HOKKAIDO UNIVERSITY

Title	A Primer on 2D Descriptors in Selectivity Modeling for Asymmetric Catalysis
Author(s)	Sidorov, Pavel; Tsuji, Nobuya
Citation	Chemistry-A European journal, 30(10), e202302837 https://doi.org/10.1002/chem.202302837
Issue Date	2023-11-27
Doc URL	https://hdl.handle.net/2115/93622
Rights	This is the peer reviewed version of the following article: P. Sidorov, N. Tsuji, Chem. Eur. J. 2024, 30, e202302837, which has been published in final form at https://doi.org/10.1002/chem.202302837. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.
Type	journal article
File Information	17870700.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Chemistry
A European
Journal

Chemistry Europe
European Chemical Societies Publishing

Accepted Article

Title: A Primer On 2D Descriptors In Selectivity Modeling For Asymmetric Catalysis

Authors: Pavel Sidorov and Nobuya Tsuji

This manuscript has been accepted after peer review and appears as an Accepted Article online prior to editing, proofing, and formal publication of the final Version of Record (VoR). The VoR will be published online in Early View as soon as possible and may be different to this Accepted Article as a result of editing. Readers should obtain the VoR from the journal website shown below when it is published to ensure accuracy of information. The authors are responsible for the content of this Accepted Article.

To be cited as: *Chem. Eur. J.* **2023**, e202302837

Link to VoR: https://doi.org/10.1002/chem.202302837

WILEY VCH

10.1002/chem.202302837

WILEY VCH

CONCEPT

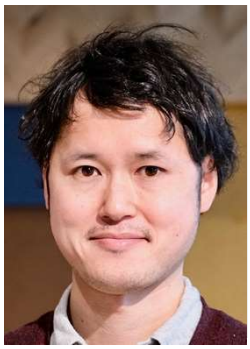
A Primer On 2D Descriptors In Selectivity Modeling For Asymmetric Catalysis

Pavel Sidorov* and Nobuya Tsuji*

[a] Dr. P. Sidorov, Dr. N. Tsuji
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD)
Hokkaido University
Sapporo 001-0021, Japan
E-mail: pavel.sidorov@icredd.hokudai.ac.jp, tsuji@icredd.hokudai.ac.jp

Abstract: Machine learning has permeated all fields of research, including chemistry, and is now an integral part of design of novel compounds with desired properties. In the field of asymmetric catalysis, the preference still lies with models based on physical understanding of the catalysis phenomenon and electronic and steric properties of catalysts. However, such models require quantum chemical calculations and are thus limited by their computational cost. Here, we highlight the recent advances in modeling catalyst selectivity using 2D structures of catalysts and substrates. While these have a less explicit mechanistic connection to the modeled property, 2D descriptors, such as topological indices, molecular fingerprints, and fragments, offer a tremendous advantage of low cost and high speed of calculations. This makes them optimal for in silico screening of large amount of data. We provide an overview of a common Quantitative Structure-Property Relationship (QSPR) workflow, model building and validation techniques, applications of these methodologies in asymmetric catalysis design, as well as the outlook on improving the understanding of 2D-based models.

Pavel Sidorov received his PhD in 2017 on the development and application of chemical space exploration methodology for drug design under supervision of Prof. Alexandre Varnek. He is currently an associate professor/Junior PI in WPI-ICReDD, Hokkaido University, and is working on applications of machine learning and artificial intelligence for modeling of chemical reactions.



Nobuya Tsuji received his Ph.D. from the University of Cologne in 2018 under the supervision of Prof. Benjamin List at the Max-Planck-Institut für Kohlenforschung. After postdoctoral studies at University of California, Berkeley in the group of Prof. Omar M. Yaghi, he moved to WPI-ICReDD at Hokkaido University where he is currently working as a specially appointed associate professor/Co-PI in the Prof. Benjamin List group. His research interests are asymmetric organocatalysis and its computational studies.

Introduction

In modern organic synthesis, asymmetric catalysis is one of the most significant development, as evidenced by the Nobel prizes awarded in 2001 and 2021[1,2]. Given that significant number of biologically active pharmaceuticals contain stereogenic center(s) in their molecules, the synthetic community has witnessed a substantial advancement in the field of asymmetric catalysis over the past decades. Nevertheless, the primary strategy for optimizing asymmetric catalysts continues to be experimental screening, guided by chemist's intuition considering the electronic, steric, and conformational effects of catalyst structures. On the other hand, the advancements in machine learning (ML) methodologies provide an opportunity to upscale the focused design of catalysts with a generality independent to the chemist involved, while reducing the associated cost of such endeavors.
Historically, the predictive modeling of selectivity followed a mechanistic view rooted in chemical intuition. Often, this approach was based on small series of catalysts and substrates, where simple steric and electronic effects could be either calculated or experimentally measured[3–5]. Some common examples include percent buried volume (%Vbur), natural bond orbitals (NBO), and energies of HOMO and LUMO[6,7]. In the case of larger, more diverse datasets, researchers turned to models based on quantum chemical[8] or 3D structural[9] features. As seen in the pioneering work provided by Oslob[10], molecular mechanics (MM) simulations are one of the avenues to estimate steric effects and interactions between solvent, catalyst, and substrates. They can be also combined with the hybrid QM/MM methods complementing them by more accurate DFT energy estimations[11]. The latter can be represented by the methods using grid-based features; as an early example, the Kozlowski and coworkers utilized a grid based on Probe Interaction Energies (PIE) to model the selectivity.[16] A more common approach would then be Comparative Molecular Field Analysis (CoMFA) which represents 3D structures as grid-based binary feature vectors.[12–15] Recently, the Denmark group described modifying this approach with non-binary features to account for conformational flexibility[9]. Nevertheless, CoMFA-based approaches require a considerable calculation cost, necessitating the 3D alignment of all structures used in training set. Some alignment-free alternatives have been reported in the literature[17–20], however, their performance is not significantly superior while still mandating a certain amount of calculations. The advances in deep learning

Accepted Manuscript

This article is protected by copyright. All rights reserved.

