INORGANIC CHEMISTRY FRONTIERS

CCS 中国化学会 CHINESE CHEMICAL SOCIETY

ROYAL SOCIETY OF CHEMISTRY

Data-Driven Ligand Field Exploration of Fe(IV)-oxo Sites for C-H Activation

Journal:	*Inorganic Chemistry Frontiers*
Manuscript ID	QI-RES-09-2022-001961.R2
Article Type:	Research Article
Date Submitted by the Author:	16-Nov-2022
Complete List of Authors:	Jones, Grier; University of Tennessee, Department of Chemistry Smith, Brett A.; The University of Tennessee Knoxville, Chemistry Kirkland, Justin; Brigham Young University, Department of Chemistry and Biochemistry Vogiatzis, Konstantinos; University of Tennessee, Department of Chemistry

Data-Driven Ligand Field Exploration of Fe(IV)-oxo Sites for C-H Activation

Grier M. Jones,[1a] Brett A. Smith,[1] [a] Justin K. Kirkland,[2] [a] Konstantinos D. Vogiatzis* [a]

[a] *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, United States*

* Corresponding Author: kvogiatz@utk.edu

ABSTRACT

High-valent Fe(IV)-oxo intermediates, found in enzyme active sites, are excellent targets for biomimetic design of molecular catalysts for C-H bond activation. C-H bonds in inert aliphatic hydrocarbons, such as methane, possess strong bonds that are resistant to chemical functionalization. To aid in the screening of potential catalysts for C-H bond activation, computational methods, such as density functional theory (DFT) and machine learning (ML), are valuable tools for performing high-throughput virtual searches of the vast chemical compound space. In this study, we have designed a database of 50 Fe(IV)-oxo species with varying coordination environments which are further functionalized for a total of approximately 181k structures. DFT calculations are then performed on a subset of the molecular database to determine spin states and C-H bond activation energies. The collected data are then curated based on a series

[1] These authors contributed equally to this work
[2] Current address: *Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah 84602, United States*

of chemically informed criteria. To avoid performing 181k DFT calculations on the total chemical compound space, we developed ML models that utilize a novel molecular representation based on persistence homology, called persistence images (PIs). In particular, we have developed a novel similarity search algorithm, followed by training a regression model to predict C-H activation energies and a classification model to predict the spin states. The priority is to provide high-fidelity predictions for C-H activation barriers. For that purpose, we divided the full database into low and high fidelity structures and we introduced a metric ($\delta\Delta G^{\ddagger}$) which evaluates the effect of a specific ligand modification with respect to the parent, unsubstituted structure. A validation step that included additional DFT calculations on 15 structures demonstrated the credibility of the proposed methodology.

1. INTRODUCTION

Methane is the main component in natural gas and methane clathrates, and with the depletion of petroleum reserves, it is estimated to become the most important hydrocarbon feedstock for the synthesis of fuels and chemicals.[1, 2] However, methane is a highly volatile and flammable gas at room temperature, two features that introduce storage and transportation issues. Methane functionalization to methanol or light hydrocarbons synthesis, which can be used as alternative fuels or feedstocks, is of high importance for petroleum chemistry.[3-5] For considering methane as a sustainable source of raw materials, it should be converted at the place of extraction. Thus, novel selective and energetically less demanding catalytic processes are needed for methane valorization.[6]

One approach is the development of new catalysts that can mimic nature's enzymes. Non-heme enzymes, such as α-ketoglutarate (αKG) dependent taurine dioxygenase (TauD)[7, 8] and syringomycin halogenase (SyrB2),[9] form high-valent Fe(IV)-oxo intermediates that are capable of abstracting an H-atom from an inert C–H bond as strong as 106 kcal/mol to initiate hydroxylation or halogenation.[10] Similar to heme chemistry, many stable non-heme Fe(IV)-oxo model complexes have been synthesized and characterized.[11-16] The high-spin quintet (S = 2) spin state is considered the most reactive spin state and important for the iron(III)-hydroxo intermediate formation, which is stabilized through an exchange controlled mechanism.[17] In 2000, Wieghardt *et al.*[18] reported the first non-heme Fe(IV)-oxo model complex $[(\text{cyclam-CH}_2\text{CO}_2)\text{Fe}^{\text{IV}}(\text{O})]^+$. Cyclam and cyclam-based ligands became a widely popular prospect for non-heme Fe(IV)-oxo chemistry. The first successful crystallographic characterization of a non-heme Fe(IV)-oxo complex, $\left[\text{Fe}^{\text{IV}}(\text{O})(\text{TMC})(\text{NCCH}_3)\right]^{2+}$, was achieved by Rohde *et al.*[19] in 2003. The Fe(IV)-oxo intermediates formed in cyclam-based complexes are relatively stable and feature a pseudo-

octahedral coordination environment. The Fe-oxygen interaction is considerably strong, leading to short bond lengths, and is sterically hindered by the bulky nature of cyclam-based ligands.[20] Numerous successful attempts by synthetic bioinorganic chemists to improve the cyclam ligand scaffolds included ring size variation and binding site modification. Tetramethylated cyclam ligands (TMC) have shown a degree of catalytic tunability based on the ring size.[21] The original cyclam ligand and its derivatives stabilize an intermediate spin state (S = 1) in the Fe(IV)-oxo intermediate. In many cases, the intermediate spin state is less reactive than the high-spin state for the hydrogen atom abstraction (HAA) reaction step which has been attributed to the increased exchange interactions present in the high-spin state.[22-24] Efforts to increase reactivity have been made through replacing one or more of the nitrogen binding sites of TMC with oxygen (TMCO).[25] A different approach is using weak field tripodal ligands to stabilize high-spin Fe(IV)-oxo sites. The first S = 2 non-heme Fe(IV)-oxo model complex was the $\left[(\mathrm{TMG_3tren})\mathrm{Fe^{IV}O}\right]^{2+}$ ($\mathrm{TMG_3tren}$ = (1,1,1-tris{2-[*N*^{2}-(1,1,3,3-tetramethylguanidino)]ethyl}amine) which has a trigonal bipyramidal geometry.[26] The HAA reactivity of this catalyst is diminished by the significant steric hindrance around the Fe-O bond. In 2015, Biswas *et al.*[27] reported one of the most reactive non-heme Fe(IV)-oxo, $[(\mathrm{TQA})\mathrm{Fe^{IV}}(\mathrm{NCCH_3})\mathrm{O}]^{2+}$. The TQA ligand (tris(2-quinolylmethyl)amine) features four weak-field quinoline ligands that stabilizes an octahedral Fe(IV)-oxo. TQA stabilizes the S = 2 spin state and prevents steric hindrance around the Fe-O bound with a vacant site that is filled by a solvent molecule. A recent example of a high-spin (S = 2) Fe(IV)-oxo formed via O_2 activation. $\mathrm{Fe^{IV}(O)(Me_3TACN)((OSi^{Ph2})_2O)}$ is the first species formed via photolytic O-O cleavage of a peroxo(diiron) complex.[28]

Computational studies have provided a comprehensive understanding of the electronic structure of the Fe(IV)-oxo site and its role in a large variety of chemical reactions.[29-37] However,

for the development of novel design strategies that lead to the improvement of catalytic performance, a systematic exploration of large number of molecular systems is mandatory. Previous studies range from the analysis of a small number of complexes for the development of structure-function relations[38] to high-throughput virtual screening of large molecular databases with the aid of quantum chemical methods and machine learning (ML).[39-45]

ML has increased in prevalence in the chemical community due to its ability to perform high-throughput evaluation of many chemical species, allowing researchers to shift from the study of individual molecules to the exploration of a larger fraction of the chemical space.[46-50] One major component of data-driven chemistry pipelines is the generation of molecular representations.[51] Common examples include Coulomb matrices (CMs),[52] the Smooth Overlap of Atomic Positions (SOAPs),[53] and the revised autocorrelation functions (RACs) developed by the Kulik group.[54] Recently, we proposed Persistence Diagrams (PDs) and Persistence Images (PIs) as alternative molecular representations[55] based on persistent homology.[56] PDs and PIs encode the topological information of molecular functional groups and offer a favorable, similar size representation—an advantage over other fingerprinting methods.

In this study, we have applied PIs as a generalizable featurization method for the exploration of Fe(IV)-oxo chemical compound space for the hydrogen atom abstraction reaction step. Methane activation was selected as a probe reaction for the extraction of important information related to the ligand field effects on the Fe center. The starting point of this work is a set of 50 Fe-containing complexes with varying coordination environments, followed by a combinatorial database generation of about 300k structures. A similarity search based on persistence homology for the removal of duplicate structures and for the selection of data for density functional theory (DFT) calculations is introduced. DFT calculations on a small molecular

subset provided reliable data for the training of machine learning models on for the prediction of C-H activation energies and spin states. The development of ML models for the high-throughput virtual screening of approximately 181k structures is presented and a detailed analysis of the results is provided in the next paragraphs. Finally, our conclusions are presented at the end of the article.

