Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry

Aditya Nandy[1,2,#], Chenru Duan[1,2,#], Jon Paul Janet[1], Stefan Gugler[1,3], 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*
[3]*Laboratorium für Physikalische Chemie, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland*
[#]These authors contributed equally.

ABSTRACT: Machine learning the electronic structure of open shell transition metal complexes presents unique challenges, including robust and automated data set generation. Here, we introduce tools that simplify data acquisition from density functional theory (DFT) and validation of trained machine learning models using the molSimplify automatic design (mAD) workflow. We demonstrate this workflow by training and comparing the performance of LASSO, kernel ridge regression (KRR), and artificial neural network (ANN) models using heuristic, topological revised autocorrelation (RAC) descriptors we have recently introduced for machine learning inorganic chemistry. On a series of open shell transition metal complexes, we evaluate set aside test errors of these models for predicting the HOMO level and HOMO-LUMO gap. The best performing models are ANNs, which show 0.15 and 0.25 eV test set mean absolute errors on the HOMO level and HOMO-LUMO gap, respectively. Poor performing KRR models using the full 153-feature RAC set are improved to nearly the same performance as the ANNs when trained on down-selected subsets of 20-30 features. Analysis of the essential descriptors for HOMO and HOMO-LUMO gap prediction as well as comparison to subsets previously obtained for other properties reveals the paramount importance of non-local, steric properties in determining frontier molecular orbital energetics. We demonstrate our model performance on diverse complexes and in the discovery of molecules with target HOMO-LUMO gaps from a large 15,000 molecule design space in minutes rather than days that full DFT evaluation would require.

1. Introduction

Transition metal complexes are promising as functional materials, e.g., as spin crossover materials[1-6] that change properties in response to light, heat or small molecules, as photosensitizers[7] for solar cells, and they are also highly active, selective homogenous catalysts[8-12]. The open-shell transition metal (TM) centers of TM complexes are characterized by unpaired electrons with differing spin and spatial symmetry that can be tuned by surrounding organic ligands. These interactions typically require a fully quantum mechanical description, and thus the resulting properties are both what gives these complexes great promise as functional materials and catalysts but also makes their rational design challenging.

Computational high-throughput screening and design have become increasingly prominent[13-17] thanks to increases in computing power and algorithmic developments. Nevertheless, challenges remain for high-throughput discovery in inorganic chemistry. Fewer affordable methods such as force fields or semi-empirical methods have been developed for inorganic chemistry[18] in comparison to organic chemistry, necessitating first principles simulation. However, transition metal chemistry is also sensitive[19-23] to the approximations made in the most widely used first principles simulation techniques (i.e., density functional theory or DFT). Inorganic chemistry discovery tools lag behind both solid state materials, where large databases are available for repurposing and discovering new materials[24], which open source codes such as AFLOW[25-26], ASE[27], and pymatgen[28] enable. A number of cheminformatics tools, such as RDkit[29] or OpenBabel[30], and large databases[31-32] of compact molecular representations assist the automated simulation of organic molecules, but they are not straightforwardly applicable to inorganic chemistry. With those challenges in mind, we introduced molSimplify[33], the first open source toolkit for automating discovery in inorganic chemistry, using a divide and

conquer approach for transition metal complexes that uses cheminformatics tools[30] and force fields for organic components and augments them with databases of quantum-mechanically-derived rules for the metal-organic bond. These approaches enabled us to carry out automated, high-throughput discovery in inorganic chemistry[11, 33-36].

Although DFT is the method of choice for computational materials discovery, efficient chemical space exploration in the large combinatorial space spanned by transition metal complexes necessitates accelerated discovery techniques. To address this challenge, we introduced the first machine learning (ML) models to predict properties of open shell transition metal complexes. Although ML has been increasingly widely used in property prediction[37-41] (e.g., potential development[42-45]) and materials discovery[46-50] especially for organic chemistry, its successful application to inorganic chemistry has not been as prevalent. Development of ML models for inorganic chemistry would offer the advantages of being able to predict properties at multiple levels of theory[51-52] or differing functional blends[53] and also making property predictions in a matter of seconds that would normally take hours or days with DFT.

We thus used our molSimplify[33] toolkit to generate DFT training data for ML models of inorganic chemistry and then incorporated these models into molSimplify. Our artificial neural networks (ANNs) and kernel ridge regression (KRR) models predicted i) spin state ordering to 1-3 kcal/mol accuracy[53], ii) sensitivity[20-21, 23, 35, 54-55] of spin state ordering to DFT functional[53], iii) redox or ionization potential to 0.2 eV accuracy[56], and iv) equilibrium DFT metal-ligand bond lengths to 0.01-0.03 Å in a spin and oxidation state dependent manner. We employed these models for materials design through a genetic algorithm optimization[57-58] that estimated model uncertainty[59-60] through heuristic distance metrics[53] we determined to be superior to ensemble-based diagnostics in discovery[58]. By relying on geometry-free representations[61-64] that we

tailored[53, 56] for inorganic chemistry, we were also able to replace our database-driven approach to structure generation[33] with a machine learning model[35].

We showed that representations tailored for inorganic chemistry[53, 56] outperformed whole-molecule, geometry-dependent descriptors[65] by as much as an order of magnitude[53, 56] on our modest (ca. 2000-3000 equilibrium geometry) data sets. We rationalized these observations by noting that large changes in system size (e.g., from 37 to 151 atoms) could have a limited effect on the target property (e.g., spin splitting) if the ligand modification was metal-distant[56]. Training sets used in organic chemistry machine learning[66-67] have been historically much more homogeneous in size (e.g., closed shell singlets comprised of no more than 9 heavy C, N, O, or F atoms[66]) than is practical for inorganic chemistry where ligand structure necessitates larger and more variable system sizes. Thus, a chief outstanding question in the field of machine learning quantum chemistry is whether developed representations are suitably transferable.

In this work, we extend our prior ML models of transition metal chemistry to study whole-complex frontier molecular orbital energetics, including the placement of the highest occupied molecular orbital (HOMO) as well as the gap between the HOMO and the lowest unoccupied molecular orbital (LUMO, i.e., HOMO-LUMO gap). Frontier molecular orbital energetics provide essential insight into chemical reactivity[68-69] and dictate optical and electronic properties[70-71]. Although in pure DFT only the HOMO strictly carries meaning[72], significant theoretical work has motivated the significance of the HOMO-LUMO gap[73-74], especially in hybrid DFT[75-76]. The rest of this manuscript is as follows. In section 2, we present an overview of new automation features of our open source software that both enable machine learning model training by streamlining high-throughput simulation and leverage machine learning models for discovery. In section 3, we present the computational details of the data sets employed in this

work. In section 4, we develop and interpret machine learning models for predicting frontier molecular orbital energies of inorganic complexes, and we compare this prediction task to other properties of inorganic complexes. Finally, in section 5, we provide our conclusions.

2. Software Developments.

2a. Automated Structure Validation.

One challenge in first-principles screening of TM complexes is confirming that large numbers of calculations completed successfully. For example, when thousands of calculations are carried out on a computing cluster, a fundamental issue for inorganic complex screening is whether the expected coordination geometry was maintained during the optimization. For example, weakly bound ligands may detach from the metal center during the optimization, either due to poor initial geometry or electronic structure that prevents identification of a stable local minimum on the potential energy surface near the expected geometry. To ensure that the calculated properties are meaningful in high-throughput screening, automated tools are needed for monitoring the geometries of inorganic complexes during and after the optimization process (see example complex in Figure 1). To perform this monitoring, we construct a set of metrics based on four aspects of complex geometry: the coordination number (CN), the shape of first coordination shell (FCS), the degree of ligand distortion, and the orientation of linear ligands. These metrics enable us to measure the difference between an observed geometry and target geometry. In our workflow, we compare this difference to threshold values to determine if a geometry has become invalid.

Determining CN. Rigorous CNs in inorganic complexes should be determined by detailed analysis of the electron density to identify atoms bonded to the metal center. However, a simpler approach that is amenable to rapid screening is to identify candidate coordinating atoms from

interatomic distances. We consider an atom, A_i, to be a candidate coordinating atom to the metal center, M, if the M-A distance is less than 1.35x the sum of the covalent radii of the two species or, at most, < 2.8 Å. In practice, this distance cutoff is intended to be overly inclusive (i.e., for F, 1.25 Å and C 0.75 Å, the cutoff is 2.72 Å). Even in the case of monohapto ligands, short intraligand bonds (e.g., C-N) can cause multiple atoms from the same ligand to be detected by this inclusive cutoff. In cases where the initial CN is higher than expected, the set of candidate coordinating atoms, $\{C_i\}$, is checked for any two atoms that are determined to be bonded and within the same ligand, as judged by 1.15x the sum of the covalent radii. The CN check is particularly useful in detecting geometries where the CN is smaller than expected due to detached ligands.

Shape of the FCS. The FCS shape can be uniquely determined by the angles formed between coordinating atoms, *C*, of *i*th and *j*th ligands with M at the vertex, i.e., $\theta(C_i\text{-M-}C_j)$. In the octahedral geometries studied in this work, $\theta(C_i\text{-M-}C_j)$ should be 90 or 180°, whereas they should all be approximately 109° in a tetrahedral geometry. Use of a reference angle guarantees the flexibility of the FCS check and does not restrict it to the octahedral complexes studied in this work. For a given FCS, the angular deviation, $\Delta\theta(C_i\text{-M-}C_j)$, can be measured by comparing the actual and ideal $\theta(C_i\text{-M-}C_j)$ values. The $\{C_i\}$ obtained from the initial CN check may contain non-bonded atoms, especially for multidentate ligands. Thus, we identify a subset $\{c_i\}$ of confirmed coordinating atoms by determining the optimal subset of $\{C_i\}$ that minimizes the average angular deviation, $\text{avg}(\Delta\theta(C_i\text{-M-}C_j))$, i.e., by varying the atoms over which the average is computed. The maximum angular deviation $\max(\Delta\theta(C_i\text{-M-}C_j))$ over the optimal subset $\{c_i\}$ is retained as an additional check (see structure example in Figure 1). From the $\{c_i\}$ subset, we obtain the metal-ligand bond lengths, $d(\text{M-}C_i)$, which are used to compute the maximum

difference in bond lengths, max(Δd), of ligands that should be equivalent. For octahedral complexes with identical equatorial ligands, the largest difference in equatorial metal-ligand bond lengths, max(Δd_{eq}), quantifies the extent to which symmetry is preserved.

