How to cite: *Angew. Chem. Int. Ed.* **2021**, *60*, 22804–22811
International Edition: doi.org/10.1002/anie.202106880
German Edition: doi.org/10.1002/ange.202106880

Data-Driven Design

Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning

*Li-Cheng Xu, Shuo-Qing Zhang, Xin Li, Miao-Jiong Tang, Pei-Pei Xie, and Xin Hong**

Abstract: *Asymmetric hydrogenation of olefins is one of the most powerful asymmetric transformations in molecular synthesis. Although several privileged catalyst scaffolds are available, the catalyst development for asymmetric hydrogenation is still a time- and resource-consuming process due to the lack of predictive catalyst design strategy. Targeting the data-driven design of asymmetric catalysis, we herein report the development of a standardized database that contains the detailed information of over 12000 literature asymmetric hydrogenations of olefins. This database provides a valuable platform for the machine learning applications in asymmetric catalysis. Based on this database, we developed a hierarchical learning approach to achieve predictive machine leaning model using only dozens of enantioselectivity data with the target olefin, which offers a useful solution for the few-shot learning problem and will facilitate the reaction optimization with new olefin substrate in catalysis screening.*

Introduction

As one of the most powerful methods in asymmetric synthesis, asymmetric hydrogenation of olefins (AHOs) has brought phenomenal impact in both academic research and industrial application.[1] Since the Nobel Prize in 2001, the decades of catalyst development have significantly expanded the efficiency, selectivity and scope of this transformation, leading to several widely applied privileged catalysts.[2] This provides the strong basis that supports the continuing interest of catalysis development for challenging olefin targets.[3] Despite the high demand for AHO development, the lack of a predictive catalyst design strategy has hindered the progress.[4] Although computational modeling,[5] 3D-QSAR,[6] multivariant linear regression,[7] and high-throughput experimentation[8] provided effective support to accelerate the discovery of asymmetric catalysis (Figure 1 a), the catalysis screening still relies heavily on empirical experience and serendipity. This issue slows down the catalyst development for the growing and diversified desires of AHO, making it a time- and resource-consuming process.

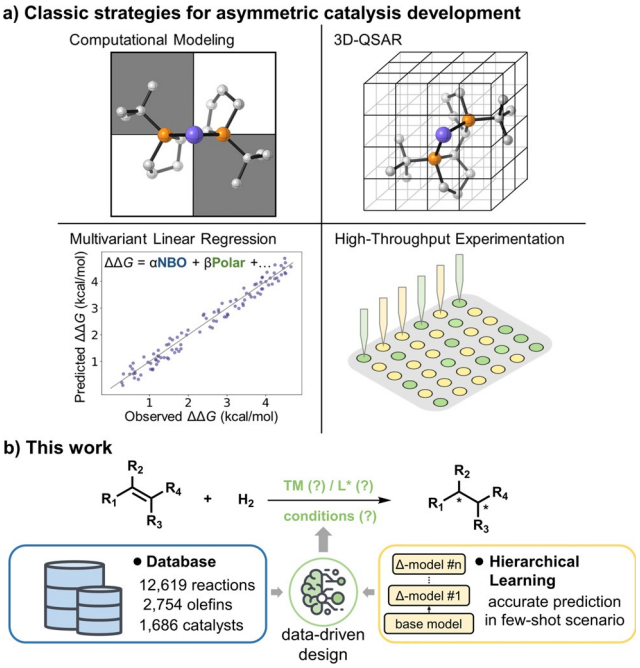

Figure 1. Classic and data-driven strategies for catalyst development in asymmetric catalysis.

Recent few years have witnessed the stimulating development of data-driven technologies for the prediction and design of organic synthesis.[9] This renaissance of chemoinformatics provides an exciting opportunity for the discovery of asymmetric catalysis.[6] In this area, Sigman's multi-variant linear regression approach[10] has reached remarkable success in the predictive modelling of asymmetric transformations. Using machine learning (ML) strategies, independent studies from Denmark,[11] Sunoj,[12] Corminboeuf[13] and Grzybowski[14] have shown that AI model is able to capture the statistical pattern of chiral induction and achieve accurate enantioselectivity prediction in stereoselective synthesis. In addition, a series of innovative ML technologies have been applied in the prediction of reactivity,[15] chemo-[16] and regioselectivity[17] of organic transformations.

Although it is established that accurate ML prediction model can be developed given the required statistics of the target chemical space,[9] it is challenging to provide such data support in the catalysis discovery process. In practical scenario of asymmetric catalysis development, it would be helpful if a predictive model can be provided in the early stage of screening process where limited data is available. This will allow efficient identification of promising scaffolds for subsequent structural engineering. In light of this challenging

[*] L.-C. Xu, Dr. S.-Q. Zhang, X. Li, M.-J. Tang, P.-P. Xie, Prof. Dr. X. Hong
Center of Chemistry for Frontier Technologies, Department of Chemistry, State Key Laboratory of Clean Energy Utilization, Zhejiang University
38 Zheda Road, Hangzhou, 310027 (China)
E-mail: hxchem@zju.edu.cn
Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202106880.

task, we envisioned that the catalysis discovery can be accelerated through AI interaction between the ongoing investigation with limited data and the existing knowledge with large quantity of related data, as a way to mimic the human chemist's approach. Using AHO as a proof of concept, herein we report the development of a database that contains over 12000 literature AHO transformations and data-driven designs of asymmetric catalysis (Figure 1b). A hierarchical learning approach was developed to harness the value of the created database, which can achieve accurate ML prediction using only dozens of data with a target olefin. This work provides the data support and a useful ML approach for the data-driven design of AHOs, which can be directly implemented in the experimental reaction optimization with new olefin substrates.

Results and Discussion

The data collection and processing of AHOs are elaborated in Figure 2. The core data are manually extracted from related synthetic method publications in a wide array of chemistry journals between 2000 and 2020.[18] The selection of the 355 papers is to capture the greatest quantity of high-quality data in an efficient fashion, which is by no means comprehensive. For each reported transformation, the details of reactant, catalyst, reaction conditions and reaction performances are extracted along with the digital object identifier (DOI) of the source publication, which creates an entity of record.

