Pacific Northwest
NATIONAL LABORATORY

PNNL-35049

Machine Learning for Prediction of Thermodynamic Descriptors

September 2023

Eric S Wiedner
Benjamin A. Helfrecht
Jeremy D. Erickson
Nancy M. Washton

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-35049

Machine Learning for Prediction of Thermodynamic Descriptors

September 2023

Eric S Wiedner
Benjamin A. Helfrecht
Jeremy D. Erickson
Nancy M. Washton

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Our objective is to apply machine learning (ML) algorithms for the prediction of molecular catalysis descriptors from geometric properties derived from experimental crystallographic databases. Catalysis is often considered a “low-data” discipline that is poorly suited for ML methods. An exception is the extensive structural information that is available for molecular catalysts through the Cambridge Structural Database (CSD), which contains atomically precise molecular structures from X-ray diffraction analysis for >600K metal complexes. As a proof-of-principle, we targeted the prediction of hydricity, a thermodynamic property that provides understanding and control of catalytic hydride transfer. We built a training set composed of ~100 molecular complexes with a known hydricity and structural information from the CSD. This data set was converted into a machine-readable format using the smooth overlap of atomic positions (SOAP) representation and further labeled with simple electronic descriptors for the metal centers. Multiple different neural networks were trained on this data set, and the accuracy of the hydricity predictions ranged from < 2 kcal/mol to 20 kcal/mol. The accuracy of each model was highly sensitive to which compounds were in the train versus test set, underscoring the challenges associated with small and chemically diverse data sets. Finally, to further augment the data set, we attempted to experimentally measure several new hydricity values, however these experiments were unsuccessful due to undesired chemical reactivity of the selected complexes.

Acknowledgments

This research was supported by the PCSD Mission Seed, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

AI	artificial intelligence
CSD	Cambridge Structural Database
ML	machine learning
PDB	Protein Data Bank
ReLU	rectified linear unit
RMSE	root mean square error
SOAP	smooth overlap of atomic positions

Contents

Figures

1.0 Introduction

The burgeoning application of artificial intelligence and machine learning (AI/ML) in scientific research heralds an ostensible sea change in methodological frameworks. However, certain scientific domains, particularly catalysis, are often categorized as "low data" regimes, ostensibly rendering them suboptimal for the integration of AI/ML paradigms. It's crucial to recognize, though, that these "low data" realms offer a fertile ground for specialized AI/ML approaches designed for data-scarce environments. Anomalously, catalysis does offer substantial data sets in the form of structural information. Molecular catalysts, for instance, have been extensively cataloged in the Cambridge Structural Database (CSD), while enzymes find their detailed representations in the Protein Data Bank (PDB). The CSD alone comprises over 600,000 atomically precise molecular structures, discerned through X-ray diffraction analysis. This repository of structural and topographical data provides an invaluable foundation for the development of ML methodologies aimed at deciphering the intricate structure-function relationships intrinsic to molecular catalysts.

Catalytic efficacy is not merely an isolated attribute of the catalyst but a complex interplay between its inherent properties and the milieu within which the reaction occurs. Typically, delineating a direct correlation between catalytic activity and structural attributes is only feasible for complexes functioning under analogous conditions. A more stable parameter—less susceptible to environmental variables—is the thermodynamic bond strength, which serves as an insightful probe into a catalyst's intrinsic reactivity. In our current undertaking, we focus on hydricity—a salient thermodynamic property—as the linchpin for understanding and modulating catalytic hydride transfer reactions. Hydricity serves as an invaluable descriptor for the conceptual design of catalysts aimed at activating small molecules. Nevertheless, the synthesis and empirical determination of hydricity values for prospective catalysts is an arduous and time-intensive endeavor.

The advent of a rigorously calibrated ML tool would significantly truncate the need for labor-intensive electronic structure calculations. Such a tool would enable a rapid pre-screening of hydricity values across a diverse array of potential catalysts. This ML-based approach serves not just as an isolated utility for hydricity prediction but sets the stage for future explorations into ML-aided forecasting of other quintessential thermodynamic attributes and more elaborate catalytic systems, such as enzymes or surface-immobilized molecular complexes.

2.0 Results

2.1 Data Set

A data set containing ~100 transition metal hydride complexes was manually curated by identifying complexes having both a known hydricity value and X-ray crystallographic structural coordinates present in the CSD. To create a machine-readable input, the Smooth Overlap of Atomic Positions (SOAP) representation[1-2] was calculated for each molecule in the training set. The SOAP representation calculates the spatial distribution of atoms and their chemical identities as a series of vectors between neighboring atoms. While the SOAP representation preserves the structural geometry of a given molecule, it does not contain intrinsic electronic information. As a result, each molecular entry was manually labeled with simple electronic descriptors of the transition metal, including its atomic number, group and period in the periodic table, formal d e^- count, and the total charge of the metal hydride complex.