10.1002/chem.202302837

also led to wide-spread application of neural networks in chemistry. The chemical data for those varies wildly, from text representations like SMILES to predict reaction yield[21], to 2D-based prediction of pharmaceutical properties[22] and 3D-based neural networks emulating quantum chemical calculations[23,24]. Their application, however, requires large volumes of training data, which is typically hard to obtain for optimizing catalyst structures for one specific reaction.

When data availability allows to go beyond one series while being insufficient for deep learning methods, we believe the application of classical ML methods is the most appropriate. Evidenced by the long history of QSPR modeling of biological properties[25], 2D representation of molecular structure, or molecular graph, allows to generate general, easy to calculate, chemically relevant descriptors. Regardless of descriptors and targeted property, QSPR modeling follows a common workflow (Figure 1) generally leading to reliable models, which has been covered in literature[26].

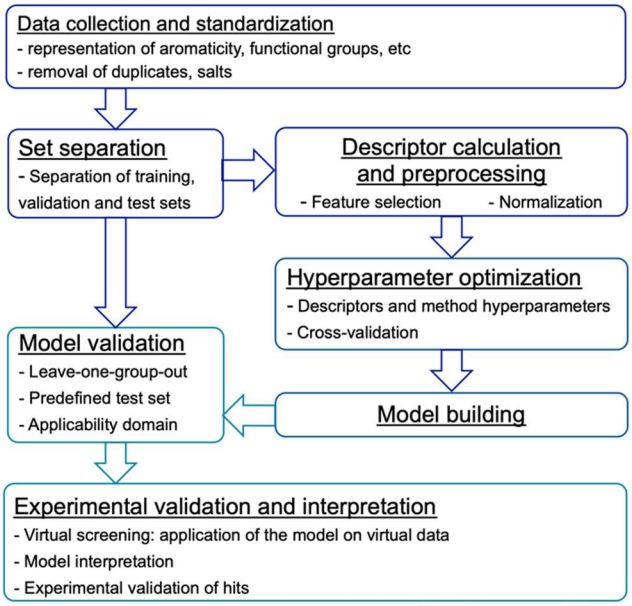


Figure 1. Generalized workflow of building and validation of a QSPR model.

The choice of descriptors is generally dictated by the property one aims to model, but it is important that they reflect the relevant structural or physico-chemical features. Moreover, some descriptors can change the information content based on the parameters used in their calculation. For example, fingerprint vectors encode a certain number of bits, and this number may improve or diminish the accuracy of predictions (see, e.g., the benchmark by Glorius and co-workers[27]). Some, such as Morgan and RDkit fingerprints[28], ISIDA[29] (In Silico Design and Data Analysis) and CircuS[30] (Circular Substructures) fragments, have parameters like size and topology (linear, circular, branched) that also affect the model's performance. Thus, the parameters of descriptors should make part of hyperparameter optimization.

The generated descriptors, however, should not be simply fed into the model. ML algorithms are sensitive to the data they learn from, and certain preprocessing steps are often necessary before training the model. One common step is feature selection, where the features deemed unimportant for the model are removed. Various algorithms exist for this purpose[31,32], including selection of features that are most correlated to the modeled property based on statistical tests, or transforming initial features into new ones, like Principal Component Analysis (PCA)[33]. Regardless of the chosen algorithm, there are common procedures to follow. For example, it is beneficial to remove features that have zero variance across the training set, as they don't provide any useful information to the model. Another typical approach is feature standardization or normalization. While tree-based algorithms are not affected by the distribution of features in the dataset, most others algorithms require all feature to be within comparable intervals[34]. Without standardization, the features with lower values may be overlooked by the algorithm.

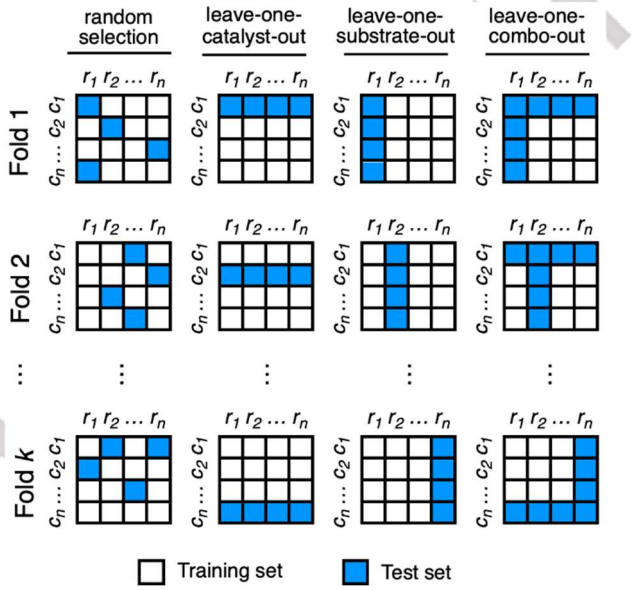


Figure 2. Examples of training and test set separation techniques in cross-validation for a dataset containing a matrix of catalysts (*c*) and substrates (*r*). In k-fold cross validation the subsets are randomly selected so that every data point is present in the test set once. The number of folds (*k*) is selected by the user, with lower *k* meaning a stricter validation. Leave-group-out validations systematically set groups related to individual catalysts, substrates, or their combinations aside to act as test sets, eliminating any influence of randomness. The number of repeats (folds) is therefore equal to the number of groups to test. These estimate the model's performance as if exposed to fully external data.

The choice of the specific ML algorithms depends on the task at hand and is ultimately decided by the researcher (Table 1), although some considerations related to the modeled property should be taken. For example, modeling activation energy expressed as ΔΔG or enantiomeric ratio (e.r.) in logarithmic scale is a case of Linear Free Energy Relationships, and most methods would have no issues with this task. On the other hand, predicting the enantiomeric excess (e.e.) requires building a non-linear relationship, so the choice of appropriate algorithms and features is limited. We also believe it is necessary to clarify the applicability of certain methods. Regression models, i.e., calculating a precise numerical value based on structure, are the most common, especially when researchers aim to identify catalysts with selectivities exceeding existing ones. In that case, ML methods such as Random Forest or K-Nearest Neighbors are inappropriate due to their reliance on averaging initial training set values, which mathematically prohibits extrapolation[35,36]. Alternatively, linear and non-linear algorithms (Elastic Net, Support Vector Machines, Gaussian Processes, Neural Networks) are more appropriate for that task[37].

