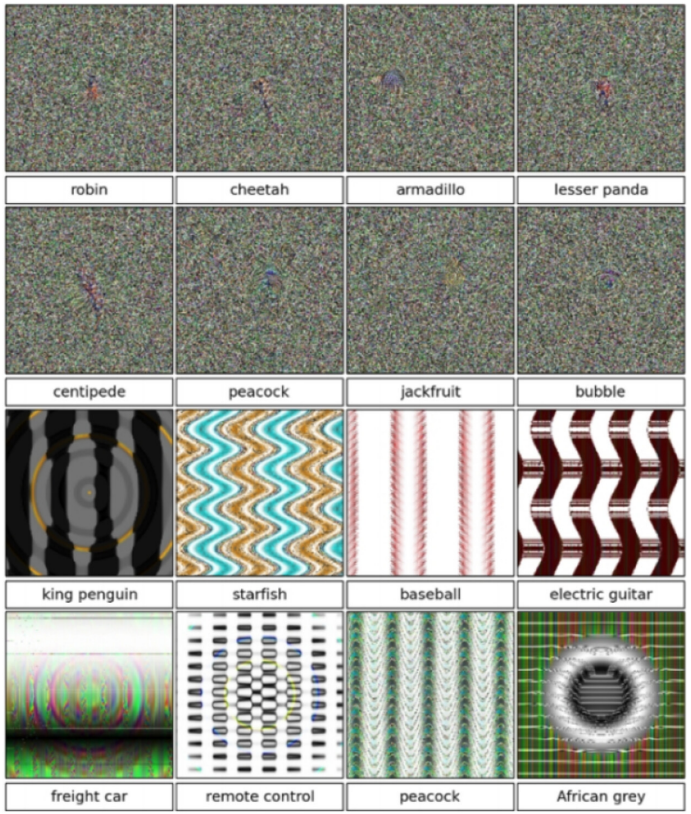
Inside applicability domain

(a) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. *SAR QSAR Environ. Res.* **2016**, *27* (11), 865–881. (b) Dutschmann, T.; Schlenker, V.; Baumann, K. *Mol. Inform.* **2024**, 10.1002/minf.202400018

“Adversarial examples”

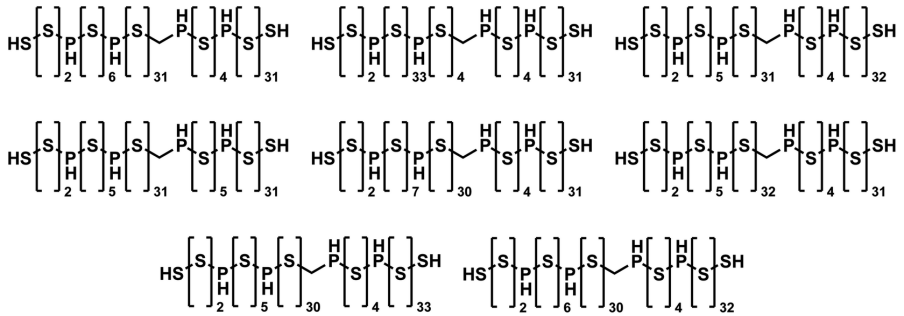
- Optimize the input to score well on classifier without connection to underlying task

Image classification

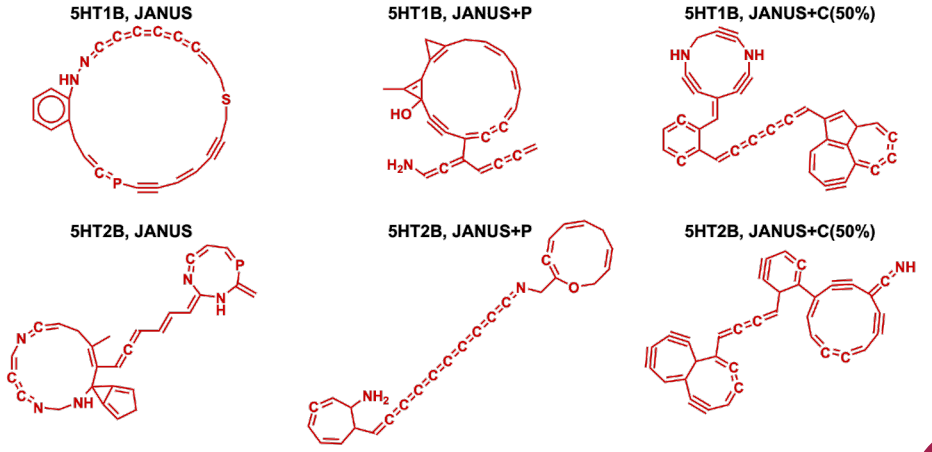


Molecular

Penalized logP



Docking

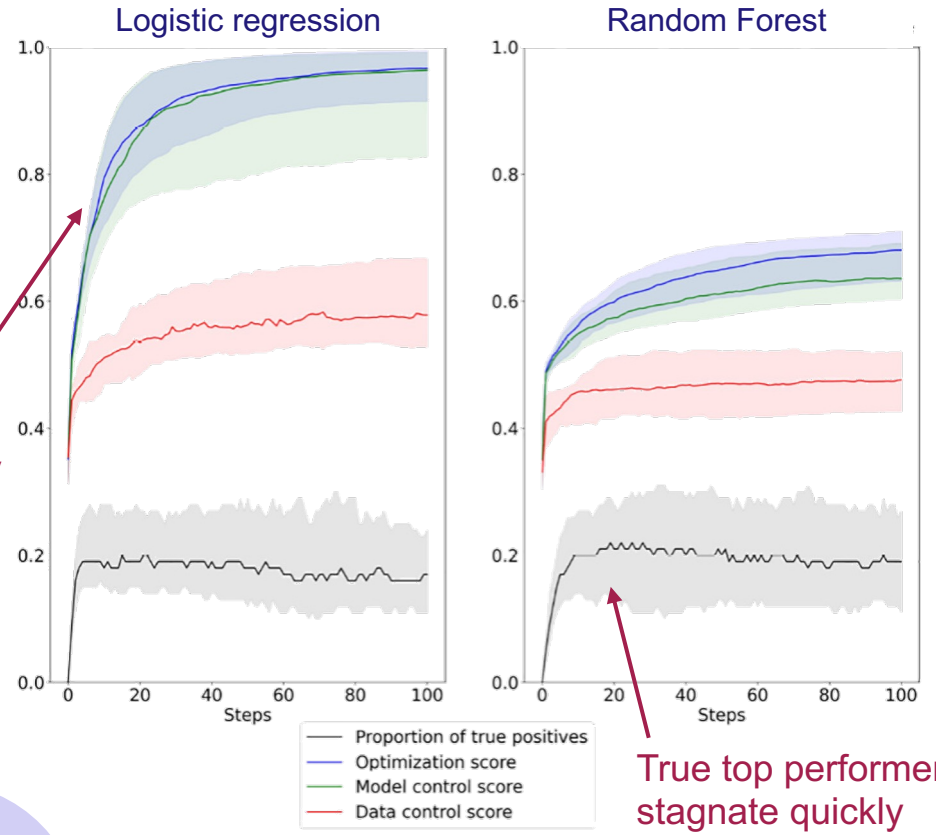


(a) Gendreau, P.; Turk, J.-A.; Drizard, N.; Ribeiro Da Silva, V. B.; Descamps, C.; Gaston-Mathé, Y. *J. Chem. Inf. Model.* **2023**. 10.1021/acs.jcim.3c00195 (b) Nigam, A.; Pollice, R.; Aspuru-Guzik, A. *Digit. Discov.* **2022**. 10.1039/D2DD00003B

Hacking the property predictor

- Generative models quickly exploit deficiencies in property predictors

Against ground truth

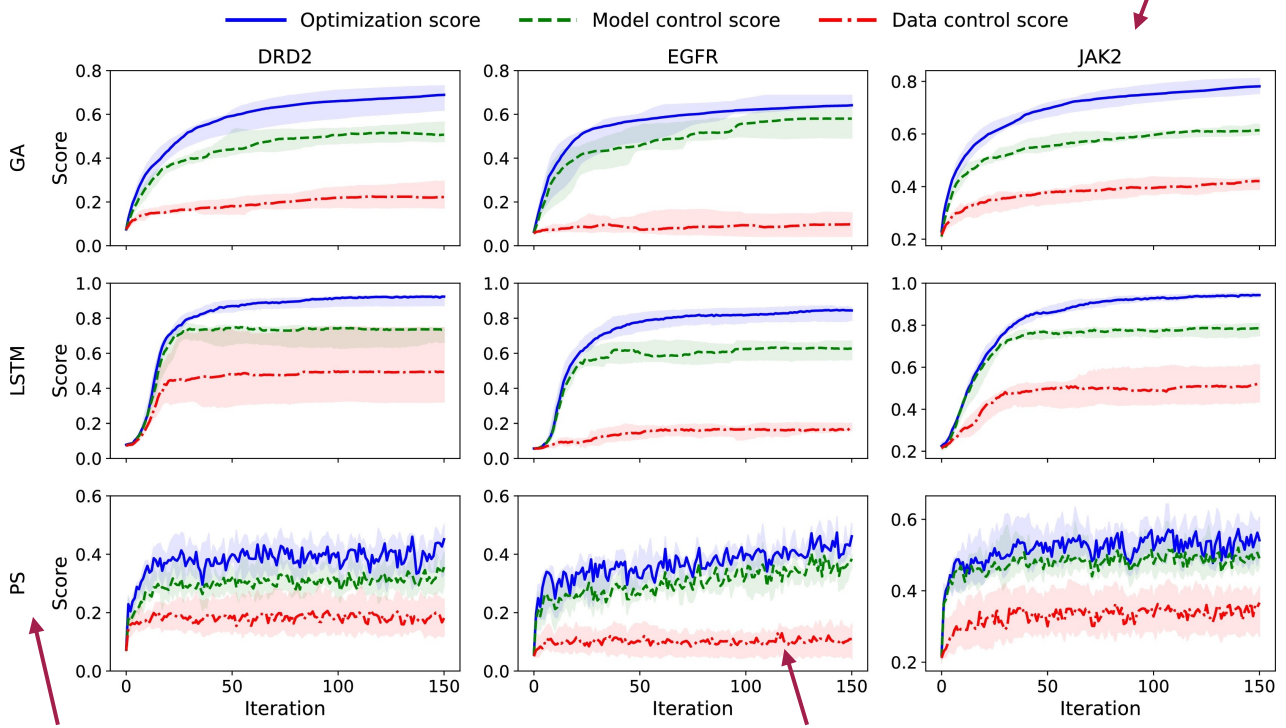


Score rises continuously

True top performers stagnate quickly

Against alternative models trained on same data

Different protein targets



Different generative algorithms

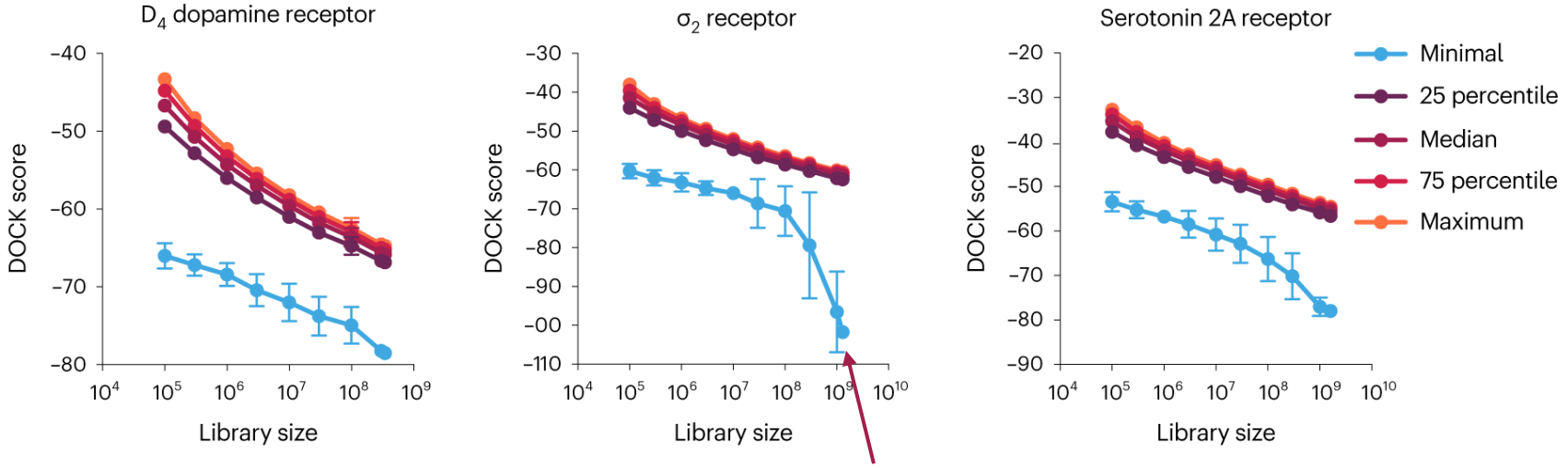
Alternative model does not improve

(a) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. *Drug Discov. Today Technol.* **2019**. 10.1016/j.ddtec.2020.09.003 (b) Gendreau, P.; Turk, J.-A.; Drizard, N.; Ribeiro Da Silva, V. B.; Descamps, C.; Gaston-Mathé, Y. *J. Chem. Inf. Model.* **2023**. 10.1021/acs.jcim.3c00195

Problem worsens in larger chemical spaces

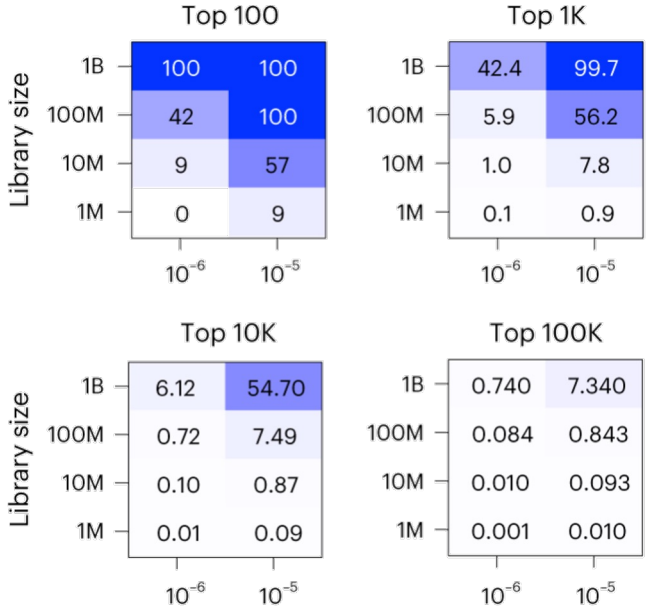
- "Artifacts" are particularly offensive molecules that exploit holes in simple predictors
- At larger chemical library sizes, they tend to dominate the top performers
- Goal-directed generative models easily latch onto these artifacts

Docking score as function of library size

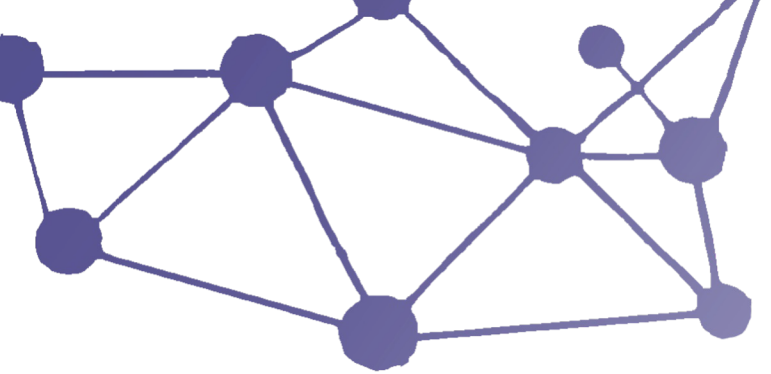


Artifacts dominate best-scoring molecules at larger sizes

Percentage of artifacts



Lyu, J.; Irwin, J. J.; Shoichet, B. K. *Nat. Chem. Biol.* **2023**. 10.1038/s41589-022-01234-w

