Navigating Transition-Metal Chemical Space:

Artificial Intelligence for First-Principles Design

Jon Paul Janet[1], Chenru Duan[1,2], Aditya Nandy[1,2], Fang Liu[1], and Heather J. Kulik[1,*]

[1]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

CONSPECTUS: The variability of chemical bonding in open-shell transition-metal complexes both motivates their study as functional materials and catalysts but also challenges conventional computational modeling tools. Here, tailoring ligand chemistry can alter preferred spin or oxidation states as well as electronic structure properties and reactivity, creating vast regions of chemical space to explore when designing new materials atom by atom. Although first-principles density functional theory (DFT) remains the workhorse of computational chemistry in mechanism deduction and property prediction, it is of limited use here. DFT is both far too computationally costly for widespread transition-metal chemical space exploration and also prone to inaccuracies that limit its predictive performance for localized *d* electrons in transition-metal complexes. These challenges starkly contrast with the well-trodden regions of small organic molecule chemical space where the analytical forms of molecular mechanics force fields and semi-empirical theories have for decades accelerated the discovery of new molecules, where accurate DFT functional performance has been demonstrated, and where gold standard methods from correlated wavefunction theory can predict experimental results to chemical accuracy.

The combined promise of transition-metal chemical space exploration and lack of established tools has mandated a distinct approach. In this Account, we will outline the path we charted in transition-metal chemical space exploration starting from the first machine learning (ML, i.e., artificial neural network, ANN, or kernel ridge regression, KRR) models and representations for the prediction of open-shell transition-metal complex properties. The distinct importance of the immediate coordination environment of the metal center as well as the lack of low-level methods to accurately predict structural properties in this coordination environment first motivated and then benefited from these ML models and representations. Once developed, the recipe for prediction of geometric, spin state, and redox potential properties was straightforwardly extended to a diverse range of other properties, including in catalysis, computational "feasibility", and in the gas separation properties of periodic metal-organic frameworks. Interpretation of selected features most important for model prediction revealed new ways to encapsulate design rules and confirmed that models were robustly mapping essential structure–property relationships. Encountering the special challenge of ensuring that good model performance could generalize to new discovery targets motivated how to best carry

out model uncertainty quantification (UQ). Distance-based approaches, whether in model latent space or in carefully engineered feature space, provided intuitive measures of domain of applicability. With all of these pieces together, ML can be harnessed as an engine to tackle the large-scale exploration of transition-metal chemical space needed to satisfy multiple objectives using efficient global optimization methods. In practical terms, these artificial intelligence tools brought to bear on the problems of transition-metal chemical space exploration have resulted in ML-model assessments of large, multi-million compound spaces in minutes and validated new design leads in weeks instead of decades.

Key references

- Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017,** *121*, 8939-8954.[1] *A systematic representation tailored for transition metal chemistry was introduced and used to train machine learning models to predict spin-splitting energies, redox potential, and bond lengths of transition-metal complexes. Analysis of feature-selected subsets revealed length and nature of design principles for each property.*
- Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018,** *9*, 1064-1071.[2] *An artificial neural network (ANN) trained on an ad hoc feature set for transition-metal chemistry was used to discover spin crossover complexes. A space of > 5k transition-metal complexes was explored with a genetic algorithm in which the objective function was modified by feature-space distance-derived uncertainty quantification (UQ) penalty.*
- Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019,** *10*, 7913-7922.[3] *A UQ metric derived from the distance of lead compounds to available training data in the latent space of a neural network was introduced, demonstrating*

superior performance to alternative quantities such as ensembles for model exploration or exploitation applications in transition-metal chemistry.

- Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020,** *6*, 513-524.[4] *The multiobjective optimization of redox potential and solubility for redox couples was carried out in a space of > 2.8M hypothetical transition-metal complexes. Efficient global optimization was employed in conjunction with the latent space distance UQ metric and a multi-task ANN trained on systematic representations for transition-metal chemistry.*

1. Introduction.

Rapid prediction of properties of transition-metal complexes with computational chemistry modeling tools is not a new objective.[5-8] Challenges for carrying out experiments on open-shell transition-metal complexes as well as in interpreting their outcomes have meant that computational and theoretical tools have been indispensible since the early days[9] of research in open-shell transition-metal chemistry. However, the first computationally affordable tools from molecular mechanics (MM)[10-12] or semi-empirical quantum mechanics (SQM)[13] models failed to generalize for open-shell transition-metal chemistry. Although some promising models have been developed to enable specific property predictions over a range of similar materials,[14] to capture ligand field and Jahn–Teller effects[15-16] (e.g., by incorporating terms from the angular overlap model[17]), or to be general enough to work across a large number of elements[10], these methods have not been widely adopted for predictive modeling of open-shell transition-metal chemistry. In more computationally demanding, first-principles (e.g., density functional theory

or DFT) methods, small errors are magnified by subtle imbalances in the treatment of localized *d* or *f* electrons[18-20] of transition-metal complexes.

The challenges for modeling open-shell transition-metal chemistry arise due to the same quantum mechanical properties that make them desirable targets for materials design, e.g., as catalysts or sensors. Mid-row transition metals have variable, low-lying electron configurations from multiple oxidation and spin states that favor a range of coordination environments (Figure 1). Changing the electron configuration of a transition-metal complex will change geometric, electronic, optical, and magnetic properties in a non-linear, often step-wise fashion without changing a graph/connectivity-based representation of the complex (Figure 1). This lack of smooth variation in properties likely explains the challenges for generalizing SQM or MM methods to open-shell transition-metal chemistry, whereas the degeneracy of slightly different electronic states explains lack of error cancellation in first-principles DFT.[18-20]

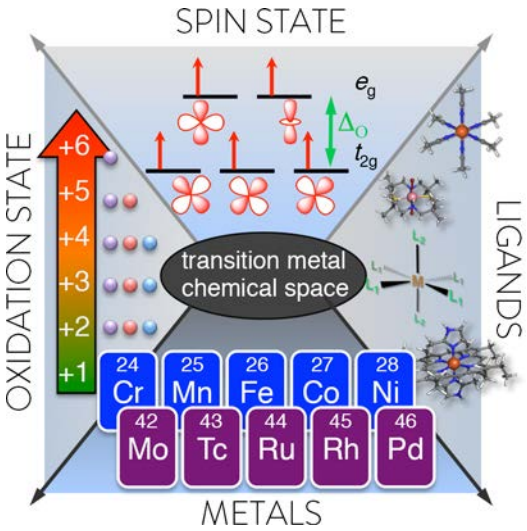

Figure 1. Combinatorial challenges for enumerating transition-metal chemical space depicting the interplay between choice of metal (bottom), oxidation state (left), spin state (top), and ligand chemistry (right) for transition-metal complexes.

Taking a step back to examine the "small molecule universe"[21-22] of closed-shell organic molecules, we see a vast space but one that can be readily traversed with computational

chemistry because efficient methods are available (i.e., SQM or MM) and more-demanding first-principles methods (i.e., DFT or wavefunction theory) can achieve chemical accuracy, in part due to the straightforward mapping between structure and function that paved the way for impressive machine learning (ML) model performance.[23] In comparison, it is evident that the number of possible metals, oxidation, spin states, and ligands that can be used to build up a transition-metal complex creates a vast chemical space with a wide range of sizes and coordination environments (Figure 1).[6,24-25] Only a small fraction of this space has likely ever been synthesized and characterized (e.g., 500k inorganic structures in the Cambridge Structural Database).

Although dependence of transition-metal complex properties on structure is highly non-linear, predictive heuristics (e.g., ligand field theory[9,26]) provide evidence that structure–property relationships are suitable tasks for machine learning. For example, weak-field ligands (e.g., ammonia) in mid-row, mononuclear octahedral 3*d* transition-metal complexes favor high-spin states, whereas strong-field ligands (e.g., carbonyl) prefer low-spin states. However, recovering this heuristic with first-principles, physics-based models (e.g., DFT) is not always straightforward.[18-19,27-29] Encountering these challenges motivated a distinct approach from our group. To build data-driven models, we compromised between the amount of data we could generate with the accuracy of the physics-based models (i.e., from DFT) by sampling a range of their parameters (i.e., exchange-correlation functionals).[30-31] We collected data and developed representations tailored for transition-metal chemistry to build the first predictive artificial neural network (ANN) models for open-shell transition-metal complex properties such as spin splitting energy, redox potential, and metal-ligand bond length.[1,31] Once this recipe was developed, we generalized it to tasks as far-ranging as predicting calculation outcomes[32-33], catalytic activity[34],

or gas uptake in metal-organic frameworks[35] (MOFs). Analysis[1,30] of the most essential features of these models recapitulated well-known heuristics[9,26] but also revealed exceptions to conclusions previously reached on small[34] or imbalanced[35] data sets.

Despite these early successes, our goal of leveraging such models to enable chemical discovery was not yet reached. Simply having a predictive ML model trained on the modest data set sizes we used early on (ca. 300–1000 complexes) does not ensure that ML models (e.g., ANNs) will generalize to new chemical spaces. We thus investigated how to quantify uncertainty in our ML models.[3,31] With evolutionary algorithms informed by model uncertainty, we set out to define and explore large spaces (i.e., thousands to millions) of theoretical compounds to discover promising new leads.[2,36] Early demonstrations on unearthing new spin crossover (SCO) complexes[2] and tailoring band gaps[36] were promising, accelerating the discovery of lead compounds by reducing the number of time-consuming calculations needed. Aiming to fully realize the opportunities of ML-accelerated discovery in transition metal chemistry, we carried out the simultaneous optimization of solubility and redox potential for redox flow battery materials in a space of millions of theoretical compounds, obtaining convergent design rules in weeks rather than decades.[4] While our efforts thus far have only sampled a small fraction of feasible transition metal chemical space, they represent a much larger portion than could ever be addressed by traditional experimental or computational approaches. In the remainder of this Account, we summarize the key considerations in demonstrating and extending our approach to outstanding challenges in open-shell transition-metal complex design.

