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

Heather J. Kulik
Associate Professor
Chemistry & Chem. Eng., MIT
March 27, 2023

IPAM
WHERE ARE THE STARS IN CHEMICAL SPACE?

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

- Catalysis
- Energy storage
- Separations
Challenges for open shell transition metal chemistry:

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)

$$P_d = \sum_{i \neq j} P_i P_j \left( d_{ij}, d \right)$$

$$P_{d} = \sum_{i \neq j} P_i P_j \left( d_{ij}, d \right)$$

$P$: nuc. charge, $Z$ electoneg., $\chi$ cov. rad., $S$ topol., $T$ ident., $I$

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

DOES IT MATTER WHICH XC WE CHOOSE?

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

Intermediate field ligands have greatest disagreement

 xc agreement is sensitive to property

Pearson’s $r = 0.81$

Pearson’s $r = 0.99$

Machine learning feature selection provides a path to universal design rules

Also applies to varying data set size and basis set!
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.

UNCERTAINTY IN DISCOVERY

Uncertainty from:
• method choice
• ML model
• calculation outcome

DFT xc choice uncertainty via ML

20% HF
5% HF

Spin splitting


ML model uncertainty quantification


Uncertainty from calculation outcome


Sources of uncertainty
AUTONOMOUS COMPUTATIONAL CHEMISTRY

computational chemist

Choose and build molecule to study

Choose calculation method and parameters

Run and check calculation for completion

Analyze final result or repeat

Consult and enrich expert knowledge

Employ and enrich data driven models

Automated selection and building of molecules

Autonomous method/parameter decision engine

Automated run monitored by decision engine

Automated data analysis or repeat

PREDICTING SUCCESS IN DISCOVERY

Defining calculation success:

What we want:  
What we get:

“good”  
“bad”

RAC/ANN classifier: 88% accurate

P | N
---|---
283 | 48
(304) | (74)

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

Transferability poor for diverse chemistry:

<table>
<thead>
<tr>
<th>data</th>
<th>diverse (100%)</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>61%</td>
<td></td>
</tr>
</tbody>
</table>

In situ calculation monitoring with new features:

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

Transferability poor for diverse chemistry:

Transferability is key in catalysis:

Descriptors make models transferable:

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

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

---

DESIGNING A DFA RECOMMENDER

A RECOMMENDER BEATS TRANSFER LEARNING

For the top-5 DFA, the recommender gives
• Similar statistics on the likelihood
• 0.95 rank ordering coefficient compared to the ground truth

DSD-BLYP-D3BJ: 9.2 kcal/mol

Co(III)(C₃H₄N₄)₄(PH₃)₂

ωB97X | ∆∆E_{H-L} | 0.1 kcal/mol
WHY THE RECOMMENDER WORKS

- DFA recommender
- → 3 kcal/mol, very competitive

DLPNO-CCSD(T) $\Delta E_{H-L}$ (kcal/mol)

weak ligand fields strong
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

DFA recommender

**Detecting Strong Correlation**

**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?

Diagnostics shed light on MR character…

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

$$r_{ND} = \frac{I_{ND}}{I_D + I_{ND}}$$

low MR  

high MR

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

“safe islands” for DFT

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


**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)

overcoming limitations of DFT
Redox flow battery redox couple design:

We must address:
- Stability/resistance to crossover
- Concentration (solubility)
- Redox potential

38 heterocycles

779 core ligands

897 functional groups

4 metals: Cr, Mn, Fe, Co

2.8 M bulky transition metal complexes

\[ E_{\text{cell}} = 0.5 \times V_{\text{cell}} \times C \times n \times F \]

ACCELERATING RFB REDOX COUPLE DESIGN

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

DFT-computed design principle and properties in weeks

ANN step: 15 minutes

DFT step: 6 weeks

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

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).
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:

RACs as features for MOFs:

Geometry only random forest:

Adding RACs improves models:

Feature analysis shows when chemistry matters:

CoRE MOF (experimental)

Hypothetical (Boyd et al.)

Hypothetical (Andersen et al.)

But….conclusions are strongly sensitive to the data set! Why is this?

WHAT’S DIFFERENT ABOUT HYPOTHETICAL MOFS?

Hypothetical sets should be enlarged to mimic diversity of the experimental (CoRE MOF) dataset

Hypothetical = colored
Expt. only = gray

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:

ML TELLS US WHY HEURISTICS FAIL

Classification for **activation** stability

Regression for **thermal** stability

**design principles:**

... ML generalizes to unseen data

ENGINEERING STABLE MOFS

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

MOF STABILITY AT YOUR FINGERTIPS

https://mofsimplify.mit.edu


Gianmarco Terrones
ChemE Ph.D.
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

ACKNOWLEDGMENTS

Chenru Duan

Aditya Nandy

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

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?

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

Engineering distortion in real materials:

Or post-synthetic modification of metal sites

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

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

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

- **uncertainty quantification**
  - feature space
  - latent space
  - distance UQ

**ML FOR INORGANIC DISCOVERY**


---

**Functional Materials**

- LUMO
- HOMO
- $\Delta E_g = 4$ eV
- $\Delta E = 3.83$ eV, $d = 4.51$
- ANN: 4.00 eV
- DFT: 3.97 eV

**Catalysis**

- $\text{Cr, Mn, Fe, Co}$
- $>5,600$ SCOs
- MAE: 4.5 kcal/mol
- $>5,600$ SCOs
- $37,128$ catalysts

---

**Model Exploitation**