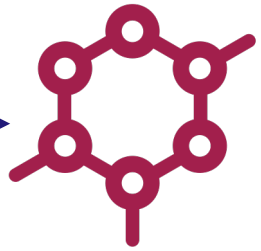


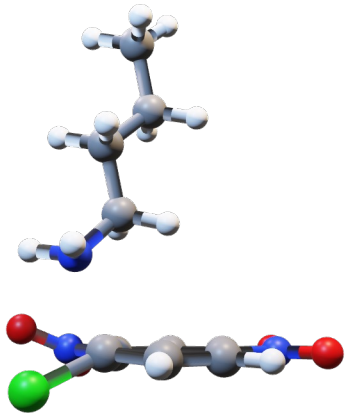
Molecular design with Hückel theory

Computational modeling of chemistry

- Translate experiment into mathematical model of atomistic model

Real world **Atomistic model** **Simulation model**



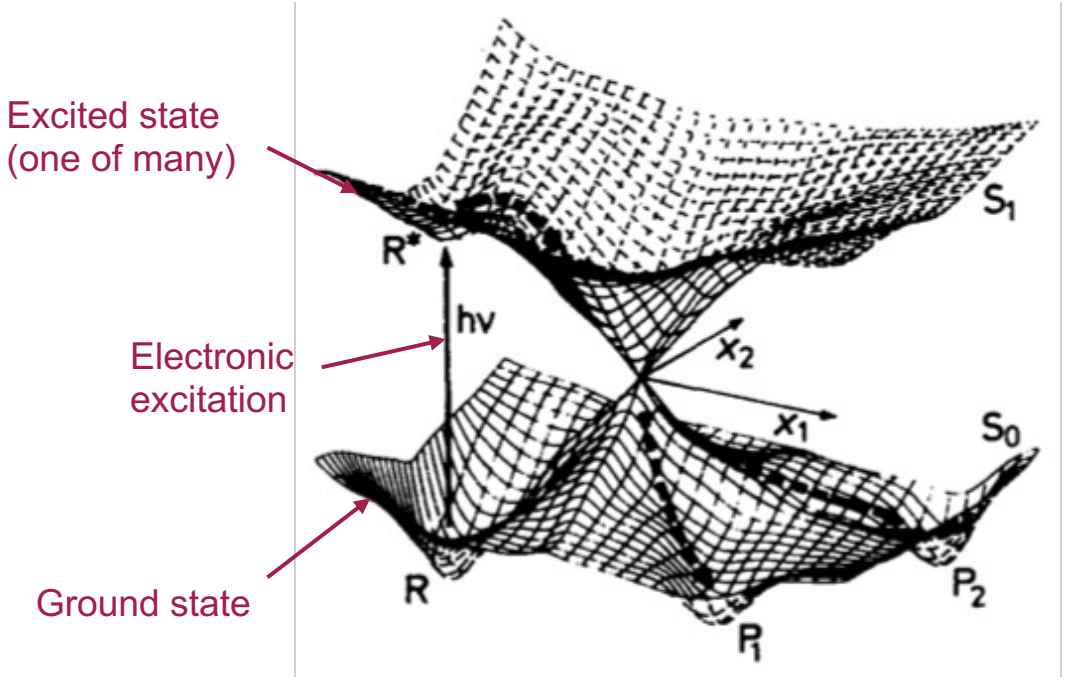


$$E[\rho(\boldsymbol{r})] = T_{\mathrm{ni}}[\rho(\boldsymbol{r})] + V_{\mathrm{ne}}[\rho(\boldsymbol{r})] + V_{\mathrm{ee}}[\rho(\boldsymbol{r})] + \Delta T[\rho(\boldsymbol{r})] + \Delta V_{\mathrm{ee}}[\rho(\boldsymbol{r})]$$

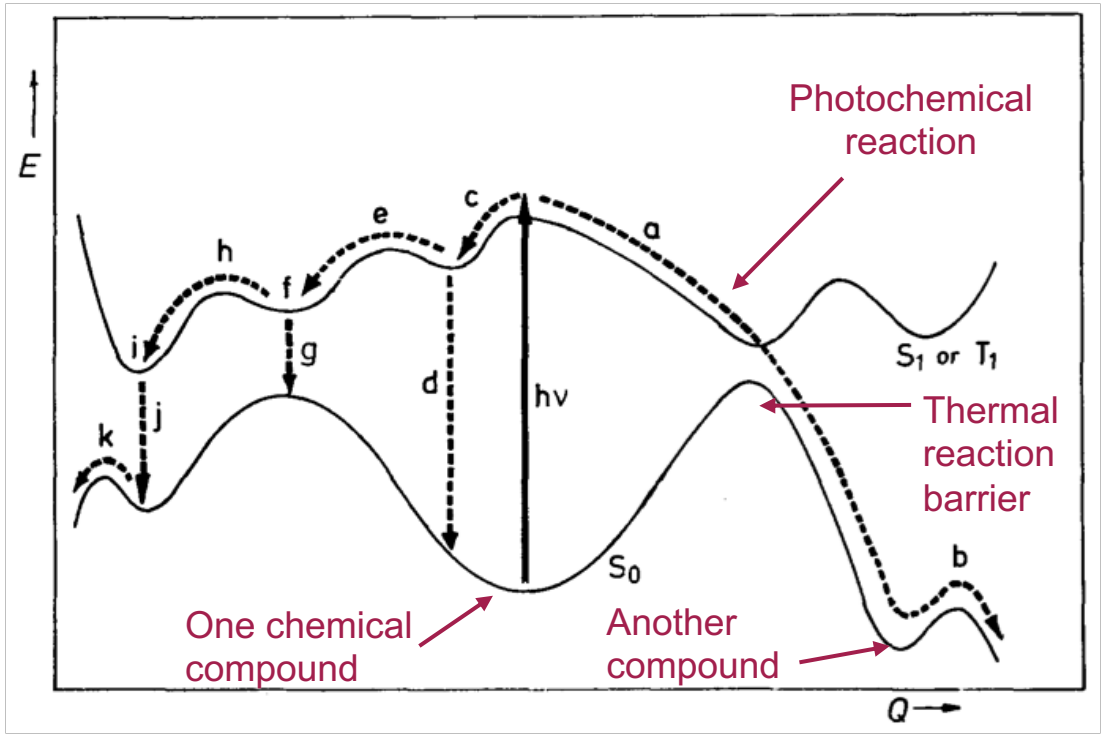
The potential energy surface (PES)

- Molecules live in 3*N*-dimensional space (*N* number of atoms)
- The energy as a function of the coordinates is called the PES

Multi-dimensional space



Reduced-coordinate "slice"



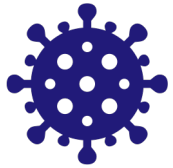
Computational methods

- Calculate the energy as function of coordinates
- A sequence of increasingly time-consuming physical methods

Increasing computational cost

Force fields	Machine learning potentials	Semi-empirical methods	Density functional theory	Ab initio
GFN-FF MMFF94 UFF …	ANI-1ccx ANI-2x AIQM1 MACE-OFF23 …	GFN2-xTB PM6 AM1 …	B3LYP-D3 ωB97XD r2SCAN-3c ωB97X-3c …	CCSD(T) ADC(2) CASSCF CASPT2 …
		Acceptable for property predictors	Too slow	

What can we simulate?



Interactions with surroundings
- Drug binding energies
- Solubility
- Redox potentials
- Thermodynamic properties of solutions
- ...

Structure & stability
- Crystal structure polymorphism
- Protein structure
- Molecular conformers
- Tautomer/protonation state
- …

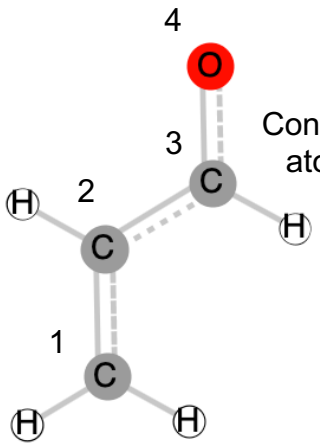
Chemical reactivity
- Rate constants
- Selectivity
- Bond dissociation energies
- Kinetic isotope effects
- Photochemistry
- …

Spectroscopic properties
- UV/vis
- IR
- Rotational
- X-ray
- Ionization potentials and electron affinities
- NMR & EPR
- Optical rotation & circular dichroism
- Mass spectra
- …

Hückel theory – extremely semi-empirical

- One of the first molecular orbital theories
- Schrödinger equation reduces to a very simple form

Molecule



Connectivity & atom types

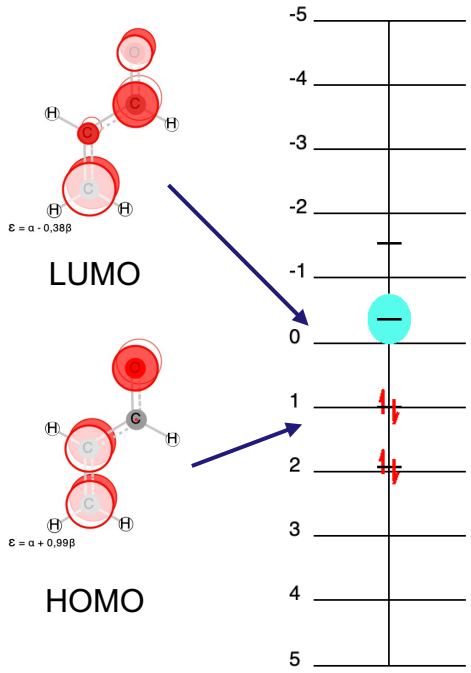
Hückel matrix

$$\begin{bmatrix} h_C & k_{CC} & 0 & 0 \\ k_{CC} & h_C & k_{CC} & 0 \\ 0 & k_{CC} & h_C & k_{CO} \\ 0 & 0 & k_{CO} & h_O \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1.06 \\ 0 & 0 & 1.06 & 0.97 \end{bmatrix}$$

One-electron integrals

Diagonalize

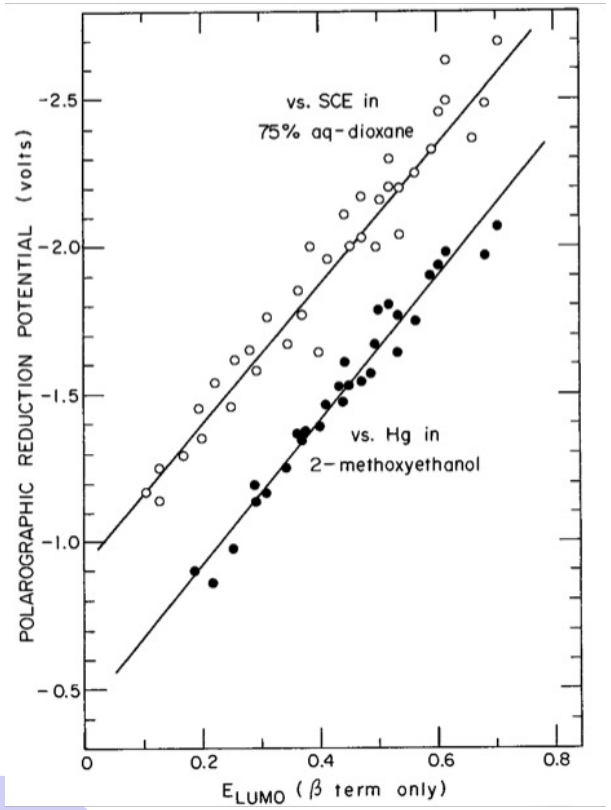
Orbitals & energies



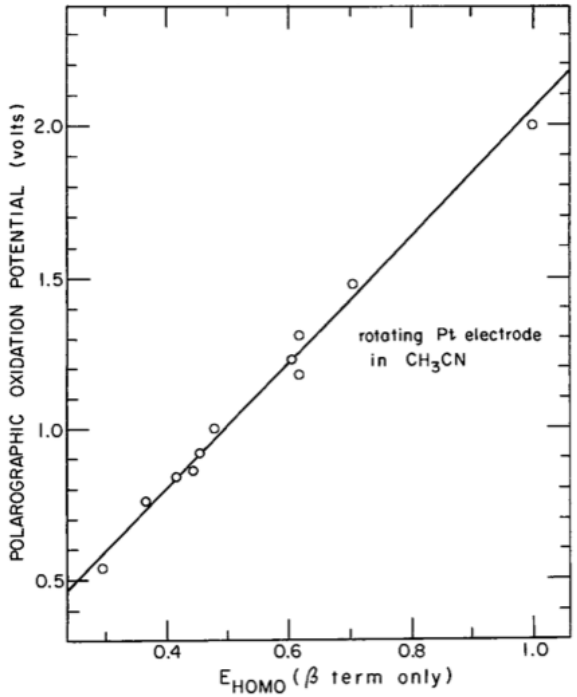
Indication of performance of Hückel theory

- Gives favorable correlations with experiment within certain compound classes

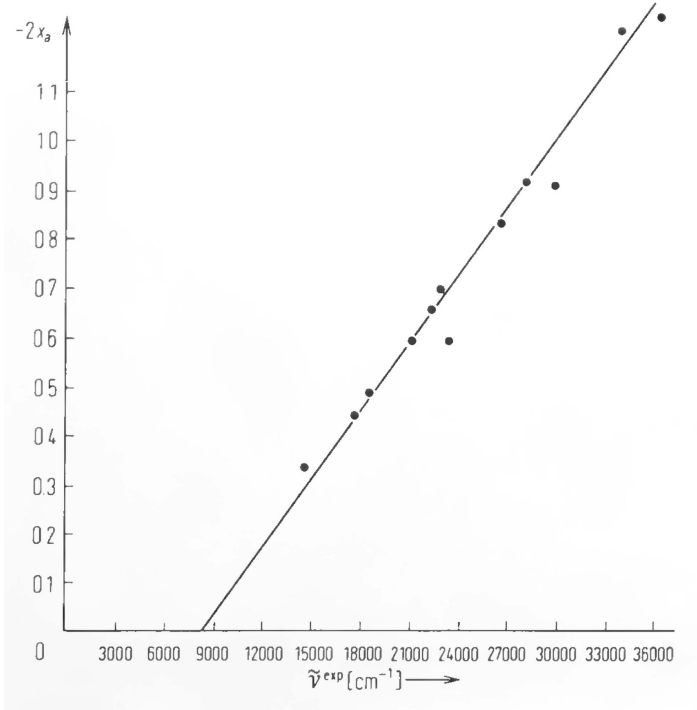
Reduction potential *vs.* LUMO



Oxidation potential *vs.* HOMO



Photoexcitation *vs.* HOMO-LUMO gap



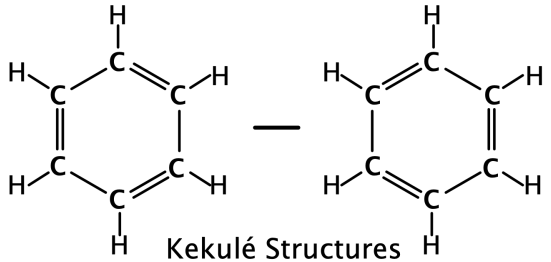
(a) Yates, K. *Hückel Molecular Orbital Theory*; Academic Press: New York, 1978. (b) Heilbronner, E.; Bock, H. *The HMO Model and Its Application: Basis and Manipulation*; Wiley: London, 1976.

Aromaticity

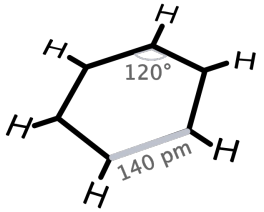
- Extra stabilization due to cyclic delocalization of π-electrons
- Applications to organic electronic materials, reactivity etc.

