Multi-Instance Learning Approach to the Modeling of Enantioselectivity of Conformationally Flexible Organic Catalysts

D. Zankov[1], T. Madzhidov[2], P. Polishchuk[3], P. Sidorov[4] and A. Varnek[1,4*]

[1] Laboratory of Chemoinformatics, University of Strasbourg, France

[2] Chemistry Solutions, Elsevier Ltd, Oxford, United Kingdom

[3] Institute of Molecular and Translational Medicine, Palacký University, Czech Republic

[4] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Japan

Abstract

The computational design of chiral organic catalysts for asymmetric synthesis is a promising technology, which may significantly reduce the material and human resources required for the preparation of enantiopure compounds. Herein, for the modelling of catalysts' enantioselectivity, we propose to use the Multi-Instance Learning (MIL) approach accounting for multiple catalyst conformers and requiring neither conformers selection nor their spatial alignment. A catalyst was represented by an ensemble of conformers, each encoded by 3D *pmapper* descriptors. A catalyzed chemical transformation was converted into a single molecular graph - Condensed Graph of Reaction (CGR) - encoded by 2D fragment descriptors. A whole chemical reaction was finally encoded by concatenated 3D catalyst and 2D transformation descriptors. The performance of the proposed method was demonstrated in the modelling of enantioselectivity of homogeneous and phase-transfer reactions and compared with some state-of-the-art approaches.

Introduction

Synthesis of enantiopure compounds is a hot topic of modern organic chemistry because highly effective drugs can be chiral, and enantiomers often have different biological activities. List and McMillan demonstrated that chiral organic molecules can effectively catalyze asymmetric reactions leading to enantiopure compounds[1,2]. Since their seminal publications, numerous chiral catalysts have been designed[3]. In most cases, the pursuit of perspective catalysts was conducted by iterative trials and errors approach, in which chemists relied on their professional experience, chemical intuition, and available experimental data. This approach, albeit often culminated in the desired result, still depends on the professional background of the researcher. At the same time, theoretical calculations may suggest the

chemical structure of promising catalysts before their synthesis and experimental tests, thus, reducing the time and material costs.

A perspective computational approach to the discovery of new catalysts is Quantitative Structure-Selectivity Relationship (QSSR) analysis, in which machine learning algorithms are applied to establish a relation between experimental enantioselectivity and a catalyst structure encoded by numerical descriptors. If such a relationship is established, the obtained predictive model can be used for the virtual screening of catalysts' candidates. The first notable example of QSSR application was reported by Norrby et al.[4], who used some geometry parameters of the metal complex (bond lengths, bond angles, and dihedral angles) and multivariate regression to analyze palladium-catalyzed allylation. Most other early studies were based on the Molecular Interaction Fields (MIF)[5] approach in which interaction energies between a given molecule with a probe species (atoms, point charges, small molecules, etc) are used as descriptors. In MIF, molecules represented by their low-energy conformers are aligned inside a rectangular box, whereas the probes are fixed in nodes of the 3D grid superposed with the box. The earliest and most popular MIF technique is the Comparative Molecular Field Analysis (CoMFA), in which steric and electrostatic energies with a probe (a neutral carbon atom and a proton, respectively) are correlated with experimental activities [6]. Lipkowitz et al.[7] reported the first application of CoMFA to the prediction of catalyst enantioselectivity in Diels-Alder reactions. Melville et al.[8,9] studied the glycine imine alkylation with quaternary ammonium ion catalysts in asymmetric phase-transfer catalysis (APTC), first considering a single catalyst conformer within CoMFA[8], then using Boltzmann-weighting of selected catalyst conformers. They demonstrated that accounting for the conformation diversity of catalysts led to some improvement in enantioselectivity predictions compared to the conventional CoMFA model. The addition of thiols to imines catalyzed by phosphoric acids was analyzed by Denmark group[10,11] using a CoMFA-like approach. Instead of interaction energies between the studied molecule and probe, they introduced novel Average Steric Occupancy (ASO) descriptors assessing the occupancy of nodes for an ensemble of aligned catalyst conformers. The ASO descriptors displayed better performance compared to single conformer descriptors[10,11].

Besides the selection of relevant conformers, another important limitation of MIF is the conformers' alignment. If considered molecules share a common scaffold, alignment is a rather simple process whereas for structurally diverse data sets it becomes problematic. This motivated development of MIF-based alignment-independent descriptors - GRid INdependent Descriptors (GRIND)[12] resulted from the transformation of interaction fields with the help of the autocorrelation function. GRIND applications to asymmetric catalysis were first reported by Sciabola et al.[13] for asymmetric reactions previously studied in references [7,14] and [15]. In general, GRIND-based models performed similarly to MIF

alignment-dependent ones[13]. On the other hand, compared to the MIF approach, GRIND models are hardly interpretable, which may explain relatively rare applications of this method.

Asahara and Miyao[16] benchmarked different 2D (ECFP6 and Mol2vec) and 3D descriptors (Dragon and MOE) in the modeling of enantioselectivity of chiral Brønsted acid catalysts. The 3D descriptors generated for the lowest energy conformers of reactants, products, and catalysts performed worse than the ECFP6 descriptors. Sandfort et al.[17] achieved a reasonable accuracy of predictions using reactants and catalysts' multiple fingerprint features (MFFs) resulting from the concatenation of 24 fingerprint sets calculated with RDKit.

Tsuji et al. [18] reported a computer-aided design of new highly enantioselective imidodiphosphorimidate catalysts for tetrahydropyran (THP) and tetrahydropyran (THP) synthesis. They used ISIDA and CircuS descriptors which provided a wide range of alternative 2D structure representations, varying by both topology (linear or atom-centered fragments, atom pairs, triplets) and fragment size, allowing for fine-tuning to the particular modeling task. Such descriptors' tunability allowed us to reach a reasonable model's performance and to propose the catalysts more enantioselective than those used for the model training.

The above studies revealed three main drawbacks of existing 3D-QSSR approaches to the modelling of catalyst enantioselectivity: (i) selection of catalyst conformers, (ii) their alignment, and (iii) relevance of 3D descriptors with respect to the enantioselectivity problem. The modeling complexity raises when a set of catalysts is applied to a set of reactions because in that case, one needs to consider the entire reaction profile[10], i.e., a pair *catalyst(i)*reaction(j)*. For this purpose, Zahrt et al. [10] suggested calculating the ASO and electronic descriptors for reactants and products and concatenating them with the ASO and electronic descriptors of a catalyst. Notice that models obtained in the above workflow are difficult to reproduce because ASO descriptors depend on both the position of the reference structure and the alignment algorithm.

Recently, we have reported an alternative approach that used Multi-Instance Learning (MIL) algorithms [18] accounting for all low-energy conformers of catalysts encoded by alignment-independent 3D *pmapper* descriptors [19] in combination with the compact representation of chemical reactions by their Condensed Graph of Reaction [22]. This approach provided with models performance similar to Zahrt et al. [10]. Unfortunately, the *short communication* format of publication [19] did not allow us to describe the modeling workflow in detail.

In this study, we demonstrate that the MIL-based 3D modelling approach successfully predicts the enantioselectivity of homogeneous and phase-transfer reactions catalyzed by structurally different catalyst families. In both cases, the obtained models outperform traditional 2D models and previously

reported 3D state-of-the-art approaches. The suggested computational protocol is reproducible and publicly available at https://github.com/Laboratoire-de-Chemoinformatique/3D-MIL-QSSR

1. Data sets

Two data sets were considered in this study: (i) asymmetric addition of thiols to imines catalyzed by chiral phosphoric acid catalysts (PAC data set) [10] and (ii) asymmetric alkylation of glycine-derived Schiff bases catalyzed by cinchona alkaloid-based ammonium salts (APTC data set) [8]. We have used both structures and enantiomeric excess (*ee* %) values reported in the original publications.

