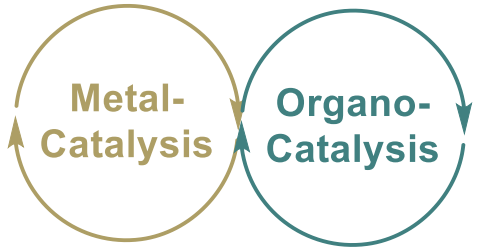
Accelerating optimization in synthetic chemistry through ML approaches

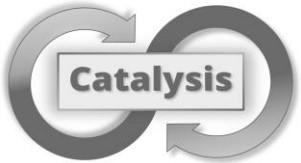
Dr Laurence Grimaud
laurence.grimaud@ens.psl.eu

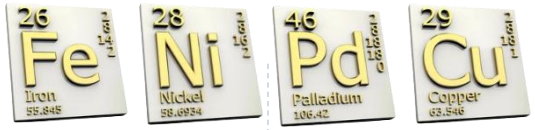
Team's research interests

Dr Maxime Vitale

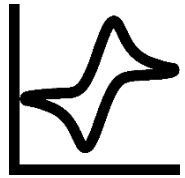
Catalysis





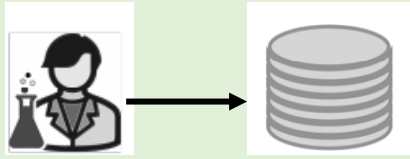


Mechanistic Studies



Isocyanides : Photocatalysis and Electrosynthesis

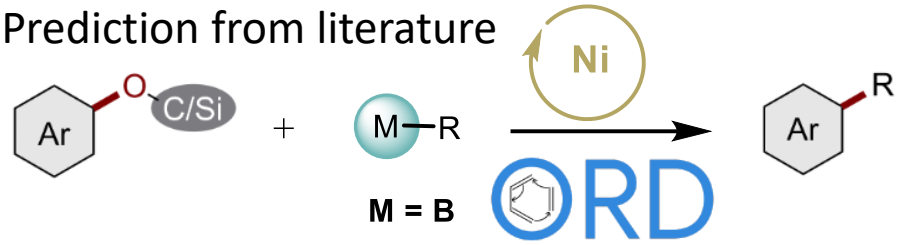
Machine Learning

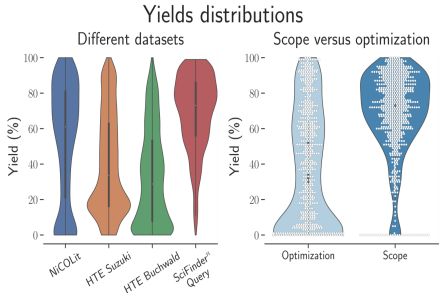


AI in our lab

Pr Rodolphe Vuilleumier (ENS)

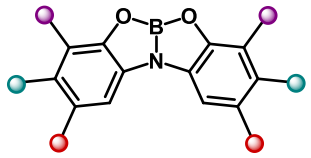
Prediction from literature

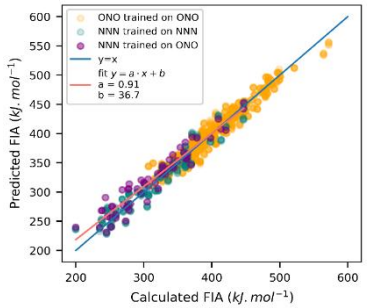


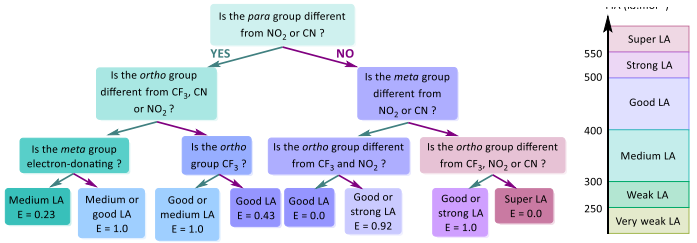


J. Am. Chem. Soc., **2022**, *144*, 14722

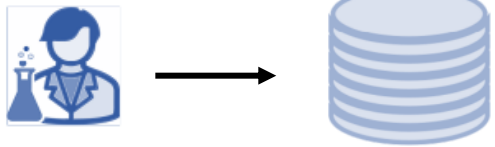
Prediction of Properties



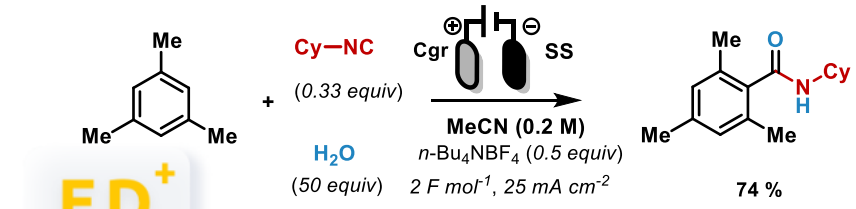


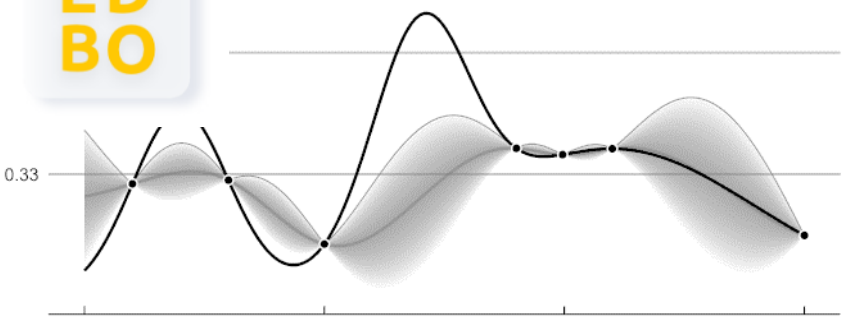


Digital Discovery **2025**, *4*, 3623



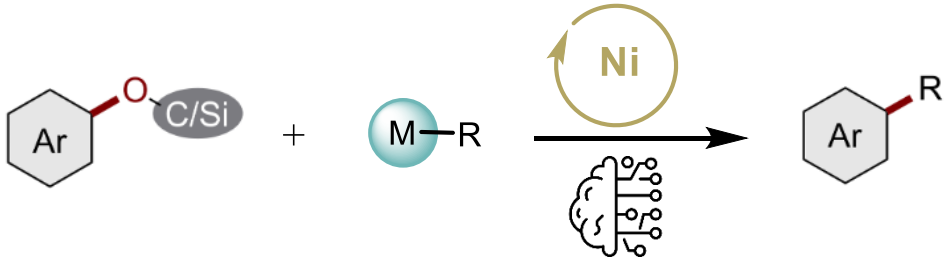
Bayesian Optimisation





C. R. Chim. **2026**, accepted. DOI: 10.5802/crchim.431

Predicting reaction yields from a literature-based reaction dataset

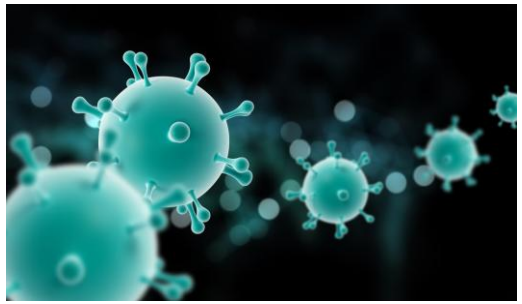


Prediction of the yield:

- ✓ Reduced cost
- ✓ Reduced time
- ✓ Reduced resources

Jules Schleinitz

Maxime Langevin

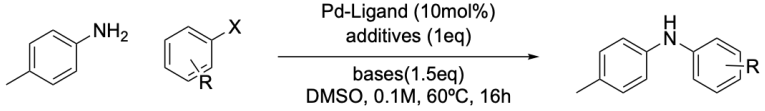


Yanis Smail & Benjamin Wehnert

Chemistry Literature for ML yield prediction ?

How to access quality data ?

HTE

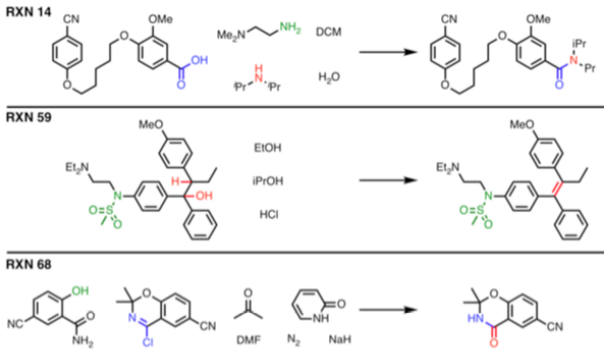


- **Very specific dataset**
  small chemical space
- **Low report bias**
- **Good ML performances**

D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson, N. W. Sach, *Science* 359, 429 (**2018**).

D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* 360, 186 (**2018**)

USPTO



- **Very broad dataset**
  Huge chemical space
- **High report bias towards high yields**
- **Poor ML performances**

*D. M. Lowe, Ph.D. thesis, University of Cambridge (**2012**)*

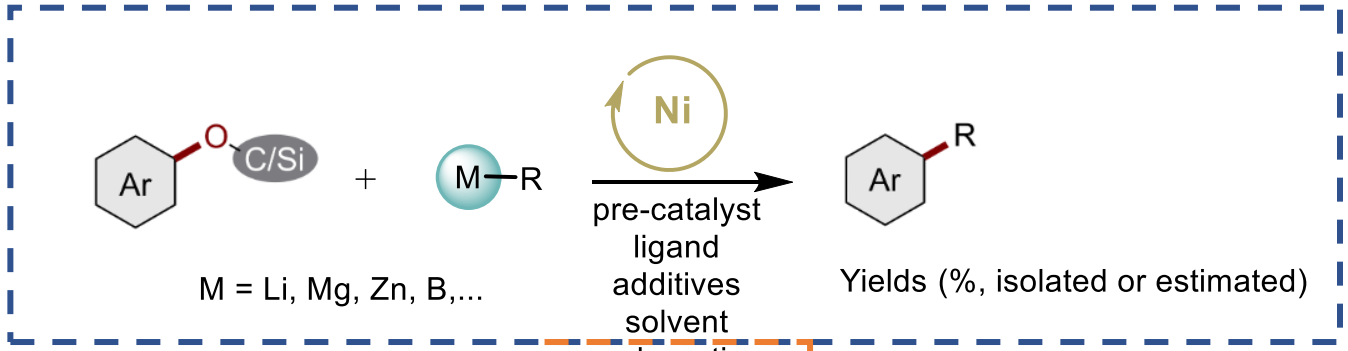
Literature data

- **Very broad dataset**
  Huge chemical space
- **High report bias**

?

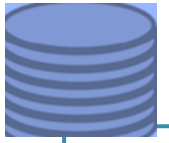
Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. *J. Am. Chem. Soc.* **2022**, 10.1021/jacs.1c12005
Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F *Angew. Chem. Int. Ed.* **2022**, 10.1002/anie.202204647

Manually extracted and curated database



From main text and SI

2000 reactions
45 publications



ORD

NiCOLit Database

Preprocessing (chemical rules + relevance criteria)

1400 reactions

Featurization

RDKit + OHE

RXNFP

DFT-based

Auto-QChem

*

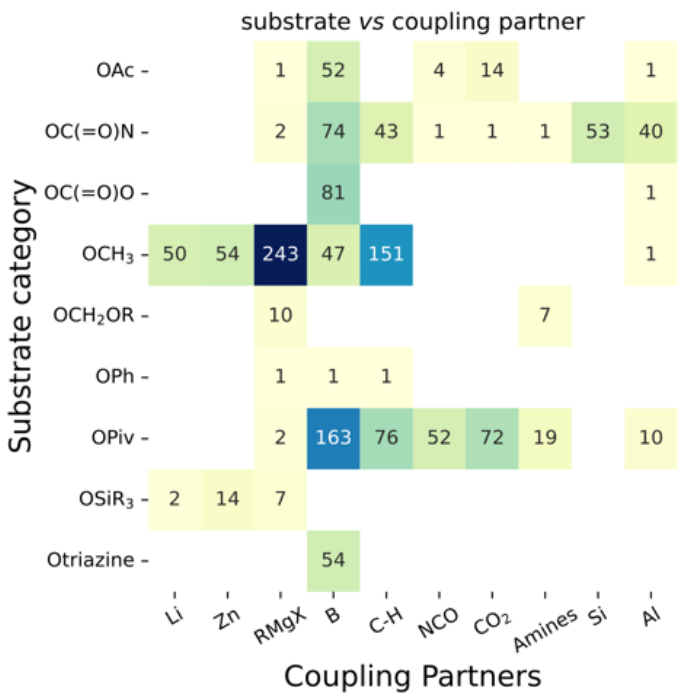
Random Forest
Regression Model

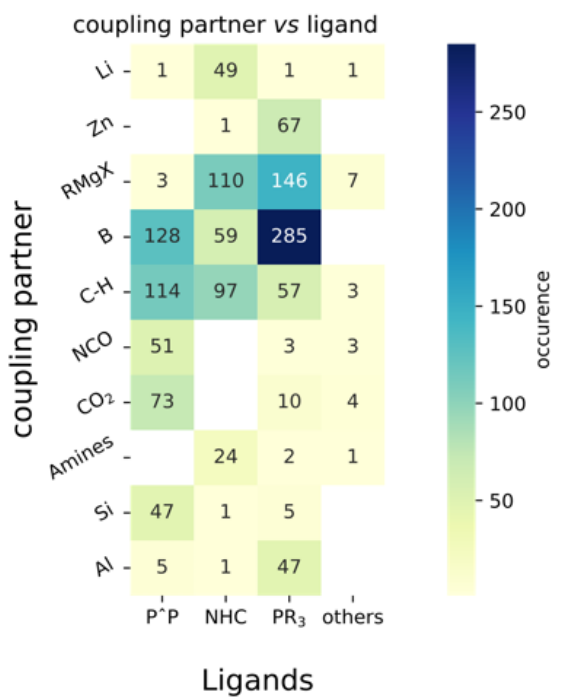
Dataset analysis
Predictive models

Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. *J. Am. Chem. Soc.* **2021**, *143*, 18820.

