ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design

Shusen Chen, Zack Meyer, Brendan Jensen, Alex Kraus, Allison Lambert, and Daniel H. Ess*

Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84604 USA

*Email: dhe@chem.byu.edu

Keywords: Catalysis, ligands, ligand library, computational design, machine learning

Abstract

Computational catalyst design requires identification of a metal and ligand that together result in the desired reaction reactivity and/or selectivity. A major impediment to translating computational designs to experiment is evaluating ligands that are likely to be synthesized. Here we provide a solution to this impediment with our ReaLigands library that contains >30,000 monodentate, bidentate (didentate), tridentate, and larger ligands cultivated by dismantling experimentally reported crystal structures. Individual ligands from mononuclear crystal structures were identified using a modified depth-first search algorithm and charge was assigned using a machine learning model based on quantum-chemical calculated features. In the library ligands are sorted based on direct ligand-to-metal atomic connections and on denticity. Representative principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) analyses were used to analyze several tridentate ligand categories, which revealed both the diversity of ligands and connections between ligand categories. We also demonstrated the utility of this library by implementing it with our building and optimization tools, which resulted in the very rapid generation of barriers for 750 bidentate ligands for Rh-hydride ethylene migratory insertion.

Introduction

While far from routine, quantum-chemical calculations, typically density functional theory (DFT), are now being used to design (or redesign) molecular catalysts.[1,2,3] Recently there has also been a surge in uniting DFT calculations with machine learning and related data science approaches for catalyst predictions.[4,5,6,7] The key to successful computational catalyst design is identification of a specific ligand framework mounted on a transition metal that will result in the desired reaction reactivity and/or selectivity. Additionally, for successful theory-to-experiment handoff designs should consider ligand synthesis in the context of previously prepared ligands. Stated another way, computational designs with highly exotic or likely-to-be unstable ligands are unlikely to be synthesized.

With the recognition that homogeneous molecular catalyst design can be framed as a ligand design problem, several groups have established methods for ligand property (without the metal center) analysis. This viewpoint hinges on the assumption that analysis of ligand properties (or relative properties) will translate to reaction reactivity and selectivity when mounted on a metal center and combined with other ligands. Additionally, there is the assumption of property transference from one group of ligands to another. Harvey and Fey pioneered evaluating ligand properties.[8,9,10,11,12,13,14] As a recent example, Fey has reported an extension of the ligand knowledge base for new descriptors of bidentate ligands and principal component analysis (PCA) provides information about ligand relationships.[15,16] Fey also recently reported a database of ligand only DFT-calculated descriptors for designing dirhodium ligands.[17]

Alternative to analysis of ligand only properties, DFT calculations can be used to directly evaluate reactivity and selectivity of possible catalyst ligands, typically through optimization of transition-state structures and intermediates. This type of approach requires mechanistic details to identify key reactivity and selectivity controlling intermediates and transition states. An advantage of this approach is that the calculations can directly provide qualitative insight into reactivity and selectivity or in certain cases with careful modeling can provide quantitatively accurate predictions. Our group recently applied this approach

for the design of Cr phosphine imine ethylene tetramerization catalysts.[18] However, major limitations of this catalyst design approach is the human time required to construct new ligands and the computer time required for DFT calculations. Another key limitation is the lack of ligand libraries that would be ideal to begin searching chemical space and that would directly translate to a plausible experimental system that can be straightforwardly synthesized.

There are now programs that automate building, optimizing, and analyzing DFT structures (ground state and transition states). For example, our Mason program automates building and optimizing transition states using a ligand library.[19,20] The most notable general program to obtain transition states through automated functionalization of specific ligand sites is the AARON toolkit.[21] Wheeler has very recently developed a powerful suite of tools for structure building, transition-state optimization, and job control, which is called QChASM.[22] Pidko and Sinha's ChemSpaX program can also perform ligand derivatization for organometallic complexes.[23] There are also a few other programs that can also rapidly build transition-metal complexes, most notably Kulik's molSimplify.[24,25,26] Other notable programs include Jensen's DENOPTIM that provides the general design of compounds, including organometallic structures.[27,28] While there are now programs that can automate catalyst building there remains a lack of extensive ligand libraries. Therefore, our goal was to develop a ligand library based on experimentally reported ligands that when used for catalyst design harmonize the handoff from computation to experiment. Additionally, an experimentally based ligand library would enable the comparison of hypothetical chemical space with currently reported synthetic chemical space.

We were inspired by Balcell's tmQM database, and more recently the updated tmQMg database, that provides a set of >80,000 single transition metal complexes extracted from the Cambridge Structure Database (CSD).[29,30] We realized that this database of structures contains most of the experimentally reported ligands mounted on a single transition metal, which if extracted and classified would provide an extremely useful starting point to computationally design new catalysts (Figure 1). While this manuscript was in preparation a similar idea was advanced by Corminboeuf who reported the program cell2mol that converts crystal structures to individual molecules.[31] Here we report the development of a modified depth-first search algorithm for ligand identification and a machine learning based model using simple quantum-chemical calculation features for charge assignment. The ligands were sorted based on direct ligand-to-metal atomic connections and denticity. This experimentally based library of >30,000 ligands is called ReaLigands. To demonstrate some of the ligand categories and information provided by the ReaLigands library we used dimension reduction techniques to analyze representative tridentate ligand categories. To demonstrate the utility of this library for catalyst screening, we used this library in conjunction with our automated building and optimization tools to generate 750 bidentate Rh-H ethylene migratory insertion transition states and barriers.

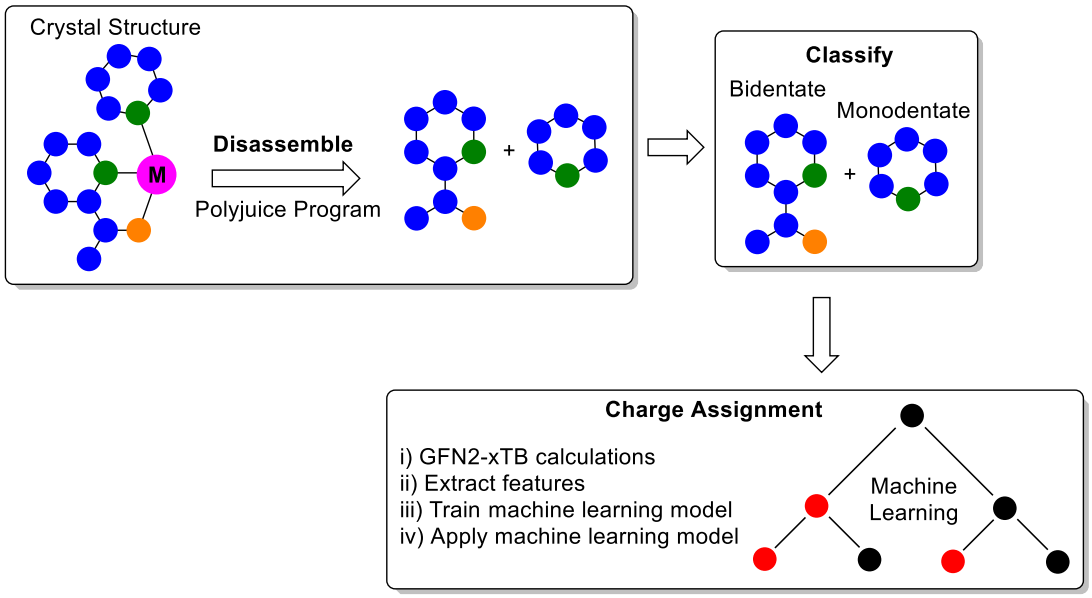


Figure 1. Diagram showing the workflow that created the ReaLigands ligand library. This involved first the development of an algorithm to dismantle experimentally reported mononuclear transition metal complexes. Each ligand was then classified according to ligand-to-metal atomic connections and denticity. Ligand charge assignment was performed using quantum-chemical calculations combined with a machine learning classification model.

