

UC Berkeley

UC Berkeley Previously Published Works

Title

3D, Structure-Based, Deep Learning Approach to Predict the Regioselectivity of Olefin Hydroformylations

Permalink

<https://escholarship.org/uc/item/88g6r4fp>

Journal

Journal of the American Chemical Society, 148(34)

ISSN

0002-7863

Authors

Hadler, Nicholas

Rinehart, N Ian

Elkin, Masha

et al.

Publication Date

2026-09-02

DOI

10.1021/jacs.6c11602

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

28 January 2026

A 3D, Structure-Based, Deep Learning Approach for Predicting the Regioselectivity of Transition-Metal Catalysis

Nicholas Hadler¹, N. Ian Rinehart^{2,1}, Masha Elkin¹, Jeremy Nicolai¹, Golsa Gheibi¹, Jiaqing Chen³, Matthew Avaylon⁴, Ross Maciejewski³, Gunther H. Weber⁴, Michael W. Mahoney^{5,4,6}, Talita Perciano⁴, John F. Hartwig^{2,1}

1. Department of Chemistry, University of California Berkeley
2. Division of Chemical Sciences, Lawrence Berkeley National Laboratory
3. School of Computing and Augmented Intelligence, Arizona State University
4. Scientific Data Division, Lawrence Berkeley National Laboratory
5. Department of Statistics, University of California at Berkeley
6. International Computer Science Institute

Abstract

Rational development of transition-metal catalysts, even when guided by theory and mechanistic knowledge, involves significant trial and error. Although ML offers the potential to accelerate catalyst discovery and optimization, accurately modeling the complex structures of catalysts and the multistep mechanisms by which they react remains challenging with the limited sets of data available. Olefin hydroformylation is a quintessential example of this challenge: its catalytic cycle involves many, often reversible, steps, and decades of study have not yielded reliable structure-selectivity relationships. We report Libra-ML, a 3D structure-based deep learning approach for predicting experimental outcomes of transition-metal catalyzed reactions. To demonstrate the ability of Libra-ML to model the outcomes of complex catalytic reactions, we predicted the regioselectivity of hydroformylation with terminal olefins catalyzed by rhodium complexes. Comparisons to existing methods demonstrate the state-of-the-art performance of Libra-ML and illustrate the importance of capturing 3D structure to predict experimental outcomes with molecular catalysts.

Keywords

Hydroformylation, Catalysis, Machine Learning

A 3D, Structure-Based, Deep Learning Approach for Predicting the Regioselectivity of Transition-Metal Catalysis

Nicholas Hadler¹, N. Ian Rinehart^{1,2}, Masha Elkin¹, Jeremy Nicolai¹, Golsa Gheibi¹, Jiaqing Chen³, Matthew Avaylon⁴, Ross Maciejewski³, Gunther H. Weber⁴, Michael W. Mahoney^{4,5,6}, Talita Perciano^{4,*}, and John F. Hartwig^{1,2,*}

¹ Department of Chemistry, University of California Berkeley, CA, United States.

² Division of Chemical Sciences, Lawrence Berkeley National Laboratory, CA, United States.

³ School of Computing and Augmented Intelligence, Arizona State University, AZ, United States

⁴ Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States.

⁵ International Computer Science Institute, CA, USA

⁶ Department of Statistics, University of California at Berkeley, CA, USA.

Hydroformylation, Catalysis, Machine Learning

ABSTRACT: Rational development of transition-metal catalysts, even when guided by theory and mechanistic knowledge, involves significant trial and error. Although ML offers the potential to accelerate catalyst discovery and optimization, accurately modeling the complex structures of catalysts and the multistep mechanisms by which they react remains challenging with the limited sets of data available. Olefin hydroformylation is a quintessential example of this challenge: its catalytic cycle involves many, often reversible, steps, and decades of study have not yielded reliable structure-selectivity relationships. We report Libra-ML, a 3D structure-based deep learning approach for predicting experimental outcomes of transition-metal catalyzed reactions. To demonstrate the ability of Libra-ML to model the outcomes of complex catalytic reactions, we predicted the regioselectivity of hydroformylation with terminal olefins catalyzed by rhodium complexes. Comparisons to existing methods demonstrate the state-of-the-art performance of Libra-ML and illustrate the importance of capturing 3D structure to predict experimental outcomes with molecular catalysts.

INTRODUCTION

The development of catalysts that are active, selective, and react with high turnover numbers has classically been guided by mechanistic information and the identification of discrete relationships between catalyst structure and reactivity. Such rational development, even when guided by theory, mechanistic knowledge, and fundamental chemical principles, requires significant trial and error. Machine learning (ML) and cheminformatics have been envisioned as an alternative approach to identify relationships between catalyst structures and reaction outcomes that are too complex or subtle for human intuition¹⁻⁵. Hallmark studies have been reported by Sigman⁶, Doyle⁷, Kozlowski⁸, Denmark⁹⁻¹¹, Hong^{12,13}, and others, wherein digital representations of catalyst structures or properties are correlated with reaction yield¹⁴ or selectivity¹⁵.

However, there are many fundamental challenges facing the application of ML to transition-metal catalysis, and synthetic chemistry more broadly. Although ML offers the potential to accelerate catalyst discovery and optimization, modeling both the complex three dimensional (3D) structures of molecular catalysts and the multistep mechanisms by which they react remains challenging with the limited datasets available for catalysis, especially when compared to the large datasets common in traditional ML. Even for larger datasets comprising diverse

catalyst structures, representing such molecular diversity as input for ML models is difficult. Finally, deriving chemical insights from commonly used "black box" ML models has been difficult, if not impossible. This inherent lack of transparency regarding the process by which model outcomes are generated makes it challenging to understand the reasoning behind predictions or derive new chemical knowledge. This lack of transparency limits the utility of such models to advance chemistry. Addressing these challenges would significantly advance the use of ML for a wide range of catalytic problems and synthetic chemistry.

We posited that a suitable dataset for olefin hydroformylation would be available, that this dataset would be amenable to augmentation by computation and experiment, and that predictions of this reaction's regioselectivity would serve as an ideal test of the ability of ML to predict outcomes in homogeneous catalysis. Hydroformylation is the addition of CO and H₂ across an olefin double bond to generate an aldehyde from an alkene typically catalyzed by Co or Rh complexes and is one of the largest-scale homogeneous catalytic reactions in the chemical industry, generating millions of tons of chemical products annually¹⁶. This reaction is a complex, multicomponent homogeneous catalytic process¹⁶ and many steps of the catalytic cycle can be reversible, with the degree of reversibility often depending on the temperature and pressure of CO and H₂¹⁷. Thus, the regioselectivity

to form linear or branched aldehydes (Figure 1) depends on both the identity of the catalyst and, due to the complex mechanism, the reaction conditions. A large body of literature data on the mechanism of this reaction exists. Although numerous proposals regarding the effects of ligand structure on catalyst activity and selectivity have emerged from decades of research that generated these data, including heuristics based on bite angle¹⁸ and electronic character¹⁹, no guidelines can reliably predict the regioselectivity of olefin hydroformylation.

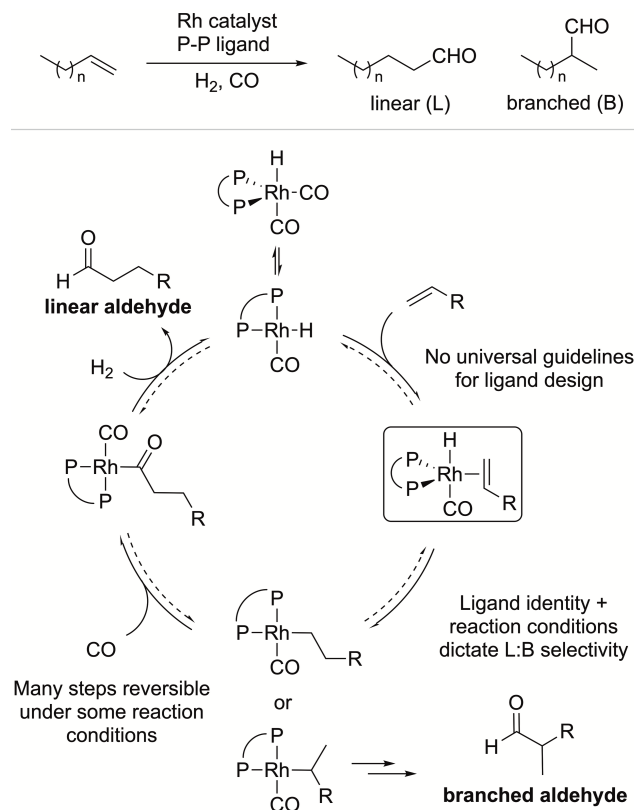


Figure 1. General mechanism of olefin hydroformylation.

A trained ML model that is valuable for catalysis would accurately predict selectivity as a function of catalyst structure and reaction conditions across a wide range of reactions. Such a model applied to hydroformylation could predict the regioselectivity and be used to design new ligands that form complexes capable of catalyzing olefin hydroformylation with counterintuitive selectivities. Existing studies using ML to predict the regioselectivity of olefin hydroformylation have established that predictions made by linear regression²⁰, random forest²¹, and extreme-gradient-boosting models^{21,22} with tabular, handcrafted descriptors are accurate within narrow training domains. Control experiments demonstrating the limitations in performance of the decision tree models and extreme-gradient-boosting models^{21,22} under rigorous evaluation are provided in the Supporting Information. Thus, a model that predicts selectivity with high accuracy and with suitable control experiments across a broad domain of ligands has not been reported.