2. Curating Data Sets.

Before embarking on the incorporation of data-driven tools into chemical discovery, it is useful to consider three interrelated concerns: data set (i.e., how much and what kind), model

(i.e., how complex or interpretable), and representation (i.e., tailored or general). While we have already reviewed some practical aspects of these tradeoffs in initial ML model training,[37] the task of finding novel materials motivates a closer look at how ligands are selected for transition-metal complex discovery and data set generation. The most straightforward strategy is to employ a range of common ligands, such as those known from the spectrochemical series (Figure 2). By restricting complexes of these ligands to reasonably high symmetry (i.e., no more than 1 or 2 equatorial or axial ligand types), even with a small number (ca. 35) of unique ligand types, one can rapidly generate hundreds of transition-metal complexes that sample a range of denticities, charges, coordinating atom types, and sizes. When first building ML models,[1,31] that was our primary strategy to ensure realism. However, by choosing common ligands, the wide range of complex sizes meant that generating equilibrium properties (i.e., from geometry optimizations) required significant computational effort. Given that metal-local effects dictate many transition-metal complex properties, such an approach does not appear to be the most efficient way to provide new data about distinct coordination environments.

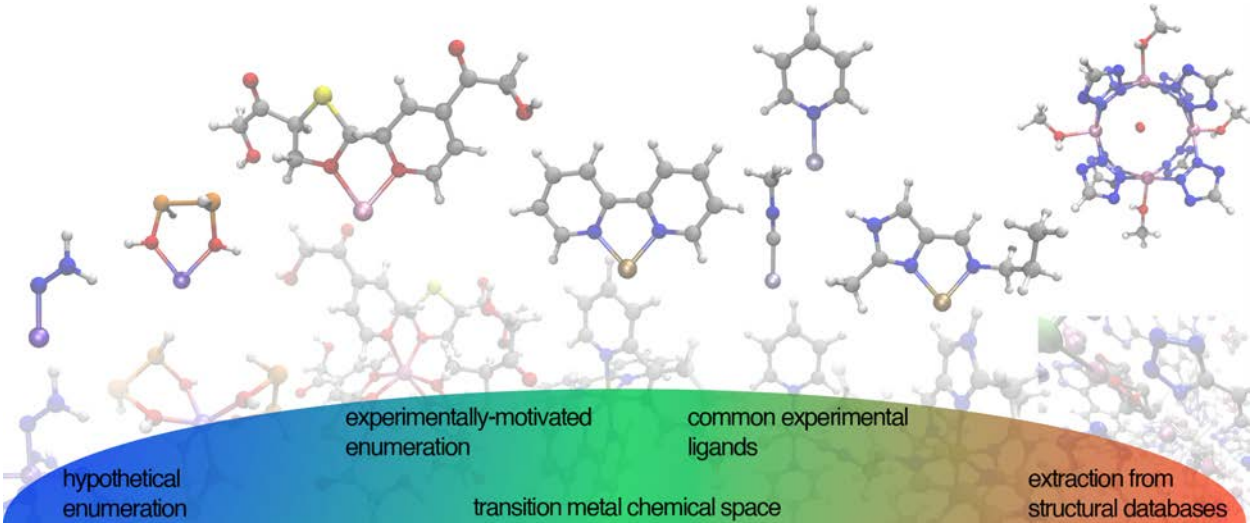

Figure 2. Depiction of four strategies for transition-metal chemical space construction using discrete ligand pools (from left to right): hypothetical small ligand enumeration from the OHLDB; experimentally motivated, rule-based ring generation in Ref. 4; use of common experimental ligands in homoleptic and heteroleptic complexes; and extraction of molecular or

periodic materials from structural databases. Representative ligand structures are shown in ball-and-stick representation with the complexed metal colored as: Cr in gray, Mn in mauve, Fe in brown, Co in purple.

Alternatively, enumerative strategies have been exploited in small molecule organic chemistry[22] for chemical space exploration. The generated database (GDB)[22] series has enumerated molecules up to discrete numbers of heavy (i.e., non-H) atom sizes under the constraints of obeying the octet rule and being in closed-shell singlet states with no net charge. Electronic structure properties of GDB sets with up to nine heavy atoms have been evaluated quantum mechanically giving rise to data sets[38-39] that have been frequent targets for ML model training. Extension of enumerative strategies to candidate transition-metal complexes requires a distinct approach. First, the smallest non-trivial octahedral complex has at least seven heavy atoms due to the ligands and has more electrons than a typical organic molecule due to the metal center. Thus, to exhaustively generate a series of very small transition-metal complexes, individual ligands should be constrained to a small number (ca. one or two) of heavy atoms per metal-coordinating site (e.g., four heavy atoms in a bidentate ligand).[25] Second, all GDB molecules are uncharged, closed shell singlets that obey the octet rule and many have fully saturated, sterically hindered bonding environments (e.g., methane and other alkanes). Such molecules are unlikely to be privileged as ligands for transition metal complexes either due to steric crowding that may prevent substantive metal coordination (e.g., methane) or due to the importance in transition metal chemistry of charged molecules (e.g., halide or hydroxyl ions) as well as those that violate the octet rule (e.g., NO_2) or those with unsaturated bonding.

We recently introduced[25] an enumerative strategy that relaxed typically-applied constraints (e.g., for the GDB) in generating ligands with one to four heavy (i.e., C, N, O, P, or S) atoms, giving more favorable scores to neutral and octet-obeying molecules but still allowing

for exceptions (Figure 3). Because the resulting ligands sampled a wider array of metal-coordinating-atom elemental identities and hybridization states, DFT-calculated properties (i.e., the octahedral homoleptic ligand database, OHLDB[25]) of around 700 homoleptic complexes of these ligands with mid-row transition metals (i.e., Cr–Co) showed a broader range of spin-splitting energies than those composed of common ligands (Figures 2 and 4). For similar reasons, enriching ML models with this training data improved ML model performance on larger transition-metal complexes from experimental databases[40] that contained similar metal-local coordination to these OHLDB[25] complexes (Figure 4). This improvement occurred in spite of the fact that many of the ligands in our set (e.g., complexes with a phosphorus nitride ligand[41]) would be difficult to isolate in experimental conditions, further highlighting the central importance of metal-local bonding.

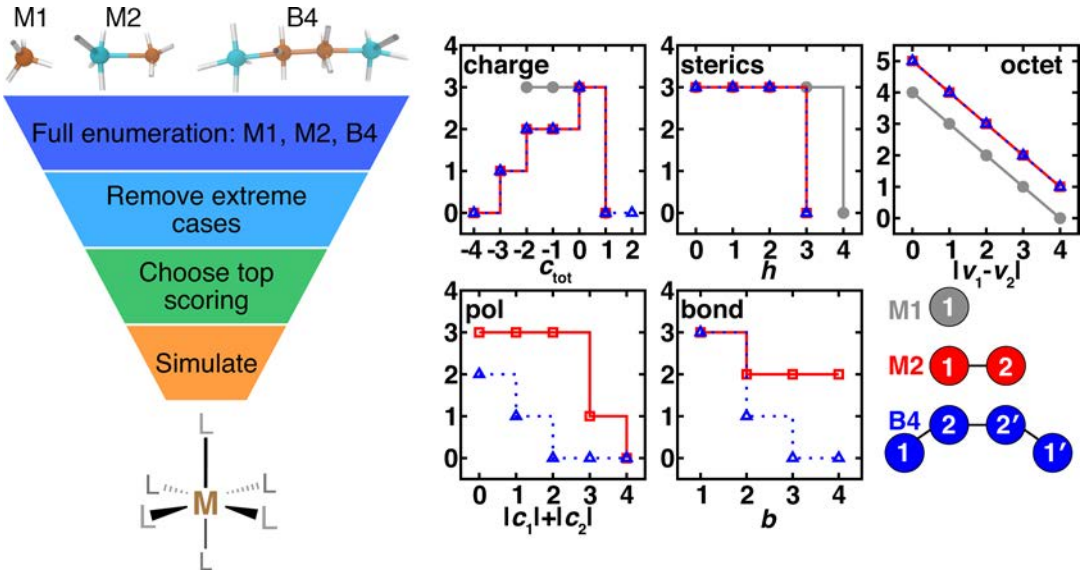

Figure 3. (left) Procedure for generating OHLDB ligands: enumerating all ligands, removing extreme cases, scoring and retaining top-scoring ligands, and simulating homoleptic M(II)/M(III) (M = Cr, Mn, Fe, or Co) octahedral complexes, as described in ref. 25. (right) Scores (i.e., M1 in gray circles, M2 in red squares, and B4 in blue triangles) of ligands. Scores for all three ligand types include: charge (c_{tot}, top left); sterics (h, top middle); and octet ($|v_1-v_2|$, where $v_2 = 8$ for M1 ligands, top right). Two scores are only for M2 and B4: pol, the polarization measured by $|c_1| + |c_2|$ (bottom, left) and bond, b, the 1-2 bond order (bottom, middle). Ligands with high charge and violation of the octet rule are penalized but not eliminated. While 71 of the OHLDB

molecules were in databases, most OHLDB ligands are not. Adapted with permission from ref. 25. Copyright 2020 Royal Society of Chemistry.

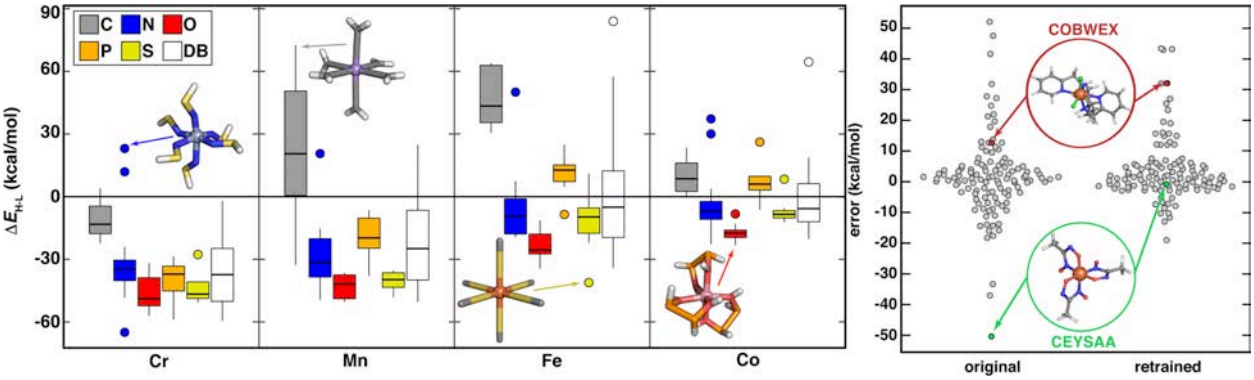

Figure 4. (left) Boxplots of M(II) (M = Cr, Mn, Fe, or Co) spin splitting energy, $\Delta E_{H\text{-}L}$, (in kcal/mol) for OHLDB complexes grouped by metal and by ligand-coordinating-atom (shown in inset legend). Each box indicates the median by a horizontal line, the interquartile range (IQR), and whiskers indicate 1.5× the IQR. The DB (white boxplot) indicates all $\Delta E_{H\text{-}L}$ values from Refs. 1-2,31,36, regardless of coordinating atom. The difference in OHLDB values from the DB highlights increased diversity in properties in the OHLDB. (right) Swarm plot of RAC/ANN signed errors (in kcal/mol) from Ref. 3 on an out-of-sample 116 CSD structure data set[3] (original) and after retraining with OHLDB data in ref. 25 (retrained). The single most improved and worsened points are shown in green and red insets, respectively. The reduction in MAE (from original: 8.6 kcal/mol to retrained: 6.7 kcal/mol) highlights how enrichment of diverse metal-ligand coordination environments from OHLDB can improve model performance on larger CSD molecules. Adapted with permission from ref. 25. Copyright 2020 Royal Society of Chemistry.