Results and Discussion

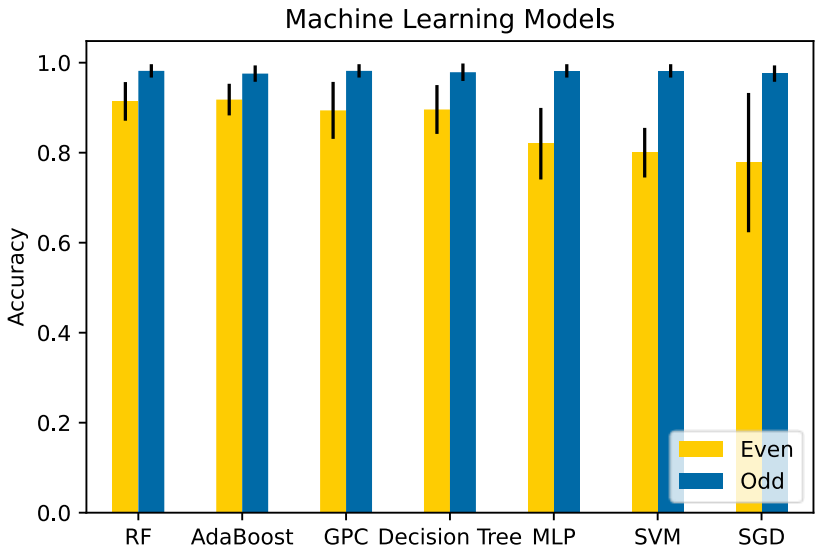
Because computational catalyst design hinges on identifying specific ligands to determine reactivity and selectivity it is surprising that there are only a few previously reported ligand library or databases, and they are often specific to a particular type of catalyst. For example, Kulik recently explored ligand additivity relationships in octahedral Fe mononuclear complexes[32] and in doing this developed a ligand library from octahedral complexes in the CSD.[33] After symmetry analysis this ligand library was used to generate ~17,000 octahedral complexes. Kulik has also presented ~5,000 ligands with corresponding charges.[34,35,36] During the writing of this manuscript, Corminboeuf reported the program cell2mol that converts crystal structures to individual molecules.[31] Similar to our work, this program has the capability of identifying and detaching ligands from a metal center. Cell2mol was able to successfully interpret about 75% of the CSD entries with a single metal. With a focus on analyzing ligands from only Cr, Mn, Fe, Co, Ni, Cu, Ru, and Re metal centers a ligand database of about 13,000 structures was reported. The ligands generated from cell2mol were not classified and we could not easily use them with our automated catalyst building program.

The most well-known large ligand database is the Kraken library that provides a massive amount (~300,000) of mostly hypothetical monodentate P^{III} structures.[37,38,39,40] This ligand database provides both structures as well as machine learning calculated properties and was developed based on the philosophy of representing the discrete ligands as continuous variables. As an important general demonstration of the utility of ligand libraries, Kraken has been used to predict cross-coupling reactions.[41] A somewhat related virtual ligand-assisted screening procedure using phosphines was recently developed by Matsuoka and Maeda.[42]

To begin we developed a program (called Polyjuice) that can identify all ligands coordinated to a single metal center of a mononuclear structure. Briefly, Polyjuice first reads the connectivity throughout a metal-ligand structure and each molecule is considered as a graph data structure with all atoms as vertices and bonds between atoms as edges. A modified depth-first search algorithm was then used to identify and extract each ligand. Our modification of the depth-first search algorithm added a condition to handle polydentate ligands because the traversal will return to the metal through a different connecting atom.

Importantly, Polyjuice was only effective with proper pre-processing of each structure from Balcell's tmQM and tmQMG databases.[29] This involved using Openbabel's molecule class to generate atomic connectivities with a modification for several atoms, such as silicon. Duplicate structures were then filtered out using the Openbabel CLI. After extraction, Polyjuice categorized each ligand based on direct ligand-to-metal connecting atoms and denticity.

The assignment of ligand charges is nontrivial because each complex in the tmQM and tmQMG databases does not have an assigned metal center oxidation state. To assign a charge to each ligand, which is necessary for each ligand to be easily integrated with software that automates building metal-ligand structures and performing quantum-chemical calculations, we assumed ligands are closed shell and only range in charge from 3- to 1+. Based on these assumptions we then carried out GFN2-xTB[43] single-point energy calculations for each ligand at each of these five charge states (3-, 2-, 1-, 0, and 1+). Based on these xTB calculations, for each charge state, we extracted the HOMO-LUMO gap, internal force, and SCF iteration number, which were used as features to build a machine learning model to assign ligand charges. These features were selected based on the hypothesis that the correct charge assignment will correlated with the largest HOMO-LUMO gap, small internal forces, and small number of SCF iterations. To ensure diversity in the training ligand set we used RDkit[44] to calculate the circular morgan fingerprints[45] of all available ligands and then calculate the dice similarity score. Model training was done by manual assignment of charge for 838 ligands that were a mixture of monodentate, bidentate, and tridentate ligands. In the training ligand dataset, there were four ligands with 3- charge, 112 ligands with 2- charge, 328 ligands with 1- charge, 391 ligands with 0 charge, and 3 ligands with 1+ charge, which represents a statistically relevant relative number of ligands for each type. Ligands were separated by charge parity of odd charge and even charge. Several classification models were generated using scikit-learn,[46] and their performance is plotted in Figure 2 (top). The random forest classification model performed best. A separate random forest model with cross validation was created for each set charge parity ligands. The average accuracy across the cross-validation trials was 90% for even charge ligands and 97% for odd charge ligands. We also characterized the accuracy of the random forest classification models in different subsets of the data. This was done by using the random forest vote percentage as a representation of the assigned charge confidence. Figure 2 (bottom) shows that for even parity charges large intervals have accuracy that exceeds the general model accuracy.