We report Libra-ML, a predictive model for hydroformylation that uses deep learning based on 3D structures to model the on-cycle rhodium species responsible for forming linear or

branched catalytic intermediates along the reaction pathway, rather than relying on computed, expert-selected descriptors. To do so, we use a 3D graph neural network (3D GNN) trained on the 3D structures of the transition-metal catalytic intermediates that contain bound reactants, to predict catalyst selectivity. Sourcing data broadly from the chemical literature resulted in a model with broad generalizability across organophosphorus ligands, while data augmentation using catalyst conformers enabled effective training on this diverse dataset. In-house experimental validation demonstrates the model's accurate prediction of olefin hydroformylation regioselectivity, its capacity to generalize across a broad domain of ligand structures, and its utility in designing new ligands with surprising selectivities. The approach used to develop Libra-ML addresses many of the challenges facing the application of ML to catalysis, such as the difficulty of developing effective data representations for diverse datasets, the small size of datasets endemic to synthetic chemistry, and the challenge of achieving model interpretability and trustworthiness. Accordingly, we present Libra-ML as a promising template that can be adapted to other catalytic systems. Libra-ML enables the prediction of catalyst selectivity with a single command-line operation and is available as a Python package and a pre-configured Docker image.

METHODS.

Data Collection.

To construct a predictive model for olefin hydroformylation, a dataset was extracted from published literature. Although the quality of data from this literature can be variable, and bias in literature-sourced datasets is common²³, these datasets do possess advantages for the training of ML models. Literature datasets are often less combinatorial than datasets obtained by high-throughput experimentation, thereby reducing the possibility of overfitting.²⁴ Literature datasets also result from decades of experimentation and, in homogeneous catalysis, often include ligands that are not commercially available or readily prepared. These ligands, as a whole, span many distinct structural types, mechanistic hypotheses, research groups, and time periods. In addition, many ligands were developed to test specific mechanistic hypotheses, and structures within individual families of ligands and within studies have frequently been varied. Finally, systematic variations in the conditions for hydroformylation are frequently reported. Based on these observations, we reasoned that a dataset on hydroformylation constructed from the literature would result in a large diversity of ligand structures and reaction conditions.

A dataset comprising published results on the hydroformylation of unactivated α -olefins catalyzed by rhodium complexes containing bidentate organophosphorus ligands was manually curated by synthetic chemists. Each paper and its supporting information were inspected, and data tables were manually added to a spreadsheet to avoid errors in translation. While the manual curation of literature data is more labor-intensive than automated exporting of data from chemical search engines like Reaxys[®] or SciFinder[®], the manual curation avoids the inclusion of mischaracterized or erroneous data, which can constitute 10% of the data in Reaxys[®] and 56% of the data in the USPTO database^{23,25}. Each reaction in the dataset includes an unactivated terminal olefin, rhodium precatalyst, bidentate organophosphorus ligand, and hydrogen and carbon monoxide gases.

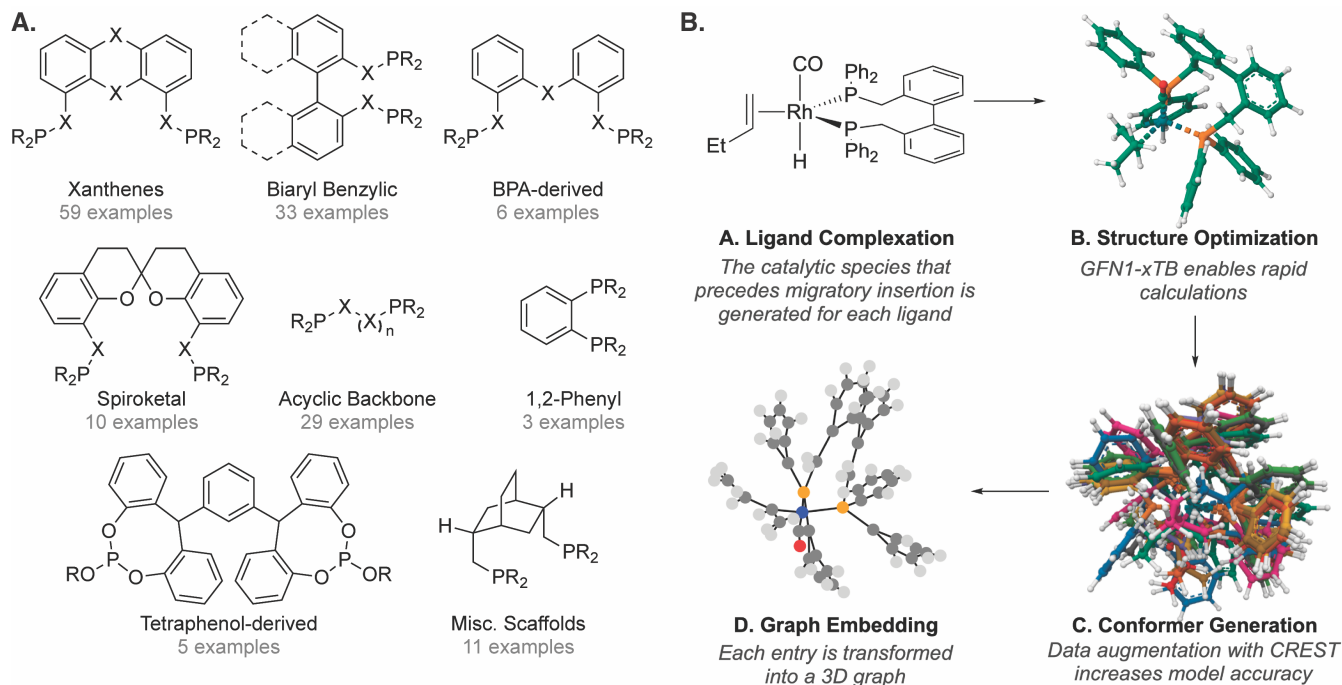


Figure 2. (A) Literature dataset encompassing a wide range of ligand structures ($X = C, O$; $R = C, N, O$). (B) Catalyst modeling, conformer generation with GFN1-xTB and CREST, and graph embedding workflow.

Reaction conditions, such as pressures of H_2 and CO , temperature, mol % of catalyst, and mol % of ligand were also tabulated. These conditions were included as descriptors to capture their effects on selectivity.

The resulting curated dataset includes 32 papers with 378 reactions that encompass 156 ligands possessing a wide range of structures. The ligands in our dataset include phosphines, phosphonites, phosphites, phosphonamides, phosphinamides, and include both C_2 -symmetric and nonsymmetric (C_1) bidentate ligands (Figure 2, A). The reported linear:branched ($L:B$) ratios range from 0.2 to 618. The reaction conditions captured in this dataset include 17 temperatures, 34 combinations of CO and H_2 pressures, and 61 combinations of precatalyst and ligand loadings. Because the hydroformylation of unactivated α -olefins is generally selective for the linear aldehyde, the data skew towards high $L:B$ ratios.

Representation of Metal-Ligand Complexes.

The need to describe the properties of a wide range of structures with just a few atoms in common creates a challenge when applying ML to make predictions about a set of potential catalysts. The application of ML to homogeneous catalysis has typically involved sets of structures with shared scaffolds and points of diversity^{26,27}. The differences between the descriptors calculated for such structures are then correlated with a reaction outcome. In that scenario, the modeling task is simpler because the domain of the dataset is limited, and the resulting models need not extrapolate to compounds lacking a shared scaffold. In contrast, the ligand library described herein, like that of many metal-catalyzed reactions, comprises a wide range of structures that only share the presence of two phosphorus atoms. An approach to address this challenge was developed in 2003 simultaneously by Kozłowski²⁸ and Lipkowitz²⁹ for modeling the enantioselectivity of catalysts that do not share a common scaffold. Both groups applied a 3D representation of the catalyst structure, but the level of predictive power was limited by the

mismatch of the high dimensionality of their molecular representation and the limited quantities of experimental data. Subsequent work from Denmark and coworkers addressed the challenges of modeling with limited data using voxel-based representations of transition-metal³⁰ and organic catalysts²⁶ with feature selection and dimensionality reduction, but the use of thousands of voxels to represent catalysts still limits the ability of such approaches to model selectivity with small experimental datasets^{27,31–33}.

We have developed an approach to model the full metal-ligand complex in a manner that captures 3D information at the level of atoms and bonds. We envisioned that directly inputting an on-cycle, selectivity-determining catalytic intermediate into a 3D GNN that captures atomic identities, spatial coordinates, and chemical bonding information, would create a model with a deeper understanding of the factors influencing regioselectivity than tabular descriptors or voxel representations, especially in scenarios with limited experimental data. A key advantage of this 3D GNN approach is its ability to learn relevant structural features automatically from the data. This approach does not require the manual feature engineering often associated with traditional descriptor-based methods and helps address complexities in feature selection and dimensionality that can arise with high-dimensional representations (*vide supra*).

