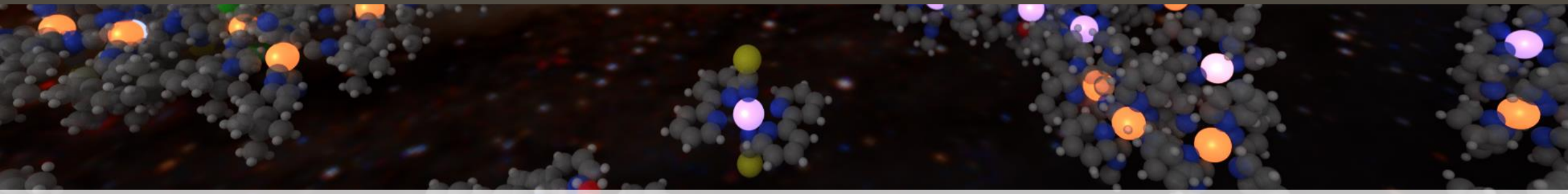
Exploring multi-million compound spaces with chemical accuracy using machine learning



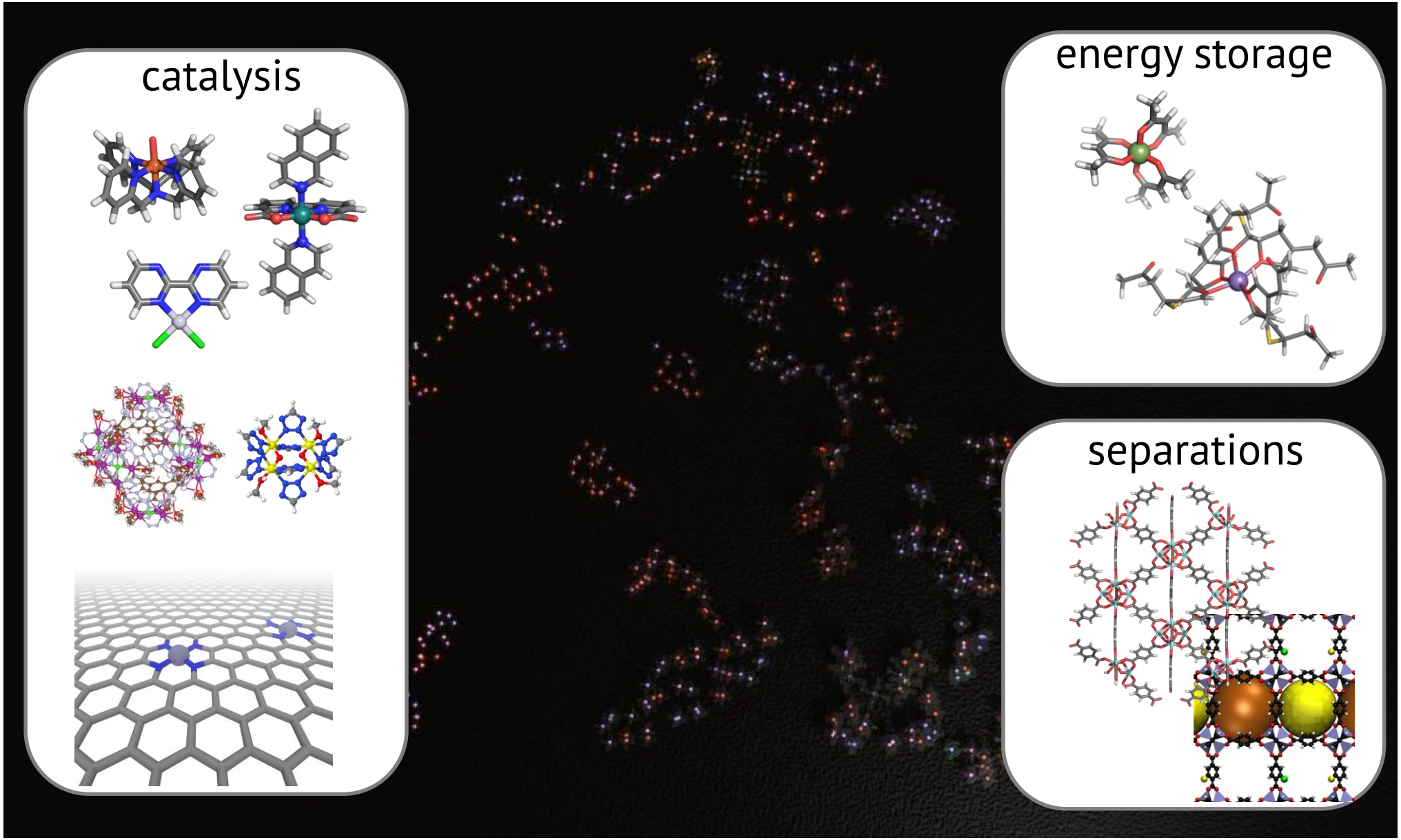
Heather J. Kulik

Associate Professor
Chemistry & Chem. Eng., MIT

IPAM

March 27, 2023

WHERE ARE THE STARS IN CHEMICAL SPACE?



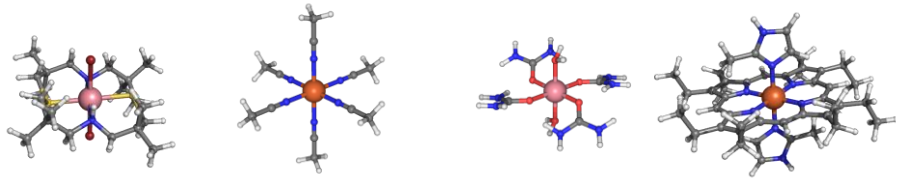
What does it take to **accelerate discovery** in inorganic chemistry with computational chemistry?

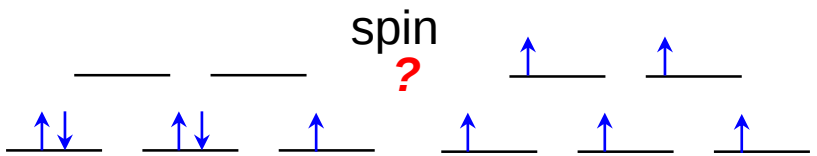
INSPIRATION FROM ORGANIC CHEMISTRY

Challenges for open shell transition metal chemistry:

J.P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin, and **HJK**, *Inorg. Chem.* (2019); A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves, and **HJK**, *Chem. Rev.* (2021).

ADDRESSING VASTNESS IN CHEMICAL SPACE

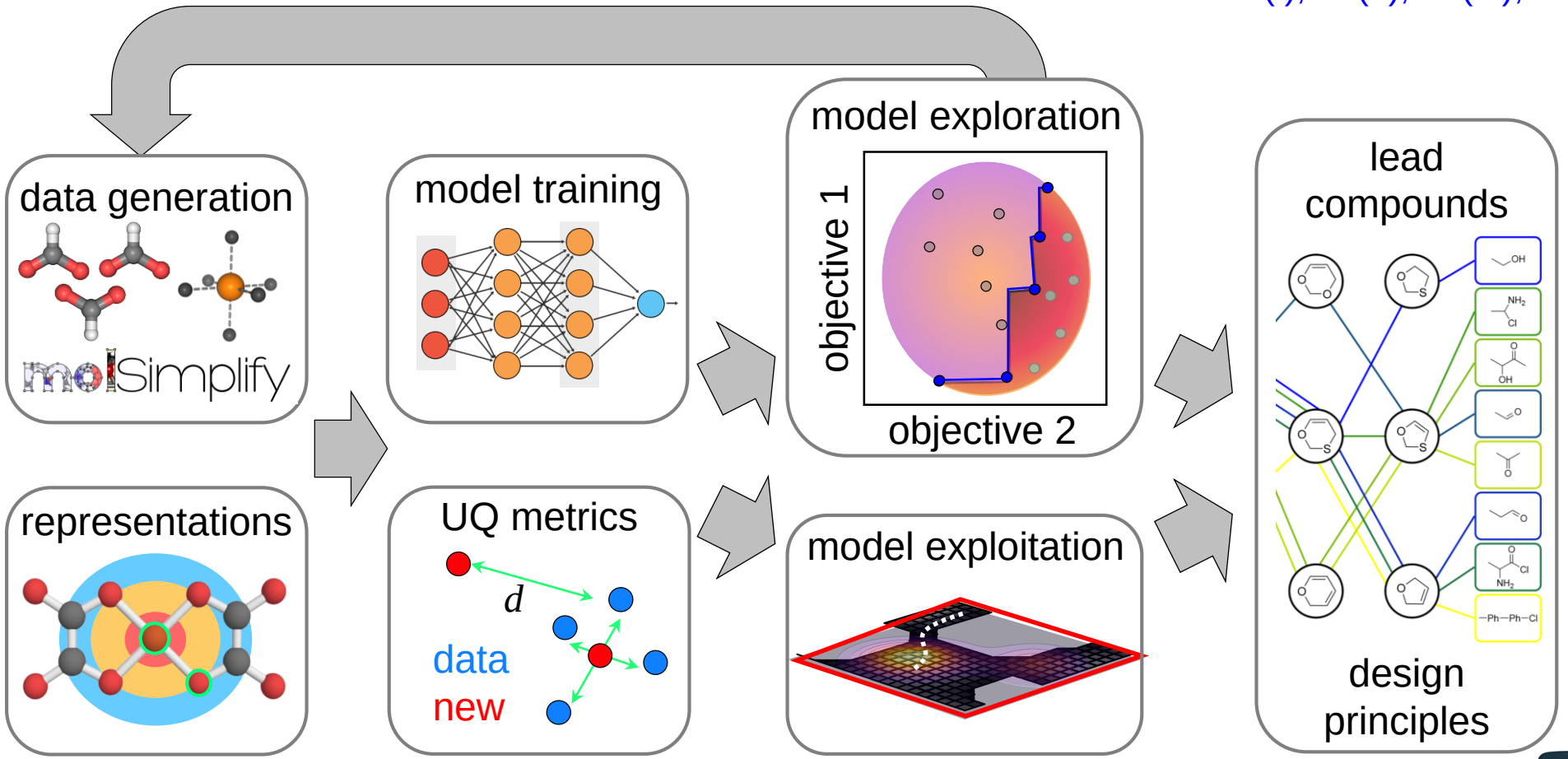




metal and oxidation state

Cr Mn Fe Co Ni

Fe(I), Fe(II), Fe(III), Fe(IV), Fe(V)…



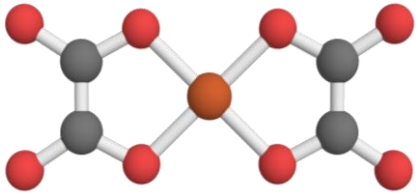
Recent perspective: J. P. Janet, C. Duan, A. Nandy, F. Liu, and **HJK** *Acc. Chem. Res.* (2021).

INORGANIC CHEMISTRY REPRESENTATIONS

Key ingredient: metal-local representations for open shell transition metal chemistry

Revised autocorrelations (RACs)

- Based on Moreau-Broto ACs
- Heuristic properties (*P*)
- Adjustable depth (*d*), ligand-centered, or metal-centered
- Product or difference on molecular graph (geometry-free)



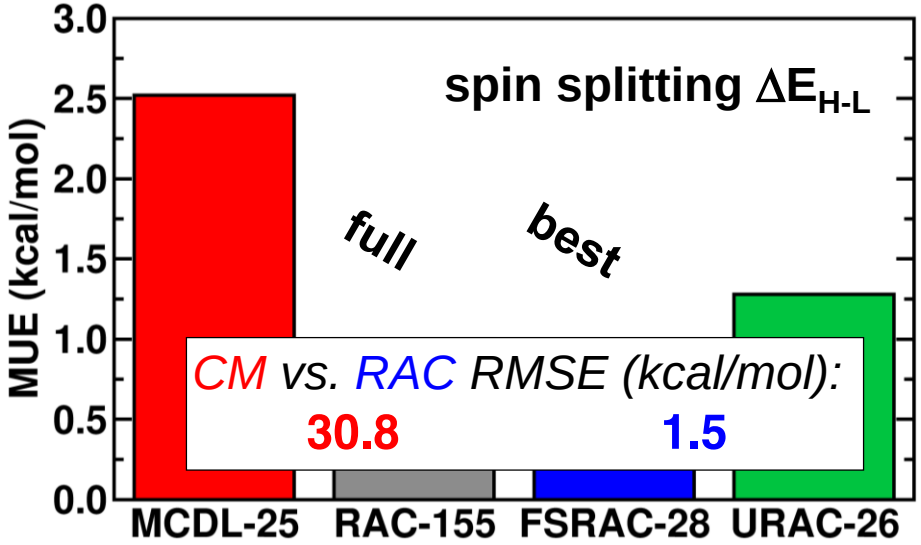
$$P_d = \sum_i \sum_j P_i P_j d\left(d_{ij}, d\right)$$

P:
nuc. charge, *Z*
electroneg., χ
cov. rad., *S*
topol., *T*
ident., *I*

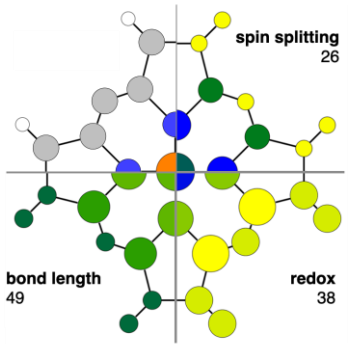
$$^{lc/mc}_{ax/eq/all}P'_d = \sum_i^{lc \text{ or } mc\, scope} \sum_j \left(P_i - P_j\right) d\left(d_{ij}, d\right)$$

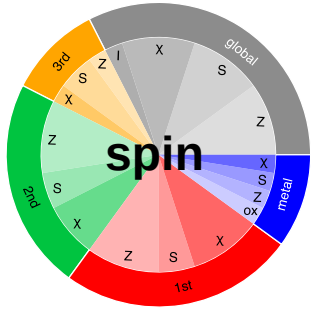
MCDL-25/ANN, 2 hidden-layer 50 nodes, dropout for regularization and credible intervals; vs. RAC-155/KRR models.

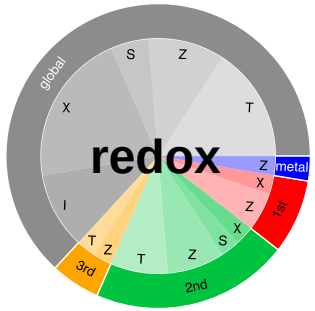
Performance:



Feature importance:





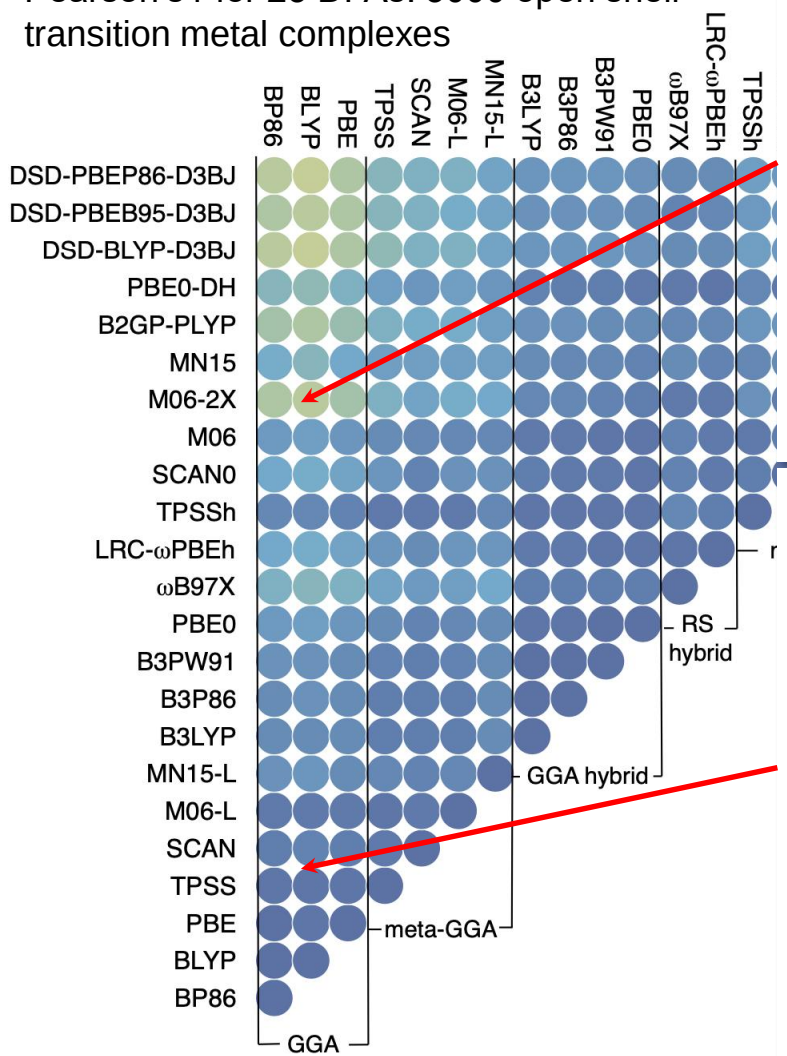


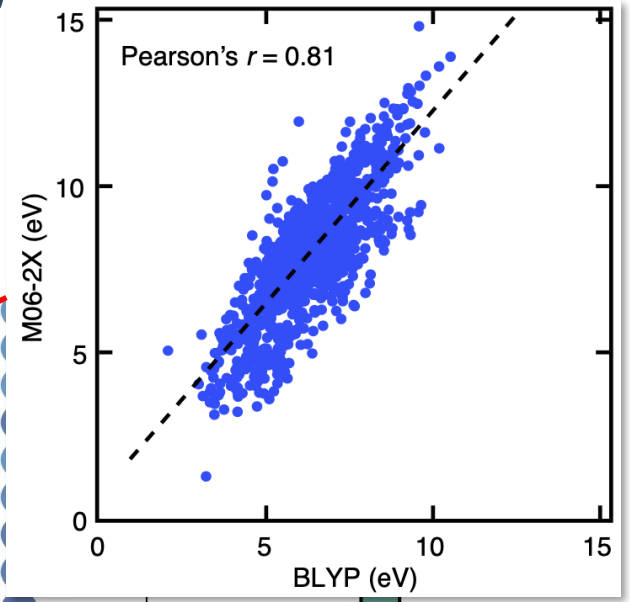
J. P. Janet and **HJK**, *Chem. Sci*. (2017), J. P. Janet and **HJK,** *J. Phys. Chem. A* (2017); J.P. Janet, T.Z.H. Gani, A.H. Steeves, E.I. Ioannidis, and **HJK** *Ind. Eng. Chem. Res.* (2017).

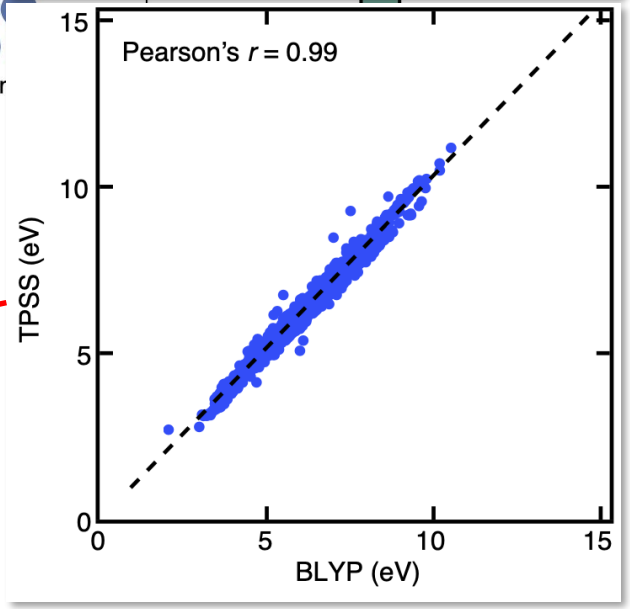
DOES IT MATTER WHICH XC WE CHOOSE?