We next performed data cleaning and standardization to ensure that the database is reliable and readily available for ML purposes. The chemical structures of the involved molecules are all drawn in full and transferred to SMILES[19] format for consistency in representation. Subsequent sanitization of the recorded SMILESs ensured that all the compound records can be recognized by RDKit,[20] and manual corrections were performed for the unrecognizable SMILESs. To resolve the issue that one molecule may have multiple SMILES representations in the records, SMILES standardization is performed for all compounds, and only the canonical SMILES[19b] is used for each molecule. In order to support the creation of the catalyst 3D structure and related AI applications, a series of customized scripts were developed to produce the SMILES representation of the ligand-coordinated transition metal catalyst from the separate records of ligand and transition metal (Figure S2). For chiral ligands that involve axial or planar chirality (i.e. biaryl and ferrocene-containing ligands), the SMILES representation cannot record the related stereochemistry information, and the complete stereochemistry information of these ligands can be retrieved from the generated 3D geometry in our database.

The finalized database contains four key entity categories, compounds, reaction conditions, reaction performances and source publication. Compounds category includes the SMILES representations and xTB[21]-optimized 3D geometries of reactant, catalyst and product with absolute stereochemistry information. Reaction conditions category includes the information of temperature, pressure, reaction time, catalyst loading, solvent and additive, if available in the original publication. Reaction performances category includes the information of enantioselectivity, reaction yield, substrate conversion, and turn over number (TON) if available in the original publication. The full data collection currently includes 12619 individual reactions, 1686 transition metal catalysts and 2754 olefin substrates, which provides a valuable resource for ML application in asymmetric catalysis and complements the available databases of organic transformation (i.e. USPTO,[22] Pistachio,[23] etc.). Considering the major data source are the publications of synthetic methods, most of the recorded olefins are the model substrates for method evaluation. To provide an overview of the structural complexity of the curated database, we also included a few widely applied indexes (molecular weight,

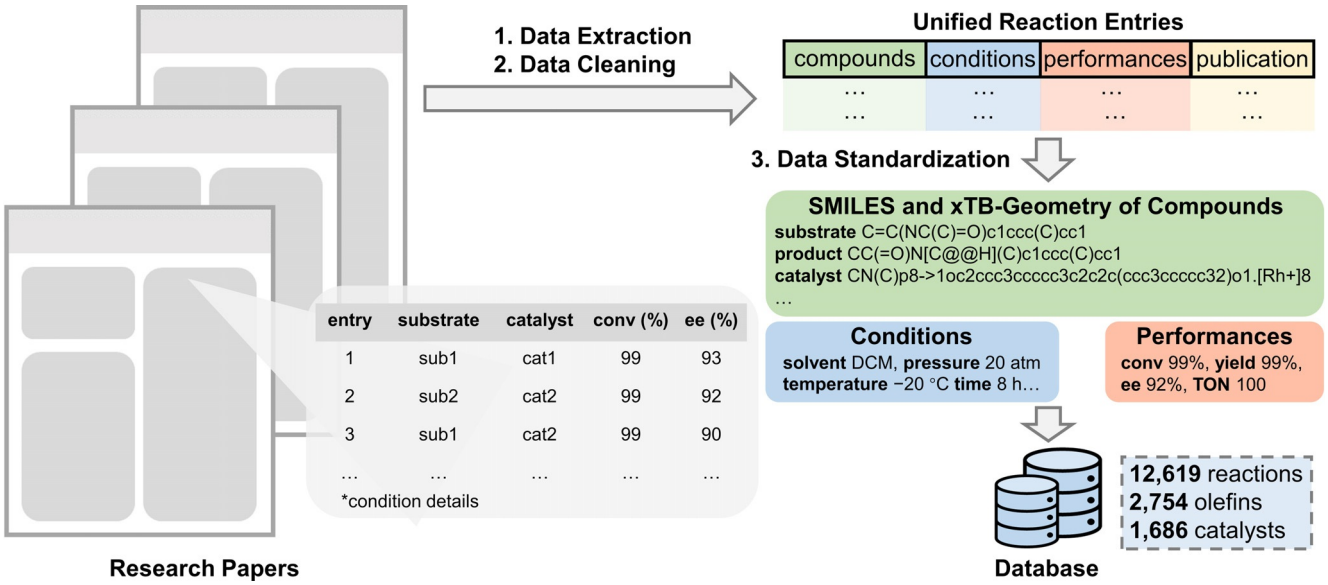

Figure 2. Workflow for the database creation of asymmetric hydrogenation of olefins.

BertzCT,[24] SCScore[25] and SAScore[26]) for potential users to identify if the curated database is suitable for the target application. We will continue to maintain and expand this database to support its ML applications. The database and our ML models are all freely available at http://asymcatml.net/, which can be directly accessed by interested users. Further details of data processing are provided in the Supporting Information (Figure S1–S3).

Our database presented a quantitative view of the AHO development for the last two decades, and several intriguing patterns were identified. Regarding the olefin type, only 7.7% of the reported substrates are tetra-substituted olefin, while 65.2% and 27.1% of the records are tri- and di-substituted olefins, respectively (Figure 3a). This corroborates the consensus that asymmetric hydrogenation of tetra-substituted olefin is still a challenging task.[27] For the distribution of transition metal (Figure 3a), rhodium[28] (53.1%) and iridium[29] (37.9%) catalysts have played a dominant role, which together accounts for over 90% of the recorded catalysts in the database. Ruthenium catalysts[2c] have also been widely used, accounting 5.3% of the recorded catalysts. Although palladium catalysts were active, limited reports were focusing on Pd-catalyzed asymmetric hydrogenation,[30] and only 0.6% of the recorded catalysts are palladium complexes. In addition to these late transition metals, there is a growing research interest in asymmetric catalysis with earth-abundant transition metals.[31] Our database identified 2.1% nickel catalysts and 1.0% cobalt catalysts. Despite Kagan's[32] pioneering studies of titanium catalysis and Buchwald's success of zirconocene-catalyzed asymmetric hydrogenation of tetra-substituted alkenes in 1999,[33] it is surprising that no records were found for zirconium and titanium catalysts in our database. We cannot find any AHO literature reports using these transition metals during 2000 to 2020.