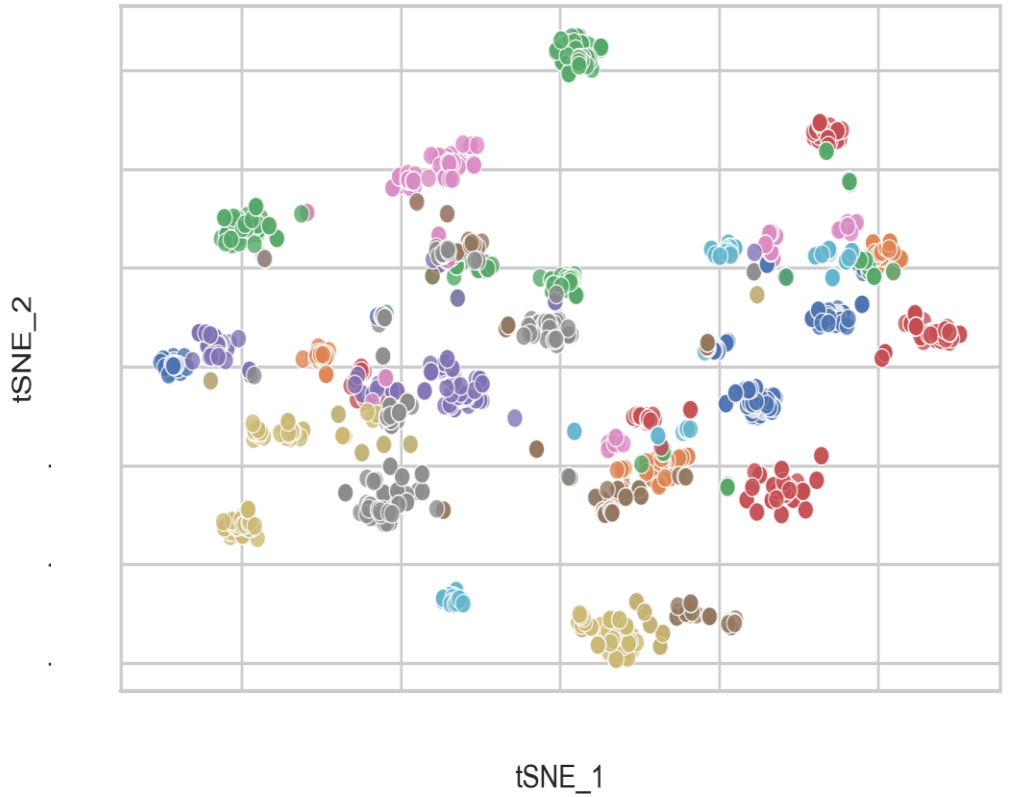
P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, Machine Learning: Science and Technology **2021**, 2, 015016

Specific combinations and minimal overlap between publications

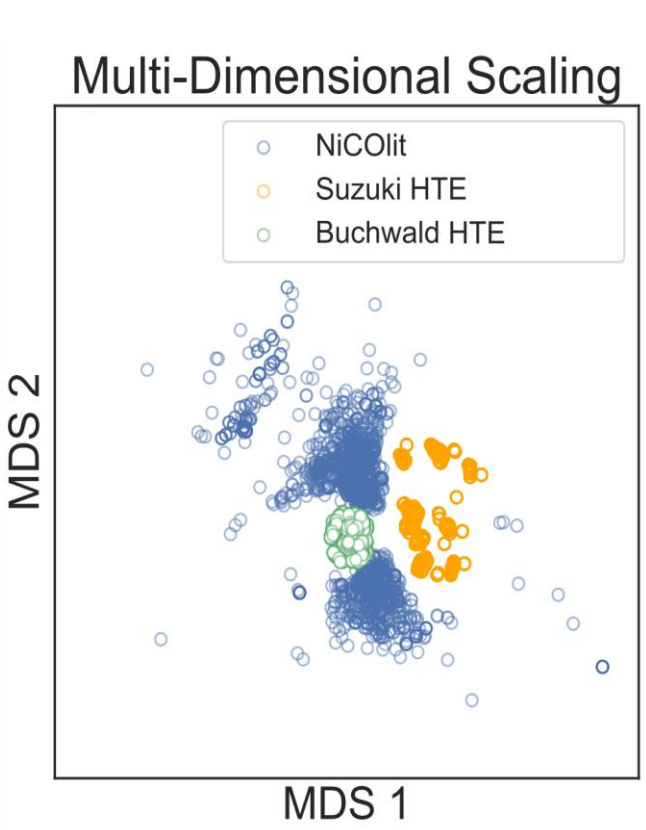




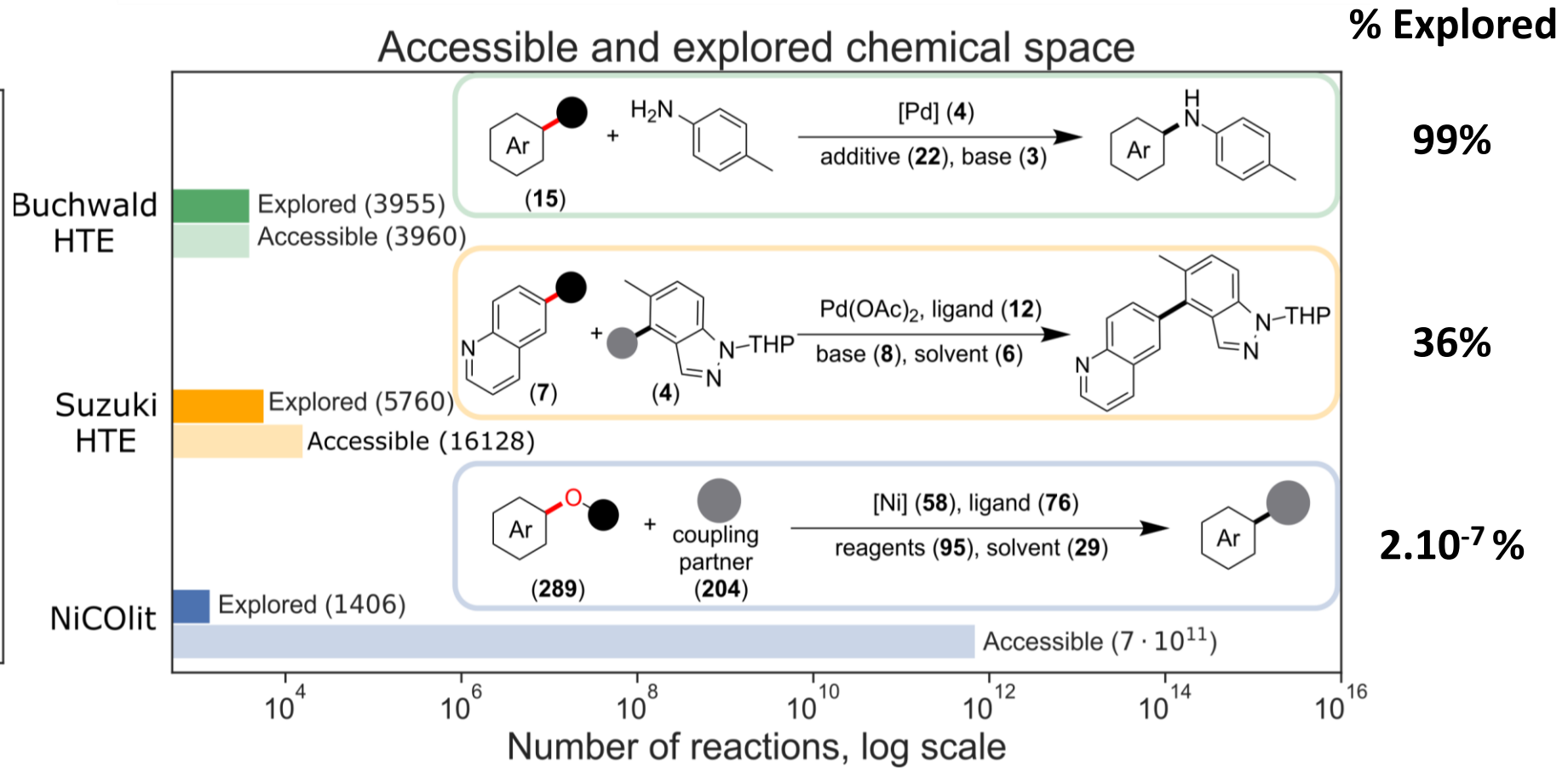
t-SNE NiCOlit coloured by publication



Comparison with HTE



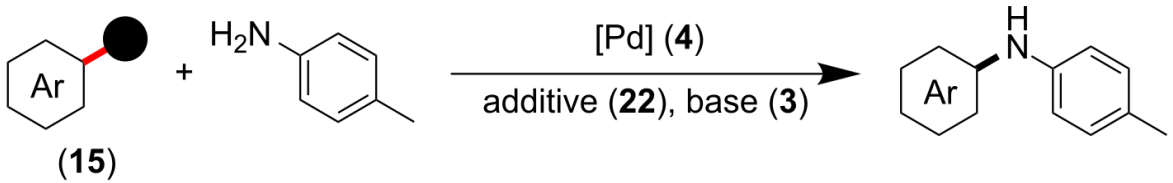
Analysis with RXNFP featurization

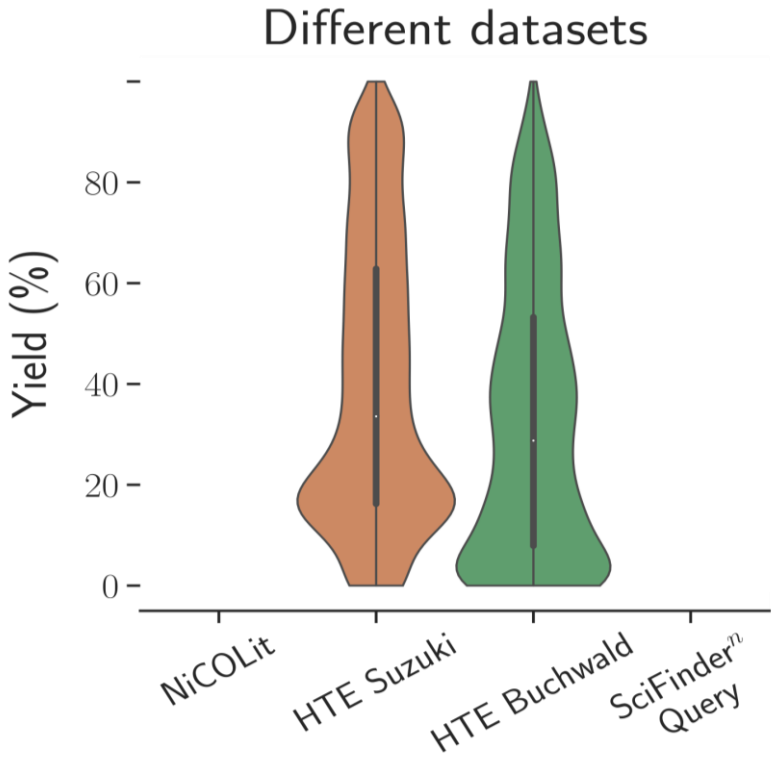


Ahneman, D.T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186

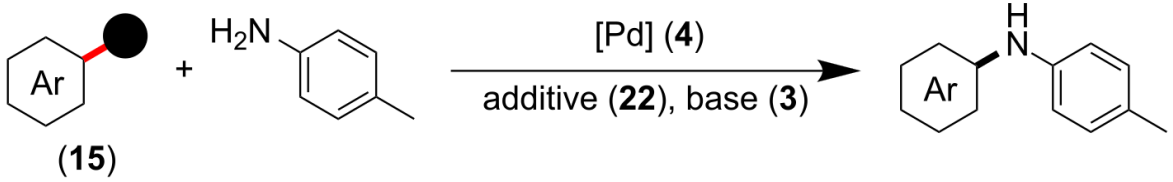
Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.;Farrell, W.; Richardson, P.; Sach, N. W. *Science* **2018**, *359*, 429.