For a given reaction, the enantiomeric excess (*ee* %) measuring catalyst enantioselectivity is defined as the difference between the amount of each enantiomer:

$$ee\ \% = \%R - \%S\ or\ ee\ \% = \%S - \%R \quad (1)$$

Predictive models were built for ΔΔG (kcal/mol) - a difference of activation free energies of competing reactions leading to different enantiomers:

$$\Delta\Delta G = -RTln\frac{[R]}{[S]} = RTln\frac{100 - ee\%}{100 + ee\%} \quad (2)$$

The phosphoric acid catalysts (PAC) dataset reported by Zahrt et al.[10] contains 43 catalysts used in 25 reactions of asymmetric addition of imine to thiol (**Figure 1a**) resulting in 43 × 25 = 1075 data points. Reported *ee* % values (in favor of R enantiomer) ranged from -34 to 99 and was converted to $\Delta\Delta G$ using eq. 2. A detailed description of the catalyst and reactants structures can be found in the original paper[10]. This data set was divided into training and several test sets, as suggested by Zahrt et al.[10] (**Figure S2** in SI). The training set consisted of 24 catalysts combined with 16 reactions resulting in 24 × 16 = 384 training catalyst/reaction pairs. Then, three test sets simulating different scenarios of the potential application of the models in real campaigns of catalyst design were prepared. The *reaction-out* test set containing 216 data points (24 training catalysts combined with 9 new reactions) was used to predict the enantioselectivity of new reactions with known (presented in the training set) catalysts. The *catalyst-out* test set containing 304 data points (19 new catalysts combined with 16 training reactions) examined the model potential to predict the enantioselectivity of known reactions with new catalysts. The *both-out* test set represents the most challenging scenario where the model was used to

predict the enantioselectivity of new reactants with new catalysts. This test set consisted of 171 data points corresponding to a combination of 19 test catalysts and 9 test reactions.

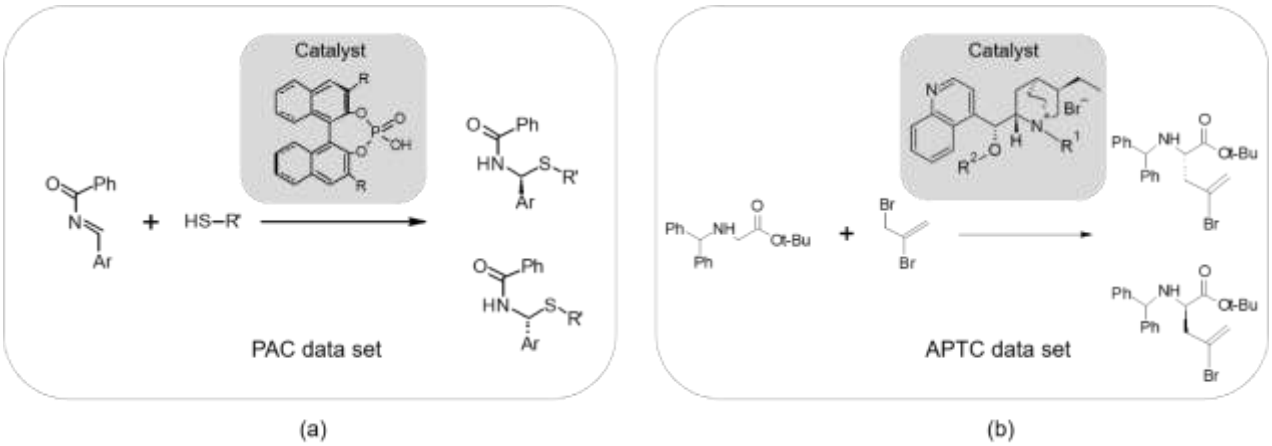


Figure 1. Data sets considered in this study: (a) asymmetric addition of thiols to imines catalyzed by chiral phosphoric acid catalysts (PAC data set) and (b) asymmetric alkylation of glycine-derived Schiff bases catalyzed by cinchona alkaloid-based ammonium salts (APTC data set).

Asymmetric phase transfer catalysis (APTC) enables reactions between reactants located in two immiscible phases with chiral catalysts to produce enantiopure substances. A classic example of APTC is the alkylation of α-amino acid derivatives catalyzed by cinchona alkaloid-based quaternary ammonium salts reported by Melville et al.[8] (**Figure 1b**). The COMFA model in publication [8] was built on 70 catalysts and validated on a set of 18 catalysts. The reported *ee* ranged from 16 to 93 % (in favor of the S enantiomer) and was converted to $\Delta\Delta G$ using eq. 2.

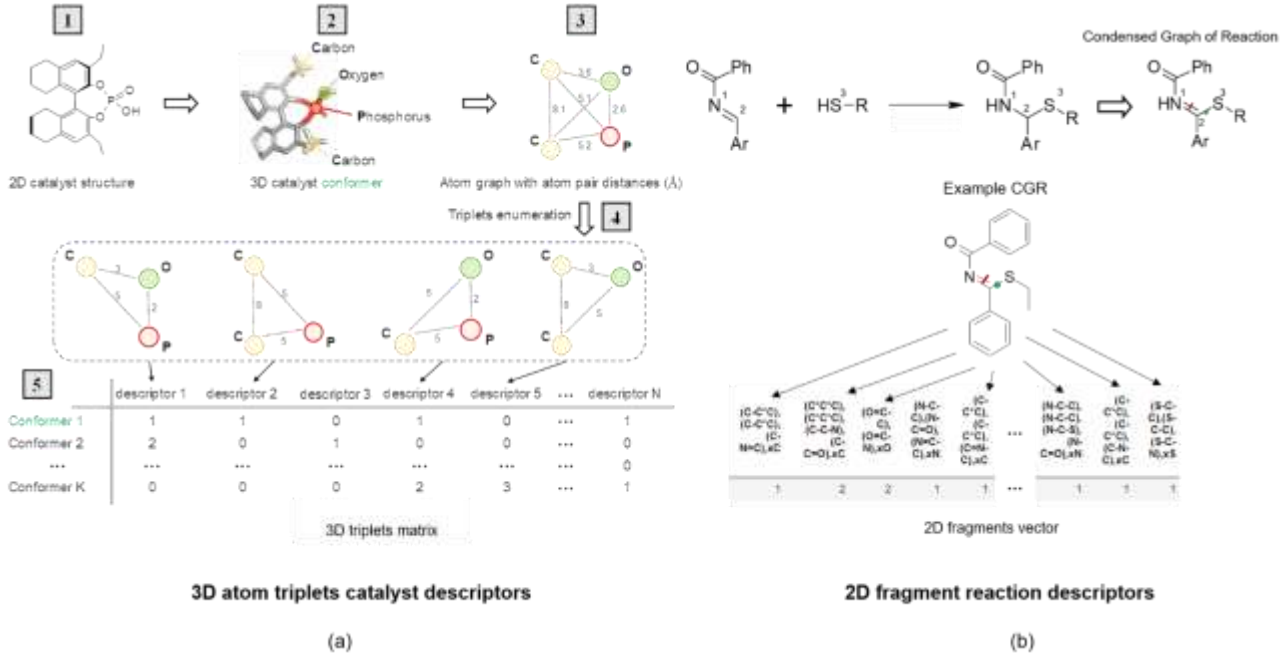


Figure 2. (a) Preparation of *pmapper* 3D descriptors for a given catalyst conformer: (1) given 2D catalyst structure; (2) generation of 3D catalyst conformer; (3) generation of a 3D fully connected graph of atoms (for demonstration, the graph of four atoms is chosen); (4) enumeration of all atom triplets; (5) counting of enumerated atom triplets in given conformer; (b) addition of thiols to imines and related Condensed Graph of Reaction (CGR) [22]. A CGR contains one created bond between the atoms S3 and C2 and one double bond transformed into a single one between the atoms N1 and C2.

2. Computational details

2.1 Reaction and catalyst descriptors

Catalyst conformers generation. Each catalyst was represented by an ensemble of conformers generated using the distance geometry algorithm implemented in the RDKit package[23]. If the RDKit algorithm failed to generate the conformers, we used a systematic conformer generator from the Open Babel package[24] and then recalculated the full energies of conformers using RDKit. For each catalyst, up to 50 conformers within an energy window of 50 kcal/mol were generated.

Catalyst 3D descriptors. Each generated catalyst conformer was encoded by *pmapper* descriptors[25] representing various combinations of 3D pharmacophore quadruplets. [20,21] In this work, instead of pharmacophore features used in our early study[19], we used quadruplets and triplets enumerating ensembles of individual atoms and/or centers of 5- and 6-membered aromatic rings. Notice that application of atoms triplets significantly reduces the number of descriptors, and related models perform similarly to those built on stereosensitive atoms quadruplets. However, if a data set contains catalysts in both R and S configurations - the application of atom quadruplets is mandatory to distinguish the two

enantiomers. In this study, atom triplets were applicable because all catalysts in the considered data sets had the same stereoconfiguration.

Atom triplets were specified by (*i*) the list of the individual atoms (C, N, O, S, P, F, Cl, Br, I) or centers of the 5-membered and 6-membered aromatic ring and (*ii*) the distances between atoms and/or center of rings in a triplet. The list of encoded atoms can be customized depending on the task. To enable fuzzy matching of atom triplets, the distances between atoms were binned with the step of 1Å (**Figure 2a**). Then the number of occurrences of each unique atom triplet is counted for each conformer, resulting in an integer descriptor matrix (**Figure 2a**).

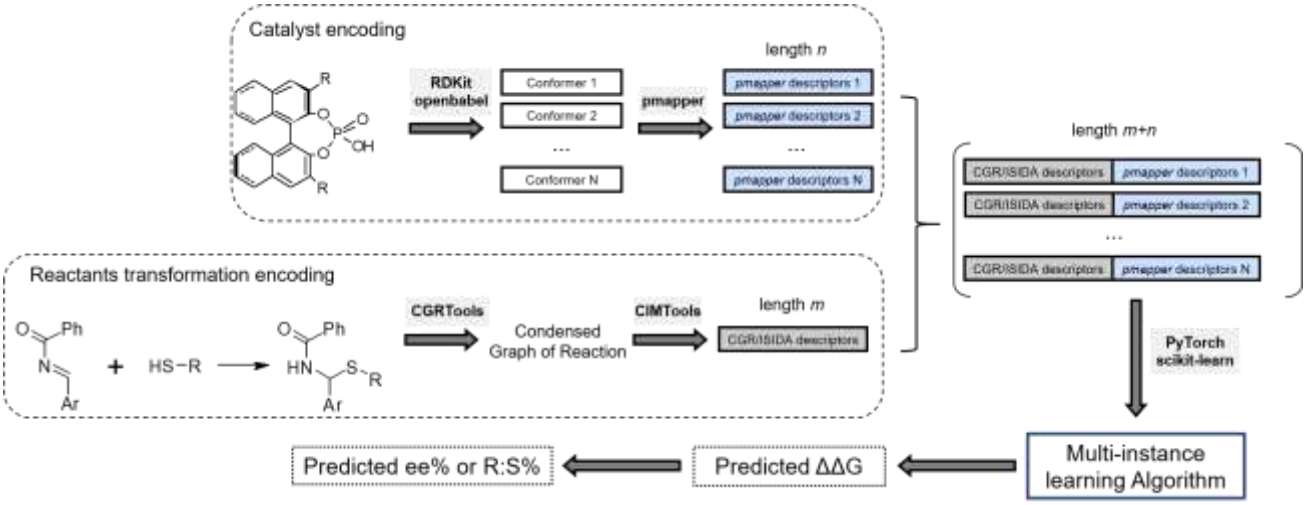


Figure 3. Modeling workflow used in this work. A chemical reaction is encoded by *m* CGR/ISIDA fragment descriptors. A catalyst is represented by its *N* conformers, each encoded by *n* 3D *pmapper* descriptors. The concatenation of *m* 2D reaction descriptors and *n* 3D catalyst descriptors results in a set of vectors of (*m* + *n*) size. The Python 3 libraries used in the modeling workflow are indicated in bold near the arrows.

2D descriptors for chemical reactions. Each chemical transformation was represented by a related Condensed Graph of Reaction (CGR)[22] using *CGRtools* and *CIMtools* package [26]. CGR represents a chemical reaction as a single molecular graph (**Figure 2b**) describing both conventional chemical bonds (e.g. single, double, triple, aromatic, etc.) and so-called "dynamic" bonds characterizing chemical transformations, i.e. breaking or forming a bond or changing bond order. Then, obtained CGRs were processed with ISIDA-Fragmentor tool [27] to generate 2D fragment descriptors. In each CGR, fragment descriptors count the occurrence of particular subgraphs (structure fragments). ISIDA-Fragmentor provides several strategies for molecule fragmentation. In this study, we used atom-centered

subgraphs (atoms with first, second, etc. coordination spheres) where the radius varied from 2 to 5 atoms. For CircuS descriptors we used atom-centered fragments with radii from 2 to 5.

Reaction profile descriptors. Vectors of 2D fragment descriptors for reactions and 3D atom triplets for catalysts were then concatenated to form reaction profile descriptor vectors (**Figure 3**). If the data set contained a single reactant transformation, there is no concatenation of catalyst and reactant descriptors.

Model availability. The final models were built using reaction profile descriptor vectors and our implementations of multi-instance learning algorithms (https://github.com/Laboratoire-de-Chemoinformatique/3D-MIL-QSSR/tree/main/miqssr/estimators). A Graphical User Interface (GUI) for predicting the catalyst enantioselectivity with the built MIC multiconformers models is available at https://chematlas.chimie.unistra.fr/Predictor/qscer.php.

2.1 Multi-instance learning algorithms

The building of 3D models with multiple catalyst conformers requires the application of special MIL algorithms. In MIL, any object (*i.e.*, a molecule) is represented by a bag of instances (*i.e.*, a set of conformers). All the considered MIL algorithms can be divided into two groups – *instance-based* and *bag-based*. *Instance-based* algorithms consider each conformer as a separate training instance. *Bag-based* algorithms, on the contrary, represent a molecule by a single vector of descriptors, which is produced from the vectors of conformer descriptors.

The learning process in *instance-based* algorithms occurs at the instance level. Instance-level learning is applicable if it is possible to assign a label to individual instances in a bag. Also, it is assumed that there is a rule that aggregates the predictions for each instance to get the prediction for the entire bag. The simplest *instance-based* machine learning (ML) algorithm is *Instance-Wrapper* (

Figure 4), where each training instance of a bag is assigned the same label as the whole bag. As a result, one obtains a data set where each conformer is an individual training object, and any conventional ML algorithms can be applied to build the model. Given a new catalyst, the enantioselectivity is predicted for each conformer, and predictions are averaged to obtain the final prediction of the enantioselectivity of the catalyst (

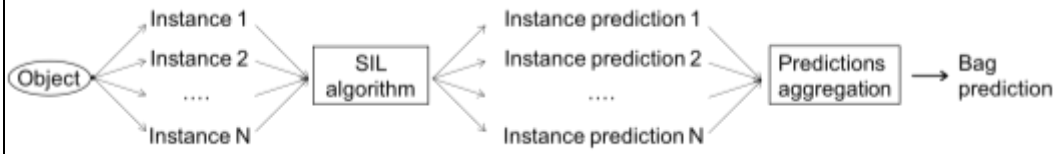
Figure 4). This approach can potentially bring some additional noise into the learning process because of assigning the same enantioselectivity to all catalyst conformers in a training set.

The learning process of *bag-based* algorithms occurs at the bag level. In *bag-based* algorithms, there is no need to identify a label for each instance in a bag. Instead, there is an operation that

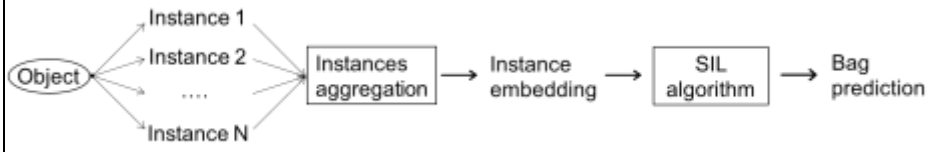
aggregates the instances to get a single vector representing the entire bag. Our implementation of the *Bag-Wrapper* algorithm (

Figure 4) averages the descriptor values across all conformers and supplies this single vector of descriptors to a conventional Single-Instance Learning (SIL) method - a three-layer fully connected neural network. The *Bag-Wrapper* algorithm has a similar drawback as the *Instance-Wrapper* – aggregation of the descriptor vectors of all conformers the resulting vector may be noised by the contribution of irrelevant conformers.

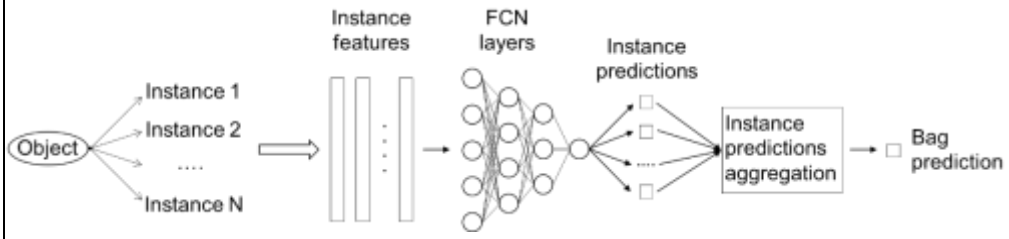
Instance Wrapper



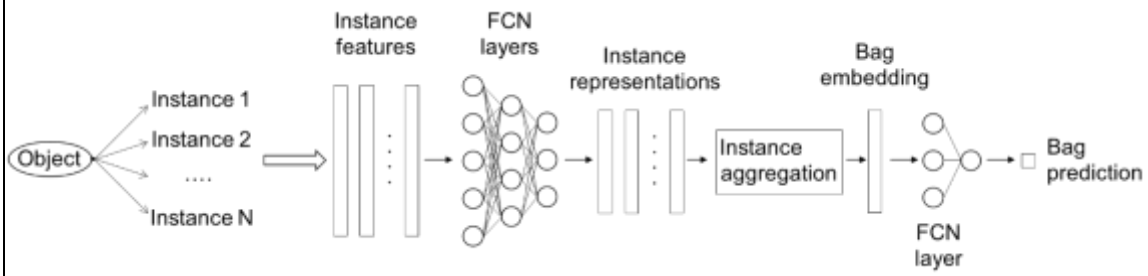
Bag Wrapper



Instance Net



Bag Net



Attention Net

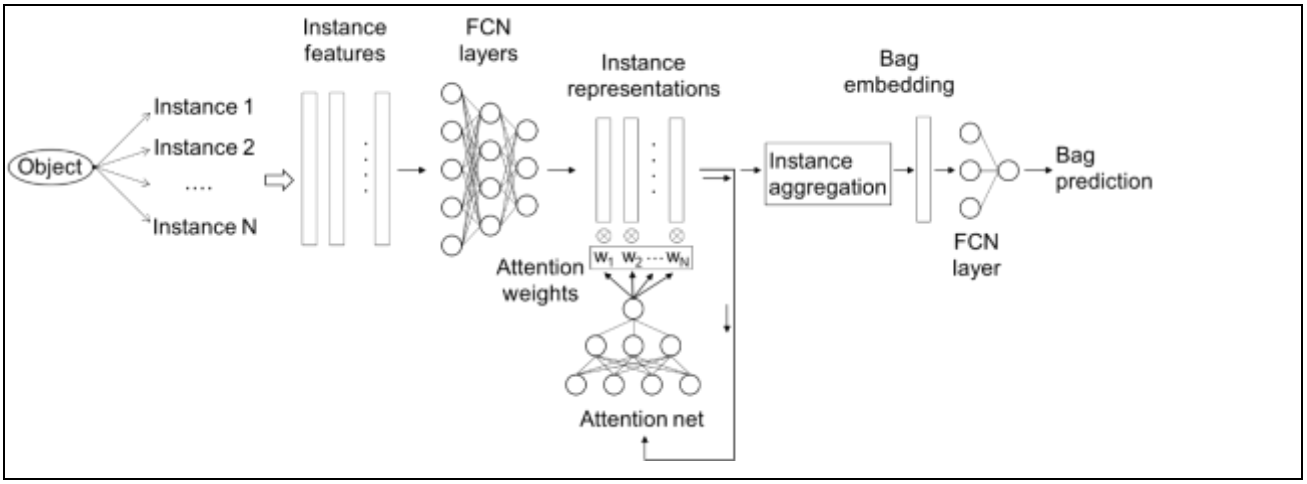


Figure 4. Multi-instance learning algorithms

There exist two types of MI neural networks: *Instance-Net* and *Bag-Net*. In *Instance-Net* (

Figure 4), the instances are running through fully connected layers and an output neuron. Then, instance predictions are averaged in the pooling layer to obtain a bag prediction, its error is calculated and backpropagated to adjust model weights. *Bag-Net* (

Figure 4) consists of three fully connected layers followed by one pooling layer. The pooling layer averages the instance representations learned by previous layers into a single embedding vector as a bag representation. The last fully connected layer takes the embedding vector as input and outputs the bag prediction.

The *Bag-Net* uses an unlearnable mean pooling function and, therefore, the irrelevant conformers can contribute noise to the prediction and reduce model performance. This drawback can be eliminated by using more flexible types of pooling, such as weighted averaging pooling, known as attention. This type of pooling was proposed in publication [28], where an additional two-layered neural network was used to obtain the weights of instances. In the *Bag-AttentionNet* (

Figure 4), all instances are first fed to three fully connected layers. Then, the learned instance representations are used by the attention network with a single hidden layer. In the attention network, the number of output neurons is equal to the number of instances. The output layer of attention has the Softmax activation function and predicts instance weights. Finally, the instance weights given by the attention network are used for weighted averaging of instance representations to get the embedding vector that is used to produce the bag prediction. Implementation of weighted pooling enables the *Bag-AttentionNet* to automatically identify probable reactive conformers.

2.2 Generation of 2D models

As an alternative to the MIL-3D approach, we also considered 2D models where the reactants and catalyst structures were encoded by different fingerprints and fragment 2D descriptors. The following fingerprints were generated using the RDKit library: Atom-Pairs (1024 bits)[29], Avalon (1024 bits)[30], and Morgan fingerprints of radius 2 (1024 bits)[31]. Fragment ISIDA[27] and CircuS[18] descriptors can be calculated with different fragmentation strategies. For ISIDA, both atom-centered and linear fragments were used. CircuS are similar to ISIDA atom-centered fragments, but explicitly consider encountered branching or cyclical structures, which makes them more efficient for catalyst structures enriched with cyclical groups and reduces the noise in the training data.

For a PAC dataset containing multiple reactant transformations, there were two encoding strategies: (a) reactant transformations were converted to CGR and then encoded by ISIDA or CircuS fragment descriptors (Imine/Thiol CGR, **Table 1**) or (b) imine and thiol were encoded by fingerprints or fragment descriptors and then concatenated to a single descriptor vector (Imine/Thiol concatenation, **Table 1**). Then the resulting reactant transformation vectors were concatenated with fingerprint or fragment descriptor vectors of the catalysts.

Fragment-based descriptors can be calculated using different strategies and fragment lengths, generating multiple sets of descriptors. In order not to be biased towards specific descriptor sets, we applied a consensus method to calculate the final predictions. First, for each descriptor type (ISIDA, CircuS, or fingerprints), we selected models with determination coefficient $R^2_{Train} > 0.7$ to discard descriptor sets that poorly describe the training set. Then the predictions of the filtered models for the test set were averaged to obtain final consensus predictions of enantioselectivity. For model training, the same fully connected neural network was used as in the *Instance-Wrapper* algorithm in multi-instance models.

The following metrics were used to assess the performance of the models: Root-Mean Squared Error (RMSE), Mean Absolute Error (MAE), determination coefficient (R^2), Spearman correlation coefficient measuring the correlation between predicted and experimental catalyst ranks (ranking accuracy, RA).

2.3 Quantile loss function

A new round of catalyst screening is expected to reveal more efficient catalysts. In this context, it is desirable to prevent under-estimation of enantioselectivity compared to its actual value. Incorrect behavior of the model in these cases can lead to ignorance of most perspective structures, which may not be chosen for experimental testing at the next rounds of screening. Thus, the predictive model should

be specially configured to avoid under-estimation ($y^{pred} < y^{obs}$) of enantioselectivity. This can be provided with a help of a special quantile loss function for the model training:

$$L = max[q \times (y^{pred} - y^{obs}), (q-1) \times (y^{pred} - y^{obs})] \quad (3)$$

Quantile loss function (3) asymmetrically penalizes overestimation ($y^{pred} > y^{obs}$) and underestimation ($y^{pred} < y^{obs}$). For q = 0.5, they both are penalized equally. The lower the value of $q < 0.5$, the more underestimation is penalized compared to overestimation. In this study, q was fixed at 0.1 which means that overestimation is penalized by a factor of 0.1, and underestimation by a factor of 0.9, and, thus, the model tries to avoid underestimation.

3. Results and Discussion

Using the described data sets and modelling protocols, various 2D and 3D models for enantioselectivity prediction were generated. The 3D single-conformer model was built on the lowest-energy catalyst conformers, while the 3D multi-conformer model included all the generated conformers.

3.1 Benchmarking of molecular descriptors and MIL algorithms

MIL algorithms benchmarking. For the benchmark of five MIL algorithms, we used the *pmapper* descriptors and the PAC data set, which was divided into 25 subsets according to the number of reactant transformations. Each subset contained 43 catalysts with experimental ΔΔG measured in a given reaction. The mean Absolute Error (MAE) of ΔΔG predictions was evaluated in a 5-fold cross-validation repeated 5 times (5×5-CV). The following values of median MAE (in kcal/mol) over 25 reactions (5×5-CV) were obtained: *Instance-Wrapper* (MAE = 0.28 kcal/mol), *Bag-Wrapper* (0.31 kcal/mol), *Instance-Net* (0.31 kcal/mol), *Bag-Net* (0.32 kcal/mol) and *BagAttention-Net* (0.35 kcal/mol). Based on these results, *Instance-Wrapper* was chosen as the main algorithm for further experiments.

The basic machine learning algorithm in *Instance-Wrapper* represented a fully connected neural network with three hidden layers of 256, 128, and 64 neurons and a ReLU activation function. The optimized hyperparameters were weight decay (0.0001, 0.001, 0.01, 0.1) and learning rate (0.001 or 0.01). The maximum number of learning epochs was 1000.

Descriptors benchmarking. Different popular descriptors were benchmarked on the same data sets as the MIL algorithms (see above). Namely, we considered ISIDA[27] and CircuS[18] fragment descriptors, 2D fingerprints, and 3D descriptors available in RDKit, as well as *pmapper* 3D atom triplets and quadruplets descriptors. A set of 3D RDKit descriptors RDF, MoRSE, WHIM, GETAWAY, and

AutoCorr3D descriptors. 3D descriptors were benchmarked in a multi-instance setting, i.e., the *pmapper* and *RDKit* 3D descriptors were generated for multiple conformers. The Instance-Wrapper MIL algorithm was used as a machine learning method to build 3D models. In the case of 2D descriptors, the MIL bag contained only one instance. The performance of the obtained models was compared to that of the baseline null model which predicts enantioselectivity always as an average value of the training experimental enantioselectivities corresponding to median MAE = 0.47 kcal/mol

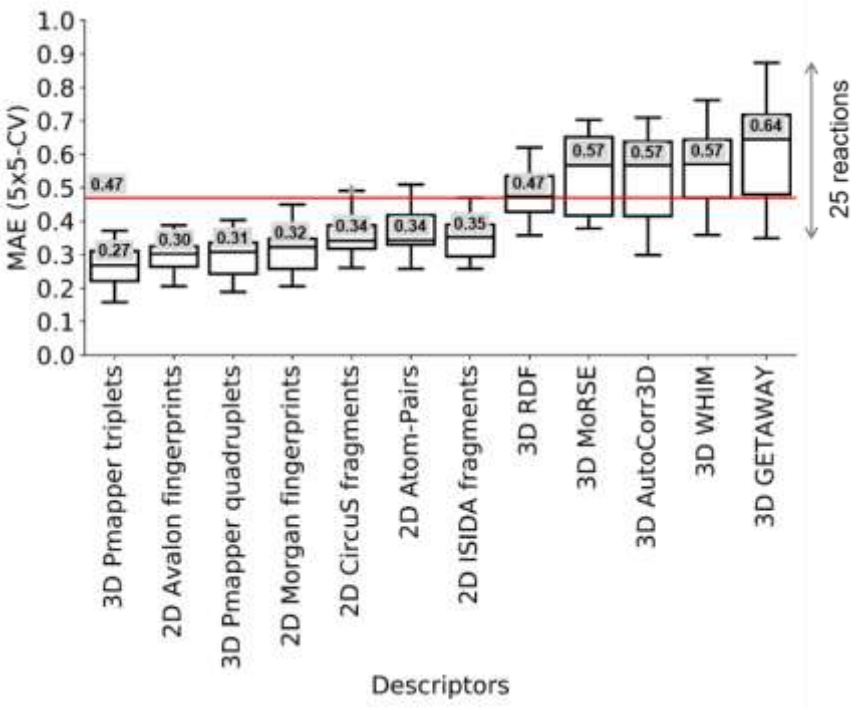


Figure 5. Performance of ΔΔG (kcal/mol) models involving 2D and 3D catalyst descriptors. Each boxplot describes a 5×5-CV-validated MAE for 43 catalysts obtained with 25 models, each corresponding to an individual chemical reaction. The red horizontal line shows the accuracy of the *null* model, which constantly predicts ΔΔG as the average experimental ΔΔG across all catalysts.

The benchmarking results show that the *pmapper* 3D triplets (median MAE_{CV} = 0.27 kcal/mol) performed better than *pmapper* 3D quadruplets (0.31 kcal/mol) and all studied 2D descriptors (0.30-0.35 kcal/mol) (**Figure 5**). The 3D RDKit descriptors were found unsuitable for modelling the catalyst enantioselectivity because they performed worse than the baseline null model. Thus, the proposed 3D atom triplets demonstrated the best performance in combination with the Instance-Wrapper MIL algorithm although their number for a set of 43 catalysts (2886) was significantly smaller than that of atom quadruplets (42824).

3.2 Asymmetric addition of thiols to imines

Table 1 reports the performance of Instance-Wrapper MIL models on three test sets described in Section 1. These results were compared with those early reported by Sandfort et al.[17], Zahrt et al. [10], and Asahara et al.[16]

In the *reaction-out test set*, all generated 2D and 3D models demonstrated good results. The 2D models predict enantioselectivity with MAE = 0.14-0.18 kcal/mol, which is even better than the 3D single-conformer model (0.21 kcal/mol). Consideration of multiple conformers significantly increases the prediction accuracy (0.13 kcal/mol).

Table 1. Mean Absolute Error (MAE, kcal/mol) of ΔΔG predictions obtained for test sets generated from phosphoric acid catalysts (PAC) data set.

Reactions representation	**Catalyst representation**	**Reaction-out**	**Catalyst-out**	**Both-out**
Imine/Thiol concatenation	2D Morgan fingerprints	0.18	0.29	0.33
	2D Avalon fingerprints	0.15	0.26	0.28
	2D Atom-Pairs fingerprints	0.16	0.36	0.33
	2D ISIDA fragments	0.14	0.27	0.28
	2D CircuS fragments	0.14	0.31	0.33
Imine/Thiol CGR	2D ISIDA fragmnets	0.15	0.27	0.30
	2D CircuS fragments	0.14	0.32	0.34
	3D Atom triplets (single conformer)	0.21	0.38	0.48
	3D Atom triplets (multiple conformers)	**0.13**	0.22	**0.21**
Alternative approaches	2D Sandfort's MFFs fingerprints [a]	0.14	0.25	0.28
	2D Mol2vec descriptors [b]	**0.13**	0.34	0.40
	2D ECFP6 descriptors [b]	0.14	0.22	**0.21**
	3D Dragon descriptors (single conformer) [b]	0.14	0.42	0.47
	3D MOE descriptors (single conformers) [b]	0.15	0.48	0.55
	3D ASO descriptors (multiple conformers) [c]	0.16	**0.21**	0.24

[a] 2D modelling approach published by Sandfort et al.[17], [b] 2D and 3D models published by Asahara and Miyao[16], and [c] 3D conformer-dependent approach published by Zahrt et al[10]. R^2 and Ranking Accuracy (RA) are reported in Table S1 and Table S2 in Supporting Information. In "Alternative approaches" reaction transformations were encoded with the same type of descriptors used for Catalyst representation.

On the *catalyst-out* test set the 3D multi-conformer model also performs significantly better (0.22 kcal/mol) than the 3D single-conformer model (0.38 kcal/mol) and 2D models (0.26-0.36 kcal/mol). A similar trend was observed for the *both-out* test set the 3D multi-conformer model (0.21 kcal/mol) outperformed the 3D single-conformer model (0.48 kcal/mol) and 2D models (0.28-0.34 kcal/mol).

Our best 3D multi-conformer models performed similarly to the best previously reported models by Asahara et al.[16] based on ECFP6 descriptors and by Zahrt et al. [10] based on ASO descriptors (**Table 1**).

3.3 Enantioselectivity prediction beyond the training set

To examine the potential of the models to predict enantioselectivity values beyond the training set, we followed the validation strategy proposed by Zahrt et al[10]. Following the above publication, the PAC data set on 1075 reactions was divided into a training set of 718 reactions with *ee* < 80% and a test set of highly selective 357 reactions with *ee* > 80%. Then, the 2D and 3D MIL models were built applying both conventional mean squared error loss (MSE) and suggested here quantile loss (see Computational details section).

All 2D models built with MSE loss failed to predict enantioselectivity beyond the training set (R^2_{Test} < 0), while the 3D single-conformer model (R^2_{Test} = 0.36) and 3D multi-conformer model (R^2_{Test} = 0.44) performs significantly better. On the other hand, training with the quantile loss function considerably improved both the 3D single-conformer model (R^2_{Test} = 0.59) and the 3D multi-conformer model (R^2_{Test} = 0.74). The 2D models built with the quantile loss function were still worse than the null model (R^2_{Test} < 0) (**Figure 6**). Notice that on the *beyond training* test set the proposed 3D MIL multi-conformer model trained with the quantile loss (MAE_{Test} = 0.19 kcal/mol) outperformed Zahrt et al. approach (0.33 kcal/mol)[10].

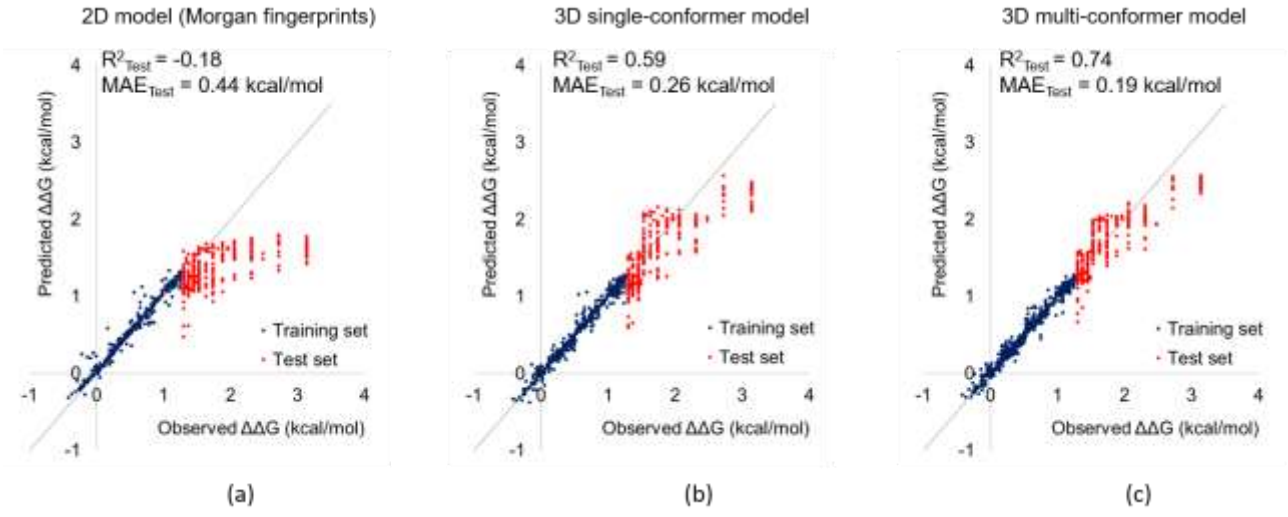


Figure 6. Predicted and observed catalyst enantioselectivity (ΔΔG, kcal/mol) for (a) 2D model, (b) 3D single-conformer model, and (c) 3D multi-conformer model trained with quantile loss. The training set included reactions with *ee* < 80% and the test set with *ee* >= 80%.

To summarize, for the PAC dataset, the 3D multi-conformer model always outperforms the 3D single-conformer models, especially in the prediction of enantioselectivity for new test catalysts, which proves the importance of accounting for conformational flexibility. We believe that the difference in the performance of 3D single-conformer and 3D multi-conformer models may still increase with the increasing flexibility of modeled catalysts. The 3D multi-conformer model outperforms the 2D models, generated with popular fingerprints and fragment descriptors, which highlights the importance of 3D information in enantioselectivity modelling.

It should also be noted that in a computational screening of candidate catalysts, the predictive model should effectively identify potentially highly selective catalysts, i.e. the model should rank them higher than the other candidates. The *ranking accuracy* (RA) estimated by the Spearman ranking correlation coefficient provided in Table S2 shows that despite large prediction error (MAE) the 2D models achieve high RA > 0.80, *i.e*., they reasonably well capture the general trend in enantioselectivity variation.

3.4 Asymmetric phase transfer catalysis

For asymmetric alkylation (APTC data set), Melville *et al*[8] reported a CoMFA model for *ee* built on the training set containing 70 catalysts. This model was validated on a test set of 18 catalysts with RMSE = 13.4 %. In our calculation, the original enantioselectivities were converted to ΔΔG, then the predictions on the test set were converted to *ee* % to be compared with the Melville *et al*[8] results. Our 3D multi-conformer model (RMSE = 8.8%) performed significantly better than the related 3D single-conformer (18.0%) and the original COMFA model. The significant difference in the performances of 3D single- and multi-conformer models can be explained by the high conformation flexibility of the catalysts – the average number of rotatable bonds in the data set was 10. The 2D models built on ISIDA and CircuS descriptors demonstrated poor performance with RMSE of 15.6 and 18.5 %, respectively (**Figure 7**).

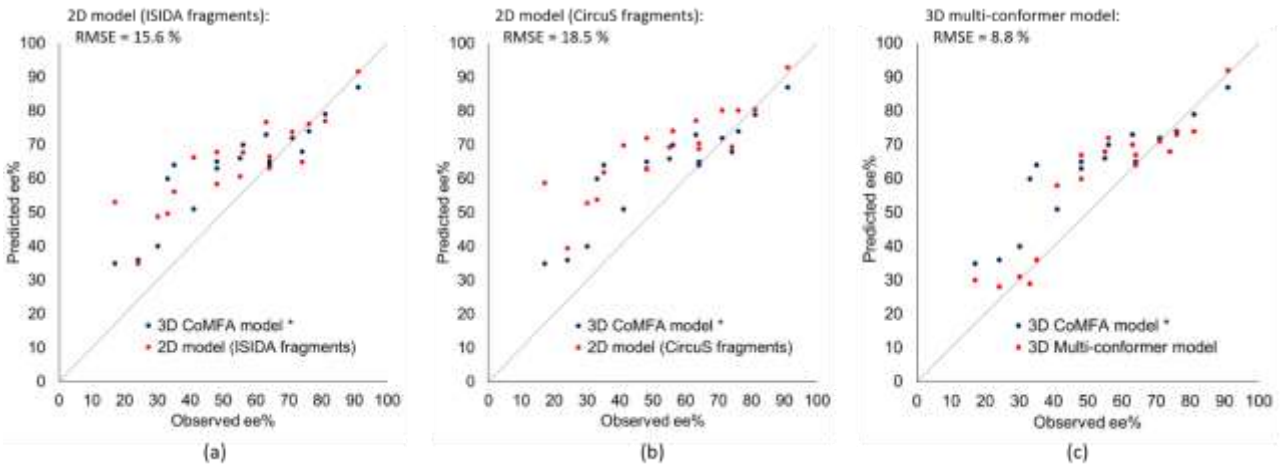


Figure 7. Observed and predicted *ee* % for 18 test catalysts from the APTC dataset comparing the performance of the 3D-CoMFA model by Melville et al[8] with: (a) 2D model (ISIDA fragments), (b) model (CircuS fragments), and (c) 3D multi-conformer model (atom triplets).

Conclusions

It has been demonstrated that the multi-instance learning approach can successfully be used to model the enantioselectivity of chiral molecules catalyzing a particular chemical reaction. Concatenation of 3D *pmapper* descriptors representing conformers of catalysts and 2D fragmental descriptors representing transformations resulted in highly predictive models. The approach was applied to two different catalyst systems (BINOL derivatives and ammonium salts) with several validation scenarios. The developed models performed similarly or better than single-instance models based on popular fingerprints or the state-of-the-art 2D or 3D descriptors. Our approach demonstrated substantial specific advantages in cases where catalysts were represented by very flexible molecules because 3D *pmapper* descriptors do not require spatial alignment of conformers which reduces ambiguity in catalyst representation and descriptor calculation. We demonstrated the importance of accounting for multiple conformers of a catalyst rather than its single conformer. In fact, selection of a chemically relevant conformer of a catalyst is a non-trivial task and may be a reason for degradation of the model's performance. Another improvement in the prediction of enantioselectivity can be achieved by applying a quintile loss function to penalize underestimated predictions. This greatly improves the extrapolation ability of models.

The developed modelling protocol is automatized and reproducible. The proposed *pmapper* 3D descriptors for the catalyst and ISIDA/CGR descriptors for the chemical reaction are easily customizable.

Funding Information

PP acknowledges the support of the European Regional Development Fund (Project ENOCH no. CZ.02.1.01/0.0/0.0/16_019/0000868) and CZ-OPENSCREEN.

Data availability.

The Python 3 source code for model building is available at https://github.com/Laboratoire-de-Chemoinformatique/3D-MIL-QSSR. A Graphical User Interface (GUI) for predicting the catalyst enantioselectivity with the developed MIC multi-conformers models is available at https://chemat-las.chimie.unistra.fr/Predictor/qscer.php.

References

(1) Jen, W. S.; Wiener, J. J. M.; MacMillan, D. W. C. New Strategies for Organic Catalysis: The First Enantioselective Organocatalytic 1,3-Dipolar Cycloaddition [20]. *J. Am. Chem. Soc.* **2000**, *122* (40), 9874–9875. https://doi.org/10.1021/ja005517p.

(2) List, B.; Lerner, R. A.; Barbas, C. F. Proline-Catalyzed Direct Asymmetric Aldol Reactions [13]. *J. Am. Chem. Soc.* **2000**, *122* (10), 2395–2396. https://doi.org/10.1021/ja994280y.

(3) Han, B.; He, X. H.; Liu, Y. Q.; He, G.; Peng, C.; Li, J. L. Asymmetric Organocatalysis: An Enabling Technology for Medicinal Chemistry. *Chem. Soc. Rev.* **2021**, *50* (3), 1522–1586. https://doi.org/10.1039/d0cs00196a.

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

(5) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857. https://doi.org/10.1021/jm00145a002.

(6) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967. https://doi.org/10.1021/ja00226a005.

(7) Lipkowitz, K. B.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68* (12), 4648–4656. https://doi.org/10.1021/jo0267697.

(8) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, *4* (12), 1410–1411. https://doi.org/10.1039/b402378a.

(9) Melville, J. L.; Lovelock, K. R. J.; Wilson, C.; Allbutt, B.; Burke, E. K.; Lygo, B.; Hirst, J. D. Exploring Phase-Transfer Catalysis with Molecular Dynamics and 3D/4D Quantitative Structure - Selectivity Relationships. *J. Chem. Inf. Model.* **2005**, *45* (4), 971–981. https://doi.org/10.1021/ci0500511.

(10) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science (80-.).* **2019**, *363* (6424). https://doi.org/10.1126/science.aau5631.

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

(12) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43* (17), 3233–3243. https://doi.org/10.1021/jm000941m.
(13) Sciabola, S.; Alex, A.; Higginson, P. D.; Mitchell, J. C.; Snowden, M. J.; Morao, I. Theoretical Prediction of the Enantiomeric Excess in Asymmetric Catalysis. An Alignment-Independent Molecular Interaction Field Based Approach. *J. Org. Chem.* **2005**, *70* (22), 9025–9027. https://doi.org/10.1021/jo051496b.
(14) Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β-Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* **2003**, *125* (22), 6614–6615. https://doi.org/10.1021/ja0293195.
(15) Hoogenraad, M.; Klaus, G. M.; Elders, N.; Hooijschuur, S. M.; McKay, B.; Smith, A. A.; Damen, E. W. P. Oxazaborolidine Mediated Asymmetric Ketone Reduction: Prediction of Enantiomeric Excess Based on Catalyst Structure. *Tetrahedron Asymmetry* **2004**, *15* (3), 519–523. https://doi.org/10.1016/j.tetasy.2003.12.013.
(16) Asahara, R.; Miyao, T. Extended Connectivity Fingerprints as a Chemical Reaction Representation for Enantioselective Organophosphorus-Catalyzed Asymmetric Reaction Prediction. *ACS Omega* **2022**, *7* (30), 26952–26964. https://doi.org/10.1021/acsomega.2c03812.
(17) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.
(18) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. **2022**.
(19) Zankov, D.; Polishchuk, P.; Madzhidov, T.; Varnek, A. Multi-Instance Learning Approach to Predictive Modeling of Catalysts Enantioselectivity. *Synlett* **2021**, *32* (18), 1833–1836. https://doi.org/10.1055/a-1553-0427.
(20) Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I. QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *J. Chem. Inf. Model.* **2021**, *61* (10), 4913–4923. https://doi.org/10.1021/acs.jcim.1c00692.
(21) Nikonenko, A.; Zankov, D.; Baskin, I.; Madzhidov, T.; Polishchuk, P. Multiple Conformer Descriptors for QSAR Modeling. *Mol. Inform.* **2021**, *40* (11), minf.202060030. https://doi.org/10.1002/minf.202060030.
(22) Hoonakker, F.; Lachiche, N.; Varnek, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20* (2), 253–270.
(23) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574. https://doi.org/10.1021/acs.jcim.5b00654.
(24) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2* (1), 1–7. https://doi.org/10.1186/1752-153X-2-5.
(25) Kutlushina, A.; Khakimova, A.; Madzhidov, T.; Polishchuk, P. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **2018**, *23* (12), 3094. https://doi.org/10.3390/molecules23123094.
(26) Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **2019**, *59* (6), 2516–2521. https://doi.org/10.1021/acs.jcim.9b00102.

(27) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aided. Mol. Des.* **2005**, *19* (9–10), 693–703. https://doi.org/10.1007/s10822-005-9008-0.
(28) Ilse, M.; Tomczak, J. M.; Welling, M. Attention-Based Deep Multiple Instance Learning. *35th Int. Conf. Mach. Learn. ICML 2018* **2018**, *5*, 3376–3391.
(29) Smith, D. H.; Carhart, R. E.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64–73. https://doi.org/10.1021/ci00046a002.
(30) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46* (5), 1924–1936. https://doi.org/10.1021/ci050413p.
(31) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

Multi-Instance Learning Approach to the Modeling of Enantioselectivity of Conformationally Flexible Organic Catalysts

D. Zankov[1], T. Madzhidov[2], P. Polishchuk[3], P. Sidorov[4] and A. Varnek[1,4*]

[1] Laboratory of Chemoinformatics, University of Strasbourg, France

[2] Chemistry Solutions, Elsevier Ltd, Oxford, United Kingdom

[3] Institute of Molecular and Translational Medicine, Palacký University, Czech Republic

[4] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Japan

Supporting information

1. Conformer generation

The catalyst conformers were generated using the distance geometry algorithm implemented in RDKit. In the default configuration of the modelling protocol, we generated a maximum of 50 conformers within an energy window of 50 kcal. We also tested other values of the maximum number of conformers per catalyst and the energy window. For the experiment, we chose a data set on the reaction of asymmetric alkylation of α-amino acid derivatives. The data set contains 88 cinchona-based catalysts. This data set was chosen because it contains flexible catalysts with an average number of rotatable bonds of 10.1.

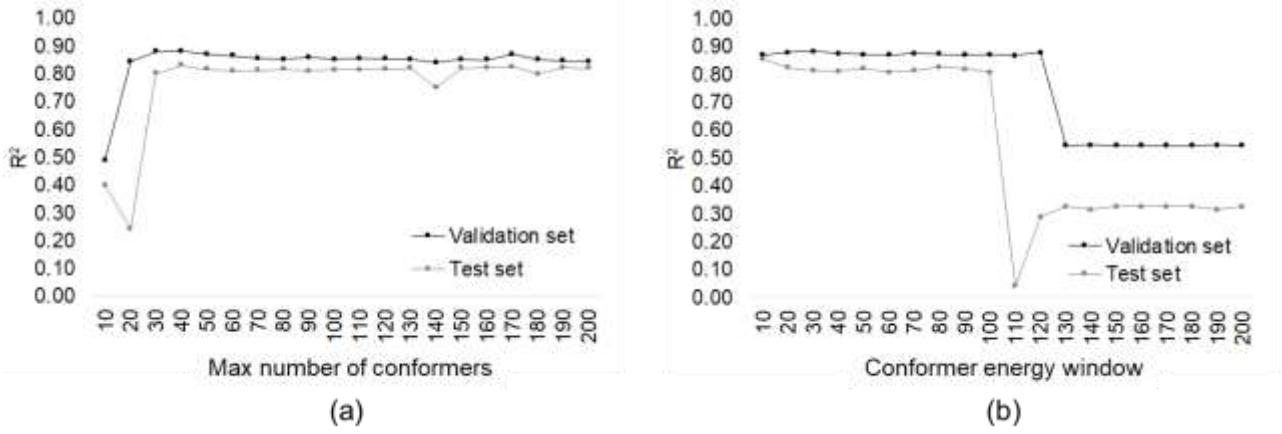


Figure S1. Prediction accuracy of catalyst enantioselectivity on the validation and test set vs. the maximum number of conformers (a) and the size of the energy window (b).

In the first experiment (**Figure S1a**), we fixed the energy window at 50 kcal and varied the maximum number of conformers from 10 to 200 with the step of 10. The results show that the prediction accuracy on the validation set increases rapidly from 10 to 20 conformers and reaches a plateau of 40 conformers.

In the second experiment (**Figure S1b**), we fixed the maximum number of conformers at 50 and varied the energy window values from 10 to 200 kcal with a step of 10 kcal. We observed that the prediction accuracy on the validation set remains constant in the range from 10 to 100 kcal, but then decreases dramatically.

2. Test and training set design

Recently, Denmark and co-workers published a data set on the enantioselectivity of phosphoric acid catalysts (PAC) for the reaction of the asymmetric addition of thiols to imines. This data set reports the enantioselectivity of 43 catalysts with 25 imine and thiol reactant combinations resulting in 43 × 25 = 1075 data points. Hereinafter we refer to this data set as the «PAC data set». Reported *ee* % (in favor of R enantiomer) ranged from -34 to 99 and for modelling were converted to ΔΔG (kcal/mol). A detailed description of the catalyst and reactant structures can be found in the original paper[11].

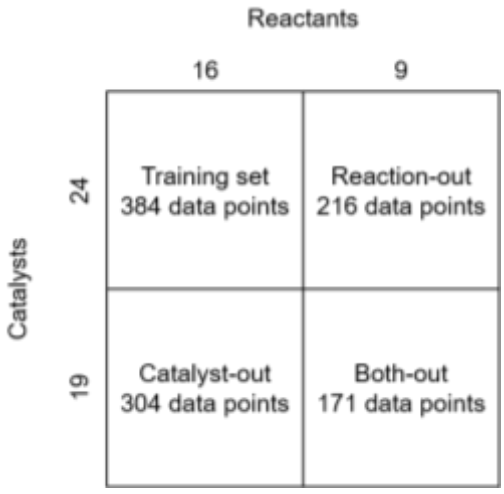


Figure S2. Training and test sets generated from phosphoric acid catalysts (PAC) data set.

Each combination of 43 catalysts and 25 reactants formed an individual reaction profile. The concatenation of reactants and catalyst descriptors in the training process produces models that can be utilized in different scenarios for the prediction of (a) enantioselectivity of known reactions with new

catalysts, (b) enantioselectivity of new reactions with known catalysts, and (c) enantioselectivity of new reactions with new catalysts.

This data set was divided into training and test set, exactly as in the original paper[11]. The training set consisted of 24 catalysts combined with 16 reactants resulting in 24 × 16 = 384 training reactions. Then, three test sets simulating different scenarios of the potential application of the models in real campaigns of catalysts design were prepared (**Figure S2**). The *reaction-out test set* simulates a scenario where the generated model is used to predict the enantioselectivity of new reactants with known (presented in the training set) catalysts. To this end, 24 training catalysts were combined with 9 new test reactants resulting in 24 × 9 = 216 test data points. The *catalyst-out test set* examines the potential of the model to predict the enantioselectivity of known reactants with new catalysts. For this purpose, 16 training reactants were combined with 19 new catalysts for a total of 16 × 19 = 304 test data points. The *both-out test set* represents the most challenging scenario where the model is used to predict the enantioselectivity of new reactants with new catalysts. This test set consists of 9 test reactants combined with 19 test catalysts providing 9 × 19 = 171 test data points.

3. Models' performance

Table S1 Coefficient of determination (R^2) of ΔΔG predictions obtained for test sets generated from phosphoric acid catalysts (PAC) data set.

Reactants representation	**Model (descriptors)**	**Reaction-out**	**Catalyst-out**	**Both-out**
Imine/Thiol concatenation	2D model (Morgan fingerprints)	0.86	0.63	0.53
	2D model (Avalon fingerprints)	0.92	0.73	0.71
	2D model (Atom-Pairs fingerprints)	0.90	0.41	0.61
	2D model (ISIDA fragments)	0.92	0.69	0.66
	2D model (CircuS fragments)	0.92	0.66	0.61
Imine/Thiol CGR	2D model (ISIDA fragments)	0.91	0.68	0.61
	2D model (CircuS fragments)	0.92	0.62	0.57
	3D single-conformer model (Atom triplets)	0.84	0.55	0.38
	3D multi-conformer model (Atom triplets)	0.93	0.83	0.86

Table S2 Ranking accuracy (RA) of ΔΔG predictions obtained for test sets generated from phosphoric acid catalysts (PAC) data set.

Reactants representation	**Model (descriptors)**	**Reaction-out**	**Catalyst-out**	**Both-out**
Imine/Thiol concatenation	2D model (Morgan fingerprints)	0.92	0.85	0.82
	2D model (Avalon fingerprints)	0.94	0.85	0.83
	2D model (Atom-Pairs fingerprints)	0.94	0.63	0.58
	2D model (ISIDA fragments)	0.94	0.85	0.88
	2D model (CircuS fragments)	0.94	0.83	0.82
Imine/Thiol CGR	2D model (ISIDA fragments)	0.94	0.86	0.88
	2D model (CircuS fragments)	0.94	0.82	0.81
	3D single-conformer model (Atom triplets)	0.91	0.76	0.71
	3D multi-conformer model (Atom triplets)	0.94	0.90	0.90

Table S3 Determination coefficient (R^2) and Mean Absolute Error (MAE) of ΔΔG predictions obtained for test sets for APTC data set.

Model (descriptors)	**RMSE (ee %)**	**R^2 (ΔΔG)**	**MAE (ΔΔG)**
2D model (ISIDA fragments)	15.6	0.63	0.20
2D model (CircuS fragments)	18.5	0.38	0.28
3D single-conformer model (Atom triplets)	18.0	0.42	0.24
3D CoMFA model	13.4	0.68	0.19
3D multi-conformer model (Atom triplets)	8.8	0.82	0.13