

Enumeration of *de novo* inorganic complexes for chemical discovery and machine learning

Stefan Gugler¹, Jon Paul Janet¹, and Heather J. Kulik^{1,*}

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

02139

ABSTRACT: Despite being attractive targets for functional materials, the discovery of transition metal complexes with high-throughput computational screening is challenged by the amount of feasible coordination numbers, spin states, or oxidation states and the potentially large sizes of ligands. To overcome these limitations, we take inspiration from organic chemistry where full enumeration of neutral, closed shell molecules under the constraint of size has enriched discovery efforts. We design monodentate and bidentate ligands from scratch for the construction of mononuclear, octahedral transition metal complexes with up to 13 heavy atoms (i.e., metal, C, N, O, P, or S). From > 11,000 theoretical ligands, we develop a heuristic score for ranking a chemically feasible 2,500 ligand subset, only 71 of which were previously included in common organic molecule databases. We characterize the top 20% of scored ligands with density functional theory (DFT) in an octahedral homoleptic ligand database (OHLDB). The OHLDB contains i) the geometry optimized structures of 1,250 homoleptic octahedral complexes obtained from the enumerated pool of ligands and an open-shell transition metal (M(II)/M(III), M = Cr, Mn, Fe, or Co), and ii) the resulting high-spin/low-spin adiabatic electronic energies (ΔE_{H-L}) obtained with hybrid DFT. Over the OHLDB, we observe structure–property (i.e., ΔE_{H-L}) relationships different from those expected on the basis of ligand field arguments or from our prior data sets. Finally, we demonstrate how incorporating OHLDB data into artificial neural network (ANN) training improves ANN out-of-sample performance on much larger transition metal complexes.

1. Introduction.

Computational, first-principles (i.e., with density functional theory or DFT) high-throughput screening¹⁻⁷ is an essential complement to experimental⁸⁻¹⁰ efforts in the discovery and design of molecules and materials. In recent years, machine learning (ML) property prediction models trained on first-principles simulation data have further accelerated this discovery process^{4, 11-18} throughout chemistry¹⁹⁻²⁴, including for catalysis^{15-16, 25-26} and materials^{4, 27-34}. Unique challenges arise in applying these tools to the discovery of open shell transition metal complexes, despite their importance as selective catalysts³⁵⁻⁴³ and functional materials (e.g., molecular switches or sensors⁴⁴⁻⁵²). The theoretical chemical space of inorganic complexes is diverse and relatively unexplored due to the variable spin, oxidation state, and coordination numbers feasible for each metal. The large size of inorganic complexes and limited applicability of more affordable semi-empirical⁵³ or force field⁵⁴⁻⁵⁵ methods in open-shell transition metal chemistry also hinders the rapid computational exploration of this space.²⁸

To accelerate discovery in open shell transition metal chemistry, we have developed^{4, 17, 27-28, 56-57} ML models for predicting quantum mechanical properties (here, computed with DFT), including spin-splitting energies^{17, 27, 56}, redox or ionization potentials^{28, 56}, metal-ligand bond lengths^{17, 28}, frontier orbital energies⁴, and reaction energetics⁵⁸. A key outstanding challenge for ML model improvement, especially in inorganic chemistry, is the generation of large data sets. While most organic chemistry ML models have been trained^{24, 59-65} on large (> 100k points) data sets^{19, 66-67} of molecules consisting of up to 9 heavy (i.e., C, N, O, or F) atoms, the higher computational cost associated with the larger number of electrons and added complexity of open shell character have limited data set sizes for inorganic chemistry²⁸. Despite these limitations in data set size, ML models for inorganic chemistry on modest data sets are predictive to 1-3

kcal/mol as judged by test set errors^{17, 56}. However, limited coverage of the wide range of chemical bonding in transition metal complexes means that ML model prediction error can rise rapidly (to ca. 6-10 kcal/mol^{17, 27}) when applied to complexes dissimilar from training data. Thus, an approach to efficiently increase data set diversity would benefit ML models in inorganic chemistry.

Here, we take inspiration from organic chemistry, where systematic enumeration of small, drug-like molecules⁶⁸⁻⁷³ paved the way for large data sets suitable for ML. However, the enumeration of possible inorganic complexes will necessarily differ from such prior organic enumerations, as the properties that make a molecule a good ligand for an inorganic complex (e.g., an unsaturated atom that can freely complex to a metal) are not the same as the characteristics that define closed shell organic molecules. To construct such a set we note that, whether through *ad hoc* feature engineering¹⁷ or systematic feature selection⁵⁶, the most predictive and transferable feature sets for open shell transition metal complex properties emphasize metal-local features. For spin-splitting energetics in particular, we found that the 24 most informative heuristic features obtained from feature selection^{28, 56} were primarily (80%) comprised of properties from atoms within two bonds of the metal on the molecular graph. Thus, by selecting small ligands and systematically varying their properties, we expect to capture the most important variations needed to improve coverage for ML model training.

In this work, we enumerate *de novo* octahedral transition metal complex ligands, study properties of homoleptic complexes of these ligands with first-principles DFT simulation, and demonstrate the improvement of our inorganic ML models through the inclusion of this data. The rest of this work is as follows. In Section 2, we describe our rules for enumerating monodentate and bidentate ligands composed of up to two and four heavy atoms, respectively. In

Section 3, we describe the Computational Details of our simulation methodology. In Section 4, we analyze the results of DFT geometry optimizations of homoleptic octahedral complexes built from these ligands and demonstrate how these data points improve ML model performance. Finally, in Section 5, we provide our outlook and conclusions.

2. Enumerating Inorganic Complex Ligands.

We enumerated candidate ligands from the elements C, N, O, P, or S, which were chosen i) for their abundance^{70-71, 74} on earth and in organisms and ii) to enable comparison between isovalent compounds (i.e., N vs. P and O vs. S). Molecules were classified into three ligand types defined by the number of heavy (i.e., non-H) atoms: monodentate ligands with one (M1) or two (M2) heavy atoms and bidentate ligands with four heavy atoms (B4) created from joining two identical M2 heavy atom pairs, all of which were variably passivated with H atoms (Figure 1). After enumeration of all theoretical ligands of these three types, we carried out two filtering steps, first excluding major violations of chemical bonding rules, then scoring the remaining cases and retaining the top-scoring ligands (Figure 1). We penalized but did not exclude ligands that do not have octet valence or neutral charge, diverging from previous organic enumeration efforts⁶⁹ that were not focused on ligand generation for inorganic chemistry. We did however remove any enumerated ligands with an odd number of electrons to avoid ligand noninnocence⁷⁵. Ligand scores were based on heuristic properties, e.g.: i) the number of H atoms bonded to any of the heavy atoms, h ; ii) the charge, c ; iii) the number of lone pairs, l ; and iv) the number of valence electrons, v , as described next.

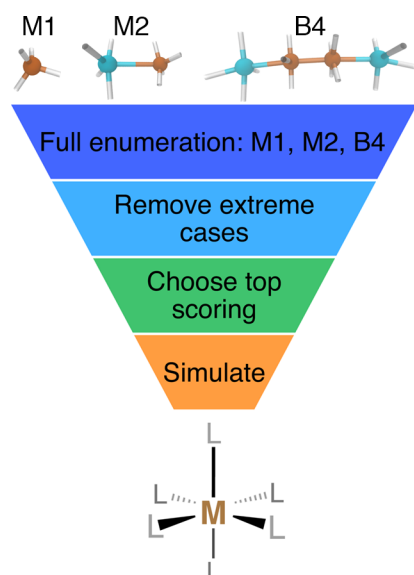


Figure 1. Schematic of M1, M2, and B4 ligands, as designated by the number of heavy atoms (i.e., C, N, O, P, or S) and number of metal coordinating atoms (top) with hypothetical places to add hydrogen atoms indicated with translucent sticks. The filtering process consists first of enumeration of all possible ligands, removal of extreme cases, scoring and retaining top-scoring ligands, and then finally the simulation of homoleptic M(II)/M(III) ($M = \text{Cr}, \text{Mn}, \text{Fe}, \text{or Co}$) octahedral complexes with DFT as indicated in the flowchart.

2.1. M1 Ligands

The simplest case corresponds to M1 ligands generated from a single heavy (i.e., C, N, O, P, S) atom with variable H atom passivation and charge. For initial enumeration, there were 5 possible heavy elements, we permitted 5 overall charge values, c , from -2 to 2 in an increment of 1, and we varied the number of H atoms added, h , over the range of 0 to 4. This combination produced $5 \times 5 \times 5 = 125$ theoretical M1 ligands. After discarding ligands with an odd number of electrons and strongly positively charged ligands (i.e., retaining only $-2 \leq c \leq 1$), we obtained 50 ligands for the second filtering step.

For M1 ligands, we assigned three scores, s , for the charge (s_{charge}), sterics (i.e., presence of H atoms, s_{sterics}), and accordance with the octet rule (s_{octet}). We favored neutral and weakly negative ligands ($s_{\text{charge}} = 3$ for $-2 \leq c \leq 0$) over positively charged ligands ($s_{\text{charge}} = 0$ for $c = 1$), due to the fact that ligands will be complexed with positively charged metal centers (Figure 2).

We penalized $h = 4$ ($s_{\text{sterics}} = 0$) passivation over all other choices ($s_{\text{sterics}} = 3$) because large numbers of H atoms hinder the heavy atom from coordinating to the metal center (Figure 2). We reduced the s_{octet} of a valence, v , in proportion to its violation of the octet rule (Figure 2):

$$s_{\text{octet}} = 4 - |8 - v| \quad (1)$$

A total score, s_{tot} , for M1 ligands was given by:

$$s_{\text{tot}} = s_{\text{charge}} + s_{\text{sterics}} + s_{\text{octet}} \quad (2)$$

which theoretically ranged from 0 to 10. In practice, over the 50 pre-filtered ligands, the scores ranged from 3 to 10.