Unlike tabular descriptors that capture 3D structures implicitly through descriptors, such as bite angle or percent-buried volume, 3D molecular representations can allow a model to learn more explicitly the structural features, at the level of atoms, bonds, and spatial coordinates, that govern the selectivity of the reaction. Furthermore, using an on-cycle metal complex, rather than a free ancillary ligand, as is commonly done in applications of ML to transition-metal catalysis, enables the model to capture both the features of the ancillary organophosphorus ligands and the influence of these features on the reactive components (hydride, CO , and olefin).

For these reasons, we hypothesized that training a 3D GNN on a catalytic intermediate containing both ancillary and reactive ligands would enable models to predict selectivity across a broad ligand domain, despite the limited size of the dataset that is typical for such catalytic reactions. 3D GNNs have been widely reported to enable ML to predict interatomic potentials,^{34,35} quantum mechanical properties,^{36–38} and both transition states and binding energies computed with DFT.³⁸ Their performance has been shown to be stronger than that of 2D GNNs when predicting DFT-computed reaction barriers for reactions catalyzed by transition-metal complexes.³⁹ They have also been used to model the properties of organic molecules and, in rare cases, reactions of them.⁴⁰ However, they have not been used to relate the structures of transition-metal catalysts to experimental outcomes or predict the outcomes of new catalysts.

To begin, we chose the trigonal bipyramidal intermediate in Figure 1 to represent the catalyst because this complex undergoes migratory insertion to either the linear or branched Rh-alkyl intermediate, and we reasoned that perturbations in its geometry and electronic properties would be sensitive to the structure of the ligand. Because the identity of the regio-determining step can be different for different ligands, substrates, and reaction conditions, the trigonal-bipyramidal intermediate serves as a standardized structure that represents the molecular recognition environment and is relevant to the selectivity. We acknowledge that a single complex does not completely control the regioselectivity. Each ligand was docked to a rhodium center bearing 1-butene, CO, and hydride ligands in a trigonal bipyramidal arrangement with both phosphorus atoms and the 1-butene in equatorial positions, and the CO and hydride ligands in the axial positions (Figure 2B, A). By modeling this reaction intermediate, the metal, butene, CO, and hydride are constant components in the dataset, while the phosphine ligand is varied. These catalyst structures are programmatically generated by a module in Libra-ML, based on molSimplify^{41,42} that uses a SMILES string of the organophosphorus ligand as the only input and returns an XYZ file of the trigonal bipyramidal complex as the output. This process allows large numbers of catalyst geometries to be generated rapidly.

The energy of the catalyst structures generated in this way are then minimized using techniques designed to enable computations to be conducted rapidly on large molecules (Figure 2B, B).^{43,44} Each catalyst structure was optimized using semi-empirical density functional theory with implicit solvation by toluene (GFN1-xTB/ALPB(Toluene))⁴⁴ because it is the most common solvent in the dataset, and all reactions in the dataset, with the exception of one, were run in toluene or benzene. GFN1-xTB was chosen over GFN2-xTB, due to its greater accuracy for the structures of transition-metal complexes.⁴³ While density functional theory (PBE0-D4/def2-SVP with CPCM(Toluene) in ORCA 6, using 16 CPU cores) minimizes the energy of a typical complex modeled in our study in approximately three hours, GFN1-xTB/ALPB(Toluene), using 16 CPU cores, minimizes the energy of the same structure in approximately three minutes, with comparable structural accuracy. These computational requirements were benchmarked for a DPPE (1,2-bis(diphenylphosphino)ethane)-ligated trigonal-bipyramidal Rh-hydride-CO-olefin complex.

Data Augmentation.

To expand the structural representations available for each catalyst, a set of conformers within 12 kcal/mol of the most

stable conformer was generated, and the 10 lowest-energy conformers were included independently as a form of data augmentation (Figure 2B, C). When predicting *L:B* ratios, the predicted values for each conformer are pooled and averaged to give a final prediction for a given catalyst. This approach can be seen as a voting or ensembling technique across accessible conformations. Prior work on catalyst parameterization included averaged or Boltzmann-weighted descriptors of conformers to create a dynamic picture of catalyst properties or averaged voxel representations of entire conformer ensembles.^{10,11,45} In our approach, each conformer is treated as a unique chemical entity with a corresponding unique data point. By doing so, the model learns that multiple conformations are correlated with the same reaction outcomes. This approach helps compensate for inaccuracies identifying global minima by computation at a level of theory that scales well with large sets of complexes. Our approach can also be viewed as a form of multi-instance learning (MIL), wherein an object is represented by a set of alternative instances and the task is to establish a correlation between the set of instances and the label.⁴⁶ Prior work has shown that the performance of MIL models containing conformer-based instances with 3D descriptors is higher than those with 2D descriptors when predicting enantioselectivity in conformationally flexible organic catalysts.⁴⁷

Conformers were generated using CREST with GFNFF/ALPB(Toluene).^{48,49} The resulting conformer ensemble was then reoptimized using GFN1-xTB/ALPB(Toluene) and reprocessed with CREST to remove degenerate structures. The 10 lowest-energy conformers, then, were used as representative structures. These conformers were subjected to a final optimization using GFN1-xTB/ALPB(Toluene) with stricter optimization criteria. This strategy provides a chemically meaningful form of data augmentation^{50,51} by increasing the number of training instances per catalyst 10-fold to improve model performance and, thereby, to address the challenges of building ML models from limited catalytic reaction data. This process resulted in a final dataset composed of 3518 entries, representing a diverse range of organophosphorus ligands and reaction conditions.

RESULTS AND DISCUSSION.

Model Training and Testing.

A series of models were trained to predict experimental regioselectivities from the literature using the combination of catalyst geometries and reaction conditions (Figure 3, A). The *L:B* ratio was transformed into the energy difference ($\Delta\Delta G^\ddagger$, in kcal/mol) between the transition states leading to the isomeric products using the reported reaction temperature. The performance of each model was evaluated using leave-one-catalyst-out (LOCO) cross-validation. In this method, a model was trained on data from all but one of the catalysts and then tested on its ability to predict regioselectivity for the remaining catalyst. Because the data are split by catalyst identity (i.e., by ligand), all conformers and all reactions associated with a given ligand are kept together in the same fold. As a result, no conformers from a ligand in the test fold can appear in the training set, eliminating conformer-level leakage. Performing LOCO cross-validation across the full set of 156 ligands therefore produced 156 distinct train/validation splits, with each split holding out one ligand and training on the remaining 155. This method to evaluate performance contrasts with random data splitting, an alternative validation approach. Because each catalyst con-

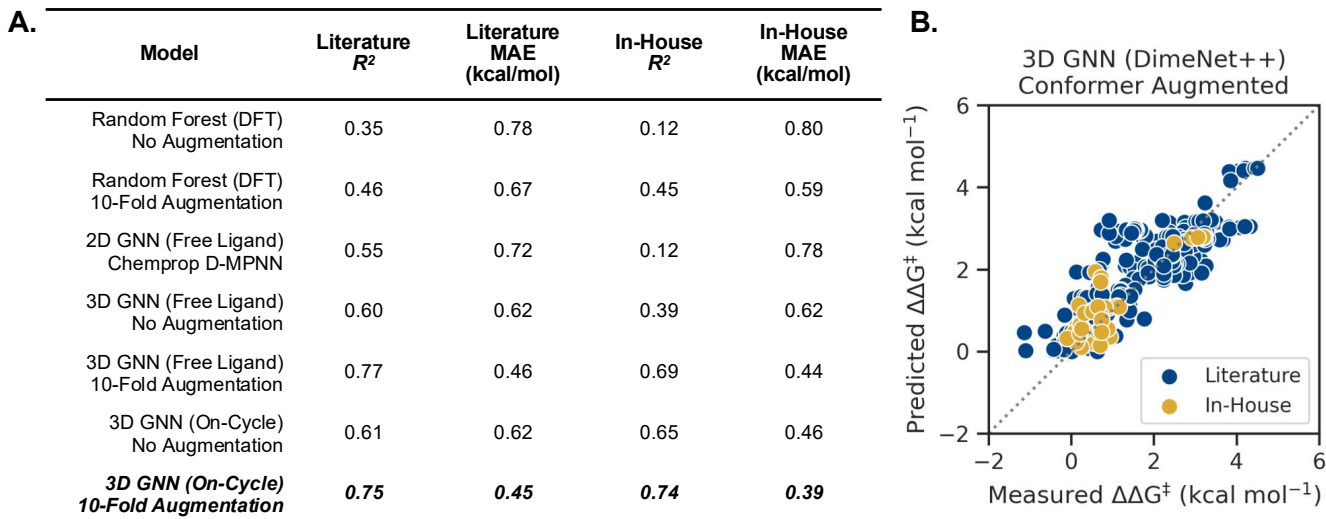


Figure 3. (A) Table containing the performance for each model on the literature leave-one-catalyst-out (LOCO) set and the in-house experimental set. Literature LOCO performance was calculated by aggregating all test predictions into a combined set and evaluating metrics on this set. For the in-house experimental set, predictions were first averaged across all ensemble members for each model, and metrics were then calculated on this set of averaged predictions. (B) Predicted versus measured $\Delta\Delta G^\ddagger$ for the conformer augmented DimeNet++ model trained on on-cycle intermediates. This model predicted the regioselectivity of olefin hydroformylation across a wide range of catalyst structures and reaction conditions.

stitutes a varying fraction of the dataset, random splitting can create partitions in which performance metrics misrepresent the model's true performance across the entire dataset, potentially leading to artificially inflated or deflated results. The LOCO approach, therefore, provides a more rigorous and accurate assessment of the model's ability to generalize its predictions of regioselectivity across the diverse range of ligand structures and reaction conditions present in the dataset. Additional experiments and statistical validation conducted using 5×2 repeated, group-aware cross-validation, can be found in the Supporting Information.

