Inorganic Chemistry

Forum Article

Cite This: Inorg. Chem. 2019, 58, 10592−10606

pubs.acs.org/IC

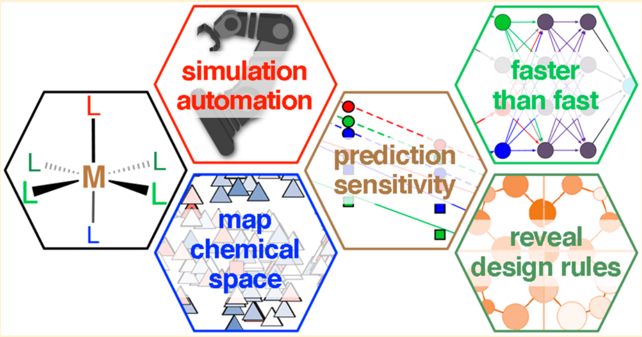
Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry

Jon Paul Janet,[†] Fang Liu,[†] Aditya Nandy,[†,‡] Chenru Duan,[†,‡] Tzuhsiung Yang,[†] Sean Lin,[†] and Heather J. Kulik*,[†]

[†]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

[‡]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Supporting Information



ABSTRACT: Recent transformative advances in computing power and algorithms have made computational chemistry central to the discovery and design of new molecules and materials. First-principles simulations are increasingly accurate and applicable to large systems with the speed needed for high-throughput computational screening. Despite these strides, the combinatorial challenges associated with the vastness of chemical space mean that more than just fast and accurate computational tools are needed for accelerated chemical discovery. In transition-metal chemistry and catalysis, unique challenges arise. The variable spin, oxidation state, and coordination environments favored by elements with well-localized d or f electrons provide great opportunity for tailoring properties in catalytic or functional (e.g., magnetic) materials but also add layers of uncertainty to any design strategy. We outline five key mandates for realizing computationally driven accelerated discovery in inorganic chemistry: (i) fully automated simulation of new compounds, (ii) knowledge of prediction sensitivity or accuracy, (iii) faster-than-fast property prediction methods, (iv) maps for rapid chemical space traversal, and (v) a means to reveal design rules on the kilocompound scale. Through case studies in open-shell transition-metal chemistry, we describe how advances in methodology and software in each of these areas bring about new chemical insights. We conclude with our outlook on the next steps in this process toward realizing fully autonomous discovery in inorganic chemistry using computational chemistry.

1. INTRODUCTION

Thanks to transformative advances in computing power and algorithms,[1−13] computational chemistry has become central to the discovery and design of new molecules[14−18] and materials.[19−25] Fully first-principles simulations of length or time scales that would have been inconceivable a little over a decade ago in the biological[13,26,27] and materials[28,29] sciences are increasingly routine. The hardware and strategies that have enabled these advances are far-ranging but include simplification of the complexity or number of quantum-mechanical electron-repulsion integral evaluations.[1−13] Regardless of the flavor of improvement, the practical effect is profound: a simulation that would have taken a week a little over a decade ago now takes less than an hour.[3] Alongside computational cost reductions, the accuracy of practical first-principles methods has improved dramatically through new functional forms that tackle long-range physics.[30−36] Within widely applied density functional theory (DFT), recent years have brought new insight about the interplay between delocalization (i.e., self-interaction) error[37,38] and static correlation error[39−41] and the relationship between density and energetic errors[42−47] relevant to transition-metal chemistry.[40,48] The same advances that have made widely employed DFT applicable to ever-larger systems have transformed gold standard correlated wave function theory (WFT) methods as well.[9−11,49] WFT methods that were once applicable to only a handful of atoms can now be employed to study hundreds.[12,50] Statistical techniques have also started[51,52] to make these WFT methods as "black box" as DFT.

Despite these great strides, the accelerated discovery of new catalysts[53−57] and materials[58−63] requires a different approach to realize computation's full potential. The orders of magnitude advance in speed and accuracy has not translated directly to an equivalent acceleration in developing new chemical insight. These observations motivate us to focus on a distinct, larger problem: only a tiny fraction (ca. 1 part in 10^{50})[64,65] of chemical space has ever been explored. This chemical space contains all of the as-yet unknown catalysts, materials, and therapeutic drugs or otherwise useful molecules. The effort to uncover these molecules and materials unites the

Special Issue: Celebrating the Year of the Periodic Table: Emerging Investigators in Inorganic Chemistry

Received: January 12, 2019
Published: March 5, 2019

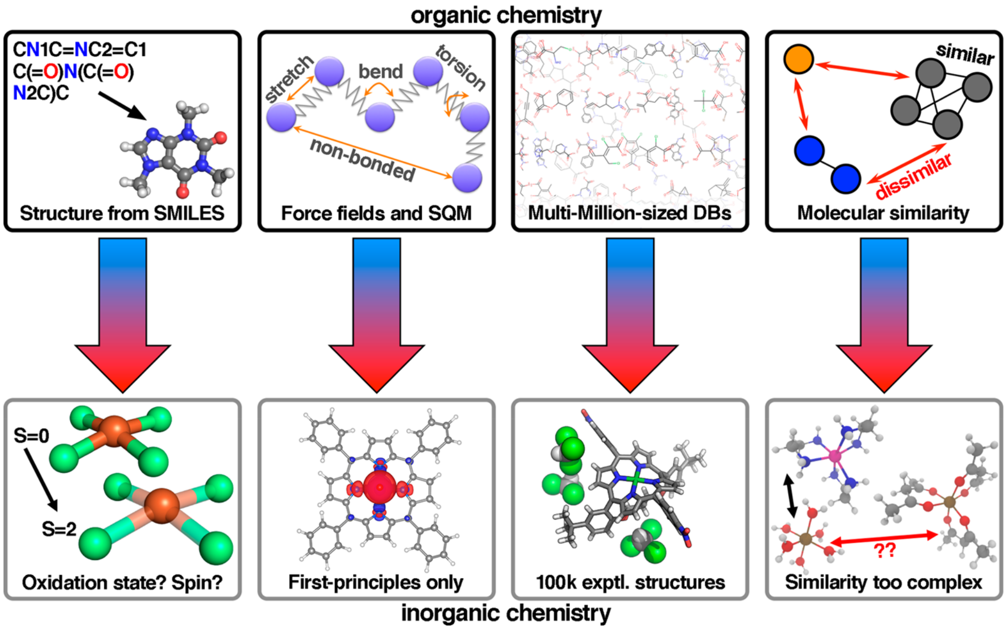


Figure 1. Differences between computational high-throughput screening in organic chemistry (top) and inorganic chemistry (bottom). From left to right: structure generation, simulation methodology, database accessibility, and concepts of molecular similarity.

efforts of thousands of researchers in chemistry, materials science, and engineering worldwide. However, over the course of the lifetimes of these researchers, only a small dent will likely be made with traditional Edisonian approaches applied to this vast unexplored chemical space. Unique challenges arise in transition-metal chemistry and catalysis. The variable spin, oxidation state, and coordination environments favored by elements with well-localized d or f electrons provide great opportunity for tailoring properties in catalysis[17,66−72] or functional (e.g., magnetic) materials.[73−81] At the same time, this combinatorial challenge increases the uncertainty in how to best explore this vast chemical space to satisfy design objectives.

Although the need for accelerated exploration of chemical space is shared by experimental and computational researchers, the recent advances outlined here have poised computational chemistry to make important contributions to discovery efforts. Over the past few years, our group's focus on how to tackle inorganic chemistry design challenges through advances in computation has been shaped by addressing **five key mandates**:

(1) **Automate the simulation of new compounds.** Until recently, it was not uncommon for new transition-metal complex and catalyst simulation candidates to be built by hand. Accelerated discovery requires tools that enable the automated generation of high-quality structures for rapid simulation, both to eliminate a potential source of human error and to remove bottlenecks to the large-scale discovery of new molecules and materials.

(2) **Quantify prediction sensitivity or accuracy.** Even as methods become more and more accurate, small changes in a computational method choice can have a substantial effect on the predicted activity of a catalyst or promise of a material. In most practical cases, where systematic improvement to chemical accuracy may be beyond reach, a design effort must operate with an awareness of the prediction sensitivity of the chosen method.

(3) **Develop faster-than-fast property predictions.** Despite orders-of-magnitude acceleration of first-principles simulation in recent years, direct combinatorial simulation will barely scratch the surface of the vast challenge that is unexplored chemical space. An alternative approach that can predict molecular or materials properties without first-principles computational cost is essential to advancing rapid chemical discovery.

(4) **Map and traverse chemical space.** To overcome combinatorial challenges, a "map" of where compounds sit with respect to each other in chemical space is needed. This map can help researchers identify what the most promising regions are for a particular target functionality and focus on only a small fraction of an otherwise unexplorable space.

(5) **Reveal design rules on the kilocompound scale.** The output of any high-throughput screen should never be a single molecule as the only promising candidate to solve an outstanding challenge. There are far too many unforeseen limitations, such as synthesizability, stability, and market-sensitive cost of materials, that cannot be anticipated completely beforehand. Computational high-throughput screening will be most valuable when it reveals the design features that improve a molecule's performance. As data set sizes get larger, the tools that can reveal and encapsulate design rules will necessarily differ from simpler models that could be used for smaller, narrower data sets studied in the past.

To begin to solve these challenges in inorganic chemistry, we first take inspiration from organic chemistry. Here, machine-readable representations such as the simplified molecular input line entry system[82] (SMILES) string tell us nearly all we need to know about a molecule. With a SMILES string, precise three-dimensional (3D) structures can be generated,[83] leading to routine force-field,[84,85] semiempirical,[86−88] or first-principles characterization with high accuracy.[89] Such representations also lend themselves to quantitative structure−property relationship (QSPR) models[90,91] that enable even more complex mappings between the chemical composition and physicochemical properties (e.g., bioavailability[92,93]). For this reason, it is not surprising that machine learning (ML) models have excelled in encapsulating organic molecule chemical bonding.[94−98] Large multimillion molecule databases of organic compounds[99,100] are an excellent source for chemical discovery. To avoid exhaustive study of that entire space, concepts of molecular similarity may be exploited to identify the most diverse subset of compounds within such databases.[101] Applications in organic chemistry also benefit from all of these concepts being distilled in open-source tools, such as *RDkit*[102] and *OpenBabel*.[83,103]

Conversely, in inorganic chemistry, accurate generation of a 3D structure from a SMILES string must be carried out in a spin- and oxidation-state-dependent manner. Few force fields[104,105] or semiempirical methods[106] have been developed to be predictive for inorganic chemistry, mandating more computationally demanding first-principles simulation with results that are very sensitive to method parameters.[107−116] Here, QSPRs are often specific to a single metal, oxidation state, and spin state, thus enabling focus on properties of the ligand rather than metal-specific properties.[58,117−120] Repositories such as the Cambridge Structural Database[121] only have thousands of inorganic complex structures, and these represent compounds that have been characterized, crystallized, and published, limiting their promise as a resource for the discovery of truly new chemistry. Concepts of molecular similarity are less well-defined: a homoleptic manganese(II) ethylenediamine complex and hexaaqua iron(III) behave more similarly to each other than either does to an $\mathrm{Fe^{III}}$(acac) complex, despite the latter two sharing the same metal, oxidation state, and immediate coordination environment (Figure 1). Advances beyond each of these limitations in inorganic chemistry are essential to addressing the broader challenges we have outlined.

