Discovery of Hemilabile Ligands Using Machine Learning

Ilia Kevlishvili[1], Chenru Duan[1,2], and Heather J. Kulik[1,2]*

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

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

ABSTRACT: Discovery of hemilabile ligands that optimally balance reactivity and stability is important for identifying novel catalyst structures. We design a workflow for identifying ligands in the Cambridge Structural Database (CSD) that have been crystalized with distinct denticities and are thus identifiable as hemilabile ligands. To overcome the difficulty of identifying negative example, non-hemilabile ligands in our data set, we implement a semi-supervised learning approach using a label-spreading algorithm together with a set of heuristic rules based on ligand frequency of appearance. We show that a heuristic based on coordinating atom identity alone is not sufficient to identify whether a ligand is hemilabile and our trained machine-learning classification models are instead needed to predict whether a bi-, tri-, or tetradentate ligand is hemilabile with high accuracy and precision. We gain deeper insight into the factors that govern ligand hemilability by conducting feature importance analysis on our models, finding that the second, third, and fourth coordination spheres all play an important role in ligand hemilability.

Ligands that can change the metal coordination environment, i.e., hemilabile ligands, are often able to address the tradeoff between catalyst activity and stability[1,2] because they can ligate and protect the transition metal while occasionally disengaging and making the catalyst site amenable for a reaction to take place. Hemilabile ligands have been used to address major challenges in organic chemistry, such as reactivity selectivity tradeoffs in enantioselective[3-6], regioselective[7], and chemoselective[8] catalysis. While hemilabile ligands have been primarily used in homogeneous catalysis, their unique properties have also been utilized in nanoparticle[9], single atom[10], and heterogeneous catalysis[11,12] in recent years.

Normally, new reaction design or reactivity improvement involves screening a large number of ligands[13-15] or costly computational mechanistic studies[16-19]. While ligand hemilability is often used as a design principle[6,18,20], these principles mainly rely on a set of heuristic rules such as distinct donor properties of coordinating atoms[21], linker flexibility between coordination atoms[20], and steric crowding[6] near the transition metal. Determining ligand hemilability in solution usually requires indirect kinetic measurements of reaction rates[22-24], trapping distinct complexes in crystal structures[25], or time-consuming computational mechanistic studies[5]. Most commonly, the design of ligands involves the trial-and-error changing and mismatching of the donor properties of coordinating atoms[21,24] in an attempt to bias ligands towards hemilability. However, symmetric, homo-functional ligands have also been shown to undergo hemilabile coordination changes[5,26,27], while some multifunctional ligands do not tend to change coordination environments[20]. Therefore, being able to tell *a priori* whether a ligand can act as a hemilabile ligand would greatly accelerate screening efforts.

Here, we employed a data-driven approach to identify factors that determine the likelihood of a ligand to be hemilabile. We curated a dataset of ligands from the Cambridge Structural

Database (CSD). We separately identified candidate hemilabile ligands from this CSD set and then used counting rules to identify non-hemilabile ligands. We trained machine learning (ML) procedures to predict ligand hemilability and used feature analysis of the trained models to show why common heuristic rules can struggle to fully account for a hemilabile character. This ML model allowed us to further expand and suggest a set of candidate hemilabile ligands from existing (i.e., in the CSD) and thus synthesizable ligands.

We first curated a dataset of hemilabile ligands from a set of all ligands that appear in mononuclear transition metal complex (TMC) crystal structures[28 29]. We identified 4,144 ligands that appear in mononuclear TMCs with different denticities, with their highest denticity conformations ranging from bidentate to nonadentate, based on the molecular graph determinants of each ligand bound to a dummy transition metal[30], as well as those of ligands with transition metal absent (Supporting Information Text S1). A ligand was labeled hemilabile if the molecular graph determinant of a ligand without the transition metal mapped to more than one molecular graph determinant of a ligand bound to a transition metal, indicating a change in the coordination environment (Figure 1, inset). For this set of ligands, we identified and separated them into distinct subsets based on the highest denticity conformation of the ligand. Since bi-, tri-, and tetradentate ligands are most widely used for catalysis and were the most common ligand types, only these ligands were studied further (Figure 1 and Supporting Information Figures S1–S2). To focus on catalytically relevant ligand types and remove trivial cases arising from agostic interactions[31] such as those with hydrogen, we eliminated any ligands where the coordinating atoms did not consist of carbon, nitrogen, oxygen, phosphorus, or sulfur (Supporting Information Table S1). Finally, any ligands with a high absolute charge, q, assignment (i.e., $|q| > 4$) were eliminated in order to remove ligands derived from alternately either highly charged or poorly resolved (i.e., missing

hydrogen atoms) complexes (Supporting Information Table S1). After each of these steps, we obtained a set of 1,531 hemilabile bidentate, 1,069 tridentate, and 492 tetradentate ligands, where we group ligands by their highest denticity observed in a transition metal complex.

