MIT Libraries | DSpace@MIT

MIT Open Access Articles

Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Massachusetts Institute of Technology

Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design

Jon Paul Janet[1,#], Terry Z.H. Gani[1,#], Adam H. Steeves[1], Efthymios I. Ioannidis[1], and Heather J. Kulik[1,*]

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

[#]These authors contributed equally.

ABSTRACT: Virtual high throughput screening, typically driven by first-principles, density functional theory calculations, has emerged as a powerful tool for the discovery of new materials. Although the computational materials science community has benefited from open source tools for the rapid structure generation, calculation, and analysis of crystalline inorganic materials, software and strategies to address the unique challenges of inorganic complex discovery have not been as widely available. We present a unified view of our recent developments in the open source molSimplify code for inorganic discovery. Building on our previous efforts in the automated generation of highly accurate inorganic molecular structures, first-principles simulation, and property analysis to accelerate high-throughput screening, we have recently incorporated a neural network that both improves structure generation and predicts electronic properties prior to first-principles calculation. We also provide an overview of how multi-million molecule organic libraries can be leveraged for inorganic discovery alongside cheminformatics concepts of molecular diversity in order to efficiently traverse chemical space. We demonstrate all of these tools on the discovery of design rules for octahedral Fe(II/III) redox couples with nitrogen ligands. Over a search of only approximately 40 new molecules, we obtain redox potentials relative to the Fc/Fc$^+$ couple ranging from -1 to 4.5 V in aqueous solution. Our new automated correlation analysis reveals heteroatom identity and the degree of structural branching to be key ligand descriptors in determining redox potential. This inorganic discovery toolkit provides a promising approach to advancing transition metal complex design.

1

1. Introduction

Over the past few years, virtual high-throughput screening (VHTS) with first-principles simulation (e.g., density functional theory, DFT) has dramatically advanced the discovery of heterogeneous catalysts and materials.[1-8] VHTS discovery efforts have particularly excelled where i) the chemical space to be explored is well-defined and low-dimensional and ii) codes are available for the rapid generation and management of geometric structures and inputs to first-principles calculations[2, 9-10]. Examples of low-dimensional chemical space exploration have included: variation of binary alloys to discover new heterogeneous catalysts[1, 11], discovery of metal-organic frameworks[12-16] through enumeration based on a small set of building blocks[17], and definition of a narrow pool of organic components for evolutionary design of organic light emitting diodes[18-19].

Screening efforts within the solid-state community have particularly benefited from open-source tools[2, 9-10, 20] to automate computational materials design[21]. These codes leverage low-dimensional design spaces by retrieving and permuting crystal structures from freely available databases and automate the calculation of properties with DFT. Open-source tools for crystal structure prediction using optimization strategies have also been developed.[22-25] For heterogeneous catalysis in particular, the Atomic Simulation Environment (ASE)[26] enables the generation of periodic slabs with adsorbates and interfaces to a number of electronic structure codes.

In the context of molecular discovery, fewer tools for discovery and optimization have been developed for the broader community wishing to carry out high-throughput, first-principles-based screening. Instead, *ad hoc* screening approaches are typically built on top of cheminformatics-oriented toolkits such as Open Babel or RDKit and their Python wrappers[27-28], the Java-based chemical discovery kit (CDK)[29], or their commercial equivalents. Successful demonstrations of the value of VHTS have included the discovery of new electrolyte materials[30-31], organic light-emitting diodes[18-19], and photovoltaics[32-33]. Advanced molecular design strategies have included application of evolutionary algorithms to organic molecule optimization[18, 34-35] or optimization in the descriptor space of a surrogate (i.e., artificial neural network) model and

subsequent decoding back to molecular representations[19]. Each of these first-principles efforts has in common a reliance on cheminformatics-derived tools; these discovery workflows harness machine-readable molecular representations (e.g., SMILES strings[36]) to carry out efficient molecule generation and characterization. These structure generation tools work quite well for building organic molecules, but they are known to perform poorly for inorganic complex generation.[37]

The difficulties of inorganic structure generation have created a situation where, despite the critical problems that may be addressed by transition metal complexes, e.g. new spin-crossover complexes[38-40], dye-sensitizers in solar cells[41], or highly-reactive, open-shell catalysts[42], fewer strategies have been developed for discovery in inorganic chemistry and almost none are available as open-source codes. Rather, inorganic molecular design efforts are usually customized with in-house codes, as was the case for electrolyte complexation with Li in the electrolyte genome project[30]. Transition metal complex optimization has been pursued through the linear combination of atomic potentials (LCAP) method[43-45], introduced by Yang and Beratan to make the molecular optimization problem continuous. Challenges here for molecular design are in the mapping of the continuous representation back to a real molecule. Alternative approaches have been focused on constructing complexes by optimizing the number and position of charges[46] or in a shellwise manner[47] for iterative construction of transition metal complexes. As first-principles simulation with density functional theory can be computationally expensive for evolutionary algorithm-based optimization, Jensen and coworkers have demonstrated catalyst design using a heuristic structure-property relationship to define the fitness function.[48]

One method to overcome challenges associated with variable and difficult-to-define coordination environments in transition-metal chemistry is to draw upon existing knowledge in the form of proprietary databases of crystal structures (e.g., the Cambridge Structural Database[49]). Fragment-based strategies have enabled catalyst or transition metal complex structure building by generating a library from such databases and then using rules for the construction of transition metal complexes[50-52]. However, these tools rely on access to proprietary databases, proper curation

of the fragment library to enable structure generation, and they are not currently available as open-source tools for researchers to apply to their own molecular design problems. We will thus describe our recently-introduced[37] open-source tools for structure generation and discovery in inorganic chemistry in the following section.

2. Approach

Structure and input generation. We recently introduced molSimplify[37], an automated, open source toolkit for the first-principles screening and discovery of new inorganic molecules and intermolecular complexes. To aid structure building in inorganic chemistry, we developed a divide-and-conquer approach to the generation of inorganic complexes that separately treats metal-proximal and metal-distant components. Ligands obtained as SMILES strings[36] or in Cartesian coordinates are preoptimized with a class II force field (MMFF94)[53] that has been parameterized to yield highly accurate bond lengths and angles for organic molecules, as implemented in Open Babel[27-28]. A database of DFT-trained metal-ligand bond lengths is then used to set the coordination distance of organic components to a metal "core", with alignment also minimizing steric repulsion between adjacent ligands (Figure 1). The code supports a variety of coordination numbers and orientations as well as the possibility to distort structures from ideal coordination geometries. For an arbitrary complex, a database of discrete training values is used to provide an initial metal-ligand bond length. Each database entry has an associated series of keys, including ligand identity, metal identity, oxidation state, spin state, exchange-correlation functional, and coordination number. The code attempts to match the most relevant keys (i.e., metal, full ligand identity) to the complex being studied to the database in order to assign the trained metal-ligand bond length. If no minimum database match can be found (i.e., metal and connecting atom identity), then the sums of covalent radii of the metal and connecting ligand atom are used to assign bonds.
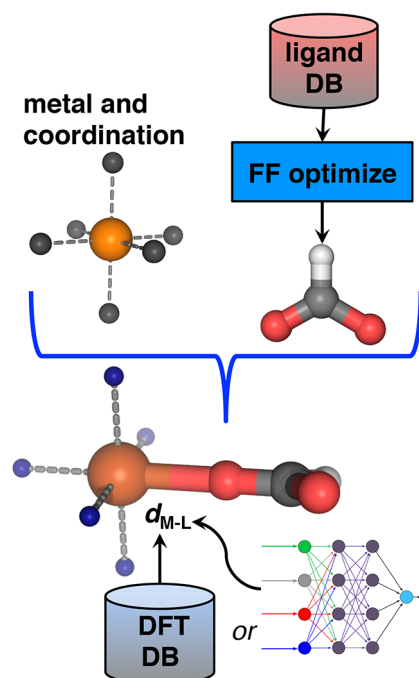
Figure 1. Schematic of molSimplify structure building: (top) the user selects a metal and coordination environment as well as ligands from a database that are force field preoptimized; (bottom) the ligand is aligned to the metal coordination site and the metal-ligand bond distance (d_{M-L}) is set from a database of values or from an ANN.