Between the limits of enumerating potentially implausible ligands or of repurposing well-studied ligands, we took yet another approach in recent work. To enumerate a space of nearly 2.8 M candidate redox flow battery (RFB) redox couples, we designed bidentate ligands to form homoleptic transition-metal complexes (Figure 2). Motivated by the observation that many well-known ligands in transition-metal chemistry consist of fused five- or six-membered rings, we devised a series of eight nitrogen- or oxygen-coordinating rings of that size, which we further permuted with in-ring changes to conjugation and addition of heteroatoms. These 38 distinct heterocycles were then fused to form nearly 800 bidentate core ligands that in turn were functionalized hierarchically with around 900 unique polar and bulky groups to create a space of 700k unique ligands or 2.8 M transition metal complexes (Figures 2 and 5).

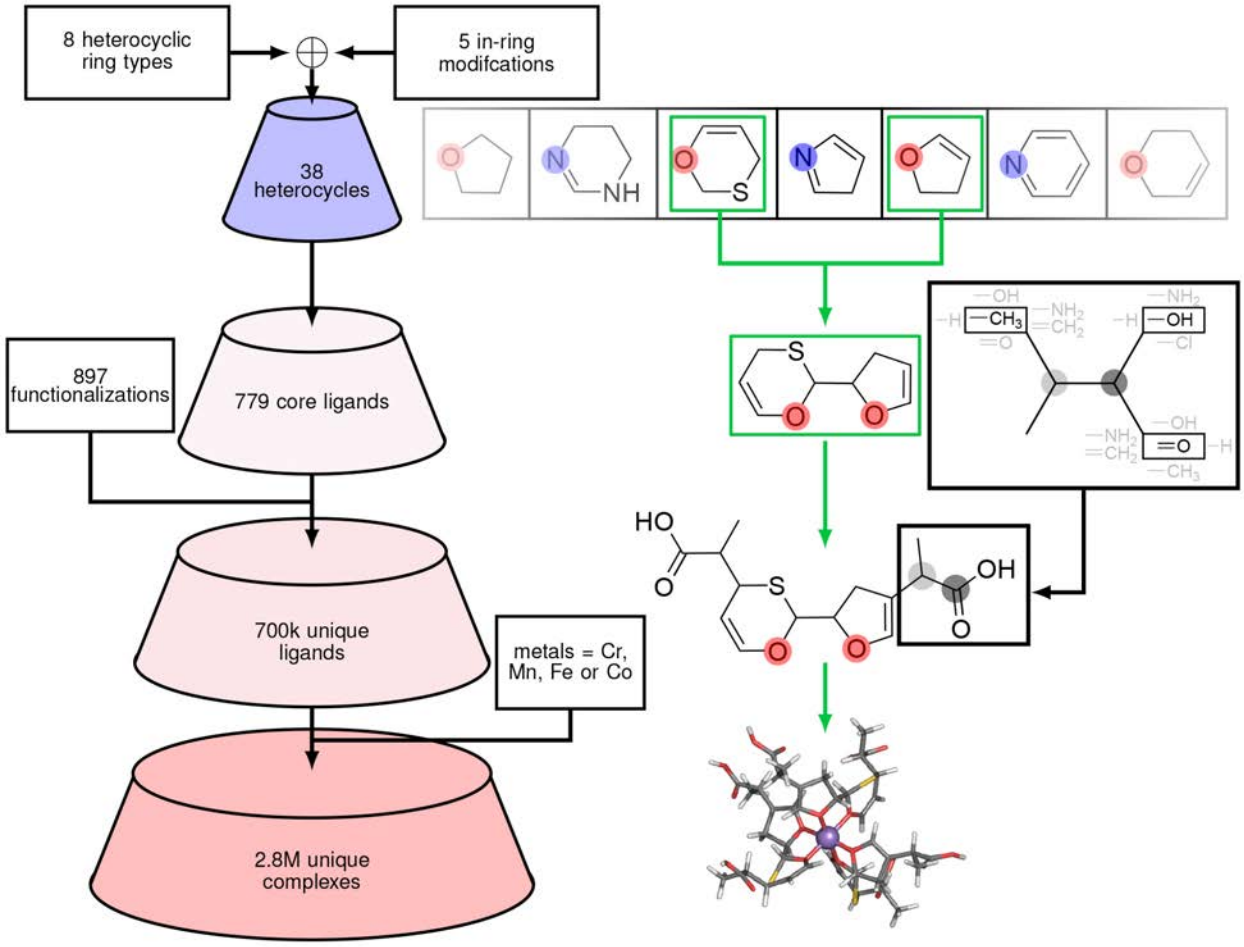

Figure 5. Approach for generating 2.8M homoleptic complex space for multi-objective RFB design in Ref. 4. Experimentally-motivated rings are modified to build a chemically reasonable, continuous and smooth design space of ligands, as shown in the flowchart at left. At right, example skeleton structure components and their assembly or modification coinciding with each step of the flowchart is shown to form a bidentate ligand and assembled complex (ball and stick structure). Adapted with permission from ref. 4. Copyright 2020 American Chemical Society.

In principle, these three strategies provide complementary information for enrichment of data-driven search and ML model development. While some properties (e.g., highest occupied molecular orbital, HOMO) cannot be expected to have similar distributions from the smallest OHLDB complexes to the large, bulky RFB complexes, we have found benefit to building models that generalize well across these sets.[33] Indeed, we recently curated[33] a subset of nearly

five thousand complexes derived from these three strategies (Figure 6). We observed improved ML model performance when using training data across the three (i.e., OHLDB, enumerative rings, and common ligands) strategies, even for application to a larger space of 187k theoretical complexes of common ligands (Figure 6).[33]

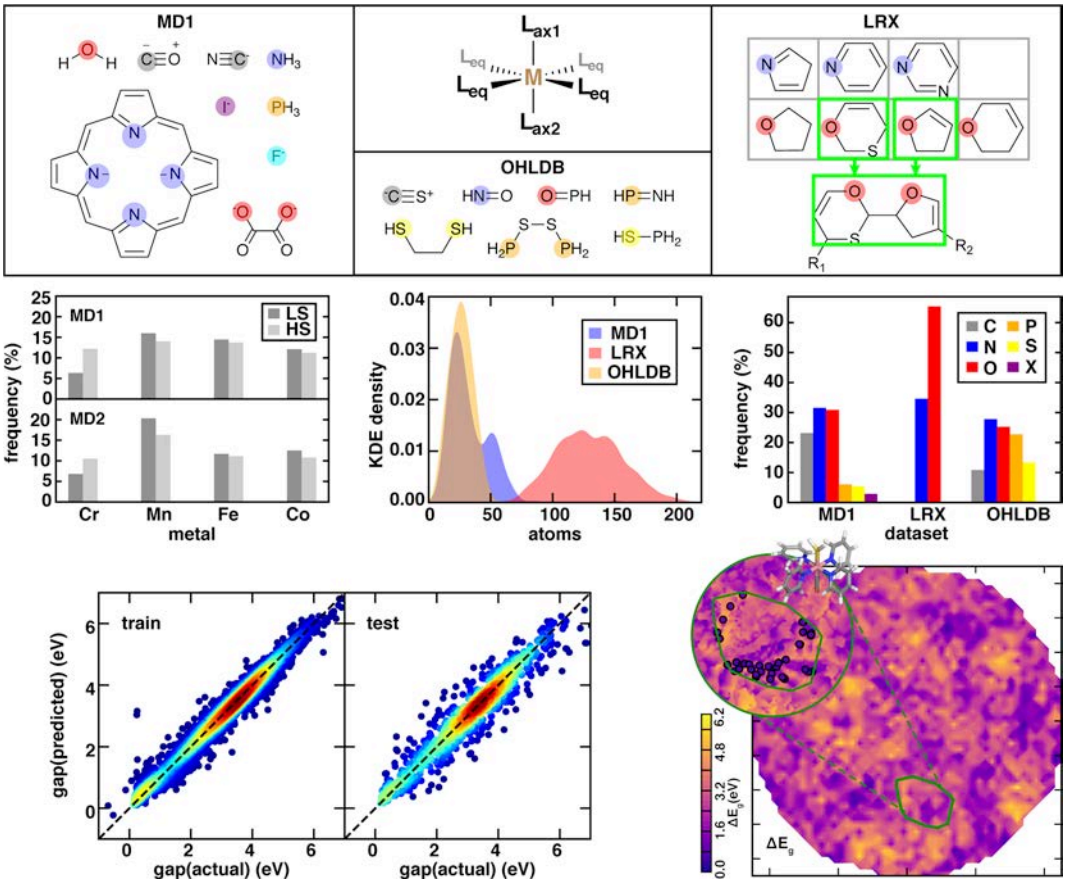

Figure 6. (top) Properties of octahedral complexes in a merged data set of common ligands (*MD1*) and *MD2* that includes the OHLDB from ref. 25 and large redox (*LRX*) RFB molecules from ref. 4, as curated in ref. 33. Bar graphs of the metal and spin state (HS: dark gray, LS: light gray) are shown along with kernel density estimation (KDE) of the size distribution of complexes and a clustered bar graph for the metal-coordinating atom identity (X indicates any halide) in *MD1, LRX,* and *OHLDB*. (bottom, left) Train and test set performance (MAE: 0.1-0.3 eV) of a RAC-155/ANN trained on *MD2* to predict the HOMO-LUMO gap. (bottom, right) Dimensionality reduction with t-distributed stochastic neighbor embedding[42] (t-SNE) of ANN-predicted HOMO-LUMO gaps (labeled ΔE_g, colored as in inset colorbar) for a hypothetical space of 187k complexes made from common ligands. Adapted with permission from ref. 33. Copyright 2020 American Chemical Society.