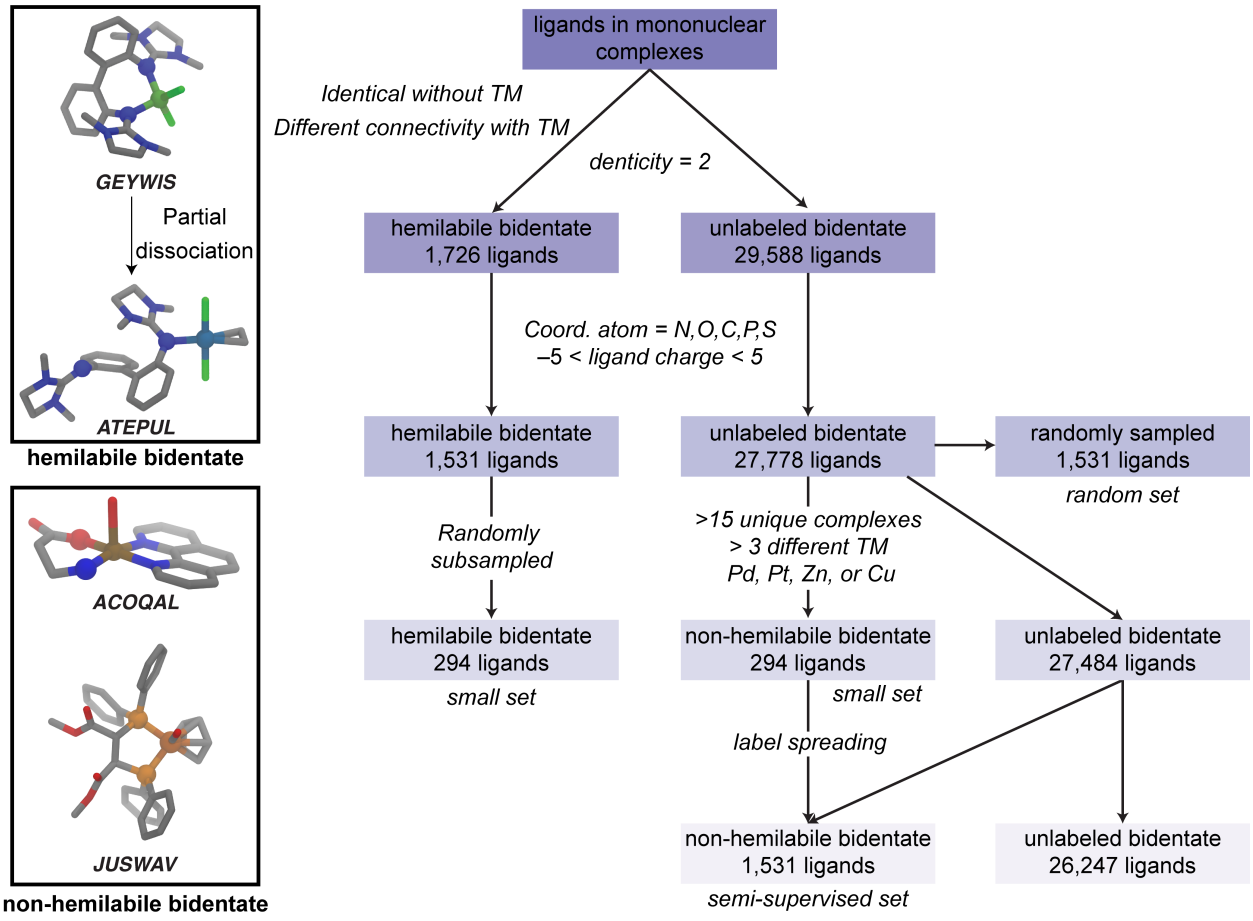

Figure 1. Description of the data curation workflow and filtering steps for defining bidentate hemilabile and non-hemilabile sets. Examples of a hemilabile ligand in high and low denticity conformation are shown as a top inset, along with refcodes (**GEYWIS** – Ni, **ATEPUL** – Pd, **ACOQAL** – Cu, **JUSWAV** – Fe) associated with a representative complex involving these ligands, where ligating atoms of hemilabile ligand and the transition metal are shown as spheres. Examples of non-hemilabile ligands are shown in the bottom inset. Hydrogens are omitted for clarity. Atoms in the insets are colored as follows: C in gray, O in red, N in blue, Ni in dark green, Cl in green, Pd in light blue, Cu in brown, P in light orange, Fe in dark orange.

To gain more insight into how transition metal elemental identity affects hemilabile ligand conformations, we analyzed how frequently each of the 4,144 hemilabile ligands appear with the ten most common transition metals both at their lowest and highest denticity conformations

(Supporting Information Figure S3). We find that while most transition metals tend to favor a higher denticity conformation of a known-hemilabile ligand, ligands that appear in complexes with palladium, platinum, and zinc tend to prefer a lower denticity conformation. We also find that the most balanced occurrence of ligands in both the lower and higher denticities occurred for copper-containing complexes. The majority (55%) of hemilabile ligands appear in both high and low denticities in at least two distinct complexes with the same transition metal center, nevertheless meaning that a significant number of ligands only appear in different denticity conformations when the identity of the transition metal is changed (Supporting Information Table S2).

To ensure we avoid introducing bias in our hemilabile ligand dataset by including ligands that strongly prefer either high or low denticity conformations, we evaluated how many times they appear in each denticity in the unique complexes of these ligands (Supporting Information Figures S4–S6). While different binding conformation changes for hemilabile tridentate (e.g., to monodentate as well) and tetradentate ligands (e.g., to bidentate) are possible, here, we only define two classes, i.e., hemilabile and non-hemilabile, to ensure sufficient dataset sizes (Supporting Information Figures S7–S8). There is a wide distribution of the hemilabile ligands occurring in the low denticity configuration relative to the total occurrences that is nevertheless centered around 0.5 (i.e., both low and high denticity are equally weighted). For the majority of ligands (i.e., 75.8% bidentate, 75.4% tridentate, 76.8% tetradentate), the ratio of low denticity to total count is between 0.2 and 0.8, indicating that these hemilabile ligands appear in higher and lower conformations with similar frequency. Focusing on the ligands that strongly prefer either high or low conformations (i.e., the ratio of low denticity to total count is < 0.01 or > 0.99), only a very minor fraction of ligands (i.e., 1.2% bidentate, 0.6% tridentate, 0.8% tetradentate) fall into this category. Thus, most

ligands in our hemilabile set can be expected to sample both denticities based on their occurrence in crystal structures.

Using this dataset of hemilabile ligands, we next devised a strategy to train machine learning (ML) models that could predict the likelihood of a ligand to exhibit hemilability for bidentate ligands. To train such a model, we require not just the hemilabile ligand dataset but also a set of negative, non-hemilabile ligands. Although positive examples of hemilability are identifiable based on the presence of complexes with ligands in multiple denticities, the absence of multiple denticities for ligands across complexes could be due to a lack of prior synthesis of diverse complexes containing a given ligand. To address this issue, we defined three different non-hemilabile sets, which were constructed by i) randomly subsampling all unlabeled ligands to obtain an equal sample size of hemilabile and non-hemilabile ligands (random dataset), ii) using frequency rules to identify a small set of non-hemilabile ligands and randomly subsampling hemilabile ligands to equalize class sizes (small dataset), and iii) using semi-supervised learning strategy that started from the small non-hemilabile dataset and employed machine learning to identify more non-hemilabile ligands (semi-supervised dataset, Figure 1 and Supporting Information text S2). Semi-supervised learning[32] encompasses a broad set of techniques that combine aspects of supervised (i.e., with labeled data) and unsupervised (i.e., with unlabeled data) learning approaches to address the challenges of only partially labeled datasets. Using this approach enables constructing a model that could benefit from the large size of the known-hemilabile set while preserving good labels for non-hemilabile ligands We used these datasets to train a classification model using the extreme gradient boosting algorithm[33] (XGBoost) for the prediction of hemilability (Figure 1). To featurize ligands, we used using ligand-based revised autocorrelations (RACs)[34], which are connectivity-based representations that have been

successfully applied to transition metal complex property prediction[34-37] (see Computational Details).