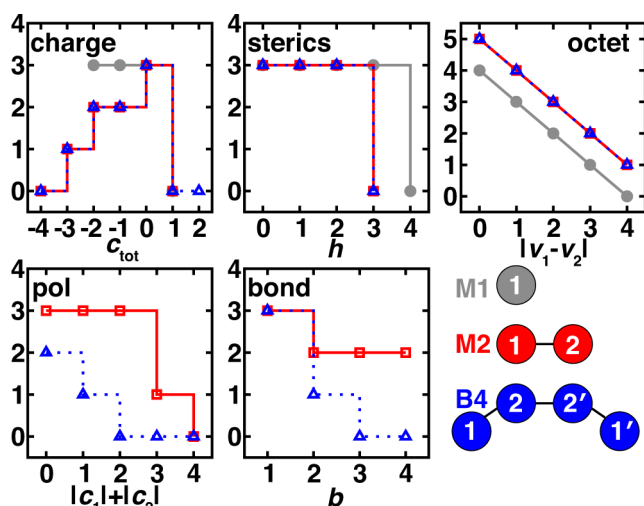


Figure 2. Scores for each of the three ligand types (i.e., M1 in gray circles and solid lines, M2 in red squares, and B4 in blue triangles and dotted lines) colored according to the bottom right inset. Three of the scores apply to all three ligand types: charge (c_{tot} , top left); sterics, as judged through number of hydrogen atoms on the metal-coordinating atom (h , top middle); and octet, as judged through valence deviations ($|v_1 - v_2|$, where $v_2 = 8$ for M1 ligands, top right). Two of the scores apply only to M2 and B4 ligands: pol, the polarization measured by $|c_1| + |c_2|$ (bottom, left) and bond, b , for the 1-2 bond order (bottom, middle).

Common M1-type ligands in the spectrochemical series⁷⁶ include OH^- , H_2O , and NH_3 , all of which were top scoring (i.e., $s_{\text{tot}} = 10$, Figure 3 and ESI Table S1). M1 ligands with $s_{\text{tot}} = 8$ that have been observed experimentally or invoked in the spectrochemical series include

methylene⁷⁷ (CH₂) and elemental sulfur⁷⁸ (Figure 3). Still lower scores (i.e., $s_{\text{tot}} = 7$) arose primarily due to penalties for: i) steric repulsion, as in methane (CH₄) or ii) charge, as in sulfonium (SH₃⁺) and hydronium (OH₃⁺) (Figure 3). The lowest scoring ligand ($s_{\text{tot}} = 3$), OH₄²⁻, simultaneously violates steric and octet rules (Figure 3). After filtering on score, we retained only the 29 M1 ligands with $s_{\text{tot}} \geq 8$, all of which were neutral (9) or negatively charged (20, see ESI Figure S1).

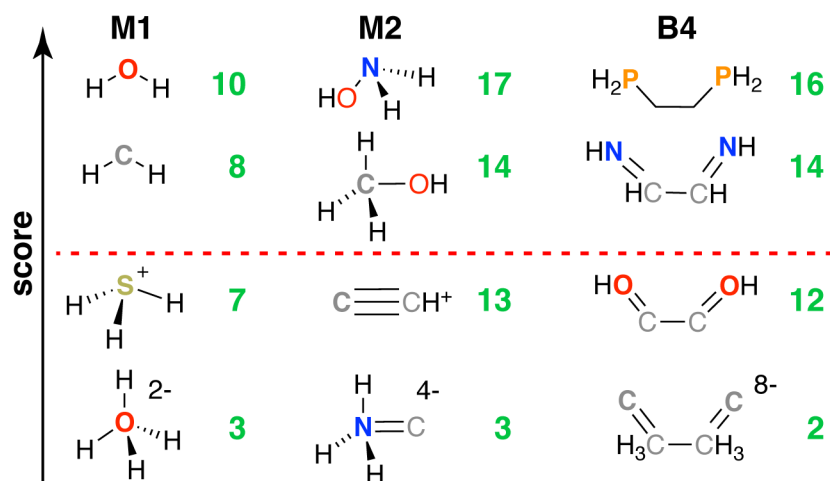


Figure 3. Four representative structures of M1 (left), M2 (middle), and B4 (right) ligand types ordered by their relative scores from top to bottom as indicated by left axis. The topmost and bottommost structures in each case correspond to ligands that score the maximum and minimum observed scores. The quantitative total scores for the respective ligand types are shown in each case at right in green, and the red dashed line indicates the separation between retained ligands and those below their respective ligand type cutoffs.

2.2. M2 Ligands.

For initial enumeration of M2 ligands, two atoms of the 5 possible heavy elements were joined, the total charge was constrained overall (i.e., retaining only $-4 \leq c_{\text{tot}} \leq 4$), and each atom was allowed to have between 0 and 4 passivating H atoms. The identities of the first and second atoms (i.e., indexed 1 and 2) were treated distinguishably because we defined the first atom as coordinating the metal, and we allowed the charges and passivating H atoms to vary between the two atoms. Thus, the initial theoretical space of M2 ligands was 5,625, from 25 combinations of

five types of atom 1 and 2 elements, 9 charge assignments, and 25 combinations of atom 1 and atom 2 H atom passivation.

In addition to eliminating ligands that produce odd numbers of electrons as in the M1 ligands, we also considered the expected bonding between the two heavy atoms in the ligand for filtering and scoring of the M2 ligands. To determine M2 candidate ligand bond order, we first assigned individual c_1 , c_2 charges by choosing from all combinations that could produce the c_{tot} value under constraint that the valence, v , of each i th atom was satisfied:

$$v_i = ve_i - c_i - 2 \cdot l_i - h_i \quad (3)$$

where l_i are the number of lone pairs of the atom (e.g., 1 for N atom), h_i are the number of hydrogen atoms, and ve_i are the maximum standard valence electrons (e.g., 5 for N atom, see ESI Table S2). From all possible values of c_1 and c_2 that satisfy eqn. (3), we chose the values that minimized $|v_1 - v_2|$. The bond order, b , was then assigned as:

$$b = \min(v_1, v_2) \quad (4)$$

with an allowable range of $0 < b \leq 4$, and b was set to 0 if eqn. (3) could not be satisfied (ESI Algorithm S1). For example, $c_{\text{tot}} = 1$ for the molecule NO^+ could be distributed as $c_i = +2/-1$ or $c_i = +1/0$; the algorithm selected $c_{\text{N}} = +1$ and $c_{\text{O}} = 0$ because this result gives $v_1 = v_2 = 2$ to minimize $|v_1 - v_2|$ and thus maximize b (ESI Algorithm S1).

After assembling these M2 ligands, we discarded i) highly positively charged ligands (i.e., $c_{\text{tot}} > 1$), ii) metal-coordinating heavy atoms (i.e., atom 1) with $h > 3$, and iii) ligands with $b = 0$ predicted between the two heavy atoms. We then scored the remaining 1,171 ligands with scores adapted and augmented from the M1 case. In comparison to M1 charge scoring, the M2 ligand charge score was biased toward neutral ligands ($s_{\text{charge}} = 3$ for $c_{\text{tot}} = 0$) and weakly

penalized intermediate, negatively charged ligands ($s_{\text{charge}} = 2$ for $c_{\text{tot}} = -1$ or -2 , $s_{\text{charge}} = 1$ for $c_{\text{tot}} = -3$, see Figure 2). Sterics of the M2 ligand were scored only for the metal-coordinating atom but we penalized $h = 3$ or higher due to the presence of the second heavy atom (i.e., $s_{\text{sterics}} = 3$ for $h < 3$, Figure 2). The M2 s_{octet} score was applied over atom 1 and 2 v values (Figure 2):

$$s_{\text{octet}} = 5 - |v_1 - v_2| \quad (5)$$

where the absolute difference of v_1 and v_2 ranged from 0 to 4 over the retained ligands. This produced a practical s_{octet} of 1 to 5 (Figure 2 and see ESI).

We introduced two bond-specific scores for the bond between the two heavy atoms in the M2 ligands: i) bond order, s_{bond} and ii) charge polarization, s_{pol} . We already excluded $b = 0$ molecules, but in scoring we favored $b = 1$ ($s_{\text{bond}} = 3$) weakly over $b > 1$ ($s_{\text{bond}} = 2$) to avoid oversampling high bond-order, few-atom M2 ligands (Figure 2). We disfavored unphysically ionic bonds in the present ligands by scoring highest the cases with low atom-wise charge assignment ($s_{\text{pol}} = 3$ for $|c_1| + |c_2| \leq 2$), assigning intermediate scores for moderate polarization ($s_{\text{pol}} = 1$ for $|c_1| + |c_2| = 3$) and penalizing the highest formal charges ($s_{\text{pol}} = 0$ for $|c_1| + |c_2| = 4$, see Figure 2).

A total M2 ligand score was thus:

$$s_{\text{tot}} = s_{\text{charge}} + s_{\text{sterics}} + s_{\text{octet}} + s_{\text{bond}} + s_{\text{pol}} \quad (6)$$

which had a range of 3 to 17 over scored ligands, due to minimum values for s_{octet} of 1 and s_{bond} of 2 (Figure 2 and ESI Table S3).