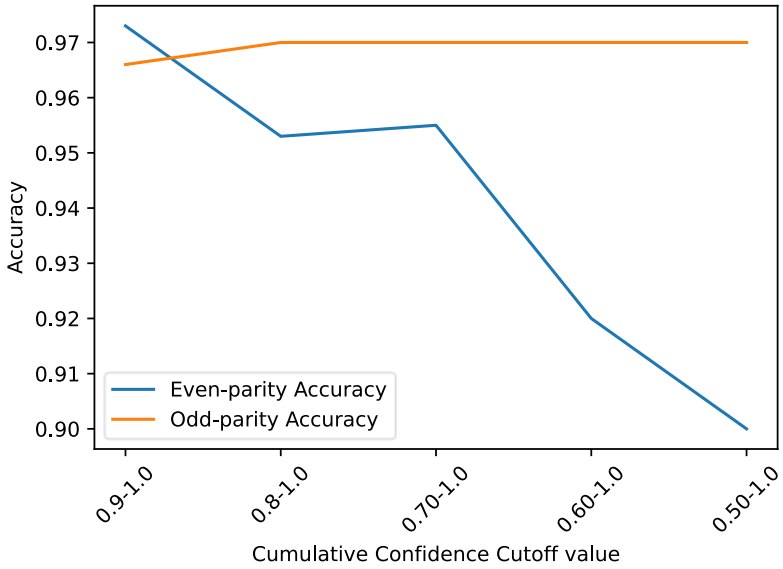
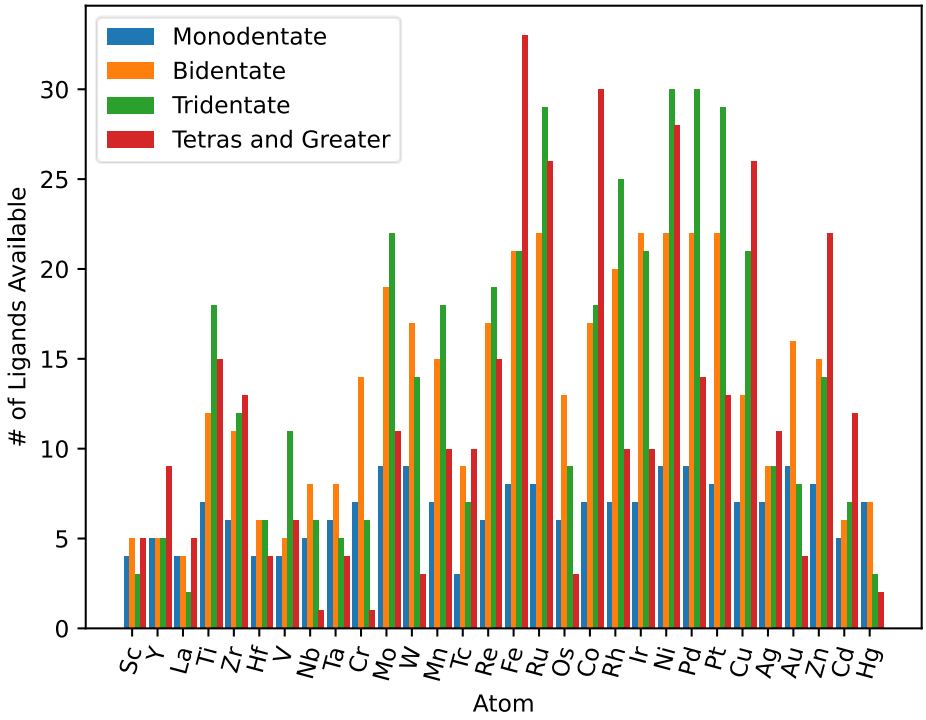


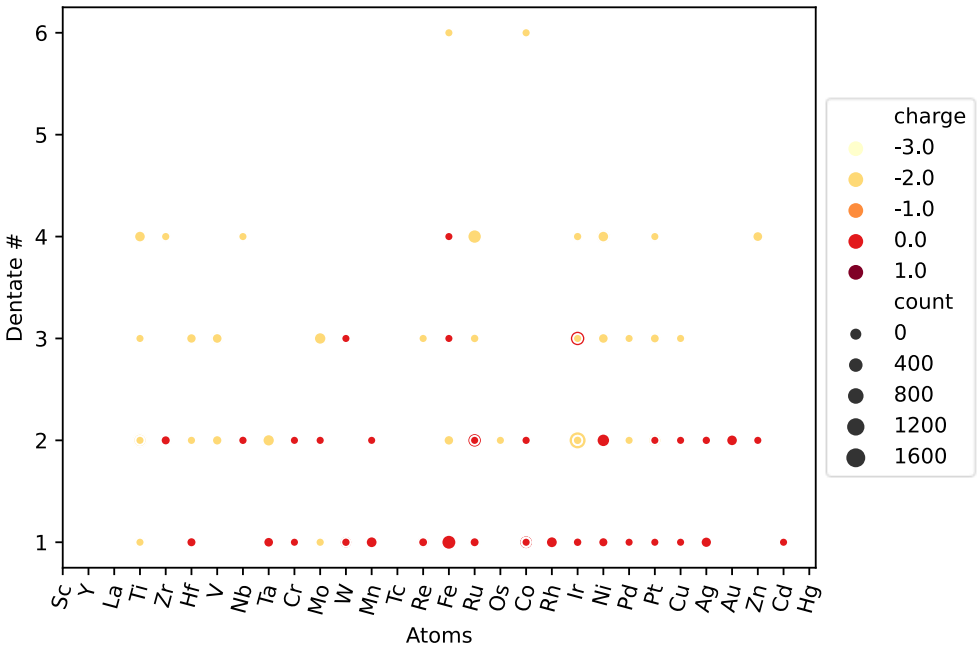
Figure 2. Top: Performance of several classifier machine learning models for ligand charge assignment. RF = Random Forest. GPC = Gaussian Process. MLP = Multi-layer Perceptron. SVM = Support Vector Machines. SGD = Stochastic Gradient Descent. Bottom: Graph representing the accuracy of the prediction based on the random forest model's confidence at a given interval.

We then applied these trained random forest classification models to charge assignment for the ~30,000 extracted ligands. This was done by calculating the xTB HOMO-LUMO gap, internal force, and SCF iteration number for all ligands and then applying the models to determine charge assignment based on the feature values. While the majority of ligands have a very high confidence for charge assignment a few ligands have low confidence, and these ligands have a flag in the file that enables users to exclude these ligands if desired. All ligand files have xyz coordinates, total charge, labeled metal connectivity, and the original CSD code. All ligand files are in folders sorted by coordination denticity and the atom type for metal connection.

It is useful to have a general view of the types of ligands in the ReaLigands library. Nearly 1/3 of the ligands are monodentate. The most common monodentate ligand-to-metal connection occurs with a carbon atom (5069 ligands). The second and third most monodentate ligand-to-metal connections are nitrogen (2625) and phosphine (1145). The top of Figure 3 plots the count of unique monodentate, bidentate, tridentate, and tetradentate ligands extracted from each transition metal center from Sc to Hg. The middle of Figure 3 plots the number of ligand connecting atoms to the transition metal center and color codes the charges. The bottom of Figure 3 plots the denticity and charge for the ligands extracted from each transition metal center.

Figures 4 and 5 provide an overview of the number of atoms in tridentate ligands. Figure 4 shows the total number of atoms (green dots) and non-hydrogen atoms (red dots) plotted versus the originating transition metal center. These plots show that for about half of the transition metal atoms there is a nearly continuous size of ligands ranging from 10-90 total atoms. Figure 5 provides a similar analysis of total ligand atoms and non-hydrogen atoms but plotted versus the type of tridentate ligand (the three atoms that directly connect to the metal center) rather than the originating transition metal center.





