Title

Predicting Enantioselectivity of Ruthenium-Catalyzed Ketone Hydrogenation with 3D Structure-Based Deep Learning

Journal

Journal of the American Chemical Society, 148(30)

ISSN

0002-7863

Authors

Hadler, Nicholas

Ramirez, Evelyn

Avaylon, Matthew

et al.

Publication Date

2026-08-05

DOI

10.1021/jacs.6c07313

Predicting Enantioselectivity of Ruthenium-Catalyzed Ketone Hydrogenation with 3D Structure-Based Deep Learning

Nicholas Hadler,[1] Evelyn Ramirez,[1] Matthew Avaylon,[2] Masha Elkin,[1] N. Ian Rinehart,[1,3] Michael W. Mahoney,[2,4,5] Gunther H. Weber,[2] Joscha Hoche,[6] Jan Gerit Brandenburg,[6] Talita Perciano,[2,*] and John F. Hartwig[1,3,*]

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

[2] Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States.

[3] Division of Chemical Sciences, Lawrence Berkeley National Laboratory, CA, United States.

[4] International Computer Science Institute, CA, United States.

[5] Department of Statistics, University of California, Berkeley, CA, United States.

[6] Merck KGaA, Frankfurter Str, Darmstadt, Germany.

ABSTRACT: We report machine-learning (ML) models that predict with high accuracy the enantioselectivity of ketone hydrogenation with Noyori-type diphosphine–diamine catalysts. To mitigate bias in published data, we combined a dataset curated from the literature with >1,000 examples from our own high-throughput experimentation. We demonstrate that 3D graph neural networks, trained on 3D representations derived from the enantiodetermining transition state, outperform models based on 2D-graphs, molecular fingerprints, and descriptors derived from semiempirical quantum calculations. These results indicate that models trained on 3D structural representations that reflect catalyst-substrate interactions relevant to enantioselectivity can accurately extrapolate to structures distinct from those in the training set.

The enantioselective hydrogenation of ketones is a valuable transformation catalyzed by organometallic complexes,[1–4] and is now widely used in both academic and industrial synthesis.[5] Noyori-type ruthenium(II) complexes containing a diphosphine ligand, a diamine ligand, and two trans chlorides often catalyze the hydrogenation of unfunctionalized ketones with high enantioselectivity, chemoselectivity, and functional-group tolerance.

Enantioselective hydrogenations with Noyori's catalysts generally proceed by a metal–ligand bifunctional mechanism, in which a metal hydride and the N–H group of the diamine directly participate in the stereodetermining, outer-sphere addition of hydride to the substrate (Figure 1).[6–10] These catalyst–substrate interactions strongly influence enantioselectivity, and no universal catalyst exists. Achieving the highest selectivity typically requires varying both the diphosphine and diamine ligands. Thus, identifying the most active and selective catalyst in this family for hydrogenation of a specific ketone is often time- and resource-intensive.

We envisioned that machine learning (ML) could help identify the optimal combination of diphosphine and diamine for a specific ketone. Prior work on predicting enantioselectivity with Noyori-type catalysts by computation has been limited to studies by Norrby[11] in 2014, who reported a bespoke force-field parameterized on seven transition-states and validated on thirteen literature reactions, and by Lei[12] in 2016, who used comparative molecular field analysis to predict enantioselectivity of five reactions after training a model on 20 literature reactions. More generally, methods for predicting enantioselectivity have used multivariate linear regression[13–19] and ML models trained on fingerprint-,[20–24] descriptor-,[25–34] voxel-,[35–40] transition-state,[41–46] and graph-based representations.[47–53]

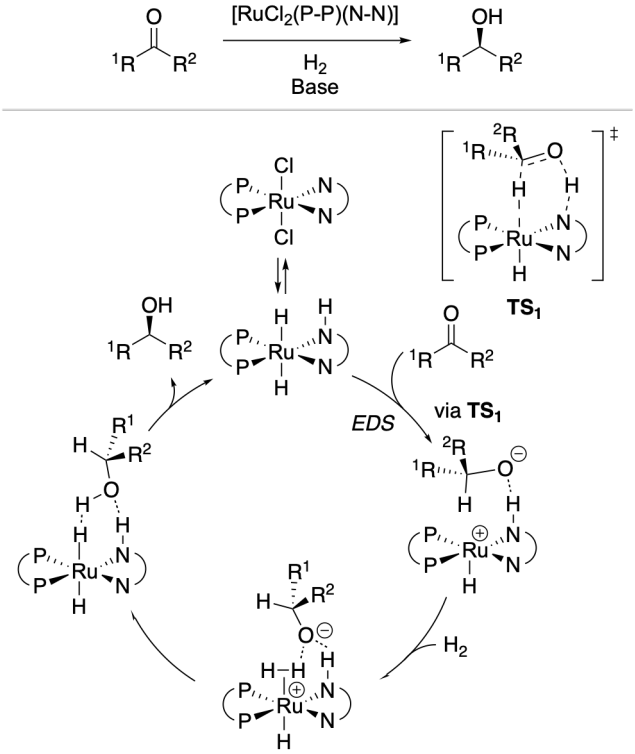


Figure 1: Mechanism of Ru-catalyzed asymmetric hydrogenation of ketones with Noyori-type catalysts.

For ML to guide catalyst selection, the models must be accurate for reactions of a diverse set of substrate–catalyst pairs

and remain accurate when applied to structures distinct from those in the training set. The large diversity of published catalysts and substrates for ketone hydrogenation makes the enantioselectivity obtained with Noyori-type catalysts a suitable yardstick for a model's ability to predict stereoselectivity in homogeneous, transition-metal catalysis.