The challenges of how to span known, synthesized chemical space and yet to reach into unexplored spaces is not specific to molecular transition-metal chemistry. Hypothetical materials

have been central to computational discovery[43] of new metal-organic frameworks (MOFs). MOFs are reticular materials consisting of secondary building units (SBUs), linkers, and functional groups, which can rapidly give rise to a combinatorial explosion (Figure 2). By repurposing relatively few SBUs and linkers, hundreds of thousands of new hypothetical materials can be generated[43]. In addition to these hypothetical materials, thousands of MOFs have been synthesized and characterized experimentally.[44] Using a series of metal-local and global materials featurizations with graph-based revised autocorrelations (RACs, see *Sec. 3*), we recently showed[35] that ML models trained on hypothetical MOFs and on experimentally characterized MOFs will "see" different structure–property relationships (Figure 7). We determined that this difference arises in part due to the expected observation that many hypothetical materials have features (e.g., large pore sizes) that are unlikely to be synthetically accessible (Figure 7). More unexpectedly, we observed[35] that hypothetical sets tended to lack diversity, completely omitting SBU chemistry that had previously been synthesized (Figure 7). Moving forward, any approach for spanning chemical space, whether for model exploitation (i.e., using a model where it is expected to perform well) or exploration (i.e., iteratively improving the model with new promising but uncertain data points, see *Sec. 4*), should make an effort to bridge the gap between the known and the unknown, and take measures to quantify this gap (*Sec. 3*).

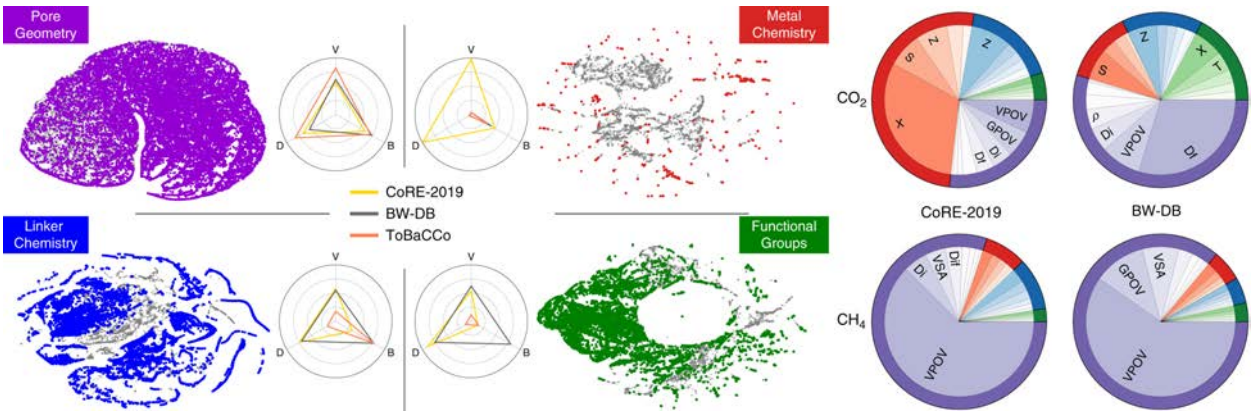

Figure 7. (left) t-SNE on subsets of MOF RACs and geometric descriptors including: (top, left) pore geometry in purple, (top, right) metal chemistry RACs in red, (bottom, left) linker RACs in blue, and (bottom, right) functional group RACs. All hypothetical and experimentally-characterized (i.e., CoRE-2019) datasets are shown in gray, whereas only hypothetical data is colored, highlighting the lack of diversity in metal chemistry in hypothetical sets. The radar charts show the three diversity metrics: variety (V), balance (B) and disparity (D), for the three databases. (right) Pie charts colored by feature subset as in left showing the feature importance for low-pressure CO_2 adsorption and CH_4 deliverable capacity for the CoRE-2019 database compared to BW-DB, highlighting how limited metal chemistry diversity in BW-DB leads to underestimation of its importance in adsorption properties. Adapted with permission from ref. 35. Copyright 2020 Springer Nature.

3. Defining Distance in Chemical Space.

Measuring similarity between chemical compounds (i.e., when features and/or properties are comparable) is essential in ML-accelerated chemical discovery because ML models primarily serve to map similar chemical compositions to similar observable properties. Similarity itself may be expressed as an understanding that materials with comparable features within a chosen chemical representation should have comparable properties, whereas those that are distinct may have very different properties. Nevertheless, adequate measures of similarity present unique challenges for open-shell transition-metal chemistry. When working with relatively small but diverse data sets in open-shell transition-metal chemistry, we showed[31] that a carefully engineered feature set (i.e., selected based on intuition) of 25 mixed continuous discrete (metal)-local (i.e., MCDL-25[31]) could enable predictive ANN training with errors about an order of

magnitude lower than could be achieved with a more standard approach of representing all atoms in the molecule equally (i.e., with a whole-molecule representation) to the model. In practical terms, this is a difference of predicting transition-metal complex properties to within a root mean squared error (RMSE) around 3 kcal/mol with the engineered feature set trained on fewer than 1,000 diverse complexes[31] versus 30 kcal/mol RMSE with whole-molecule features (Figure 8).[45]

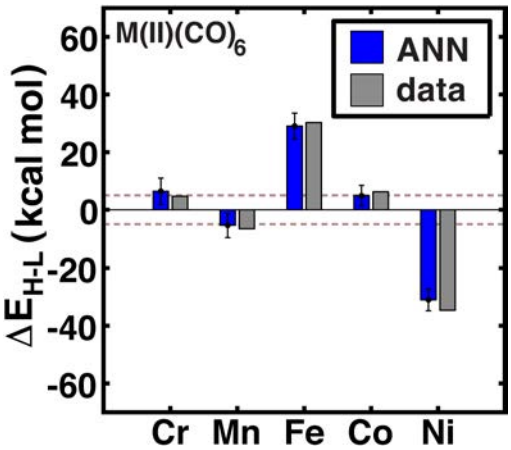

Figure 8. MCDL-25/ANN model predictions (ANN, blue bars) and computed (data, gray bars) spin-state splittings, $\Delta E_{H\text{-}L}$, from hybrid DFT (i.e., with B3LYP) in kcal/mol on $M(II)(CO)_6$ complexes, where M = Cr, Mn, Fe, Co, or Ni, from ref. 31. The error bars represent an estimated ±1 standard deviation credible interval from the mean prediction, and brown dashed lines correspond to a ±5 kcal/mol range around zero $\Delta E_{H\text{-}L}$, corresponding to near-degenerate spin states. This model outperforms a whole molecule representation-trained model (not shown) by an order of magnitude. Adapted with permission from ref. 31. Copyright 2017 Royal Society of Chemistry.

When the heavily engineered feature sets enable simple models (e.g., regularized multiple linear regression, KRR, and shallow ANNs) to be predictive, the features themselves (i.e., how the molecule is represented to the model) in turn provide guidance about the most essential structural variations that give rise to property variations.[1] Comparing the most important features to predict different properties provides guidance on when the properties may be independently tuned.[1,30] The feature-engineering step need not be manual (i.e., from intuition alone). We have found it beneficial[1,36] to start from systematic feature sets tailored for transition-metal chemistry known as revised autocorrelations[1,21,46] (RACs), which are geometry-free products and

differences of heuristic properties (e.g., nuclear charge, electronegativity, or covalent radius) on the molecular graph that can be metal- or ligand-focused as well as products over the whole molecule. By carrying out automated feature selection on RACs (e.g., by selecting the set of RACs that give the lowest prediction error of the model), the most essential RACs to define this distance in chemical space are revealed.

Although graph-based (i.e., based only on chemical bonds, not on distance or position in space) and thus lacking any explicit 3D geometric information, RACs have been successfully used to learn properties as far ranging as gas adsorption in MOFs (see Figure 7);[35] spin state, redox, and catalytic properties of transition-metal complexes (Figure 9);[1,31,34,36,47] and even the spin-state-dependent metal–ligand bond length (Figure 9).[31,48] RACs generally outperform both the *ad hoc* feature sets (i.e., MCDL-25) we first built, (i.e., RACs achieve < 1 kcal/mol mean absolute error on spin splitting vs 2-3 kcal/mol with MCDL-25) as well as the standard whole-molecule descriptors we mentioned earlier (Figure 9). Systematically feature-selected RAC subsets recapitulate known design principles and reveal new ones (Figure 10). For instance, RACs selected on spin-state properties tend to be more metal-focused than those[1,30] predicting bond lengths (Figure 10). Feature selection on the formation of high-valent metal-oxo catalytic intermediates and on spin-state prediction[34] showed a number of features in common, confirming[49] the importance of spin state in determining reactivity (Figure 10). Conversely, strongly non-local dependence of redox potential or frontier orbital properties[36] reveals why frontier-orbital-based heuristics for metal-oxo formation[50] should fail when studies encompass[34] a wide range of compounds (Figure 10).

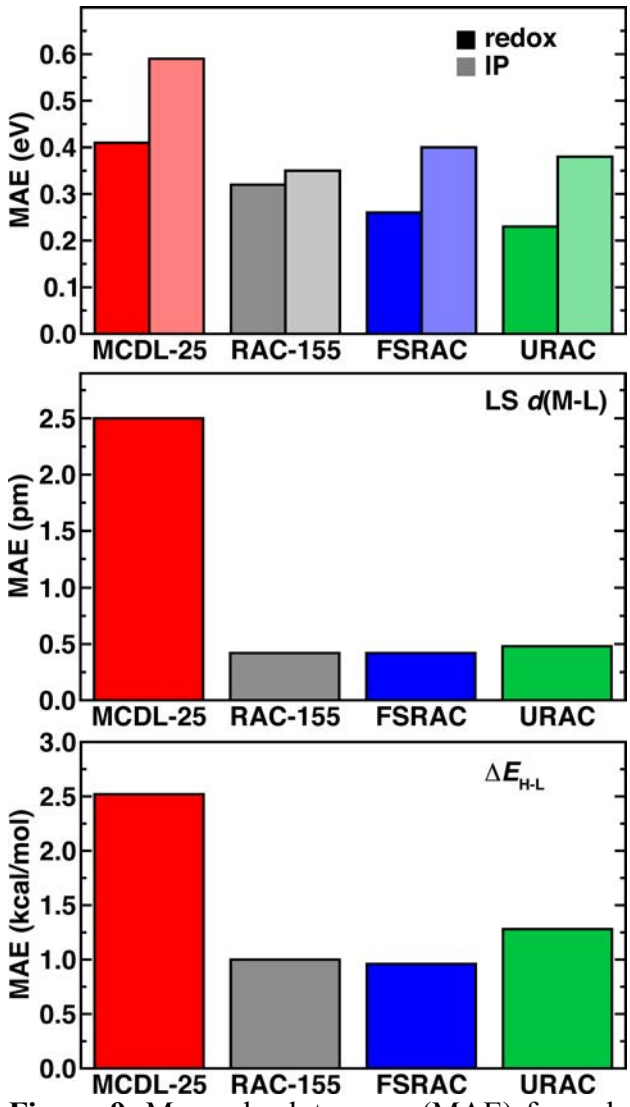

Figure 9. Mean absolute error (MAE) for redox and ionization potential (top) in eV, low spin (LS) metal-ligand bond length (middle) in pm, and $\Delta E_{H\text{-}L}$ (bottom) in kcal/mol. Comparisons are for the MCDL-25/ANN from ref. 31 along with KRR models trained with RAC-155, a feature-selected (FS) RAC subset for each property from ref. 1, and for the best-overall-performing URAC 26 feature set in ref. 1. These results highlight how the systematic RAC-155 outperforms *ad hoc* MCDL-25. Adapted with permission from ref. 30. Copyright 2019 American Chemical Society.