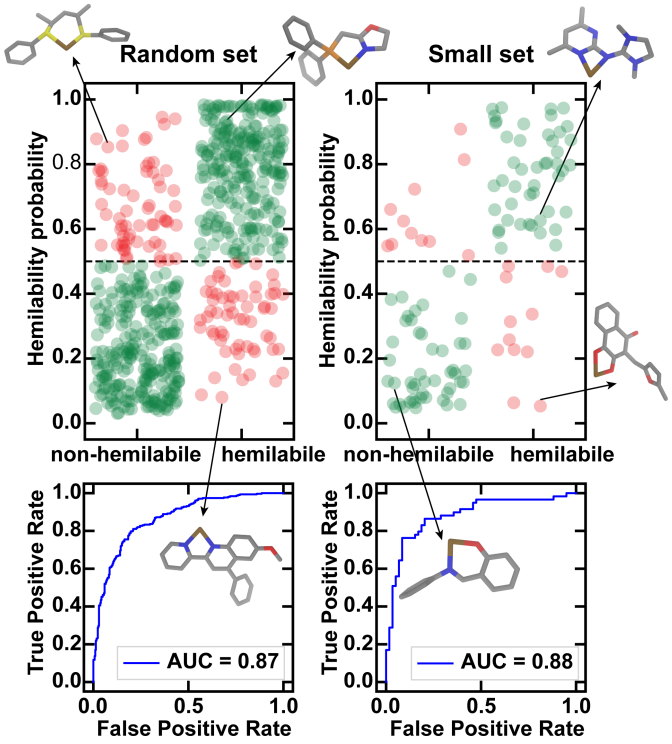

Figure 2. ML classifier (i.e., XGBoost) prediction probability (top) and ROC for random (left), small (middle), and semi-supervised (right) datasets. All data points are represented as translucent circles to depict data density and colored by classification correctness: correct (green) and incorrect (red). Examples of correct and incorrect classifications of ligands are shown as insets, bound to a metal. Hydrogens are omitted for clarity. Atoms are colored as C in gray, N in blue, O in red, P in orange, S in sulfur, and metal in brown.

The model trained on the random dataset shows promising performance on a set-aside test set with good separation between two classes with a receiver-operating characteristic area under the curve (ROC-AUC) of 0.86 as well as good accuracy (0.80) and recall (0.80), despite our expectation of potential label contamination due to likely incorrectly assigned negative labels (Figure 2 and Supporting Information Figures S9–S10 and Table S3). Despite a significantly diminished training set size of the small dataset (i.e., an 80% training partition of 3,062 ligands

versus of 588 ligands in the small set), the model trained on this dataset still shows a slight improvement over the randomly sampled set in the predictive power on the test set, including an improved ROC-AUC of 0.88, accuracy of 0.81, and a comparable recall of 0.80 (Figure 2 and Supporting Information Figures S11–S12 and Table S3). Finally, we trained an ML model classifier (i.e., again XGBoost) on the original set of hemilabile ligands along with the balanced class of new non-hemilabile ligands and preserved the initial test set that was also used during semi-supervised learning. This model shows by far the best overall performance, with a marked improvement that includes an ROC-AUC of 0.96, accuracy of 0.90, and recall of 0.89 (Supporting Information Figures S18–S19 and Table S3). In order to test the limits of this encouraging performance, we carried out a more stringent test of a grouped split in which we nearly eliminated specific coordinating atom elements from the training set. Specifically, we removed 90% of the ligands that contained at least one phosphorus atom as a coordinating atom from the training set. This split largely preserves our label balance (i.e., 51:49 hemilabile:non-hemilabile in training and 45:55 in the test set). Although performance is expectedly reduced, this model still shows encouragingly good performance, including an ROC-AUC of 0.94, accuracy of 0.87, and a recall of 0.84 (Supporting Information Figures S20–S22 and Table S3).

Given the good performance we observed on bidentate hemilabile ligands, we repeated our analysis and ML model training for tri- and tetradentate ligands. Given the somewhat smaller dataset sizes, we reduced the requirement for the number of unique complexes to confidently label negative examples (Supporting Information Figures S1–S2). For both tridentates and tetradentates, we trained XGBoost ML models using all three protocols we demonstrated on the bidentate set. The XGBoost ML models trained to predict the hemilability of tridentate ligands on the randomly selected set (2,138 ligands total) show relatively poor performance, with an ROC-AUC of 0.72

and accuracy of 0.67, but this is either improved by using the small set (ROC-AUC of 0.79 and accuracy of 0.73, 354 ligands total) or even more substantially by using a semi-supervised set (ROC-AUC of 0.94 and accuracy of 0.87, 2,138 ligands, Figure 3 and Supporting Information Table S4). For the tetradentate models, smaller dataset sizes (984 ligands for randomly selected set) mean that we do not see the performance improvement from the randomly selected dataset (ROC-AUC of 0.81, accuracy 0.73, and recall 0.83) to the small dataset (ROC-AUC of 0.82, accuracy of 0.75, and recall 0.77, 222 ligands total), and we attribute this relatively comparable performance to the small size (178 ligands) of the training set (Figure 3 and Supporting Information Table S5). Thus, the semi-supervised approach is particularly critical in this case, giving by far the best model performance (ROC-AUC of 0.97, accuracy of 0.93, and recall 0.96). We also carried out the same grouped split test for tri- and tetradentate ligands, but we held out oxygen for tetradentate ligands due to both the limited number and class imbalance of phosphorus-coordinating ligands among the tetradentate set. These grouped split models using the semi-supervised labeled data still show good performance, with ROC-AUC of 0.88 and 0.92 for tri- and tetradentate ligands, respectively. To confirm the approach is not strongly sensitive to the ML model, we also trained support vector classifiers, random forest models, and multilayer perceptrons which all have comparable performance to the XGBoost model across all three ligand types (Supporting Information Tables S6–S8).

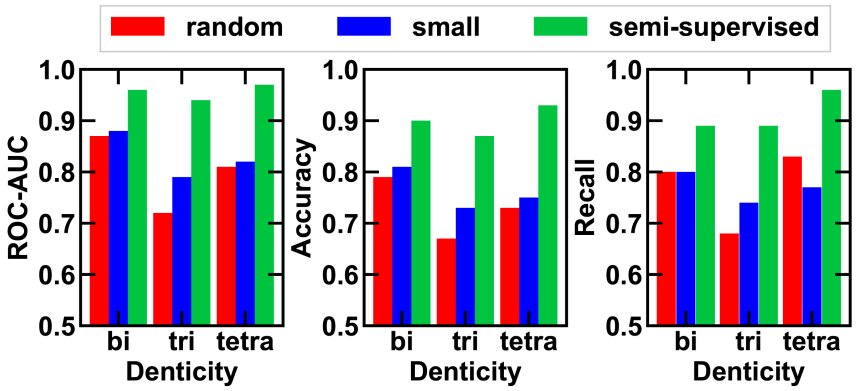

Figure 3. Accuracy and recall of the XGBoost classifier model on the test sets of random, small, and semi-supervised sets.

We next aimed to understand better what chemical mapping we were developing in the semi-supervised learning task by comparing differences in chemical and structural diversity of the two classes. We first analyzed the coordination environment of these ligands and calculated common geometric features of ligands bound to a representative transition metal center (here, copper) in their highest denticity conformations. Analysis of the coordination atom environment shows some differences and similarities between the two sets (Figure 4). As expected, hetero-donating ligands, where the identity of coordinating ligands differ, are more common among hemilabile ligand sets, which is a common design principle for hemilabile ligands. However, the non-hemilabile ligand set also features many hetero-donating ligands. Furthermore, homo-donating ligands are still frequent in the hemilabile ligand set. In particular, the hemilabile ligand set shows an increased number of bis-oxygen coordinating configurations, which can be attributed to the generally weaker donor ability of oxygen-coordinating ligands. While bis-nitrogen ligands are more common among non-hemilabile sets, they are still abundant within hemilabile sets. From the geometric analysis, we find that hemilabile ligands tend to have a slightly lower steric crowding near the metal based on the common steric descriptors such as buried volume[38] or solvent-

accessible surface area[39], which can be attributed to the decrease in ligand bite angle and weaker binding to the transition metal, based on the metal-ligand bond distances (Supporting Information Figures S13–S17). These similarities and differences highlight that while the two classes, obtained through semi-supervised learning, are different, their separation based on one or two geometric or chemical characteristics alone is not trivial.

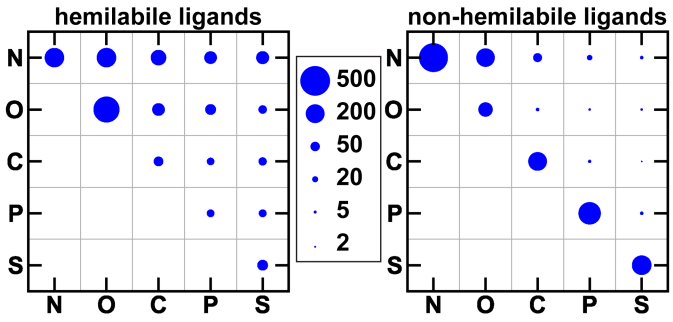

Figure 4. Upper left triangular coordinating-atom matrix showing the frequency of different coordinating environments observed in the hemilabile (left) and non-hemilabile ligand datasets of bidentate ligands. The area of each circle represents the total count of unique ligands, as indicated qualitatively by the inset legend of representative circle sizes.

Analysis of the coordinating atom environment of tri- and tetradentate ligands shows that, similar to the bidentate ligands, the frequency of hetero-donating ligands increases within the hemilabile set for both tri- and tetradentate ligands (Supporting Information Figures S23–S24). Unlike the bidentate set, we find that the all-nitrogen-donating ligand becomes the predominant class not only for the non-hemilabile ligand set but also for the hemilabile ligand set. Similarly, we see an increase in the total number of oxygen-donating ligands in hemilabile class, which can be attributed to weaker donor strength of oxygen-donating ligands. Furthermore, there is a marked increase in the number of bis-carbon donating ligands among the hemilabile set for both tri- and tetradentate ligands, which can be attributed to the π-coordinating alkene ligands. Similar to the

bidentate ligand set, steric crowding around non-hemilabile ligands tends to be lower (Supporting Information Figures S25–S26). While these differences across two classes for three ligand types demonstrate that there is some internal consistency with our label assignment, similarities between the two classes still demonstrate the difficulty of identifying hemilabile ligands and demonstrate the need for classification models.

One potential limitation with our ML models to predict hemilability is that they would require difficult experiments to validate. As an alternative strategy to validate our models, we carried out electronic structure calculations with density functional theory (DFT) to discern differences in ligand dissociation energies from our hemilabile and non-hemilabile (i.e., either from the small or semi-supervised) sets. We selected 100 total tridentate ligands that were neutral (i.e., to avoid issues with charge separation during dissociation) and had been crystalized with Cu in the CSD. In total, 50 tridentate ligands were obtained from the hemilabile set, and 50 tridentate ligands were from the non-hemilabile set (25 small, 25 semi-supervised). A complete list of the ligand refcodes and structures of the ligands are provided in the Supporting Information. We selected tridentate ligands for this stringent test because tridentate ligands were the most challenging for our ML models to classify. The focus on copper is motivated by the fact that copper is the metal with the greatest balance in lower and higher denticity ligand sampling. We then constructed complexes with the hemilabile or non-hemilabile ligand bound to a Cu complex that also contained chloride, in a four-coordinate tetrahedral or square planar geometry, depending on the ligand geometry. We computed partial dissociation energies of the three Cu-L bonds in these complexes. For these partial ligand dissociation energies (see Computational Methods) we observe that partial ligand dissociation is more favorable for hemilabile ligands, with a mean partial

dissociation energy of –1.8 kcal/mol for the bidentate configuration relative to the tridentate bound conformation, whereas the partial dissociation energy of non-hemilabile ligands was found to be 2.3 kcal/mol (Supporting Information Figures S27–S28), implying that the partial dissociation for the hemilabile set is more favorable. Independent t-test analysis showed that these two sets are statistically different, whereas the same test applied to the two non-hemilabile sets are statistically indistinguishable (Supporting Information Table S9). Thus, our semi-small subset supervised labeling strategies quantitatively distinguish features important to separating hemilabile and non-hemilabile ligands.