The preponderance of high enantioselectivity in reported examples[54] leads to the well appreciated overrepresentation of catalysts that are highly active or selective, or both. Thus, models trained solely on published data fail to learn what leads to low activity or selectivity, and they often incorrectly predict the outcomes of reactions comprising substrates or catalysts that are structurally distinct from those represented in the training set. This issue was underscored by Kumar in 2023[55] with a dataset of ester hydrogenations collated from the literature.

Recently, our group reported the use of 3D graph neural networks (3D-GNNs) trained on catalytic intermediates to predict the regioselectivity of the hydroformylation of terminal olefins.[56] Building on this strategy, we sought to determine whether structure-based deep learning could predict enantioselectivity controlled by a single transition-state and whether a 3D representation of this enantiodetermining transition-state would lead to more accurate predictions than approaches followed previously.

We report ML models, based on 3D representations of the transition-state, that predict enantioselectivity for the asymmetric hydrogenation of ketones with high accuracy. To train these models, we curated a dataset of published results and, to address the scarcity of low and moderate enantioselectivities, obtained >1,000 datapoints with a variety of catalysts and substrates by high-throughput experimentation (HTE). We compared several state-of-the-art ML methods using train–test splits designed to evaluate both model interpolation and extrapolation and found that our model based on 3D-GNN representations of the enantiodetermining transition-state consistently outperformed models based on 2D-graphs, molecular fingerprints, and descriptors derived from semiempirical quantum calculations.

To begin, we curated a dataset comprising published results on the asymmetric hydrogenation of aryl ketones catalyzed by Noyori-type ruthenium(II) complexes bearing a diphosphine ligand and a diamine ligand. Each paper and its Supporting Information were inspected, and data tables were collated manually. Each reaction in the dataset included only a ketone, a ruthenium complex, a bidentate organophosphine ligand, a bidentate amine ligand, base, solvent, and hydrogen gas. Reactions not meeting this criterion were excluded. Reported enantioselectivities were rounded to the third decimal place and transformed into the corresponding absolute free-energy difference between the transition-states leading to the enantiomeric products ($\Delta\Delta G^{\ddagger}$, kcal/mol) calculated from the reported reaction temperature. The resulting curated dataset was derived from 33 papers describing 725 reactions that encompass 163 unique catalysts and 249 unique substrates.

The set of reactions from the literature contained a disproportionate number of examples with high enantioselectivity (Figure 2A); over half of the reported reactions occurred with enantioselectivities above 90%. Although skewed, such datasets do offer advantages because they often include a broader range of structures by including molecules that are not commercially available or readily prepared. Rather than discarding these training examples, we followed a synergistic strategy that combined data from the literature with data generated by HTE in our laboratory and deliberately designed to sample reactions that occur with low to moderate enantioselectivity.

We constructed a library of Noyori-type catalysts by systematically varying commercially available diphosphines (P–P) and diamines (N–N) in 96-well plates, enabling rapid, modular assembly of hundreds of ruthenium complexes. We assembled Ru(P–P)(N–N)Cl_2 complexes by combining Ru(nbd)(py)$_2Cl_2$ (nbd = norbornadiene; py = pyridine) with 48 diphosphines and 32 diamines. The ligands were divided into two independent sets of 24 diphosphines and 16 diamines with no overlapping diphosphines or diamines. For each set, 24 Ru(P–P)(py)$_2Cl_2$ stock solutions were dispensed across one dimension of a 96-well plate and 16 diamine stock solutions were dispensed along the other dimension, generating 384 Ru(P–P)(N–N)Cl_2 complexes (24 × 16) per set. The resulting 768 Ru(P–P)(N–N)Cl_2 complexes were used without further purification. Full experimental details are provided in the Supporting Information.

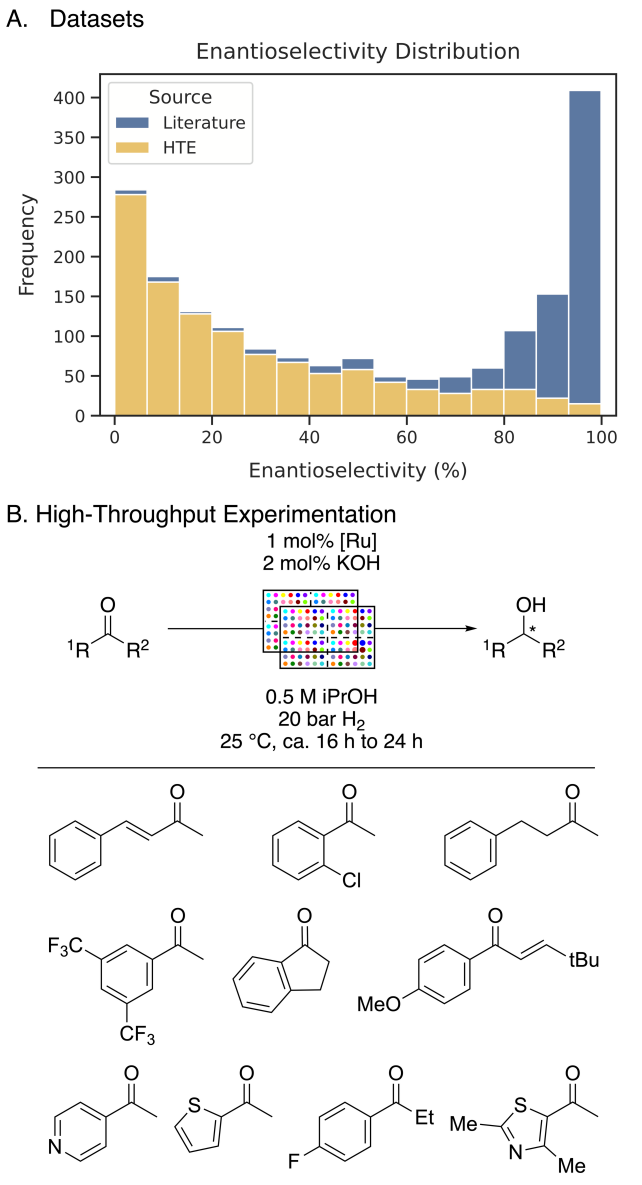


Figure 2: (A) Distribution of enantioselectivity (%) for both the literature and HTE data. (B) Scheme and substrates of the HTE campaign.