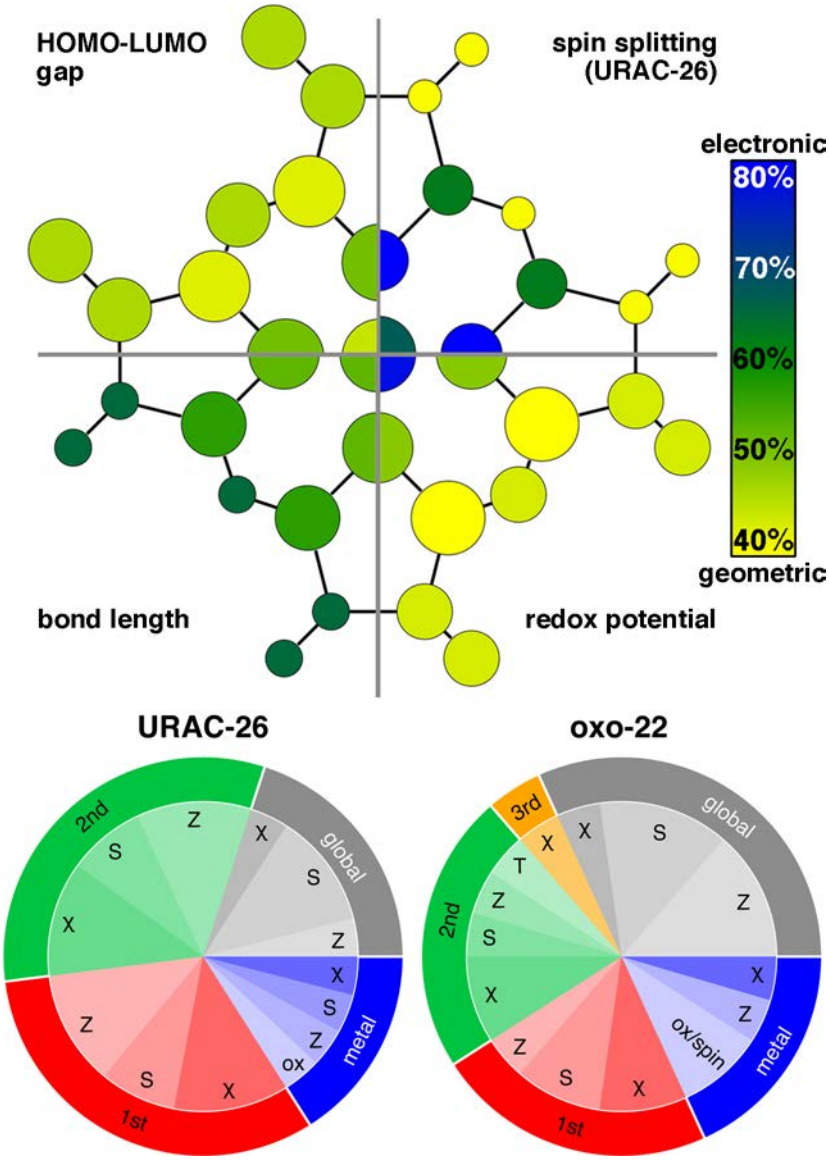

Figure 10. (top) Schematic of proximity and character (electronic in blue to geometric in yellow) character of feature-selected subsets of RACs on a representative molecular abstraction for: HOMO-LUMO gap (top, left) from ref. 36 and spin splitting (URAC-26, top, right), metal-ligand bond length (bottom, left) and redox potential in aqueous solvent (bottom, right) all from ref. 1. Atom sizes are scaled relative to the metal by the number of RAC dimensions involving that atom (divided into 1st, 2nd, and 3rd coordination sphere). (bottom) Pie charts of features in ref. 1 for predicting spin splitting (URAC-26, left) and metal-oxo formation (oxo-22, right) in ref. 34. Features are grouped by the most distant atoms: metal in blue; coordination sphere: 1st in red, 2nd in green, 3rd in orange; or global in gray. The property (i.e., electronic χ or Z and geometric S, T, or I) is indicated, and oxidation state (ox) and spin are assigned as metal-local. Adapted with permission from refs. 1, 34, and 36. Copyright 2017, 2018, and 2019 American Chemical Society.

A key advantage of having intuitive representations for transition metal chemistry is the relationship between definitions of distance in chemical space (i.e., chemical similarity) and ML model performance. If distance between compounds in the engineered feature space is predictive of target property similarity, then compounds highly distant from available training data are those for which trained ML model performance will be eroded (Figure 11). This behavior is implicit in KRR models: a KRR model prediction on a test molecule is informed based on the most proximal compounds, controlled only by a single hyperparameter that controls how the model behaves (i.e., the kernel width) that determines how quickly the contribution to a test point prediction decays for more distant training points. If the test point is too distant, then the KRR model makes no prediction at all. In chemical discovery applications (*Sec. 4*), it may be preferable to have a model (e.g., an ANN) make predictions with a fixed uncertainty budget, rather than making no prediction like a KRR model would. In either application, if the feature space representation provides chemically useful measures of distance, we can avoid making model predictions for new points distant in feature space (i.e., model exploitation)[2] or, alternatively, we can choose distant, promising points as new molecules to acquire for model exploration (i.e., by retraining the model in an approach known as active learning) applications[4] (*Sec. 4*).

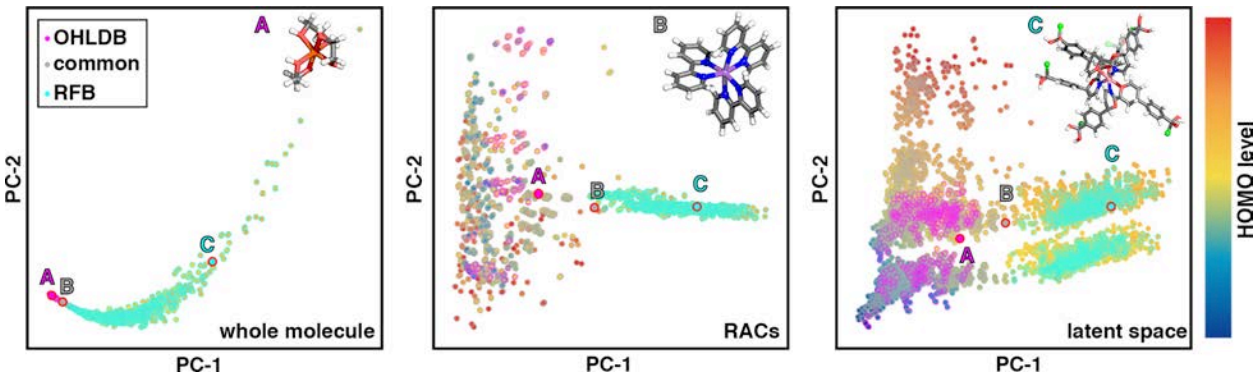

Figure 11. Dimensionality reduction with principal component (PC) analysis (PCA) for OHLDB complexes (magenta outlines, structure A inset), transition-metal complexes with common ligands (gray outlines, structure B inset), and RFB (cyan outlines, structure C inset) using a

whole molecule Coulomb matrix representation (left), RACs (middle), and latent space of a HOMO level ANN[33] (right) colored by HOMO (colorbar, right). The whole molecule representation overemphasizes differences by complex size and elemental composition (i.e., A and B are much closer to each other than B and C), whereas distinct compounds are more evenly distributed in the RACs feature space (i.e., A, B, and C are more distant from each other). The latent space PCA represents the model's interpretation of compounds, with more comparable distances of A, B, and C observed for both RACs and the latent space PCA. The PCA of the latent space highlights a slightly different data distribution than observed with RACs as a result of model training.

Although feature engineering or feature selection can lead to a subset of easily interpreted critical features, the popularity and flexibility of ANNs with numerous hidden layers (i.e., between the input layer where features are provided to the model and the output layer where the model makes predictions) motivate alternative approaches to uncertainty quantification (UQ). In these models, the action of the hidden layers can be interpreted as successive mathematical operations that discard uninformative features (i.e., carry out a form of automated feature selection) prior to the final layer at which point a linear mapping (i.e., as in multiple linear regression) typically occurs to the property space. Thus, rather than the feature space, the distance in the latent space[3,51-52] (i.e., the last hidden layer) can be used as a UQ metric[3] (Figures 11 and 12). Indeed, the distance in latent space[3] outperforms alternatives such as Monte Carlo dropout[53] (i.e., uncertainty estimates from different predictions obtained by varying the nodes in ANNs) or widely employed[54-57] ensembles (i.e., uncertainty estimates from different predictions obtained by varying which data is provided to the ANNs) in detecting high-error points, as are encountered in discovery applications, where these standard approaches are generally over-confident (Figure 12). This distance in latent space can also be calibrated[3] to property prediction units to provide property-unit-based error bars (Figure 12). For classification tasks, we have shown[32] that a closely related metric, the latent space entropy,[32] can capture both distance to available training data as well as proximity only to points from a single class that both are

important in ensuring a model knows how to assign the relevant class. Thus, both feature space and latent space distance are suitable tools for measuring the similarity of new compounds to known compounds (Figure 12). What remains is to consider how to best navigate transition metal chemical space (*Sec. 4*).