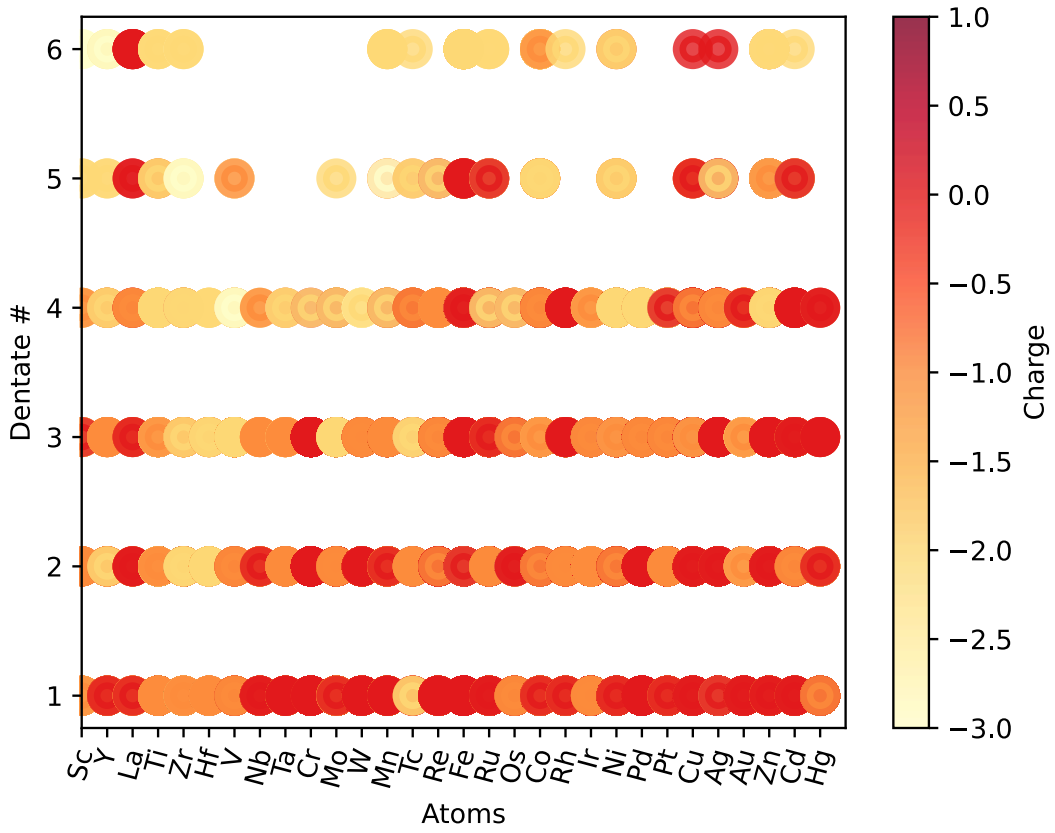


Figure 3. Top: Histogram showing the number of monodentate, bidentate, tridentate, and tetradentate/larger than tetradentate ligands of versus transition metal center that each ligand was extracted from. Bottom: Relationship between charge assignment, denticity (connecting number of atoms) and transition metals each ligand originated from.

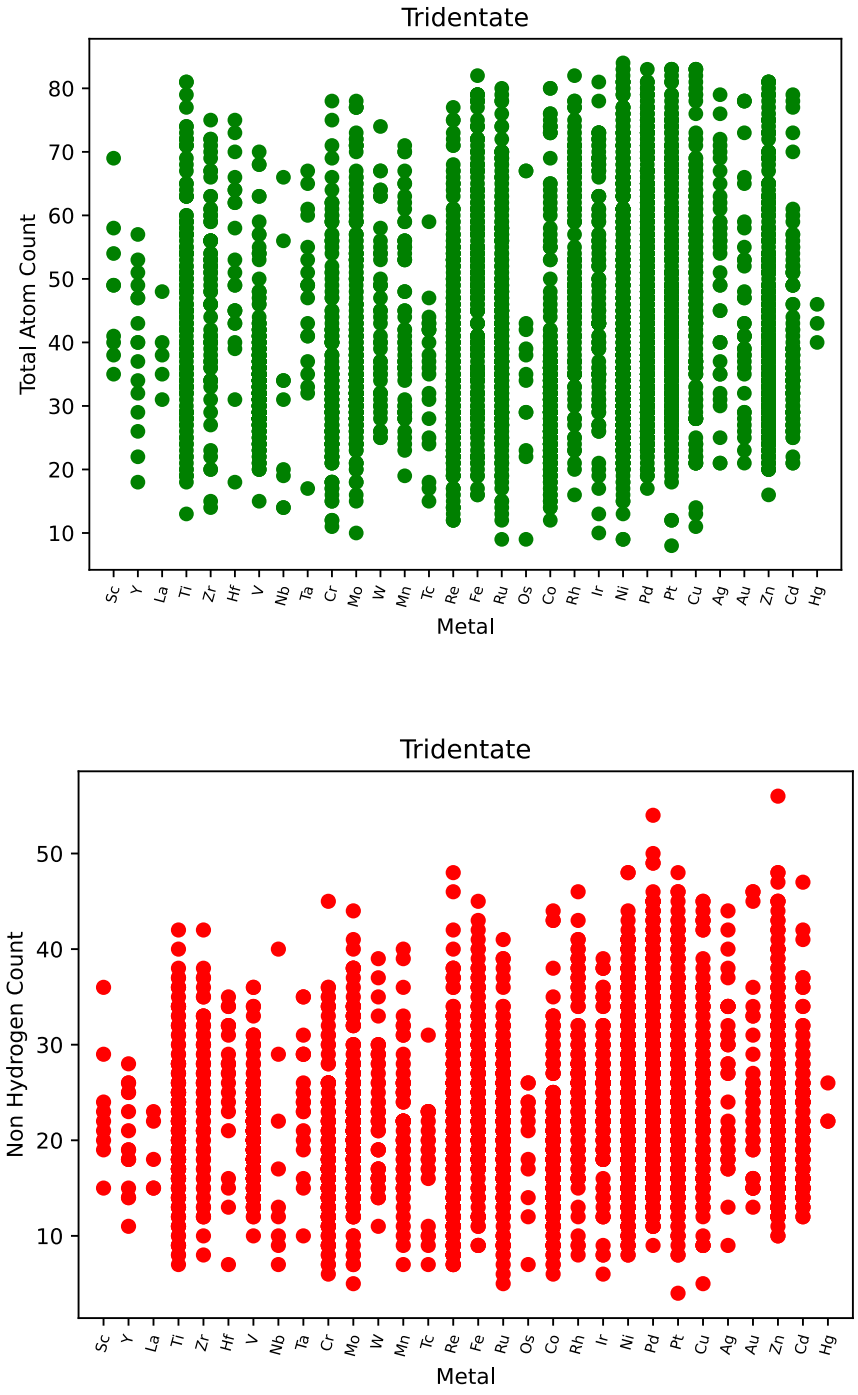


Figure 4. Top: Green dots represent the total number of atoms in tridentate ligands plotted versus originating transition metal center. Bottom: Red dots represent the total number of all non-hydrogen atoms in tridentate ligands plotted versus originating transition metal center.