A potential pitfall in using structural coordinates from X-ray crystallography is that they represent a single "snapshot" of the molecular conformation. In solution, the ligands have a low energy barrier to bend and deform, and hence the complexes can access a wide range of geometric conformational space. To better capture this conformational complexity and further augment the amount of data, different conformations were generated for each entry using tight-binding-based molecular dynamics and metadynamics simulations.[3] The conformer generation aims to explore as much of the conformational space as possible for each complex, yielding between one and a few thousand conformers for each molecule, depending on the degree of ligand flexibility.

2.2 Neural Networks for Predicting Hydricity

We employed two distinct neural networks for our hydricity predictions: an autoencoder used to define a chemical and structural latent space, and a simple feed-forward network used to predict hydricities from the latent space. The rationale behind this design is that a latent space informed by all of the conformers across all of the (training) complexes may define a more relevant data representation from which property predictions can be made. The encoder portion of the autoencoder comprises two hidden layers with 75 and and 50 nodes, with input and output layers with 100 and 25 nodes. The decoder portion comprises the same architecture, but reversed in order. ReLU activation functions were applied to all but the final encoder and decoder layers. The autoencoder was trained for 200 epochs with the Adam optimizer, using a learning rate of 5.0×10^{-4} and a weight decay of 1×10^{-4}. The feed-forward network for predicting hydricities from the latent space comprised an input layer with 25 nodes, an output layer with a single node, and three hidden layers with 50, 50, and 25 nodes, respectively. ReLU activation functions and dropout with probability of 0.3 was applied to all but the final layer. The feed-forward network was trained for 500 epochs using the same optimizer and parameters as the autoencoder. To reduce the computational cost of the model, we reduced the dimensionality of the feature vectors through a farthest-point sampling scheme, retaining only the 100 most diverse features (out of more than 8000).

We trained the neural network on the donor forms of the organometallic complexes in our data set, except for those of $[HNi(MeIm(CH_2)_2PPh_2)_2]^+$,[4] $HRh(triphos)(PPh_3)$,[5] $HRh(triphos)(PMe_3)$,[5] and $[HPd(depp)_2]^+$,[6] which were set aside as a test set. The predictive capability of the neural network was examined using these four test molecules. The proximity of these test molecules to the training set in the latent space is shown in Figure 1 for two of the latent variables. Three of the test molecules are metal complexes with four phosphine ligands, a class which is wellrepresented in the training set. The hydricity predictions for these complexes ranged from 3.4 to 10.8 kcal/mol difference from the actual hydricity values. The fourth test molecular contained two phosphine ligands and two carbene ligands, which are underrepresented in the training set. As might be expected, this complex showed a higher error of 18.5 kcal/mol in its predicted hydricity. From these results, it can be concluded that the neural network has difficulty in predicting hydricities for complexes that inhabit regions of chemical space that are sparsely populated or are populated by complexes with dissimilar hydricities.

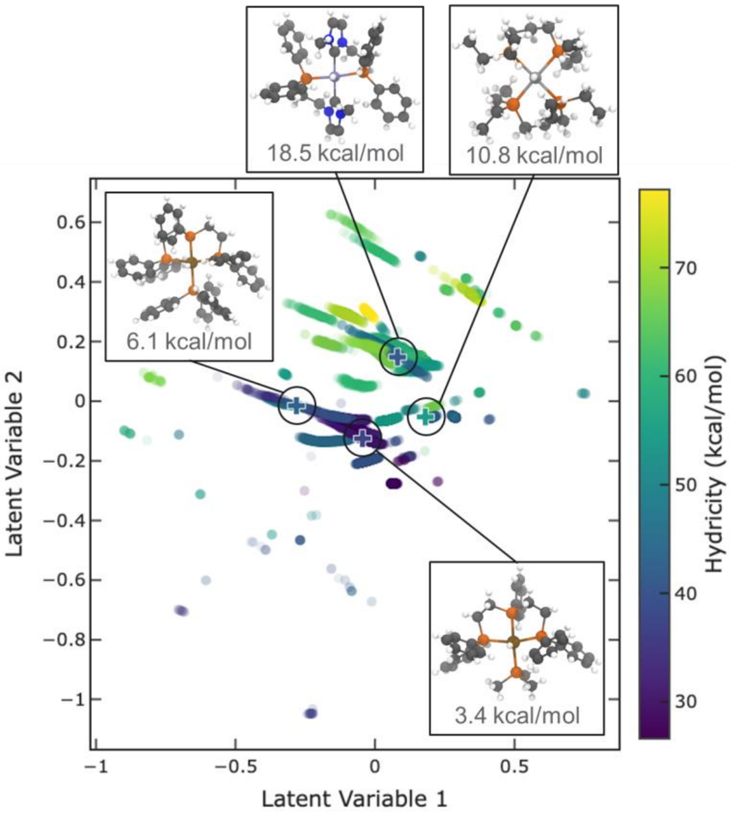


Figure 1. A neural network shows good to moderate success in predicting hydricity of test compounds, with the numerical value indicating the prediction error relative to the known value. Only 2 out of 25 dimensions of the latent space are shown for clarity.

