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

Stefan Gugler, i Jon Paul Janet i and Heather J. Kulik i *

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 2500 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 1250 homoleptic octahedral complexes obtained from the enumerated pool of ligands and an open-shell transition metal (M(I)/M(II), M = Cr, Mn, Fe, or Co) and ii) the resulting high-spin/low-spin adiabatic electronic energy differences (ΔEHS/LS) obtained with hybrid DFT. Over the OHLDB, we observe structure-property (i.e., ΔEHS/LS) 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.

Design, System, Application

We develop a strategy for octahedral transition metal complex design by enumerating potential ligands from scratch. Such complexes are relevant as molecular sensors and switches that can change their ground state spin in response to light, heat, or other stimuli. They are also models of catalytic active sites. We show how our de novo enumeration both produces complexes with novel electronic structure and provides training data that improves the baseline performance of our machine learning models (here, artificial neural networks) in out-of-sample tests.

1. Introduction

Computational, first-principles (i.e., with density functional theory, 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 process4,11–18 throughout chemistry,19–24 including for catalysis15,16,25,26 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 states, oxidation states, and coordination numbers feasible for each metal. The large sizes 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 developed3,4,17,27,28,56,57 ML models for
predicting quantum mechanical properties (here, computed with DFT), including spin-splitting energies, redox or ionization potentials, metal-ligand bond lengths, frontier orbital energies, and reaction energetics. 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 on large (>100k points) data sets 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 wavefunctions have limited data set sizes for inorganic chemistry. Despite these limitations in data set size, ML models for inorganic chemistry on modest data sets are predictive to 1–3 kcal mol$^{-1}$ as judged by test set errors. 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$^{-1}$) 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 molecules 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 engineering or systematic feature selection, 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 were primarily 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 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 (Fig. 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 (Fig. 1). We penalized but did not exclude ligands that do not have octet valence or neutral charge, diverging from previous organic enumeration efforts 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 noninnocence. Ligand scores were based on heuristic properties, i.e.: 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.

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 (\(s_{\text{charge}}\)), steric (i.e., presence of H atoms, \(s_{\text{sterics}}\)), and according to the octet rule (\(s_{\text{octet}}\)). We favored neutral and weakly negative ligands (\(s_{\text{charge}} = 0\) for \(-2 \leq c \leq 0\)) over positively charged ligands (\(s_{\text{octet}} = 4 - |8 - v|\), a total score, \(s_{\text{tot}}\), for M1 ligands was given by:

\[
s_{\text{tot}} = s_{\text{charge}} + s_{\text{sterics}} + s_{\text{octet}}
\]

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

Common M1-type ligands in the spectrochemical series include methylene (\(\text{CH}_2\)) and elemental sulfur (\(\text{S}\)). Still lower scores (i.e., \(s_{\text{tot}} = 7\)) arose primarily due to penalties for: i) steric repulsion, as in methane (\(\text{CH}_4\)) or ii) charge, as in sulfonium (\(\text{SH}_3\)) and hydronium (\(\text{OH}_2^+\)) (Fig. 3). The lowest scoring ligand (\(s_{\text{tot}} = 3\), \(\text{OH}_2^+\), simultaneously violates steric and octet rules (Fig. 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† Table S1).

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 5625, 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 \( c_{\text{tot}} \) value under constraint that the valence, \( v_i \), of each \( i \)th atom was satisfied:

\[
v_i = v_{\text{e}} - c_i - 2l_i - h_i \tag{3}
\]

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 \( v_{\text{e}} \) 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:

\[
b = \min (v_1, v_2) \tag{4}
\]

with an allowable range of 0 < \( b \) ≤ 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 \( v_1 = v_2 = 2 \) 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 1171 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 Fig. 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 \( h = 3 \)). The M2 \( s_{\text{octet}} \) score was applied over atom 1 and 2 \( v \) values (Fig. 2):

\[
s_{\text{octet}} = 5 - |v_1 - v_2| \tag{5}
\]

where the absolute difference of \( v_1 \) and \( v_2 \) ranged from 0 to 4 over the retained ligands. This produced a practical \( s_{\text{octet}} \) of 1 to 5 (Fig. 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, \( s_{\text{bond}} \) and ii) charge polarization, \( s_{\text{pol}} \). 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 (Fig. 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 Fig. 2). A total M2 ligand score was thus:

\[
s_{\text{tot}} = s_{\text{charge}} + s_{\text{sterics}} + s_{\text{octet}} + s_{\text{bond}} + s_{\text{pol}} \tag{6}
\]