Over a test set of 150 molecules, we found[37] trained metal-ligand bond lengths and force-field preoptimization to improve structure generation by reducing initial gradients and relative energies of starting structures with respect to structures built using sums of covalent radii for the metal-ligand bond. These generated structures also substantially outperformed the Universal Force Field (UFF)[54], a force field developed for inorganic complexes, 75% of the time, as assessed by relative energies of the generated structures. UFF outperformed molSimplify[37] only for bidentate structures in which ligand rigidity prevented satisfying desired coordination geometries. Examining monodentate structures reveals molSimplify[37] is superior to UFF in 95% of 21 cases. An additional advantage of this structure generation approach is that it greatly simplifies VHTS efforts in inorganic chemistry by automating both the generation of coordination complexes as well as intermolecular complexes, e.g. for binding energies[55], or adsorbate placement on surfaces with our supercell builder[56]. At runtime, electronic structure input files are generated as well as a molSimplify input file that can exactly regenerate the

5

structure at some later point in time. The code comes with numerous built-in ligands in a prepopulated database, and the code interfaces to alternate databases, as we will discuss shortly.

Neural network predictions. Although generally superior to UFF, one shortcoming of our first-generation structure-building toolkit was the requirement that sufficient data be prepopulated in a DFT-trained metal-ligand bond length database. If a structure is not present in that database, then covalent radii are used to guess the initial metal-ligand bond length. Transition metal complexes often have multiple oxidation and spin states that further increase the specificity requirements for such a database. In order to overcome some of these limitations, we have recently incorporated an artificial neural network (ANN) for the prediction of transition metal complex properties[57] in molSimplify[37]. This ANN predicts a number of properties, including metal-ligand bond lengths for structure generation. The metal-ligand bond lengths are predicted to within 0.02-Å accuracy in a spin-state- and oxidation-state dependent manner for molecules well-represented by the training data. The initial training of the ANN was based on around 2700 structures[57], and we continue to update the ANN in the public release of molSimplify as more training data is acquired.

Feature selection for the neural network included metal identity, charge, and oxidation state as well as topology of the ligands (Figure 2). Specifically, we focus on the direct connecting atom of the ligand, electronegativity differences and bond order of atoms around the connecting atom, and the truncated Kier shape index[58] to measure topology of the ligand. Importantly, all of these properties are heuristics that are readily calculated by molSimplify and require no first-principles calculations. Regularized multiple linear regression (i.e., LASSO) was used to select these key features, and future refinement will focus on improving the balance between metal-focused and ligand-focused descriptors. The use of topological and chemical heuristics allows us to predict geometric properties for structure generation.
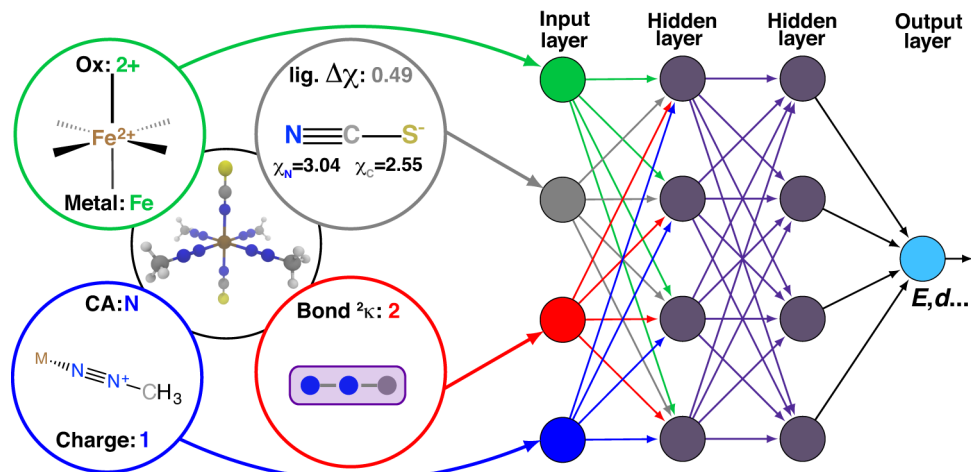
Figure 2. Schematic of descriptors that go into an ANN prediction of octahedral transition metal complex properties: metal identity and oxidation state, connecting atom identity, ligand charge, electronegativity differences of atoms surrounding the connecting atom, and the truncated Kier shape index, all depicted on a representative transition metal complex.

In order to provide a measure of reliability of the ANN bond length predictions, the code computes the Euclidean distance of the heuristic features of the generated molecule of interest to the original training data, where large values indicate low robustness of the ANN prediction. In addition to a new approach to provide credible intervals on ANN predictions[59], we have found this simple metric to be useful to identify when the ANN predictions are likely to be robust. Another key determinant of bond length is exchange-correlation functional choice, especially through variation of the percent of exact exchange[60]. The ANN predicts bond length variation with functional choice to further improve initial structure generation in VHTS of inorganic complexes. Importantly, the inorganic complex heuristics used in ANN training and prediction of structural properties are highly transferable to other quantities of interest, including prediction of results from first-principles simulation, as we will show shortly.

Predicting energetics and exchange-correlation sensitivity from structure. The importance of exchange-correlation functional sensitivity has been increasingly recognized[61-64] as a key challenge in screening efforts that rely on DFT. Energetic predictions are highly-sensitive to the choice of exchange-correlation functional, but the optimal choice is often system-dependent and seldom known for an arbitrary system of interest. Both non-self-consistent[64] functional families and variation of functional choice[61-62] have been incorporated into select

7

catalyst studies to identify how quantified uncertainty in DFT model choice propagates to predictions in catalysis. However, a non-self-consistent approach is insufficient when the density changes dramatically with change in functional choice[65], and uncertainty quantification in inorganic complexes[66] has only been applied to a model catalyst. We have taken a different approach to generalizing functional sensitivity by focusing on how chemical/structural properties of inorganic complexes correlate to underlying sensitivity to exchange correlation choice.

We have identified[60] a strong relationship between the nature of the direct connecting atom of a ligand to the metal and the degree to which exchange alters the spin-state ordering in inorganic complexes. More generally, we have found ligand field arguments to explain both exact exchange sensitivity as well as the effect of incorporating meta-GGA exchange[67] and the degree to which exchange-correlation functional tuning alters the electron density[65]. We have encapsulated[57] these observations in a trained ANN using the same descriptors as were used for bond length training (see Figure 2). The ANN predicts both absolute DFT-derived spin-state splitting at an arbitrary amount of HF exchange[57] and also predicts HF exchange sensitivity of spin-state splitting[60, 68-70]. Spin-state splittings for molecules well-represented by training set data are predicted with 3 kcal/mol average errors[57]. Larger errors in direct ANN prediction of molecules not well-represented in the training set motivates the use of the HF exchange sensitivity ANN model, e.g. in extrapolation from one fraction of exact exchange to another with on average 4 kcal/mol error over a 20% exchange variation interval.

ANN-predicted high functional sensitivity can in turn be verified straightforwardly with a handful of calculations, as electronic and energetic properties vary linearly with HF exchange mixing[60, 68-73]. High functional sensitivity can also be used to identify when beyond-DFT assessment is necessary. We are in the process of extending the ANN that predicts spin-state splitting and its exchange sensitivity to key catalytic properties of interest, including redox potential and catalytic barriers. Preliminary results on these properties have indicated good transferability of our curated descriptor set.

Repurposing cheminformatics tools for discovery and iterative design. We

have thus far introduced an approach to efficient structure generation for VHTS of inorganic complexes. When paired with a trained ANN, this approach also enables us to predict some energetic properties, such as the ground state spin of a molecule, without requiring a first-principles calculation. However, since inorganic chemical space is vast, a key challenge is in identifying which portions of inorganic chemical space should be explored to maximize new knowledge. Evolutionary algorithms (EA) provide one strategy[74] to molecular property optimization and have occasionally been employed for the design of inorganic complexes[48] when paired with a heuristic model for the fitness function. However, even when an EA is employed, a critical question is how to select the initial pool of ligand properties, as EAs generally solve a local but not global optimization problem.