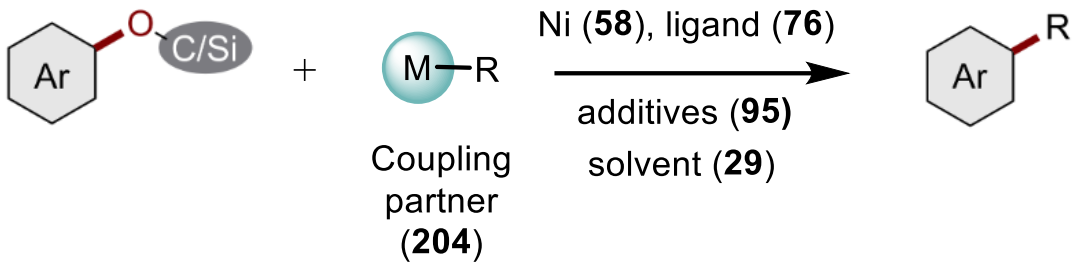
Yields distribution

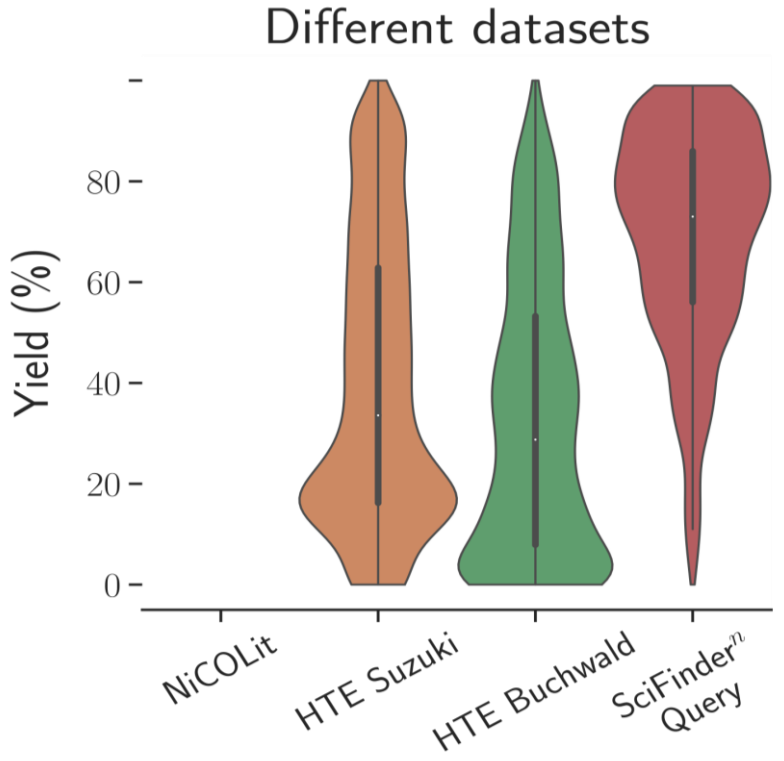




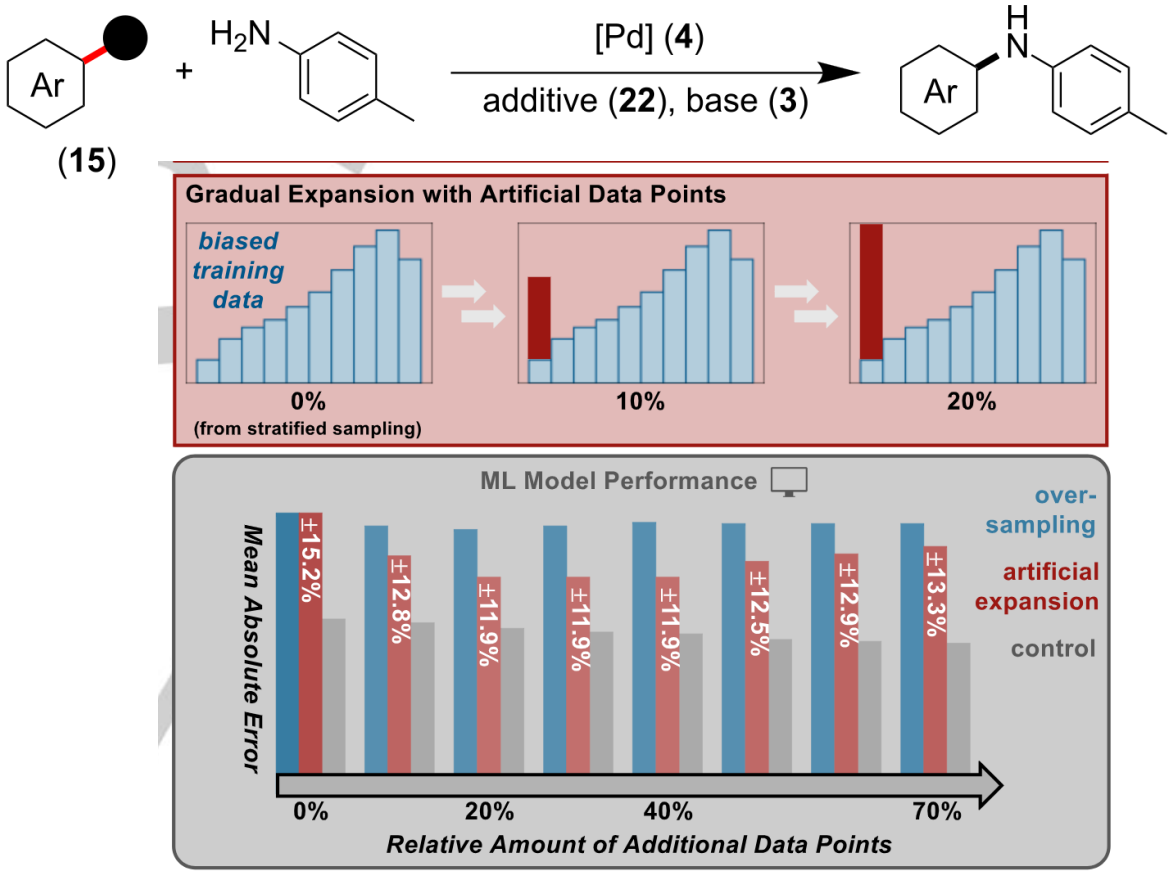
Yields distribution

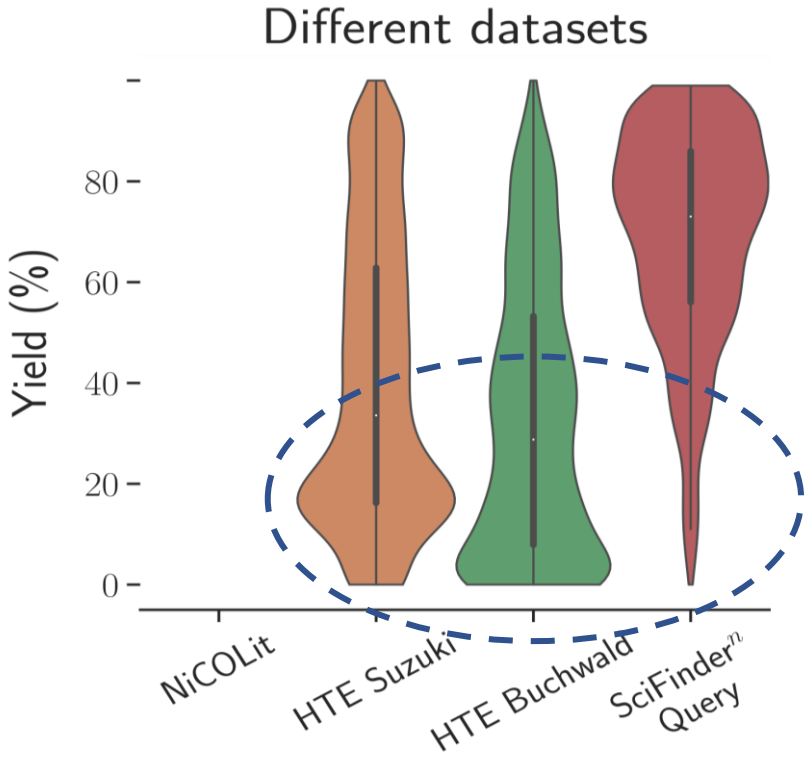






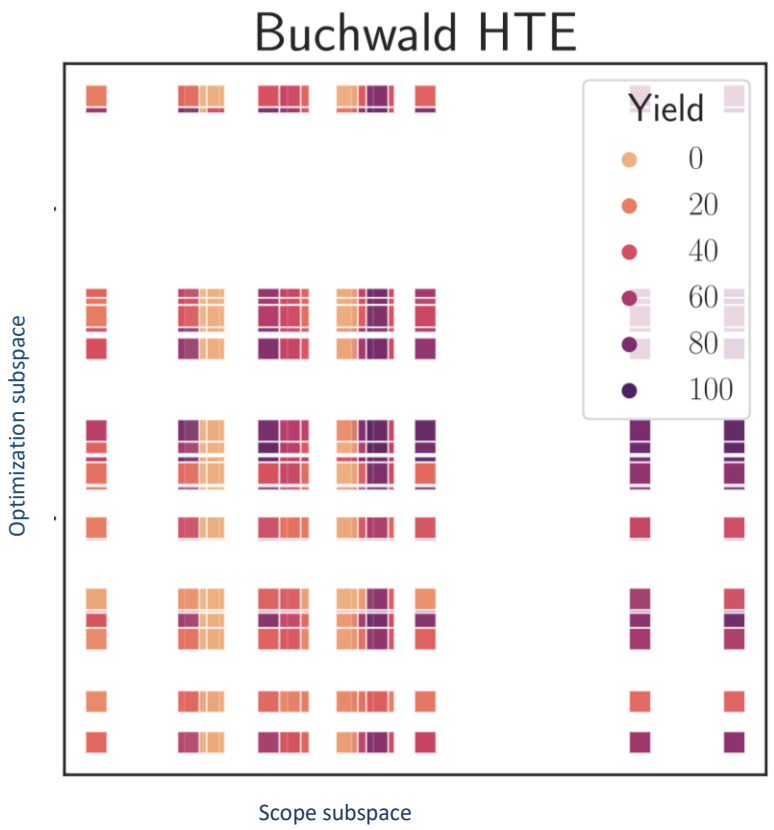
Yields distribution

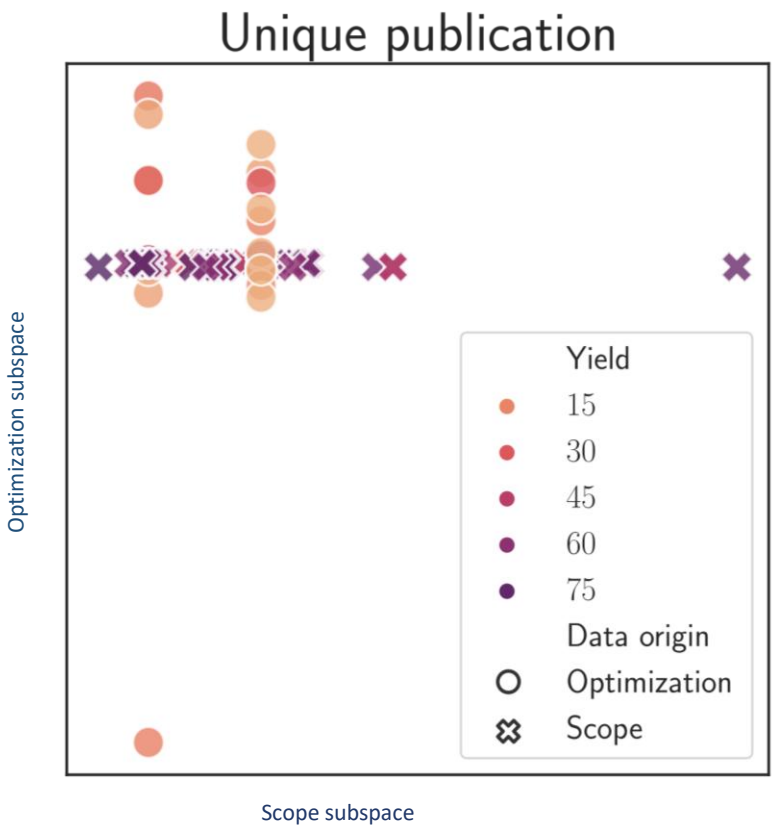




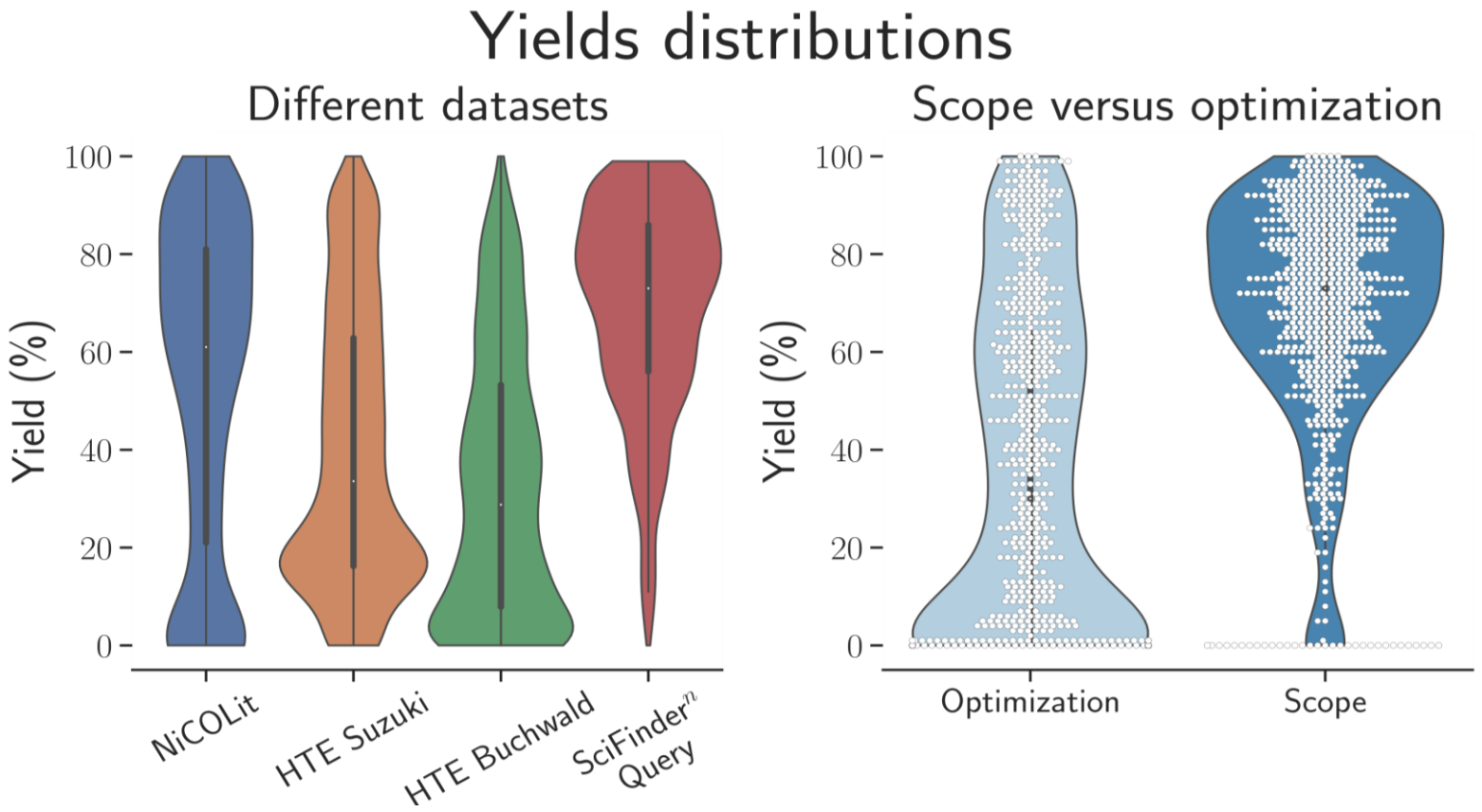
Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. *Angew. Chem., Int. Ed.* **2022**, *61*, e20220467

Comparison with HTE

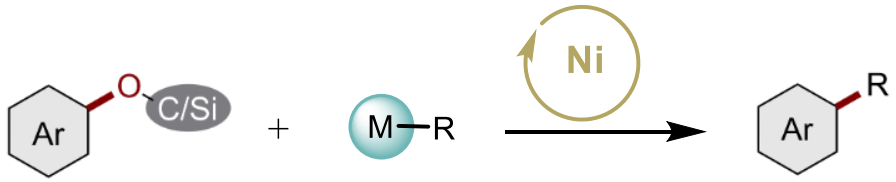




Scope and Optimization data



Performances of the model on a random split



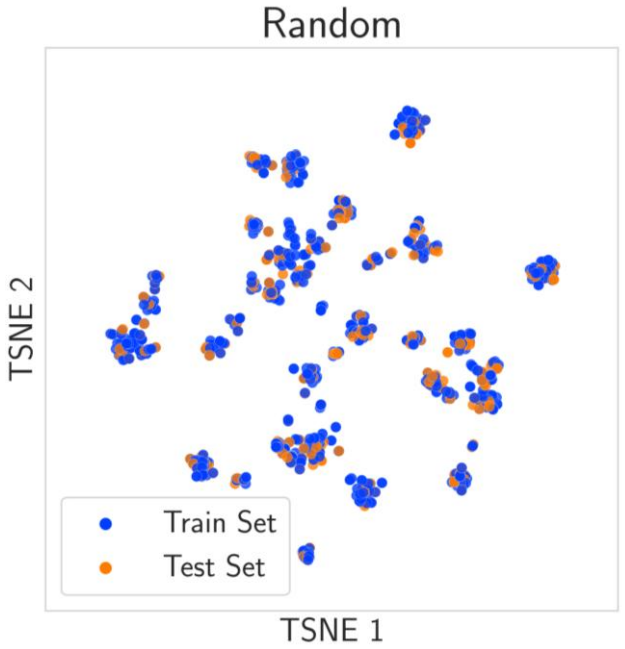
M = Li, Mg, Zn, B,...