C_6H_6

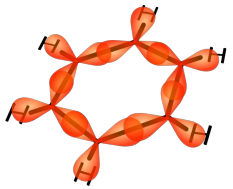
Benzene
Molecular formula



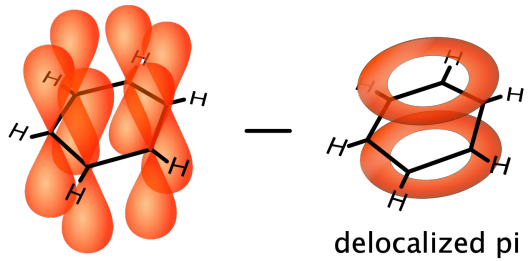
Kekulé Structures
(Resonance Forms)



Planar Hexagon
Bond Length 140 pm

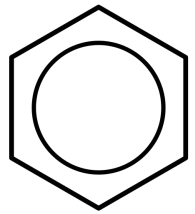


Sigma Bonds
sp^2 Hybridized orbitals



6 p_z orbitals

delocalized pi
system

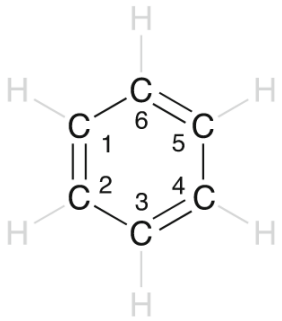


Benzene ring
Simplified depiction

The characteristic polynomial

- Solving Hückel theory in the case of only carbon atoms

Molecule



Adjacency matrix, $A(G)$

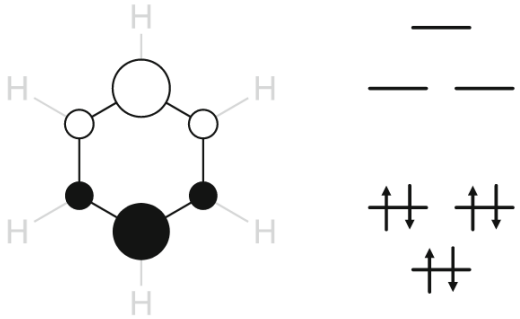
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$A(G) - X\mathbf{I}$

$$\begin{bmatrix} -x & 1 & 0 & 0 & 0 & 1 \\ 1 & -x & 1 & 0 & 0 & 0 \\ 0 & 1 & -x & 1 & 0 & 0 \\ 0 & 0 & 1 & -x & 1 & 0 \\ 0 & 0 & 0 & 1 & -x & 1 \\ 1 & 0 & 0 & 0 & 1 & -x \end{bmatrix}$$

$\det |A(G) - X\mathbf{I}|$

$x^6 - 6x^4 + 9x^2 - 4 = 0$



Orbitals & energies

Characteristic polynomial, $P_G(X)$

Secular determinant, $D_G(X)$

Jorner, K. *J. Phys. Org. Chem.* **2023**. 10.1002/poc.4460

Graph theory of aromaticity

- Invented independently by Aihara in Japan and Trinajstić & Gutman in Croatia
- Builds on Coulson's graph treatment of Hückel molecular orbital theory
- Calculates aromatic properties as difference between
  - E_π: Energy of actual π system
  - E_π^{ref}: Energy of "olefinic" reference system (lacking cyclic conjugation)

Jun-ichi Aihara

Nenad Trinajstić

Ivan Gutman

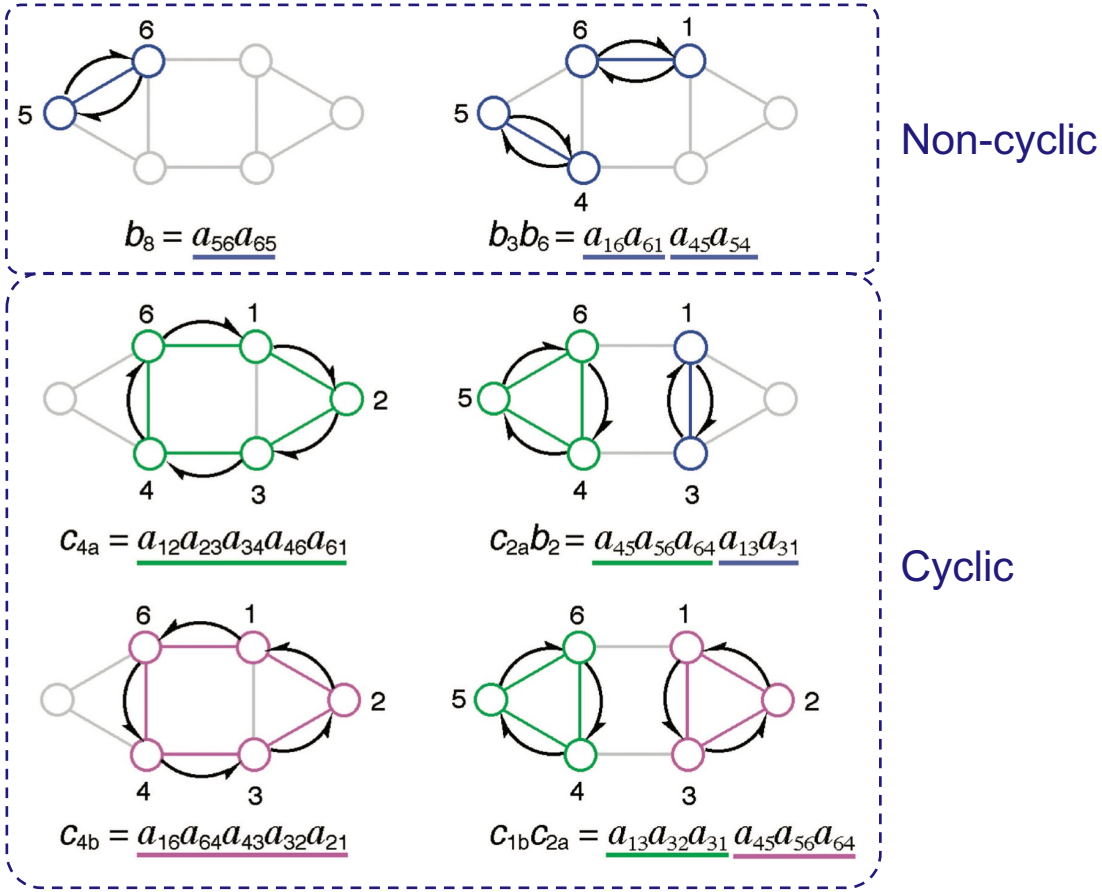
(a) Aihara, J. *J. Am. Chem. Soc.* **1976**, *98* (10), 2750–2758. (b) Gutman, I.; Milun, M.; Trinajstic, N. *J. Am. Chem. Soc.* **1977**, *99* (6), 1692–1704. (c) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, 1992.

Constructing the reference polynomial

Expand the characteristic polynomial of the Hückel matrix

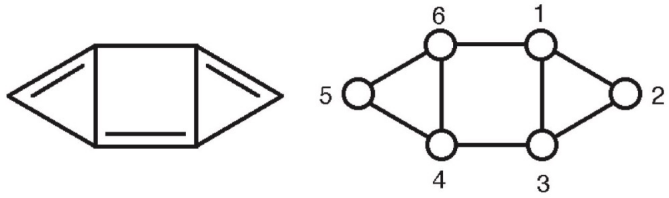
$$
\begin{aligned}
P_G(X) = X^6 &+ (-b_1 - b_2 - b_3 - b_4 - b_5 - b_6 - b_7 - b_8)X^4 \\
&+ (-c_{1a} - c_{1b} - c_{2a} - c_{2b})X^3 + (b_1b_5 + b_1b_6 \\
&+ b_1b_7 + b_1b_8 + b_2b_6 + b_2b_7 + b_2b_8 + b_3b_4 \\
&+ b_3b_5 + b_3b_6 + b_4b_6 + b_4b_7 + b_4b_8 + b_5b_8 \\
&- c_{3a} - c_{3b})X^2 + (c_{1a}b_6 + c_{1a}b_7 + c_{1a}b_8 + c_{1b}b_6 \\
&+ c_{1b}b_7 + c_{1b}b_8 + c_{2a}b_1 + c_{2a}b_4 + c_{2a}b_2 + c_{2b}b_1 \\
&+ c_{2b}b_4 + c_{2b}b_2 - c_{4a} - c_{4b} - c_{5a} - c_{5b})X \\
&+ (-b_1b_5b_8 - b_4b_6b_3 + c_{1a}c_{2a} + c_{1a}c_{2b} + c_{1b}c_{2a} \\
&+ c_{1b}c_{2b} - c_{6a} - c_{6b})
\end{aligned}
$$

Sachs subgraphs

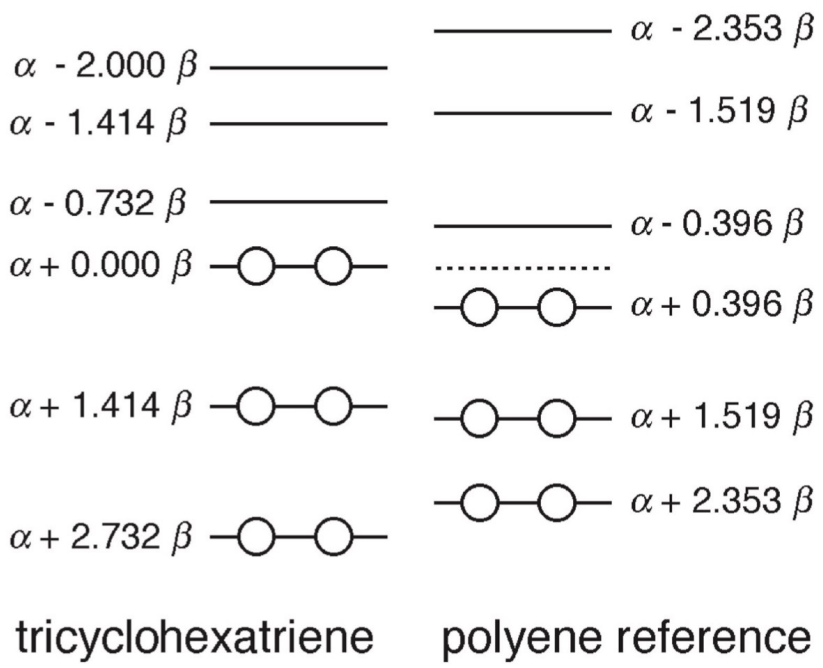


Topological resonance energy

- $P_G(X) \Rightarrow$ roots $X \Rightarrow$ orbital energies ϵ
- $R_G(X) \Rightarrow$ roots $X \Rightarrow$ orbital energies ϵ
- $\mathrm{TRE} = -\left(E_\pi - E_\pi^{\mathrm{ref}}\right) = -0.2425|\beta|$
  - Positive TRE ⇒ aromatic
  - Negative TRE ⇒ anti-aromatic
- Normalized metric
  - $\%\mathrm{TRE} = \frac{\mathrm{TRE}}{E_\pi^{\mathrm{ref}}}$

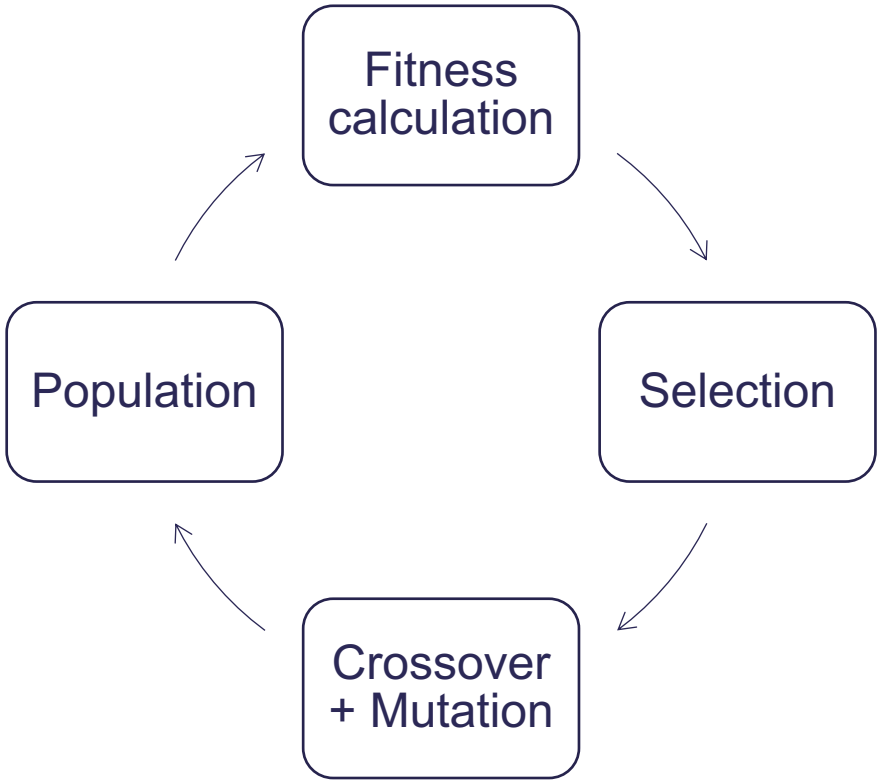


Tricyclo[3.1.0.0^{2,4}]hexatriene



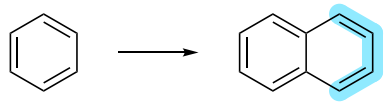
Genetic algorithm

Algorithm

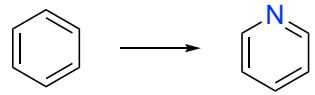


Genetic operations

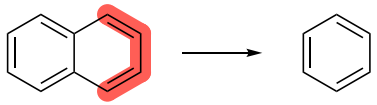
Add ring



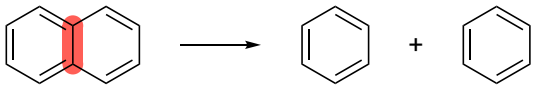
Replace atom



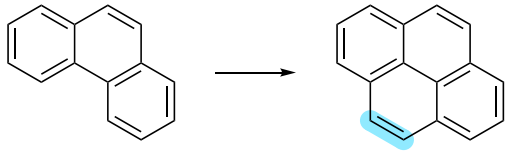
Remove ring



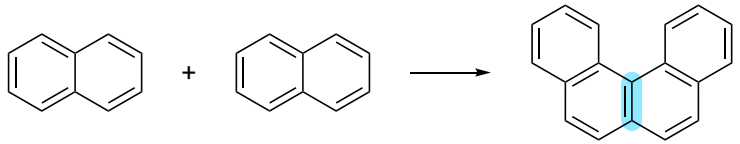
Fragment molecule



Fill out bay area



Combine molecules (crossover)

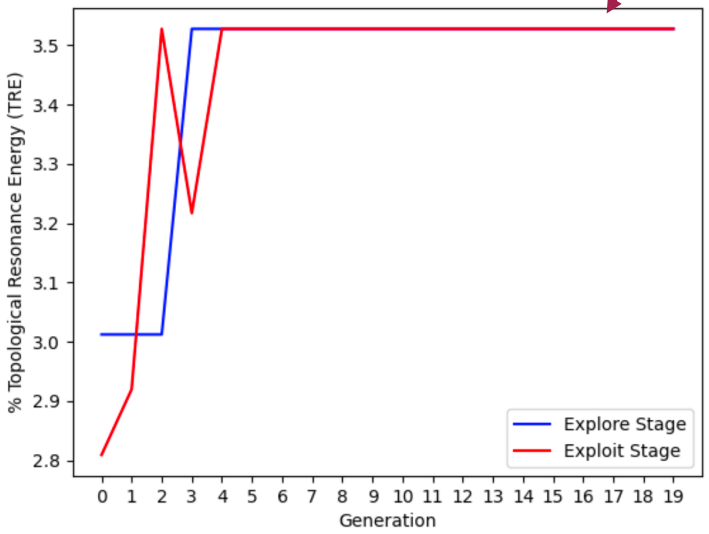


(a) Nigam, A.; Pollice, R.; Aspuru-Guzik, A. *Digit. Discov.* **2022**, *1* (4), 390–404 (b) Cheng, C. Y.; Campbell, J. E.; Day, G. M. *Chem. Sci.* **2020**, *11* (19), 4922–4933