Motivated by our observation of good separation of ligands that have hemilability from those that do not along with good ML model performance, we further analyzed what features our models trained on the semi-supervised labeled set emphasize the most in making this classification. We carried out a feature importance analysis of the final XGBoost model by examining the total gain function of each feature, where we only considered features that contributed at least 1% to the total gain. Consistent with our earlier analysis on the set, we find that metal-local features (i.e., 1st or 2nd coordination sphere) contribute less (i.e., 25-50%) compared with more distal features (i.e., 3rd coordination sphere and global) that contribute ~50-75% to the total prediction for the three different ligand types (Figure 5). The significant contribution of metal-distal and global features explains the difficulty associated with predicting ligand hemilability based on heuristics and donor-ability alone that had been previously emphasized in the literature[3,20,24,40]. Furthermore, hemilability is a highly balanced property that depends both on structural features as well as electronic features in comparison to other properties such as spin state that depend much more strongly on electronic features[34] (Figure 5). Because we might assume this feature importance was sensitive to our choice of an XGBoost classifier, we also compared feature importance based on

Gini impurity in a random forest classifier. We indeed observe similar feature importances, with metal-distal features contributing ~50-75% of the total prediction (Supporting Information Figure S29).

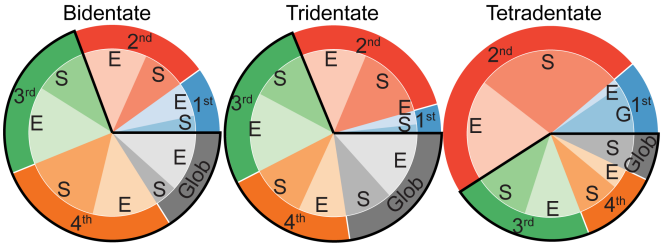

Figure 5. Feature importance of bidentate (left), tridentate (middle) and tetradentate (right) ligands based on the total gain of the XGBoost classifier. Only features that contributed more than 1% to the classifier were retained. *S* refers to structural (topology, identity, radius) and *E* refers to electronic (electronegativity, nuclear charge) features. First through fourth refers to the coordination shell relative to the transition metal based on ligand-centered RACs depth, and Glob refers to global (ligand-scope) features.

Finally, to demonstrate the promise of our ML models for ligand discovery in catalyst design, we use the best-performing XGBoost models to make hemilability predictions for all unlabeled ligands in our original dataset. Our model assigns many ligands as candidates for hemiability. We obtain a bimodal distribution of a similar number of bi- and tetradentate ligands labeled as labile and non-hemilabile, whereas the majority of the tridentate ligands are labeled as hemilabile, which appears to likely be an overprediction of hemilability (Figure 6). These trends by ligand denticity are consistent with the performance of the classifier with the random set, where bidentate and tetradentate ligands showed better performance on the randomly sampled set, but the tridentate classifier has poorer performance, which can potentially be attributed to a larger degree of label contamination if we assume the majority of the unlabeled tridentate ligands are in fact hemilabile.

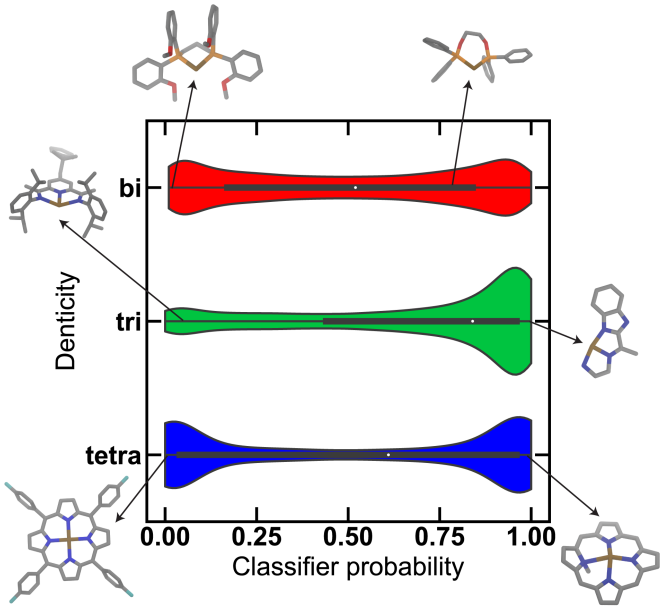

Figure 6. Distribution of the classifier probability on the unlabeled set for bidentate (red), tridentate (green), and tetradentate (blue) ligands. The total areas of each distribution are scaled relative to the size of each set. Examples of hemilabile and non-hemilabile ligands within each set are shown as insets. Hydrogens are omitted for clarity. Representative ligand structures (Refcodes: **BUGVOP**, **AXOLEE**, **COBHIL**, **BODZEA**, **FUGCES**, **CMPORZ** – from left to right starting at top) are shown, with atoms are colored as C in gray, N in blue, O in red, P in orange, and metal in brown.

We further analyzed the predictions by the model to gain insight into the confidence we should have in its predictions and to identify where it could be used in ligand design. For example, a bidentate, bisphosphine ligand with a short, rigid linker between two coordinating atoms is confidently classified as non-hemilabile. However by changing the bisphosphine ligand's electronic character to bisphosphinite, which has a more flexible ethane diol linker between two coordinating atoms, the resulting ligand is classified as hemilabile (Figure 6, inset-top). Similarly, we find that the structure of a tridentate N,N,N-coordinating ligand consisting of rigid sp^2 hybridized linkages between coordinating atoms and bulky substituents that could constrain free atom movement results in the classification of a ligand as non-hemilabile. Whereas a ligand possessing a free-rotating ethylene linker in place of a rigid linker is confidently classified as hemilabile (Figure 6, inset-middle). Finally, we find that macrocyclic tetradentate ligands, such as

porphyrin-derived ligands, are confidently classified as non-hemilabile. On the other hand, when one of the coordinating nitrogen atoms is alkylated, leading to a significant reduction in its donor-ability, the classifier confidently assigns this ligand as hemilabile, highlighting how the model is sensitive to small alterations in the overall structure, including those two bonds or more from the metal center (Figure 6, inset-bottom). Thus, the ML models, especially those trained on bidentate and tetradentate ligands, should provide new pathways to discovering novel hemilabile ligands. We propose that tridentate model predictions could be paired with high-throughput DFT to further strengthen confidence in the model predictions given limitations in tridentate dataset quality from labels obtained purely from the CSD.