2. RESULTS AND DISCUSSION

2.1 The Fe(IV)-Oxo Molecular Subspace

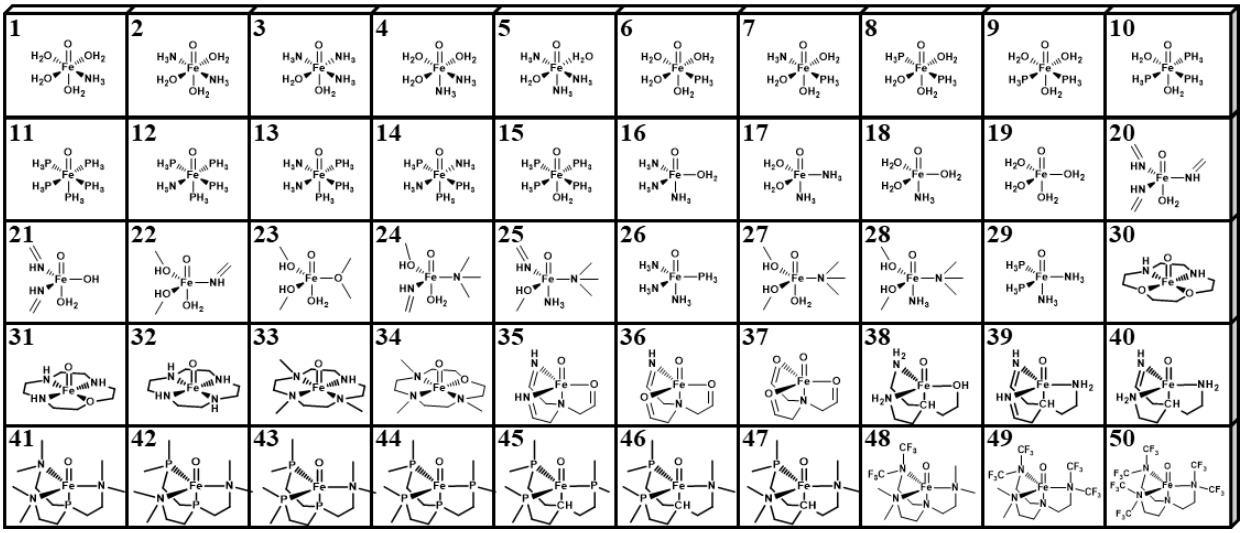


Figure 1: The 50 parent Fe-containing molecular structures used in this study.

In this study, we are introducing a database of 50 Fe(IV)-oxo complexes (Figure 1) that include octahedral (**1-15**), κ^1-trigonal bipyramidal (**16-29**), square pyramidal (**30-34**), and κ^4-trigonal bipyramidal or tetradentate tripodal (**35-50**) molecular structures. This database consists of 15 octahedral complexes containing only monodentate water, ammonia, and phosphane ligands (**1-15**), which introduce ligand field strength effects in a systematic manner (H_2O, NH_3, PH_3 as

weak, intermediate, and strong ligands, respectively). Additionally, there are 16 trigonal bipyramidal complexes containing only monodentate aqua, amino, methanimine, methanol, and trimethylamine ligands (**16-29**). Complexes **1-29** are important for capturing the ligand field effects in both octahedral and trigonal bipyramidal geometries. These complexes will serve to calibrate the ligand field effects in the evaluation of the extended database (*vide infra*). Complexes **30-34** include derivatives of the cyclam and tetramethyl cyclam ligands. These ligands support an Fe(IV)-oxo with pseudo-C_{4v} symmetry with a vacant axial binding site. Cyclam and tetramethylated cyclams stabilize an S = 1 Fe(IV)-oxo, while many of the TMC and TMCO complexes have a transition state with a S = 2 spin state. The remaining complexes **(35-50)** utilize tetradentate tripodal tren-based ligands. The TMG_3tren ligand scaffold supports the notion that trigonal bipyramidal geometries can stabilize a high spin Fe(IV)-oxo, and may lead to high HAA reactivity.

A brute-force combinatorial expansion of the 50 parent structures yields approximately 300,000 molecular structures (Figure 2(a)). Single and double substitutions of hydrogen atoms with fluorine, chlorine, bromine, methyl, or amino groups are introduced in each of the 50 parent structures. Additionally, we developed a novel molecular similarity algorithm based on persistent homology to remove duplicate structures from the database, which is briefly described here. First, each molecular entry is grouped based on their given parent structure and stoichiometry, and their respective persistent diagram (PD) and persistent image (PI) are generated (ESI, Section S1). The PIs of structures within the same group are compared, and if they have a mean squared error (MSE) below a given threshold (here, 10^{-15}), then the structures are considered identical (Figure 2(b)). This process significantly reduces the total number of structures in the diluted database as it

removes molecules that are identical. The refined database consists of 181,436 unique and diverse Fe(IV)-oxo structures (*vide infra*).

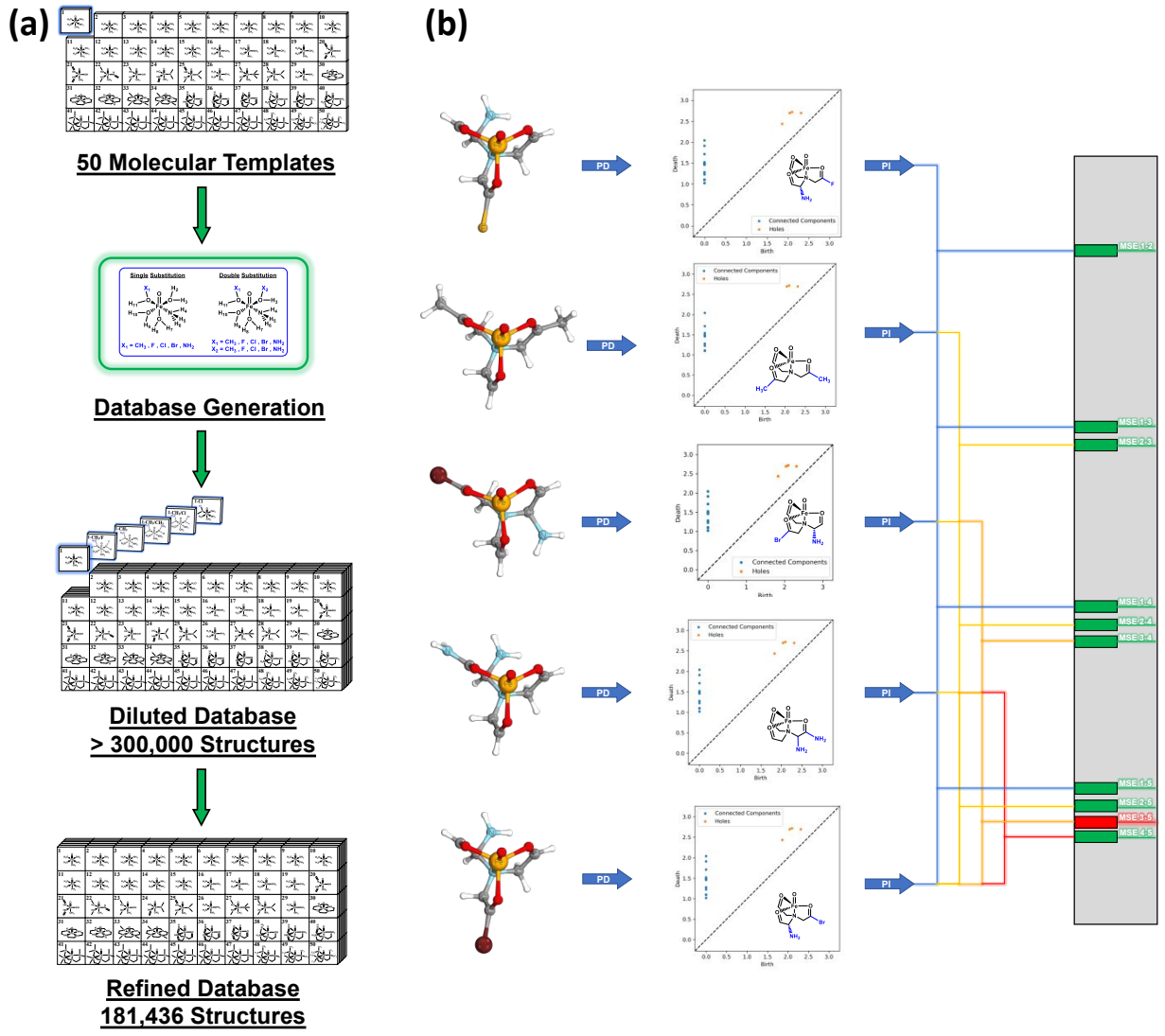


Figure 2: Database generation flowchart. (a) A combinatorial expansion of the 50 parent structures provides a database of approximately 300,000 molecular geometries with many redundant structures. (b) Structure refinement based on persistent homology yields 181,436 unique entries.

2.2 Data Generation from DFT Calculations

Once redundancies are removed, a fraction of the compound space is chosen for explicit calculation of their activation barrier using DFT. We sought to select the most unique structures by identifying 30 systems with the largest MSEs from the comparison of PIs within a given parent structure (1,500 total). Once structures are selected, subsequent DFT calculations are performed for the quintet (S = 2) and triplet (S = 1) spin states of the Fe(IV)-oxo intermediate, as well as for the quintet (S = 2) transition state (TS) of the C-H activation step. We chose two different spin states for the reaction intermediate to properly consider single-state and two-state reactivity channels, which are the two most prominent reaction schemes in Fe(IV)-oxo chemistry. Overall, 4,500 DFT calculations were performed. From these calculations, we were able to determine the activation energy $(\Delta G^{\ddagger})$ for each molecular structure. The computed $\Delta G^{\ddagger}$ values and the corresponding spin-states were used to train and test ML models.

To ensure the integrity of the DFT results, we impose a strict set of four criteria for the automated selection of reliable computational data. Two geometric criteria related to the F-O-H structural core of the HAA have been applied on the TS and the two Fe(IV)-oxo intermediates (S = 1, 2). First, structures are flagged if at least one of the three optimized geometries have an iron-oxygen bond length that is less than 1.45 Å or greater than 1.8 Å, and second, if the oxygen-hydrogen bond distance is less than 0.98 Å or greater than 1.5 Å for the optimized TS structure. The first geometric constraint identifies when a geometry optimization does not converge to an iron-oxo species, while the second identifies cases where the C-H bond remains unactivated. The third and fourth criteria ensure the correct TS was found based on the imaginary vibrational modes. These consider that the primary imaginary frequency (ν_1) must be between 100i and 1600i cm^{-1}, and for cases where there are two imaginary frequencies, the secondary frequency must be greater than 50i cm^{-1}, which typically correspond to a methyl rotational mode. It is important to note that

numerical frequencies were explicitly calculated for each structure in the DFT training set, which allows the implementation of the last two criteria at no additional cost. Structures that did not meet the criteria were manually examined and removed if they did not correspond to the correct structures for HAA.

