This article can be cited before page numbers have been issued, to do this please use: A. Hoque, N. Jain, D. R. Chenna and R. B. Sunoj, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00483G.

View Article Online
DOI: 10.1039/D5DD00483G

Machine Learning Models for Catalytic Asymmetric Reactions of Simple Alkenes: From Enantioselectivity Predictions to Chemical Insights

Ajnabiul Hoque,[a] Nupur Jain,[a] Divya Chenna,[a] and Raghavan B. Sunoj[a,b,*]

[a]Department of Chemistry, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

[b]Centre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

E-mail: sunoj@chem.iitb.ac.in

Abstract: Increasing number of applications of machine learning (ML) in chemical catalysis has engendered considerable confidence in predicting reaction outcomes. Despite the successful applications of ML to high-throughput experimentation (HTE) datasets, extension to small real-world datasets prevalent in organic synthesis remained more difficult, primarily due to their imbalanced and sparse distribution. Herein, we present a new chemical reaction dataset curated from published literature that bears class imbalance (CI) with a skewness of –1.37. The reactions in focus belong to an important class of transition metal-catalysed asymmetric transformation of alkens such as cyclopropanation, aziridination, and arylation. Such reactions are indispensable for the construction of three-membered structural motifs, a versatile building block found in complex bioactive molecules. In cognizance of the CI in the reaction outcome, measured in terms of enantiomeric excess (%*ee*), we employ the AttentiveFP-CI model to predict %*ee*. This class-imbalance aware graph-based model with an attention mechanism exhibits commendable performance, as evidenced by the root mean square error (RMSE) of 9.80±1.40. Evaluation across various molecular representations of these reactions (OHE, fingerprints, SMILES, graphs) and ML algorithms (DNN, T5Chem,

View Article Online
DOI: 10.1039/D5DD00483G

Transformer, MPNN), the AttentiveFP-CI emerged as the best model distinguished by its minimal overfitting (train-test RMSE difference of 3.59, compared to up to 5.40 for other CI-aware models). When extended to other important reaction datasets such as the N,S-acetylation, asymmetric hydrogenation of alkenes, and the USPTO, the improved predictions could be obtained by using the AttentiveFP-CI. Furthermore, attention visualization identifies key atoms and substructures contributing to high enantioselectivity, offering valuable chemical insights for planning the synthesis of new molecular targets. Harnessing insights derived from ML model could serve an efficient and cost-effective approach for expedited developments in asymmetric catalysis.

Introduction: In the domain of organic synthesis, efforts toward generating chiral molecules have remained a vibrant activity for several decades. Such developments were motivated by the utility of target compounds as drugs and pharmaceuticals.[1,2] Historically, the discovery and optimization of catalytic reactions have relied heavily on empirically gathered domain knowledge, the intuition of experienced chemists, and an inevitable series of trial-and-error explorations.[3] Such traditional and heuristic approaches might often encounter limitations while predicting the behaviour of specific substrates/catalysts or when fine-tuning reaction conditions to achieve a desired transformation.[4,5] The intricate interactions between various parameters of a reaction, some obvious and some subtle, could alter the course of reaction in unpredictable ways.[6,7] A change in reaction outcome, such as the yield or selectivity, could stem from a number of factors, such as the choice of catalysts, ligands, solvents, and additives, besides temperature and other parameters pertaining to the reaction conditions. Since yield and/or selectivity depend on multiple components, it becomes an inherently complex high-dimensional problem. This very complexity associated with reaction outcome predictions makes them an all the more interesting research problem to consider.

View Article Online
DOI: 10.1039/D5DD00483G

There have been a good number of previous efforts aimed at predicting reaction outcomes. The predictive capabilities of quantum chemically derived molecular descriptors have been exploited to build bespoke linear regression models for catalytic reactions.[8,9] Molecular descriptors such as charge, NMR chemical shifts, vibrational frequencies and intensities, sterimol parameters, buried volumes, etc., bearing electronic and steric features of the participating molecules, have served as useful inputs for modelling reactions.[10,11] The use of such descriptors comes with their own challenges such as higher computational cost, particularly in the case of complex molecular systems as well as the requirement of annotation/curation by domain experts.[12,13] Given these challenges, an increasing interest in exploring alternative approaches became more prominent in current literature. Modern ML algorithms, capable of handling complex and diverse reaction data, can offer promising solutions in this front.[14,15] In practice, when one attempts reaction optimization by way of changing various controllable parameters (as described above), sparsely distributed reaction data becomes available. It would be of interest to try and see whether such data could possibly present potential opportunities for applying suitable ML algorithms for reaction modelling.[16,17,18] The key advantage of an early ML intervention would be to help make an informed choice of substrates/catalyst/solvent during the reaction development phase.

The unprecedented growth in computational capabilities has rendered the applications of ML to chemical reactivity problems increasingly more feasible.[19,20] Deployment of incredibly complex language models such as BERT for yield prediction became possible in very recent years.[21,22] The use of hybrid graph neural networks on molecular graphs to derive features for selectivity predictions in the case of chiral phosphoric acid catalysed thiol addition to N-acyl imines is now available.[23] These contemporary ML models for reaction outcome prediction have offered robust performance on High Throughput Experimental (HTE) datasets.[24,25,26] Analyses have shown that HTE datasets, as used in many of the recent

View Article Online
DOI: 10.1039/D5DD00483G

studies, exhibited reduced variability in data quality, high internal consistency and high fidelity.[27,28] Another aspect of the HTE settings is that exhaustive permutations of reactants/reagents are affordable under uniform reaction conditions. However, in real-life reaction development only a few combinations between the reactants and the associated conditions can be practically explored. For example, the dataset sourced from the AstraZeneca electronic laboratory notebooks (ELNs) potentially encompassed approximately 470 M possible combinations of reactants. However, in practice, only 1000 reactions were experimentally examined by engaging 340 aryl halides, 260 amines, 24 ligands, 15 bases, and 15 solvents for the Buchwald-Hartwig reaction.[29] In this context, we consider it highly timely to develop accurate ML models for small-sized reaction datasets with different distribution characteristics.

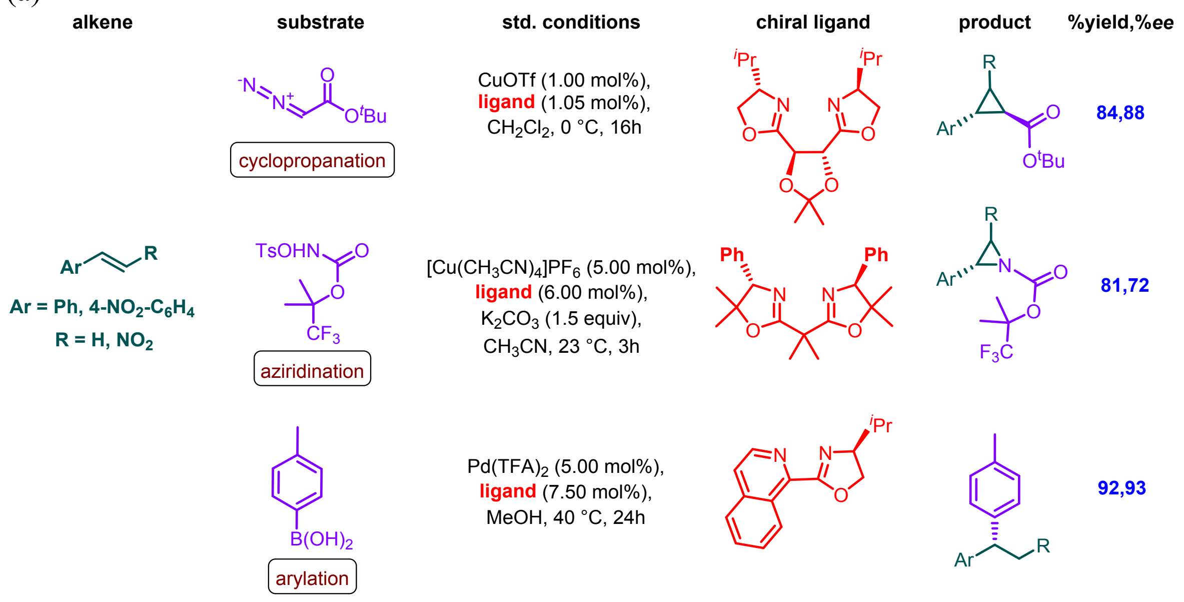
Recent years witnessed several successful applications of ML in predicting yields or enantioselectivities of various catalytic reactions such as the Buchwald-Hartwig cross-coupling,[24] Lewis base-catalysed propargylation,[30] β-C–H activation,[31] and asymmetric hydrogenation,[32] relay Heck,[33] Negishi cross-coupling,[34] and palladaelectro-catalyzed C–H annulation reactions.[35] Needless to say, most of these studies are early examples of implementing deep learning (DL) methods and were confined to only a few reaction types, leaving out a large family of important asymmetric catalytic reactions. One of the key reasons for such exclusions from ML studies can be traced to the lack of good datasets. One such important catalytic asymmetric reaction that has not received attention is shown in Scheme 1a, which employs simple alkenes as the core substrate. Alkenes are abundant precursors that can participate in a wide array of reactions to provide valuable products. For example, under a suitably chosen Cu/Pd catalytic condition, alkenes can react with (a) diazoester to form cyclopropane, (b) aryl boronic acid to yield important 1,1-diaryl compounds, and (c) aliphatic or aromatic N-tosyloxycarbamates to access key structural motifs such as aziridines. This

View Article Online
DOI: 10.1039/D5DD00483G

class of reaction holds promise as it can help synthesize stereochemically well defined cyclopropanes and aziridines, which are key constituents in medicinal and agrochemical compounds.[36,37,38,39] A few representative examples bearing these substructures are shown in Scheme 1b to convey the significance of these small ring containing molecules.[40,41,42,43]

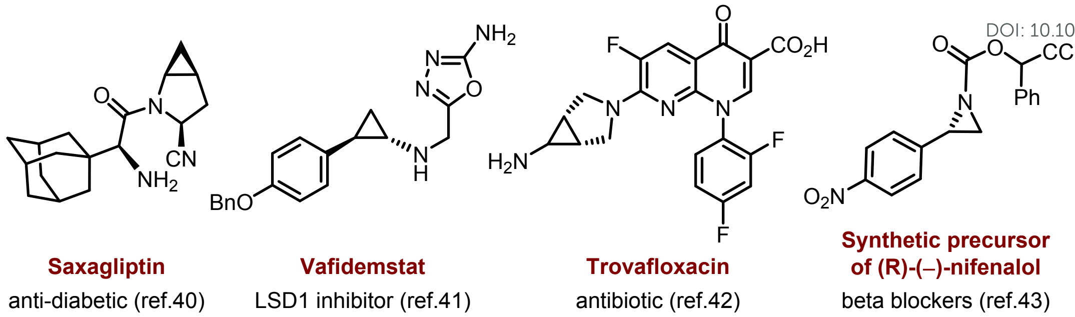
Apart from the synthetic utility of this class of reactions, the use of one of the most widely found ligands, such as the chiral bis(oxazoline), is to be taken cognizance of, as the applications of such chiral motifs go well beyond these reactions.[44,45,46] The conformationally rigid framework of bis(oxazoline) metal chelates, bearing chiral centres close to the donor nitrogen atoms, can provide the desired chiral environment nearer to the catalytic site. The modular architecture of these ligands can help create desirable variations in both steric and electronic attributes, thus allowing fine-tuning of their catalytic activity for specific applications.[47,48] It would therefore be of importance to identify the key regions in the chiral catalyst that impact the stereochemical outcome of such reactions, potentially using ML tools (*vide infra*).

(a)



(b)

View Article Online
DOI: 10.1039/D5DD00483G



Scheme 1. (a) A representative example of catalytic asymmetric cyclopropanation, aziridination, and arylation of alkene, (b) select examples of pharmaceuticals/drugs containing cyclopropane/aziridine structural motifs.