We have taken a divide-and-conquer approach to address this challenge: using techniques that work in organic chemistry and devising new syntax and tools where conventional approaches would fail because of the uncertainty of inorganic complex modeling. The rest of this manuscript is as follows. In section 2, we provide the computational details of the calculations employed in this work. In section 3, we present case studies illustrating successes and remaining opportunities for improvement that arise in tackling the five key challenges that we have outlined here. Finally, in section 4, we provide our conclusions and outlook for the most important obstacles that remain toward realizing the goal of fully autonomous, accelerated computational discovery in inorganic chemistry.

2. COMPUTATIONAL DETAILS

In this work, we carry out original analysis of DFT data sets and trained ML models generated in prior work.[18,122−124] We concisely summarize some of the details of these efforts here but refer the reader to the original work for more detail. The compounds in section 3a are homoleptic complexes generated in ref 124. The 5664 compound space was generated in ref 18, characterized with an artificial neural network (ANN) from ref 123, as well as with DFT, as outlined in sections 3b and 3d. The revised autocorrelation (RAC) feature selection and kernel ridge regression (KRR) models detailed in sections 3c and 3e are from ref 122. ML models introduced in this work include (i) an ANN that separately predicts equatorial and axial metal−ligand (M−L) bond lengths, trained on data from refs 122 and 123, and (ii) ANNs trained on MCDL-25 descriptors that predict the redox and ionization (IP) potentials. These new ML models that facilitate analysis are freely available online as part of the *molSimplify*[125] code and are detailed further in the Supporting Information (SI). The complete-active-space perturbation theory (CASPT2) benchmarks and Perdew−Burke−Ernzerhof (PBE) functional tuning results are derived from prior work[126] and outlined in the SI.

For all other simulations, a consistent workflow was employed. The *molSimplify*[125] toolkit was used to generate octahedral transition-metal complex structures from a pool of organic ligands common in inorganic chemistry with enforced equatorial symmetry but allowing up to two distinct axial ligands. DFT geometry optimizations were then carried out using *TeraChem*[1,127] with the B3LYP[128−130] hybrid DFT functional, occasionally varying the fraction of Hartree−Fock (HF) exchange from its default 20% value, as indicated in the text. The LANL2DZ[131] effective core potential was employed for transition metals with the 6-31G* basis set 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.[132] The metals studied throughout were Cr, Mn, Fe, and Co in $\mathrm{M^{II}}$ and $\mathrm{M^{III}}$ oxidation states. The high-spin/low-spin definitions are generally quintet-singlet for d^6 $\mathrm{Co^{III}/Fe^{II}}$, sextet-doublet for d^5 $\mathrm{Fe^{III}/Mn^{II}}$, quintet-singlet for d^4 $\mathrm{Mn^{III}/Cr^{II}}$, and quartet-doublet for both d^3 $\mathrm{Cr^{III}}$ and d^7 $\mathrm{Co^{II}}$. The ground-state spin of the $\mathrm{M^{II}}$ ion was used to select a $\mathrm{M^{III}}$ spin state that differed by a single electron ionization, and the relaxed $\mathrm{M^{III}}$ complex energy was used to compute the adiabatic IPs reported in this work. Reported redox potentials refer to these IPs corrected by thermodynamic corrections and solvent single-point energies with the conductor-like solvent model,[133,134] as evaluated through a thermodynamic cycle approach.[135−137] Relevant details for specific data sets and ML models are also provided in the SI.

3. RESULTS AND DISCUSSION

3a. Automating Simulation.

To conduct a high-throughput computational screen of transition-metal complexes, thousands of precise structures must be generated. The universal force field (UFF)[105] is one of the few force fields[104,105] to provide support for transition-metal coordination environments, e.g., for preoptimization prior to DFT simulation. Even then, the UFF parameters are only available for a single spin state in limited metal, oxidation state, and coordination geometries; e.g., $\mathrm{Cr^{III}}$, $\mathrm{Mn^{II}}$, $\mathrm{Fe^{II}}$, and $\mathrm{Co^{III}}$ are in the UFF, but $\mathrm{Cr^{II}}$, $\mathrm{Mn^{III}}$, $\mathrm{Fe^{III}}$, and $\mathrm{Co^{II}}$ are not. Force-field optimization requires that all M−L bonds are properly sensed from reading in a structure or SMILES string and have valid force-field parameters associated with them. Thus, for automation in inorganic chemistry, UFF optimization following structure generation can be expected to be unsatisfactory because of (i) difficulty in assigning bond order in metal complexes that will limit the scope of automation and (ii) a dearth of oxidation- and spin-state-dependent parameters that will limit accuracy.

To this end, we have developed a divide-and-conquer approach in our open-source toolkit *molSimplify*.[124,125,132,138] In this approach, force fields are applied only to the organic ligands, whereas the M−L bond lengths are assigned in a data-driven manner to reproduce final DFT-optimized bond lengths as much as possible. The code generates both starting coordinates and input files to enable fully automated simulation of transition-metal complexes. In its original incarnation, the data-driven approach corresponded to a database of prior DFT calculations.[125] We have since generalized this approach to use an ANN to predict the M−L bond length in a spin- and oxidation-state-dependent manner.[123]

To illustrate the benefit of this divide-and-conquer approach, we now compare the structural properties of a set of 66 homoleptic octahedral complexes for which we also have full

DOI: 10.1021/acs.inorgchem.9b00109
Inorg. Chem. 2019, 58, 10592−10606

DFT geometry optimizations. These complexes consist of $Cr^{II/III}$, $Mn^{II/III}$, $Fe^{II/III}$, and $Co^{II/III}$ metals in their low and high spin states complexed with water, carbonyl, methylisocyanide (misc), furan, and pyridine (pyr) ligands, as outlined in prior work[122,124] (see the Computational Details). For each metal and ligand combination, a single UFF bond length is available, and M−L bond-length absolute (abs.) errors are distributed over a wide range from 0.00 to 0.37 Å (Figure 2). There is little correlation between the metal oxidation state for which the parameters were defined in the UFF and the complexes for which the M−L bond-length errors are lowest (Tables S1−S6). For example, the UFF has parameters for Fe^{II}, but the Fe−C bond length predicted by the UFF agrees best with the doublet iron(III) hexacarbonyl complex and most poorly with the quintet iron(II) hexacarbonyl complex (Table S2).

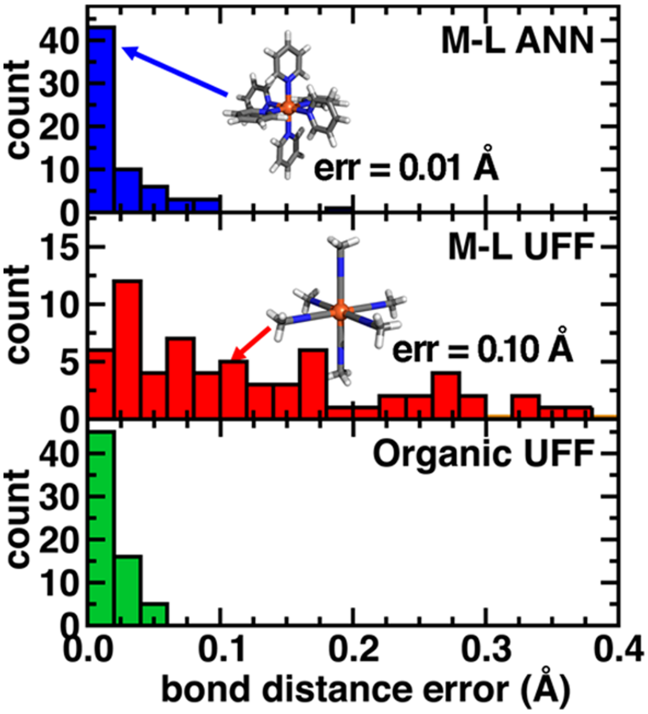


Figure 2. Unnormalized histogram of geometric structure absolute errors (in Å) for 66 $M^{II/III}$ homoleptic octahedral complexes with M = Cr, Mn, Fe, or Co. (top) Absolute errors in M−L bond-length prediction with an ANN, (middle) absolute errors in M−L bond-length prediction with the UFF, and (bottom) absolute errors for the organic bond lengths in ligands over the same subset with the UFF. Representative compounds with median errors for M−L prediction are shown in the inset: 0.01 Å abs. error for the ANN example of quintet $Fe^{II}(pyr)_6$ and 0.10 Å abs. error for the UFF example of singlet $Fe^{II}(misc)_6$.

A typical abs. bond-length error for the UFF in this data set of 0.10 Å is observed for singlet $Fe^{II}(misc)_6$ (Figure 2). Here, the UFF overestimates the uniform, DFT-optimized bond length of 1.93 Å, instead predicting a 2.03 Å bond, whereas the ANN predicts the 1.93 Å bond length to within model precision (Table S3). One of the key advantages of the ANN introduced in this work is that it separately predicts equatorial and axial ligand bond lengths, which will be identical in symmetric complexes or differ in cases of Jahn−Teller distortion. For example, the ANN correctly predicts that the shortest and longest M−L bonds differ by 0.18 Å (0.20 Å for DFT) in quintet $Mn^{III}(CO)_6$, but it also predicts a much smaller distortion in the singlet $Mn^{III}(CO)_6$ of 0.05 Å (0.03 Å for DFT; Table S2). This leads to much smaller abs. errors from the ANN with a narrower 0.00−0.09 Å range and a lower median abs. error (0.01 vs 0.10 Å for the UFF; Figure 2). A compound with an ANN M−L bond length prediction error close to the median abs. error of 0.01Å is quintet $Fe^{II}(pyr)_6$ (Figure 2). The ANN predicts distinct equatorial 2.29 Å and axial 2.36 Å bond lengths in excellent agreement with DFT values (2.29 and 2.35 Å), whereas the UFF predicts a symmetric compound with much too short 2.13 Å M−L bonds for an abs. error of 0.18 Å (Table S5). Overall, significant UFF M−L errors can be attributed to both over- and underestimation, and large errors are observed in both symmetric and Jahn−Teller distorted complexes.