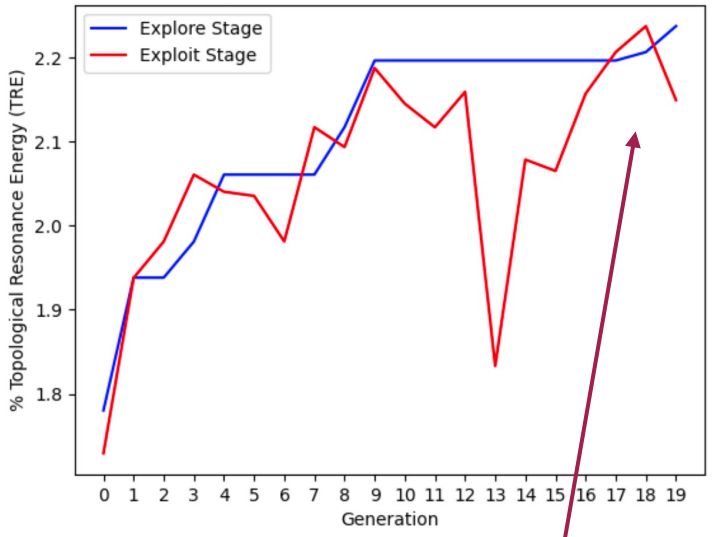
Maximizing %TRE – more aromatic

Finds benzene

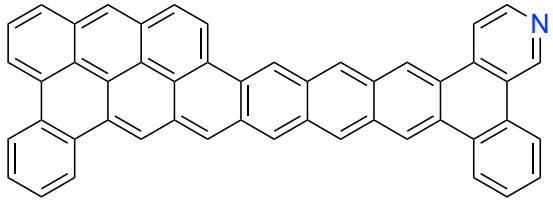
Ground state



Excited state

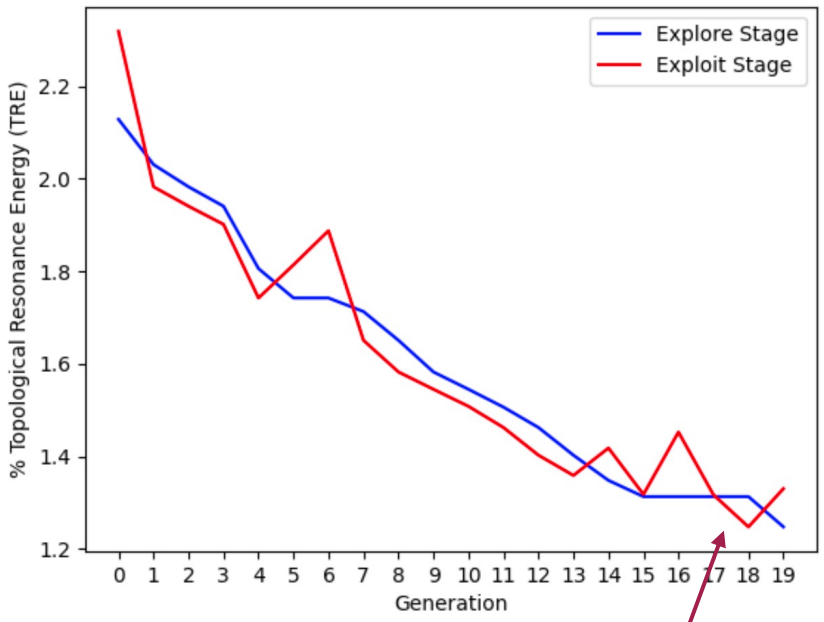


Not yet clear (not converged)

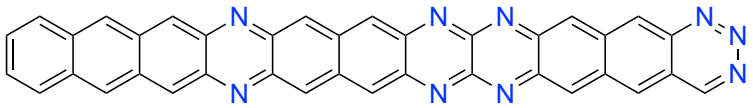


Minimizing %TRE – less aromatic

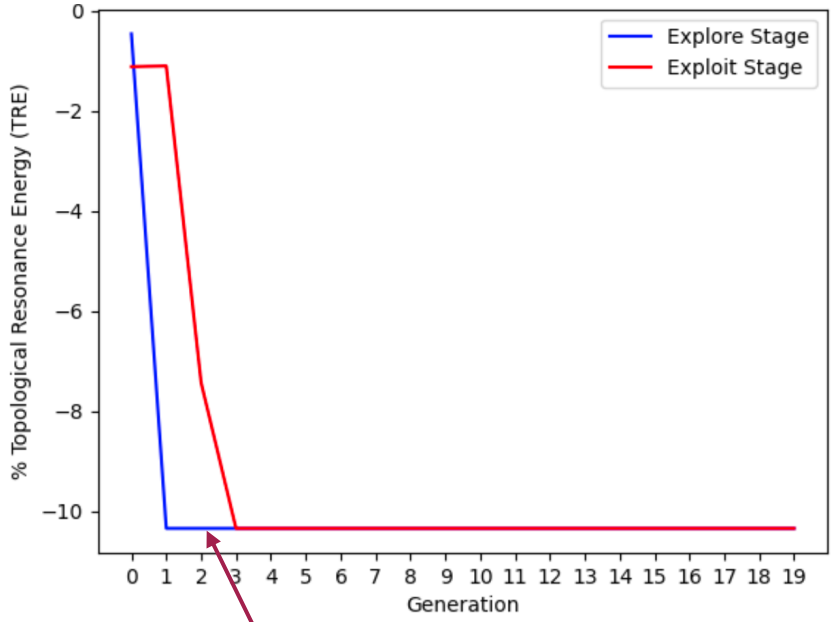
Ground state



Finds longer acenes (not converged)



Excited state

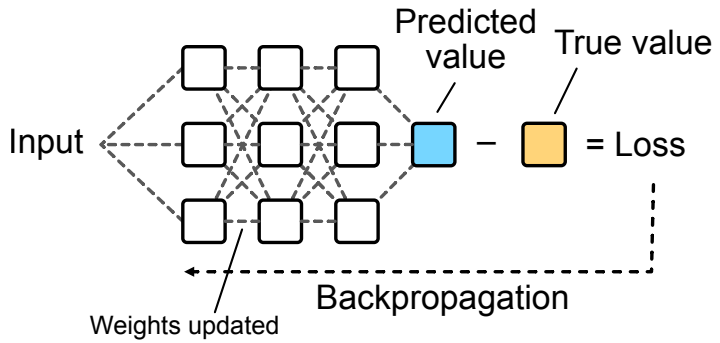


Finds benzene

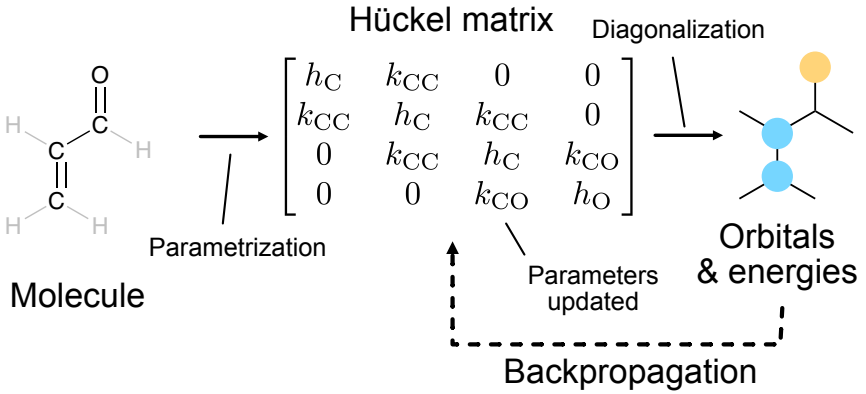
Hybrid ML & physical models

- Physical models can be coded within auto-differentiable frameworks
- Hybrids are regularized by the physical model
- Can in principle be legacy code via Enzyme LLVM compiler

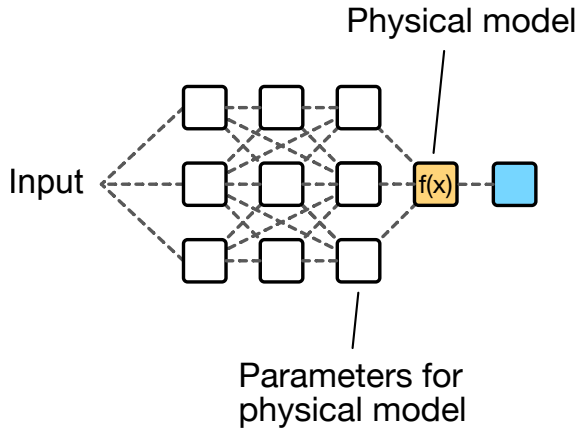
Neural network



Differentiable model



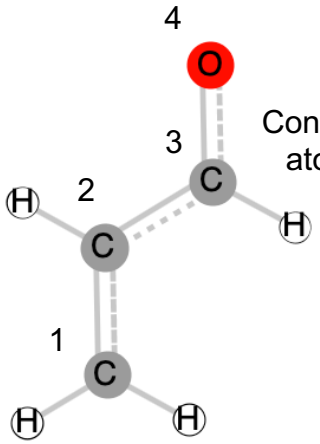
Hybrid



Auto-differentiable Hückel theory

- One of the first molecular orbital theories
- Schrödinger equation reduces to a very simple form

Molecule



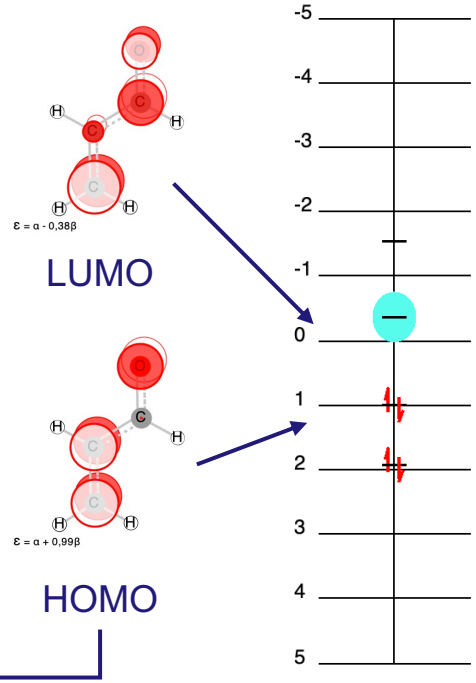
Connectivity & atom types

Hückel matrix

$$\begin{bmatrix} h_C & k_{CC} & 0 & 0 \\ k_{CC} & h_C & k_{CC} & 0 \\ 0 & k_{CC} & h_C & k_{CO} \\ 0 & 0 & k_{CO} & h_O \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1.06 \\ 0 & 0 & 1.06 & 0.97 \end{bmatrix}$$

Diagonalize

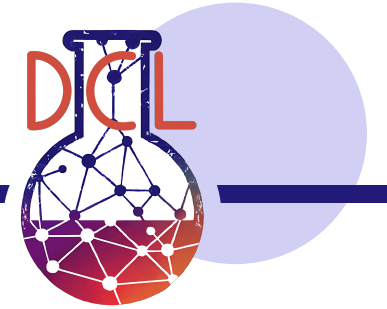
Orbitals & energies



Backpropagate error to parameters

Automatic differentiation

Translating NumPy code to JAX



- `jax.numpy & jax.scipy`

COULSON

```
import numpy as np
eig_vals, eig_vects = np.linalg.eigh(huckel_matrix)
```

HÜXEL

```
import jax.numpy as jnp
eig_vals, eig_vects = jnp.linalg.eigh(huckel_matrix)
```

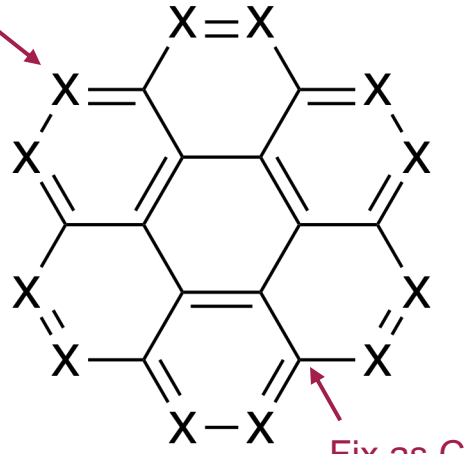
+

Code for parameter optimization

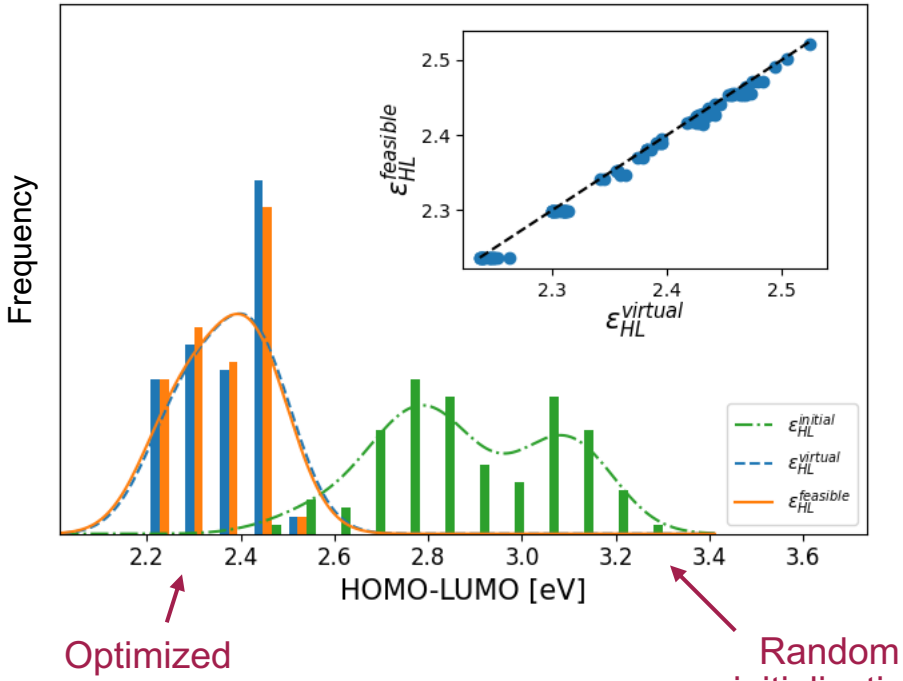
Inverse design

- "Alchemical" optimization of atom identity
- Optimize atom "weights" w_i with constraint $\sum w_i = 1$