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

The 55 highest scoring ligands (\( s_{\text{tot}} = 17 \)) include the common metal-complexing ligands methylamine (\( \text{NH}_2\text{CH}_3 \)),\(^79\) hydroperoxide (\( \text{H}_2\text{O}_2 \)), and both hydroxylamine (\( \text{NH}_2\text{OH} \))\(^80\) and its experimentally-observed analogue, thiophosphinous acid\(^81\) (\( \text{PH}_2\text{SH} \)), (Fig. 3). Both reactive peroxide\(^82\) (\( \text{O}_2 \)) and nitrosyl\(^83\) (\( \text{NO}^\text{−} \)) score highly as do stable small molecules such as \( \text{N}_2 \) and HCN (\( s_{\text{tot}} = 16 \)) and methanol (\( \text{CH}_3\text{OH} \)) (O-coordinating \( s_{\text{tot}} = 17 \), C-coordinating \( s_{\text{tot}} = 14 \), see Fig. 3). Most M2-type ligands in the spectrochemical series\(^84\) have high scores (e.g., \( \text{CN}^- \), \( s_{\text{tot}} = 15 \) and \( \text{CO} \), \( s_{\text{tot}} = 16 \)), with \( \text{NO}^\text{−} \) (\( s_{\text{tot}} = 13 \)) scoring the lowest due to its positive charge. Other \( s_{\text{tot}} = 13 \) ligands include sterically hindered ligands (e.g., \( \text{NH}_3\text{NH}_3^− \)) or likely unstable, positively charged species, such as \( \text{CCH}^- \) (Fig. 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., \( \text{NH}_3\text{C}^\text{++} \); \( s_{\text{octet}} = 1 \), \( s_{\text{bond}} = 2 \), all other scores are zero, see Fig. 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† Fig. S2).

### 2.3 B4 ligands

We generated symmetric bidentate B4 ligands solely by joining two identical M2 ligands from the original pool of 5625 theoretical M2 ligands, and so there were also 5625 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 ligand, 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 \( b_{1'-2'} \), Fig. 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 \( c_{\text{tot}} \) 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 \( b_{1-2} \) and \( b_{2-2} \) bond orders simultaneously by iterating over allowed values between single and triple bonds:

\[
\{ b_{1-2}, b_{2-2} \} = \arg\min \| v_1 - b_{1-2} + v_2 - b_{2-2} \| \tag{7}
\]

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 \( b_{1-2} \) and \( b_{2-2} \) 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 \( \text{NH}_2\text{CH} \) has \( b_{1-2} = 2 \) with \( c_1 = -1 \) and \( c_2 = 1 \). The resulting B4 ligand formed from two of these M2 ligands is \( \text{NH}_2\text{CH} = \text{CH}-\text{NH}_2 \), 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 1356 B4 ligands suitable for further scoring with a five-component score similar to that applied to the M2 ligands. Both \( s_{\text{charge}} \) and \( s_{\text{series}} \) 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 (Fig. 2). The B4 \( s_{\text{octet}} \) was also scored only for a single building block using eqn (5) but using the revised valency after charge redistribution (Fig. 2). After redistribution of charges, we scored a single 1-2 pair, penalizing strong polarization \( s_{\text{polar}} = 0 \) for \( |c_1| + |c_2| > 1 \), favoring completely neutral atomic charges \( s_{\text{polar}} = 2 \) for \( |c_1| + |c_2| = 0 \), and assigning an intermediate score \( s_{\text{polar}} = 1 \) for slight polarization (i.e., \( |c_1| + |c_2| = 1 \), see Fig. 2). We also computed \( s_{\text{bond}} \) 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 Fig. 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 (Fig. 1). Ethylenediamine (en, \( \text{C}_2\text{H}_4\text{N}_2 \)) is the only B4-type ligand in the spectrochemical series, and it has a maximum score \( s_{\text{tot}} = 16 \) as does its phosphorus analogue, 1,2-ethanediylphosphine \( (\text{C}_2\text{H}_8\text{P}_2, \ s_{\text{tot}} = 16, \text{Fig. 3} \text{ and ESI† Table S4}) \). Other common ligands that score highly include ethylene glycol \( (\text{C}_2\text{H}_4\text{O}_2, \ s_{\text{tot}} = 16) \) 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 (Fig. 3). Stable organic molecules that are sterically hindered and would make poor ligands, e.g., butane \( (\text{C}_4\text{H}_{10}, \ s_{\text{tot}} = 13) \), score more poorly than unstable but likely metal-coordinating molecules, e.g., tetraoxide \( (\text{H}_2\text{O}_4, \ s_{\text{tot}} = 16, \text{Fig. 3}) \). 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., \( \text{Sn}_4 \), or \( \text{HOC}_2\text{OH} \), Fig. 3). The lowest scoring ligands are charged, have polarized bonds, and are octet violating (e.g., \( \text{CCH}_2\text{CH}_3\text{C}^\text{+} \), \( s_{\text{tot}} = 2 \), Fig. 3). Following these holistic observations, we retained only the 47 B4 ligands with \( s_{\text{tot}} \geq 14 \) for subsequent DFT calculations (ESI† Fig. S3).