To create benchmarks for the performance of a 3D GNN, we evaluated the ability of published approaches to predict the regioselectivity from the same dataset. The predictive performance of a linear regression and lasso regression model using tabular chemical descriptors derived from the lowest energy conformer (see Table S2 in the Supporting Information for details) was poor (LOCO $R^2 = -0.04$, MAE = 0.99 kcal/mol and LOCO $R^2 = -4.83$, MAE = 1.39 kcal/mol, respectively). The predictive power of a random forest regressor was higher with the same tabulated descriptors (LOCO $R^2 = 0.35$, MAE = 0.78 kcal/mol) but still resulted in poor predictive performance. Improved but still weak predictive performance was observed for the linear regression (LOCO $R^2 = 0.55$, MAE = 0.68 kcal/mol), lasso regression (LOCO $R^2 = 0.53$, MAE = 0.69 kcal/mol), and random forest regressor (LOCO $R^2 = 0.46$, MAE = 0.67 kcal/mol) models when using tabular chemical descriptors with the data augmentation described above.

We also tested a 2D GNN. To do so, we trained Chemprop⁵², a directed-message-passing graph neural network (D-MPNN), on the free ligand structures represented as 2D graphs with reaction conditions concatenated to the graph embeddings prior to prediction. The performance metrics of this approach were the similar to those of the models trained on data augmented tabular descriptors (LOCO $R^2 = 0.55$, MAE = 0.72 kcal/mol), indicating only modest predictive power.

To test our hypothesis that a model trained directly on the optimized 3D geometries of a catalytic intermediate would gain a richer understanding of the structural features of a ligand that influence regioselectivity, we assessed the performance of a 3D GNN on both the geometry of the free ligand and on-cycle catalytic intermediate. We chose a modified architecture based on DimeNet++^{53,54}, an E(3)-invariant, directed-message-passing 3D GNN that uses interatomic distances and angles. The input for DimeNet++ is only atomic identities and coordinates. We modified DimeNet++ for our application to catalysis by adding a multi-layer perceptron (MLP) to predict the experimental $\Delta\Delta G^\ddagger$ from the concatenated graph embedding and reaction conditions.

Training DimeNet++ with only the lowest-energy conformer for both the free ligand and catalyst as input resulted in better performance (LOCO $R^2 = 0.60$, MAE = 0.62 kcal/mol and LOCO $R^2 = 0.61$, MAE = 0.62 kcal/mol, respectively) than the models trained on tabular chemical descriptors and the 2D GNN, despite the small dataset and simple embeddings. These higher values for R^2 and RMSE show that the 3D structure of the free ligand and a key catalytic intermediate are both rich sources of predictive information about the complex relationship between the structure of the catalyst and regioselectivity of olefin hydroformylation.

Subsequent training using the data augmentation described above led to remarkable performance for both the 3D GNN trained on the free ligand and the catalytic intermediate (LOCO $R^2 = 0.77$, MAE = 0.46 kcal/mol and LOCO $R^2 = 0.75$, MAE = 0.45 kcal/mol, respectively). Overall, both 3D GNN models surpassed the performance of the previous models across a wide range of catalyst structures and reaction conditions (Figure 3, B). However, importantly, when the performance of each model is stratified into branched-selective, near-equipolar, and linear-selective regimes (defined as above 20:1 L:B), the performance of the model trained on the on-cycle intermediate was much stronger in the branched-selective regime (branched MAE = 0.33 kcal/mol) than was that of the model trained on the free

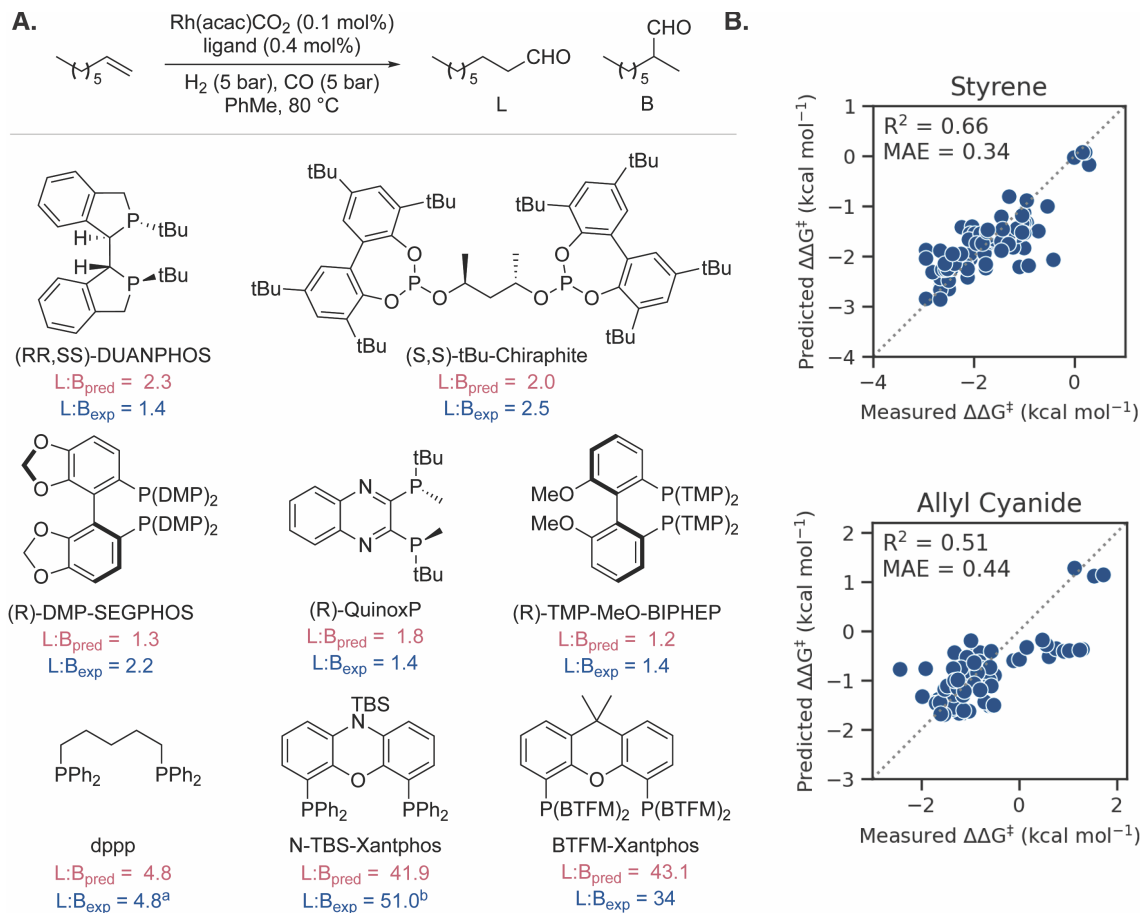


Figure 4. (A) Experimental validation demonstrating the ability of Libra-ML to make predictions in diverse chemical space. ^aReaction and prediction were run at 2 bar of H₂ and 2 bar of CO. ^bReaction and prediction were run at 100 °C. DMP = 3,5-dimethylphenyl; TMP = 3,4,5-trimethoxyphenyl; BTfM = 3,5-trifluoromethylphenyl; dppp = diphenylphosphinopentane. (B) Model performance for the prediction of styrene and allyl cyanide regioselectivity.

ligand geometries (branched MAE = 0.46 kcal/mol); improvements in the linear-selective regime (linear MAE = 0.43 kcal/mol versus linear MAE = 0.44 kcal/mol) were smaller than those in the branched-selective regime. This overall stronger performance of the model built on the rhodium complex suggests that models trained on on-cycle intermediates generalize more accurately to challenging predictive regimes. Full details of stratified model performance can be found in Table S3 in the Supporting Information. Ultimately, the ability to incorporate a conformer-augmented dataset into our 3D GNN model was particularly important because such conformers cannot augment a 2D GNN dataset. They cannot augment this type of dataset because 2D graphs capture only atomic information and connectivity, not the three-dimensional spatial relationships that define conformations.

The data augmentation we used for this model resembles noise augmentation techniques used in protein structure prediction with GNNs⁵⁵ and CNNs⁵⁶ for which random noise added to atomic positions leads to more robust models by preventing overfitting. However, our conformer-based method introduces chemically meaningful structural perturbations derived from molecular dynamics simulations, rather than random noise. Of course, these conformers are easier to obtain for transition-metal complexes than for entire proteins. We hypothesize that conformer augmentation using the 3D geometries of these complexes improves model performance by reducing overfitting by

providing chemically meaningful conformational information that enables the model to learn more generalizable structure–selectivity relationships, rather than memorizing specific catalyst structures or substructures. Before further evaluating that hypothesis, we first assessed the validity of our 3D GNN models. To do so, we performed control experiments designed to determine whether the observed predictive performance was genuinely reliant on chemically meaningful inputs, or if it could be attributed to data leakage or spurious patterns in the data.