We have taken a complementary approach to screening, where we focus on maximizing the information from initial screens in a strategy that is tailored to inorganic complex building. We hold fixed or select few variations for the chemistry directly proximal to the metal, such as the direct connecting ligand atom and its nearest neighbors as well as the coordination environment[55, 75]. We address the question of how to vary the rest of the ligand by leveraging freely available multi-million molecule libraries of organic compounds[76]. These libraries are most commonly employed to identify synthetically accessible and biologically active molecules, e.g. for therapeutic drug design. However, by repurposing organic ligand libraries in inorganic chemistry, we both increase the likelihood of computationally designed compounds being synthetically accessible and the chances of discovering new compounds that have not been used before. In parallel with our computational work[75], recent experimental efforts have shown the value of this approach in catalyst discovery as well[77].

For our computational design strategy, we search a database for compounds that have a substructure (i.e., SMARTS pattern) match to the desired direct ligand connecting atom and its environment, while we also filter for ligand size, charge, and elemental composition (Figure 3). The general database search and interface to large-molecule databases is implemented in Open Babel[27-28], and we use these tools along with additional filtering steps implemented in molSimplify

for inorganic ligand screening. We then also maximize diversity[78-79] in the resultant search using a molecular fingerprint[80] of each ligand result to ensure broad search of chemical space. Molecular similarity metrics and diversity are well-defined on the ligands themselves but would be ambiguous for the whole-complex, thus our divide and conquer approach again enables us to leverage cheminformatics strategies where they are already successful but impart knowledge of inorganic chemistry where necessary. All of the described database search and filtering approaches described are implemented in molSimplify.
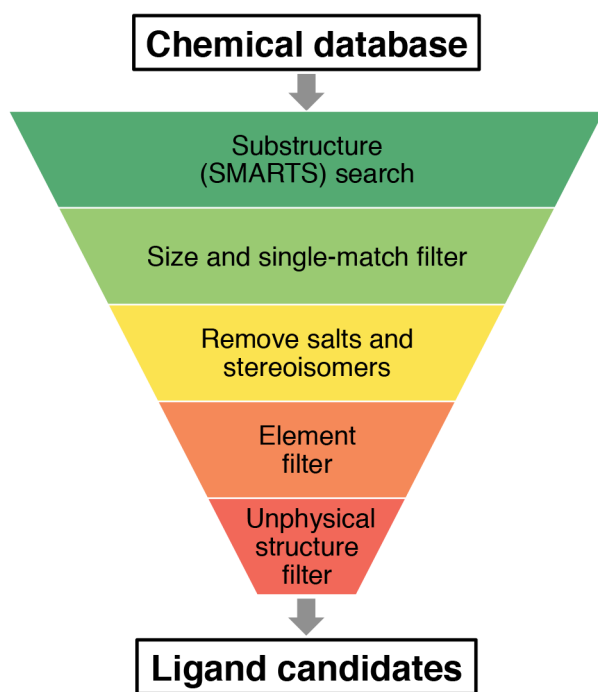
Figure 3. Schematic of order of ligand candidate retrieval from large chemical databases using filtering steps implemented in molSimplify.

In one example, we employed this design strategy for quantum dot precursors, reducing the 110,000 carboxylate-containing molecules present in the database to a few hundred ligands of reasonable size for rapid DFT calculation (i.e., less than 30 atoms per ligand) and maximizing diversity. Computational evaluation of the resulting compounds, with initial structures provided by molSimplify, revealed design rules for quantum dot precursors that extended beyond those molecules in the initial screen[75]. Strategies for continued screening and optimization are also essential. We have taken an approach that is inspired by therapeutic drug design to identify a

spatial and chemical representation of critical characteristics of an inorganic complex. We recently identified[55] optimal characteristics of ferrocenium-derived polymer building blocks that have been demonstrated to be a promising strategy for electrochemical ion separation[61]. An initial broad screen revealed optimal placement of hydrogen bonding groups and their electronegativity, which we transformed into a 3D representation we call the materiaphore[55]. By screening new functionalizations that match this 3D representation, we identified new complexes with significantly improved characteristics over the initial screen. Such methods could be used to augment traditional fitness functions in EAs, e.g. through multi-objective optimization[81-82].

Taken together, automated structure generation, exchange-correlation sensitivity, surrogate prediction of DFT properties, and database-derived compound generation with diversity as a driver for efficient discovery provide a new direction to help advance VHTS as a tool for inorganic complex design. Although challenges remain, these open source tools are already beginning to change how our group discovers new design rules[55, 75] in inorganic chemistry. The rest of this paper is as follows. In section 3, we provide the Computational Details of the demonstration of our inorganic discovery tools for Fe(II/III) redox couples. In section 4, we provide the Results and Discussion of our screening approach, including details on the application and performance of our strategies as well as demonstration of new tools for automated discovery of correlations in data sets. Finally, in section 5 we provide our conclusions and outlook on computational screening for inorganic complexes.

3. Computational Details

Initial electronic structure calculations, including single point energies and geometry optimizations, were carried out using the TeraChem[83-84] graphical processing unit (GPU)-accelerated quantum chemistry package with the B3LYP[85-87] hybrid exchange-correlation functional. The default definition of B3LYP in TeraChem employs the VWN1-RPA form for the LDA VWN[88] component of LYP[85] correlation. Iron, bromine, and iodine were treated with the LANL2DZ effective core potential[89], and the 6-31G* basis was used for the remaining atoms. Diffuse functions were included for select representative molecules and found to not alter

relative redox potentials, the focus of this work (Supporting Information Table S1). Dispersion corrections, which are neglected here but potentially important for studying physisorbed intermediates in catalysis applications of our toolkit, are not incorporated due to their minimal effect on computed redox potentials (Supporting Information Table S1). All structures were simulated in low-spin and high-spin states for Fe(II) and Fe(III), which corresponds to a singlet and doublet or quintet and sextet, respectively. All calculations were spin-unrestricted with virtual and open-shell orbitals level-shifted[90] by 1.0 eV and 0.1 eV, respectively, to aid self-consistent field (SCF) convergence to an unrestricted solution. The aqueous solvent environment was modeled using an implicit polarizable continuum model (PCM) with the conductor-like solvation model (COSMO[91-92]) and ε=78.39. The solute cavity was built using Bondi's van der Waals radii[93] scaled by a factor of 1.2 for available elements and 2.05 Å for iron.

Geometry optimizations in the gas phase of molSimplify-generated[37] structures were carried out using the L-BFGS algorithm in translation and rotation internal coordinates[94] as implemented in a development version of TeraChem[83-84] to the default tolerances of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in SCF energy between steps.

Redox potentials on Fe(II)/Fe(III) redox couples were computed using a standard protocol. The absolute redox potential (E^0) is related to the Gibbs energy of reaction in solution(ΔG_{solv}):

$$E^0 = -\frac{\Delta G_{solv}}{nF} \tag{1}$$

where n is the number of electrons transferred in the redox reaction (i.e., 1 for the Fe(II)/Fe(III) redox pairs studied in this work) and F is Faraday's constant. Following common practices[95-97], we employ a thermodynamic cycle to obtain ΔG_{solv} (Figure 4). In this approach, thermodynamic (i.e., vibrational enthalpy and zero point vibrational energy) corrections at 300 K are obtained on each of the gas phase structures. These differences are paired with the differences in the gas-phase-

optimized geometry solvation free energies to obtain the overall difference in solvated Fe(II) and Fe(III) complexes:

$$\Delta G_{solv} = G_{gas}(Fe(III)) - G_{gas}(Fe(II)) + \Delta G_s(Fe(III)) - \Delta G_s(Fe(II)) \qquad (2)$$