Ten representative ketones (Figure 2B) were then hydrogenated with 384 different catalysts by dispensing stock solutions of each substrate into the prepared plates and running the reactions in a custom high-pressure reactor (20 bar H_2) for 24 h at room temperature. After venting, reactions were analyzed by chiral HPLC to measure enantioselectivity and the ratio of the product peak area to reactant peak area. Reactions with peak area ratios reflecting yields < 80% were excluded, resulting in an experimental dataset of 1,154 reactions that encompasses 500 discrete catalysts and 10 discrete substrates. Combined with the literature dataset, the final dataset contained 1,877 reactions with 660 catalysts and 254 substrates. This dataset provided results of reactions occurring with low and moderate enantioselectivity that were absent from the literature dataset (Figure 2A).

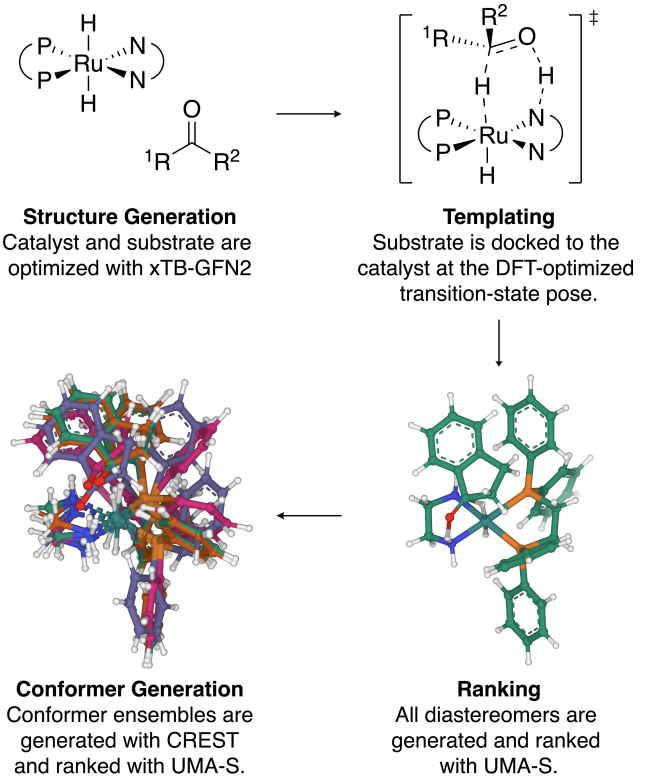


Figure 3: Workflow for rapid generation of 3D transition-state representations.

Next, we generated the 3D representations needed to train a 3D-GNN that would predict enantioselectivity. We hypothesized that the geometry of the enantiodetermining transition-state (Figure 1, TS1), which includes both the catalyst and the substrate, would best capture the non-covalent interactions underlying enantioselectivity. Because the optimization of thousands of transition states with accurate levels of theory can be prohibitively expensive, we developed a workflow (Figure 3) that templates catalyst–substrate pairs onto a constrained, ground-state geometry derived from a DFT-optimized transition-state.[9] This structure captures relevant non-covalent interactions and can be optimized more quickly than a transition-state.

All catalysts were built in Molli[57] to ensure correct absolute configurations and were then optimized with xTB-GFN2.[58] Substrates were docked to the catalyst by placing each substrate 10 Å above the templated position and running a constrained optimization that converged to the target pose. For each catalyst–substrate pair, both diastereomers were generated, single-point energies were computed with UMA-S-1.1,[59,60] and the lowest-energy diastereomer was retained. Conformer searches with CREST[61,62] were performed, and conformers were re-ranked with single-point energies calculated using UMA-S. To increase the effective size of the training set, up to 10 lowest-energy conformers were included as separate entries. Full details can be found in the Supporting Information. This approach conducted on all 1,877 reactions yielded a final dataset of 8,100 entries.

To evaluate the accuracy of ML models that predict experimental enantioselectivity, we assessed performance using three methods to split the data: (i) interpolative, in which catalysts and substrates are shared between the training and test sets; (ii) catalyst-extrapolative, in which the catalysts in the test set were excluded from the training set; and (iii) substrate-extrapolative, in which the substrates in the test set were excluded from the training-set. Interpolative splits typically yield the highest apparent model performance, but they can overestimate the generality of the model because they emphasize patterns represented in the training data. The extrapolative splits more stringently test the generality of the model because they assess whether the model can predict outcomes accurately when the catalyst or substrate lies outside the types of structures in the training set.

We compared our 3D-GNN approach to a series of ML methods commonly applied to transition-metal catalysis: 2D-GNNs, such as Chemprop,[63] the 2D-GNN foundation model CheMeleon,[64] and tabular regressors trained on either molecular fingerprints or descriptors derived from semiempirical quantum calculations of transition-state structures. These descriptors include the bite angle, percent buried volume, Sterimol values, partial charges, and other steric and electronic features detailed in the Supporting Information. To assess whether performance differences among the evaluated ML methods were statistically significant, we followed the procedure recommended by Walters et al.[65] We conducted 5×5 repeated cross-validation to assess interpolative performance, in which 5-fold cross-validation was repeated five times with different random partitions, and repeated cross-validation split by catalyst (5×5) or substrate (5×3) to assess extrapolative performance.