Accepted Manuscript

This article is protected by copyright. All rights reserved.

Table 1. Popular ML algorithms used in QSPR modeling, their advantages and disadvantages[38].

Algorithm	Pros	Cons
Linear regression (LR)	• No parameterization required; • Low complexity and easy to interpret;	• Assumes linear relationship between features and property; • Outliers impact the model considerably; • Requires a low number of non-collinear features (unless using Ridge Regression or Elastic Net);
Random Forest (RF)	• Very few parameters with low impact on the model; • Non-linear, can capture more complex relationships; • Any features can be combined (continuous, binary, etc) without preprocessing;	• Long training times depending on parameters; • Changes in training data affect the model considerably; • Cannot predict property outside of the training range;
Boosting algorithms (XGBoost, LightGBM, CatBoost)	• Higher accuracy than RF due to the boosting algorithms; • Flexibility in the function fitting, complexity; • No data preprocessing required;	• Longer training times; • Outliers impact the model considerably;
Support Vector Machines (SVM)	• Very flexible function fitting, kernels allow for choice between linear and non-linear functions; • Does not require large datasets; • Generally, high accuracy of models without risk of overfitting;	• Model parameters must be optimized, parameter choice is not trivial; • Models are difficult to interpret; • Requires data preprocessing, cannot handle various data types in one dataset;
Gaussian Processes (GP)	• Flexible function fitting due to kernel choice; • Provides estimation of uncertainty for predictions; • Performs very well for smaller sets;	• Kernel hyperparameters require extensive tuning; • Computational cost is extremely high in the basic implementation, but state-of-the-art approaches[39] allow to reduce it significantly[40];
Neural Networks (NN)	• Can fit very complex linear and non-linear relationships between features and property; • Generally, offers very high accuracy; • Depending on the type of the NN, can work with various types of data (numerical, text, graphs, etc); • Graph NN[41,42] allow to learn features directly from molecular graphs, without precomputed features.	• Very prone to overfitting, requires good control during the training process; • Requires very large amounts of data to train effectively; • Models are very difficult to interpret;

The most important part in building a reliable model is its validation. To verify that the model is reliable when applied to external data, it is usually tested on a separate set that is not used in training. Often, such a set is taken from the initial data before the training. A great example of rigorous preparation of validation sets for selectivity modeling is set by Denmark and co-workers [9]. They divide the initial data into a training set and three validation sets – for new catalysts, for new substrates, and one fully external where both catalysts and substrates are new to the model. Other works using this data set follow these validation guidelines[27,30].

An even more systematic estimation of model reliability would come from using cross-validation (CV) – dividing the initial data into several sets of equal size, and then using each subset to test the model built on the rest. This approach ensures that all available data points are used for testing, offering a more comprehensive assessment of the model's performance. CV is widely used during the model hyperparameter optimization. While random separation of CV sets is the most common approach, it is often undesirable due to possible data leakage: since the separation into training and test sets is not controlled, same catalysts and/or substrates may end up in both, thus yielding overly optimistic results. It is thus recommended to systematically withhold subsets associated with distinct catalysts, substrates, solvents, *etc.*[43], imitating exploration of never-before-seen data (Figure 2). Withholding several groups simultaneously (e.g., leave-one-combo-out in Figure 2) provides the most challenging and most realistic validation of the model; however, it is often difficult to implement due to data scarcity. Another method of cross-validation is Leave-One-Out (LOO), where individual data points are withheld successively for testing purposes. This method is particularly suitable for smaller datasets (30 entries or fewer), although it tends to yield more optimistic scores since more data is utilized for training each individual model.

Based on these general concepts and premises, in this manuscript we aim to highlight the major developments related to the Quantitative Structure-Selectivity Relationship (QSSR) modeling using the 2D representation of reactions and catalysts in the context of asymmetric catalysis.

This article is protected by copyright. All rights reserved.

Accepted Manuscript

10.1002/chem.202302837

CONCEPT

2D descriptors in Quantitative Structure-Selectivity Relationship modeling

Earlier works on QSSR modeling using 2D-based descriptors followed the same principles as the 3D-based models: finding linear correlations between the property and the selected features. A series of studies by You and co-workers[44–46] described the application of topological indices to predict enantioselectivity in various asymmetric addition reactions. Topological indices represent the connectivity and other graph features (in this case, paths of certain lengths between atoms) as numerical values. While initially the authors utilized traditional Kier-Hall and Randic indices[44], they later turned to augmented indices developed by Yao *et al*[47]. These incorporate the atom type via the use of Van der Waals radii in the calculation of each index. The models built on single-layer NN have demonstrated good performance in cross-validation for a range of catalyst types (Figure 3).

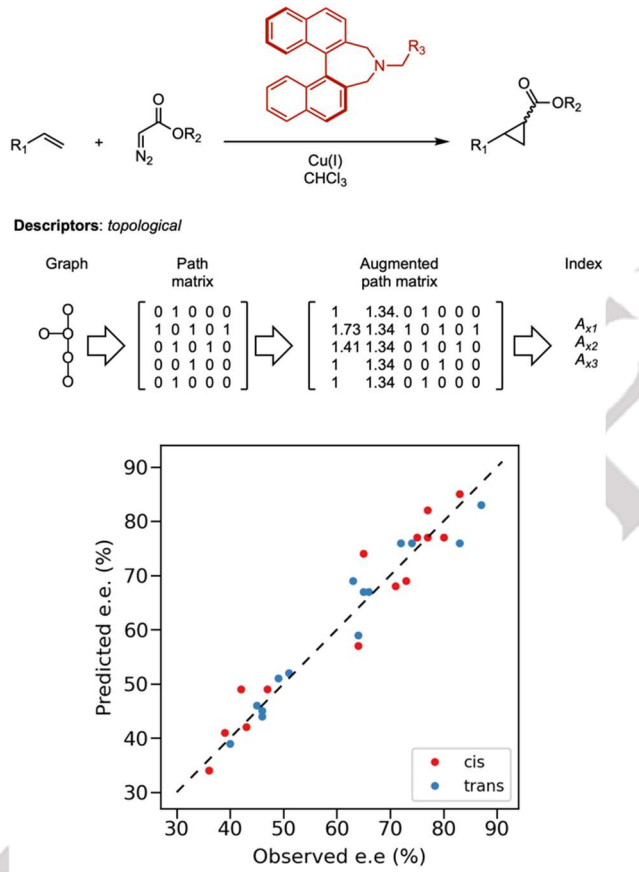


Figure 3. Modeling of selectivity of cyclopropanation reaction of olefins with diazoacetates by Jiang et al[46]. Using the topological indices calculated from augmented path matrices for substituents R_1, R_2, and R_3, as well as an indicator variable for absolute conformation of R_2, they train models that successfully predict selectivity of trans and cis products of the reaction. Note that two models are trained separately. Adapted with permission from [46]. Copyright 2003 American Chemical Society.