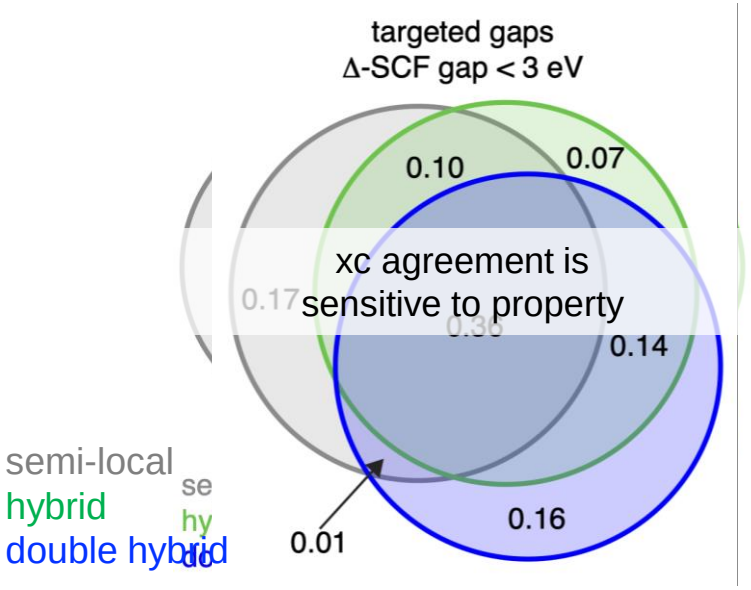
Δ–SCF gap

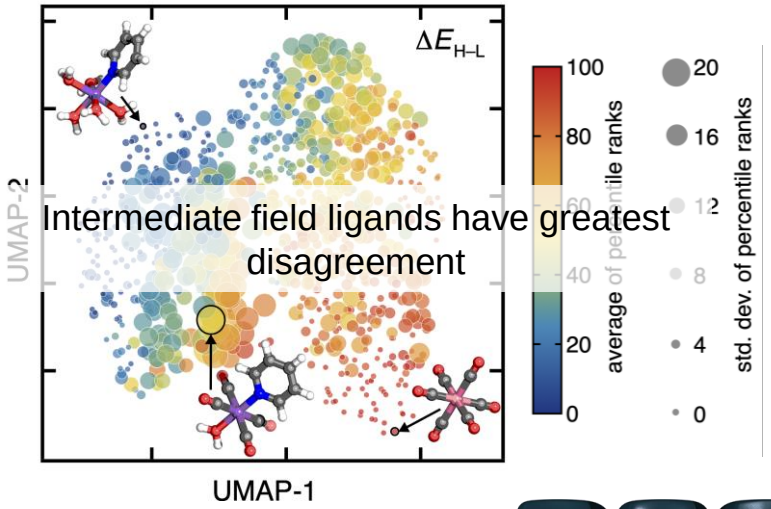
Pearson's r for 23 DFAs: 3000 open shell transition metal complexes









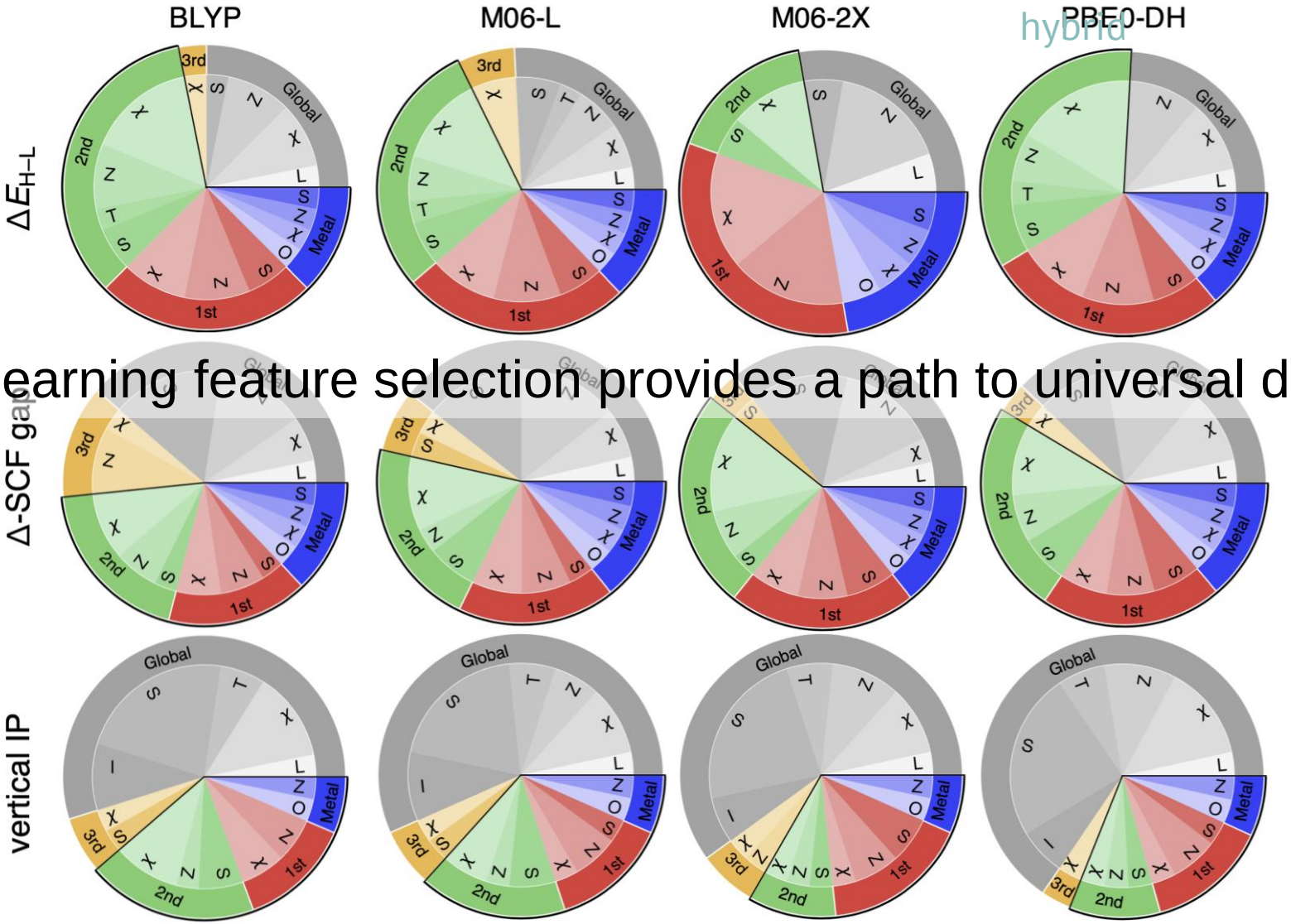


C. Duan, S. Chen, M. G. Taylor, F. Liu, and HJK, Chem. Sci. (2021).

UNIVERSAL DESIGN PRINCIPLES

GGA

double hybrid



Machine learning feature selection provides a path to universal design rules

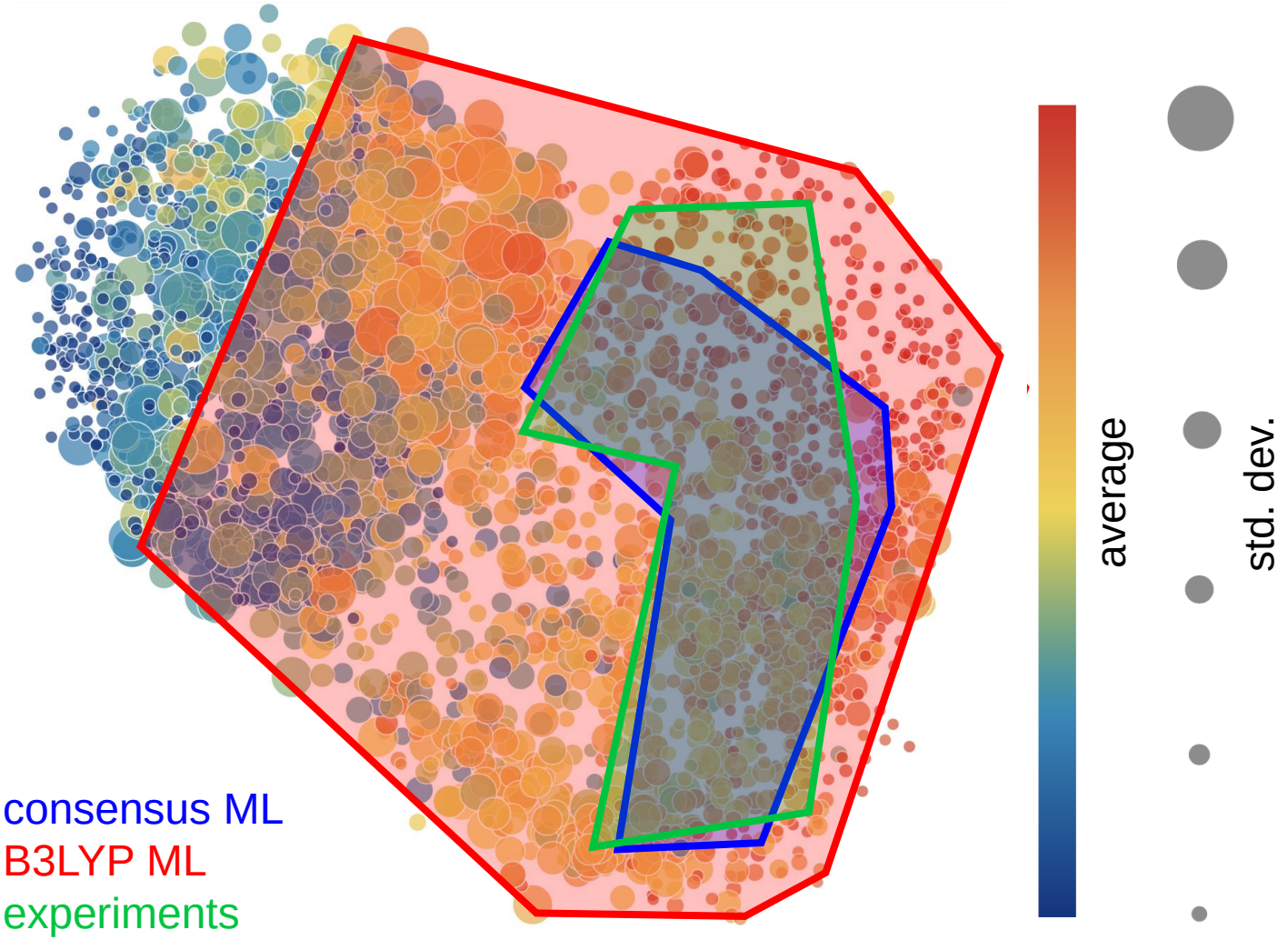
Chenru Duan
Ph.D. '22

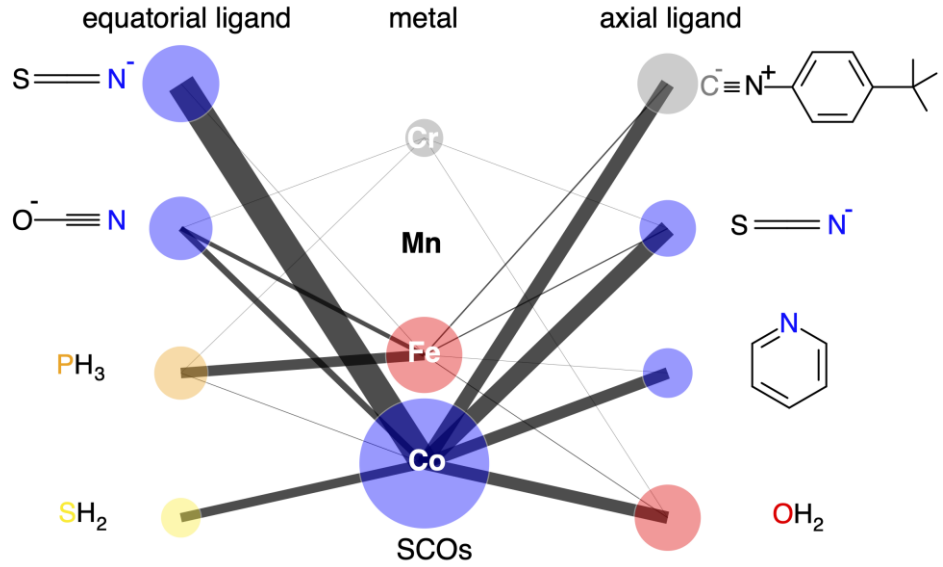
Also applies to varying data set size and basis set!

C. Duan, S. Chen, M. G. Taylor, F. Liu, and **HJK**, *Chem. Sci.* (2021).

DESIGN BY CONSENSUS

Lead SCOs from 1 DFA *vs* consensus of 23 DFA-trained ANNs (>50% agree) in 187k design space:

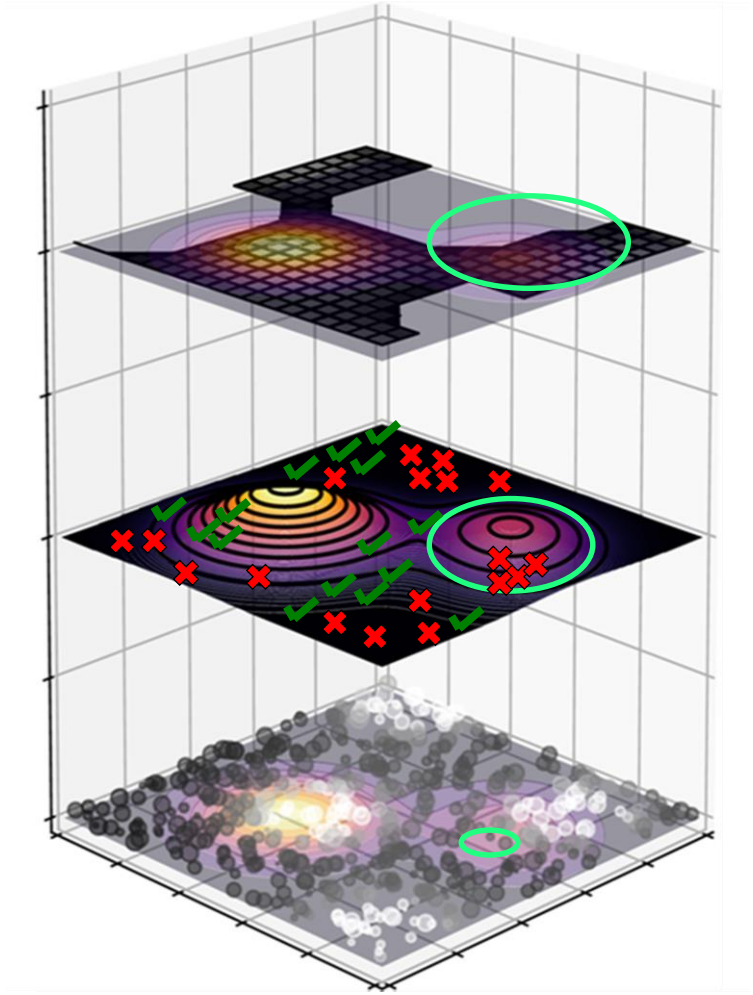




Consensus leads support experimental focus on Fe(II)/N but also suggests Co/N for SCO study.

C. Duan, S. Chen, M. G. Taylor, F. Liu, and **HJK**, *Chem. Sci.* (2021).

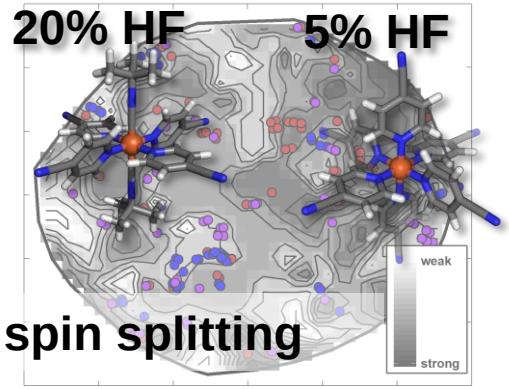
UNCERTAINTY IN DISCOVERY



Uncertainty from:

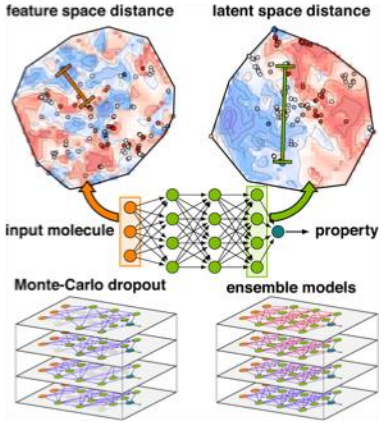
- method choice
- ML model
- calculation outcome

DFT xc choice uncertainty via ML



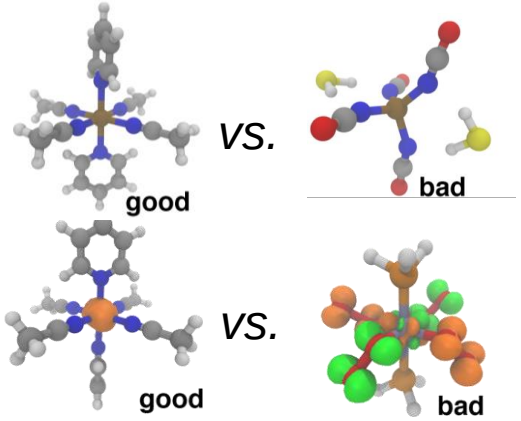
J. P. Janet and **HJK**, *Chem. Sci.* (2017), J.P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, and **HJK**, *Inorg. Chem.* (2019).