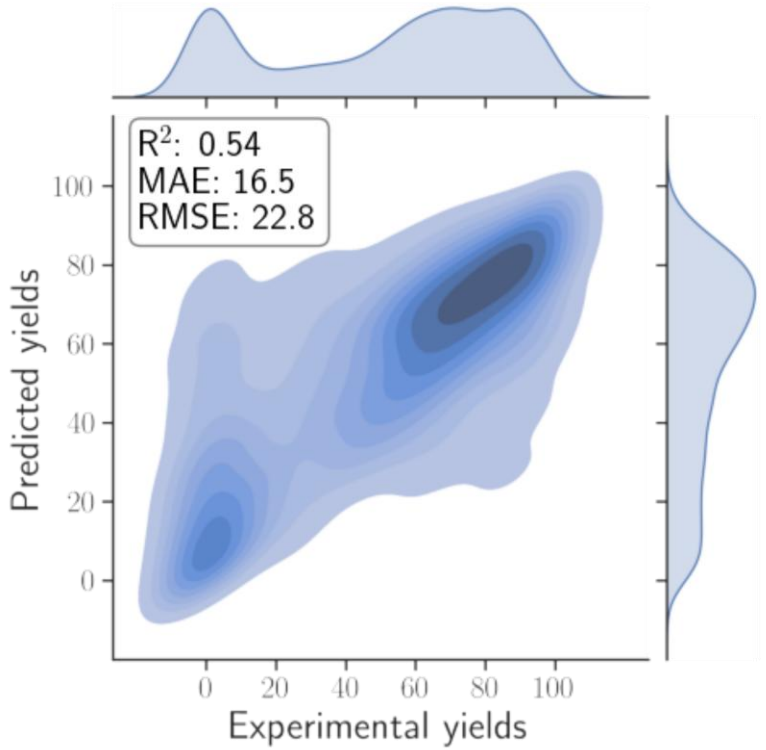
DFT featurization

Comparison to literature
HTE R^2 = 0.8
USPTO R^2 < 0.2

Training set ≈80%

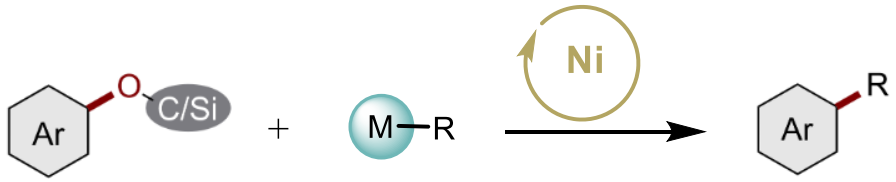
Test set ≈20%





RDKit FP and DRFP R^2 0.50
RXNFP R^2 0.37

The importance of failed experiments

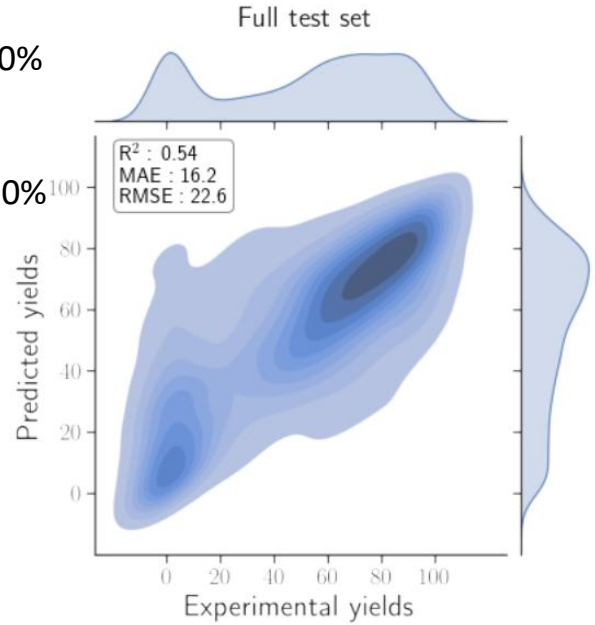


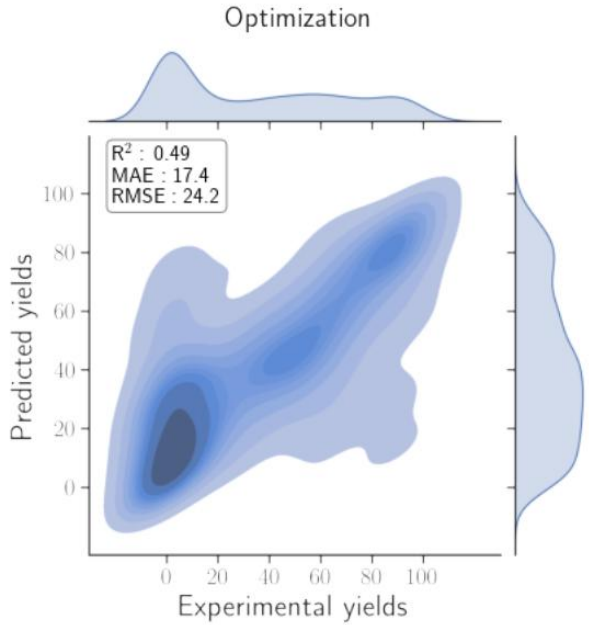
M = Li, Mg, Zn, B,...

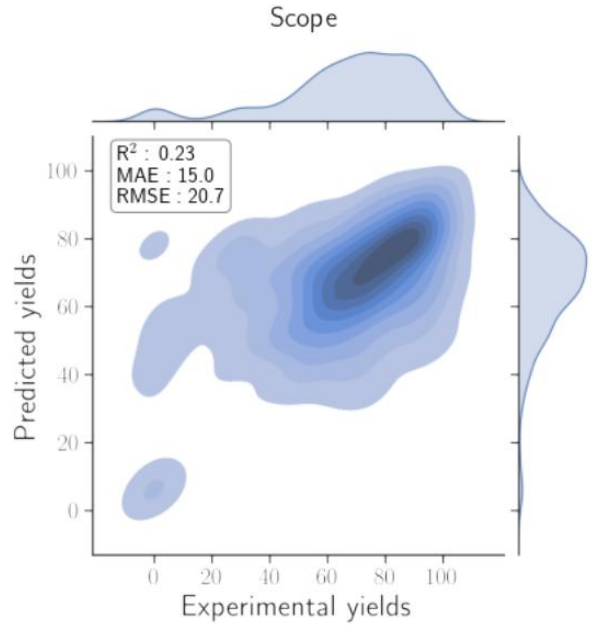
DFT featurization

Training set ≈80%

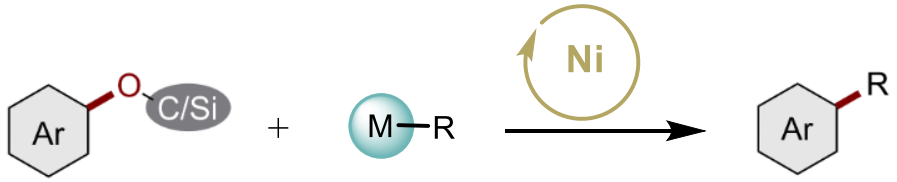
Test set ≈20%







Performances for a out-of-sample publication

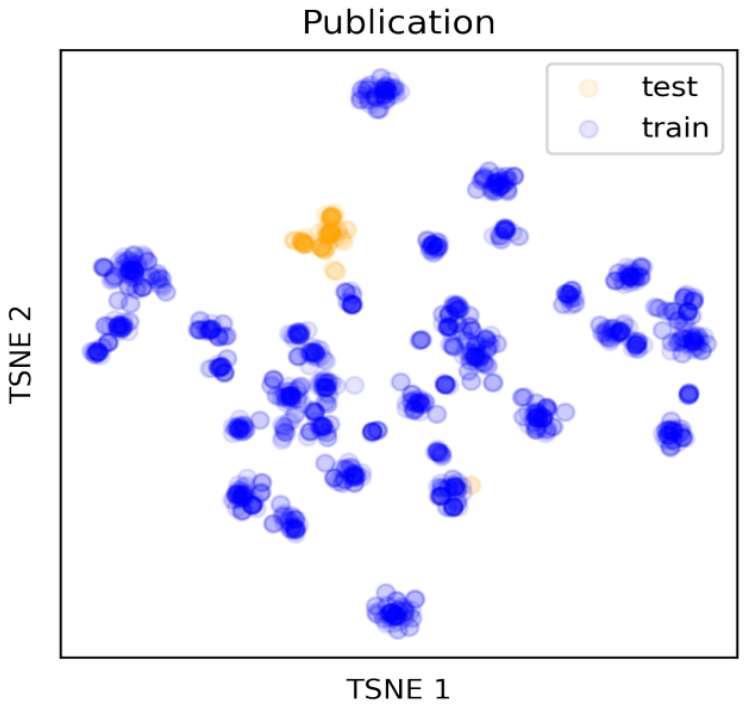


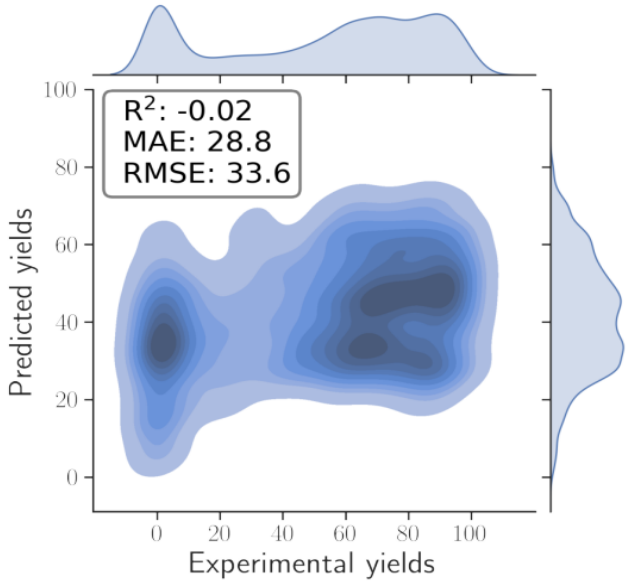
M = Li, Mg, Zn, B,...

DFT featurization

Training set ≈80%

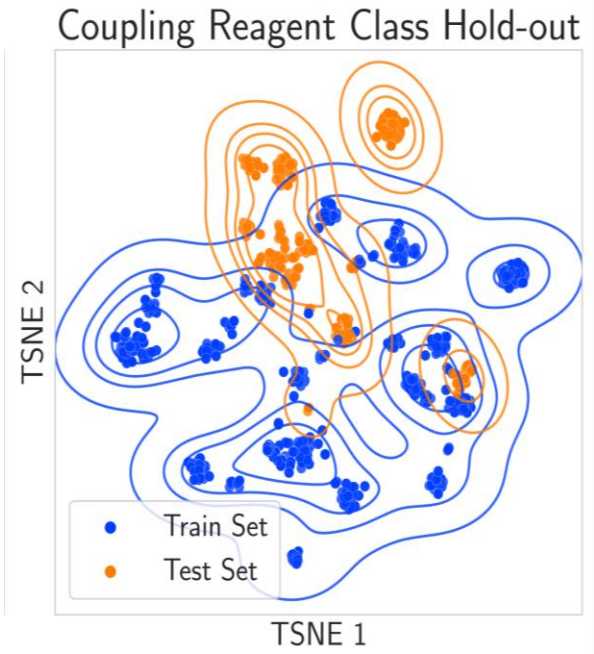
Test set ≈20%

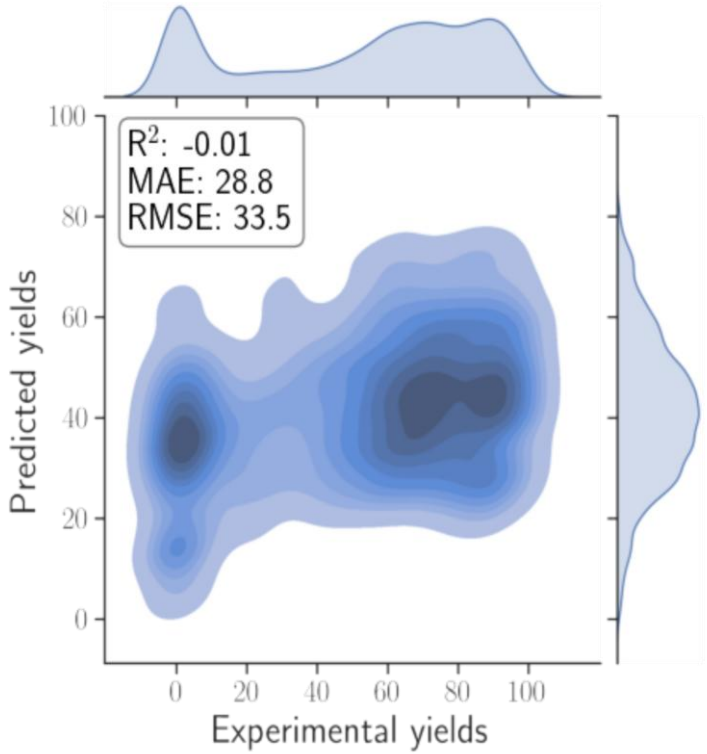




Performances for a out-of-sample coupling partner

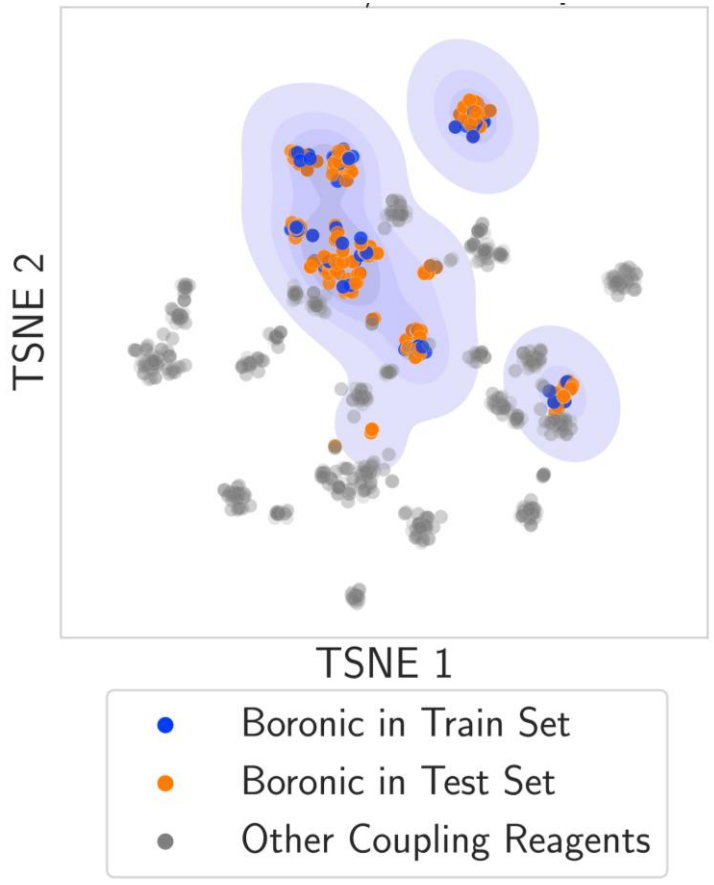
DFT featurization





Performances on a restricted dataset with similar coupling partners

DFT featurization



Coupling partner	R^2	Number of reactions	Number of publications
B	**0.47**	472	11
C-H	**0.59**	271	3
RMgX	**0.51**	266	5
CO_2	**0.52**	87	1
Zn	0.54	68	2
NCO	**0.39**	57	1
Al	0.20	53	1
Si	**0.64**	53	1

Similar to the prediction on the whole NiCOlit database

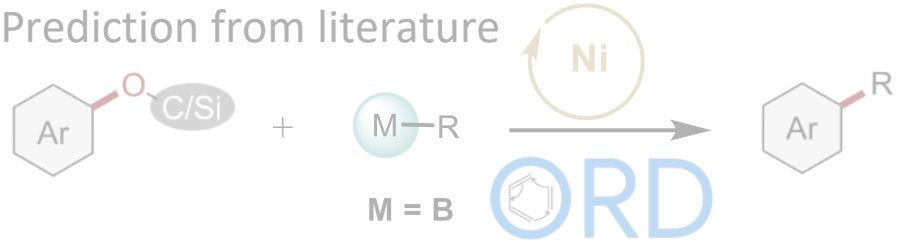
on Buchwald HTE data: R^2 = 0.59 for 98 data points