2.4. Overall ligand analysis

From an original combinatorial space of 11325 theoretical M1, M2, and B4 ligands, we scored 2577 ligands and identified the top 20% (570 ligands) as good candidates for DFT characterization (Fig. 1 and see sec. 4). The other 8748 ligands were eliminated prior to scoring due to disqualifications ranging from unpaired electrons on the ligand, zero bond order between heavy atoms or unsuitable bond order for bidentate coordination, high net positive charge that would prevent coordinating a positively charged metal, or strong steric repulsion from a high number of hydrogen atoms on the metal-coordinating atom that would prevent metal coordination. All such excluded ligands are provided in the ESI† and can be interpreted as scoring zero in comparison to the retained ligands that all score higher. 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: ChEBI,86 the generated database of enumerated organic molecules with up to 9 heavy atoms, GDB-9,69 and a database of experimentally and computationally characterized diatomic molecules, DiRef.87 Across these databases, 71 of the 2577 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, \( s_{\text{scaled}} \), between the minimum and maximum observed \( s_{\text{tot}} \) values within each ligand type:

\[
s_{\text{scaled}} = \frac{s_{\text{min}} - s_{\text{tot}}}{s_{\text{tot}} - s_{\text{max}}} \times 100 \tag{8}
\]
The scaled threshold for retention is lowest for M1 ligands (>71%) and highest for the B4 cases (>86%, see Fig. 4). Over all ligands, 75% (53 of 71 overall; 36 of 48 unique) of compounds found in DBs were above the relevant ligand cutoff (Fig. 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* ES† Table S5). Although such ligands are relevant for heteroleptic transition metal complexes, our focus on homoleptic octahedral complexes with high valent metals motivated penalizing positively charged ligands (see Fig. 2). Our restriction to octahedral complexes is motivated by their relevance in catalysis and functional materials, but study of heteroleptic or lower coordination number complexes in future work will motivate alternate scoring. The remaining database ligands below threshold typically exhibit steric hindrance (e.g., M1: CH4 or B4: C6H10, see Fig. 4). Conversely, known good inorganic ligands such as M1 ammonia, M2 methyamine, or B4 ethylenediamine are all among the top scoring database ligands (Fig. 4 and ES† Table S5).

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.

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 results and ML models17,27,56 (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 methods,107–109 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 generation. The molSimplify3 code uses OpenBabel27,88 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, \( \Delta E_{\text{H-L}} \), was computed as the electronic energy difference between the two geometry-optimized states (i.e., the adiabatic energy difference). This choice is motivated by the fact that spin-state change is slower than other processes (e.g., optical excitations) for which a vertical energy evaluation may be more appropriate. The high-spin/low-spin definitions for the metals studied in this work are: quintet-singlet for d5 Co(II)/Fe(II), sextet-doublet for d5 Fe(III)/Mn(III), quintet-singlet for d5 Mn(III)/Cr(III), and quartet-doublet for both d5 Cr(III) and d5 Co(III). 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 effects,96,97 and we therefore focused on relationships between ligand identity and DFT adiabatic, electronic \( \Delta E_{\text{H-L}} \) energies. All open-shell calculations (i.e., non-singlet spin states) were carried out with level shifting98 using spin-up and spin-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 \( \times \) 10\(^{-4}\) Hartree/Bohr for the maximum gradient and 1 \( \times \) 10\(^{-6}\) Hartree for the change in energy between steps.