ML model uncertainty quantification



J. P. Janet, C. Duan, T. Yang, A. Nandy, and **HJK** *Chem. Sci.* (2019).

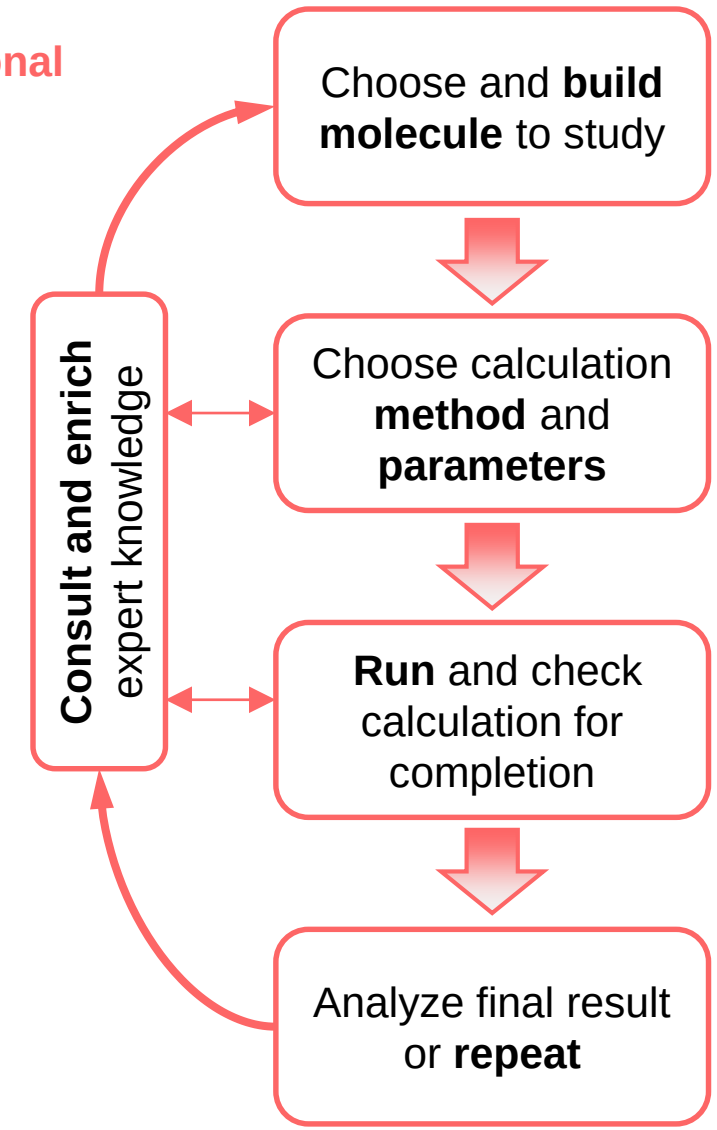
Uncertainty from calculation outcome

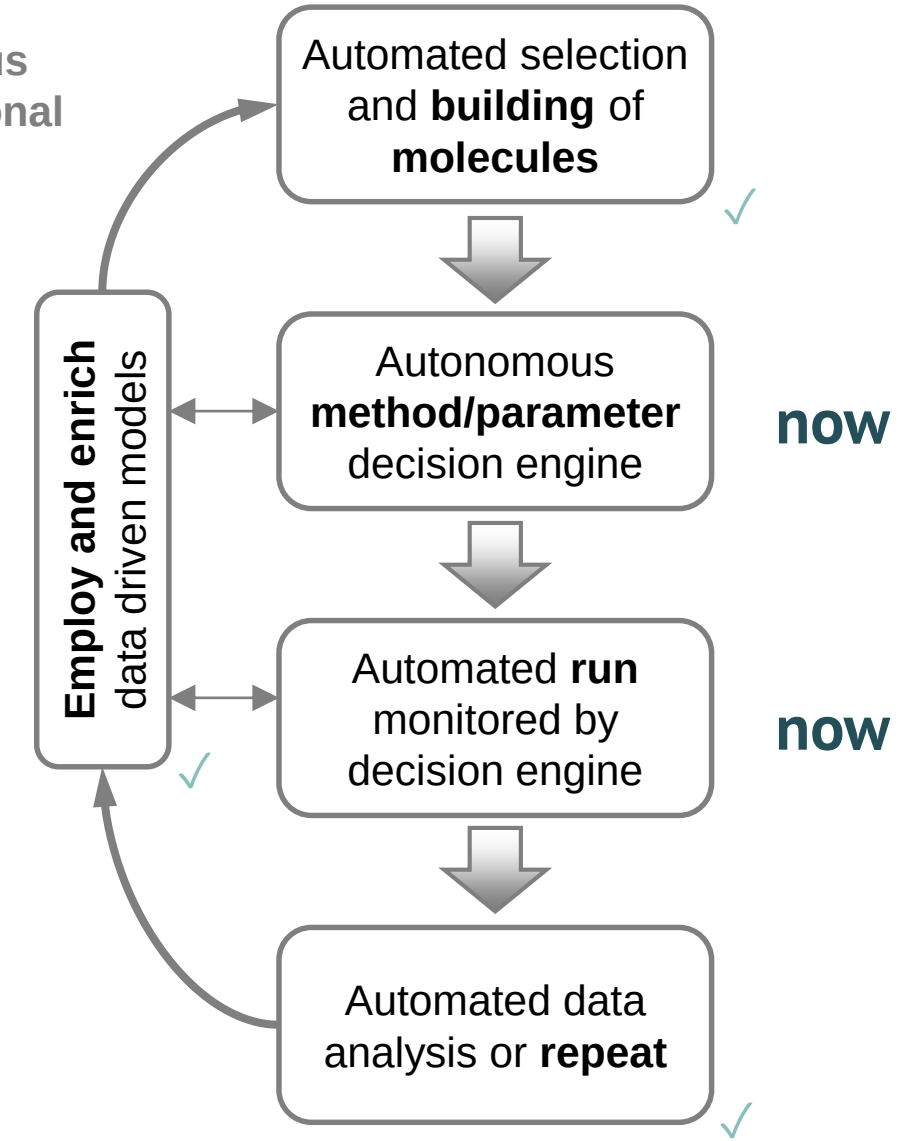


C. Duan, J. P. Janet, F. Liu, A. Nandy, and **HJK**, *J. Chem. Theory Comput.* (2019).

HJK, *WIRES Comput. Mol. Sci.* (2020).

AUTONOMOUS COMPUTATIONAL CHEMISTRY



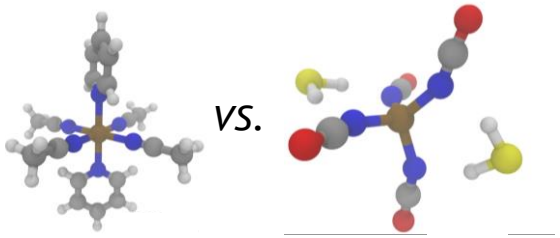


HJK, *WIRES Comput. Mol. Sci.* (2020).

PREDICTING SUCCESS IN DISCOVERY

Defining calculation success:

What we want: What we get:



"good" "bad"

RAC/ANN classifier: 88% accurate

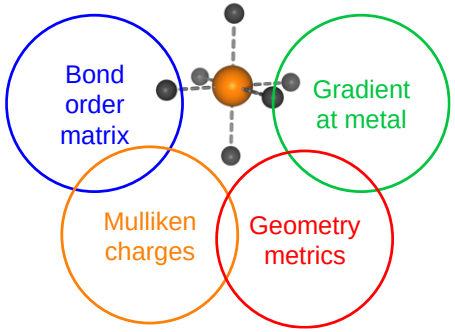
	P	N
P	**283** (304)	**48** (74)
N	**35** (14)	**317** (291)

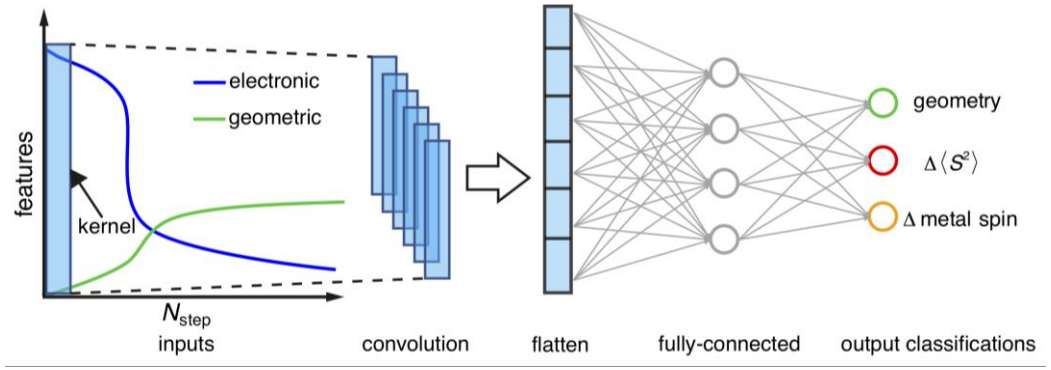
"Bad" takes longer than "good": **88%** of the space in **1/3** the time!

Transferability poor for diverse chemistry:

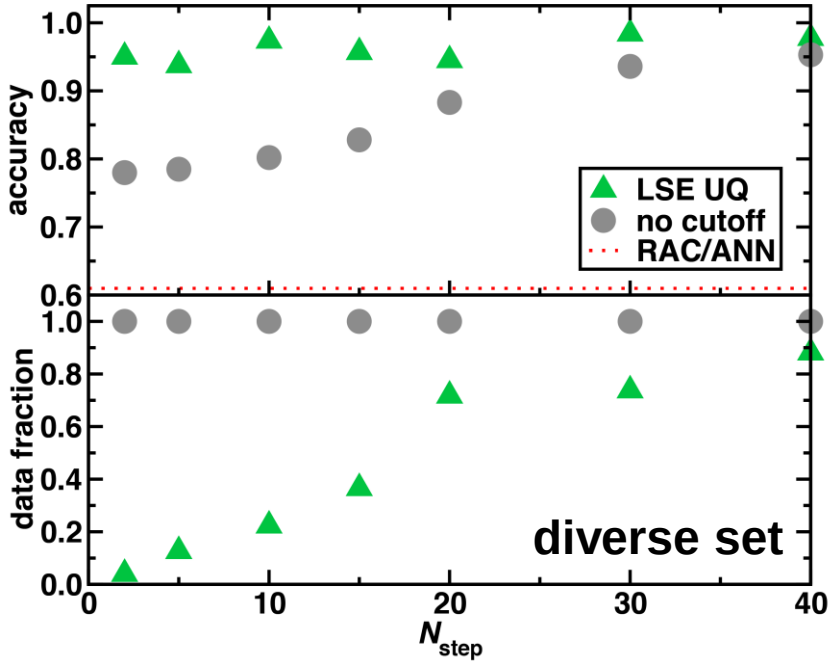
data	diverse (100%)
accuracy	61%

In situ calculation monitoring with new features:





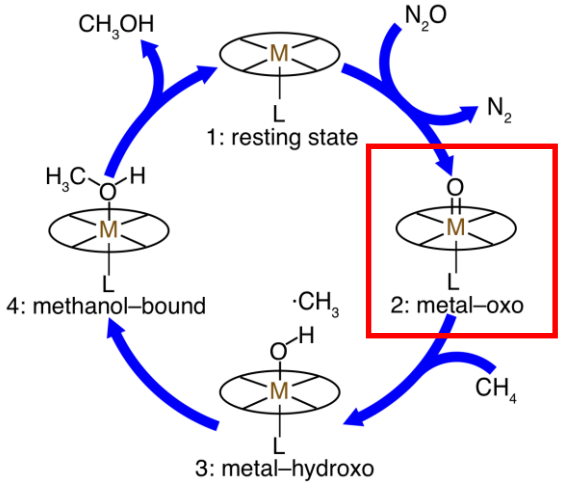
CNN on descriptors from 2-40 steps is better for diverse data, 99% accurate with LSE UQ:

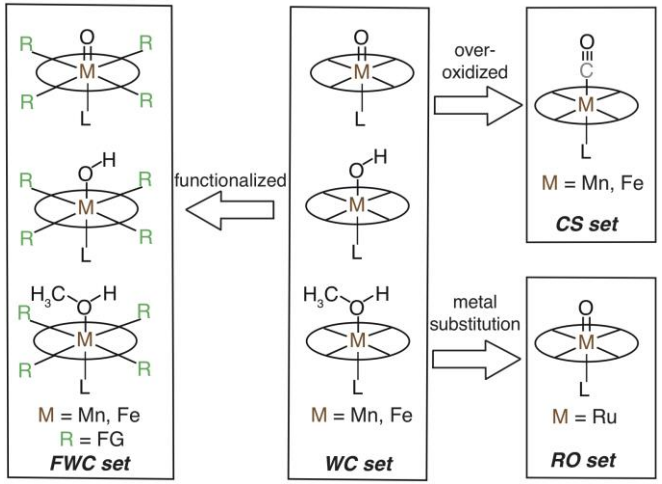


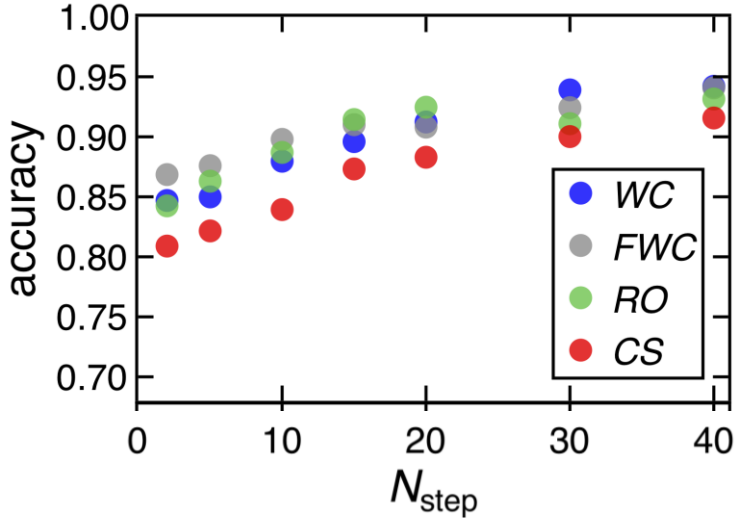
C. Duan, J. P. Janet, F. Liu, A. Nandy, and **HJK**, *J. Chem. Theory Comput.* (2019). A. Nandy, C. Duan, J. P. Janet, S. Gugler, and **HJK** *I&ECR* (2018).

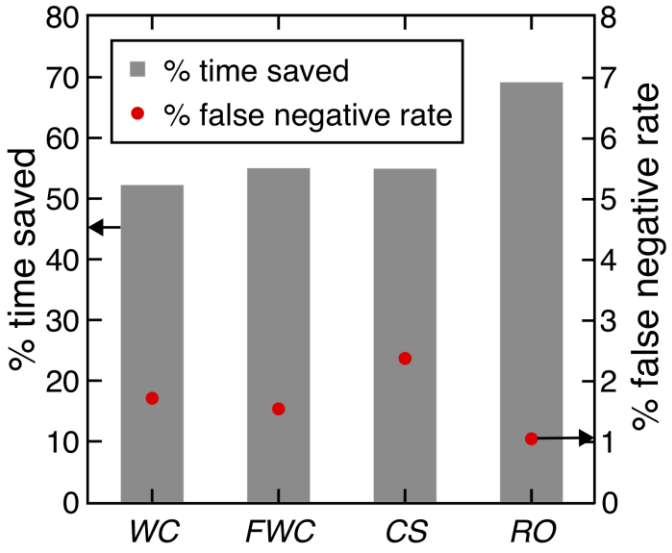
PREDICTING SUCCESS IN DISCOVERY

Transferability is key in catalysis:

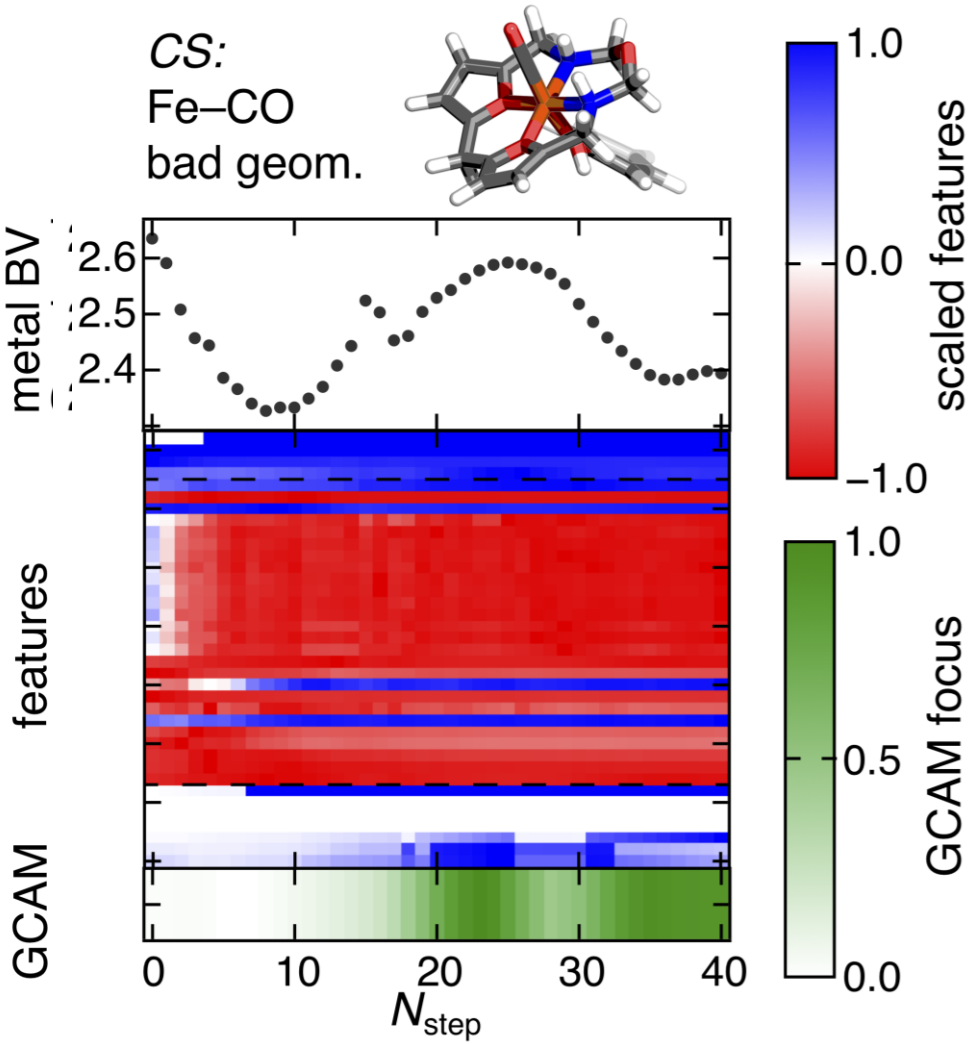








Descriptors make models transferable:

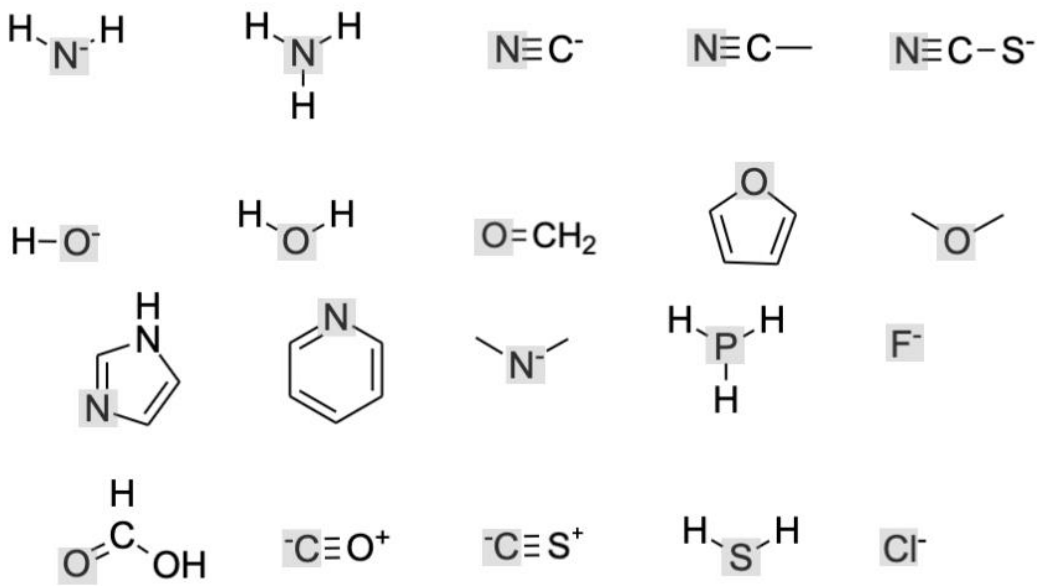


C. Duan, A. Nandy, H. Adamji, Y. Roman-Leshkov, and **HJK** *J. Chem. Theory Comput.* (2022).

SPIN SPLITTING LACKS A UNIVERSAL DFA

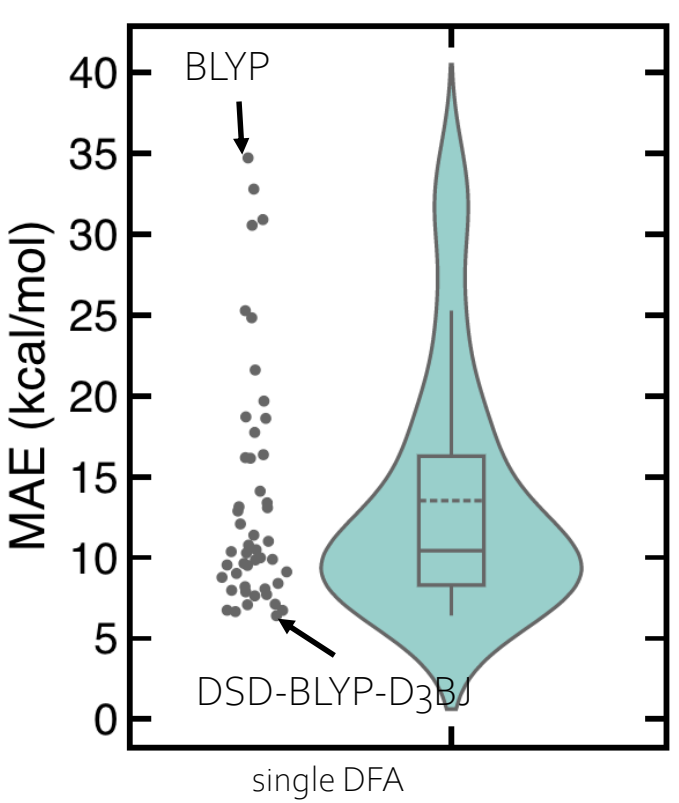
452 octahedral monodentate TMCs (VSS-452)
Property of interest: vertical spin splitting energy
Reference: DLPNO-CCSD(T)/def2-TZVP