where G_{gas} is the gas phase energy with thermodynamic corrections and ΔG_s is the solvation free energy of the gas phase structure. Redox potentials relative to a reference redox couple (RC)[95-96] further minimize systematic errors in approximate DFT. Here, we select the ferrocene/ferrocenium (Fc/Fc·) RC as our reference, as has previously been suggested[95-96], because it undergoes the same transition metal redox reaction (i.e., Fe(II) to Fe(III)) as the species of interest. The calculated relative redox potential (E_{rel}^0) is then:

$$E_{rel}^0 = E^0 - E^0(RC) = \frac{\Delta G_{solv}(RC) - \Delta G_{solv}}{nF} \qquad (3)$$

where ΔG_{solv} in both cases is computed from the thermodynamic cycle in Figure 4.
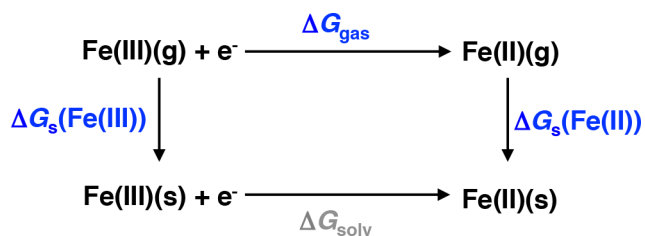
Figure 4. Thermodynamic cycle approach to calculation of redox potentials. The relationship between directly calculated quantities (shown in blue) and indirectly calculated quantities (shown in gray) is indicated by the arrows.

4. Results and Discussion

4a. Ligand Curation

We begin our Fe(II/III) redox couple design by defining the features of the ligand environment to be held constant in our screen: here, octahedral coordination with nitrogen ligand connecting atoms. This coordination environment was selected due to the prevalence of Fe(II/III) nitrogen octahedral complexes in previous theoretical[95-96, 98] and experimental[99-100] redox studies and

the coordination-sensitivity of ground state spin in these complexes, which should give rise to differing redox properties. Homoleptic complexes are built either from monodentate or bidentate ligands through separate database queries. For monodentate structures, we focus on substructure matching direct-connecting N atoms likely to form dative bonds through undercoordination (i.e., a nitrogen atom that is connected to at most 1 or 2 other species) that do not carry a net positive charge (Figure 5). Hydrogen atoms are kept attached to the nitrogen atoms in these cases, except for aromatic and amide nitrogens due to difficulties with SMARTS-based bond order assignment. The code searches for these special cases and removes hydrogen atoms where necessary to ensure one dative bond from the nitrogen to the metal. This non-restrictive query would return much of the database, nearly 1.3M of 1.9M molecules in ChEMBL-21[76], but for the purpose of illustrating the database search we make further restrictions (see cartoon in Figure 3). First, we require a single pattern match, which eliminates the presence of other undercoordinated nitrogen groups that would make choosing the connecting atom for coordination ambiguous. We also limited the number of atoms in each ligand to 14, a strategy we previously employed successfully[75] thanks to the fact that inorganic complex properties are most sensitive to the choice of immediate environment of the metal. The size and single-match restrictions have the greatest effect on our ligand pool, bringing 1.3M candidate nitrogen monodentate ligands down to 563. Further filtering steps carried out by molSimplify's database search module include removing counterions, stereoisomers or duplicates (527 matches) and filtering to remove undesired elements. Here, we allow organic (H, C, N, O) and halide (F, Cl, Br, I) elements, which may be applied in the search either by explicit specification of each element allowed in the filter or by keyword, reducing the total number of matches to 355.
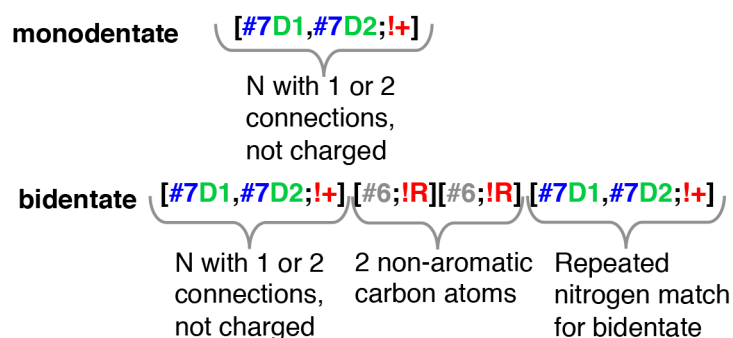
Figure 5. Representative SMARTS pattern matching for nitrogen-containing monodentate (top) or bidentate (bottom) ligands used in database search.

A final concern beyond substructure matching, size, charge, and elemental composition in choosing inorganic complexes from a database of individual organic fragments is the matter of feasibility of complex construction. That is, even if individual organic ligands match the above criteria, interligand distances may still be too low, even after molSimplify routines reduce interligand steric repulsion. To avoid this scenario, we have developed an unphysical complex detector that excludes a complex if any atom pairwise distances in the complex are less than 70% of the sum of covalent radii. The 70% value was determined by trial and error and leads to elimination of 44 compounds in the monodentate case to produce 311 candidate structures. This filter is necessary to avoid structures that would either be challenging for first-principles simulation or would not preserve the ligand structure we intended to simulate: 70% of the sums of covalent radii of two adjacent, non-bonding carbon atoms is 0.94 Å, which is 0.6 Å shorter than a typical C-C single bond. Although it is feasible to screen all of these complexes, we maximize efficiency in computational screening by further reducing this set to 10 molecules through a greedy algorithm to maximize fingerprint-based[27] molecular diversity (i.e., their Tanimoto[101] distance) among the matches (select structures shown in Figure 6, full list in Supporting Information Figure S1). In this approach, we set a target number of molecules (i.e., 10) and repeatedly compared Tanimoto distances[101], selecting the most dissimilar structures at each step until the target number of molecules is obtained. Representative monodentate structures range from bulkier groups in which the chelating nitrogen is part of a ring (M1-M2) or the nitrogen has varying bond order to its other substituents (M3-M5).
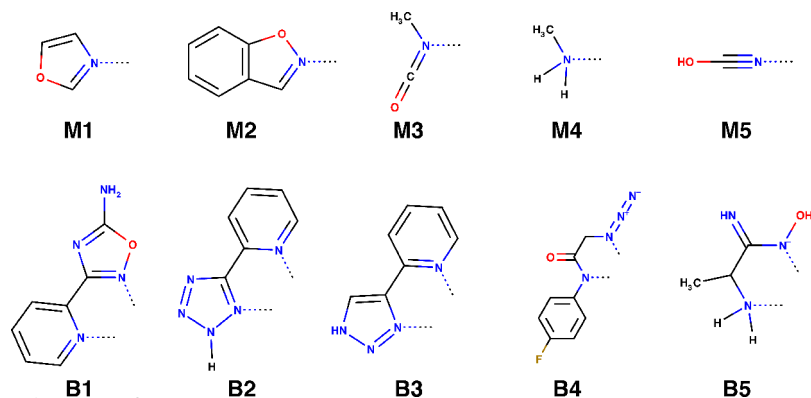
Figure 6. Representative monodentate (top) and bidentate (ligands) from database search with metal-coordinating sites indicated with dashed lines. Carbon and hydrogen atoms are shown in black, nitrogen atoms in blue, oxygen or hydroxyl functional groups in red, and a fluoride is shown in brown.

We repeat the same procedure on bidentate nitrogen ligands using two related filters (one shown in Figure 5). In the first, we search for undercoordinated nitrogen atoms connected by two aliphatic carbons. The focus on aliphatic carbons increases the likelihood that the bond is rotatable and allows for a cis orientation of the two chelating nitrogen atoms (e.g., ethylenediamine). In future work, a more general search could encompass nitrogen atoms separated by a variable number of carbon atoms. A second bidentate search aims to identify aromatic compounds with at least one nitrogen participating in a six-membered ring, common to many[77] polycyclic chelating ligands, such as bipyridine and phenanthroline. The same procedure for database filtering in the monodentate case is then applied here. A large pool of nearly 90,000 compounds is reduced by single-match and size-based (here, 22 atoms) filters to only 189 candidates. Further filtering of duplicates reduces this number to 158, and limiting elements to the organohalide set produces 115 total molecules. In order to ensure the correct bidentate orientation is generated, we have implemented a rotatable bond search to identify optimal coordinating conformers for bidentate ligands. Using both this approach and the unphysical structure filter, a final set of 80 bidentate homoleptic complexes are generated. These structures are then reduced to a 31-molecule subset after maximizing diversity (representative structures in Figure 6, full list in Supporting Information Figure S2 and summary of filtering in Supporting Information Table S2). Representative examples include both nitrogen on aromatic rings with

various heteroatom substitutions (B1-B3) as well as cases where the aliphatic carbon substitution varies alongside functional groups on the nitrogen atoms themselves (B4-B5).