Because we only employ the force field for the organic ligands in automated structure generation with *molSimplify*,[124,125] more sophisticated force fields designed only for organic chemistry, such as MMFF94,[139] can also be employed. We first consider whether the UFF performance on organic bonds in the inorganic complexes is more robust than the M−L bond-length prediction. Indeed, abs. errors of the UFF on the organic ligand bond lengths are much lower, spanning a 0.00−0.05 Å range with a 0.01 Å median (Figure 2). For each inorganic complex, deviations of the organic ligand bond lengths are modest and fall within the expected ranges for organic bond lengths, meaning that one set of UFF parameters can perform well for all structures (Tables S1−S6).

We can also systematically improve upon the UFF results with MMFF94, which outperforms the already good performance of the UFF for organic bonds except in the case of the carbonyl triple bond (Tables S1−S6). For example, MMFF94 correctly differentiates a longer 1.41 Å C−C and a shorter 1.37 Å C═C bond in furan (1.40 and 1.35 Å in DFT), where the UFF predicts identical 1.38 Å bond lengths for both the C═C and C−C bonds (Table S4). MMFF94 also performs better at predicting heteroatom angles (e.g., C−O−C in furan or C−N−C in pyridine). In pyridine, the C−N−C angle should be below the 120° value for benzene, a phenomenon captured by DFT (117°) and MMFF94 (119°), but it is instead larger with the UFF at 122° (Table S5). Overall, a data-driven divide-and-conquer approach automates structure generation with the accuracy needed for inorganic complex simulation with DFT.

3b. Quantifying Prediction Sensitivity. Within high-throughput computational screening, a simulation method should be black box and efficient, often making DFT the method of choice for materials where a single functional can be expected to perform reliably. At the same time, DFT predictions in transition-metal chemistry are highly sensitive to the exchange-correlation approximation.[109−116] In studies that cover a large number of metals and ligands, it can be challenging to confidently select an exchange-correlation approximation because the best one can vary strongly with the coordination environment.[109−113,140,141] Uncertainty quantification is one technique that has been pursued, e.g., through Bayesian[142−144] and other statistical[145−148] methodologies, but the transition-metal chemistry of interest in this work motivates us to focus instead on sensitivity analysis. Although a large number of functional parameters can influence predictions in transition-metal chemistry, the degree of incorporation of HF exchange generally has the most significant effect[42,109,110,112,113,115,116,140,149,150] on transition-metal complex property prediction because of its penalization of unphysical delocalization.[40,42,48,127,151]

As a demonstration, we computed the properties of transition-metal complexes using DFT functionals with varying HF exchange fractions and with higher-level, multireference CASPT2, a correlated WFT approach. We selected homoleptic hexaaqua and hexaammine octahedral M^{II} (M = Cr, Mn, or Fe) and M^{III} (M = Mn, Fe, or Co) complexes with four to six nominal 3d electrons in the three accessible high, intermediate, and low spin states (see the Computational Details). These complexes were selected for their modest size, which makes large-active-space CASPT2 tractable (see the Computational Details and Table S7). For all complexes, we computed the exchange fraction in a hybrid functional needed to recover the CASPT2 high-spin/low-spin or high-spin/intermediate-spin splitting energies (Figure 3 and Table S7). Aside from the high-spin/low-spin gaps of $Cr^{II}(H_2O)_6$ and $Mn^{III}(NH_3)_6$, nearly all complexes require a significant HF exchange fraction to recover the CASPT2 spin splitting (Figure 3). However, the optimal degree of exchange is very metal- and spin-state-specific. For example, the optimal exchange fractions to predict high-spin/low-spin and high-spin/intermediate-spin energetics in $Mn^{II}(NH_3)_6$ and $Fe^{II}(NH_3)_6$ are similar, but the two values differ substantially for $Fe^{III}(NH_3)_6$ (Figure 3). Similarly, within a single nominal d filling, optimal exchange fractions are comparable at around 40% for all d^4 high-spin/intermediate spin splitting energies, but they have a large 40−85% range for the same quantity in d^5 complexes (Figure 3).

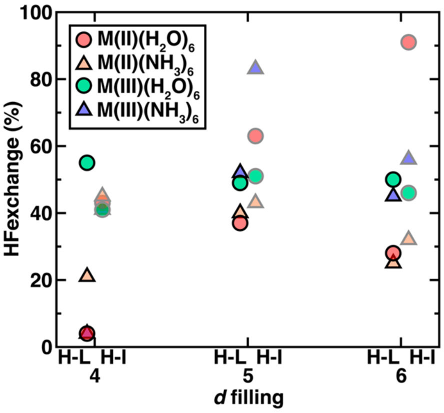


Figure 3. Optimal exchange fractions (HF exchange, from 0 to 100%) in a modified PBE0 functional with the def2-TZVP basis set to match the CASPT2 results for ΔE_{H-L} (left symbols) or ΔE_{H-I} (right symbols). The results are shown for hexaaqua (circles) or hexaammine (triangles) complexes in M^{II} (i.e., Cr, Mn, and Fe) and M^{III} (i.e., Mn, Fe, and Co) oxidation states grouped by nominal d filling and colored as in the inset legend.

Because no single functional will predict optimal energetics, quantifying the sensitivity of identified leads to functional parameters provides essential insight into the uncertainties inherent in a DFT-led computational screen. We recently[18] studied a space of 5664 complexes to discover candidate spin-crossover (SCO) complexes. This pool of candidates consisted of octahedral complexes with $M^{II/III}$ centers, where M = Cr, Mn, Fe, or Co, which we studied using both DFT and an ANN that predicts high-spin/low-spin splitting,[123] i.e., ΔE_{H-L}. Now, we leverage the fact that the ANN is trained to predict ΔE_{H-L} over a range of HF exchange fractions and can make these predictions over all 5664 compounds in minutes, allowing us to understand how lead SCOs differ with varying HF exchange.

As in previous work,[18] we define SCO lead compounds as those with $|\Delta E_{H-L}| < 5$ kcal/mol, as judged through electronic energy differences. Incorporating thermodynamic and solvent corrections will shift some of these compounds from this SCO definition, but it generally preserves most SCO leads observed in the gas phase.[18] We can evaluate where leads sit on a map of the full compound space obtained through dimensionality reduction. Specifically, we reduce a predictive ca. 40-dimensional representation of the complexes to the two most informative dimensions with t-stochastic neighbor embedding (t-SNE)[152] (Table S8). t-SNE takes all complexes defined by their multidimensional representation and attempts to preserve the pairwise distribution of distances between complexes in a two-dimensional mapping.[152] This approach improves upon principal-component analysis, which takes only the first two combinations of descriptors in a representation but can make complexes that are distant only in higher components appear close to each other. When we color the t-SNE map by the properties of the complexes, we can observe trends. With this map, we observe that lead compounds are distributed in different portions of the space as the HF exchange fraction is varied (Figure 4). Higher HF exchange (20%) leads populate stronger field (i.e., ligands with a coordinating carbon) regions of space, whereas low HF exchange (5%) leads reside more frequently in weak field (i.e., ligands with a coordinating oxygen) regions of the space (Figure 4). These observations are to be expected: pure DFT functionals or those with low HF exchange fractions have a low-spin bias,[44,115,138,153] which would cause only weak ligand fields to correspond to SCOs. Regardless, even a small change in HF exchange from 20% to 15% (e.g., as in B3LYP*[109]) also shifts the distribution of leads, demonstrating high sensitivity of ΔE_{H-L} to HF exchange (Figure 4).

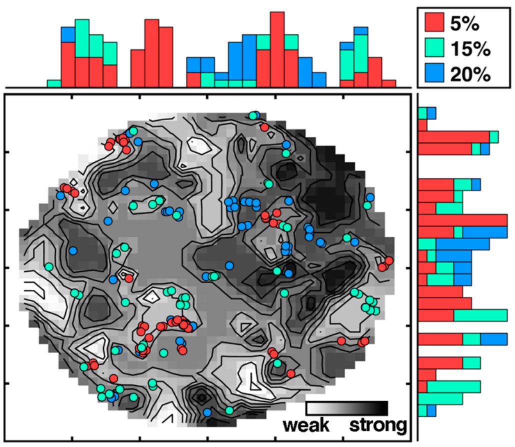


Figure 4. t-SNE[152] plot of the full compound space colored by the connecting-atom ligand field from weak (white) to strong (black). Lead SCO complexes (i.e., $|\Delta E_{H-L}| < 5$ kcal/mol) are shown as circle symbols at three HF exchange fractions: 5% (red circles), 15% (green circles), and 20% (blue circles). One-dimensional stacked histograms of lead compounds are shown for a projection of the two t-SNE dimensions with the same coloring by the exchange fraction as the circle symbols and also shown in the inset legend.

An example 5% exchange SCO lead is a homoleptic complex with weak-field 4-cyanopyridine ligands, whereas at 20% exchange, this SCO lead has been replaced by a similar compound with axial ligands substituted for strong-field *tert*-butylisocyanide (Figure 5). On the basis of our analysis of the

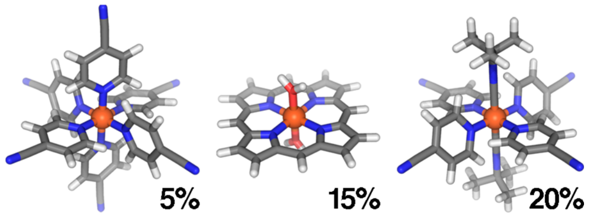


Figure 5. Example Fe^{II} complex leads at different percentages of HF exchange: a homoleptic 4-cyanopyridine complex at 5% (left), equatorial porphine with two axial water ligands at 15% (middle), and a heteroleptic complex with equatorial 4-cyanopyridine and axial *tert*-butylisocyanide ligands at 20% (right).

CASPT2 agreement with hybrid functionals, we anticipate the 20% exchange result to be in better agreement for Fe^{II} complexes, but across a broader compound space, the optimal functional will vary. For example, 15% exchange in B3LYP* has been previously motivated for other Fe^{II} complexes.[109] One 15% exchange lead has a porphine equatorial macrocyclic ligand, which is a stronger field ligand than 4-cyanopyridine, but this effect is balanced by weaker-field water axial ligands, in line with the expected intermediate behavior of 15% HF exchange (Figure 5).