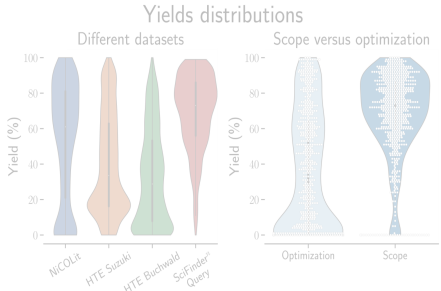
J. Am. Chem. Soc., **2022**, *144*, 14722

AI in our lab

Pr Rodolphe Vuilleumier (ENS)

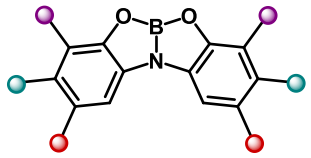
Prediction from literature

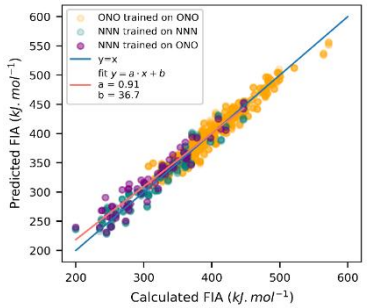


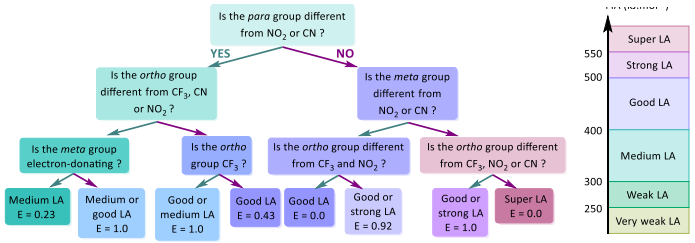


J. Am. Chem. Soc., **2022**, *144*, 14722

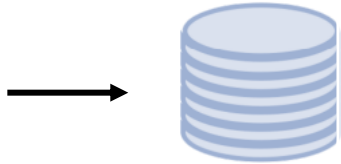
Prediction of Properties



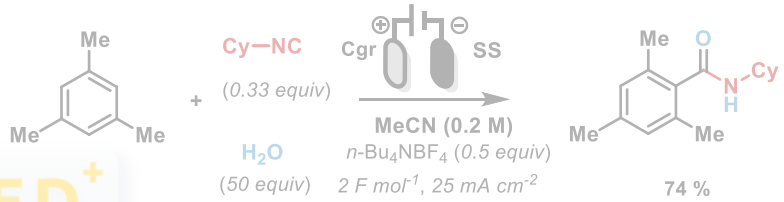


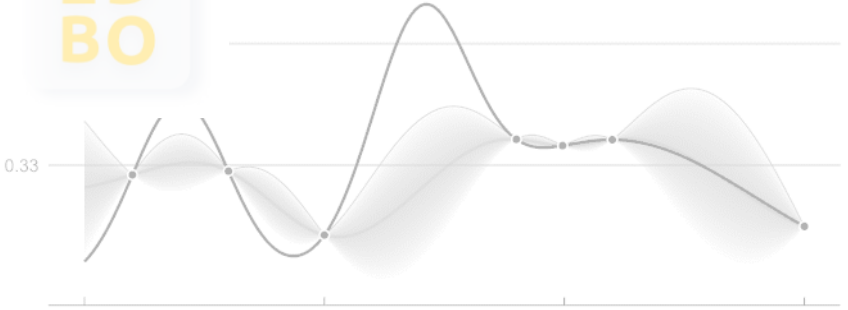


Digital Discovery **2025**, *4*, 3623



Bayesian Optimisation



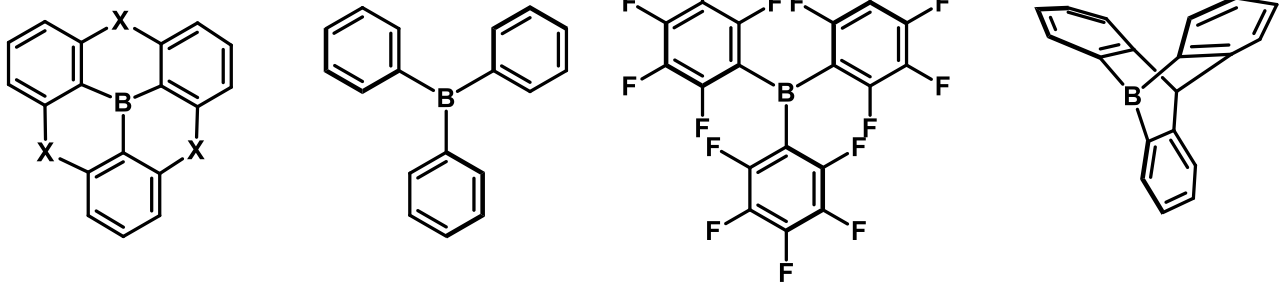


C. R. Chim. **2026**, accepted. DOI: 10.5802/crchim.431

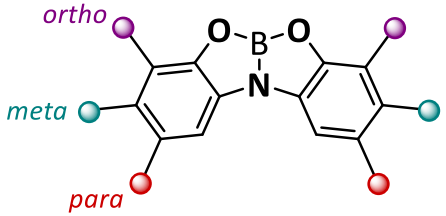
Predicting properties in low data regime: Lewis acidity

Juliette Fenogli

Boron-based Lewis acids



Using ML in low data regime



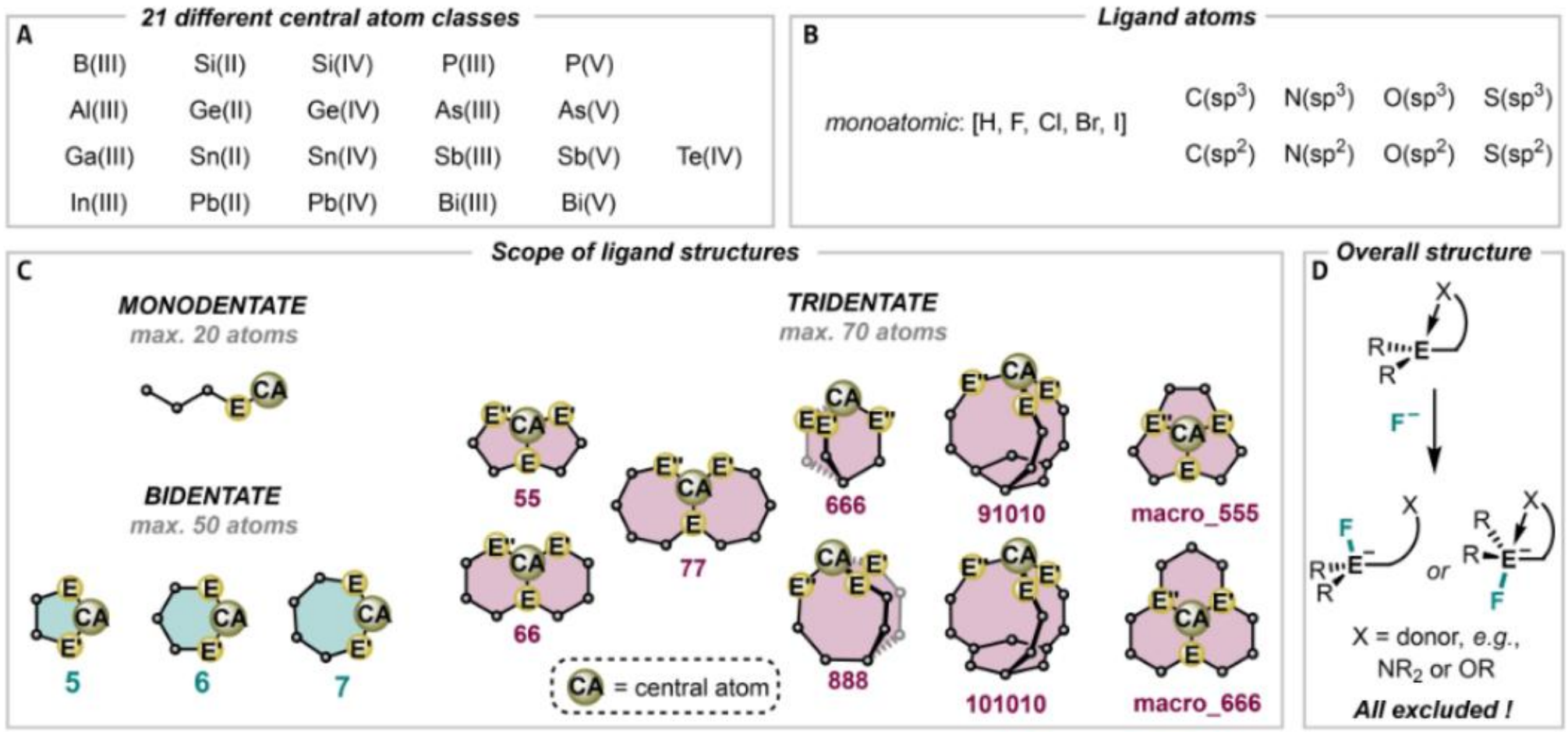
With models close to statistical analysis to perform XIA

- To get more insights into the qualitative concept of LA
- To design new LA

Massey, A. G. ; Park, A. J. *J. Organomet. Chem.*, 1964, **2** , 245
Alharbi, M. M.; Van Ingen, Y.; Roldan, A.; Kaehler, T.; Melen, R. L. Dalton Trans. 2023, 52, 1820
Ben Saida, A. ; Chardon, A. ; Osi, A. ; Tumanov, N. ; Wouters, J. ; Adjieufack, A. I. ; Champagne, B. ; Berionni, G. *Angew. Chem. Int. Ed.* **2019**, *131*, 17045
Huang, K.; Dutton, J. L.; Martin, C. D. *Chem. Eur. J.* **2017**, *23*, 10532

ML for predicting Lewis acidity

Predicting FIA using graph neural network

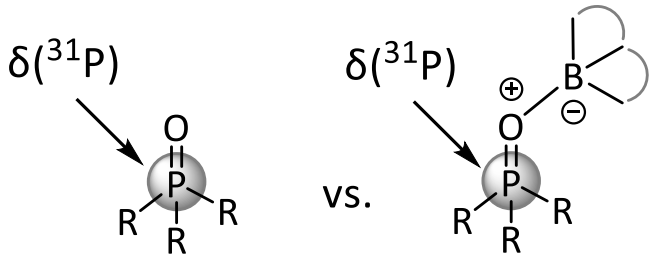


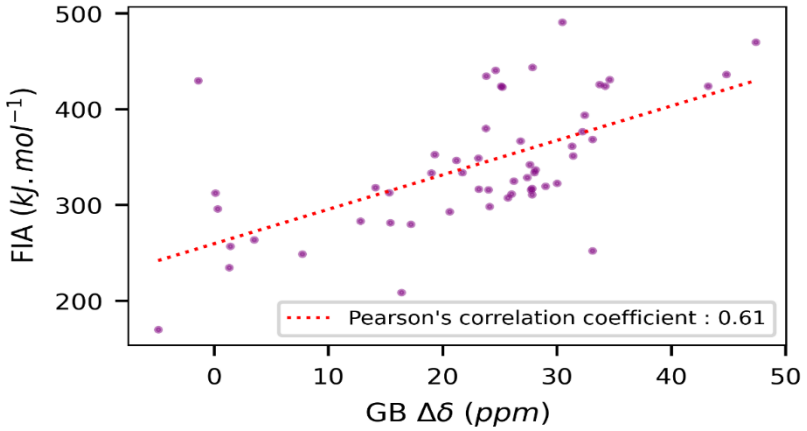
49 000 molecules !

Sigmund, L. M.; S., S. S.; Albers, A.; Erdmann, P.; Paton, R. S.; Greb, L. *Angew. Chem. Int. Ed.* **2024**, e202401084.

Lewis acidity determination

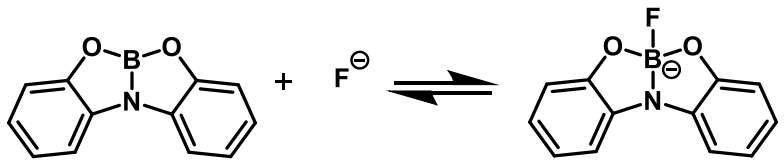
Experimental determination: Gutmann-Beckett NMR displacement



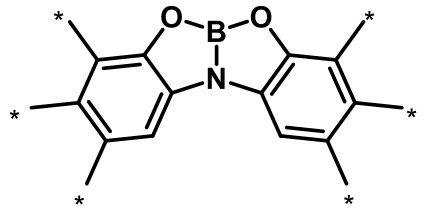


Theoretical determination: Fluoride ion affinity

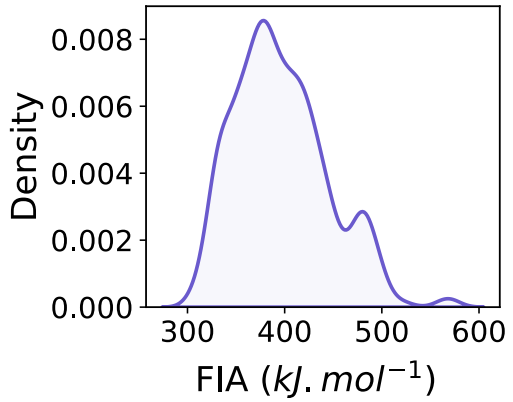
For a reliable database based on
DFT calculations (M062X/6-31G(d))



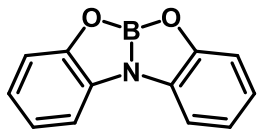
$$\Delta_r H° = -\,\mathrm{FIA}$$



-H, –F, –OH, –NH_3, –Cl, –Br, –CH_3,
–OMe, –$N(Me)_2$, –CN, –*t*Bu,
–CF_3, –NO_2



Molecular descriptors



SMILES : c12c(OB3Oc4c(N32)cccc4)cccc1

Molecular representation

Molecular descriptors

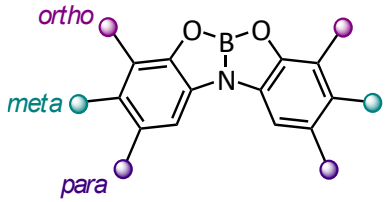
Training data

Chemoinformatics

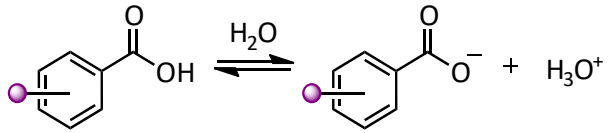
DFT-calculated

Method : M06-2X/6-31g

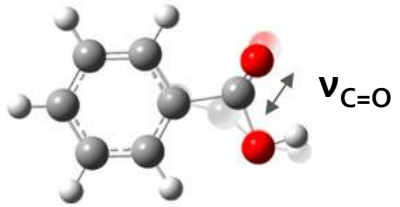
Hammett-extended

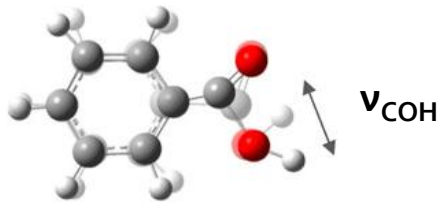


Global parameters	Dipolar moment, frontier orbitals energies, absolute electronegativity…
Atomic parameters for B	Charge, valence, buried volume…