4b. Structure Building and Pre-screening with an ANN

In order to accelerate VHTS, it is essential to generate high-quality structures. As mentioned in Sec. 2, our molSimplify toolkit[37] initially generated high-quality starting structures by matching metal-ligand bond lengths from a DFT-trained database (DB). These structures were shown to be superior to the best closest alternative for high-throughput studies, optimization with the UFF force field[54], particularly for monodentate structures. To compare the utility of these approaches to our recently introduced ANN[57], we generated starting structures across a set of representative bidentate and monodentate (five each) octahedral complexes in each oxidation state (Fe^{2+} or Fe^{3+}) and both high- and low-spin states (see Figure 6). UFF parameters are only available for select coordination environments and oxidation states for each metal, and this information must be fed to an optimization program in order for UFF structure generation to be successful. Thus, UFF optimizations will predict bond lengths for fixed spin and oxidation state only. Some effort has been made to expand supported coordination environments in UFF to address common motifs in solid-state metal-organic frameworks based on new semi-local DFT training data[102]. However, two key factors that determine bond length are the i) strong spin-state dependence, which could only be improved through incorporation of an advanced but not widely-available or parameterized ligand field molecular mechanics model[103] and ii) functional dependence that is only present in a force field by virtue of a specific training recipe. In theory, our metal-ligand DB supports variable spin and oxidation state, but that relies on a database entry that is a close match for each spin. In the case of nitrogen complexes, our metal-ligand DB contained only high-spin data.

The shortcomings of the two original approaches (UFF and DB) become apparent when comparing average unsigned errors in bond lengths versus optimized DFT geometries for UFF, the DB, and ANN (Figure 7 and Supporting Information Table S3). As in previous work[37], the UFF is observed to produce small errors for bidentate structures, particularly in the low-spin

state where errors are lower than those with the DB. In all other categories, the metal-ligand DB outperforms the UFF approach. Finally, the ANN produces the smallest errors of all 3 methods, particularly for low-spin bidentate structures (Figure 7). The improvements of the ANN over the DB are likely due to successful generalization of ANN training data coming from bidentate structures (e.g., bipyridine), and incorporated spin- and oxidation-state dependence. Some errors as high as 0.08 Å in the HS monodentate structures by the ANN are larger than the 0.02-0.03 Å average we observed previously during ANN training, suggesting future improvement of the ANN through descriptor refinement or enriching training data.
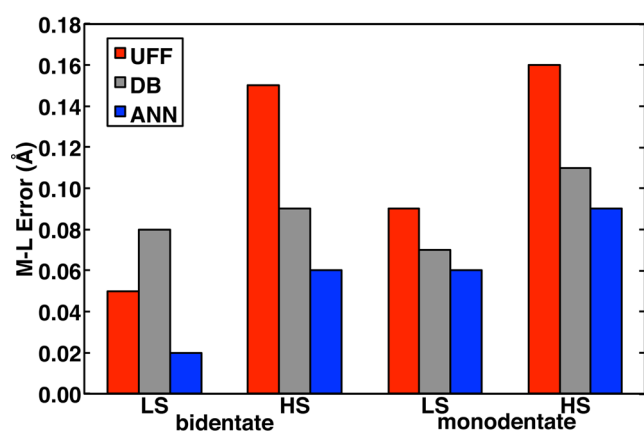
Figure 7. Unsigned average error in bond length prediction (in Å) compared against a DFT B3LYP optimized structure over representative bidentate (5) and representative monodentate (5) octahedral Fe(II) and Fe(III) low-spin and high-spin complexes from the UFF (red bars), metal-ligand database (grey bars), and an ANN (blue bars).

A second question in screening open-shell transition metal complexes is the qualitative (e.g., ground state) and quantitative spin-state ordering. For redox processes, knowledge of the ground state spin of the reduced system is essential to determine the spin state from which the oxidation process occurs. Since spin-state ordering is highly-sensitive to fraction of exact exchange incorporated in the functional[60, 68-69, 104-105], we demonstrate our ANN trained to predict exchange-sensitivity of spin-state ordering on a select group of representative bidentate and monodentate complexes (see Figure 6). The high-spin/low-spin, gas phase, adiabatic splitting (ΔE_{H-L}) of these bidentate and monodentate structures for 20% HF exchange (black circles in Figure 8) varies primarily by oxidation state: four out of five bidentate Fe(II) complexes and all

five Fe(II) monodentate complexes have high-spin ground states (Figure 8 and Supporting Information Table S4). Conversely, Fe(III) complexes are predominantly low spin (four out of five bidentate and three out of five monodentate) at 20% HF exchange. Incorporation of the ANN exchange sensitivity extrapolated in either direction from the 20% value provides a range of possible spin-state splittings: four of the bidentate cases and six of the monodentate cases are projected to change ground state spin between 0% and 20% exchange (Figure 8). Comparison of explicit calculation of 0% HF exchange spin-state ordering to the ANN-projected value demonstrates good correspondence in all cases except for M2 in either oxidation state. For M2, the ANN model sufficiently overestimates exchange sensitivity to predict that 0% exchange results should produce a low-spin Fe(III) complex where the calculated ground state is instead high-spin.
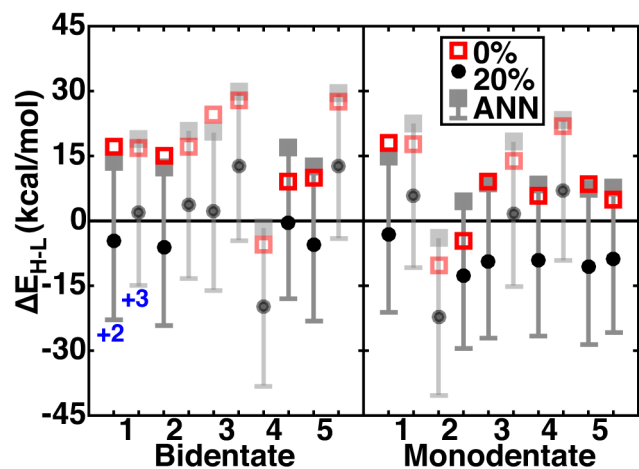
Figure 8. Spin-state splitting of five representative bidentate (left) and monodentate (right) at 20% HF exchange (black filled circles) and 0% HF exchange (red hollow squares) compared to 0 (filled grey square) to 40% (horizontal line) splitting ranges extrapolated from ANN predicted sensitivities (gray line). Each structure is grouped into alternating Fe(II) and Fe(III) results, as labeled on the first bidentate structure as +2 or +3 in blue labels. Cases for which the ground state spin changes from low-spin at 0% to high-spin at 20% HF exchange are highlighted with saturated colors.

Nevertheless, in 90% of the representative cases, the ANN provides a very good guide on how sensitive gas phase spin-state splitting is to HF exchange. Overall mean absolute errors are around 3.5 kcal/mol for the set, and our previously introduced distance-to-training-data heuristic is large for the cases where errors are above 5 kcal/mol (Supporting Information Figure S3 and

S4). Notably, our ANN predicts a narrow range of exchange sensitivities from -80 to -91 kcal/mol from 0 to 100% HF exchange due to the fact that the metal, coordination number, and direct ligand atom identity, the main determinants of sensitivity, are all unchanged across the representative set. Lower or higher exchange sensitivities are well predicted on representative molecules of more variable composition (Supporting Information Table S5). Given observations that optimal exchange mixing for iron-nitrogen complexes is around 20-30%[67], we use the spin state assignment at 20% exchange after inclusion of solvent and entropic effects in this work (ca. 3-4 kcal/mol for both effects, see Supporting Information Figure S5 and Table S6). However, we note that for cases where the spin-state splitting is close to zero and sensitive to HF exchange (i.e., several of the bidentate cases in this work), it is likely beneficial to compute redox properties from both spin states (see sec. 4c). Our conclusion from this analysis is that the ANN exchange sensitivity tool can provide efficient and valuable insight into how sensitive energetic predictions are to one of the most commonly adjusted DFT functional parameters.