Motivated by recent advancements in machine learning approaches for reaction outcome prediction,[49,50,51,52,53] including contributions from our laboratory,[31,32,33] we became interested in the catalytic enantioselective reactions of alkenes as shown in Scheme 1a.[54,55] The availability of reliable predictive ML models can help identify optimal reactant triads comprising of alkene, chiral ligand, and substrate that are likely to offer higher %*ee*. Such ML models might help reduce the typical timelines involved in reaction discovery. Given these motivations, we set the following major objectives in this work; (a) evaluation of the effectiveness of DL methods in enantioselectivity predictions in a transition metal-catalysed asymmetric reactions of alkenes, (b) identification of optimal featurisation strategy from among One-Hot Encoding (OHE), molecular fingerprints, SMILES, and Graph representations, (c) address the issues associated with data imbalance consisting of more samples in the high %*ee* region by implementing cost-sensitive training loss, (d) identification the better combinations of reactants (alkene, chiral ligand, and substrate) that are likely to offer superior reaction outcome, and (e) examination of learning ability of the DL models by using attention mechanism to identify the critical regions in chiral ligands and substrates that can influence the reaction outcomes. Utilization of a trained DL model can streamline and help expedite the reaction discovery pipeline by identifying and eliminating low selectivity reactions in the initial screening, which would save time and effort.

View Article Online
DOI: 10.1039/D5DD00483G

Results and discussion

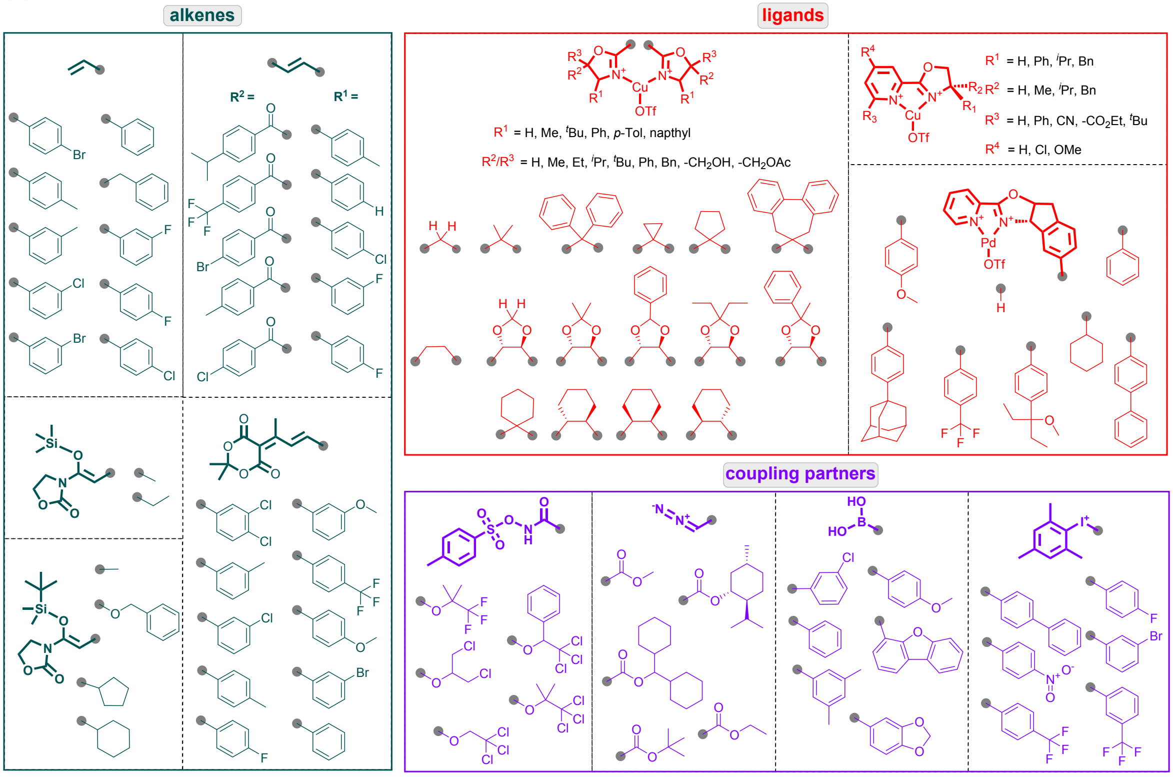
We have arranged the discussions into eight subsections. First, an overview of the reaction dataset is provided to shed light on reaction-specific details such as the type and diversity of reactions considered as well as the distribution characteristics in the dataset. Next, we describe data encoding modality, followed by an outline of the AttentiveFP model. Subsequently, the model performance across different featurization and further refinements to the model to address the class imbalance (CI) issues is provided. Then, we discuss model interpretability by drawing parallels between graph attention and chemical intuition. Next section is on the prospects of exploiting the model interpretability and predictive abilities to identify optimal catalysts potentially useful for the synthesis of important drug molecules. Finally, we extend our AttentiveFP-CI framework to some of the commonly used reaction datasets as additional case studies —an asymmetric hydrogenation,[32] an N,S-acetylation reaction,[56] and the USPTO[57] to evaluate the applicability of our model across different reaction classes.[58]

1. Reaction Dataset: The dataset was manually curated by collecting reaction details from peer-reviewed publications, which led to a comprehensive collection of 376 reactions.[59,60,61,62,63,64,65] The three classes of catalytic asymmetric reactions of alkenes considered in this study are cyclopropanation, aziridination, and arylation. The dataset is abbreviated as ART that stands for *AlkeneReactionTriad* to reflect the three classes of alkene reactions.[66] The reactions in the ART dataset differ in terms of ligands bound to the transition metal (Pd or Cu), alkenes, and substrates. Each sample represents a distinct combination of reactants and the corresponding %*ee* as the output value. The diversity of each reaction component can be gleaned from Fig. 1a. In the transition metal-ligand complexes, the major difference arises from the decorations in the chiral ligands, ranging from bis(oxazoline) combined with Cu and pyridine-oxazolines bound to Pd or Cu (see Tables S1-S3 in section-1

View Article Online
DOI: 10.1039/D5DD00483G

of ESI). Either of these two chiral ligand families, when bound to a suitable transition metal can serve as the catalyst with intriguingly diverse steric and electronic characteristics. They form a sufficiently wide range of catalysts in active use today. Additionally, the reactions exhibit differences in the substrate involved, comprising of 73 distinct entities, encompassing electron-rich to electron-deficient aryl boronic acids, aliphatic or aromatic N-tosyloxycarbamates, carbenoids, diazo compounds, diaryliodonium salts, and other species. The alkenes used in these reactions span a wide array of compounds, including mono- and di-substituted alkenes, alkyl/aryl-substituted silylketenimides, and Meldrum's acid-derived dienes. All these diverse combinations across each reaction partner together makes the ART dataset richly diverse and, hence, can be considered as a representative of ground truth situations one would typically encounter in reaction discovery. Thus, the utility of the ML approach, designed for such a family of reaction with sparse and imbalanced data distribution (Fig. 1b), should remain good for other reactions as well.

(a)


alkenes
ligands
coupling partners
R^2 =
R^1 =
R^1 = H, Me, ^{t}Bu, Ph, p-Tol, napthyl
R^2/R^3 = H, Me, Et, ^{i}Pr, ^{t}Bu, Ph, Bn, -CH$_2$OH, -CH$_2$OAc
R^1 = H, Ph, ^{i}Pr, Bn
R^2 = H, Me, ^{i}Pr, Bn
R^3 = H, Ph, CN, -CO$_2$Et, ^{t}Bu
R^4 = H, Cl, OMe
Cu
OTf
Pd

(b)

View Article Online
DOI: 10.1039/D5DD00483G

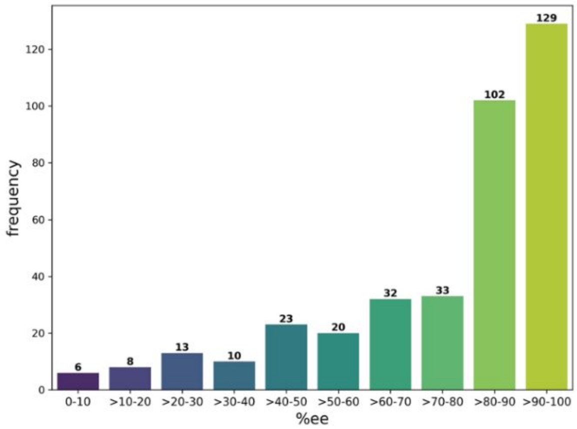


Fig. 1. (a) Details of various substituents present in the individual reacting partner, (b) yield distribution in the ART dataset.

The chemical space spanned by the ART dataset is very sparse compared to the combinatorial possibilities arising from the number of reactions between the compatible partners. For instance, in the cyclopropanation subset there are 130 examples (67 catalysts, 10 substrates, and 23 alkenes) while the aziridination reaction comprises only 91 reactions (14 catalysts, 13 substrates, and 44 alkenes), and the arylation reaction class contains as many as 155 reactions (19 catalysts, 50 substrates, and 55 alkenes). Given the possible combinations between the reactants, the theoretically likely reactions for cyclopropanation, aziridination, and arylation are respectively 15410, 52250, and 8008, totaling 75,668 possibilities. However, the ART dataset contains only 376 experimentally known reactions from among these combinations, indicating a highly sparse distribution. Furthermore, distribution of the %*ee* values is also skewed towards the high *ee* regions (Fig. 1b).[67,68] The diversity of chemical structures, skewed distribution of reaction outcome, and sparsity in the dataset can together make the ML model building rather challenging.

2. Data pre-processing and training protocol: Since each reaction variable can play a role in influencing the outcome to different extent, their meaningful representation is central to the input data. To achieve this, we encoded all the molecular partners as linear strings using the Simplified Molecular Input Line-Entry System (SMILES) notation.[69] These individual

View Article Online
DOI: 10.1039/D5DD00483G

molecules are concatenated to form a complete reaction. Since the target %*ee* values are continuous, the ML task can be considered as a regression problem. Within the ART dataset, each catalyst scaffold consistently delivers preferred enantiomeric outcome, allowing the model to focus on predicting the numerical %*ee* values. The dataset is randomly divided into train, validation, and test sets in a 70:10:20 ratio for training purposes. We have conducted hyperparameter tuning on the validation set based on the criterion of achieving the lowest mean validation loss and then employ such optimal hyperparameters for prediction on the test set. To mitigate potential bias due to sample distribution, while creating the test-train splits (70:10:20), 30 independent runs with randomised splits were considered. The model performance is evaluated using root mean squared error (RMSE) on the test sets as the average RMSE over these runs.

3. AttentiveFP Model: In this study, we use the AttentiveFP-based DL model as the primary model to predict the output of the reaction (i.e., the enantioselectivity expressed in %*ee*).[70] This model uses molecular graphs as the input, where nodes and edges of the graphs respectively represent atoms and bonds. For a full representation of the reaction, the three major reaction entities i.e., alkene, chiral ligand, and coupling partner, are concatenated to form a single composite graph. These molecular graphs remain disconnected within this composite representation, implying no explicit edges linking different reactants involved in the reaction. The bond features and atom features are obtained from the RDkit and encoded as described in section 2.5 in ESI. The model utilizes a graph-attention mechanism to focus on the most relevant regions in the composite graph of the concatenated ‘reactants’. Fig. 2 outlines the AttentiveFP network with an attention mechanism.

View Article Online
DOI: 10.1039/D5DD00483G

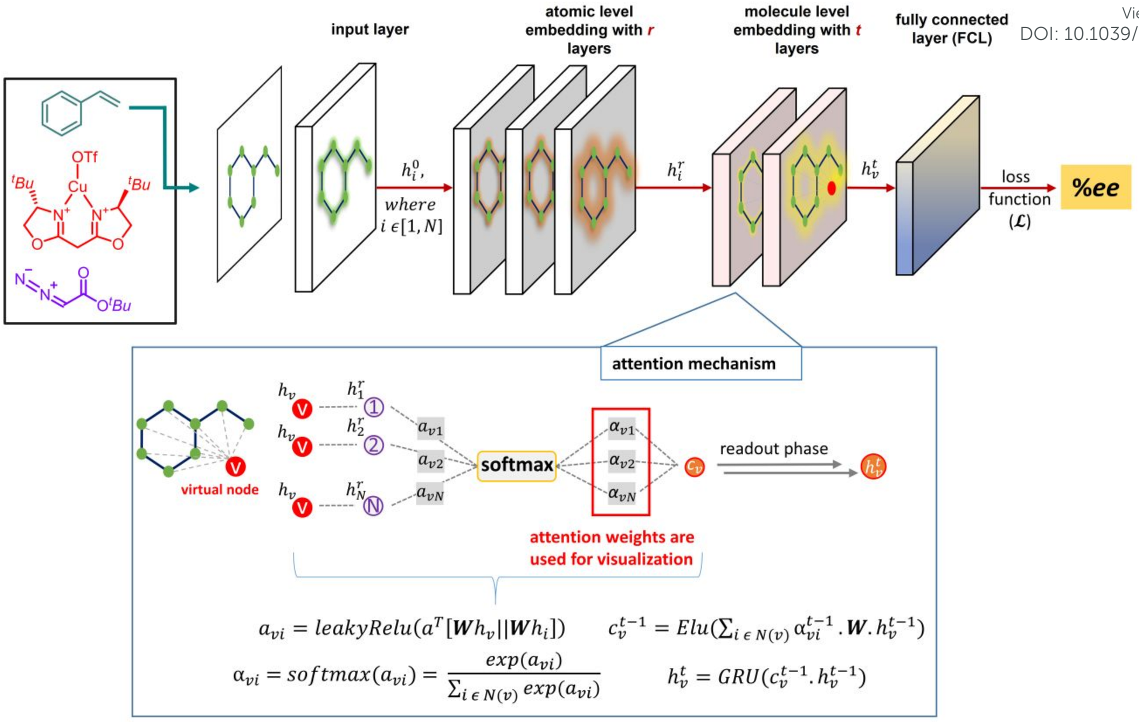

input layer
atomic level embedding with r layers
molecule level embedding with t layers
fully connected layer (FCL)
OTf
Cu
^{t}Bu
^{t}Bu
N$^+$
N$^+$
O
O
^{t}Bu
h_i^0,
where $i \in [1, N]$
h_i^r
h_v^t
loss function ($\mathcal{L}$)
%ee
attention mechanism
virtual node
h_v
h_1^r
h_2^r
h_N^r
a_{v1}
a_{v2}
a_{vN}
softmax
α_{v1}
α_{v2}
α_{vN}
c_v
readout phase
h_v^t
attention weights are used for visualization
$a_{vi} = leakyRelu(a^T[\boldsymbol{W}h_v||\boldsymbol{W}h_i])$
$c_v^{t-1} = Elu(\sum_{i \in N(v)} \alpha_{vi}^{t-1}.\boldsymbol{W}.h_v^{t-1})$
$\alpha_{vi} = softmax(a_{vi}) = \frac{exp(a_{vi})}{\sum_{i \in N(v)} exp(a_{vi})}$
$h_v^t = GRU(c_v^{t-1}.h_v^{t-1})$

Fig. 2. Major components of the AttentiveFP model architecture. While the input sample is a composite molecular graph formed by concatenating all the three reaction components (alkene, chiral ligand, and coupling partner), only a representative alkene (styrene) is shown here for brevity.

Since the model focuses on individual atoms, each atom includes features from its neighbouring atoms and the bonds connecting them to form the respective initial state vectors h_i^0 for each atom. These initial vectors are further embedded with ***r*** number of stacked attentive layers, allowing atoms to aggregate relevant "messages" from their neighborhood. This step is expected to capture the nuances of atomic local environments by effectively propagating node information over various distances. For molecule-level embedding, the entire molecule is treated as a super-virtual node (*V*), connecting every atom, and embedded using the attention mechanism as shown in Fig. 2. This process, over ***t*** stacked layers, produces a state vector h_v^t for the whole molecule. In this mechanism, the first step is to concatenate the state vectors of the virtual node (h_v) with all connected nodes (h_1^r), followed by a linear transformation (**W**) and nonlinear activation (leakyRelu) to produce a_{vi}. This a_{vi}

View Article Online
DOI: 10.1039/D5DD00483G

is then normalised using a softmax function over the neighbour nodes, resulting in α_{vi} that captures the importance (weight) of each neighbour node to the virtual node. These attention weights are then passed through the message and readout functions to obtain the final state vector h_v^t, which encodes structural information about the molecular graph. Finally, h_v^t is fed through the fully connected layer (FCL) for the regression task.

We have optimised the model using the Adam optimization algorithm, in conjunction with Bayesian optimization (BO) for model-specific parameters such as the number of graph layers for atom embedding, the number of time steps for molecule embedding (denoted respectively using *r* and *t* in Fig. 2), graph feature size, dropout rate, and optimizer parameters such as learning rate. Tuning hyperparameters in DNN can be challenging as they are used for model parameter estimation rather than the model directly assessing them. Hence, BO, as implemented in the Optuna python package,[71] is used to optimize model-specific parameters. The optimal set of validation hyperparameters for all 30 runs are provided in section 2.5 in ESI.

4. Effect of different featurization methods and DL models on enantioselectivity prediction: Since ML model performance can depend on the nature of chemical featurisation, we examined the influence of different featurization methods such as one-hot encoding (OHE), molecular fingerprints (FP), SMILES, and graphs in this work. This aspect has received relatively limited attention in previous works on %*ee* prediction, prompting us to assess the three traditional featurisation modalities to determine their efficacy on the ART dataset. We have encoded the samples using one-hot encoded vectors (OHEs) to capture the presence or absence of chemical entities such as catalyst, ligand, and substrates in a given reaction. This approach therefore treats each reaction component as an independent categorical variable without explicitly encoding the corresponding chemical structures. While such representations may lack chemical relevance, they might help unveil statistical patterns

View Article Online
DOI: 10.1039/D5DD00483G

within complex datasets.[72] The second approach utilizes molecular fingerprints (FP), a well-established molecular encoding technique in the form of fixed-length binary strings.[73] These FPs capture molecular topology by delineating atom neighbourhoods within a specified radius, providing a compact yet informative representation of the molecular structure. We used a variety of FPs, such as circular, atom pair, and layered fingerprints, each furnishing unique structural details. For instance, circular FPs encode sub-structural patterns around each atom, atom pair FPs capture pairwise interactions, and layered FPs bring structural information at multiple abstraction levels. Leveraging the open-source cheminformatics package RDKit, we generated molecular FPs with specific parameters tailored to our experimental setup. Additionally, we explored the utility of molecular SMILES, a string-based representation that not only encapsulates structural details like atoms, bonds, and connectivity but also captures nuanced features such as stereochemistry and unique bonding patterns. Our fourth approach involved molecular graphs, offering a robust representation of molecules where atoms correspond to nodes and bond to edges in an undirected graph. By augmenting nodes and edges respectively with atomic identity and bond order as features, these attributed molecular graphs can exhibit versatile predictive capabilities in DL endeavours.[74]

In addition to examining the effects of featurisation on the ART dataset, we aim to compare the performance of the AttentiveFP model with some of the state-of-the-art (SOTA) DL models commonly used in enantioselectivity predictions. We have conducted comprehensive evaluations of various DNN models employing both OHE and molecular FPs.[75] Additionally, we explored transformer-based language models such as Transformer,[76] ULMFiT,[77] and T5Chem,[78] which utilize reaction SMILES, and graph neural networks like MPNN[79] and AttentiveFP, which leverage graphs for reactant featurisation. Each

View Article Online
DOI: 10.1039/D5DD00483G

combination between a DL model and featurization is then evaluated on the basis of the corresponding train and test RMSEs (see section 2 in ESI for more details).

Table 1: Performance comparison in terms of RMSE (in %*ee*) for different models as well as various featurization techniques obtained as the averaged over 30 independent runs[a]

featurization	model	train	test
OHE	DNN	3.67±1.94	14.43±3.05
Fingerprint	DNN	5.54±1.50	9.55±1.31
SMILES	T5Chem	6.74±0.39	10.83±1.73
	ULMFiT	10.94±0.51	11.30±1.30
	Transformer	5.28±1.18	12.26±2.02
Graph	**AttentiveFP**	**7.41±1.77**	**10.56±1.86**
	MPNN	8.01±1.18	11.00±2.22

[a] The datasets are randomly divided into 70:10:20 train, validation, and test sets.

The model performance compiled in Table 1 highlight the effect of different featurization of the ART dataset. With OHE, the test RMSE of the DNN is found to be 14.43±3.05 (The details of the DNN architecture are given in section 2.2 in ESI). These OHE-based models serve as a statistical probe, offering an internal baseline performance for other models built using chemically meaningful descriptors. In contrast to OHE, models that utilize FPs offered improved performance.[71,80] It is worth noting that although the fingerprint-based DNN model exhibited a relatively lower test RMSE (9.55±1.31) compared to AttentiveFP, a larger gap with the train RMSE (5.54±1.50) suggests overfitting, hence should be considered with caution when applied to out-of-bag situations. The use of SMILES representations in conjunction with advanced DL architectures, including T5Chem, ULMFiT, and Transformer, yielded slightly higher test RMSEs of 10.83±1.73, 11.30±1.30, and 12.26±2.02, respectively.[81] The graph-based MPNN model showed a high test RMSE of 11.00±2.00.[82] Thus, despite not being the top performer in terms of the lowest test RMSE, the

View Article Online
DOI: 10.1039/D5DD00483G

balanced performance of AttentiveFP suggests it is a robust model with a lower susceptibility to overfitting compared to the other models considered here.[83] In addition to the good performance, the graph attention mechanism inherent to the AttentiveFP model that allows for chemically meaningful interpretability (*vide infra*), has made the AttentiveFP as our primary framework for the *ee* prediction task.[84]

5. Class imbalanced aware reweighting strategies

While the AttentiveFp model built on molecular graphs showed good predictive performance (Table 1), it does not address the class imbalance issue, given the inherently skewed distribution toward the high *ee* values found in several real-world cases, including that in the ART dataset. Although class imbalance mitigation methods are commonly applied to classification problems, regression tasks like the *ee* prediction have received little attention toward building ML models that address distribution imbalances. Previous research has often relied on the SMOTE technique to generate synthetic data in the minority class for certain regression models.[31,33] The ART dataset bears a skewed distribution with a majority of samples in the higher end of the %*ee* values (Fig. 1b), thus necessitating the development of a customized regression model aware of this class imbalance.

AttentiveFP

$$\mathcal{L} = \frac{1}{M}\sum_{j=1}^{M}(y_j - \hat{y}_j)^2; \quad \hat{y}_j = \text{FCL}\,(h_v^t)$$

AttentiveFP-CI

$$\mathcal{L}_{\text{class\_imb}} = \frac{1}{M}\sum_{j=1}^{M}\begin{cases}(y_j - \hat{y}_j)^2 & if\ \ y_j < 50\\ 0.5 \times (y_j - \hat{y}_j)^2 & if\ \ y_j \geq 50\end{cases}$$

boundary set at 50, ≥ 50 majority samples and <50 minority samples

Fig. 3. Different types of loss are used in the AttentiveFP and AttentiveFP-CI model.

Herein, we propose a customized model for class imbalance, namely, AttentiveFP-CI. Unlike conventional Mean Squared Error (MSE) loss, our model incorporates a class imbalance loss, assigning different weights to training samples based on their actual *ee* values (Fig. 3). The idea is to reduce the influence of the majority class samples while prioritising the more challenging minority class instances during training.[85] We have examined the effect of using different class boundaries, from 30 to 60, as well as by placing the boundary at

View Article Online
DOI: 10.1039/D5DD00483G

statistically important points of the dataset such as the mean (μ) value of 76 and μ-σ of 54. In most cases, the AttentiveFP-CI models performed better than the AttentiveFP model, as evident from the corresponding test RMSE (see Tables S61-S69 in ESI). The model with a class boundary of 30 achieved the best test RMSE of 9.80±1.40 as compared to other class boundaries considered.[86] Moreover, the *t*-test resulted in a p-value < 0.05, indicating that the gain in performance is statistically significant as compared to the model without the CI (test RMSE for AttentiveFP is 10.56 ± 1.86). Incorporation of the CI-aware loss into other DL models also improved the respective test RMSEs, except in the case of the ULMFiT.[87] A comparison between different deep learning architectures reveals that the AttentiveFP-CI outperforms Transformer-CI, and ULMFiT, with p-values < 0.05 endorsing their statistical significance.[88] However, most of these models tend to exhibit overfitting issues, evident from the train-test RMSE differences as follows; MPNN-CI (4.59), T5Chem-CI (5.4), and DNN-CI (3.93) as opposed to the AttentiveFP-CI (3.59).[89] Additionally, the number of model parameters in the AttentiveFP-CI is of the order of 1.93M, which are much lesser than those in T5Chem-CI (14.71M), assuring us of better computational scalability. Given the lower RMSE, reduced overfitting, and computational efficiency, the AttentiveFP-CI stands out as the optimal choice from among all the SOTA models for the %*ee* predictions on the ART dataset.[90]

Efforts were also expended to assess whether the performance issues could be traced to the sparse and imbalanced distribution in the ART dataset. We have compared the model performances on more balanced and denser datasets such as the Buchwald–Hartwig Amination (BHA), which is a catalytic transformation of high practical utility. The high throughput experimental (HTE) dataset of the BHA reaction, denoted as BHA-HTE is a commonly used dataset for baseline comparisons for yield prediction tasks.[24] The BHA-HTE comprises of 3,955 labeled reactions and their corresponding experimentally measured yields.

View Article Online
DOI: 10.1039/D5DD00483G

The AttentiveFP model offered a good test performance, with an RMSE of 6.49±0.33 and a coefficient of determination (R^2) of 0.94±0.01 (See Table S72 in ESI), surpassing the previously reported R^2 of 0.92 obtained using physical-organic descriptors.[29]

In the present context, we have done random sampling of the full BHA-HTE dataset to create a few sparser subsets, denoted as BHA-LTE (low throughput), each containing about 500 reactions. The idea is to induce skewness to produce an imbalance in labels such that the distribution (μ and σ) in the BHA-LTE subsets resembles that of the ART dataset. These subsets are then employed for evaluating the baseline performance of various deep learning models as considered in this study.[91] In general, the BHA-LTE subsets have a μ of 75 and a σ of 14 (see Fig. S1 in ESI), similar to that in our ART dataset (μ=76 and σ=22). When the AttentiveFP model was trained using these subsets bearing an induced sparse distribution, the test RMSE dropped from 6.49±0.33 with the original BHA-HTE dataset (see Table S72 in ESI) to 9.14±0.80 (or higher, depending on the BHA-LTE subset used). The lower performance of the same AttentiveFP model can be attributed to the induced sparse distribution and CI in the BHA-LTE subsets. Interestingly, inclusion of the CI loss with a class boundary of 50 improved the test RMSE to 8.70±0.52. Such test performances are analogous in quality of predictions by the same model on the ART dataset bearing comparable distribution characteristics (μ of 76 and σ of 22). Additional details on the model performance with varying class boundaries, spanning 30 to 70, as well as with the μ of 75 are provided in section 3 of ESI. An improvement in RMSE (9.14±0.80 for AttentiveFP versus 8.70±0.52 for AttentiveFP-CI) could possibly be due to the use of a customized loss function to mitigate CI issues.

A similar performance trend is conspicuous in our ART dataset as well. For instance, the test RMSE of the AttentiveFP-CI model (with a class boundary of 30) is found to be 9.80 over 10.56 obtained with AttentiveFP model without the CI loss. On the basis of the model

View Article Online
DOI: 10.1039/D5DD00483G

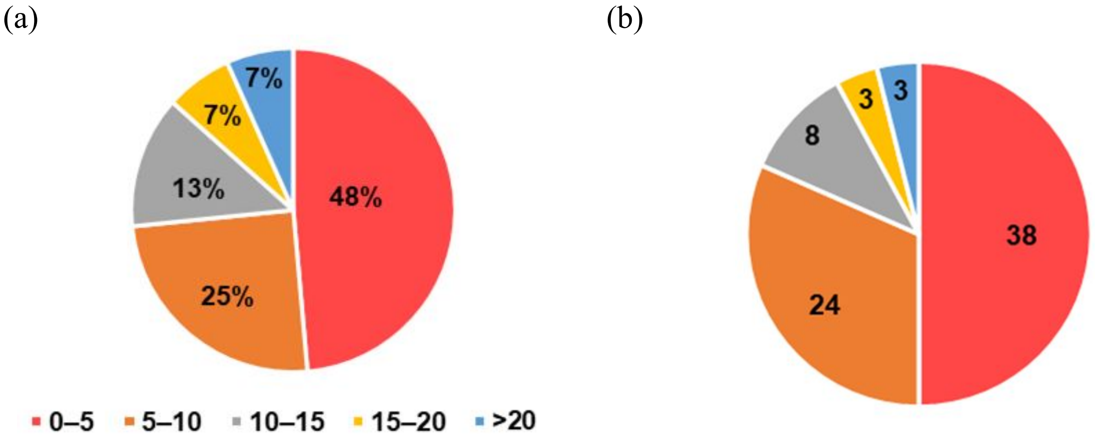
performance, with and without the inclusion of the CI loss, on the ART and BHT-LTE datasets, we could conclude that the data sparsity is primarily responsible for the higher RMSEs. These insights would be valuable in developing suitable deep learning models with customized loss functions for chemical reaction datasets bearing skewed distribution.

An alternative for imbalanced and sparsely distributed chemical dataset is to consider a two-step model,[92] wherein a classification of samples is done first on the basis of a predefined class label. Subsequently, separate regressors are developed for the major and minor classes. This approach, termed as classification-followed-by-regression (CFR), is likely of relevance to the ART dataset.[93] In the first step, reactions are classified as 'major' or 'minor' using a statistically meaningful class boundary set at the (μ−σ) of 54 %*ee*.[94] We found that a hyperparameter optimized custom built DNN classifier could achieve a very good accuracy of 0.98±0.003.[95] The reactions thus got classified are employed in training two separate AttentiveFP regression models, one for the major class and the other for the minor class. The AttentiveFP regressor achieved test RMSEs of 10.74±1.98 for major class and 8.73±0.90 for the minor class, outperforming our direct regression in the case of minor class reactions (i.e., reactions with less than 54 %*ee* as their true label).[96] To ensure a balanced assessment of the overall model performance, we have also considered the use of weighted RMSE, which accounts for class imbalance by combining error contributions proportionate to the sample size, is calculated.[97] AttentiveFP model, the weighted RMSE is 8.97±1.28, which is poorer than our AttentiveFP-CI with a class boundary of 30 (RMSE =9.80±1.40, R^2 = 0.80±0.05) (See Table S64 in ESI). Similarly, an ULMFiT regression model showed a test RMSE of 10.15±2.26 and 8.48±1.26 respectively for the major and minor classes with the corresponding weighted RMSE of 8.75±1.05 (R^2 = 0.40±0.29).[98] Since no significant improvement is found with CFR-major and CFR-minor classes, our original AttentiveFP-CI

View Article Online
DOI: 10.1039/D5DD00483G

with its interpretable characteristics, can be considered as a more appropriate model for the ART dataset.

AttentiveFP-CI showed good performance in predicting %*ee*, achieving a test RMSE of 9.80±1.40. Importantly, the difference between the training and validation RMSEs suggests a lesser overfitting, which is good for model generalizability when predicting on unseen samples. In the 30 independent runs, the model predicts %*ee* thousands of times for the 76 reactions present in the test set. Further, every reaction gets predicted multiple times whenever it appears in the test set. A comparison of the predicted %*ee* with the experimentally known ground truth values revealed good correlation as shown in Fig. 4. In fact, ~87% of the predictions remain within 15 units of the actual values (Fig. 4a). In the optimal run with an RMSE of 8.2 %*ee*, as many as 70 out of 76 test samples are predicted well within an error limit of 15 units with respect to the corresponding true values (Fig. 4b). In a typical run (RMSE=10.1), only 12 out of the 76 samples incurred prediction errors in excess of 15 units (Fig. 4c). The parity plot also conveys a good correlation between the %*ee* predicted by the AttentiveFP-CI model and the corresponding experimental values with an R^2 of 0.84 (Fig. 4d).[99] These assuring findings highlight the efficacy of AttentiveFP-CI in learning from the sparse ART dataset for catalytic asymmetric reactions of alkenes.



(c) (d)

View Article Online
DOI: 10.1039/D5DD00483G

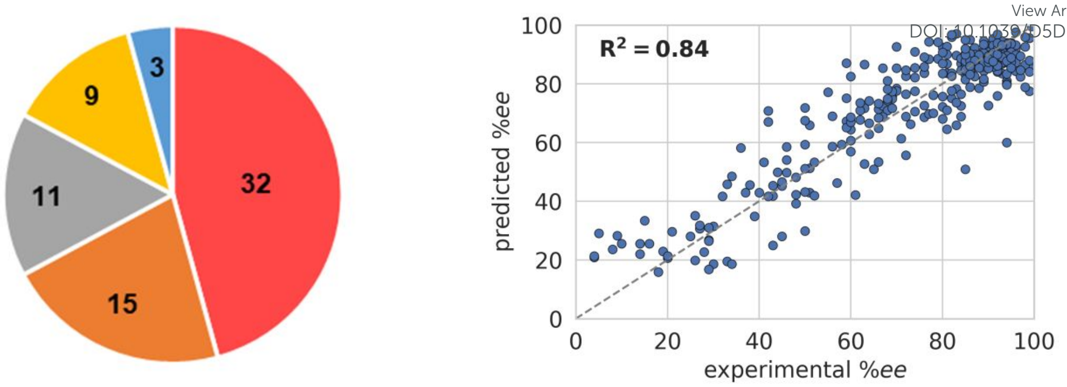


Fig. 4. Pie charts summarize the prediction errors (%*ee* differences) obtained using the AttentiveFP-CI model across 76 test samples: (a) aggregated over 30 independent runs, (b) from the best run, and (c) in a typical run whose performance is similar to the average RSME over 30 runs. (d) Parity plot showing the correlation between experimental and predicted %*ee* values for all test samples.

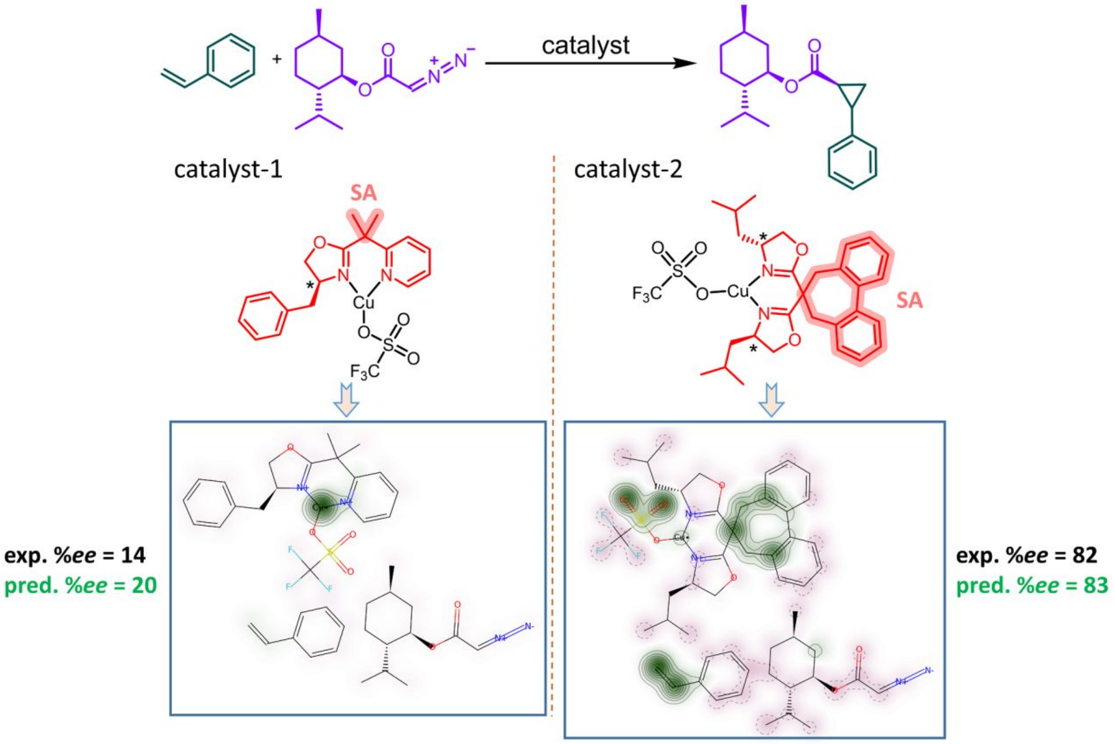
To evaluate the learning ability of an ML model and to examine its robustness, control experiments are required. For this purpose, the dependence of the model performance on the quality of the input data is assessed using techniques such as Y-scrambling. We created a straw model of AttentiveFP-CI, which intentionally breaks the potential connections between the input features and the output variable. Here, each sample is incorrectly mapped to an output value belonging to some other sample within the dataset. The considerably worse test RMSE of 25.2±2.1 obtained with the Y-scrambling run shows that the model learns from the true features it was provided with in the correct training campigns. The inferior performance also highlights the effectiveness of the AttentiveFP-CI in learning the chemically meaningful aspects of the catalytic reaction investigated in this work (*vide infra*).

6. Interpretability and graph attention visualization: With these promising findings on %*ee* prediction, it becomes all the more relevant to gather additional insights on what the model would have learned from the molecular representations as provided. We sought details such as the key atoms that are likely to exert more impact on the reaction output expressed in %*ee*. Leveraging on the interpretability of attention weights (see α_{vi} in Fig. 2) learned by the

View Article Online
DOI: 10.1039/D5DD00483G

model, we probed the latent connections between various molecules present in a given reaction sample and the predicted outcome. Here, we aim to visualize the atomic attention weights using the similarity map as implemented in the RDKit program.[100] In this mode of visualization shown in Fig. 5, the regions with a green glow indicate a positive influence on %*ee*, a pink glow suggests a negative influence, and a grey colour conveys no significant overall effect. As a representative example, we have compared the attention weights for two of the chosen reactions involving the same reactants but different chiral ligands (pyridine-oxazoline and bis(oxazoline) in Fig. 5a. It shall be noted that in both these reactions, the predicted %*ee* is comparable to the experimentally reported %*ee*.

(a)



(b)

View Article Online
DOI: 10.1039/D5DD00483G

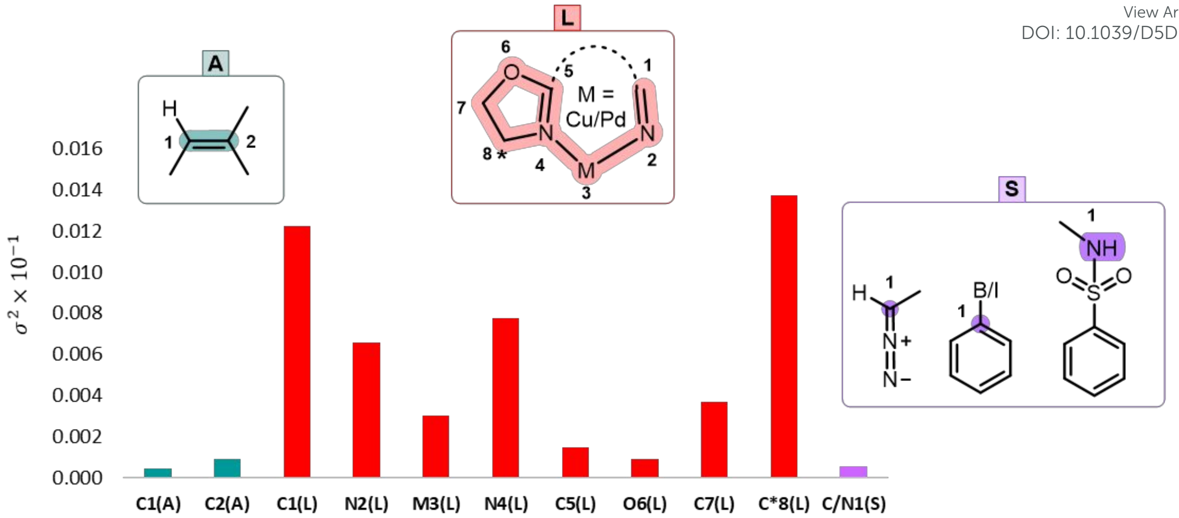


Fig. 5. Visualization of the attention weights of two representative reactions involving same substrates and different catalysts. (a) In the predicted probability maps, atoms with positive contributions to the reaction outcome appear in green, while red indicates that the corresponding attention weight is negative. The larger the absolute value, darker is the colour in the attention map. (b) Atoms within the shared regions in each reaction partner involved in the overall reaction are shown highlighted, along with the variance (σ^2) of the attention value (y-axis) for the highlighted atoms (x-axis). Different colours such as cyan for alkene (A), red for ligand (L), and purple for substrates (and diazo compound, boronic acid, or N-tosyloxycarbamate) (S) is used in the bar plots.

A perusal of the attention map shown in the inset of Fig. 5a as well as the bar plot in Fig. 5b for the reaction involving catalyst-1 (pyridine-oxazoline catalyst), it can be learned that the Cu centre, pyridine N with two nearest C atoms make a positive contributions while all other atoms or substructures have relatively lower negative or negligible contribution to the %*ee*. In the case of catalyst-2, the side arm (SA) on the bridge carbon of the bis(oxazoline) ligands positively contribute to high %*ee*, consistent with the trends observed with this family of ligands.[101] It is further evident that the other positive contributors to the reaction outcome are (i) SA on the chiral ligand, (ii) the transition metal-bound triflate ligand, (iii) the styrenyl double bond, and (iv) one of the carbon atoms of the cyclohexyl diazo

View Article Online
DOI: 10.1039/D5DD00483G

compound. These positive attention values are suggestive of their synergistic role in the enantioselectivity of cyclopropanation reaction. One of the significances of this analysis is that installation of suitable substituents on the SA group could be key to achieving enhanced enantioselectivity. This prediction by the model is chemically meaningful and intuitively appealing, as it aligns with the fact that most variants of reported bis(oxazoline) ligands rely on modifications of the SA.[101]

After visualizing attention for two representative examples, we have analysed the global effects of the critical regions/atoms that are likely exert a significant contribution toward the quality of the %*ee* prediction. To accomplish this, it is essential to identify a common region present in each reaction partner across all the samples. Fig. 5b highlights such shared regions in all the reaction components, along with their atom numbering. The steps involved in estimating the effect of each atom are as follows: First, attention values for each atom in the shared region are extracted using the corresponding SMART pattern. Second, the variance of these attention values is plotted, since the variance is crucial in assessing the feature importance as it captures the most significant and informative variations in the data.[102] Interestingly, the bar diagram shown in Fig. 5b indicates a higher importance of the chiral ligand (atom numbers are given with L in parentheses to denote the chiral ligand) as compared to the reactants such as the alkene and other substrates. It is gratifying to see that our attention-based model deciphered chemically intuitive patterns present in the chiral ligand as the most relevant contributor to asymmetric induction. The variance in the attention values exhibited by the atoms in the chiral ligands are found to be much higher than those of the alkene and other substrate. This observation is in line with one's chemical intuition that the chiral ligands play a pivotal role in transferring the chiral information to the developing product.[103] Notably, the chiral carbon centre, denoted as [C*8(L)] in the ligand exhibits the highest variance in the attention, corroborating the domain knowledge that the substituents at

View Article Online
DOI: 10.1039/D5DD00483G

this centre are largely known to influence enantioselectivity. Additionally, the carbon atom near the bridge or SA [C1(L)] shows the second-highest variance, suggesting that modifications to this atom, where branching from the bridge carbon begins, could help in fine-tuning the reaction outcome. The identification of these atoms as important ones indicates that AttentiveFP-CI accurately captures the relationship between the molecular factors and the desired outcome. Exploiting this protocol by fine-tuning these key features, particularly during reaction development, or while expanding the scope of this reaction family, could prove advantageous.

7. Application of AttentiveFP-CI regressor for identifying potential catalysts:

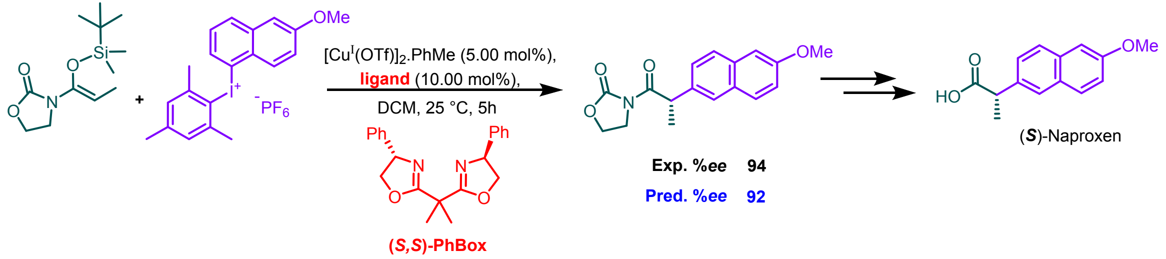
We wish to demonstrate one of the potential applications of the attention-based regression model here in terms of identifying suitable substrates or catalysts to expedite the synthesis of target molecules of interest, such as a drug molecule. To illustrate this aspect, we have used the AttentiveFP-CI model to identify an efficient catalyst for synthesizing (*S*)-naproxen, which is an oral non-steroidal anti-inflammatory drug. One of the experimental reports suggests the use of (*S,S*)-PhBox as the chiral ligand in an enantioselective Cu-catalyzed arylation reaction as shown in Fig. 6a.[104,105] The observed *ee* was 94%, which is very close to the predicted *ee* of 92% by the AttentiveFP-CI model. Given this encouraging agreement, we have screened a library of 35 chiral Box ligands from the PubChem database using the AttentiveFP-CI model and identified several promising chiral ligands with high predicted %*ee* values, as shown using a heatmap representation in Fig. 6b (see Table S88 in ESI for more details).[106,107] These candidates prioritized by the model can be utilized to streamline future screening efforts. However, their activity remains to be established through wet-lab validation.

It is worth noting that although these ligands were previously reported, their use in copper-catalyzed arylation reactions involving silylketenes and diaryliodonium salts, remains

View Article Online
DOI: 10.1039/D5DD00483G

unexplored (Fig. 6a). Thus, the potential of these ligands for such alkene-substrate pairs are novel, even as the reaction conditions are retained as is.[63,64,65] A heatmap representation of the predicted %*ee* for different chiral ligands shown in Fig. 6b conveys the significance of fast (virtual) screening of chiral ligands by using our regression model. It can be seen (top left) that a ligand with only one of the oxazoline rings chiral is predicted to exhibit a lower %*ee* of 65. Interestingly, the attention analysis identifies the groups on the side arm at the bridge carbon (SA) as well as the chiral carbon of the bis(oxazoline) ligand as the dominant contributors to %*ee*. In light of the attention as noted, we considered two representative variants of the bis(oxazoline) family of ligands such as the (*S*,*S*)-Ph-Box for further illustration as shown in Fig. 6b. One of these ligands is obtained by replacing the Ph group at the chiral carbon with 4-^{t}Bu-Ph, and the other by replacing the 1,1-dimethyl on the side arm with 1,1-diisopropyl group. Both of these ligands are predicted to show high %*ee*. More importantly, a higher attention value noted for the 1,1-diisopropyl as well as in the Ph regions (green color contours) indicates their positive contribution to enantioselectivity. These can be considered as indicative of how attention-based approach could be utilized in catalyst design for asymmetric reactions.

(a)



(b)

View Article Online
DOI: 10.1039/D5DD00483G

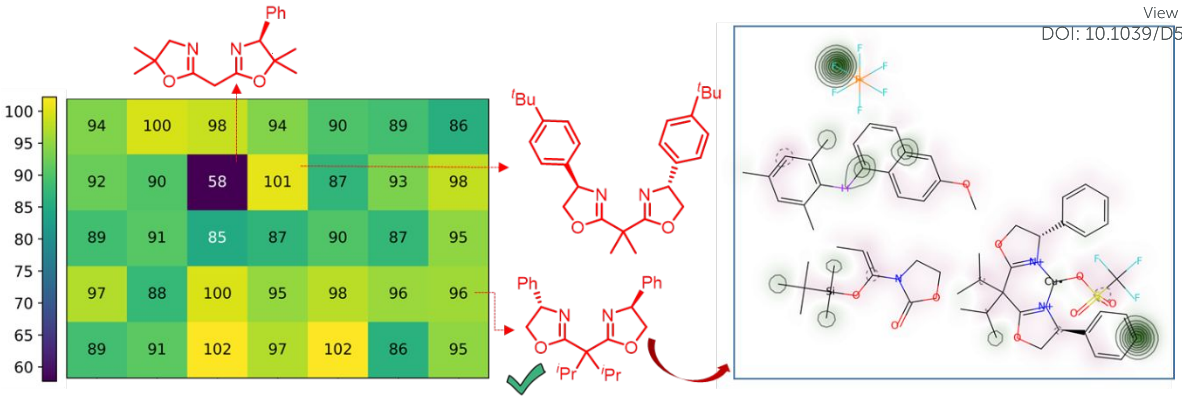


Fig. 6. (a) Experimental condition for making a precursor involved in the synthesis of (*S*)-naproxen using an asymmetric Cu-catalyzed arylation reaction of alkene, (b) representative examples of some chiral bis(oxazoline) ligands with their predicted %*ee* obtained using the AttentiveFP-CI model. Shown in the inset is attention weight visualization for the reaction catalyzed by [Cu-(*S*,*S*)-Ph-Box] with 1,1-diisopropyl groups at the bridging carbon.

8. Extension of AttentiveFP model to asymmetric hydrogenation, N,S-acetylation, and USPTO datasets: Motivated by the promising interpretability of the AttentiveFP model, we have considered two more important reactions from the domain of asymmetric catalysis involving axially chiral ligands/phosphoric acid (CPAs). These are N,S-acetylation[56] and asymmetric hydrogenation of alkenes,[32] both known for their high industrial importance.[108,109,110] The N,S-acetylation involves the enantioselective coupling between an imine and a thiol, while the hydrogenation reaction encompasses the reduction of imines and alkenes using BINOL- and BINAP-derived chiral ligands. The original N,S-acetylation dataset contains 1075 reactions, wherein the enantioselectivity is expressed using the corresponding ΔΔG‡ values, ranging from positive to negative values, or used a signed *ee* value to denote the experimentally observed major enantiomer (*R* or *S*) in the reaction. Hence, it shall be noted that in the present work, we formulate the problem as a regression task to predict %*ee* values on a 0-100 scale, focusing on the magnitude of enantioselectivity in line general practice followed in organic chemistry literature. Consequently, 48 reactions bearing negative *ee* values (used to indicate the opposite enantiomer) are excluded during

View Article Online
DOI: 10.1039/D5DD00483G

data curation to maintain a consistent target value, making it 1027 reactions. For consistency in benchmarking and comparison with the prior studies, the AttentiveFP-CI model is also evaluated on the full 1075 N,S-acetylation reaction dataset using the same 80:10:10 train, validation, test split as used before besides a μ-based class boundary of 0.98. Our model achieved a test R^2 of 0.90±0.02 and an RMSE of 0.21±0.02.[111] These results are comparable to those of the SEMG-MIGNN (R^2 = 0.915, RMSE = 0.197)[23] and ChemAH (R^2 = 0.918, RMSE = 0.209).[112] Under the regression setting with 1027 reactions, the AttentiveFP-CI achieves a good predictive performance of 8.06 ±1.00 (R^2 of 0.92±0.10) on the 1,027 N,S-acetylation reaction, which is found to be a statistically significant improvement over the corresponding AttentiveFP model devoid of CI loss.[113a] In the case of asymmetric hydrogenation, although the performance of AttentiveFP-CI as indicated by an RMSE of 10.48±1.10 (R^2 of 0.60±0.17) is good, the gain as compared to the AttentiveFP model is not statistically significant.[113b] Notably, for the N,S-acetylation reaction, the AttentiveFP-CI significantly outperformed a often used baseline such as the ULMFiT model (p = 0.0036), with a test RMSE of 8.88±1.03.[114] Although in the hydrogenation case, baseline ULMFiT provides better predictive accuracy (test RMSE of 8.56±1.46), it lacks the advantage of interpretability as afforded by AttentiveFP-CI.

In addition to the enantioselectivity prediction on three important chemical reaction datasets, the utility of the AttentiveFP-CI framework is also evaluated for an yield prediction tasks on the USPTO (grams) reaction dataset.[21] This dataset comprises 1.9×10^5 reactions, each annotated with the corresponding yield values. The yield distribution exhibits a skewness of −0.86, indicating the presence of CI in the USPTO dataset. Furthermore, the yield values in this dataset are reported as scaled values to fit in the interval of 0 to 1. In light of this skewed distribution, we used the AttentiveFP-CI model with a statistically relevant class boundary of μ+σ (0.94) to note a test RMSE of 0.20±0.01 and a marginally better R^2 of

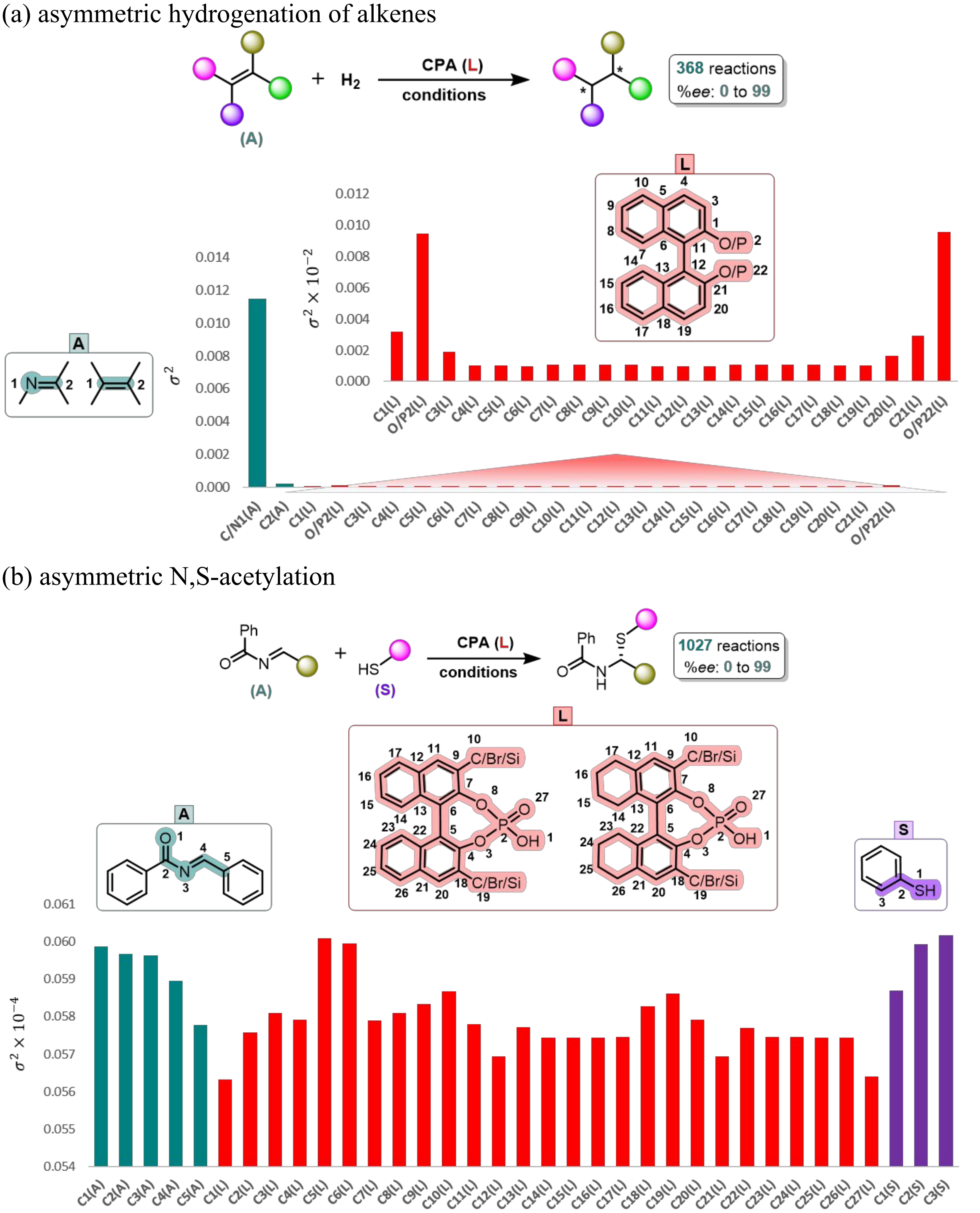
View Article Online
DOI: 10.1039/D5DD00483G

0.08±0.00, as compared to the AttentiveFP without the CI consideration (test RMSE = 0.21±0.01, $R^2 = 0.04\pm0.01$).[115] However, the *t*-test resulted in a p-value > 0.05, indicating that the numerical improvement is not statistically significant in the case of USPTO dataset. Notably, the performance of the AttentiveFP-CI is even comparable to the previously reported RMSE of 0.195 obtained using a more complex transformer-based model on the same dataset.[116] Overall, these results indicate that the AttentiveFP-CI could be useful in addressing CI issues in chemical datasets, even if it does not always surpass state-of-the-art benchmark performances.

Similar to the approach employed earlier in this manuscript for global attention analysis, we have visualized the attention weights by using the computed variance in the attention values of atoms within the shared region as shown in Fig. 7, for both asymmetric N,S-acetylation and asymmetric hydrogenation reactions.[100,117,118,119] The variance in the attention values of the shared region atoms of the catalysts and substrates across different samples (i.e., reactions in the dataset) can be considered as a measure of sensitivity of the reaction outcome to the environment of such atoms. Therefore, such analysis might help decipher how the changes in the local substituents are likely to influence the enantioselectivity. A relatively larger change in variance implies a higher attention on such atoms, which might stem from the changes in its substituents or local environment. A modest change indicates that the atom concerned consistently gets similar attention weights. Interestingly, in the context the reactions in our ART dataset, we notice that reactive positions on the alkene and the vital regions around the chiral centre exhibit high variances in their attention. These variances both in the ART reactions as well as in the case of asymmetric hydrogenation are in line with our chemical intuition, where alkene is the key substrate undergoing the reaction. Since this analysis collectively reveals mechanistically valuable insights consistent with the domain knowledge on the origin of enantioselectivity catalyzed

View Article Online
DOI: 10.1039/D5DD00483G

by axially chiral ligands,[101,120,121] we consider that the AttentiveFP-CI model is meaningfully interpretable.



(a) asymmetric hydrogenation of alkenes

(b) asymmetric N,S-acetylation

Fig. 7. Highlighted atoms in the alkene (A), chiral ligand (L), and substrate (S) from the shared region across different reaction/samples are identified as exerting a positive/negative impact on the reaction outcome. The bar plot shows the variance in

View Article Online
DOI: 10.1039/D5DD00483G

attention values (σ^2) obtained from the AttentiveFP-CI model (y-axis) for each atom as shown labeled in the x-axis. Colors indicate different reaction components: cyan for alkene, red for ligand, and purple for substrate for better visualization.

In the asymmetric hydrogenation reaction, the atoms belonging to the imine or alkene (e.g., C/N1(S)) exhibited consistently higher variance in attention, conveying that they might play a key role in enantioselectivity. While the atoms in the chiral ligand showed relatively lower variance compared to those of the substrate, the P/O center in BINOL- and BINAP-derived ligands (e.g., O/P2(L) or O/P22(L)) have maximal attention variance. This is an interesting aspect, which is in line with the chemical intuition that these positions in the chiral BINOL/BINAP frameworks are expected to influence effectiveness of enantioinduction. Thus, suitable substitution at the ortho positions in the biphenyl ring can potentially impact enantioselectivity values. For the N,S-acetylation reaction, high attention variance is found across all three components: the thiol (C3(S)), imine (C1(A)), and ligand (C5(L)), reflecting the impact of local substituent changes on enantioselectivity. Notably, within the ligand, attention values corresponding to the atoms C5(L) and C6(L), which bridge the biphenyl units, showed substantial variance, linked to the dihedral angle fluctuations (e.g., C4–C5–C6–C7) that can modulate the chiral environment. Additionally, variance in attention scores suggest that the meta-positions on the biphenyl rings (C19(L), C10(L)) are important where a change in substituents is likely to influence the reaction outcome. Thus, the attention variance analysis suggests that introducing suitable substituents at the high-variance sites (e.g., P/O centers, dihedral-sensitive atoms) might shift the enantioselectivity. These findings offer a firm basis for the interpretability of the AttentiveFP-CI by way of identifying hotspots for enantioselectivity tuning. The results could become useful for rational catalyst design and for making informed choice of substrates during reaction scope investigations in asymmetric catalysis.

View Article Online
DOI: 10.1039/D5DD00483G

Conclusions

We have curated a comprehensive dataset, named ART (AlkeneReactionTriad) that contains 376 catalytic asymmetric reactions of alkenes, as a representative of the real-world chemical reaction dataset suitable for machine learning model building. We have employed an attention-based class imbalance aware GNN model (AttentiveFP-CI) on the ART dataset for predicting enantioselectivity in terms of %*ee* as the label. Our ML model is found to be effective in addressing one of the known challenges in reaction outcome prediction, arising due to the sparse and imbalanced distributions often found in chemical reaction datasets. Through a comparative analysis of various featurization techniques, including one-hot encoding, molecular fingerprints, SMILES, and molecular graphs, we found that the graph-based AttentiveFP-CI is the most suitable model offering a test RMSE of 9.80±1.40 on the ART dataset. Importantly, our findings indicate that an RMSE in this range is comparable to the performance of the same AttentiveFP-CI on a sparsity-induced Buchwald–Hartwig amination dataset (10.86±1.32 test RMSE as opposed to 6.49±0.33 with the full BHA dataset). Lower performance on the sparse subset of the BHA dataset can be considered to stem from the sparse distribution and CI issues, similar to that in the ART dataset.

Visualization of the atomic attention weights could identify the pivotal regions in the reaction partners, such as the chiral centre as a high attention spot in the chiral catalyst. Similarly, critical atoms/substructures in the reactant(s) are identified as an important contributor to high enantioselectivity. Thus, the AttentiveFP-CI not only serves as a good predictive model, it also offers chemically meaningful insights for reaction optimization. This method can therefore pave way for informing ligand design and reaction development, as exemplified by the identification of the (*S*,*S*)-PhBox ligand variant, featuring 1,1-diisopropyl groups on the side arm as potentially effective catalyst relevant to the synthesis of (*S*)-naproxen. When extended to an important enantioselective reaction, such as the axially chiral

View Article Online
DOI: 10.1039/D5DD00483G

phosphoric acid (CPA) catalyzed N,S-acetalylation, the AttentiveFP-CI offered a very good RMSE of 8.06±1.00. The interpretability of our model sheds light on the factors governing enantioselectivity in the form of identifying the reactive olefinic sites in imine and alkenes in asymmetric hydrogenation reaction as the key contributors and the binaphthyl axis of the axially chiral ligands in the case of asymmetric acetylation reaction. Overall, the AttentiveFP-CI model not only offers as a robust predictive framework but also as a chemically interpretable tool that complements intuition. Interpretable models can therefore be exploited in data-driven discovery of chiral ligands and substrates in asymmetric catalysis.

Data availability

Data and codes related to this work are publicly available through our GitHub repository at https://github.com/alhqlearn/ART-AttentiveFP-CI. A citable, versioned snapshot of the codebase is archived on Zenodo with an assigned DOI: https://doi.org/10.5281/zenodo.18256995

Author contributions

AH and NJ contributed equally, DC prepared the ART dataset. AH, NJ and RBS wrote the manuscript. RBS supervised the research and analysis.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We are thankful to Institution of Eminence (IoE) Data and Information Science computing facility for generous computational resources. N.J. acknowledges the Prime Minister Research Fellowship (PMRF).

References

1. K. R. Campos, P. J. Coleman, J. C. Alvarez, S. D. Dreher, R. M. Garbaccio, N. K. Terrett, R. D. Tillyer, M. D. Truppo and E. R. Parmee, *Science*, 2019, **363**, 6424.

View Article Online
DOI: 10.1039/D5DD00483G

2. G. M. Whitesides, *Angew. Chem. Int. Ed.,* 2015, **54**, 3196–3209.

3. I. W. Davies, *Nature*, 2019, **570**, 175–181.

4. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.

5. S. Lin, S. Dikler, W. D. Blincoe, R. D. Ferguson, R. P. Sheridan, Z. Peng, D. V. Conway, K. Zawatzky, H. Wang, T. Cernak, I. W. Davies, D. A. DiRocco, H. Sheng, C. J. Welch and S. D. Dreher, *Science*, 2018, **361**, 6402.

6. S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.

7. P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf and T. Laino, *WIREs Comput. Mol. Sci.*, 2022, **12**, 5.

8. J. Werth and M. S. Sigman, *J. Am. Chem. Soc.*, 2020, **142**, 16382–16391.

9. J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang and S. E. Denmark, *J. Am. Chem. Soc.*, 2020, **142**, 11578–11592.

10. A. Milo, A. J. Neel, F. D. Toste and M. S. Sigman, *Science*, 2015, **347**, 737–743.

11. S. Zhao, T. Gensch, B. Murray, Z. L. Niemeyer, M. S. Sigman and M. R. Biscoe, *Science*, 2018, **362**, 670–674.

12. K. H. Hopmann, *Int. J. Quantum Chem.*, 2015, **115**, 1232–1249.

13. A. S. K. Tsang, I. A. Sanhueza and F. Schoenebeck, *Chem. Eur. J.*, 2014, **20**, 16432–16441.

14. W. Beker, R. Roszak, A. Wołos, N. H.Angello, V. Rathore, M.D. Burke and B. A. Grzybowski, *J. Am. Chem. Soc.*, 2022, **144**, 4819−4827.

15. M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218–10239.

16. S. Singh and R. B. Sunoj, *Acc. Chem. Res.*, 2023, **56**, 402–412.

17. A. Hoque, M. Das, M. Baranwal and R. B. Sunoj, *arXiv*, 2024, DOI:10.48550/arXiv.2407.10090.

View Article Online
DOI: 10.1039/D5DD00483G

18. P.-X. Hua, Z. Huang, Z.-Y. Xu, Q. Zhao, C.-Y. Ye, Y.-F. Wang, Y.-H. Xu, Y. Fu and H. Ding, *Commun. Chem.*, 2025, **8**, 42.

19. S. Shilpa, G. Kashyap and R. B. Sunoj, *J. Phys. Chem. A*, 2023, **127**, 8253–827.

20. A. Hoque, M. Surve, S. Kalyanakrishnan and R. B. Sunoj, *J. Am. Chem. Soc.*, 2024, **146**, 28250-28267.

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

22. M. Das, A. Ghosh and R. B. Sunoj, *J. Comput. Chem.*, 2024, **45**, 1160–1176.

23. S.-W. Li, L.-C. Xu, C. Zhang, S.-Q. Zhang and X. Hong, *Nat. Commun.*, 2023, **14**, 1.

24. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.

25. D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429–434.

26. J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.

27. V. Voinarovska, M. Kabeshov, D. Dudenko, S. Genheden and I. V. Tetko, *J. Chem. Inf. Model.*, 2023, **64**, 42–56.

28. M. J. Gaunt, J. M. Janey, D. M. Schultz and T. Cernak, *Chem*, 2021, **7**, 2259–2260.

29. M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle, N. V. Chawla and O. Wiest, *Chem. Sci.*, 2023, **14**, 4997–5005.

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

31. A. Hoque and R. B. Sunoj, *Digit. Discov.*, 2022, **1**, 926–940.

32. S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1339–1345.

33. M. Das, P. Sharma and R. B. Sunoj, *J. Chem. Phys.*, 2022, **156**, 114303.

View Article Online
DOI: 10.1039/D5DD00483G

34. A. E. Cuomo, S. Ibarraran, S. Sreekumar, H. Li, J. Eun, J. P. Menzel, P. Zhang, F. Buono, J. J. Song, R. H. Crabtree, V. S. Batista and T. R. Newhouse, *ACS Cent. Sci.*, 2023, **9**, 1768–1774.

35. X. Hou, S. Li, J. Frey, X. Hong and L. Ackermann, *Chem*, 2024, **10**, 1–12.

36. Y. Imai, W. Zhang, T. Kida, Y. Nakatsuji and I. Ikeda, *Tetrahedron Lett.*, 1997, **38**, 2681–2684.

37. A. V. Bedekar and P. G. Andersson, *Tetrahedron Lett.*, 1996, **37**, 4073–4076.

38. A. V. Bedekar, E. B. Koroleva and P. G. Andersson, *J. Org. Chem.*, 1997, **62**, 2518–2526.

39. K. Alexander, S. Cook and C. L. Gibson, *Tetrahedron Lett.*, 2000, **41**, 7135–7138.

40. A. Ramirez, V. C. Truc, M. Lawler, Y. K. Ye, J. Wang, C. Wang, S. Chen, T. Laporte, N. Liu, S. Kolotuchin, S. Jones, S. Bordawekar, S. Tummala, R. E. Waltermire and D. Kronenthal, *J. Org. Chem.*, 2014, **79**, 6233–6243.

41. T. Nakamura, T. Yoshihara, C. Tanegashima, M. Kadota, Y. Kobayashi, K. Honda, M. Ishiwata, J. Ueda, T. Hara, M. Nakanishi, T. Takumi, S. Itohara, S. Kuraku, M. Asano, T. Kasahara, K. Nakajima, T. Tsuboi, A. Takata and T. Kato, *Mol. Psychiatry*, DOI:10.1038/s41380-024-02479-8.

42. S. Kurosawa, F. Hasebe, H. Okamura, A. Yoshida, K. Matsuda, Y. Sone, T. Tomita, T. Shinada, H. Takikawa, T. Kuzuyama, S. Kosono and M. Nishiyama, *J. Am. Chem. Soc.*, 2022, **144**, 16164–16170.

43. T. Norris, T. F. Braish, M. Butters, K. M. DeVries, J. M. Hawkins, S. S. Massett, P. R. Rose, D. Santafianos and C. Sklavounos, *J. Chem. Soc., Perkin trans. 1*, 2000, 1615–1622.

44. K. Kikushima, J. C. Holder, M. Gatti and B. M. Stoltz, *J. Am. Chem. Soc.*, 2011, **133**, 6902–6905.

45. G. Chelucci, M. G. Sanna and S. Gladiali, *Tetrahedron*, 2000, **56**, 2889–2893.

View Article Online
DOI: 10.1039/D5DD00483G

46. J. I. Garcia, B. Lopez-Sanchez, J. A. Mayoral, E. Pires and I. Villalba, *J. Catal.*, 2008, **258**, 378–385.

47 . J. I. Garcia, G. Jimenez-Oses, B. Lopez-Sanchez, J. A. Mayoral and A. Velez, *Dalton Trans.*, 2010, **39**, 2098-2107.

48. C. Mazet, V. Köhler and A. Pfaltz, *Angew. Chem. Int. Ed.*, 2005, **44**, 4888–4891.

49. A. Button, D. Merk, J. A. Hiss and G. Schneider, *Nat. Mach. Intell.*, 2019, **1**, 307–315.

50 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.

51. A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields and A. G. Doyle, *Acc. Chem. Res.*, 2021, **54**, 1856–1865.

52.Y. Ma, X. Zhang, L. Zhu, X. Feng, J. A. H. Kowah, J. Jiang, L. Wang, L. Jiang and X. Liu, *Molecules*, 2023, **29**, 5995.

53. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.

54. J. Li, S. H. Liao, H. Xiong, Y. Y. Zhou, X. L. Sun, Y. Zhang, X. G. Zhou and Y. Tang, *Angew. Chem. Int. Ed.*, 2012, **51**, 8838-8841.

55. D. A. Evans, K. A. Woerpel, M. M. Hinman and M. M. Faul, *J. Am. Chem. Soc.*, 1991, **113**, 726-728.

56. A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.

57. D. M. Lowe, Extraction of Chemical Structures and Reactions from the Literature. Ph.D. thesis, University of Cambridge, 2012.

58. Additional details about each of the reaction dataset used in this study are provided in Section 1.4 of the ESI.

59. R. E. Lowenthal, A. Abiko and S. Masamune, *Tetrahedron Lett.*, 1990, **31**, 6005–6008.

View Article Online
DOI: 10.1039/D5DD00483G

60. R.E. Lowenthal and S. Masamune, *Tetrahedron Lett.*, 1991, **32**, 7373-7376.

61. M.B. France, A.K. Milojevich, T.A. Stitt and A.J. Kim, *Tetrahedron Lett.*, **44**, 9287-9290.

62. M. Itagaki, K. Masumoto, K Suenobu and Y. Yamamoto, *Org. Process Res. Dev.*, **10**, 245-250.

63. A. Ebinger, T. Heinz, G. Umbricht and A. Pfaltz, *Tetrahedron*, 1998, **54**, 10469–10480.

64. H. Lebel, M. Parmentier, O. Leogane, K. Ross and C. Spitz, *Tetrahedron*, 201, **68,** 3396-3409.

65. S. Chen, L. Wu, Q. Shao, G. Yang and W. Zhang, *Chem. Commun.*, 2018, **54***,* 2522-2525.

66. As the ART dataset aggregates reactions reported across different peer reviewed literature, it inherently reflects the reporting practices, which may introduce biases such as an over representation of successful or high-yielding reactions. Since the experimental data were not generated within this work, such biases cannot be fully eliminated.

67. W. Gao, P. Raghavan, R. Shprints and C. W. Coley, *J. Am. Chem. Soc.* 2025, **147**, 8959−8968.

68. To meaningfully capture the distribution of the label (%*ee*) in our ART dataset, skewness is calculated, which is a statistical measure of how the data deviate from symmetry about the mean. The fact that a skewness of zero implies a symmetric distribution and a negative value indicates left-skewness, the skewness of -1.37 found for the ART dataset indicates a pronounced deviation toward the lower end.

69. D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

70. Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang and M. Zheng, *J. Med. Chem.*, 2019, **63**, 8749–8760.

71. T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, In *KDD*, 2019.

72. K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, 6416.

View Article Online
DOI: 10.1039/D5DD00483G

73. F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379-1390.

74. Additional details on the featurization strategies are provided in Section 1.5 of the ESI.

75. In addition to DNN model building, we built various tree-based models (Random Forest, Decision Tree, and Gradient Boosting) as well as other models like Support Vector Machine to predict %*ee* on our ART dataset. Although performances were comparable among these models, they are found to be less effective than DNN. Each of these model performances with their hyperparameters are compiled in Tables S4-S11 in ESI.

76. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Kaiser and I. Polosukhin, *Adv. Neural Inf. Process.*, 2017, **30**, 5998-6008.

77. S. Singh and R. B. Sunoj, *Digit. Discov.*, 2022, **1**, 303–312.

78. J. Lu and Y. Zhang, *J. Chem. Inf. Model.*, 2022, **62**, 1376–1387.

79. Y. Kwon, D. Lee, Y.-S. Choi and S. Kang, *J. Cheminform.*, 2022, **14**, 2.

80. For each of the 30 splits, hyperparameter optimization of the DNN architecture is performed using the Optuna framework (ref. 71). The optimal hyperparameters are selected based on the lowest validation RMSE. Details on the optimal hyperparameters and corresponding model performance are provided in Tables S12-S14 in ESI.

81. Details of the hyperparameters and corresponding model performances for T5Chem, Transformer encoder, and ULMFiT are respectively provided in Tables S23-S24, S31-S32, and S39-S46 in ESI.

82. Details of the hyperparameters and corresponding model performances for Yield-BERT and MPNN are respectively provided in Tables S21-S22 and Tables S53-S54 in ESI.

83. To evaluate alternative target values, we computed $\Delta\Delta G^{\ddagger}$ from the reported %*ee* values and trained the AttentiveFP model accordingly. We note that training on %*ee* offers better predictive performance (RMSE = 10.56±1.86, R^2 = 0.77±0.08) than on $\Delta\Delta G^{\ddagger}$ (RMSE =

View Article Online
DOI: 10.1039/D5DD00483G

0.43±0.05, $R^2 = 0.58 \pm 0.10$). This highlights that the model captures enantioselectivity trends more effectively when learning from %*ee* directly. See section 5 in ESI for more details.

84. Additional justification for the use of AttentiveFP as the primary framework is provided in the ESI (Section 2.7).

85. In this case, a true *ee* < 50, the standard squared error (SE) between the predicted and true *ee* is taken into consideration just as in the case of AttentiveFP implementation. For *ee* > 50, on the other hand, the SE is halved, to enhance the model sensitivity to minority class samples.

86. The AttentiveFP-CI model maintains strong predictive performance (MAE ≈ 7.10, RMSE ≈ 9.80, RSE ≈ 0.16, $R^2 \approx 0.84$) with low relative error across all class boundaries, confirming that it is free from significant systematic bias (see Table S71 in the ESI).

87. In the ULMFiT regressor, each SMILES in the reaction dataset is augmented by 125 non-canonicalized/randomized form (hyperparameter tuning using various degree of SMILES augmentation (from 25, 50, 75, 100, 125, to 150), helped us 125 as the optimal choice). For further details on hyperparameter tuning and model performance, see section 2.4 in ESI. Additionally, Gaussian noise is added to the corresponding output of each randomized SMILES. This augmentation, with slight modifications to their *ee* values effectively serves to increase samples in the low *ee* region. This is likely the reason why the addition of CI loss did not enhance performance, as the augmentation inherently addresses CI. The performance of the ULMFiT-CI Algorithm with various class boundaries are provided in Tables S47-S52 in ESI.

88. The *t*-test results resulted *p*-values of 9.0×10^{-6} for the transformer model and 0.0002 for the ULMFiT.

89. Details and performance of all other CI-aware models considered in this study are provided in section 2 in ESI. The results with various class boundaries for DNN-CI are

View Article Online
DOI: 10.1039/D5DD00483G

provided in Tables S15-S20, T5Chem-CI in Tables S25-S30, Transformer encoder-CI in Tables S33-S38 and MPNN-CI in Table S55-S60.

90. To assess model robustness evaluated using random partitioning, we have additionally implemented a scaffold-based spliting using the Murcko scaffolds. the AttentiveFP-CI model with a class boundary of 30 exhibits an increase in test RMSE from 9.80±1.40 under random splitting to 13.85±3.41 under scaffold splitting (See Table S70 of the ESI). Importantly, these results confirm that the AttentiveFP-CI algorithm is able to generalise to reactions with unseen core scaffolds, rather than relying solely on close structural analogues.

91. (a) Additional details about dataset creation, performance of the AttentiveFP model are provided in section 3 in ESI. (b) The AttentiveFP-CI model on different sparsity induced subsets of BHA dataset, denoted as BHA-LTE, revealed a similar performance as that obtained using the same model on our ART dataset (See Table S73 in ESI).

92. (a) A conceptually related two-step strategy by Chung et al., first partioned the reactions on the basis of their input chemical diversity and used cluster-specific regressors to improve accuracy across varied reaction types. See, J. Chung, J. Li, A. I. Saimon, P. Hong and Z. Kong, *Sci. Rep.*, 2024, **14**, 12131. (b) In contrast, we employ the classification followed by regression (CFR) framework, which partitions reactions according to their output %*ee* labels to directly address class imbalance and data sparsity in the ART dataset. Thus, although both approaches share a two-step structure, they target distinct modelling challenges.

93. S. Ghosh, N. Jain and R. B. Sunoj, *ACS Catal.*, 2025, **15**, 20251–20269.

94. Using alternative boundaries as μ and (μ+σ), the maximum attainable classification accuracies are found to be inferior to that obtained with (μ−σ). Further details are provided in Table S74 in ESI.

95. See section 4.1 in ESI for more details on the DNN architecture and their performances in Tables S75-S76.

View Article Online
DOI: 10.1039/D5DD00483G

96. See Tables S81-S83 in ESI for details of performance of the AttentiveFP model for the CFR-major and minor classes.

97. Details of calculation of weighted average of RMSE and R^2 is provided in Table S84 in section 4.4 of ESI.

98. See Tables S77-S80 in ESI for more details about the ULMFiT model performances for the CFR-major and minor classes.

99. See Fig. S4 and section 6 of ESI for more details.

100. S. Riniker and G. A. Landrum, *J. Cheminform.*, 2013, **5**, 1.

101. S. Liao, X.-L. Sun and Y. Tang, *Acc. Chem. Res.*, 2014, **47**, 2260–2272.

102. S. Sadeghyan, *arXiv*, 2018, DOI: 10.48550/arXiv.1804.05092.

103. R. B. Sunoj, *Acc. Chem. Res.*, 2016, **49**, 1019–1028.

104. J. S. Harvey, S. P. Simonovich, C. R. Jamison and D. W. C. MacMillan, *J. Am. Chem. Soc.*, 2011, **133**, 13782–13785.

105. R. Connon, B. Roche, B. V. Rokade and P. J. Guiry, *Chem. Rev.*, 2021, **121**, 6373–6521.

106. The additional details on the extraction of chiral Box ligands from the PubChem library as well as the computational filtering protocol, are provided in Section 7 of the ESI.

107. The predicted %*ee* exceeding 99% is likely due to data sparsity and the use of linear output layer, which, in most of the DL regression tasks, inherently allow extrapolation beyond the training distribution. Mathematically, a linear output layer applies a transformation of the kind $y = Wx + b$, where W and b are the parameters learned by the model. Without constraints, if $W = 0.98$ and $b = 4.98$, an input ($x = 99$) could yield ($y = 102$), exceeding the expected range of 0 to 99. Thus, the model can return overestimated predictions, especially when trained on imbalanced datasets with sparse data near the extrema. See H. Shimakawa, A. Kumada, and M. Sato, *npj Comput. Mater.*, 2024, **10**, 11.

108. L. Simón and J. M. Goodman, *J. Am. Chem. Soc.*, 2008, **130**, 8741–8747.

View Article Online
DOI: 10.1039/D5DD00483G

109. D. J. Ager, A. H. M. de Vries and J. G. de Vries, *Chem. Soc. Rev.*, 2012, **41**, 3340–3380.

110. J. F. Teichert and B. L. Feringa, *Angew. Chem. Int. Ed. Engl.*, 2010, **49**, 2486–2528.

111. The details of the benchmarking on the full 1075 N,S-acetylation reaction dataset are provided in Section 5 of the ESI.

112. L. Cheng, P.-L. Shao, J. Lv, H. Xiao, Y. Sun, J. Yang, Z. Xu, M. Lv, G. Wang, S. Zhao, J. Li, Z. Jin, X. Tan, G. Xing and B. Zhang, *Nat. Comput. Sci.*, DOI:10.1038/s43588-025-00920-8.

113. (a) The statistical significance tests are carried out for an effective comparison of performance obtained with AttentiveFP-CI and AttentiveFP models for both N,S-acetylation and asymmetric hydrogenation reactions. For N,S-acetylation, the AttentiveFP model yielded RMSE = 8.57±0.82, R^2 = 0.91±0.02. (b) In the case of hydrogenation, the difference in RMSE between AttentiveFP-CI (11.00±0.52) and AttentiveFP (10.48±1.10) is not statistically significant ($p > 0.05$) (see Table S90 in ESI). However, the change in R^2 (AttentiveFP-CI: 0.52±0.03, $p = 0.0355$; AttentiveFP: 0.60±0.17) indicates a meaningful gain in performance.

114. The dataset is randomly divided into 70:10:20 train, validation, and test sets. The hyperparameter tuning for the ULMFiT model is performed on the validation set using the Optuna framework (ref. 71), and the resulting optimal hyperparameters are applied to the model for predictions on the test set. The model performance is reported in temrs of RMSE and R^2 obtained as the average over 30 different runs each using a randomly created train-test split.

115. The dataset is randomly divided into training, validation, and test sets into 70:10:20 ratio, and trained on five such independent random splits.

116. X. Yin, C.-Y. Hsieh, X. Wang, Z. Wu, Q. Ye, H. Bao, Y. Deng, H. Chen, P. Luo, H. Liu, T. Hou and X. Yao, *Research (Wash. D.C.)*, 2024, **7**, 0292.

View Article Online
DOI: 10.1039/D5DD00483G

117. X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou and X. Yao, *Research (Wash. D.C.)*, 2023, **6**, 0231.

118. E. Mathew, K. A. Emmitte and J. Liu, *ACS Omega*, 2025, **10**, 32968–32986.

119. C. Liu, Y. Sun, R. Davis, S. T. Cardona and P. Hu, *J. Cheminform.*, 2023, **15**, 29.

120. K. Ermanis, A. C. Colgan, R. S. J. Proctor, B. W. Hadrys, R. J. Phipps and J. M. Goodman, *J. Am. Chem. Soc.*, 2020, **142**, 21091–21101.

121. M. N. Grayson, *J. Org. Chem.*, 2021, **86**, 13631–13635.

View Article Online
DOI: 10.1039/D5DD00483G

Data and Code Availability

Data and codes related to this work are publicly available through our GitHub repository at https://github.com/alhqlearn/ART-AttentiveFP-CI.

A citable, versioned snapshot of the codebase is archived on Zenodo with an assigned DOI: https://doi.org/10.5281/zenodo.18256995