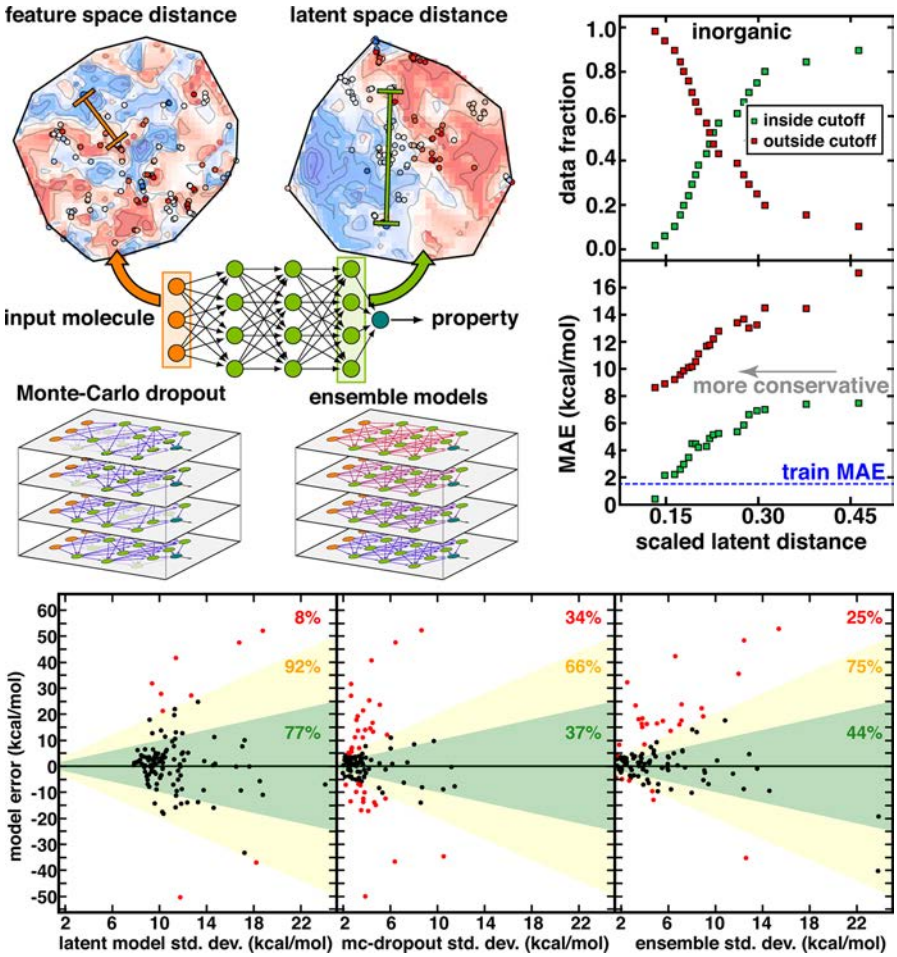

Figure 12. (top, left) Schematic of an ANN annotated with: feature space distance on a t-SNE map of data in the input layer (in orange), the latent space distance (in green), Monte-Carlo dropout (mc-dropout, i.e., zeroed out nodes), and ensemble models (i.e., varied model weights). (top, right) Demonstration of adjusting latent space distance cutoffs on a RAC/ANN for predicting the spin splitting on an 116-molecule CSD set curated in ref. 3. (bottom) Comparison of the calibrated latent space distance, mc-dropout, and ensemble standard deviation (std. dev.) prediction of errors (in kcal/mol) vs model error over 116-molecule CSD set. The one (two) std. dev. bound is shown in green (yellow), and points outside both are shown in red with percentages of each in inset. Latent space distance performs best, monotonically decreasing errors and bound more (92%) points. Adapted with permission from ref. 3. Copyright 2019 Royal Society of Chemistry.

4. Accelerating Discovery.

Machine learning (e.g., ANN or KRR) models trained on tailored representations[1,31] predict transition-metal complex properties with low errors (ca. 1-3 kcal/mol) with respect to the ground truth training data, e.g., from DFT calculations[1,31,34,36,47-48] or classical[35] simulation (Figure 9). Once trained, these ML models can be evaluated in a fraction of a second where the direct calculation would have required minutes to days. The promise of these savings in computational effort motivates model exploitation in the discovery of new materials with targeted properties, i.e., by applying a trained model on a large pool of diverse compounds (*Sec. 2*). A particular challenge for transition-metal chemistry is that the variation in elements, ligand chemistry, spin, and oxidation state is large enough that models trained on small data sets can be expected to erode when applied to diverse chemistry (Figure 12).[3-4,31] In such cases, it is still possible to exploit a model (i.e., use the model where it is confident) for property optimization but this should be done in conjunction with an uncertainty budget (*Sec. 3*) to avoid introducing large errors in chemical discovery (Figure 12).

To balance finding an optimal property from a large number of candidate complexes with an uncertainty budget on model error, we first combined[2] the two in the fitness function (i.e., the measure of what a good compound is) for an evolutionary algorithm (Figure 13). Genetic algorithms (GAs) are a natural and popular[8,58-61] choice for discrete optimization of transition-metal complex properties, where genes are assigned to encode a metal and ligands, and properties are optimized by selecting candidates from a discrete design space.[2,8,34,36,60-61] Although typically employed with physics-based (i.e., DFT or SQM) models[8,60-61], we showed[2] that ANNs trained on the feature-engineered MCDL-25 descriptors could be used with a GA to optimize properties of spin crossover (SCO) complexes without eroding ANN model accuracy

(Figure 14). Here, the composite fitness function favored complexes with degenerate spin states but penalized complexes with high feature space distance to training data to avoid making predictions where the ANN would be incorrect. This approach was used to explore a space of 5,600 theoretical compounds of which only a small number (2%) had been previously seen by the model[2] (Figures 13 and 14). While large feature space distances were penalized, the GA optimization was still permitted to explore significant differences in chemistry and discover new compounds (e.g., a change of ligand from a small water molecule to a larger furan, Figure 14).

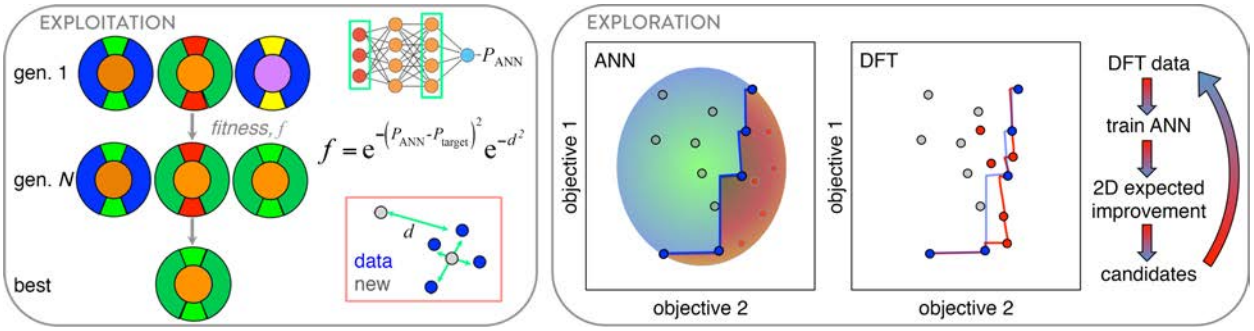

Figure 13. (left) Model exploitation with a genetic algorithm composite fitness function of optimal properties, *P*, and feature or latent space distance, *d*. (right) Model exploration for multi-objective optimization with efficient global optimization using the 2D expected improvement criterion. Calculated points are shown as solid circles with black outline, predicted properties are shown as shaded regions or circles with gray outline, and the Pareto front of computed values is shown as a solid line (the red line is the generation after the blue line).

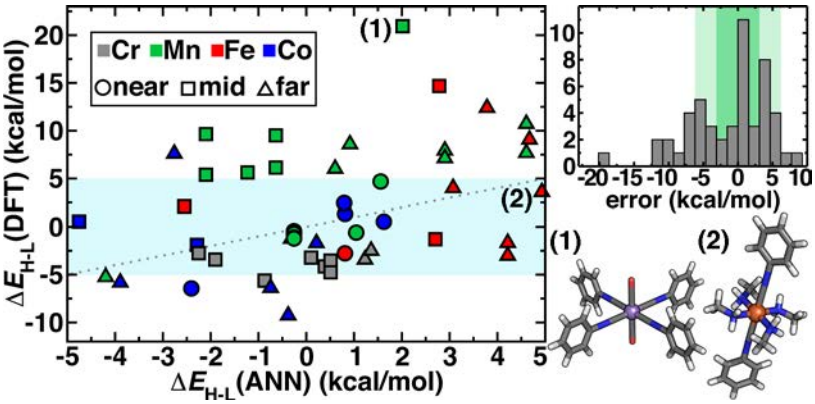

Figure 14. DFT $\Delta E_{H\text{-}L}$ vs. MCDL-25/ANN for 50 complexes obtained from a distance-controlled GA search of a 5600 complex space for SCOs grouped by metal (Cr gray, Mn green, Fe red, and Co blue) and feature space distance (near in circles, mid in squares, and far in triangles). A parity line is shown (gray, dotted), and $|\Delta E_{H\text{-}L}| < 5$ kcal/mol is shaded light blue. Over 2/3 of complexes are confirmed SCOs with DFT after ANN selection. (right, top) Error

histogram of ANN predictions. (right, bottom) Representative complexes corresponding to labels at left. Adapted with permission from ref. 2. Copyright 2018 American Chemical Society.

Overall, this approach ensured limited errors on discovered compounds to only about 1.5–2x (ca. 4.5 kcal/mol) the original test set error, and over two-thirds of candidate SCO complexes predicted by the ANN were validated to be SCO complexes with DFT (Figure 14).[2] We also showed how this approach could be used to find optimized HOMO–LUMO gaps from 10k candidate transition-metal complexes,[36] achieving average ANN-targeted values within chemical accuracy of the DFT-validated band gaps (ANN: 4.00 eV vs DFT: 3.97 eV). We also showed[34] that latent space (i.e., from the last hidden layer of the ANN instead of the feature space input layer) distance could be used similarly in a composite GA fitness function, in this case enabling the error-controlled discovery of single-site catalysts with counterintuitive properties.[34]

While uncertainty quantification (i.e., through feature space or latent space distance measures) can control model error in chemical discovery, model exploitation-based approaches require extensive calculations to generate data and train a model prior to its application in the design space (i.e., where we'd like to discover new compounds). When design principles are truly unknown, the most promising materials may be significantly distant from the original training data (Figures 4 and 12). Therefore, a trained ML model may not be predictive of these new spaces and a UQ-controlled optimization approach may not visit the most fruitful regions of the design space. As an alternative, model exploration through active learning can take advantage of these high uncertainty points in chemical discovery.[62-68] In this approach, the points judged to be simultaneously the most promising and the most uncertain are sought out for further characterization (here, with DFT) and then used to iteratively retrain the model.