Metsänen *et al* investigated the enantioselective aza-Michael conjugate addition catalyzed by bistriflamides in the synthesis of Letermovir[48]. They elucidated the use of combined 2D descriptors (Carhart atom pairs) with physical organic descriptors such as NBO charges or logP, hypothesizing that their synergy would benefit the model. Indeed, upon comparing such hybrid models to both pure 2D and 3D models, they observed an improved performance in virtual screening, leading to the discovery of a better catalyst, albeit accompanied by an increased calculation cost (Figure 4).

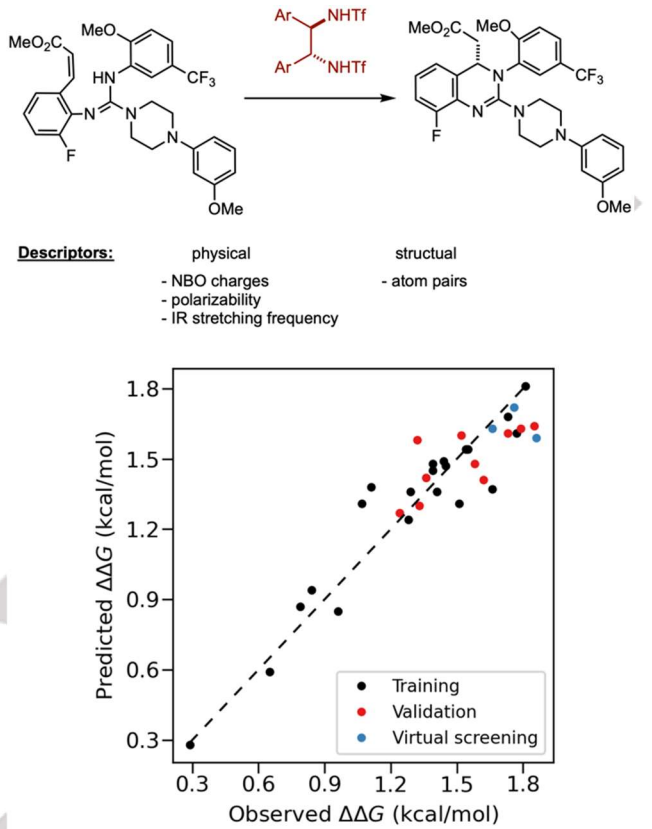


Figure 4. Modeling aza-Michael conjugate addition with bistriflamide catalysts by Metsänen et al[48]. A linear model combining physico-chemical parameters (NBO charges, logP) and Carhart atom pairs led to great performance in cross-validation on the training set (black dots in plot), external validation (red) and virtual screening for a novel catalyst (blue). Plot is adapted from the original paper[48].

The same reaction was further studied in the work of Lexa and co-workers[49]. They developed the iterative workflow for improvement of the enantioselectivity with cinchona alkaloid-derived phase-transfer catalysts using 2D and 3D models (Figure 5). Starting from the initial dataset of 177 catalysts, they iteratively increased the coverage of the training compounds, to increase chemical diversity, property range or to confirm the prospective predictions of previous rounds. In a ML benchmark authors described that Random Forest models based on Carhart atom pair counts, property-layered atom pair counts and MOE 2D descriptors outperformed binary descriptors. The iterative process allowed them to enrich the dataset by diverse catalysts with high selectivities and find novel catalysts outside of the initial selectivity range. Authors also note that application of 2D descriptors significantly accelerates and reduces the computational cost of model creation, while the accuracy of the model remains on par with 3D-based ones.

Accepted Manuscript

This article is protected by copyright. All rights reserved.

10.1002/chem.202302837

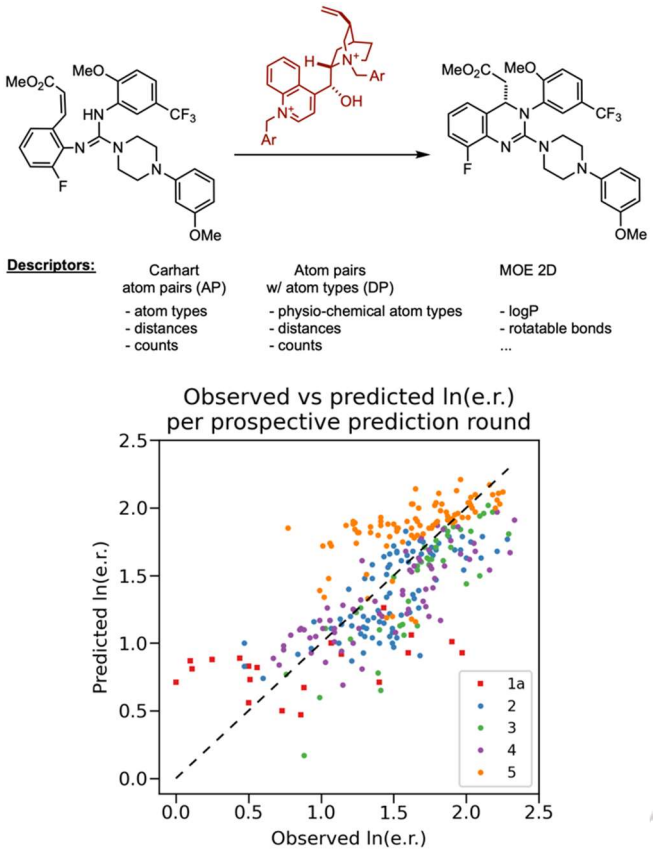


Figure 5. Iterative modeling aza-Michael conjugate addition with cinchona alkaloid-derived catalysts by Lexa et al[49]. Authors report a RF model based only on 2D parameters (atom pairs, MOE 2D). The plot shows the performance of model in prospective prediction each round (indicated by color). Adapted with permission from [49]. Copyright 2022 American Chemical Society.