Vary between C, N and P



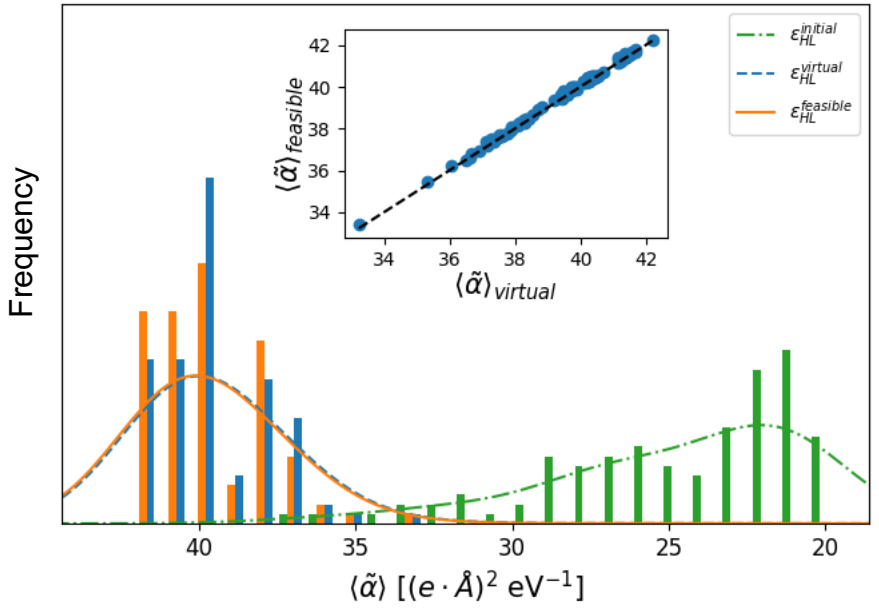
HOMO-LUMO gap



Optimized

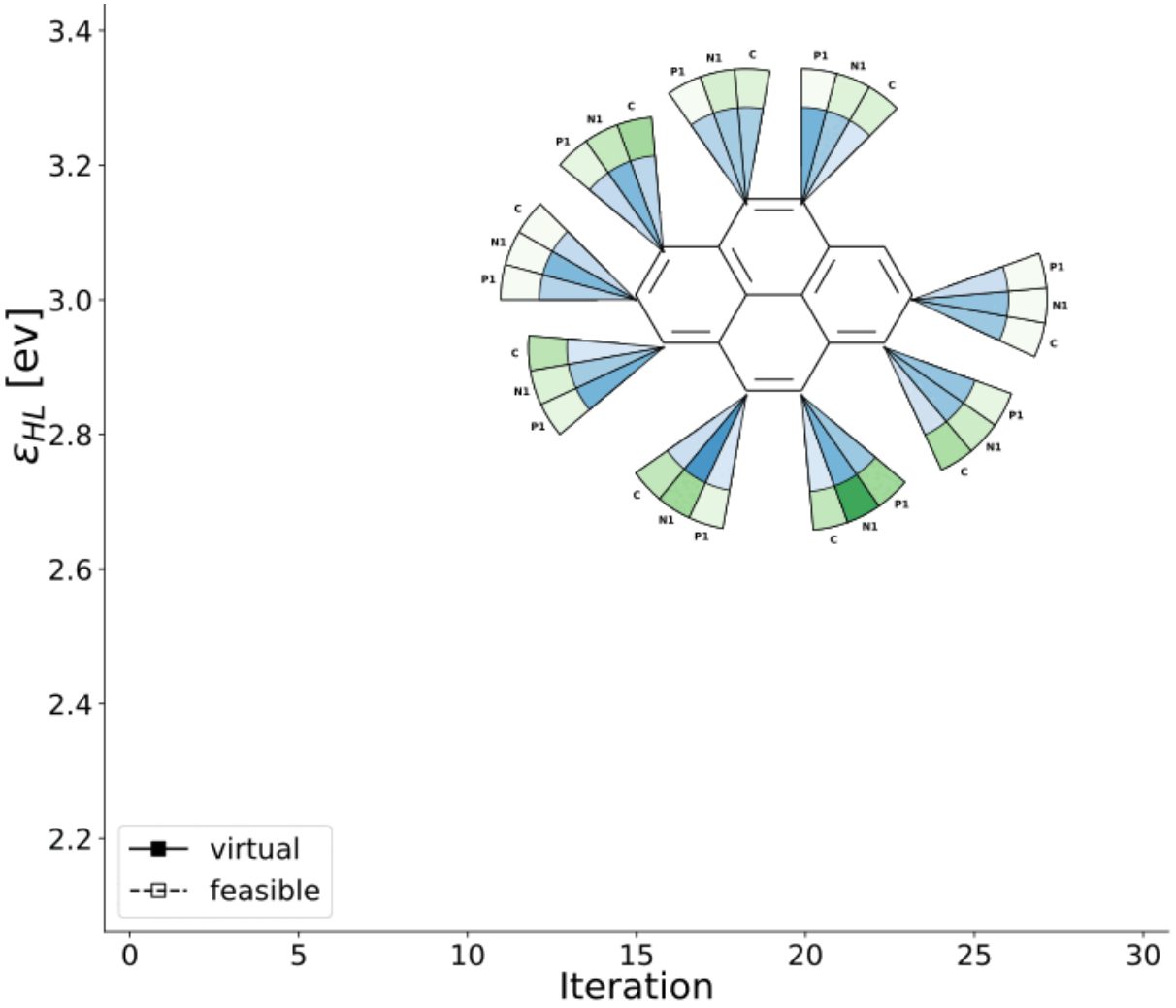
Random initialization

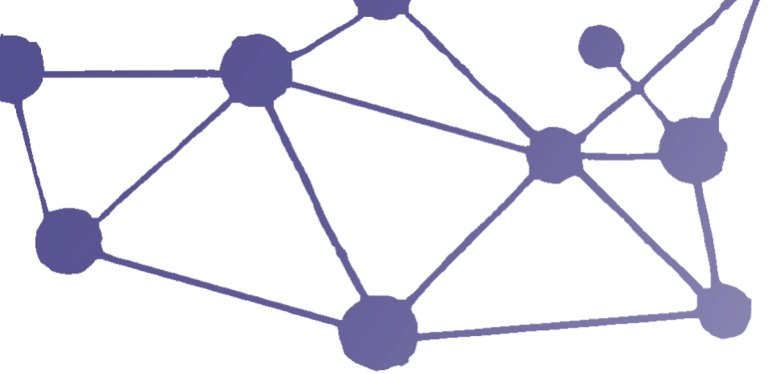
Second order derivative → **Polarizability**



Xiao, D.; Yang, W.; Beratan, D. N. J. *Chem. Phys.* **2008**, *129* (4), 044106.

Inverse design in action

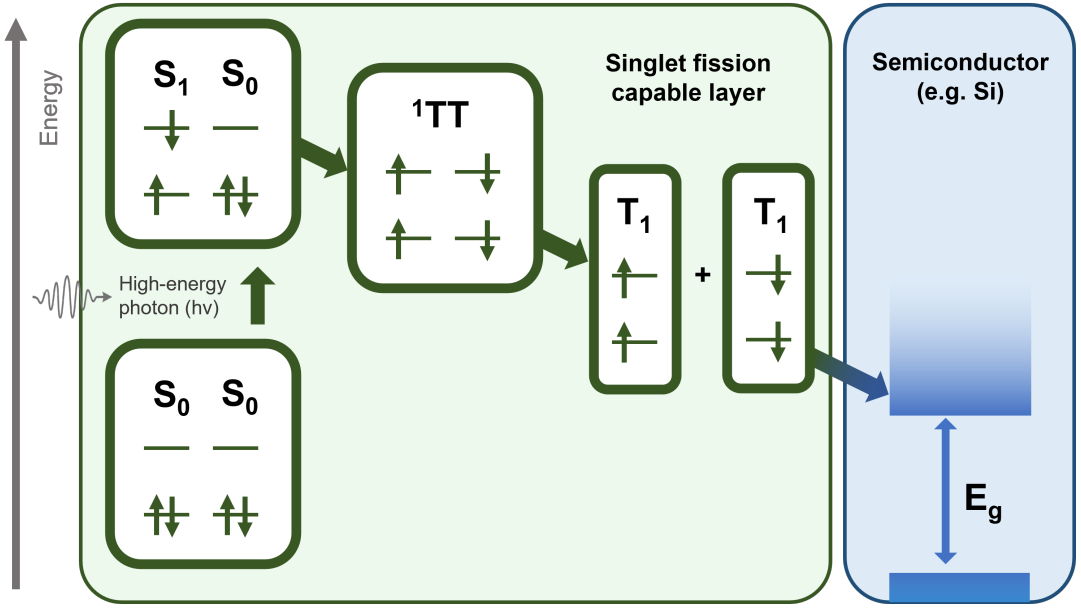




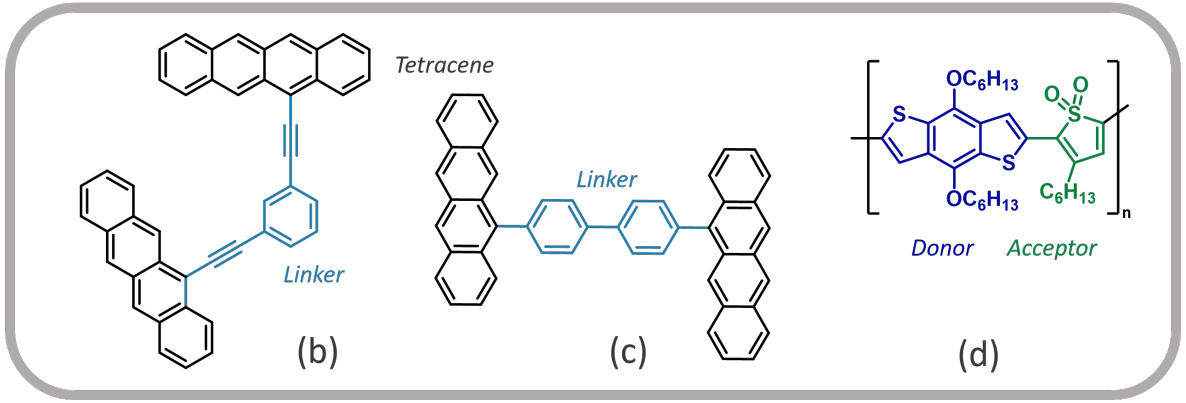
Designing singlet fission materials with uncertainty-aware genetic algorithms

Singlet fission

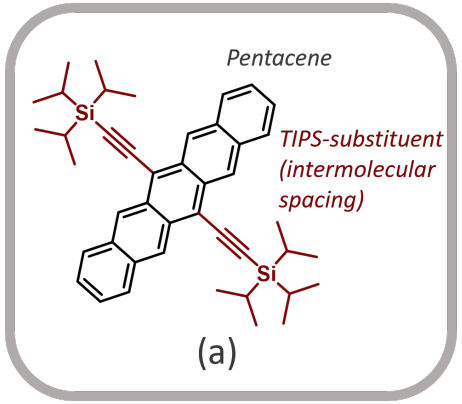
- Allow harvesting of high-energy photons to match band-gap of solar cell
- Relatively few molecules known