3. Computational details

Homoleptic, mononuclear octahedral transition metal complexes were generated from de novo ligands (sec. 2) with the molSimplify toolkit and molSimplify Automatic Design (mAD),4,27 which automated both structure and input file generation.
(i.e., −9 or −10) of the homoleptic octahedral complex that cannot be treated well within approximate DFT. For the remaining 396 ligands, 16 combinations of metal, oxidation, and spin state mean that 6336 geometry optimizations were carried out for 3168 possible $\Delta E_{II}$ evaluations (see sec. 3).

To streamline and improve the quality of data ingested during high-throughput simulation of transition metal complexes, we recently introduced 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 $mAD^3$ 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 6336 initial geometry optimizations, 22% (1387) 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 correspond to those that failed geometry checks initially (214 or 3%), during resubmission (3816 or 60%), or on the fully optimized structure (919 or 15%). Such ligand detachment or strong asymmetry in metal–ligand bond lengths can be attributed to Jahn–Teller distortion, which in extreme cases would lead to unstable transition metal complexes. Of all excluded calculations, 1537 exhibit strong bond asymmetry and 1037 exhibit ligand detachment, although many of these calculations had at least one other failure mode as well. 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 1387 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† Fig. 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 (Fig. 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 average score, M2 15.8 vs. 15.4 average score, and B4 15.7 vs. 15.4 average score. Because the applied geometry check penalized strong metal–ligand distortions or ligand detachment, 97% (1347) of optimized structures have metal–ligand bond asymmetries (i.e., the difference between maximum and minimum metal–ligand bond lengths) below 0.4 Å and 80% (1137) have differences below 0.2 Å (structures provided in the ESI†).

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 1387 converged geometries, we also discarded 134 for $\Delta S^2$ 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 $\Delta E_{II}$ evaluation requires two successful geometry optimizations (the high-spin and low-spin states) of a given octahedral complex, a total of 343 new $\Delta E_{II}$ evaluations were obtained for 106 unique ligands (see ESI† for all energies and structures).

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 (Fig. 5 and ESI† Fig. 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 (Fig. 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 (Fig. 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, e.g., NH$_3$, PH$_3$, H$_2$S, and H$_2$O. One exception was a Co(iii) complex of the ammonia analogue, CH$_3$NH$^+$. (Fig. 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 (Fig. 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 (Fig. 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 1901 ΔE₁⁺− values we had previously obtained for artificial neural network (ANN) model training. To compare diversity in the chemical structures, we featurized each new complex with the revised autocorrelation (RAC-155) representation (ESI† Text S1). The RAC-155 representation consists of products and differences of heuristic properties on the molecular graph and has shown good performance for predicting inorganic chemistry properties, including ΔE₁⁺−. 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 (Fig. 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%, Fig. 6 and ESI† Fig. S5).

An alternative measure of data diversity is in property space, which we assessed first by determining if a previously trained RAC-155/ANN model could have predicted the ΔE₁⁺− values exhibited by the OHLDB complexes (Fig. 6 and ESI† Table S6). Overall, although a large number of complexes were well predicted, significant (e.g., >60 kcal mol⁻¹) over- and underestimations of ΔE₁⁺− are indicative of limited prior knowledge by the ANN (mean absolute error, MAE = 14.3 kcal mol⁻¹) of the chemistry of the OHLDB complexes (Fig. 6 and ESI† Table S6). Indeed, high error points are both chemically distinct and exhibit unexpected spin-state order-
Unsaturated carbon-coordinating ligands (e.g., CCH₂ in Mn(II)(CCH₂)₂, \( \Delta E_{H\leftrightarrow L} = 72.6 \text{ kcal mol}^{-1} \)) are known to be low-spin directing ligands but most were absent from our earlier data set, as were more unusual low-spin-directing ligands such as CHOH or CHNH₂ (Mn(II) \( \Delta E_{H\leftrightarrow L} = 47 \) to 62 kcal mol\(^{-1}\), Fig. 7). The sulfur-coordinated Fe(II) complexes in our new data set span from low-spin (e.g., monodentate SHOH: \( \Delta E_{H\leftrightarrow L} = 11.0 \text{ kcal mol}^{-1} \)) or bidentate SC₂H₂S: \( \Delta E_{H\leftrightarrow L} = 10.3 \text{ kcal mol}^{-1} \)) to high-spin (e.g., SC: \( \Delta E_{H\leftrightarrow L} = -41.1 \text{ kcal mol}^{-1} \)), corresponding to a range that we had not observed in prior Fe(II) data (Fig. 7). Since sulfur is considered a soft element, low-spin sulfur complexes are somewhat surprising. Examination of the OHLDDB confirms that the bidentate SC₂H₂S ligand also forms low-spin complexes with Fe(II) and Co(II) but forms high-spin complexes with Cr(II) and Mn(II) (see ESI†). Saturating the sulfur (i.e., SHC₂H₄SH) 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 OHLDDB, in addition to those previously observed, unexpectedly high-spin complexes e.g., Co(II)(OHP₂-H₂OH)₃ (\( \Delta E_{H\leftrightarrow L} = -22.8 \text{ kcal mol}^{-1} \), Fig. 7). This bidentate ligand and the isoelectronic OH₃O ligand produce among the most high-spin-favoring Cr(II)/O-coordinating complexes (\( \Delta E_{H\leftrightarrow L} \approx 49 \) to 52 kcal mol\(^{-1}\), see ESI†).