To further examine the chemical boundaries for accurate prediction of hydricities, a regression-only neural network was constructed using 100 independent 90/10 train/test splits. The performance of this model varied widely based on which compounds were assigned to the training and test sets, with a minimum RMSE of 2.03 kcal/mol, a maximum RMSE of 10.25 kcal/mol, and an average of 6.04 kcal/mol. This behavior is consistent with a data-limited model, making it difficult to accurately predict hydricity values for complexes that lie outside of the chemical space occupied by the training set. In future studies, we plan to systematically examine the outliers in this model in order to better define the chemical features that lead to good or poor prediction of hydricity. We expect such an analysis will help provide indicators for molecules that are expected to be predicted well and to help identify improved model features that will improve the range of chemical space that can be predicted.

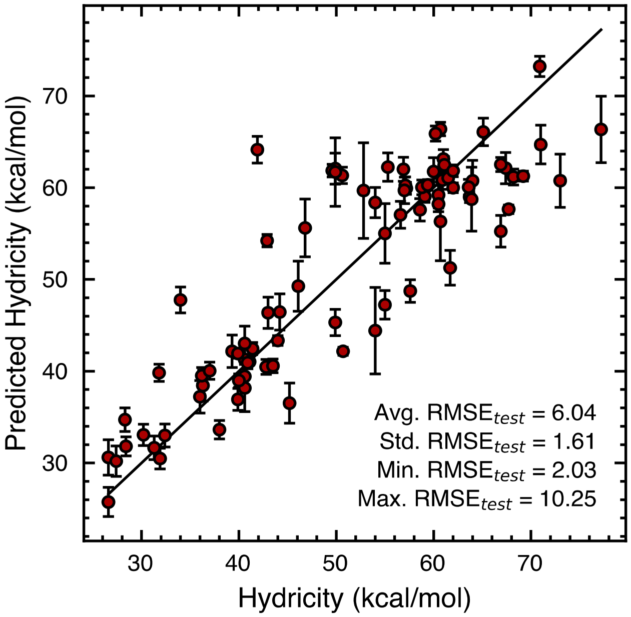


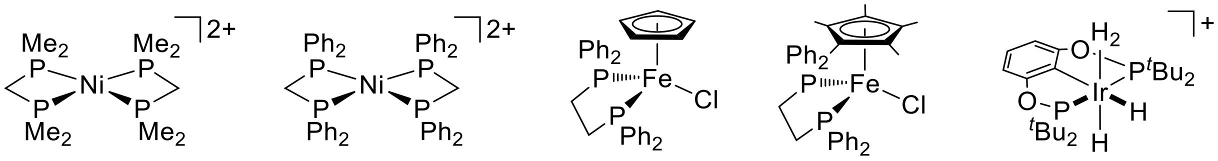
Figure 2. Output from regression-only neural network using 100 independent 90/10 train/test splits. The predicted hydricity values are the average for each of the models in which the compound was included in the test set.

2.3 Experimental Hydricity Measurements to Augment the Training Model

To augment the chemical space of the hydricity data set, we targeted four complexes to synthesize and experimentally measure the hydricity (Figure 3). The first two complexes, $[Ni(dmpm)_2]^{2+}$ and $[Ni(dppm)_2]^{2+}$, have phosphine ligands with a much smaller chelate angle (P-Ni-P) than the complexes in the training set, a structural feature which is known to have a strong influence on the hydricity.[7] These complexes were readily synthesized by reaction the precursor $[Ni(CH_3CN)_6]^{2+}$ with two equivalents of the commercial diphosphine ligands. To measure the hydricities, these complexes were treated with dihydrogen (H_2) and a series of organic bases with known basicity values in order to measure the equilibrium between the Ni(II) and Ni(II)H states. However, we were unable to identify a base that was strong enough to generate the Ni(II)H without also binding to the Ni center. Base binding results in the formation of high-spin Ni(II) species that give rise to broadened and paramagnetically shifted resonances in the NMR spectra, thereby preventing accurate calculation of the reaction equilibrium.