Control Experiments.

To establish whether our 3D GNN models are learning chemically meaningful information or inherent patterns in the data, we performed a control experiment to test if chemically meaningful inputs are necessary for accurate predictions from the data (Figure 3, A). If the dataset contains patterns that are not chemically meaningful and that the model learned to make its predictions, then high performance should be observed when the models are trained using inputs that are not chemically meaningful²⁴. Training DimeNet++ on a dataset with shuffled selectivities (the Y-shuffle test) resulted in a model lacking any predictive power (LOCO R² = -0.20, MAE = 1.18 kcal/mol).

We also conducted a control experiment to determine whether biases and patterns within the literature data could, by themselves, lead to predictive ability, even in the absence of chemically meaningful catalyst input. For this control, a random forest model trained solely with one-hot encoded ligand

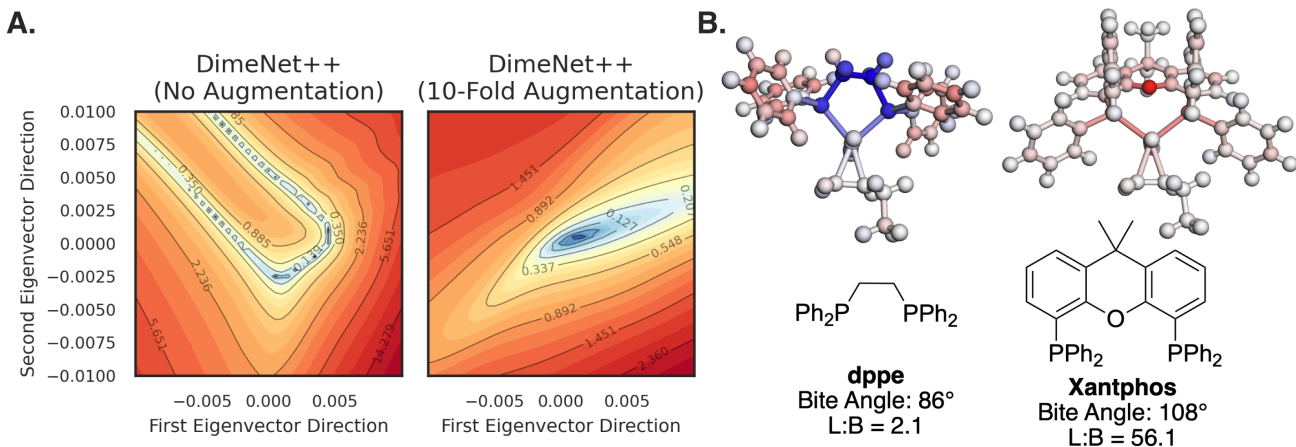


Figure 5. (A) Loss landscape visualizations comparing the baseline DimeNet++ model trained on lowest-energy catalyst geometries with a DimeNet++ model trained on the conformer-augmented dataset. The smoother landscape of the augmented model compared to the baseline suggests that conformer augmentation reduces overfitting and enhances the model’s ability to generalize. (B) SHAP feature importance visualization for dppe (1,2-Bis(diphenylphosphino)ethane) and Xantphos. Red indicates these substructures contribute to a higher predicted $L:B$ value, and blue indicates these substructures contribute to a lower predicted $L:B$ value.

identities and reaction conditions yielded poor, yet measurable, predictive performance (LOCO $R^2 = 0.47$, MAE = 0.68 kcal/mol). We interpret this result to mean that the modest predictive power from this control model stems from the model exploiting inherent biases within the literature data, not from learning genuine chemical principles. For example, it could learn correlations between specific reaction conditions commonly reported alongside certain ligands and the regioselectivities typically obtained with such combinations of ligand and reaction condition. While reaction conditions influence selectivity, they are insufficient by themselves to predict selectivity because the conditions leading to the highest selectivity vary with the identity of the ligand. However, literature data often underrepresent pairings of less-optimal conditions with various ligands. Consequently, the control model is inadequately penalized for over-relying on patterns in reaction conditions, rather than the ligand’s structural features. This outcome underscores the critical need for in-house experimental validation of models trained on literature data.

Experimental Validation with New Ligands.

Having developed a series of ML models to predict the regioselectivity of hydroformylation, we evaluated the performance of each model (Figure 3, A) with a test set of experiments in our laboratory. To obtain this set, 29 bidentate organophosphorus ligands previously unreported for the hydroformylation of unactivated α -olefins were selected (Figure 4, A). The unreported ligands include many chiral bidentate organophosphorus ligands because they are not frequently used to form the achiral linear aldehyde products from hydroformylation reactions of α -olefins. The set of ligands for these experiments also cover a diverse chemical space, including ligands with acyclic, biaryl, and xanthene backbones, as well as varied phosphorus substituents.

For each ligand, we measured the experimental $L:B$ ratio for the hydroformylation of 1-octene. We then used each model to predict the $L:B$ ratio and compared their predictive performances (Figure 3, A). Each reaction was conducted under the same conditions (0.1 mol% Rh, 0.4 mol% ligand, 5 bar H_2 , 5 bar CO, 80 °C, 4 h) and then with varying reaction conditions to probe the ability of the model to predict the effect of reaction conditions on selectivity (see Supporting Information for full

details). Because the reactions in our set were evaluated under standardized and unoptimized conditions, this reaction set is less prone to patterns and biases that might allow a model to predict outcomes without learning chemically meaningful principles than is typical for literature data. Therefore, evaluating the model’s performance on this set provides a rigorous assessment of its ability to generalize to practically relevant conditions and to predict outcomes for ligands previously untested in a reaction of interest.

Among the models we used in this work, the conformer-augmented 3D GNN trained on on-cycle catalytic intermediates predicted regioselectivities for the in-house experimental test set with the highest accuracy ($R^2 = 0.74$, MAE = 0.39 kcal/mol). The 3D GNN trained on free ligand geometries was less accurate ($R^2 = 0.69$, MAE = 0.44 kcal/mol) than the 3D GNN trained on on-cycle intermediates, even though the performance of the two models was similar on the literature dataset. Consistent with our results from the literature dataset, predictions by 3D GNNs trained without conformer augmentation were less accurate on our experimental test set than were those trained with conformer augmentation. In particular, the accuracy of the model trained on free ligand structures was much lower ($R^2 = 0.39$, MAE = 0.62 kcal/mol) than that of the model trained on on-cycle intermediates ($R^2 = 0.65$, MAE = 0.46 kcal/mol). This result mirrors the trend of performance by the conformer-augmented models, but with a larger gap in the quality of the performance.

The predictive performance of the random forest model (trained on tabular chemical descriptors with and without augmentation) on our in-house experimental set was poor ($R^2 = 0.12$, MAE = 0.80 kcal/mol and $R^2 = 0.45$, MAE = 0.59 kcal/mol, respectively) and similar to its performance on the literature dataset. The performance of the 2D GNN, Chemprop, on our experimental set was extremely poor ($R^2 = 0.12$, MAE = 0.78 kcal/mol), especially in comparison to its performance on the literature dataset. This large difference in performance demonstrates a failure of this model to learn generalizable connections, presumably chemical principles, beyond the biases and patterns inherent in the literature data.

Thus, the combination of a 3D GNN with conformer augmentation, trained on on-cycle intermediates, was the only approach

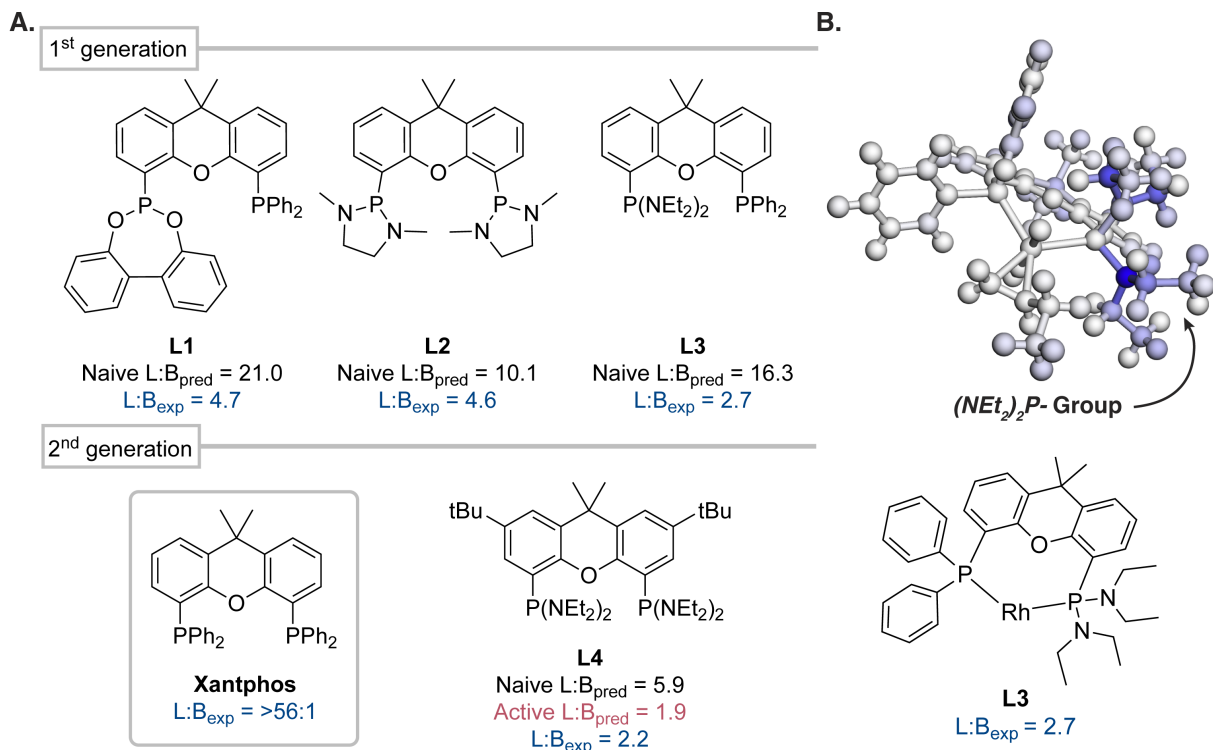


Figure 6: (A) Predicted and experimental results for the reprogramming of Xantphos selectivity. Naïve refers to the baseline Libra-ML model predictions. Retrained refers to the model trained on the experimental results from the previous generation. (B) SHAP value visualization of **L3**, filtered to only visualize negative contributions, indicate the impact of the (NEt₂)₂P- group on the model predictions.