In summary, we developed a data-driven workflow for identifying hemilabile and non-hemilabile ligands that can accelerate catalyst screening. We used a semi-supervised learning approach to leverage a combination of labeled and unlabeled data to confidently identify examples of non-hemilabile ligands. We trained ML models that can predict ligand hemilability for bi-, tri-, and tetradentate ligands with high accuracy. We showed that coordinating atom identity alone fails to account for ligand hemilability. Feature importance analysis of machine learning models highlights why conventional design principles can be insufficient for the identification of hemilabile ligands, due to the high significance of metal-distant and structural features. We used trained machine learning models to identify a large number of ligands that are predicted to be hemilabile, that can be used for accelerated discovery of new catalytic reactions.

Computational Methods

Dataset curation: A set of ligands present in mononuclear transition metal complexes was curated from the Cambridge Structural Database (CSD)[41] version 5.41 (November 2019). The procedure employed the Conquest graphical interface and the CSD Python API, with the v5.41

dataset including complexes from the November 2019 dataset with March 2020 and May 2020 updates. A dummy atom with identical connectivity to the metal with an atomic number of 0 was introduced to identify ligands without preserving metal identity. For each ligand with a dummy atom, the atomic number and bond-order weighted connectivity matrix determinant were calculated to identify unique ligands, including their metal-ligand connectivity, as described in ref 28. Atomic number and bond-order weighted connectivity matrix determinant in the absence of a dummy atom was also calculated to identify the same ligands with differing transition metal connectivity.

Feature set: Ligands were featurized using ligand-based revised autocorrelations (RACs),[34] which are connectivity-based representations that have been successfully applied to transition metal complex property predictions.[34-37] Ligand-based RAC features are generated from molecular graphs of a ligand bound to the same dummy transition metal, where each atom is represented by a vertex and each bond is represented by an unweighted edge. Each RAC feature is the sum of products or the sum of differences of heuristic atom properties at depth d (i.e., the number of bonds separating two atoms) on a molecular graph. The ligand-based RACs in this work include features that both span the entire ligand bound to a transition metal, where every atom is used as a starting atom in RACs, as well as features that are centered around only coordinating atoms with a maximum depth $d = 3$. Overall, ligand-based RACs consist of 52 total features (Supporting Information Text S3).

Machine learning models: Three different models were trained per ligand dataset, where the assignment of negative labels was different. For each model, we used the 80/20 random train/test split for the entire database with stratified labels. For random and semi-supervised sets, that contain identical positive labels, the identities of the positive train/test groups were preserved.

We trained classification models using XGBoost v1.5.0, a gradient boosting ensemble model, to classify ligands as either hemilabile or non-hemilabile. Hyperparameters were optimized using Hyperopt v.0.2.7[42] (Supporting Information Table S10). Cross-validation was done using a stratified k-fold, with three folds of random splits. Machine learning model feature importance analysis was conducted with the feature scores of the XGBoost model based on the total gain. We employed the label-spreading semi-supervised learning approach implemented in scikit-learn[43] to identify non-hemilabile ligands. Ligands that were assigned to the negative class with high confidence (>0.995) based on label-spreading, were assigned a negative label. The pseudo-label set was then randomly sampled to supplement the original dataset to obtain an equal number of positive and negative examples for further examination.

Electronic structure calculations: We employed a developer version of the GPU-accelerated TeraChem v1.9[44,45] code to carry out DFT calculations. All calculations were carried out using the B3LYP[46-48] functional with the semi-empirical D3[49] dispersion correction and using Becke-Johnson damping.[50-52] The LACVP* basis set was used, employing the LANL2DZ[53] effective core potential for Cu and 6-31G* for other atoms. All calculations were carried out as closed shell singlets in a restricted formalism. All initial geometries of ligands bound to metal were obtained from the CSD, and chloride atom was added manually, followed by universal force field[54] optimization. All structures were initially optimized to the tridentate-bound conformation with the translation rotation internal coordinate (TRIC) optimizer[55], using the BFGS algorithm with default convergence thresholds of maximum energy gradient of 4.5×10^{-4} hartree/bohr and energy difference between steps of 10^{-6}. To systematically calculate partial dissociation energies, we carried out a series of constrained scans, where each of the three metal-ligand bond was extended by 2 Å from the ground state geometry, in 10 incremental steps, using the TRIC optimizer, while

all other internal coordinates were allowed to relax. The final structure from the scan was used to carry out another optimization using the TRIC optimizer and same convergence threshold as described above, which converged to a lower denticity conformation minimum. Out of the three conformers, the lowest energy conformation was chosen to calculate the partial dissociation energy.

ASSOCIATED CONTENT

Supporting Information. The identification of hemilabile ligands using molecular graph determinants; Statistics on hemilabile ligand dataframe curation; Workflow for defining tridentate ligand sets; Workflow for defining tetradentate ligand sets; Statistics on transition metal counts of hemilabile ligands; Frequency of hemilabile ligands to share a common transition metal; Distribution of low denticity ratio for bidentate ligands; Distribution of low denticity ratio for tridentate ligands; Distribution of low denticity ratio for tetradentate ligands; Tridentate ligand denticity in lower denticity conformations; Tetradentate ligand denticity in lower denticity conformations; Curation of non-hemilabile ligand datasets; ROC curve of XGB classifier for a randomly selected bidentate set; PR curve of XGB classifier for a randomly selected bidentate set; ROC curve of XGB classifier for a small bidentate set; PR curve of XGB classifier for a small bidentate set; Buried volume of bidentate ligands; Solvent accessible surface area of bidentate ligands; Bite angle of bidentate ligands; Metal ligand bond distances of bidentate ligands; Scaled metal ligand bond distances of bidentate ligands; ROC curve of XGB classifier for a semi-supervised bidentate set; PR curve of XGB classifier for a semi-supervised bidentate set; XGBoost classifier prediction probability on the group split set; ROC curve of XGB classifier for a bidentate group split set; PR curve of XGB classifier for a bidentate group split set; Coordinating atom matrix for tridentate ligands; Coordinating atom matrix for tetradentate ligands; Buried volume of tridentate ligands; Buried volume of tetradentate ligands; Performance metrics of XGBoost classifier on bidentate ligands; Performance metrics of XGBoost classifier on tridentate ligands; Performance metrics of XGBoost classifier on tetradentate ligands; Performance metrics of different classifiers on bidentate ligands; Performance metrics of different classifiers on tridentate ligands; Performance metrics of different classifiers on tetradentate ligands; Partial dissociation energies of tridentate ligands combined non-hemilabile; Partial dissociation energies of tridentate ligands split non-hemilabile; Independent T-test for DFT calculated partial dissociation energies; Feature importance of random forest classifier; Hyperparameters for XGBoost model; Description of RACs feature set. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*email:hjkulik@mit.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under grant number CBET-1846426 (to I.K. and H.J.K.). Initial database development was supported by the Office of Naval Research under grant number N00014-20-1-2150 (to C.D.). H.J.K. holds an Alfred P. Sloan Fellowship in Chemistry and is the recipient of a Simon Family Faculty Research Innovation Fund, which supported this work. The authors acknowledge Adam H. Steeves for providing a critical reading of the manuscript.