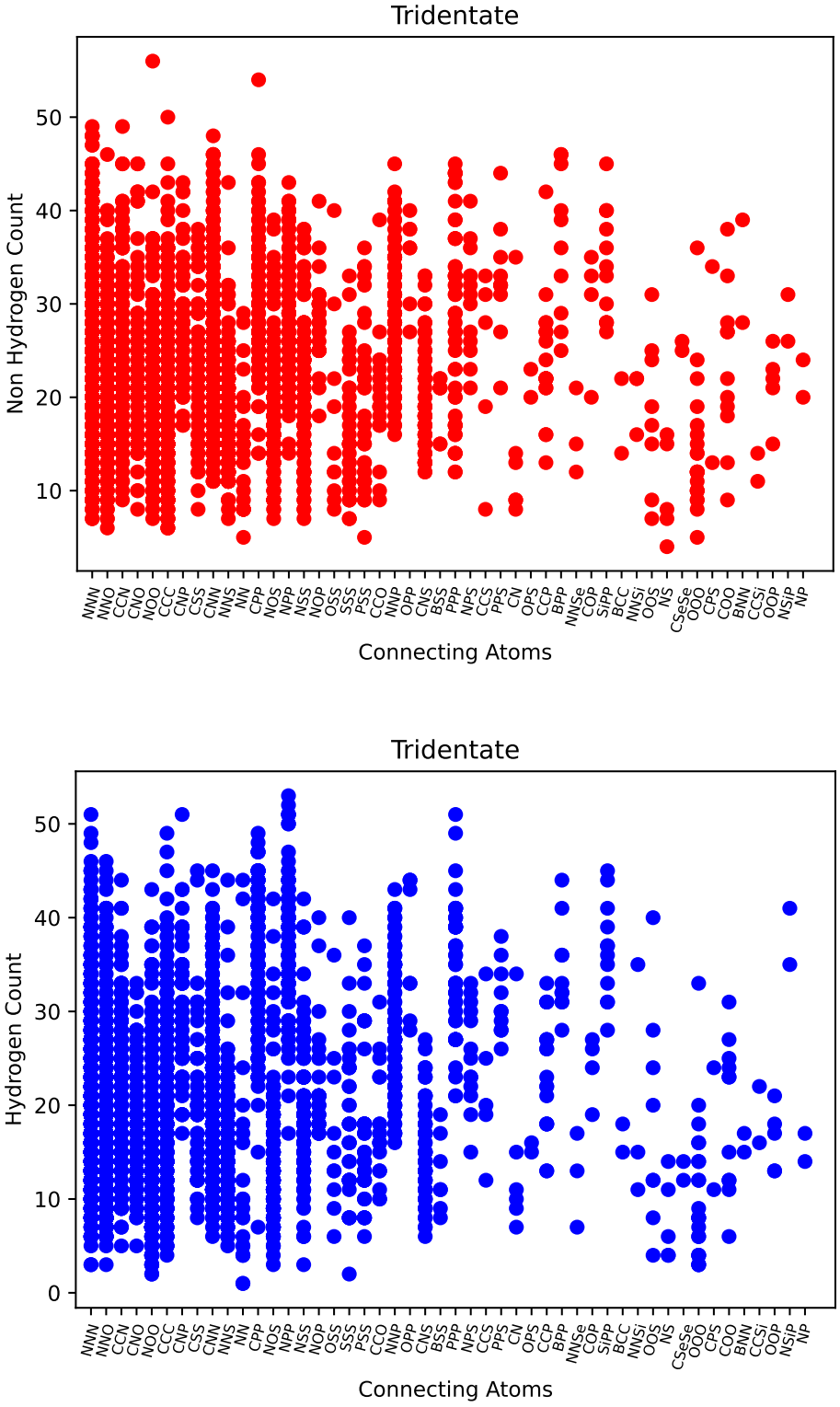


Figure 5. Top: Red dots represent the total number of all non-hydrogen atoms in tridentate ligands plotted versus the type of tridentate ligand (connecting atoms to metal center). Bottom: Blue dots represent the

total number of all hydrogen atoms in tridentate ligands plotted versus the type of tridentate ligand (connecting atoms to metal center).

Dimensionality reduction is often a useful way to visualize property space of ligands. Fey and coworkers have shown the value of condensing several molecular descriptors to only a few dimensions through PCA, especially to visualize connections and differences between different ligands. Here we used PCA and UMAP to analyze several categories of tridentate ligands. Figure 6 shows PCA and UMAP plots of 11 different tridentate ligand sets and about 300 total ligands within the ReaLigands library. These ligands have C, N, O, and P atom direct connections with the metal center. The PCA and UMAP analyses were performed in sci-kit learn[46] using features generated by Mordred.[47] To visualize the PCA analysis, the top of Figure 6 shows a plot of the two most important components and color coded for the type of tridentate ligand. The bottom of Figure 6 shows a similar component plot with UMAP. While both analyses demonstrate separation between labeled subclasses of tridentate ligands, UMAP provides the most segregation. For example, ligands with a metal connection of type NNN or NNO are distanced from ligands with connection type PPP and NPP. However, while there is some separation between different types of tridentate ligands there is often a smooth connection and overlay between multiple ligand subclasses. There are also some ligands that span nearly the entire spectrum. For example, CCC type ligands are often found throughout the entire chemical space, which likely illustrates the flexibility for designing these types of ligands with different properties.

