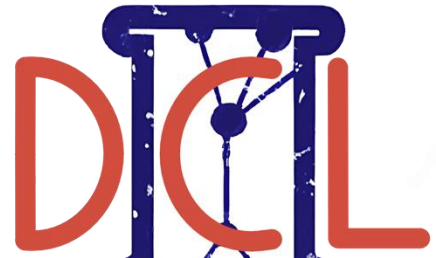
Digital tools for reaction prediction and optimization

Prof. Dr. Kjell Jorner

Assistant Professor of Digital Chemistry



ETH zürich

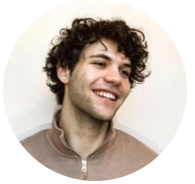
The Digital Chemistry Laboratory

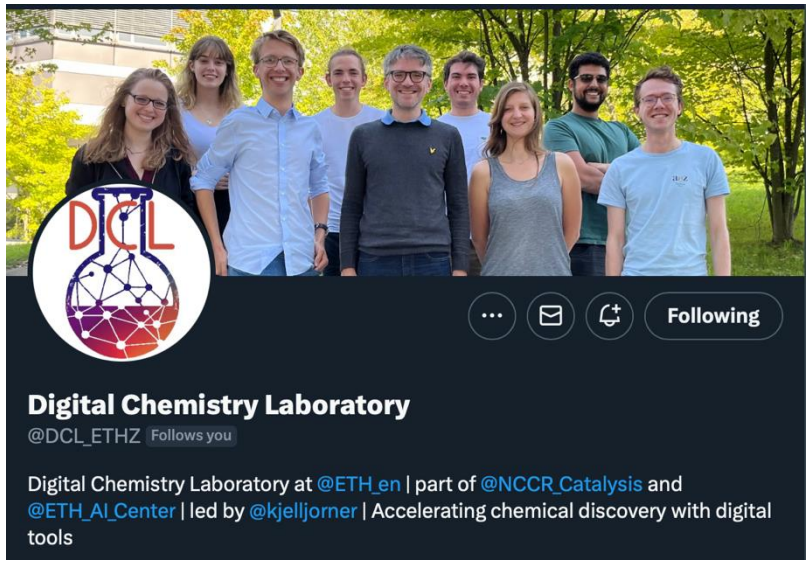
Department of Chemistry and Applied Biosciences (D-CHAB)
Institute of Chemical and Bioengineering (ICB)

Principal investigator
Prof. Dr. Kjell Jorner

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

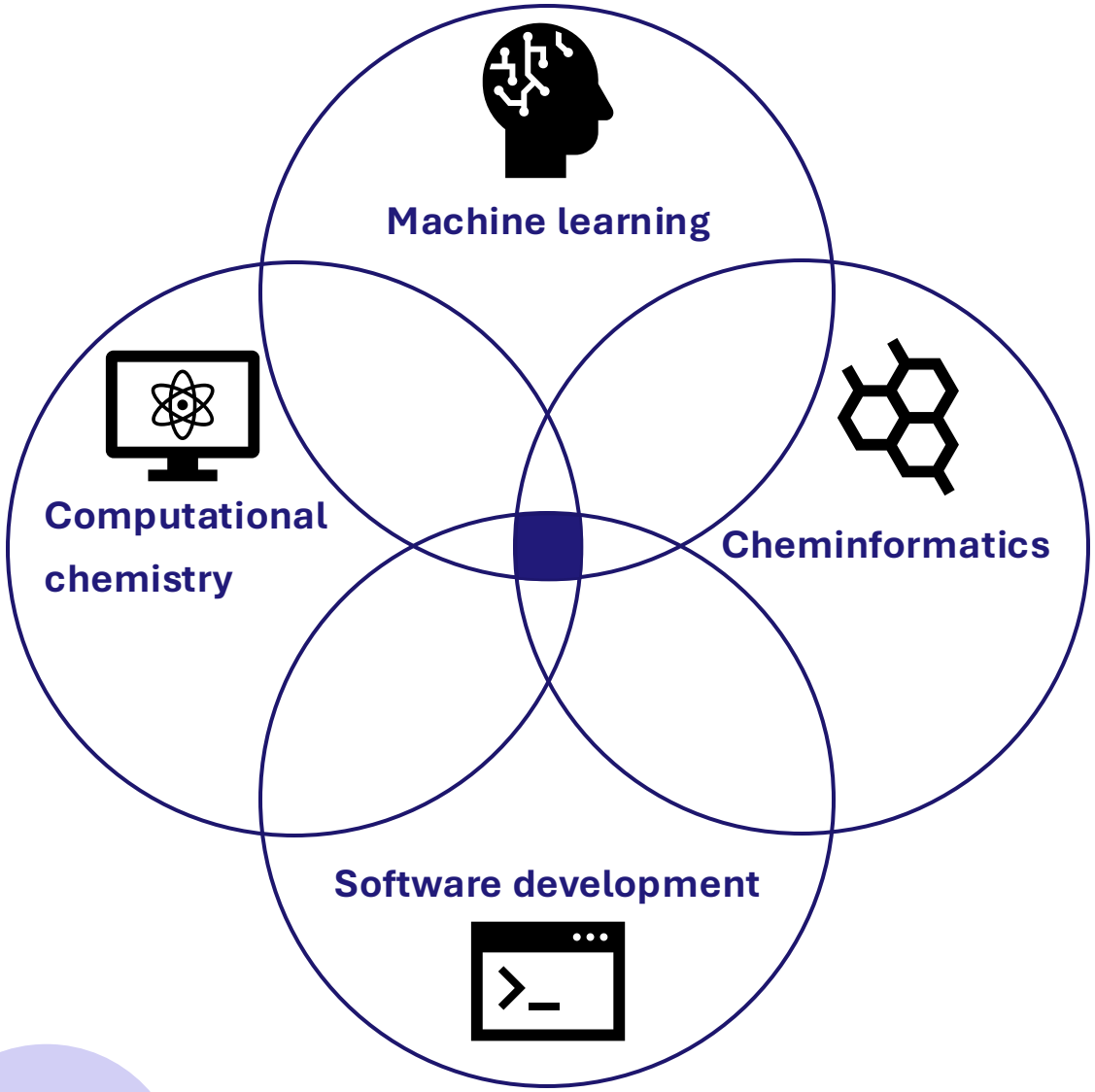
Co-supervisor
Vignesh Ram Somnath (w. Krause)
Riccardo De Santi (w. Krause, He)





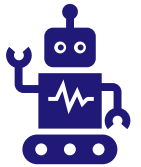
@DCL_ETHZ

Digital Chemistry



Lab chemist

High-throughput experimentation



Self-driving labs

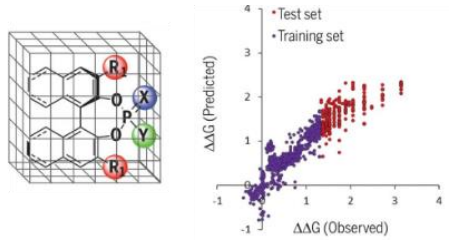
Mission: Accelerate chemical discovery with digital tools

Strategic core research

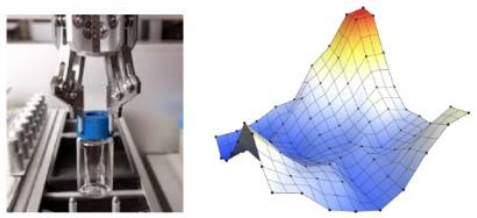
Opportunity-driven

Reactivity and catalysis

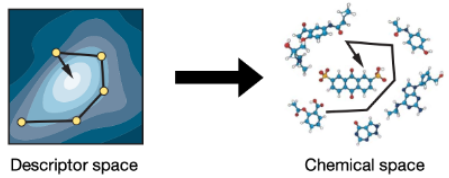
Reaction outcome prediction



Reaction optimization

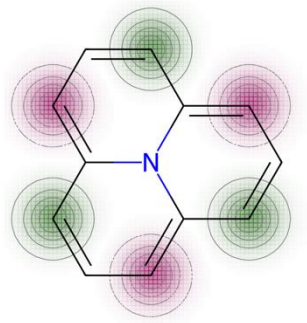


Catalyst & reaction design

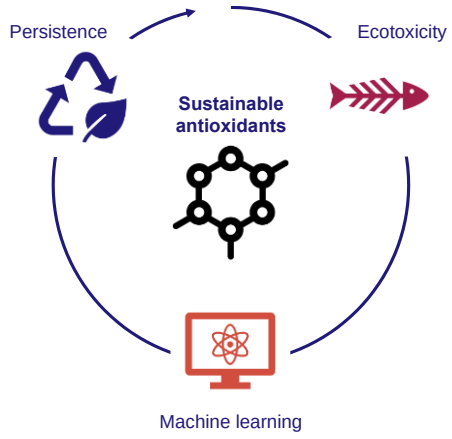


Computer-aided molecular design

Organic electronic materials



Safe and sustainable by design



NCCR Catalysis

Our niche: Models in the low-data regime

Simulations

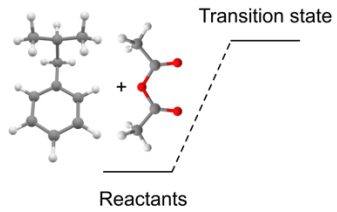
Chemistry-informed machine learning

Machine learning

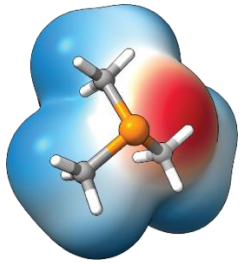
No data

Lots of data

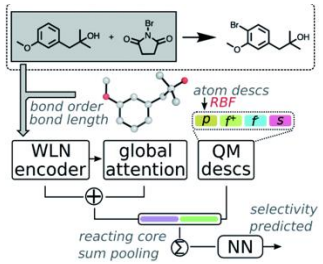
Mechanistic DFT



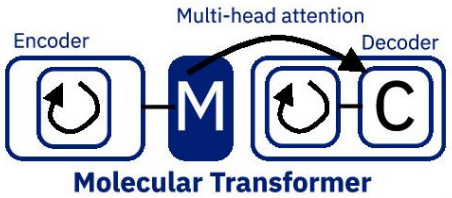
Chemical descriptors



Chemistry-augmented deep learning



Deep learning



Sinergia project – Safe & sustainable by design

Kathrin Fenner
University of Zurich
& Eawag

Persistence

Sustainable antioxidants

Ecotoxicity

Beate Escher
Helmholtz Centre for
Environmental Research

Kjell Jorner
ETH Zurich

Machine learning

High-throughput experimentation

Machine learning models

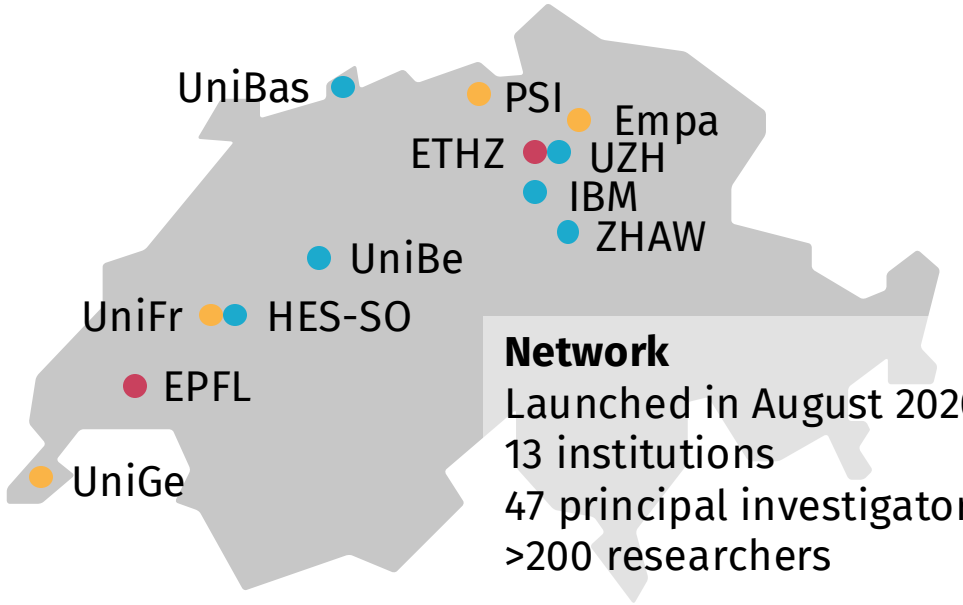
Computer-aided molecular design

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

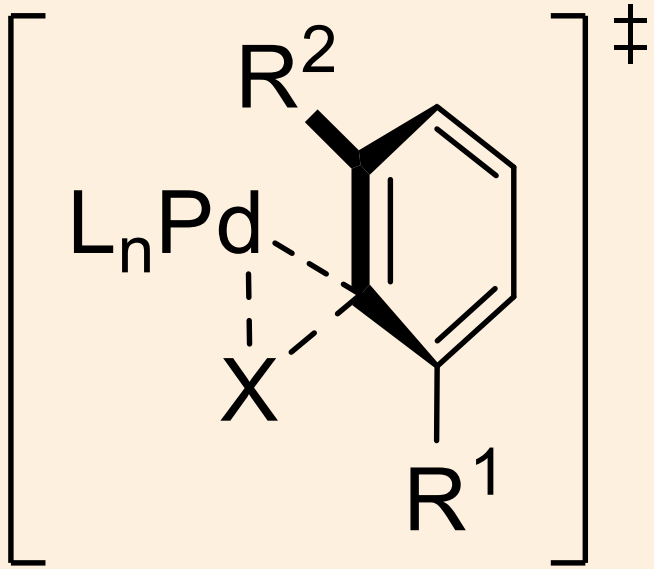
Outline

- Reaction prediction with transition state descriptors
- Calculating descriptors with MORFEUS
- Generality-oriented Bayesian optimization of reaction conditions

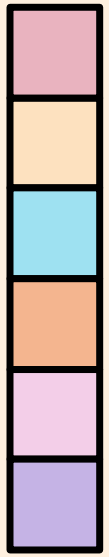
Reaction prediction with transition state descriptors