REFERENCES

(1) Zeng, M.; Li, L.; Herzon, S. B. A Highly Active and Air-Stable Ruthenium Complex for the Ambient Temperature Anti-Markovnikov Reductive Hydration of Terminal Alkynes. *Journal of the American Chemical Society* **2014,** *136*, 7058-7067.

(2) Weissman, H.; Shimon, L. J. W.; Milstein, D. Unsaturated Pd(0), Pd(I), and Pd(Ii) Complexes of a New Methoxy-Substituted Benzyl Phosphine. Aryl−X (X = Cl, I) Oxidative Addition, C−O Cleavage, and Suzuki−Miyaura Coupling of Aryl Chlorides. *Organometallics* **2004,** *23*, 3931-3940.

(3) Chintawar, C. C.; Bhoyare, V. W.; Mane, M. V.; Patil, N. T. Enantioselective Au(I)/Au(Iii) Redox Catalysis Enabled by Chiral (P,N)-Ligands. *Journal of the American Chemical Society* **2022,** *144*, 7089-7095.

(4) Ye, X.; Wang, C.; Zhang, S.; Tang, Q.; Wojtas, L.; Li, M.; Shi, X. Chiral Hemilabile P,N-Ligand-Assisted Gold Redox Catalysis for Enantioselective Alkene Aminoarylation. *Chemistry – A European Journal* **2022,** *28*, e202201018.

(5) Apolinar, O.; Kang, T.; Alturaifi, T. M.; Bedekar, P. G.; Rubel, C. Z.; Derosa, J.; Sanchez, B. B.; Wong, Q. N.; Sturgell, E. J.; Chen, J. S.; Wisniewski, S. R.; Liu, P.; Engle, K. M. Three-Component Asymmetric Ni-Catalyzed 1,2-Dicarbofunctionalization of Unactivated Alkenes Via Stereoselective Migratory Insertion. *Journal of the American Chemical Society* **2022,** *144*, 19337-19343.

(6) Wang, P.-F.; Yu, J.; Guo, K.-X.; Jiang, S.-P.; Chen, J.-J.; Gu, Q.-S.; Liu, J.-R.; Hong, X.; Li, Z.-L.; Liu, X.-Y. Design of Hemilabile N,N,N-Ligands in Copper-Catalyzed Enantioconvergent Radical Cross-Coupling of Benzyl/Propargyl Halides with

Alkenylboronate Esters. *Journal of the American Chemical Society* **2022,** *144*, 6442-6452.
(7) Ros, A.; Estepa, B.; López-Rodríguez, R.; Álvarez, E.; Fernández, R.; Lassaletta, J. M. Use of Hemilabile N,N Ligands in Nitrogen-Directed Iridium-Catalyzed Borylations of Arenes. *Angewandte Chemie* **2011,** *123*, 11928-11932.
(8) Hale, L. V. A.; McGarry, K. A.; Ringgold, M. A.; Clark, T. B. Role of Hemilabile Diamine Ligands in the Amine-Directed C–H Borylation of Arenes. *Organometallics* **2015,** *34*, 51-55.
(9) Yan, N.; Yuan, Y.; Dyson, P. J. Nanometallic Chemistry: Deciphering Nanoparticle Catalysis from the Perspective of Organometallic Chemistry and Homogeneous Catalysis. *Dalton Transactions* **2013,** *42*, 13294-13304.
(10) Chen, Z.; Liu, Z.; Xu, X. Dynamic Evolution of the Active Center Driven by Hemilabile Coordination in Cu/Ceo2 Single-Atom Catalyst. *Nature Communications* **2023,** *14*, 2512.
(11) Peralta, R. A.; Lyu, P.; López-Olvera, A.; Obeso, J. L.; Leyva, C.; Jeong, N. C.; Ibarra, I. A.; Maurin, G. Switchable Metal Sites in Metal–Organic Framework Mfm-300(Sc): Lewis Acid Catalysis Driven by Metal–Hemilabile Linker Bond Dynamics. *Angewandte Chemie International Edition* **2022,** *61*, e202210857.
(12) Peralta, R. A.; Huxley, M. T.; Lyu, P.; Díaz-Ramírez, M. L.; Park, S. H.; Obeso, J. L.; Leyva, C.; Heo, C. Y.; Jang, S.; Kwak, J. H.; Maurin, G.; Ibarra, I. A.; Jeong, N. C. Engineering Catalysis within a Saturated in(Iii)-Based Mof Possessing Dynamic Ligand–Metal Bonding. *ACS Applied Materials & Interfaces* **2023,** *15*, 1410-1417.
(13) Blume, F.; Zemolka, S.; Fey, T.; Kranich, R.; Schmalz, H.-G. Identification of Suitable Ligands for a Transition Metal-Catalyzed Reaction: Screening of a Modular Ligand Library in the Enantioselective Hydroboration of Styrene. *Advanced Synthesis & Catalysis* **2002,** *344*, 868-883.
(14) Sun, H.-Y.; Kubota, K.; Hall, D. G. Reaction Optimization, Scalability, and Mechanistic Insight on the Catalytic Enantioselective Desymmetrization of 1,1-Diborylalkanes Via Suzuki–Miyaura Cross-Coupling. *Chemistry – A European Journal* **2015,** *21*, 19186-19194.
(15) van Dijk, L.; Haas, B. C.; Lim, N.-K.; Clagg, K.; Dotson, J. J.; Treacy, S. M.; Piechowicz, K. A.; Roytman, V. A.; Zhang, H.; Toste, F. D.; Miller, S. J.; Gosselin, F.; Sigman, M. S. Data Science-Enabled Palladium-Catalyzed Enantioselective Aryl-Carbonylation of Sulfonimidamides. *Journal of the American Chemical Society* **2023**, DOI:10.1021/jacs.3c06674 10.1021/jacs.3c06674.
(16) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Accounts of Chemical Research* **2017,** *50*, 605-608.
(17) Thomas, A. A.; Speck, K.; Kevlishvili, I.; Lu, Z.; Liu, P.; Buchwald, S. L. Mechanistically Guided Design of Ligands That Significantly Improve the Efficiency of Cuh-Catalyzed Hydroamination Reactions. *Journal of the American Chemical Society* **2018,** *140*, 13976-13984.
(18) Burrows, L. C.; Jesikiewicz, L. T.; Lu, G.; Geib, S. J.; Liu, P.; Brummond, K. M. Computationally Guided Catalyst Design in the Type I Dynamic Kinetic Asymmetric Pauson–Khand Reaction of Allenyl Acetates. *Journal of the American Chemical Society* **2017,** *139*, 15022-15032.