4c. Redox Potential Tuning Trends

We now compare redox potentials of Fe(II/III) complexes relative to the Fc/Fc+ couple, as evaluated from the ground state spin of each reduced molecule to the closest Fe(III) complex with an electron configuration that differs by single electron removal (Supporting Information Tables S6-S8). The ground state spin for the reduced species is chosen after incorporation of modest combined solvation, zero point energy, and entropic contributions to spin-state ordering (ca. 3-4 kcal/mol) that preserve the gas phase spin-state ordering for all but two cases (Supporting Information Figure S5 and Table S9). Over the full data set, a large range of redox potentials is observed from slightly below the ferrocenium couple (around -0.8 V) to substantially higher at 4.5 V (Figure 9). Bidentate complexes generally span the lower redox potentials (up to 2 V), whereas the monodentate species are all at the higher range (1 to 4.5 V). The large range sampled is partly enabled by our diversity-oriented screen, which ensures that chemically distinct ligands are sampled (see examples in Figure 6).
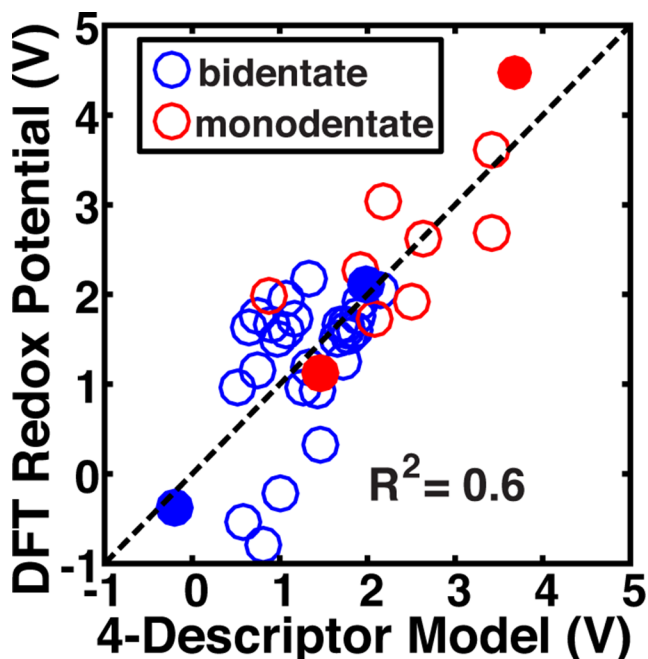
Figure 9. Correlation of redox potential (in V) relative to Fc/Fc⁺ reference vs. four descriptor model with parity line shown (dashed black line), as described in the main text for bidentate (blue circles) and monodentate (red circles) structures. Select high and low redox potential examples described in the main text are indicated in filled circles, and the remainder are indicated as hollow circles.

Having surveyed a diverse set of ligands, we wish to extract key molecular descriptors that are well-correlated (e.g., through multiple-linear regression) to redox potential in order to reveal potential design rules. In order to avoid overfitting, we implemented in molSimplify[37] a recursive feature elimination (RFE)[106] approach to automated correlation discovery. Here, we start with a large space of possible heuristic descriptors derived from molecular topology/size, elemental identity, and electronegativity, to be described next, and remove the worst performing ones until an optimal number of descriptors is reached. This greedy fashion of feature removal in RFE is not guaranteed to find the best subset, but we have chosen this approach for its ability to provide a ranked list of features. Alternative methods for feature selection including LASSO[107] regression or methods based on random forests[108] will be pursued in future work. We use the leave-one-out cross-validation (LOOCV) error[106] on withheld data points as a measure of model robustness: descriptors are automatically removed from the feature set until the square of the LOOCV error, as evaluated with sklearn[109], reaches a minimum. Out-of-sample error could also

be estimated by bootstrapping[110-111] but there is evidence[112-113] that CV is sufficient for feature selection.

In combination with this approach, we define a descriptor set from which molSimplify selects the most significant correlations to redox potential (Supporting Information Table S10). Here, we employ heuristic, chemical/topological descriptors of the ligands, in an extension of our prior ANN training[57], in order to identify quantities that can be employed in subsequent screens of large molecule databases without direct first-principles simulation. We employ autocorrelation functions[114-115] (AC), which are a pairwise sum of a products of given key property over all atoms that are a fixed number of bonds, d, away from each other:

$$\text{AC}_p(d) = \Sigma_{i,j} p_i p_j \delta\left(d_{i,j} - d\right) \tag{4}$$

In this work, we allow d to vary from 0 up to 5 for autocorrelations of four properties, p: 1) Pauling electronegativity, χ; 2) nuclear charge, Z; 3) topology, T, as indicated through the atom's coordination number; and 4) the identity, I, that is 1 for all atoms, as suggested in Ref. [115]. Several of these autocorrelation functions correspond to recognizable descriptors, e.g., I_0 is the number of atoms, and T_1 is closely related to the Randić index[116] used in cheminformatics. A complete list of all 26 autocorrelation descriptors after normalization and centering to have a zero mean and unit variance is provided in Supporting Information Table S11.

The single best descriptor set, as judged by minimum LOOCV error, is revealed to be a combination of topological and nuclear charge-dependent properties: atomic T_0 and nearest neighbor Z_1, I_1, and T_1 (see Supporting Information Text S1, Table S12, and Figures S6-S7). These descriptors exhibit a good correlation ($R^2 = 0.6$, see Figure 9) with balanced mean squared error on the full data set and on the withheld points (Supporting Information Text S1). In order to interpret these descriptors further, we lump together the structural terms weighted and normalized by their coefficients, S, into the final expression for the predicted relative redox potential:

$$E^0_{\text{rel,pred}} = 3.62 Z_1 - 4.10 S + 1.51 \tag{5}$$

This expression reveals a balance between increased redox potential with increasing weight of substituents (large Z_i) and decreased redox potential for branched (large, positive S) structures (Supporting Information Figure S8). Quantitative analysis of the distribution of the Z_i and S descriptors reveals that redox potential is maximized for ligands with large positive Z_i and negative (i.e., relatively low) S best encapsulated by evaluating Z_i-S (Supporting Information Figure S9). For instance, reviewing representative high redox potential structures reveals a monodentate structure with N_2O ligands, which has a low Z_i due to the small ligand size but even lower S due to the higher bond order in the ligands (structures in Figure 10 and filled circles in Figure 9). In comparison, the low redox potential monodentate structure contains sp^3 nitrogen ligands with greater branching (Figure 10). Similar analysis captures the trend of higher redox potential bidentate structure with rigid, aromatic nitrogen-containing ligands compared to the low redox example with both chelating sp^3 nitrogen atoms and greater branching (Figure 10). Thus, these descriptors represent useful metrics that can be applied in future screens to tailor redox potentials of iron complexes.
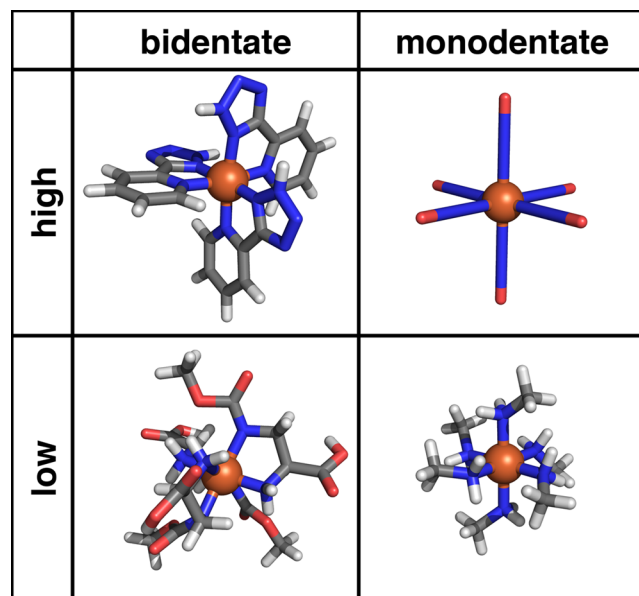
Figure 10. Representative high (top) and low (bottom) relative redox potential cases for bidentate (left) and monodentate (right) transition metal complexes, which were indicated as filled circles in Fig. 8. Iron is shown as a brown sphere, and the remaining atoms (blue nitrogen, gray carbon, red oxygen, and white hydrogen) are shown as sticks.