We recently demonstrated[4] the power of model exploration to accelerate the discovery of new redox couples for redox flow batteries (RFBs) from a design space of nearly three million candidates (Figures 5, 13, and 15). The candidate complexes in our design space were large and bulky to prevent crossover in RFBs[69] and thus did not resemble any training data we had generated in prior redox applications (Figures 2 and 5).[1,36,47] We sought to identify how to improve RFB candidate material solubility in polar solvents while increasing their redox potentials.[4] These two properties are challenging to optimize simultaneously because redox potential is maximized by minimizing size (e.g., as we learned from RACs structure-property maps,[1,30] see *Sec. 3* and Figure 10), but solubility can be expected to be increased by adding polar functional groups and increasing size. Thus, we employed an efficient global optimization (EGO) approach with the 2D expected improvement (EI) criterion[70-71] to ensure sampling along the Pareto front defined by the set of complexes with optimal solubility and redox properties[4] (Figures 13 and 15). The EGO approach with 2D-EI identifies the simultaneously most promising (i.e., beyond the prior calculated Pareto front) but most uncertain points, balancing the likelihood of obtaining new lead compounds with also acquiring data that can improve the model. Although active learning with an approach like EGO is most naturally carried out with Gaussian process regression (GPR), a model similar to a KRR but one that has built-in UQ estimation,[72] we showed that on our data sets the errors were lowest for a multitask (i.e., that predicts solubility and redox at the same time) ANN with UQ from calibrated[3] latent space distances (Figure 12),[4] leading to faster identification of the best RFB leads.

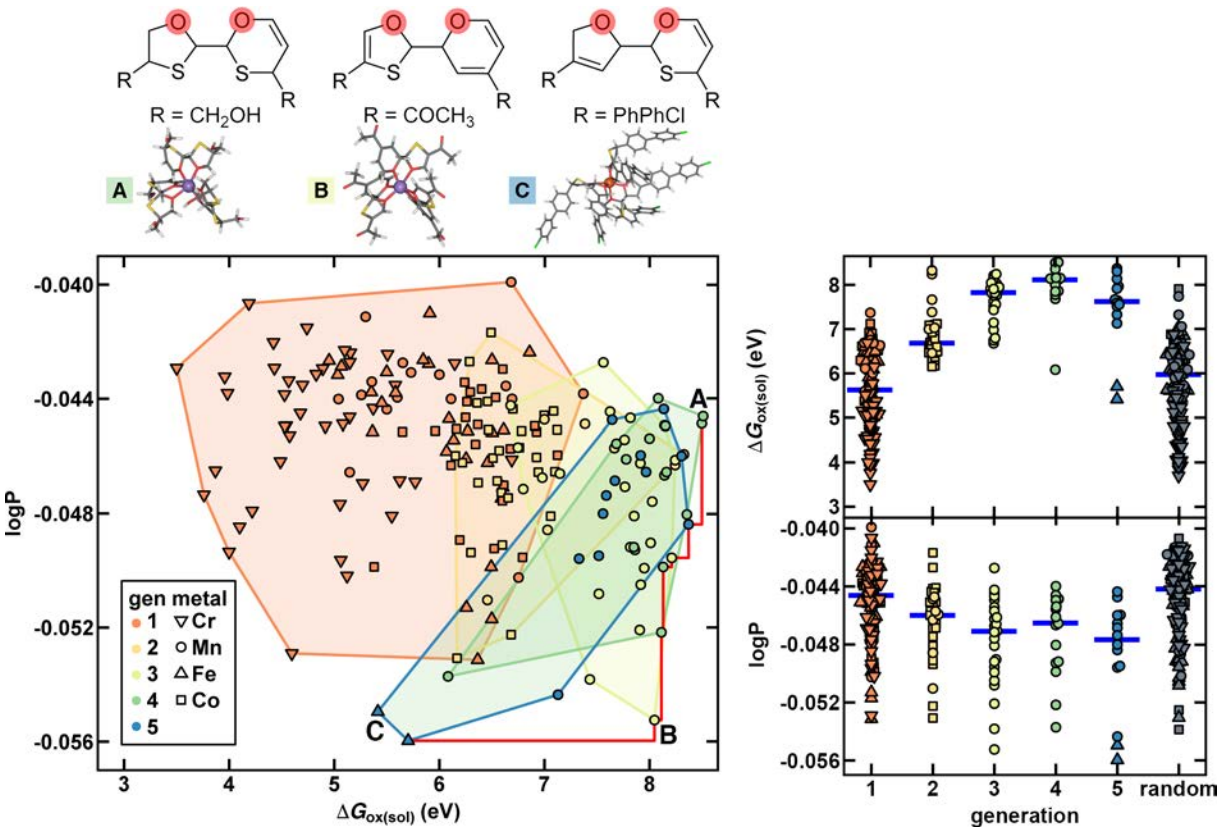

Figure 15. (Left) Redox potential ($\Delta G_{ox(sol)}$, in eV) and solubility (logP, unitless) for complexes from EGO with 2D-EI using a multitask ANN in ref. 4, colored as in in inset legend. The final Pareto front is indicated by a red line with three representative complexes labeled and shown at top: the highest $\Delta G_{ox(sol)}$ (A), the balanced trade-off (B), and the highest logP (C). (Right) Distribution of $\Delta G_{ox(sol)}$ and logP for each generation (as in inset legend) alongside a random sample (gray symbols). EGO/2D-EI yields a 500-fold acceleration over random search. Adapted with permission from ref. 4. Copyright 2020 American Chemical Society.

In only five generations of EGO with 2D-EI, the number of complexes expected by the ANN to exceed the Pareto front rapidly diminished,[4] and all of the DFT-validated points on the ultimate Pareto front were sampled from the final generations of the optimization (Figure 15). Given the high cost of each generation, corresponding to hundreds of DFT calculations, we elected to terminate our optimization once the ANN lacked confidence it could improve significantly (i.e., at the fifth generation). Importantly, design principles were convergent in this small number of generations: the best materials all were high-spin Mn complexes with oxygen-coordinating ligands that contained a sulfur heteroatom in a five- or six-membered ring paired

with small polar groups that increased solubility without sacrificing high redox potentials (Figure 15). Even taking into account acceleration through parallelization of DFT calculations, the 2D-EI approach achieved new leads in around five weeks where a fully random search would have required decades to reach the same conclusions (Figure 15). For higher-dimensional searches (i.e., optimizing three or more properties simultaneously), the acceleration benefits will be even greater because this acceleration generally is multiplicative in each dimension.[62-64] While impressive, the acceleration achieved by ML models for navigating transition metal chemical space that we have demonstrated thus far can be improved further through developments in representation and improvements in ML model architecture.

5. Summary.

Although the importance of discovery and design of new functional materials and catalysts is undisputed, large-scale discovery efforts had been hampered by the complexity of describing the electronic properties of open-shell transition metals. In this Account, we described the path we charted to overcome the limitations imposed by this complexity. Starting from the first ML models and representations for the prediction of open-shell transition-metal complex properties, we predicted spin splitting, redox potential, and structure in seconds instead of the days that computational chemistry (i.e., DFT) would have required. This paradigm shift was made possible by recognition of the importance of near-sighted, metal-local coordination environments to transition-metal complex properties as well as the utility of geometry-free representations to sidestep the lack of accurate structural predictions. Beyond speed, we showed how feature selection provided new ways to develop and interpret structure–property relationships, including the identification of opportunities for orthogonal property design and the interpretation of the robustness of data set diversity.

Accelerated discovery of transition-metal complexes requires not just fast ML models but also an approach to detect when model prediction accuracy will erode in new regions of transition-metal chemical space. Distance-based approaches, whether in model latent space or in carefully engineered feature space, provided intuitive measures of just how far is too far for model extrapolation. With all of these pieces together, we showed how ML models could be exploited or adaptively retrained to tackle the large-scale exploration of transition-metal chemical space. In practical terms, these artificial intelligence tools enabled multi-objective optimization from ML-model assessments of large, multi-million compound spaces in minutes to produce first-principles, DFT-validated design leads in weeks instead of decades.

From this progress, a number of natural next steps emerge. The approach we developed is general across inorganic materials and properties. The benefits of accelerated multi-objective optimization will increase with search dimensionality. Thus, a key target will be addressing outstanding challenges in catalysis, where computational design of selective, active, stable, and earth abundant materials has yet to be realized. A closely related concern is the realism that models trained on DFT predictions can provide in such challenging spaces. We have recently demonstrated[33] how to identify regions of chemical space that are "DFT-safe" islands. Key questions that remain are if derived structure–property relationships are more or less sensitive to the DFT functional that generated them than individual predictions, as well as how to best incorporate beyond-DFT methods when DFT is insufficiently accurate for design. Furthermore, some laboratory conditions and properties may still be too challenging to simulate. Extraction of experimental data from the literature, as has been demonstrated in inorganic materials synthesis,[73-74] can be expected to be fruitful. We have recently shown[48] that the geometry-free nature of RACs are particularly amenable to ML model comparison to literature data, but

challenges remain in ascertaining the availability, reliability, and mapping of extracted data to a single chemical structure. With ever-improving tools at our fingertips, computational exploration of transition-metal chemical space is poised to address some of the most pressing technological challenges.

ASSOCIATED CONTENT

Supporting Information. Glossary of specialized technical terms and acronyms (PDF)

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*email: hjkulik@mit.edu phone: 617-253-4584

Notes

The authors declare no competing financial interest.

Biographies

Jon Paul Janet completed his Ph.D. in Chemical Engineering and Computational Science and Engineering from MIT in 2019 in the group of Professor Heather Kulik. Prior to that, he obtained his B.Sc. in Chemical Engineering from the University of Cape Town in 2012 along with M.Sc. degrees in Scientific Computing and Applied Mathematics from the Technical University of Berlin and the Royal Institute of Technology in Stockholm in 2015.

Chenru Duan is a Ph.D. candidate in Chemistry in the group of Professor Heather Kulik at MIT. Prior to that, he obtained his B.S. in Physics from Zhejiang University in China in 2017.

Aditya Nandy is a Ph.D. candidate in Chemistry in the group of Professor Heather Kulik at MIT. Prior to that, he obtained his B.S. in Chemical Engineering from UC Berkeley in 2017, with a minor in Chemistry.