Hammett σ constants	σ_m, σ_p
Steric parameters	Sterimol **B1**, **B5** and **L** for *o*, *m* & *p* Torsional angle : $\theta_{tor,o}$
IR parameters	$\nu_{C=O}$, ν_{COH}, $I_{C=O}$, I_{COH} for *o*, *m* & *p*
NBO charges	NBO_C, $NBO_{=O}$, NBO_O, NBO_H for *o*, *m* & *p*



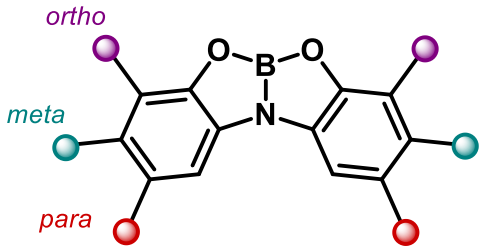


⟹ 244 concatenated features

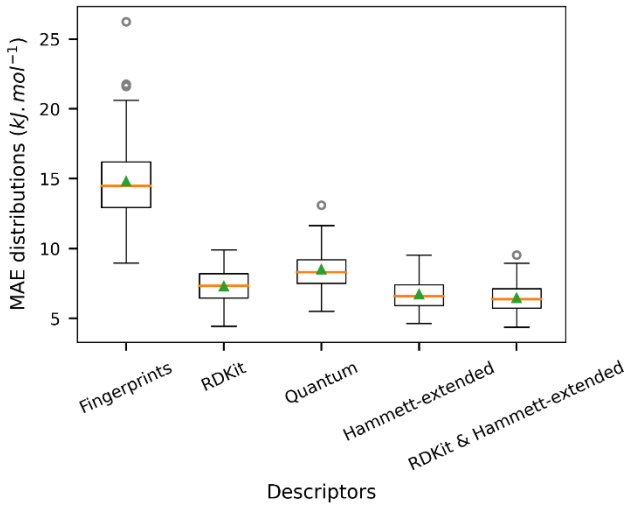
Żurański, & Doyle, *React. Chem. Eng.* **2022**, *7*, 1276

Santiago, C. B.; Milo, A.; Sigman, M. S. *J. Am. Chem. Soc.* **2016**, 138, 13424

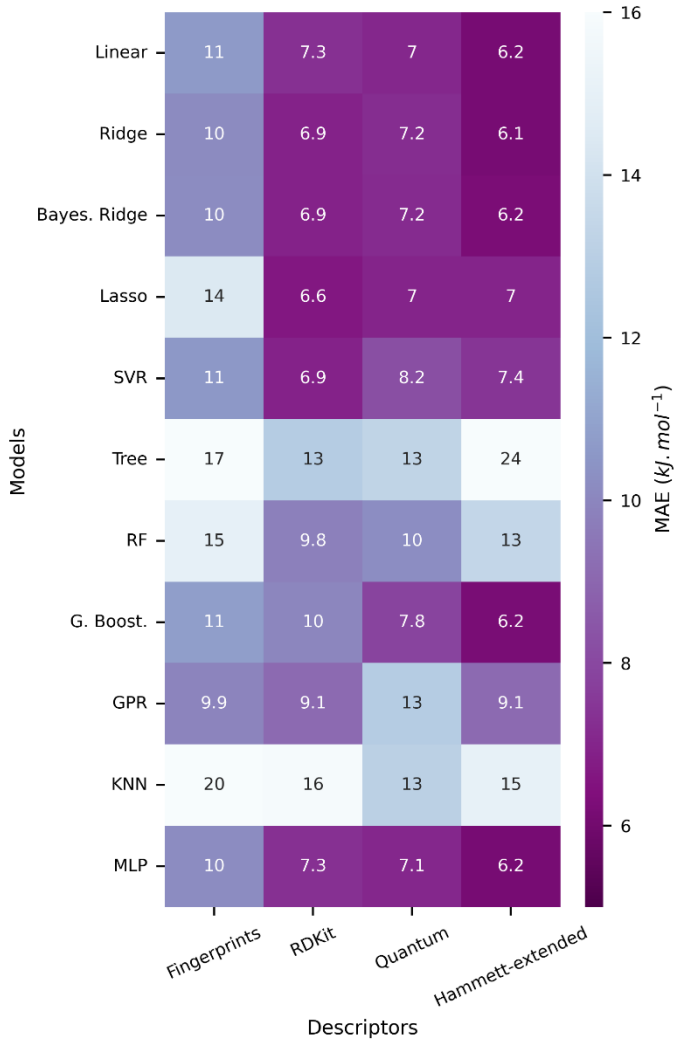
Model construction



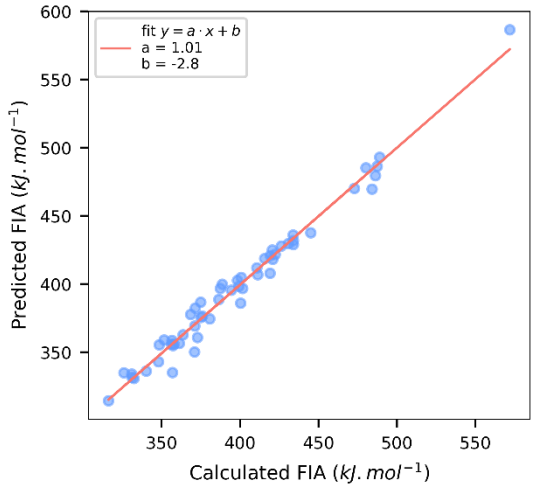
Descriptors evaluation



Test set



Optimized model: Linear Regression



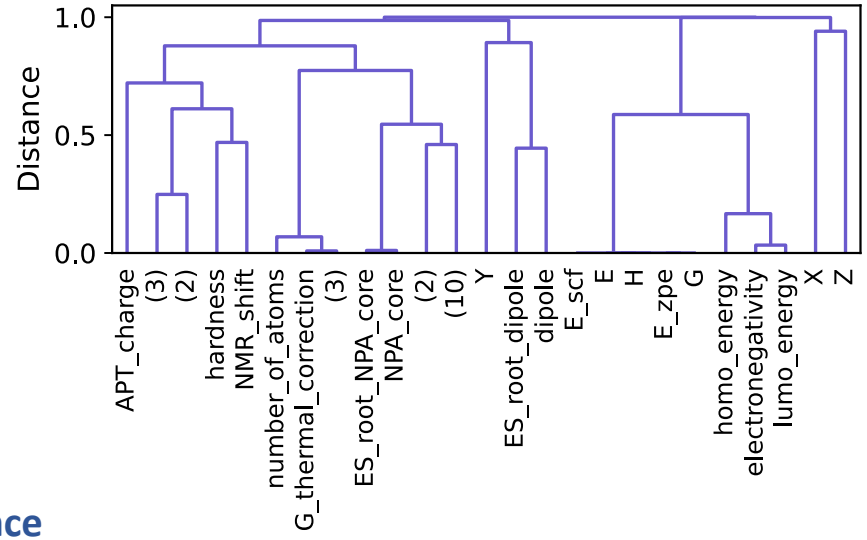
MAE about 5 kJ.mol^{-1}

⇨ Regression models offer an optimal balance between interpretability and prediction accuracy

Interpretability: finding the key feature of LA

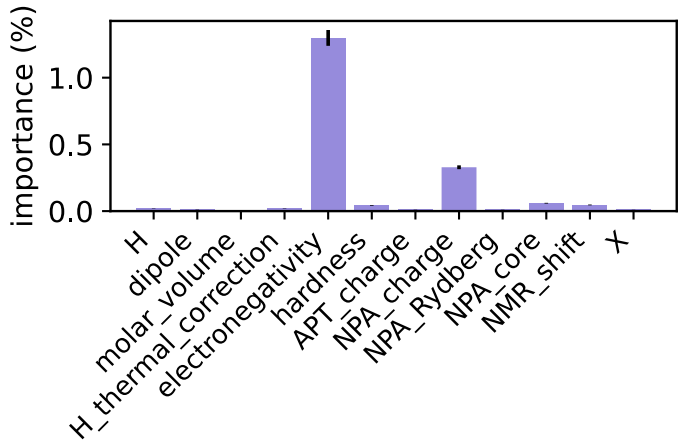
Hierachical clustering

12 uncorrelated features



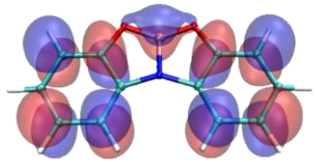
With DFT descriptors

Permutation importance



Absolute electronegativity

$$\chi = {}^{-1}/_{2}\,(E_{HOMO} + E_{LUMO})$$

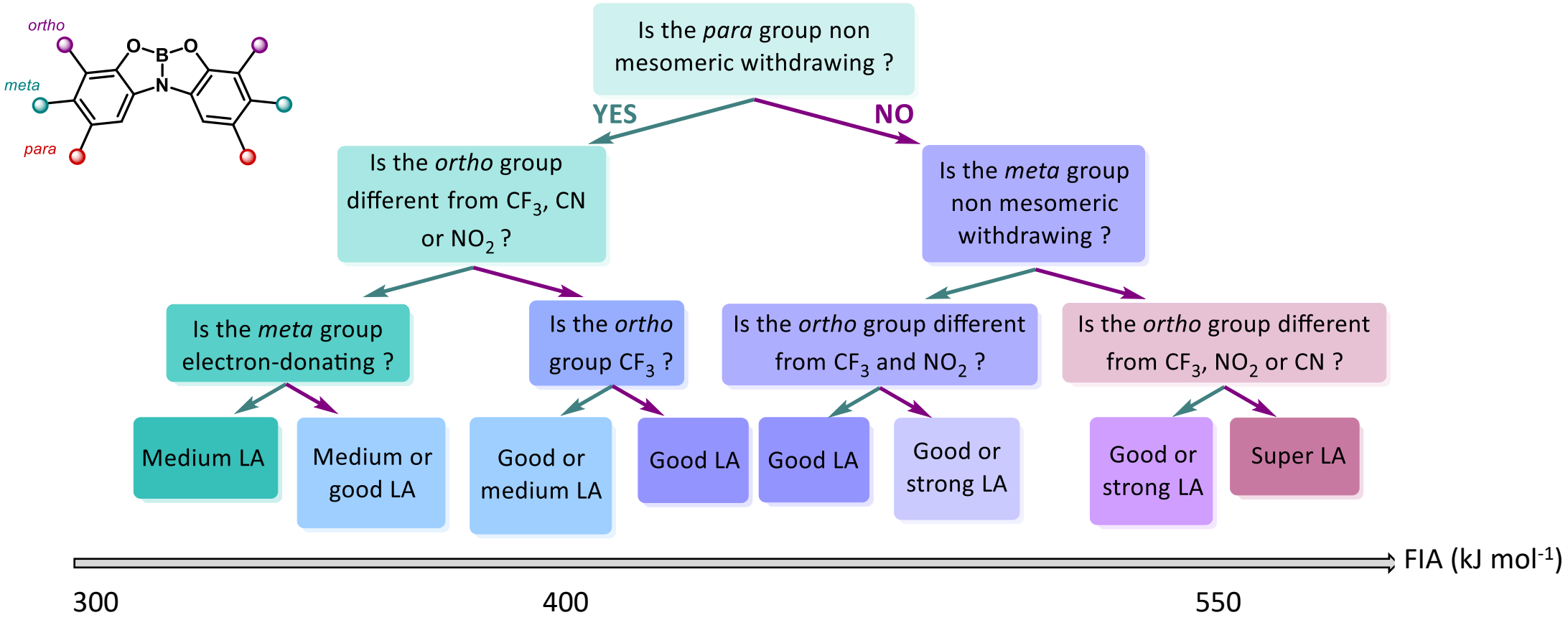


⇨ Unexpectedly Lewis Acidity is more influenced by MO interactions than by Coulombic interactions (Hard acidity)

A classification to help design

Decision tree with 3 questions

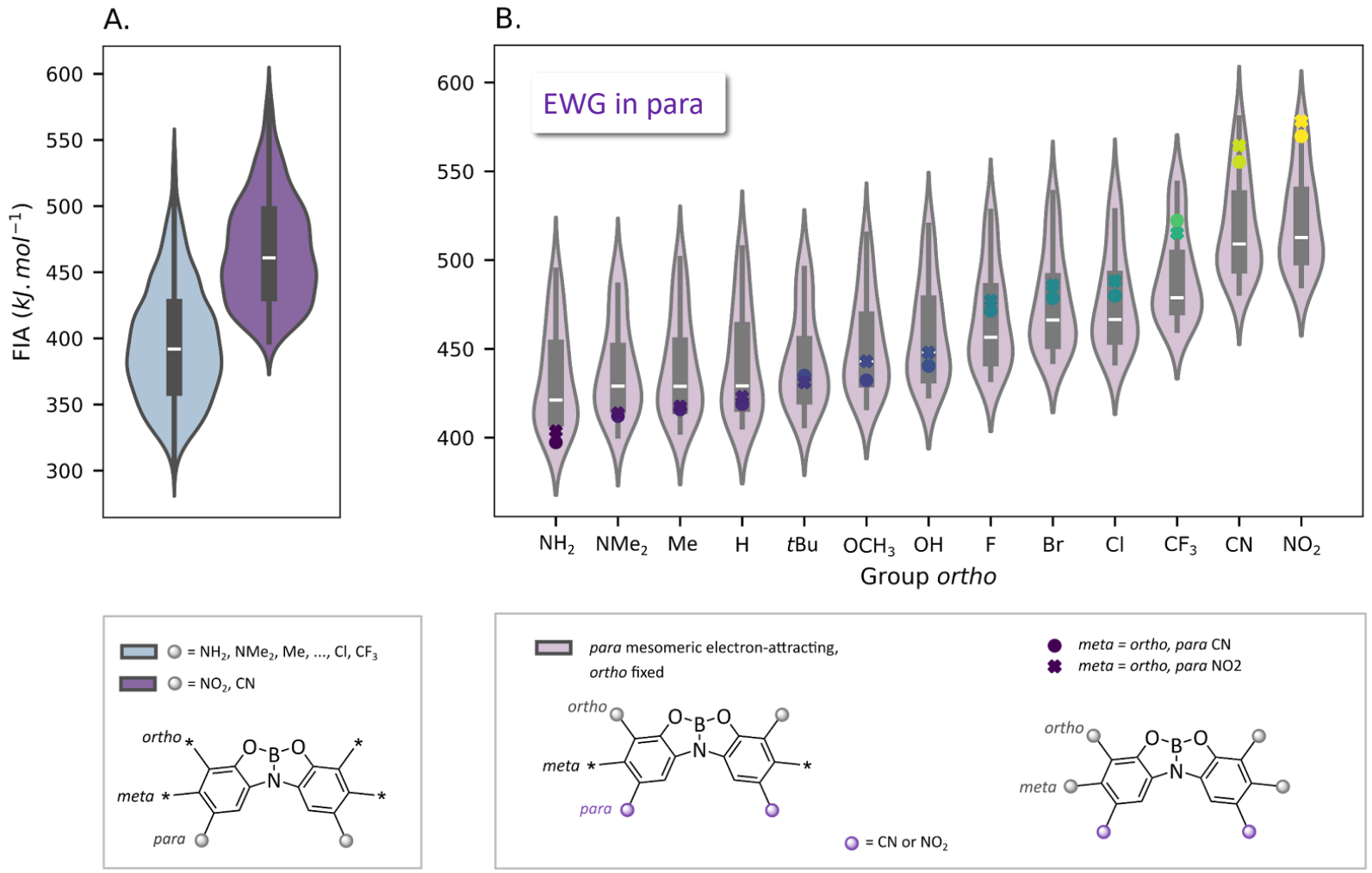
With Hammett-extended descriptors



⇨ In low-data regime, ML models can accurately and interpretably predict Lewis acidity, bridging the gap between advanced ML techniques and traditional chemical intuition