Spin splitting remains one of the properties most sensitive to functional choice, but some catalyst energetics have comparable sensitivities when strong changes in delocalization occur across the reaction coordinate.[113,154] Although there is no universally optimal functional, the relationship between the functional choice and chemical composition is inherently learnable in these data-driven models.[113,116,123] For any design objective, data-driven methods such as those described here can both predict the uncertainty of the design landscape with changing functional and estimate derivatives of the properties to functional parameter variation.[123]

3c. Faster-Than-Fast Property Prediction. Now we consider the extent to which these ML models can replace DFT during high-throughput screening efforts in open-shell transition-metal chemistry. A DFT simulation that would require at least an hour for a single evaluation should require less than a second for evaluation with an ML model, enabling faster-than-fast property prediction. For data-driven property prediction, three essential ingredients are (i) training data-set size, (ii) ML model architecture, and (iii) the representation provided as inputs to the ML model. In inorganic chemistry ML, it is challenging to generate data-set sizes as large as small molecule sets that have been developed[155] for organic chemistry ML. The smallest nontrivial octahedral transition-metal complex has seven heavy atoms, whereas many organic chemistry data sets consist of C-, N-, O-, and F-containing compounds with only nine heavy atoms or less and many fewer electrons than a transition-metal complex. Thus, it becomes essential to make careful choices[122,123] of the representation and ML model to ensure that models can be predictive when trained on smaller data sets.

We recently used two distinct approaches to develop representations specifically for open-shell transition-metal chemistry. In both cases, we required that only heuristic information related to the connectivity and composition be part of the representation so that the structure could be predicted (see section 3a) and to account for the fact that no low-level theory could be used to generate an initial structure to provide as input, unlike organic chemistry. First, we developed a combination of 25 mixed continuous and discrete local (MCDL-25) features[123] (Table S9). Guided by chemical intuition, we selected features that focused on the metal and the surrounding ligand-field environment using simple, regularized linear regression models to evaluate the predictive capability of candidate features. The selected features included metal identity; connecting atom identity, electronegativity, and bond order; and the truncated Kier shape index,[156] a nonlocal feature that characterizes the topology of atoms within three bonds of the atom coordinating to the metal. Most organic chemistry representations focus on the whole molecule, but transition-metal chemistry properties are inherently local. For example, iron(II) methylisocyanide octahedral complexes have spin splitting within 3 kcal/mol of the analogous phenylisocyanide complexes, despite the latter molecule having around 4 times as many atoms.[122]

Using this representation, we trained[123] our first-generation ANN to predict ΔE_{H-L}, in addition to the M−L bond lengths and sensitivity to exchange on a data set of five metals (Cr−Ni) in 2+ and 3+ oxidation states in high and low spin states at a number of HF exchange fractions. By using varying dropout realizations, we regularized our models (i.e., avoided overfitting) and also generated credible intervals on predictions.[123] Overall, the MCDL-25-trained ANN could predict ΔE_{H-L} to within 2.5 kcal/mol on set-aside test complexes and the correct qualitative ground-state spin (i.e., high or low) in 97% of the cases, where the remaining 3% of the cases were challenging because of near-degeneracy of the two spin states.[123] We note here that beyond the ability of the ANN to predict properties at arbitrary percent HF exchange, these varying exchange fractions also served to improve the model accuracy. These distinct energetic results on similar structures serve as a perturbation, much as geometric distortions are carried out in organic chemistry ML model training.[94,157] Predictions at intermediate exchange are supported by lower or higher fractions (abs. errors <2 kcal/mol), whereas errors for extreme exchange values are higher (abs. errors >2 kcal/mol; Figure S1). The model readily reproduces the DFT results, such as the spectrochemical series of homoleptic Fe^{II} complexes (Figure 6). Here, the ML model learns the B3LYP*[109] DFT answers, not the “exact” experimental observations, as indicated by the nonmonotonic spin splitting observed in Fe^{II} when the ligands are ordered by their nominal field strength (Figure 6). The MCDL-25/ANN model is predictive[18] for chemical discovery applications when paired with a constraint on high feature-space-distance-to-train points[18,123] in a genetic algorithm. Such an approach allowed us to score new compounds with an ANN in minutes instead of the days or weeks that full DFT evaluation would have required.[18] Despite applying the ANN to nearly all new complexes, i.e., only 2% were in the training set, the mean absolute errors (MAEs) remained close to the baseline (4.5 kcal/mol) in this discovery effort compared to the DFT results.[18]

In an attempt to remove bias from the feature set, we introduced modified autocorrelation[158] descriptors that we refer to as RACs.[122] RACs were designed to be a complete set of features that ranged from metal-local (as in MCDL-25) to whole-complex (as in typical organic representations) properties. Through feature selection or model-based feature engineering, the most relevant subset of the features from RACs could then be used without a priori assumptions about the most relevant length scales for property prediction.

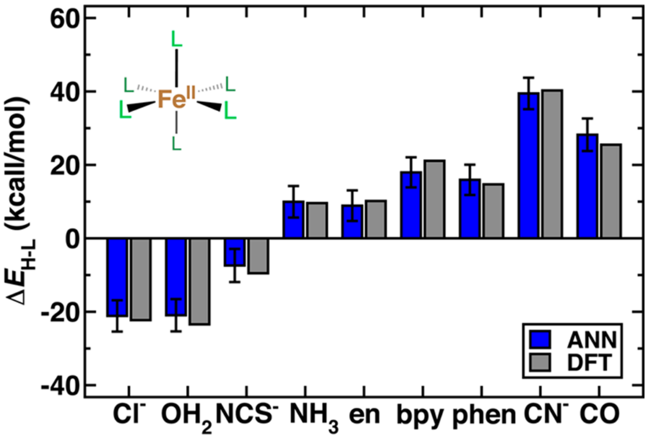


Figure 6. High-spin/low-spin splitting ($\Delta E_{\mathrm{H-L}}$, in kcal/mol) for representative ligands sorted from left to right by increasing experimental ligand-field strength obtained from an ANN and from DFT. Here, DFT corresponds to 15% exchange in B3LYP, i.e., B3LYP*.[109] All complexes are homoleptic $\mathrm{Fe^{II}}$ complexes, as shown schematically in the inset. The error bars are credible intervals on the ANN predictions.

Specifically, RACs have been defined[122] as the sum of products and differences on the molecular graph of five quantities: nuclear charge, electronegativity, covalent radius, topology, and identity. For inorganic chemistry, we define metal- and ligand-centered RACs, in which one of the atoms in the product or difference is the metal or ligand connecting atom, respectively. These products or differences involve two atoms separated by a certain number of bond paths; e.g., zero depth is simply the sum of the products of atomic properties. In prior work, we found no benefit to correlating atoms more than three bond paths away and therefore typically terminated RAC depths at 3, leading to a total of 155 features for octahedral complexes (i.e., RAC-155). However, this truncation does not mean that the whole molecule is not included in the representation because any RAC not centered on the metal or ligand is defined as the sum of the products or differences of atomic properties over the full complex.

To test the RAC performance and feature subsets, we employed the KRR ML model.[122] If specific complexes are close in the representation to a test complex, they will generally contribute most strongly to the KRR model property prediction on this test complex. The disadvantage of the KRR is that if a test complex is too far from any available training data, the KRR model will make no prediction, unlike an ANN. Thus, KRR is a good model for testing representations because it may be thought of as a data-clustering model. We determined the RAC/KRR performance for $\Delta E_{\mathrm{H-L}}$, gas-phase adiabatic IP, solvent-corrected and thermodynamically corrected redox potential, and M−L bond lengths (Figure 7).

Although both representations were previously developed,[122,123] we provide a direct comparison between RAC-155 and MCDL-25 here for the first time. In all cases, the RAC-155/KRR performance improves upon equivalent MCDL-25/ANN models (Figure 7). Spin-splitting MAEs of 2.5 kcal/mol in MCDL-25/ANN were reduced to subkilocalories per mole with RAC-155/KRR, and M−L bond-length errors of 0.025 Å were reduced to around 0.005 Å (Figure 7). We now present newly trained MCDL-25/ANN models to predict the redox potential and IP and observe these to also perform more poorly at around 0.41 and 0.53 eV MAE, respectively, than the 0.32 and 0.35 eV errors for the RAC-155/KRR model (Figure 7 and Table S10). We observe that the redox potential, which is evaluated by combining gas-phase adiabatic IP with solvent and thermodynamic corrections, is no more challenging to predict than the individual IP, although the RAC-155/KRR relative performance improvement over MCDL-25/ANN is larger for the IP prediction.

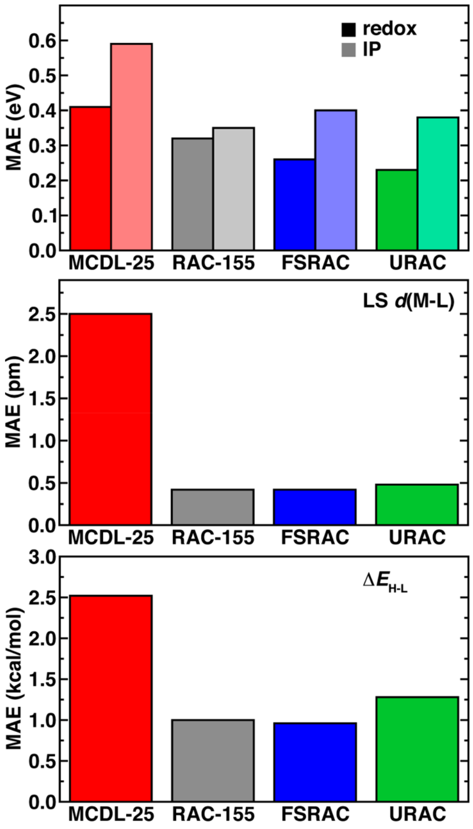


Figure 7. MAE for the redox potential and IP (eV; top), low-spin (LS) M−L bond length (pm; middle), and $\Delta E_{\mathrm{H-L}}$ (kcal/mol; bottom). Comparisons are for the MCDL-25 feature set with an ANN along with KRR models trained with the full RAC-155, a FSRAC subset that performs best for each property as described in ref 122, and URAC, which was a 26 feature set selected by random forest on $\Delta E_{\mathrm{H-L}}$ and found to perform best overall for multiple predictions. The redox potential and IP predictions are shaded with high and low saturation, as indicated in inset legend.

By combining RAC-155 with feature-selection techniques, we could identify the most predictive subset of the RAC features. Using such feature-selection techniques, we were able to down-select to the ca. 40 or so most important features (feature-selected RAC, FSRAC) for each property without significant loss in predictive accuracy, ensuring that RAC-155 performance improvement was not simply due to the larger number of features (Figure 7 and Table S11). We have recently trained models[124] on related properties, specifically orbital energies, such as the highest occupied molecular orbital (HOMO) or HOMO−lowest unoccupied molecular orbital (LUMO) gap. We showed that these orbital energies could be predicted with comparable accuracy to the IP and redox potential models as well.[124] We also found[124] that RAC-155/ANN and FSRAC/KRR models performed comparably, highlighting the complementarity of the two approaches outlined here.

Rather than carrying out a feature selection on one property to build a single model, feature-selected subsets are often transferable to other properties. After a comparison of the test set performance[122] of features selected on bond length, redox potential, and $\Delta E_{\mathrm{H-L}}$, the best performing (i.e., most-transferable) set was determined to be the 26 features most relevant for $\Delta E_{\mathrm{H-L}}$[122] (Figure 7 and Table S11). A KRR model trained with this universal RAC (URAC) set outperforms the full RAC-155 for redox potential prediction and performs comparably on other prediction tasks (Figure 7). Examining the metal-proximal versus whole-complex nature of MCDL-25, RAC-155, and URAC helps to rationalize these observations (Figure 8). On the molecular graph, we define proximal features as those that only involve either the metal or its first coordination sphere, middle features as those involving atoms up to two bonds away from the metal, and distal features as anything involving atoms three bonds or more away (Figure 8). The MCDL-25 subset we created from intuition has more than 50% proximal features and the fewest distal features of the three categories. Conversely, RAC-155, by definition, is heavily weighted toward metal-distal, whole-complex features. Feature selection to form the 26 features in the URAC subset recovers a distribution consistent with MCDL-25: nearly 50% of the features are proximal, and the distal fraction is still the smallest (Figure 8). Thus, URAC systematically recovers metal-centric characteristics of MCDL-25 at the same time reducing the uncertainty in the representation choice through the use of rigorous feature-selection techniques.

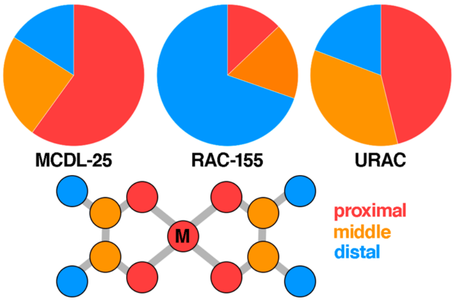


Figure 8. Pie-chart distribution of features in representative feature sets: proximal, which includes the metal and first coordination sphere; middle, which includes up to the second coordination sphere; distal, which includes the third coordination sphere and beyond. The three feature sets compared are MCDL-25, full RAC-155, and URAC, which is a 26 feature set selected by a random forest for its balanced performance across several features. The three feature classes are also shown schematically on the equatorial plane of a complex with two oxalate ligands, where atoms are colored by the feature type, as indicated in the inset legend.

3d. Mapping Transition-Metal Chemical Space. Careful tailoring of representations not only improves property predictions in data-clustering KRR models but also defines molecular similarity and provides a map of relevant chemical space. In prior work,[122] we showed that ligand substitution to transform homoleptic complexes into each other through related heteroleptic complexes followed simple paths in reduced dimensionality representations (i.e., with principle-component analysis) based on the feature sets described earlier (section 3c). When these reduced dimensions capture essential features about the transition-metal complex, they provide a path to visualizing molecular similarity.

Returning to the 5664 compounds[18] that we described earlier, we can examine the reduced dimensionality (i.e., t-SNE[152]) map of properties distinguished by the metal or spin splitting to identify trends among SCO complexes (Figure 9).

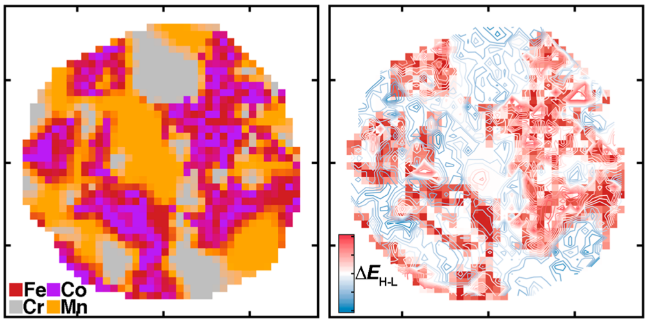


Figure 9. Side-by-side t-SNE plot of the full compound space. The plot is colored in the left panel by the metal center: chromium in gray, manganese in yellow, iron in maroon, and cobalt in purple. The right panel shows the same t-SNE plot with only Fe compounds shown in the background in maroon and full data spin-splitting contours overlaid. The spin-splitting contours range from +40 kcal/mol in red to −40 kcal/mol in blue, passing through white at 0 kcal/mol, as shown in the inset colorbar.

This analysis reveals that Fe and Co cluster together surrounded by islands of more distinct Mn and Cr complexes (Figure 9). For example, $\mathrm{Fe^{II}(acac)_2(tbisc)_2}$ and $\mathrm{Co^{II}(acac)_2(benzisc)_2}$ are both predicted[18] by the ANN to be SCO complexes (−3.9 and 0.1 kcal/mol) at a good level of agreement with DFT (−5.8 and −3.2 kcal/mol; Table S12). These Fe and Co complexes contain the same coordinating environment and differ by only metal-distant functionalization of the axial isocyanide ligands. Because regions of high and low spin are generally smoothly localized, it suggests that complexes are well clustered according to their primary properties by the t-SNE map (see section 3b). Although Fe complexes span regions of both low spin and high spin, Cr complexes are more universally low spin (Figure 9). Thus, regions of Mn and Cr that are intermingled with the associated Fe/Co complexes are similar to the Fe/Co complexes in both properties and representations (Figure 9). For example, an $\mathrm{Mn^{II}(CO)_4(misc)_2}$ SCO (ANN, 3.1 kcal/mol; DFT, 4.7 kcal/mol) has properties comparable to those of an $\mathrm{Fe^{III}(4\text{-}CN\text{-}pyr)_4(pisc)_2}$ complex (ANN, 3.1 kcal/mol; DFT, 4.1 kcal/mol) with only slightly weaker field equatorial ligands (Table S12). The 2−3 kcal/mol error of the ANN with respect to the DFT reference on these example complexes is comparable to the performance described for the overall test set errors in section 3c, but the model performance can degrade if the complex is very different from the training data (see Figures 6 and 7). As long as complexes not too distinct from the training data are being visualized, these maps can be used to reduce the uncertainty associated with meeting the design objectives, for instance, by suggesting substitutions of metals to those with reduced functional sensitivity.

3e. Revealing Design Rules. Analysis of the most essential atomic and connectivity features that give rise to a property in large data sets, e.g., from feature selection, enables us to reveal design rules on a kilocompound scale. Here, our goal is not simply to obtain accurate predictions but to understand which substituent RAC length scale and character is most essential for predicting the redox potential or $\Delta E_{\mathrm{H-L}}$.

The KRR model is invariant to the choice of feature sets because molecules still cluster similarly enough to give comparable test set performances (see section 3c), but a random forest model provides a ranked importance of the features toward making these predictions that can provide additional chemical insights (Tables S8 and S13). Both selected subsets are about 25% of the original RAC feature set (38 for the redox potential and 39 for $\Delta E_{\mathrm{H-L}}$). We first divide the selected RACs by the most metal-distant atom participating in the feature: metal-only, first, second, and third coordination spheres, or more distant features (Figure 10).

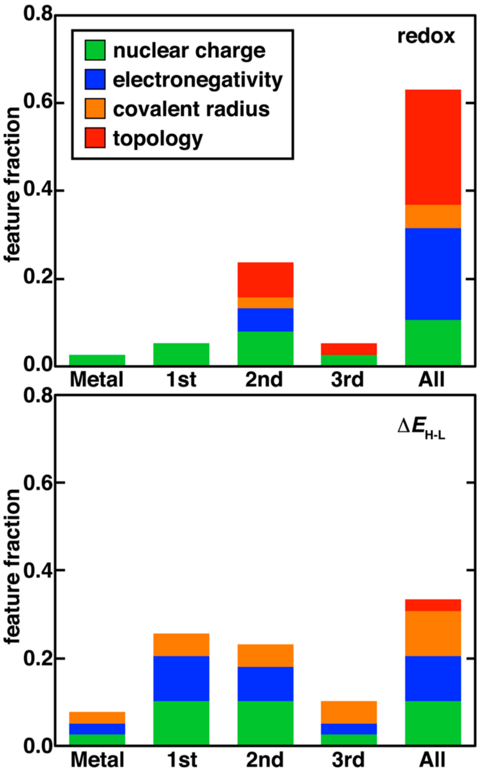


Figure 10. Normalized distribution of selected features for the redox potential (top) and $\Delta E_{\mathrm{H-L}}$ (bottom). Features are grouped by the most distant atoms present: metal-only, first, second, and third coordination sphere or beyond, and fully nonlocal features (All). Within each length scale, features are decomposed into nuclear charge in green, electronegativity in blue, covalent radius in orange, and topology in red (i.e., connectivity-only), as shown in the inset legend.

Although both spin-splitting and redox subsets have features that range from metal-only to whole-complex, the relative weight of these features differs dramatically (Figure 10). Where 70% of the spin-splitting features are within the first three coordination spheres, the opposite is true for the redox potential, with over 60% of the features being wholly nonlocal. Observations similar to those for the redox potential hold for the related IP[122] or frontier orbital energetics[124] quantities.

Beyond length scales, we can also review the nature of each RAC descriptor to identify the most essential heuristic atomic properties that relate back to the overall property prediction task (Figure 10). The spin-splitting subset contains almost no features that depend only on the connectivity (i.e., topology or identity), whereas nearly half of the fully nonlocal RACs in the redox subset are of this type as well as a significant fraction of the second coordination sphere redox RACs (Figure 10 and Tables S8 and S13). The majority of spin-splitting-selected RACs are instead correlations of nuclear charges or electronegativity, with the third most common fraction corresponding to the covalent radius. The metal identity also has a larger overall weight in spin splitting because the metal-only nuclear charge, electronegativity, and covalent radius all contribute, whereas the nuclear charge is the single metal-only RAC present in the redox subset (Figure 10 and Table S13). The nuclear charge, electronegativity, and covalent radius, in principle, encode information about the elemental identity of the atoms in the complex, but the nuclear charge linearly distinguishes each element, whereas moving across the periodic table leads to nonmonotonic changes in the electronegativity or covalent radius.