Main idea

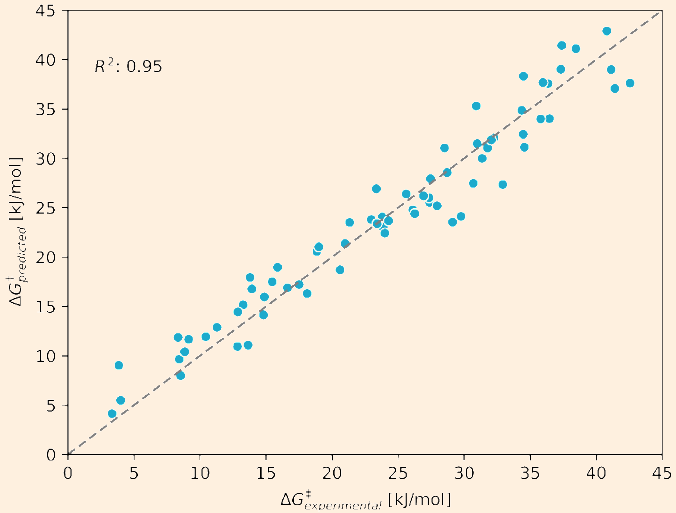
TS geometries



TS descriptors

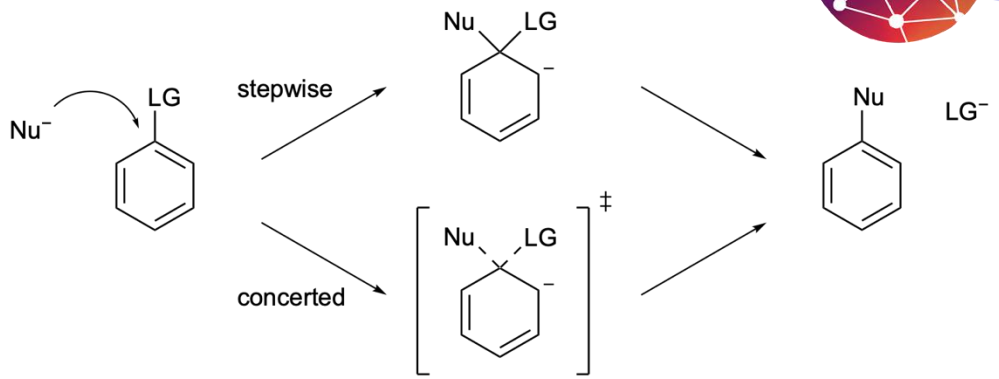


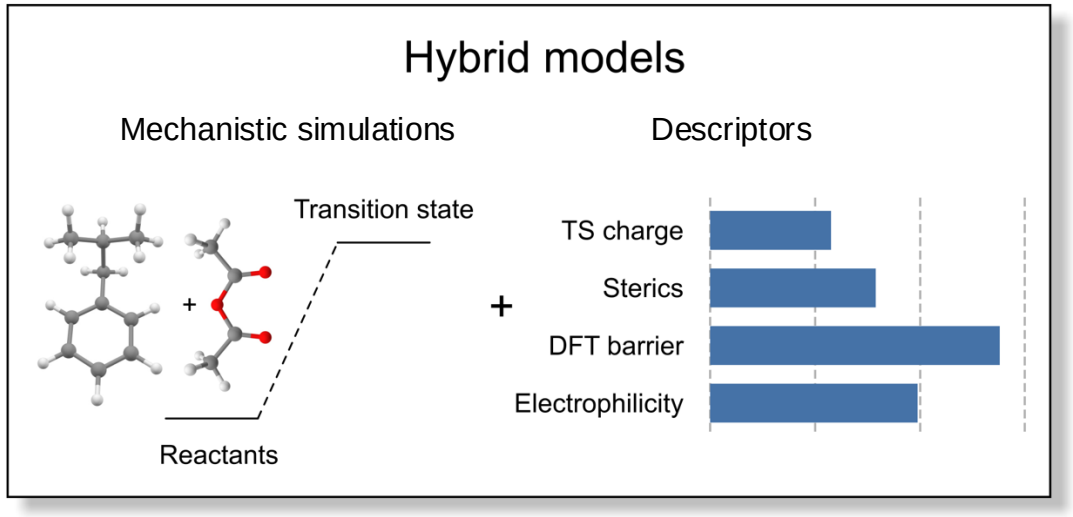
ML predictions

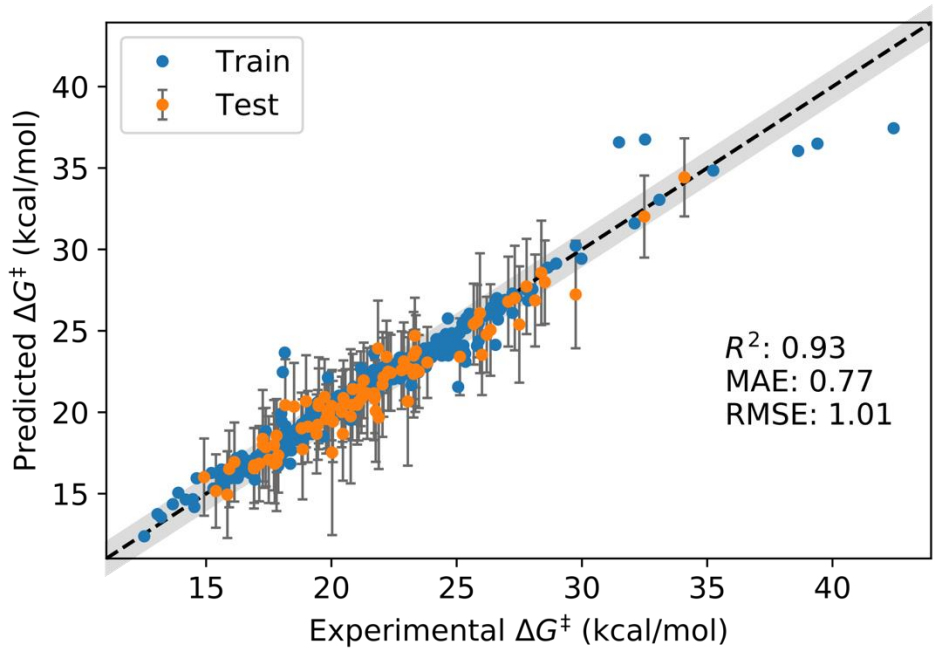


Hybrid mechanistic & ML models – Predict-S_NAr

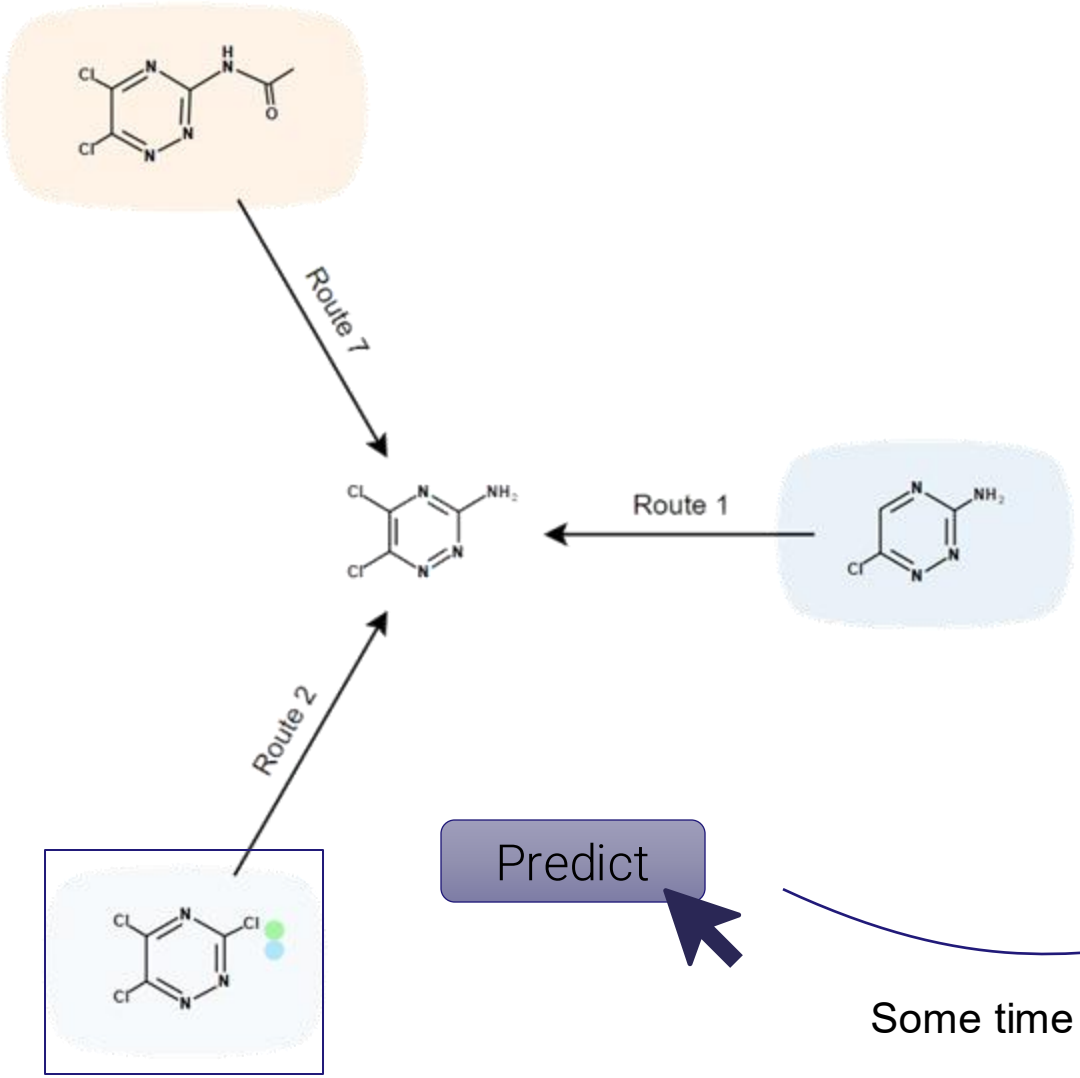
- Nucleophilic aromatic substitution
- 10% of all reactions in pharma
- Automated simulation and ML workflow
- Predict activation energies
- ~450 reaction rates from the literature



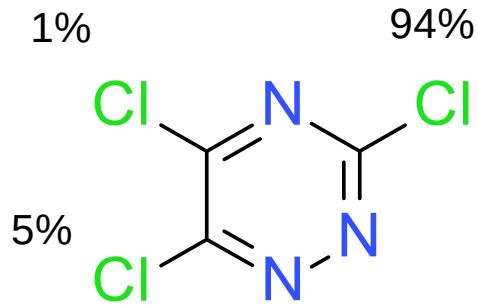




Future AI-powered route designer software



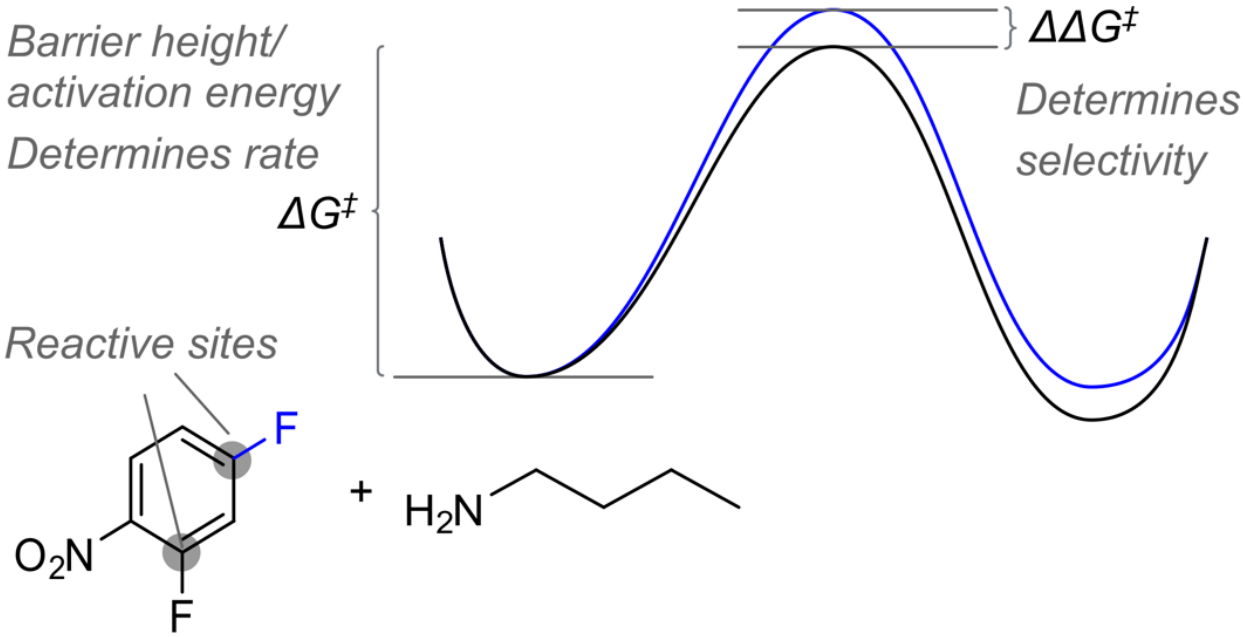
Yield prediction



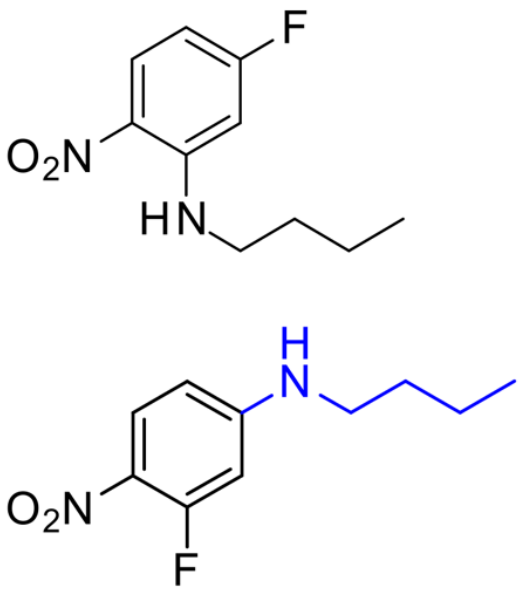
Indications:

- Mediated by general base
- Selectivity best in acetonitrile

Activation energies as target



Kinetically preferred isomer



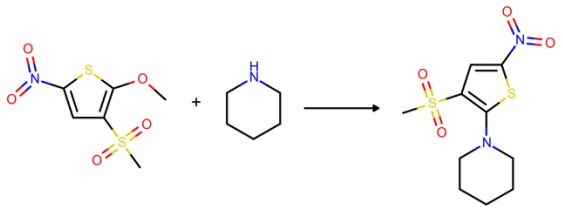
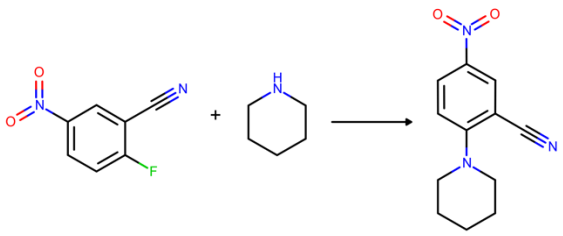
Can predict

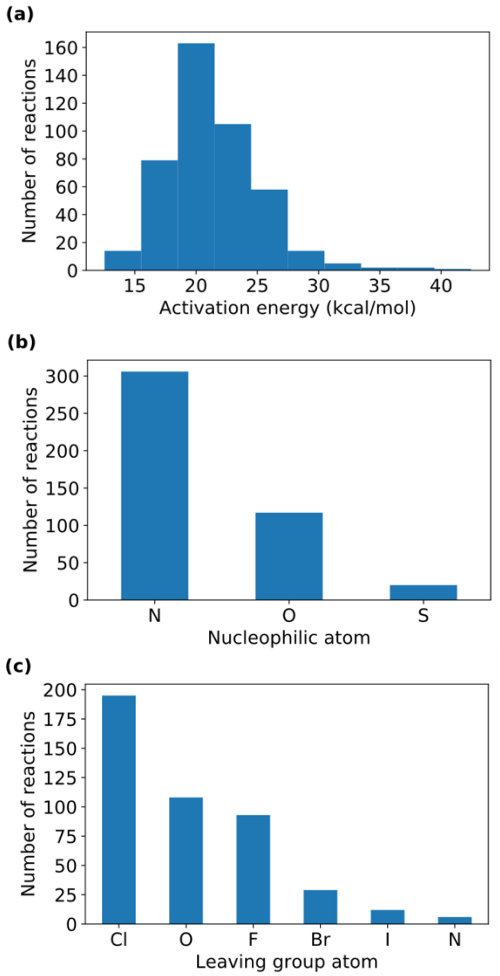
- Absolute reactivity
- Selectivity
- Temperature & solvent dependence

Compare work of the Varnek group: *SAR QSAR Environ. Res.* **2021**, *32* (3), 207–219 and references therein.

Dataset

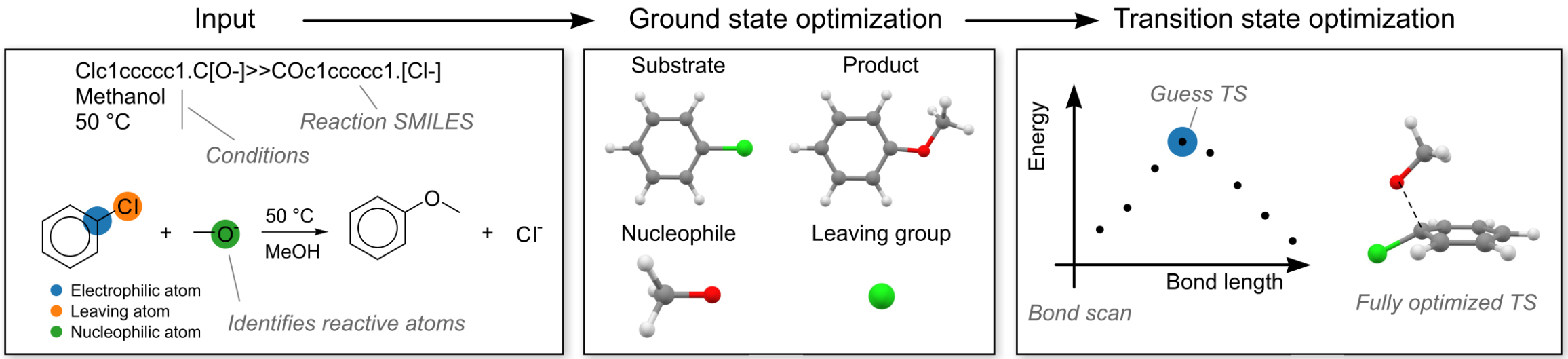
- Train on reaction barriers
- Kinetic data (ca 450 points)
- Collected from the literature

Reaction	Temperature	Solvent	Barrier
	20 °C	DMSO	16.3
	25 °C	Acetonitrile	17.9



Predict-S_NAr reaction platform

Automated reaction path workflow

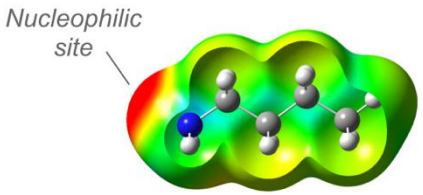


GFN2-xTB
ωB97X-D/6-311+G(d,p)//6-31+G(d)

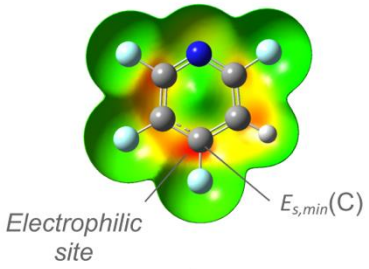
Automated descriptor calculation

Electronic structure

Local ionization energy ($I_{s,min}$)



Local electron attachment energy ($E_{s,min}$)

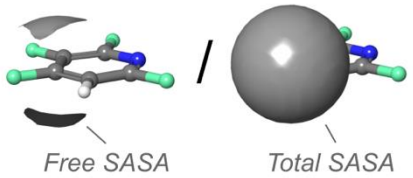


+

Surface electrostatic potential ($\bar{V}_s$)
Electrostatic potential at nuclei (V_N)
DDEC6 atomic charges (q)
DDEC6 bond orders (BO)
Local nucleophilicity (ℓ_N)
Local electrophilicity (ℓ_ω)
Global electrophilicity (ω)
Global nucleophilicity (N)

Sterics & dispersion

Ratio of free to total atomic solvent accessible surface area ($SASA_r$)



+

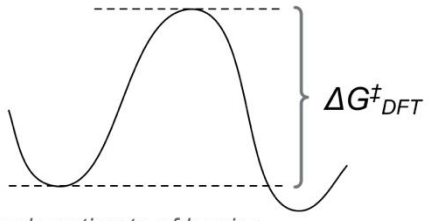
P_{int} dispersion descriptor

Conditions

Solvent PCA components (PC_i)

Transition state features

DFT activation energy ($\Delta G^{\ddagger}_{DFT}$)
DDEC6 atomic charges (q_{TS})
DDEC6 bond orders (BO_{TS})
Electrostatic potential at nuclei ($V_{N,TS}$)

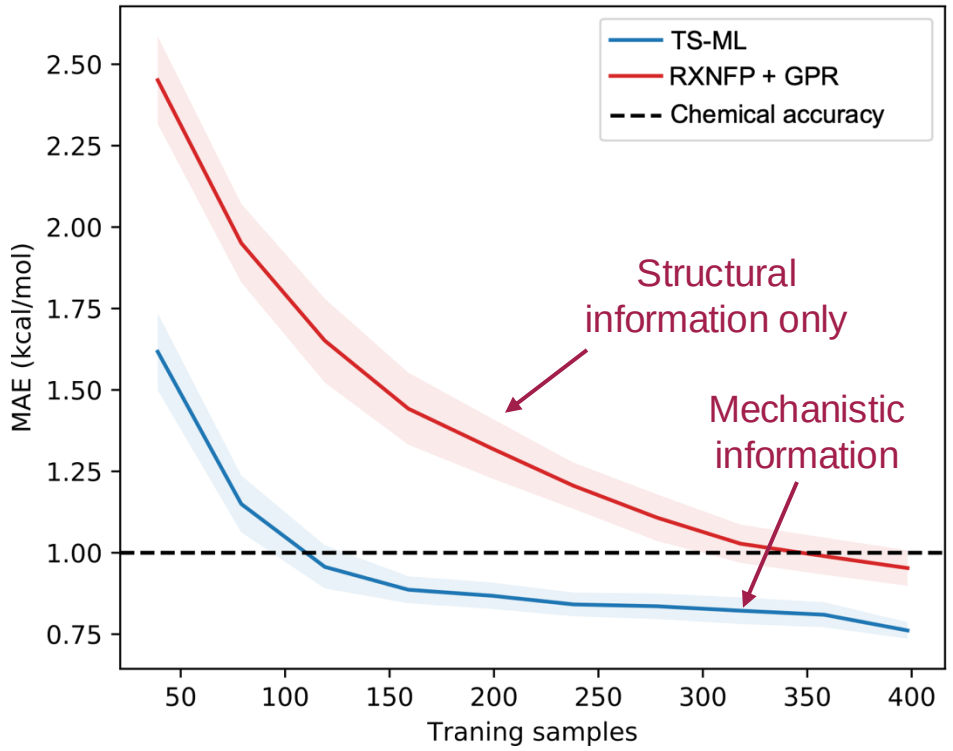


Crude estimate of barrier

Predict activation energies

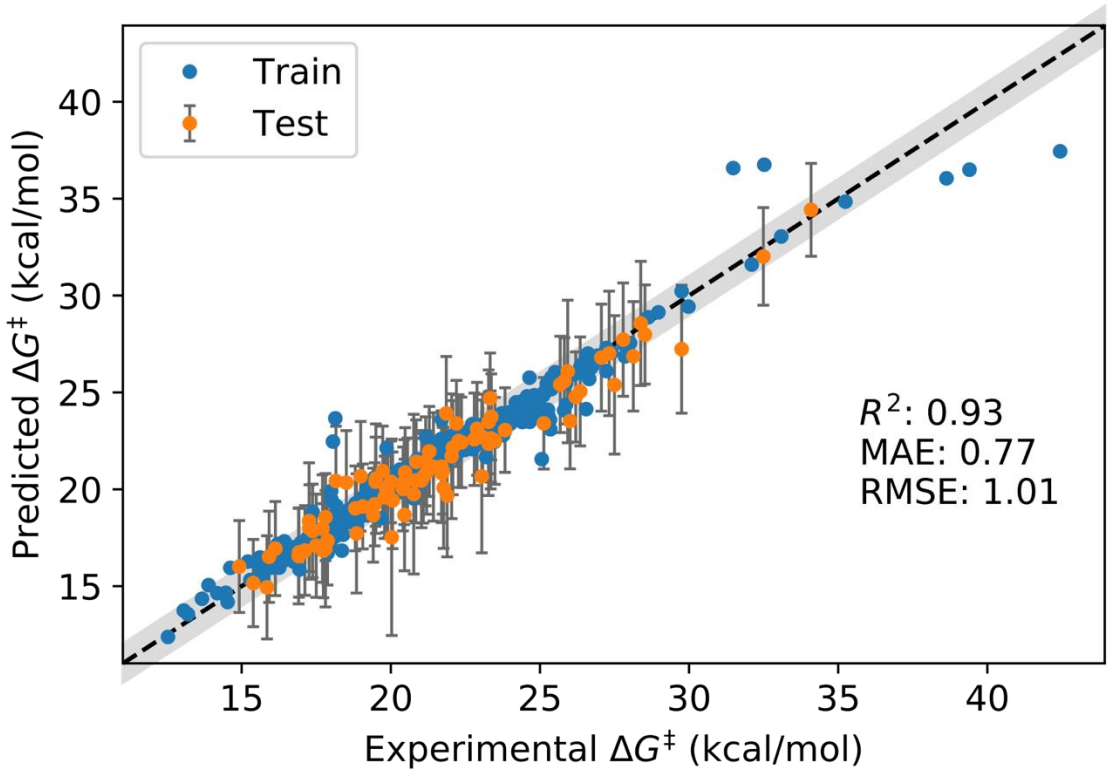
- 450 points from kinetics literature

Learning curves



Gaussian Process Regression, Matern 3/2 kernel

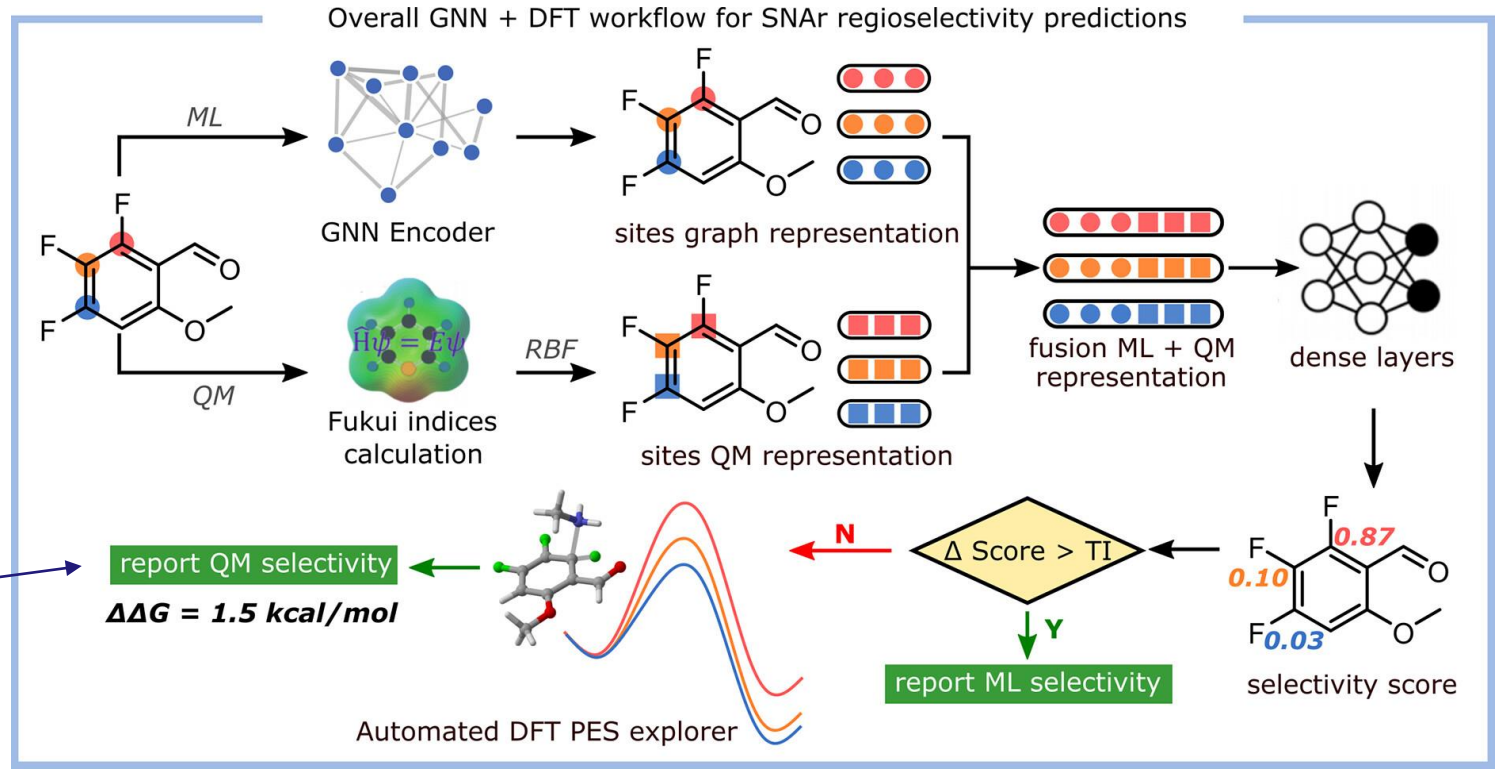
Prediction on test set



Prediction uncertainty • Model validation • Applicability domain • Regio-/chemoselectivity • Feature importances done right

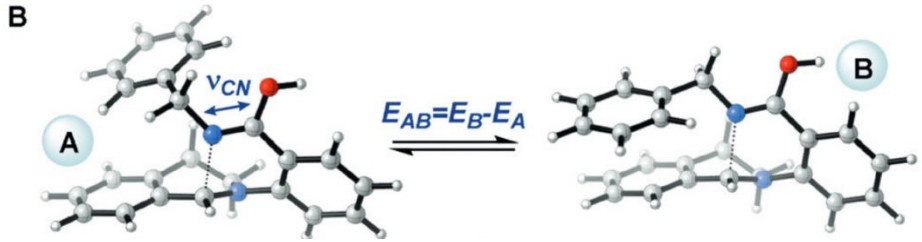
S_NAr workflow from Pfizer

- Chemistry-informed deep learning prediction of selectivity
- Use ML uncertainty to trigger launch of DFT workflow
- 96.3% on in-house data and 94.7% on USPTO data

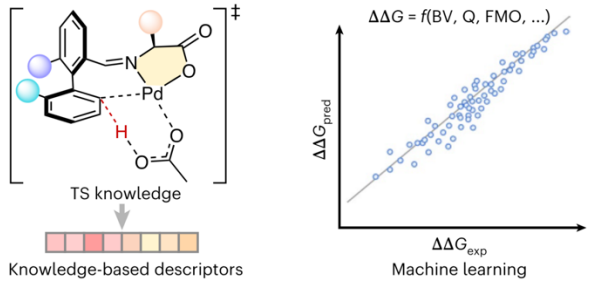


Based on Predict-S_NAr workflow

Transition state descriptors

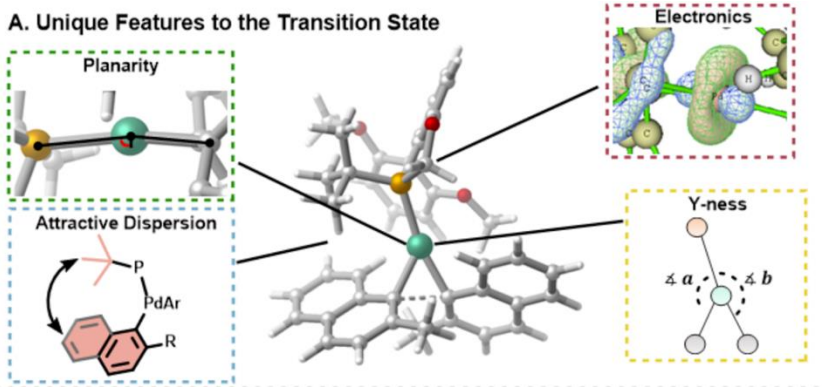


Orlandi, M.; Toste, F. D.; Sigman, M. S. *Angew. Chem. Int. Ed.* **2017**, *56* (45), 14080–14084.

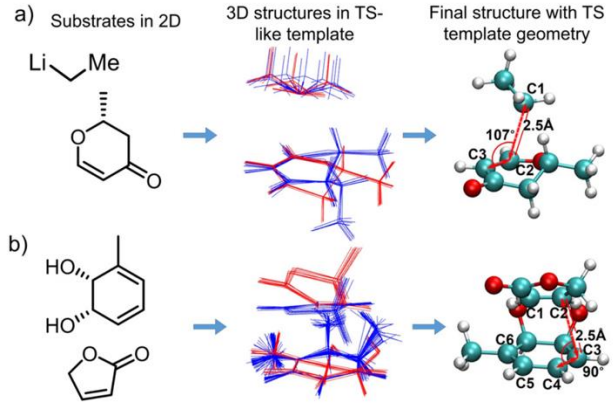


Descriptors from MORFEUS

Xu, L.-C.; Frey, J.; Hou, X.; Zhang, S.-Q.; Li, Y.-Y.; Oliveira, J. C. A.; Li, S.-W.; Ackermann, L.; Hong, X. *Nat. Synth.* **2023**



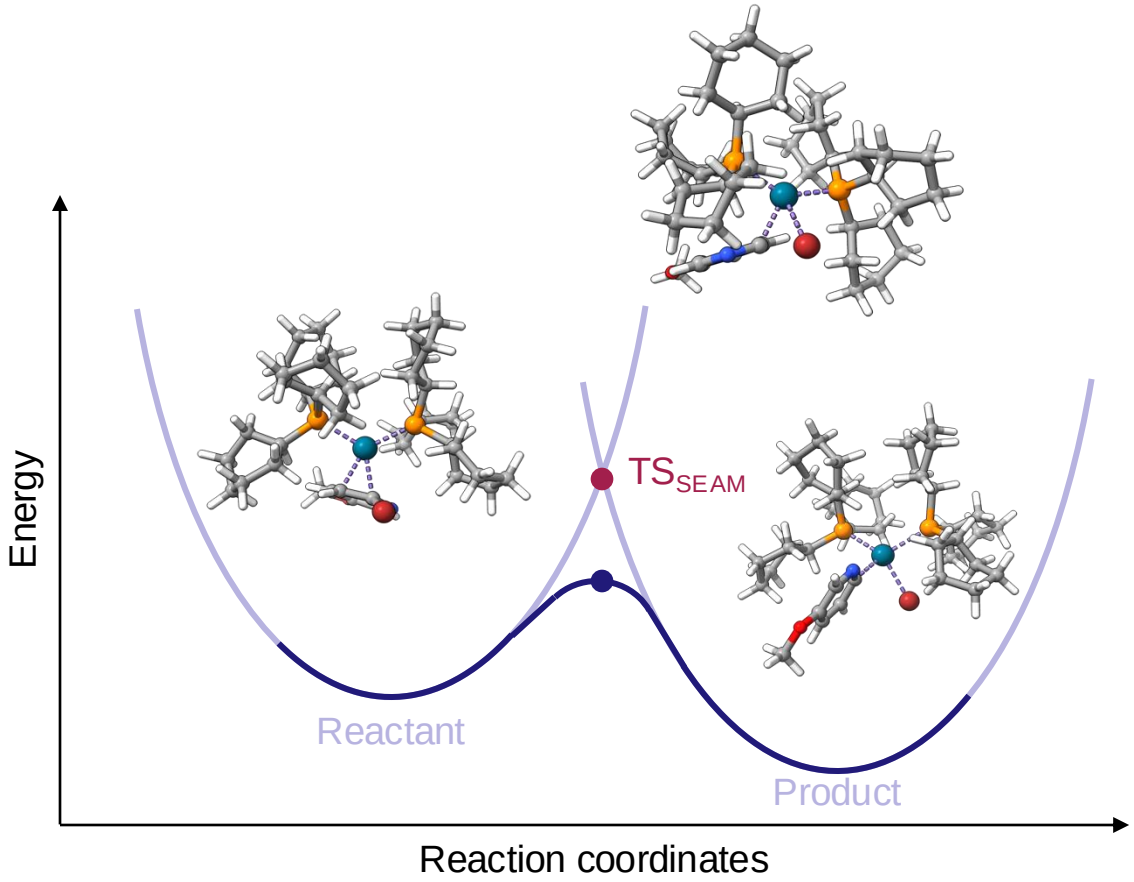
Cuomo, A. E.; Ibarraran, S.; Sreekumar, S.; Li, H.; Eun, J.; Menzel, J. P.; Zhang, P.; Buono, F.; Song, J. J.; Crabtree, R. H.; Batista, V. S.; Newhouse, T. R. *ACS Cent. Sci.* **2023**, *9* (9), 1768–1774.



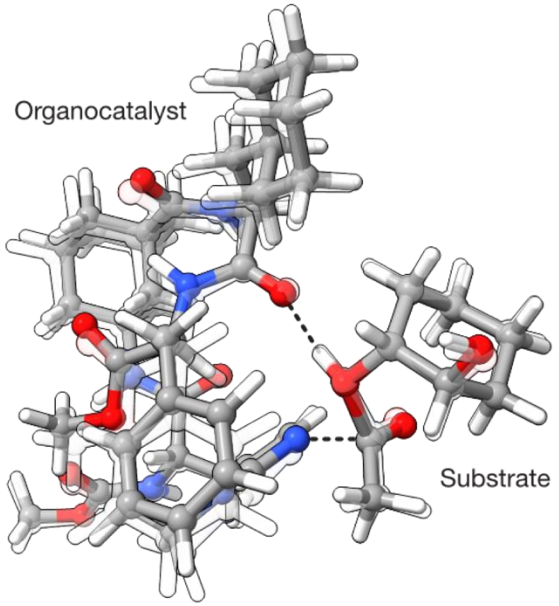
Moskal, M.; Beker, W.; Szymkuć, S.; Grzybowski, B. A. *Angew. Chem. Int. Ed.* **2021**, *60* (28), 15230–15235.

Luo, S.; Liu, L.; Lyu, C.-J.; Sim, B.; Liu, Y.; Gong, H.; Nie, Y.; Zhao, Y.-L. *Cell Rep. Phys. Sci.* **2022**, 101128.

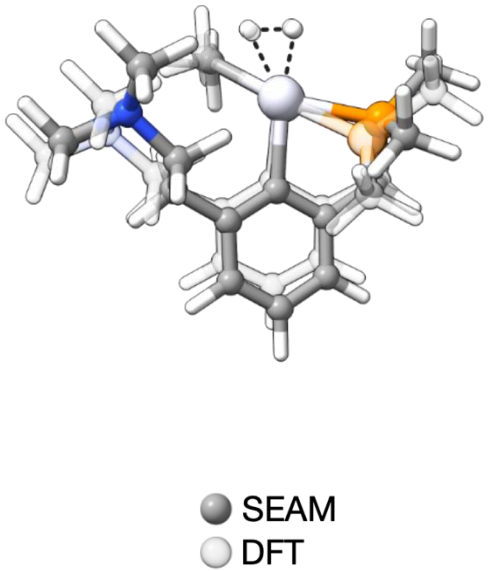
The SEAM method



Organocatalyst



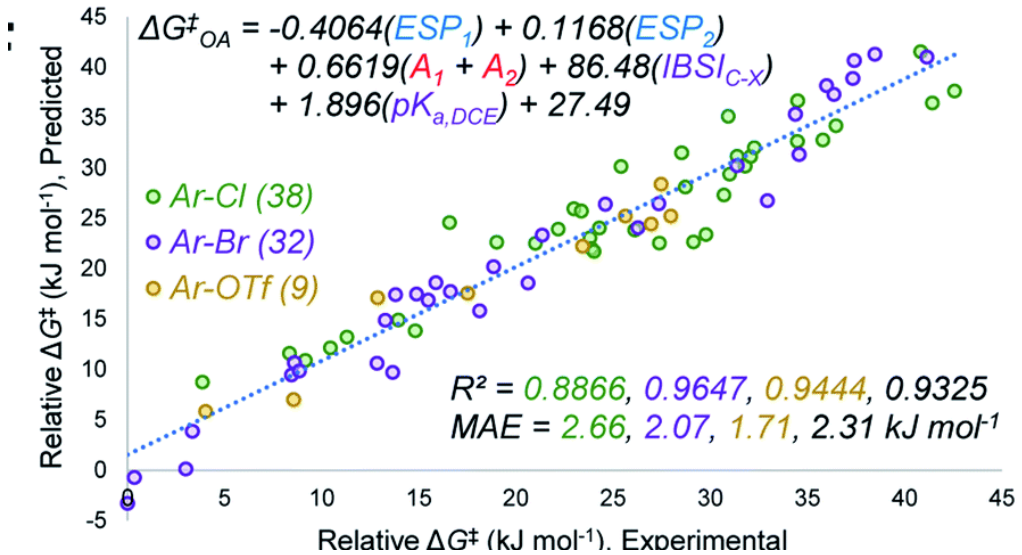
Transition metal catalyst



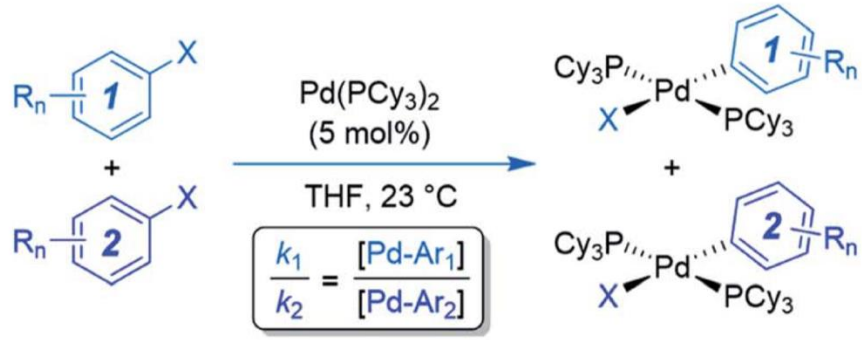
Jensen, F. *J. Am. Chem. Soc.* **1992**, *114* (5), 1596–1603.