For each method and fold, predictions were obtained by averaging outputs from five independent replicates, each trained on the corresponding training partition with a different random initialization seed. Performance metrics were then computed on the averaged predictions for each test partition. Differences were assessed using a one-way ANOVA or Friedman test on the MAE computed for each fold, followed by Tukey's HSD or Conover–Holm post hoc pairwise comparisons. Results tables for each splitting strategy, which include all evaluated methods, performance metrics (MAE, R^2, RMSE, Spearman's ρ, $MCC_{90\%}$, stratified MAE), and the outcomes of the pairwise statistical analyses, are provided in the Supporting Information.

Under the interpolative train-test split (Figure 4A), the MAE of the 3D-GNN DimeNet++ trained on 3D transition-state representations with 10-fold conformer augmentation was lower and the R^2 higher (MAE: 0.299 kcal/mol, R^2: 0.817) than those of the other methods evaluated by statistically significant amounts. The performance of CheMeleon (MAE: 0.312 kcal/mol, R^2: 0.810) was second-best and was significantly better than that of Chemprop (MAE: 0.447 kcal/mol, R^2: 0.660). Random-forest regressors trained on descriptors derived from semiempirical quantum calculations with and without

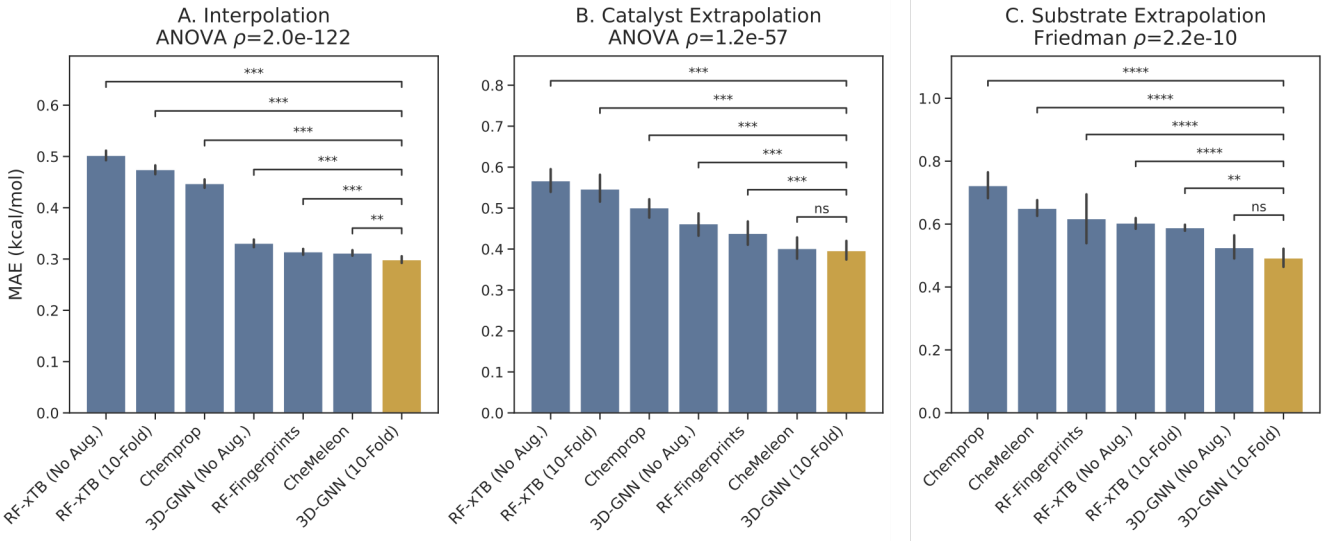


Figure 4. Mean absolute error (MAE) averaged across models for (A) interpolative, (B) catalyst-extrapolative, and (C) substrate-extrapolative splits. Brackets denote pairwise comparisons; asterisks indicate significant differences in mean MAE (ns, not significant; *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001).

conformer-based data augmentation were among the least accurate models evaluated (MAE: 0.547 kcal/mol, R^2: 0.483; MAE: 0.567 kcal/mol, R^2: 0.468, respectively). We observed a similar trend in model performance for the catalyst-extrapolative split (Figure 4B). The MAEs from DimeNet++ with 10-fold conformer-based augmentation (MAE: 0.396 kcal/mol, R^2: 0.680) and CheMeleon (MAE: 0.401 kcal/mol, R^2: 0.678) were the lowest and the R^2 higher than those of the methods tested.

The substrate-extrapolative split (Figure 4C) is the most challenging evaluation because it is less combinatorial than the catalyst-extrapolative split. In this case, the MAE for DimeNet++ trained with conformer augmentation (MAE: 0.492 kcal/mol, R^2: 0.614) and without conformer augmentation (MAE: 0.525 kcal/mol, R^2: 0.565) was lower and the R^2 was higher than those of the other methods evaluated. CheMeleon (MAE: 0.650 kcal/mol, R^2: 0.417), which was the next-best-performing model in both the interpolative and catalyst-extrapolative splits, was significantly less accurate when extrapolating to substrates outside the training set.

Thus, 3D-GNNs trained on 3D representations derived from the enantiodetermining transition state gave the most accurate predictions across all three data splits, including the most challenging substrate-extrapolative split. We hypothesize that the strong performance of this approach arises from its ability to capture catalyst-substrate interactions relevant to enantioselectivity. Consistent with this interpretation, the 3D-GNN appears to learn structure-selectivity relationships that extrapolate more readily to substrates distinct from those in the training data than models trained on 2D or descriptor-based representations.