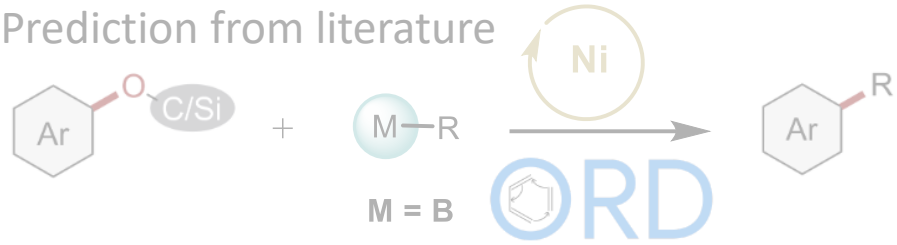
Digital Discovery **2025**, *4*, 3623

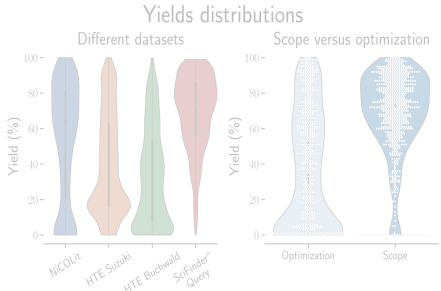
Screening and Analysing the database



AI in our lab

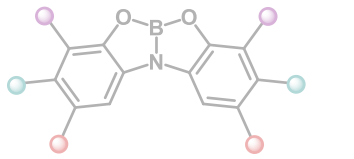
Prediction from literature

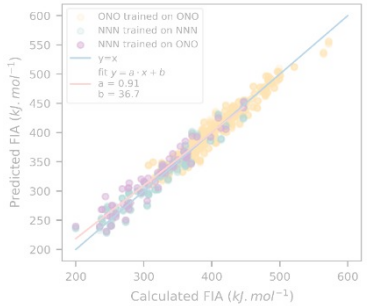


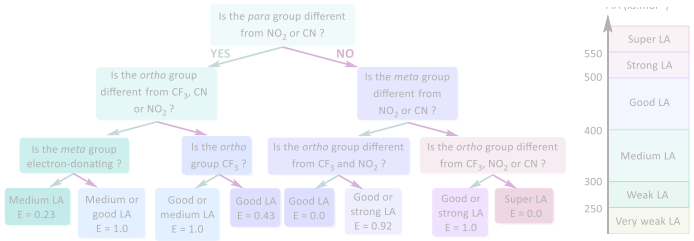


J. Am. Chem. Soc., **2022**, *144*, 14722

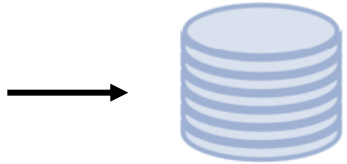
Prediction of Properties



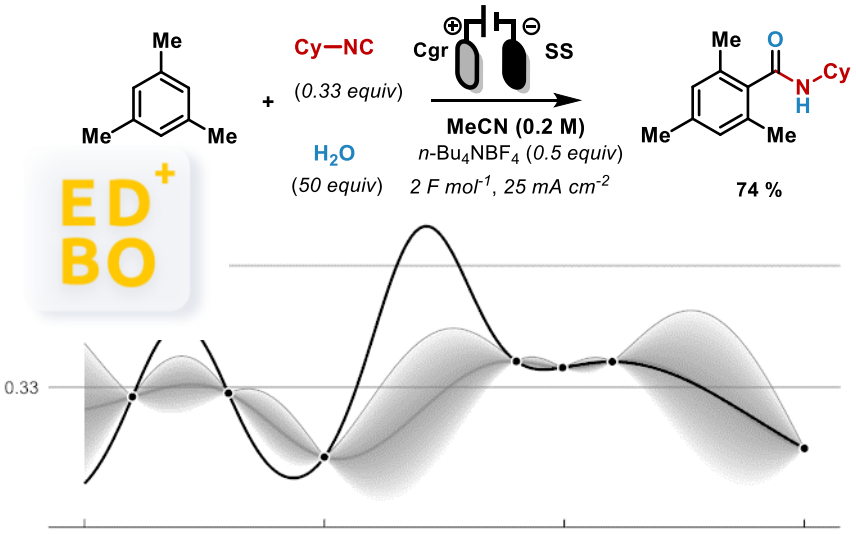




Digital Discovery **2025**, *4*, 3623



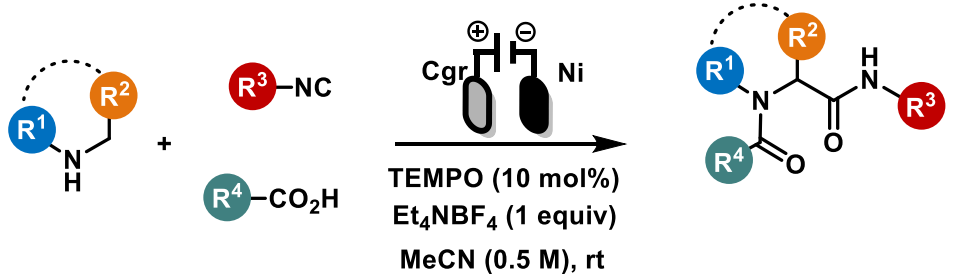
Bayesian Optimisation



C. R. Chim. **2026**, accepted. DOI: 10.5802/crchim.431

Electrosynthesis and IMCRs

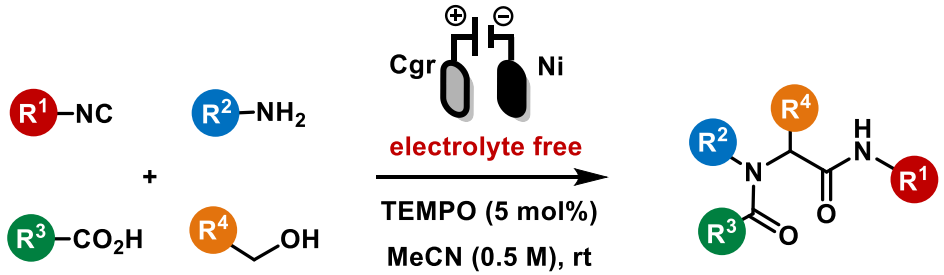
Electro-induced Ugi-Joullié coupling



Pan, N.; Ling, J.; Zapata, R.; Pulicani, J.-P.; Grimaud, L.; Vitale, M. R. *Green Chem.* **2019**, *21*, 6194

Undivided cell

Electro-oxidative Ugi Reaction

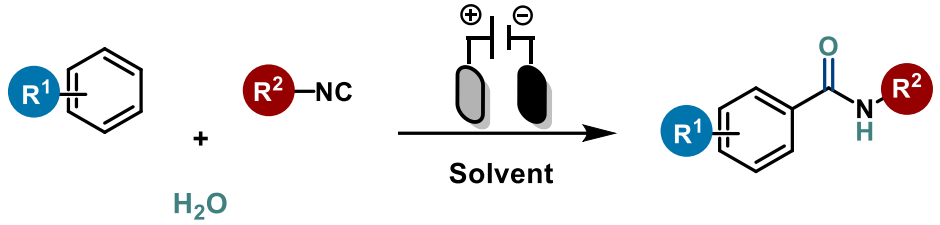


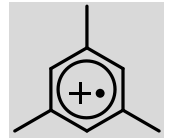
Pan, N.; Xinen Lee, M.; Bunel, L.; Grimaud, L.; Vitale, M. R. *ACS Org. Inorg. Au* **2021**, *1*, 21

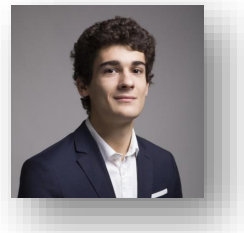
For the electro-Passerini, see: Simon, A.; Bachollet, S. P. J. T.; Pan, N.; Grimaud, L.; Vitale, M. R. *Synlett* **2024**, *35*, A-E

New electrophiles in IMCRs

Carbocations







Virgile Rouffeteau

For cation-pool: Yoshida *J. Am. Chem. Soc.* **2013**, *135*, 5000

For previous cyanation: Gooßen *Chem. Eur. J.* **2018**, *24*, 11288

Dr Maximilian Fleck

Experimental conditions

Substrates' substituants

Stoichiometry

Concentration

Solvent(s)

Temperature

etc…

+

Electrosynthesis

Electrodes materials

Supporting electrolyte

(nature/quantity)

mA or Volt

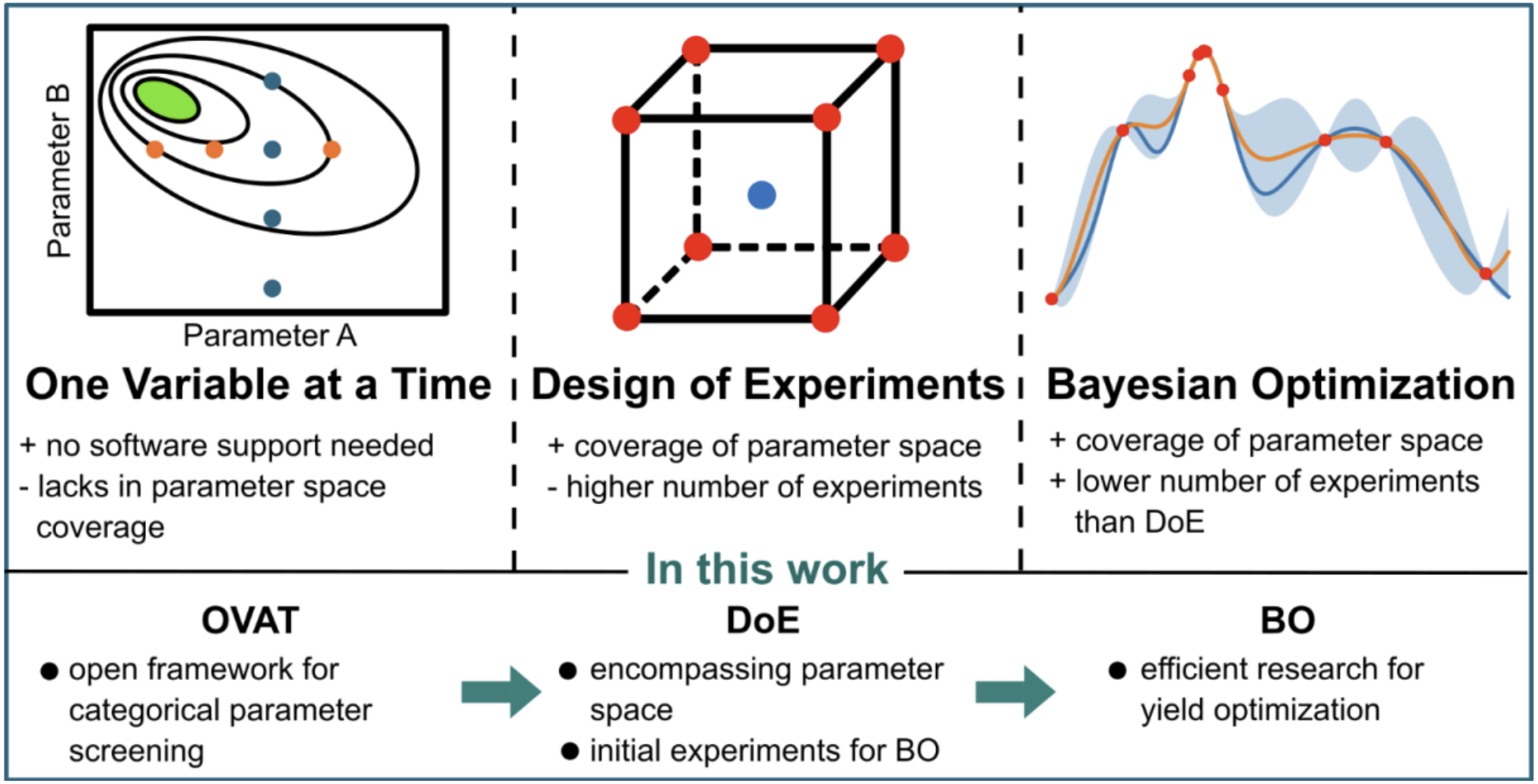
Charge (e^- injected per substrate)

etc…

> 14 parameters

> 10^9 possible combinations

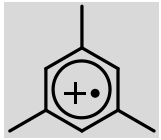
Different optimization methods

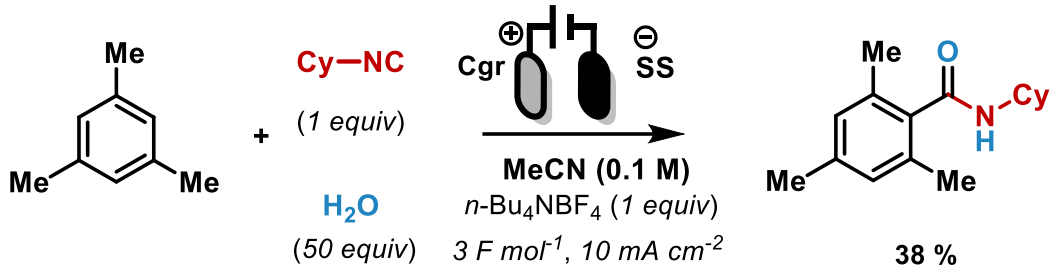


Alán Aspuru-Guzik,[b, c, e, g, h, i, j, k, l] and Siegfried R. Waldvogel ChemRXIV 2026 10.26434/chemrxiv-2026-8rd5c

Chemical Space for Bayesian Optimization

Electro-carbamoylation of electron-rich arenes using Bayesian optimisation





- Output Yield (based on limiting Substrate)
- Definition of space

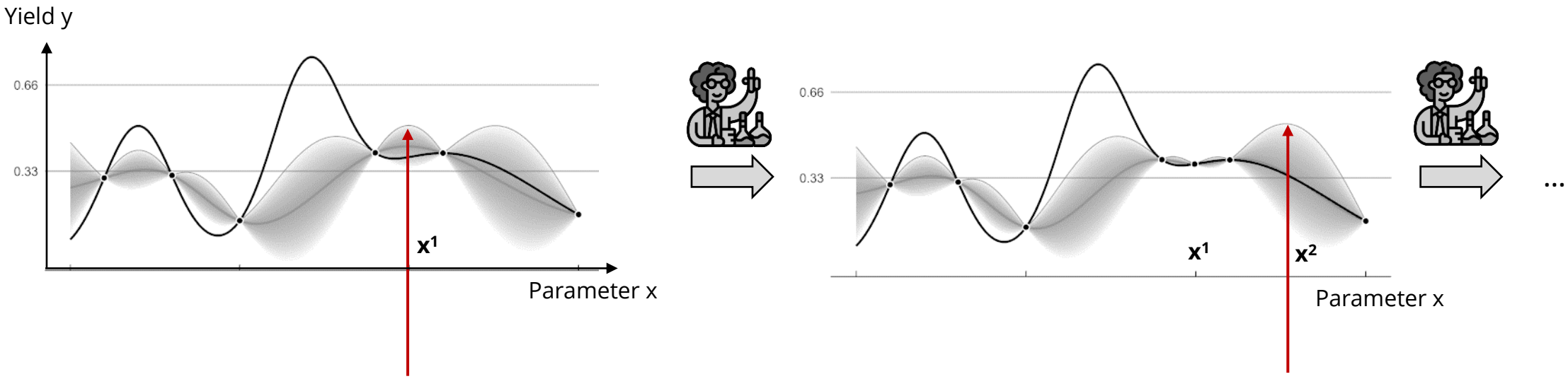
6 Parameters	Values						
Cy-NC equivalent(s)	0.33	0.5	0.75	1	1.5	2	3
Water equivalent(s) %limiting	1	5	10	25	50		
Salt equivalent(s) %limiting	0.5	1	2				
Current Density (mA.cm^{-2})	5	10	15	20	25		
Limiting Reagents' Concentration(M)	0.05	0.1	0.15	0.2			
Faraday per mole of arene	2	3	4	5	6		

- Evaluation of the efficiency (NMR with ext. Std)
- Algorithm: EDBO+

10 500 combinations !

