Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors

Authors: Nobuya Tsuji\textsuperscript{1,}\textsuperscript{*}, Pavel Sidorov\textsuperscript{1,}\textsuperscript{*}, Chendan Zhu\textsuperscript{2}, Yuuya Nagata\textsuperscript{1}, Timur Gimadiev\textsuperscript{1}, Alexandre Varnek\textsuperscript{1,3,}\textsuperscript{*}, Benjamin List\textsuperscript{1,2,}\textsuperscript{*}

Affiliations:
\textsuperscript{1}Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo 001-0021, Japan
\textsuperscript{2}Max-Planck-Institut für Kohlenforschung, D-45470 Mülheim an der Ruhr, Germany
\textsuperscript{3}Laboratory of Chemoinformatics, UMR 7140, CNRS, University of Strasbourg, 67081 Strasbourg, France
\textsuperscript{*}Corresponding author.

Email: tsuji@icredd.hokudai.ac.jp, pavel.sidorov@icredd.hokudai.ac.jp, varnek@unistra.fr, and list@mpi-muelheim.mpg.de

†These authors contributed equally to this work.

Abstract: Catalyst optimization process is typically relying on an inductive and qualitative assumption of chemists based on screening data. While machine learning models using molecular properties or calculated 3D structures enable quantitative data evaluation, costly quantum chemical calculations are often required. In contrast, readily available binary fingerprint descriptors are time- and cost-efficient, but their predictive performance remains insufficient. Here, we describe a machine learning model based on fragment descriptors, which are fine-tuned for asymmetric catalysis and represent cyclic or polyaromatic hydrocarbons, enabling robust and efficient virtual screening. Using training data with only moderate selectivities, we designed theoretically and validated experimentally new catalysts showing higher selectivities in a previously unaddressed transformation.

Main text:

In asymmetric catalysis, screening processes toward satisfactory selectivity require considerable effort and time, and typically relies on trial-and-error experimentation. Chemists often identify optimal catalysts with good enantioselectivities using “educated guessing”, in other words, by considering steric, electronic, and conformational effects of the catalysts. However, the efficiency of this process largely depends on the scientists involved, and even when such “predictions” work, the outcome is typically more qualitative in nature. The development of tools which enable a quantitative evaluation of screening data is highly desirable.

Machine learning is recognized as a practical tool to quantify molecular properties in the contexts, such as biological activity or chemical synthesis. Historically, quantitative explanations and predictions of reactivity and selectivity from structural information follow Hammett’s approach, where steric and electronic properties of substituents at certain positions in a series of substrates or catalysts act as equation parameters. The variables are obtained by either experiment\textsuperscript{1,2} or quantum chemical computation,\textsuperscript{3} and are widely applied to asymmetric catalysis using multivariate linear regression models or non-linear models.\textsuperscript{4-9} While this approach can provide insights into the reaction mechanism, the choice of those descriptors typically relies on the knowledge of chemists, thus rendering the model difficult to consider general features of the catalyst structure.\textsuperscript{6}
Alternatively, 3D descriptors represent general structural information, and thus models on 3D structures may identify correlations without requiring knowledge of the detailed reaction mechanism. These methods have already been applied in asymmetric catalysis, using calculated 3D structures with grid-based binary feature vectors, such as Comparative Molecular Field Analysis (CoMFA). The Denmark group reported a modification of grid-based methods with non-binary features based on distribution of conformations to account for their flexibility, and demonstrated that the model can predict higher selectivity even in the absence of such examples in the training dataset. The disadvantage of such methods is the requirement of costly quantum chemical calculations and, in the case of grid-based methods, the necessity of alignment of core structures. Several 3D structure-based methods which do not require alignment have also been described in the literature, however, their performances, especially when it comes to extrapolation, are limited.

In contrast, 2D descriptors, such as fragment counts or binary fingerprints, also represent general structural features, albeit often implicitly, and allow to avoid costly calculations, as they are derived directly from a 2D representation of a molecule. Using 2D descriptors, therefore, has a clear advantage in the speed. Nonetheless, fingerprint descriptors by themselves are considered insufficient in representing the structures to construct reliable predictive models in asymmetric catalysis. Recently, multiple fingerprint features were employed together to build predictive models. Despite the relatively good performance of the test set on substrates, the validation sets for catalysts still had room for improvement, which confirms the inherent difficulty of representing complicated catalyst structures solely relying on the binary 2D fingerprints.