Fang Liu is an Assistant Professor in Chemistry at Emory University. Prior to that, she did postdoctoral research in the group of Professor Heather J. Kulik at MIT (2017–2020) and obtained her Ph.D. in Chemistry at Stanford in the group of Todd J. Martínez in 2017.

Heather J. Kulik is an Associate Professor in the Department of Chemical Engineering at MIT. She received her B.E. in Chemical Engineering from the Cooper Union in 2004. She obtained her Ph.D. from the Department of Materials Science and Engineering at MIT in 2009 in the

group of Nicola Marzari. She completed postdocs at Lawrence Livermore with Felice Lightstone and Stanford with Todd Martínez, prior to joining MIT as a faculty member in 2013.

ACKNOWLEDGMENT

The authors acknowledge support by the Office of Naval Research under grant numbers N00014-17-1-2956, N00014-18-1-2434, and N00014-20-1-2150, DARPA grant D18AP00039, the Department of Energy under grant numbers DE-SC0018096 and DE-SC0012702 as well as a MolSSI fellowship (grant no. ACI-1547580) to F.L., and the National Science Foundation under grant numbers CBET-1704266 and CBET-1846426. A.N. was partially supported by a National Science Foundation Graduate Research Fellowship under Grant #1122374. This work was supported by an AAAS Marion Milligan Mason Award. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

References

1. Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017,** *121*, 8939-8954.
2. Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018,** *9*, 1064-1071.
3. Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019,** *10*, 7913-7922.
4. Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020,** *6*, 513-524.
5. Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A. Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chem. Rev.* **2018,** *119*, 2453-2523.
6. Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019,** *119*, 6561-6594.
7. Weymuth, T.; Reiher, M. Inverse Quantum Chemistry: Concepts and Strategies for Rational Compound Design. *Int. J. Quantum Chem.* **2014,** *114*, 823-837.
8. Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020,** *10*, 2354-2377.

9. Griffith, J. S.; Orgel, L. E. Ligand-Field Theory. *Q. Rev., Chem. Soc.* **1957,** *11*, 381-393.
10. Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992,** *114*, 10024-10035.
11. Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. Molecular Mechanics for Coordination Complexes: The Impact of Adding d-Electron Stabilization Energies. *J. Am. Chem. Soc.* **1995,** *117*, 8407-8415.
12. Li, P.; Merz, K. M. Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **2017,** *117*, 1564-1686.
13. Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Mueller-Westerhoff, U. T. An Intermediate Neglect of Differential Overlap Technique for Spectroscopy of Transition-Metal Complexes. Ferrocene. *J. Am. Chem. Soc.* **1980,** *102*, 589-599.
14. Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R. Mof-Ff – a Flexible First-Principles Derived Force Field for Metal-Organic Frameworks. *Phys. Status Solidi B* **2013,** *250*, 1128-1141.
15. Comba, P.; Zimmer, M. Molecular Mechanics and the Jahn-Teller Effect. *Inorg. Chem.* **1994,** *33*, 5368-5369.
16. Deeth, R. J. The Ligand Field Molecular Mechanics Model and the Stereoelectronic Effects of d and S Electrons. *Coord. Chem. Rev.* **2001,** *212*, 11-34.
17. Schäffer, C. E.; Jørgensen, C. K. The Angular Overlap Model, an Attempt to Revive the Ligand Field Approaches. *Mol. Phys.* **1965,** *9*, 401-412.
18. Cramer, C. J.; Truhlar, D. G. Density Functional Theory for Transition Metals and Transition Metal Chemistry. *Phys. Chem. Chem. Phys.* **2009,** *11*, 10757-10816.
19. Kulik, H. J. Perspective: Treating Electron over-Delocalization with the DFT+U Method. *J. Chem. Phys.* **2015,** *142*, 240901.
20. Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L. Beyond Density Functional Theory: The Multiconfigurational Approach to Model Heterogeneous Catalysis. *ACS Catal.* **2019,** *9*, 8481-8502.
21. Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013,** *135*, 7296-7303.
22. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012,** *52*, 2864-2875.
23. von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem., Int. Ed.* **2018,** *57*, 4164-4169.
24. Fey, N.; Orpen, A. G.; Harvey, J. N. Building Ligand Knowledge Bases for Organometallic Chemistry: Computational Description of Phosphorus (III)-Donor Ligands and the Metal–Phosphorus Bond. *Coord. Chem. Rev.* **2009,** *253*, 704-722.
25. Gugler, S.; Janet, J. P.; Kulik, H. J. Enumeration of De Novo Inorganic Complexes for Chemical Discovery and Machine Learning. *Mol. Sys. Des. Eng.* **2020,** *5*, 139-152.
26. Hauser, A. Ligand Field Theoretical Considerations. In *Spin Crossover in Transition Metal Compounds I*; Springer: 2004; pp 49-58.
27. Deeth, R. J.; Fey, N. The Performance of Nonhybrid Density Functionals for Calculating the Structures and Spin States of Fe (II) and Fe (III) Complexes. *J. Comput. Chem.* **2004,** *25*, 1840-1848.

28. Ioannidis, E. I.; Kulik, H. J. Towards Quantifying the Role of Exact Exchange in Predictions of Transition Metal Complex Properties. *J. Chem. Phys.* **2015,** *143*, 034104.
29. Ganzenmüller, G.; Berkaïne, N.; Fouqueau, A.; Casida, M. E.; Reiher, M. Comparison of Density Functionals for Differences between the High- (T2g5) and Low- (A1g1) Spin States of Iron(II) Compounds. IV. Results for the Ferrous Complexes [Fe(L)('NHS4')]. *J. Chem. Phys.* **2005,** *122*, 234321.
30. Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.* **2019,** *58*, 10592-10606.
31. Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017,** *8*, 5137-5152.
32. Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput.* **2019,** *15*, 2331-2345.
33. Liu, F.; Duan, C.; Kulik, H. J. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **2020,** *11*, 8067-8076.
34. Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal- Oxo Intermediate Formation. *ACS Catal.* **2019,** *9*, 8243-8255.
35. Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Frameworks Ecosystems. *Nat. Commun.* **2020,** *11*, 4068.
36. Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018,** *57*, 13973-13986.
37. Kulik, H. J. Making Machine Learning a Useful Tool in the Accelerated Discovery of Transition Metal Complexes. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020,** *10*, e1439.
38. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014,** *1*, 140022.
39. Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017,** *4*, 170193.
40. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B* **2016,** *72*, 171-179.
41. Martinez, J. L.; Lutz, S. A.; Beagan, D. M.; Gao, X.; Pink, M.; Chen, C.-H.; Carta, V.; Moënne-Loccoz, P.; Smith, J. M. Stabilization of the Dinitrogen Analogue, Phosphorus Nitride. *ACS Cent. Sci.* **2020**.
42. van der Maaten, L.; Hinton, G. Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008,** *9*, 2579-2605.
43. Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal–Organic Frameworks. *Nat. Chem.* **2012,** *4*, 83.
44. Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: Core Mof 2019. *J. Chem. Eng. Data* **2019,** *64*, 5985-5998.

45. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012,** *108*, 058301.
46. Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980,** *4*, 359.
47. Janet, J. P.; Gani, T. Z. H.; Steeves, A. H.; Ioannidis, E. I.; Kulik, H. J. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* **2017,** *56*, 4898-4910.
48. Taylor, M. G.; Yang, T.; Lin, S.; Nandy, A.; Janet, J. P.; Duan, C.; Kulik, H. J. Seeing Is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *J. Phys. Chem. A* **2020,** *124*, 3286-3299.
49. Schröder, D.; Shaik, S.; Schwarz, H. Two-State Reactivity as a New Concept in Organometallic Chemistry. *Acc. Chem. Res.* **2000,** *33*, 139-145.
50. Liao, P.; Getman, R. B.; Snurr, R. Q. Optimizing Open Iron Sites in Metal−Organic Frameworks for Ethane Oxidation: A First-Principles Study. *ACS Appl. Mater. Interfaces* **2017,** *9*, 33484-33492.
51. Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018,** *4*, 268-276.
52. Iovanac, N. C.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**.
53. Gal, Y.; Ghahramani, Z. In *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, international conference on machine learning, Year; pp 1050-1059.
54. Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2018**.
55. Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019,** *15*, 906-915.
56. Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017,** *8*, 3192-3203.
57. Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017,** *19*, 10978-10985.
58. Leardi, R. Genetic Algorithms in Chemistry. *J. Chromatogr. A* **2007,** *1158*, 226-233.
59. Venkatraman, V.; Abburu, S.; Alsberg, B. K. Artificial Evolution of Coumarin Dyes for Dye Sensitized Solar Cells. *Phys. Chem. Chem. Phys.* **2015,** *17*, 27672-27682.
60. Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for De Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012,** *134*, 8885-8895.
61. Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019,** *119*, 6595-6612.
62. Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of Compounds for Machine-Learning Prediction of Physical Properties. *Phys. Rev. B* **2017,** *95*, 144110.
63. Herbol, H. C.; Hu, W.; Frazier, P.; Clancy, P.; Poloczek, M. Efficient Search of Compositional Space for Hybrid Organic–Inorganic Perovskites via Bayesian Optimization. *npj Comput. Mater.* **2018,** *4*, 51.

64. Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A. In *Parallel and Distributed Thompson Sampling for Large-Scale Accelerated Exploration of Chemical Space*, Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research; Doina, P.; Yee Whye, T., Eds.; PMLR, Year; pp 1470-1479.
65. Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018,** *4*, 1134-1145.
66. Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016,** *7*, 11241.
67. Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning. *Adv. Mater.* **2018,** *30*, 1702884.
68. Okamoto, Y. Applying Bayesian Approach to Combinatorial Problem in Chemistry. *J. Phys. Chem. A* **2017,** *121*, 3299-3304.
69. Matsuda, Y.; Tanaka, K.; Okada, M.; Takasu, Y.; Morita, M.; Matsumura-Inoue, T. A Rechargeable Redox Battery Utilizing Ruthenium Complexes with Non-Aqueous Organic Electrolyte. *J. Appl. Electrochem.* **1988,** *18*, 909-914.
70. Forrester, A. I. J.; Keane, A. J. Recent Advances in Surrogate-Based Optimization. *Prog. Aeronaut. Sci.* **2009,** *45*, 50-79.
71. Keane, A. J. Statistical Improvement Criteria for Use in Multiobjective Design Optimization. *AIAA J.* **2006,** *44*, 879-891.
72. Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press: 2006.
73. Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017,** *29*, 9436-9444.
74. Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019,** *573*, 251-255.