After this initial screening, structures are dropped from the dataset if there is a negative C-H activation barrier, or if DFT did not converge. All reaction barriers presented herein correspond to thermally corrected ΔG values (in kcal/mol). Negative activation barriers were observed in a few cases and have been confirmed as artifacts derived from the high-throughput data generation. These artifacts led to ligand decoordination and/or rearrangement, incorrect transition state structures, and migration of the transition state active oxo from an axial position to an equatorial position. To maintain the integrity of the calculated data, those cases were removed from the dataset. While our criteria manage to remove most of the undesirable cases, we also removed structures which exhibited ligand decoordination as it was traced to spin contamination (difference between ideal and computed expectation value of the S^2 operator larger than 0.5) and based on the Euclidean distance between the initial and optimized structure (deviations larger than 2 Å). While there is no linear relationship between these two values and it is difficult to derive explicit boundary values for these metrics, we found that the outliers of both metrics tended to be structures where the first coordination sphere rearranges during the geometry optimization. Using our data curation protocol, we refined our DFT database from 1,500 data points down to 486 data points.

2.3 Machine Learning Model Analysis and Validation

Two ML models have been developed for the prediction of spin states and reaction barriers with the DFT data collected from the 486 complexes. Before scanning the database of 181,436 unique structures, we wanted to determine whether a structure has a triplet or a quintet reactant intermediate without explicitly performing 181,436 additional DFT calculations. For that purpose, we train and test a spin classification model using ridge classification. The classification step provides an overall accuracy of 96.47% for the prediction of triplet/quintet spin states for the training set and 91.09% for the test set (Figure 3). Correctly predicted triplets (true positives) make up 29.41% and 19.18% of the training and test set data, respectively, while correctly predicted quintets (true negatives) are 67.06% and 71.92% of the training and test set, respectively. Upon inspection of the triplet-quintet gap ($\Delta E_{\text{T-Q}}$) of the false positive/false negative cases from the training set, we found a very low mean $\Delta E_{\text{T-Q}}$ of 2.6 kcal/mol, which is the source of the misclassification of the 13 cases in the test set (mean $\Delta E_{\text{T-Q}}$ of 7.7 kcal/mol). From the erroneously predicted spin states we conclude that the model is more likely to predict a false negative (predicting a triplet for a true quintet) than a false positive (predicted quintet for a true triplet), which can be attributed to the overall distribution of quintets (70.99%) versus triplets (29.01%) in the DFT data (Figure 4 bottom left). The spin states of the full database are predicted to be 57.77% quintet and 42.23% triplets (Figure 4 bottom right).

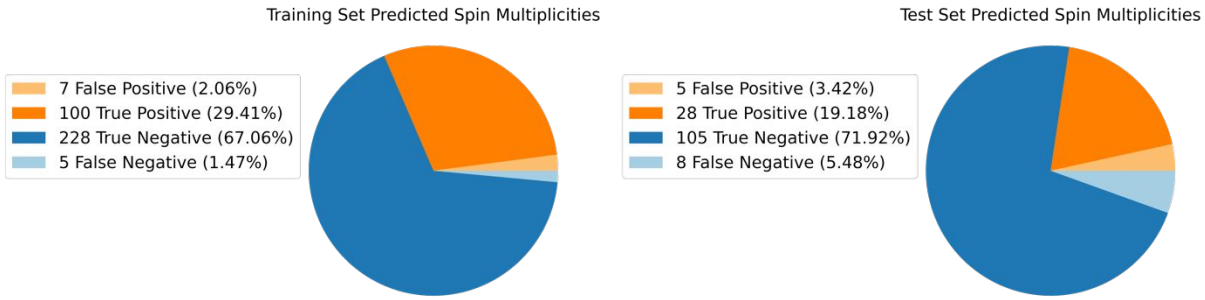


Figure 3: Confusion matrices represented as pie charts of training data (left) and test data (right). A true positive denotes a correctly predicted triplet spin state, whereas a false positive denotes an incorrectly predicted triplet spin state instead of a quintet. True negative denotes a correctly predicted quintet spin state, while a false negative denotes a wrongly predicted quintet instead of a triplet.

Figure 4 shows the geometric composition of the DFT dataset and database along with the spin composition of the DFT dataset and the predicted spin states of the ~181k database. The resulting DFT dataset is composed of 38% κ^4-trigonal bipyramidal, 27% octahedral, 24% κ^1-trigonal bipyramidal, and 12% square pyramidal geometries. The full database consists of 54% κ^4-trigonal bipyramidal, 17% octahedral, 16% κ^1-trigonal bipyramidal, and 14% square pyramidal geometries which demonstrates the heterogeneity of the dataset.

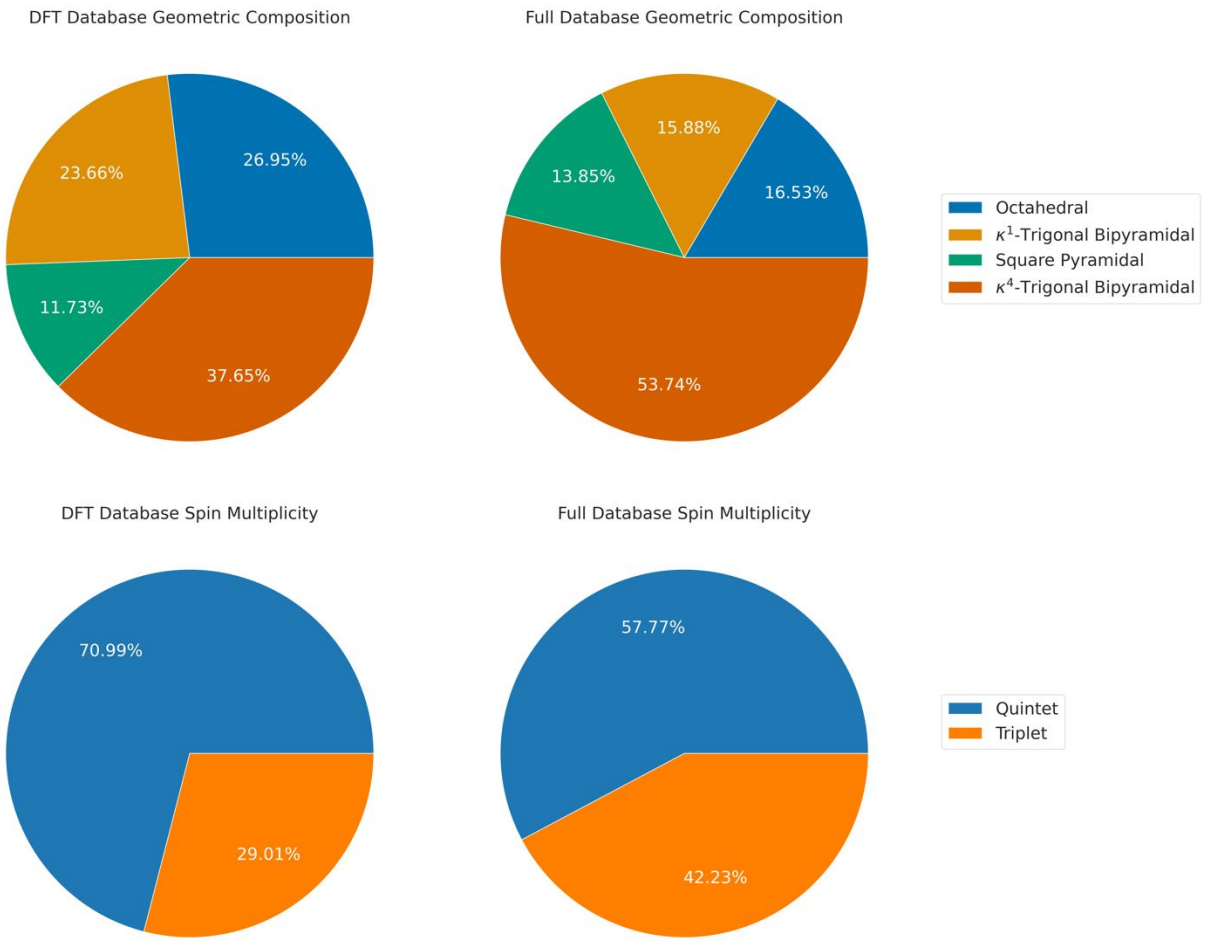


Figure 4: Breakdown of the geometric (top) composition of the DFT database (left) and the full database (right) of 181,436 structures. The spin states (bottom) of the DFT database (left) and the predicted spin states for the full database (right).

Once the classification model was calibrated, we used kernel ridge regression to predict the reaction barriers of the remaining 181,436 molecular complexes within the database. Figure 5 shows the regression parity plot for the reaction barriers of the trained ML model, where the x-axis corresponds to the calculated DFT values, and the y-axis represents the ML (predicted) values. The regression model is evaluated using the coefficient of determination (R^2), mean absolute error

(MAE), and root-mean-square error (RMSE). The training set (70% of the data) has an R^2 of 0.99 with a MAE of 0.73 kcal/mol and a RSME of 1.07 kcal/mol. The test set (30% of the data) has an R^2 of 0.93 with a MAE of 1.69 kcal/mol and a RSME of 2.55 kcal/mol. To further evaluate the model over the full DFT dataset, we use 10-fold cross-validation to get average R^2 values and RMSEs. The average training RMSE is 1.07 ± 0.04 kcal/mol with an average R^2 of 0.99 ± 0.00. The average test RMSE is 2.24 ± 0.40 kcal/mol with an average R^2 of 0.95 ± 0.02. The four largest outliers in Figure 5 correspond to structures 706-(**9**), 884-(**9**), 155-(**32**), and 554-(**20**) (see ESI, section S6 for a more detailed discussion on the outliers). The commonality between these four systems is the ligand functionalization with halogens. We believe that this discrepancy in the predicted barriers is attributed to ligand combinations that were underrepresented in the training set.