Such unsatisfactory results would partly stem from their binary nature; they often do not contain enough structural information to construct a reasonable model, as a key corresponding to a certain fragment is only determined by its presence. Nevertheless, simply employing conventional non-binary fingerprints is still insufficient for predictions; thus a relatively large number of training data with multiple optimization cycles were required as described by Belyk, Sherer, and co-workers. Alternatively, fragment count descriptors, represented

![Figure 1](image.png)

**Figure 1.** (A) The data set (THF) used in the initial study. (B) The general modeling workflow. Each catalyst/reaction combination has been encoded by different types of molecular descriptors forming several descriptors pools. Each descriptors pool was used for the building of related Support Vector Regression model, performance of which (determination coefficient, $R^2$) was optimized in 10-fold cross-validation. Best performing model (model with maximum $R^2$) is selected among those for external predictions.
by ISIDA (In Silico design and Data Analysis) platform, encode the fragments into the non-binary vectors based on the number of their occurrences in a molecule without limiting the number of features. ISIDA descriptors in particular offer a wide variety of possible representations of the chemical structures, including topology (linear, atom-centered fragments, atom pairs, or triplets) and size of fragments, which allows to fine-tune them to the problem at hand. Furthermore, ISIDA platform enables to calculate fragments for reaction schemes as a whole via Condensed Graphs of Reaction (CGR), a special representation that combines reactants and products into a single pseudo-molecule containing dynamic (i.e., changing during the reaction) bonds, which is an unique property of this approach. Although these descriptors have been utilized in modeling of pharmacological properties, materials, and chemical transformations, their application to asymmetric catalysis has never been described.

Here, we report a predictive model for enantioselectivity of structurally diverse and flexible catalysts using fragment count descriptors based on ISIDA platform. For more precise representation of cyclic or polyaromatic hydrocarbon substituents which are common in asymmetric catalysis, another type of fragments is introduced. Further, its application to an actual synthetic challenge is demonstrated; highly selective catalysts of a previously unsolved asymmetric catalytic reaction have been predicted from the training data consisting of catalysts showing only moderate selectivities.

We initiated a preliminary study on the stereoselective hydroalkoxylation reaction of simple alkenes to construct 2,2-disubstituted tetrahydrofuran rings using imidodiphosphorimidate (IDPi) catalysts under several different conditions (35 examples, e.r. from 15:85 to 91:9; referenced further as THF set). These IDPi catalysts possess a common scaffold, yet have a variety of substituents on the 3,3’-positions of BINOLs and nitrogen substituents, represented as Ar and R (Figure 1A, only R=Ts in THF set for the sake of simplicity). The workflow is summarized in Figure 1B. The structural information of catalysts, substrates, and products is encoded by either fingerprints or fragments descriptors. In the case of fragments, the reaction schemes are transformed into CGRs. Structures of catalysts can be represented either as full structures or only by the modulable substituents depicted

![Figure 2. (A) Descriptor benchmarking results (as MAE, kJ/mol) on THF set in Leave-One-Out (LOO) cross-validation. The SVM model was optimized by genetic algorithms, and the best model’s results are shown for several descriptor types (see Figure S9 for full data). (B) Fragments generated by CircuS at given distance from center atoms. (C) LOO cross-validation results for the best model (CircuS fragments, solvent descriptors, catalysts represented by the Ar substituent) on the THF dataset. Statistical parameters (MAE, R²) are indicated on the plot.](image)
as Ar and R; the latter would reduce noise from the common scaffold and enable to apply different descriptors which are suitable for each substituent. Physico-chemical parameters of solvents are also included as descriptors into the model. Other reaction conditions, namely temperatures and concentrations, are separately considered; all data points are calibrated to an identical condition before constructing a predictive model, and the predicted free energy differences are calibrated back to each experimental condition, providing the predicted enantiomeric ratio in the given conditions (see SI for details). Thus, this workflow represents a generalized approach to handling catalysts, reaction schemes, and reaction conditions simultaneously.

With these strategies, preliminary benchmark studies were performed to compare the performance of various fingerprints and ISIDA fragment counts with which both atom-centered and linear fragments are considered. Due to the modest size of the dataset, the Support Vector Machine (SVM) method was chosen as a widely used non-linear machine learning algorithm.42 Descriptor space and model hyperparameters were optimized by genetic algorithms in 10-fold cross-validation. In our preliminary studies, ISIDA fragments have demonstrated better predictivity than any conventional fingerprints (Figure 2A). However, there is still a concern of their intrinsic linearity: for instance, atom-centered fragments are represented as a collection of linear fragments sharing the same central atom, and thus they do not explicitly account for cyclic structures. While this is normally considered to make the descriptors more general and insensitive to subtle changes in structures,44 it may introduce ambiguity in the case of structures such as IDPi catalysts, which contain a large number of condensed aromatic rings, or certain substituents, like polycyclic hydrocarbons.

To account for this complexity, augmented substructure type descriptors called CircuS (Circular Substructures) were developed in the framework of the CGRtools library.45 They function similarly to ISIDA atom-centered fragments, yet explicitly consider encountered full substructures with closed rings, representing

![Figure 3. (A) Target reaction (THP synthesis) to predict more selective catalyst. (B) 10-fold cross-validation results for the best model (CircuS fragments, solvent descriptors, catalysts represented by the Ar and R substituent, and the reaction is represented as CGR) on the THF dataset (black dot) and THP dataset (black triangle). Dash line represents perfect prediction. Statistical parameters (MAE, R²) are indicated on the plot. Validation sets for THP are indicated in red triangle. (C) Selected examples for comparison between experimental and predicted values under different reaction conditions. ΔΔG is represented in kJ/mol.](image)
complex structural motifs more specifically (Figure 2B). As expected, the CircuS performed even better than ISIDA descriptors; the fully optimized model built on CircuS fragments of catalyst substituents outperformed others with the satisfactory $R^2=0.905$, $MAE=0.741$ kJ/mol in LOO (Figure 2A and 2C).