A systematic benchmark of various ML methods has been reported by Glorius' group[27]. They studied the data on asymmetric N,S-acetal formation with CPA catalysts[9] and built models using various popular fingerprints. Fingerprints represent a chemical structure by binary vectors encoding the presence or absence of structural features. Different FP types account for different features: circular for Morgan FP, atom pairs and torsion angles, etc. The authors also introduced Multiple Fingerprint Features (MFF), represented by concatenation of several fingerprint vectors, that led to high accuracy of predictions in random validation sets and sets of new substrates outside of the training, even compared to the original models based on 3D structures of catalysts (Figure 6). Nevertheless, despite these advancements, the predictive capability of the model for new catalysts was limited, likely due to the inherent loss of information caused by binary 2D fingerprints when used to represent complicated catalyst structures. Moreover, the choice of Random Forest limited the prediction of selectivity beyond the training property range.

A unique difficulty of reaction property modeling is encoding a reaction by descriptors. Reactions consist of several molecules (at least one reactant and one product); moreover, other factors such as additives, catalysts, and reaction conditions affect the reactivity and selectivity, and thus cannot be ignored by the model. The common approach is to concatenate the descriptors vector of all participants, as in the work by Glorius' group[27].

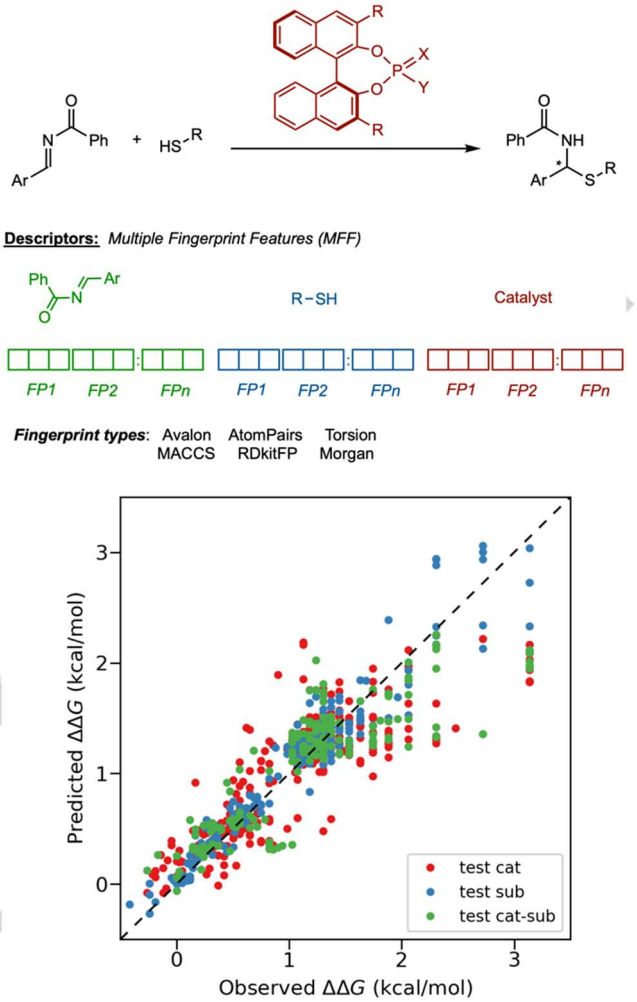


Figure 6. Modeling selectivity in N,S-acetal formation with CPA catalysts by Sandfort et al[27]. Multiple Fingerprint Features are introduced as a descriptor vector that combines several types of representations of 2D structure. These features allowed accurate performance comparable to 3D models in prediction of new catalysts (test cat, in red), substrates (test sub, in blue) or catalyst-substrate combinations (test cat-sub, in green); cross-validation scores and predictions for the training set are not indicated in the original work. Plot is reproduced using the data and code provided in the original paper[27].

While concatenation of fingerprints creates a vector that accounts for all factors that a researcher deemed important, there are several shortcomings. First, the vector limits the scope of the model: it can only consider the reactions with the same number of molecules/additives, and they must always be presented to the model in the same order. Therefore, if structural features are used as descriptors, this prevents the model from learning the contribution of the same features in different molecules. In this context, Condensed Graph of Reaction (CGR)[50] is an alternative approach of reaction representation. A reaction is represented as a single graph, or a pseudo-molecule, with special dynamic bonds and atoms that indicate the changes occurring during the reaction.

Accepted Manuscript

This article is protected by copyright. All rights reserved.