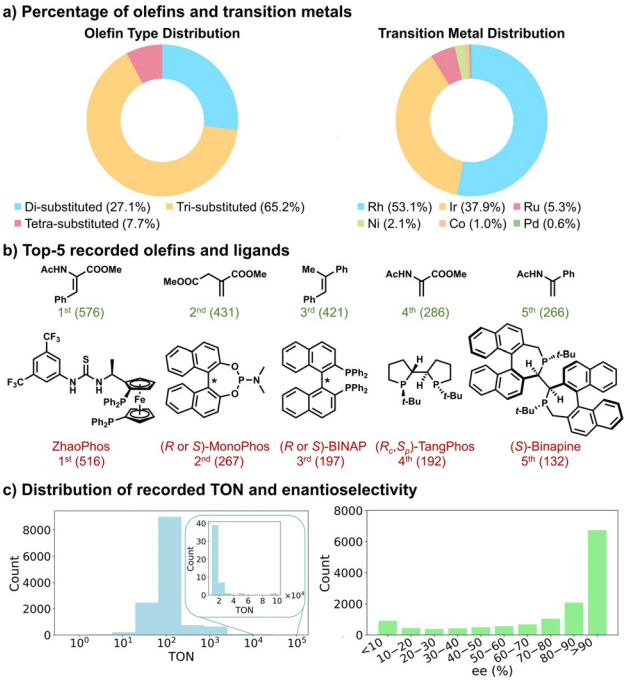

Figure 3. Key statistics of the created AHO database.

The top five most recorded olefins and ligands in the database are shown in Figure 3b. Methyl (*Z*)-2-acetamido-3-phenylacrylate (**1**), dimethyl itaconate (**2**), methyl 2-acetamidoacrylate (**4**) and *N*-(1-phenylvinyl) acetamide (**5**) are the classical benchmark substrates for the catalytic properties of AHO, which appeared 576, 431, 286 and 266 times in the database, respectively. Trans-α-methylstilbene (**3**) is the only all-carbon olefin and a model substrate for unfunctionalized olefins among the top five compounds, which has 421 recorded transformations in our database. For the ligands, ZhaoPhos **L1**, which is a privileged ligand for rhodium catalyst developed by the Zhang group,[34] has the highest number of records (516). MonoPhos **L2**, developed by Feringa,[35] is a well-known phosphoramidite ligand for rhodium and is recorded 267 times. 197 recorded transformations were associated with Noyori's BINAP ligand **L3**,[36] which is widely applied in Ru-catalyzed asymmetric hydrogenations.[37] Zhang group also developed the powerful bisphosphine ligands TangPhos **L4**[38] and (*S*)-Binapine **L5**,[39] and these ligands have 192 and 132 associated records in our database, respectively.

The statistics of the recorded reaction performances are shown in Figure 3c. Most of the reported TONs are in the scale of 10^2. Only 1.3% are over 1000, and 0.2% are over 10000. The highest TON in our database is 10^5, which is from Ding's report of Rh-catalyzed asymmetric hydrogenation of dimethyl itaconate using monodentate phosphoramidite DphenPhos.[40] Due to the literature nature of the extracted data, the distribution of the enantioselectivity data is not even, and a significant portion of the recorded enantioselectivities is over 90% (Figure 3c). It would be difficult for ML purposes if the variation of the training enantioselectivity data is limited. Luckily, our database does contain a substantial amount of low to medium enantioselectivity data (4261 data below 80% *ee*). These data, although not appealing in the original reports, are the "hidden treasure" of the literature studies. Including these data will allow the ML model to recognize the patterns for the enantioselectivity change, in order to create ML model for the prediction of highly selective AHO catalysts.

With the database in hand, we next explored the ML prediction for catalysis discovery in AHOs. Our ML task was focused on the early screening scenario that only dozens of enantioselectivity data is available for the target olefin substrate, which is a typical few-shot learning problem.[41] We envisioned that the key to solve this problem is to harness the value of the large quantity of reported data and build a ML model that can benefit from the diversified data sources with varying priorities.

A hierarchical learning strategy was surmised to connect a particular target olefin with the AHO database (Figure 4a). By matching the substitutions of the target olefin and the recorded olefin substrates, related datasets can be extracted from the database. The data with olefins that matches one substitution with the target substrate led to the dataset a. This dataset contains the largest quantity of data that is relevant to the target olefin, which supports the training of the base model. This base model provides a base prediction of enantioselectivity, which is further improved by subsequent