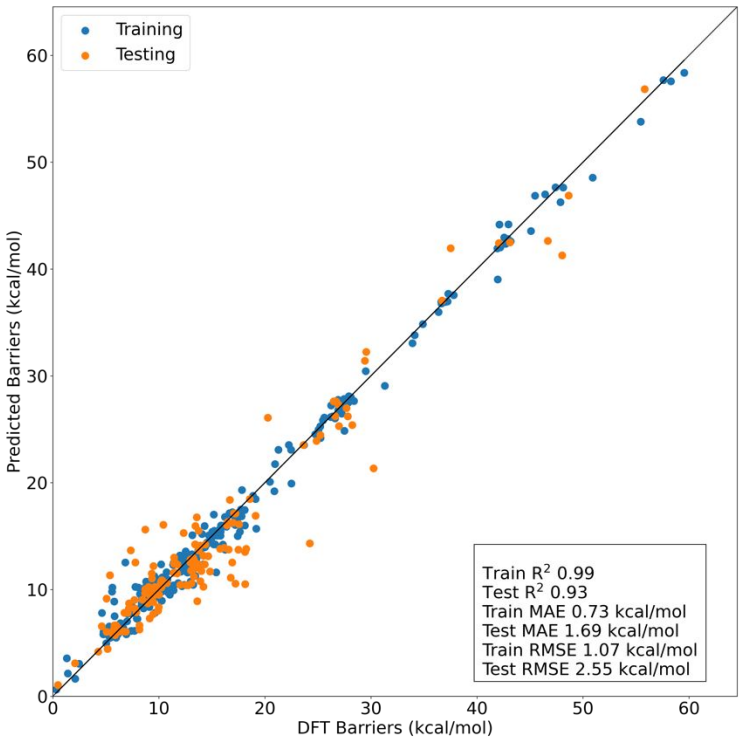


Figure 5: Predicted (kernel ridge regression) vs. DFT(OPBE-D3(BJ)/def2-SVP) barrier heights (in kcal/mol) for the training data points (blue) and the test data points (orange).

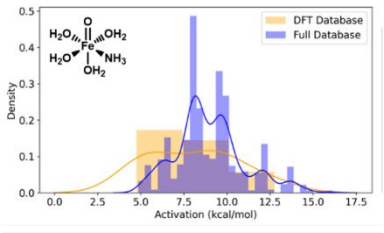
The trained model was used for the high-throughput computational screening of the remaining 181,436 structures. The largest predicted activation barrier is 58.1 kcal/mol (parent structure **11** with double amine substituents on two phosphanes in *cis* positions) while the lowest predicted activation energy is 2.1 kcal/mol (parent structure **37** with double chlorine substitutions on the two adjacent carbons that connect the equatorial oxygen binding site to the axial nitrogen binding site). We have partitioned the range of the predicted activation barriers into three regions: (**A**) activation barriers below 5 kcal/mol, (**B**) activation barriers between 5 and 25 kcal/mol and (**C**) activation barriers above 25 kcal/mol. Group **A** consists of low fidelity structures with unphysically low C-H activation barrier and is sparsely populated in the DFT dataset (less than 3%). The cutoff value of 5 kcal/mol was chosen based on the analysis and curation of the data.

The value of 5 kcal/mol was utilized during data curation as the structures associated with barriers below this threshold were characterized as unphysical. We found that these structures exhibit ligand decoordination and/or rearrangement or in some other cases, migration of the oxo from its axial position to an equatorial position. Overall, group **A** consists only of 0.5% of the 181,436 structures. Group **B** is the area with the greatest fidelity and the most potential insight for C-H bond activation with Fe(IV)-oxo complexes (67.2% of the total database). The complexes in this subspace are predicted to have accessible activation barriers. The largest barriers in this subspace remain below 25 kcal/mol, which is a threshold for potential catalytic activity at reasonable reaction conditions and are above 5 kcal/mol making them more physically meaningful than **A**. The last subspace **C** (32.3% of the total database) is in an area that is not particularly useful for catalyst development and of average fidelity, considering this region populates roughly 16% of the DFT dataset used in training.

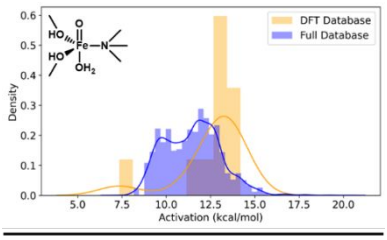
Going forward, we will focus our analysis on subspace **B** (high fidelity/low reaction barriers) for the reliable extraction of structure-function relations for the purposes of HAA. The most commonly observed coordination environment within this subspace is the κ^4-trigonal bipyramidal geometry (43% of **B**). The rest of subspace **B** consists of 14% octahedral, 22% κ^1-trigonal bipyramidal and 21% square pyramidal complexes. The square pyramidal and κ^4-trigonal bipyramidal complexes are of particular interest as they most resemble experimentally characterized molecular Fe(IV)-oxo complexes. The optimal candidates, which can be further computationally or experimentally examined, are identified based on the previous analysis in combination with validation metrics. For that purpose, we have utilized a kernel density estimate together with bar plots of the DFT calculated barriers and the predicted barriers of the full 181,436 database. A plot for each parent structure **1-50** can be found in ESI, Section S2. Here, we analyze

three representative cases from the high-fidelity region of subspace **B** which correspond to parent structures **1**, **27**, and **30**. The selected plots (Figure 6) provide metrics and descriptors that highlight the performance of the ML model on a given parent structures chemical space, i.e., the reliability of prediction. These plots can be used in identifying areas where the machine learning model is likely to provide more reliable predictions, indicated by the overlap of the orange (DFT data) and blue (predicted data) densities and bars. For complexes **1**, **27** and **30**, the difference in mean errors is 0.94, 1.12 and 4.10 kcal/mol, respectively. For validation of our computational methodology and extracting useful information, we have selected 15 functionalized complexes from parent structures **1**, **2**, **4**, **27** and **30** (high fidelity regions), and DFT calculations were employed to explicitly calculate the activation barriers (see ESI, Table S6). The ML predicted barriers show good agreement with DFT, with a MAE of 2.18 kcal/mol, a minimum error of 0.17 kcal/mol, and a maximum error of 8.60 kcal/mol. The MAE of the validation step (2.18 kcal/mol) is comparable to the MAE from of the full model (1.64 kcal/mol), which indicates that the reliability metrics can identify areas where our model provides high accuracy. Additionally, the tested complexes have relatively small activation barriers and may provide additional insights for catalyst development. The change in ligand field effects observed in each entry varies based on the parent structure. In complexes **1** and **2**, the activation barriers increase by 3-4 kcal/mol through the addition of an electron withdrawing group (halides) *cis* to an electron donating group (amino or methyl). The barriers in complex **4** are lowered by 3-4 kcal/mol with the addition of two electron withdrawing groups to *trans* aqua ligands. In complex **30**, the activation barrier is lowered when electron donating groups (methyl or amino) are added into the backbone of the ligand (i.e., to one of the carbon atoms that connect the binding sites). This model exhibits the ability to predict subtle

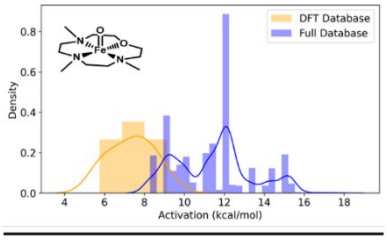
changes to Fe(IV)-oxo complexes. In the next section, a detailed analysis of the model predictions with respect to ligand field effects is presented.



	DFT Database	Full Database	Difference
Mean	8.05	8.90	0.94
SD	2.64	1.98	0.67
Min	4.75	4.86	0.11
Max	12.76	16.12	3.36



	DFT Database	Full Database	Difference
Mean	12.61	11.50	1.12
SD	1.90	1.59	0.31
Min	7.35	7.24	0.10
Max	14.20	20.30	6.10



	DFT Database	Full Database	Difference
Mean	7.35	11.45	4.10
SD	1.14	1.92	0.78
Min	5.74	8.28	2.53
Max	9.14	17.84	8.70

Figure 6: Spread of the DFT activation barriers (orange) and predicted database activation barriers (purple) using histograms and kernel density estimates. Complexes **1** (left), **27** (center) and **30** (right) were selected as representative examples.

2.4 Ligand Field Exploration

First and second coordination effects on the electronic structure and oxidation strength of the Fe(IV)-oxo site vary significantly based on the identity and position of functional groups and the overall ligand architecture. To quantify the functionalization effects, we introduce the $\delta\Delta G^{\ddagger}$ metric to denote the change in the C-H activation energy as the result of the functional groups introduced in the parent molecular structures. Thus, $\delta\Delta G^{\ddagger}$ is computed from the difference between the $\Delta G^{\ddagger}$ of the unsubstituted, parent complex and a given functionalized complex. The average $\delta\Delta G^{\ddagger}$ as a function of the substituent type is shown in Figure 7. The most favorable changes in $\delta\Delta G^{\ddagger}$ derive from substituting hydrogen atoms with fluorine and bromine while the least favorable changes in $\delta\Delta G^{\ddagger}$ result from addition of amino group.

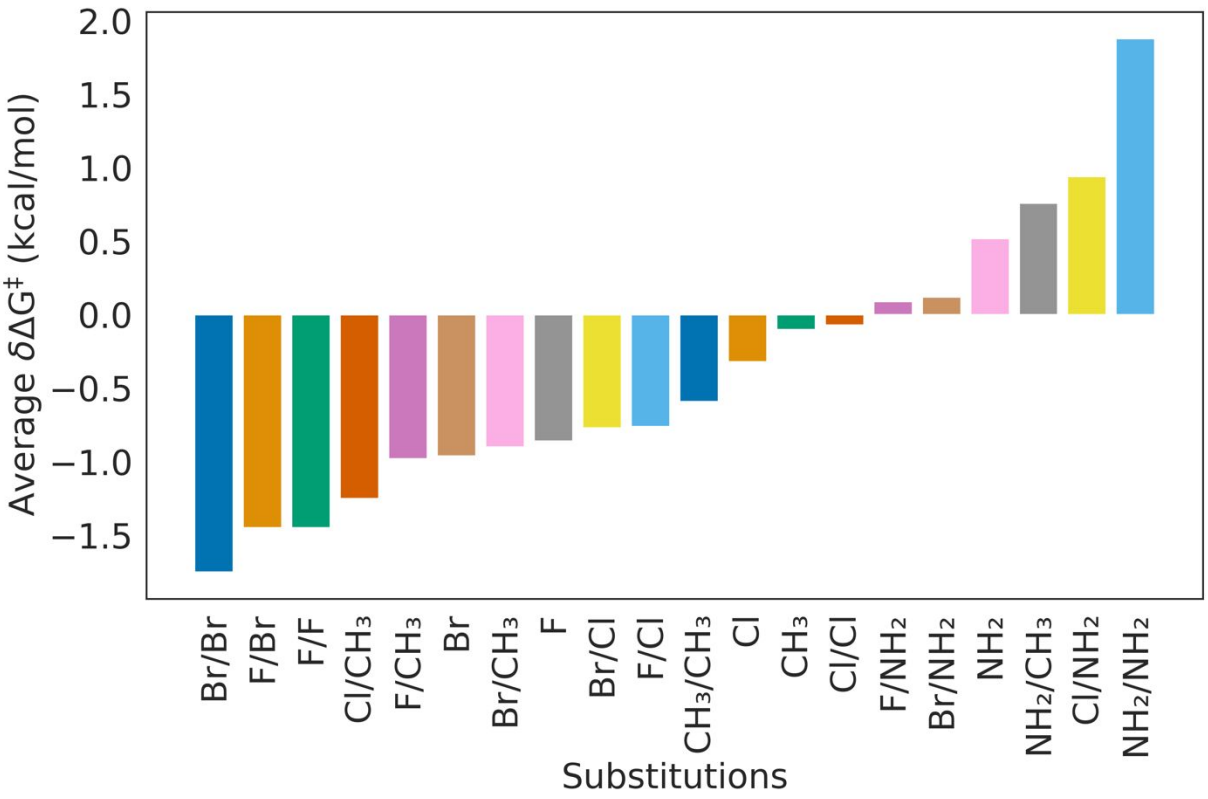


Figure 7: Average changes in the activation barrier (in kcal/mol) between the 50 parent structures and the 181,436 unique entries. Negative values correspond to reduced C-H reaction barriers.

Figure 8 includes the largest positive and negative changes in activation energy for each of the 20 possible group combinations considered in this study. The largest negative $\delta\Delta G^{\ddagger}$ for pure halide (X/X) substitutions were found for parent structure **23,** which is a trigonal bipyramidal coordination complex with three equatorial methoxy ligands and an axial aqua ligand. The incorporation of halides to this ligand architecture is predicted to have favorable changes in activation energy. This is a consequence of their electron withdrawing character which reduces the electron density of atoms coordinated on the Fe center. Therefore, the functionalized complexes have weaker ligand fields with respect to their parent structure equivalent, an effect that yields more reactive Fe(IV)-oxo active sites. In a previous study, we explained this effect by monitoring

the electronic structure of Fe(IV)-oxo complexes with strong and weak ligand fields[31]. The Fe(IV)-oxo reaction intermediates evolve into a Fe(III)-oxyl radical in order to activate a C-H bond, an effect that is enhanced under a weaker coordination ligand field.[27] Additionally, we found common trends regarding the halide position in the functionalized ligands. Seven of the nine most favorable mono- or dihalide substitutions utilize the same position, on a carbon atom adjacent to the nitrogen that is directly coordinated to the Fe center (for example, see structure 3558-(**23**) in Figure 8, first column, first row).

The largest favorable changes in activation energy for the amino/methyl, amino/halide or methyl/halide substitutions correspond to parent structures **26**, and **29**, which are trigonal bipyramidal complexes with amino and phosphane ligands. Even if the effect on the C-H activation barrier for these cases is lower *on average* than the pure halide cases, on an individual basis they yielded some of the most favorable $\delta\Delta G^{\ddagger}$ values. The four largest favorable changes derive from methyl/methyl and methyl/halide substitutions on parent structure **12** (2535-(**12**), 837-(**12**), 836-(**12**) and 4845-(**12**)). Parent structure **12** is comprised of four phosphane ligands and one equatorial amino ligand arranged in an octahedral coordination environment. All four of these complexes share a methyl substituent at the equatorial amino group. In the methyl/methyl case (Figure 8, third column, first row), the second methyl is found on a *cis* equatorial phosphane group. Methyl functionalized amino and phosphane ligands become weaker π-accepting ligands, which leads to increased electron density in the π-antibonding orbital and weakens iron(IV)-oxo bond. All three of the methyl/halide substitutions occur in the same position, with the methyl group added to the amino ligand, and the halide added to the axial phosphane ligand.

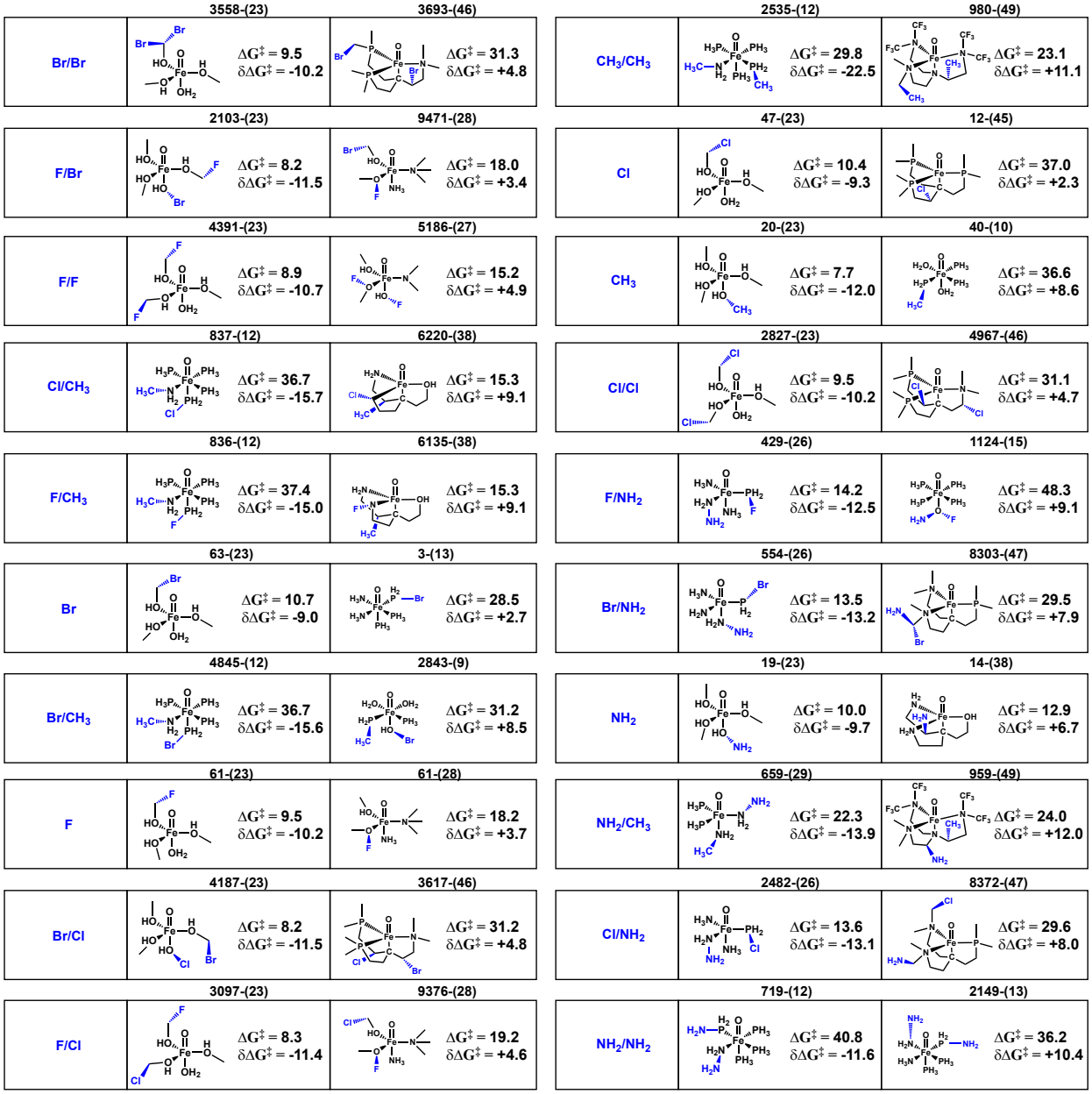


Figure 8: Functionalization type, database ID (integer-letter combination) with the parent structure in parenthesis, largest positive (unfavorable) and negative (favorable) changes in activation energy ($\delta\Delta G^{\ddagger}$ in kcal/mol) and activation energy of the complex of interest ($\Delta G^{\ddagger}$ in kcal/mol)). Groups and bonds displayed in blue indicate the location of functionalization and identity of the group.

The distributions of the $\delta\Delta G^{\ddagger}$ values for each of the 50 different parent structures are shown as violin plots in Figure 9. The violin plots show a quantitative distribution of the data using symmetric kernel density estimates, where more densely populated areas are shown as wider areas on the graph, and the minimum (bottom line), median (middle line), and maximum (top line) from a box and whisker plot are shown as horizontal lines. The octahedral complexes (**1-15**) have the largest range (13.97 kcal/mol) and the most favorable $\delta\Delta G^{\ddagger}$ values on average (-1.72 kcal/mol). The κ^1-trigonal bipyramidal complexes (**16-29**) and κ^4-trigonal bipyramidal complexes (**35-50**) exhibit similar ranges of $\delta\Delta G^{\ddagger}$ (10.40 and 10.18 kcal/mol, respectively). However, on average, the κ^1-trigonal bipyramidal complexes experience a more negative $\delta\Delta G^{\ddagger}$ than κ^4-trigonal bipyramidal complexes (-1.47 and 0.68 kcal/mol, respectively), with the κ^4 complexes exhibiting the least favorable $\delta\Delta G^{\ddagger}$ of any group. These tetradentate complexes are likely the most susceptible to change in the activation energy due to the proximity and connectivity to the axial ligand. A large majority of the functionalization sites are within 1-2 atoms of the axial ligand, where the effects of varying ligand fields are experienced more directly by the oxo ligand. The square pyramidal complexes (**30-34**) are likely the least susceptible to changes in the activation energy, as they experience the tightest spread of values, with an average range of 8.33 kcal/mol. The average $\delta\Delta G^{\ddagger}$ within this group is -0.47 kcal/mol, which is less favorable than both the octahedral and monodentate trigonal bipyramidal complexes but is an improvement to the κ^4-trigonal bipyramidal.

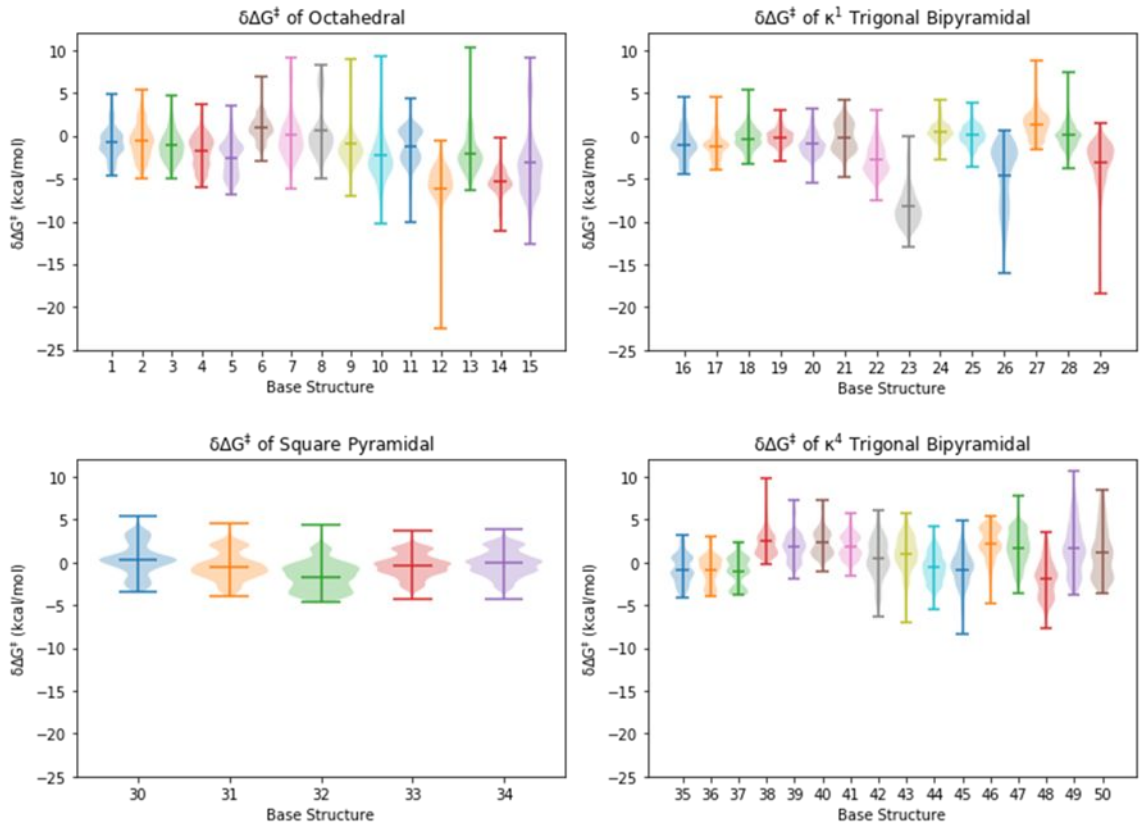


Figure 9: Violin plots of the change in C-H activation energy ($\delta\Delta G^{\ddagger}$) in kcal/mol, to highlight the range, spread and average changes in each parent structure. Complexes grouped based on their molecular geometry.

3. CONCLUSIONS

In this work, we have explored a large ligand space of Fe(IV)-oxo complexes by combining quantum chemical data and novel chemical machine learning methodologies. Our primary aim was to extract relations between molecular structure and C-H activation for catalyst optimization. Our previous work on Fe(IV)-oxo chemistry and how the ligand field strength gives access to multiple reaction channels was the starting point of this project. Here, we began with 50 Fe(IV)-oxo parent structures and we developed a molecular database via combinatorial expansion. The ~300,000

structures were examined via a novel similarity metric that we developed for this project. This metric allowed the further refinement of the chemical space of interest that includes a total of 181,436 unique entries. DFT computations were performed on a representative subspace, which were further reduced to 486 by following an automated data curation protocol. Two ML models were developed based on the 486 data points, a spin state classification model and a C-H activation barrier regression model. Input data were introduced in both models by utilizing the PI molecular representation that encodes geometric and topological information of molecular structures. The trained classification and regression models achieved high accuracy, with 91% confidence for spin state prediction and 1.69 kcal/mol mean absolute error for C-H activation barriers (test data). Both models were further applied for the high-throughput virtual screening of the full ~181k database. We divided the molecular structures into three groups based on the C-H activation barriers: low fidelity/low barrier predictions (group **A**, barrier less than 5 kcal/mol), high fidelity/low barrier predictions (group **B**, between 5 and 25 kcal/mol), and high fidelity/high barrier predictions (group **C**, more than 25 kcal/mol). We then selected 15 structures from group **B** and validated the prediction with additional DFT calculations. Indeed, the largest deviation was 8.6 kcal/mol, while the lowest was only 0.17 kcal/mol. Thus, our model was able to reproduce DFT C-H activation barriers within a small statistical deviation. Based on these findings, we were able to provide an examination of the average change in activation energy and functionalization type, which enable us to highlight key substituents and their most and least effective positions. For example, in strong ligand fields methyl functionalization to phosphanes and other strong field ligands are heavily favorable in octahederal, and both κ^1- and κ^4-trigonal bipyramidal geometries. Conversely, with weak to moderate field ligands halide functionalization to the second coordination sphere are heavily favorable for each of the four explored coordination environments. In moderate to strong

field ligands a combination of halide/methyl substituents were found to be most favorable in *anti* positions, whereas halide/halide groups are generally more favorable in *syn* positions. For that purpose, we introduced the $\delta\Delta G^{\ddagger}$ metric which describes the net difference on the C-H activation reaction step between the initial parent structure and a particular molecular modification of the ligand.

Overall, the topology-based tools and metrics introduced in this work have helped us explore the target chemical space of the Fe(IV)-oxo complexes and enable us to develop a computational methodology for the fast and reliable examination of ligand field effects for improved C-H bond activation. We are currently exploring the transferability of the PI-based tools for other chemical applications, including catalyst optimization for more complex chemical reactions and ligand environments.

4. EXPERIMENTAL SECTION

4.1 Persistent Homology for Molecular Representations

In this work, we have extensively applied the persistence image (PI) molecular representation, which is based on persistent homology, a mathematical tool that allows the computation of topological features of a space at different spatial resolutions. A detailed description of the method can be found in Ref. 55, and a representative example is given in the ESI, Section S1. In short, the geometrical and topological information of a molecular structure are encoded into a 2D diagram, the persistent diagram (PD). The generation (birth) and the elimination (death) of the topological features included in a PD are tracked though a filtration parameter. For this project, we have chosen to include zeroth- and first-order topological features. The zeroth-

order features are called “connected components” and encode the atom connectivity (bonds) of a molecule. The first-order topological features are called “holes” and encode the topology of atoms that form rings in a molecular structure. Each molecular functional group has distinct “hole” birth and death coordinates in the PD which are used as molecular fingerprints. The PDs are then converted into vectorized 2D images, the persistence images (PIs), that are used as input in ML models.

4.2 Database Generation

Molecular substitutions were performed on each of the 50 parent structures using molSimplify,[57] which utilizes OpenBabel in the backend,[58, 59] to generate force field optimized structures. Either one or two of the ligand’s hydrogen atoms were substituted with a fluorine, chlorine, or bromine atom(s) or a methyl or amino group(s). This was performed for each hydrogen atom and for each combination of hydrogen atom pairs (Figure 10). Structure generation in this manner is highly comprehensive, but also generates redundancies. The naming conventions utilized in the full database (181,436 molecular complexes) for each entry is listed as an integer-integer pair, e.g. 0-(**1**), 1-(**2**) …, etc. The first integer number corresponds to a nondescript identifier that was used to provide a unique *x*-(***y***) pair for every substituted parent structure (*y* = **1**, **2**, …, **50**).

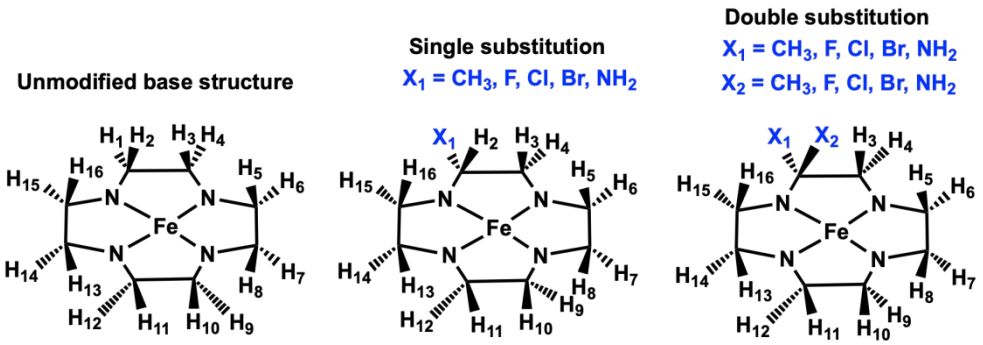


Figure 10: An example of modifications applied for structure generation (parent structure **32**). Unmodified parent structure (left), singly substituted structure (center) and doubly substituted (right). Text in blue highlights the hydrogen that is replaced with one of the 5 atoms/groups used (-CH_3, -F, -Cl, -Br and/or -NH_2)

4.3 Computational Details

All density functional theory calculations were performed with ORCA 4.2.1[60, 61] using the OPBE functional,[62, 63] the def2-SVP basis set,[64, 65] and the Grimme's D3 dispersion correction[66] with the Becke-Johnson damping function.[67] Due to the large number of computations required for the generation of the database, we have selected a relatively small basis set, similarly to previous high-throughput computational screening studies of transition metal complexes for catalytic applications.[68] The resolution of identity chain-of-spheres approximation was used in the two-electron integral evaluation with Lebedev's 590-point integration grid.[69, 70] For each structure included in the training set, a minimum of three calculations were performed: (1) transition state optimization, (2) optimization of the high-spin (S = 2) Fe(IV)-oxo intermediate and (3) optimization of the low/intermediate spin (S = 1) Fe(IV)-oxo intermediate. The hydrogen atom abstraction transition state optimization began with an initial transition state scan along the O-H internal coordinate, from 1.5 to 1.1 Å in increments of 0.02 Å. In a few cases, the preliminary

internal coordinate scan led to an unsuccessful transition state optimization. In these cases, the scanned geometries were manually inspected for potential saddle points and the Cartesian coordinates were generated from one of the intermediate scan structures. Full transition state optimizations were then performed without a constraining the O-H coordinate. The OPBE functional was selected since previous benchmark studies have shown that it has excellent agreement with unrestricted coupled-cluster singles, doubles and perturbative triples (UCCSD(T)).[71, 72] In addition, it has been extensively used in previous computational studies of Fe(IV)-oxo complexes.[38, 73, 74]

4.4 Machine Learning Model Development

The PIs[55] of each Fe(IV)-oxo species were generated using a pixelation of 20 by 20, a minBD of -0.1 Å (lower boundary of the PI), a maxBD of 2.5 Å (upper boundary of the PI), and a spread of 0.06 (standard deviation of the gaussian kernel). The classification and regression models are performed using Python 3.6 with the Sci-kit Learn package.[75] For both models, 70% (340 structures) of the DFT data is used for training and 30% (146 structures) of the data is used for testing.

The classification model is performed using linear ridge classification, with an alpha regularization parameter of 1e-2. We evaluate the effectiveness of our model using the accuracy score, classification report, and confusion matrices as they are implemented in Sci-kit Learn. The training set has an accuracy of 0.96 and the test set has an accuracy of 0.91. Figure 6 shows a confusion matrix represented as a pie chart where negative values, or 0, denote quintet structures and positive values, or 1, denote triplet structures. A true negative (TN) indicates a negative value,

or quintet, is predicted correctly and a false negative (FN) indicates that a positive value is predicted incorrectly to be a negative value. For the triplets, a true positive (TP) designates triplets being classified correctly and a false positive (FP) designates a negative value being classified incorrectly as a positive value. The training set has a TN of 67.06%, FN of 1.47%, TP of 29.41%, and FP of 2.06% and the test set has a TN of 71.92%, FN of 5.48%, TP of 19.18%, and FP of 3.42%. Further evaluation of the model includes metrics such as the precision, recall, and F1-score (see ESI, Section S3).

After the calibration of the classification model, we performed an extensive search of the kernel ridge regression parameters with three-fold cross-validation using GridSearchCV in Sci-kit Learn (the parameters can be found in ESI, Section S4). The optimal parameters were a Laplacian kernel with an alpha regularization parameter of 1e-2 and a gamma kernel parameter of 1e-2.

Electronic supplementary information (ESI) available: Examples of Persistent Images, structural and database distribution, classification, and regression model metrics, DFT energies, validation energies. Cartesian coordinates of the full 181,436 database and DFT optimized geometries will be provided on our GitHub: **XXX**.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We would like to thank Dr. Jacob Townsend for fruitful discussions during the early stages of this project. This material is based upon work supported by the National Science Foundation under Grant No. CHE-1800237. The authors acknowledge the Infrastructure for Scientific Applications and Advanced Computing (ISAAC) of the University of Tennessee for computational resources.

References

1. R. A. Kerr, Natural gas from shale bursts onto the scene, *Science*, 2010, **328**, 1624-1626.
2. Y.-S. Yu, X. Zhang, J.-W. Liu, Y. Lee and X.-S. Li, Natural gas hydrate resources and hydrate technologies: a review and analysis of the associated energy and global warming challenges, *Energy & Environmental Science*, 2021, **14**, 5611-5668.
3. G. A. Olah, Beyond Oil and Gas: The Methanol Economy, *Angewandte Chemie International Edition*, 2005, **44**, 2636-2639.
4. G. A. Olah, Towards Oil Independence Through Renewable Methanol Chemistry, *Angewandte Chemie International Edition*, 2013, **52**, 104-107.
5. S. Verhelst, J. W. G. Turner, L. Sileghem and J. Vancoillie, Methanol as a fuel for internal combustion engines, *Progress in Energy and Combustion Science*, 2019, **70**, 43-88.
6. D. Saha, H. A. Grappe, A. Chakraborty and G. Orkoulas, Postextraction Separation, On-Board Storage, and Catalytic Conversion of Methane in Natural Gas: A Review, *Chemical Reviews*, 2016, **116**, 11436-11499.
7. J. M. Elkins, M. J. Ryle, I. J. Clifton, J. C. Dunning Hotopp, J. S. Lloyd, N. I. Burzlaff, J. E. Baldwin, R. P. Hausinger and P. L. Roach, X-ray Crystal Structure of *Escherichia coli* Taurine/α-Ketoglutarate Dioxygenase Complexed to Ferrous Iron and Substrates, *Biochemistry*, 2002, **41**, 5185-5192.
8. J. C. Price, E. W. Barr, T. E. Glass, C. Krebs and J. M. Bollinger, Evidence for Hydrogen Abstraction from C1 of Taurine by the High-Spin Fe(IV) Intermediate Detected during Oxygen Activation by Taurine:α-Ketoglutarate Dioxygenase (TauD), *Journal of the American Chemical Society*, 2003, **125**, 13008-13009.

9. M. L. Matthews, C. S. Neumann, L. A. Miles, T. L. Grove, S. J. Booker, C. Krebs, C. T. Walsh and J. M. Bollinger, Substrate positioning controls the partition between halogenation and hydroxylation in the aliphatic halogenase, SyrB2, *Proceedings of the National Academy of Sciences*, 2009, **106**, 17723-17728.

10. C. Krebs, D. Galonić Fujimori, C. T. Walsh and J. M. Bollinger, Non-Heme Fe(IV)–Oxo Intermediates, *Accounts of Chemical Research*, 2007, **40**, 484-492.

11. V. A. Larson, B. Battistella, K. Ray, N. Lehnert and W. Nam, Iron and manganese oxo complexes, oxo wall and beyond, *Nature Reviews Chemistry*, 2020, **4**, 404-419.

12. J. Chen, Z. Jiang, S. Fukuzumi, W. Nam and B. Wang, Artificial Nonheme Iron and Manganese Oxygenases For Enantioselective Olefin Epoxidation and Alkane Hydroxylation Reactions, *Coordination Chemistry Reviews*, 2020, **421**, 213443.

13. S. Kal, S. Xu and L. Que, Bio-inspired Nonheme Iron Oxidation Catalysis: Involvement of Oxoiron(V) Oxidants in Cleaving Strong C−H Bonds, *Angewandte Chemie International Edition*, 2020, **59**, 7332-7349.

14. M. Guo, T. Corona, K. Ray and W. Nam, Heme and Nonheme High-Valent Iron and Manganese Oxo Cores in Biological and Abiological Oxidation Reactions, *ACS Central Science*, 2019, **5**, 13-28.

15. X. Engelmann, I. Monte-Perez and K. Ray, Oxidation Reactions with Bioinspired Mononuclear Non-Heme Metal-Oxo Complexes, *Angewandte Chemie International Edition*, 2016, **55**, 7632-7649.

16. M. Milan, M. Bietti and M. Costas, Enantioselective aliphatic C-H bond oxidation catalyzed by bioinspired complexes, *Chemical Communications*, 2018, **54**, 9559-9570.

17. S. Shaik, H. Chen and D. Janardanan, Exchange-enhanced reactivity in bond activation by metal-oxo enzymes and synthetic reagents, *Nature Chemistry*, 2011, **3**, 19-27.

18. C. A. Grapperhaus, B. Mienert, E. Bill, T. Weyhermuller and K. Wieghardt, Mononuclear (nitrido)iron(V) and (oxo)iron(IV) complexes via photolysis of [(cyclam-acetato)FeIII(N3)]+ and ozonolysis of [(cyclam-acetato)FeIII(O3SCF3)]+ in water/acetone mixtures, *Inorganic Chemistry*, 2000, **39**, 5306-5317.

19. J. U. Rohde, J. H. In, M. H. Lim, W. W. Brennessel, M. R. Bukowski, A. Stubna, E. Munck, W. Nam and L. Que, Jr., Crystallographic and spectroscopic characterization of a nonheme Fe(IV)-O complex, *Science*, 2003, **299**, 1037-1039.

20. S. P. de Visser, J.-U. Rohde, Y.-M. Lee, J. Cho and W. Nam, Intrinsic properties and reactivities of mononuclear nonheme iron–oxygen complexes bearing the tetramethylcyclam ligand, *Coordination Chemistry Reviews*, 2013, **257**, 381-393.

21. S. Hong, H. So, H. Yoon, K. B. Cho, Y. M. Lee, S. Fukuzumi and W. Nam, Reactivity comparison of high-valent iron(IV)-oxo complexes bearing N-tetramethylated cyclam ligands with different ring size, *Dalton Transactions*, 2013, **42**, 7842-7845.

22. G. Xue, R. De Hont, E. Munck and L. Que, Jr., Million-fold activation of the [Fe(2)(micro-O)(2)] diamond core for C-H bond cleavage, *Nature Chemistry*, 2010, **2**, 400-405.

23. H. Hirao, L. Que, Jr., W. Nam and S. Shaik, A two-state reactivity rationale for counterintuitive axial ligand effects on the C-H activation reactivity of nonheme FeIV=O oxidants, *Chemistry*, 2008, **14**, 1740-1756.

24. M. S. Seo, N. H. Kim, K.-B. Cho, J. E. So, S. K. Park, M. Clémancey, R. Garcia-Serres, J.-M. Latour, S. Shaik and W. Nam, A mononuclear nonheme iron(iv)-oxo complex which

is more reactive than cytochrome P450 model compound I, *Chemical Science*, 2011, **2**, 1039-1039.