among those we tested that predicted the selectivity of this complex hydroformylation reaction with the highest precision for both the literature and in-house experimental datasets, especially in challenging selectivity regimes important for use in designing new ligands. We anticipate that this approach will be valuable beyond hydroformylation because the model was created from an experimental dataset with a size that is characteristic of those available for complex chemical systems, including other chemical reactions catalyzed by transition-metal complexes.

Validation with Distinct Substrates.

To gain initial information on the general ability of a 3D GNN to predict the outcomes of related but distinct chemical reactions, we collected a separate set of literature data comprising 150 hydroformylation reactions of styrene and 85 reactions of allyl cyanide. We then applied the same computational workflow as used for reactions of the unactivated α -olefins to obtain representative structures of the hydridorhodium carbonyl unit bound to each ligand-substrate combination within these two additional datasets. Subsequently, a new 3D GNN was trained and tested on each of these two new datasets using the established LOCO cross-validation strategy (*vide supra*). The performance (Figure 4, B) of models built to predict the regioselectivity for the hydroformylation of styrene and allyl cyanide was strong, despite the small size of their datasets (LOCO R^2 = 0.66, MAE = 0.34 kcal/mol for styrene; LOCO R^2 = 0.51, MAE = 0.44 kcal/mol for allyl cyanide).

This strong performance highlights the generalizable nature of this computational workflow and modeling approach. This accuracy was achieved, even though the datasets for reactions of these olefins were much smaller than were those for the unactivated α -olefins, the reactions of styrene are biased to form

the branched isomer, and the mechanistic origin of the regioselectivity is fundamentally different than that for 1-octene. The success of this approach in predicting outcomes for these two alkenes demonstrates its potential for application to additional reactions catalyzed by transition-metal complexes.

Analysis of the Loss Landscape.

Having established the validity of our model by control experiments and experimental validation, we evaluated whether conformer augmentation improves performance by reducing overfitting while providing chemically meaningful conformational information that enables the model to learn more generalizable structure-selectivity relationships. To evaluate this hypothesis, we analyzed the loss landscape^{57–59}. The loss landscape is a visual representation of the model's loss as a function of the perturbations in its trainable parameters. In this landscape, the x and y axes represent two perturbation directions, while the z corresponds to the loss value. Analysis of the topology of the resulting landscape provides insights into the optimization of the model during training and its subsequent predictive ability. Broadly, landscapes with single, wide minima have been found to correlate with less overfitting and better generalization performance⁶⁰.

The loss landscape (Figure 5, A) of the DimeNet++ model trained without augmentation is fragmented with multiple minima below an RMSE of 0.885 kcal/mol. A loss landscape with multiple local minima can be indicative of overfitting, reflecting many distinct parameter configurations that fit the training data well but may generalize poorly. In contrast, the landscape of the DimeNet++ model trained with conformer augmentation is smoother, with a single, broad minimum below an RMSE of 0.892 kcal/mol. This difference indicates a more stable loss landscape and an improved ability to generalize. Full details of

this experiment can be found in the Supporting Information, and further work is underway to analyze the loss landscapes of all models in this work to gain a deeper understanding of how these models learn and generalize.

Model Interpretability.

While 3D GNNs offer a powerful way to learn important features directly from the chemical structures on which they are trained, they are fundamentally black-box models. An ability to generate insights into the structural features that drive the model’s predictions would allow chemists to combine human chemical intuition with data-driven selectivity patterns in a manner that has not yet been demonstrated for the application of deep learning to the development and optimization of new catalysts.

To understand which catalyst structural features influence the model’s prediction, we employed Shapley value analysis (SHAP), a post hoc interpretation technique⁶¹. Using the *PermutationExplainer* function from the SHAP package, we evaluated feature contributions by systematically masking combinations of atoms and observing the resulting impact on the prediction. This process yields Shapley values for each atom, quantifying their individual importance to the model’s output. Positive Shapley values signify contributions favoring linear selectivity (higher *L:B*), whereas negative values signify contributions favoring branched selectivity (lower *L:B*). We then developed a visualization method to overlay these Shapley values onto the 3D structure of the catalyst, highlighting the molecular substructures responsible for the model’s predictions. This approach allows us to begin to understand the reasoning of the model’s predictions and assess its alignment with chemical intuition. By doing so, we obtain a post-hoc, local, atom-level attribution of the chemical origins of the predicted reaction outcome, thereby increasing the interpretability and confidence in model predictions.

We chose to interrogate the reasoning of the model’s predictions for the regioselectivity with dppe and Xantphos as examples of ligands leading to lower and higher *L:B* selectivity, respectively. These ligands have been studied extensively to determine the structural features responsible for their observed regioselectivities^{62,63}. Libra-ML identified the ligand substructure (Figure 5, B) that determines the bite angle, associating the smaller bite angle (85°) of dppe with increased branched selectivity and the larger bite angle (108°) of Xantphos’s with increased linear selectivity. This correlation aligns with established empirical heuristics that ligands with larger bite angles lead to an increase in linear selectivity, indicating that the model is learning chemically meaningful information from its training data, rather than making spurious correlations. In addition to validating the predictions made by the model, this approach to visualization offers a powerful tool to assist the design of new ligands. By highlighting the molecular substructures predicted to influence selectivity most strongly, it can help chemists choose the regions of a ligand to modify to achieve a desired goal.

Use of Libra-ML for Accelerated Catalyst Development.

To evaluate the ability of Libra-ML to aid in the design of new ligands, we envisioned a workflow in which an experimentalist imagines a new ligand structure, uses Libra-ML to predict the selectivity of the resulting catalyst, interprets these predictions across many candidates to form a structure/selectivity hypothesis, and then synthesizes a ligand that could lead to a

desired outcome. Given the difficulty synthesizing organophosphorus compounds, this ability to prioritize candidates with Libra-ML would substantially accelerate the design process over the synthesis of ligand libraries or traditional trial-and-error.

We considered the potential to convert the classic Xantphos ligand, with its wide bite angle and accompanying high *L:B* selectivity (Figure 6, B) to a ligand that leads to a much different selectivity that would be difficult to predict with existing heuristics. Studies with preliminary models not discussed here suggested this possibility, and rapid *in silico* screening of hypothetical ligand structures with Libra-ML predicted selectivity patterns for catalysts containing heteroatoms on phosphorus that most practicing chemists would not anticipate: the catalyst with a bidentate ligand containing phenoxy or amino groups on phosphorus and the same wide bite angle of Xantphos produces nearly equal ratios of linear and branched isomers.

Because non-symmetrical xanthene ligands and ligands with amino groups on phosphorus were rare in Libra-ML’s training dataset, we envisioned an active learning scenario in which we would iteratively update the model by retraining with new experimental *L:B* ratios (each evaluated under 6 reaction conditions) as new ligands containing more heteroatom-based groups on phosphorus were identified and synthesized. An initial *in silico* exploration with Libra-ML led us to investigate ligands L1-L3 in Figure 6A. The model predicted *L:B* ratios of 21.0:1 for L1, 10.1:1 for L2, and 16.3:1 for L3. Consistent with this prediction, the new ligands formed catalysts that reacted with much lower *L:B* ratios than Xantphos and even more branched product than predicted (*L:B* ratios of 4.7:1 with L1, 4.6:1 with L2, and 2.7:1 with L3). While Libra-ML did not predict the exact experimental *L:B* ratios, it correctly predicted that these new ligands would form catalysts that react with *L:B* ratios that are substantially lower than those with catalysts containing Xantphos.

To help understand the model’s reasoning behind these results and improve model predictions within this class of ligands, we retrained Libra-ML with these new data points and generated the SHAP value visualization in Figure 6B. This visualization indicates that the $(NEt_2)_2P-$ group significantly contributed to the predicted greater amount of branched isomer formed with the catalysts containing L3. Using the newly trained model, we conducted a second *in silico* exploration campaign. Ligand L4 retains the $(NEt_2)_2P-$ group that is shown to be important by feature visualization and was targeted for synthesis and experimental assessment. The catalyst formed from this ligand was predicted to form the aldehyde with an *L:B* ratio of 1.9:1. It reacted with a regioselectivity of 2.2:1, a value remarkably close to the prediction and unlikely to be predicted with existing heuristics based on the well-known high regioselectivity of Xantphos containing Ph_2P- groups on the backbone.