In summary, we report ML models trained to predict the enantioselectivity of ketone hydrogenation with Noyori-type catalysts. To train these models, we combined a set of curated data from the literature with >1,000 experimental datapoints from HTE enriched with reactions occurring with low to moderate enantioselectivity. These results indicate that 3D-GNNs can predict enantioselectivity with high accuracy for a broad range of catalysts and substrates. We expect this strategy will be valuable to predict selectivity and potentially activity for additional catalytic reactions, and such work is underway in our laboratory.

ASSOCIATED CONTENT

DATA AVAILABILITY

The data used to generate the results in this paper can be found on GitHub at this link: https://github.com/Hartwig-Group/NAH-Predict

SUPPORTING INFORMATION

All experimental procedures, details of the computational workflow, procedures and hyperparameters for training the ML models, are provided in the Supporting Information.

AUTHOR INFORMATION

Corresponding Authors

Talita Perciano – Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States 94720; email: tperciano@lbl.gov

John F. Hartwig – Department of Chemistry, University of California, Berkeley, CA, United States 94720; Division of Chemical Sciences, Lawrence Berkeley National Laboratory, CA, United States 94720; email: jhartwig@berkeley.edu

Authors

Nicholas Hadler – Department of Chemistry, University of California, Berkeley, CA, United States 94720

Evelyn Ramirez – Department of Chemistry, University of California, Berkeley, CA, United States 94720

Matthew Avaylon – Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States 94720

Masha Elkin – Department of Chemistry, University of California, Berkeley, CA, United States 94720

N. Ian Rinehart – Department of Chemistry, University of California, Berkeley, CA, United States 94720; Division of Chemical Sciences, Lawrence Berkeley National Laboratory, CA, United States 94720

Michael W. Mahoney – Department of Statistics, University of California, Berkeley, CA, United States 94720; International Computer Science Institute, CA, United States 94704; Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States 94720

Gunther H. Weber – Scientific Data Division, Lawrence Berkeley National Laboratory, CA, United States 94720

Joscha Hoche – Merck KGaA, Frankfurter Strasse 250 Darmstadt, 64293, Germany.

Jan Gerit Brandenburg – Merck KGaA, Frankfurter Strasse 250 Darmstadt, 64293, Germany.

Author Contributions

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

ACKNOWLEDGMENT

We thank Merck KGaA for funding the development of 3D GNN algorithms. We thank the NIH for funding the collection of the dataset at UC Berkeley and for support of model training (1R35GM130387). 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. E.R. was supported by the Berkeley Amgen Scholars Program. M.E. was supported by an NIH Kirschstein-NRSA postdoctoral fellowship (F32-GM134579). 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. T.P. and M.A. were supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory. M.W.M., T.P., and G.H.W. were supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program, both under Contract No. DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. Under these awards, M.W.M. acknowledges the DOE Competitive Portfolios grant and the DOE SciGPT grant, and T.P. and G.H.W. acknowledge support from the grant "Visualizing High-Dimensional Functions in Scientific Machine Learning."

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.

REFERENCES

(1) Noyori, R.; Ohkuma, T. Asymmetric Catalysis by Architectural and Functional Molecular Engineering: Practical Chemo- and Stereoselective Hydrogenation of Ketones. *Angew. Chem. Int. Ed.* **2001**, *40* (1), 40–73. https://doi.org/10.1002/1521-3773(20010105)40:1%3C40::AID-ANIE40%3E3.0.CO;2-5.

(2) Ohkuma, T.; Koizumi, M.; Doucet, H.; Pham, T.; Kozawa, M.; Murata, K.; Katayama, E.; Yokozawa, T.; Ikariya, T.; Noyori, R. Asymmetric Hydrogenation of Alkenyl, Cyclopropyl, and Aryl Ketones. $RuCl_2$(Xylbinap)(1,2-Diamine) as a Precatalyst Exhibiting a Wide Scope. *J. Am. Chem. Soc.* **1998**, *120* (51), 13529–13530. https://doi.org/10.1021/ja983257u.

(3) Ohkuma, T.; Doucet, H.; Pham, T.; Mikami, K.; Korenaga, T.; Terada, M.; Noyori, R. Asymmetric Activation of Racemic Ruthenium(II) Complexes for Enantioselective Hydrogenation. *J. Am. Chem. Soc.* **1998**, *120* (5), 1086–1087. https://doi.org/10.1021/ja972897e.

(4) Ohkuma, T.; Ooka, H.; Hashiguchi, S.; Ikariya, T.; Noyori, R. Practical Enantioselective Hydrogenation of Aromatic Ketones. *J. Am. Chem. Soc.* **1995**, *117* (9), 2675–2676. https://doi.org/10.1021/ja00114a043.

(5) Noyori, R. Asymmetric Catalysis: Science and Opportunities (Nobel Lecture) Copyright© The Nobel Foundation 2002. We Thank the Nobel Foundation, Stockholm, for Permission to Print This Lecture. *Angew. Chem. Int. Ed.* **2002**, *41* (12), 2008. https://doi.org/10.1002/1521-3773(20020617)41:12%3C2008::AID-ANIE2008%3E3.0.CO;2-4.

(6) Sandoval, C. A.; Ohkuma, T.; Muñiz, K.; Noyori, R. Mechanism of Asymmetric Hydrogenation of Ketones Catalyzed by BINAP/1,2-Diamine−Ruthenium(II) Complexes. *J. Am. Chem. Soc.* **2003**, *125* (44), 13490–13503. https://doi.org/10.1021/ja030272c.