10.1002/chem.202302837

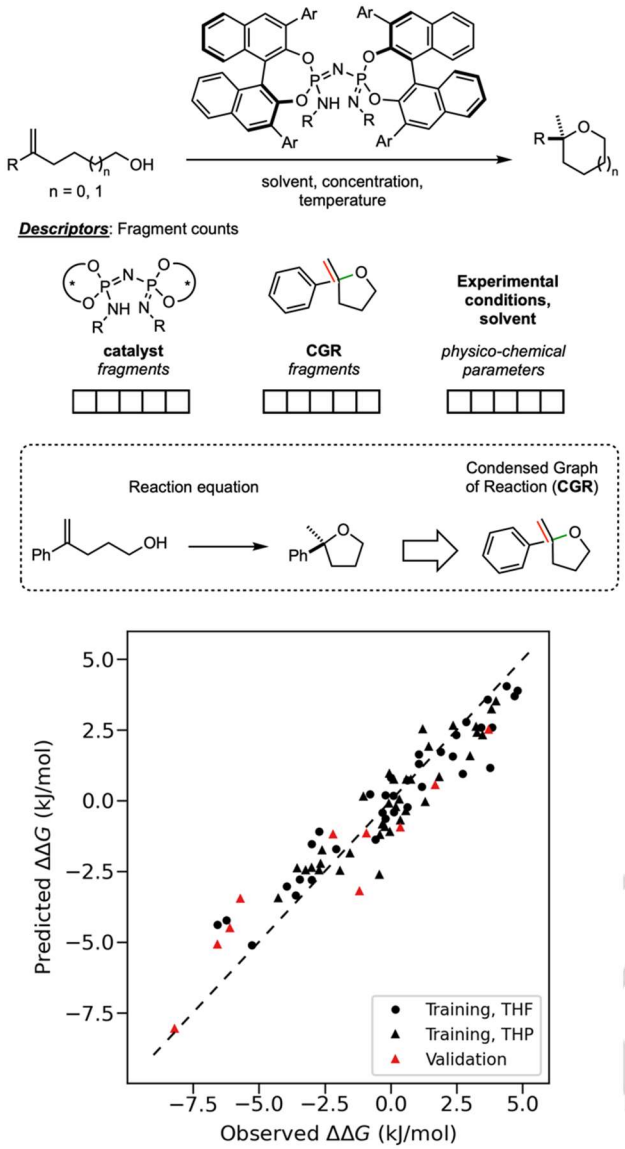


Figure 7. Prediction of novel IDPi catalysts in asymmetric hydroalkoxylation reactions by Tsuji et al[30]. Authors utilize molecular fragments derived from catalysts and reactions, as well as experimental conditions as parameters in SVM model. The model has succeeded in predicting a catalyst with a higher selectivity than any in the initial training set (validation set in red). Plot is adapted from the original paper[30].

This way, the concerns related to the number of reactants, their order in the concatenated vector and its length are alleviated. CGRs have been used in plethora of works related to modeling of various reaction properties[51–53]. Most notably, in the recent work Tsuji et al[30] opted for fragment count descriptors on CGRs in prediction of enantioselectivity of imidodiphosphorimidate (IDPi) catalysts in hydroalkoxylation reaction[30]. CGRs aided in integrating the datasets related to different substrates, as well as representing the catalysts by the substituents rather than the full structure, which led to both better performance and better interpretability of the model. Molecular fragments such as ISIDA represent counts of occurrences of various substructures (linear, circular, etc) in the molecule, allowing for high adaptability of the model. The authors also introduced Circular Substructures (CircuS) descriptors, to account for the complex polycyclic structure of the catalysts, which led to a performant model capable of predicting catalysts with higher selectivity than any used for training (Figure 7).

While numerous works demonstrated the accuracy of models based on 2D molecular structures, some disadvantages of them are often reported. One of the main weaknesses that is attributed to 2D-based models is their apparent disconnection from the physical interpretation of the modeled phenomenon. While topological indices have a degree of inherent interpretability as they encode how the connectivity, branching and rings in the structure affect the predicted property, fingerprints and fragment counts represent molecules by a large number of substructures that may only have an implicit connection to the property. Fingerprints are especially hard to interpret due to hashing algorithms obfuscating the features represented by each column. Furthermore, both fingerprints and fragments descriptors are usually used in tandem with non-linear ML methods, rendering the model even more complicated. However, some methods of interpreting the model's decisions still exist for those.

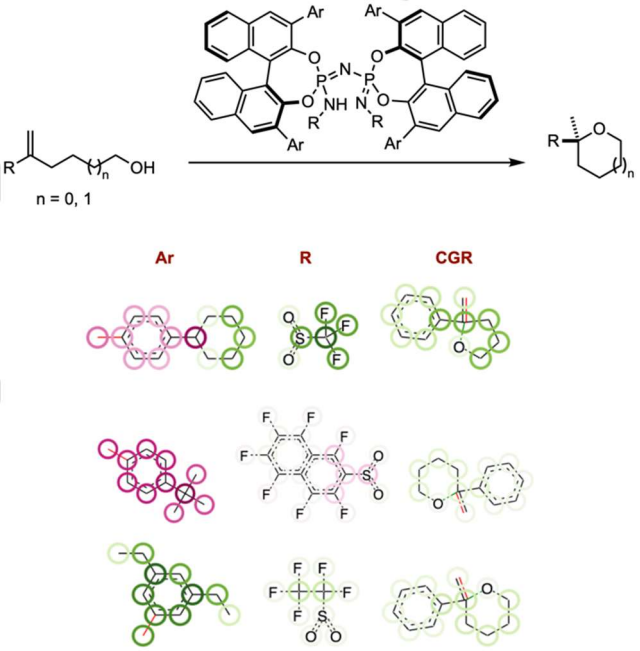


Figure 8. ColorAtom used for interpretation of IDPI catalyst selectivity prediction model[30]. Relative atom contributions are indicated by colors: green indicates that the atom contributes to a more positive prediction, red – to a more negative. The intensity of the color shows the relative importance of the atomic contribution.

ColorAtom is a method developed specifically for ISIDA fragment counts[54], although its application with other fragments, and, most importantly, in catalyst enantioselectivity model, has also been reported[30]. In this approach, the weights of all fragments are calculated as a partial derivative of the predicted value, and atom contributions are estimated as the sum of weights for all fragments where these atoms are present. The approach works in both classification and regression and allows for a simple and visual interpretation of the predictions (Figure 8).

Models based on fingerprints are more difficult to explain structurally. Recently, some approaches like SHAP values have

Accepted Manuscript

This article is protected by copyright. All rights reserved.

been reported for the QSAR models on molecules. SHAP values represent the relative importance of each feature, and methodology to relate these back to structure via fingerprints have been described by Bajorath et al[55,56]. However, no application of that approach in prediction of reaction properties has been reported yet.

It is also important to consider the applicability domain (AD) of the model. By the definition of Netzeva et al[57], AD is the part of chemical space and property space where a model predict with a given reliability. Note that AD is defined for the property space, too, as most ML methods only work reliably in interpolation. Some metrics for AD (leverage, zKNN, etc) are based on distance to model, or how similar are the tested compounds to the training set[58,59]. Other, more categorical measures, also exist: bounding box checks whether the descriptor values of the test molecules are within the boundaries of the training data, while fragment control, unique to fragments, considers all molecules with new substructures to be outliers[60]. However, no systematic studies of AD for enantioselectivity prediction have been carried out yet.