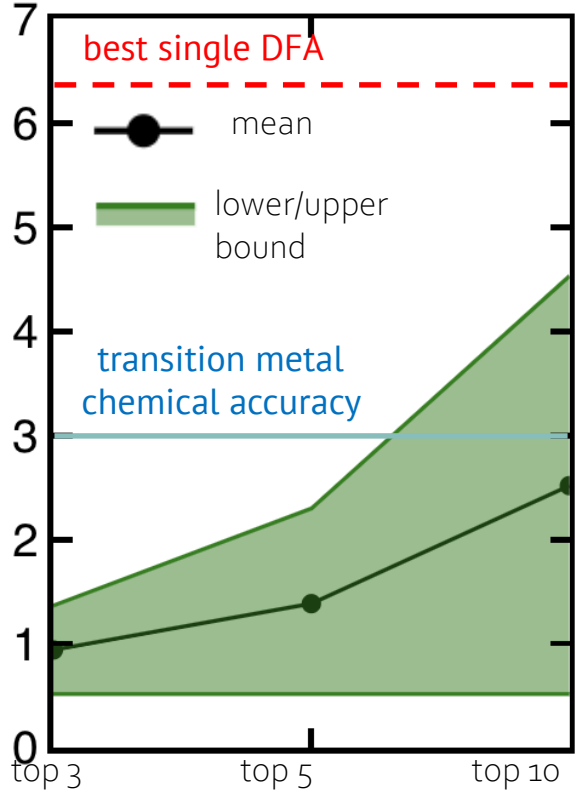
24	25	26	27
Cr	**Mn**	**Fe**	**Co**
Chromium	Manganese	Iron	Cobalt
51.996	54.938	55.845	58.933



Pool of DFAs:

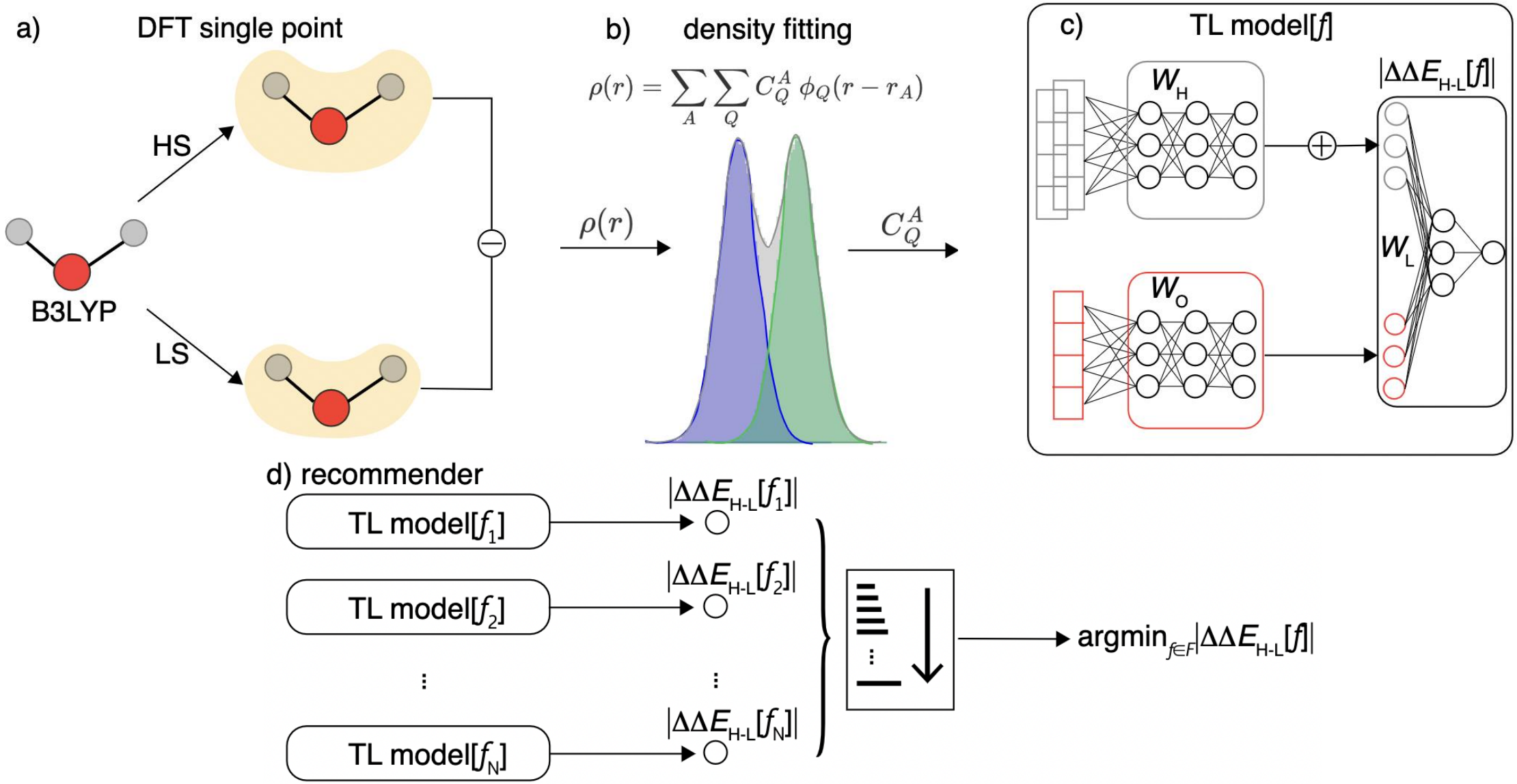
1) 23 DFAs from previous work[1] that evenly spans multiple rungs of the Jacob's ladder (GGA to double hybrid)
2) HF sampling (10% to 50%) of 5 representative GGA and meta-GGAs (BLYP, PBE, SCAN, M06, and MN15) – 48 functionals in total





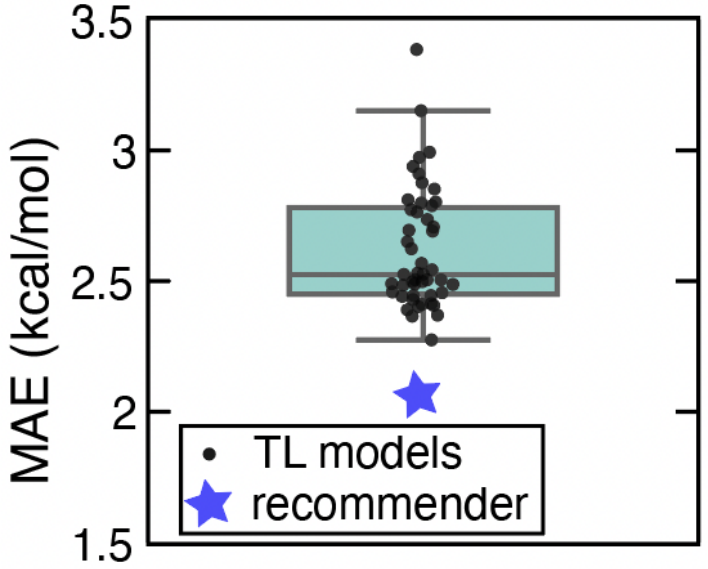
[1]C. Duan, S. Chen, M. G. Taylor, F. Liu, and HJK *Chem. Sci.* (2021). C. Duan, A. Nandy, and **HJK** ICML 2022 AI4Science Workshop; C. Duan et al., *Nat. Comput. Sci.* (2023).

DESIGNING A DFA RECOMMENDER

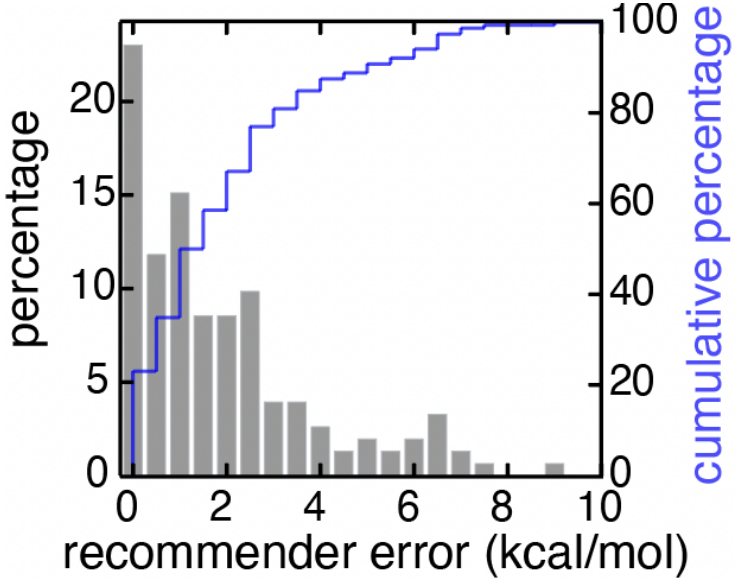


C. Duan, A. Nandy, and **HJK** ICML 2022 AI4Science Workshop; C. Duan et al., *Nat. Comput. Sci.* (2023).

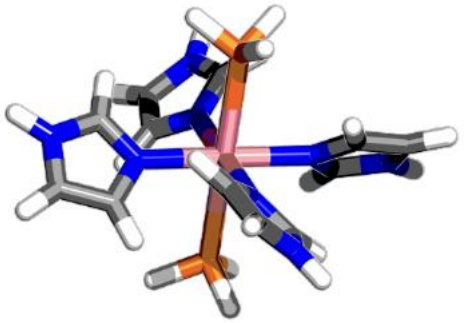
A RECOMMENDER BEATS TRANSFER LEARNING



TL models: 2.3-3.4 kcal/mol
recommender: 2.1 kcal/mol



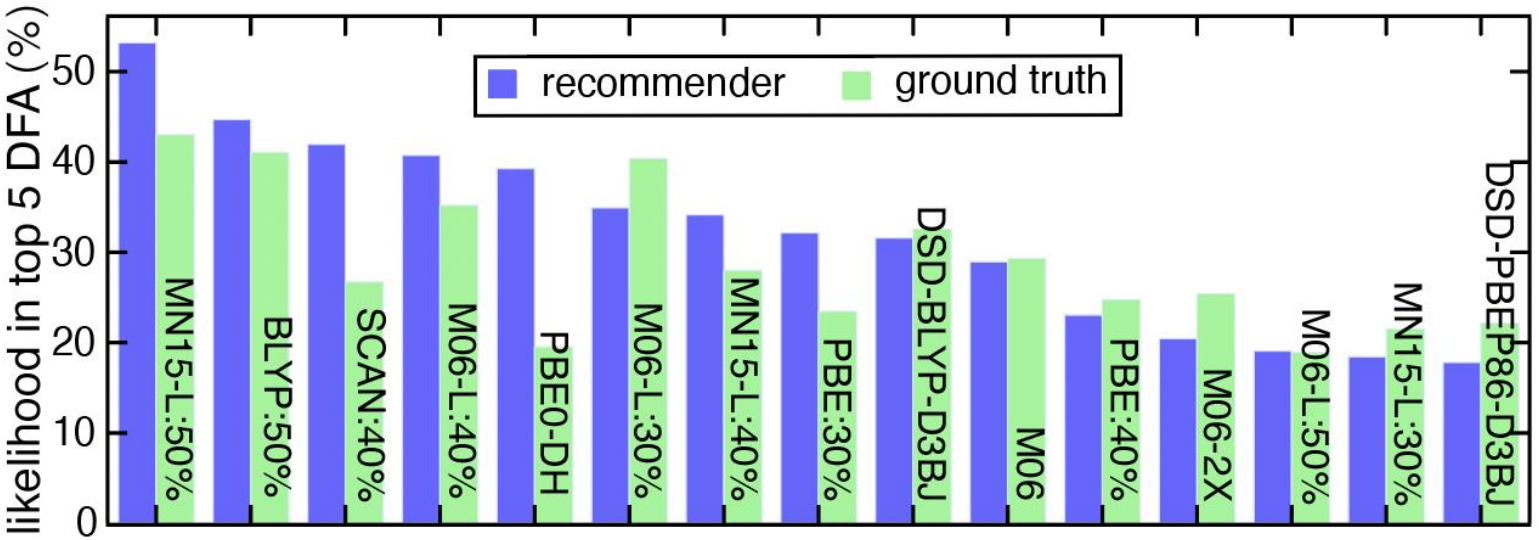
77% of complexes have errors < 3 kcal/mol



$Co(III)(C_3H_4N_2)_4(PH_3)_2$

recommend	$\lvert\Delta\Delta E_{H\text{-}L}\rvert$
ωB97X	0.1 kcal/mol

DSD-BLYP-D3BJ: 9.2 kcal/mol



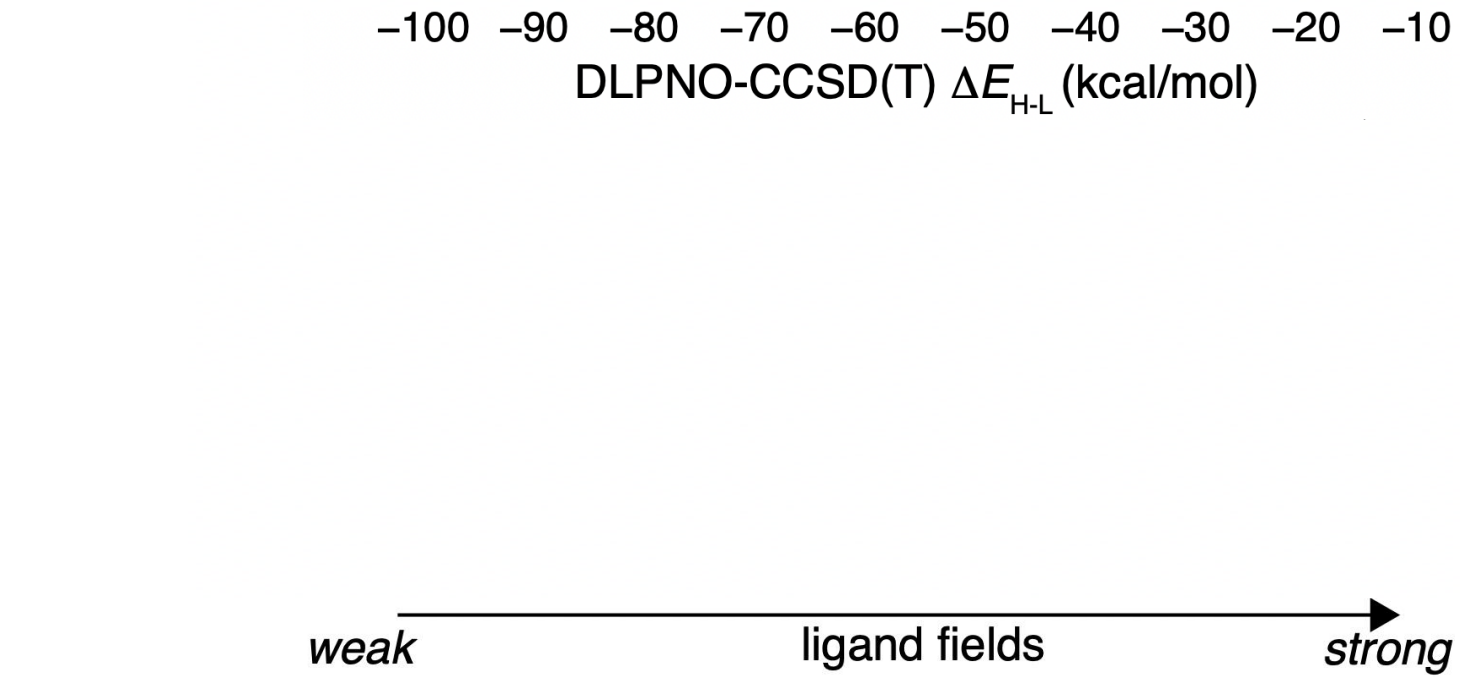
For the top-5 DFA, the recommender gives

- Similar statistics on the likelihood
- 0.95 rank ordering coefficient compared to the ground truth

WHY THE RECOMMENDER WORKS

PBE:30% SCAN:40% M06-L:40% MN15-L:50% BLYP:50%

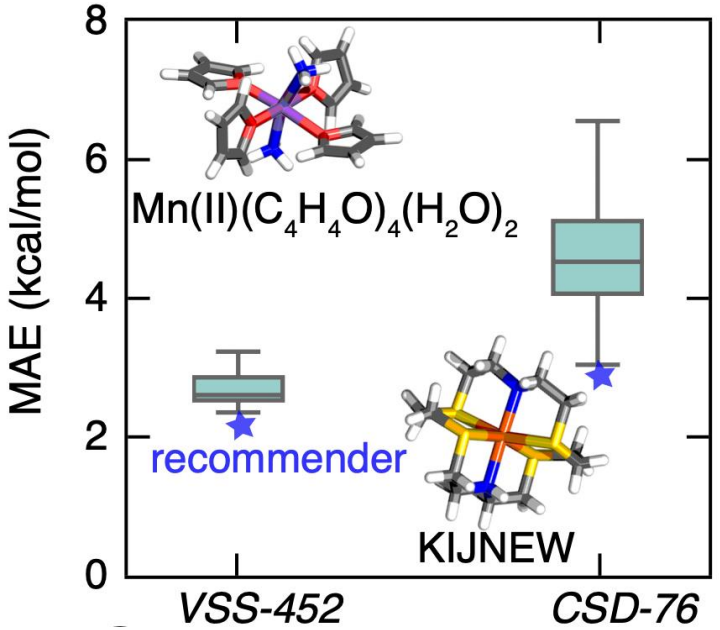
→ 3 kcal/mol, very competitive

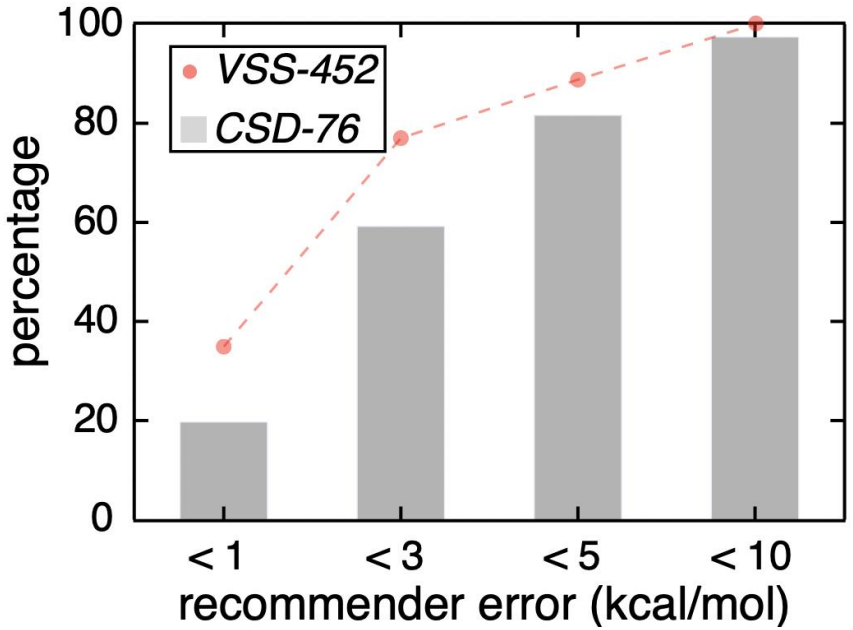


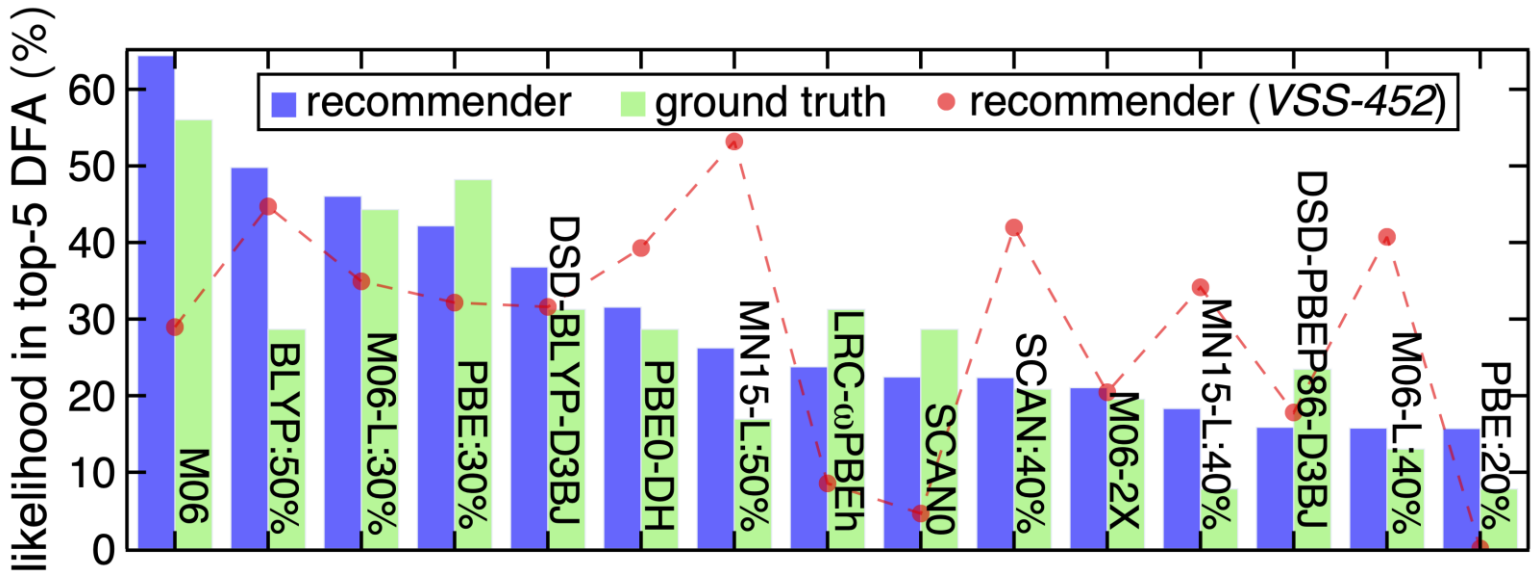
THE RECOMMENDER IS TRANSFERABLE

Diverse and unseen ligand chemistry and connectivity in Cambridge Structural Database for CSD-76

- MAE lower than the best TL model
- Still 60% of the CSD complexes < 3 kcal/mol err.