Does SEAM work for oxidative addition to Pd?

- Can SEAM produce reliable TSs for oxidative addition on palladium?
- Can TS descriptors predict activation energies better than just ground state descriptors?
- Reference dataset from Leitch of 79 rates based on competition experiments



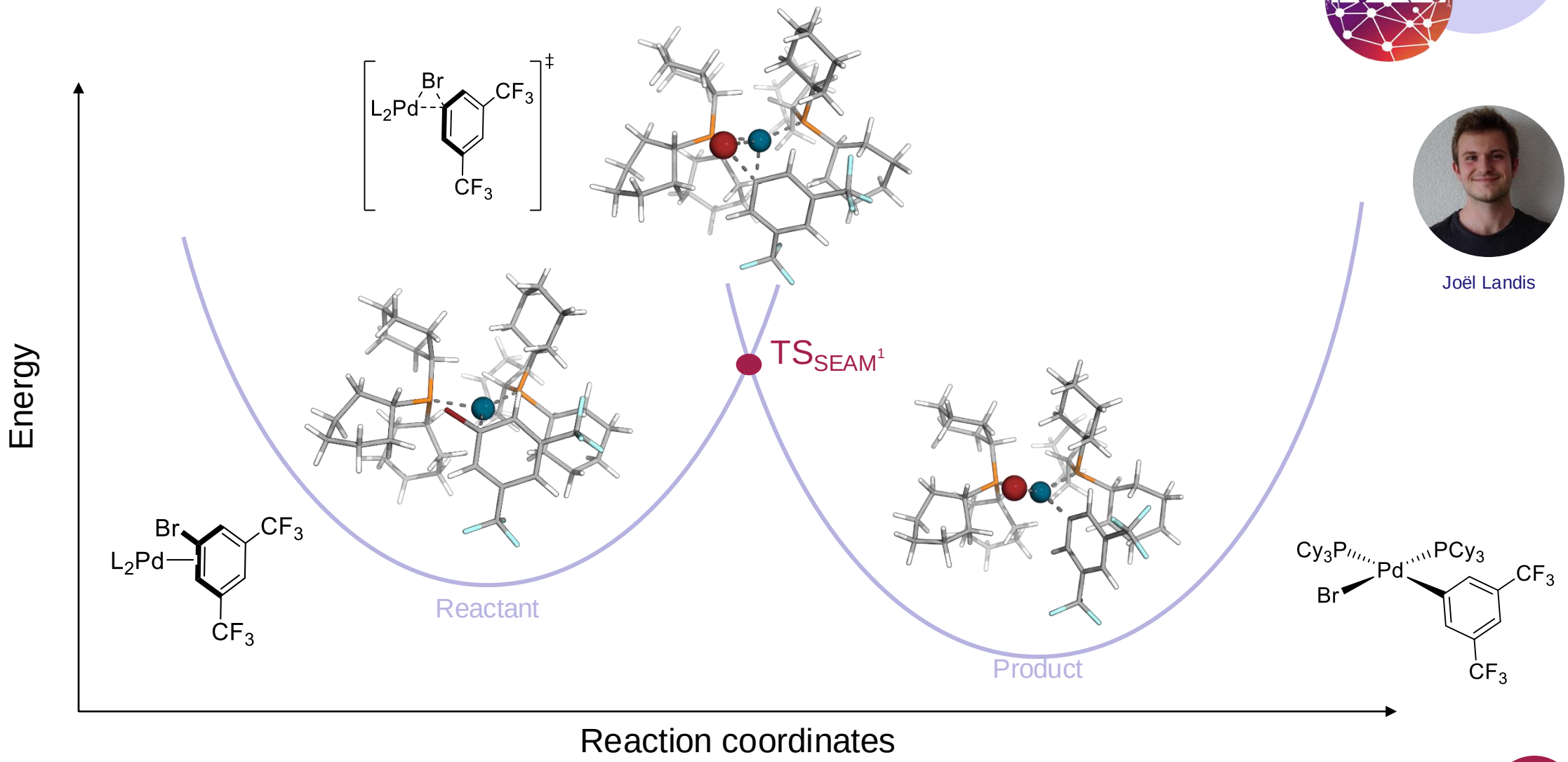
B ***Relative rates via competition experiments:***



- Product ratio determined by ^{31}P NMR spectroscopy
- Relative rates measured for 79 substrates
- Substrate set includes Ar-Cl, Ar-Br, and Ar-OTf
- Heteroaromatics and varied subsitution patterns studied
- Oxidative addition rate constants span 7 orders of magnitude

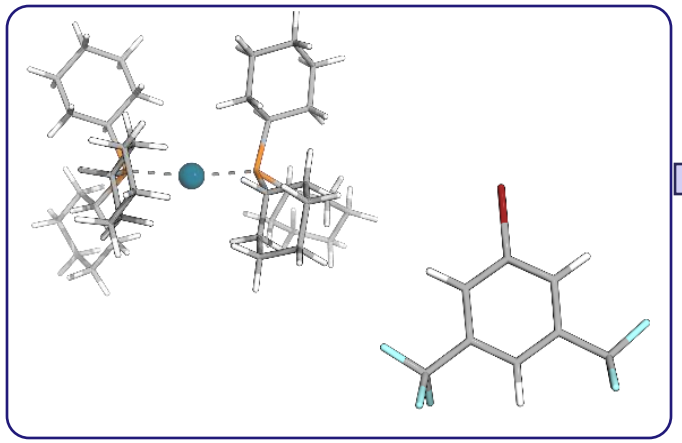
Lu, J.; Donnecke, S.; Paci, I.; Leitch, D. C. *Chem. Sci.* **2022**, 10.1039/D2SC00174H