Ligand distortion. During optimization, changes in electronic structure can have an unexpected effect on an individual ligand geometry. To detect these cases, each ligand is extracted from the optimized structure and superimposed with its initial coordinates using the Kabsch algorithm[77], which calculates the optimal rotation matrix that minimizes the root mean squared deviation (RMSD). The maximum RMSD, max(RMSD), over all coordinating ligands is used as a metric for ligand distortion (Figure 1).

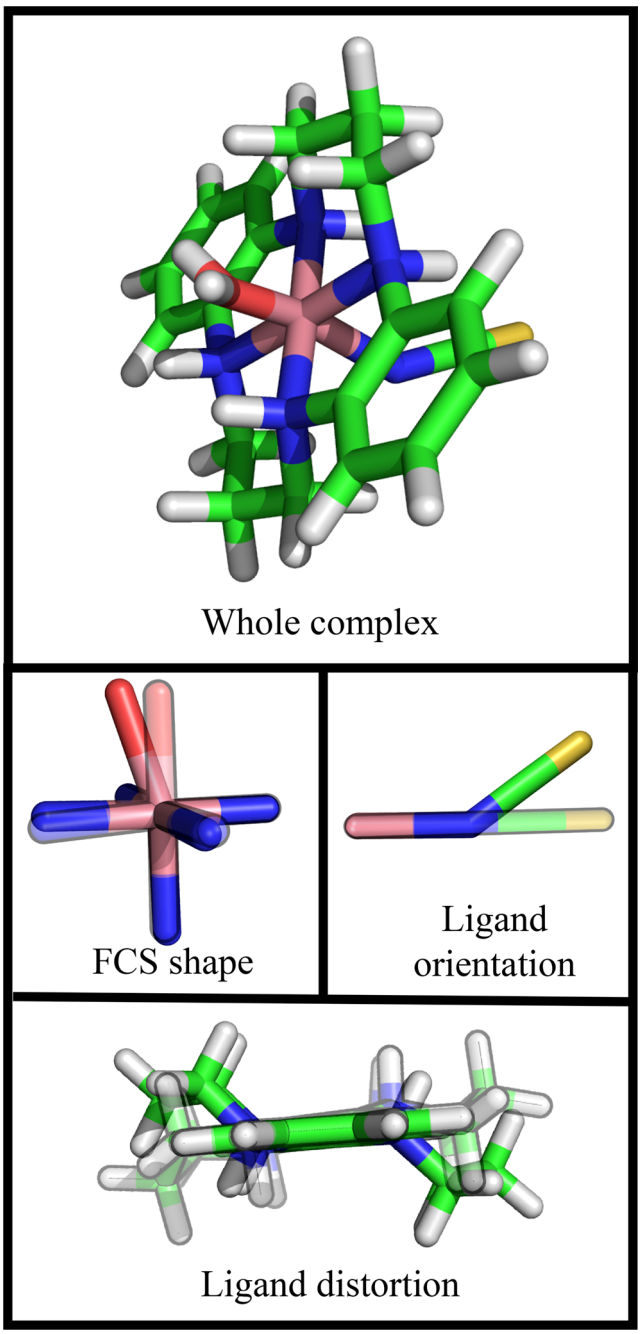


Figure 1. (top) Example octahedral transition metal complex to illustrate common failure modes for the geometry check. (middle and bottom) FCS shape, ligand orientation and ligand distortion checks. The FCS and ligands are extracted from the whole complex (opaque sticks shown in middle and bottom panes) to compare with idealized counterparts (translucent, black outline).

Linear ligand orientation. For a ligand with a triple bond between the coordinating atom, A, and its nearest neighbor, B (e.g., as determined by the isolated ligand's SMILES notation), we expect linear coordination, i.e., θ(M-A-B) = 180°. However, this initially linear geometry is not

constrained in any way during a DFT calculation and can easily deviate from this expected value (Figure 1). We compute the deviation from linearity, $\Delta\theta$(M-A-B), and quantify any deviations by the averaged, avg($\Delta\theta$(M-A-B)), and maximum, max($\Delta\theta$(M-A-B)), deviations over coordinating linear ligands.

The combined mAD geometry check workflow begins by obtaining a set of candidate coordinating atoms $\{C_i\}$ to assess ligand dissociation. For intact cases, a subset of confirmed coordinating atoms is obtained by comparing to the reference angle of an idealized geometry. (Figure 2). At this step, metrics characterizing the shape of the FCS, e.g., max($\Delta\theta(C_i\text{-M-}C_j)$), avg($\Delta\theta(C_i\text{-M-}C_j)$), max($\Delta d$), and max($\Delta d_{eq}$) are calculated (Figure 2). Once the $\{c_i\}$ is determined, we separate the complex into the metal center and ligands and compute ligand structural properties, i.e., max(RMSD), avg($\Delta\theta$(M-A-B)), and max($\Delta\theta$(M-A-B)) (Figure 2). Note that geometries with $\{c_i\}$ smaller than the expected value (i.e., due to ligand detachment) are abandoned prior to these additional checks (Figure 2).

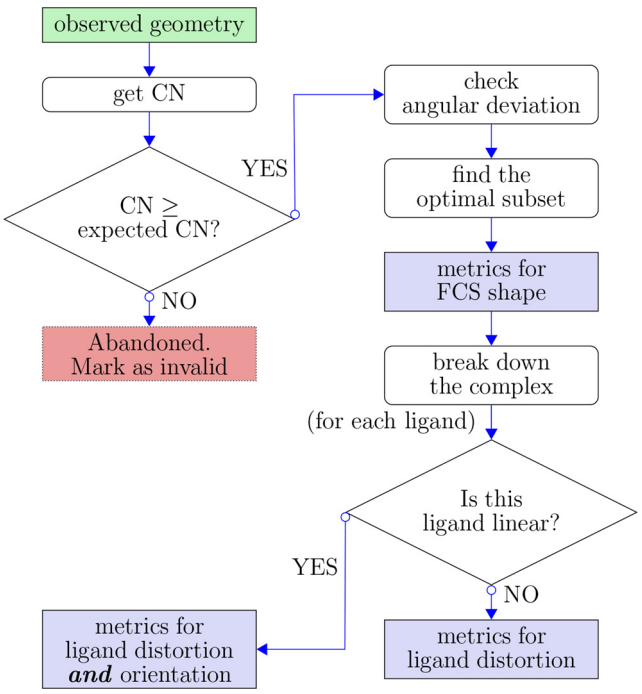


Figure 2. Workflow for mAD geometry check function. Starting from the observed geometry

(shown in green), four different checks (CN, FCS shape, ligand distortion, linear ligand orientation) are carried out to compute the corresponding metrics (shown in light blue). Structures found with CN smaller than the expected CN are directly categorized as bad and abandoned for further checks (shown in red).

To assign a *bad* or *good* label for each geometry, a cutoff for each metric is required. A geometry is categorized as *bad* if there exists one metric with a value larger than the cutoff, otherwise it is labeled as *good*. The current cutoffs were empirically determined for octahedral complexes based on trial and error categorization of hundreds of structures generated in prior work[53, 56] (Table 1). Although these cutoffs are based on chemical intuition, they are consistently applied to the data set without bias. Further refinement to this approach is to go beyond structure and identify signatures of *bad* structures directly from electronic properties, which is currently underway in our group.

Table 1. Geometry check metric cutoffs used for an octahedral transition metal complex. The angular ligand distortion quantities are only computed for ligands that are expected to be linear.

Coordination number			
CN			
6			
Shape of the first coordination sphere			
avg($\Delta\theta$(C_i-M-C_j))	**max($\Delta\theta$(C_i-M-C_j))**	**max(Δd)**	**max(Δd_{eq})**
12°	22.5°	1.00 Å	0.35 Å
Ligand distortion			
max(RMSD)	**avg($\Delta\theta$(M-A-B))**	**max($\Delta\theta$(M-A-B))**	
0.30 Å^2	20°	28°	

2b. Automated Design Software.

We have developed open-source Python packages for first-principles and ML-driven screening of TM complexes: i) molSimplify[33, 35, 78], a 3D-structure generation and manipulation toolbox that is tailored for TM complexes, and ii) molSimplify Automatic Design (mAD), a

molSimplify extension for materials design. The molSimplify code enables commandline generation of accurate initial geometries of inorganic complexes by preoptimizing organic ligands with force fields and attaching ligands to predefined coordination geometries at database[33, 35, 78] or machine learning (i.e., neural network) predicted[53] bond lengths. Both codes are available on github[79-80] and on our website[81]. Although we recently demonstrated the mAD code for surrogate model-driven spin crossover (SCO) design[58], we now introduce the logic and core functions of mAD that enabled that approach. mAD can be used as a general purpose tool to streamline and mitigate some of the challenges in high-throughput inorganic complex screening (Figure 3).

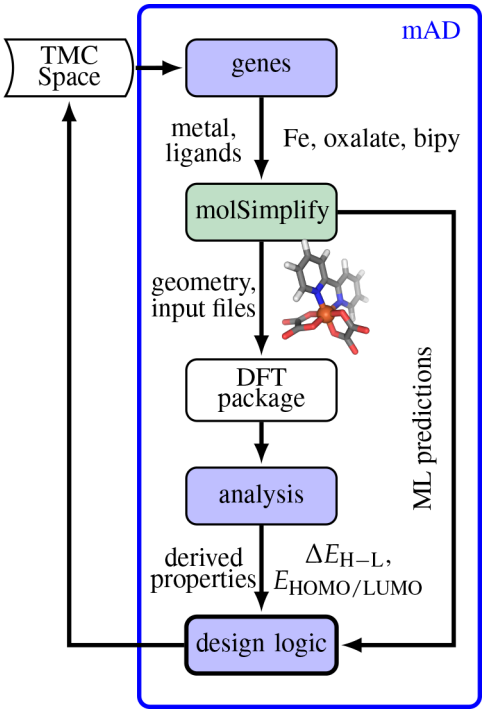


Figure 3. Workflow for mAD with key mAD components shown in blue: transition metal chemical (TMC) space is discretized into genes, converted to 3D geometries and optional input files using molSimplify (shown in green), and properties are either calculated using an external DFT package using molSimplify-generated input files or internal ML models. Results are analyzed and guide iterative discovery using the design logic in mAD. An example 3D geometry is shown in inset.