Two iron complexes, CpFe(dppe)Cl and Cp*Fe(dppe)Cl, were synthesized and tested for hydricity measurements. In a modification of a literature procedure,[8] Fe(dppe)Cl2 was reacted with either CpLi or Cp*Li to afford the target complexes. In acetonitrile solution, the solvent displaces the chloride ligand on the Fe complexes. Due to tight binding of the acetonitrile ligand, attempts to generate Fe hydride species by treating with H_2 and organic base were unsuccessful and resulted in one of three outcomes: (i) no reaction was observed, (ii) the base coordinated to Fe by displacing the dppe ligand, or (iii) the complex decomposed into multiple unidentified species. In principle the hydricity of these complexes could be measured in a non-coordinating solvent like tetrahydrofuran, but this was not attempted since the neural network was not trained on hydricity values measured in non-coordinating solvents.

An iridium complex, $[(H)_4Ir(POCOP)]^+$, was synthesized by literature methods[9] and was tested for hydricity measurement. Similar to the Fe complexes, acetonitrile was observed to bind tightly to the Ir complex by displacing H_2. Strong organic bases were observed to deprotonate the acetonitrile ligand, again precluding measurement of the hydricity in acetonitrile.



$[Ni(dmpm)_2]^{2+}$ $[Ni(dppm)_2]^{2+}$ CpFe(dppe)Cl Cp*Fe(dppe)Cl $[(H)_4Ir(POCOP)]^+$

Figure 3. Complexes that were targeted for synthesis and experimental measurement of hydricity.

3.0 Conclusion

In the present study, we pursued a multifaceted approach to model and predict the hydricity of transition metal hydride complexes, employing neural networks alongside computational simulations and experimental measurements. The data set comprised ~100 transition metal hydride complexes with known hydricity values and X-ray crystallographic structural coordinates were enriched with additional conformational variants via tight-binding-based molecular dynamics and metadynamics simulations. This provided a more nuanced exploration of the complexes' conformational space, thereby addressing the intrinsic limitation associated with X-ray crystallography-derived structures.

The neural network architecture employed was bifurcated: an autoencoder for latent space definition and a feed-forward network for hydricity predictions. The results suggest moderate predictive accuracy; however, the model's performance exhibited sensitivity to the composition of the training and test sets. This was particularly noticeable for complexes that were underrepresented or occupied sparsely populated regions in the chemical space, corroborated by discrepancies in hydricity prediction errors and Root Mean Square Error (RMSE) variations in regression-only neural networks.

Experimental endeavors to augment the training data set unveiled complications. Particularly, issues related to base binding interfered with the hydricity measurements, highlighting the pertinence of the electronic environment in governing hydric properties. This underscores the need for further experimental work that not only expands the compositional diversity of the data set but also addresses the limitations related to the choice of base and ligand conformation in the measurement of hydricity.

4.0 References

1. Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-density representations for machine learning. *J. Chem. Phys.* **2019,** *150*.
2. Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013,** *87*, 184115.
3. Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020,** *22*, 7169-7192.
4. Galan, B. R.; Wiedner, E. S.; Helm, M. L.; Linehan, J. C.; Appel, A. M. Effects of Phosphine–Carbene Substitutions on the Electrochemical and Thermodynamic Properties of Nickel Complexes. *Organometallics* **2014,** *33*, 2287-2294.
5. Fernandez, W.; Hudson, D. B.; Arachchilage, H. J.; Zall, C. M. When 3+1 is Less Than 2+2: Surprisingly Moderate Hydricities in Heteroleptic Rhodium Complexes Containing Triphosphine and Monophosphine Ligands. *Organometallics* **2023,** *42*, 1465-1476.
6. Raebiger, J. W.; Miedaner, A.; Curtis, C. J.; Miller, S. M.; Anderson, O. P.; DuBois, D. L. Using Ligand Bite Angles to Control the Hydricity of Palladium Diphosphine Complexes. *J. Am. Chem. Soc.* **2004,** *126*, 5502-5514.
7. Berning, D. E.; Miedaner, A.; Curtis, C. J.; Noll, B. C.; Rakowski DuBois, M.; DuBois, D. L. Free-Energy Relationships Between the Proton and Hydride Donor Abilities of $[HNi(diphosphine)_2]^+$ Complexes and the Half-Wave Potentials of Their Conjugate Bases. *Organometallics* **2001,** *20*, 1832-1839.
8. Long, E. M.; Brown, N. J.; Man, W. Y.; Fox, M. A.; Yufit, D. S.; Howard, J. A. K.; Low, P. J. The synthesis, molecular and electronic structure of cyanovinylidene complexes. *Inorg. Chim. Acta* **2012,** *380*, 358-371.
9. Göttker-Schnetmann, I.; White, P. S.; Brookhart, M. Synthesis and Properties of Iridium Bis(phosphinite) Pincer Complexes (*p*-XPCP)IrH_2, (*p*-XPCP)Ir(CO), (*p*-XPCP)Ir(H)(aryl), and $\{(p\text{-XPCP})Ir\}_2\{\mu\text{-}N_2\}$ and Their Relevance in Alkane Transfer Dehydrogenation. *Organometallics* **2004,** *23*, 1766-1776.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov