

Electronic Supplementary Information

for

Deep Learning for Enantioselectivity Predictions in Catalytic Asymmetric β -C-H Bond Activation Reactions

Ajnabiul Hoque^a and Raghavan B. Sunoj^{a,b,*}

^aDepartment of Chemistry, Indian Institute of Technology Bombay, Powai, Mumbai 400076,
India

^bCentre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay,

Powai, Mumbai 400076, India

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

Section	Content	Page No.
1	Details of all reaction components	3-9
1.1	Ligands	3-6
1.2	Coupling partner	7-8
1.3	Substrate	8
1.4	Base	8
1.5	Pd-catalyst precursor	8
1.6	Solvent	8-9
1.7	Additive	9
2	Selection of chemically relevant model for feature extraction	9-10
3	Parameter selection	10-14
4	Computational methods and programming details	14-15
5	Model building, results and analysis for various machine learning methods	15-25
5.1	Synthetic data generation using SMOTE technique	15-16
5.2	The cross-validation procedure for ML models other than DNN	16-17
5.3	DNN hyperparameter optimization	17-20
5.4	Details of various ML methods	21-23
5.5	Predictive performance with different ML algorithms for various subsets	23-25
5.6	Performance of DNN in terms of the R-squared values	25
6	Performance with the real dataset using DNN algorithm	26
7	A comparison of test and train RMSEs over 100 runs for different ML models	26-27
8	Performance analysis of ML models in different class intervals	27-28
9	Effect of train-test splitting	27-28
10	Assessment of feature importance	28-29
10.1	Randomization of features	28
10.2	Normally distributed set of random numbers	28
10.3	One-hot encoding	28-29
11	Correlation analysis	29-37
12	Details of parameters and performance of different ML models for unbound model	37-42
12.1	Performance of different ML models for different subsets	40-42
13	Performance of the DNN algorithm for MLS model with different binary and ternary combinations of samples	42
14	Details of out-of-bag sets	42-48
14.1	Set-1	42-44
14.2	Set-2	44-45
14.3	Set-3	45-48
15	Selective reduction of features	48
16	ML model interpretability using SHAP	48-50
17	Conformational sampling using CREST	50-53
18	Analyses of performance of DNN in the low % <i>ee</i> region	53-54
19	Workflow for planning new experiments	54-55

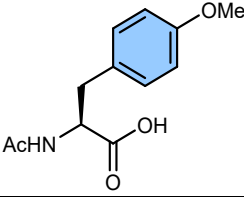
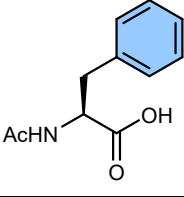
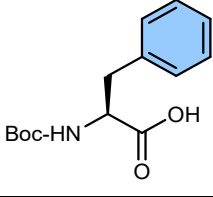
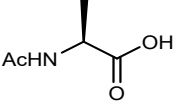
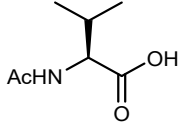
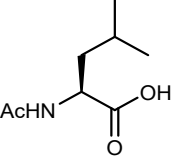
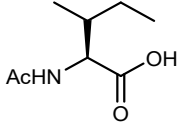
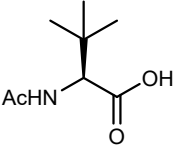
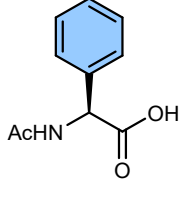
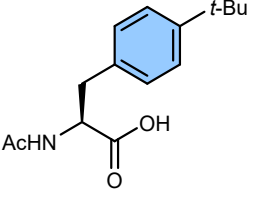
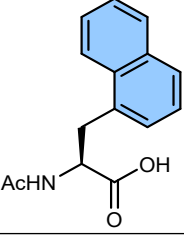
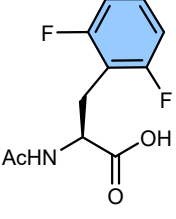
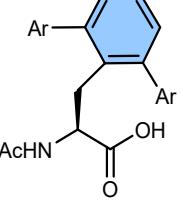
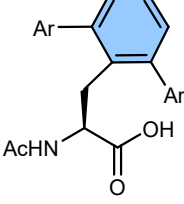
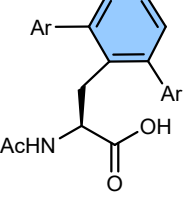
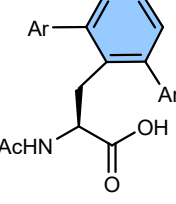
(1) Details of all reaction components

(1.1) Ligand library¹

The ligands can be chemically grouped into four categories as listed in Table S1, which are (a) mono-N-protected amino acid (MPAA) [L_A], (b) mono-N-protected α -amino-O-alkyl hydroxamic acid (MPAHA) [L_B], (c) mono-N-protected amino alkyl amine (MPAAM) [L_C], and (d) N-acyl-protected amino oxazoline (APAO) [L_D]. There are 77 different ligands in total.

Table S1. Identities and Notations of the Ligands

Table S1.A. Subset L_A

L_A (mono-protected amino acid [MPAA]) (number of ligands = 27)			
			
L_A-1	L_A-2	L_A-3	L_A-4
			
L_A-5	L_A-6	L_A-7	L_A-8
			
L_A-9	L_A-10	L_A-11	L_A-12
 Ar = 3,5- <i>t</i> -Bu-C ₆ H ₃	 Ar = C ₆ H ₅	 Ar = 4-F-C ₆ H ₄	 Ar = 4-Me-C ₆ H ₄
L_A-13	L_A-14	L_A-15	L_A-16

<p>Ar = 4-OMe-C₆H₄</p>	<p>Ar = 4-Ph-C₆H₄</p>	<p>Ar = 4-<i>t</i>-Bu-C₆H₄</p>	<p>Ar = 3-Me-C₆H₄</p>
LA-17	LA-18	LA-19	LA-20
<p>Ar = 3,5-OMe-C₆H₃</p>			
LA-21	LA-22	LA-23	LA-24
	<p>R' = </p>	<p>Ar = 4-F-C₆H₄</p>	
LA-25	LA-26	LA-27	

Table S1.B. Subset **L_B**

L_B (mono-N-protected α -amino-O-alkyl hydroxamic acid (MPAHA)) (number of ligands = 28)			
LB-1	LB-2	LB-3	LB-4
LB-5	LB-6	LB-7	LB-8

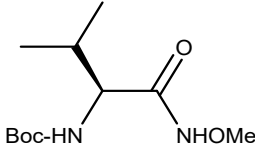
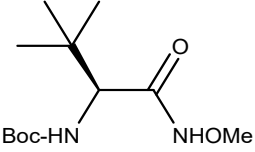
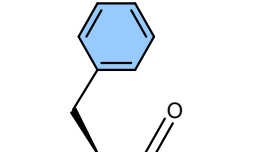
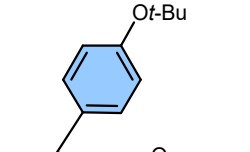
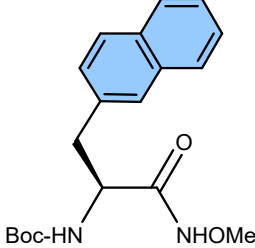
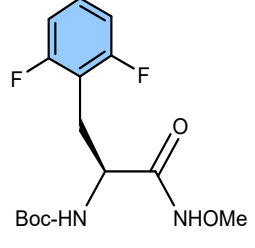
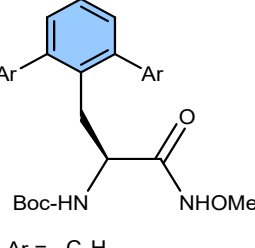
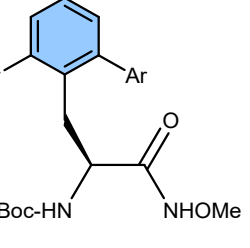
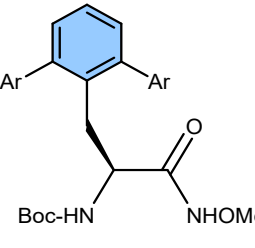
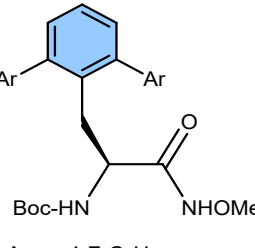
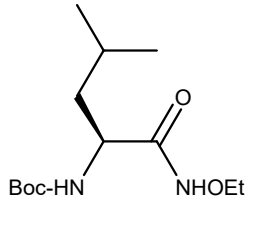
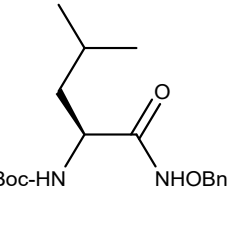
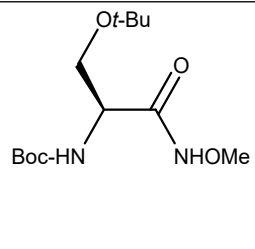
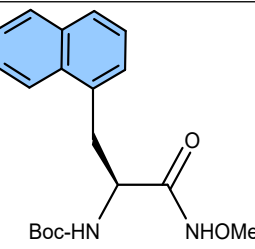
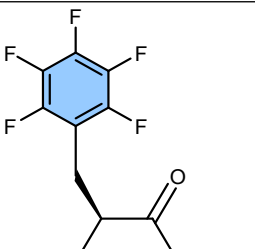
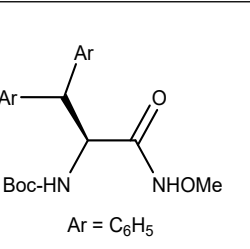
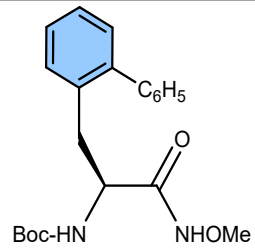
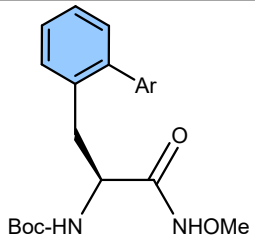
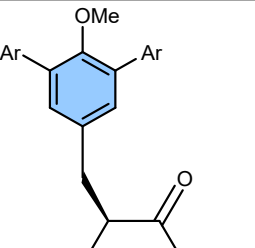
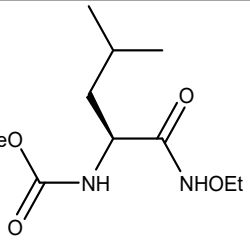
			
LB-9	LB-10	LB-11	LB-12
			
LB-13	LB-14	LB-15	LB-16
<p>Ar = 4-OMe-C₆H₄</p>	<p>Ar = 4-F-C₆H₄</p>	<p>Ar = C₆H₅</p>	<p>Ar = 4-Ph-C₆H₄</p>
			
LB-17	LB-18	LB-19	LB-20
			
LB-21	LB-22	LB-23	LB-24
			
LB-25	LB-26	LB-27	LB-28
<p>Ar = C₆H₅</p>	<p>Ar = 4-F-C₆H₄</p>	<p>Ar = C₆H₅</p>	

Table S1.C. Subset L_C

L _C (mono-N-protected amino alkyl amine (MPAAM)) (number of ligands = 20)			
LC-1	LC-2	LC-3	LC-4
LC-5	LC-6	LC-7	LC-8
LC-9	LC-10	LC-11	LC-12
LC-13	LC-14 Ar = 4-F-C ₆ H ₄	LC-15 Ar = 4- <i>t</i> -Bu-C ₆ H ₄	LC-16 Ar = 3,5- <i>t</i> -Bu-C ₆ H ₃
LC-17	LC-18	LC-19	LC-20

Table S1.D. Subset L_D

L _D (N-acyl-protected amino oxazoline (APA0)) (number of ligands = 2)	
LD-1	LD-2

(1.2) Coupling partner library

Table S2. Identities and Notations of the Coupling Partners

CP1	CP2	CP3	CP4	CP5
CP6	CP7	CP8	CP9	CP10
CP11	CP12	CP13	CP14	CP15
CP16	CP17	CP18	CP19	CP20
CP21	CP22	CP23	CP24	CP25
CP26	CP27	CP28	CP29	CP30
CP31	CP32	CP33	CP34	CP35
CP36	CP37	CP38	CP39	CP40

CP41	CP42	CP43	CP44	CP45
CP46	CP47	CP48	CP49	CP50
CP51				

(1.3) Substrate library

Table S3. Identities and Notations of the Substrates

Ar _F = 4-CF ₃ C ₆ F ₄			Ar _F = 4-CF ₃ C ₆ F ₄	Ar _F = 4-CNC ₆ F ₄
S1	S2	S3	S4	S5

(1.4) Base library

NaTFA, Na₂CO₃, NaHCO₃, K₂HPO₄, Li₃PO₄, Na₃PO₄, K₃PO₄, LiH₂PO₄, Li₂CO₃, K₂CO₃, Cs₂CO₃, LiOAc, NaOAc, KOAc, CsOAc, NaH₂PO₄, Na₂HPO₄, KHCO₃, KH₂PO₄, K₂HPO₄.3H₂O

(1.5) Metal-catalyst precursor library

Pd(MeCN)₂Cl₂, Pd(TFA)₂, Pd(C₃H₅)Cl₂, Pd(PhCN)₂Cl₂, Pd(OTf)₂(MeCN)₄, Pd(OAc)₂, Pd(BF₄)₂(MeCN)₄, Pd(PPh₃)₂Cl₂, Pd(OPiv)₂

(1.6) Solvent library

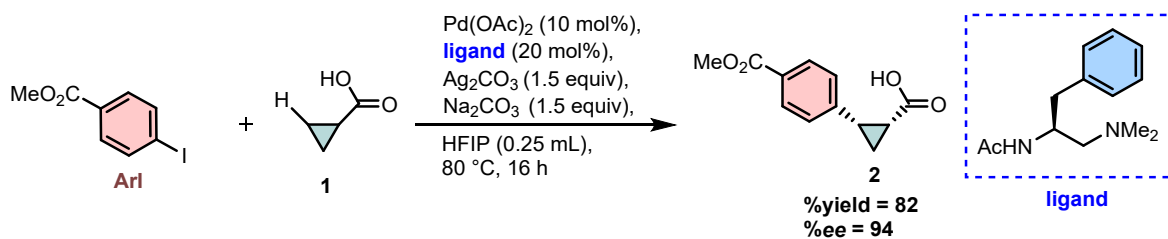
Toluene, CHCl₃, HFIP, *t*-AmylOH + H₂O, *t*-BuOH + H₂O, *i*-PrOH + H₂O, *i*-BuOH + H₂O, THF + H₂O, HFIP + H₂O, *t*-AmylOH, DCE, DCM, DMF, C₆F₆, TBME, MeCN, THF, Dioxane, Et₂O, CCl₄.

(1.7) Additive library

Ag₂CO₃, AgOAc, Ag₂O

(2) Selection of chemically relevant model for feature extraction

Based on available mechanistic studies on similar C(sp³)-H functionalization reaction,² probable mechanistic pathway could follow the following series of steps- formation of active catalyst, N-acyl group on the ligand act as a base to deprotonate the C(sp³)-H bond, oxidative addition of **ArI**, subsequent reductive elimination, and regeneration of the active mono-ligated catalyst (Fig. S1). Close mechanistic and structural approximation of intermediate **4** with C(sp³)-H activation TS [4-5][‡] makes it reasonable choice as a chemically relevant model for feature extraction.



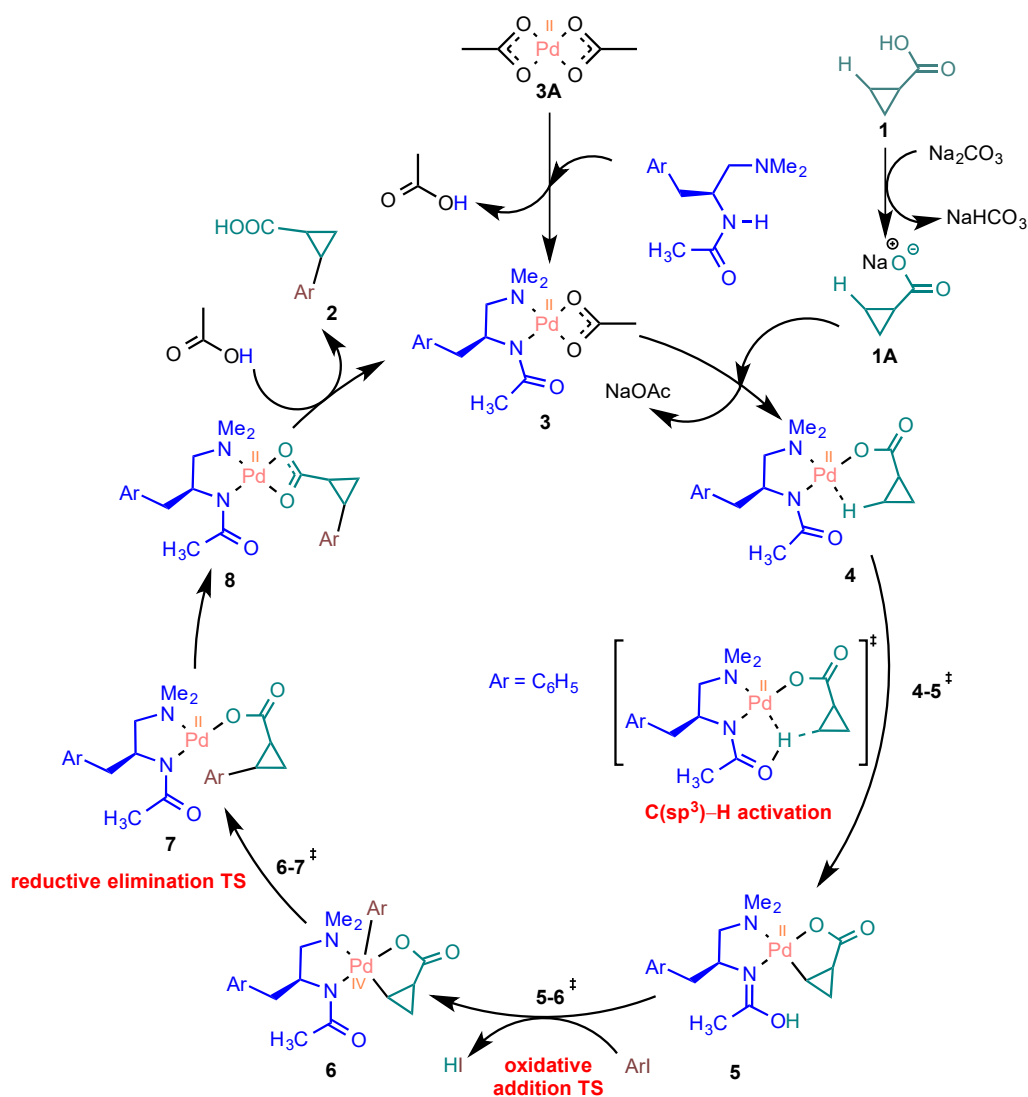


Fig S1. An illustrative example of (A) β -C(sp³)-H activation reaction, (B) proposed mechanism promoted by the Pd(OAc)₂-MPAAM catalytic system.

(3) Parameter selection

We employed physically meaningful descriptors derived from optimized molecule geometries.³ For site-specific properties, various local parameters such as NBO charge, vibrational frequencies, bond length, bond angle, dihedral angle, and so on are taken into account. Global parameters such as HOMO and LUMO energies, rotational constants, polar surface area, volume, and so on are used to represent the entire structural and geometrical properties. To account for the influence of the solvent, we used the continuum solvation model in our calculations. Furthermore, as the reactions are used in a wide range of reaction

conditions, we used them as descriptors. Experimental conditions like reaction temperature, time, amount of ligand/base, solvent dielectric are selected as descriptors.⁴ All these descriptors for MLS model are enlisted in Table S4 and Table S5.

Table S4. Parameter Details of Various Reacting Components in the MLS Model

metal-ligand-substrate (MLS) complex			
local parameters			
bond length (BL)	1-2, 5-6, 6-7, 11-12, 4-20, 4-19, 6-22, 10-15	bond angle (BA)	1-2-3, 3-4-20, 1-5-4, 5-4-20, 5-6-7, 7-6-22, 10-11-12, 10-11-14, 1-12-11, 5-1-12
dihedral angle (DA)	3-4-5-6, 4-5-1-2, 4-5-6-7, 9-10-11-14, 9-10-11-12, 1-5-6-7, 1-12-11-14, 1-5-4-20, 6-5-4-20	non-bonded distance (NB)	1-12, 7-12
charge (q)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20, 22	NMR shift (NMR)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20, 22
vibrational frequency (VF) & intensity (VI)	6-7, 11-12, 9-13	sterimol(L, B1, B5)	R ¹⁴ , R ¹⁵ , R ¹⁶ , X, Y
global parameters			
HOMO energy, LUMO energy, Dipole moment (DM), Rotational constant (R _x), Area, Volume, PSA, Ovality			
additional parameter			
% buried volume (BV) ⁹ : The radius of the sphere has been set to 5.0 Å in this case, five different parameters are taken into consideration. These are BV, BV in NW quadrant, BV in SW quadrant, BV in NE quadrant and BV in SE quadrant.			
coupling partner (CP)			
local parameters			
bond length (BL)	1-2, 1-6, 1-7	NMR shift (NMR): 1, 2, 6	charge (q): 1, 2, 6

global parameters
HOMO energy, LUMO energy, Dipole moment (DM), Rotational constant (R_x), Area, PSA metal-catalyst precursor (MC)
local parameter
q(Pd)
global parameters
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y), Volume, PSA, Ovality
base (B)
global parameters
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y , R_z), Area, Volume, PSA, Ovality
additive (A)
local parameter
q(Ag)
global parameters
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y , R_z), Area, Volume, PSA, Ovality
secondary parameters
dielectric constant of solvent (DC)
experimental parameters
amount of ligand - Eqv(L), amount of base - Eqv(B), amount of additive - Eqv(A), reaction time and reaction temperature.

Table. S5. List of Features (P1 - P153) in the MLS Model

P1	Rx-MLS	P78	VF(C9=O13)-MLS
P2	DA3-4-5-6-MLS	P79	VI(C9=O13)-MLS
P3	DA4-5-1-2-MLS	P80	VF(C11-H12)-MLS
P4	DA4-5-6-7-MLS	P81	VI(C11-H12)-MLS
P5	DA9-10-11-14-MLS	P82	L-Y-MLS
P6	DA9-10-11-12-MLS	P83	B1-Y-MLS
P7	DA1-5-6-7-MLS	P84	B5-Y-MLS
P8	DA1-12-11-14-MLS	P85	L-X-MLS
P9	DA1-5-4-20-MLS	P86	B1-X-MLS
P10	DA6-5-4-20-MLS	P87	B5-X-MLS
P11	BA1-2-3-MLS	P88	L-R16-MLS
P12	BA3-4-20-MLS	P89	B1-R16-MLS
P13	BA1-5-4-MLS	P90	B5-R16-MLS
P14	BA5-4-20-MLS	P91	BV-MLS
P15	BA5-6-7-MLS	P92	BV-SW-MLS
P16	BA7-6-22-MLS	P93	BV-NW-MLS
P17	BA10-11-12-MLS	P94	BV-NE-MLS
P18	BA10-11-14-MLS	P95	BV-SE-MLS
P19	BA1-12-11-MLS	P96	Area-MLS
P20	BA5-1-12-MLS	P97	Volume-MLS
P21	NB1-12-MLS	P98	PSA-MLS
P22	NB7-12-MLS	P99	Ovality-MLS
P23	BL1-2-MLS	P100	HOMO-MC

P24	BL5-6-MLS	P101	LUMO-MC
P25	BL6-7-MLS	P102	DM-MC
P26	BL11-12-MLS	P103	Rx-MC
P27	BL4-20-MLS	P104	Ry-MC
P28	BL4-19-MLS	P105	Volume-MC
P29	BL6-22-MLS	P106	PSA-MC
P30	BL10-15-MLS	P107	Ovality-MC
P31	HOMO-MLS	P108	q(Pd)-MC
P32	LUMO-MLS	P109	DM-CP
P33	DM-MLS	P110	HOMO-CP
P34	q1-MLS	P111	LUMO-CP
P35	q2-MLS	P112	Rx-CP
P36	q3-MLS	P113	BL1-2-CP
P37	q4-MLS	P114	BL1-6-CP
P38	q5-MLS	P115	BL1-7-CP
P39	q6-MLS	P116	NMR1-CP
P40	q7-MLS	P117	NMR2-CP
P41	q8-MLS	P118	NMR6-CP
P42	q9-MLS	P119	NMR7-CP
P43	q10-MLS	P120	q1-CP
P44	q11-MLS	P121	q2-CP
P45	q12-MLS	P122	q6-CP
P46	q13-MLS	P123	Area-CP
P47	q14-MLS	P124	Volume-CP
P48	q15-MLS	P125	PSA-CP
P49	q19-MLS	P126	Ovality-CP
P50	q20-MLS	P127	Rx-B
P51	q22-MLS	P128	Ry-B
P52	NMR1-MLS	P129	Rz-B
P53	NMR2-MLS	P130	DM-B
P54	NMR3-MLS	P131	HOMO-B
P55	NMR4-MLS	P132	LUMO-B
P56	NMR5-MLS	P133	Area-B
P57	NMR6-MLS	P134	Volume-B
P58	NMR7-MLS	P135	PSA-B
P59	NMR8-MLS	P136	Ovality-B
P60	NMR9-MLS	P137	q(Ag)-A
P61	NMR10-MLS	P138	Rx-A
P62	NMR11-MLS	P139	Ry-A
P63	NMR12-MLS	P140	Rz-A
P64	NMR13-MLS	P141	HOMO-A
P65	NMR14-MLS	P142	LUMO-A
P66	NMR15-MLS	P143	DM-A
P67	NMR19-MLS	P144	Area-A
P68	NMR20-MLS	P145	Volume-A
P69	NMR22-MLS	P146	PSA-A
P70	L-R14-MLS	P147	Ovality-A
P71	B1-R14-MLS	P148	Eqv(L)

P72	B5-R14-MLS	P149	Eqv(B)
P73	L-R15-MLS	P150	Eqv(A)
P74	B1-R15-MLS	P151	Time (H)
P75	B5-R15-MLS	P152	T (°C)
P76	VF(C6=O7)-MLS	P153	DC(Solvent)
P77	VI(C6=O7)-MLS		

(4) Computational methods and programming details

All quantum chemical calculations in this study were done using the Gaussian 09 program.⁵ All the geometries were optimized in the condensed phase using the dispersion-corrected hybrid density functional B3LYP-D3 with the 6-31G** basis set for all atoms except for palladium.⁶ We used the Stuttgart-Dresden double-basis (SDD) basis set with an effective core potential (ECP) for Pd, Ag, and Cs. For the Pd atom, 28 core electrons were represented using an ECP,⁷ while standard basis sets were used to explicitly treat 18 valence electrons. The fully optimized geometries of all stationary points were characterized by frequency calculations in order to verify that the optimized geometries have all positive Hessian indices. The Truhlar-Cramer SMD solvation model, which uses the full solute electron density without defining partial atomic charges, was used to incorporate the effect of continuum solvation.⁸ We used the continuum dielectric of applicable solvents in our computations as reported in the corresponding experimental studies which used different solvents. The optimized geometries as described above were then used to derive all of the stereo-electronic parameters. For the calculation of multi-dimensional Sterimol parameters: L, B1 and B5, we used the Python program developed by the Paton group.⁹ The percentage of buried volume was calculated by using SambVca 2.1 program developed by the Cavallo group.¹⁰

The code, data, and instructions are available at <https://github.com/alhqlearn/ML-for-Asymmetric-C-sp3-H-Reaction>. Instructions for installing software used in the study are as follows. Download and install the following programs: (a) Spartan'16 Parallel Suite: We applied for 30-day Spartan'16 Parallel Suite demo license <https://www.wavefun.com/> (Accessed on April 24th, 2018), (b) Python 3.6 (The anaconda distribution is recommended,

as it has packages required for the software to run: Download at <https://www.anaconda.com/download/>) (Accessed on June 23rd, 2018), (c) PyTorch (Download at <https://pytorch.org/>)

(5) Model building, results and analysis of various machine learning methods

(5.1) Synthetic data generation using SMOTE technique

Among the 240 reactions considered in this study, the experimental %*ee* distribution (Fig. S2) is such that only 57 samples are <80 %*ee* constituting the minority class while a large majority of them (183) are >80 %*ee*. The output values therefore indicate a class-imbalance, necessitating the use of synthetic data for improved training of the ML algorithms. We have generated synthetic samples using the SMOTE (SVM) method.¹¹ SMOTE is an oversampling technique that adds synthetic data to the minority class. The SMOTE procedure consists of the following steps; (i) selects a sample (**m**) from the minority class, (ii) identifies the *k*-nearest neighbors (**k1**, **k2**, **k3**, **k4**, **k5**) (default value is 5) of the chosen data point, (iii) synthetic sample (**n1**) is then created from the line connecting data point **m** and neighbor **k1**, (iv) the difference of the feature vector of the selected data point (f_m) and neighbors (f_{k1}) are calculated, (v) features of this new sample (f_{n1}) is obtained as $f_{n1} = f_m + (f_m - f_{k1}) * r$; $r \in [0,1]$, where the features of data point f_m are added with the difference multiplied with a random number (*r*) that ranges from 0 to 1. Similarly, several such synthetic data points are generated by iterating over all the minority class samples.

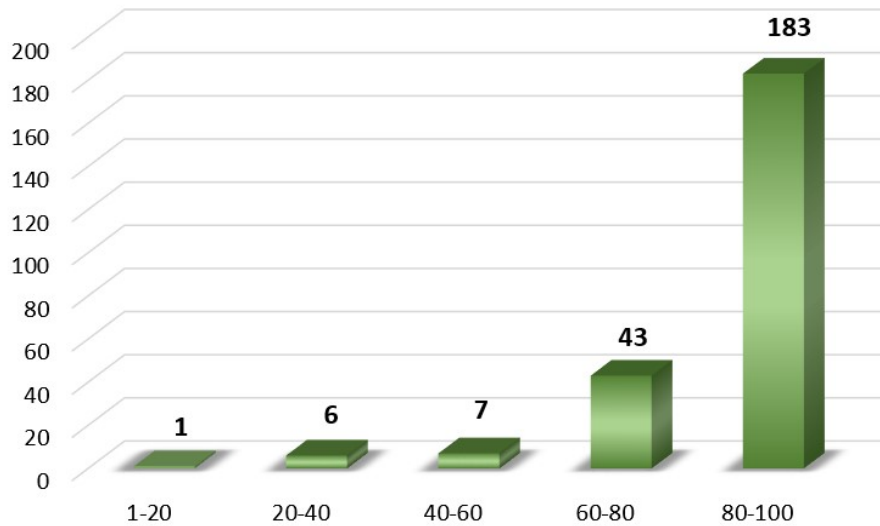


Fig S2. Experimentally reported %ee distributions in various class intervals

The dataset is divided into major and minor classes based on their output values (%ee), with a class boundary set to ≤ 80 %ee for the minority class. With the inclusion of the synthetic data in the minority class, the dimension of the feature matrix changes from 240x153 to 342x153. This dataset containing both the real and synthetic samples are used in training our ML models. The *Smote.ipynb* Python file provided in the Github link can be used to generate synthetic data. The README file of the Github repository contains step-by-step instructions for replicating this.

(5.2) The cross-validation procedure for ML models other than DNN

The full dataset was randomly divided in a 80:20 ratio, with 80% of samples placed in the training and the remaining 20% in the test sets. The following 7-fold cross-validation strategy was used to identify the optimal hyperparameter for each machine learning method used in this study (Fig S3). Seven validation runs were used for each hyperparameter H_i (assuming the hyperparameter set is $H = \{H_1, H_2, H_3, H_4, \dots, H_n\}$). After doing 7 cross-validation runs, the average RMSE was calculated as $\text{Avg. RMSE}[H_i] = \sum(\text{RMSE}[F_i])/7$ for each of the hyper-parameters where $[F_i]$ is the i^{th} fold in the validation set (Fig S3). The hyperparameter with the lowest average RMSE was then chosen, which was subsequently used for model

building. The performance of the resulting model, on the basis of the test set RMSE is reported. We repeated this technique for 100 separate runs, with randomly chosen training and test samples, to get an unbiased estimation of generalization error. The construction of different partitions was controlled by seeding the random number generator with a seed value to ensure the reproducibility of the results. The final RMSE reported in the following sections is averaged over all these 100 runs.

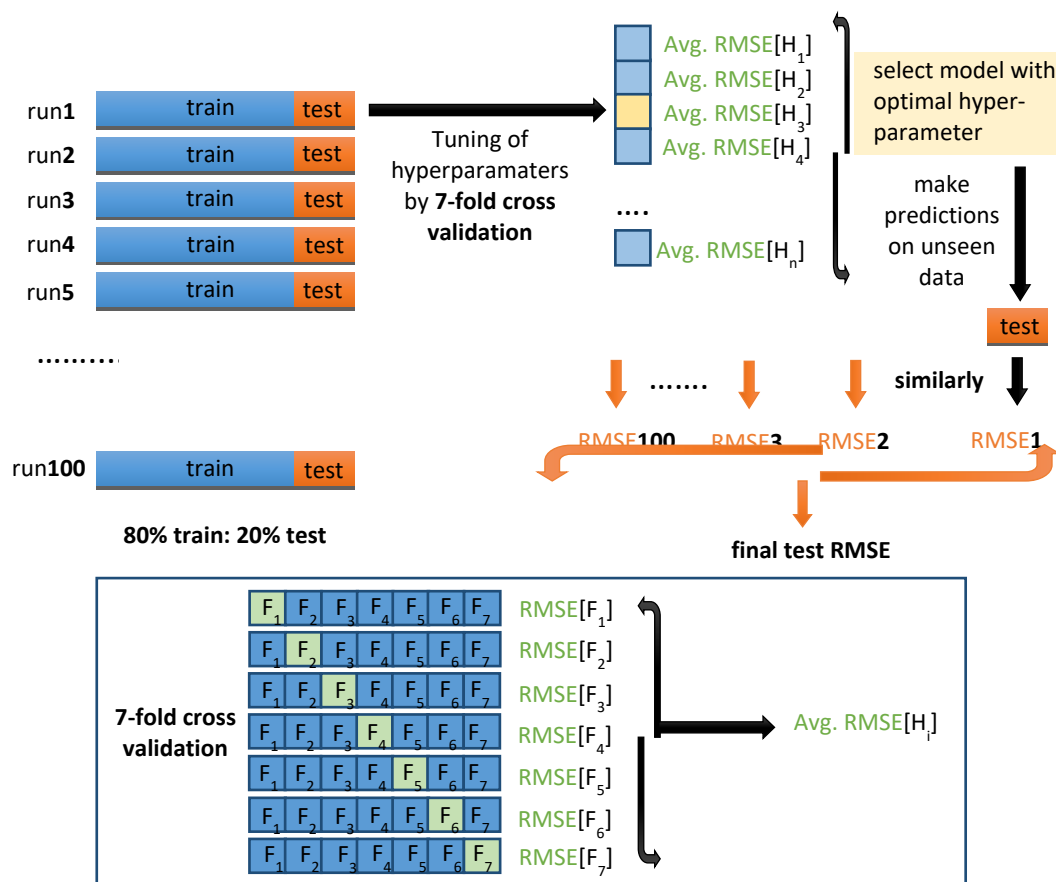


Fig S3. A general procedure for ML model building. Hyperparameter tuning was done on the validation set. Finally, the trained model was used to assess the test set.

(5.3) Hyperparameter optimization of DNN

A deep neural network (DNN) is made up of several fully connected input layer, one or more hidden layers, and a single output layer. A deep neural network (DNN) is typically trained by using a supervised learning back-propagation algorithm, with Adam optimization technique

to update and adjust the weights and biases of neurons. The performance of a DNN depends on several hyperparameters such as the number of hidden layers, number of neurons, learning rate, dropout rate, number of epochs etc. Hence, it is important to examine what combination of these hyperparameters is likely to provide a good model. The optimal architecture of DNN is essential for improved accuracy and faster convergence.

In this study, various DNN network architecture were examined. The data matrix used here has a sample x feature dimension of 342x153, where 240 were real and 102 were synthetic samples generated using the SMOTE technique (see section 5.1) and molecular features were collected from the MLS model (see section 3). The dataset is partitioned into train-validation-test sets with 64:16:20 ratio. The model was trained on the train set, and hyperparameter tuning was done on the validation set. Finally, the trained model was used to evaluate the model performance on the test set. To get an unbiased estimate of generalization error, we repeated this technique 100 times with randomly chosen training and test samples. The final train, validation, and test RMSEs tabulated in Table S6, are the average over all the 100 runs. A default settings of our DNN use (a) Adam optimization function, (b) a learning rate of 0.001, (c) number of epochs 1000, and (d) Rectified Linear Unit (ReLU) as the activation function. The data provided in Table S6 help in comparing effect of different numbers of hidden layers and numbers of neurons in each hidden layer.

Table. S6. Train, Validation, and Test RMSEs Obtained by Varying the Number of Hidden Layers and Neurons. *i* Shown in Bold Font is the Optimal Model

DNN layers/neurons in each layer	train	validation	test
[153, 128, 1]	7.58 ±0.43	7.25±1.12	8.14±1.21
[153, 300, 1]	5.2±0.27	6.11±1.00	7.05±1.44
[153, 53, 1]	12.56±0.26	10.54±1.22	10.91±1.53
[153, 400, 128, 1]	5.1±0.75	6.48±1.05	7.33±1.7
[153, 200, 128, 1]	4.87±0.83	6.33±1.04	7.08±1.35
[153, 70, 128, 1]	4.87±0.55	6.3±1.15	7.03±1.46
[153, 500, 250, 128, 1]	4.88±0.89	6.19±1.11	6.87±1.53

[153, 500, 168, 128, 1]	4.97±0.82	6.17±1.02	6.84±1.39
[153, 500, 168, 70, 1]	4.89±1.02	6.24±1.23	7.04±1.6
[153, 75, 250, 168, 128, 1]	4.57±0.68	6.09±1.03	6.85±1.23
[153, 500, 250, 168, 128, 1]	4.86±0.76	6.17±1.23	6.91±1.34
[153, 700, 350, 168, 128, 1]	4.51±0.48	6.04±1.01	6.59±1.39
[153, 500, 150, 400, 168, 128, 1]	4.68±0.65	6.04±1.08	6.77±1.41
[153, 700, 400, 250, 168, 128, 1]	4.75±1.03	5.96±1.17	6.70±1.43
[153, 33, 150, 400, 168, 128, 1]	4.42±0.49	5.91±0.86	6.53±1.04
[153, 120, 150, 400, 168, 128, 1]	4.61±0.57	6.04±0.89	6.78±1.21
[153, 250, 150, 400, 168, 128, 1]	4.83±0.95	6.12±1.22	6.84±1.51
[153, 33, 50, 400, 168, 128, 1]	4.63±1.65	6.13±1.22	6.80±1.50
[153, 33, 150, 300, 168, 128, 1]	4.68±0.7	6.23±0.98	6.98±1.45
[153, 33, 150, 400, 200, 128, 1]	4.52±1.03	6.31±1.25	6.81±1.44

ⁱ learning rate=0.001, epoch=1000, activation function= ReLU, optimizer=Adam, dropout rate = 0.0

Varying the dropout rate is recommended as a technique to reduce over-fitting and improve generalization. The data provided in Table S7 can be used for comparing the effect of various dropout rates.

Table. S7. Train, Validation, and Test RMSEs Obtained by Varying the Dropout Rates ⁱ

Shown in Bold Font is the Optimal Model

dropout rate	train	validation	test
0.0	4.42±0.49	5.91±0.86	6.53±1.04
0.1	3.96±0.38	5.86±0.82	6.74±1.11
0.2	3.94±0.47	5.88±0.88	6.79±1.12
0.3	3.88±0.46	5.83±0.88	6.64±1.09
0.4	3.88±0.38	5.91±0.79	6.77±1.12
0.5	4.01±0.49	6.03±0.98	6.82±1.23
0.6	4.10±0.44	6.11±0.89	6.87±1.21
0.7	4.14±0.43	5.90±0.87	6.74±1.23
0.8	4.02±0.57	6.00±0.97	6.79±1.22
0.9	4.22±0.39	6.13±0.77	6.76±0.97

ⁱ learning rate=0.001, epoch=1000, activation function= ReLU, optimizer=Adam, DNN architecture = [153, 33, 150, 400, 168, 128, 1]

Three initial learning rates (0.1, 0.01, and 0.0001) were also considered (Table S9).

Tables S9 and S10 show the results with different number of epochs and activation functions.

Table. S8. Train, Validation, and Test RMSEs Obtained by Varying the Number of Epoch ⁱ

Shown in Bold Font is the Optimal Model

epoch	train	validation	test
500	5.36±0.82	6.85±1.29	7.14±1.11
1000	4.42±0.49	5.91±0.86	6.53±1.04
1500	3.86±0.49	5.69±0.79	6.38±0.97
2000	3.44± 0.50	5.70± 0.90	6.35± 0.94

ⁱ learning rate=0.001, activation function= ReLU, optimizer=Adam, dropout ratio = 0.0, DNN architecture = [153, 33, 150, 400, 168, 128, 1]**Table. S9.** Train, Validation, and Test RMSEs Obtained by Varying the Learning Rate ⁱ

Shown in Bold Font is the Optimal Model

learning rate	train	validation	test
0.1	17.71±0.38	15.67±1.32	15.03±1.72
0.01	13.38±4.44	11.55±3.58	11.91±3.23
0.001	4.42±0.49	5.91±0.86	6.53±1.04
0.0001	5.59±0.22	6.62±0.80	6.94±0.99

ⁱ epoch=1000, activation function= ReLU, optimizer=Adam, dropout ratio = 0.0, DNN architecture = [153, 33, 150, 400, 168, 128, 1]**Table. S10.** Train, Validation, and Test RMSEs Obtained by Varying Activation Functions ⁱShown in Bold Font is the Optimal Model ⁱ

activation function	train	validation	test
ReLU	4.42±0.49	5.91±0.86	6.53±1.04
LeakyReLU	4.41±0.76	5.97±1.11	6.76±1.33
Tanh	17.93±0.36	15.76±1.12	16.35±1.35
Sigmoid	24.01±0.19	21.88±0.68	25.96±0.94
Tanh (2000) ^j	17.70±0.38	15.66±1.30	15.04±1.71
Sigmoid (2000) ^j	17.71±0.38	15.64±1.27	15.24±1.64

ⁱ learning rate=0.001, epoch=1000, optimizer=Adam, dropout ratio = 0.0, DNN architecture = [153, 33, 150, 400, 168, 128, 1].^j A total of 2000 epochs were used to train the DNN model.

In line with the current practices in ML community, we built DNN model with a 70:10:20 train-validation-test ratio. The train, validation, and test RMSEs were found to be 4.4±0.7, 5.9±1.0, and 6.5±1.1 %*ee*, respectively, which is nearly similar to the result obtained using 64:16:20 train-validation-test split.

(5.4) Details of various ML methods

In the current study, we used the ML methods enlisted below. The direct use of default parameters as in the *scikit-learn* package led to overfitting with all of these techniques, hence demanded rigorous hyperparameter optimization. The chosen hyperparameters for each method are listed below. We have used *scikit-learn*¹² (python machine learning package) for all methods, and "pytorch" for deep neural network.

(I) Random Forest (RF)

Default parameters and hyperparameters¹³ used in the RF algorithm are provided in Table S11.

Table S11. List of Parameters used for the RF

method	default parameters	hyperparameters
RF	bootstrap=True, criterion='mse', max_depth=10000,max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_jobs=1, oob_score=False, random_state=42, verbose=0, warm_start=False	n_estimators {100,200,...,1900, 2000}

(II) k-Nearest Neighbors (kNN)

The default parameters and hyperparameters used in the kNN algorithm are provided in Table S12.

Table S12. List of Parameters used for the kNN

method	default parameters	hyperparameters
kNN	weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=1	n_neighbors {2,3,...,30}

(III) Gradient Boosting (GB)

Default parameters and hyperparameters used in the gradient boosting algorithm are provided in Table S13.

Table S13. List of Parameters used for GB

method	default parameters	hyperparameters
GB	loss='huber', learning_rate=0.1, n_estimators=800, criterion='mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=1, min_impurity_decrease=0.0, min_impurity_split=None, init=None, random_state=42, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, presort='auto'	Subsample (0.1,0.2,.....,0.9) Alpha {10-5 ,10-4 ,10-3 , 10-2 ,0.1,0.5,0.9}

(IV) Decision Tree (DT)

The default parameters and hyperparameters used in the DT algorithm are provided in Table S14.

Table S14. List of Parameters used for DT

method	default parameters	hyperparameters
DT	criterion='mse', splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=42, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, presort=False	max_depth {1000,2000,.....,10000}

(V) Deep Neural Network (DNN)

Neural networks are composed of (multiple) layers of interconnected computation modules (so-called neurons). Each neuron uses parameters or weights, as well as a nonlinear activation function, to process its input (i.e., the values received from previous neurons). The architecture of neural network is given in Table S15.

Table S15. List of Parameters used for DNN

DNN	number of hidden layers = 6, number of neurons in each layer = [153, 33, 150, 400, 168, 128, 1], learning rate = 0.001, epoch = 1000, activation function= ReLU, optimizer=Adam
-----	---

(VI) Gaussian Process Regression (GPR)

Gaussian process regression models are nonparametric kernel-based probabilistic models. A kernel is used to define the covariance of a prior distribution over the target functions in GPR. We have considered radial basis function (RBF) as kernel functions. The final kernels were created by multiplying the constant kernel with a kernel function (RBF) of the GPR and then adding a white kernel. The default parameters and hyperparameters used in the GPR algorithm are provided in Table S16.

Table S16. List of Parameters used for GPR algorithm with different kernel

method	default parameters	hyperparameters
RBF	kernel = ConstantKernel() * RBF(length_scale_bounds=(1e-05, 100000.0)) + WhiteKernel() kernel=kernel, alpha=1e-10, optimizer='fmin_l_bfgs_b', n_restarts_optimizer=0, normalize_y=False, copy_X_train=True, random_state=42	length_scale { 0.01, 0.1, 1, 2, 3, 5, 10, 20, 50, 100}

(5.5) Predictive performance of different ML algorithms for various subsets

We have created individual ML algorithms for each of the four ligand families denoted as L_A , L_B , L_C , and L_D . Only reactions from that family are included in each of these individual models. The training set contains both real and synthetic samples from that family, while the test set contains only real reactions. The individual performances are tabulated in the tables below.

In the combined set, test set is made up of reactions, chosen at random, from one or more ligand sets. Subsequently, we have combined the reactions from two different ligand families to create a new dataset (L_A-L_B). Similarly, more diverse models were created by combining reactions from the L_A , L_B , and L_C families to make a new ML model. All the 240

reactions were combined to create the final unified set representing maximum diversity in the dataset composed of L_A , L_B , L_C , and L_D .

Table S17. Train and Test RMSEs Obtained using the RF Algorithm for Various Subsets

set	train	test
L_A	2.19±0.09	4.64±1.03
L_B	2.66±0.19	6.64±2.55
L_C	3.05±0.34	6.16±2.35
L_D	2.79±0.25	7.96±3.25
L_A-L_B	3.49±0.11	5.93±1.36
$L_A-L_B-L_C$	4.39±0.09	6.12±0.89
$L_A-L_B-L_C-L_D$	4.85±0.10	6.31±0.71

Table S18. Train and Test RMSEs Obtained using the kNN Algorithm for Various Subsets

set	train	test
L_A	2.69±0.30	5.67±1.67
L_B	7.17±1.77	8.18±3.82
L_C	3.15±0.34	5.71±1.95
L_D	4.87±0.75	8.57±2.46
L_A-L_B	3.68±0.28	6.90±2.13
$L_A-L_B-L_C$	3.68±0.66	6.41±1.37
$L_A-L_B-L_C-L_D$	3.48±0.25	6.36±1.04

Table S19. Train and Test RMSEs Obtained using the GB Algorithm for Various Subsetsⁱ

set	train	test
L_A	2.09±0.46	5.70±1.23
L_B	1.17±0.42	7.48±2.41
L_C	1.93±0.78	5.77±2.11
L_D	1.68±0.43	8.11±2.85
L_A-L_B	3.65±0.55	7.31±1.59
$L_A-L_B-L_C$	3.68±0.77	6.50±1.01
$L_A-L_B-L_C-L_D$	3.43±0.21	6.47±0.77

ⁱLigand subsets with very small number of samples (L_B , number of samples = 56) and (L_D , number of samples = 36) results in significant over-fitting.

Table S20. Train and Test RMSEs Obtained using the DT Algorithm for Various Subsetsⁱ

set	train	test
L_A	2.44±0.11	5.53±1.62
L_B	1.53±0.19	7.76±3.33
L_C	2.45±0.14	6.73±2.88
L_D	1.18±0.19	7.52±3.81
L_A-L_B	3.82±0.14	6.18±1.64
$L_A-L_B-L_C$	5.01±0.13	6.98±1.25
$L_A-L_B-L_C-L_D$	5.81±0.15	7.46±1.26

ⁱ Ligand subsets with very small number of samples (L_B , number of samples = 56) and (L_D , number of samples = 36) results in significant over-fitting.

Table S21. Train and Test RMSEs Obtained using the DNN Algorithm for Various Subsets ⁱ

set	train	test
L_A	2.22±0.37	5.27±1.49
L_B	3.56±0.84	7.56±2.91
L_C	2.96±0.50	4.78±1.45
L_D	5.40±0.70	8.02±2.28
L_A-L_B	3.58±0.59	6.05±1.76
$L_A-L_B-L_C$	3.54±0.45	5.82±0.93
$L_A-L_B-L_C-L_D$	4.48±0.46	6.32±0.90

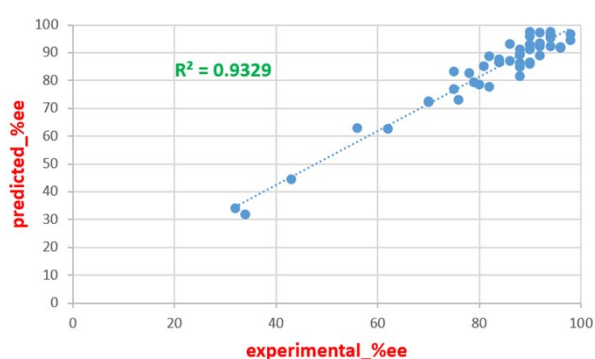
ⁱ All of these results were obtained using an 80:20 train-test ratio, with 80% of samples in the training set and 20% in the test set, this ensures an equitable comparison of different ML models.

Table S22. Train and Test RMSEs Obtained using the GPR_{RF} Algorithm for Various Subsets

set	train	test
L_A	1.97±0.8	6.09±2.3
L_B	4.91±0.46	7.19±2.43
L_C	1.60±0.15	4.51±1.91
L_D	6.49±0.45	8.71±2.4
L_A-L_B	3.55±0.58	6.35±1.49
$L_A-L_B-L_C$	2.22±0.2	6.26±1.17
$L_A-L_B-L_C-L_D$	3.64±0.14	6.32±0.89

(5.6) Performance of the DNN model in terms of the R-squared value

best run



a run closest to the average performance

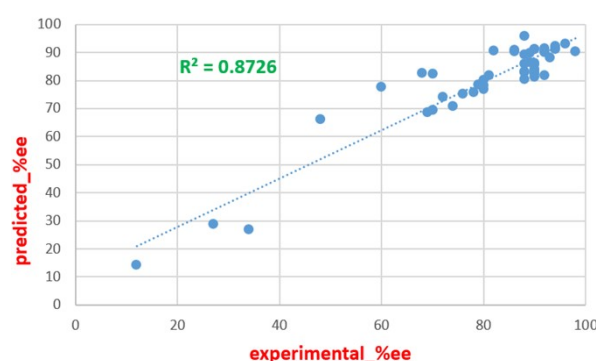


Fig S4. Performance of the DNN expressed in terms of the R-squared value for 48 test samples in the best run with an RMSE of 4.1 %ee and a typical run of RMSE 6.3 %ee which is the closest to the to the average RMSE over 100 runs.