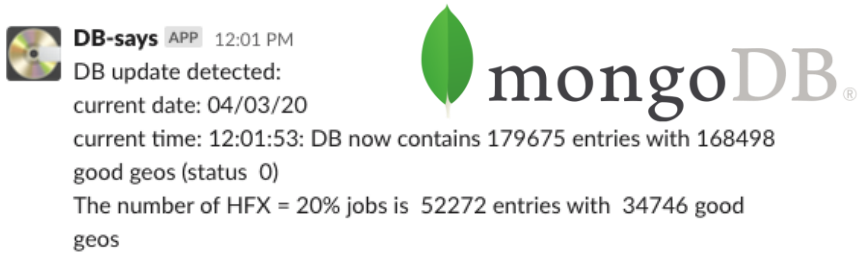


CSD complexes favors a distinct set of DFAs compared to VSS-452

The recommender is still able to capture the DFAs that are most likely to be in top 5

SHEDDING NEW LIGHT ON OLD DATA

mAD centralizes all runs in a database:

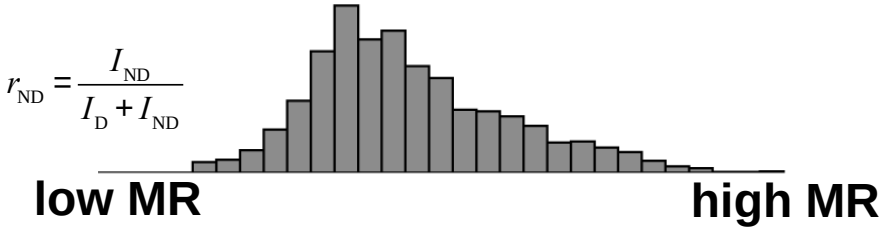


180k TMCs and counting…

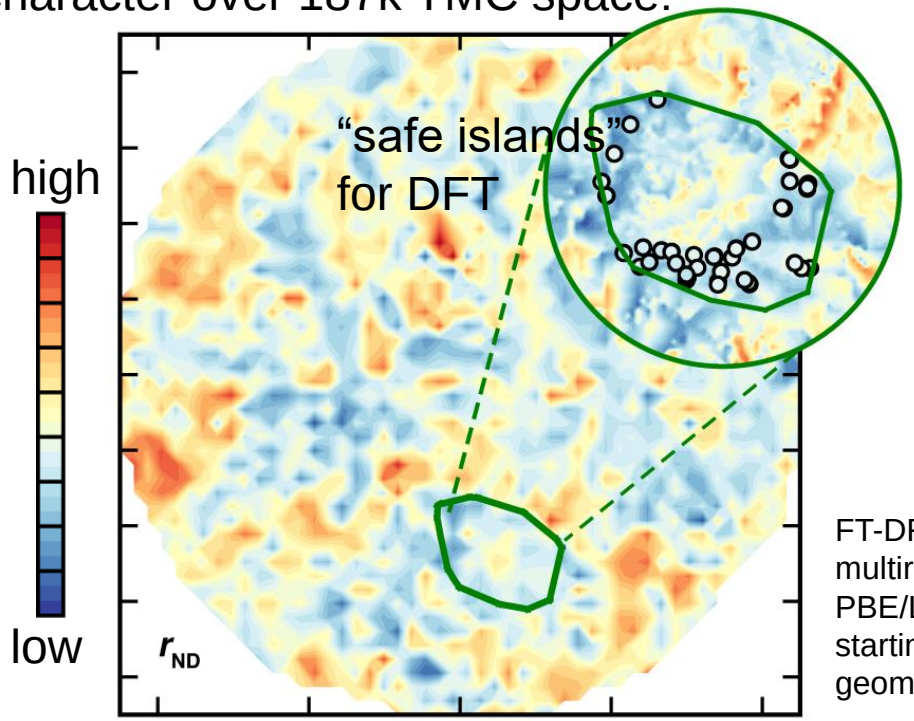
Some questions we should ask:
Where shouldn't we have used DFT?
Where is multi-reference character highest?

F. Liu, C. Duan, and **HJK**, *J. Phys. Chem. Lett.* (2020).

There is a range! (FT-DFT r_{ND} over 5k TMCs):



Train on this data and use RACs/ANN to predict MR character over 187k TMC space:

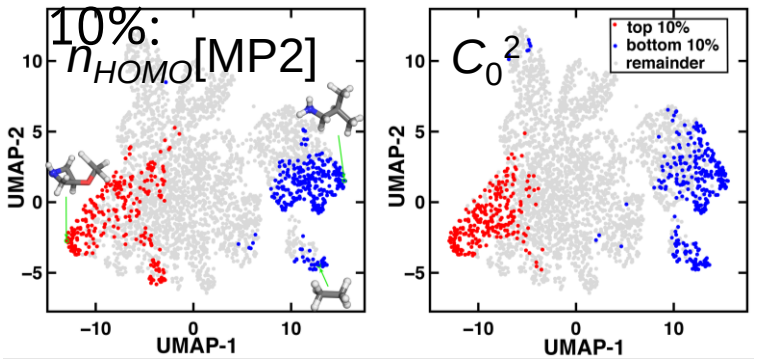


FT-DFT automated with multirefpredict using B3LYP or PBE/LACVP* in TeraChem from starting wavefunctions and geometries.

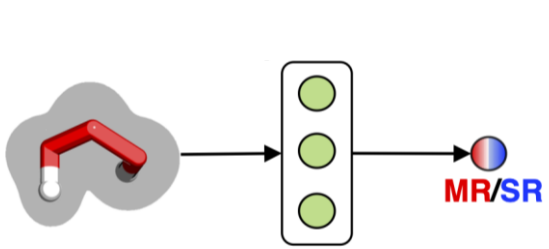
TACKLING ELECTRONIC STRUCTURE CHALLENGES

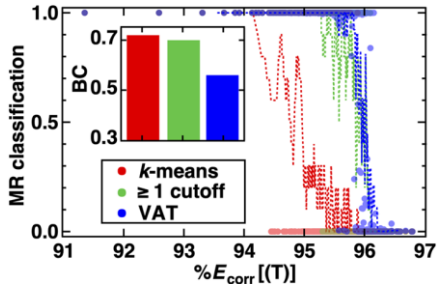
detecting MR character

Semi-supervised learning: label top/bottom



Dr. Fang Liu
PD '20, now Asst. Prof. Emory

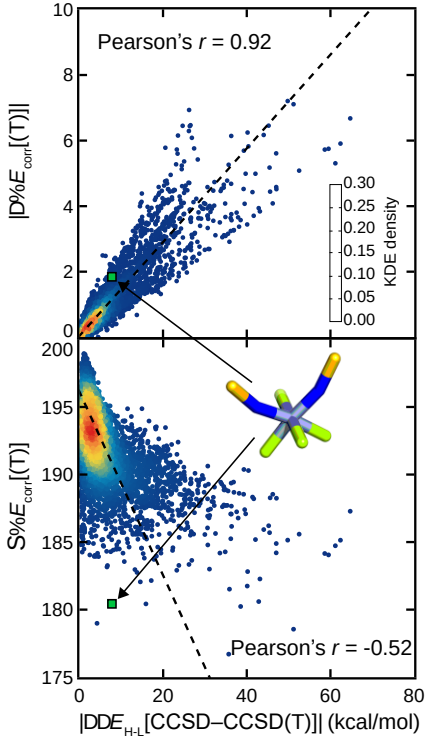


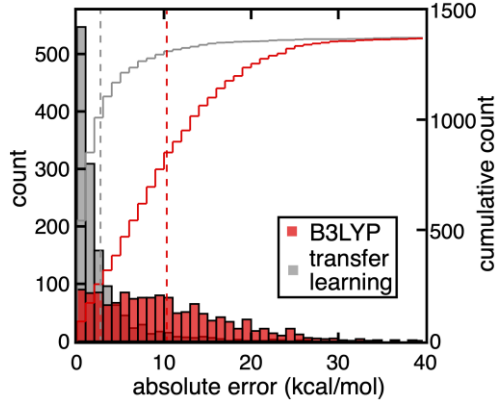


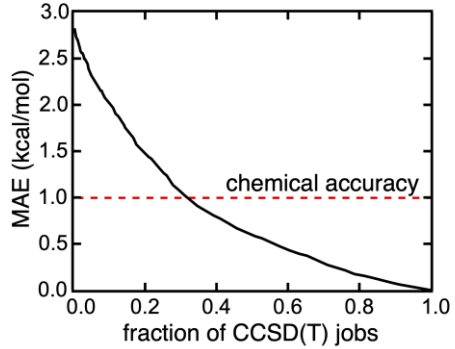
C. Duan, F. Liu, A. Nandy, and H. J. Kulik, *J. Chem. Theory Comput.* (2020); C. Duan, F. Liu, A. Nandy, and H. J. Kulik, *J. Phys. Chem. Lett.* (2020); F. Liu, C. Duan, and H. J. Kulik, *J. Phys. Chem. Lett.* (2020).

correcting for MR effect

Difference matters most and can be learned:



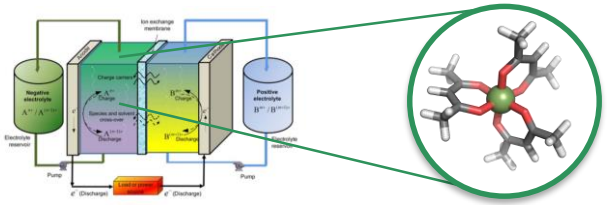




C. Duan, D. B. K. Chu, A. Nandy, and H. J. Kulik *Chem. Sci.* (2022)

TACKLING REAL WORLD DESIGN CHALLENGES

Redox flow battery redox couple design:



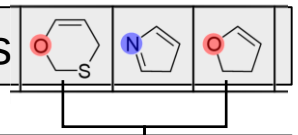
$$E_{cell} = 0.5 \times V_{cell} \times C \times n \times F$$

We must address

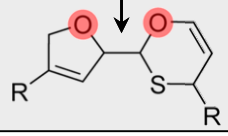
- Stability/resistance to crossover
- Concentration (solubility)
- Redox potential

candidate pool | uncertainty quantification | optimization algorithm

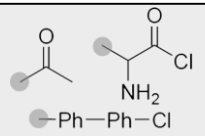
38 heterocycles



779 core ligands



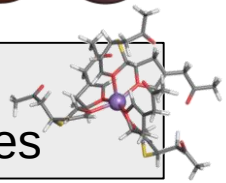
897 functional groups

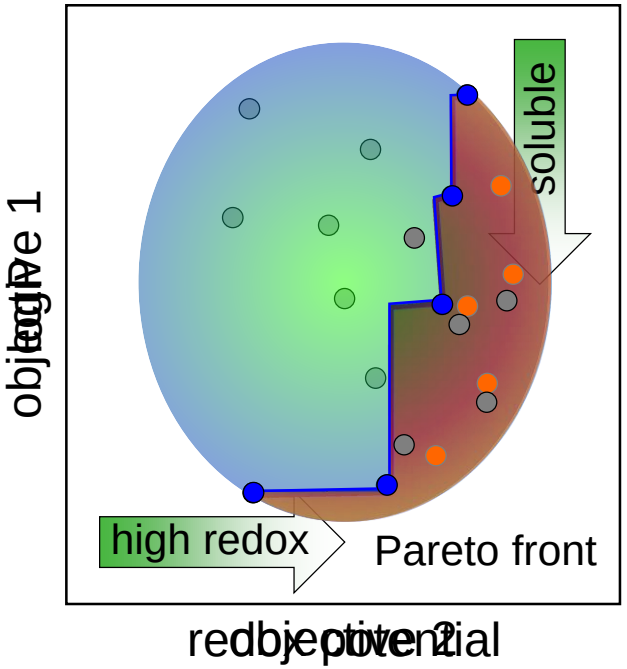


4 metals Cr Mn Fe Co

=

2.8 M bulky transition metal complexes





objective 1

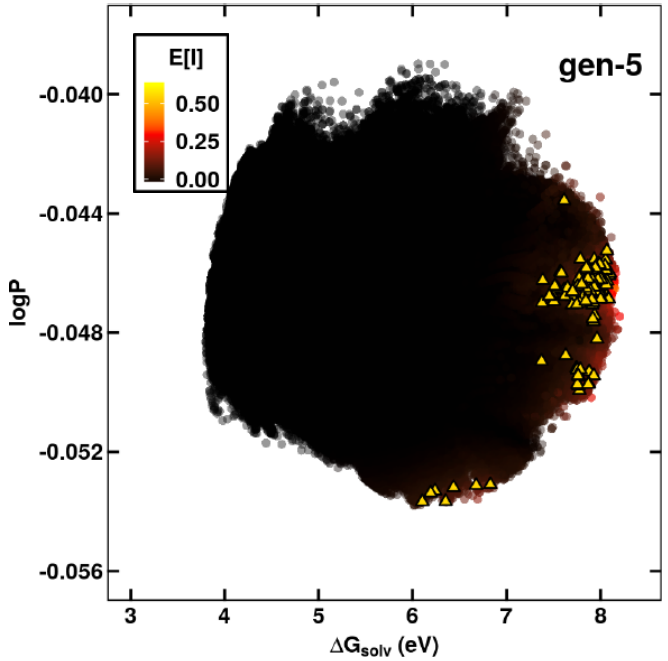
objective 2

DFT data → train ANN → 2D expected improvement → candidates

M. L. Perry and A. Z. Weber. *J. Electrochem. Soc.* (2016); J. P. Janet, S. Ramesh, C. Duan, and **HJK**, *ACS Central Science* (2020).

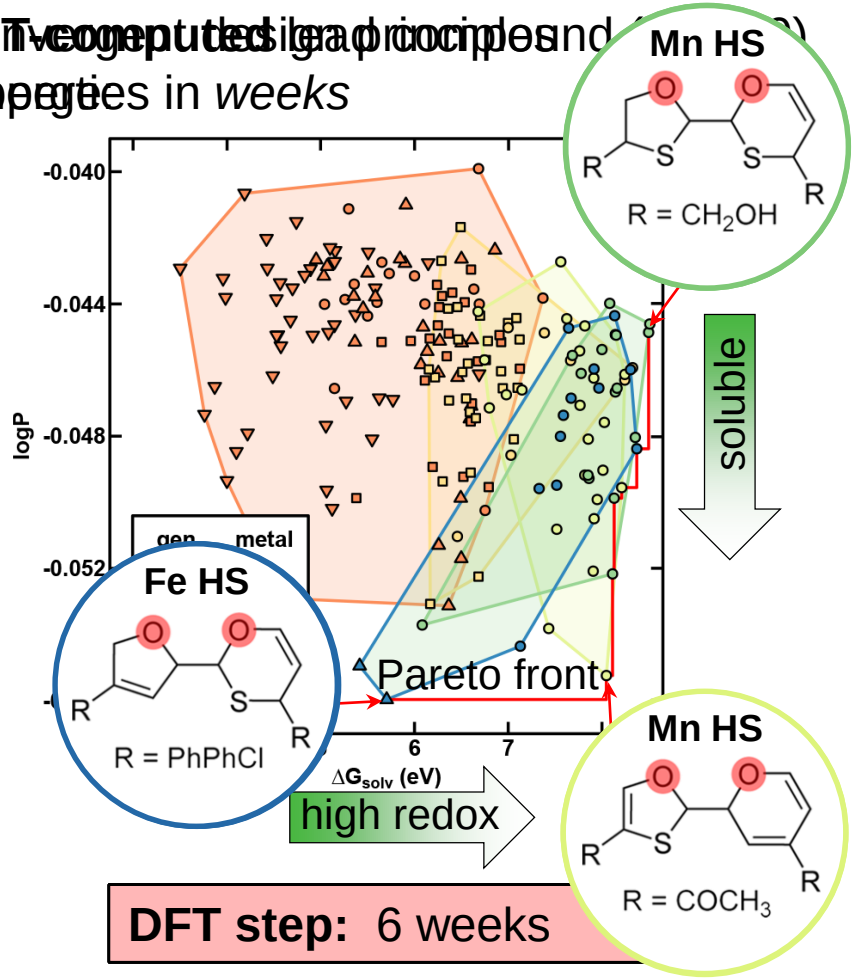
ACCELERATING RFB REDOX COUPLE DESIGN

EI with ANN predictions/UQ of 2.8M complexes in *minutes*



ANN step: 15 minutes

DFT-computed [illegible] properties in *weeks*



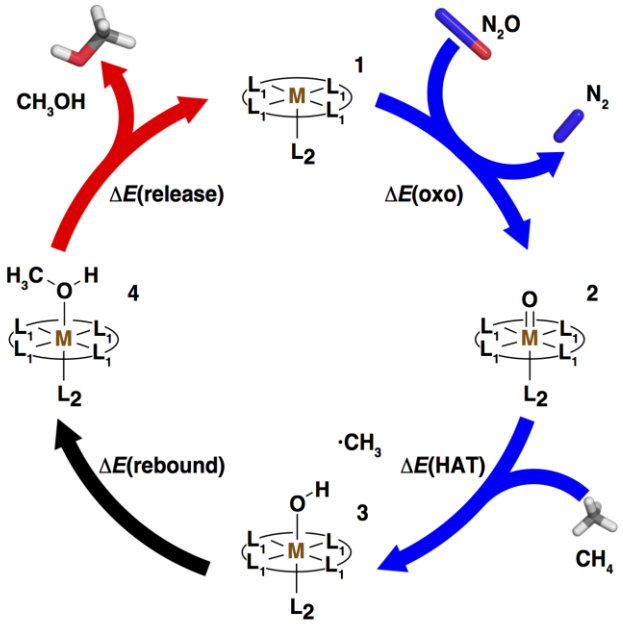
DFT step: 6 weeks

Comparison to random search reveals **at least** 500-fold acceleration:

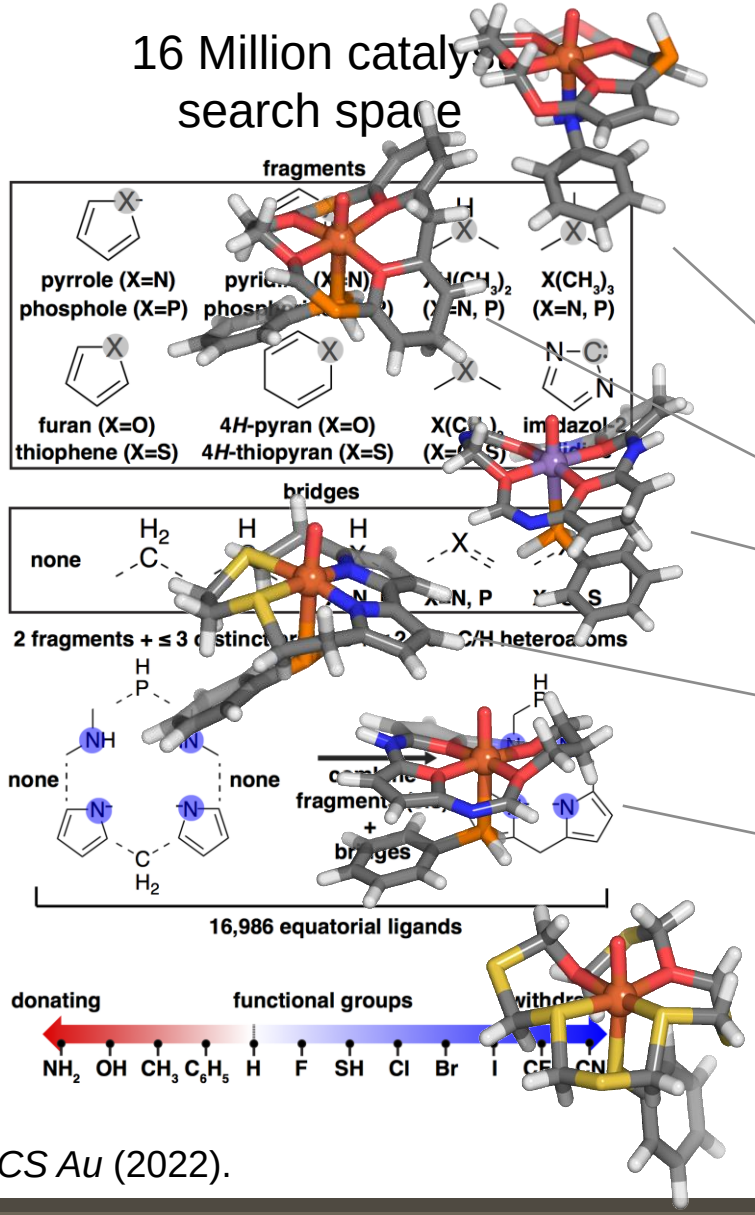
J. P. Janet, S. Ramesh, C. Duan, and **HJK**, *ACS Central Science* (2020).

ACCELERATING CATALYST DISCOVERY

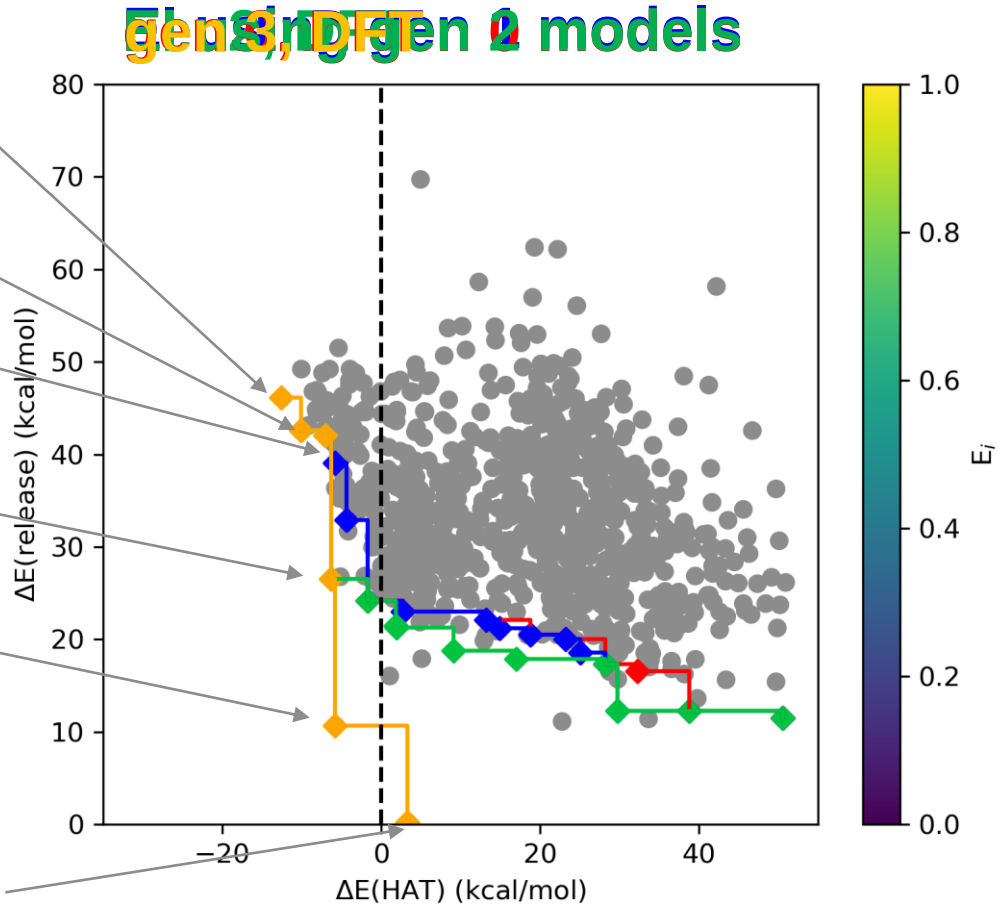
Back to methane to methanol: optimizing HAT and release



16 Million catalyst search space



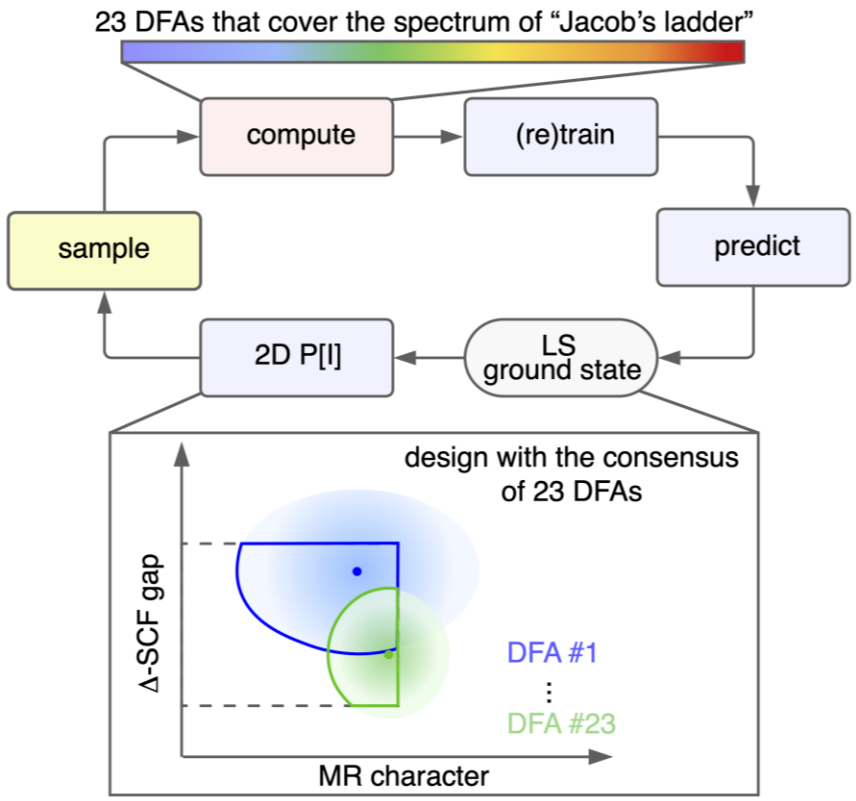
Novel catalysts and design principles



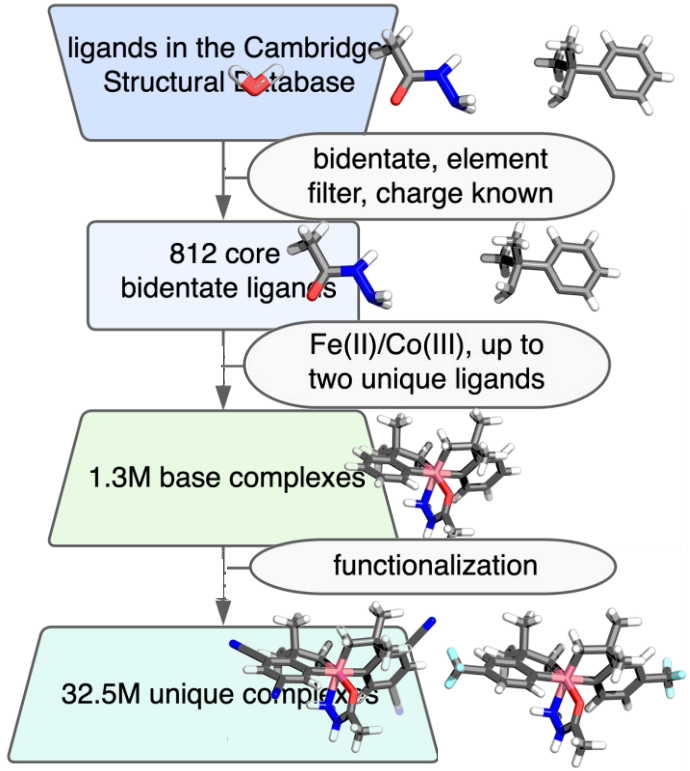
A. Nandy, C. Duan, C. Goffinet, and **HJK** *JACS Au* (2022).

DISCOVERING LIGHT-HARVESTING COMPLEXES

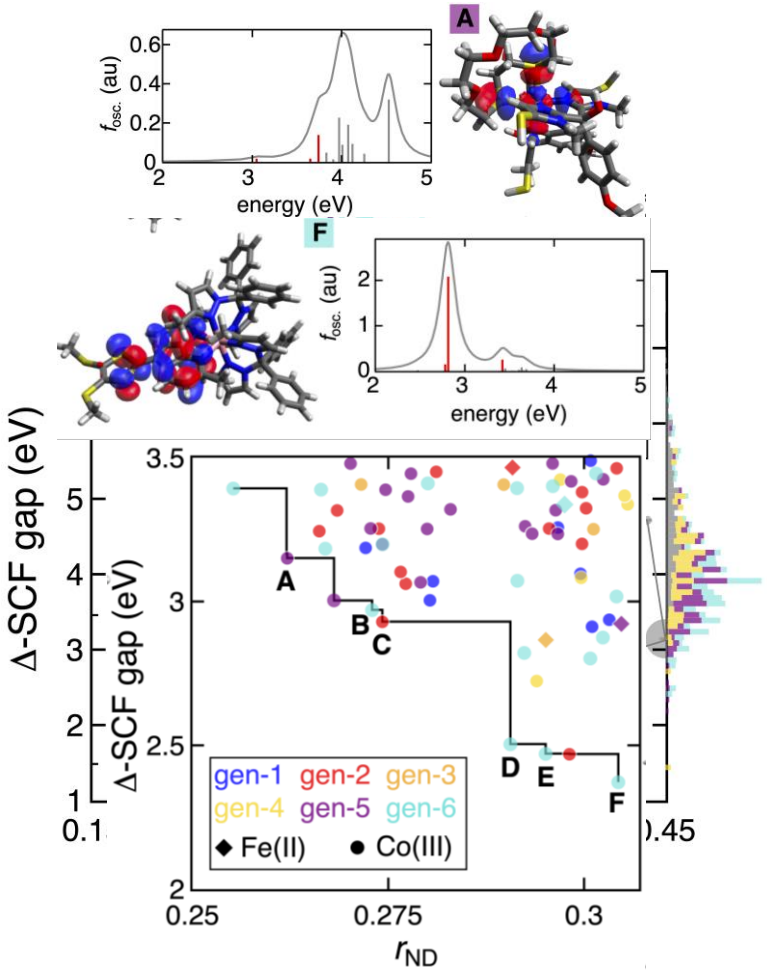
Optimize the HOMO-LUMO gap and minimize DFT model uncertainty



Construct a space of 32.5 M CSD-derived structures



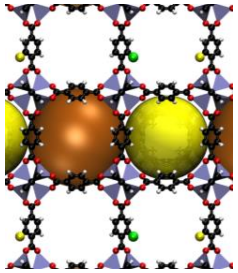
1000x speedup of discovering method-insensitive chromophores



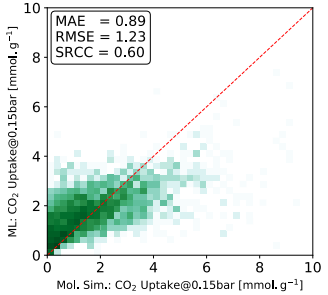
C. Duan, A. Nandy, G. Terrones, D. W. Kastern, and **HJK**, *JACS Au* (2023).

THE NATURAL EXTENSION TO SEPARATIONS

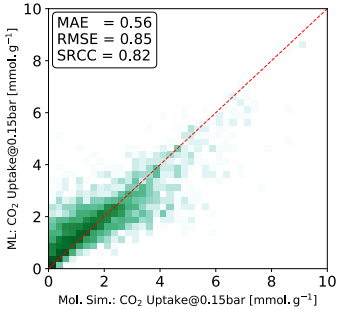
Typical MOF ML with pore geometry features:



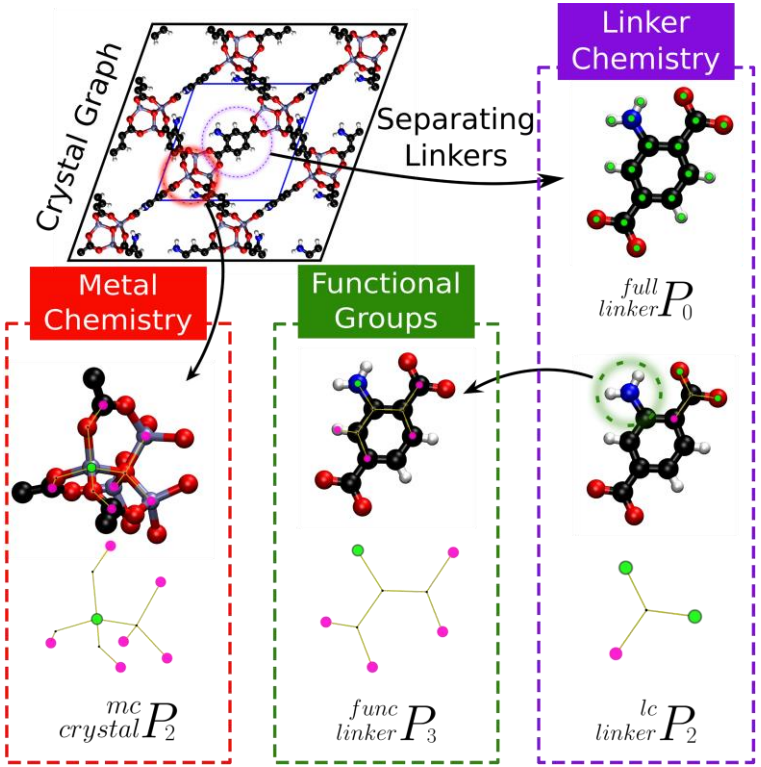
Geometry only random forest:



Adding RACs improves models:

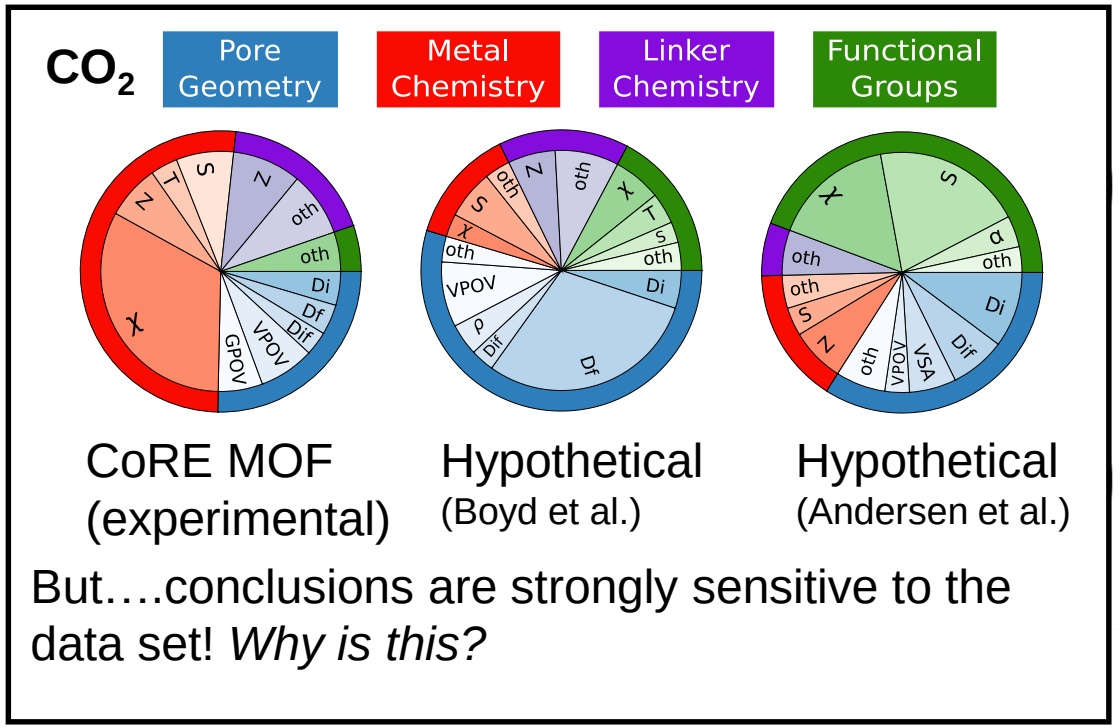


RACs as features for MOFs:



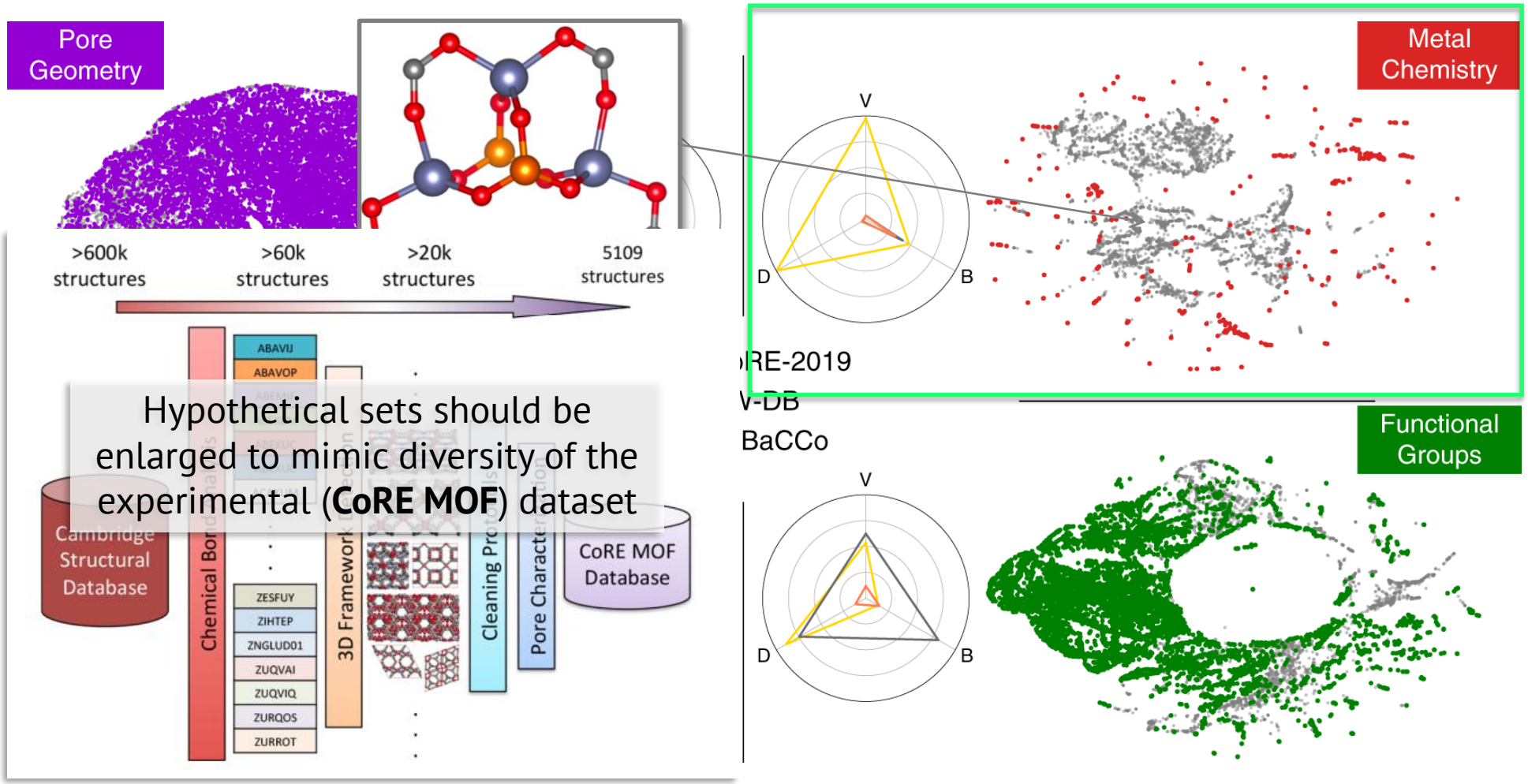
Aditya Nandy
Chemistry Ph.D.

Feature analysis shows when chemistry matters:



Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B. and **HJK**, *Nat. Commun.* (2020).

WHAT'S DIFFERENT ABOUT HYPOTHETICAL MOFS?

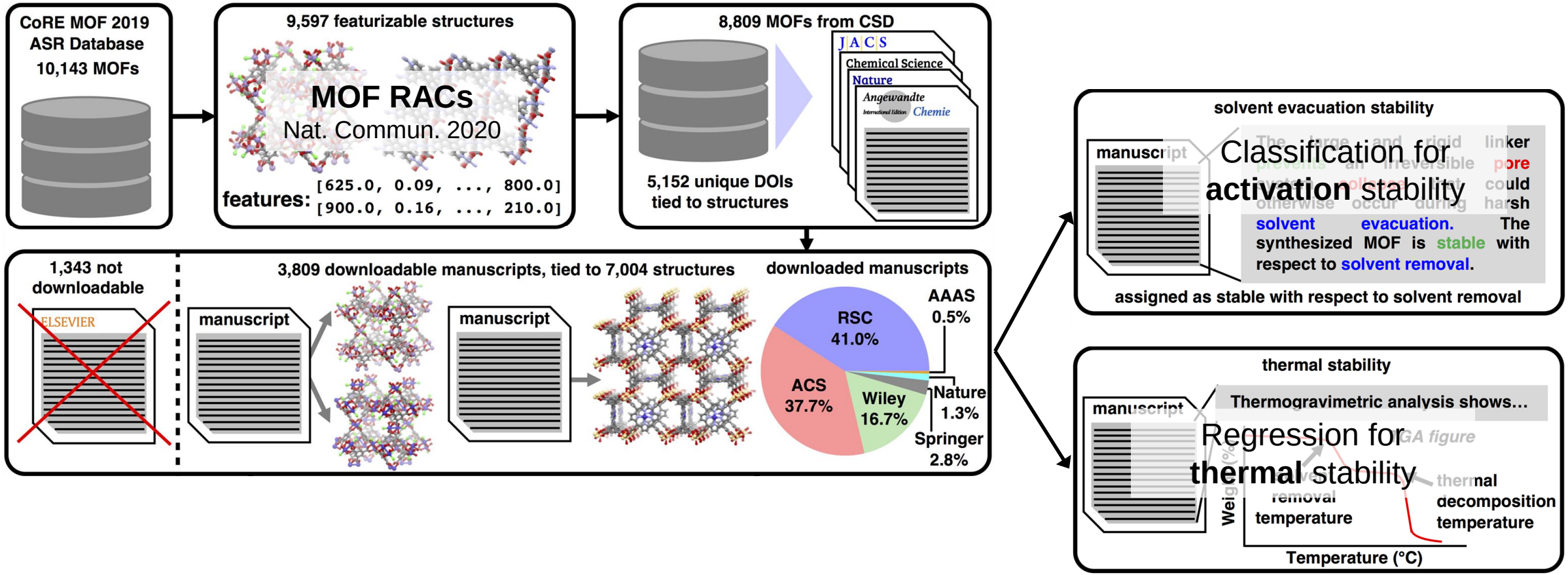


Hypothetical = colored
Expt. only = gray

Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B. and **HJK**, *Nat. Commun.* (2020).

HOW DO WE EXTRACT EXPERT KNOWLEDGE?

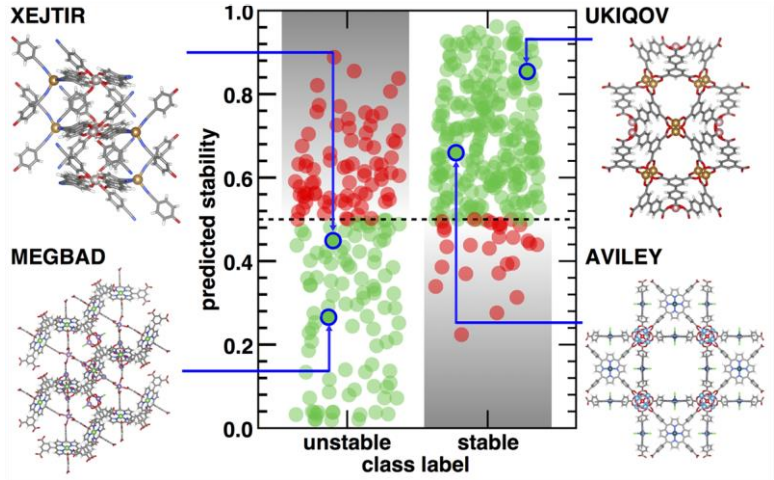
Thousands of MOFs have been experimentally characterized, but a lack of consistent naming and reporting makes it challenging to leverage this knowledge:



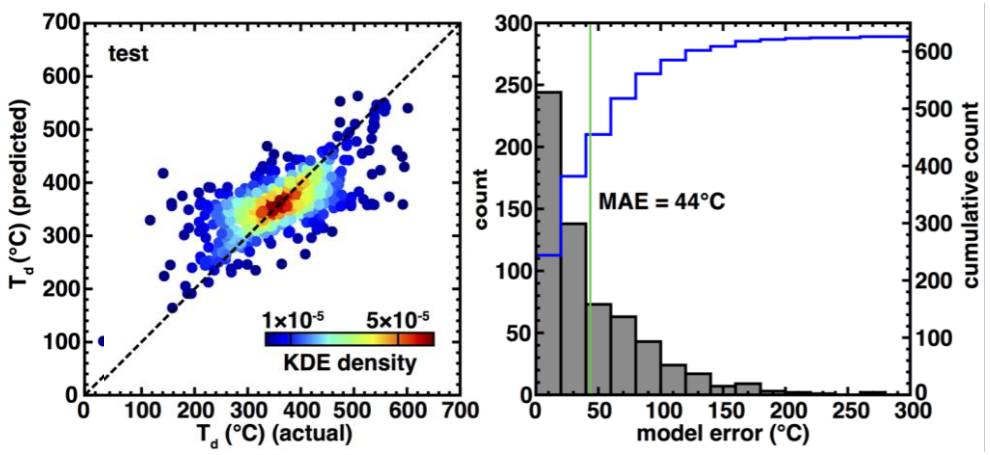
S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, and **HJK** *Nat. Commun.* (2020). A. Nandy, C. Duan, and **HJK**, *J. Am. Chem. Soc.* (2021).

ML TELLS US WHY HEURISTICS FAIL

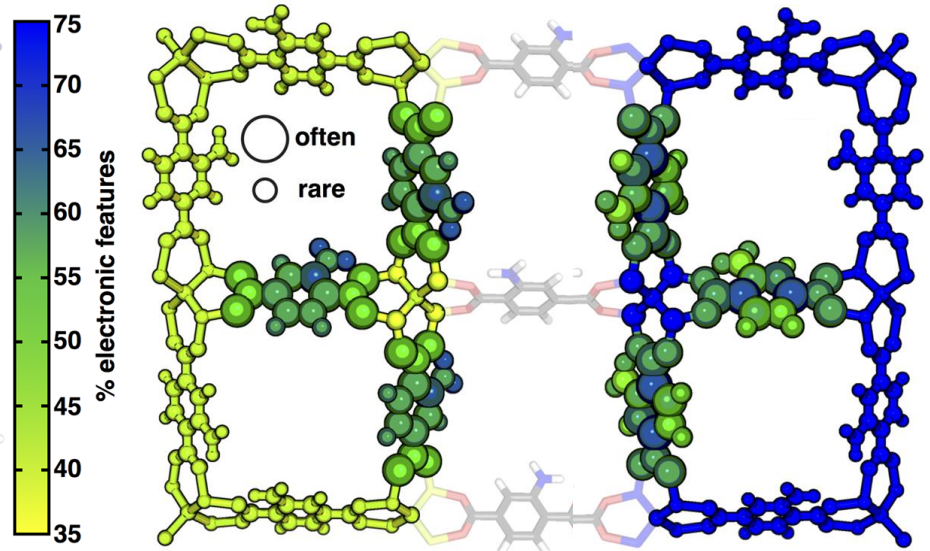
Classification for **activation** stability

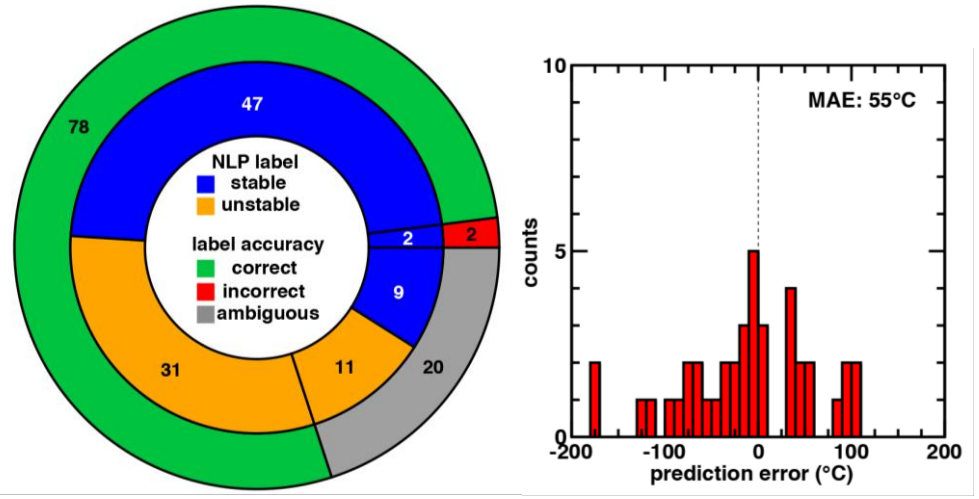


Regression for **thermal** stability



design principles:



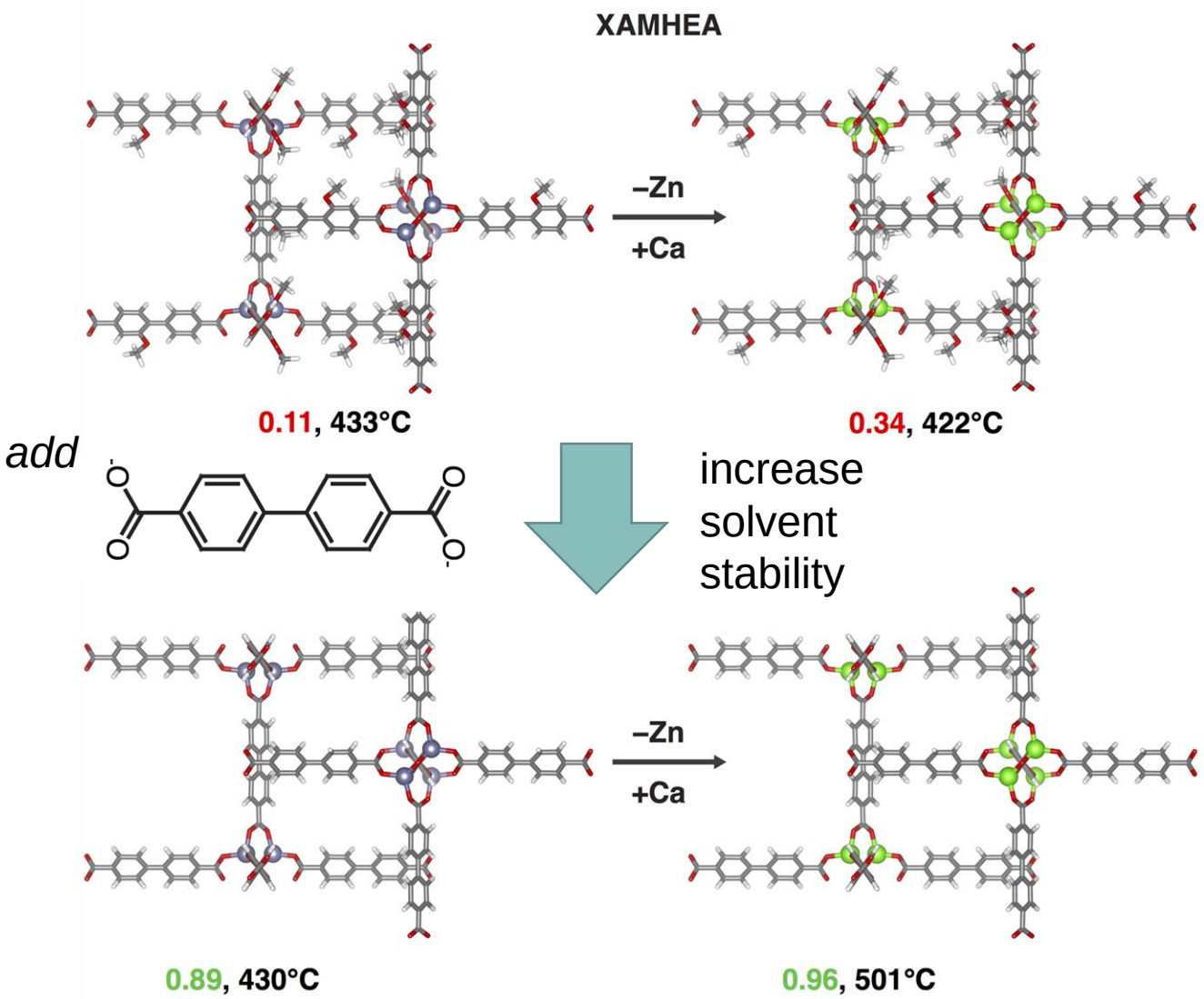


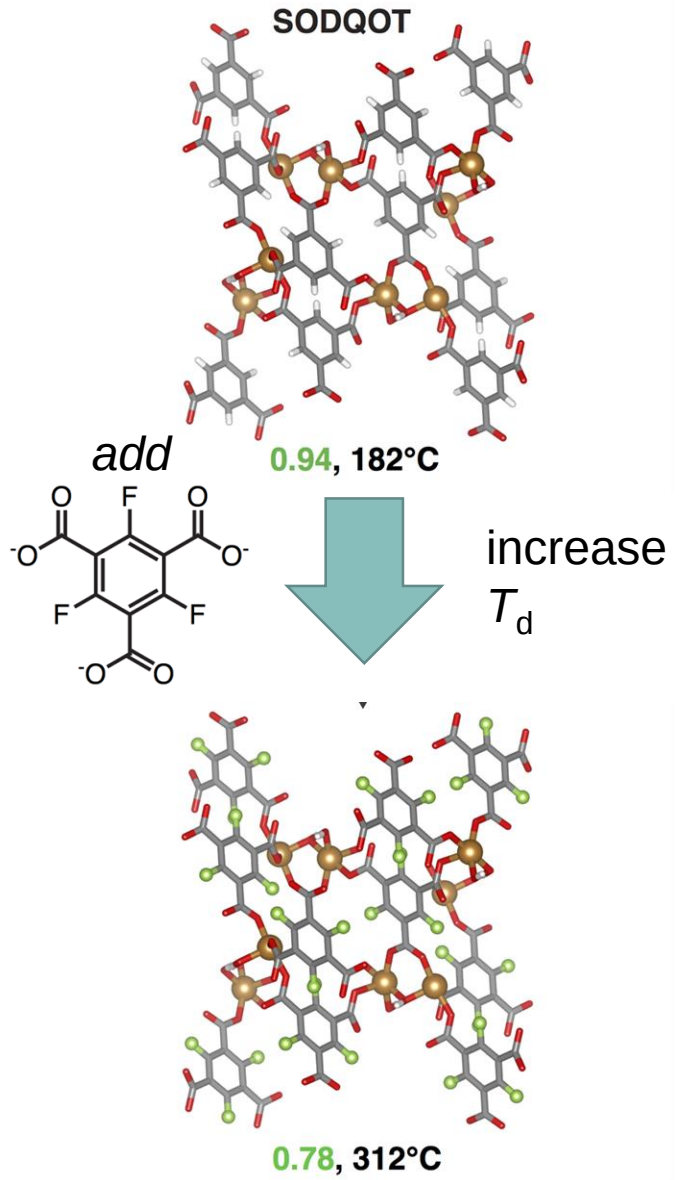
... ML generalizes to unseen data

A. Nandy, C. Duan, and **HJK**, *J. Am. Chem. Soc.* (2021).

ENGINEERING STABLE MOFS

ML models reveal strategies for redesigning MOFs to be more stable:

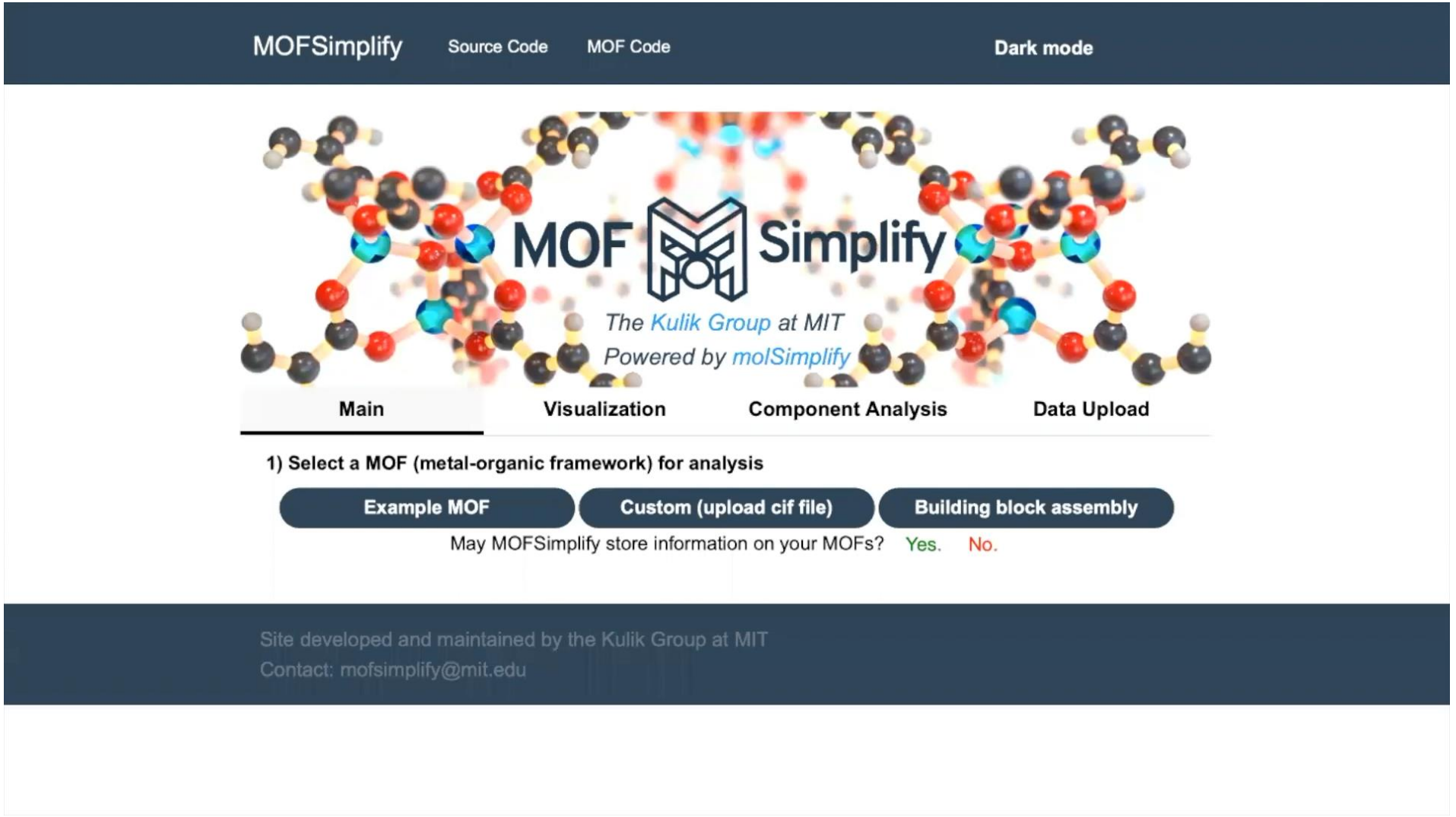




A. Nandy, C. Duan, and **HJK**, *J. Am. Chem. Soc.* (2021).

MOF STABILITY AT YOUR FINGERTIPS

https://mofsimplify.mit.edu

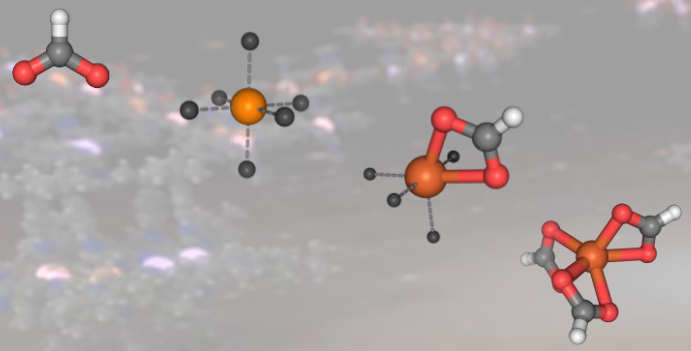


Gianmarco Terrones
ChemE Ph.D.