The 55 highest scoring ligands ($s_{\text{tot}} = 17$) include the common metal-complexing ligands methylamine (NH_2CH_3),⁷⁹ hydrogen peroxide (H_2O_2), and both hydroxylamine (NH_2OH)⁸⁰ and its experimentally-observed analogue, thiophosphinous acid⁸¹ (PH_2SH , Figure 3). Both reactive

peroxide⁸² (O_2^{2-}) and nitrosyl⁸³ (NO^-) ($s_{\text{tot}} = 15$) score highly as do stable small molecules such as N_2 and HCN ($s_{\text{tot}} = 16$) and methanol (CH_3OH) (O-coordinating $s_{\text{tot}} = 17$, C-coordinating $s_{\text{tot}} = 14$, see Figure 3). Most M2-type ligands in the spectrochemical series⁸⁴ have high scores (e.g., CN^- , $s_{\text{tot}} = 15$ and CO , $s_{\text{tot}} = 16$), with NO^+ ($s_{\text{tot}} = 13$) scoring the lowest due to its positive charge. Other $s_{\text{tot}} = 13$ ligands include sterically hindered ligands (e.g., $\text{NH}_3\text{NH}_3^{2-}$) or likely unstable, positively charged species, such as CCH^+ (Figure 3). The lowest score ($s_{\text{tot}} = 3$) was only assigned to 2 ligands when steric repulsion, octet rule violation, and unfavorable charges were all present in a single molecule (e.g., NH_3C^{4-} : $s_{\text{octet}} = 1$, $s_{\text{bond}} = 2$, all other scores are zero, see Figure 3). From the scored subset of M2 ligands, we therefore retained the 494 M2 ligands with $s_{\text{tot}} \geq 14$ for further characterization with DFT (Sec. 4). Most of the retained ligands have single or double bonds between the two heavy atoms and are neutral or negatively charged (ESI Figure S2).

2.3. B4 Ligands.

We generated symmetric bidentate B4 ligands solely by joining two identical M2 ligands from the original pool of 5,625 theoretical M2 ligands, and so there were also 5,625 theoretical B4 ligands. To simplify our algorithmic approach, we did not remove hydrogen atoms or use fragments with unpaired electrons to generate B4 ligands, as might be done in an intuitive dimerization procedure. If an atom was labeled as the metal-coordinating atom (i.e., 1) in the M2 ligand, it remained so in the B4 ligands, and the identical atoms in the ligand were assigned the indices 1' and 2'. To construct the B4 ligands from the M2 substituents, we defined the 2-2' bond order ($b_{2-2'}$) by reassigning the value for b_{1-2} (and equivalently $b_{1'-2'}$, Figure 2). This reassignment was necessary because the number of electrons and hydrogen atoms in the B4 ligand is simply twice that of the original M2 ligand. To determine the reassigned bond orders, we took the c_{tot}

from the parent M2 molecule and computed all possible c_i and v_i , requiring that $v_1 > 0$ and $v_2 > 1$, where v_i was assigned from eqn. (3).

We next determined the b_{12} and $b_{22'}$ bond orders simultaneously by iterating over allowed values between single and triple bonds:

$$\{b_{12}, b_{22'}\} = \underset{\substack{b_{12} \in \{1,2,3\} \\ b_{22'} \in \{1,2,3\}}}{\text{argmin}} \left\{ |v_1 - b_{12}| + |v_2 - b_{12} - b_{22'}| \right\} \quad (7)$$

where the first term is the difference between atom 1 valence electrons and the number of electrons used in the 1-2 bond, and the second term is the difference between the atom 2 valence electrons and those used in either the 1-2 or 2-2' bond (ESI Algorithm S2). If multiple choices of b_{12} and $b_{22'}$ minimized the argument, we selected the one with lower charge polarization (i.e., lower $|c_1| + |c_2|$), and zero bond order was again assigned if no result satisfied the equation (ESI Algorithm S2). For example, the neutral M2 ligand NH_2^-CH^+ has $b_{12} = 2$ with $c_1 = -1$ and $c_2 = 1$. The resulting B4 ligand formed from two of these M2 ligands is $\text{NH}_2\text{-CH=CH-NH}_2$, which has the same net charge but neutral atom-wise charge assignments with $b_{1-2} = b_{1'-2'} = 1$ and $b_{2-2'} = 2$.

After assembling all theoretical B4 ligands, we discarded i) ligands with two consecutive double bonds or any cases with triple bonds that would be unable to form a 1-2-2'-1' dihedral necessary to enable bidentate metal coordination, ii) strongly charged ligands with total charge (i.e., twice the charge of the original M2 ligand) higher than 4 or more negative than -8, iii) cases in which any bond orders were assigned to be zero, and iv) ligands with connecting atoms that had three or more passivating hydrogen atoms. These down-selection steps left a pool of 1,356 B4 ligands suitable for further scoring with a five-component score similar to that applied to the M2 ligands. Both s_{charge} and s_{sterics} were unchanged from the M2 case but were evaluated, respectively, on the charge only of the original M2 fragment (i.e., half of the total B4 ligand

charge) or a single relevant connecting atom (Figure 2). The B4 s_{octet} was also scored only for a single building block using eqn. (5) but using the revised valency after charge redistribution (Figure 2). After redistribution of charges, we scored a single 1-2 bond, penalizing strong polarization ($s_{\text{pol}}=0$ for $|c_1|+|c_2| > 1$), favoring completely neutral atomic charges ($s_{\text{pol}}=2$ for $|c_1|+|c_2| = 0$), and assigning an intermediate score ($s_{\text{pol}}=1$) for slight polarization (i.e., $|c_1|+|c_2|=1$, see Figure 2). We also computed s_{bond} only on the 1-2 bond and reduced it with respect to M2 scoring for higher bond orders ($s_{\text{bond}} = 3$ for $b = 1$, $s_{\text{bond}} = 1$ for $b = 2$, see Figure 2).

The five B4 metrics were combined as in eqn. (6) for a total score with a range of 2 to 16 across ligands retained for scoring (Figure 1). Ethylenediamine (en, $\text{C}_2\text{H}_8\text{N}_2$) is the only B4-type ligand in the spectrochemical series⁸⁴, and it has a maximum score ($s_{\text{tot}} = 16$) as does its phosphorus analogue, 1,2-ethanediyldiphosphine ($\text{C}_2\text{H}_8\text{P}_2$, $s_{\text{tot}} = 16$, Figure 3 and ESI Table S4). Other common ligands that score highly include ethylene glycol ($\text{C}_2\text{H}_6\text{O}_2$, $s_{\text{tot}} = 16$) and 1,2-ethanediimine ($\text{C}_2\text{H}_4\text{N}_2$, $s_{\text{tot}} = 14$), which contains a bonding pattern analogous to that observed in the common bipyridine ligand (Figure 3). Stable organic molecules that are sterically hindered and would make poor ligands, e.g., butane (C_4H_{10} , $s_{\text{tot}} = 13$), score more poorly than unstable but likely metal-coordinating molecules, e.g., tetraoxidane⁸⁵ (H_2O_4 , $s_{\text{tot}} = 16$, Figure 3). Other molecules with intermediate scores ($s_{\text{tot}} = 12$) typically score lower due to a combination of steric hindrance, charge, and violations of the octet rule (e.g., S_4 , or HOC_2OH , Figure 3). The lowest scoring ligands are charged, have polarized bonds, and are octet violating (e.g., $\text{CCH}_3\text{CH}_3\text{C}^{8-}$, $s_{\text{tot}} = 2$, Figure 3). Following these holistic observations, we retained only the 47 B4 ligands with $s_{\text{tot}} \geq 14$ for subsequent DFT calculations (ESI Figure S3).

2.4. Overall Ligand Analysis.

From an original combinatorial space of 11,325 theoretical M1, M2, and B4 ligands, we

scored 2,577 ligands and identified the top 20% (570 ligands) as good candidates for DFT characterization (Figure 1 and see Sec. 4). We next compared characteristics of these ligands to molecules in other curated data sets that satisfy M1, M2, or B4 ligand definitions. We focused on three representative databases: ChEMBL⁸⁶; the generated database of enumerated organic molecules with up to 9 heavy atoms, GDB-9⁶⁹; and a database of experimentally and computationally characterized diatomic molecules, DiRef⁸⁷. Across these databases, 71 of the 2,577 ligands were observed in at least one database, 17 occurred in more than one database, and most are diatomic molecules found only in DiRef⁸⁷ (ESI Table S5). The majority (55) of database ligands are M2 type, and 46 of these M2-type molecules are present twice in the enumerated ligand set, distinguished only by the metal coordinating atom and corresponding to a single chemical species. Thus, the 71 ligands identified in existing databases (DBs) are 48 chemically distinct species (ESI Table S5).

To simplify comparison of scores across the three ligand types, we obtained scaled scores, s_{scaled} , between the minimum and maximum observed s_{tot} values within each ligand type:

$$s_{\text{scaled}} = \frac{s_{\text{tot}} - s_{\text{tot}}^{\min}}{s_{\text{tot}}^{\max} - s_{\text{tot}}^{\min}} \times 100 \quad (8)$$

The scaled threshold for retention is lowest for M1 ligands (> 71%) and highest for the B4 cases (> 86%, see Figure 4). Over all ligands, 75% (53 of 71 overall; 36 of 48 unique) of compounds found in DBs were above the relevant ligand cutoff (Figure 4). The majority (13 of 18, 7 of 12 unique) of below-cutoff ligands are M2 type and correspond to positively charged diatomics from DiRef⁸⁷ (e.g., NO⁺ and PS⁺ ESI Table S5). Although such ligands are relevant for heteroleptic transition metal complexes, our focus on homoleptic complexes with high valent metals motivated penalizing positively charged ligands (see Figure 2). The remaining database

ligands below threshold typically exhibit steric hindrance (e.g., M1: CH₄ or B4: C₄H₁₀, see Figure 4). Conversely, known good inorganic ligands such as M1 ammonia, M2 methylamine, or B4 ethylenediamine are all among the top scoring database ligands (Figure 4 and ESI Table S5).