25. I. Monte Perez, X. Engelmann, Y. M. Lee, M. Yoo, E. Kumaran, E. R. Farquhar, E. Bill, J. England, W. Nam, M. Swart and K. Ray, A Highly Reactive Oxoiron(IV) Complex Supported by a Bioinspired N3 O Macrocyclic Ligand, *Angewandte Chemie International Edition*, 2017, **56**, 14384-14388.

26. J. England, M. Martinho, E. R. Farquhar, J. R. Frisch, E. L. Bominaar, E. Munck and L. Que, Jr., A synthetic high-spin oxoiron(IV) complex: generation, spectroscopic characterization, and reactivity, *Angewandte Chemie International Edition*, 2009, **48**, 3622-3626.

27. A. N. Biswas, M. Puri, K. K. Meier, W. N. Oloo, G. T. Rohde, E. L. Bominaar, E. Munck and L. Que, Jr., Modeling TauD-J: a high-spin nonheme oxoiron(IV) complex with high reactivity toward C-H bonds, *Journal of the American Chemical Society*, 2015, **137**, 2428-2431.

28. J. B. Gordon, T. Albert, A. Dey, S. Sabuncu, M. A. Siegler, E. Bill, P. Moenne-Loccoz and D. P. Goldberg, A Reactive, Photogenerated High-Spin (S = 2) Fe(IV)(O) Complex via O2 Activation, *Journal of the American Chemical Society*, 2021, **143**, 21637-21647.

29. P. Verma, K. D. Vogiatzis, N. Planas, J. Borycz, D. J. Xiao, J. R. Long, L. Gagliardi and D. G. Truhlar, Mechanism of Oxidation of Ethane to Ethanol at Iron(IV)-Oxo Sites in Magnesium-Diluted Fe2(dobdc), *Journal of the American Chemical Society*, 2015, **137**, 5770-5781.

30. G. Ricciardi, E. J. Baerends and A. Rosa, Charge Effects on the Reactivity of Oxoiron(IV) Porphyrin Species: A DFT Analysis of Methane Hydroxylation by Polycationic Compound I and Compound II Mimics, *ACS Catalysis*, 2015, **6**, 568-579.

31. J. K. Kirkland, S. N. Khan, B. Casale, E. Miliordos and K. D. Vogiatzis, Ligand field effects on the ground and excited states of reactive FeO(2+) species, *Physical Chemistry Chemical Physics*, 2018, **20**, 28786-28795.

32. K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu and E. A. Pidko, Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities, *Chemical Reviews*, 2019, **119**, 2453-2523.

33. A. Nandy and H. J. Kulik, Why Conventional Design Rules for C–H Activation Fail for Open-Shell Transition-Metal Catalysts, *ACS Catalysis*, 2020, **10**, 15033-15047.

34. R. Kumar, M. Sundararajan and G. Rajaraman, A six-coordinate high-spin Fe(IV)=O species of cucurbit[5]uril: a highly potent catalyst for C-H hydroxylation of methane, if synthesised, *Chemical Communications*, 2021, **57**, 13760-13763.

35. J. E. Schneider, M. K. Goetz and J. S. Anderson, Statistical analysis of C-H activation by oxo complexes supports diverse thermodynamic control over reactivity, *Chemical Science*, 2021, **12**, 4173-4183.

36. A. Nandy, H. Adamji, D. W. Kastner, V. Vennelakanti, A. Nazemi, M. J. Liu and H. J. Kulik, Using Computational Chemistry To Reveal Nature's Blueprints for Single-Site Catalysis of C-H Activation, *ACS Catalysis*, 2022, **12**, 9281-9306.

37. A. Katoch and D. Mandal, Effect of the substituent on C-H activation catalyzed by a non-heme Fe(IV)O complex: a computational investigation of reactivity and hydrogen tunneling, *Dalton Transactions*, 2022, **51**, 11641-11649.

38. P. C. Andrikopoulos, C. Michel, S. Chouzier and P. Sautet, In Silico Screening of Iron-Oxo Catalysts for CH Bond Cleavage, *ACS Catalysis*, 2015, **5**, 2490-2499.

39. T. Z. H. Gani and H. J. Kulik, Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane to Methanol Conversion by FeIV═O, *ACS Catalysis*, 2018, **8**, 975-986.

40. G. H. Gu, C. Choi, Y. Lee, A. B. Situmorang, J. Noh, Y. H. Kim and Y. Jung, Progress in Computational and Machine-Learning Methods for Heterogeneous Small-Molecule Activation, *Advanced Materials*, 2020, **32**, e1907865.

41. J. P. Janet, L. Chan and H. J. Kulik, 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.

42. H. J. Kulik, Making machine learning a useful tool in the accelerated discovery of transition metal complexes, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2019, **10**, 1-14.

43. A. Nandy, C. Duan, C. Goffinet and H. J. Kulik, New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts, *JACS Au*, 2022, **2**, 1200-1213.

44. D. R. Harper, A. Nandy, N. Arunachalam, C. Duan, J. P. Janet and H. J. Kulik, Representations and strategies for transferable machine learning improve model performance in chemical discovery, *The Journal of Chemical Physics*, 2022, **156**, 074101.

45. C. Duan, A. Nandy, H. Adamji, Y. Roman-Leshkov and H. J. Kulik, Machine Learning Models Predict Calculation Outcomes with the Transferability Necessary for

Computational Catalysis, *Journal of Chemical Theory and Computation*, 2022, **18**, 4282-4292.

46. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**, 547-555.

47. R. Gomez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D. G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, 2016, **15**, 1120-1127.

48. J. G. Freeze, H. R. Kelly and V. S. Batista, Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists, *Chemical Reviews*, 2019, **119**, 6595-6612.

49. K. V. Chuang and M. J. Keiser, Comment on "Predicting reaction performance in C-N cross-coupling using machine learning", *Science*, 2018, **362**, 16-19.

50. J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, Response to Comment on "Predicting reaction performance in C-N cross-coupling using machine learning", *Science*, 2018, **362**, eaat8763-eaat8763.

51. F. Musil, A. Grisafi, A. P. Bartok, C. Ortner, G. Csanyi and M. Ceriotti, Physics-Inspired Structural Representations for Molecules and Materials, *Chem Rev*, 2021, **121**, 9759-9815.

52. M. Rupp, A. Tkatchenko, K. R. Muller and O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Physical Review Letters*, 2012, **108**, 058301.

53. A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Physical Review B*, 2013, **87**, 1-16.

54. J. P. Janet and H. J. Kulik, Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships, *The Journal of Physical Chemistry A*, 2017, **121**, 8939-8954.

55. J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas and K. D. Vogiatzis, Representation of molecular structures with persistent homology for machine learning applications in chemistry, *Nature Communications*, 2020, **11**, 3230.

56. L. Wasserman, Topological Data Analysis, *Annual Review of Statistics and Its Application, Vol 5*, 2018, **5**, 501-532.

57. E. I. Ioannidis, T. Z. Gani and H. J. Kulik, molSimplify: A toolkit for automating discovery in inorganic chemistry, *Journal of Computational Chemistry*, 2016, **37**, 2106-2117.

58. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *Journal of Cheminformatics*, 2011, **3**, 33.

59. N. M. O'Boyle, C. Morley and G. R. Hutchison, Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, *Chemistry Central Journal*, 2008, **2**, 5.

60. F. Neese, The ORCA program system, *Wiley Interdisciplinary Reviews-Computational Molecular Science*, 2012, **2**, 73-78.

61. F. Neese, Software update: the ORCA program system, version 4.0, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2017, **8**, 4-9.

62. W.-M. Hoe, A. J. Cohen and N. C. Handy, Assessment of a new local exchange functional OPTX, *Chemical Physics Letters*, 2001, **341**, 319-328.

63. J. P. Perdew, K. Burke and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Physical Review Letters*, 1996, **77**, 3865-3868.

64. F. Weigend and R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy, *Physical Chemistry Chemical Physics*, 2005, **7**, 3297-3305.

65. F. Weigend, Accurate Coulomb-fitting basis sets for H to Rn, *Physical Chemistry Chemical Physics*, 2006, **8**, 1057-1065.

66. S. Grimme, J. Antony, S. Ehrlich and H. Krieg, 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.

67. S. Grimme, S. Ehrlich and L. Goerigk, Effect of the damping function in dispersion corrected density functional theory, *Journal of Computational Chemistry*, 2011, **32**, 1456-1465.

68. P. Friederich, G. Dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex, *Chemical Science*, 2020, **11**, 4584-4601.

69. O. Treutler and R. Ahlrichs, Efficient molecular numerical integration schemes, *The Journal of Chemical Physics*, 1995, **102**, 346-354.

70. V. I. Lebedev and D. N. Laikov, Quadrature formula for the sphere of 131th algebraic order of accuracy, *Dokl Akad Nauk+*, 1999, **366**, 741-745.

71. M. Feldt, Q. M. Phung, K. Pierloot, R. A. Mata and J. N. Harvey, Limits of Coupled-Cluster Calculations for Non-Heme Iron Complexes, *Journal of Chemical Theory and Computation*, 2019, **15**, 922-937.

72. H. Chen, W. Lai and S. Shaik, Exchange-Enhanced H-Abstraction Reactivity of High-Valent Nonheme Iron(IV)-Oxo from Coupled Cluster and Density Functional Theories, *Journal of Physical Chemistry Letters*, 2010, **1**, 1533-1540.

73. E. Andris, J. Jasik, L. Gomez, M. Costas and J. Roithova, Spectroscopic Characterization and Reactivity of Triplet and Quintet Iron(IV) Oxo Complexes in the Gas Phase, *Angewandte Chemie International Edition*, 2016, **55**, 3637-3641.

74. M. Swart, Accurate Spin-State Energies for Iron Complexes, *Journal of Chemical Theory and Computation*, 2008, **4**, 2057-2066.

75. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, 2011, **12**, 2825-2830.