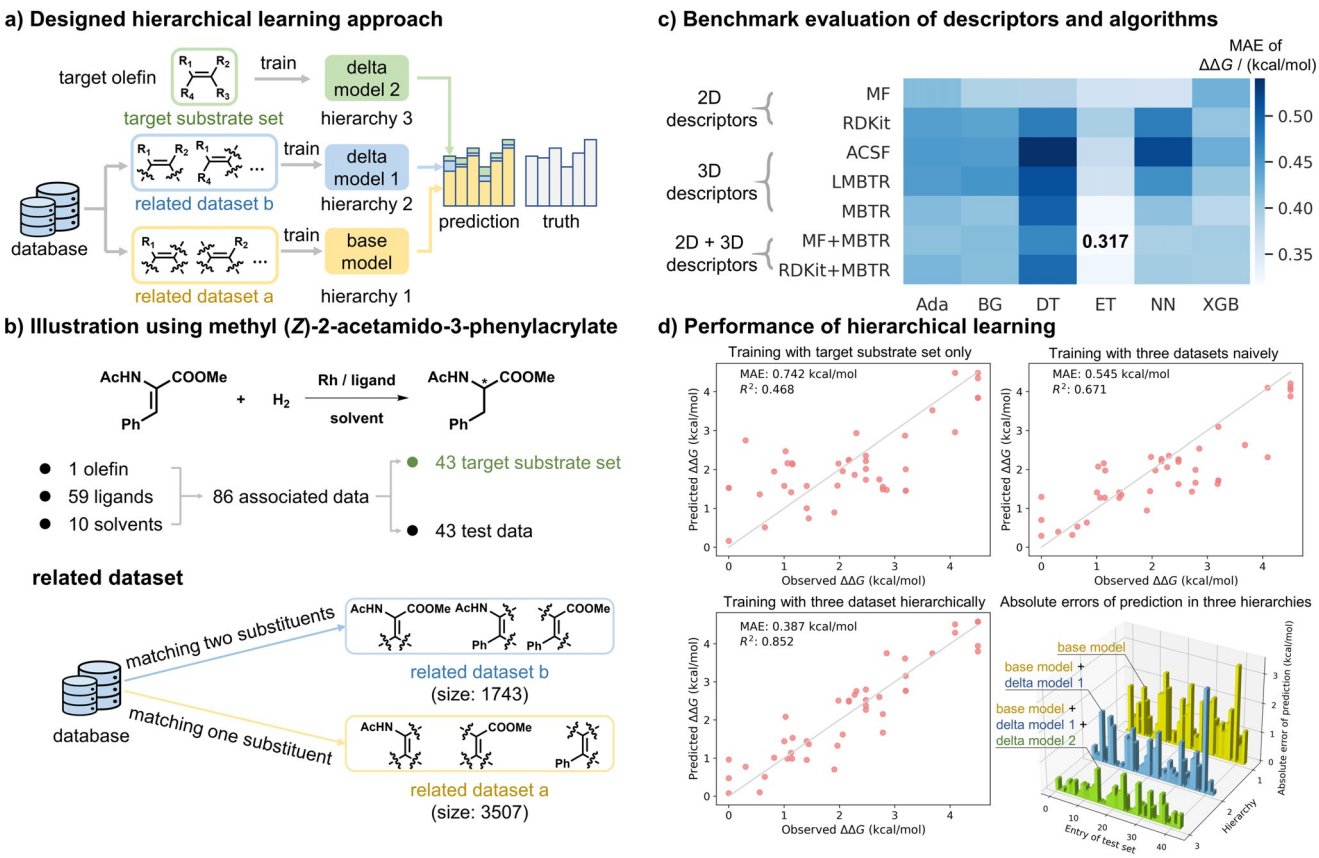

Figure 4. Design and performance of the proposed hierarchical learning strategy. Benchmark evaluation includes molecular fingerprint (MF), molecular descriptors in RDKit (RDKit), atom-centered symmetry functions (ACSF), local many-body tensor representation (LMBTR) and many-body tensor representation (MBTR) as molecular descriptor, and AdaBoost (Ada), Bagging (BG), Decision Tree (DT), Extra-Trees (ET), Neural Network (NN) and XGBoost (XGB) as ML algorithms.

delta learnings.[42] The first delta learning relies on dataset b, which involves the olefins that have two identical substitutions with the target substrate. The delta learning with dataset b is to create a delta model 1 that learns the difference between the reported enantioselectivities and the predicted values from the base model for the olefins in dataset b. Similarly, the second delta learning can be applied using the limited data with the target olefin. Therefore, the final enantioselectivity prediction is the sum of the predicted values from the base model and the two delta models.

Through the hierarchical learning strategy, the base model is improved sequentially to reach the target chemical space with limited data. This offers a practical solution to the model training during the catalysis screening, especially in the early stage of hit identification. We want to emphasize that the substituent matching is just a native way to evaluate the structural similarity; a wide array of molecular features (i.e. molecular fingerprints,[43] molecular descriptors in RDKit,[20] physical organic descriptors,[16,17c,e] etc.) can be utilized to identify the related data for the designed hierarchical learning strategy based on the application purposes (vide infra). A detailed workflow of the hierarchical learning strategy is provided in the Supporting Information (Figure S8).

The top-1 recorded olefin in our AHO database, methyl (*Z*)-2-acetamido-3-phenylacrylate, was used as an example to demonstrate the application of the designed hierarchical learning strategy (Figure 4b). By randomly splitting the 86 enantioselectivity data of this olefin under rhodium catalysis, 43 data were used for model training (target substrate set), and the other 43 cases are included in the test set for performance evaluation. By matching with the substitutions of methyl (*Z*)-2-acetamido-3-phenylacrylate, the related datasets a and b were located to allow the hierarchical machine learnings. To avoid data leakage, all the datasets are mutually exclusive. Dataset b contains 1743 data with olefins which have two matching substituents, without any data of the target olefin. Dataset a contains 3507 data with olefins that match one of the three substitutions of methyl (*Z*)-2-acetamido-3-phenylacrylate, without dataset b and the data of target olefin.

To identify the optimal combination of molecular descriptor and ML algorithm for model training, we next performed a benchmark evaluation using the related dataset a. Through random splitting, 90% of the related dataset a was used as training set, and the test set includes the other 10%. A wide array of molecular descriptors based on 2D topological structure (i.e. RDKit[20] descriptors, MF,[43] etc.) or 3D coordinate (i.e. ACSF,[44] MBTR,[45] etc.) were evaluated. The test set performances of the representative combinations are shown in Figure 4c. Comparing the molecular descriptors,

the predictive abilities of 3D descriptors are generally better than those of 2D descriptors. This is understandable considering the nature of chiral induction. Combining both 2D and 3D descriptors can further improve the regression performance limitedly. For the ML algorithms, the tree-based models generally have superior performances, in which the Extra-Trees model[46] gave the best results. The Extra-Trees model with MBTR and MF descriptors is the best combination with a mean absolute error (MAE) of 0.317 kcal mol^{-1}, which was used in the subsequent hierarchical trainings. Full details of additional tested descriptors, ML algorithms and 3D structure generation are included in the Supporting Information (Figure S9).

The effectiveness of the designed hierarchical learning strategy is elaborated in Figure 4d. Using only the 43 data with the target olefin substrate (target substrate set), the model performance is unreliable with a R^2 of 0.468 due to the lack of data. Simply combining the small target substrate set with the large amount of data from the related datasets a and b, the naive way of training only improved the regression performance limitedly even with over 5000 additional data (MAE = 0.545 kcal mol^{-1}; R^2 = 0.671). This is because the limited data with valuable labels of the target olefin is "drowned" in the thousands of data from the related datasets, which emphasizes the necessity to create a smart strategy to treat these data with varying priorities. To our delight, the hierarchical learning strategy is able to provide a significantly better model by effectively using both the large quantity of related data (datasets a and b) and the small number of valuable data (target substrate set); the hierarchical learning model has a R^2 of 0.852 and MAE of 0.387 kcal mol^{-1}, whose predictive ability can provide useful support in the early stage of catalysis discovery given only dozens of data of the target olefin. The hierarchical improvement is clearly demonstrated by comparing the absolute error of predictions in the test set (Figure 4d). Through the increasing hierarchies, the model performance is improved steadily with the sequential delta learnings. This highlights the importance of logically connecting the large amount of available related data and the small number of data from ongoing investigation for ML-assisted reaction optimization. The developed hierarchical learning model can be easily transferred to any target olefin substrate for ongoing AHO investigations.