(7) Dub, P. A.; Gordon, J. C. The Role of the Metal-Bound N–H Functionality in Noyori-Type Molecular Catalysts. *Nat. Rev. Chem.* **2018**, *2* (12), 396–408. https://doi.org/10.1038/s41570-018-0049-z.

(8) Abdur-Rashid, K.; Clapham, S. E.; Hadzovic, A.; Harvey, J. N.; Lough, A. J.; Morris, R. H. Mechanism of the Hydrogenation of Ketones Catalyzed by *Trans*-Dihydrido(Diamine)Ruthenium(II) Complexes. *J. Am. Chem. Soc.* **2002**, *124* (50), 15104–15118. https://doi.org/10.1021/ja016817p.

(9) Dub, P. A.; Henson, N. J.; Martin, R. L.; Gordon, J. C. Unravelling the Mechanism of the Asymmetric Hydrogenation of Acetophenone by [RuX_2(Diphosphine)(1,2-Diamine)] Catalysts. *J. Am. Chem. Soc.* **2014**, *136* (9), 3505–3521. https://doi.org/10.1021/ja411374j.

(10) Dub, P. A.; Gordon, J. C. The Mechanism of Enantioselective Ketone Reduction with Noyori and Noyori–Ikariya Bifunctional Catalysts. *Dalton Trans.* **2016**, *45* (16), 6756–6781. https://doi.org/10.1039/C6DT00476H.

(11) Limé, E.; Lundholm, M. D.; Forbes, A.; Wiest, O.; Helquist, P.; Norrby, P.-O. Stereoselectivity in Asymmetric Catalysis: The Case of Ruthenium-Catalyzed Ketone Hydrogenation. *J. Chem. Theory Comput.* **2014**, *10* (6), 2427–2435. https://doi.org/10.1021/ct500178w.

(12) Li, L.; Pan, Y.; Lei, M. The Enantioselectivity in Asymmetric Ketone Hydrogenation Catalyzed by RuH_2(Diphosphine)(Diamine) Complexes: Insights from a 3D-QSSR and DFT Study. *Catal. Sci. Amp Technol.* **2016**, *6* (12), 4450–4457. https://doi.org/10.1039/C5CY01225B.

(13) Kozlowski, 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.

(14) Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1876–1881. https://doi.org/10.1021/ci034119d.

(15) Jensen, K. H.; Sigman, M. S. Systematically Probing the Effect of Catalyst Acidity in a Hydrogen-Bond-Catalyzed Enantioselective Reaction. *Angew. Chem. Int. Ed.* **2007**, *46* (25), 4748–4750. https://doi.org/10.1002/anie.200700298.

(16) Harper, K. C.; Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. *Science* **2011**, *333* (6051), 1875–1878. https://doi.org/10.1126/science.1206997.

(17) 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.

(18) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-Catalyzed C–C Bond Formation Enabled through Ligand Parameterization. *Science* **2018**, *362* (6415), 670–674. https://doi.org/10.1126/science.aat2299.

(19) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(20) Betinol, I. O.; Demchenko, A.; Reid, J. P. Evaluating Predictive Accuracy in Asymmetric Catalysis: A Machine Learning Perspective on Local Reaction Space. *ACS Catal.* **2025**, *15* (8), 6067–6077. https://doi.org/10.1021/acscatal.5c01051.

(21) Gao, Y.; Hu, K.; Rao, J.; Zhu, Q.; Liao, K. Artificial Intelligence-Driven Development of Nickel-Catalyzed Enantioselective Cross-Coupling Reactions. *ACS Catal.* **2024**, *14* (24), 18457–18468. https://doi.org/10.1021/acscatal.4c04277.

(22) Baczewska, P.; Kulczykowski, M.; Zambroń, B.; Jaszczewska-Adamczak, J.; Pakulski, Z.; Roszak, R.; Grzybowski, B. A.; Mlynarski, J. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions. *Angew. Chem. Int. Ed.* **2024**, *63* (37), e202318487. https://doi.org/10.1002/anie.202318487.

(23) Yang, Q.; Liu, Y.; Cheng, J.; Li, Y.; Liu, S.; Duan, Y.; Zhang, L.; Luo, S. An Ensemble Structure and Physicochemical (SPOC) Descriptor for Machine-Learning Prediction of Chemical Reaction and Molecular Properties. *ChemPhysChem* **2022**, *23* (14), e202200255. https://doi.org/10.1002/cphc.202200255.

(24) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.

(25) Xu, X.-Y.; Liu, L.-G.; Xu, L.-C.; Zhang, S.-Q.; Hong, X. Transfer Learning-Enabled Ligand Prediction for Ni-Catalyzed Atroposelective Suzuki–Miyaura Cross-Coupling Based on Mechanistic Similarity: Leveraging Pd Knowledge for Ni Discovery. *J. Am. Chem. Soc.* **2025**, *147* (18), 15318–15328. https://doi.org/10.1021/jacs.5c00838.

(26) Ocampo, B. E.; Altundas, B.; Bock, M. J.; Feiz, S.; Denmark, S. E. Data-Driven Prediction of Enantioselectivity for the Sharpless Asymmetric Dihydroxylation: Model Development and Experimental Validation. *ACS Cent. Sci.* **2025**, *11* (9), 1640–1650. https://doi.org/10.1021/acscentsci.5c00900.

(27) Nakajima, H.; Murata, C.; Noto, N.; Saito, S. Database Construction for the Virtual Screening of the Ruthenium-Catalyzed Hydrogenation of Ketones. *J. Org. Chem.* **2025**, *90* (2), 1054–1060. https://doi.org/10.1021/acs.joc.4c02347.