Intramolecular SF



Intermolecular SF

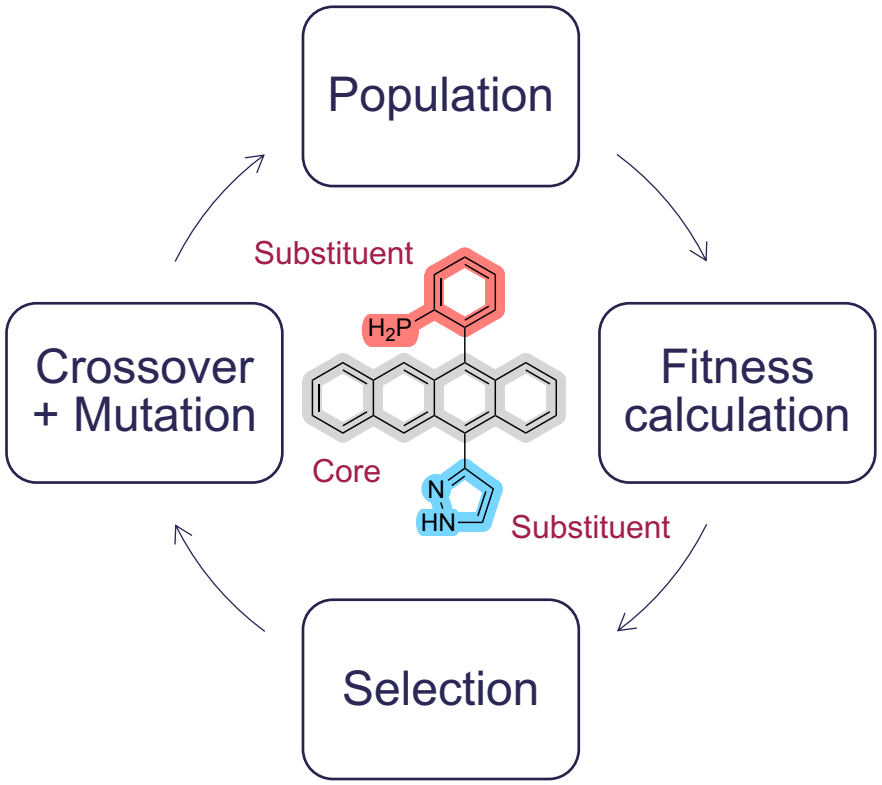


Luca Schaufelberger

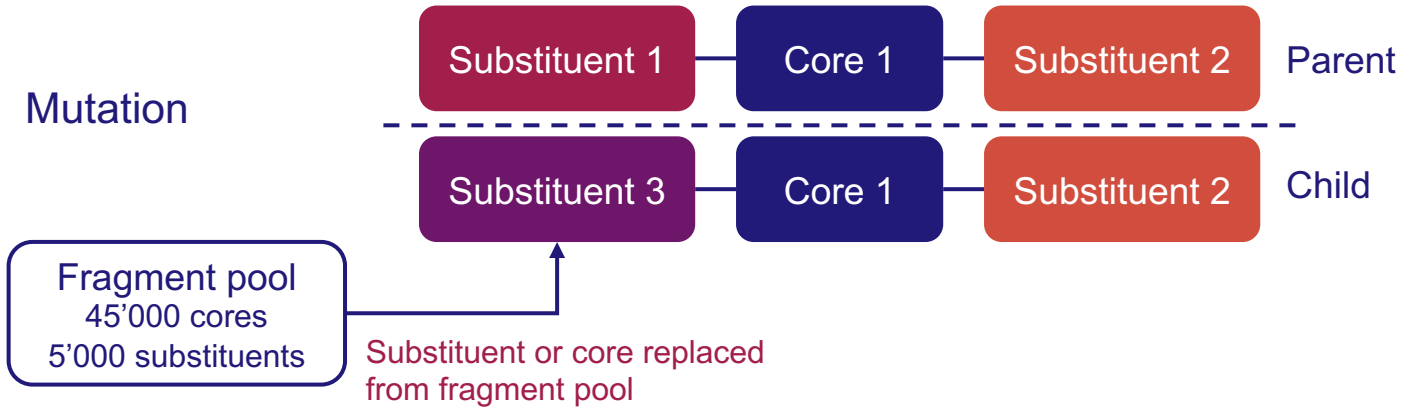
Clémence Corminboeuf

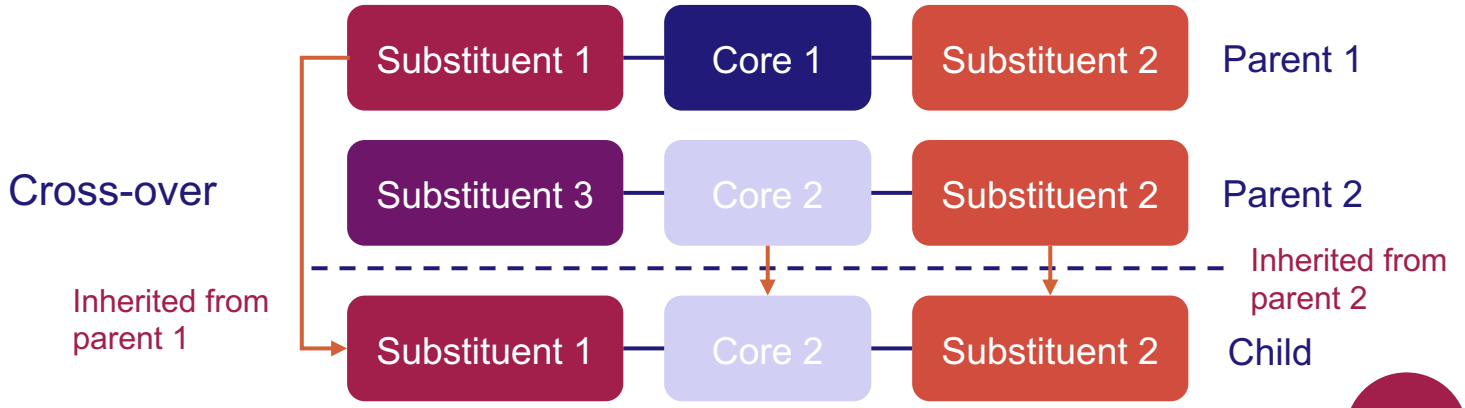
Genetic algorithm

Algorithm



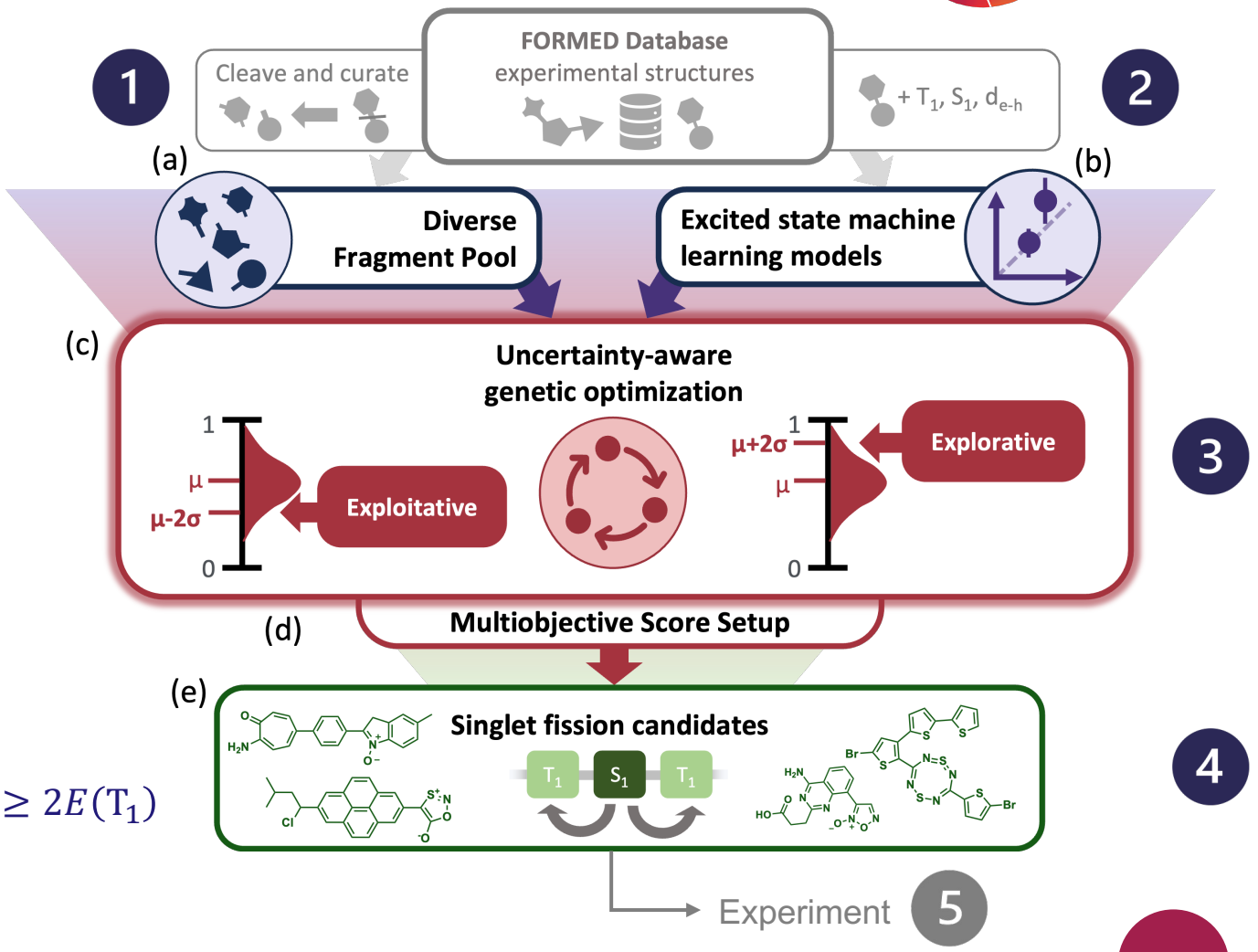
Genetic operations





Optimization with genetic algorithm

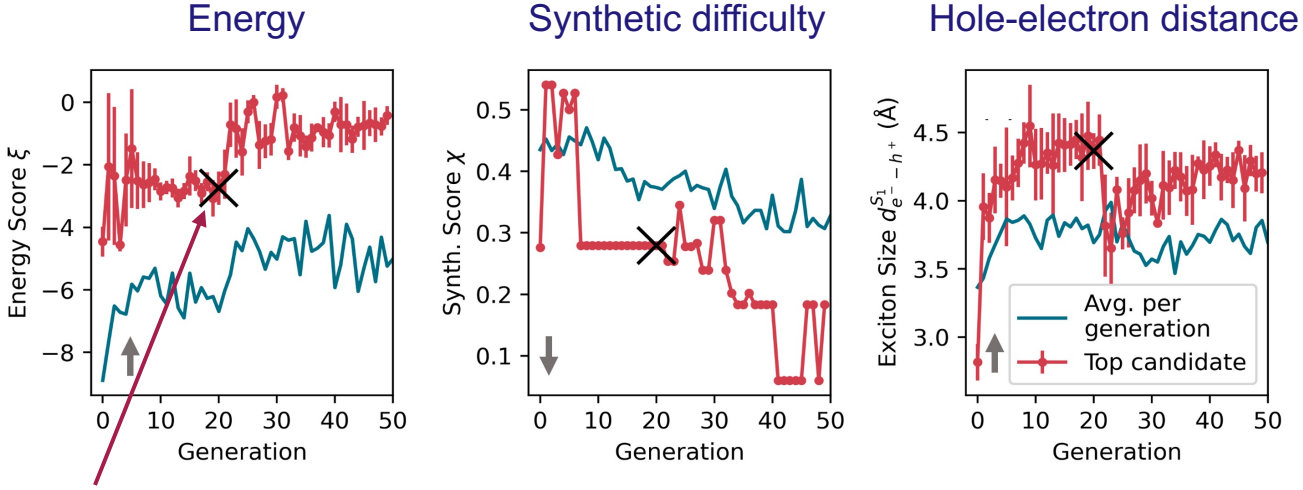
1. Create a pool of fragments from FORMED database
2. Build machine learning models for molecular properties
3. Run genetic algorithm in explorative or exploitative mode
4. Collect candidates and further study them with DFT simulations
5. (Experimental testing)



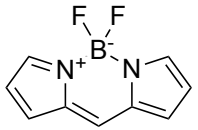
Optimization criteria

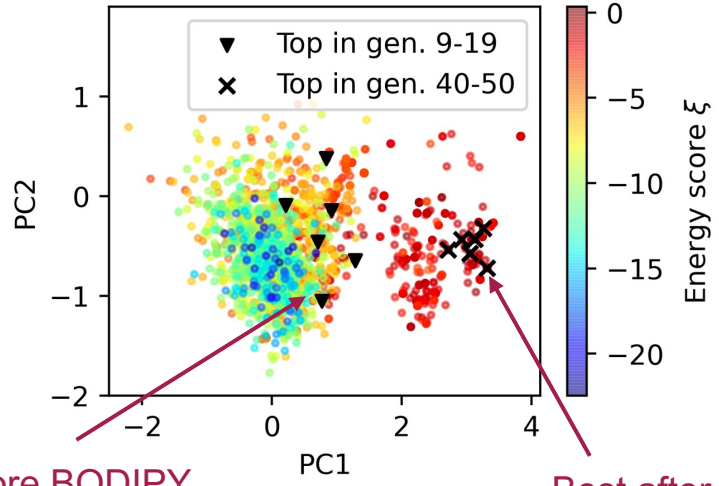
1. Singlet fission energy score $E(S_1) \geq 2E(T_1)$
2. Synthetic difficulty
3. Electron-hole separation

Results



"Discovery" of BODIPY core

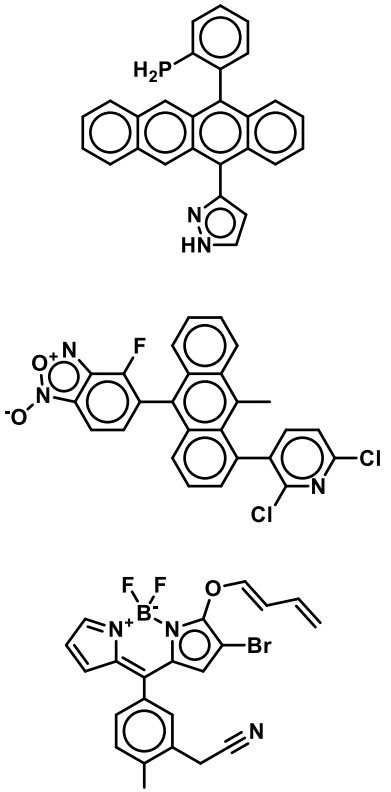




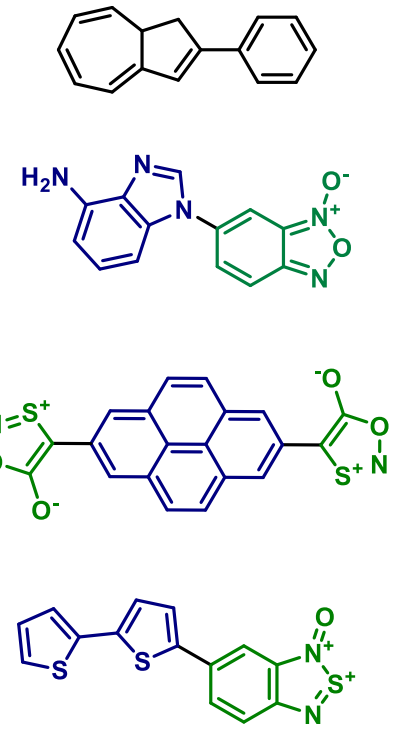
Best before BODIPY

Best after BODIPY

Rediscovered
Exploitation

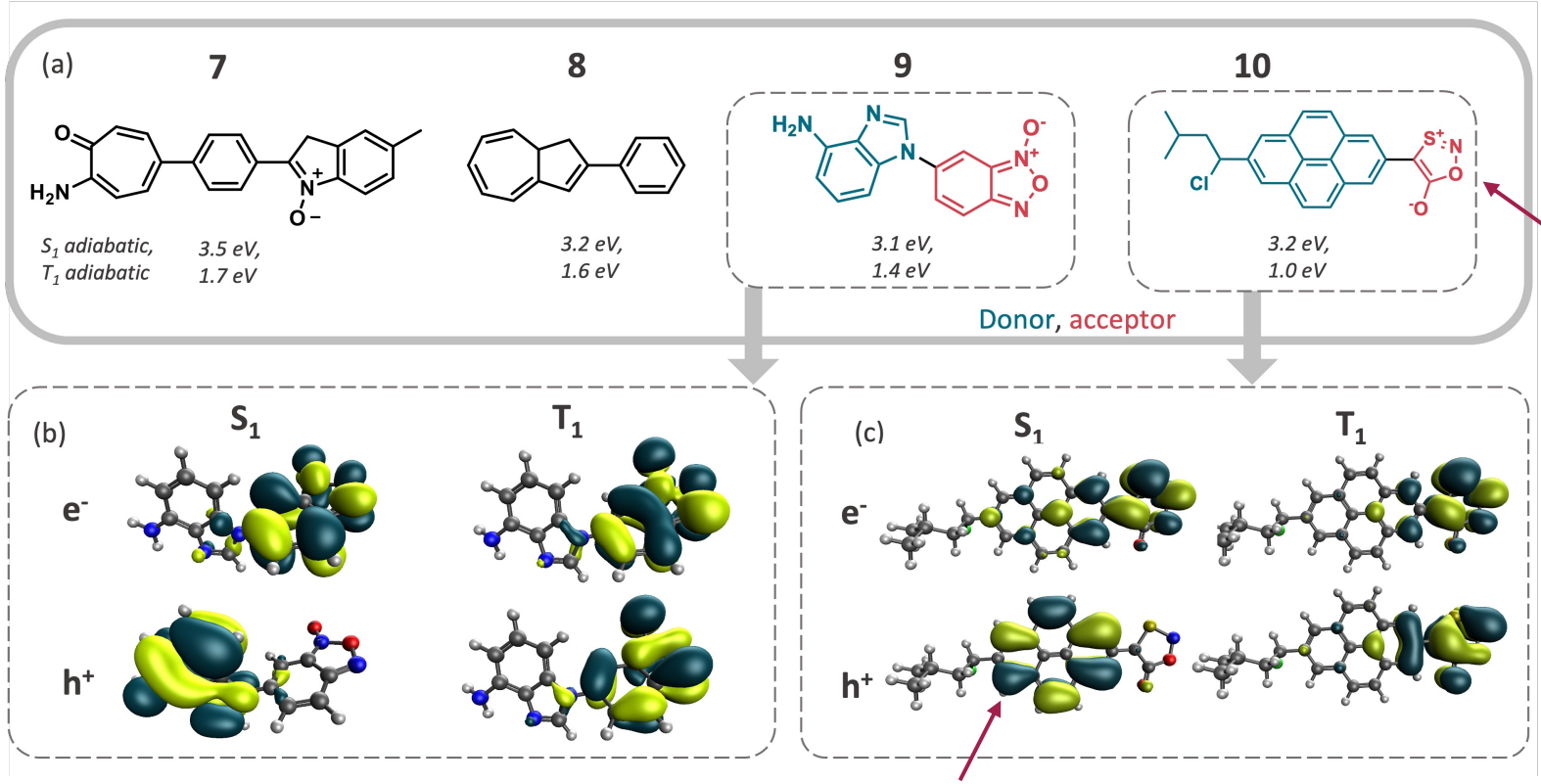


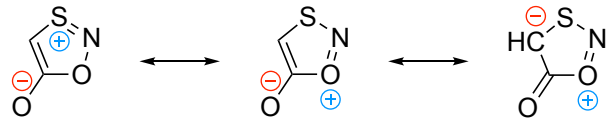
Previously unknown (?)
Exploration



Mesoionic acceptors

- Exotic electronic structure where no charge-neutral resonance forms can be drawn
- Feature charge-transfer S_1 states and locally excited T_1 states





"Mesoionic compounds are one in which a heterocyclic structure is dipolar and where both the negative and the positive charges are delocalized"

Electron and hole densities in T_1 are localized on same fragment (local excitation)

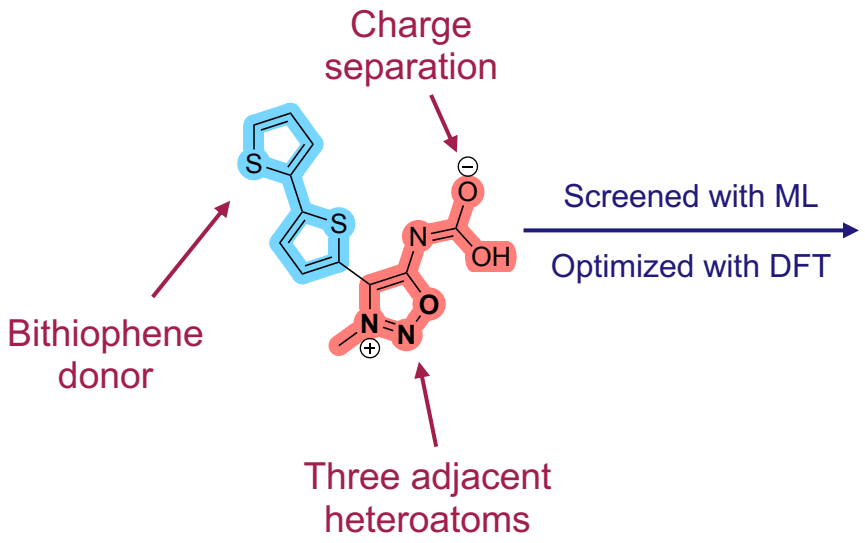
Electron and hole densities in S_1 are localized on different fragments (charge transfer excitation)

(a) https://en.wikipedia.org/wiki/Mesoionic_compounds. 2024-04-21. (b) https://doi.org/10.1351/goldbook.M03842