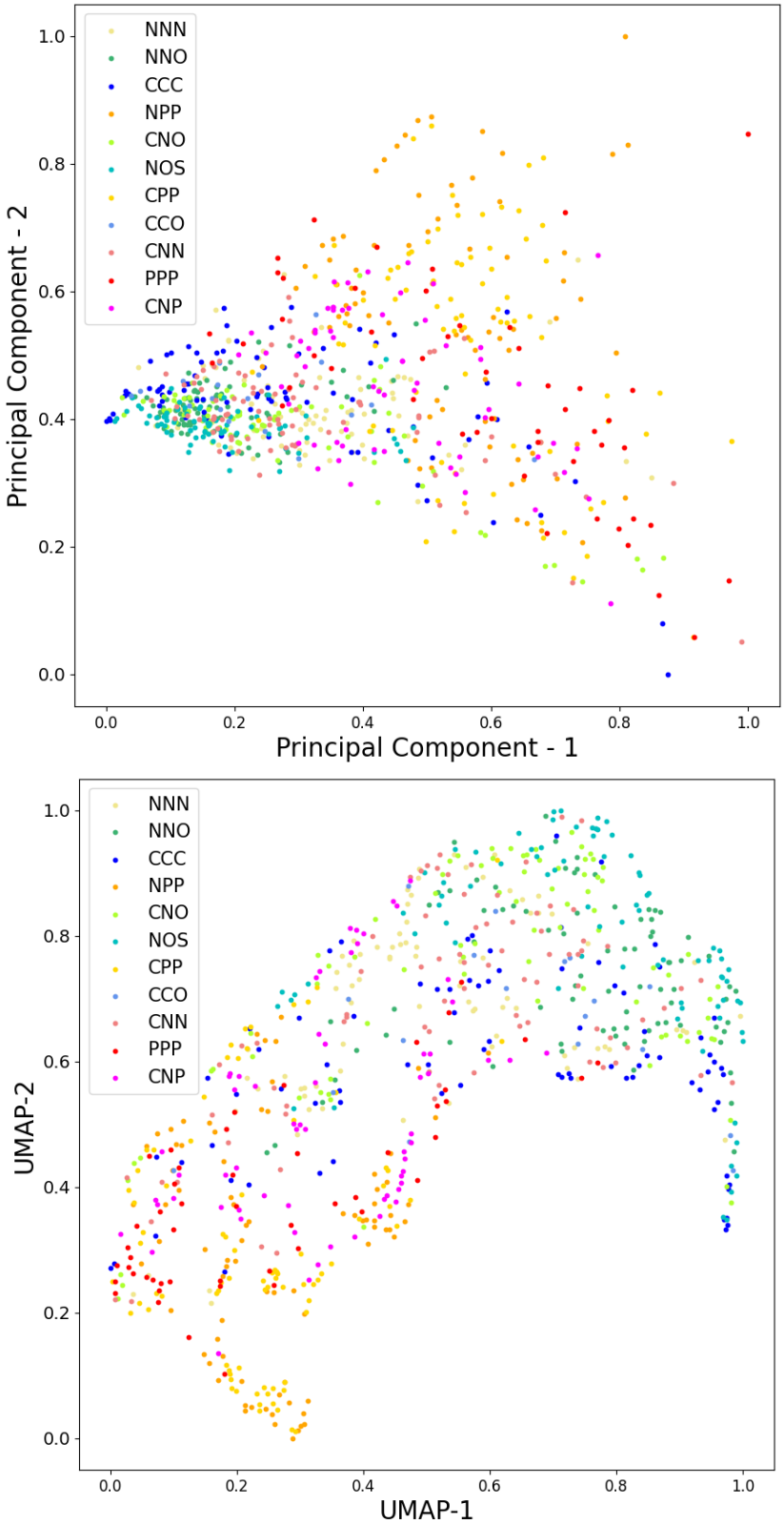
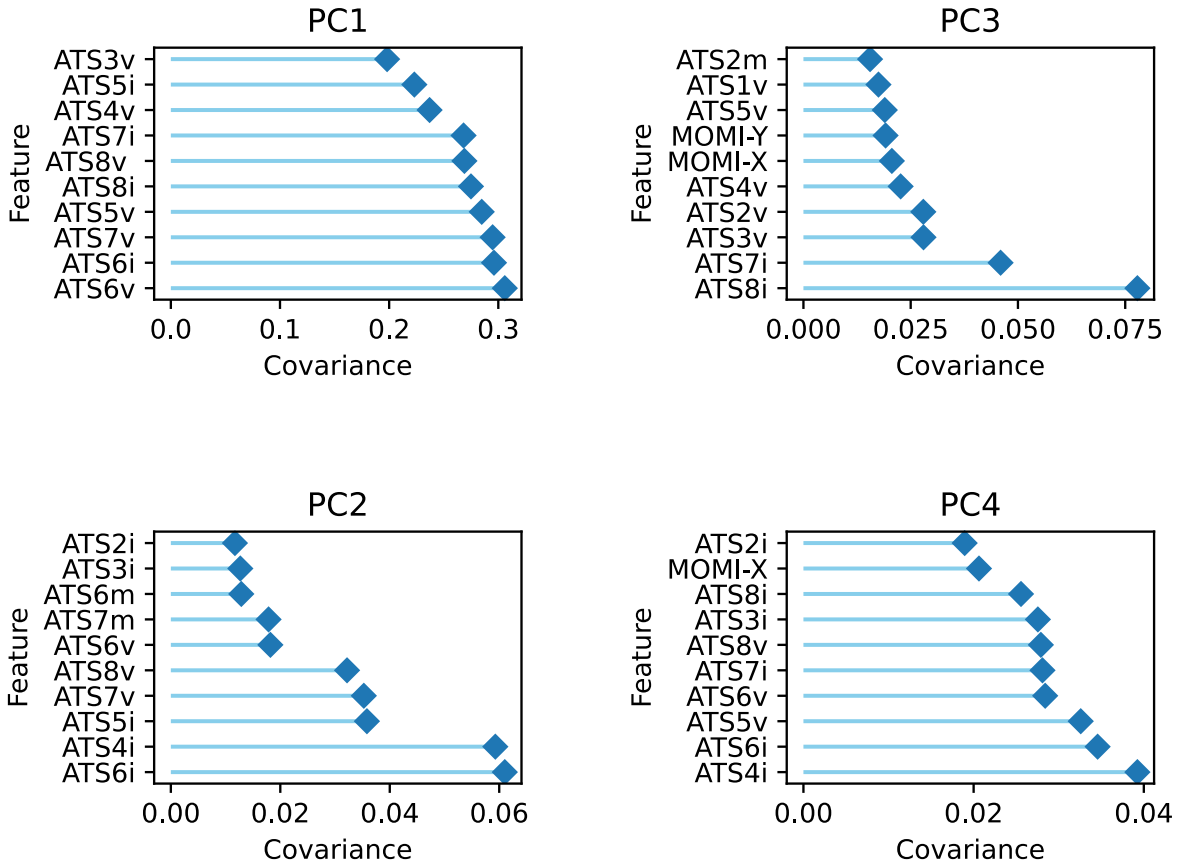


Figure 6. Top: PCA analysis of ~300 tridentate ligands based on features generated with Mordred. Bottom: UMAP analysis of ~300 tridentate ligands based on features generated with Mordred. The axes

represent the two most important components of these normalized reduced chemical space dimensions. Color coding is according to direct atomic connections with a metal center.

Figure 7 shows the breakdown of feature contributions to the principal components used for PCA and UMAP analyses. ATS (autocorrelation of topological structure (moreau-broto)) features dominate the first two principal components. The lag 6 weight is likely due to the tridentate ligand scaffold connections. The bottom of Figure 6 plots the ATS values versus the PCA and UMAP components. These plots demonstrate the variability in ATS values versus components that provides continuous separation of ligands.