Overall, the design rules implied by these feature-selected subsets indicate that it should be possible to tailor metal-proximal features, especially those that relate to atomwise electronegativity, to alter the spin-splitting properties. Such an observation is simply a recasting of well-known principles in ligand-field theory. In the case of the redox potential, our descriptors suggest that the relative rigidity and branching of the complex should play an important role in controlling the redox potential. For example, $\mathrm{Mn^{II}(CO)_6}$ and $\mathrm{Fe^{II}(CO)_6}$ have high, comparable redox potentials of 10.3 and 10.4 eV, respectively, owing to the rigid, small CO ligand (Figure 11).

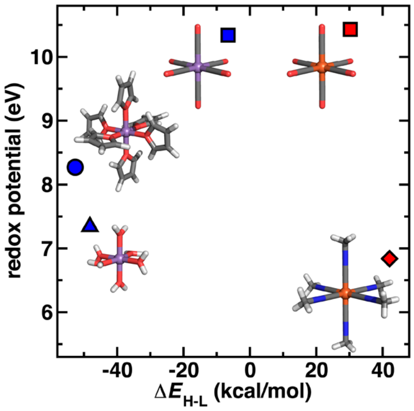


Figure 11. Redox potential (eV) and gas-phase $\Delta E_{\mathrm{H-L}}$ (kcal/mol) values for representative homoleptic complexes: $\mathrm{Mn^{II}(CO)_6}$ (blue square), $\mathrm{Fe^{II}(CO)_6}$ (red square), $\mathrm{Fe^{II}(misc)_6}$ (red diamond), $\mathrm{Mn^{II}(furan)_6}$ (blue circle), and $\mathrm{Mn^{II}(H_2O)_6}$ (blue triangle). The structures are shown in the insets.

However, the overriding influence of the metal- and ligand-field interactions means that the spin splitting of these same two compounds differs much more significantly because the $\mathrm{Fe^{II}}$ complex is low spin by 30 kcal/mol and the $\mathrm{Mn^{II}}$ complex has a $\Delta E_{\mathrm{H-L}}$ of −6.6 kcal/mol (Figure 11 and Table S14). Replacing the carbonyl ligands with misc ligands in a homoleptic $\mathrm{Fe^{II}}$ complex preserves the low-spin character ($\Delta E_{\mathrm{H-L}}$ = 42 kcal/mol) by leaving the first two coordination sphere atoms largely unchanged, whereas the increased branching and number of atoms at the third and fourth coordination sphere lower the redox potential to 6.8 eV (Figure 11 and Table S14).

Strongly high-spin-directing $\mathrm{Mn^{II}}$ combined with weak-field ligands leads to comparable spin-splitting energetics for furan ($\Delta E_{\mathrm{H-L}}$ = −52 kcal/mol) or water ($\Delta E_{\mathrm{H-L}}$ = −48 kcal/mol) again because of the similarity in the direct M−L coordination spheres. Unsaturation of the C atoms in the furan ligand means that a higher redox potential is observed than that for water (8.3 vs 7.3 eV), despite the larger ligand size, highlighting that topological descriptors are relevant for distinguishing not just the overall size but also the overall connectivity (Figures 10 and 11). Thus, using these design principles, one could tailor ligand chemistry to tune the redox potential while making more limited metal-centered substitutions to tailor the spin state in an orthogonal multiobjective design approach.

4. CONCLUSIONS AND OUTLOOK

In this work, we laid out five key mandates for computationally accelerated discovery in inorganic chemistry: (i) to automate the simulation of new compounds including through structure generation and calculation automation, (ii) to quantify the sensitivity or accuracy of any computational predictions made, (iii) to develop models that enable faster-than-fast property prediction to overcome bottlenecks in first-principles characterization, (iv) to devise ways to map and rapidly traverse interesting regions of chemical space, and (v) to reveal design rules on the kilocompound scale.

We have described how progress in each of these areas has provided new chemical insights and pathways to accelerated transition-metal complex design in the face of uncertainty. We described how ML models can improve structure prediction, enable an understanding of the prediction sensitivity to the DFT functional, and replace DFT entirely for chemical space exploration. We also showed how data science tools such as new representations tailored for inorganic chemistry yield new chemical insight. These representations combined with dimensionality reduction revealed chemical similarity in open-shell transition-metal chemistry as well as design rules for independently tailoring multiple properties atom-by-atom in transition-metal complexes.

Despite this progress in achieving the five mandates laid out here, outstanding challenges remain in the effort toward the accelerated computational discovery of new inorganic complexes and materials. We now highlight just a few that we believe to be most essential toward furthering this goal. Beyond automated simulation, fully autonomous workflows become essential in reducing human intervention in computational design efforts. On its own, simply automating simulation can lead to an increased number of failed and uninformative calculations because the researcher is no longer supervising and interacting with each simulation. We envision that the same data-driven models developed for property predictions could instead be developed to replace the basic decisions employed by a computational researcher when deciding which simulations to carry out. This could involve training a model to predict if a simulation will fail or detect a simulation failure in progress, for instance, as judged through whether a molecule stays intact. More sophisticated models could be developed to detect when one method (e.g., DFT) is not sufficiently accurate or predictive for a region of chemical space of interest, prompting automatic adaptation to systematically improved but more computationally costly methods. Here, additional challenges will arise in developing sufficient training sets that contain molecules small enough for even the most accurate methods. Alternatively, experimental data sets could be curated to guide data-driven method accuracy models. With these efforts, additional uncertainty metrics will necessarily be developed for both ML and theoretical models. With sufficient data, many of the decisions that require years of training in computational chemistry could instead be encapsulated in an artificial intelligence engine that drives simulation.

Despite the challenges outlined here, the most straightforward application of computational chemistry is in the tailoring of energetics, e.g., to search for catalysts with optimal activity by tuning an activation energy, to optimize materials to have a specific band gap, or to design new SCO materials. In most practical design challenges, the requirements of a new material are much more multifaceted. Synthetic feasibility or cost may be a concern, both in terms of the use of rare elements or in terms of requiring many laborious steps. Environmental conditions also remain a challenge: the best redox couple for a redox flow battery may lack solubility in the optimal solvent in which it would operate. Interactions with a multitude of species generated during operating conditions could lead to catalyst inactivation and materials degradation. Thus, descriptors of the stability and feasibility are as essential as the energetic descriptors of activity. These, in turn, could be used for multiobjective strategies to design materials that are not just active but also stable or inexpensive. To minimize computational overhead when searching such large spaces, active learning approaches that maximize new information to the model will also be necessary. Although multiobjective optimization strategies are well developed in other research fields, embracing and adapting these strategies in computational chemistry will become essential as computing power and new models make it increasingly feasible to explore large regions of chemical space.

We have outlined just a handful of next steps toward realizing the goal of end-to-end design of transition-metal complexes. We envision the challenge of discovery in inorganic chemistry as one best addressed by a broad software and methodological toolset that can readily adapt to the properties and design objectives beyond those discussed in this work. As new critical challenges in resource and energy utilization arise, we expect such a flexible toolset to be readily adapted.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.inorgchem.9b00109.

Structural properties of 66 octahedral transition-metal complexes with water, carbonyl, misc, furan, and pyridine ligands obtained from DFT geometry optimization, ANN predictions, and UFF or MMFF94 force fields, errors of the force fields or ANN with respect to DFT geometry optimization, tuning of DFT functionals to reproduce CASPT2 results, MCDL-25 descriptors, MCDL-25 errors averaged by an exchange fraction, hyperparameters of additional ML models, feature selected subsets for KRR comparisons, comparison of SCO leads with varying metal and oxidation states using both ANN and DFT, redox potential and spin-splitting RAC featuring subset characteristics, and spin and redox properties of representative complexes (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hjkulik@mit.edu. Phone: 617-253-4584.

ORCID

Jon Paul Janet: 0000-0001-7825-4797

Fang Liu: 0000-0003-1322-4997
Aditya Nandy: 0000-0001-7137-5449
Chenru Duan: 0000-0003-2592-4237
Tzuhsiung Yang: 0000-0002-6751-9806
Heather J. Kulik: 0000-0001-9342-0191

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge support by the Office of Naval Research under Grants N00014-17-1-2956 and N00014-18-1-2434, DARPA under Grant D18AP00039, the Department of Energy under Grant DE-SC0018096, the National Science Foundation under Grant CBET-1704266, and MIT Energy Initiative seed grants (2014 and 2017). H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. 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) 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.

(2) Ufimtsev, I. S.; Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **2008**, *4*, 222−231.

(3) Ufimtsev, I. S.; Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *J. Chem. Theory Comput.* **2009**, *5*, 1004−1015.

(4) Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S. Linear-Scaling Methods in Quantum Chemistry. *Rev. Comput. Chem.* **2007**, *23*, 1.

(5) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. Auxiliary Basis Sets for Main Row Atoms and Transition Metals and Their Use to Approximate Coulomb Potentials. *Theor. Chem. Acc.* **1997**, *97*, 119−124.

(6) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283−290.

(7) Libisch, F.; Huang, C.; Carter, E. A. Embedded Correlated Wavefunction Schemes: Theory and Applications. *Acc. Chem. Res.* **2014**, *47*, 2768−2775.

(8) Challacombe, M.; Schwegler, E. Linear Scaling Computation of the Fock Matrix. *J. Chem. Phys.* **1997**, *106*, 5526−5536.

(9) Hampel, C.; Werner, H. J. Local Treatment of Electron Correlation in Coupled Cluster Theory. *J. Chem. Phys.* **1996**, *104*, 6286−6297.

(10) Schütz, M.; Hetzer, G.; Werner, H.-J. Low-Order Scaling Local Electron Correlation Methods. I. Linear Scaling Local MP2. *J. Chem. Phys.* **1999**, *111*, 5691−5705.

(11) Hohenstein, E. G.; Parrish, R. M.; Martínez, T. J. Tensor Hypercontraction Density Fitting. I. Quartic Scaling Second-and Third-Order Møller-Plesset Perturbation Theory. *J. Chem. Phys.* **2012**, *137*, 044103.

(12) Song, C.; Martínez, T. J. Reduced Scaling CASPT2 Using Supporting Subspaces and Tensor Hyper-Contraction. *J. Chem. Phys.* **2018**, *149*, 044108.

(13) Andermatt, S.; Cha, J.; Schiffmann, F.; VandeVondele, J. Combining Linear-Scaling DFT with Subsystem DFT in Born−Oppenheimer and Ehrenfest Molecular Dynamics Simulations: From Molecules to a Virus in Solution. *J. Chem. Theory Comput.* **2016**, *12*, 3214−3227.

(14) Shu, Y.; Levine, B. G. Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015**, *142*, 104104.

(15) 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. L.; 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.

(16) 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.

(17) 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**, DOI: 10.1021/acs.chemrev.8b00361.

(18) 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.

(19) 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.

(20) Curtarolo, S.; Setyawan, W.; Hart, G. L.; 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.

(21) 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.

(22) Nørskov, J. K.; Bligaard, T. The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776−777.

(23) Han, S.; Huang, Y.; Watanabe, T.; Dai, Y.; Walton, K. S.; Nair, S.; Sholl, D. S.; Meredith, J. C. High-Throughput Screening of Metal−Organic Frameworks for CO2 Separation. *ACS Comb. Sci.* **2012**, *14*, 263−267.

(24) 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*, 83−89.

(25) Witman, M.; Ling, S.; Anderson, S.; Tong, L.; Stylianou, K. C.; Slater, B.; Smit, B.; Haranczyk, M. In Silico Design and Screening of Hypothetical Mof-74 Analogs and Their Experimental Synthesis. *Chem. Sci.* **2016**, *7*, 6263−6272.

(26) Ufimtsev, I. S.; Luehr, N.; Martínez, T. J. Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011**, *2*, 1789−1793.

(27) Kulik, H. J. Large-Scale QM/MM Free Energy Simulations of Enzyme Catalysis Reveal the Influence of Charge Transfer. *Phys. Chem. Chem. Phys.* **2018**, *20*, 20650−20660.

(28) Fales, B. S.; Levine, B. G. Nanoscale Multireference Quantum Chemistry: Full Configuration Interaction on Graphical Processing Units. *J. Chem. Theory Comput.* **2015**, *11*, 4708−4716.

(29) Zhao, Q.; Kulik, H. J. Electronic Structure Origins of Surface-Dependent Growth in III-V Quantum Dots. *Chem. Mater.* **2018**, *30*, 7154−7165.

(30) Kümmel, S.; Kronik, L. Orbital-Dependent Density Functionals: Theory and Applications. *Rev. Mod. Phys.* **2008**, *80*, 3−60.

(31) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(32) Livshits, E.; Baer, R. A Well-Tempered Density Functional Theory of Electrons in Molecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932−2941.

(33) Stein, T.; Kronik, L.; Baer, R. Reliable Prediction of Charge Transfer Excitations in Molecular Complexes Using Time-Dependent Density Functional Theory. *J. Am. Chem. Soc.* **2009**, *131*, 2818−2820.

(34) Körzdörfer, T.; Brédas, J.-L. Organic Electronic Materials: Recent Advances in the DFT Description of the Ground and Excited States Using Tuned Range-Separated Hybrid Functionals. *Acc. Chem. Res.* **2014**, *47*, 3284−3291.

(35) Autschbach, J.; Srebro, M. Delocalization Error and "Functional Tuning" in Kohn−Sham Calculations of Molecular Properties. *Acc. Chem. Res.* **2014**, *47*, 2592−2602.

(36) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van Der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(37) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Many-Electron Self-Interaction Error in Approximate Density Functionals. *J. Chem. Phys.* **2006**, *125*, 201102.

(38) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792−794.

(39) Bajaj, A.; Janet, J. P.; Kulik, H. J. Communication: Recovering the Flat-Plane Condition in Electronic Structure Theory at Semi-Local DFT Cost. *J. Chem. Phys.* **2017**, *147*, 191101.

(40) Srebro, M.; Autschbach, J. Does a Molecule-Specific Density Functional Give an Accurate Electron Density? The Challenging Case of the CuCl Electric Field Gradient. *J. Phys. Chem. Lett.* **2012**, *3*, 576−581.

(41) Brumboiu, I. E.; Prokopiou, G.; Kronik, L.; Brena, B. Valence Electronic Structure of Cobalt Phthalocyanine from an Optimally Tuned Range-Separated Hybrid Functional. *J. Chem. Phys.* **2017**, *147*, 044301.

(42) Gani, T. Z. H.; Kulik, H. J. Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT +U. *J. Chem. Theory Comput.* **2016**, *12*, 5931.

(43) Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. Density Functional Theory Is Straying from the Path toward the Exact Functional. *Science* **2017**, *355*, 49−52.

(44) Kulik, H. J. Perspective: Treating Electron over-Delocalization with the DFT+U Method. *J. Chem. Phys.* **2015**, *142*, 240901.

(45) Zhao, Q.; Kulik, H. J. Where Does the Density Localize in the Solid State? Divergent Behavior for Hybrids and DFT+U. *J. Chem. Theory Comput.* **2018**, *14*, 670−683.

(46) Kim, M.-C.; Sim, E.; Burke, K. Understanding and Reducing Errors in Density Functional Calculations. *Phys. Rev. Lett.* **2013**, *111*, 073003.

(47) Kim, M.-C.; Park, H.; Son, S.; Sim, E.; Burke, K. Improved DFT Potential Energy Surfaces via Improved Densities. *J. Phys. Chem. Lett.* **2015**, *6*, 3802−3807.

(48) Duignan, T. J.; Autschbach, J. Impact of the Kohn−Sham Delocalization Error on the 4f Shell Localization and Population in Lanthanide Complexes. *J. Chem. Theory Comput.* **2016**, *12*, 3109−3121.

(49) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.

(50) Saitow, M.; Becker, U.; Riplinger, C.; Valeev, E. F.; Neese, F. A New Near-Linear Scaling, Efficient and Accurate, Open-Shell Domain-Based Local Pair Natural Orbital Coupled Cluster Singles and Doubles Theory. *J. Chem. Phys.* **2017**, *146*, 164105.

(51) Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760−1771.

(52) Sayfutyarova, E. R.; Sun, Q.; Chan, G. K.-L.; Knizia, G. Automated Construction of Molecular Active Spaces from Atomic Valence Orbitals. *J. Chem. Theory Comput.* **2017**, *13*, 4063−4078.

(53) Xiao, D.; Martini, L. A.; Snoeberger, R. C., III; Crabtree, R. H.; Batista, V. S. Inverse Design and Synthesis of Acac-Coumarin Anchors for Robust TiO_2 Sensitization. *J. Am. Chem. Soc.* **2011**, *133*, 9014−9022.

(54) Weymuth, T.; Reiher, M. Gradient-Driven Molecule Construction: An Inverse Approach Applied to the Design of Small-Molecule Fixating Catalysts. *Int. J. Quantum Chem.* **2014**, *114*, 838−850.

(55) Krausbeck, F.; Sobez, J.-G.; Reiher, M. Stabilization of Activated Fragments by Shell-Wise Construction of an Embedding Environment. *J. Comput. Chem.* **2017**, *38*, 1023−1038.

(56) 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.

(57) Gani, T. Z. H.; Kulik, H. J. Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane-to-Methanol Conversion by Fe(IV)═O. *ACS Catal.* **2018**, *8*, 975−986.

(58) 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.

(59) Keinan, S.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules with Optimal Properties Using the Linear Combination of Atomic Potentials Approach in an AM1 Semiempirical Framework. *J. Phys. Chem. A* **2007**, *111*, 176−181.

(60) Keinan, S.; Therien, M. J.; Beratan, D. N.; Yang, W. Molecular Design of Porphyrin-Based Nonlinear Optical Materials. *J. Phys. Chem. A* **2008**, *112*, 12203−12207.

(61) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228−3232.

(62) 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.

(63) 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.

(64) 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.

(65) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(66) Grajciar, L.; Heard, C. J.; Bondarenko, A. A.; Polynski, M. V.; Meeprasert, J.; Pidko, E. A.; Nachtigall, P. Towards Operando Computational Modeling in Heterogeneous Catalysis. *Chem. Soc. Rev.* **2018**, *47*, 8307−8348.

(67) Arockiam, P. B.; Bruneau, C.; Dixneuf, P. H. Ruthenium(II)-Catalyzed C-H Bond Activation and Functionalization. *Chem. Rev.* **2012**, *112*, 5879−5918.

(68) Prier, C. K.; Rankic, D. A.; MacMillan, D. W. C. Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chem. Rev.* **2013**, *113*, 5322−5363.

(69) Rouquet, G.; Chatani, N. Catalytic Functionalization of C(sp2)-H and C(sp3)-H Bonds by Using Bidentate Directing Groups. *Angew. Chem., Int. Ed.* **2013**, *52*, 11726−11743.

(70) Schultz, D. M.; Yoon, T. P. Solar Synthesis: Prospects in Visible Light Photocatalysis. *Science* **2014**, *343*, 1239176.

(71) Shaffer, D. W.; Bhowmick, I.; Rheingold, A. L.; Tsay, C.; Livesay, B. N.; Shores, M. P.; Yang, J. Y. Spin-State Diversity in a Series of Co(II) PNP Pincer Bromide Complexes. *Dalton Trans.* **2016**, *45*, 17910−17917.

(72) Tsay, C.; Yang, J. Y. Electrocatalytic Hydrogen Evolution under Acidic Aqueous Conditions and Mechanistic Studies of a Highly Stable Molecular Catalyst. *J. Am. Chem. Soc.* **2016**, *138*, 14174−14177.

(73) 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.

(74) Bowman, D. N.; Bondarev, A.; Mukherjee, S.; Jakubikova, E. Tuning the Electronic Structure of Fe(II) Polypyridines via Donor

Atom and Ligand Scaffold Modifications: A Computational Study. *Inorg. Chem.* **2015**, *54*, 8786−8793.

(75) Yella, A.; Lee, H. W.; Tsao, H. N.; Yi, C. Y.; Chandiran, A. K.; Nazeeruddin, M. K.; Diau, E. W. G.; Yeh, C. Y.; Zakeeruddin, S. M.; Gratzel, M. Porphyrin-Sensitized Solar Cells with Cobalt (II/III)-Based Redox Electrolyte Exceed 12% Efficiency. *Science* **2011**, *334*, 629−634.

(76) Czerwieniec, R.; Yu, J. B.; Yersin, H. Blue-Light Emission of Cu(I) Complexes and Singlet Harvesting. *Inorg. Chem.* **2011**, *50*, 8293−8301.

(77) Dias, F. B.; Bourdakos, K. N.; Jankus, V.; Moss, K. C.; Kamtekar, K. T.; Bhalla, V.; Santos, J.; Bryce, M. R.; Monkman, A. P. Triplet Harvesting with 100% Efficiency by Way of Thermally Activated Delayed Fluorescence in Charge Transfer OLED Emitters. *Adv. Mater.* **2013**, *25*, 3707−3714.

(78) Kuttipillai, P. S.; Zhao, Y. M.; Traverse, C. J.; Staples, R. J.; Levine, B. G.; Lunt, R. R. Phosphorescent Nanocluster Light-Emitting Diodes. *Adv. Mater.* **2016**, *28*, 320−326.

(79) Leitl, M. J.; Kuchle, F. R.; Mayer, H. A.; Wesemann, L.; Yersin, H. Brightly Blue and Green Emitting Cu(I) Dimers for Singlet Harvesting in Oleds. *J. Phys. Chem. A* **2013**, *117*, 11823−11836.

(80) Linfoot, C. L.; Leitl, M. J.; Richardson, P.; Rausch, A. F.; Chepelin, O.; White, F. J.; Yersin, H.; Robertson, N. Thermally Activated Delayed Fluorescence (TADF) and Enhancing Photoluminescence Quantum Yields of Cu-I(Diimine)(Diphosphine)(+) Complexes-Photophysical, Structural, and Computational Studies. *Inorg. Chem.* **2014**, *53*, 10854−10861.

(81) Zink, D. M.; Bachle, M.; Baumann, T.; Nieger, M.; Kuhn, M.; Wang, C.; Klopper, W.; Monkowius, U.; Hofbeck, T.; Yersin, H.; Brase, S. Synthesis, Structure, and Characterization of Dinuclear Copper(I) Halide Complexes with PAN Ligands Featuring Exciting Photoluminescence Properties. *Inorg. Chem.* **2013**, *52*, 2292−2305.

(82) Weininger, D. Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(83) 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.

(84) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(85) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(86) Brandenburg, J. G.; Grimme, S. Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *J. Phys. Chem. Lett.* **2014**, *5*, 1785−1789.

(87) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931−948.

(88) Korth, M.; Thiel, W. Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost as Accurate and Robust as DFT-GGA Methods for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7*, 2929−2936.

(89) Gallandi, L.; Marom, N.; Rinke, P.; Körzdörfer, T. Accurate Ionization Potentials and Electron Affinities of Acceptor Molecules II: Non-Empirically Tuned Long-Range Corrected Hybrid Functionals. *J. Chem. Theory Comput.* **2016**, *12*, 605−614.

(90) Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 1: Methodology. *Drug Discovery Today* **1997**, *2*, 457−467.

(91) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.

(92) Turner, J. V.; Glass, B. D.; Agatonovic-Kustrin, S. Prediction of Drug Bioavailability Based on Molecular Structure. *Anal. Chim. Acta* **2003**, *485*, 89−102.

(93) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(94) 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.

(95) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower Than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(96) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−96.

(97) 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.

(98) 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.

(99) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(100) 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.

(101) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(102) Landrum, G.*Rdkit: Open-Source Cheminformatics*, 2006.

(103) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the Openbabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2*, 5.

(104) 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.

(105) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. Uff, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(106) Minenkov, Y.; Sharapa, D. I.; Cavallo, L. Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *J. Chem. Theory Comput.* **2018**, *14*, 3428−3439.

(107) Husch, T.; Freitag, L.; Reiher, M. Calculation of Ligand Dissociation Energies in Large Transition-Metal Complexes. *J. Chem. Theory Comput.* **2018**, *14*, 2456−2468.

(108) Simm, G. N.; Reiher, M. Systematic Error Estimation for Chemical Reaction Energies. *J. Chem. Theory Comput.* **2016**, *12*, 2762−2773.

(109) Reiher, M.; Salomon, O.; Artur Hess, B. Reparameterization of Hybrid Functionals Based on Energy Differences of States of Different Multiplicity. *Theor. Chem. Acc.* **2001**, *107*, 48−55.

(110) 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.

(111) Fouqueau, A.; Casida, M. E.; Daku, L. M. L.; Hauser, A.; Neese, F. Comparison of Density Functionals for Energy and Structural Differences between the High-[5t2g:(T2g) 4 (Eg) 2] and Low-[1a1g:(T2g) 6 (Eg) 0] Spin States of Iron (II) Coordination Compounds. II. More Functionals and the Hexaminoferrous Cation, [Fe (NH3) 6] 2. *J. Chem. Phys.* **2005**, *122*, 044110.

(112) 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.

(113) Gani, T. Z. H.; Kulik, H. J. Unifying Exchange Sensitivity in Transition Metal Spin-State Ordering and Catalysis through Bond Valence Metrics. *J. Chem. Theory Comput.* **2017**, *13*, 5443−5457.

(114) Cramer, C. J.; Truhlar, D. G. Density Functional Theory for Transition Metals and Transition Metal Chemistry. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757−10816.

(115) 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.

(116) 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.

(117) Tortorella, S.; Marotta, G.; Cruciani, G.; De Angelis, F. Quantitative Structure-Property Relationship Modeling of Ruthenium Sensitizers for Solar Cells Applications: Novel Tools for Designing Promising Candidates. *RSC Adv.* **2015**, *5*, 23865−23873.

(118) Cruz, V. L.; Martinez, S.; Ramos, J.; Martinez-Salazar, J. 3D-QSAR as a Tool for Understanding and Improving Single-Site Polymerization Catalysts. A Review. *Organometallics* **2014**, *33*, 2944−2959.

(119) 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.

(120) 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.

(121) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.

(122) 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.

(123) Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137−5152.

(124) 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.

(125) 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.

(126) Liu, F.; Yang, T.; Yang, J.; Xu, E.; Bajaj, A.; Kulik, H. J., Bridging the Homogeneous−Heterogeneous Divide: Modeling Spin and Reactivity in Single Atom Catalysis. *Submitted.*

(127) Pritchard, B.; Autschbach, J. Theoretical Investigation of Paramagnetic NMR Shifts in Transition Metal Acetylacetonato Complexes: Analysis of Signs, Magnitudes, and the Role of the Covalency of Ligand−Metal Bonding. *Inorg. Chem.* **2012**, *51*, 8340−8351.

(128) 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.

(129) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(130) 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: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(131) 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.

(132) 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.

(133) Klamt, A.; Schuurmann, G. Cosmo: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, *2*, 799−805.

(134) Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J. Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.* **2015**, *11*, 3131−3144.

(135) Konezny, S. J.; Doherty, M. D.; Luca, O. R.; Crabtree, R. H.; Soloveichik, G. L.; Batista, V. S. Reduction of Systematic Uncertainty in DFT Redox Potentials of Transition-Metal Complexes. *J. Phys. Chem. C* **2012**, *116*, 6349−6356.

(136) Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. Calculation of One-Electron Redox Potentials Revisited. Is It Possible to Calculate Accurate Potentials with Density Functional Methods? *J. Phys. Chem. A* **2009**, *113*, 6745−6750.

(137) Baik, M.-H.; Friesner, R. A. Computing Redox Potentials in Solution: Density Functional Theory as a Tool for Rational Design of Redox Agents. *J. Phys. Chem. A* **2002**, *106*, 7407−7412.

(138) Janet, J. P.; Zhao, Q.; Ioannidis, E. I.; Kulik, H. J. Density Functional Theory for Modeling Large Molecular Adsorbate-Surface Interactions: A Mini-Review and Worked Example. *Mol. Simul.* **2017**, *43*, 327−345.

(139) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(140) Reiher, M. Theoretical Study of the Fe (Phen) 2 (NCS) 2 Spin-Crossover Complex with Reparametrized Density Functionals. *Inorg. Chem.* **2002**, *41*, 6928−6935.

(141) Fouqueau, A.; Mer, S.; Casida, M. E.; Lawson Daku, L. M.; Hauser, A.; Mineva, T.; Neese, F. Comparison of Density Functionals for Energy and Structural Differences between the High- [5t2g: (T2g)4(Eg)2] and Low- [1a1g: (T2g)6(Eg)0] Spin States of the Hexaquoferrous Cation [Fe(H2O)6]2. *J. Chem. Phys.* **2004**, *120*, 9473−9486.

(142) Mortensen, J. J.; Kaasbjerg, K.; Frederiksen, S. L.; Nørskov, J. K.; Sethna, J. P.; Jacobsen, K. W. Bayesian Error Estimation in Density-Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 216401.

(143) Proppe, J.; Reiher, M. Reliable Estimation of Prediction Uncertainty for Physicochemical Property Models. *J. Chem. Theory Comput.* **2017**, *13*, 3297−3317.

(144) Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14*, 5238−5248.

(145) Cailliez, F.; Pernot, P. Statistical Approaches to Forcefield Calibration and Prediction Uncertainty in Molecular Simulation. *J. Chem. Phys.* **2011**, *134*, 054124.

(146) Pernot, P.; Civalleri, B.; Presti, D.; Savin, A. Prediction Uncertainty of Density Functional Approximations for Properties of Crystals with Cubic Symmetry. *J. Phys. Chem. A* **2015**, *119*, 5288−5304.

(147) Pernot, P.; Savin, A. Probabilistic Performance Estimators for Computational Chemistry Methods: The Empirical Cumulative Distribution Function of Absolute Errors. *J. Chem. Phys.* **2018**, *148*, 241707.

(148) Weymuth, T.; Proppe, J.; Reiher, M. Statistical Analysis of Semiclassical Dispersion Corrections. *J. Chem. Theory Comput.* **2018**, *14*, 2480−2494.

(149) Bowman, D. N.; Jakubikova, E. Low-Spin Versus High-Spin Ground State in Pseudo-Octahedral Iron Complexes. *Inorg. Chem.* **2012**, *51*, 6011−6019.

(150) Salomon, O.; Reiher, M.; Hess, B. A. Assertion and Validation of the Performance of the B3LYP* Functional for the First Transition Metal Row and the G2 Test Set. *J. Chem. Phys.* **2002**, *117*, 4729−4737.

(151) Duignan, T.; Autschbach, J.; Batista, E.; Yang, P. Assessment of Tuned Range Separated Exchange Functionals for Spectroscopies and Properties of Uranium Complexes. *J. Chem. Theory Comput.* **2017**, *13*, 3614−3625.

(152) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(153) Wilbraham, L.; Verma, P.; Truhlar, D. G.; Gagliardi, L.; Ciofini, I. Multiconfiguration Pair-Density Functional Theory Predicts Spin-State Ordering in Iron Complexes with the Same Accuracy as Complete Active Space Second-Order Perturbation Theory at a Significantly Reduced Computational Cost. *J. Phys. Chem. Lett.* **2017**, *8*, 2026−2030.

(154) Mahler, A.; Janesko, B. G.; Moncho, S.; Brothers, E. N. When Hartree-Fock Exchange Admixture Lowers DFT-Predicted Barrier Heights: Natural Bond Orbital Analyses and Implications for Catalysis. *J. Chem. Phys.* **2018**, *148*, 244106.

(155) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Quantum Chemistry Structures and Properties of 134 Kilo Molecules. 2014, *1*, 140022.

(156) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(157) Herr, J. E.; Yao, K.; McIntyre, R.; Toth, D. W.; Parkhill, J. Metadynamics for Training Neural Network Model Chemistries: A Competitive Assessment. *J. Chem. Phys.* **2018**, *148*, 241710.

(158) 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.