The OHLDDB also enables examination of how isovalent and isoelectronic variations in ligands alter spin-state ordering (Fig. 8). Here, we focus on the widely-studied CO ligand and related isovalent and isoelectronic species, including those in the spectrochemical series²⁶ (e.g., CN⁻) and other common molecules (e.g., HCN and N₂, Fig. 8). Given ligand definitions, \( \Delta E_{H\leftrightarrow L} \) can in principle be obtained for either orientation of asymmetric M₂ ligands, but only 14 of these ligands in practice yielded at least one \( \Delta E_{H\leftrightarrow L} \) value for any metal or oxidation state (Fig. 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 (Fig. 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 (Fig. 8). Although CO is often invoked as one of the strongest field ligands (\( \Delta E_{H\leftrightarrow L} = 20-30 \text{ kcal mol}^{-1} \) 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⁻, CNH₂ or CS, Fig. 8). The trends observed for iron complexes generally hold for other metals, with Co(II) complexes...
exhibiting more low-spin bias for the same ligands, Mn(II) complexes remaining uniformly high spin, and all other metals and oxidation states generally residing within these two bounds (Fig. 8). Thus, a very wide range of $\Delta E_{H-L}$ values (ca. $-60$ to $+80$ kcal mol$^{-1}$) can be obtained simply by adjusting the charge and elemental identities in M2-type ligands isoelectronic or isovalent to a common ligand.

Finally, we considered the extent to which OHLDB data could be used to improve ML model predictions on large, diverse complexes$^{112}$ by improving the chemical coverage of the ML model training data. We recently curated$^{112}$ a 116 complex out-of-sample test set from the Cambridge Structural Database (CSD)$^{113}$ for testing $\Delta E_{H-L}$ predictions with a RAC-155/ANN model. Because the CSD complexes were chosen to be distinct from the 1901 complexes used in the training of the ANN, the CSD set $\Delta E_{H-L}$ MAE of 8.6 kcal mol$^{-1}$ was much poorer than set-aside test set errors (ca. $1-3$ kcal mol$^{-1}$ (ref. 17 and 56)) or uncertainty-controlled, out-of-sample prediction errors (ca. 4.5 kcal mol (ref. 27)). Notably, very high $\Delta E_{H-L}$ prediction errors, either due to over or underestimation, were observed on the order of 20–50 kcal mol$^{-1}$ (Fig. 9).

Incorporating OHLDB data and retraining the RAC-155/ANN eliminated many of these highest error points and reduced CSD set MAE to 6.7 kcal mol$^{-1}$ (Fig. 9, ESI Text S1, Table S6, and Fig. S8 and 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, Fig. 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, Fig. 9 and ESI† Fig. S9 and S10). Given that most of the CSD curated set$^{112}$ is multideterminate 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.

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 2500-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 2500) 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, homoletic octahedral transition metal complexes formed from combinations of the 396 ligands in complex with a choice of eight metal/oxidation state combinations 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 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 both in spin-splitting energy predictions and beyond.

Fig. 9 Swarm plot of RAC-155/ANN signed errors (in kcal mol$^{-1}$) on an out-of-sample 116 CSD structure data set$^{112}$ (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.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

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