We next performed a series of additional explorations to verify the effectiveness and rationalize the chemical basis of the designed hierarchical learning approach. Using the same molecular descriptors (MBTR and MF) and ML algorithm (ET), the averaged regression performances of ten trials for the tested hierarchical learning models in methyl (*Z*)-2-acetamido-3-phenylacrylate are shown in Figure 5. The original substituent-based hierarchical learning model with three hierarchical orders achieved an averaged MAE of 0.398 kcal mol^{-1} and R^2 of 0.842. Using the same data, random data shuffling among the three orders led to significantly worse performance (MAE = 0.649 kcal mol^{-1}; R^2 = 0.585). Consistently, random data selection from the entire database resulted in even worse performance (MAE = 0.684 kcal mol^{-1}; R^2 = 0.519). Particularly, including one additional hierarchical order using more training data also led to worse

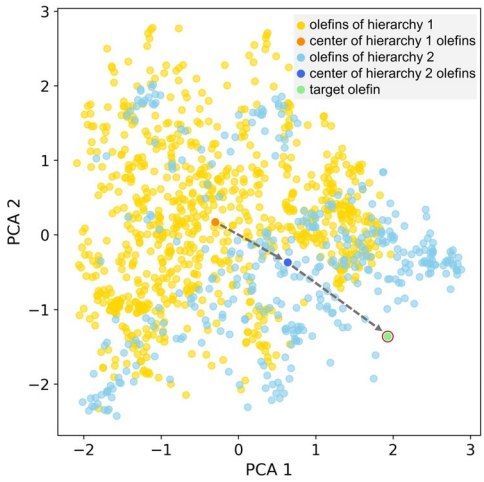

Figure 5. Regression performances of hierarchical models with varying data treatments.

predictive ability. This model maintains the datasets of the original three hierarchical orders, but utilizes all the rest 6793 data of the AHO database for training in the new hierarchy one (four total hierarchies). Despite the extra hierarchical order and additional training data, this model only achieved a MAE of 0.597 kcal mol^{-1} and R^2 of 0.669. This result suggested that simply including more hierarchical order and additional data without the chemical relevance does not improve the model performance. These comparisons collectively support that the performance improvement of substituent-based hierarchical learning approach is underlain by chemical heuristics, not simply due to the hierarchical data splitting. Detailed regression results are provided in the Supporting Information (Table S6 and Figure S10).

We believe that the chemical basis of the hierarchical learning approach is that the target transformation follows a high-dimensional structure–activity relationship. To recreate this relationship with desired scope and accuracy is challenging because these two requirements are contradictory from a data-driven point of view. The amount of available data cannot match the complexity of the enormous chemical space. However, in the scenario of catalysis discovery, chemists are usually focused on a specific and relatively smaller chemical space (i.e. one novel olefin substrate). The hierarchical learning approach applies the large amount of related data to learn a base model that comprehends the general catalysis behavior, and uses the valuable data from the target space sampling to learn the perturbance of the general relationship in the specific target chemical space. This strategy allows the ensemble model to approach the target space based on chemical heuristics.

The above rationalization of the hierarchical learning strategy is supported by additional analysis of the olefin chemical space and the ML models. Based on the selected olefins of the substituent-based hierachical model (Figure 4b), we performed a principal component analysis (PCA) of the Morgan fingerprints of the selected olefins. The selected olefins are projected in the two PCA dimensions (Figure 6). Yellow dots are the 803 olefins in hierarchy one;

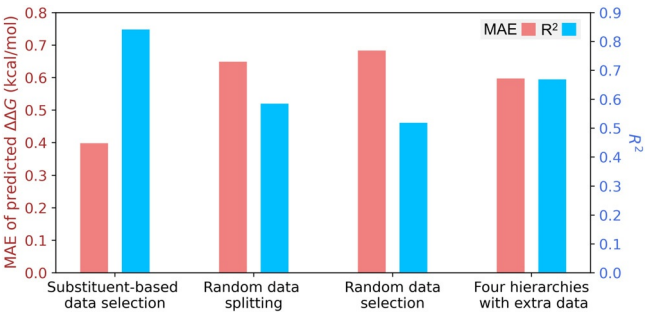

Figure 6. Projection of the selected olefins in the substituent-based hierarchical learning model to the PCA dimensions of Morgan fingerprints.

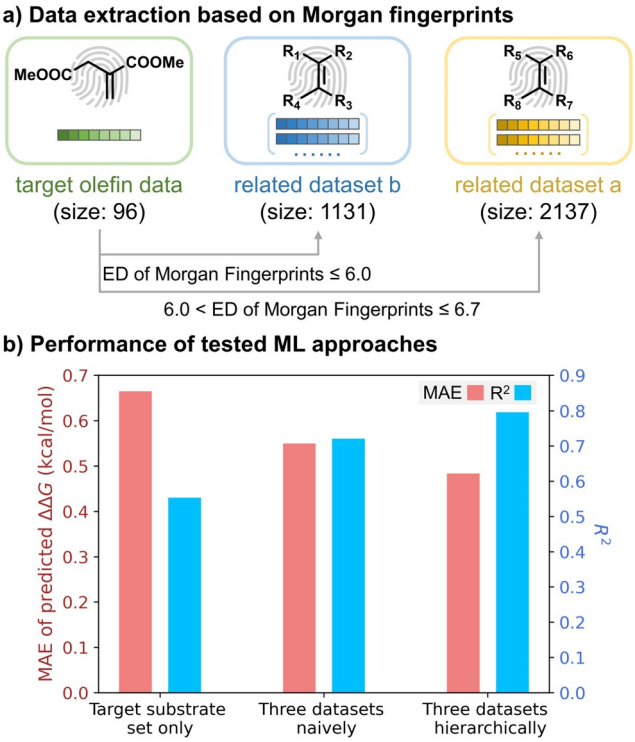

Figure 7. Data extraction using Euclidean distance (ED) of Morgan fingerprints and performance comparisons of tested ML approaches for dimethyl itaconate.

blue dots refer to the 376 olefins in hierarchy two; green dot represent the target olefin methyl (*Z*)-2-acetamido-3-phenylacrylate. The results in Figure 6 support the proposed chemical space approaching; with increasing hierarchical orders, the olefins are approaching the target olefin from a structural point of view (Figure 6). The analysis of the errors of prediction of the hierachical models also led to consistent results. The substituent-based hierarchical model achieved the sequential error reductions due to the space approaching (Figure 4d). After randomly shuffling the same data among the three hierarchical orders, the space approaching no longer exists, and the error reduction disappears (Figure S11).