(19) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chemical Science* **2020,** *11*, 4584-4601.
(20) Pérez García, P. M.; Ren, P.; Scopelliti, R.; Hu, X. Nickel-Catalyzed Direct Alkylation of Terminal Alkynes at Room Temperature: A Hemilabile Pincer Ligand Enhances Catalytic Activity. *ACS Catalysis* **2015,** *5*, 1164-1171.
(21) Weng, Z.; Teo, S.; Hor, T. S. A. Metal Unsaturation and Ligand Hemilability in Suzuki Coupling. *Accounts of Chemical Research* **2007,** *40*, 676-684.
(22) Knebel, W. J.; Angelici, R. J. Mechanism of Chelate Ring-Opening in Metal Carbonyl Complexes. *Inorganic Chemistry* **1974,** *13*, 627-631.
(23) Buckingham, D. A.; Clark, C. R. Kinetics and Mechanism of Ring Opening in the Hydrolysis of Cobalt(Iii) Carbonato Chelates. *Inorganic Chemistry* **1994,** *33*, 6171-6179.
(24) Bassetti, M. Kinetic Evaluation of Ligand Hemilability in Transition Metal Complexes. *European Journal of Inorganic Chemistry* **2006,** *2006*, 4473-4482.
(25) García-Antón, J.; Pons, J.; Solans, X.; Font-Bardia, M.; Ros, J. Synthesis of New Pdii Complexes Containing Thioether−Pyrazole Hemilabile Ligands − Structural Analysis by 1h and 13c Nmr Spectroscopy and Crystal Structures of [Pdcl2(Bddo)] and [Pd(Bddo)](Bf4)2 [Bddo = 1,8-Bis(3,5-Dimethyl-1-Pyrazolyl)-3,6-Dithiaoctane]. *European Journal of Inorganic Chemistry* **2002,** *2002*, 3319-3327.
(26) Duarte, F. J. S.; Poli, G.; Calhorda, M. J. Mechanistic Study of the Direct Intramolecular Allylic Amination Reaction Catalyzed by Palladium(Ii). *ACS Catalysis* **2016,** *6*, 1772-1784.
(27) Higman, C. S.; Nascimento, D. L.; Ireland, B. J.; Audörsch, S.; Bailey, G. A.; McDonald, R.; Fogg, D. E. Chelate-Assisted Ring-Closing Metathesis: A Strategy for Accelerating Macrocyclization at Ambient Temperatures. *Journal of the American Chemical Society* **2018,** *140*, 1604-1607.
(28) Arunachalam, N.; Gugler, S.; Taylor, M. G.; Duan, C.; Nandy, A.; Janet, J. P.; Meyer, R.; Oldenstaedt, J.; Chu, D. B. K.; Kulik, H. J. Ligand Additivity Relationships Enable Efficient Exploration of Transition Metal Chemical Space. *The Journal of Chemical Physics* **2022,** *157*, 184112.
(29) Nandy, A.; Taylor, M. G.; Kulik, H. J. Identifying Underexplored and Untapped Regions in the Chemical Space of Transition Metal Complexes. *The Journal of Physical Chemistry Letters* **2023,** *14*, 5798-5804.
(30) Taylor, M. G.; Yang, T.; Lin, S.; Nandy, A.; Janet, J. P.; Duan, C.; Kulik, H. J. Seeing Is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *The Journal of Physical Chemistry A* **2020,** *124*, 3286-3299.
(31) Brookhart, M.; Green, M. L. H.; Parkin, G. Agostic Interactions in Transition Metal Compounds. *Proceedings of the National Academy of Sciences* **2007,** *104*, 6908-6914.
(32) Seeger, M. "Learning with Labeled and Unlabeled Data," 2000.
(33) Chen, T.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, California, USA, 2016, DOI:10.1145/2939672.2939785 10.1145/2939672.2939785.
(34) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *The Journal of Physical Chemistry A* **2017,** *121*, 8939-8954.

(35) 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.
(36) 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.
(37) 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.
(38) Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P. A Combined Experimental and Theoretical Study Examining the Binding of N-Heterocyclic Carbenes (Nhc) to the Cp*Rucl (Cp* = H5-C5me5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated Nhc Ligands. *Organometallics* **2003,** *22*, 4322-4326.
(39) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies. *Journal of Computational Chemistry* **1995,** *16*, 273-284.
(40) Curley, J. B.; Townsend, T. M.; Bernskoetter, W. H.; Hazari, N.; Mercado, B. Q. Iron, Cobalt, and Nickel Complexes Supported by a Iprpnphp Pincer Ligand. *Organometallics* **2022,** *41*, 301-312.
(41) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B* **2016,** *72*, 171-179.
(42) Bergstra, J.; Yamins, D.; Cox, D.; p 115-123.
(43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **2011,** *12*, 2825-2830.
(44) Petachem. Http://Www.Petachem.Com/. (Accessed April 17, 2023).
(45) 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.
(46) 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.
(47) Becke, A. D. Density-Functional Thermochemistry. Iii. The Role of Exact Exchange. *Journal of Chemical Physics* **1993,** *98*, 5648-5652.
(48) 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.
(49) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (Dft-D) for the 94 Elements H-Pu. *The Journal of chemical physics* **2010,** *132*, 154104.
(50) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *The Journal of Chemical Physics* **2005,** *123*, 154101.
(51) Johnson, E. R.; Becke, A. D. A Post-Hartree–Fock Model of Intermolecular Interactions. *The Journal of Chemical Physics* **2005,** *123*, 024101.

(52) Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *The Journal of Chemical Physics* **2006,** *124*, 174104.
(53) 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.
(54) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. Uff, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society* **1992,** *114*, 10024-10035.
(55) Wang, L.-P.; Song, C. Geometry Optimization Made Simple with Translation and Rotation Coordinates. *The Journal of Chemical Physics* **2016,** *144*, 214108.