

PAPER

[View Article Online](#)
[View Journal](#) | [View Issue](#)Cite this: *Digital Discovery*, 2022, 1, 926Deep learning for enantioselectivity predictions in catalytic asymmetric β -C–H bond activation reactions†Ajnabiul Hoque^a and Raghavan B. Sunoj *^{ab}

The growth of catalytic asymmetric C–H bond activation reactions, as well as that in a seemingly disparate domain like machine learning (ML), has been unprecedented. In due cognizance of the potential of such technologies, we herein examine the utility of modern ML for one of the most recent Pd-catalyzed enantioselective β -C(sp³)–H functionalization reactions using chiral amino acid ligands. Focus is on a practically relevant small data regime problem consisting of 240 such reactions, wherein substituted cycloalkanes undergo enantioselective to form arylated/alkenylated products. The molecular descriptors from a mechanistically important metal–ligand–substrate complex are used for the first time to build various ML models to predict % ee. The Deep Neural Network (DNN) offers accurate predictions with a root mean square error (RMSE) of $6.3 \pm 0.9\%$ ee. The RMSEs of out-of-bag predictions on three different reactions, namely, the enantioselective arylation of cyclobutyl carboxylic amide, the alkenylation of isobutyric acid, and the C(sp³)–H arylation of free cyclopropylmethylamine are found to be 7.8, 5.0, and 7.1% ee. This high generalizability of the DNN model suggests that it could be deployed for planning and designing of asymmetric catalysis on small data settings. The application of explainable tools using feature attribution methods on the DNN has identified important molecular features that impact the % ee. The chemical insights gathered can effectively be employed in planning the synthesis of new molecular targets.

Received 6th August 2022
Accepted 2nd November 2022

DOI: 10.1039/d2dd00084a

rsc.li/digitaldiscovery

Introduction

Over the past few decades, asymmetric catalysis has emerged as an increasingly powerful platform for the construction of chiral molecules of importance to the domains of pharmaceuticals, agrochemicals, materials, natural products and so on.¹ The ever-growing requirements for efficient methods for chiral synthesis led to the development of newer catalysts and reaction protocols.² Quite a few enabling tools, such as the use of computational approaches, data-driven mathematical modeling, and machine learning (ML), have found applications in expediting catalysis research.³ While all these techniques have certain advantages when applied to a given class of reactions, generalizability across different catalytic reactions seems formidable at this time.⁴ In this context, studies directed toward examining the potential utility of various ML methods for catalytic reactions are very timely.⁵

The impact of ML on almost all domains of science and technology is increasingly more visible now.⁶ Prodigious applications in retrosynthetic planning,⁷ *de novo* drug design,⁸ inverse design of materials,⁹ reaction condition predictions¹⁰ *etc.*, are just a handful of applications of ML in the chemical space. The successful exploitation of ML pivotally depends on the availability of quality data.¹¹ Improved access to various molecular datasets has helped in developing good ML models in predicting molecular properties as well as reactions.¹² It should be noted that in the early phase of discovering a new asymmetric catalytic method, access to the data pertaining to only a few hundreds of reactions is affordable. Building ML models using a modest number of samples is therefore of higher practical relevance, although it might turn out to be an arduous task. In a proof-of-concept study from our laboratory comprising 368 asymmetric hydrogenation reactions, designed to predict the enantiomeric excess (% ee) using the random forest (RF) ML model, gave a test root mean square error (RMSE) of 8.4% ee.¹³ In their effort to predict the yield of the Buchwald–Hartwig reaction, the Doyle group has obtained an RMSE of 7.5 for the test set with the RF model.^{14a} These performance matrices, on rather complex reactivity problems involving multiple participating molecules besides a wide range of reaction conditions, can be considered as a baseline for ML models for chemical reactivity problems (*vide infra*).

^aDepartment of Chemistry, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India. E-mail: sunoj@chem.iitb.ac.in^bCentre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India† Electronic supplementary information (ESI) available: Details of chemical featurization, computational methods, various ML models, hyperparameter tuning, and other analyses. See DOI: <https://doi.org/10.1039/d2dd00084a>

While interesting demonstrations on the use of ML methods for yield and selectivity predictions are available,¹⁴ one of the most active domains in catalysis, namely C–H bond activation reactions,¹⁵ has not received much attention yet. Taking cognizance of the importance of this genre of catalytic reactions, we became interested in examining a set of most recent Pd catalyzed C(sp³)–H bond activation reactions¹⁶ promoted by mono-protected chiral amino acid (MPAA) ligands (Scheme 1A).^{19a} Apart from the synthetic potential of this reaction, it exhibits interesting mechanistic characteristics. For instance, (a) minor changes in the electronic and/or steric attributes of the α -side chain of the chiral MPAA ligand is known to affect large changes in the enantioselectivity,¹⁷ (b) the catalytic efficacy is known to be dependent on the nature of the transition metal catalyst, additive, base, and other factors (Scheme 1B).¹⁸ Thus, the identification and even tuning of important molecular features might help in finding better combinations of the reaction components capable of providing high enantioselectivity. The key would be to map the stereochemical outcome to a set of mechanistically important local and global molecular descriptors.

In this work, we examine the suitability of ML methods for predicting the enantioselectivity of β -C(sp³)–H functionalization reactions. A well-trained ML model can be deployed to identify untested reactions that could potentially offer higher enantioselectivities. The insights into how different molecular features of the chiral ligand, substrates, *etc.*, influence the outcome of the reaction could help in predicting the efficacies of various ligand scaffolds for a given pair of substrates.

Utilizing such ML protocols can save time and effort and can therefore accelerate the reaction discovery workflow.

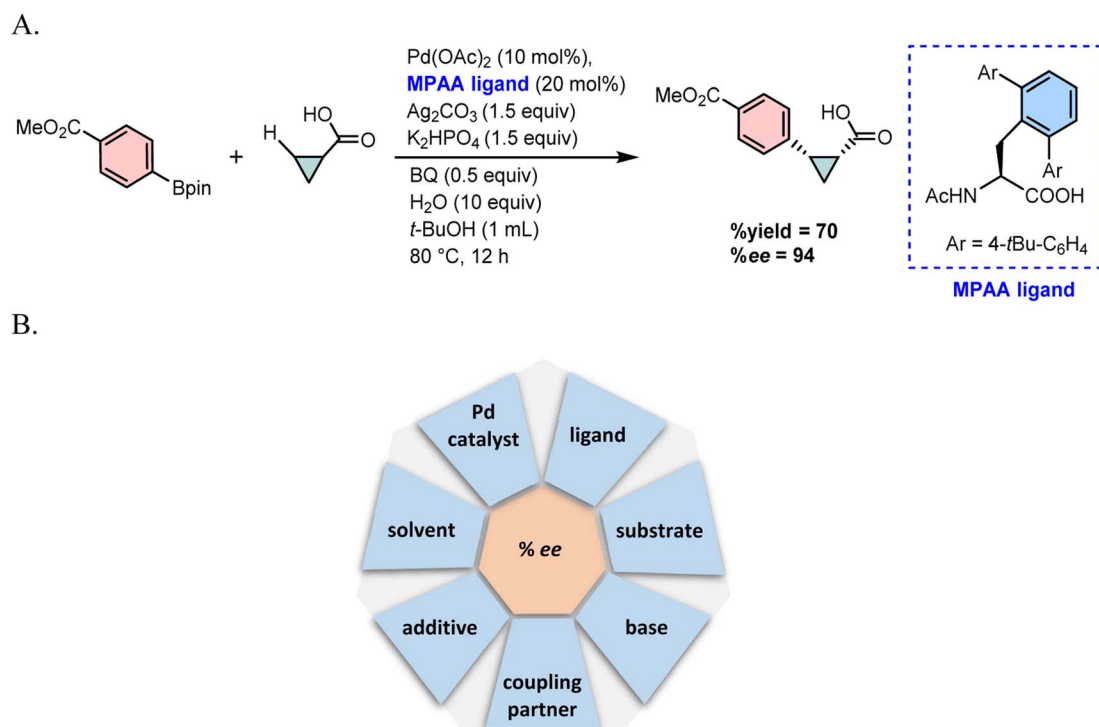
Results and discussion

For ease of comprehension, discussions are organized into (i) the preparation of the reaction dataset, (ii) the choice of a suitable reactivity model, (iii) chemical featurization, (iv) ML model building, (v) performance comparison between different ML models, (vi) out-of-bag prediction, and (vii) model interpretability and guidelines for future experiments.

Reaction dataset

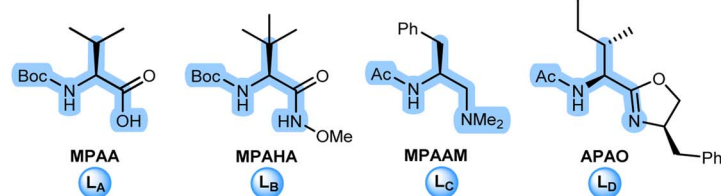
In the present study, manually curated data relevant to the reaction are first collected. The details of each reaction and the corresponding enantioselectivity from previously reported experimental studies have provided us with a total of 240 reactions.¹⁹ These reactions differ in terms of the nature of the ligands bound to the Pd center, catalyst precursor, substrates, coupling partner, additive, base, solvent, and the reaction conditions (temperature, time, relative proportion of ligand/base *etc.*).²⁰ Various combinations between these participating entities constitute a given reaction (sample), which in turn is associated with an output value expressed in % ee. The diversity of all the reaction components can be gathered from the generalized representations shown in Fig. 1 and 2 respectively for the chiral ligands and for the other components.

First, the diversity in the reaction stemming from the choice of the chiral ligand needs attention. As can be learned



Scheme 1 A representative example of (A) β -C(sp³)–H activation by the Pd(OAc)₂–MPAA catalytic system and (B) a schematic representation of various likely components that may influence the enantioselectivity (% ee).

(A) chiral ligands

 L_A = mono-N-protected amino acid (MPAA) L_C = mono-N-protected amino alkyl amine (MPAAM) L_B = mono-N-protected α -amino-O-alkyl hydroxamic acid (MPAHA) L_D = N-acyl-protected amino oxazoline (APAO)

(B) substituents on different ligand classes

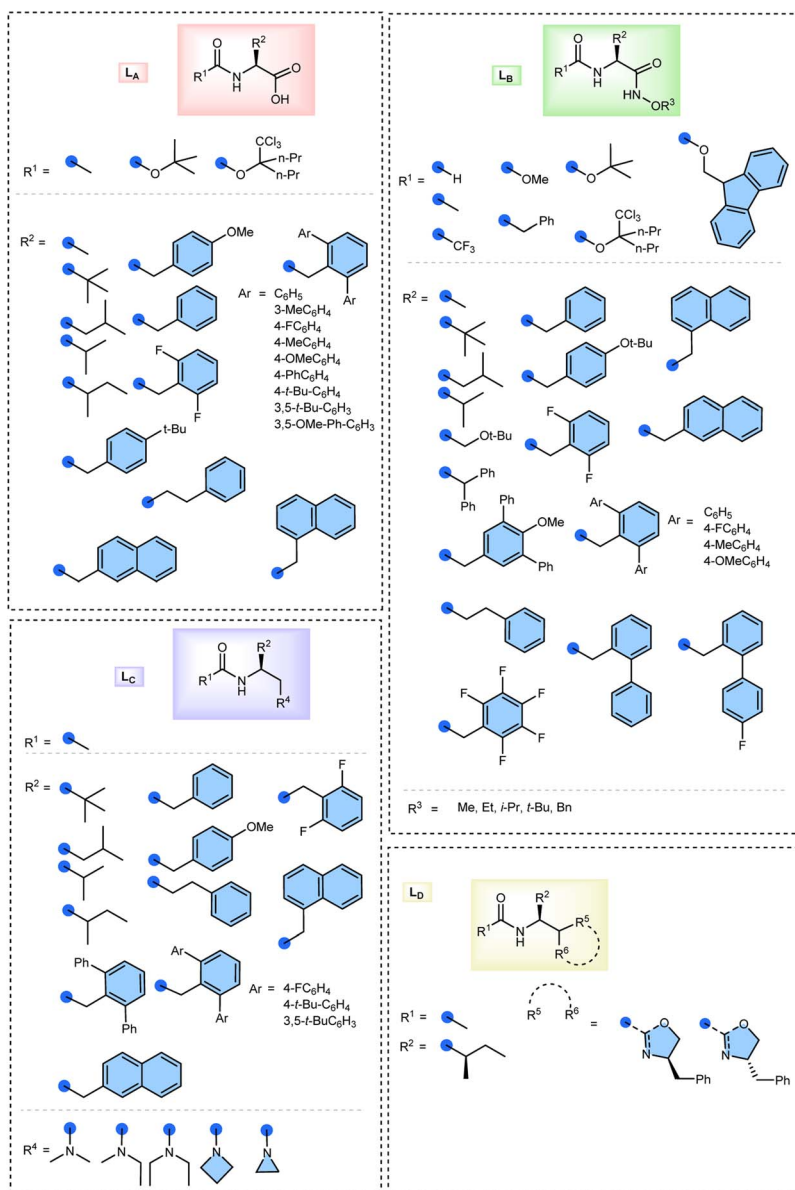


Fig. 1 (A) Generalized representations of individual chiral ligand families (highlighted atoms/bonds imply the common regions in each class of ligands), and (B) details of various substituents in each ligand family.



from Fig. 1A, there are four different chiral ligand families denoted as L_A , L_B , L_C , and L_D . In these ligands, the α -stereogenic carbon is decorated with different aryl or alkyl substituents besides the differences in the protecting groups (Ac, Boc, Fmoc, *etc.*), as shown in Fig. 1B. These variations together provide a total of 77 chiral ligands with interesting variations in their steric as well as electronic characteristics. Similarly, a range of five types of substrates are present in the dataset, which differ in terms of the nature of substituents and the weakly coordinating directing groups in them (Fig. 2). Some of these substrates

contain valuable cycloalkanes as found in certain bioactive molecules.²¹ As far as the coupling partners are concerned, aryl/vinyl iodides and aryl boronic acids, bearing a range of electron-donating/-withdrawing substituents, together form about 51 unique possibilities.

Similar diversity in the reaction space can be gleaned from the palladium catalyst precursor, base, additive, and solvent used (Fig. 2). It is important to note that each of these components have one or more distinctive role(s) in the mechanism of the reaction. For instance, nature of the Pd-bound ligands is

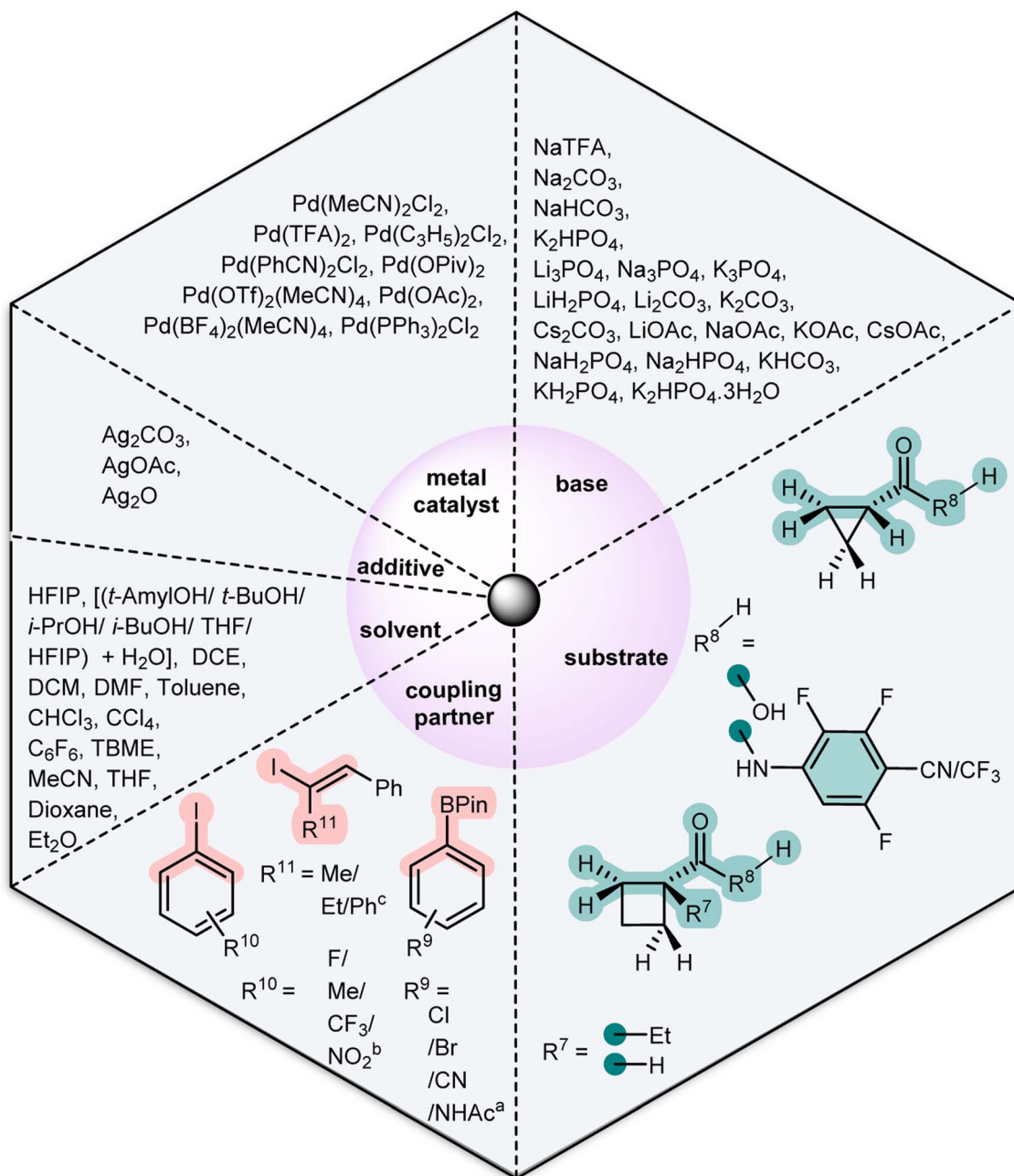


Fig. 2 Details of the substrates, coupling partner and other species involved in the reaction. The shared/common regions are highlighted. Only a representative set of examples of substituents in the coupling partners are shown here (^aR⁹ and ^{b,c}R¹¹). Full details can be found in Section 1.2 in the ESI.†



critical to the C–H activation step,²² whereas the base is generally regarded as involved in ligand de-protonation at some stage during the course of the reaction. Changing the additive/solvent can as well impact the reaction performance.²³ Hence, adequate representation of all these species in the dataset is important for meaningful featurization of the reaction (*vide infra*). All the above details indicate that the chosen dataset may be diverse enough for building a reasonably good and general ML model suitable for enantioselectivity predictions on asymmetric C(sp³)–H functionalization reactions.

Choice of a chemically relevant model for feature extraction

A simplified way of looking at this reaction is as a combination between the substrate and the coupling partner in the presence of the catalyst. The reaction performance is therefore expected to depend primarily on the nature of the catalyst, substrate, and coupling partner. In building an ML model, one can consider each species such as the ligand, catalyst precursor, substrate, *etc.*, as a free entity in its native form or as a composite intermediate that may carry high-level mechanistic information. The available mechanistic insights into similar C(sp³)–H functionalization reactions, including those with the MPAA ligand, could be made use of at this point.²⁴ For instance, a direct participation of the MPAA ligand in the C–H bond de-protonation²⁵ and the knowledge that C(sp³)–H bond activation is the enantioselectivity controlling step in an APAO (*N*-acyl-protected

amino oxazoline) ligand assisted borylation of amides²⁶ both convey the importance of the C–H bond activation step in these reactions.²⁷

In light of the above mechanistic insights, it seems cogent to consider that the preferential formation of the major enantiomer is more likely to be dictated by the energy difference between the diastereomeric transition states (TS) for the C–H bond activation step. It would therefore be of interest to examine the suitability of an intermediate, closer to that of the enantiocontrolling transition state, as a molecular entity from where chemically meaningful features could be collected for the downstream ML tasks. An intermediate, such as the pre-reacting complex for the C–H bond activation, might carry valuable mechanistic information, making it a suitable candidate for developing a reaction model. In doing so, molecular features can be extracted with much lower computing costs as compared to those required for locating the corresponding transition states. In this study, we have collected most of the features from a composite model, denoted as the metal–ligand–substrate (MLS) model as shown in Fig. 3, wherein the substrate undergoing the C–H bond activation and the chiral ligand are bound to the Pd center. It should, however, be noted that the fuller set of features in this study includes the molecular features of the coupling partner involved in the arylation/alkenylation step of the reaction and those from other species such as the base, additive and so on.

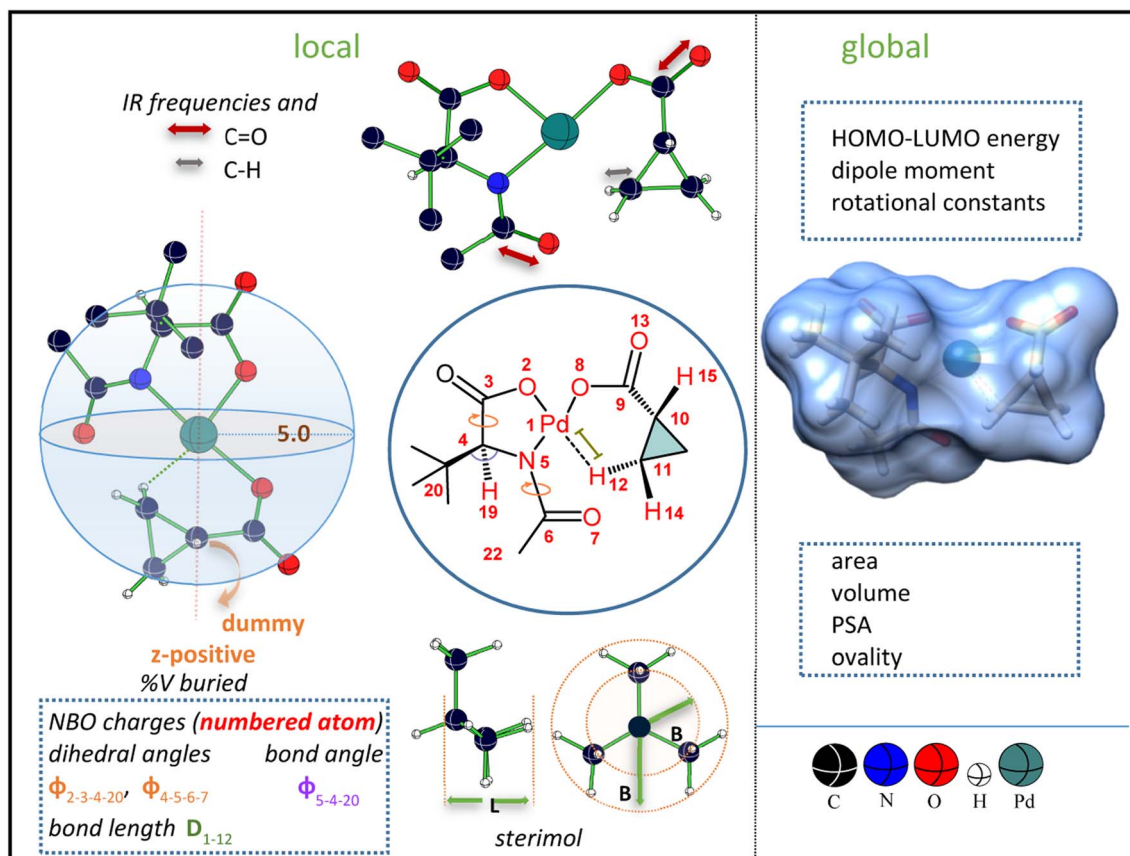


Fig. 3 Illustration of certain important local and global features in the metal–ligand–substrate (MLS) model employed for feature extraction.



The following aspects can be considered as the potential advantages of the composite MLS model used for the chemical featurization; (a) the bidentate binding of the chiral ligand to Pd makes it conformationally less flexible, (b) the directing group on the substrate (carboxylate in the example shown in Fig. 3) places the C–H bond in closer proximity to Pd, thereby helping capture the important agostic interaction likely between the Pd and the C–H bond, and (c) the ability of the C=O group to serve as an internal base in a concerted metalation de-protonation (CMD) mechanism is implicitly accommodated. These mechanistically important characteristics of the composite MLS model make it a reasonable alternative to the enantiocontrolling C–H activation TS. It would therefore be intriguing to see whether the features collected from the MLS model are effective in predicting the enantioselectivity of these C(sp³)–H bond activation reactions.

Chemical featurization using molecular descriptors

The success of building an ML model for reactivity predictions would demand good and adequate featurization of the participating components. These features should capture the structural, electronic, and global characteristics of the molecules, in addition to carrying useful mechanistic information.²⁸ We have used the DFT(B3LYP-D3) optimized geometries of the MLS model (Fig. 3) in the SMD continuum solvation model²⁹ for feature extraction. It has earlier been shown that the DFT derived molecular features can serve as the representation for obtaining good performance with various ML models for chemical reactions.^{13,14} In this approach, the molecular features across all the samples are likely to be of homogeneous quality, thus making the trends, subtleties, and variance of any given feature across the whole range of samples more reliable.

The steric effect of the substituents on the ligand is represented by the corresponding sterimol parameters (Fig. 3).³⁰ In other words, the differences in the spatial extent of various substituents of the protecting group/side chain of the catalyst/substrate (as seen in the MLS model) are captured using the sterimol parameters. In addition, we have used percent buried volume (% V) as another descriptor that provides a measure of the steric bulk of the ligand.³¹ Local features such as bond lengths and bond angles connecting important atoms are considered from the common regions of the substrates. Similarly, IR frequencies and intensities are included for improved representation of the electronic characteristics of the key bonds.³² Atomic descriptors, such as the natural charges and NMR chemical shifts, provide the site-specific electronic properties of specific atoms of high importance. A number of global descriptors (HOMO, LUMO energy, dipole moment, CPK area, *etc.*) are as well considered for adequate representation of the full molecular space. Likewise, the coupling partners are represented by using the local parameters collected from the highlighted region as well as by using certain global descriptors. For the other reaction components, such as the base, additive, and Pd pre-catalyst, we employed global parameters as listed in Fig. 3. To accommodate the experimental conditions in the data matrix, we have included reaction temperature, time, quantity

of ligand/base, solvent dielectric as descriptors. Together, each of the 240 reactions is described using 153 descriptors in total, thus giving a data matrix having a sample with feature dimensions of 240 × 153.³³ This kind of intrinsic featurization is expected to represent the rich and diverse samples as considered in our study and is more likely to be suitable for developing ML models. In view of previous chemical featurization strategies,^{13,14a} wherein the catalyst and substrates were considered as independent entities, we have also examined the performance of our ML models by using the non-interacting or independent participant approach.^{43a}

The ML protocol

With the target value expressed as % ee being a continuous variable, we have considered the present task as a regression problem. The unevenly distributed output values seen in the actual experiments indicate a class imbalance in the dataset, with a scarcity of samples in the low ee range. To address this issue, we have included synthetic data in the minority class in the 0–80 class boundary, following the standard recommendation of the synthetic minority oversampling technique (SMOTE).³⁴ The primary data, henceforth used for training various ML models, are inclusive of real and synthetic data.³⁵ The dataset is randomized and divided into training and test sets with an 80 : 20 ratio. The trained model with the optimal hyperparameter combination is deployed to make predictions on the test set.³⁶ The model performance on the test set is expressed in terms of the root mean square error (RMSE). It measures the error between the predicted % ee and experimentally reported % ee. To get an unbiased estimate of the generalization error, 100 different randomized test–train splits are constructed. The final RMSE is reported as the average RMSE over all the 100 runs using as many different test–train partitions.

Performance of different ML models

The choice of an ML model depends on the nature of the regression task at hand. Deep Neural Networks (DNNs) are a type of non-linear method that has recently gained popularity in data rich fields such as image analysis or natural language processing.³⁷ On relatively smaller data problems, DNN models have outperformed traditional ML methods for quantitative structure–activity relationship (QSAR) tasks.³⁸ Denmark and coworkers recently demonstrated that predictive modelling based on deep feed-forward neural networks (DNNs) was about as accurate as support vector machines (SVMs) in predicting % ee for chiral phosphoric acid-catalyzed thiol addition to *N*-acylimines.^{14b} In this work, we have employed a DNN as the primary ML model for predicting the % ee for Pd-catalyzed asymmetric functionalization through β-C(sp³)–H bond activation. The model architecture consists of 153 input neurons and one output, with five hidden layers with a composition of 33, 150, 400, 168, and 128 neurons.³⁹

We have developed independent DNN models for four different chiral ligand families such as **L_A**, **L_B**, **L_C**, and **L_D**, as shown in Fig. 1A.⁴⁰ For the MPAA ligand (**L_A**) with 69 samples,



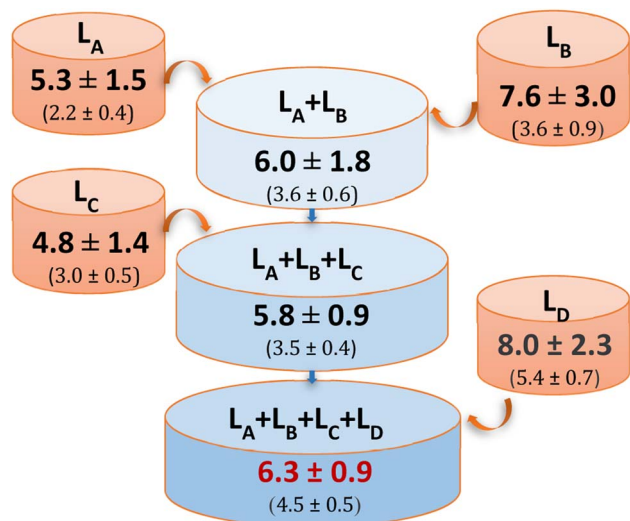


Fig. 4 The DNN model performance on different subsets and combined sets expressed using the corresponding test RMSEs in % ee units. The values shown in parentheses are the corresponding train RMSEs. The number of samples in each ligand subset is L_A (69), L_B (56), L_C (79), and L_D (36).

a test RMSE of $5.3 \pm 1.5\%$ ee is found, albeit with notable overfitting (Fig. 4). The test RMSEs of 7.6 ± 3.0 (L_B), 4.8 ± 1.4 (L_C), and 8.0 ± 2.3 (L_D) % ee are all found to be good for its practical deployment. The ability of the DNN algorithm in predicting well for the individual ligands prompted us to combine multiple ligand families together. This approach would improve both the diversity and the number of samples in the training set. Such composite models are likely to be broader in scope and more generalizable. The results of representative combinations such as L_A-L_B , $L_A-L_B-L_C$, and $L_A-L_B-L_C-L_D$ are provided in Fig. 4.⁴¹

In a practically more likely situation during reaction development, such as for enantioselective $C(sp^3)-H$ functionalization, the experimental results may be available early on, which belong to different chiral ligand scaffolds. Hence, the performance of the DNN model on such combinations of subsets would be of interest. Based on the similarity of the chiral ligand scaffolds, we have considered reactions drawn from L_A and L_B

families, to develop a new DNN model. In this two-ligand model, the test sets consist of randomly chosen reactions from either of the chiral ligand families. The trained DNN model predicted the % ee with a very good test RMSE of $6.0 \pm 1.8\%$ ee, which is encouraging.

Next, to construct a more inclusive set, we have considered all the reactions involving three chiral ligands L_A , L_B , and L_C together. Differences in the stereoelectronic characteristics of the ligand donor atoms make this dataset much more diverse (Fig. 1). Using this expanded sample space, the trained DNN model offered very good predictions with an RMSE of $5.8 \pm 0.9\%$ ee. One of the reasons for this improved performance of the combined model could possibly be due to an increase in the number of samples in the minority class. For instance, the number of samples in the 0–60% ee range was 2 and 3 respectively in the L_B and L_C subsets. In the composite set $L_A + L_B + L_C$, the presence of 13 samples in the minority class appears to strengthen the predictive capability of the model. In the final DNN model, the full data matrix, encompassing all the chiral ligand families, is taken into consideration.⁴² An impressive RMSE of $6.3 \pm 0.9\%$ ee is obtained for this unified model.⁴³ The key advantage of such a unified model is that it could predict well on samples belonging to any of the subsets.

In the final unified model consisting of a total of 240 samples, predictions on 48 randomized reactions belonging to the test set are made in every run.⁴⁴ Thus, thousands of predictions over 100 such independent runs would help assess the overall model performance. In this process, for every given sample, more than one predicted value becomes available. Analysis reveals that most of these predictions are in excellent agreement with the experimental % ee values (Fig. 5). For instance, the predicted values of 97% of samples are within 15 units of the actual values. In a typical run, an RMSE of 6.3% ee would mean that only 5 out of the 48 samples have an error of 10 units or higher. In the best run (RMSE of 4.1% ee), 47 samples are within 10 units of the experimental % ee value. These are promising indicators of the ability of the DNN model to efficiently learn on the data for the enantioselective $C(sp^3)-H$ functionalization reaction.

A comparison of the DNN performance with the other commonly used ML models in the domain of chemical

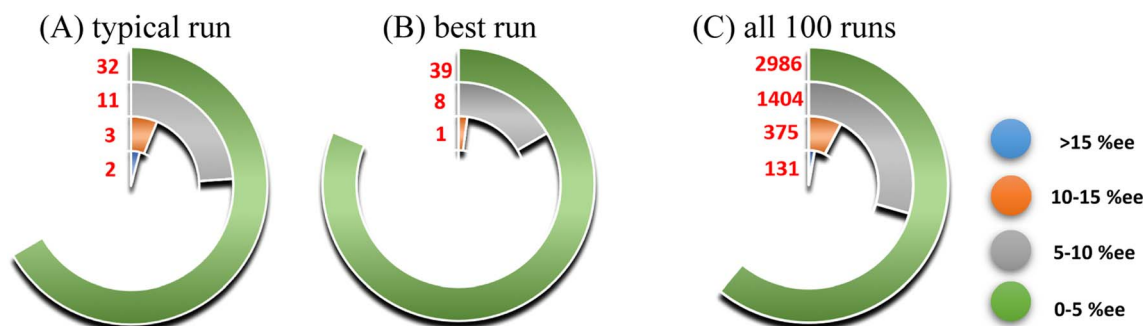


Fig. 5 Pie charts representing the actual difference between the experimental and predicted % ee using the unified DNN model for the 48 test samples as seen in (A) the best run with an RMSE of 4.1% ee, and (B) in a typical run of RMSE 6.3% ee (which is the closest to the overall performance of the model), and (C) in all 100 runs. The numbers in red color shown adjacent to the respective colored strips (for a given range of quantitative agreement with the experimental % ee) indicate the total samples in a given interval.



reactivity is as well made here. Interestingly, perusal of Fig. 6 reveals that the DNN offered superior performance to the other such alternatives, given the expected generalization issues due to over-fitting arising from the use of small data with an inherent class imbalance. Different ML models can be compared using their test RMSEs, which are as follows; k -nearest neighbors (kNN, 6.4 ± 1.0), random forest (RF, 6.3 ± 0.7), gradient boosting (GB, 6.5 ± 0.8), Gaussian process regression with RBF kernel (GPR_{RBF}, 6.3 ± 0.9), and decision tree (DT, 7.5 ± 1.3).⁴⁵ Despite the comparable performance of the RF model as well as a slightly lower overfitting than the DNN, we consider the DNN as our best model here. It is instructive to note that the output values predicted by the RF were a result of two-level averaging.⁴⁶ An upper bound of the predicted % ee by the RF model is found to fold into a value of 91 due to such averaging, whereas a higher predicted value could be seen in the DNN.

After having established the ability of the DNN model built using the physical organic molecular descriptors obtained from the metal–ligand–substrate (MLS) complex involved in diverse enantioselective C(sp³)–H functionalization reactions, we wanted to examine the importance of chemical featurization. On this front, we have carried out three additional experiments using the DNN algorithm.⁴⁷ First, the output values are randomly shuffled, such that the respective features are not associated with their true output. When trained with such a scrambled dataset, a test set RMSE of $17.8 \pm 3.1\%$ ee is obtained. Second, we have generated a dataset in which the values of the features are replaced with random numbers that follow a normal distribution (*i.e.*, mean of 0 and standard deviation of 1). Such generated numbers would have neither any chemical significance nor similarity to the chemical descriptors. A test RMSE of $18.2 \pm 4.1\%$ ee obtained in these runs indicates a worsened performance as compared to that obtained using the actual chemical descriptors. In the third control

experiment, we have used one-hot encoded vectors (OHEs) as the descriptor, which simply connotes the presence or absence of a chemical entity in any given reaction.⁴⁸ The trained algorithm with this binary encoded dataset yields a test RMSE of $15.0 \pm 4.1\%$ ee. The strikingly lower performances, as noted with the above-mentioned alternative featurization techniques as compared to the chemically meaningful physical organic descriptors, implicitly endorses the importance of the latter for studying the chemical reactivity problem at hand.

As another approach toward ascertaining the sufficiency of our chemical featurization, we have employed an unsupervised learning technique on the feature space. A correlation analysis of the features is carried out, without exposing the output values, with an intent of reducing the number of features considered in the model building. The removal of the correlated features with correlation coefficients of 0.9 or higher resulted in 114 features from a total of 153.⁴⁹ The freshly trained DNN model with these 114 features returned a test RMSE of $8.6 \pm 1.8\%$ ee, which is obviously higher than $6.3 \pm 0.9\%$ ee obtained with the full feature space. With further reduction in the number of features to 83 (by setting the correlation coefficient down to 0.8 or higher), the test RMSE deteriorated to $8.3 \pm 1.4\%$ ee. Similarly, we have performed a couple of additional experiments, wherein DNN models are built using a subset of features obtained by deleting a handpicked set of features that are expected to be of higher mechanistic importance. For instance, atomic descriptors of the key atoms of the chiral ligand/substrate, which are bound to Pd, as well as those belonging to the bonds around the site of reaction, are kept aside from the feature matrix. The test RMSEs of these DNN models are found to be much higher (10% ee).⁵⁰ A comparatively poorer performance noted upon various feature reduction methods as compared to those with the full feature list could be considered as evidence that the chosen number of chemical descriptors provides adequate features for describing the reactivity problem as studied.