Substituent matching is a native way to evaluate the proximity of chemical space, and other molecular features can also be utilized to identify the related data. To prove that additional classification methods are feasible and the hierarchical learning strategy is applicable for olefins which may not be suitable for substituent matching, we further tested the Morgan fingerprints-based data extraction using the dimethyl itaconate (Figure 7). Dimethyl itaconate is a di-substituted olefin, which cannot allow three hierarchical orders using substituent matching. By randomly splitting the 96 data with dimethyl itaconate, 48 data are used for target substrate set in hierarchy three, and the other 48 cases are included in the test set. The Euclidean distances between the Morgan fingerprints of dimethyl itaconate and those of the rest olefins in the database are measured, which allow the identification of the related datasets a and b for hierarchical learning (Figure 7a). The olefins with Euclidean distance smaller than 6.0 are included in the related dataset b (1131 data) for hierarchy two, and those with Euclidean distance between 6.0 and 6.7 are included in the related dataset a (2137 data) for hierarchy one. These datasets are mutually exclusive, as in the case of methyl (*Z*)-2-acetamido-3-phenylacrylate. The Euclidean distance of Morgan fingerprints provides a quantitative measurement of the structural similarity. Its cut off value is a customized number depending on the target olefin and the available data in the curated database. We are working on a fully automated process to achieve the data selection based on automated ML technology.[47]

The performances of the tested machine learning approaches in dimethyl itaconate are compared in Figure 7b. Using only the 48 data in the target substrate set, the model performance is poor with a R^2 less than 0.6. Training with the three related datasets naively led to improved model performance with an averaged MAE of 0.550 kcal mol^{-1} and R^2 of 0.721. Through hierarchical learning approach, the model achieved the best results with MAE of 0.484 kcal mol^{-1} and R^2 of 0.796. Detailed regression performances are provided in the Supporting Information (Table S7 and Figure S12). These results are consistent with the substituent-based data extraction in methyl (*Z*)-2-acetamido-3-phenylacrylate (Figure 4d), which demonstrated that the hierarchical learning approach can apply to additional structural evaluation methods. Using Morgan fingerprints-based data extraction, we also tested the hierarchical learning approach with out-of-sample examples from a new AHO reference in 2021.[48] Consistently, the hierarchical learning model led to improved predictive performance as compared to naive training and training with target substrate set (Table S8 and Figure S13–S15).

Conclusion

In summary, we developed the database and machine learning strategy for data-driven design of asymmetric hydrogenation of olefins. The database contains the standardized information of 12619 literature transformations from 2000 to 2020, involving 2754 olefins and 1686 chiral transition metal catalysts. This provides a valuable and readily available data resource for the machine learning applications in organic synthesis, especially for asymmetric catalysis. Benefiting from this database, a hierarchical learning approach was designed to build predictive machine learning model using only dozens

of enantioselectivity data with the target olefin substrate. This approach effectively connects the large amount of related data from existing knowledge and the small number of data from ongoing investigation based on chemical heuristics, offering a practical solution to the model training in the early stage of reaction optimization with new olefin substrate. Data collection for additional types of asymmetric transformations and further applications of the hierarchical learning strategy in AHO development with challenging target compounds are currently under investigation in our laboratory.

Acknowledgements

Financial support from National Natural Science Foundation of China (21702182 and 21873081), the Fundamental Research Funds for the Central Universities (2020XZZX002-02), the State Key Laboratory of Clean Energy Utilization (ZJUCEU2020007), and the Center of Chemistry for Frontier Technologies is gratefully acknowledged. Calculations were performed on the high-performance computing system at Department of Chemistry, Zhejiang University. All the Hong research group members are acknowledged for their contributions in the data collection process.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: asymmetric hydrogenation · database · data-driven design · enantioselectivity prediction · hierarchical learning

[1] a) W. S. Knowles, *Angew. Chem. Int. Ed.* **2002**, *41*, 1998 – 2007; *Angew. Chem.* **2002**, *114*, 2096 – 2107; b) R. Noyori, *Angew. Chem. Int. Ed.* **2002**, *41*, 2008 – 2022; *Angew. Chem.* **2002**, *114*, 2108 – 2123.

[2] For reviews, see: a) L. Eberhardt, D. Armspach, J. Harrowfield, D. Matt, *Chem. Soc. Rev.* **2008**, *37*, 839 – 864; b) J. J. Verendel, O. Pamies, M. Dieguez, P. G. Andersson, *Chem. Rev.* **2014**, *114*, 2130 – 2169; c) Z. Zhang, N. A. Butt, W. Zhang, *Chem. Rev.* **2016**, *116*, 14769 – 14827; d) D. Janssen-Muller, C. Schlepphorst, F. Glorius, *Chem. Soc. Rev.* **2017**, *46*, 4845 – 4854.

[3] For reviews, see: a) L. Massaro, J. Zheng, C. Margarita, P. G. Andersson, *Chem. Soc. Rev.* **2020**, *49*, 2504 – 2522; b) J. Wen, F. Wang, X. Zhang, *Chem. Soc. Rev.* **2021**, *50*, 3211 – 3237.

[4] A. Pfaltz, W. J. Drury, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5723.