We then turned our attention to demonstrate the generality of the developed method and envisioned to apply it for computer-aided design of more selective catalysts for tetrahydropyran (THP) synthesis (Figure 3A). Taking the advantage of the modularity of IDPi catalysts, a variety of catalysts were prepared, providing a diversity on the $3,3'$-positions of BINOL backbones represented as Ar and the sulfonyl groups on nitrogen atoms represented as R. For more efficient screening and data consistency, a synthesis robot was employed, streamlining the process from experiments to data generation (Figure S1). To build a predictive model for the THP synthesis, we then constructed a common data comprising with 35 examples for THF and 35 examples for THP to ensure the diversity of the catalyst structures (Figure S7).

The fully optimized model built on CircuS fragment descriptors provided a reasonable predictivity with $R^2_{CV}=0.878$, $MAE_{CV}=0.754$ kJ/mol (Figure 3B). Based on this, a virtual screening has been conducted. As the training data does not cover all the combinations of Ar and R groups of IDPi catalysts, additional 190 catalysts representing experimentally unseen combinations of those were prepared in silico, and their selectivities were virtually evaluated. As expected, among all the predictions, some catalysts were predicted to provide higher enantioselectivities than those experimentally obtained in the training set. Following these results, additional catalysts were prepared and evaluated as a validation set. Overall, predictions worked well (Figure 3B), and the most selective catalyst 61 indeed provided an enantiomeric ratio of 91.5:8.5, which is significantly higher than any examples from the training set (up to 82:18 e.r.). Additionally, by changing the temperature and concentration, an even higher enantiomeric ratio (96:4) was predicted, and this predicted value was almost identical to the experimentally observed value (Figure 3C).

Having been able to predict the optimal catalysts and conditions, we were keen on exploring other substrates. Phenyl-substituted substrate 1b furnished an excellent isolated yield and enantioselectivity even on a larger scale. This result is superior to the previously reported one, regarding both reactivity and selectivity, which underlines the efficiency of the presented method. Naphthalene derivative 2c was also obtained with excellent yield and enantioselectivity. The more challenging aliphatic substrates, both with primary and secondary alkyl groups,

![Image](image_url)

**Figure 4.** Substrate scope with the optimal catalyst 61. Isolated yields are presented. Enantiomeric ratio (e.r.) determined either via high-performance liquid chromatography (HPLC) or gas chromatography (GC). *0.025 M at 40 °C.
underwent the desired cyclization to give products 2d and 2e in excellent yields and enantioselectivities (Figure 4).

The methodology delineated here enables quantitative evaluation of catalyst selectivities, including the predictions of extrapolated examples. This approach is solely based on 2D structures and reaction conditions and does not require costly quantum mechanical calculation. Our approach appears to be of utility to optimize various catalytic reactions enabling fast and robust predictions.

Acknowledgement:
We thank the generous support from the Max Planck Society, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Leibniz Award to B.L.) and under Germany’s Excellence Strategy (EXC 2033-390677874-RESOLV), the European Research Council (ERC, European Union’s Horizon 2020 research and innovation program “C–H Acids for Organic Synthesis, CHAOS” Advanced Grant Agreement No. 694228). This work was also 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, and by JSPS KAKENHI Grants 21H01925, 22K14672, and 21H01924. The authors thank Dr. Thomas Buyck and Dr. Jennifer L. Kennemur for screening data and preliminary studies. We also thank Dr. Roberta Properzi, Dr. Ravindra Raut, Dr. Yowei Xie, and Stefanie Dehn for analytical data of catalysts, and all group members for sharing IDPi catalysts. We also thank the technicians and the members of our NMR, MS, and chromatography groups for their excellent service. We also thank Instrumental Analysis Division, Global Facility Center, Creative Research Institution, Hokkaido University for HRMS with Thermo Fisher Scientific Exactive Plus.

Code and data availability:
Data used in the study (structures and enantioselectivities of catalysts, experimental conditions, calibration curves, prediction results) is available in the Supporting Information. Python libraries developed in the study are open to public and are available on GitHub: https://github.com/POSidorov/ChemInfoTools (CircuS fragment descriptors), https://github.com/POSidorov/HPLCAnalysis (analysis of experimental HPLC data). Full protocol of data processing and modeling is available as Jupyter Notebook in Supporting materials.

Competing interests: The authors declare the following competing financial interest(s): A patent, WO2017037141 (A1), has been filed by the MPI fur Kohlenforsohung covering the IDPi catalyst class and their applications in asymmetric synthesis.