Summary and Outlook

Machine learning models are a significant aid in the rational design of compounds with desired properties, including highly selective asymmetric catalysts. The use of experimental or calculated steric and electronic parameters derived from 3D structures of catalysts and substrates allows to build a model with mechanistic interpretation; however, the high computational cost of these features limits the applicability of such models. Descriptors derived from 2D structures, such as fingerprints, fragments, or topological indices, offer a practical alternative, capable of building highly accurate models at a fraction of the cost. While 2D descriptors generally lack explicit connection to the physical phenomena of the catalysis, certain techniques for interpretation of 2D models, such as ColorAtom or SHAP values analysis, are actively being developed and applied in various fields. Therein lies an opportunity for a practical virtual screening approach that combines high data capacity, low computational cost, and the chemical insights that can guide the synthesis towards the desired selectivity.

The scarcity of data remains, however, a significant challenge in the application of machine learning in asymmetric catalysis. The most common task in this field is the optimization of catalyst structure for a specific reaction. On one hand, this restricts the model to the prediction of magnitude of selectivity, and often parameters such as the configuration of catalyst itself are irrelevant, rendering the modeling process simpler. On the other hand, this limits the opportunities for building a general, all-encompassing model for selectivity that would consider a wide range of reactions, mechanisms, and catalysts. While there are examples of the application of 3D QC features to such tasks[8], 2D representations might struggle as special descriptors that would account for chirality in substrates and catalysts would be necessary. Similarly, modeling enantioselectivity in regio-, chemo-, and diastereoselective transformations is challenging, especially for 2D descriptors, as it is crucial to consider selectivity towards several isomers and their co-dependence. Thus, the field of asymmetric catalysis is still rife with opportunities for innovations in machine learning.

Acknowledgements

This work was financially supported by the Institute for Chemical Reaction Design and Discovery (ICReDD), which was established by the World Premier International Research Initiative (WPI), MEXT, Japan, List Sustainable Digital Transformation Catalyst Collaboration Research Platform offered by Hokkaido University, and by JSPS KAKENHI Grants 21H01925 and 23K16936.

Keywords: asymmetric catalysis • machine learning • quantitative structure-selectivity relationships

[1] "The Nobel Prize in Chemistry 2001," can be found under https://www.nobelprize.org/prizes/chemistry/2001/summary/, **2001**.

[2] "The Nobel Prize in Chemistry 2021," can be found under https://www.nobelprize.org/prizes/chemistry/2021/summary/, **2021**.

[3] K. C. Harper, M. S. Sigman, *Science (80-.).* **2011**, *333*, 1875–1878.

[4] K. C. Harper, E. N. Bess, M. S. Sigman, *Nat. Chem.* **2012**, *4*, 366–374.

[5] C. B. Santiago, A. Milo, M. S. Sigman, *J. Am. Chem. Soc.* **2016**, *138*, 13424–13430.

[6] T. Tang, A. Hazra, D. S. Min, W. L. Williams, E. Jones, A. G. Doyle, M. S. Sigman, *J. Am. Chem. Soc.* **2023**, *145*, 8689–8699.

[7] S. K. Nistanaki, C. G. Williams, B. Wigman, J. J. Wong, B. C. Haas, S. Popov, J. Werth, M. S. Sigman, K. N. Houk, H. M. Nelson, *Science (80-.).* **2022**, *378*, 1085–1091.

[8] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.

[9] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science (80-.).* **2019**, *363*, eaau5631.

[10] J. D. Oslob, B. Åkermark, P. Helquist, P.-O. Norrby, *Organometallics* **1997**, *16*, 3015–3021.

[11] J. M. Brown, R. J. Deeth, *Angew. Chemie Int. Ed.* **2009**, *48*, 4476–4479.

[12] K. B. Lipkowitz, M. Pradhan, *J. Org. Chem.* **2003**, *68*, 4648–4656.

[13] S. E. Denmark, N. D. Gould, L. M. Wolf, *J. Org. Chem.* **2011**, *76*, 4337–4357.

[14] S. Yamaguchi, M. Sodeoka, *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701–1706.

[15] H. Chen, S. Yamaguchi, Y. Morita, H. Nakao, X. Zhai, Y. Shimizu, H. Mitsunuma, M. Kanai, *Cell Reports Phys. Sci.* **2021**, *2*.

[16] M. C. Kozlowski, S. L. Dixon, M. Panda, G. Lauri, *J. Am. Chem. Soc.* **2003**, *125*, 6614–6615.

[17] M. Urbano-Cuadrado, J. J. Carbó, A. G. Maldonado, C. Bo, *J. Chem. Inf. Model.* **2007**, *47*, 2228–2234.

[18] V. L. Cruz, S. Martinez, J. Ramos, J. Martinez-Salazar, *Organometallics* **2014**, *33*, 2944–2959.

[19] D. Zankov, P. Polishchuk, T. Madzhidov, A. Varnek, *Synlett* **2021**, *32*, 1833–1836.

[20] D. Zankov, T. Madzhidov, P. Polishchuk, P. Sidorov, A. Varnek, *J. Chem. Inf. Model.* **2023**, *63*, 6629–6641.

[21] P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016.

[22] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, F. Wang, *Brief. Bioinform.* **2020**, *21*, 919–935.

Accepted Manuscript

This article is protected by copyright. All rights reserved.

[23] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nat. Commun.* **2019**, *10*, 2903.

[24] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, K.-R. Müller, *Nat. Commun.* **2021**, *12*, 7273.

[25] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2014**, *57*, 4977–5010.

[26] A. Tropsha, *Mol. Inform.* **2010**, *29*, 476–488.

[27] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **2020**, *6*, 1379–1390.

[28] G. Landrum, **2006**.

[29] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 693–703.

[30] N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, B. List, *Angew. Chemie - Int. Ed.* **2023**, *62*, e202218659.

[31] Y. Liu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823–1828.

[32] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, *TrAC Trends Anal. Chem.* **2013**, *42*, 49–63.