(28) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. Probing Machine Learning Models Based on High Throughput Experimentation Data for the Discovery of Asymmetric Hydrogenation Catalysts. *Chem. Sci.* **2024**, *15* (34), 13618–13630. https://doi.org/10.1039/D4SC03647F.

(29) Chen, G.-M.; Ye, Z.-H.; Li, Z.-M.; Zhang, J.-L. Universal Descriptors of Quasi Transition States for Small-Data-Driven Asymmetric Catalysis Prediction in Machine Learning Model. *Cell Rep. Phys. Sci.* **2024**, *5* (7), 102043. https://doi.org/10.1016/j.xcrp.2024.102043.

(30) Zhang, Z.-J.; Li, S.-W.; Oliveira, J. C. A.; Li, Y.; Chen, X.; Zhang, S.-Q.; Xu, L.-C.; Rogge, T.; Hong, X.; Ackermann, L. Data-Driven Design of New Chiral Carboxylic Acid for Construction of Indoles with C-Central and C–N Axial Chirality via Cobalt Catalysis. *Nat. Commun.* **2023**, *14* (1), 3149. https://doi.org/10.1038/s41467-023-38872-0.

(31) 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.

(32) 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.

(33) Xu, L.; Zhang, S.; Li, X.; Tang, M.; Xie, 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.

(34) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A Unified Machine-Learning Protocol for Asymmetric Catalysis as a Proof of Concept Demonstration Using Asymmetric Hydrogenation. *Proc. Natl. Acad. Sci.* **2020**, *117* (3), 1339–1345. https://doi.org/10.1073/pnas.1916392117.

(35) 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.

(36) 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.

(37) 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**, *144* (50), 22950–22964. https://doi.org/10.1021/jacs.2c08820.

(38) Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E. Computational Methods for Training Set Selection and Error Assessment Applied to Catalyst Design: Guidelines for Deciding Which Reactions to Run First and Which to Run Next. *React. Chem. Amp Eng.* **2021**, *6* (4), 694–708. https://doi.org/10.1039/D1RE00013F.

(39) 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.

(40) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142* (26), 11578–11592. https://doi.org/10.1021/jacs.0c04715.

(41) 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.

(42) Gallarati, S.; Bucci, E. M.; Doyle, A. G.; Sigman, M. S. Transferable Enantioselectivity Models from Sparse Data. *Nature* **2026**, *651* (8106), 637–646. https://doi.org/10.1038/s41586-026-10239-7.

(43) Cuomo, A. E.; Ibarraran, S.; Sreekumar, S.; Li, H.; Eun, J.; Menzel, J. P.; Zhang, P.; Buono, F.; Song, J. J.; Crabtree, R. H.; Batista, V. S.; Newhouse, T. R. Feed-Forward Neural Network for Predicting Enantioselectivity of the Asymmetric Negishi Reaction. *ACS Cent. Sci.* **2023**, *9* (9), 1768–1774. https://doi.org/10.1021/acscentsci.3c00512.

(44) Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-O. Steric Influences on the Selectivity in Palladium-Catalyzed Allylation. *Organometallics* **1997**, *16* (13), 3015–3021. https://doi.org/10.1021/om9700371.

(45) Reid, J. P.; Hu, M.; Ito, S.; Huang, B.; Hong, C. M.; Xiang, H.; Sigman, M. S.; Toste, F. D. Strategies for Remote Enantiocontrol in Chiral Gold(<scp>iii</Scp>) Complexes Applied to Catalytic Enantioselective γ,δ-Diels–Alder Reactions. *Chem. Sci.* **2020**, *11* (25), 6450–6456. https://doi.org/10.1039/D0SC00497A.

(46) Mukai, M.; Nagao, K.; Yamaguchi, S.; Ohmiya, H. Molecular Field Analysis Using Computational-Screening Data in Asymmetric *N* -Heterocyclic Carbene-Copper Catalysis toward Data-Driven *In Silico* Catalyst Optimization. *Bull. Chem. Soc. Jpn.* **2022**, *95* (2), 271–277. https://doi.org/10.1246/bcsj.20210349.

(47) Hoque, A.; Jain, N.; Chenna, D.; Sunoj, R. B. Machine Learning Models for Catalytic Asymmetric Reactions of Simple Alkenes: From Enantioselectivity Predictions to Chemical Insights. *Digit. Discov.* **2026**, *5* (4), 1718–1735. https://doi.org/10.1039/D5DD00483G.

(48) Varnek, A.; Zankov, D.; Polishchuk, P.; Madzhidov, T. Multi-Instance Learning Approach to Predictive Modeling of Catalysts Enantioselectivity. *Synlett* **2021**, *32* (18), 1833–1836. https://doi.org/10.1055/a-1553-0427.

(49) 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.

(50) Tang, M.; Zhang, T.; Huang, Q.; Li, S.; Liu, R.; Li, H.; Chen, X.; Dong, S.; Liu, X.; Feng, X.; Hong, X. Data-Driven Modeling of *N,N′* -Dioxide/Metal-Catalyzed Asymmetric Michael Additions. *Angew. Chem. Int. Ed.* **2025**, *65* (12), e18560. https://doi.org/10.1002/anie.202518560.

(51) Aguilar-Bejarano, E.; Özcan, E.; Rit, R. K.; Li, H.; Lam, H. W.; Moore, J. C.; Woodward, S.; Figueredo, G. Homogeneous Catalyst Graph Neural Network: A Human-Interpretable Graph Neural Network Tool for Ligand Optimization in Asymmetric Catalysis. *iScience* **2025**, *28* (3), 111881. https://doi.org/10.1016/j.isci.2025.111881.