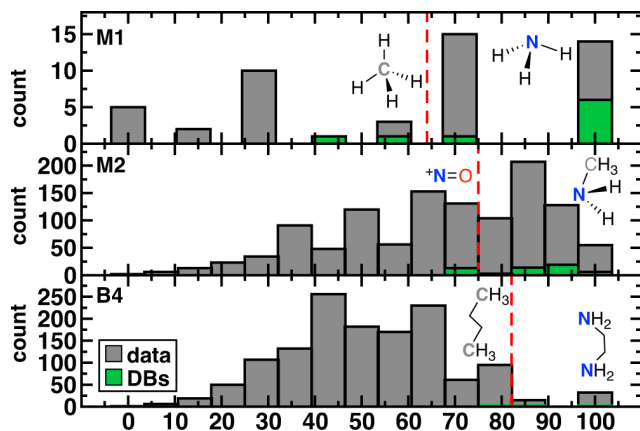


Figure 4. Distribution of scaled scores for 50 M1 (top), 1,171 M2 (middle), and 1,356 B4 (bottom) ligands over all data described in this work (bars shown in gray) as well as 71 ligands from three databases (DBs, in green). Representative molecules above and below the cutoff for ligand retention (red vertical dashed line) are shown in inset with the metal-coordinating atom in bold.

Overall, the enumeration recovered small molecules previously observed in other databases that are likely ligands in inorganic chemistry. Beyond validation of scoring heuristics, the large number of enumerated and retained ligands not present in other databases (i.e., around 90% or 517 of 570) suggests that our data set contains unique chemical bonding configurations. Thus, it will be useful to identify through first principles DFT simulation the extent to which complexes could be formed from these ligands.

3. Computational Details.

Homoleptic, mononuclear octahedral transition metal complexes were generated from *de novo* ligands (Sec. 2) with the molSimplify³ toolkit and molSimplify Automatic Design (mAD)^{4, 27}, which automated both structure and input file generation. The molSimplify³ code uses OpenBabel^{5, 88} as a backend for force field-based transition metal complex preoptimization prior to first-principles simulation with DFT. DFT geometry optimizations were carried out using

TeraChem⁸⁹⁻⁹⁰ using the B3LYP⁹¹⁻⁹³ hybrid DFT functional. The default definition of B3LYP in TeraChem employs the VWN1-RPA⁹⁴ form for the local density approximation correlation component. The LANL2DZ⁹⁵ effective core potential was employed for transition metals with the 6-31G* basis for all other atoms. These choices were made due to the limited effect of a modest basis set on the relative energies of interest⁹⁶ and to enable comparison to previously generated data sets^{17, 27, 56}.

All complexes were generated with four metals (M = Cr, Mn, Fe, and Co) in M(II) and M(III) oxidation states. The differences between high (H) and low (L) spin states, ΔE_{H-L} , was computed as the electronic energy difference between the two geometry-optimized states (i.e., the adiabatic energy difference). The high-spin/low-spin definitions for the metals studied in this work are: quintet-singlet for d^6 Co(III)/Fe(II), sextet-doublet for d^5 Fe(III)/Mn(II), quintet-singlet for d^4 Mn(III)/Cr(II), and quartet-doublet for both d^3 Cr(III) and d^7 Co(II). Although thermodynamic and solvent corrections are known to be important in making direct comparison with experimental spin state ordering⁹⁷, the two corrections typically have compensating effects⁹⁶⁻⁹⁷, and we therefore focused on relationships between ligand identity and DFT adiabatic electronic energies. All open-shell calculations (i.e., non-singlet spin states) were carried out with level shifting⁹⁸ using spin up and down level shifts of 1.0 and 0.1 Ha, respectively. Geometry optimizations used the L-BFGS algorithm in translation rotation internal coordinates (TRIC)⁹⁹ as implemented in TeraChem to the default tolerances of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in energy between steps.

4. Properties of *de novo* Transition Metal Complexes.

We next computed with DFT properties of homoleptic octahedral transition metal complexes containing the curated *de novo* ligands. Here, we focus on the high-spin to low-spin

adiabatic energy splitting, $\Delta E_{\text{H-L}}$, of M(II) and M(III) (M = Cr, Mn, Fe, or Co) complexes, which we obtained with hybrid DFT for comparison to both prior DFT and ML models^{17, 27, 56} (see Sec. 3). Although quantitative spin-state assignment remains an outstanding challenge for DFT^{97, 100-106} with no one-size-fits-all functional for spin-state energetics motivating more advanced methods¹⁰⁷⁻¹⁰⁹, ligand-dependent trends in relative spin-state ordering are expected to be less sensitive to method choice. From the 570 high-scoring ligands in Sec. 2, we excluded 10 M1 and 164 M2 ligands with net -2 charge from calculation due to the high, negative complex charge (i.e., -9 or -10) of the homoleptic octahedral complex that cannot be treated well within approximate DFT¹¹⁰⁻¹¹¹. For the remaining 396 ligands, 16 combinations of metal, oxidation, and spin state mean that 6,336 geometry optimizations were carried out for 3,168 possible $\Delta E_{\text{H-L}}$ evaluations (see Sec. 3).

To streamline and improve the quality of data ingested during high-throughput simulation³ of transition metal complexes, we recently introduced⁴ automated checks of geometry and properties of the wavefunction. The geometric checks, as outlined in Ref. 4, focus on preservation of metal-ligand bond lengths in the coordination sphere, ligand detachment, and ligand distortion. In practice, all simulations run with mAD⁴ are run in 24 hour increments, with geometry checks being carried out at each resubmission as well as on the final optimized structure. From the 6,336 initial geometry optimizations, 22% (1,387) of all geometry optimizations completed successfully, a somewhat reduced success rate with respect to the range reported in our prior work on transition metal complexes^{17, 56-57}. The majority of unsuccessful calculations corresponded to those that failed geometry checks initially (214 or 3%), during resubmission (3,816 or 60%) or on the fully optimized structure (919 or 15%). Some cases showed bond rearrangement within the ligand, which could lead to an alternative feasible

complex, but we judged success here as only cases where the original connectivity in the ligands was preserved.

Over the 1,387 data set of converged complex results, completion rates are roughly evenly distributed over the M(II) (716) and M(III) (671) oxidation states as well as metals (Cr: 348, Mn: 341, Fe: 361, and Co: 337, see ESI Figure S4). Some bias is observed for successful convergence of low spin states (792 singlets or doublets) vs. their high spin counterparts (595 quartets, quintets, or sextets), likely due to the weaker bonding in high-spin complexes. Separating convergence by ligand reveals that of the 396 ligands we initially selected, only 185 converged successfully in at least one metal, oxidation state, or spin state (Figure 5). The full ranges of retained scores are observed for successful ligands of all types, but the average score among the 185 ligand set is slightly higher than in the original 396 ligand set: M1 9.3 vs. 8.9 to average score, M2 15.8 vs. 15.4 average score, and B4 15.7 vs. 15.4 average score.

Dividing further by charge of the individual ligand, we observe that none of the 45 positively charged ligands of the M2 type in our original set led to a productive geometry optimization, further justifying our penalties on positively charged ligands during initial scoring (Sec. 2). The highest success rate (37% or 77 of 210) is observed for neutral (M1, M2, or B4) ligands, but approximately 20% of the negatively charged ligands (29 of 141) also produced at least one stable complex. Of the 1,387 converged geometries, we also discarded 134 for $\langle S^2 \rangle$ values that deviated from the anticipated value by more than $1 \mu_B$, as in prior work, to ensure limited symmetry breaking and localization of the spin on the metal. Finally, because ΔE_{H-L} evaluation requires two successful geometry optimizations for both the high and low spin state of a given octahedral complex, a total of 343 new ΔE_{H-L} evaluations were obtained for 106 unique ligands (see ESI for all energies and structures).

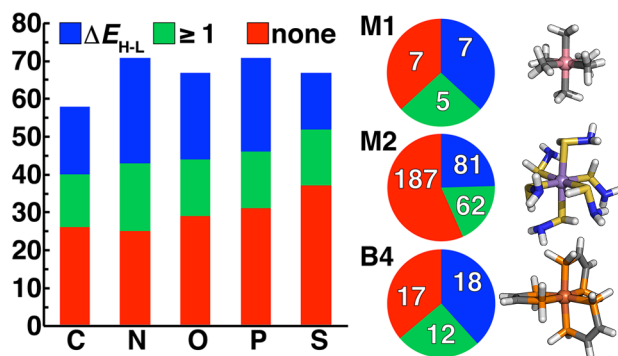


Figure 5. Comparison of ligands grouped by failed (red), at least one successful optimization (green), or at least one ΔE_{H-L} value (blue) separated by ligand connecting atom (C, N, O, P, or S) shown in bars (left) and divided by ligand type in pie charts (right, top to bottom: M1, M2, and B4). At right, an example complex for which ΔE_{H-L} evaluation was successful is shown for each ligand type (from top to bottom): M1 Co(III)(CH₃)₆, M2 Mn(II)(SHNH₂)₆, and B4 Fe(III)(PH₂CH=CHPH₂)₃.