The mAD code enables evolutionary-algorithm based materials optimization[58] of TM

complexes. A TM complex is described as a set of *genes*, where each *gene* refers to a distinct design element to evaluate and optimize. For spin splitting energetics, the genes of a TM complex correspond to a metal, oxidation state, and ligands. For example, if the design space is restricted to homoleptic complexes, there will be only one ligand gene, whereas relaxing symmetry (e.g., to distinct axial and equatorial ligands) increases the number of ligand genes to two or more. For catalysis, the appropriate gene choice could correspond to a fixed choice of metal and equatorial ligands while the axial species are obtained from a list of relevant reactive intermediates. The user provides to the program a list of possible identities for each ligand gene, which can be either drawn from the common ligands built into molSimplify or input as SMILES. From the list of possible ligands, a population of genes that will encode a set of complexes can be generated randomly or by user-specified indices for the desired ligands.

Once a population of genes is assembled, mAD can either use the molSimplify neural network to estimate properties (e.g. spin splitting or bond lengths) or to generate initial geometries and input files for DFT calculations, with native support for TeraChem (Figure 3). In developing mAD, we have used a highly modular structure that supports easy extension. Since molSimplify supports generating other electronic structure input files (e.g., Q-Chem and Orca), the only requirements to switch mAD to other codes is to modify job submission scripts and revise post-processing Python functions. Generated jobs are tracked in plain text files within a job subfolder and submitted to a computing queue, with native support for SGE and SLURM. The job script and run parameters are both customizable by the user either by pointing to a custom script in the mAD input file or by editing the initially generated queue interface files, which are then reused by mAD. Given the flexibility of the input structure, mAD runs can be carried out with either single point energy evaluations at guessed geometries, as in our previous

work[38], or with full geometry optimizations. For time-consuming full geometry optimizations, mAD contains a number of helpful features to ensure that these jobs complete successfully, as described next. Ongoing efforts in our group are focused on increasing code and queue support of molSimplify and mAD, and interested users are encouraged to keep up to date on these improvements through github[79-80] and our website[81].

The mAD process can be run persistently in the background or with a user-scheduled cron and requires only standard user permissions. Once running, the mAD code monitors the submitted jobs by periodically querying the queuing system and inspecting completed jobs for convergence and completion at a valid structure (see sec. 2a). The monitoring intervals and maximum time for mAD to run can be specified by the user. Unfinished calculations that pass geometry inspection are automatically restarted by mAD up to a maximum number of three times by default, but this can be customized by the user in the mAD input file. Completed calculations that have *good* geometries are inspected for deviations of the expectation value of the $\hat{S}^2$ operator (i.e., $\langle \hat{S}^2 \rangle$) from the expected value, indicating broken symmetry between spin up and spin down orbitals that suggests the total spin assigned in the calculation is not well-defined within approximate DFT. Once a calculation passes both of those tests, it is grouped with related calculations needed to calculate a target property, e.g., high- and low-spin energies of the same complex are needed to calculate a spin-splitting energy. The specified target properties for each complex are collected and used to inform future calculations, e.g., through fitness function evaluation. The mAD code generates descriptor representations and computes the molSimplify-built-in ML model prediction and distance-based uncertainty measures, where applicable, for each attempted calculation to enable easy model retraining or extension. Data provenance is established by storing the configuration and software versions used in all calculations in the

mAD run output files.

Once properties are available for all sets of genes (i.e., TM complexes) in the current generation or have been excluded due to calculation failures, mAD can propose new candidate materials. The current version of mAD supports a basic genetic algorithm that has the capacity to use either DFT or ML predictions. For ML predictions, mAD supports balancing property optimization with distance from training data to conduct design in large, varied design spaces while maintaining model confidence[58]. This approach could also be used for active learning, which has shown promise[82-83] in materials optimization.

The mAD code includes automation of subsequent calculations that depend on an initial geometry optimization. Some examples of supported calculations are single point energies with implicit solvent incorporated, Hessian calculations, or calculations carried out with a different DFT functional (e.g., varied HF exchange fraction). These derivative jobs can be automatically generated and submitted by mAD using both the parent optimized geometry as well as converged wave functions as initial guesses, where appropriate. In HF exchange fraction resampling, for example, the default behavior after converging an initial optimization at B3LYP (20% HF) is to scan the range 25% to 30% and then 15% to 0% in 5% increments, always using the converged wave function of the previous step in order to ensure the same electronic state is converged, regardless of HF exchange fraction used.

3. Computational Details.

We trained all ML models on electronic structure properties of octahedral inorganic complexes obtained from previous work[56]. Here, we expand on our prior predictions[53, 56] of spin-state splitting, ionization/redox potential, and bond lengths to now predict highest occupied (and lowest unoccupied molecular orbital (HOMO and LUMO) energies and gaps. We selected the

redox data set used in prior work to predict adiabatic ionization and redox potentials rather than larger data sets used for spin splitting prediction[53, 56]. The redox prediction models are physically most similar to our present goal of predicting HOMO and LUMO energies, where the former should correspond to the vertical ionization energy[72] in an exact functional.

The original data set consists of geometry optimized inorganic complexes with M(II)/M(III) oxidation states of Co, Cr, Fe, and Mn metals in high-spin (HS) and low-spin (LS) states initially intended for ionization potential and redox potential calculation[56]. The HS-LS definitions in prior work were: quintet-singlet for d^6 Co(III)/Fe(II), sextet-doublet for d^5 Fe(III)/Mn(II), quintet-triplet for d^4 Mn(III)/Cr(II), and quartet-doublet for both d^3 Cr(III) and d^7 Co(II). The lack of isolated Mn(III) singlet gas phase ions[84] originally motivated[53] our definition for the d^4 complexes. However, we have since revised the d^4 LS definition to singlet to increase correspondence with the d^6 systems. The ligands in this redox data set[56] were CO, pyridine, water, furan, and methyl isocyanide. The symmetry was restricted to a single equatorial ligand identity with up to two distinct axial ligands (i.e., $M(L_1)_4L_2L_3$) for a theoretical total of 1200 compounds. Considering the multiple states required to obtain a redox potential, this space corresponded to 300 redox potential evaluations defined in terms of the M(II) ground state and ionization to an M(III) electronic state with a spin multiplicity that differed only by one. In that work[56], only 185 of the 300 theoretical redox potential evaluations were possible due to challenges in completing one of the three needed calculations.

In addition to using a revised singlet definition for Mn(III)/Cr(II) in the present work, we also perform the newly introduced geometry check to ensure the fidelity of the DFT geometry optimized structures (see sec. 2a). Complexes with geometries that failed the geometry check (217 of 1200) were eliminated from the set (see Supporting Information). Structures with

deviations of $\langle S^2 \rangle$ > 1.0 from expected values (109 of 1200) were also eliminated (see Supporting Information). Our final revised redox data set consists of 874 structures, 555 of which were also used for ionization and redox potential prediction in Ref. [56]. The 319 additional data points correspond to those that either differ in spin state definition or were not part of the necessary calculations for the ionization potential prediction.

Although the HOMO-LUMO properties reported in this work were generated during those prior simulations, we briefly review the computational details of those calculations. Structures were generated by first building an initial geometry with molSimplify[91] and then carrying out DFT geometry optimizations using the TeraChem[85-86] graphical processing unit (GPU)-accelerated quantum chemistry package. The B3LYP[87-89] hybrid DFT functional was used in combination with the LANL2DZ[90] effective core potential for transition metals and the 6-31G* basis for all other atoms. The effect of using a modest basis set, which enables larger data set generation for ML models, was found to be limited in prior work on the relative energies of interest[55]. All calculations were spin-unrestricted, with virtual and occupied orbitals level-shifted[91] by 1.0 and 0.1 eV, respectively.

It is well known that pure Kohn-Sham DFT underestimates HOMO-LUMO gaps[92-93] due to both delocalization error and limits of the exact theory.[94-95] However, HOMO-LUMO gaps are on a firmer foundation in generalized Kohn-Sham (KS) DFT[76] (i.e., the hybrid functionals used in this work). In open shell systems, ionization processes to which FMOs and their energies should correspond can originate in either majority or minority spin. In this work, we adopt the convention of training models on the strict energetic definition of HOMO and LUMO. In practice, this means that in some metals, we are training on the gap between a spin up (α) HOMO and spin down (β) LUMO (quintet Fe(III)$(CO)_6$), whereas in others, we are learning the

gap between the α HOMO and LUMO (quartet Mn(III)$(CO)_6$). This approach is consistent with our observations that ML test set errors on redox and ionization potentials from a mixture of spin states were no worse than models trained on only a single spin state[56].

4. Machine Learning Inorganic Chemistry Properties.

4a. Overview of Feature Sets and Models.

Feature sets. The manner in which molecular structures are converted into numerical inputs is a critically important aspect of atomistic machine learning[96-98]. We have observed good performance of machine learning models trained on metal-centric (i.e., containing only information about atoms in the first few coordination spheres[99]) topological heuristic representations of inorganic complexes[53]. Avoiding incorporation of explicit geometry dependence enables metal-ligand bond length prediction by molSimplify[33] structure generation[35] routines and makes straightforward the prediction of properties that depend on multiple geometries.

We recently introduced a systematic approach to inorganic chemistry featurization that blends metal-centric and whole-complex topological properties in a feature set referred to as revised autocorrelation functions (RACs)[56]. These RACs, variants of graph autocorrelations[61-64], are sums of products and differences of atomic properties (i.e., electronegativity, nuclear charge, topology, covalent radius, and identity) (Figure 4). We demonstrated these RACs to be predictive featurizations for inorganic chemistry properties, such as spin-state splitting and ionization/redox potential. Over all possible origins (i.e., metal-centered, mc, or ligand-centered, lc) and definitions, there are 42*d*+30 theoretical RAC features, where *d* is the maximum distance in bonds through which two atoms are correlated in a single descriptor (mc examples shown in Figure 4).[56] As described in previous work[56], this number of features arises from the fact that

there are 6*d*+6 RACs for each of the five atomic property product RACs (i.e., 30*d*+30), and there are no zero-depth difference descriptors, only three non-trivial start/scope definitions, and *I* is excluded, giving 12*d* difference RACs for a total of 42*d*+30 productive and difference RACs.