Doyle, A., *et al*. *Nature*, **2021**, 89

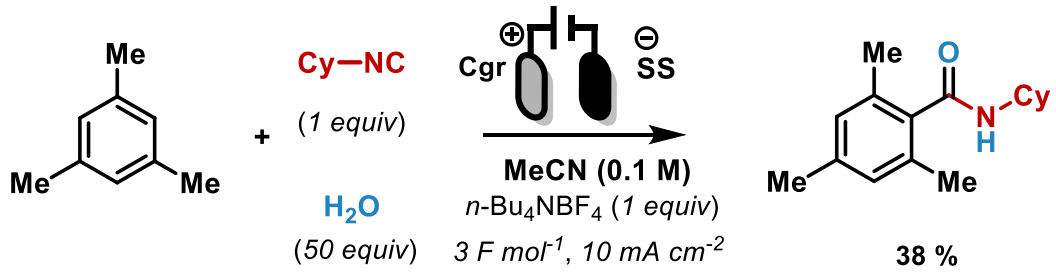
Bayesian Optimization

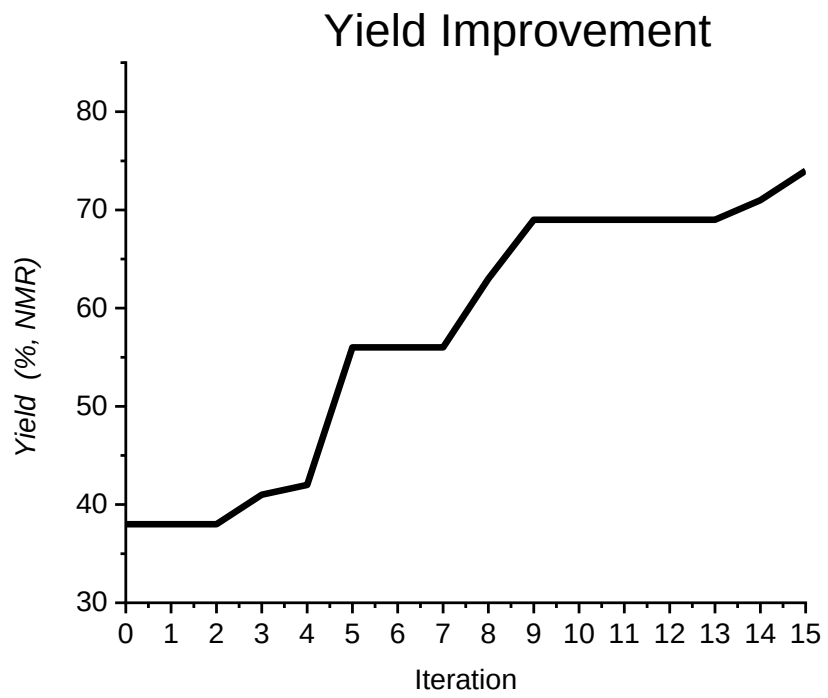


Highest peak at x^1,
new theroretical highest yield $y^{1\prime}$

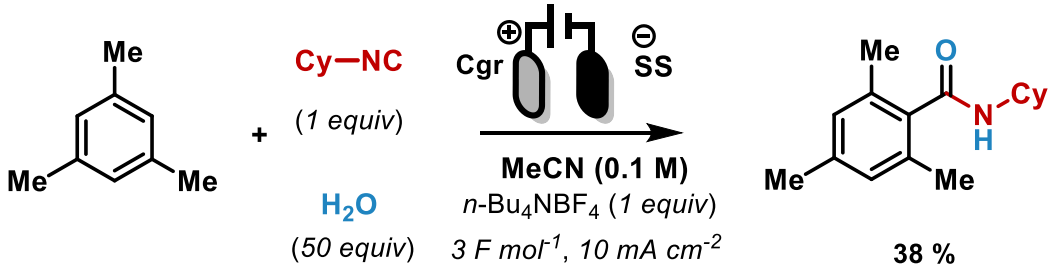
New highest peak at x^2,
new theoretical highest yield $y^{2\prime}$

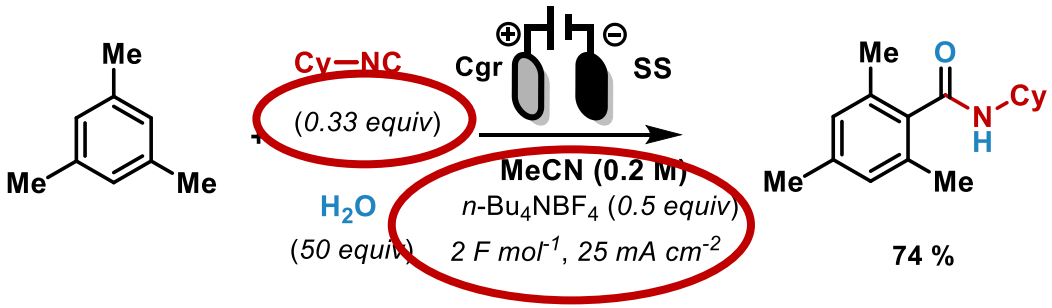
Bayesian Optimization for electrosynthesis





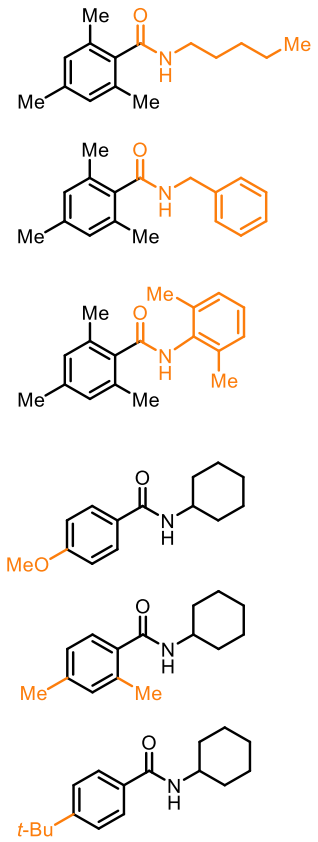
Optimized conditions



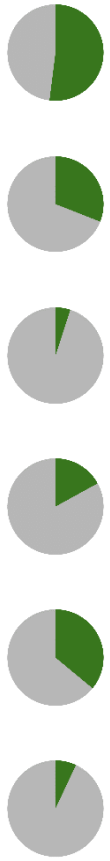


High faradic efficiency

Product



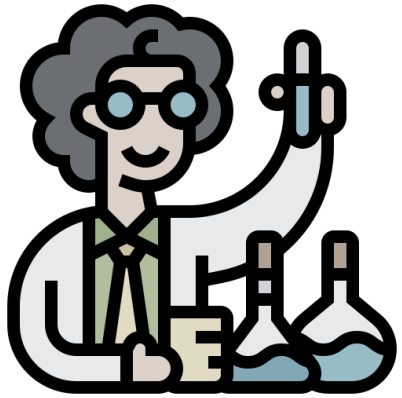
EDBO

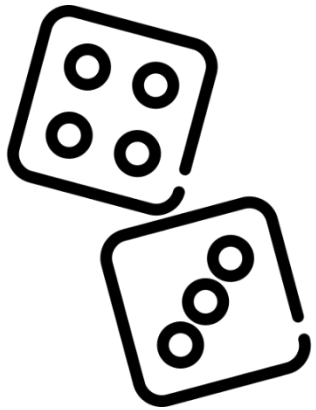


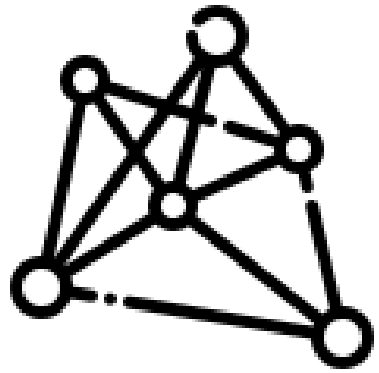
C. R. Chim. **2026**, accepted. DOI: 10.5802/crchim.431

Training Dataset Design

Precedents







- Intuitive sampling

Majority of the literature

- Random sampling

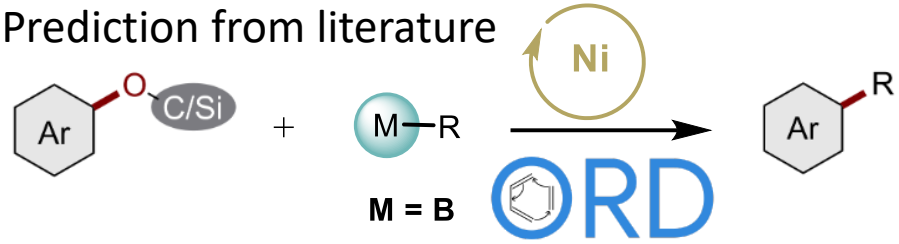
Doyle *et coll*, *Nature*, **2021**, *590*, 89.

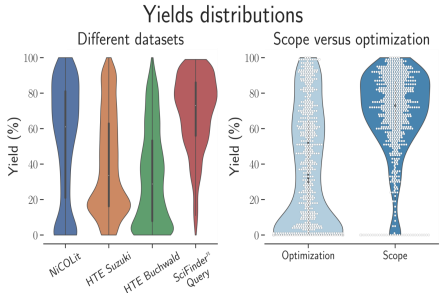
- Orthogonal sampling

Felpin *et coll*, *Org. Process Res. Dev.* **2024**, *28*, 1597.

AI in our lab

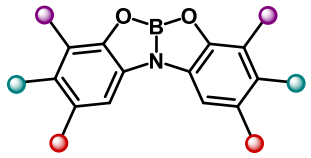
Prediction from literature

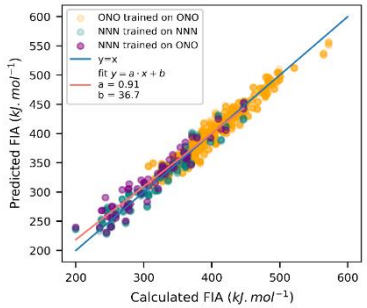


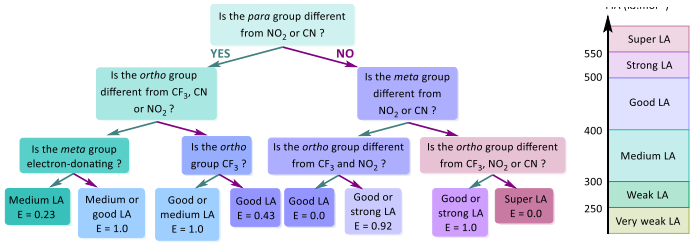


J. Am. Chem. Soc., **2022**, *144*, 14722

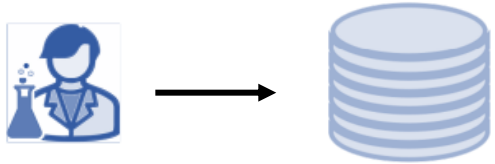
Prediction of Properties



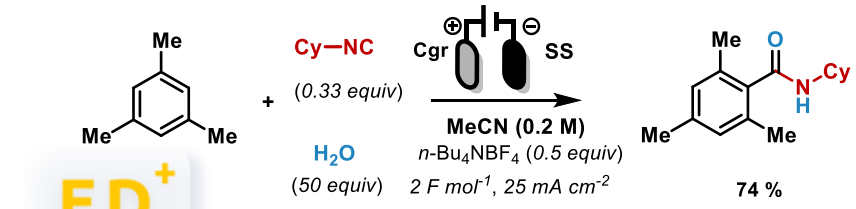


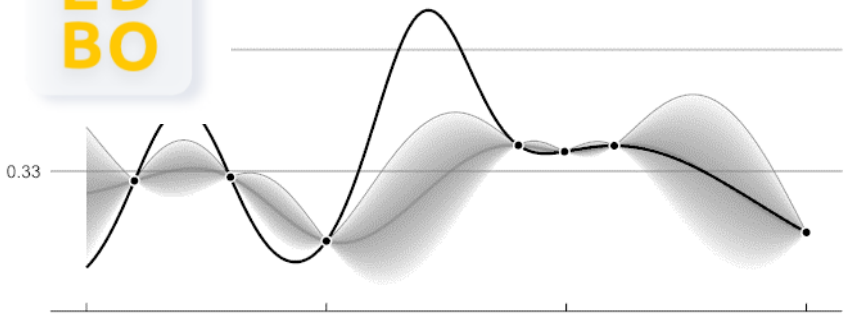


Digital Discovery **2025**, *4*, 3623



Bayesian Optimisation





C. R. Chim. **2026**, accepted. DOI: 10.5802/crchim.431

Dr Maxime Vitale

Pr Rodolphe Vuilleumier

Dr Maximilian Fleck

Thank you !

PhD students

Dr Na Pan
Dr Johanne Ling
Dr Khaoula Jaouadi
Dr Aurélien Bailly
Dr Jules Schleinitz
Dr Emile Escoudé
Dr Alexandre Simon
Juliette Fenogli
Virgile Rouffeteau
Mohktar Ben Ghanem
Louis Lengagne
Sabry Zekhenine

Interns

Dr Francesca Brunelli
Dr Sylvestre Bachollet
Lyse Kastner
Elyssa Caron
Clara Perrier
Lilian Bourqui
Salomé Nicolas
Ramiro Zapata
Louis Bunel
Maegan Xinen Lee...