5. Conclusions

We have introduced and extended our discovery approach for inorganic complexes.

Based on an open source toolkit, molSimplify, our computational strategy focuses on a divide-and-conquer approach to inorganic complex structure building and screening. Here, we leverage tools that work in organic chemistry, such as the transformation of simple strings to 3D coordinates and concepts of molecular similarity and diversity, for the organic components of transition metal complexes. We harness knowledge we have developed from direct simulation and established rules in inorganic chemistry, as encapsulated in our databases and machine learning models, to define aspects of the metal-ligand coordination environment. Building on our initial efforts to automate generation of highly accurate inorganic molecular structures, first-principles simulation, and property analysis to accelerate high-throughput screening, we have recently incorporated an ANN trained only on heuristic descriptors that both improves structure generation and predicts electronic properties prior to or in lieu of first-principles simulation. We demonstrated the value of some of these machine learning tools in improved bond length prediction and exchange-sensitivity of spin state ordering on representative complexes.

We provided an overview of how multi-million molecule organic libraries can be leveraged for inorganic discovery alongside cheminformatics concepts of molecular diversity in order to efficiently traverse chemical space. We demonstrated all of these tools on the discovery of design rules for Fe(II/III) redox couples with octahedral coordinating nitrogen ligands. Over a search of only approximately 40 new molecules, we obtained a variation of redox potential relative to the Fc/Fc⁺ couple from around -1 to 4.5 V in aqueous solution. Our new automated correlation analysis reveals heteroatom identity and the degree of structural branching to be key ligand descriptors in determining redox potential. The use of leave one out cross validation in our approach avoids overfitting and helps identify the four most relevant parameters to redox potential prediction from an initial set of 28 ligand descriptors.

This inorganic discovery toolkit and suite of strategies offers a promising approach to advancing inorganic complex design and provides a much needed complement to existing tools for high-throughput screening of solid state materials. Ongoing efforts in our lab are focused on extending these tools for high-throughput initial guesses for transition states relevant for catalysis

as well as the development of general purpose frameworks for both local and global property optimization.

ASSOCIATED CONTENT

Supporting Information. Optimized structures of all Fe(II/III) complexes in high-spin and low-spin states; chemdraw structures of ligands for monodentate and bidentate octahedral complexes; results of database filtering; results with empirical dispersion and diffuse functions; bond length prediction details; spin-state splitting HFX sensitivity predicted with ANN and explicit calculation as well as analysis of errors; gas phase versus solvated free energy spin-state splittings; extended ANN exchange sensitivity predictions; ferrocenium redox potential reference calculations; relative redox potentials and components; spin-state specific redox potential details for near-degenerate Fe(II) complexes; scaling and ranking for descriptors and a full list of descriptor values for each complex; extended discussion of multiple linear regression model; predicted redox potentials from optimal descriptors; LOOCV and MSE/correlation coefficient versus number of features; analysis of S vs. Z descriptors with redox. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

References

1. Nørskov, J. K.; Bligaard, T., The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776-777.
2. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A., Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.
3. Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nørskov, J. K., Computational High-Throughput Screening of Electrocatalytic Materials for Hydrogen Evolution. *Nat. Mater.* **2006**, *5*, 909-913.
4. Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H., Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37-46.
5. Jensen, P. B.; Bialy, A.; Blanchard, D.; Lysgaard, S.; Reumert, A. K.; Quaade, U. J.; Vegge, T., Accelerated DFT-Based Design of Materials for Ammonia Storage. *Chem. Mater.* **2015**, *27*, 4552-4561.
6. Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G., Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chem. Mater.* **2010**, *22*, 3762-3767.
7. Jain, A.; Hautier, G.; Moore, C. J.; Ong, S. P.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G., A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295-2310.
8. Hautier, G.; Miglio, A.; Ceder, G.; Rignanese, G.-M.; Gonze, X., Identification and Design Principles of Low Hole Effective Mass P-Type Transparent Conducting Oxides. *Nat. Commun.* **2013**, *4*, 2292.
9. Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O., Aflow: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218-226.
10. Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B., Aiida: Automated Interactive Infrastructure and Database for Computational Science. *Comput. Mater. Sci.* **2016**, *111*, 218-230.
11. Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Norskov, J. K., Toward Computational Screening in Heterogeneous Catalysis: Pareto-Optimal Methanation Catalysts. *J. Catal.* **2006**, *239*, 501-506.
12. Watanabe, T.; Sholl, D. S., Accelerating Applications of Metal–Organic Frameworks for Gas Adsorption and Separation by Computational Screening of Materials. *Langmuir* **2012**, *28*, 14114-14128.
13. 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.
14. Simon, C. M.; Kim, J.; Gomez-Gualdron, D. A.; Camp, J. S.; Chung, Y. G.; Martin, R. L.; Mercado, R.; Deem, M. W.; Gunter, D.; Haranczyk, M.; Sholl, D. S.; Snurr, R. Q.; Smit, B., The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* **2015**, *8*, 1190-1199.
15. Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M., What Are the Best Materials to Separate a Xenon/Krypton Mixture? *Chem. Mater.* **2015**, *27*, 4459-4475.
16. Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q., Evaluating Topologically Diverse Metal–

Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci.* **2016**, *9*, 3279-3289.

17.	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.

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

19.	Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120-1127.

20.	Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C., Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501-1509.

21.	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.

22.	Glass, C. W.; Oganov, A. R.; Hansen, N., Uspex—Evolutionary Crystal Structure Prediction. *Comput. Phys. Commun.* **2006**, *175*, 713-720.

23.	Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q., New Developments in Evolutionary Structure Prediction Algorithm Uspex. *Comput. Phys. Commun.* **2013**, *184*, 1172-1182.

24.	Lonie, D. C.; Zurek, E., Xtalopt: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Comput. Phys. Commun.* **2011**, *182*, 372-387.

25.	Wang, Y.; Lv, J.; Zhu, L.; Ma, Y., Calypso: A Method for Crystal Structure Prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063-2070.

26.	Atomic Simulation Environment. https://wiki.fysik.dtu.dk/ase/about.html (accessed March 31, 2017).

27.	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.

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

29.	Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L., Recent Developments of the Chemistry Development Kit (Cdk)-an Open-Source Java Library for Chemo-and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111-2120.

30.	Cheng, L.; Assary, R. S.; Qu, X.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A., Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 283-291.

31.	Qu, X.; Jain, A.; Rajput, N. N.; Cheng, L.; Zhang, Y.; Ong, S. P.; Brafman, M.; Maginn, E.; Curtiss, L. A.; Persson, K. A., The Electrolyte Genome Project: A Big Data Approach in Battery Materials Discovery. *Comput. Mater. Sci.* **2015**, *103*, 56-67.

32.	Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Roman-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S., Lead Candidates for High-Performance Organic Photovoltaics from High-Throughput Quantum Chemistry–the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698-704.

33.	Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A., The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241-2251.

34.	van der Horst, E.; Marqués-Gallego, P.; Mulder-Krieger, T.; van Veldhoven, J.; Kruisselbrink, J.; Aleman, A.; Emmerich, M. T.; Brussee, J.; Bender, A.; IJzerman, A. P., Multi-

Objective Evolutionary Design of Adenosine Receptor Ligands. *J. Chem. Inf. Model.* **2012**, *52*, 1713-1721.