A given depth cutoff does not mean that whole-molecule information is excluded, but it does allow the user to choose not to directly correlate in a single feature the product of properties of two atoms farther apart than a certain topological distance. Indeed, we found benefit in limiting *d* to three bond paths in KRR model training, thereby making the theoretical RAC space 156 features in size, 5 of which are constant for the octahedral complexes studied in this and prior work. The full definition of the RAC representation also included oxidation state, spin state, denticity and Hartree-Fock (HF) exchange for a total of 155 features.[56]

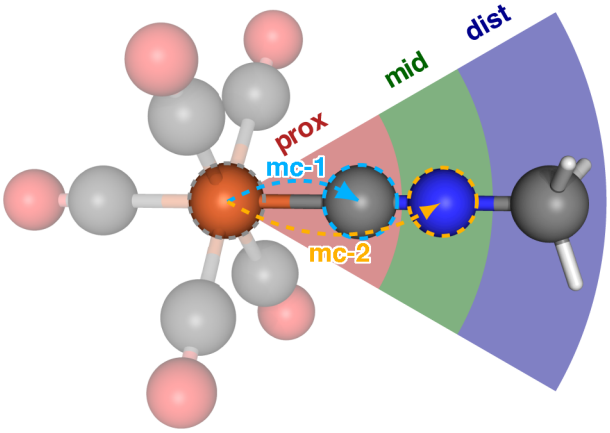


Figure 4. Example RACs depicted on the structure of $Fe(CO)_5$(misc) in ball and stick representation (iron is brown, oxygen is red, nitrogen is blue, carbon is gray and hydrogen in white sticks, CO ligands are semitransparent). The example paths shown are for depth one and two mc RACs (mc-1 in blue or mc-2 in orange, respectively) shown only on the misc ligand. We characterize RACs by locality relative to metal center: proximal (prox in red for metal and first shell as in mc-1), intermediate (mid in green for second shell as in mc-2), and distal (dist in blue for third shell or beyond).

In the present work, all complexes in the training set contain identical denticity (i.e., monodentate) ligands so we exclude the denticity descriptor, and all training data uses the B3LYP (i.e., 20% exchange) functional so we exclude the HF exchange feature (see

Computational Details). These omissions leave a full RAC set consisting of 153 features (Supporting Information Table S1). We now compare the performance of RAC featurization for predicting HOMO-LUMO energetics with i) a linear LASSO regression model, ii) a KRR model with a Gaussian kernel, and iii) an ANN. We partition the data set into training and test sets using an 80%/20% random split and fix the training/test definitions to enable comparison across all ML models.

Linear model. In linear models, RAC features are weighted by coefficients in a linear combination to be correlated with the output variable. We employed L1-norm-regularized linear regression (LASSO[100]) as implemented in the scikit-learn software package[101]. LASSO prevents overfitting by using regularization to reduce the coefficients of the least-predictive variables to zero, and the hyperparameter (i.e., adjustable parameter) that is associated with the regularization strength was selected by 10-fold cross-validation (CV) error (Supporting Information Figure S1).

KRR model. In kernel based ML methods, inputs are non-linearly transformed into a higher dimensional space, more flexibly fitting data sets than is possible in a linear model. As it is impractical to work directly in a high dimensional space, a so-called kernel trick is used to yield the kernel matrix from the inner products of original inputs, thus encoding the geometric similarities in the original space. KRR is a method that combines the kernel trick with the least squares loss and L2-norm regularization[102], and we employ a Gaussian kernel using the scikit-learn software package[101]. The two adjustable hyperparameters in a KRR model are the regularization coefficient and kernel width (i.e., decay length by which distant points contribute to predicting a specific point). This hyperparameter selection was accelerated using the Bayesian optimization Python library Hyperopt[103], which optimizes the expected improvement during hyperparameter optimization rather than using an exhaustive grid search (Supporting Information

Figure S2). As with the linear model, we employed the 10-fold CV mean absolute error (MAE) for hyperparameter selection.

ANN model. Our previous work[53] demonstrated ANNs for inorganic chemistry, especially for prediction on diverse test molecules. The ANN models in the present work were trained using the keras software package[104] with TensorFlow[105], and hyperparameters were obtained with the Hyperopt[103] package. The 80% training partition of the full data set was further partitioned randomly into a 90% train and 10% validation set to ensure that the ANN was not overfit (Supporting Information Figure S3). The optimal ANN topology for HOMO level (HOMO-LUMO gap) was determined to be an input layer, two fully connected hidden layers with 500 (300) nodes each, and an output layer. Each hidden layer of an ANN transforms the input features through non-linear amplification (here, with rectified linear unit[106] non-linearities) and a linear activation function (i.e., linear combinations of feature weights arising from prior layers) is used in the output layer. To avoid overfitting, dropout regularization, i.e., zeroing out nodes within a network at a fixed 22% (31%) probability, was used to regularize the network, as in previous work[53]. Although increasing dropout increases training errors by eliminating network dependence on specific nodes, it generally improves test errors and ANN generalization, as we have previously shown on inorganic complex data sets[53]. To further reduce overfitting, L2-norm weight regularization was used at every layer of the network, with the regularization hyperparameter set to be $1x10^{-4}$ ($1x10^{-3}$). The ANN was trained with batch optimization with batch size 50 using an Adam[107] optimizer (see Supporting Information Table S2).

Feature selection. Reduction of the dimensionality of the original representation of the data set by techniques broadly referred to as feature selection can provide insight into data sets[108]. By eliminating less informative features, simple linear or kernel based models (e.g.,

KRR) can achieve higher out-of-sample performance. Based on our prior comparison of several feature selection techniques[56], we employ random forest as a starting point because it provides a low cost estimate of feature importance. Applying feature selection (for example with random forest models[109]) allows the relative importance of these different descriptors to be assessed for different prediction targets, for example, in previous work revealing that spin splitting is more locally controlled by first-shell effects as compared to redox potential[56]. Here, we employ random forest to rank the descriptors by an importance score[109] and add them sequentially to the target feature set using recursive feature addition (RFA)[108]. This differs from our prior work by combining RFA with random forest feature sets, whereas we previously used error cutoffs in the random forest model to directly select features. At each iteration, a KRR model is trained with the new feature set, and a descriptor is kept only if it improves performance. The first KRR model is trained only on oxidation state and spin multiplicity, and it and subsequent KRR models are judged by the 10-fold CV MAE of KRR after hyperparameter optimization. We continue RFA until the performance of the selected feature set levels off and stops improving (Supporting Information Figure S4). Although the non-linearity of ANNs is expected to obviate explicit feature selection, we also employ these selected subsets to test whether feature selection has any benefit for ANN performance.

4b. Model Performance.

We first trained linear (i.e., LASSO) and KRR models with RAC-153 and feature selected subsets for HOMO level prediction. The HOMO values in the data set range from -27 to -11 eV and have a mean of -15.81 eV. The inherently regularized LASSO model produces balanced train (0.80 eV) and test (0.71 eV) MAEs (Table 2 and Supporting Information Table S3). However, the lack of coupling between features or higher order terms in the linear model

limits its predictive accuracy especially at low HOMO values (Figure 5 and Table 2). Moving beyond linear models to a KRR model trained with RAC-153 features improves test MAE to 0.25 eV (Table 2). As expected, incorporation of non-linear dependences reduces the MAE of the KRR model with respect to LASSO, although we note that the lower train MAE of 0.12 eV is suggestive of overfitting that is common with KRR models (Table 2). Applying RFA to RAC-153 retains only 29 features (RFA-29). Consistent with prior work[56], this reduction improves KRR test MAE to 0.18 eV (Table 2 and Supporting Information Table S4). This improved performance can be understood as arising from improved distribution of training complexes in a space that has fewer uninformative features.

Table 2. Mean absolute errors (MAEs) for LASSO, KRR (RAC-153 or RFA-selected subsets), and ANN ML models: training set, CV partition, test set, and OH64.

Model	Train MAE (eV)	CV MAE (eV)	Test MAE (eV)	OH64 MAE (eV)
		HOMO level		
LASSO	0.80	0.82	0.71	1.85
KRR (RAC-153)	0.12	0.23	0.25	6.10
KRR (RFA-29)	0.04	0.17	0.18	1.55
ANN	0.05	0.17	0.17	1.87
		ΔE_g		
LASSO	0.68	0.70	0.47	9.47
KRR (RAC-153)	0.16	0.35	0.33	11.58
KRR (RFA-22)	0.08	0.26	0.23	2.60
ANN	0.06	0.23	0.22	2.56

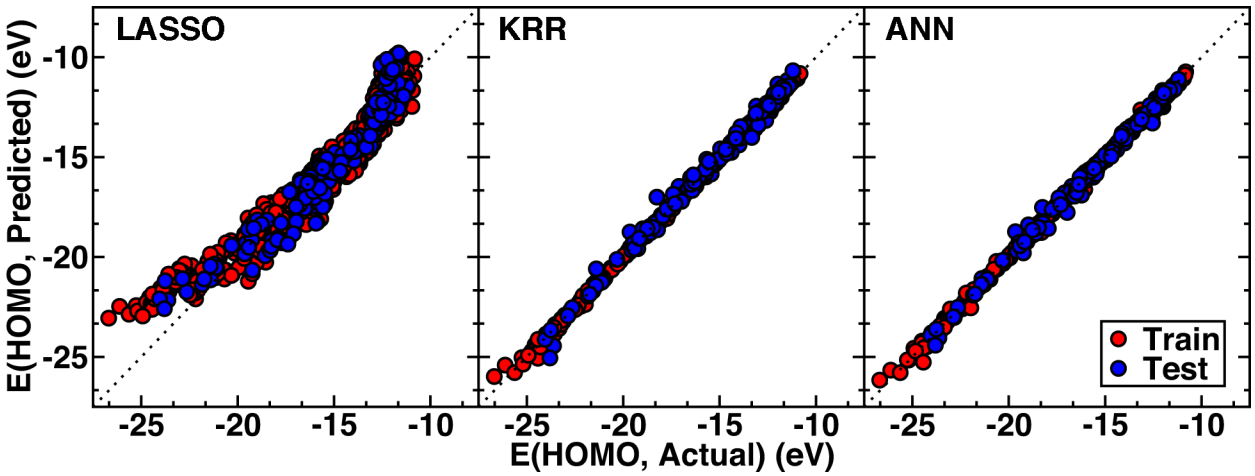


Figure 5. Train (red filled circles) and test (blue filled circles) HOMO level prediction model performance for LASSO (left), KRR (middle), and ANN (right). All results are shown as parity plots with the actual HOMO values and a black dotted parity line.

We then trained an ANN model using RAC-153 features and obtained a comparable test MAE of 0.17 eV (0.05 eV train MAE) to the RFA-29 KRR (0.18 eV) (Figure 5 and Table 2). Although ANN structure essentially incorporates feature selection via nonlinear activation that zeroes out less-predictive features, we tested the performance of an ANN using RFA-29 features. We kept fixed all hyperparameters aside from the number of inputs and found that the performance of the ANN trained on RFA-29 was comparable to or only weakly improved over the RAC-153 ANN, with train and test MAEs of 0.06 and 0.15 eV, respectively. As expected, feature selection is likely unnecessary with an ANN because the ANN is robust to uninformative features, but feature selection does provide the added advantage that complex similarity can more readily be interpreted through distances in a smaller feature space[53, 58].

Overall, train and test errors appear balanced in all models across the range of HOMO level values, except in the case of LASSO where train and test errors both appear disproportionately large at both high and low HOMO levels, indicative of the insufficiency of the linear model for capturing the phenomena that give rise to extreme HOMO values (Figure 5). For instance, LASSO significantly overestimates the DFT HOMO of -23.76 eV for a singlet

$[Co(H_2O)_5(CO)]^{3+}$ complex by 2.57 eV, whereas KRR and ANN errors are smaller at -0.80 eV and 0.00 eV, respectively (Supporting Information Figure S5). However, it is difficult to generalize these observations across all deep HOMO level compounds: LASSO conversely overestimates the DFT HOMO (-23.80 eV) of quintet $[Mn(CO)_6]^{3+}$ by a comparable amount to the underestimate error in KRR and ANN (1.21 eV vs. -1.25 and -1.08 eV, respectively, see Supporting Information Figure S5). All models perform well (0.1 eV absolute error) on certain higher HOMO level (DFT value: -13 eV) compounds, such as Fe(II) and Mn(II) complexes that are a heteroleptic mixture of equatorial pyridines with strong field axial CO ligands (Supporting Information Figure S5).

In addition to the HOMO level, we trained LASSO, KRR, and ANN models to predict the HOMO-LUMO gap, ΔE_g (Table 2 and Supporting Information Figure S6). The ΔE_g values range from 0 to 8 eV with a mean of 3.10 eV. Although the LASSO gap prediction test MAE of 0.47 eV is smaller in an absolute sense than the HOMO prediction error, it is a larger percentage of the range (6% of the ΔE_g 8 eV range vs. 4% of the 16 eV HOMO range), indicating somewhat degraded performance in comparison to HOMO level prediction (Table 2). The test MAE of a KRR model trained on RAC-153 also shows worsened test MAE of 0.35 eV, and the higher train MAE of 0.16 eV for ΔE_g (vs. 0.12 eV for the HOMO level) is suggestive of more difficulty fitting the training data with the RAC feature set. Feature selection with RFA retains 22 features and significantly improves test MAE to 0.23 eV (Table 2 and Supporting Information Table S5). We note that even after feature selection the larger train MAE of 0.08 eV (vs. 0.04 eV) and larger relative test MAE indicates that the ΔE_g prediction task is more challenging than direct HOMO level prediction (Table 2). Such challenges in predicting ΔE_g are also observed in degraded ANN model performance (train MAE: 0.06; test MAE: 0.22 eV) that is comparable to

the RFA-22 KRR (Table 2).

The increased difficulty in predicting ΔE_g over the HOMO level is at odds with our observations of comparable prediction accuracy between ionization potential and redox potential[56], suggesting there may be more noise arising from opposing effects in HOMO and LUMO variation that go into the gap prediction. In the present work, we constructed the training set for HOMO level and ΔE_g prediction to maximize the similarity to training sets used in previous work for predicting gas phase, adiabatic ionization potential and redox (i.e., solvent and thermodynamically-corrected) potential[56]. In prior work[56], we obtained 0.2-0.3 eV test set MAE for redox potentials (or 3-4% of the 6.7 eV mean value) and comparable relative MAEs of 0.4-0.6 eV (3-4% of the 14.4 eV mean value) for adiabatic ionization potentials. These ranges corresponded to feature selection with random forest generally improving redox potential prediction but slightly worsening ionization potential predictions[56]. In the present work, KRR test set MAEs using features ranked with random forest and then confirmed with RFA are comparable to or slightly better than the prior work for the HOMO: the 0.18 eV test set MAE is 1% of the -15.81 eV mean value, and slightly worse for ΔE_g: the 0.23 eV test set MAE is 7% of the 3.10 eV mean value. Analysis of the effect of the feature selection on principal component analysis (PCA) distributions of the data sheds light on this difference in model error (Supporting Information Figures S7-S8). Specifically, feature selection distributes the data more evenly and clusters deep HOMO level compounds in a distinct spatial location from shallow HOMO level compounds (Supporting Information Figure S7). Conversely, the effect is more muted for ΔE_g data, with some very small gap compounds still near large gap compounds, although feature selection does spread out the compounds more significantly (Supporting Information Figure S8).

Analysis of the character of selected feature sets is useful for understanding the length

scale and character of substituent atomic interactions that give rise to properties, even when sophisticated models (e.g., ANNs) are not very sensitive to input features. Here we again[56] distinguish the features by relative locality of the atoms in the property: proximal, intermediate, and distal (Figure 4). We also again[56] classify atomic properties as either electronic (i.e., electronegativity and nuclear charge) or steric (i.e., size, identity, and topology) in nature. The RFA-selected 29 and 22 feature sets obtained for the HOMO level and ΔE_g, respectively, contain higher distal and steric feature weights in comparison to the spin-splitting (26 features) feature set obtained from random forest previously[56] (Figure 6 and Supporting Information Tables S1 and S6). Even though the 38-feature set selected by random forest for redox potential prediction also contains higher distal and steric feature weights than the spin splitting set, it surprisingly has a strong electronic contribution from the metal center that is particularly absent in the HOMO level set and also reduced in the ΔE_g set (Figure 6 and Supporting Information Table S7). A comparison to 28 features selected on gas phase ionization potential data by random forest shows strong dissimilarity to the frontier orbital models, weighting the steric contribution even more strongly (Supporting Information Figure S9). However, we note that the random forest feature set was not as predictive on gas phase ionization potential as it was on other quantities in that work[56].

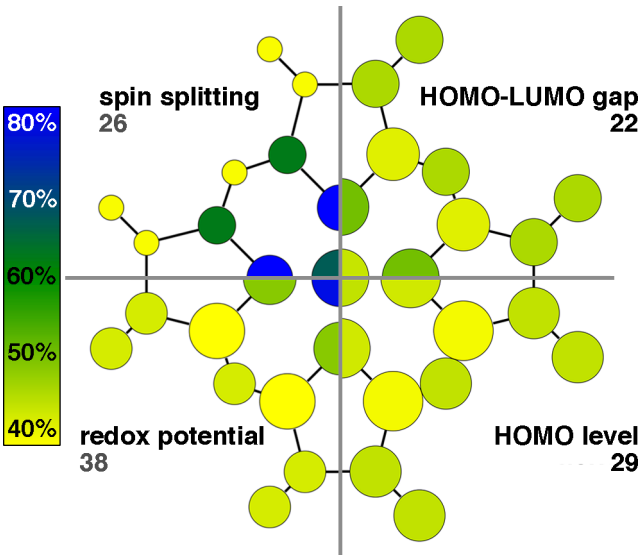


Figure 6. Schematic of relative proximity and electronic (electronegativity and nuclear charge RACs, blue) or steric (topology, identity, and size RACs, yellow) character of feature sets on a metalloporphyrin abstraction. Feature sets are designated by their training data: spin splitting (top left), redox potential (bottom left), HOMO-LUMO gap (top right), and HOMO level (bottom right). Retained features from random forest are also indicated, with spin splitting and redox from prior work[56] indicated in dark gray, and the new feature sets for HOMO and HOMO-LUMO are indicated in black. Atom sizes of the first, second shell, and beyond are scaled by the number of descriptor dimensions involving that shell relative to the metal center, which is kept the same size in all sets. The color bar and absolute percentages of electronic and topological descriptors, as defined in the main text, is shown in the left inset.

Although differences in down-selected feature sets can be informative for guiding iterative design of materials properties, we have previously demonstrated good transferability between feature sets for differing properties. We may expect such observations to hold here for HOMO level and ΔE_g prediction, particularly because the weight of steric and distal features appears comparable between the two sets (Figure 6). Indeed, KRR with the ΔE_g-selected RFA-22 produces a HOMO level test MAE of 0.19 eV, only slightly increased over the RFA-29 set (0.18 eV). Using the larger RFA-29 feature set to train a HOMO level KRR model degrades performance (0.26 eV test MAE) slightly more over the original model (0.23 eV test MAE). The electronically-weighted and metal-centric features of the redox and spin splitting feature sets further degrade test MAEs for HOMO level (0.21-0.24 eV test MAE) and ΔE_g (0.29 eV test

MAE), although both still outperform the models trained on RAC-153 (Supporting Information Tables S8-S9).

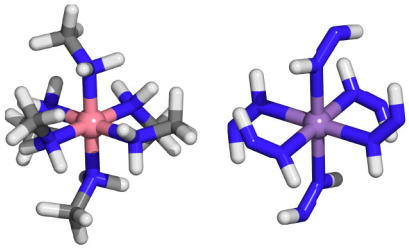
To test the ability to extrapolate in discovery with our models, we evaluated the HOMO level and ΔE_g models on a new set of 64 octahedral homoleptic complexes (OH64) with monodentate ligands containing up to two heavy C, N, O atoms (Supporting Information Table S10 and structures and more details provided in the Supporting Information). These OH64 complexes contain the same metal centers (Co, Cr, Fe, Mn) in the same oxidation and spin states complexed with 8 unique but similar composition (i.e., C, N, O, and H-containing) ligands. Specifically, the ligands include NH_2CH_3, $NHCH_2$, NCH, N_2, OCH_2, NH_3, N_2H_2, and NHO, where the first atom is the one that coordinates the metal (Supporting Information Table S10). Despite some similarities, these complexes were absent from training data, and only 8% (54 of 699) of the training data contains homoleptic complexes. The differences between OH64 and the redox set produces large distances of the newly generated molecules to training data (5-nearest-neighbor, 5-NN, distance of OH64 to training: 9.6 vs. 5-NN redox test to train: 2.2) potentially limiting prediction by interpolative (e.g., KRR) or overfit models. Properties of the two sets also differ, likely due to the smaller size of OH64 complexes: ΔE_g (HOMO level) averages 5.81 eV (-18.79 eV) for OH64 vs. 3.10 eV (-15.81 eV) for the training set (Supporting Information Figure S10).

As expected, all model MAEs increase on the OH64 set in comparison to the test set MAE (Table 2). Somewhat surprisingly, the best performing model on the HOMO level is the feature-selected KRR (OH64 MAE: 1.55 eV), although both the ANN (1.87 eV) and LASSO (1.85 eV) perform reasonably (Table 2). In particular, LASSO error from test set MAE to OH64 MAE increases the least (0.71 eV vs. 1.85 eV), but the LASSO model performs much more

poorly on ΔE_g prediction (OH64 MAE 9.47 eV vs. 0.47 eV) (Table 2). Relative model performance (i.e., test vs. OH64) worsens in all cases on ΔE_g over the HOMO, potentially due to relatively strong size dependence of the HOMO level vs. more subtle effects that dictate the gap. Without feature selection, KRR models that use the full RAC-153 space perform the worst of all models, including LASSO, giving OH64 MAEs of 6.10 eV for the HOMO and 11.58 eV for ΔE_g (Table 2). This large error is due to the large Euclidean distance between the training set and OH64 molecules in the latent space, leading to unsupported (i.e., mean value) predictions by the KRR model. Thus, for extrapolation to diverse molecules, feature selection is essential to enhance the role of the most important chemical factors in model prediction when using small training sets as we have in this work.

We examined OH64 HOMO level performance and identified two N-coordinating ligand complexes on which the ANN model performed alternately well or poorly (Figure 7). In the training set only a single N-coordinating ligand, pyridine, is present, meaning that the most similar complexes to the OH64 cases may instead have non-N-coordinating ligands with more similar topology. Nevertheless, we observe that singlet $[Co(NH_2CH_3)_6]^{3+}$ is well predicted by our model (-19.91 eV from the ANN vs. -20.00 eV for DFT) (Figure 7). Conversely, the quintet $[Mn(HNNH)_6]^{3+}$ complex HOMO level is underestimated by 4.9 eV, which is almost double the MAE for the ANN on the OH64 set (Figure 7 and Table 2). We attribute this performance difference due to the similarity of methylamine ligands to other training set chemistry and relative dissimilarity of the HNNH ligand that involves a N-N double bond not present in the training set. In order to avoid such large errors as observed for the HNNH ligand in the future, a more diverse training set, better metrics for model uncertainty (i.e., to limit prediction on uncertain compounds), and a better match between training and prediction molecules will all be

beneficial. For example, an ANN trained to predict spin splitting[53] on a broader data set from our prior work[53, 56] with RACs[56] only has a 3 kcal/mol error for the same compound. Overall, these results reinforce observations that ML model test set error is an insufficient indicator of the likely errors[53] on extrapolative complexes, and that relative similarity of diverse or discovery target molecules must be taken into account when applying ML models to new compounds.



	$[Co(NH_2CH_3)_6]^{3+}$	$[Mn(HNNH)_6]^{3+}$
DFT	-20.00 eV	-18.64 eV
ANN	-19.91 eV	-23.55 eV

Figure 7. Example molecules from the diverse OH64 test set: (left) small ANN error (0.09 eV) with respect to DFT in a singlet $[Co(NH_2CH_3)_6]^{3+}$ complex and (right) large ANN error (-4.91 eV) with respect to DFT for a quintet $[Mn(HNNH)_6]^{3+}$ complex. Metals are shown as spheres and coordinating atoms as sticks, with C atoms in gray, N atoms in blue, and H atoms in white.

4c. Optimizing Frontier MO Properties.

Given the trained HOMO level and ΔE_g models, we now demonstrate how such models can accelerate materials design when used for scoring in combination with a genetic algorithm (GA), as implemented in mAD[58]. Here, we use the ANN model that predicts ΔE_g to design a TM complex with a target ΔE_g of 4 eV in its ground spin state, as indicated by a spin splitting ANN[53]. The design space is constructed from a series of genes: i) one for any of four metals, Co, Cr, Mn, Fe, in two oxidation states, +2 or +3, and ii) 15 monodentate ligands with up to three unique coordinating species per complex, with C, N, or O connecting atoms (Supporting Information Table S11). Specifically, 1800 ligand combinations are possible: 15 homoleptic, 210 with a single unique axial species (i.e., $M(L_1)_5(L_2)$), 210 with unique equatorial and axial species

(i.e., $M(L_1)_4(L_2)_2$), and 1365 with the same equatorial ligands but two unique axial ligands (i.e., $M(L_1)_4(L_2)(L_3)$). In total, the metal and ligand combinations give rise to 14,400 possible complexes, 699 of which (4.9%) were in the training data, and the space is larger than in previous work despite a smaller (i.e., 15 vs. 35[58]) ligand pool, owing to the inclusion of more asymmetric complexes. The fitness function used here to target ΔE_g values is modified from the form we used for targeting spin splitting in previous work[58]:

$$F_{s,d} = e^{\left(-\left(\max\left(0,\, 3.75-\Delta E_g\right)+\max\left(0,\, \Delta E_g-4.25\right)\right)\right)}\, e^{-\left(\frac{d}{d_{\text{opt}}}\right)^2} \quad (1)$$

where the first term is a flat bottom exponential penalty for ΔE_g values that are distant (i.e., > ± 0.25 eV) from the target value of 4.0 eV and the second term is a penalty on distance to training data. Although we previously introduced a distance penalty approach to enable discovery in an ANN with knowledge of uncertainty[58], in the present work, we select d_{opt} = 30 to encourage discovery of previously unseen complexes, penalizing values of d > 10 in RAC-153 descriptors.

Although d values in RAC-153 are not equivalent to the prior work in MCDL-25[58], this weak penalty is generally more encouraging of new complex discovery than in our prior work. The motivation for this weaker penalty is also that, without feature selection, RAC-153 exaggerates small differences in ligand connectivity, producing much larger distances in feature space than would be observed in the nearsighted MCDL-25 and lessening the value of distance control in limiting model uncertainty (Figure 8). A moderate distance in RAC-153 for metal-distant differences (e.g., d = 5 for cyanopyridine ligands in place of pyridine in a training set complex) is due to the higher weight of distal features in the full RAC-153 space over MCDL-25 (Figures 4 and 8). Larger distances (d = 10) are observed when both connectivity and spin state change, as is the case for quintet $[Cr(CO)_5(tbisc)]^{2+}$ (tbisc = t-butyl-isocyanide), which is quite

far from the closest training structure, singlet $[Cr(CO)_5(pyridine)]^{2+}$ (Figure 8). Thus, penalizing distances above $d = 10$, which typically correspond to substantial differences in multiple ligands, oxidation, or spin states, encourages most exploration by the ANN but avoids cases where molecules are extremely distant from training data (Figure 8). In all cases, the ANN predicts comparable ΔE_g (ca. 4-5 eV) for these Cr(II) complexes, but we should expect the ANN to be more reliably predictive (i.e., closer to the DFT result) on cases with distances closer to the training data.

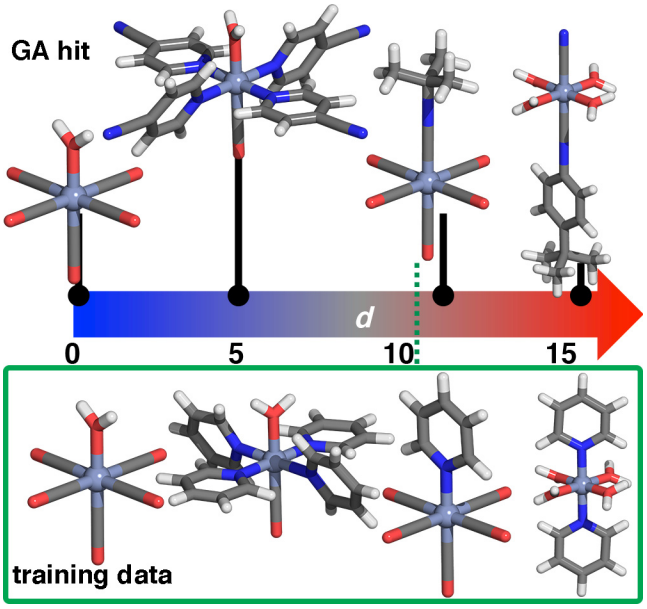


Figure 8. Depiction of Euclidean norm distance of GA hit complexes (top) to closest available training data (bottom) illustrated on a family of related Cr complexes. All GA hit complexes are quintet Cr(II), and training data points change oxidation/spin state in the following cases only: the $d = 5$ training point is Cr(III) high spin, $d = 10$ is Cr(II) low spin, $d = 15$ is Cr(III) high spin. The distance at which distances are penalized is indicated with a dashed green vertical line.

Starting from a pool of 20 randomly selected complexes, the GA is run for 21 generations with both distance control as in eq. 1 as well as diversity control[58] on the mutation probability, p_{mut}. Here, p_{mut} is increased from its default value of 0.15 to 0.65 whenever the diversity (i.e., unique complexes in a generation) falls below 25%, and it is returned to 0.15 once diversity increases to 25% or higher. The mean fitness function rises rapidly to 0.85 in the first five

generations and remains there for the rest of the GA run, leading to local exploration around similar compounds in the remaining generations (Supporting Information Figure S11). At the end of the GA run, we recover both complexes from the training set (e.g., quartet Co(II)(pyr)$_4$(H$_2$O)(misc), with ANN ΔE_g = 4.17 eV) and new complexes (e.g., quintet Fe(II)(CN-pyr)$_4$(misc)(furan), with ANN ΔE_g = 3.83 eV) (Figure 9).

In total, 105 new compounds are discovered during the GA run with an average distance of 7.9 to training data, and only 9 of these compounds were previously employed in training (Supporting Information). Over a 10 complex subset of these new compounds with an average ANN-predicted gap of 4.00 eV, the average ΔE_g from DFT with full geometry optimizations is 3.97 eV (Supporting Information Table S12). The MAE over all compounds for the ANN with respect to DFT is 0.27 eV, a modest increase from the 0.22 eV ΔE_g test set MAE for the ANN (Table 2 and Supporting Information Table S12). Errors ranged from as small as 0.01 eV with respect to DFT values and as large as 0.98 eV in one case (Supporting Information Table S12). In this design space, the objective function is easily fulfilled by a large number of compounds, therefore motivating the use of multi-objective optimization in future studies, e.g., by optimizing both the ΔE_g as well as placement of HOMO and LUMO levels.

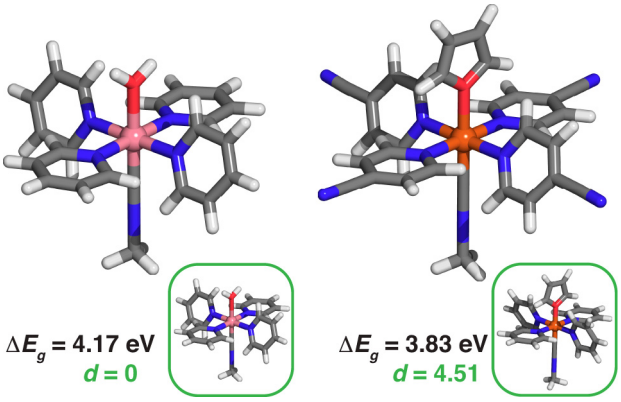


Figure 9. Leads from GA run with the ANN predicted ΔE_g in eV and closest training data in inset green rectangle with Euclidean norm distance indicated in green: (left) lead quartet $[Co(pyr)_4(H_2O)(misc)]^{2+}$ (where misc = methylisocyanide and pyr = pyridine) is also in the

training set and (right) lead quintet $[Fe(CN\text{-}pyr)_4(furan)(misc)]^{2+}$ (where CN-pyr = 4-cyanopyridine) is closest to a training point singlet $[Fe(pyr)_4(furan)(misc)]^{2+}$.

5. Conclusions.

We introduced software automation tools that enable the training of and exploit the use of machine learning models. The geometry checks we introduced ensure that new simulation data for machine learning models is robust without requiring manual validation. The molSimplify automatic design (mAD) workflow enables automated property optimization using genetic algorithms (GAs) with custom fitness functions. When using machine learning model energetics are used in mAD fitness functions, we showed that knowledge of model uncertainty, as judged through distance to training data, can be used fruitfully for chemical discovery.

To demonstrate the power of our RAC topological representation for inorganic chemistry, we train three types of machine learning models, LASSO, KRR, and ANNs. We had previously developed models to predict adiabatic gas phase ionization potential, redox potential, bond length, and spin splitting. In this work, we demonstrated the performance of RACs for predicting HOMO levels and HOMO-LUMO gaps for the first time. These models included the first ANNs trained on the RAC representation, which showed the best (ca. 0.15-0.20 eV) test set MAE performance on predicting the HOMO level of a diverse set of open shell transition metal complexes in varying spin and oxidation state. Performance on HOMO-LUMO gaps was slightly poorer (test set MAE ca. 0.25 eV) for all models. Although KRR performance using the full RAC-153 data set was inferior to the ANN, KRR models trained on RFA-selected feature sets that included 22 and 29 features for HOMO and HOMO-LUMO gap, respectively, showed nearly comparable performance to the ANN. The proximal/distal and electronic/steric blend of HOMO and HOMO-LUMO feature sets were comparable: both emphasized non-local, steric

properties even more than prior redox or ionization potential data sets. The feature sets demonstrated good transferability between the HOMO and HOMO-LUMO properties but previously selected electronic and metal-focused sets, e.g., from spin splitting or redox potential performed less well.

Overall, diverse molecule performance was tested on a series of 64 small transition metal complexes, and HOMO and HOMO-LUMO errors increased by around an order of magnitude, particularly for unusual complexes containing types of bonds not present in the training data. With this model uncertainty in mind, a mAD GA run with fitness function designed to enable discovery of complexes with a 4 eV HOMO-LUMO gap was carried out in a design space of nearly 15,000 complexes. This screen recovered both complexes from our training set as well as new lead candidates in a matter of minutes, demonstrating the power of ML models for rapid pre-screening of a large design space. The next step will be to pursue active learning strategies that exploit, rather than avoid, regions of uncertainty for machine learning models, which is currently the focus of work underway in our group.

ASSOCIATED CONTENT

Supporting Information. LASSO model hyperparameter selection; KRR model hyperparameter selection; ANN train and validation curves for HOMO level and ΔE_g; ANN Adam optimizer hyperparameters; features selected by LASSO for HOMO level; RFA feature selection details; LASSO, KRR, and ANN comparison for ΔE_g; RAC-153, random forest-26 (for spin), random forest-38 (for redox), RFA-29 (for ΔE_g), and RFA-22 (for HOMO level) MAE and hyperparameter comparison for HOMO level and ΔE_g; principal component analysis before and after feature selection for HOMO level and ΔE_g; RFA-22 features; RFA-29 features; RAC-153

features; GA design space ligand list; GA fitness vs. generations; HOMO level and HOMO-LUMO gap value distributions; DFT vs. ANN comparison of GA runs. (PDF)

List of eliminated structures, sorted by geometric failure or spin contamination; ΔE_g energy categorization by type; summary of OH64 prediction results and details of ligand assignment; summary of GA run results (ZIP)

Geometries for 874 TM complexes used in the present work in training and test data; geometries for OH64 data set (ZIP)

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.

ACKNOWLEDGMENT

The authors acknowledge support by the Office of Naval Research under grant numbers N00014-17-1-2956 and N00014-18-1-2434, DARPA grant D18AP00039, the Department of Energy under grant number DE-SC0018096, the National Science Foundation under grant number CBET-1704266, and MIT Energy Initiative seed grants (2014, 2017). H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830). The

authors thank Adam H. Steeves for providing a critical reading of the manuscript.

References

1. Bousseksou, A.; Molnár, G.; Matouzenko, G., Switching of Molecular Spin States in Inorganic Complexes by Temperature, Pressure, Magnetic Field and Light: Towards Molecular Devices. *Eur. J. Inorg. Chem.* **2004,** *2004*, 4353-4369.
2. Decurtins, S.; Gütlich, P.; Köhler, C.; Spiering, H.; Hauser, A., Light-Induced Excited Spin State Trapping in a Transition-Metal Complex: The Hexa-1-Propyltetrazole-Iron (II) Tetrafluoroborate Spin-Crossover System. *Chem. Phys. Lett.* **1984,** *105*, 1-4.
3. Hauser, A., Light-Induced Spin Crossover and the High-Spin→ Low-Spin Relaxation. In *Spin Crossover in Transition Metal Compounds II*, Springer: 2004; pp 155-198.
4. Reed, D. A.; Xiao, D. J.; Gonzalez, M. I.; Darago, L. E.; Herm, Z. R.; Grandjean, F.; Long, J. R., Reversible CO Scavenging via Adsorbate-Dependent Spin State Transitions in an Iron (II)–Triazolate Metal–Organic Framework. *J. Am. Chem. Soc* **2016,** *138*, 5594-5602.
5. Groizard, T.; Papior, N.; Le Guennic, B.; Robert, V.; Kepenekian, M., Enhanced Cooperativity in Supported Spin-Crossover Metal–Organic Frameworks. *J. Phys. Chem. Lett.* **2017,** *8*, 3415-3420.
6. Neville, S. M.; Halder, G. J.; Chapman, K. W.; Duriska, M. B.; Moubaraki, B.; Murray, K. S.; Kepert, C. J., Guest Tunable Structure and Spin Crossover Properties in a Nanoporous Coordination Framework Material. *J. Am. Chem. Soc.* **2009,** *131*, 12106-12108.
7. Bignozzi, C. A.; Argazzi, R.; Boaretto, R.; Busatto, E.; Carli, S.; Ronconi, F.; Caramori, S., The Role of Transition Metal Complexes in Dye Sensitized Solar Devices. *Coord. Chem. Rev.* **2013,** *257*, 1472-1492.
8. Harvey, J. N.; Poli, R.; Smith, K. M., Understanding the Reactivity of Transition Metal Complexes Involving Multiple Spin States. *Coord. Chem. Rev.* **2003,** *238*, 347-361.
9. Gani, T. Z. H.; Kulik, H. J., Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane to Methanol Conversion by Feiv═O. *ACS Catal.* **2018,** *8*, 975-986.
10. Schilling, M.; Patzke, G. R.; Hutter, J.; Luber, S., Computational Investigation and Design of Cobalt Aqua Complexes for Homogeneous Water Oxidation. *J. Phys. Chem. C* **2016,** *120*, 7966-7975.
11. Kim, J. Y.; Kulik, H. J., When Is Ligand pKa a Good Descriptor for Catalyst Energetics? In Search of Optimal CO2 Hydration Catalysts. *J. Phys. Chem. A* **2018,** *122*, 4579-4590.
12. Pellizzeri, S.; Jones, I. A.; Doan, H. A.; Snurr, R. Q.; Getman, R. B., Using Gas-Phase Clusters to Screen Porphyrin-Supported Nanocluster Catalysts for Ethane Oxidation to Ethanol. *Catal. Lett.* **2016,** *146*, 2566-2573.
13. Nørskov, J. K.; Bligaard, T., The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013,** *52*, 776-777.
14. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A., Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013,** *1*, 011002.
15. Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O., The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013,** *12*, 191-201.
16. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A., The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011,** *2*, 2241-2251.

17. Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R., Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013,** *4*, 1613-1623.
18. 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.
19. Huang, W.; Xing, D.-H.; Lu, J.-B.; Long, B.; Schwarz, W. H. E.; Li, J., How Much Can Density Functional Approximations (DFA) Fail? The Extreme Case of the FeO4 Species. *J. Chem. Theory Comput.* **2016,** *12*, 1525-1533.
20. 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.
21. Ioannidis, E. I.; Kulik, H. J., Ligand-Field-Dependent Behavior of Meta-GGA Exchange in Transition-Metal Complex Spin-State Ordering. *J. Phys. Chem. A* **2017,** *121*, 874-884.
22. Ashley, D. C.; Jakubikova, E., Ironing out the Photochemical and Spin-Crossover Behavior of Fe (II) Coordination Compounds with Computational Chemistry. *Coord. Chem. Rev.* **2017,** *337*, 97-111.
23. 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.
24. Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L.; Sanvito, S.; Buongiorno-Nardelli, M., AFLOWLIB. Org: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012,** *58*, 227-235.
25. Calderon, C. E.; Plata, J. J.; Toher, C.; Oses, C.; Levy, O.; Fornari, M.; Natan, A.; Mehl, M. J.; Hart, G.; Buongiorno Nardelli, M.; Curtarolo, S., The AFLOW Standard for High-Throughput Materials Science Calculations. *Comput. Mater. Sci.* **2015,** *108*, 233-238.
26. Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D., AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012,** *58*, 218-226.
27. Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Jensen, P. B.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, E. L.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Maronsson, J. B.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W., The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017,** *29*, 273002.
28. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013,** *68*, 314-319.
29. Landrum, G. Rdkit: Open-Source Cheminformatics Software. http://www.rdkit.org (accessed August 17, 2018).
30. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011,** *3*, 33.
31. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E., The Chembl Database in 2017. *Nucleic acids research* **2016,** *45*, D945-D954.