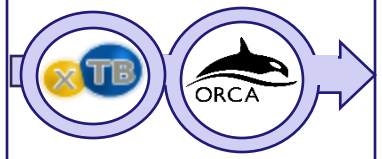
SEAM for oxidative addition



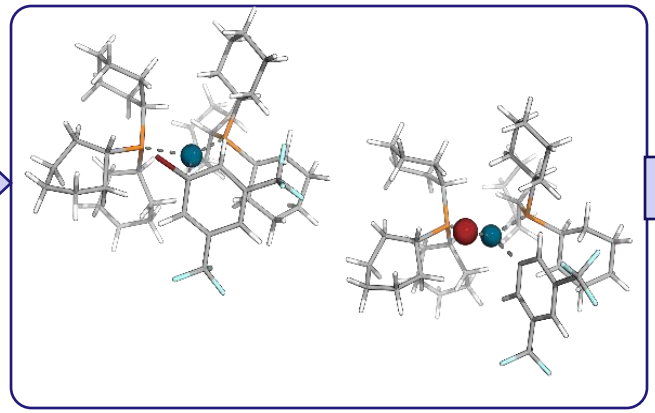
Joël Landis

Workflow to validate against DFT

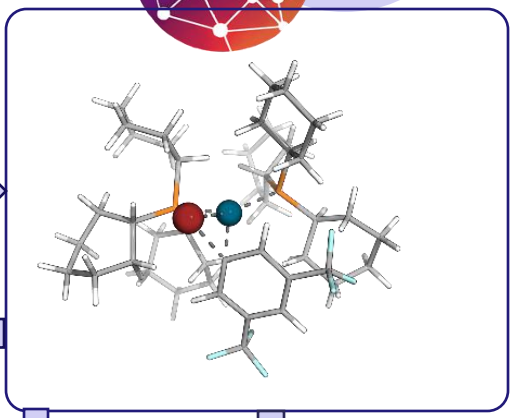




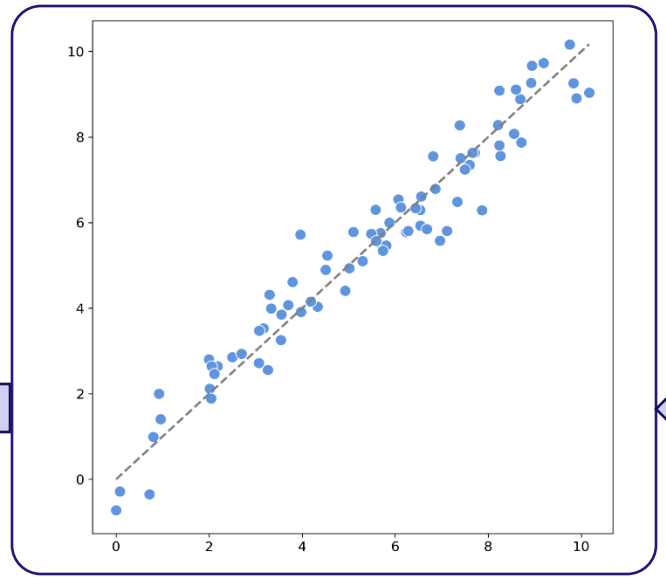
R^2SCAN-3c



POLANYI



DFT TS



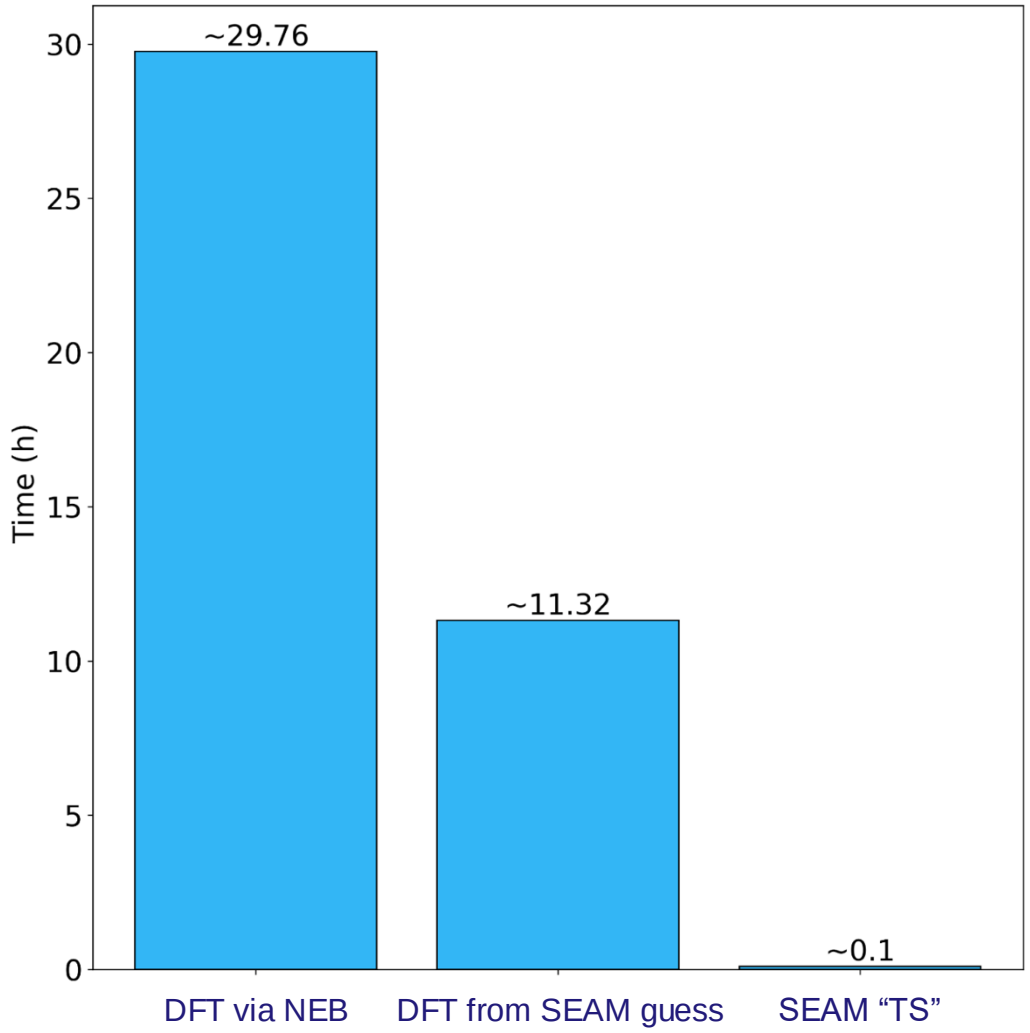
Ridge

TS descriptors

Compare with Leitch

Polanyi data set: 78
DFT data set: 42

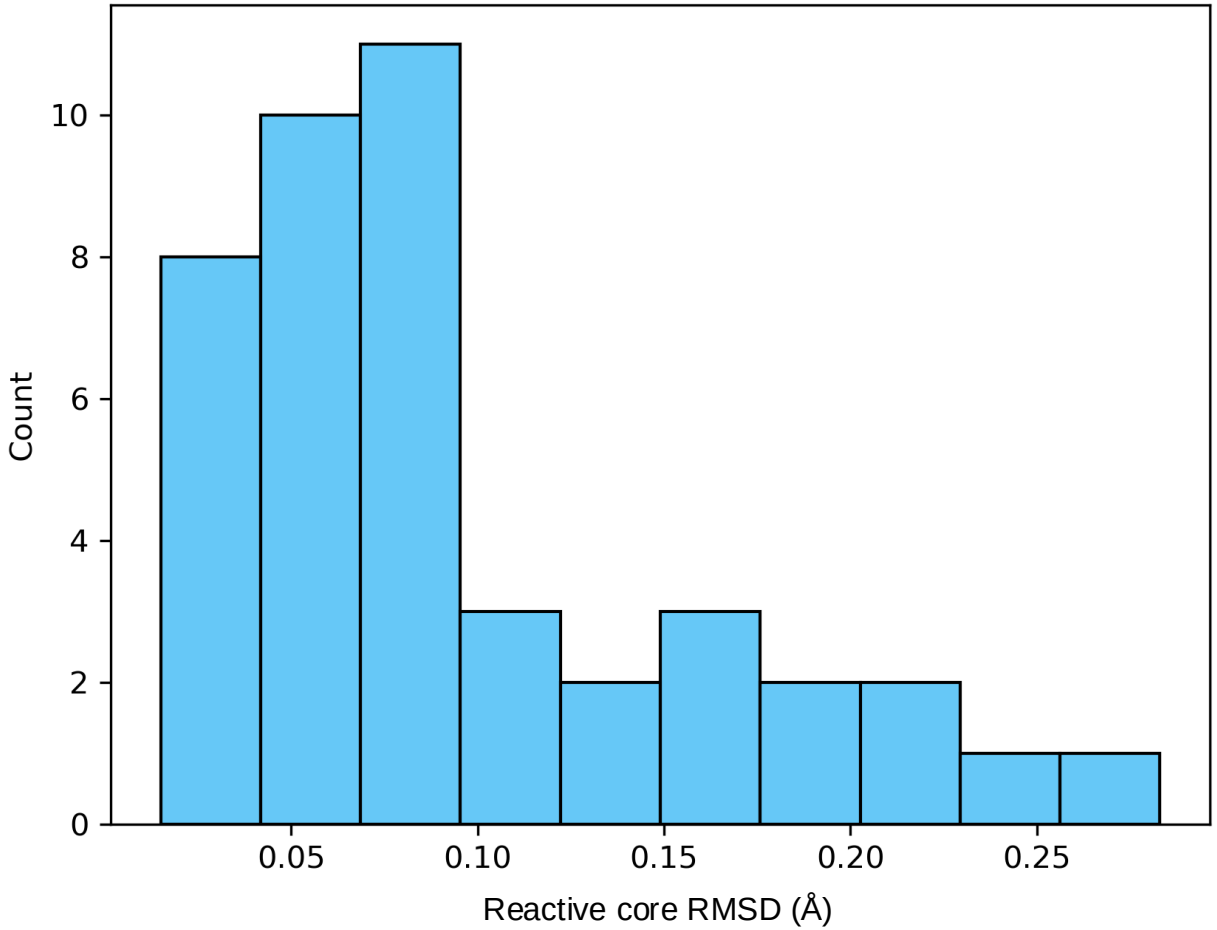
Approximate time savings

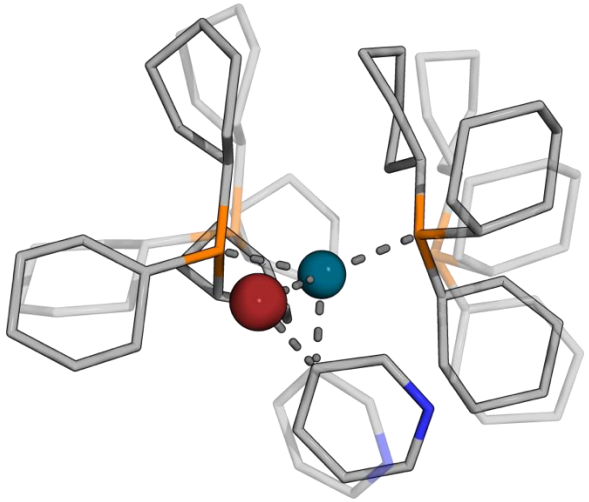


DFT optimizations with

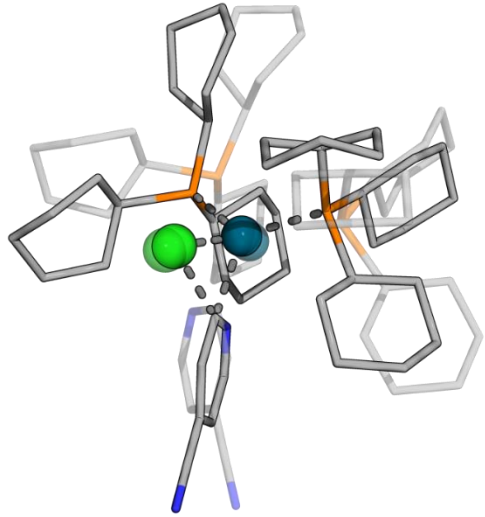
- ORCA 5.0.4
- r^2SCAN-3c
- 24 cores

Comparison with DFT reference





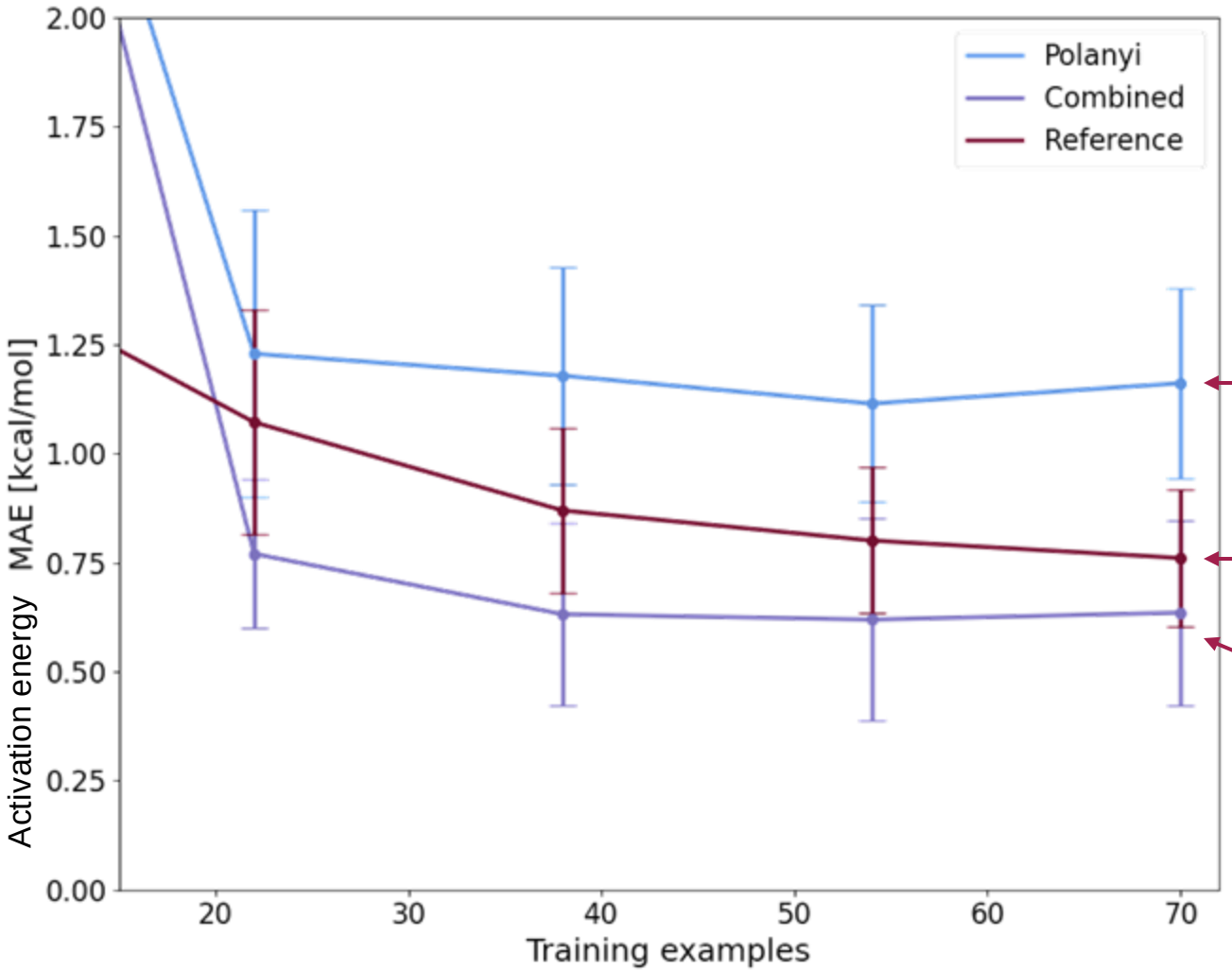
Best case
RMSD: 0.02 Å



Worst case
RMSD: 0.28 Å

r^2SCAN-3c

Learning curves for oxidative addition



SEAM transition state descriptors (only)

Ground state descriptors from Leitch

Combined TS and GS descriptors

→ No statistically significant advantage of TS descriptors
→ Ground state model already very good for this dataset (experimental error at least 0.1–0.3 kcal/mol?)

MORFEUS

Where can I get some descriptors too?

MORFEUS – molecular features for machine learning

- Calculates descriptors for homogeneous catalysis
- Easy to install
- Helpful documentation with examples
- Usable by both non-experts and specialists
- Continuously expanded in collaborations

"I am seeing that the group of Evgeny Pidko is using MORFEUS quite heavily."

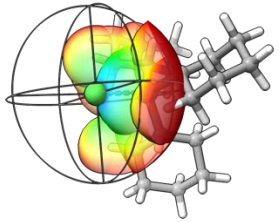
"…the people in Peng Liu's group [...] LOVE MORFEUS"

"We steal stuff from MORFEUS all the time"

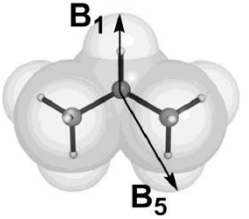
Selected offerings

Steric

Buried volume

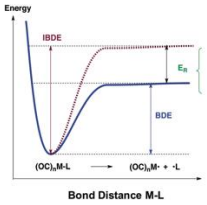


Sterimol

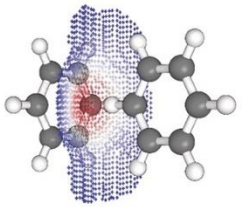


Electronic

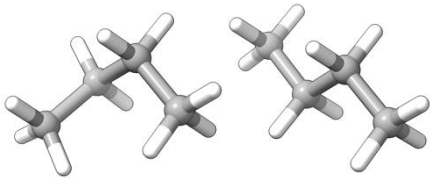
Local force constant



P_{int} dispersion



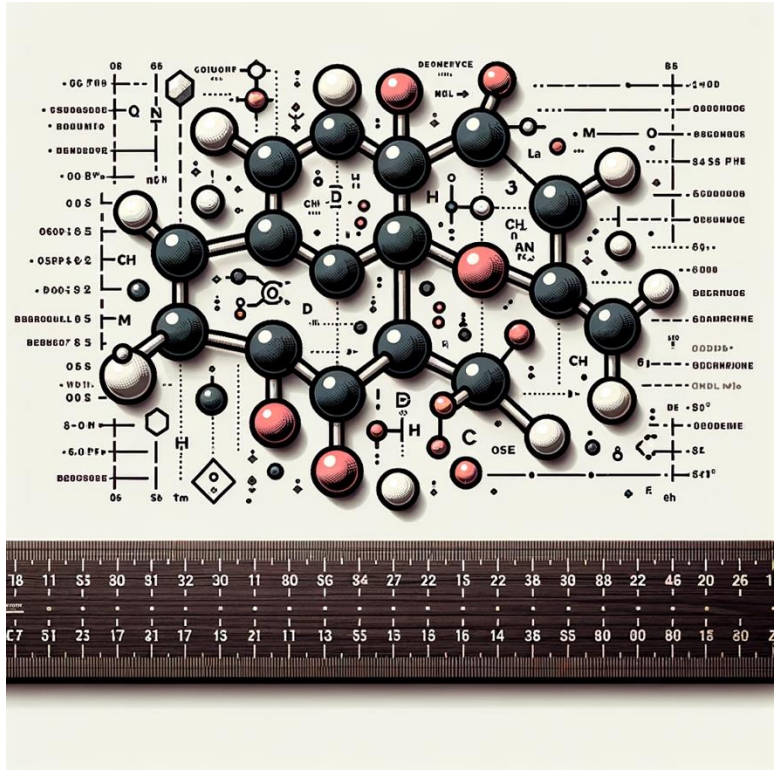
Conformer tools



Boltzmann-averaged descriptors

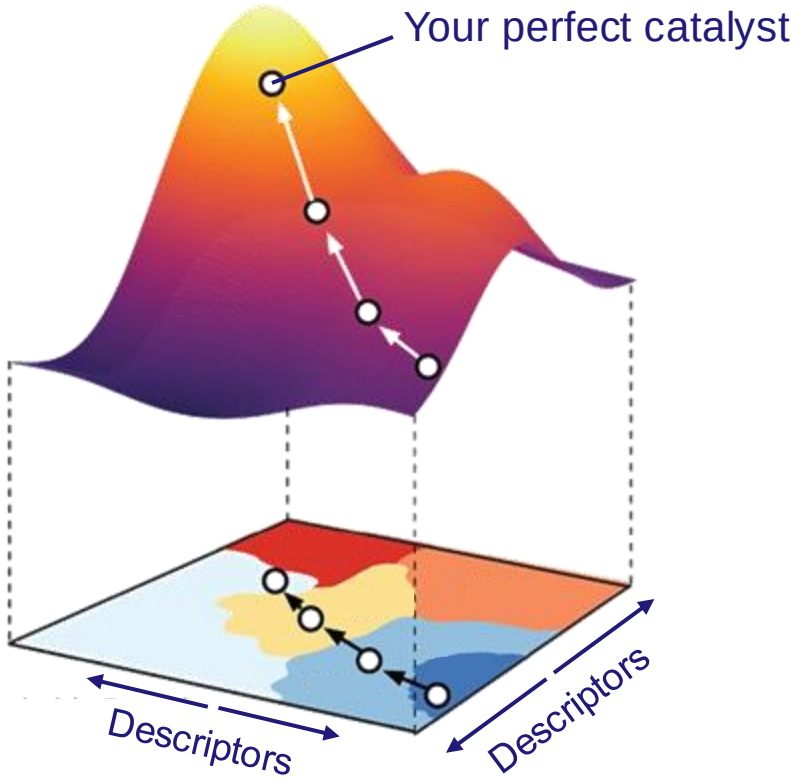
Tailor-made molecules

Descriptors



80	1.5	20	0.1	11
Size	Reactivity	Stability	Hazard	Cost

Chemical space



MORFEUS – a molecular tailor

molecular features for machine learning

Library

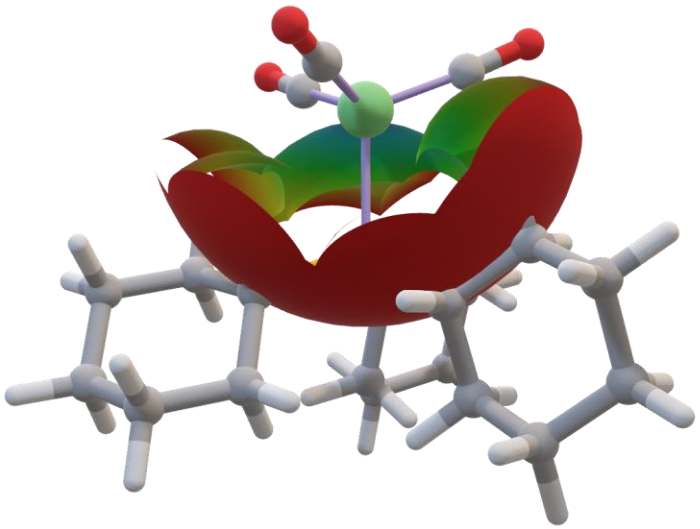
```
>>> from morfeus import ConeAngle, read_xyz
>>> elements, coordinates = read_xyz("PdPMe3.xyz")
>>> cone_angle = ConeAngle(elements, coordinates, 1)
>>> print(cone_angle.cone_angle)
117.11012922937584
```