[33] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, E. Tuzhilina, *Nat. Rev. Methods Prim.* **2022**, *2*, 100.

[34] H. Brink, J. Richards, M. Fetherolf, *Real-World Machine Learning*, Simon And Schuster, **2016**.

[35] G. Biau, E. Scornet, *TEST* **2016**, *25*, 197–227.

[36] G. Biau, L. Devroye, *J. Multivar. Anal.* **2010**, *101*, 2499–2518.

[37] P. F. Smith, S. Ganesh, P. Liu, *J. Neurosci. Methods* **2013**, *220*, 85–91.

[38] A. Varnek, *Tutorials in Chemoinformatics*, John Wiley & Sons, **2017**.

[39] F. X. Yu, A. T. Suresh, K. Choromanski, D. Holtmann-Rice, S. Kumar, **2016**, DOI https://doi.org/10.48550/arXiv.1610.09072.

[40] N. J. Browning, F. A. Faber, O. Anatole von Lilienfeld, *J. Chem. Phys.* **2022**, *157*, 214801.

[41] M. Sakai, K. Nagayasu, N. Shibui, C. Andoh, K. Takayama, H. Shirakawa, S. Kaneko, *Sci. Rep.* **2021**, *11*, 525.

[42] 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.

[43] A. Rakhimbekova, T. N. Akhmetshin, G. I. Minibaeva, R. I. Nugmanov, T. R. Gimadiev, T. I. Madzhidov, I. I. Baskin, A. Varnek, *SAR QSAR Environ. Res.* **2021**, *32*, 207–219.

[44] J. Chen, W. Jiwu, L. Mingzong, T. You, *J. Mol. Catal. A Chem.* **2006**, *258*, 191–197.

[45] C. Jiang, D. Li, J. Wen, T. You, *J. Mol. Model.* **2006**, *13*, 91–97.

[46] C. Jiang, Y. Li, Q. Tian, T. You, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1876–1881.

[47] Y. Yao, L. Xu, Y. Yang, X. Yuan, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 590–594.

[48] T. T. Metsänen, K. W. Lexa, C. B. Santiago, C. K. Chung, Y. Xu, Z. Liu, G. R. Humphrey, R. T. Ruck, E. C. Sherer, M. S. Sigman, *Chem. Sci.* **2018**, *9*, 6922–6927.

[49] K. W. Lexa, K. M. Belyk, J. Henle, B. Xiang, R. P. Sheridan, S. E. Denmark, R. T. Ruck, E. C. Sherer, *Org. Process Res. Dev.* **2022**, *26*, 670–682.

[50] F. Hoonakker, N. Lachiche, A. Varnek, A. Wagner, *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

[51] G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch, A. Varnek, *J. Chem. Inf. Model.* **2015**, *55*, 239–250.

[52] T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, I. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin, A. Varnek, *Mol. Inform.* **2019**, *38*, 1800104.

[53] T. Gimadiev, R. Nugmanov, A. Khakimova, A. Fatykhova, T. Madzhidov, P. Sidorov, A. Varnek, *J. Chem. Inf. Model.* **2022**, *62*, 2015–2020.

[54] G. Marcou, D. Horvath, V. Solov'Ev, A. Arrault, P. Vayer, A. Varnek, *Mol. Inform.* **2012**, *31*, 639–642.

[55] R. Rodríguez-Pérez, J. Bajorath, *J. Comput. Aided. Mol. Des.* **2020**, *34*, 1013–1026.

[56] C. Feldmann, J. Bajorath, *iScience* **2022**, *25*, 105023.

[57] T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. W. Roberts, T. W. Schultz, D. T. Stanton, J. J. M. van de Sandt, W. Tong, G. Veith, C. Yang, *Altern. to Lab. Anim.* **2005**, *33*, 155–173.

[58] I. V Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

[59] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, V. V Kovalishyn, V. V Prokopenko, I. V Tetko, *J. Chemom.* **2010**, *24*, 202–208.

[60] A. Rakhimbekova, T. I. Madzhidov, R. I. Nugmanov, T. R. Gimadiev, I. I. Baskin, A. Varnek, *Int. J. Mol. Sci.* **2020**, *21*, DOI 10.3390/ijms21155542.

Accepted Manuscript

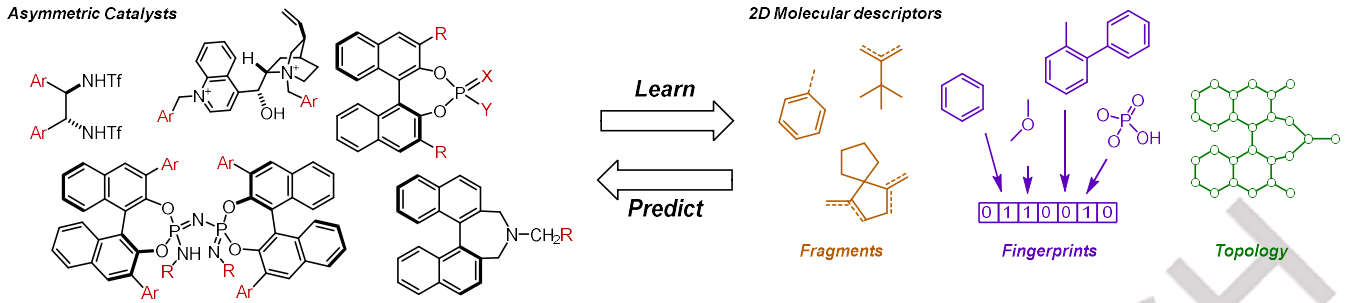
This article is protected by copyright. All rights reserved.

10.1002/chem.202302837

WILEY VCH

CONCEPT

Entry for the Table of Contents



This brief review highlights recent advances in modeling catalyst selectivity using 2D structures of catalysts and substrates. 2D descriptors offer the advantage of low cost and high speed in calculations compared to quantum chemical parameters. We provide an overview of a typical Quantitative Structure-Property Relationship workflow, including model building and validation techniques, as well as applications of these methodologies in the design of asymmetric catalysis.

Institute and/or researcher Twitter usernames: @ICReDDconnect

Accepted Manuscript

This article is protected by copyright. All rights reserved.