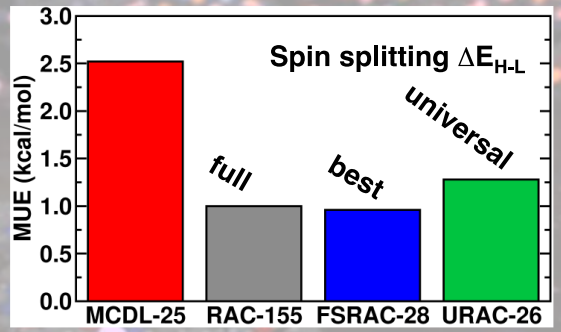
A. Nandy, C. Duan, and **HJK**, *J. Am. Chem. Soc.* (2021); A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, and **HJK**, *Sci. Data* (2022).

OUTLOOK

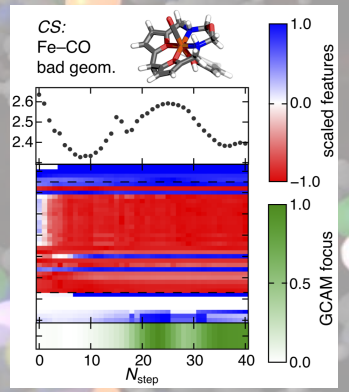
High-throughput inorganic chemistry



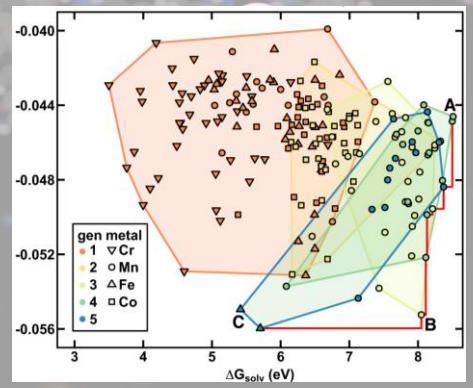
Machine learning property prediction



All models available in molSimplify:
https://molsimplify.mit.edu
also on Conda and Github



Autonomous tools



Accelerating discovery

Recent perspective: J.P. Janet, C. Duan, A. Nandy, F. Liu, and **HJK**, *Acc. Chem. Res.* (2021).

ACKNOWLEDGMENTS

On the web: http://hjkgrp.mit.edu
Group news: @KulikGroup on Twitter

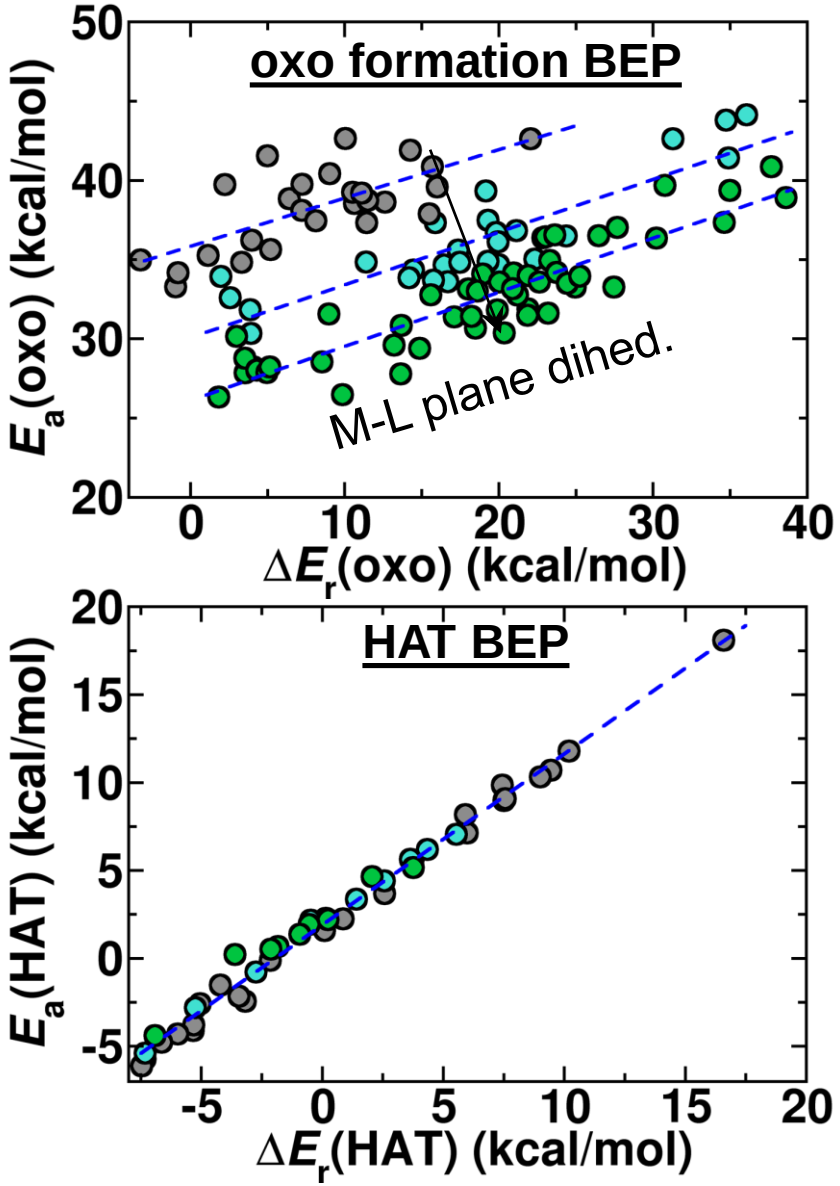
Chenru Duan

Aditya Nandy

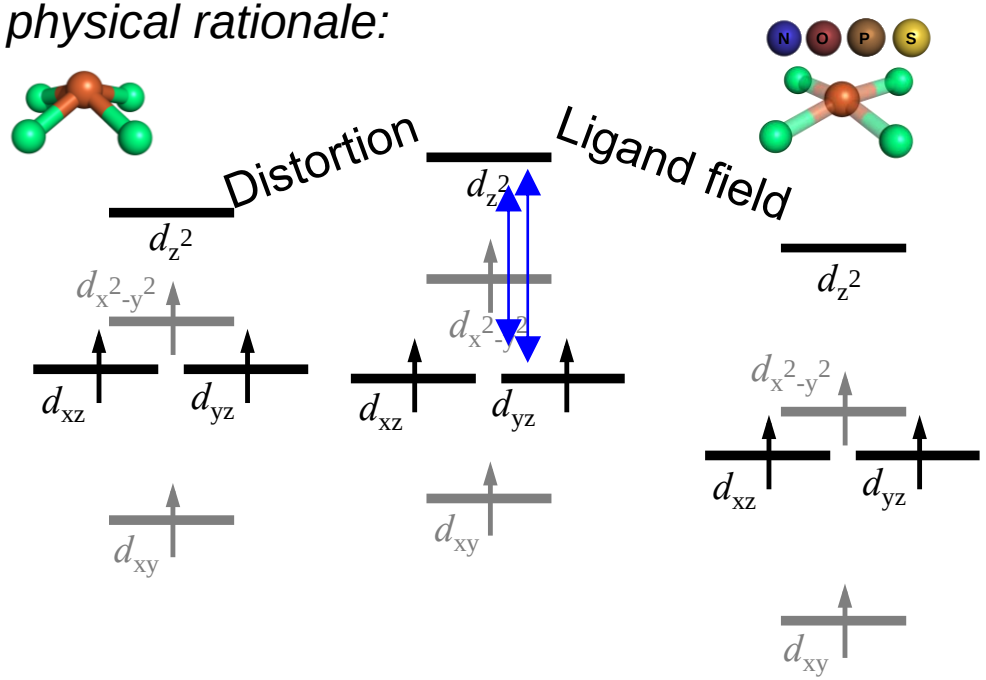
And… Yeongsu Cho, Ralf Meyer, Jacob Toney, Melissa Manetsch, Akash Ball, Roland St. Michel
Alumni: Daniel Harper, Zhongyue Yang, Fang Liu (now TT AP Emory), Jing, Yang, Mengyi Wang, J. P. Janet (Ph.D. ‘19), Jit Ramesh (MS Oxford), S. Mohamad Moosavi (EPFL, Smit group), Stefan Gugler (MS ETHZ, Reiher group), Helena W. Qi (Ph.D. ’19), Qing Zhao (Ph.D. ’18, now TT AP Northeastern), Dr. Efthymios I. Ioannidis (Ph.D. ’16), Ms. Lydia Chan (Troy H.S. ’18)

Thanks for listening! …Any questions?

BREAKING AND EXPLOITING SCALING RELATIONS

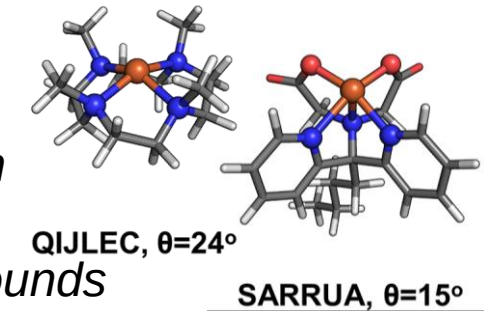


That was thermo. What about kinetics?
A physical rationale:



Engineering distortion in real materials:

Known (CSD) compounds

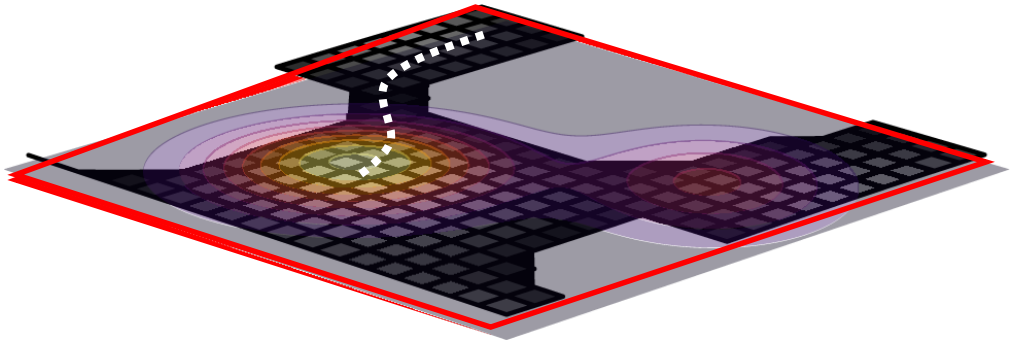


Or post-synthetic modification of metal sites

T. Z. H. Gani and **HJK**, *ACS Catalysis* (2018).

EXPLOITING ANNS FOR CHEMICAL DISCOVERY

With an ANN, we can score in seconds but must be aware of model uncertainty to **exploit** fruitful predictions:

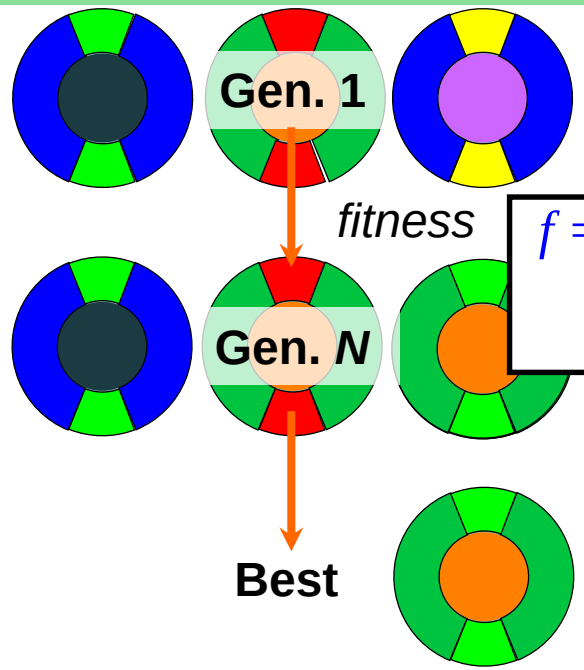


candidate pool

uncertainty quantification

optimization algorithm

optimization algorithm



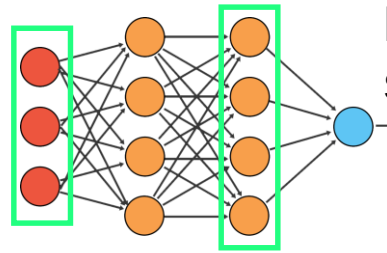
optimize 1000s of complexes, e.g., with genetic algorithm

fitness

$$f = \exp(-(P_{ANN}-P_{target})^2) * \exp(-(d_{data})^2)$$

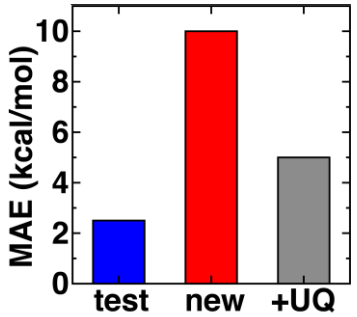
uncertainty quantification

feature space

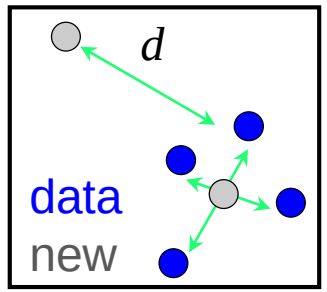


latent space

performance



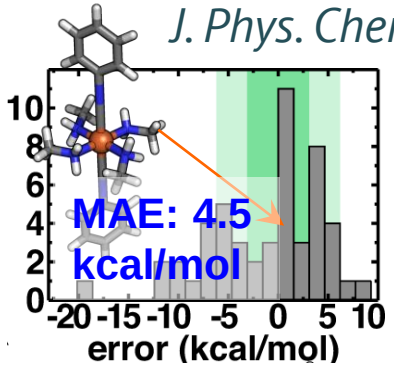
distance UQ

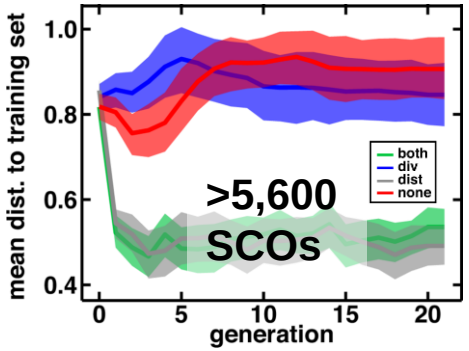


J. P. Janet, L. Chan, and **HJK**, *J. Phys. Chem. Lett.* (2018); J. P. Janet, C. Duan, T. Yang, A. Nandy, and **HJK** *Chem. Sci.* (2019).

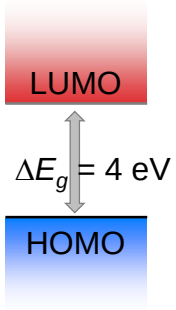
ML FOR INORGANIC DISCOVERY

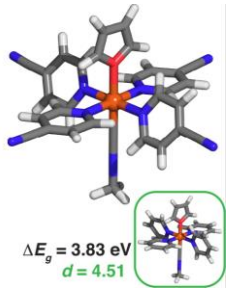
J. P. Janet, L. Chan, and H.J. Kulik
J. Phys. Chem. Lett. (2018).





A. Nandy, C. Duan, J. P. Janet, S. Gugler, and H.J. Kulik *Ind. Eng. Chem. Res.* (2018).



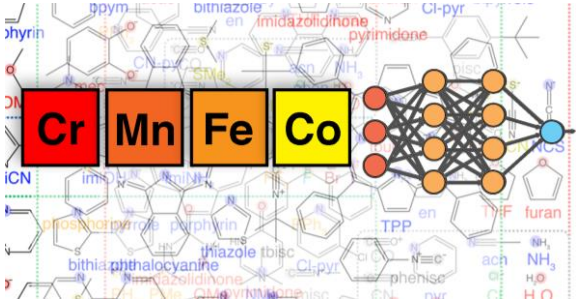


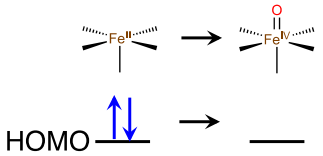
14,400 complexes

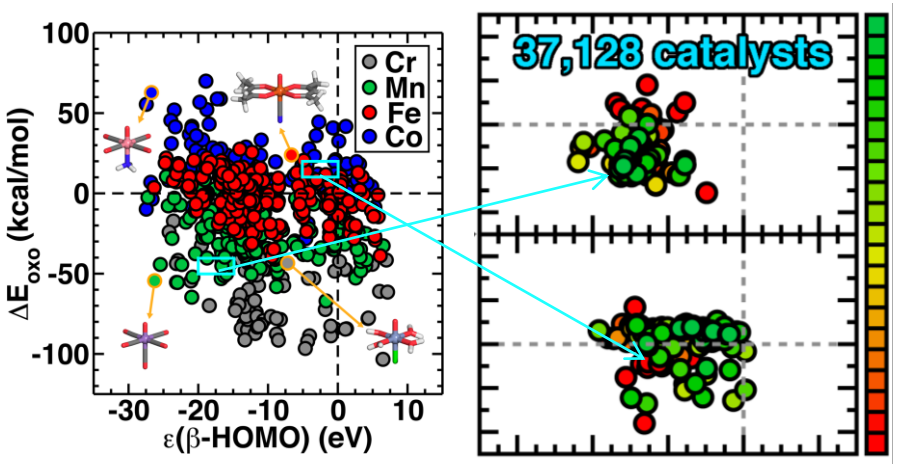
ANN	4.00 eV
DFT	3.97 eV

functional materials

A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman, and H.J. Kulik
ACS Catal. (2019).







catalysis