Command line script

```
$ morfeus cone_angle PdMe3.xyz - 1 - cone_angle
117.11012922937584
```

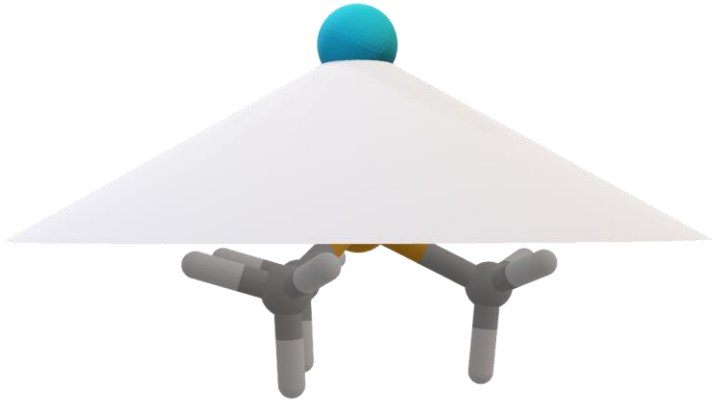
Molecular descriptors for phosphines

Buried volume



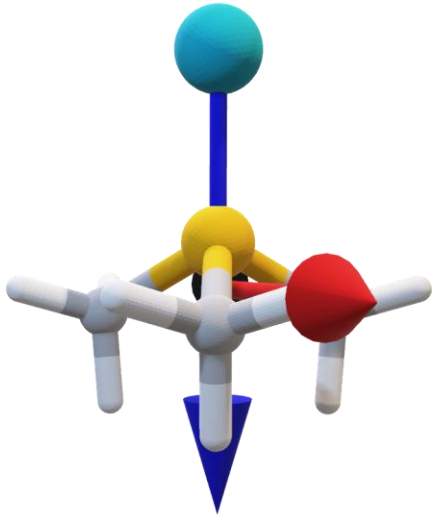
Close to metal atom

Cone angle



As viewed from metal

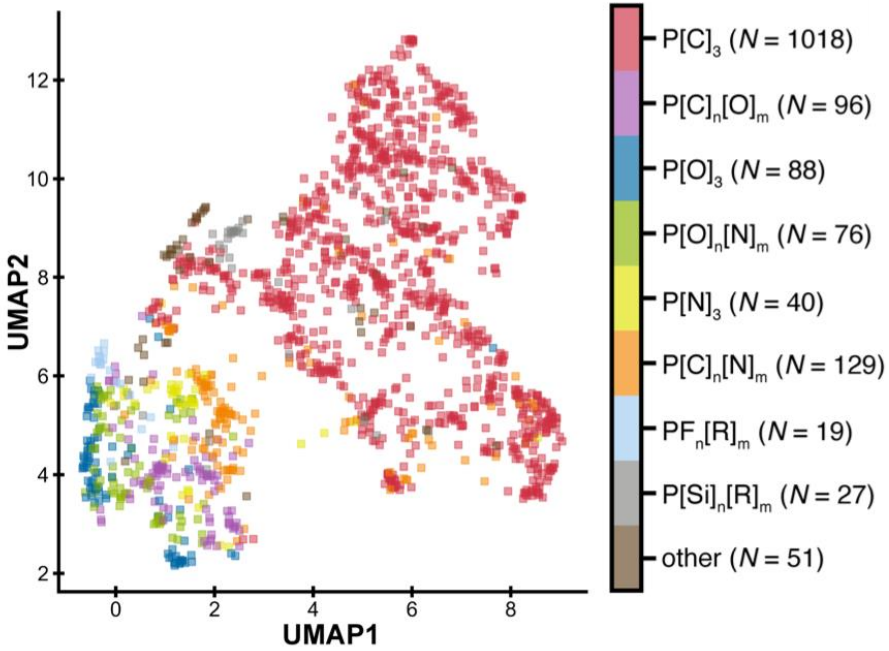
Sterimol



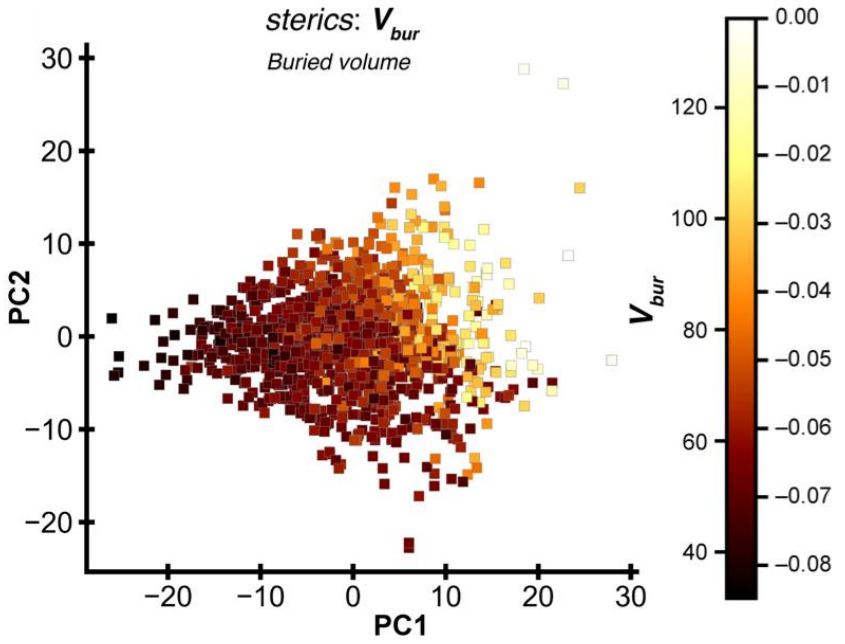
Rotational size and length

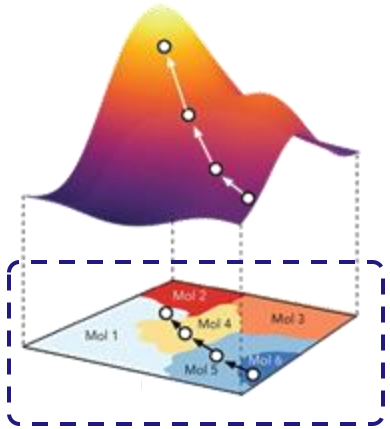
KRAKEN maps of phosphines

UMAP



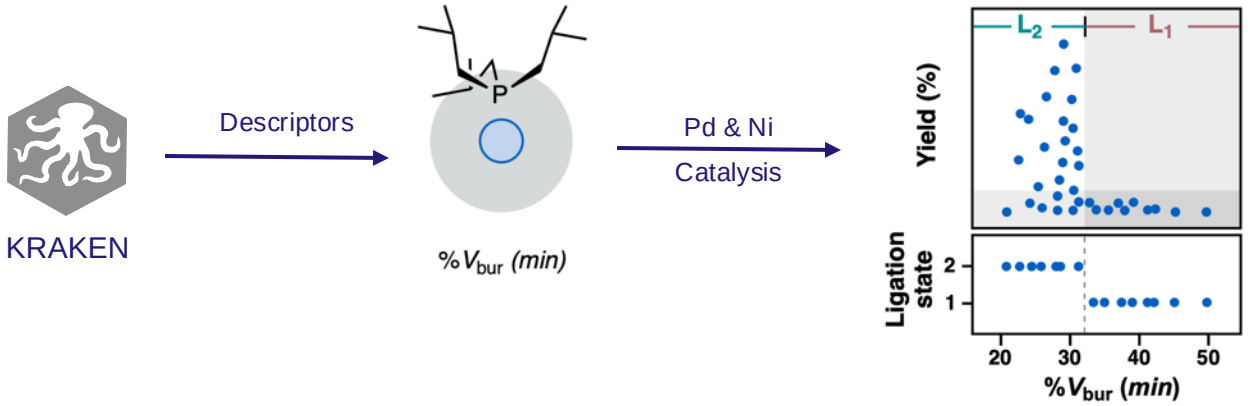
PCA



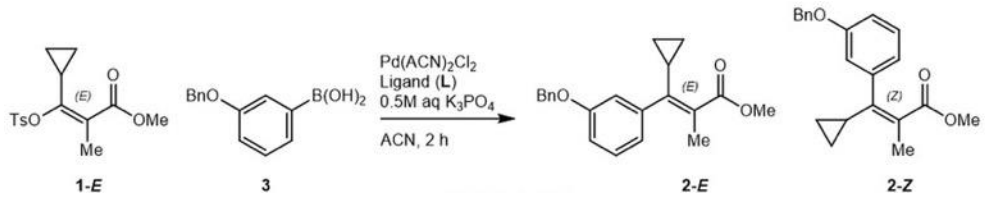


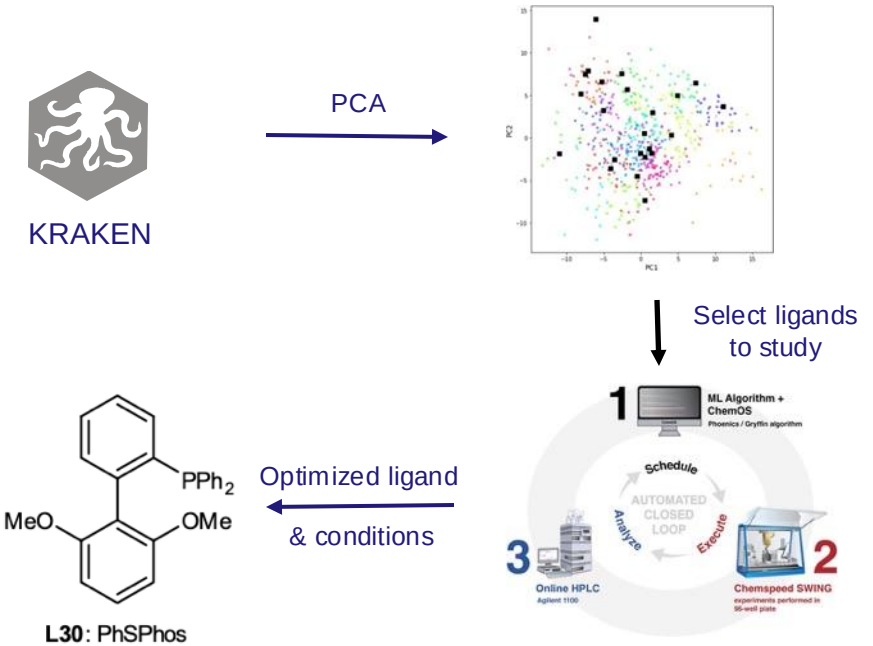
Applications of KRAKEN

Effect of ligation state on Pd & Ni catalysis

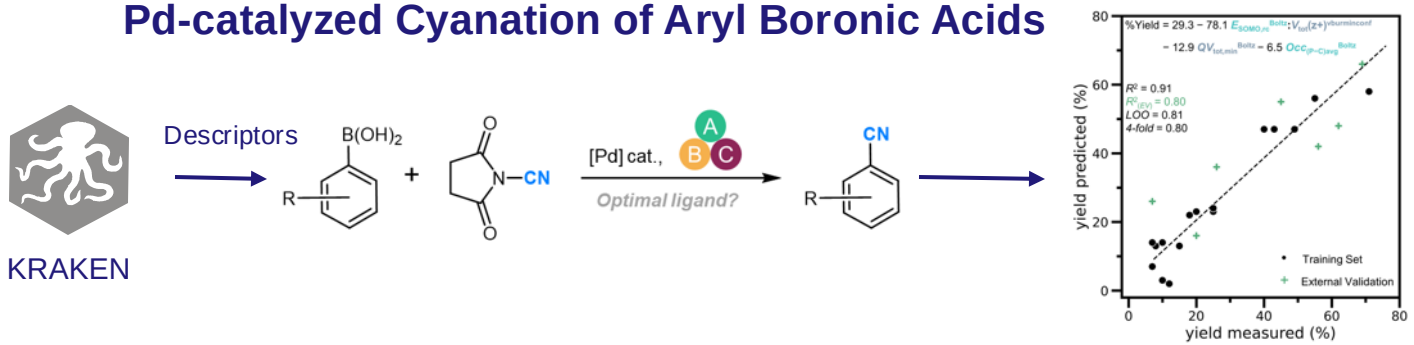


Optimization of stereoselective Suzuki-Miyaura





Pd-catalyzed Cyanation of Aryl Boronic Acids



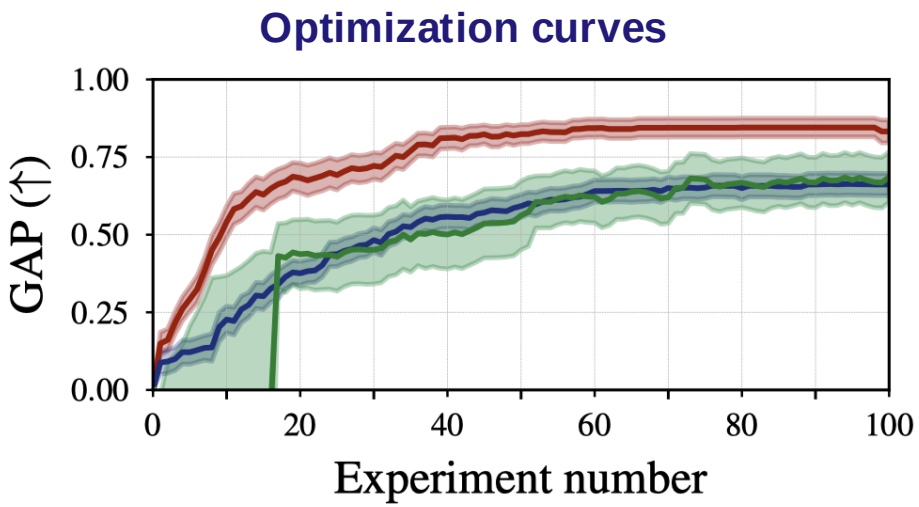
Generality-oriented Bayesian optimization

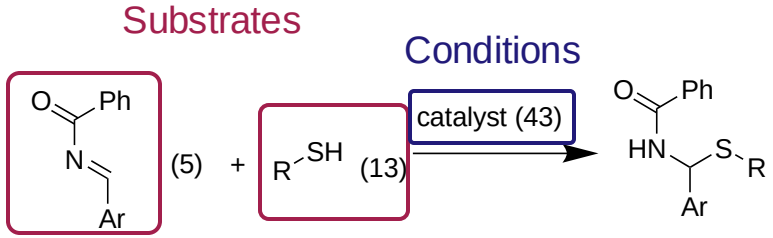
How can we optimize conditions that work well across the board?