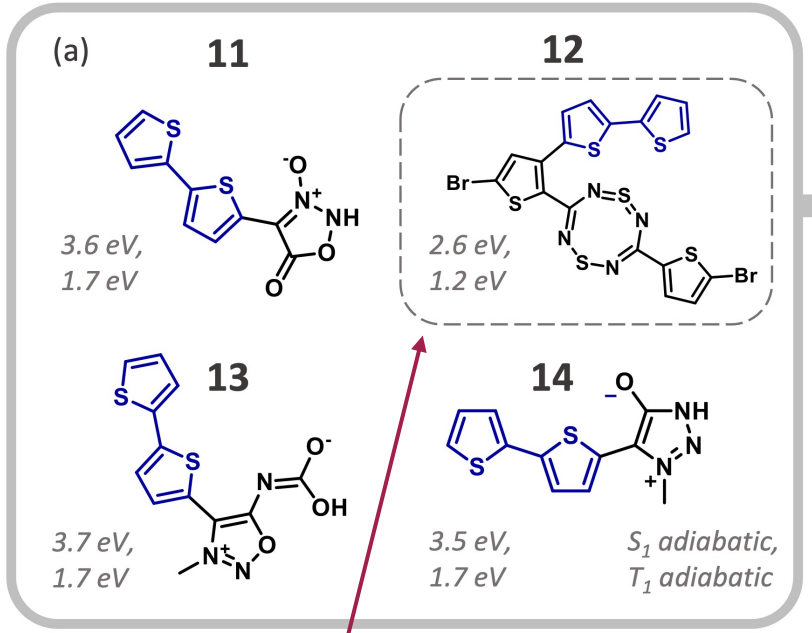
Further exploration from identified trends

- Exploiting all mesoionic-like fragments in the original FORMED database
- Combine them with bithiophene donors and screening with machine learning and DFT

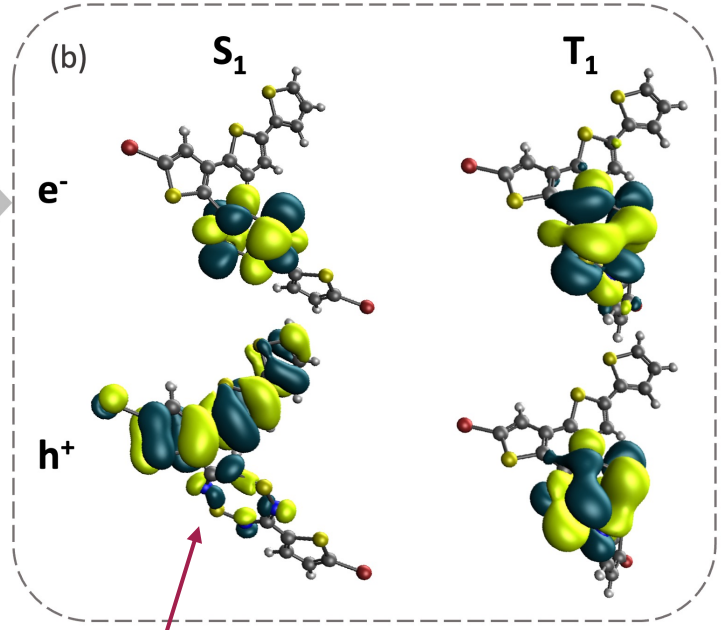
Combination of fragments



Most promising candidates

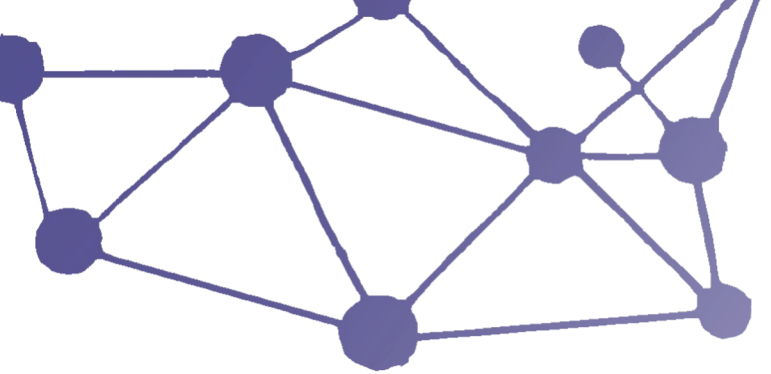


Excited state characterization



Collaboration for experimental testing of candidate

Also displays charge-transfer S_1 and locally excited T_1



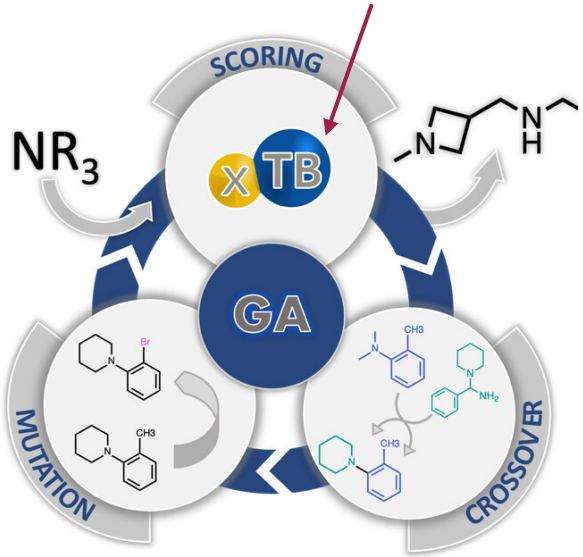
Reaction design with reactive force field methods

Simulating reactive chemistry

- Density functional theory is the go-to-method for simulations and dataset generation
- Computational cost is prohibitively high to use as property predictor for generative models
- Current work use reactive force fields or semi-empirical methods

Typical reaction simulations	This work
High computational cost	Fast force fields
High failure rate (20–70%)	Robustness of >99.9%
High variance of outputs	Exhaustive sampling
Doubtful correlation to experiment	Initial validation to experiment

Semi-empirical methods can be used in favorable cases

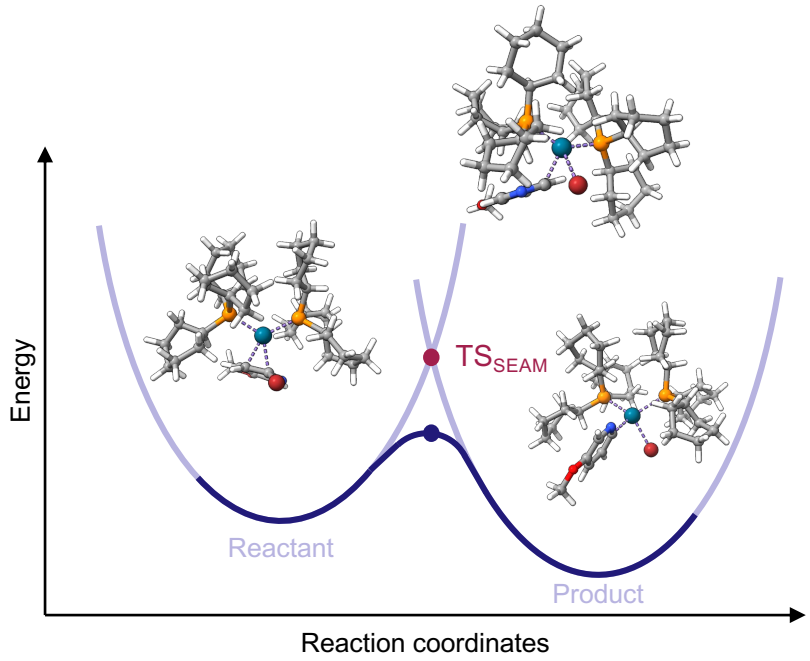


Seumer, J.; Kirschner Solberg Hansen, J.; Brøndsted Nielsen, M.; Jensen, J. H. *Angew. Chem. Int. Ed.* **2023**, *62* (18), e202218565.

Reactive force fields

- Simulate chemical reactivity approximately with the cost of force fields
- The SEAM method avoids reaction-specific parametrization

SEAM method



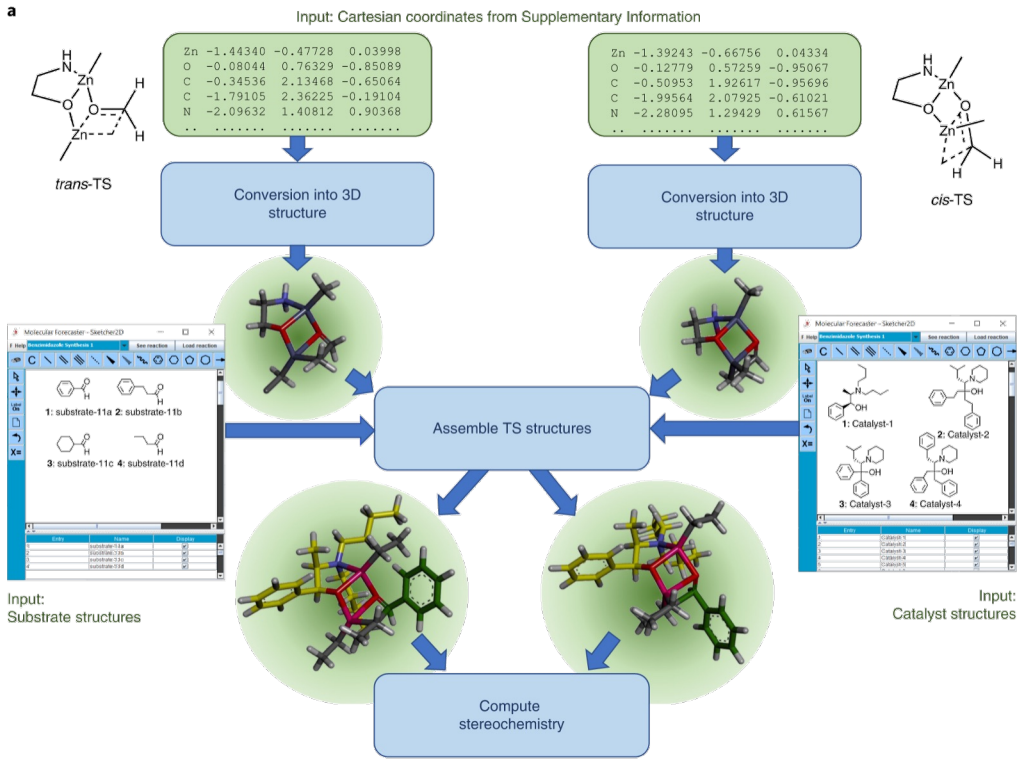
Simplified code with our POLANYI package

```
from polanyi.workflow import opt_ts_python
from polanyi.io import read_xyz

elements_r, coordinates_r = read_xyz("reactant.xyz")
elements_p, coordinates_p = read_xyz("product.xyz")

results = opt_ts_python(
        elements_r,
        [coordinates_r, coordinates_p]
        )
coordinates_ts = results.coordinates_opt
```

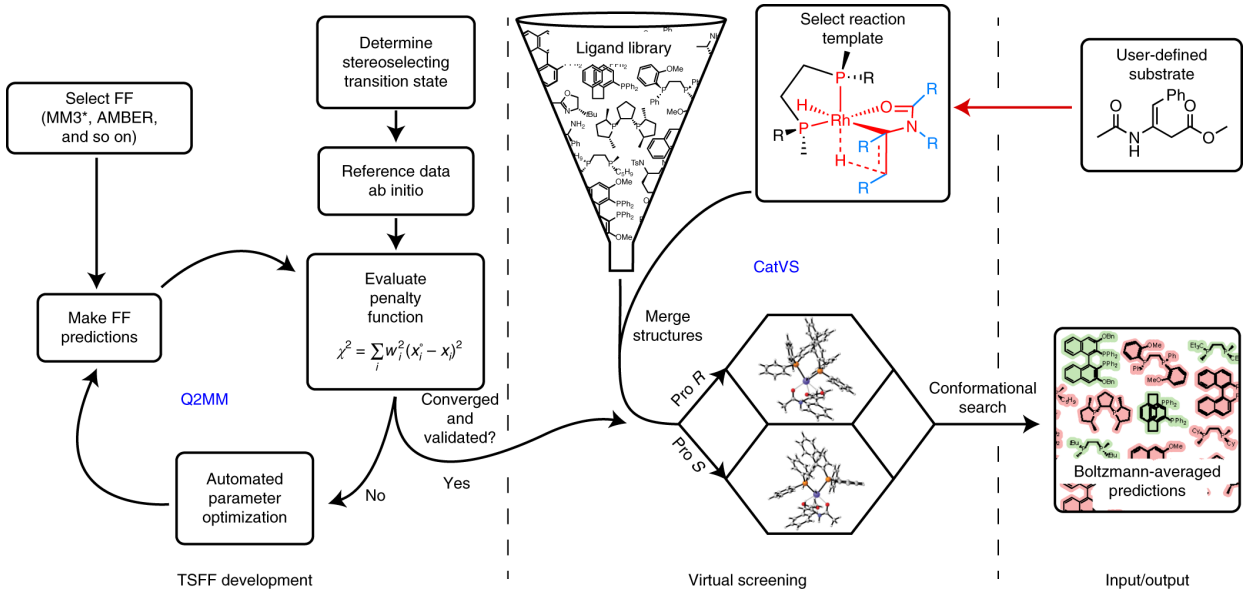
Similar methods

- Asymmetric catalyst evaluation (ACE) from Moitessier
- Quantum-Guided Molecular mechanics (Q2MM) from Norrby, Wiest

Virtual Chemist



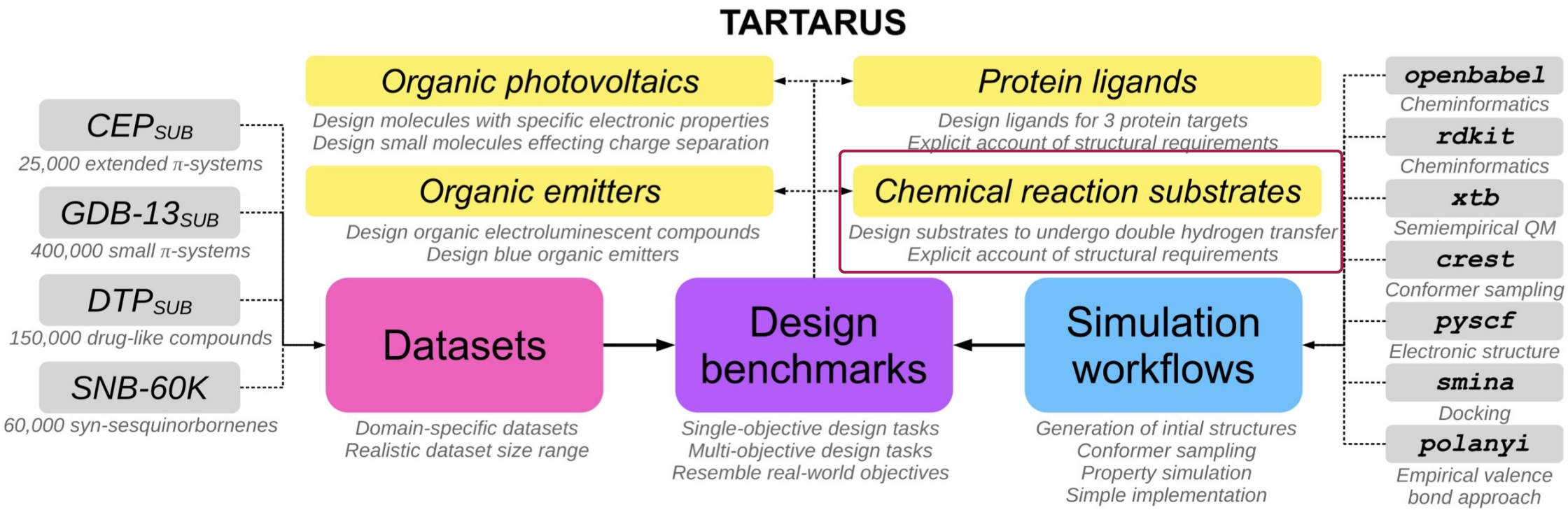
CatVS



(1) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. *Nature Catalysis* **2019**, *2* (1), 41–45. (2) Burai Patrascu, M.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P.-O.; Moitessier, N. *Nat Catal* **2020**, *3* (7), 574–584.