This sequence shows that the model can make predictions that are distinct from those in the literature and convert the Xantphos ligand to one generating a catalyst reacting with much different selectivity, even though it still contains the backbone leading to a large bite angle that was thought to dominate control over the *L:B* ratio. Moreover, this sequence shows how just a few new datapoints can strongly impact the model’s predictions of selectivity and rapidly expand the chemical domain in which predictions can be trusted. In our case, this expansion was realized for ligands absent in the original training set (non-symmetric xanthene ligands and ligands with amino groups on phosphorus). We anticipate that this computational workflow

and modeling approach, when combined with insights from human-interpretable, data-driven selectivity patterns, may be adapted to other catalytic systems. This workflow should help prioritize candidates that generalize beyond the immediate training domain, reduce the number of active learning cycles and synthetic effort, and, ultimately, reducing the amount of trial-and-error in traditional catalyst development.

Conclusion

This strong performance of Libra-ML, which is founded on our mechanistically guided approach to represent a catalytic process with 3D structures, demonstrates that deep learning based on these 3D structures can accurately predict the experimental outcome of organic reactions catalyzed by transition-metal complexes whose complicated mechanisms have prevented traditional approaches from developing reliable, quantitative structure-selectivity relationships and can begin to reveal the origin of the predictions by visualization methods. By incorporating domain knowledge and a fundamental understanding of the underlying chemical transformations, this approach helps address key challenges facing the application of ML to chemistry and the natural sciences, such as the need for improved model generality, interpretability, and performance. We expect that the approach described herein can be broadly applied to numerous problems in catalysis, and efforts to this end are underway in our laboratory.

ASSOCIATED CONTENT

SUPPORTING INFORMATION

All experimental procedures, details of the computational workflow, procedures and hyperparameters for training the ML models, are provided in the Supporting Information. The Libra-ML package, containing all structure generation and computational workflows and featuring a command-line interface to predict regioselectivity from the SMILES string of an organophosphorus ligand, is available on GitHub (<https://github.com/Hartwig-Group/Libra-ML>), PyPI, and as a preconfigured Docker image. All code and datasets required to replicate this work are available on GitHub (<https://github.com/Hartwig-Group/Libra-ML-Modeling>) and archived on Zenodo.

AUTHOR INFORMATION

Corresponding Authors

Talita Perciano – Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States; email: tperciano@lbl.gov

John F. Hartwig – Department of Chemistry, University of California, Berkeley, Berkeley, CA, United States; Division of Chemical Sciences, Lawrence Berkeley National Laboratory, CA, United States; email: jhartwig@berkeley.edu

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

We thank the NIH for funding the collection of the dataset and development of the initial model at UC Berkeley (1R35GM130387). We thank the Laboratory Directed Research and Development (LDRD) Program of Lawrence Berkeley National Laboratory Program for providing computational resources for structure

generation and model training, as well as for funding studies on the generation and evaluation of new ligands. N.H. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. M.E. was supported by an NIH Kirschstein-NRSA postdoctoral fellowship (F32-GM134579). G.G. was supported by a Pfizer La Jolla Academic Industrial Relations (AIR) Diversity Research Fellowship. This work utilized MGCF resources supported by NIH grant S10OD034382. Instruments used in this work were supported by the PMRC Core and in part by NIH grant S10OD024998. This work was also supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory and Award No. DE-SC0023328 and to Arizona State University (“Visualizing High-dimensional Functions in Scientific Machine Learning”). This research used resources at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy, Office of Science, User Facility under NERSC Award Number ASCR-ERCAP0026937.

ACKNOWLEDGMENT

We thank Dr. Cooper Citek at the LBNL Catalysis Laboratory for experimental and analytical assistance. We thank Drs. Kathleen Durkin and Dave Smalls at the UC Berkeley Molecular Graphics and Computation Facility (MGCF) for computational advice and assistance. We thank Drs. Hasan Celik, Raynald Giovine, and the Pines Magnetic Resonance Center’s Core NMR Facility (PMRC Core) for spectroscopic assistance. We thank Drs. Caleb Geniesse and Tiankai Xie for insightful discussions regarding model visualization techniques.

REFERENCES

- (1) 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* (10), 1622–1637. <https://doi.org/10.1021/acscentsci.1c00535>.
- (2) 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* (4), 837–848. <https://doi.org/10.1021/acs.accounts.0c00807>.
- (3) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. Organic Reactivity from Mechanism to Machine Learning. *Nat. Rev. Chem.* **2021**, *5* (4), 240–255. <https://doi.org/10.1038/s41570-021-00260-x>.
- (4) Yang, W.; Fidelis, T. T.; Sun, W.-H. Machine Learning in Catalysis, from Proposal to Practicing. *ACS Omega* **2020**, *5* (1), 83–88. <https://doi.org/10.1021/acsomega.9b03673>.
- (5) Morán-González, L.; Burnage, A. L.; Nova, A.; Balcells, D. AI Approaches to Homogeneous Catalysis with Transition Metal Complexes. *ACS Catal.* **2025**, *15* (11), 9089–9105. <https://doi.org/10.1021/acscatal.5c01202>.
- (6) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A Data-Intensive Approach to Mechanistic Elucidation Applied to Chiral Anion Catalysis. *Science* **2015**, *347* (6223), 737–743. <https://doi.org/10.1126/science.1261043>.
- (7) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**,

- 360 (6385), 186–190. <https://doi.org/10.1126/science.aar5169>.
- (8) Kozłowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β -Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* **2003**, *125* (22), 6614–6615. <https://doi.org/10.1021/ja0293195>.
 - (9) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. A Machine-Learning Tool to Predict Substrate-Adaptive Conditions for Pd-Catalyzed C–N Couplings. *Science* **2023**, *381* (6661), 965–972. <https://doi.org/10.1126/science.adg2114>.
 - (10) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), eaau5631. <https://doi.org/10.1126/science.aau5631>.
 - (11) Zahrt, A. F.; Rinehart, N. I.; Denmark, S. E. A Conformer-Dependent, Quantitative Quadrant Model. *Eur. J. Org. Chem.* **2021**, *2021* (17), 2343–2354. <https://doi.org/10.1002/ejoc.202100027>.
 - (12) Xu, L.-C.; Zhang, S.-Q.; Li, X.; Tang, M.-J.; Xie, P.-P.; Hong, X. Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed.* **2021**, *60* (42), 22804–22811. <https://doi.org/10.1002/anie.202106880>.
 - (13) Xu, L.-C.; Frey, J.; Hou, X.; Zhang, S.-Q.; Li, Y.-Y.; Oliveira, J. C. A.; Li, S.-W.; Ackermann, L.; Hong, X. Enantioselectivity Prediction of Pallada-Electrocatalysed C–H Activation Using Transition State Knowledge in Machine Learning. *Nat. Synth.* **2023**, *2* (4), 321–330. <https://doi.org/10.1038/s44160-022-00233-y>.
 - (14) Żurański, A. M.; Martínez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. <https://doi.org/10.1021/acs.accounts.0c00770>.
 - (15) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120* (3), 1620–1689. <https://doi.org/10.1021/acs.chemrev.9b00425>.
 - (16) Franke, R.; Selent, D.; Börner, A. Applied Hydroformylation. *Chem. Rev.* **2012**, *112* (11), 5675–5732. <https://doi.org/10.1021/cr3001803>.
 - (17) Hartwig, J. *Organotransition Metal Chemistry: From Bonding to Catalysis*, 1st ed.; University Science books: Sausalito, Calif, 2009.
 - (18) Bronger, R. P. J.; Kamer, P. C. J.; van Leeuwen, P. W. N. M. Influence of the Bite Angle on the Hydroformylation of Internal Olefins to Linear Aldehydes. *Organometallics* **2003**, *22* (25), 5358–5369. <https://doi.org/10.1021/om034012f>.
 - (19) Yu, S.; Zhang, X.; Yan, Y.; Cai, C.; Dai, L.; Zhang, X. Synthesis and Application of Tetraphosphane Ligands in Rhodium-Catalyzed Hydroformylation of Terminal Olefins: High Regioselectivity at High Temperature. *Chem. – Eur. J.* **2010**, *16* (16), 4938–4943. <https://doi.org/10.1002/chem.200903109>.
 - (20) Dotson, J. J.; Van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **2023**, *145* (1), 110–121. <https://doi.org/10.1021/jacs.2c08513>.
 - (21) Wang, H.; Chen, Y.; Yu, H.; Qi, M.; Xia, D.; Qin, M.; Lv, X.; Lu, B.; Gao, R.; Wang, Y.; Mao, S. Personalized Machine Learning Models of Terminal Olefin Hydroformylation for Regioselectivity Prediction. *Chem Catal.* **2024**, *4* (9), 101079. <https://doi.org/10.1016/j.checat.2024.101079>.
 - (22) Shi, H.; Shen, C.; Huang, Z.; Dong, K. Machine Learning-Guided Prediction of Hydroformylation. *Chem-PhysChem* **2024**, *n/a* (n/a), e202400773. <https://doi.org/10.1002/cphc.202400773>.
 - (23) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. <https://doi.org/10.1021/jacs.1c12005>.
 - (24) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22* (11), 586–591. <https://doi.org/10.1021/acscmbosci.0c00118>.
 - (25) Szymkuć, S.; Badowski, T.; Grzybowski, B. A. Is Organic Chemistry Really Growing Exponentially? *Angew. Chem. Int. Ed.* **2021**, *60* (50), 26226–26232. <https://doi.org/10.1002/anie.202111540>.
 - (26) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**. <https://doi.org/10.1126/science.aau5631>.
 - (27) Rinehart, N. I.; Zahrt, A. F.; Denmark, S. The Leveraging Machine Learning for Enantioselective Catalysis: From Dream to Reality. *CHIMIA* **2021**, *75* (7–8), 592–592. <https://doi.org/10.2533/chimia.2021.592>.
 - (28) Kozłowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β -Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* **2003**, *125* (22), 6614–6615. <https://doi.org/10.1021/ja0293195>.
 - (29) Lipkowitz, K. B.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels–Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68* (12), 4648–4656. <https://doi.org/10.1021/jo0267697>.
 - (30) Olen, C. L.; Zahrt, A. F.; Reilly, S. W.; Schultz, D.; Emerson, K.; Candito, D.; Wang, X.; Strotman, N. A.; Denmark, S. E. Chemoinformatic Catalyst Selection Methods for the Optimization of Copper–Bis(Oxazoline)-Mediated, Asymmetric, Vinylogous Mukaiyama Aldol Reactions. *ACS Catal.* **2024**, *14* (4), 2642–2655. <https://doi.org/10.1021/acscatal.3c05903>.
 - (31) Lexa, K. W.; Belyk, K. M.; Henle, J.; Xiang, B.; Sheridan, R. P.; Denmark, S. E.; Ruck, R. T.; Sherer, E. C. Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts. *Org. Process Res. Dev.* **2022**, *26* (3), 670–682. <https://doi.org/10.1021/acs.oprd.1c00155>.