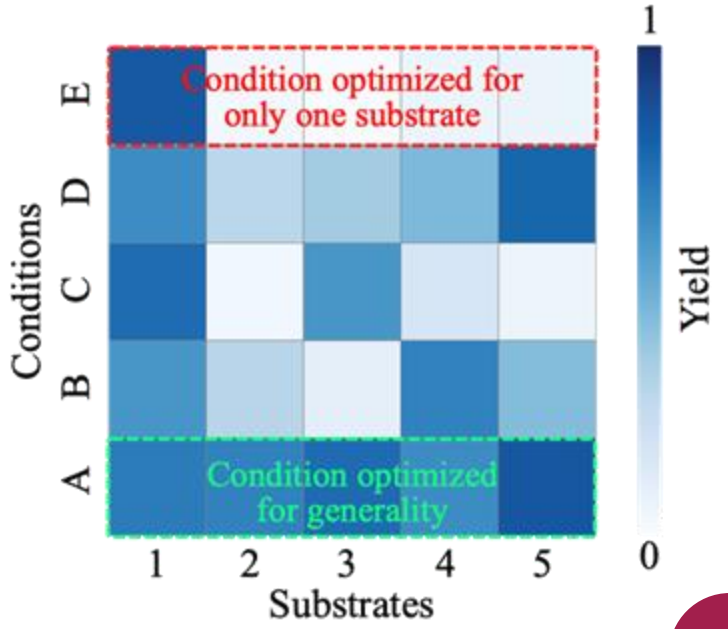
Introduction

- Find general reaction conditions that work well across multiple substrates
- Important for, *e.g.*, automated or diversity-oriented synthesis
- Choose conditions **and** substrates for each experiment
- Each substrate has a conditions → outcome relationship
- We optimize for the best conditions on average

Stefan P. Schmid

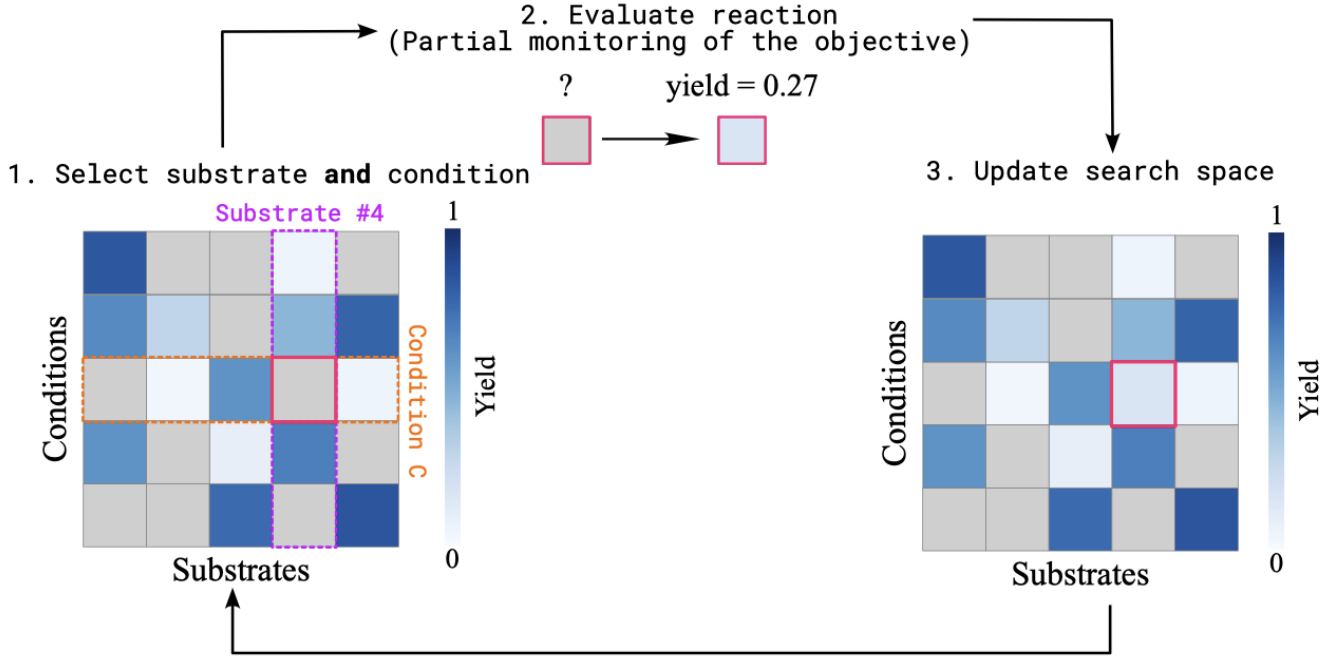






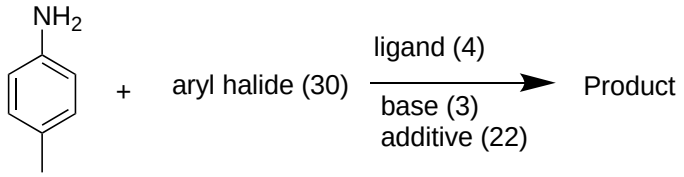
Bayesian optimization loop

1. Select (a) substrate and (b) conditions for experiment
2. Evaluate outcome experimentally
3. Repeat until target reached or budget exhausted

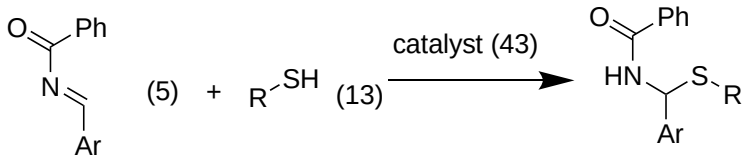


Retrospective validation on literature data

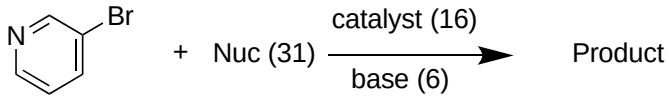
- Buchwald-Hartwig aminations (3955 experiments) → yields



- N,S-acetal formation (1075 experiments) → e.e.



- Pd-catalyzed carbon-heteroatom coupling (1536 experiments) → conversion

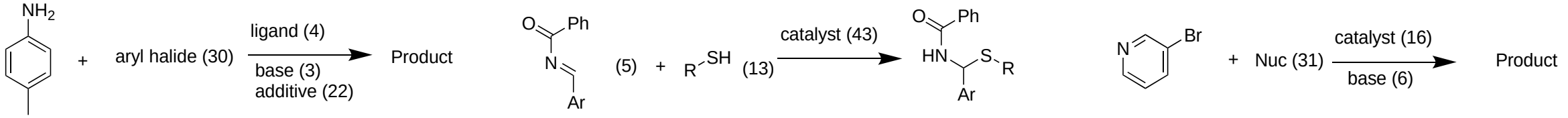


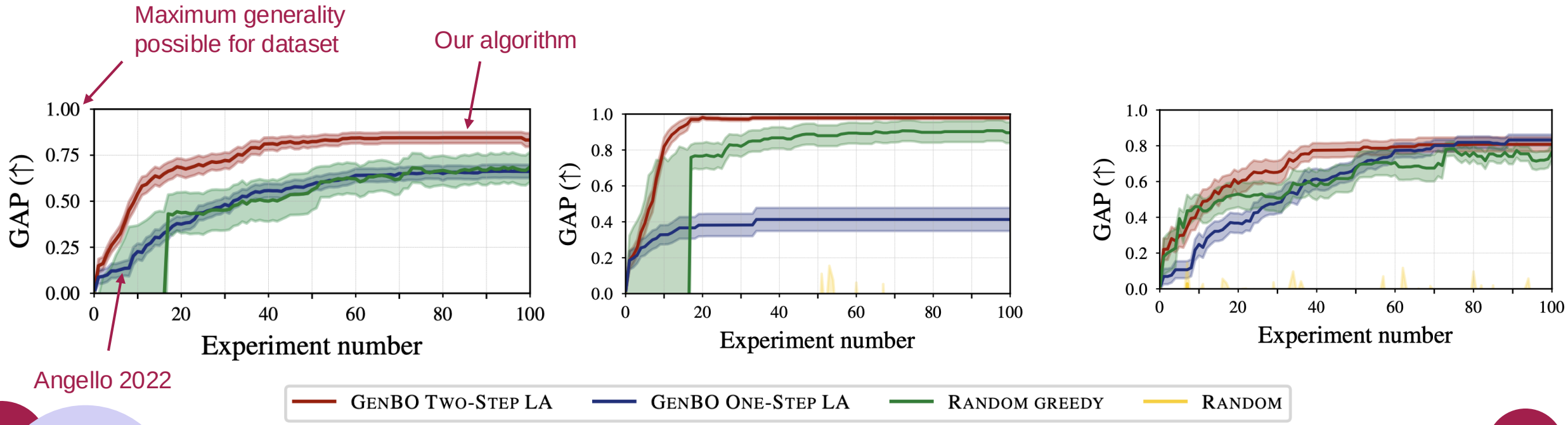
- Data augmentation with low-outcome reactions

(a) Ahneman, D. T. *et al*. *Science* **2018**, 10.1126/science.aar5169 (b) Zahrt, A. F. *et al. Science* **2019**, 10.1126/science.aau5631 (c) Buitrago Santanilla, A. *et al*. *Science* **2015**, 10.1126/science.1259203

Results

- Better than previous algorithms and random baseline

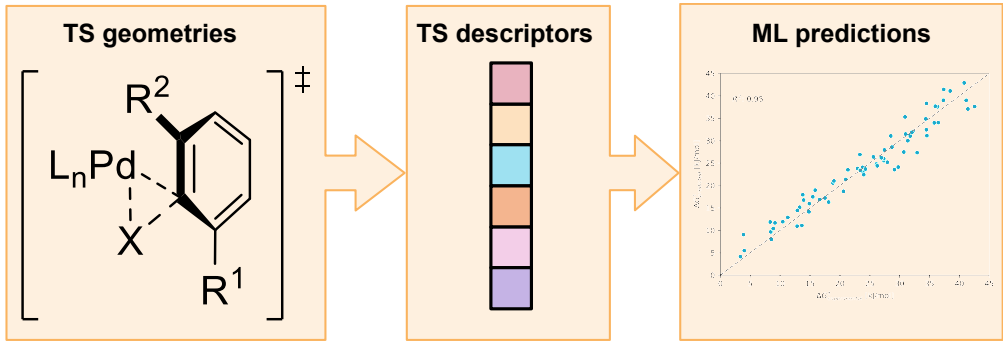


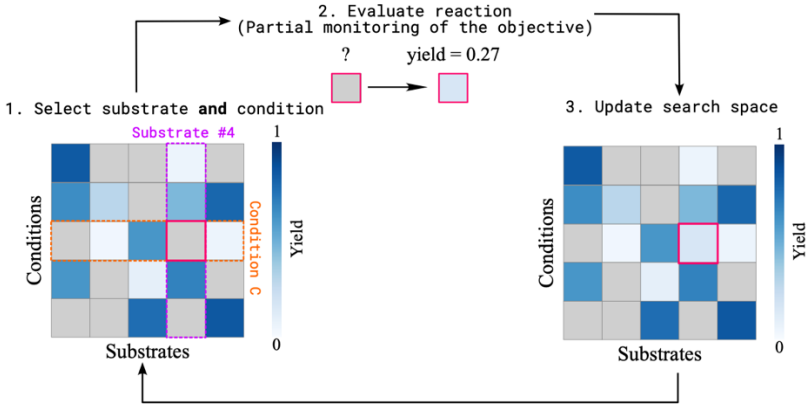


(a) Angello, N. H. *et al. Science* **2022**. 10.1126/science.adc8743 (b) ongoing: Wang, J. Y. *et al. Nature* **2024**, 10.1038/s41586-024-07021-y

Conclusions & outlook

- Reaction prediction with transition state descriptors for low data situations
- Calculate descriptors easily with MORFEUS or get them from repositories like KRAKEN
- Generality-oriented Bayesian optimization for conditions that work across the board → webapp coming





Acknowledgements

AstraZeneca


David Buttar


Per-Ola Norrby

KTH


Tore Brinck

University of Utah


Matt Sigman


Ellyn Peters

TU Berlin


Tobias Gensch

ETH


Joël Landis


Stefan P. Schmid


Lauriane Jacot-Descombes

University of Toronto


Alán Aspuru-Guzik


Cyrille Lavigne


AkshatKumar Nigam


Ella Miray Rajaonson


Théophile Gaudin (also IBM)


Shi Xuan Leong


Cher Tian Ser


Michael Lindner D'Addario


Robert Pollice

NCCR Catalysis

Swiss National Science Foundation

Swedish Research Council

Vector Institute


Agustinus Kristiadi

Uni Wuppertal


Felix Streith-Kalthoff

Havard University


Mohammad Haddadnia

Questions