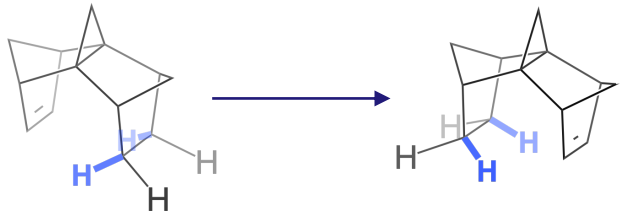
“Realistic” benchmark set for generative models

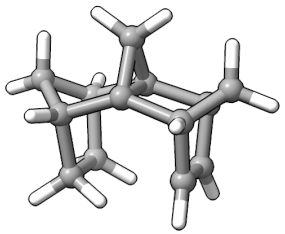
- Moderately complex tasks related to important applications
- Benchmarks based on fast simulations rather than surrogate ML models

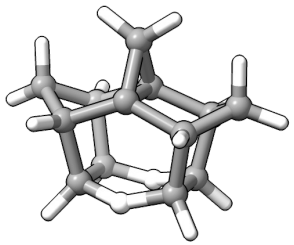


Reaction benchmark for generative models

- Double hydrogen atom transfer of *syn*-sesquinorbornenes
- Benchmarked against experiment with SEAM
- Very robust optimization







Reactant

Geometry from openbabel
CREST conformer search

Product

Geometry from reactant
CREST conformer search

SEAM optimization

Geometry from interpolation
POLANYI-optimization

Transition state

Constrained CREST search
POLANYI optimization

E_R

Reactant energy

E_{TS}

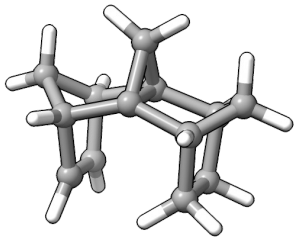
Transition state energy

E_P

Product energy

Activation energy

$\Delta E^{\ddagger} = E_{TS} - E_R$

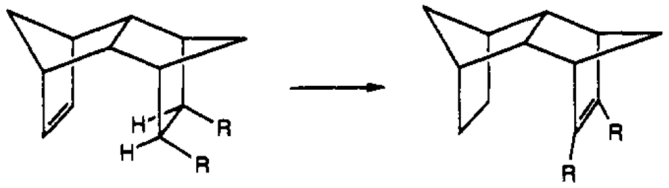


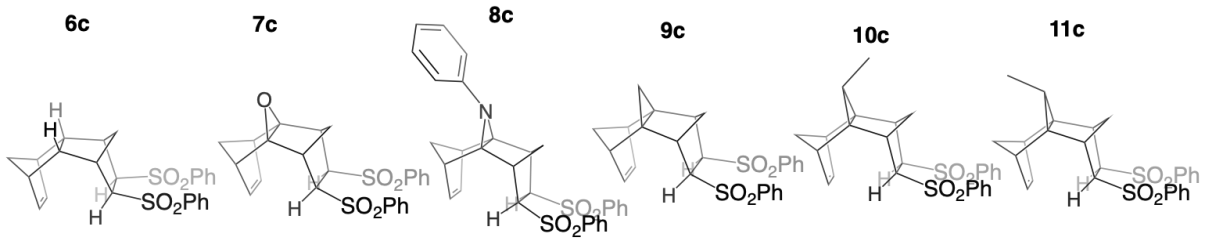
$\Delta E_r = E_P - E_R$

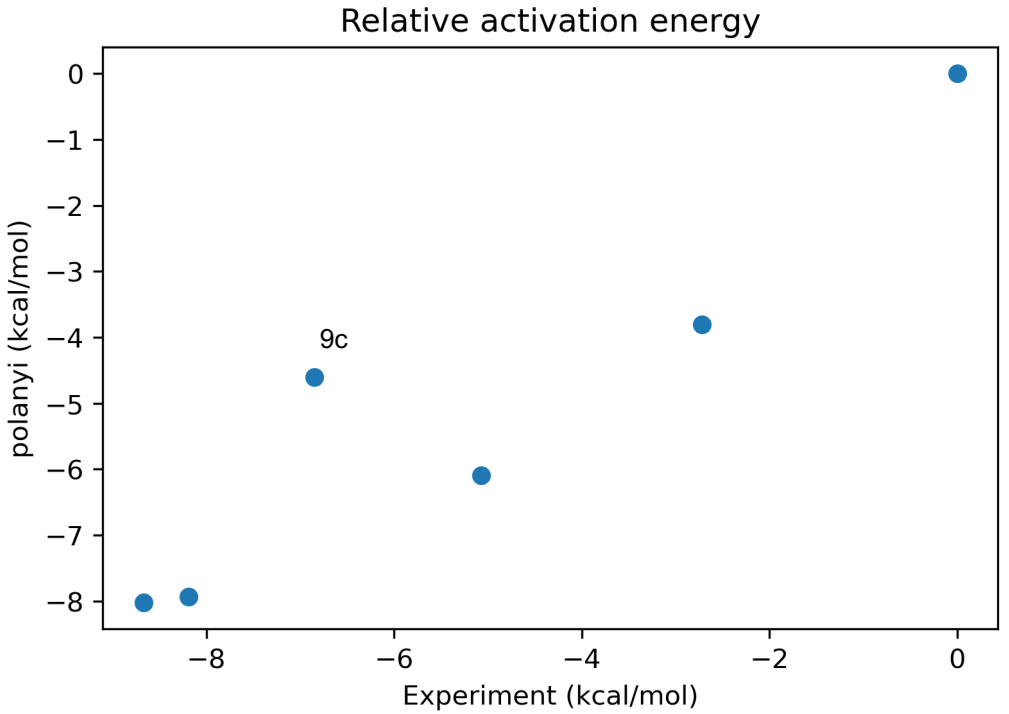
Reaction energy

Reaction benchmark for generative models

- Decent correlation with (very limited) experimental data
- No solvent model (exp: bromobenzene)



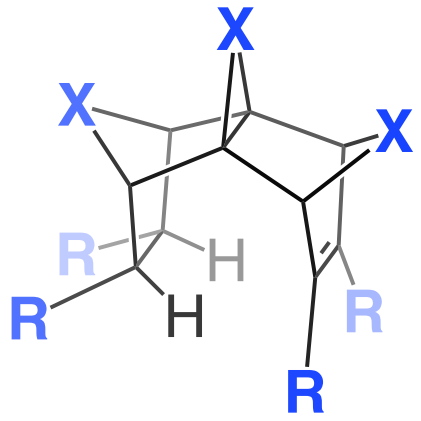




Jensen, F. *J. Am. Chem. Soc.* **1992**, *114* (5), 1596–1603.

Dataset generation

- Start from unsubstituted compound
- Perform STONED SELFIES mutations
- Filter for original fragment
- Obtain 60’850 examples
- Filter out ”unwanted” fragments

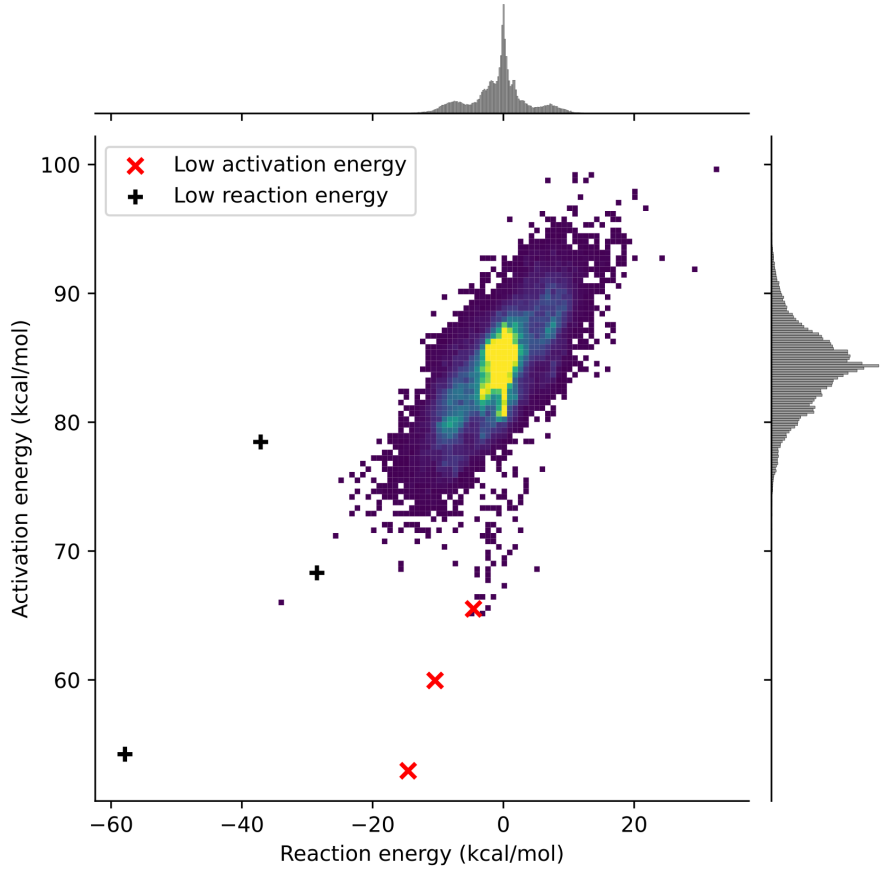



```
[C-], [S-], [O-], [N-], ‘[*+], [*-] [PH], [pH], [N&X5], *=[S,s;!R], [S&X3], [S&X4],
[S&X5], [S&X6], [P,p], [B,b,N,n,O,o,S,s]~[F,Cl,Br,I], *=*=*, *#*,
[O,o,S,s]~[O,o,S,s], [N,n,O,o,S,s]~[N,n,O,o,S,s]~[N,n,O,o,S,s],
[N,n,O,o,S,s]~[N,n,O,o,S,s]~[C,c]=, :[O,o,S,s,N,n;!R], *=N-[*;!R], *~[N,n,O,o,S,s]-
[N,n,O,o,S,s;!R]
```


Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. dos P.; Aspuru-Guzik, A. *Chem. Sci.* **2021**, *12* (20), 7079–7090.

Dataset & generative modelling

- Reference dataset of 60,850 reactions
- 99.94% success rate
- ~ 12 min/molecule/CPU
- Limited by conformational sampling
- Interfaced with generative models

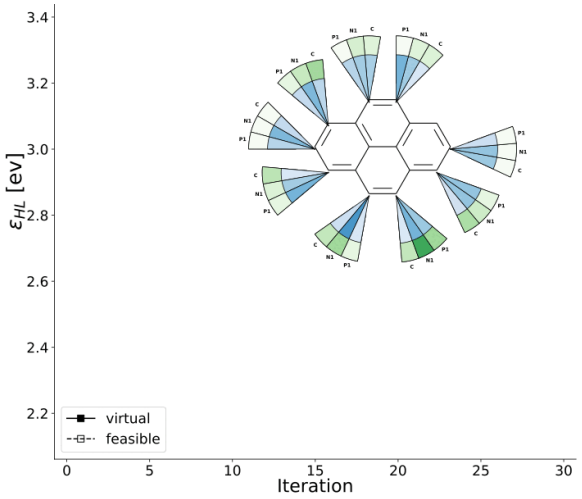


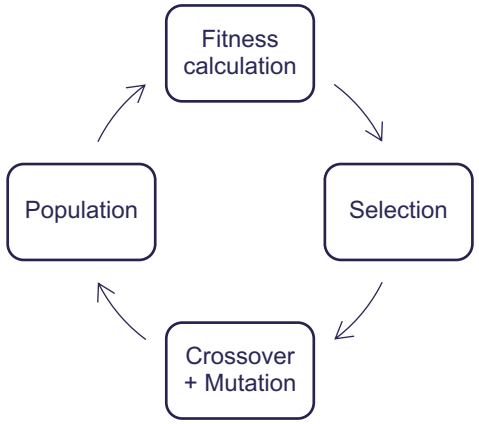
	$\Delta E^{\ddagger}$	ΔE_r	$\Delta E^{\ddagger} + \Delta E_r$	$-\Delta E^{\ddagger} + \Delta E_r$
Parent Substrate	85.16	0.00	85.16	-85.16
Dataset	64.94	-34.39	56.48	-95.25
SMILES-VAE	76.81 ± 0.25	-10.96 ± 0.71	71.01 ± 0.62	-90.94 ± 1.04
SELFIES-VAE	72.45 ± 3.79	-10.45 ± 3.83	72.05 ± 0.00	-87.82 ± 2.13
SMILES-LSTM-HC	59.64 ± 4.10	-31.03 ± 16.15	71.81 ± 1.56	-91.58 ± 2.14
SELFIES-LSTM-HC	63.17 ± 4.34	-21.02 ± 4.95	68.06 ± 5.74	-96.59 ± 4.59
REINVENT	68.38 ± 2.00	-24.35 ± 6.46	55.25 ± 5.88	-94.52 ± 1.20
GB-GA	56.04 ± 3.07	-41.39 ± 5.76	45.20 ± 6.78	**-100.07 ± 1.35**
JANUS	**47.56 ± 2.19**	**-45.37 ± 7.90**	**39.22 ± 3.99**	-97.14 ± 1.13

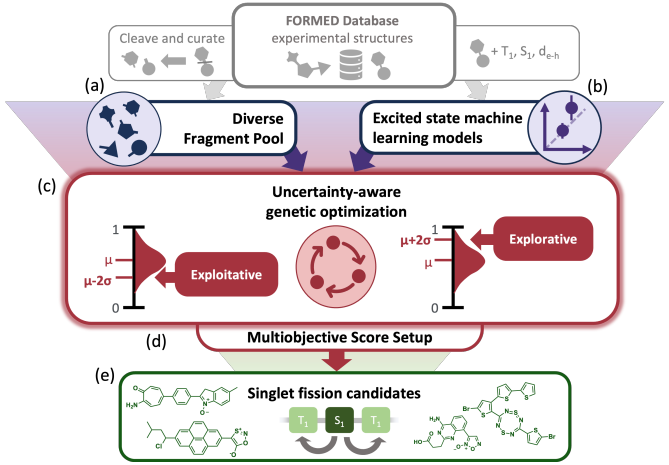
Nigam, A.; Pollice, R.; Tom, G.; Jorner, K.; Thiede, L. A.; Kundaje, A.; Aspuru-Guzik, A. *arXiv*, **2022**.

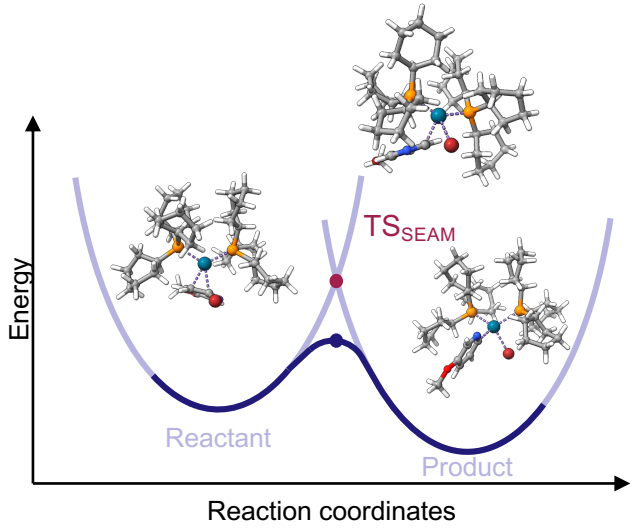
Conclusions and outlook

- Computer-aided molecular design promises to accelerate chemical discovery
- Reliable property predictors that generalize out-of-domain are essential
- Cheap physics-based predictors with limited number of parameters work well
- Alternative design algorithms work in-domain by, *e.g.*, fragment approaches









Acknowledgements

University of Toronto


Alán Aspuru-Guzik


Gary Tom


Rajvi Rana


Cyrille Lavigne


Robert Pollice


Luca Thiede


John Willes


Rodrigo Vargas

Stanford University


AkshatKumar Nigam


Anshul Kundaje

Swedish Research Council

EPFL


Clémence Corminboeuf


Rubén Laplaza


Terence Blaskovits

ETH


Luca Schaufelberger