In addition to being the most abundant ligands among the set selected for DFT characterization, N-coordinating ligands had the highest overall success rate, followed by P-coordinating ligands (Figure 5 and ESI Figure S4). Generally, the second row analogues (P or S) of first row (N or O) complexes had lower overall success rates, but the final ligand set remained relatively balanced over all coordinating-atom types (Figure 5). Because the majority of the selected ligand set is of the M2 type, this is also the greatest share of the successful ligands, but M2 ligands do not have the highest success rate (Figure 5). For both M2 and B4 ligands, around 20% of ligands had at least one successful geometry optimization but no spin splitting energy pair, likely due to the greater flexibility of these ligands that increases the probability that at least one spin state fails to pass geometry checks. Within M1 ligand types, only 7 of the 19 retained ligands converged with spin-splitting energies, and most were complexes that are well known or that we had previously incorporated into ML data sets^{4, 17, 56-57} (e.g., NH₃, PH₃, H₂S, and H₂O). One exception was a Co(III) complex of the ammonia analogue, CH₃⁻ (Figure 5). More diversity is observed in the successful B4 ligands, despite the relatively small size of the retained B4 ligand set, with the phosphorus analogue of the bipyridine core converging for multiple metal

centers (Figure 5). Finally, a large number of stable M2 ligand chemistries are observed for which spin-splitting energies were obtained, including a wide array of neutral (e.g., SHNH₂) and negatively charged complexes (Figure 5).

From the successfully converged complexes that make up our curated octahedral ligand database (OHLDB), we next quantified the extent to which these systematically enumerated complexes reflected chemistry divergent from the 1,901 ΔE_{H-L} values we had previously obtained for artificial neural network (ANN) model training^{4, 56}. To compare diversity in the chemical structures, we featurized each new complex with the revised autocorrelation (RAC-155) representation⁵⁶ (ESI Text S1). The RAC-155 representation consists of products and differences of heuristic properties on the molecular graph and has shown good performance^{4, 28, 56, 58, 96} for predicting inorganic chemistry properties, including ΔE_{H-L} . Although OHLDB complexes primarily lie within the convex hull of the first two principal components (PCs) in the RAC-155 representation, the overall Euclidean norm distance in feature space averaged over the ten nearest neighbors in existing data is quite large (> 20) for a number of the complexes (Figure 6). The complexes indeed fall outside the convex hull of the pre-existing data but do so especially at higher PCs (i.e., 7-8), where the first eight PCs generally contain the vast majority of the variance (89%, Figure 6 and ESI Figure S5).

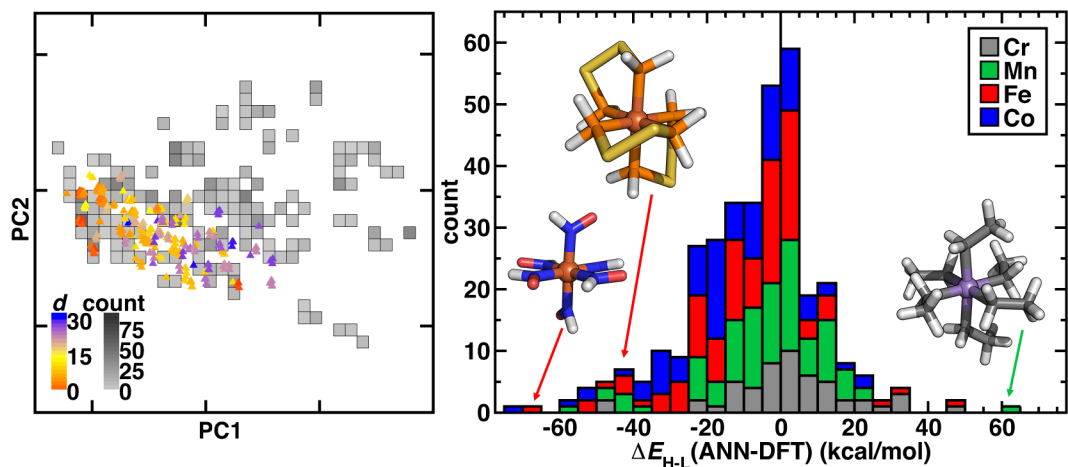


Figure 6. (left) Principal component analysis of new OHLDB data in the RAC-155 representation colored by Euclidean norm distance to available training data (d , colored according to inset colorbar) and overlaid on top of a 2D histogram of available data, with bins colored by count as indicated in grayscale colorbar. (right) Stacked histogram of errors (bin width: 5 kcal/mol) colored by metal type for the RAC-155/ANN prediction on OHLDB molecules with successful DFT $\Delta E_{\text{H-L}}$ evaluations. Representative large error complexes are shown in the histogram inset (left to right): Fe(II)(HNO)_6 , $\text{Fe(II)(PH}_2\text{SSPH}_2)_3$, and $\text{Mn(III)(CH}_2\text{CH}_3^-)_6$.

An alternative measure of data diversity is in property space, which we assessed first by determining if a previously trained RAC-155/ANN model¹¹² could have predicted the $\Delta E_{\text{H-L}}$ values exhibited by the OHLDB complexes (Figure 6 and ESI Table S6). Overall, although a large number of complexes were well predicted, significant (e.g., > 60 kcal/mol) over- and underestimations of $\Delta E_{\text{H-L}}$ are indicative of limited prior knowledge by the ANN (MAE = 14.3 kcal/mol) of the chemistry of the OHLDB complexes (Figure 6 and ESI Table S6). Indeed, high error points are both chemically distinct and exhibit unexpected spin-state ordering, such as an Fe(II)(HNO)_6 complex ($\Delta E_{\text{H-L}}$ ANN: -17.1, DFT: 50.1 kcal/mol), which contains an NO motif adjacent to the metal that had been absent from prior training complexes and is erroneously predicted by the ANN to be weak field in nature (Figure 6). Similarly, no phosphorus-coordinating metal complexes and few sulfur-containing ligands had been in training data, leading to large errors for an Fe(II) complex with bidentate PH_2SSPH_2 ligands ($\Delta E_{\text{H-L}}$ ANN: -27.8, DFT: 15.2 kcal/mol, Figure 6). Although phosphorus ligands are known to be low-spin directing, their absence from our training data means that accurate ANN predictions on such complexes cannot be expected. Finally, in some cases, the coordinating atom may be present in training data, but the chemistry is still unusual, as is the case for a strongly high-spin favoring $\text{Mn(III)(CH}_2\text{CH}_3^-)_6$ complex (Figure 6). Although the ANN correctly predicts this complex to be high spin, it cannot predict the strong high-spin stabilization observed in the DFT calculation ($\Delta E_{\text{H-L}}$ ANN: -11.8, DFT: -72.0 kcal/mol) for this saturated, negatively charged carbon ligand

that is distinct from other C-coordinating ligands (e.g., CO) in our prior training data sets.

Indeed, across a broad range of metals, oxidation states, and ligand coordinating atoms, the range of OHLDB $\Delta E_{\text{H-L}}$ values exceeds that seen in our prior data sets (Figure 7 and ESI Figures S6-S7). Expected trends are observed, such as carbon- and phosphorus-coordinating complexes generally corresponding to low-spin-directing, strong field ligands, especially for Mn(II), Fe(II/III), or Co(II/III) complexes (Figure 7 and ESI Figure S7). Although N-coordinating ligands generally form high-spin complexes, especially with Cr(II/III) or Mn(II/III) metals, notable exceptions are observed including low-spin Cr(II)(NSH)₆ ($\Delta E_{\text{H-L}} = 23.1$ kcal/mol) and Mn(II)(NNH₂)₆ ($\Delta E_{\text{H-L}} = 20.6$ kcal/mol) complexes (Figure 7). Given the dearth of low-spin Cr database complexes, the OHLDB therefore can be expected to enhance ML model predictions of $\Delta E_{\text{H-L}}$ (Figure 7).

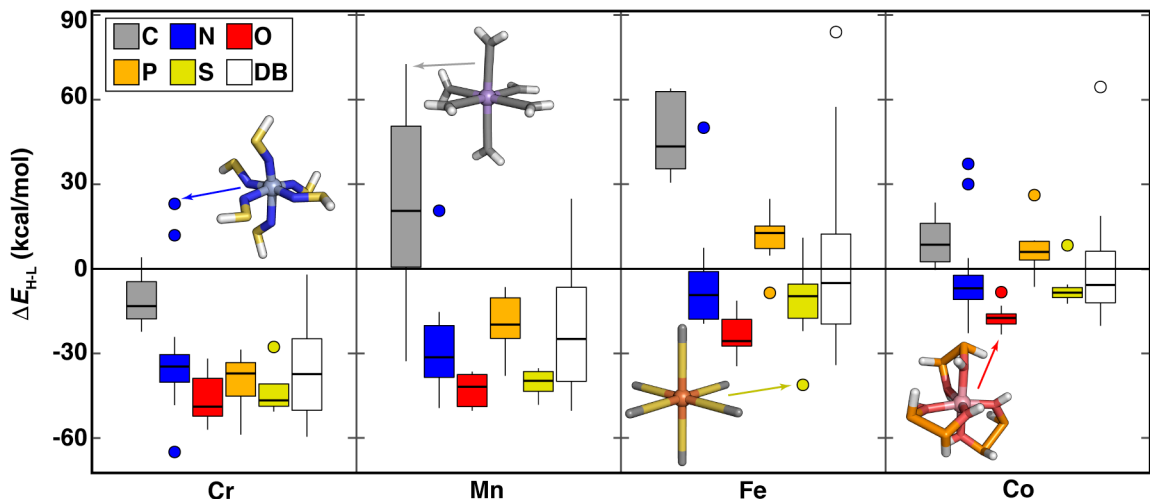


Figure 7. Boxplots of M(II) (M = Cr, Mn, Fe, or Co) $\Delta E_{\text{H-L}}$ (in kcal/mol) for ligands grouped by metal and by ligand-coordinating-atom (C in gray, N in blue, O in red, P in orange, and S in yellow, as shown in inset legend). Each box indicates the median by a horizontal line, the interquartile range (IQR), and whiskers indicate $1.5\times$ the IQR. Here, DB (white boxplot) corresponds to range of $\Delta E_{\text{H-L}}$ values from prior work in Refs. 4, 56 for the relevant metal, regardless of coordinating atom.