- (32) Rose, B. T.; Timmerman, J. C.; Bawel, S. A.; Chin, S.; Zhang, H.; Denmark, S. E. High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2-Amino-6-Arylpyridines. *J. Am. Chem. Soc.* **2022**. <https://doi.org/10.1021/jacs.2c08820>.
- (33) Schnitzer, T.; Schnurr, M.; Zahrt, A. F.; Sakhaee, N.; Denmark, S. E.; Wennemers, H. Machine Learning to Develop Peptide Catalysts—Successes, Limitations, and Opportunities. *ACS Cent. Sci.* **2024**, *10* (2), 367–373. <https://doi.org/10.1021/acscentsci.3c01284>.
- (34) Batatia, I.; Kovács, D. P.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *arXiv* January 26, 2023. <https://doi.org/10.48550/arXiv.2206.07697>.
- (35) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13* (1), 2453. <https://doi.org/10.1038/s41467-022-29939-5>.
- (36) Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.* **2022**, *4* (2), 127–134. <https://doi.org/10.1038/s42256-021-00438-4>.
- (37) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; Bin, R. D.; Balcells, D. Deep Learning Metal Complex Properties with Natural Quantum Graphs. *Digit. Discov.* **2023**, *2* (3), 618–633. <https://doi.org/10.1039/D2DD00129B>.
- (38) Zhu, Y.; Hwang, J.; Adams, K.; Liu, Z.; Nan, B.; Stenfors, B.; Du, Y.; Chauhan, J.; Wiest, O.; Isayev, O.; Coley, C. W.; Sun, Y.; Wang, W. Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks. *arXiv* July 28, 2024. <https://doi.org/10.48550/arXiv.2310.00115>.
- (39) van Gerwen, P.; Briling, K. R.; Bunne, C.; Somnath, V. R.; Laplaza, R.; Krause, A.; Corminboeuf, C. 3DReact: Geometric Deep Learning for Chemical Reactions. *J. Chem. Inf. Model.* **2024**, *64* (15), 5771–5785. <https://doi.org/10.1021/acs.jcim.4c00104>.
- (40) Nippa, D. F.; Atz, K.; Hohler, R.; Müller, A. T.; Marx, A.; Bartelmus, C.; Wuitschik, G.; Marzuoli, I.; Jost, V.; Wolfard, J.; Binder, M.; Stepan, A. F.; Konrad, D. B.; Grether, U.; Martin, R. E.; Schneider, G. Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning. *Nat. Chem.* **2024**, *16* (2), 239–248. <https://doi.org/10.1038/s41557-023-01360-5>.
- (41) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37* (22), 2106–2117. <https://doi.org/10.1002/jcc.24437>.
- (42) Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57* (42), 13973–13986. <https://doi.org/10.1021/acs.iecr.8b04015>.
- (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* (3), 1652–1671. <https://doi.org/10.1021/acs.jctc.8b01176>.
- (44) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All Spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017**, *13* (5), 1989–2009. <https://doi.org/10.1021/acs.jctc.7b00118>.
- (45) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313–2323. <https://doi.org/10.1021/acscatal.8b04043>.
- (46) Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. Chemical Complexity Challenge: Is Multi-Instance Machine Learning a Solution? *WIREs Comput. Mol. Sci.* **2024**, *14* (1), e1698. <https://doi.org/10.1002/wcms.1698>.
- (47) Zankov, D.; Madzhidov, T.; Polishchuk, P.; Sidorov, P.; Varnek, A. Multi-Instance Learning Approach to the Modeling of Enantioselectivity of Conformationally Flexible Organic Catalysts. *J. Chem. Inf. Model.* **2023**, *63* (21), 6629–6641. <https://doi.org/10.1021/acs.jcim.3c00393>.
- (48) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.* **2020**, *59* (36), 15665–15673. <https://doi.org/10.1002/anie.202004239>.
- (49) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160* (11), 114110. <https://doi.org/10.1063/5.0197592>.
- (50) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Data Augmentation Strategies to Improve Reaction Yield Predictions and Estimate Uncertainty. *ChemRxiv* November 26, 2020. <https://doi.org/10.26434/chemrxiv.13286741.v1>.
- (51) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes. *Org. Chem. Front.* **2021**, *8* (7), 1415–1423. <https://doi.org/10.1039/D0QO01636E>.
- (52) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17. <https://doi.org/10.1021/acs.jcim.3c01250>.
- (53) Gasteiger, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* April 5, 2022. <https://doi.org/10.48550/arXiv.2003.03123>.
- (54) Gasteiger, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* April 5, 2022. <https://doi.org/10.48550/arXiv.2011.14115>.
- (55) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust Deep Learning-Based Protein

- Sequence Design Using ProteinMPNN. *Science* **2022**, 378 (6615), 49–56. <https://doi.org/10.1126/science.add2187>.
- (56) Pun, M. N.; Ivanov, A.; Bellamy, Q.; Montague, Z.; La-Mont, C.; Bradley, P.; Otwinowski, J.; Nourmohammad, A. Learning the Shape of Protein Microenvironments with a Holographic Convolutional Neural Network. *Proc. Natl. Acad. Sci.* **2024**, 121 (6), e2300838121. <https://doi.org/10.1073/pnas.2300838121>.
- (57) Li, H.; Xu, Z.; Taylor, G.; Studer, C.; Goldstein, T. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2018; Vol. 31.
- (58) Xie, T.; Chen, J.; Yang, Y.; Geniesse, C.; Shi, G.; Chaudhari, A.; Cava, J. K.; Mahoney, M. W.; Perciano, T.; Weber, G. H.; Maciejewski, R. LossLens: Diagnostics for Machine Learning through Loss Landscape Visual Analytics. arXiv December 17, 2024. <https://doi.org/10.48550/arXiv.2412.13321>.
- (59) Yang, Y.; Hodgkinson, L.; Theisen, R.; Zou, J.; Gonzalez, J. E.; Ramchandran, K.; Mahoney, M. W. Taxonomizing Local versus Global Structure in Neural Network Loss Landscapes. arXiv December 12, 2021. <https://doi.org/10.48550/arXiv.2107.11228>.
- (60) Iyer, N.; Thejas, V.; Kwatra, N.; Ramjee, R.; Sivathanu, M. Wide-Minima Density Hypothesis and the Explore-Exploit Learning Rate Schedule. arXiv June 1, 2021. <https://doi.org/10.48550/arXiv.2003.03977>.
- (61) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. arXiv November 25, 2017. <https://doi.org/10.48550/arXiv.1705.07874>.
- (62) Kamer, P. C. J.; van Leeuwen, P. W. N. M.; Reek, J. N. H. Wide Bite Angle Diphosphines: Xantphos Ligands in Transition Metal Complexes and Catalysis. *Acc. Chem. Res.* **2001**, 34 (11), 895–904. <https://doi.org/10.1021/ar000060+>.
- (63) van der Veen, L. A.; Keeven, P. H.; Schoemaker, G. C.; Reek, J. N. H.; Kamer, P. C. J.; van Leeuwen, P. W. N. M.; Lutz, M.; Spek, A. L. Origin of the Bite Angle Effect on Rhodium Diphosphine Catalyzed Hydroformylation. *Organometallics* **2000**, 19 (5), 872–883. <https://doi.org/10.1021/om990734o>.