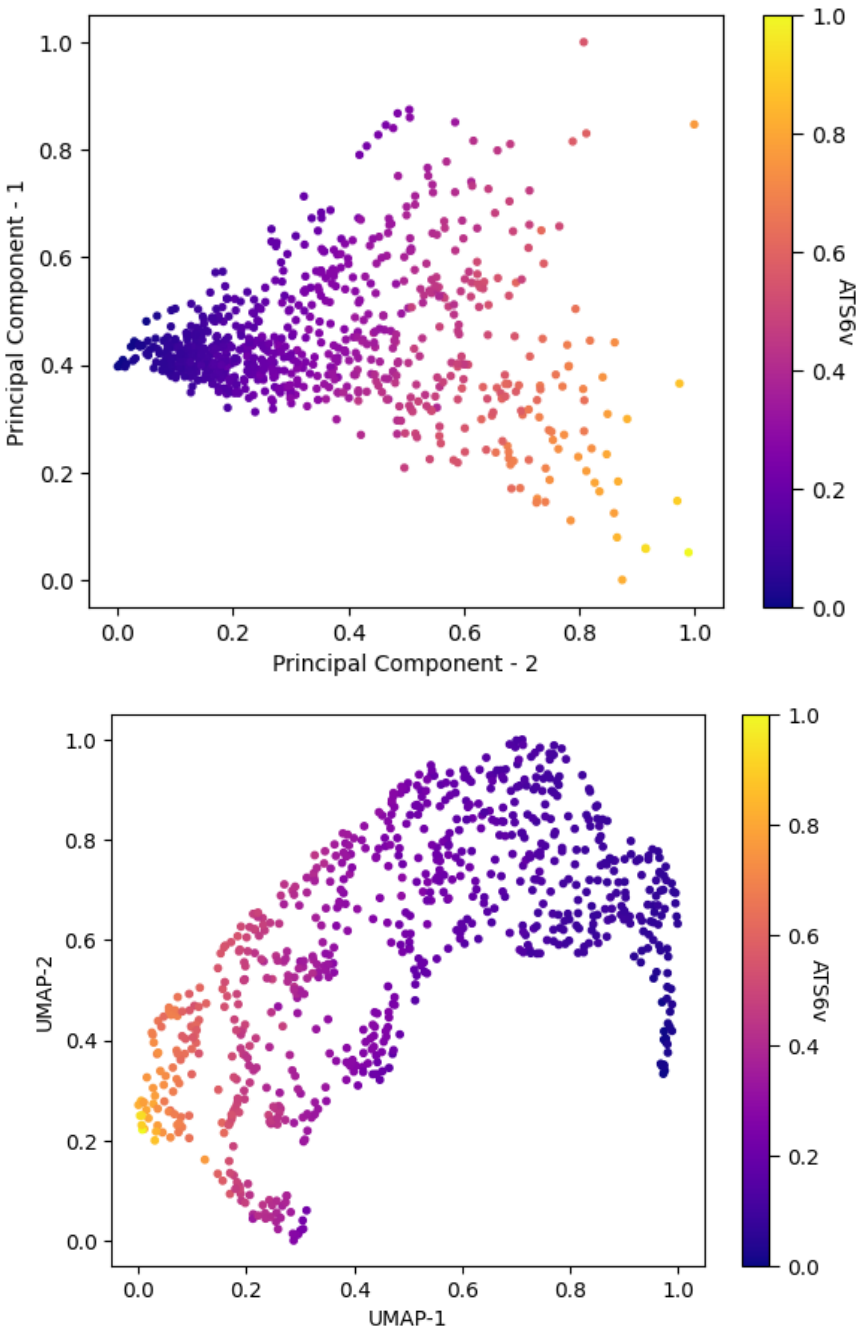


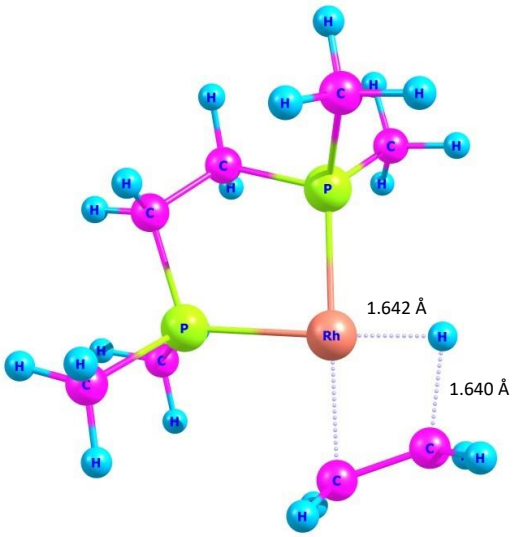
Figure 7. Top: Descriptor importance of for reduced components. Middle: Trends of ATS6 descriptor values for PCA components. Bottom: Trends of ATS6 descriptor values for PCA components. The ATS6 descriptor values have been normalized.

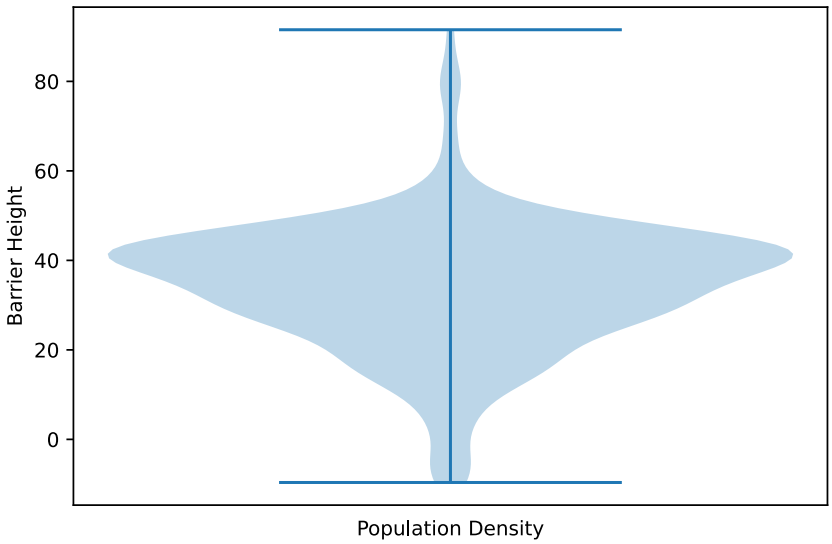
For us the most important aspect of the ReaLigands library is the ability to use it for rapid evaluation of catalysts with quantum-chemical calculations. Therefore, to demonstrate the utility of this library we used a subset set of the bidentate ligands to generate many Rh-H ethylene migratory insertion transition states and ground states to generate barrier heights. This was done by importing the ReaLigands library into our program Mason[19] that uses Open Babel APIs to add ligands to a frozen transition-state core (one-by-one) and only requires defined connection points between the ligand and metal, which are

contained in the ReaLigands files. These built structures are then piped to an automated program to handle structure optimization in Gaussian 16.[48]

A diverse random subset of 750 bidentate ligands was generated by calculating the Morgan fingerprints and using a dice similarity score. The M06[49]/def2-SVP method was used to optimize all transition-state and ground-state structures using based on the Rh(H)(ethylene) core obtained from the fully optimized migratory insertion transition state for $(Me_2PCH_2)_2Rh(H)(ethylene)$ (Figure 8). Ground state calculations have ethylene removed and therefore barrier heights are relative to the optimized Rh-H intermediate and separated ethylene. All transition states were verified to have a singlet negative vibrational frequency with the correct reaction coordinate motion.

The middle plot in Figure 8 shows a very large range in barrier heights for these 750 bidentate ligands. Barriers range from slightly negative barriers due to a stabilized ethylene π complex to barriers >80 kcal/mol. Most bidentate ligands have barriers between 15-55 kcal/mol. The bottom of Figure 8 shows similar barrier height distribution plots broken categorized by the two atoms from the ligand directly coordinated to the Rh metal center. While not elaborated further here, these quantitative barrier heights can either be used to design a catalytic cycle or there can be an in-depth qualitative assessment of barrier heights with the assistance of machine-learning models.





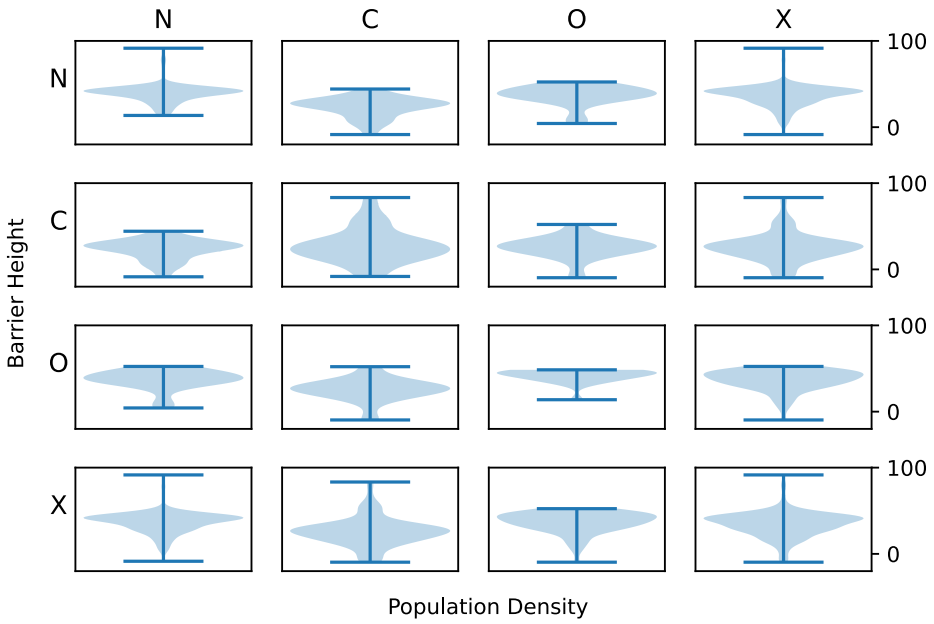


Figure 8. Top: Transition-state structure for migratory insertion involving $(Me_2PCH_2)_2Rh(H)(ethylene)$. Middle: Distribution of all 750 bidentate barrier heights for Rh-H migratory insertion. Barriers are in kcal/mol. Bottom: Distribution of barrier heights plotted by which atoms of the bidentate ligand directly coordinate to the Rh metal center. Barriers are in kcal/mol. X = Si, S, P, or Cl.