32. Irwin, J. J.; Shoichet, B. K., Zinc− a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45*, 177-182.
33. Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J., molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016,** *37*, 2106-2117.
34. Gani, T. Z. H.; Ioannidis, E. I.; Kulik, H. J., Computational Discovery of Hydrogen Bond Design Rules for Electrochemical Ion Separation. *Chem. Mater.* **2016,** *28*, 6207-6218.
35. 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.
36. Kim, J. Y.; Steeves, A. H.; Kulik, H. J., Harnessing Organic Ligand Libraries for First-Principles Inorganic Discovery: Indium Phosphide Quantum Dot Precursor Design Strategies. *Chem. Mater.* **2017,** *29*, 3632-3643.
37. Bleiziffer, P.; Schaller, K.; Riniker, S., Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018,** *58*, 579-590.
38. Li, Z.; Omidvar, N.; Chin, W. S.; Robb, E.; Morris, A.; Achenie, L.; Xin, H., Machine-Learning Energy Gaps of Porphyrins with Molecular Graph Representations. *J. Phys. Chem. A* **2018,** *122*, 4571-4578.
39. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., Moleculenet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018,** *9*, 513-530.
40. Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M., Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018,** *9*, 1289-1300.
41. Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R., Schnet – a Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018,** *148*, 241722.
42. Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J., The Tensormol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018,** *9*, 2261-2269.
43. Behler, J.; Parrinello, M., Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007,** *98*, 146401.
44. Behler, J., Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016,** *145*, 170901.
45. 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.
46. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C., Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017,** *3*, 54.
47. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016,** *15*, 1120-1127.
48. Kitchin, J. R., Machine Learning in Catalysis. *Nat. Catal.* **2018,** *1*, 230-232.
49. Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C., Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018,** *64*, 2311-2323.

50. Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N., Machine Learning-Based Screening of Complex Molecules for Polymer Solar Cells. *J. Chem. Phys.* **2018,** *148*, 241735.
51. Pilania, G.; Gubernatis, J. E.; Lookman, T., Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017,** *129*, 156-163.
52. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015,** *11*, 2087-2096.
53. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017,** *8*, 5137-5152.
54. Droghetti, A.; Alfè, D.; Sanvito, S., Assessment of Density Functional Theory for Iron (II) Molecules across the Spin-Crossover Transition. *J. Chem. Phys.* **2012,** *137*, 124303.
55. Mortensen, S. R.; Kepp, K. P., Spin Propensities of Octahedral Complexes from Density Functional Theory. *J. Phys. Chem. A* **2015,** *119*, 4041-4050.
56. 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.
57. Shu, Y.; Levine, B. G., Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015,** *142*, 104104.
58. 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.
59. Peterson, A. A.; Christensen, R.; Khorshidi, A., Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017,** *19*, 10978-10985.
60. Liu, R.; Glover, K. P.; Feasel, M. G.; Wallqvist, A., General Approach to Estimate Error Bars for Quantitative Structure–Activity Relationship Predictions of Molecular Activity. *J. Chem. Inf. Model.* **2018,** *ASAP*, DOI:10.1021/acs.jcim.8b00114.
61. 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.
62. Broto, P.; Moreau, G.; Vandycke, C., Molecular Structures: Perception, Autocorrelation Descriptor and Sar Studies: System of Atomic Contributions for the Calculation of the N-Octanol/Water Partition Coefficients. *Eur. J. Med. Chem.* **1984,** *19*, 71-78.
63. Broto, P.; Devillers, J., *Autocorrelation of Properties Distributed on Molecular Graphs*. 1990; p 105-127.
64. Devillers, J.; Domine, D.; Guillon, C.; Bintein, S.; Karcher, W., Prediction of Partition Coefficients (Log P Oct) Using Autocorrelation Descriptors. *SAR QSAR Environ. Res.* **1997,** *7*, 151-172.
65. 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.
66. 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.
67. 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.

68. Houk, K. N., Frontier Molecular Orbital Theory of Cycloaddition Reactions. *Acc. Chem. Res.* **1975,** *8*, 361-369.
69. Fukui, K.; Yonezawa, T.; Shingu, H., A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons. *J. Chem. Phys.* **1952,** *20*, 722-725.
70. Joachim, C.; Gimzewski, J.; Aviram, Electronics Using Hybrid-Molecular and Mono-Molecular Devices. *Nature* **2000,** *408*, 541.
71. Li, Y., Molecular Design of Photovoltaic Materials for Polymer Solar Cells: Toward Suitable Electronic Energy Levels and Broad Absorption. *Acc. Chem. Res.* **2012,** *45*, 723-733.
72. Janak, J. F., Proof That dE/dni=Epsilon in Density-Functional Theory. *Phys. Rev. B* **1978,** *18*, 7165-7168.
73. Stein, T.; Eisenberg, H.; Kronik, L.; Baer, R., Fundamental Gaps in Finite Systems from Eigenvalues of a Generalized Kohn-Sham Method. *Phys. Rev. Lett.* **2010,** *105*, 266802.
74. Zhang, G.; Musgrave, C. B., Comparison of DFT Methods for Molecular Orbital Eigenvalue Calculations. *J. Phys. Chem. A* **2007,** *111*, 1554-1561.
75. Perdew, J. P.; Yang, W.; Burke, K.; Yang, Z.; Gross, E. K.; Scheffler, M.; Scuseria, G. E.; Henderson, T. M.; Zhang, I. Y.; Ruzsinszky, A., Understanding Band Gaps of Solids in Generalized Kohn–Sham Theory. *Proc. Natl. Acad. Sci. U. S. A.* **2017,** *114*, 2801-2806.
76. Kümmel, S.; Kronik, L., Orbital-Dependent Density Functionals: Theory and Applications. *Rev. Mod. Phys.* **2008,** *80*, 3-60.
77. Kabsch, W., Solution for Best Rotation to Relate 2 Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976,** *32*, 922-923.
78. Janet, J. P.; Zhao, Q.; Ioannidis, E. I.; Kulik, H. J., Density Functional Theory for Modelling Large Molecular Adsorbate–Surface Interactions: A Mini-Review and Worked Example. *Mol. Simul.* **2017,** *43*, 327-345.
79. Janet, J. P.; Duan, C.; Nandy, A.; Kulik, H. J. molSimplify Github. https://github.com/hjkgrp/molSimplify (accessed August 17, 2018).
80. Janet, J. P.; Duan, C.; Nandy, A.; Kulik, H. J. molSimplifyAD Github. https://github.com/hjkgrp/AutomaticDesign (accessed August 17, 2018).
81. Janet, J. P.; Kulik, H. J. molSimplify Web Tutorials. http://molsimplify.mit.edu (accessed August 17, 2018).
82. Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A., Parallel and Distributed Thompson Sampling for Large-Scale Accelerated Exploration of Chemical Space. *eprint arXiv:1706.01825* **2017**, arXiv:1706.01825.
83. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018,** *148*, 241733.
84. Kramida, A., Ralchenko, Yu., Reader, J. and NIST ASD Team NIST Atomic Spectra Database (Version 5.3). http://physics.nist.gov/asd (accessed August 17, 2018).
85. Terachem, Petachem, Llc. http://www.petachem.com (accessed August 17, 2018).
86. Ufimtsev, I. S.; Martinez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009,** *5*, 2619-2628.
87. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994,** *98*, 11623-11627.
88. Becke, A. D., Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

89. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988,** *37*, 785-789.
90. Hay, P. J.; Wadt, W. R., Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for the Transition Metal Atoms Sc to Hg. *J. Chem. Phys.* **1985,** *82*, 270-283.
91. Saunders, V. R.; Hillier, I. H., Level-Shifting Method for Converging Closed Shell Hartree-Fock Wave-Functions. *Int. J. Quantum Chem.* **1973,** *7*, 699-705.
92. Mori-Sánchez, P.; Cohen, A. J.; Yang, W., Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008,** *100*, 146401.
93. Cohen, A. J.; Mori-Sánchez, P.; Yang, W., Fractional Charge Perspective on the Band Gap in Density-Functional Theory. *Phys. Rev. B* **2008,** *77*, 115123.
94. Perdew, J. P.; Levy, M., Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.* **1983,** *51*, 1884-1887.
95. Sham, L. J.; Schlüter, M., Density-Functional Theory of the Energy Gap. *Phys. Rev. Lett.* **1983,** *51*, 1888-1891.
96. Bartók, A. P.; Kondor, R.; Csányi, G., On Representing Chemical Environments. *Phys. Rev. B* **2013,** *87*, 184115.
97. Huang, B.; von Lilienfeld, O. A., Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016,** *145*, 161102.
98. Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M., Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015,** *114*, 105503.
99. Kier, L. B., A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985,** *4*, 109-116.
100. Tibshirani, R., Regression Shrinkage and Selection via the Lasso. *J. Royal Stat. Soc.: Ser. B* **1996,** *58*, 267-288.
101. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E., Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011,** *12*, 2825-2830.
102. Hastie, T.; Tibshirani, R.; Friedman, J. H., *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2nd ed.; Springer: New York, 2009; p xxii, 745 p.
103. Bergstra, J. Y., D.; Cox, D. D, Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proceedings of the 12th Python in science conference* **2013**, 13-20.
104. Keras. https://github.com/keras-team/keras (accessed August 17, 2018).
105. Abadi, M., Tensorflow: A System for Large-Scale Machine Learning.
106. Nair, V.; Hinton, G. E. In *Rectified Linear Units Improve Restricted Boltzmann Machines*, International Conference on Machine Learning, 2010; pp 807-814.
107. Kingma, D. P.; Ba, J. L. In *Adam: A Method for Stochastic Optimization*, International Conference for Learning Representations, San Diego, San Diego, 2015.
108. Guyon, I. E., An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003,** *3*, 1157−1182.
109. Breiman, L., Random Forests. *Mach. Learn.* **2001,** *45*, 5-32.

TOC Graphic