Unsaturated carbon-coordinating ligands (e.g., CCH₂ in Mn(II)(CCH₂)₆ $\Delta E_{\text{H-L}} = 72.6$ kcal/mol) are known to be low-spin directing ligands but they were absent from our earlier data

set, as were more unusual low-spin-directing ligands such as CHOH or CHNH₂ (Mn(II) $\Delta E_{\text{H-L}} = 47$ to 62 kcal/mol, Figure 7). The sulfur-coordinated Fe(II) complexes in our new data set span from low-spin (e.g., monodentate SHOH: $\Delta E_{\text{H-L}} = 11.0$ kcal/mol or bidentate SC₂H₂S: $\Delta E_{\text{H-L}} = 10.3$ kcal/mol) to high-spin (e.g., SC: $\Delta E_{\text{H-L}} = -41.1$ kcal/mol), corresponding to a range that we had not observed in prior Fe(II) data (Figure 7). Since sulfur is considered a soft element, low-spin sulfur complexes are somewhat surprising. Examination of the OHLDB confirms that the bidentate SC₂H₂S ligand also forms low-spin complexes with Fe(III) and Co(II/III) but forms high-spin complexes with Cr(II) and Mn(III) (see ESI). Saturating the sulfur (i.e., SHC₂H₂SH) and the carbon backbone (i.e., SHC₂H₄SH) instead yields the expected, uniformly high-spin complexes regardless of metal center and oxidation state (see ESI). Although oxygen-coordinating ligands are known to be weak-field, high-spin directing in nature, diverse ligand chemistry in the OHLDB yields, in addition to those previously observed, unexpectedly high-spin complexes e.g., Co(II)(OHP₂H₂OH)₃ ($\Delta E_{\text{H-L}} = -22.8$ kcal/mol, Figure 7). This bidentate ligand and the isoelectronic OHS₂OH ligand produce among the most high-spin-favoring Cr(II)/O-coordinating complexes ($\Delta E_{\text{H-L}}$ ca. -49 to -52 kcal/mol, see ESI).

The OHLDB also enables examination of how isovalent and isoelectronic variations in ligands alter spin-state ordering (Figure 8). Here, we focus on the widely-studied CO ligand and related isovalent and isoelectronic species, including those in the spectrochemical series⁷⁶ (e.g., CN⁻) and other common molecules (e.g., HCN and N₂, Figure 8). Given ligand definitions, $\Delta E_{\text{H-L}}$ can in principle be obtained for either orientation of asymmetric M2 ligands, but only 14 of these ligands in practice yielded at least one $\Delta E_{\text{H-L}}$ value for any metal or oxidation state (Figure 8). High-spin Fe(II) or Fe(III) complexes are formed from either homonuclear ligands (e.g., N₂ and P₂) or cases where the weaker-field element coordinates the metal (e.g., SC, OC), despite being

isovalent or isoelectronic with low-spin directing CO (Figure 8). These effects are not additive by element, where the NP ligand is low-spin directing, in spite of the high spin preferences of N₂ and P₂ (Figure 8). Although CO is often invoked as one of the strongest field ligands ($\Delta E_{\text{H-L}} = 20\text{-}30$ kcal/mol for Fe(II) or Fe(III) complexes), five ligands form even more low-spin-favoring complexes, including those where O is replaced by anionic (e.g., CN⁻, CP⁻) or less electron withdrawing species (e.g., CCH⁻, CNH, or CS, Figure 8). The trends observed for iron complexes generally hold for other metals, with Co(III) complexes exhibiting more low-spin bias for the same ligands, Mn(III) complexes remaining uniformly high spin, and all other metals and oxidation states generally residing within these two bounds (Figure 8). Thus, a very wide range of $\Delta E_{\text{H-L}}$ values (ca. -60 to +80 kcal/mol) can be obtained simply by adjusting the charge and elemental identities in M2-type ligands isoelectronic or isovalent to a common ligand.

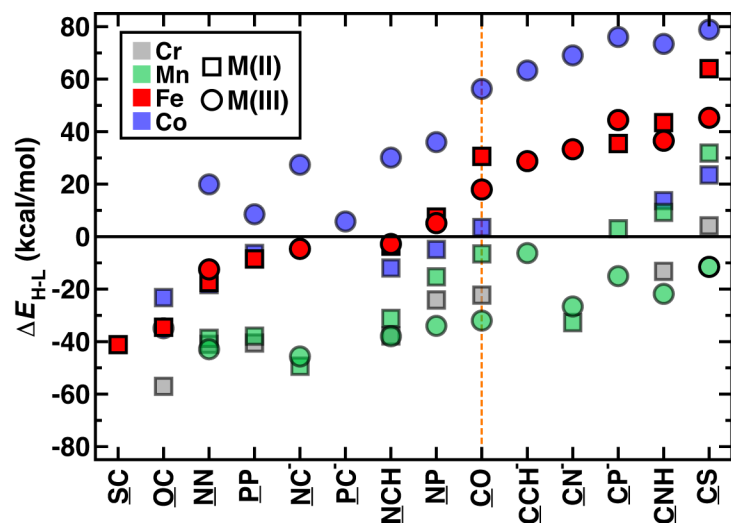


Figure 8. Adiabatic gas phase spin splitting, $\Delta E_{\text{H-L}}$, in kcal/mol for octahedral complexes with CO and isovalent or isoelectronic ligands. Complex energetics are shown for all converged DFT results in M(II) (squares) and M(III) (circles) oxidation states with Cr (gray), Mn (green), Fe (red), and Co (blue) metals. Ligands are ordered to be monotonically increasing for Fe(II/III) complexes, which are shown as solid red symbols, whereas all other metals and oxidation states are shown as translucent symbols to aid comparison. The metal coordinating atom in the ligand is underlined.

Finally, we considered the extent to which OHLDB data could be used to improve ML

model predictions on large, diverse complexes¹¹² by improving the chemical coverage of the ML model training data. We recently curated¹¹² a 116 complex out-of-sample test set from the Cambridge Structural Database (CSD)¹¹³ for testing $\Delta E_{\text{H-L}}$ predictions with a RAC-155/ANN model. Because the CSD complexes were chosen to be distinct from the 1,901 complexes used in the training of the ANN, the CSD set $\Delta E_{\text{H-L}}$ MAE of 8.6 kcal/mol was much poorer than set-aside test set errors (ca. 1-3 kcal/mol^{17, 56}) or uncertainty-controlled, out-of-sample prediction errors (ca. 4.5 kcal/mol²⁷). Notably, very high $\Delta E_{\text{H-L}}$ prediction errors, either due to over or underestimation, were observed on the order of 20-50 kcal/mol (Figure 9). Incorporating OHLDB data and retraining the RAC-155/ANN eliminated many of these highest error points and reduced CSD set average error to 6.7 kcal/mol (Figure 9, ESI Text S1, Table S6, and Figures S8-S9). Despite the fact that most of the CSD complexes are much larger in size, significant improvements are observed for complexes that had metal-adjacent coordination environments present in the OHLDB but absent in our prior data, such as coordination by NO species (CSD ID: CEYSAA, Figure 9). In most cases model performance improved, but for select complexes model performance remained the same or worsened slightly in a manner that is not dependent on the metal center (CSD ID: COBWEX, Figure 9 and ESI Figures S9-S10). Given that most of the CSD curated set¹¹² is multidentate in nature, whereas the OHLDB is weighted toward monodentate ligands, further improvement could likely be achieved through continued systematic enumeration of a greater number of ligands of higher denticity.

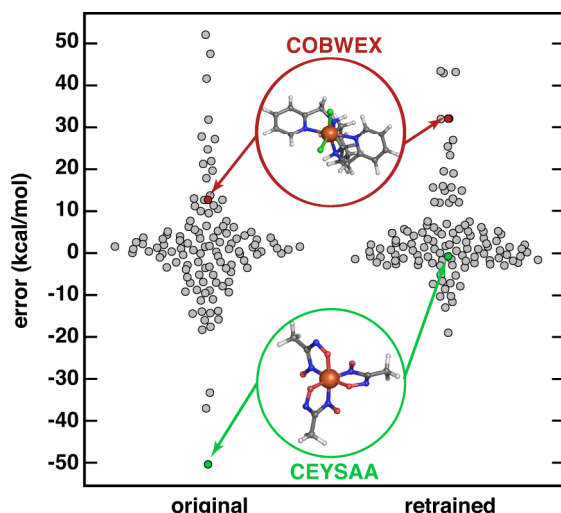


Figure 9. Swarm plot of RAC-155/ANN signed errors (in kcal/mol) on an out-of-sample 116 CSD structure data set¹¹² (original, left) and after retraining with OHLDB data (retrained, right). The single most improved (CSD ID: CEYSAA) and worsened (CSD ID: COBWEX) points are shown in green and red insets, respectively, and have data points colored in the same manner.