[5] For reviews, see: a) P. H. Cheong, C. Y. Legault, J. M. Um, N. Celebi-Olcum, K. N. Houk, *Chem. Rev.* **2011**, *111*, 5042 – 5137; b) T. Sperger, I. A. Sanhueza, I. Kalvet, F. Schoenebeck, *Chem. Rev.* **2015**, *115*, 9532 – 9586; c) S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M. H. Baik, *Chem. Rev.* **2019**, *119*, 6509 – 6560.

[6] For reviews, see: A. F. Zahrt, S. V. Athavale, S. E. Denmark, *Chem. Rev.* **2020**, *120*, 1620 – 1689.

[7] For reviews, see: a) M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292 – 1301; b) J. P. Reid, M. S. Sigman, *Nat. Rev. Chem.* **2018**, *2*, 290 – 305; c) C. B. Santiago, J. Y. Guo, M. S. Sigman, *Chem. Sci.* **2018**, *9*, 2398 – 2412.

[8] a) C. Jäkel, R. Paciello, *Chem. Rev.* **2006**, *106*, 2912 – 2942; b) Y. Chen, W. L. Tang, J. Mou, Z. Li, *Angew. Chem. Int. Ed.* **2010**, *49*, 5278 – 5283; *Angew. Chem.* **2010**, *122*, 5406 – 5411; c) A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak, S. D. Dreher, *Science* **2015**, *347*, 49 – 53; d) K. Troshin, J. F. Hartwig, *Science* **2017**, *357*, 175 – 181; e) N. J. Gesmundo, B. Sauvagnat, P. J. Curran, M. P. Richards, C. L. Andrews, P. J. Dandliker, T. Cernak, *Nature* **2018**, *557*, 228 – 232.

[9] a) C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281 – 1289; b) P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Rev. Chem.* **2019**, *3*, 119 – 128; c) C. W. Coley, N. S. Eyke, K. F. Jensen, *Angew. Chem. Int. Ed.* **2020**, *59*, 22858 – 22893; *Angew. Chem.* **2020**, *132*, 23054 – 23091; d) Y. Shen, J. E. Borowski, M. A. Hardy, R. Sarpong, A. G. Doyle, T. Cernak, *Nat. Rev. Methods Primers* **2021**, *1*, 23; e) A. M. Zuranski, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Acc. Chem. Res.* **2021**, *54*, 1856 – 1865; f) K. Jorner, A. Tomberg, C. Bauer, C. Sköld, P.-O. Norrby, *Nat. Rev. Chem.* **2021**, *5*, 240 – 255.

[10] a) K. C. Harper, E. N. Bess, M. S. Sigman, *Nat. Chem.* **2012**, *4*, 366 – 374; b) J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343 – 348; c) J. Werth, M. S. Sigman, *ACS Catal.* **2021**, *11*, 3916 – 3922; and ref. [7a].

[11] a) A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631; b) A. F. Zahrt, S. E. Denmark, *Tetrahedron Lett.* **2019**, *75*, 1841 – 1851; c) J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang, S. E. Denmark, *J. Am. Chem. Soc.* **2020**, *142*, 11578 – 11592.

[12] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, R. B. Sunoj, *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1339 – 1345.

[13] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2021**, *12*, 6879 – 6889.

[14] M. Moskal, W. Beker, S. Szymkuc, B. Grzybowski, *Angew. Chem. Int. Ed.* **2021**, *60*, 15230 – 15235; *Angew. Chem.* **2021**, *133*, 15358 – 15363.

[15] a) D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186 – 190; b) M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *J. Am. Chem. Soc.* **2018**, *140*, 5004 – 5008; c) C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370 – 377; d) F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **2020**, *6*, 1379 – 1390; e) K. Jorner, T. Brinck, P.-O. Norrby, D. Buttar, *Chem. Sci.* **2021**, *12*, 1163 – 1175; f) Y. Chen, B. Tian, Z. Cheng, X. Li, M. Huang, Y. Sun, S. Liu, X. Cheng, S. Li, M. Ding, *Angew. Chem. Int. Ed.* **2021**, *60*, 4199 – 4207; *Angew. Chem.* **2021**, *133*, 4245 – 4253.

[16] S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chem. Sci.* **2020**, *11*, 9665 – 9674.

[17] a) S. Banerjee, A. Sreenithya, R. B. Sunoj, *Phys. Chem. Chem. Phys.* **2018**, *20*, 18311 – 18318; b) A. Tomberg, M. J. Johansson, P. O. Norrby, *J. Org. Chem.* **2019**, *84*, 4695 – 4703; c) W. Beker, E. P. Gajewska, T. Badowski, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2019**, *58*, 4515 – 4519; *Angew. Chem.* **2019**, *131*, 4563 – 4567; d) T. J. Struble, C. W. Coley, K. F. Jensen, *React. Chem. Eng.* **2020**, *5*, 896 – 902; e) X. Li, S. Q. Zhang, L. C. Xu, X. Hong, *Angew. Chem. Int. Ed.* **2020**, *59*, 13253 – 13259; *Angew. Chem.* **2020**, *132*, 13355 – 13361; f) G. Pesciullesi, P. Schwaller, T. Laino, J. L. Reymond, *Nat. Commun.* **2020**, *11*, 4874; g) Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green, K. F. Jensen, *Chem. Sci.* **2021**, *12*, 2198 – 2208.

[18] The detailed list of chemistry journals for the data collection is provided in the Supporting Information.

[19] a) D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31 – 36; b) D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97 – 101; c) D. Weininger, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237 – 243.