Conclusion

The ReaLigands library reported here represents an important tool for computational catalyst design using quantum-chemical structure and energy calculations. This >30,000 ligand library was generated by dismantling ligands from experimental crystal structure followed by classification based on denticity. A Random Forest machine learning model was used to make ligand charge assignments. PCA and UMAP analyses provided a glimpse at the diversity and range of ligand properties. We demonstrated how this library can be used in conjunction with automated building programs to evaluate a large collection of ligands for barrier heights. Perhaps most important, catalyst designs based on ligands from this library have a direct connection to experiment that will likely facilitate translation to the lab, which is the goal of computational-based catalyst evaluation.

Notes

The authors declare no conflict of interest.

Acknowledgements

We thank Brigham Young University and the Office of Research computing, especially the Fulton Supercomputing Lab. This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Catalysis Science Program, under Award # DE-SC0018329. We thank David Balcells and Bastian Skjelstad for helpful conversations and creation of their tmQM dataset.

References

1. Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354−2377.
2. Jover, J.; Fey, N. The Computational Road to Better Catalysts. *Chem. Asian J.* **2014**, *9*, 1714−1723.
3. Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Romina Oliva, R.; Scarano, V.; Cavallo L. Towards the online computer-aided design of catalytic pockets. *Nat. Catalysis.* **2019**, *11*, 872−879.

4. Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54*, 3136−3148.
5. Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.* **2019**, *58*, 10592−10606.
6. Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I. Shimizu, K-I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.
7. Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622−1637.
8. Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands. *Chem. Eur. J.* **2006**, *12*, 291–302.
9. Fey, N.; Harris, S. E.; Harvey, J. N.; Orpen, A. G. Adding value to crystallographically-derived knowledge bases. *J. Chem. Inf. Model.* **2006**, *46*, 912–929
10. Mansson, R. A.; Welsh, A. H.; Fey, N.; Orpen, A. G. Statistical modeling of a ligand knowledge base. *J. Chem. Inf. Model.* **2006**, *46*, 2591–2600.
11. Fey, N.; Orpen, A. G.; Harvey, J. N. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal–phosphorus bond. Coord. Chem. Rev. **2009**, *253*, 704–722.
12. Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational Descriptors for Chelating P,P- and P,N-Donor Ligands. *Organometallics* **2008**, *27*, 1372–1383.
13. Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561−6594.
14. Durand, D. J.; Fey, N. Building a Toolbox for the Analysis and Prediction of Ligand and Catalyst Effects in Organometallic Catalysis *Acc. Chem. Res.* **2021**, *54*, 837−848.
15. Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E. Mapping the properties of bidentate ligands with calculated descriptors (LKB-bid). *Dalton Trans.* **2020**, *49*, 8169−8178.
16. Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29*, 6245–6258.
17. Green, A. L.; Tinworth, C. P.; Warriner, S.; Nelson, A.; Fey, N. Computational Mapping of Dirhodium(II) Catalysts. *Chem - Eur. J.* **2021**, *27*, 2402–2409.
18. Kwon, D-H.; Fuller, J. T. III; Kilgore, U. J.; Sydora, O. L.; Bischof, S. M.; Ess, D. H. Computational Transition-State Design Provides Experimentally Verified Cr(P,N) Catalysts for Control of Ethylene Trimerization and Tetramerization. *ACS Catal.* **2018**, *8*, 1138–1142.
19. Chen, S.; Nielson, T.; Zalit, E.; Skjelstad, B. B.; Borough, B.; Hirschi, W. J.; Yu, S.; Balcells, D. Ess, D. H. Automated Construction and Optimization Combined with Machine Learning to Generate Pt(II) Methane C–H Activation Transition States. *Topics in Catalysis* **2022**, 65, 312-324.
20. https://github.com/DanielEss-lab/ReaLigands
21. Guan, Y.; Ingman, V. M.; Rooks B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. J. *Chem. Theory. Comput.* **2018**, *14*, 5249–5261.
22. Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. QChASM: Quantum chemistry automation and structure manipulation. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1510.
23. Kalikadien, A. V.; Pidko, E. A.; Sinha, V. ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold. *Digital Discovery*, **2022**, 1, 8–25.
24. Ioannidis, E. I.; Gani, T. Z.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106–2117.

25. Janet, J. P.; Ramesh, S.; Duan C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513–524.
26. Friederich, P.; Gomes, G. D. P.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.
27. Foscato, M.; Venkatraman, V.; Jensen, V. R. DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59*, 4077–4082.
28. Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 767–780.
29. Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *12*, 6135–6146.
30. Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; Bin, R. D.; and Balcells, D. Deep learning metal complex properties with natural quantum graphs. *Digital Discovery*, **2023**, *2*, 618–633.
31. Vela, S.; Laplaza, R.; Cho, Y.; Corminboeuf, C. cell2mol: encoding chemistry to interpret crystallographic data. *npj computational materials*, **2022**, *8*, 188.
32. 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. *J. Chem. Phys.* **2022**, *157*, 184112.
33. Taylor, R.; Wood, P. A. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chem. Rev.* **2019**, *119*, 9427–9477.
34. Duan, C.; Ladera, A. J.; Liu, J. C.-L.; Taylor, M. G.; Ariyarathna, I. R.; Kulik, H. J. *J. Chem. Theory Comput.* **2022**, *18*, 4836–4845.
35. Janet, J. P. Duan, C.; Nandy, A.; Liu, F.; Kulik, H. J. Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design *Acc. Chem. Res.* **2021**, *54*, 532–545.
36. Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121*, 9927–10000.
37. Gensch, T.; Gomes, G. d. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.
38. M. Crawford, J. M.; Gensch, T.; Sigman, M. S.; Elward, J. M.; Steves, J. E. Impact of Phosphine Featurization Methods in Process Development. *Org. Process Res. Dev.* **2022**, *26*, 1115–1123.
39. Gillespie, J. A.; Dodds, D. L.; Kamer, P. C. J. Rational design of diphosphorus ligands – a route to superior catalysts. *Dalton Trans.* **2010**, *39*, 2751–2764.
40. Fey, N.; Howell, J. A. S.; Lovatt, J. D.; Yates, P. C.; Cunningham, D.; McArdle, P.; Gottlieb, H. E.; Coles, S. J. A molecular mechanics approach to mapping the conformational space of diaryl and triarylphosphines. *Dalton Trans.* **2006**, 5464–5475.
41. Gensch, T.; Smith, S. R.; Colacot, T. J.; Timsina, Y. N.; Xu, G.; Glasspoole, B. W.; Sigman, M. S. Design and Application of a Screening Set for Monophosphine Ligands in Cross-Coupling. *ACS Catal.* **2022**, *12*, 7773–7780.
42. Matsuoka, W.; Harabuchi, Y.; Maeda, S. Virtual Ligand-Assisted Screening Strategy to Discover Enabling Ligands for Transition Metal Catalysis. *ACS Catal.* **2022**, *12*, 3752–3766.
43. Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
44. https://www.rdkit.org/.
45. Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc.* **1965**, *5*, 107–113.
46. https://scikit-learn.org/stable/.

47. Moriwaki, H.; Tian, Y-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4.
48. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16, revision B.01; Gaussian, Inc., Wallingford CT, 2016
49. Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

Graphical Abstract