(52) Wang, S.; Zhao, W.; Liu, Y.; Li, Y. Multi-Modal Homogeneous Chemical Reaction Performance Prediction with Graph and Chemical Language Information. *Chin. J. Chem.* **2025**, *43* (11), 1230–1238. https://doi.org/10.1002/cjoc.202401186.

(53) Cheng, L.; Shao, P.-L.; Lv, J.; Xiao, H.; Sun, Y.; Yang, J.; Xu, Z.; Lv, M.; Wang, G.; Zhao, S.; Li, J.; Jin, Z.; Tan, X.; Xing, G.; Zhang, B. Chemistry-Informed Deep Learning Model for Predicting Stereoselectivity and Absolute Configuration in Asymmetric Hydrogenation. *Nat. Comput. Sci.* **2025**, *6* (2), 145–155. https://doi.org/10.1038/s43588-025-00920-8.

(54) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *J. Org. Chem.* **2023**, *88* (9), 5239–5241. https://doi.org/10.1021/acs.joc.3c00844.

(55) Mishra, C.; Wolff, N. von; Tripathi, A.; Brodie, C. N.; Lawrence, N. D.; Ravuri, A.; Brémond, É.; Preiss, A.; Kumar, A. Predicting Ruthenium Catalysed Hydrogenation of Esters Using Machine Learning. *Digit. Discov.* **2023**, *2* (3), 819–827. https://doi.org/10.1039/D3DD00029J.

(56) Hadler, N.; Rinehart, N. I.; Elkin, M.; Nicolai, J.; Gheibi, G.; Chen, J.; Avaylon, M.; Maciejewski, R.; Weber, G. H.; Mahoney, M. W.; Perciano, T.; Hartwig, J. F. A 3D, Structure-Based, Deep Learning Approach for Predicting the Regioselectivity of Transition-Metal Catalysis. *ChemRxiv 2026* (0128). https://doi.org/10.26434/chemrxiv.10001648/v1.

(57) Shved, A. S.; Ocampo, B. E.; Burlova, E. S.; Olen, C. L.; Rinehart, N. I.; Denmark, S. E. Molli: A General Purpose Python Toolkit for Combinatorial Small Molecule Library Generation, Manipulation, and Feature

Extraction. *J. Chem. Inf. Model.* **2024**, *64* (21), 8083–8090. https://doi.org/10.1021/acs.jcim.4c00424.

(58) 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.

(59) Wood, B. M.; Dzamba, M.; Fu, X.; Gao, M.; Shuaibi, M.; Barroso-Luque, L.; Abdelmaqsoud, K.; Gharakhanyan, V.; Kitchin, J. R.; Levine, D. S.; Michel, K.; Sriram, A.; Cohen, T.; Das, A.; Rizvi, A.; Sahoo, S. J.; Ulissi, Z. W.; Zitnick, C. L. UMA: A Family of Universal Models for Atoms. arXiv June 30, 2025. https://doi.org/10.48550/arXiv.2506.23971.

(60) Levine, D. S.; Shuaibi, M.; Spotte-Smith, E. W. C.; Taylor, M. G.; Hasyim, M. R.; Michel, K.; Batatia, I.; Csányi, G.; Dzamba, M.; Eastman, P.; Frey, N. C.; Fu, X.; Gharakhanyan, V.; Krishnapriyan, A. S.; Rackers, J. A.; Raja, S.; Rizvi, A.; Rosen, A. S.; Ulissi, Z.; Vargas, S.; Zitnick, C. L.; Blau, S. M.; Wood, B. M. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. arXiv May 13, 2025. https://doi.org/10.48550/arXiv.2505.08762.

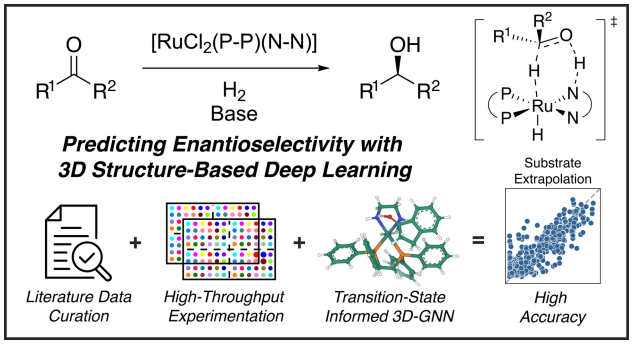
(61) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160* (11), 114110. https://doi.org/10.1063/5.0197592.

(62) 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* (14), 7169–7192. https://doi.org/10.1039/C9CP06869D.

(63) 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.* **2023**, *64* (1), 9–17. https://doi.org/10.1021/acs.jcim.3c01250.

(64) Burns, J. W.; Zalte, A. S.; Abreu, C. R. A.; Sieg, J.; Feldmann, C.; Mathea, M.; Green, W. H. Deep Learning Foundation Models from Classical Molecular Descriptors. arXiv February 9, 2026. https://doi.org/10.48550/arXiv.2506.15792.

(65) Ash, J. R.; Wognum, C.; Rodríguez-Pérez, R.; Aldeghi, M.; Cheng, A. C.; Clevert, D.-A.; Engkvist, O.; Fang, C.; Price, D. J.; Hughes-Oliver, J. M.; Walters, W. P. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. *J. Chem. Inf. Model.* **2025**, *65* (18), 9398–9411. https://doi.org/10.1021/acs.jcim.5c01609.


[RuCl2(P-P)(N-N)]
H2
Base
Predicting Enantioselectivity with
3D Structure-Based Deep Learning
Substrate
Extrapolation
Literature Data
Curation
High-Throughput
Experimentation
Transition-State
Informed 3D-GNN
High
Accuracy