[20] RDKit: open-source chemoinformatics and machine learning. http://www.rdkit.org.
[21] a) S. Grimme, C. Bannwarth, P. Shushkov, *J. Chem. Theory Comput.* **2017**, *13*, 1989 – 2009; b) P. Pracht, E. Caldeweyher, S. Ehlert, S. Grimme, *ChemRxiv* **2019**, ChemRxiv preprint, https://doi.org/10.26434/chemrxiv.8326202.v1.
[22] D. M. Lowe, PhD thesis, University of Cambridge (UK), **2012**.
[23] J. Mayfield, D. Lowe, R. Sayle, Pistachio. 2.0 edn, NextMove Software (UK), **2018**.
[24] S. H. Bertz, *J. Am. Chem. Soc.* **1981**, *103*, 3599 – 3601.
[25] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2018**, *58*, 252 – 261.
[26] P. Ertl, A. Schuffenhauer, *J. Cheminf.* **2009**, *1*, 8.
[27] S. Kraft, K. Ryan, R. B. Kargbo, *J. Am. Chem. Soc.* **2017**, *139*, 11630 – 11641.
[28] P. Etayo, A. Vidal-Ferran, *Chem. Soc. Rev.* **2013**, *42*, 728 – 754.
[29] a) S. J. Roseblade, A. Pfaltz, *Acc. Chem. Res.* **2007**, *40*, 1402 – 1411; b) T. L. Church, P. G. Andersson, *Coord. Chem. Rev.* **2008**, *252*, 513 – 531; c) S. F. Zhu, Q. L. Zhou, *Acc. Chem. Res.* **2017**, *50*, 988 – 1001.
[30] For selected studies, see: a) D.-S. Wang, J. Tang, Y.-G. Zhou, M.-W. Chen, C.-B. Yu, Y. Duan, G.-F. Jiang, *Chem. Sci.* **2011**, *2*, 803 – 806; b) C.-B. Yu, K. Gao, D.-S. Wang, L. Shi, Y.-G. Zhou, *Chem. Commun.* **2011**, *47*, 5052 – 5054.
[31] For selected reviews, see: a) P. J. Chirik, *Acc. Chem. Res.* **2015**, *48*, 1687 – 1695; b) J. Chen, Z. Lu, *Org. Chem. Front.* **2018**, *5*, 260 – 272.
[32] E. Cesarotti, R. Ugo, H. B. Kagan, *Angew. Chem. Int. Ed. Engl.* **1979**, *18*, 779 – 780; *Angew. Chem.* **1979**, *91*, 842 – 843.
[33] M. V. Troutman, D. H. Appella, S. L. Buchwald, *J. Am. Chem. Soc.* **1999**, *121*, 4916 – 4917.
[34] a) Q. Zhao, S. Li, K. Huang, R. Wang, X. Zhang, *Org. Lett.* **2013**, *15*, 4014 – 4017; b) Q. Zhao, C. Chen, J. Wen, X. Q. Dong, X. Zhang, *Acc. Chem. Res.* **2020**, *53*, 1905 – 1921.
[35] a) R. Hulst, N. K. de Vries, B. L. Feringa, *Tetrahedron: Asymmetry Tetrahedron Asymmetry* **1994**, *5*, 699 – 708; b) A. H. M. de Vries, A. Meetsma, B. L. Feringa, *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2374 – 2376; *Angew. Chem.* **1996**, *108*, 2526 – 2528.
[36] A. Miyashita, A. Yasuda, H. Takaya, K. Toriumi, T. Ito, T. Souchi, R. Noyori, *J. Am. Chem. Soc.* **1980**, *102*, 7932 – 7934.
[37] a) S. Akutagawa, *Appl. Catal. A Gen.* **1995**, *128*, 171 – 207; b) M. Berthod, G. Mignani, G. Woodward, M. Lemaire, *Chem. Rev.* **2005**, *105*, 1801 – 1836; c) H. Shimizu, I. Nagasaki, K. Matsumura, N. Sayo, T. Saito, *Acc. Chem. Res.* **2007**, *40*, 1385 – 1393.
[38] W. Tang, X. Zhang, *Angew. Chem. Int. Ed.* **2002**, *41*, 1612 – 1614; *Angew. Chem.* **2002**, *114*, 1682 – 1684.
[39] W. Tang, W. Wang, Y. Chi, X. Zhang, *Angew. Chem. Int. Ed.* **2003**, *42*, 3509 – 3511; *Angew. Chem.* **2003**, *115*, 3633 – 3635.
[40] Y. Liu, C. A. Sandoval, Y. Yamaguchi, X. Zhang, Z. Wang, K. Kato, K. Ding, *J. Am. Chem. Soc.* **2006**, *128*, 14212 – 14213.
[41] For selected studies, see: a) H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283 – 293; b) J. Ma, S. H. Fong, Y. Luo, C. J. Bakkenist, J. P. Shen, S. Mourragui, L. F. A. Wessels, M. Hafner, R. Sharan, J. Peng, T. Ideker, *Nat. Cancer* **2021**, *2*, 233 – 244.
[42] For selected studies, see: a) R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087 – 2096; b) M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K. R. Muller, K. Burke, *Nat. Commun.* **2020**, *11*, 5223; c) E. M. Collins, K. Raghavachari, *J. Chem. Theory Comput.* **2020**, *16*, 4938 – 4950; d) P. A. Unzueta, C. S. Greenwell, G. J. O. Beran, *J. Chem. Theory Comput.* **2021**, *17*, 826 – 840.
[43] For selected studies, see: a) D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742 – 754; b) A. Cereto-Massague, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallve, G. Pujadas, *Methods* **2015**, *71*, 58 – 63; c) I. Muegge, P. Mukherjee, *Expert Opin. Drug Discovery* **2016**, *11*, 137 – 148; d) M. Yang, B. Tao, C. Chen, W. Jia, S. Sun, T. Zhang, X. Wang, *J. Chem. Inf. Model.* **2019**, *59*, 5002 – 5012; and ref. [15d].
[44] a) J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106; b) J. Behler, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930 – 17955; c) J. Behler, *Angew. Chem. Int. Ed.* **2017**, *56*, 12828 – 12840; *Angew. Chem.* **2017**, *129*, 13006 – 13020.
[45] H. Huo, M. Rupp, **2017**, arXiv preprint arXiv:1704.06439 [physics.chem-ph].
[46] P. Geurts, D. Ernst, L. Wehenkel, *Mach. Learn.* **2006**, *63*, 3 – 42.
[47] X. He, K. Zhao, X. Chu, *Smithson. Contrib. Knowl. Knowl. Based Syst.* **2021**, *212*, 106622.
[48] J. Margalef, M. Biosca, P. Cruz-Sánchez, X. Caldentey, C. Rodríguez-Escrich, O. Pàmies, M. A. Pericàs, M. Diéguez, *Adv. Synth. Catal.* **2021**, https://doi.org/10.1002/adsc.202100069.

Manuscript received: May 23, 2021
Revised manuscript received: July 14, 2021
Accepted manuscript online: August 9, 2021
Version of record online: September 12, 2021