35.   Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P., The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545-552.

36.   Weininger, D., Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.

37.   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.

38.   Halcrow, M. A., Structure: Function Relationships in Molecular Spin-Crossover Complexes. *Chem. Soc. Rev.* **2011**, *40*, 4119-4142.

39.   Létard, J.-F.; Guionneau, P.; Goux-Capes, L., Towards Spin Crossover Applications. In *Spin Crossover in Transition Metal Compounds III*, Springer: 2004; pp 221-249.

40.   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**.

41.   Bignozzi, C. A.; Argazzi, R.; Boaretto, R.; Busatto, E.; Carli, S.; Ronconi, F.; Caramori, S., The Role of Transition Metal Complexes in Dye-Sensitized Solar Devices. *Coord. Chem. Rev.* **2013**, *257*, 1472-1492.

42.   Harvey, J. N.; Poli, R.; Smith, K. M., Understanding the Reactivity of Transition Metal Complexes Involving Multiple Spin States. *Coord. Chem. Rev.* **2003**, *238*, 347-361.

43.   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.

44.   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.

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

46.   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.

47.   Krausbeck, F.; Sobez, J.-G.; Reiher, M., Stabilization of Activated Fragments by Shell-Wise Construction of an Embedding Environment. *J. Comput. Chem.* **2017**, n/a-n/a.

48.   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.

49.   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.

50.   Foscato, M.; Houghton, B. J.; Occhipinti, G.; Deeth, R. J.; Jensen, V. R., Ring Closure to Form Metal Chelates in 3D Fragment-Based De Novo Design. *J. Chem. Inf. Model.* **2015**, *55*, 1844-1856.

51.   Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R., Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 767-780.

52.   Foscato, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R., Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 1919-1931.

53.   Halgren, T. A., Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of Mmff94. *J. Comput. Chem.* **1996**, *17*, 490-519.

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

55.	Gani, T. Z.; Ioannidis, E. I.; Kulik, H. J., Computational Discovery of Hydrogen Bond Design Rules for Electrochemical Ion Separation. *Chem. Mater.* **2016,** *28*, 6207-6218.

56.	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.* **2016,** *43*, 327-345.

57.	Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *arXiv:1702.05771.*

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

59.	Gal, Y.; Ghahramani, Z., Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv preprint arxiv:1506.02142* **2015.**

60.	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.

61.	Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G., Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016,** *8*, 331-337.

62.	Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A., Uncertainty Quantification Framework Applied to the Water–Gas Shift Reaction over Pt-Based Catalysts. *J. Phys. Chem. C* **2016,** *120*, 10328-10339.

63.	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.

64.	Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W., Density Functionals for Surface Science: Exchange-Correlation Model Development with Bayesian Error Estimation. *Phys. Rev. B* **2012,** *85*, 235149.

65.	Gani, T. Z.; Kulik, H. J., Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+ U. *J. Chem. Theory Comput.* **2016,** *12*, 5931-5945.

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

67.	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.

68.	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.

69.	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.

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

71.	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.

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

73.	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.

74.	Clark, D. E.; Westhead, D. R., Evolutionary Algorithms in Computer-Aided Molecular Design. *J. Comput.-Aided Mol. Des.* **1996,** *10*, 337-358.

75.   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.* `just accepted` http://dx.doi.org/10.1021/acs.chemmater.7b00472.

76.   Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S., The Chembl Bioactivity Database: An Update. *Nucleic Acids Res.* **2014,** *42*, D1083-D1090.

77.   Hansen, E. C.; Pedro, D. J.; Wotal, A. C.; Gower, N. J.; Nelson, J. D.; Caron, S.; Weix, D. J., New Ligands for Nickel Catalysis from Diverse Pharmaceutical Heterocycle Libraries. *Nat. Chem.* **2016,** *8*, 1126-1130.

78.   Bender, A.; Glen, R. C., Molecular Similarity: A Key Technique in Molecular Informatics. *Org Biomol Chem* **2004,** *2*, 3204-3218.

79.   Eckert, H.; Bojorath, J., Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discov Today* **2007,** *12*, 225-233.

80.   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.

81.   Coello, C. A. C.; Lamont, G. B., *Applications of Multi-Objective Evolutionary Algorithms.* World Scientific: 2004; Vol. 1.

82.   Nicolaou, C. A.; Brown, N.; Pattichis, C. S., Molecular Optimization Using Computational Multi-Objective Methods. *Curr. Opin. Drug Discovery Dev.* **2007,** *10*, 316.

83.   Petachem. http://www.petachem.com. (accessed February 23, 2017).

84.   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.

85.   Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988,** *37*, 785--789.

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

87.   Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994,** *98*, 11623-11627.

88.   Vosko, S. H.; Wilk, L.; Nusair, M., Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980,** *58*, 1200-1211.

89.   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.

90.   Saunders, V. R.; Hillier, I. H., A "Level–Shifting" Method for Converging Closed Shell Hartree–Fock Wave Functions. *Int. J. Quantum Chem.* **1973,** *7*, 699-705.

91.   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.

92.   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.

93.   Bondi, A., Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964,** *68*, 441-451.

94.   Wang, L.-P.; Song, C., Geometry Optimization Made Simple with Translation and Rotation Coordinates. *J. Chem. Phys.* **2016,** *144*, 214108.

95.   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.

96.   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.

97.   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.

98.   Kim, H.; Park, J.; Lee, Y. S., A Protocol to Evaluate One Electron Redox Potential for Iron Complexes. *J. Comput. Chem.* **2013**, *34*, 2233-2241.

99.   Tokel-Takvoryan, N. E.; Hemingway, R. E.; Bard, A. J., Electrogenerated Chemiluminescence. Xiii. Electrochemical and Electrogenerated Chemiluminescence Studies of Ruthenium Chelates. *J. Am. Chem. Soc.* **1973**, *95*, 6582-6589.

100.   Saji, T.; Aoyagui, S., Polarographic Studies on Bipyridine Complexes: I. Correlation between Reduction Potentials of Iron (II), Ruthenium (II) and Osmium (II) Complexes and Those of Free Ligands. *J. Electroanal. Chem. Interfacial Electrochem.* **1975**, *58*, 401-410.

101.   Bajusz, D.; Racz, A.; Heberger, K., Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7*, 20.

102.   Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T., Extension of the Universal Force Field to Metal–Organic Frameworks. *J. Chem. Theory Comput.* **2014**, *10*, 880-891.

103.   Foscato, M.; Deeth, R. J.; Jensen, V. R., Integration of Ligand Field Molecular Mechanics in Tinker. *J. Chem. Inf. Model.* **2015**, *55*, 1282-1290.

104.   Mortensen, S. R.; Kepp, K. P., Spin Propensities of Octahedral Complexes from Density Functional Theory. *J. Phys. Chem. A* **2015**, *119*, 4041-4050.

105.   Chandrasekhar, N.; Sholl, D. S., Large-Scale Computational Screening of Binary Intermetallics for Membrane-Based Hydrogen Separation. *J. Phys. Chem. C* **2015**, *119*, 26319-26326.

106.   Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning*. Springer New York: 2009.

107.   Tibshirani, R., Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B* **1996**, *58*, 267-288.

108.   Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C., Variable Selection Using Random Forests. *Pattern Recogn. Lett.* **2010**, *31*, 2225-2236.

109.   Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825-2830.

110.   Efron, B., Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316-331.

111.   Efron, B.; Tibshirani, R., Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Am. Stat. Assoc.* **1997**, *92*, 548-560.

112.   Kohavi, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th international joint conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada, 1995; Vol. 2, pp 1137-1143.

113.   Molinaro, A. M.; Simon, R.; Pfeiffer, R. M., Prediction Error Estimation: A Comparison of Resampling Methods. *Bioinformatics* **2005**, *21*, 3301-3307.

114.   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.

115.   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.

116.   Randic, M., On Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 57-77.

TOC Graphic