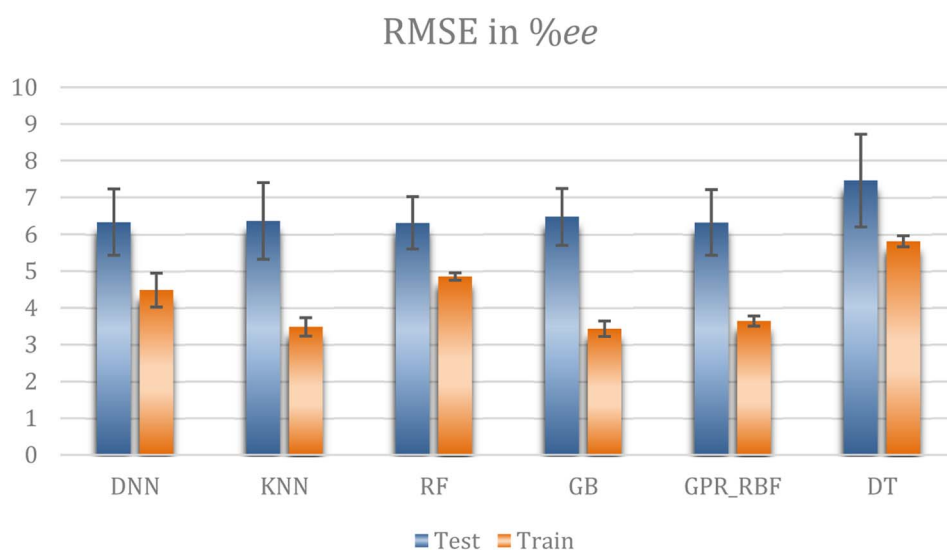


Fig. 6 Performance comparison between different ML models, RMSEs for different ML models such as the deep neural network (DNN), k -nearest neighbor (k -NN), random forest (RF), gradient boosting (GB), GPR with the RBF kernel (GPR_{RBF}), and decision tree (DT). The error bars denote the corresponding standard deviation.

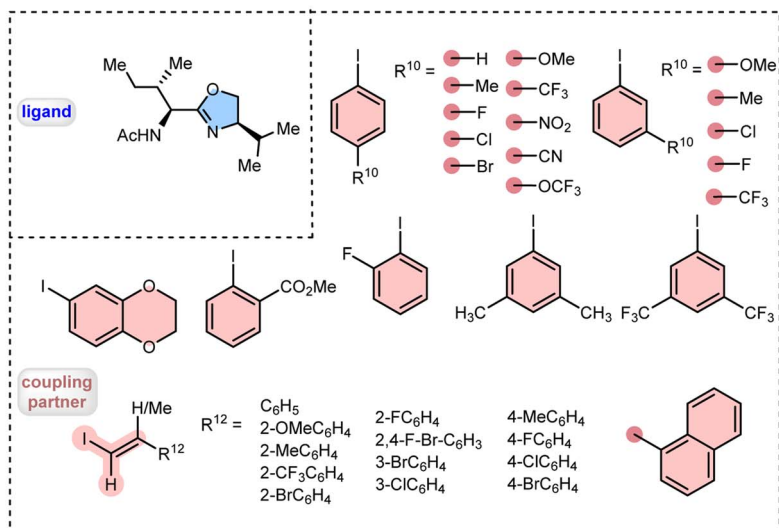


Prediction on out-of-bag samples

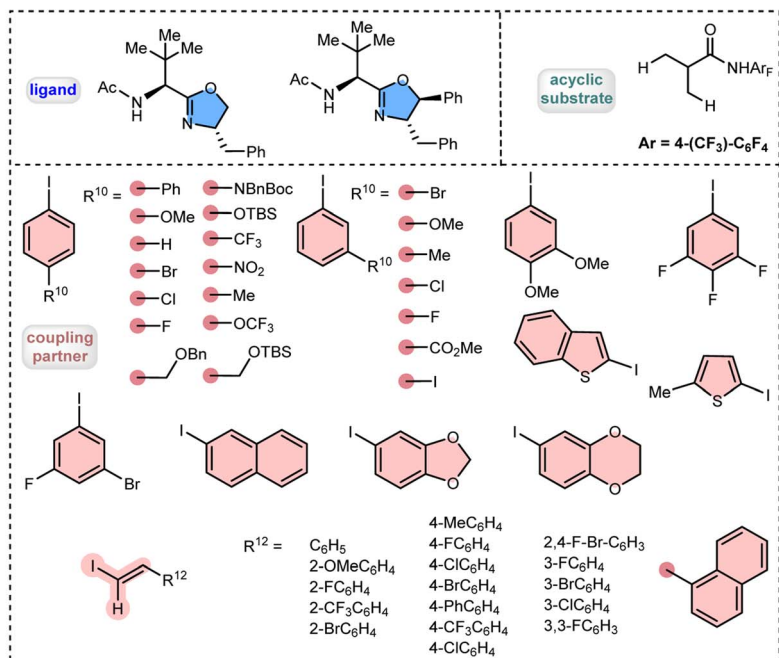
Internal validation, such as cross-validation, does not always ensure the quality of an ML model. One of the standard

practices in determining the efficacy and model generalizability is to test on out-of-bag (OOB) samples. This entails the use of separate sample sets (which are not present in the training and

Set-1



Set-2



Set-3

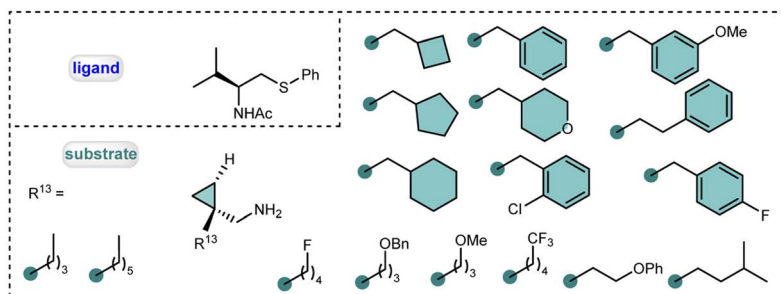


Fig. 7 Details of various reacting partners for all three sets of out-of-bag samples considered for examining the generalizability of the DNN model.



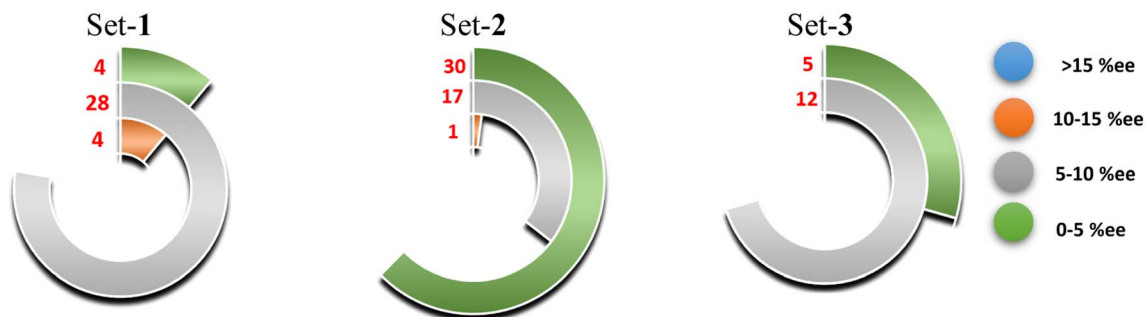


Fig. 8 The goodness of the predicted % ee expressed in terms of the number of samples across different ranges of % ee for Set-1, Set-2, and Set-3 out-of-bag samples. The numbers in red color shown adjacent to the respective colored strips (for a given range of quantitative agreement with the experimental % ee) indicate the total samples in a given interval.

test sets) to validate the performance of the ML model. We used three different sets of Pd(II)-catalyzed enantioselective β -C(sp³)-H bond activation reactions as the OOB samples, details of which are shown in Fig. 7.⁵¹ Set-1 contains 36 new asymmetric arylation/vinylation reactions of cyclobutyl carboxylic amide,^{19d} while 48 arylation/alkenylation reactions of isobutyric acid constitute Set-2, both involving the chiral APAQ ligands.⁵² In Set-3, a total of 17 γ -C(sp³)-H arylation reactions of free cyclopropylmethylamine using mono-N-protected aminoethyl thioether (MPATHio) as the ligand are considered.⁵³ It may be recollected that the original DNN model has been developed using reactions containing cyclic substrates and different variants of amino acids as the chiral ligand (**L_A**, **L_B**, **L_C**, and **L_D** as shown in Fig. 1), which are different from the OOB sets with newer acyclic substrate, a different variant of the MPAA ligand with S-donating atoms besides the alkenyl coupling partners.

A comparison of these OOB samples and those in the original training set is prudent at this juncture to highlight the key details of the sample diversity. Set-1 reactions undergoing the β -C(sp³)-H bond activation differ from the training set primarily in the nature of the coupling partner as well as the chiral ligand (Fig. 7), latter bearing an *i*-Pr group on the stereogenic center of the oxazoline ring (compare with **L_D** in Fig. 1B). In Set-2 acyclic substrates are involved while those in the training set were all cyclic substrates. In addition, the newer coupling partners in Set-2 consist of heteroaryl iodides, *para*-substituted aryl iodides, and a number of *E*-styrenyl iodides all contributing to the desirable sample diversity in the OOB test (Fig. 7). More interestingly, in the case of Set-3, both the ligand and the substrates are very new to the trained model. A comparison between the initially used training samples (shown in Fig. 1) and that in the OOB set in Fig. 7 reveals the contrast. For instance, the chiral ligand is a mono-N-protected thioether and the substituents on the substrate cyclopropylmethylamine contain various aliphatic chains as well as aryl group decorations on such aliphatic chains (*e.g.*, **R¹³** as shown in Fig. 7C). The above aspects of the diversity of OOB test samples that are used in this study, as compared to the original training set samples, engenders a similar confidence to that of a prospective validation of our ML model. Furthermore, current practices in the use of ML in chemical space indicate that model generalizability checks by

way of using OOB tests are widely seen⁵⁴ compared to a potential alternative of experimental verification of ML predictions.⁵⁵

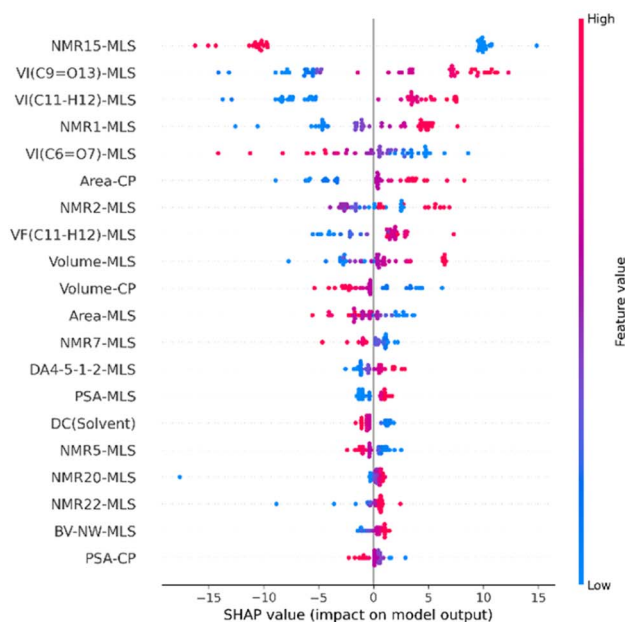
The DNN model trained using all the reactions are deployed for predicting on OOB samples.⁴² Interestingly, very good predictions for Set-1, Set-2 and Set-3 with an RMSE of 7.8, 5.0 and 7.1% ee respectively are obtained. Given the notable diversity of samples in Set-3 compared to those in the training set, we wondered whether improved representation of such samples in the training might become beneficial. To examine this hypothesis, we have moved just about four randomized samples from the 17 OOB samples to train a new DNN model. This new model offered a remarkably superior performance with an RMSE of 1.95% ee for the remaining 13 OOB samples.⁵⁶ A summary of the quality of predictions on the OOB samples across all three sets can be gleaned from Fig. 8. It can be noted that the difference between the experimental and predicted % ee for a large majority of samples is <10% ee units. The ability of the DNN model in providing good quality predictions for unseen samples demonstrates its potential to serve as an important tool for assessing the performance for newer reactions prior to their actual experimental validation. In summary, the DNN model built on mechanistically meaningful molecular descriptors could be efficiently employed in real-world applications such as for enantioselectivity predictions on an important contemporary reaction like the Pd-catalyzed asymmetric β -C(sp³)-H bond activation.

ML model interpretability and clues for planning future experiments

With these promising results, it would be more interesting to rationalize ML model decisions. Extracting the key insights from the model would gain better acceptance of ML-driven predictions in asymmetric catalysis.⁵⁷ Herein, we used a conceptual appealing feature attribution method known as Shapley additive explanations (SHAP), for decoding the complex DNN model employed in our study.⁵⁸ We have utilized the ability of the SHAP⁵⁹ method in recognizing some of the geometric and electronic features of the MLS system or their combinations that may have a significant influence on the desired output expressed in % ee.⁶⁰



(A) SHAP summary plot



(B) SHAP bar plot

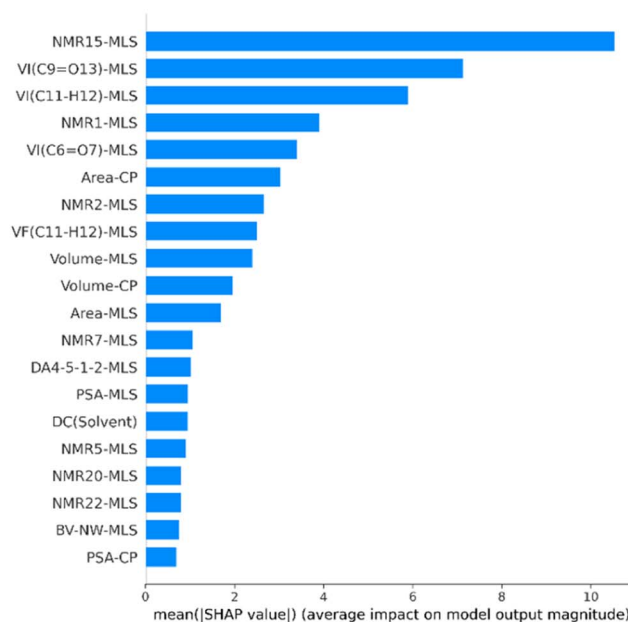


Fig. 9 Plots representing global feature importance as obtained using the SHAP method. (A) Summary plot where each point on the plot represents a SHAP value for a feature that is attributed to a particular reaction. The red, blue and purple colors respectively denote high, low, and intermediate values of features. The y-axis carries the feature list in the order of decreasing importance from top to bottom, whereas the SHAP values are on the x-axis. Positive SHAP values indicate model output promotion, while negative values indicate model output repression. A SHAP value of +5 for a specific reaction indicates that the value of that feature increases the output of the model by 5% ee. (B) Bar plot represents the average SHAP value across all samples for all features (full details of various features can be found in Fig. 3).

We have considered the best test run (RMSE of 4.1% ee) of the trained DNN model built on the molecular features collected from the MLS system. Fig. 9A depicts the distribution of the SHAP interaction values across the top 20 most effective features for the 48 test samples. This summary plot describes the importance of features as well as their effects. It can be noticed that an electronic parameter such as the NMR chemical shift of the first atom of the α -substituent of the substrate (NMR15-MLS) appears as the most important feature influencing the % ee. This parameter may be tuned by way of introducing suitable α -substituents on the substrate. Intuitively appealing are the high ranks found for the other top features such as the vibrational intensities of (a) the directing group (VI(C9=O13)-MLS), (b) the very β -C(sp³)-H bond of the substrate (VI(C11-H12)-MLS) that undergoes the C-H bond activation, and (c) the C=O bond of the N-protecting group of the chiral ligand (VI(C6=O7)-MLS). The emergence of these features as the prominent ones implies that our MLS model is able to capture how the reaction outcome depends on a set of comprehensible electronic factors arising from the protecting group, directing group, and nature of the Pd-bound chiral ligand.⁶¹

Another aspect of this study is to examine how the output will be affected by the magnitude of various features. A positive SHAP value indicates promotion, while a negative value conveys repression. For example, a substrate with a higher intensity for the C9=O13 or C11-H12 stretching (red points in Fig. 9A) is more likely to give high % ee. Notably, top 20 of these features

are not just from the MLS entity, but also include those of the coupling partner, and even the solvent. This suggests that the features of the other reaction partners are equally important and therefore a cumulative influence is most likely to be seen in the desired % ee. The average impact of each feature and their mean SHAP values are shown in Fig. 9B. The average SHAP value of about 8 and 7 units respectively for VI(C9=O13)-MLS and VI(C11-H12)-MLS suggests that a change in % ee by 8 and 7 units is likely by changing these features. Modifications to other features such as NMR15-MLS, VI(C6=O7)-MLS, *etc.*, from the list of important features shown in Fig. 9B could similarly help in making an informed choice of the reactants/chiral ligand toward designing new reactions.⁶² We have provided a complete and simplified workflow for planning new experiments that can make best use of ML-enabled asymmetric catalysis. Following such a protocol, tuning of these key features, particularly during the reaction development or for furthering the scope of this family of reaction, could be beneficial. The recommended approach can save time and sources in choosing an appropriate chiral ligand and potential substrates for a desired target molecule.

Conclusions

We have developed machine learning protocols for a series of synthetically important Pd-catalyzed asymmetric β -C(sp³)-H bond activation reactions promoted by chiral mono-N-protected amino acid (MPAA) ligands. The reaction is well known for its



ability to couple cycloalkanes with aryl coupling partners, through an enantioselective β -C(sp³)-H bond activation as the key mechanistic step, wherein the MPAA ligand serves as the primary source of chirality. The machine learning models, built using the quantum mechanically derived molecular descriptors of the respective catalyst-substrate complexes of hundreds of reactions, have been able to offer very good performance in predicting the enantioselectivity of these asymmetric β -C(sp³)-H bond activation reactions. In particular, the deep neural network (DNN) performed significantly well with a test RMSE of $6.3 \pm 0.9\%$ ee. The model generalizability has been assessed by predicting on several unseen reactions, drawn from three different sets of out-of-bag samples comprising new and diverse substrates, coupling partners, and chiral ligands. The trained DNN model offered promising predictive capabilities as indicated by the test RMSEs of 7.8, 5.0 and 7.1% ee respectively for enantioselective arylation of cyclobutyl carboxylic amide, alkenylation of isobutyric acid, and γ -C(sp³)-H arylation of free cyclopropylmethylamine. Thus, the deployment of the DNN model, built on the initial set of substrates/coupling partners/chiral ligands, can serve as a valuable guide toward identifying the combinations of substrates/coupling partners/chiral ligands that are likely to offer high enantioselectivities. We further illustrated that feature attribution methods can help in understanding how important molecular features impact the % ee and how such chemical insights obtained from the DNN can be made use of in planning the synthesis of novel target compounds. The promising combinations could then be subjected to experimental validation, which in turn would help expedite the reaction discovery.

Data availability

Data and codes related to this work are publicly available through our Github repository at <https://github.com/alhqlern/ML-for-Asymmetric-C-sp3-H-Reaction>.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

Generous computing time from the *SpaceTime* supercomputing at IIT Bombay is acknowledged.

References

- (a) B. M. Trost, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 5348–5355; (b) M. S. Taylor and E. N. Jacobsen, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 5368–5373; (c) S. Mukherjee, J. W. Yang, S. Hoffmann and B. List, *Chem. Rev.*, 2007, **107**, 5471–5569.
- (a) G. Schneider, *Nat. Rev. Drug Discovery*, 2018, **17**, 97–113; (b) T. Melanie and M. D. Burke, *Angew. Chem., Int. Ed.*, 2018, **57**, 4192–4214.
- (a) P.-O. Norrby, in *Transition State Modeling for Catalysis*, ACS Symposium Series 721, ed. D. G. Truhlar and K. Morokuma, American Chemical Society, Washington, DC, 1999, pp. 163–172; (b) R. R. Knowles and E. N. Jacobsen, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 20678–20685; (c) W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622–1637; (d) Y. Reddi, C. C. Tsai, C. M. Avila, F. D. Toste and R. B. Sunoj, *J. Am. Chem. Soc.*, 2019, **141**, 998–1009; (e) R. B. Sunoj, *Acc. Chem. Res.*, 2016, **49**, 1019–1028.
- K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P.-O. Norrby, *Nat. Rev. Chem.*, 2021, **5**, 240–255.
- (a) 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; (b) S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman and A. G. Doyle, *Science*, 2021, **374**, 301–308; (c) A. L. Haywood, J. Redshaw, M. W. D. Hanson-Heine, A. Taylor, A. Brown, A. M. Mason, T. Gärtner and J. D. Hirst, *J. Chem. Inf. Model.*, 2022, **62**, 2077–2092.
- (a) K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555; (b) F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, *Annu. Rev. Phys. Chem.*, 2020, **71**, 361–390; (c) S. Chmiela, H. E. Sauceda, K.-R. Müller and A. Tkatchenko, *Nat. Commun.*, 2018, **9**, 3887; (d) Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins and Q. Zhou, *J. Chem. Inf. Model.*, 2021, **61**, 3197–3212.
- (a) T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651; (b) S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937; (c) B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
- M. Popova, O. Isayev and A. Tropsha, *Sci. Adv.*, 2018, **4**, eaap7885.
- (a) J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, *Matter*, 2019, **1**, 1370–1384; (b) B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
- J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
- E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
- (a) Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530; (b) P. Friederich, G. dos Passos Gomes, R. D. Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601; (c) K. Jorener, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175; (d) M. Das, P. Sharma and R. B. Sunoj, *J. Chem. Phys.*, 2022, **156**, 114303; (e) S. Singh and R. B. Sunoj,



- Digital Discovery*, 2022, **1**, 303–312; (f) S. Singh and R. B. Sunoj, *iScience*, 2022, **25**, 104661.
- 13 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.
 - 14 (a) D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. Doyle, *Science*, 2018, **360**, 186–190; (b) A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
 - 15 (a) B. A. Arndtsen, R. G. Bergman, T. A. Mobley and T. H. Peterson, *Acc. Chem. Res.*, 1995, **28**, 154–162; (b) R. H. Crabtree, *Chem. Rev.*, 2010, **110**, 575; (c) X. Chen, K. M. Engle, D.-H. Wang and J.-Q. Yu, *Angew. Chem., Int. Ed.*, 2009, **48**, 5094–5115; (d) I. A. I. Mkhalid, J. H. Barnard, T. B. Marder, J. M. Murphy and J. F. Hartwig, *Chem. Rev.*, 2010, **110**, 890–931; (e) O. Daugulis, H.-Q. Do and D. Shabashov, *Acc. Chem. Res.*, 2009, **42**, 1074–1086; (f) S. K. Sinha, S. Guin, S. Maiti, J. P. Biswas, S. Porey and D. Maiti, *Chem. Rev.*, 2022, **122**, 5682–5841.
 - 16 (a) B.-F. Shi, N. Maugel, Y.-H. Zhang and J.-Q. Yu, *Angew. Chem., Int. Ed.*, 2008, **47**, 4882–4886; (b) M. Wasa, K. M. Engle, D. W. Lin, E. J. Yoo and J.-Q. Yu, *J. Am. Chem. Soc.*, 2011, **133**, 19598–19601; (c) K.-J. Xiao, D. W. Lin, M. Miura, R.-Y. Zhu, W. Gong, M. Wasa and J.-Q. Yu, *J. Am. Chem. Soc.*, 2014, **136**, 8138–8142; (d) L. Hu, P.-X. Shen, Q. Shao, K. Hong, J. X. Qiao and J.-Q. Yu, *Angew. Chem., Int. Ed.*, 2019, **58**, 2134–2138; (e) K. S. L. Chan, M. Wasa, L. Chu, B. N. Laforteza, M. Miura and J.-Q. Yu, *Nat. Chem.*, 2014, **6**, 146; (f) Q. Shao, J. He, Q.-F. Wu and J.-Q. Yu, *ACS Catal.*, 2017, **7**, 7777–7782; (g) K. S. L. Chan, H.-Y. Fu and J.-Q. Yu, *J. Am. Chem. Soc.*, 2015, **137**, 2042–2046; (h) Q. Shao, Q.-F. Wu, J. He and J.-Q. Yu, *J. Am. Chem. Soc.*, 2018, **140**, 5322–5325.
 - 17 (a) T. G. Saint-Denis, R.-Y. Zhu, G. Chen, Q.-F. Wu and J.-Q. Yu, *Science*, 2018, **359**, eaao4798; (b) Q. Shao, K. Wu, Z. Zhuang, S. Qian and J.-Q. Yu, *Acc. Chem. Res.*, 2020, **53**, 833–851.
 - 18 T. Rogge, N. Kaplaneris, N. Chatani, J. Kim, S. Chang, B. Punji, L. L. Schafer, D. G. Musaev, J. Wencel-Delord, C. A. Roberts, R. Sarpong, Z. E. Wilson, M. A. Brimble, M. J. Johansson and L. Ackermann, *Nat. Rev. Methods Primers*, 2021, **1**, 43.
 - 19 (a) P.-X. Shen, L. Hu, Q. Shao, K. Hong and J.-Q. Yu, *J. Am. Chem. Soc.*, 2018, **140**, 6545–6549; (b) L. Hu, P.-X. Shen, Q. Shao, K. Hong, J. X. Qiao and J.-Q. Yu, *Angew. Chem., Int. Ed.*, 2019, **58**, 2134–2138; (c) K.-J. Xiao, D. W. Lin, M. Miura, R.-Y. Zhu, W. Gong, M. Wasa and J.-Q. Yu, *J. Am. Chem. Soc.*, 2014, **136**, 8138–8142; (d) Q.-F. Wu, X.-B. Wang, P.-X. Shen and J.-Q. Yu, *ACS Catal.*, 2018, **8**, 2577–2581.
 - 20 See Section 1 in the ESI† for the full list of ligands, substrates, coupling partners, additives, etc.
 - 21 (a) T. V. Hansen and Y. Stenström, Naturally Occurring Cyclobutanes, in *Organic Synthesis: Theory and Applications*, ed. T. Hudlicky, Elsevier Science, Oxford, U.K., 2001, vol. 5, p. 1; (b) W. R. Gutekunst and P. S. Baran, *J. Am. Chem. Soc.*, 2011, **133**, 19076–19079; (c) W. R. Gutekunst and P. S. Baran, *J. Org. Chem.*, 2014, **79**, 2430–2452; (d) R. A. Panish, S. R. Chintala and J. M. Fox, *Angew. Chem., Int. Ed.*, 2016, **55**, 4983–4987.
 - 22 T. Gensch, M. N. Hopkinson, F. Glorius and J. Wencel-Delord, *Chem. Soc. Rev.*, 2016, **45**, 2900–2936.
 - 23 (a) M. Anand, R. B. Sunoj and H. F. Schaefer, *ACS Catal.*, 2016, **6**, 696–708; (b) M. Anand, R. B. Sunoj and H. F. Schaefer, *J. Am. Chem. Soc.*, 2014, **136**, 5535–5538; (c) L. Hong, W. Sun, D. Yang, G. Li and R. Wang, *Chem. Rev.*, 2016, **116**, 4006–4123.
 - 24 (a) D. Balcells, E. Clot and O. Eisenstein, *Chem. Rev.*, 2010, **110**, 749–823; (b) D. L. Davies, S. A. Macgregor and C. L. McMullin, *Chem. Rev.*, 2017, **117**, 8649–8709; (c) R. Giri, Y. Lan, P. Liu, K. N. Houk and J.-Q. Yu, *J. Am. Chem. Soc.*, 2012, **134**, 14118–14126; (d) D. G. Musaev, A. Kaledin, B.-F. Shi and J.-Q. Yu, *J. Am. Chem. Soc.*, 2012, **134**, 1690–1698; (e) B. E. Haines, J.-Q. Yu and D. G. Musaev, *ACS Catal.*, 2017, **7**, 4344–4354; (f) Y.-F. Yang, X. Hong, J.-Q. Yu and K. N. Houk, *Acc. Chem. Res.*, 2017, **50**, 2853–2860.
 - 25 (a) G.-J. Cheng, Y.-F. Yang, P. Liu, P. Chen, T.-Y. Sun, G. Li, X. Zhang, K. N. Houk, J.-Q. Yu and Y.-D. Wu, *J. Am. Chem. Soc.*, 2014, **136**, 894–897; (b) G.-J. Cheng, P. Chen, T.-Y. Sun, X. Zhang, J.-Q. Yu and Y.-D. Wu, *Chem.-Eur. J.*, 2015, **21**, 11180–11188.
 - 26 Y.-Y. Xing, J.-B. Liu, Q.-M. Sun, C.-Z. Sun, F. Huang and D.-Z. Chen, *J. Org. Chem.*, 2019, **84**, 10690–10700.
 - 27 Additional mechanistic details on how different Pd-bound ligands are involved in the catalytic cycle are provided in Fig. S1 in the ESI†
 - 28 F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154–6168.
 - 29 (a) M. J. Frisch, *et al.*, *Gaussian 09, D.01*, Gaussian, Wallingford, CT, 2009; (b) S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104–154119; (c) P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222; (d) V. A. Rassolov, J. A. Pople, M. A. Ratner and T. L. Windus, *J. Chem. Phys.*, 1998, **109**, 1223–1229.
 - 30 (a) A. Verloop, *Drug Design*, ed. E. J. Ariens, Academic Press, New York, 1976, vol. III; (b) A. V. Brethome, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313–2323; (c) Three sterimol parameters (B_1 , B_5 , and L) are commonly used, which capture the size of the substituents with respect to the axis of attachment. The shortest distance perpendicular to the axis of attachment is the B_1 parameter while the longest distance is taken as the B_5 parameter. The maximum distance from the point of attachment is the L parameter.
 - 31 L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo, *Nat. Chem.*, 2019, **11**, 872–879.
 - 32 (a) A. Milo, E. N. Bess and M. S. Sigman, *Nature*, 2015, **507**, 210–214; (b) C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398–2412; (c) Y. Park, Z. L. Niemeyer, J.-Q. Yu and M. S. Sigman, *Organometallics*, 2018, **37**, 203–210.
 - 33 See Tables S4 and S5 in the ESI† for details of the feature space.
 - 34 (a) N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *J. Artif. Intell.*, 2002, **16**, 321–357; (b)



- E. Sara, C. Laila and I. Ali, The Impact of SMOTE and Grid Search on Maintainability Prediction Models, in *IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 2019, pp. 1–8, DOI: [10.1109/AICCSA47632.2019.9035342](#); (c) L. Demidova and I. Klyueva, SVM Classification: Optimization with the SMOTE Algorithm for the Class Imbalance Problem, in *6th Mediterranean Conference on Embedded Computing (MECO)*, 2017, pp. 1–4, DOI: [10.1109/MECO.2017.7977136](#); (d) I. A. Jimoh, I. Ismaila and M. Olalere, Enhanced Decision Tree-J48 with SMOTE Machine Learning Algorithm for Effective Botnet Detection in Imbalance Dataset, in *15th International Conference on Electronics, Computer and Computation (ICECCO)*, 2019, pp. 1–8, DOI: [10.1109/ICECCO48375.2019.9043233](#).
- 35 See Section 5.1 in the ESI† for additional details on the generation of synthetic data. It should be noted that the synthetic data were only used for training the model, while the test set contains only the real reactions.
- 36 See Sections 5.3 and 5.2 in the ESI† for full details about the choice of the optimal hyperparameter combination and cross-validation procedure.
- 37 (a) Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–444; (b) A. Krizhevsky, I. Sutskever and G. E. Hinton, *Commun. ACM*, 2017, **60**, 84–90; (c) G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, *IEEE Signal Process. Mag.*, 2012, **29**, 82–97.
- 38 (a) I. I. Baskin, D. Winkler and I. V. Tetko, *Expert Opin. Drug Discov.*, 2016, **11**, 785–795; (b) J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl and V. Svetnik, *J. Chem. Inf. Model.*, 2015, **55**, 263–274; (c) E. B. Lenselink, N. ten Dijke, B. Bongers, G. Papadatos, H. W. T. van Vlijmen, W. Kowalczyk, A. P. IJzerman and G. J. P. van Westen, *J. Cheminf.*, 2017, **9**, 45.
- 39 (a) A balance between low test RMSE and overfitting (caused by small sized data with an intrinsic class imbalance) is considered while choosing this architecture of the DNN as compared to the other comparable choices; (b) See section Section 5.3 in the ESI† for more details of various other architectures of DNNs as well as hyperparameter tuning strategies employed.
- 40 See Table S21 in the ESI† for details of the train and test RMSEs obtained using the DNN model.
- 41 (a) See Section 13 in the ESI† for additional details about the performance of the DNN model with different binary and ternary combinations of samples; (b) The performance of additional possible binary and ternary combinations (L_A-L_D , L_B-L_D , $L_A-L_B-L_D$, etc.) is shown in Table S38†; (c) It is important to note that, in a combined set like L_A-L_B , training and test sets include samples from both the L_A and L_B sets.
- 42 The total number of samples in the data matrix is 342, with 240 real and 102 synthetic samples.
- 43 (a) In an alternative featurization, each component of the chemical space is considered as a free substrate, ligand, catalyst etc., in their unbound state. Such a featurization method has been known to provide good performance in previous ML studies as applied to catalytic reactions (ref. 13 and 14a) The feature extraction from the unbound model has the key advantage of faster feature extraction from a set of relatively simpler molecules, as compared to locating an intermediate (Fig. 3) bearing high mechanistic significance in the MLS model. In the present study, the unbound/free ligand model has generally been found to give inferior performance; (b) See Tables S30 and S31 in the ESI† for more details of the feature space; (c) See Section 12.1 in the ESI† for details of the train and test RMSEs for all ML models employed in this study.
- 44 The full data matrix consists of 342 samples (240 real + 102 synthetic), from where 20% of randomized real samples are considered in the test set.
- 45 See Sections 5.4 and 5.5 in the ESI† for details of the train and test RMSEs obtained for all the ML models employed in this study.
- 46 A random forest consists of an ensemble of decision trees. For a given decision tree, the output % ee value in a specific leaf node represents the average over all the reactions that fall in that specific leaf node. In addition, the % ee of each reaction is predicted using all of the trees in the forest, and the average value of the % ee across all of these decision trees is calculated. Thus, the RF predicted values represent two-level averaging.
- 47 See Section 10 in the ESI† for additional details of these three experiments.
- 48 (a) K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, 6416; (b) J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, *Science*, 2018, **362**, eaat8763.
- 49 Correlation analysis is described in detail in Section 11 of the ESI†
- 50 (a) See Table S45 in the ESI† for additional details of the control experiments performed by removing the features that are likely of higher mechanistic importance; (b) To ascertain whether our initial geometry optimization that took into account a number of noncovalent interactions is sufficient, we have used conformational sampling using the CREST algorithm. See P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192; (c) Most of our original conformers were found to be of lower energy than the CREST conformers for the representative MLS systems examined. See Section 17 in the ESI† for more details.
- 51 (a) Details of the three OOB sets with their representative examples are shown in Section 14 of the ESI†; (b) It should be noted that some of the OOB reactions are devoid of a base. To maintain the homogeneity in the feature space, we have included the features of the additive (or solvent) in place of the base features; (c) There are very few reactions with low % ee reported in the experimental literature. We have also considered a few additional randomized OOB sets with <80% ee, drawn from the original 240 training reactions to test the model generalizability in the low % ee region. These new OOB sets are then evaluated with the trained DNN model as



- shown in Section 18 of the ESI† Test RMSEs of around 7.8% ee could be obtained.
- 52 Q.-F. Wu, P.-X. Shen, J. He, X.-B. Wang, F. Zhang, Q. Shao, R.-Y. Zhu, C. Mapelli, J. X. Qiao, M. A. Poss and J.-Q. Yu, *Science*, 2017, **355**, 499–503.
- 53 Z. Zhuang and J.-Q. Yu, *J. Am. Chem. Soc.*, 2020, **142**, 12015–12019.
- 54 (a) Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2020, **12**, 2198–2208; (b) C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. A. Jensen, *Chem. Sci.*, 2019, **10**, 370–377; (c) S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich and C. Corminboeuf, *Chem. Sci.*, 2021, **12**, 6879–6889; (d) Y. Gong, D. Xue, G. Chuai, J. Yu and Q. Liu, *Chem. Sci.*, 2021, **12**, 14459–14472; (e) D. Kreutter, P. Schwaller and J.-L. Reymond, *Chem. Sci.*, 2021, **12**, 8648–8659; (f) P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098; (g) F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, *Chem. Sci.*, 2018, **9**, 1289–1300; (h) O. Egorova, R. Hafizi, D. C. Woods and G. M. Day, *J. Phys. Chem. A*, 2020, **124**, 8065–8078; (i) D. McDonagh, C.-K. Skylaris and G. M. Day, *J. Chem. Theory Comput.*, 2019, **15**, 2743–2758.
- 55 (a) S. Moon, S. Chatterjee, P. H. Seeberger and K. Gilmore, *Chem. Sci.*, 2020, **12**, 2931–2939; (b) J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
- 56 (a) The augmented dataset of 350 samples consisting of both real and synthetic samples is used for training and for subsequent predictions on the out-of-bag samples; (b) Analysis of the feature variance, with and without the inclusion of the four randomly chosen samples, revealed notable differences. See Table S44 in the ESI† for additional details; (c) To examine whether the identity of the reactions being moved to the training set would influence the quality of these OOB predictions, six more additional runs were conducted. In each such run, a different group of four samples was used in the training to learn that the average test RMSE was 3.28% ee with the best value of 1.72% ee and the least of 4.69% ee; (d) Inclusion of four random samples in the original model helps to learn from the higher feature variance as noted (Table S44†). The diversity-included-training set strengthens the ML model ability to predict on unseen samples.
- 57 (a) W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 22071–22080; (b) F. Oviedo, J. L. Ferres, T. Buonassisi and K. Butler, *Acc. Mater. Res.* 2022, **3**, 597–607.
- 58 (a) L. S. Shapley, *Contrib. Theor. Games*, 1953, **2**, pp. 307–317; (b) M. Sundararajan and A. Najmi, arXiv, 2019, preprint, arXiv [cs.AI], <https://arxiv.org/abs/1908.08474>; (c) D. Janzing, L. Minorics and P. Blöbaum, arXiv, 2019, preprint, arXiv [stat.ML], <https://arxiv.org/abs/1910.13413>; (d) S. Lundberg and S.-I. Lee, *Adv. Neural Information Processing*, Curran Associates, 2017, pp. 4765–4774.
- 59 (a) R. Rodríguez-Pérez and J. Bajorath, *J. Med. Chem.*, 2020, **63**, 8761–8777; (b) R. Kronberg, H. Lappalainen and K. Laasonen, *J. Phys. Chem. C*, 2021, **125**, 15918–15933; (c) S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nat. Mach. Intell.*, 2020, **2**, 56–67.
- 60 See Section 16 in the ESI† for more details on feature attributions using the SHAP method.
- 61 B. E. Haines and D. G. Musaev, *ACS Catal.*, 2015, **5**, 830–840.
- 62 See Section 19 in the ESI†.