5. Conclusions.

We developed an approach for *de novo* ligand enumeration for the discovery of octahedral transition metal complexes. Our effort diverged from prior enumeration studies that had focused on neutral and stable organic molecules both by requiring that the individual ligands be smaller in size to form mononuclear octahedral transition metal complexes with no more than 13 heavy atoms as well as by relaxing prior constraints on charge or in satisfying the octet rule. From a space of over 11,000 theoretical monodentate or bidentate ligands comprised of C, N, O, P, or S heavy atoms, we identified a 2,500-ligand subset for scoring. Based on analysis of ligand feasibility by score, we identified cutoffs and retained a high-scoring, 570-ligand subset as most promising for subsequent calculations. Only a small number (71 of over 2,500) of our ligands were in prior databases, and most (75%) of those ligands remained within our high-scoring cutoff.

We next characterized with DFT all of the feasible mononuclear, homoleptic octahedral transition metal complexes formed from combinations of the 396 ligands in complex with eight

choices of metals and oxidation states in each of two spin states. Over the calculations that comprise the OHLDB, we obtained and analyzed nearly 350 spin-splitting energies. We observed unexpected combinations of metal/coordinating atom and spin-state ordering, including those that extended the ranges sampled in our prior databases of octahedral transition metal complexes. We showed how these complexes both reflected chemical compositions previously absent from our machine learning (i.e., artificial neural network) models for predicting spin splitting. After enriching machine learning models with OHLDB data, we showed improved machine learning model prediction performance on an out-of-sample test set consisting of transition metal complexes much larger in size. We anticipate that the OHLDB will be a good testbed both for the application of high-scaling, correlated wavefunction theory methods and for representation development for machine learning in inorganic chemistry.

ASSOCIATED CONTENT

Electronic Supplementary Information. SMILES strings and components of scores for M1, M2, and B4 ligands; details of M1, M2, and B4 ligand scoring along with properties of retained ligands; comparison of ligand scores to properties of ChEMBL, DiRef, and GDB-9 molecules; principal component analysis of OHLDB data and comparison to prior calculations; success rate of DFT calculations; spin-splitting properties of new complexes, including M(III) boxplot results; analysis of improvement in out-of-sample CSD ANN performance with retrained ANN (PDF)

Geometry optimized structures and properties of OHLDB complexes; geometry optimized structures and properties of 1,901 training complexes, geometry optimized structures and properties of 116 CSD test case complexes; ANN predictions on OHLDB structures; ANN

predictions on CSD structures before and after retraining; ANN models, ANN models, weights and scaling information before and after retraining (ZIP)

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors acknowledge support by the Office of Naval Research under grant numbers N00014-17-1-2956 and N00014-18-1-2434 and a 2017 MIT Energy Initiative seed grant. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This work was supported in part by an AAAS Marion Milligan Mason award (to H.J.K.). This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830). The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

References

1. 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. *Computational Materials Science* **2012**, *58*, 218-226.

2. 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. *Computational Materials Science* **2013**, *68*, 314-319.
3. Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J., molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *Journal of Computational Chemistry* **2016**, *37*, 2106-2117.
4. Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J., Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Industrial & Engineering Chemistry Research* **2018**, *57*, 13973-13986.
5. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An Open Chemical Toolbox. *Journal of cheminformatics* **2011**, *3*, 33.
6. Martínez, T. J., Ab Initio Reactive Computer Aided Molecular Design. *Accounts of chemical research* **2017**, *50*, 652-656.
7. Lubber, S., Recent Progress in Computational Exploration and Design of Functional Materials. *Computational Materials Science* **2019**, *161*, 127-134.
8. Caruthers, J.; Lauterbach, J. A.; Thomson, K.; Venkatasubramanian, V.; Snively, C.; Bhan, A.; Katare, S.; Oskarsdottir, G., Catalyst Design: Knowledge Extraction from High-Throughput Experimentation. *Journal of Catalysis* **2003**, *216*, 98-109.
9. Katare, S.; Caruthers, J. M.; Delgass, W. N.; Venkatasubramanian, V., An Intelligent System for Reaction Kinetic Modeling and Catalyst Design. *Industrial & Engineering Chemistry Research* **2004**, *43*, 3484-3512.
10. Corma, A.; Díaz-Cabanas, M. J.; Moliner, M.; Martínez, C., Discovery of a New Catalytically Active and Selective Zeolite (ITQ-30) by High-Throughput Synthesis Techniques. *Journal of Catalysis* **2006**, *241*, 312-318.
11. Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J., Predicting the Band Gaps of Inorganic Solids by Machine Learning. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1668-1673.
12. De, S.; Bartok, A. P.; Csanyi, G.; Ceriotti, M., Comparing Molecules and Solids across Structural and Alchemical Space. *Physical Chemistry Chemical Physics* **2016**, *18*, 13754-13769.
13. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C., A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Computational Materials* **2016**, *2*, 16028.
14. Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R., Accelerating Materials Property Predictions Using Machine Learning. *Scientific Reports* **2013**, *3*, 2810.
15. Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C., Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chemical Science* **2018**, *9*, 7069-7077.
16. Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H., Machine-Learning-Augmented Chemisorption Model for CO₂ Electroreduction Catalyst Screening. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3528-3533.
17. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chemical Science* **2017**, *8*, 5137-5152.
18. Li, Z.; Omidvar, N.; Chin, W. S.; Robb, E.; Morris, A.; Achenie, L.; Xin, H., Machine-Learning Energy Gaps of Porphyrins with Molecular Graph Representations. *The Journal of Physical Chemistry A* **2018**, *122*, 4571-4578.
19. 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. *Chemical science* **2018**, *9*, 2261-2269.
20. Behler, J., Perspective: Machine Learning Potentials for Atomistic Simulations. *The Journal of Chemical Physics* **2016**, *145*, 170901.
21. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chemical science* **2017**, *8*, 3192-3203.

22. Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E., Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Physical Review Letters* **2018**, *120*, 143001.
23. Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R., Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Science Advances* **2017**, *3*, e1603015.
24. 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. *Journal of chemical theory and computation* **2017**, *13*, 5255-5264.
25. Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C., Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE Journal* **2018**, *64*, 2311-2323.
26. Kitchin, J. R., Machine Learning in Catalysis. *Nature Catalysis* **2018**, *1*, 230.
27. 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. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1064-1071.
28. Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J., Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorganic Chemistry* **2019**, ASAP.
29. Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J., Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nature communications* **2018**, *9*, 3405.
30. Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T., Accelerated Discovery of Large Electrostrains in BaTiO₃-Based Piezoelectrics Using Active Learning. *Advanced Materials* **2018**, *30*, 1702884.
31. Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J., Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Molecular Systems Design & Engineering* **2018**, *3*, 819-825.
32. Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A., Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-Throughput Experiments. *Science advances* **2018**, *4*, eaq1566.
33. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360.
34. Zhao, Q.; Kulik, H. J., Where Does the Density Localize in the Solid State? Divergent Behavior for Hybrids and DFT+U. *Journal of Chemical Theory and Computation* **2018**, *14*, 670-683.
35. 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**.
36. 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. *Chemical Society Reviews* **2018**, *47*, 8307-8348.
37. Arockiam, P. B.; Bruneau, C.; Dixneuf, P. H., Ruthenium(II)-Catalyzed C-H Bond Activation and Functionalization. *Chem. Rev.* **2012**, *112*, 5879-5918.
38. 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.
39. Rouquet, G.; Chatani, N., Catalytic Functionalization of C(sp²)-H and C(sp³)-H Bonds by Using Bidentate Directing Groups. *Angew. Chem.-Int. Edit.* **2013**, *52*, 11726-11743.
40. Schultz, D. M.; Yoon, T. P., Solar Synthesis: Prospects in Visible Light Photocatalysis. *Science* **2014**, *343*, 985-+.

41. 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 Transactions* **2016**, *45*, 17910-17917.
42. 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.
43. Schilling, M.; Patzke, G. R.; Hutter, J.; Luber, S., Computational Investigation and Design of Cobalt Aqua Complexes for Homogeneous Water Oxidation. *The Journal of Physical Chemistry C* **2016**, *120*, 7966-7975.
44. Ashley, D. C.; Jakubikova, E., Ironing out the Photochemical and Spin-Crossover Behavior of Fe (II) Coordination Compounds with Computational Chemistry. *Coordination Chemistry Reviews* **2017**, *337*, 97-111.
45. 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. *Inorganic Chemistry* **2015**, *54*, 8786-8793.
46. 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 Percent Efficiency. *Science* **2011**, *334*, 629-634.
47. Czerwieniec, R.; Yu, J. B.; Yersin, H., Blue-Light Emission of Cu(I) Complexes and Singlet Harvesting. *Inorganic Chemistry* **2011**, *50*, 8293-8301.
48. 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. *Advanced Materials* **2013**, *25*, 3707-3714.
49. Kuttipillai, P. S.; Zhao, Y. M.; Traverse, C. J.; Staples, R. J.; Levine, B. G.; Lunt, R. R., Phosphorescent Nanocluster Light-Emitting Diodes. *Advanced Materials* **2016**, *28*, 320-326.
50. Leidl, 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. *Journal of Physical Chemistry A* **2013**, *117*, 11823-11836.
51. Linfoot, C. L.; Leidl, 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. *Inorganic Chemistry* **2014**, *53*, 10854-10861.
52. Zink, D. M.; Bachle, M.; Baumann, T.; Nieger, M.; Kuhn, M.; Wang, C.; Kloppe, 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. *Inorganic Chemistry* **2013**, *52*, 2292-2305.
53. Minenkov, Y.; Sharapa, D. I.; Cavallo, L., Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *Journal of Chemical Theory and Computation* **2018**, *14*, 3428-3439.
54. Deeth, R. J., The Ligand Field Molecular Mechanics Model and the Stereoelectronic Effects of d and s Electrons. *Coordination Chemistry Reviews* **2001**, *212*, 11-34.
55. Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W., Uff, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024-10035.
56. Janet, J. P.; Kulik, H. J., Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *Journal of Physical Chemistry A* **2017**, *121*, 8939-8954.
57. Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J., Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *Journal of Chemical Theory and Computation* **2019**, *15*, 2331-2345.

58. Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J., Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal- Oxo Intermediate Formation. *chemRxiv*: <https://doi.org/10.26434/chemrxiv.8182025.v1> 2019.
59. Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J., Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *The Journal of Chemical Physics* **2018**, *148*, 241718.
60. Huang, B.; von Lilienfeld, O. A., Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *The Journal of Chemical Physics* **2016**, *145*, 161102.
61. Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J., Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *The Journal of Physical Chemistry Letters* **2017**, *8*, 2689-2694.
62. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *Journal of Physical Chemistry Letters* **2015**, *6*, 2326-2331.
63. Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V., Machine Learning of Molecular Properties: Locality and Active Learning. *The Journal of Chemical Physics* **2018**, *148*, 241727.
64. Bjørn Jørgensen, P.; Wedel Jacobsen, K.; Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials *arXiv e-prints* [Online], 2018. <https://ui.adsabs.harvard.edu/-/abs/2018arXiv180603146B> (accessed June 01, 2018).
65. Lubbers, N.; Smith, J. S.; Barros, K., Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *The Journal of Chemical Physics* **2018**, *148*, 241715.
66. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A., Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific data* **2014**, *1*, 140022.
67. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Scientific data* **2017**, *4*, 170193.
68. 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.
69. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Raymond, J.-L., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864-2875.
70. Sterner, R. W.; Elser, J. J., *Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere*. Princeton University Press: 2002.
71. Bowen, H. J. M., *Environmental Chemistry of the Elements*. Academic Press.: 1979.
72. Fink, T.; Bruggesser, H.; Raymond, J.-L., Virtual Exploration of the Small-Molecule Chemical Universe Below 160 Daltons. *Angewandte Chemie International Edition* **2005**, *44*, 1504-1508.
73. Wester, M. J.; Pollock, S. N.; Coutsiyas, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I., Scaffold Topologies. 2. Analysis of Chemical Databases. *J. Chem. Inf. Model.* **2008**, *48*, 1311-1324.
74. Heymsfield, S. B.; Waki, M.; Kehayias, J.; Lichtman, S.; Dilmanian, F. A.; Kamen, Y.; Wang, J.; Pierson, R. N., Chemical and Elemental Analysis of Humans in Vivo Using Improved Body Composition Models. *American Journal of Physiology-Endocrinology and Metabolism* **1991**, *261*, E190-E198.
75. Jørgensen, C. K., Differences between the Four Halide Ligands, and Discussion Remarks on Trigonal-Bipyramidal Complexes, on Oxidation States, and on Diagonal Elements of One-Electron Energy. *Coordination Chemistry Reviews* **1966**, *1*, 164-178.
76. Tsuchida, R., Absorption Spectra of Co-ordination Compounds. I. *BCSJ* **1938**, *13*, 388-400.
77. Herrmann, W. A.; Krüger, C.; Goddard, R.; Bernal, I., Transition Metal Methylene Complexes: III. Methylene (CH₂), a Carbonyl Analogous Ligand; Crystal Structure of M-Methylenebis(Carbonyl-H₅-Cyclopentadienylrhodium)(Rh—Rh. *Journal of Organometallic Chemistry* **1977**, *140*, 73-89.

78. Vahrenkamp, H., Sulfur Atoms as Ligands in Metal Complexes. *Angewandte Chemie International Edition in English* **14**, 322-329.
79. Miller, J.; Balch, A. L.; Enemark, J. H., Addition of Methylamine to Hexakis(Methyl Isocyanide)Iron(II). Formation of an Unusual Chelating Ligand. *J. Am. Chem. Soc.* **1971**, *93*, 4613-4614.
80. Hughes, M. N.; Shrimanker, K., Metal Complexes of Hydroxylamine. *Inorganica Chimica Acta* **1976**, *18*, 69-76.
81. Barbaro, P.; Di Vaira, M.; Peruzzini, M.; Seniori Costantini, S.; Stoppioni, P., Hydrolysis of Dinuclear Ruthenium Complexes [CpRu(PPh₃)₂2(M,H1:1-L)][CF₃SO₃]₂ (L=P₄, P₄S₃): Simple Access to Metal Complexes of P₂H₄ and PH₂SH. *Chemistry – A European Journal* **2007**, *13*, 6682-6690.
82. Mimoun, H., Transition-Metal Peroxides—Synthesis and Use as Oxidizing Agents. In *Peroxides* (1983), Wiley-Blackwell: 2010; pp 463-482.
83. Hayton, T. W.; Legzdins, P.; Sharp, W. B., Coordination and Organometallic Chemistry of Metal–NO Complexes. *Chem. Rev.* **2002**, *102*, 935-992.
84. Shimura, Y.; Tsuchida, R., Absorption Spectra of Co(III) Complexes. II. Redetermination of the Spectrochemical Series. *BCSJ* **1956**, *29*, 311-316.
85. McKay, D. J.; Wright, J. S., How Long Can You Make an Oxygen Chain? *J. Am. Chem. Soc.* **1998**, *120*, 1003-1013.
86. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45*, D945-D954.
87. Bernath, P. F.; McLeod, S., Diref, a Database of References Associated with the Spectra of Diatomic Molecules. *J Mol Spectrosc* **2001**, *207*, 287.
88. O'Boyle, N. M.; Morley, C.; Hutchison, G. R., Pybel: A Python Wrapper for the Openbabel Cheminformatics Toolkit. *Chemistry Central Journal* **2008**, *2*, 5.
89. 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. *Inorganic chemistry* **2012**, *51*, 8340-8351.
90. Ufimtsev, I. S.; Martinez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *Journal of Chemical Theory and Computation* **2009**, *5*, 2619-2628.
91. 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. *The Journal of Physical Chemistry* **1994**, *98*, 11623-11627.
92. Becke, A. D., Density-Functional Thermochemistry. III. The Role of Exact Exchange. *Journal of Chemical Physics* **1993**, *98*, 5648-5652.
93. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Physical Review B* **1988**, *37*, 785--789.
94. Vosko, S. H.; Wilk, L.; Nusair, M., Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Canadian Journal of Physics* **1980**, *58*, 1200-1211.
95. Hay, P. J.; Wadt, W. R., Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for the Transition Metal Atoms Sc to Hg. *The Journal of Chemical Physics* **1985**, *82*, 270-283.
96. 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. *Industrial & Engineering Chemistry Research* **2017**, *56*, 4898-4910.
97. Mortensen, S. R.; Kepp, K. P., Spin Propensities of Octahedral Complexes from Density Functional Theory. *The Journal of Physical Chemistry A* **2015**, *119*, 4041-4050.
98. Saunders, V. R.; Hillier, I. H., Level-Shifting Method for Converging Closed Shell Hartree-Fock Wave-Functions. *Int J Quantum Chem* **1973**, *7*, 699-705.

99. Wang, L.-P.; Song, C., Geometry Optimization Made Simple with Translation and Rotation Coordinates. *The Journal of Chemical Physics* **2016**, *144*, 214108.
100. 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.
101. Ioannidis, E. I.; Kulik, H. J., Towards Quantifying the Role of Exact Exchange in Predictions of Transition Metal Complex Properties. *The Journal of Chemical Physics* **2015**, *143*, 034104.
102. 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.
103. Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N., Density Functional Theory in Transition-Metal Chemistry: A Self-Consistent Hubbard U Approach. *Phys. Rev. Lett.* **2006**, *97*, 103001.
104. 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')]. *The Journal of Chemical Physics* **2005**, *122*, 234321.
105. Droghetti, A.; Alfè, D.; Sanvito, S., Assessment of Density Functional Theory for Iron (II) Molecules across the Spin-Crossover Transition. *The Journal of chemical physics* **2012**, *137*, 124303.
106. Verma, P.; Varga, Z.; Klein, J. E.; Cramer, C. J.; Que, L.; Truhlar, D. G., Assessment of Electronic Structure Methods for the Determination of the Ground Spin States of Fe (II), Fe (III) and Fe (IV) Complexes. *Physical Chemistry Chemical Physics* **2017**, *19*, 13049-13069.
107. 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. *The journal of physical chemistry letters* **2017**, *8*, 2026-2030.
108. Phung, Q. M.; Feldt, M.; Harvey, J. N.; Pierloot, K., Toward Highly Accurate Spin State Energetics in First-Row Transition Metal Complexes: A Combined CASPT2/Cc Approach. *Journal of chemical theory and computation* **2018**, *14*, 2446-2455.
109. Zhou, C.; Gagliardi, L.; Truhlar, D. G., Multiconfiguration Pair-Density Functional Theory for Iron Porphyrin with CAS, RAS, and DMRG Active Spaces. *The Journal of Physical Chemistry A* **2019**.
110. Kim, M. C.; Sim, E.; Burke, K., Communication: Avoiding Unbound Anions in Density Functional Calculations. *J Chem Phys* **2011**, *134*, 171103.
111. Jensen, F., Describing Anions by Density Functional Theory: Fractional Electron Affinity. *J Chem Theory Comput* **2010**, *6*, 2726-35.
112. Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J., A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *chemrxiv* <https://doi.org/10.26434/chemrxiv.7900277.v2> **2019**.
113. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72*, 171-179.