(6) Performance with the real dataset consisting of 240 reactions as obtained using the DNN algorithm

Table S23. Train and Test RMSEs Obtained using DNN Algorithm using Real Data for Various Subsets

set	train	test
L_A	2.82 ± 0.41	6.62 ± 1.99
L_B	3.4 ± 0.57	8.32 ± 3.16
L_C	3.04 ± 0.63	9.99 ± 5.77
L_D	5.34 ± 0.66	9.20 ± 2.94
L_A-L_B	3.4 ± 0.43	8.05 ± 2.43
$L_A-L_B-L_C$	3.41 ± 0.62	8.67 ± 2.65
$L_A-L_B-L_C-L_D$	4.47 ± 0.78	8.58 ± 2.53

The reduction in the test RMSE after adding synthetic data for all the subsets clearly indicates that the problem of class imbalance could be reasonably addressed (Fig S5).

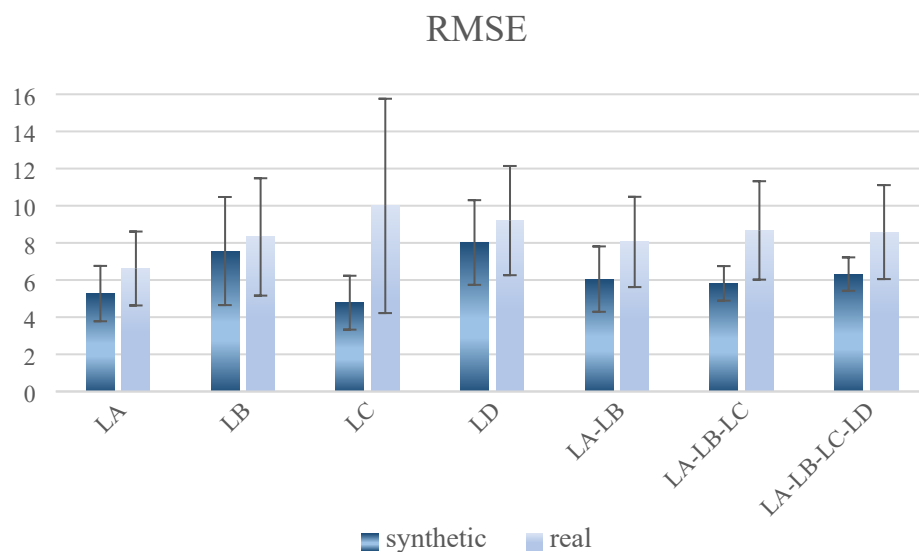


Fig S5. Comparison between test RMSEs of only real and real + synthetic datasets obtained using DNN algorithm. The error bars denote the standard deviations.

(7) A comparison of test and train RMSEs noted in different ML models

Overfitting is used to describe when a ML model fits its training data much superior whereas the trained algorithm is unable to make accurate predictions on the unseen data. As low-data ML models are prone to overfitting, it is important to evaluate overfitting for all models. This

is done by comparing train and test RMSEs. The following plots shows the variation train and test RMSEs for all 100 runs.

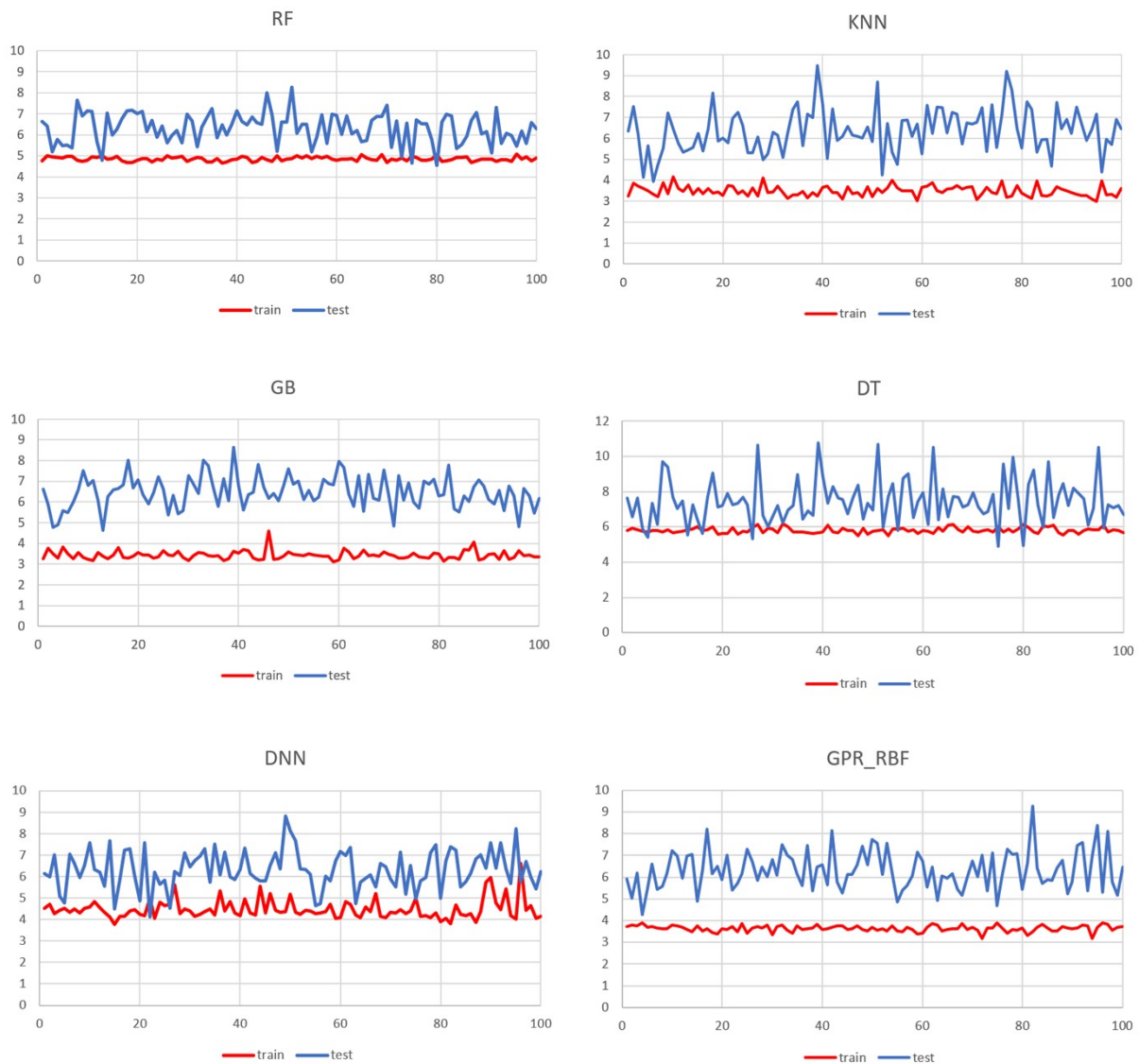


Fig S6. Plots of train and test RMSEs as obtained in different ML models across 100 runs.

(8) Performance analysis of ML models in different class intervals

During the model development phase, we considered 100 distinct runs with different test samples. The effect of the 80:20 train-test split ratio on the total number of 240 real samples results in 48 test samples, on which predictions were done. The prediction error in various

ranges such as 0-5, 5-10, 10-15, and >15 %*ee* for these 48 samples are shown below. The prediction error is the difference between the actual (experimental) and ML predicted %*ee*.

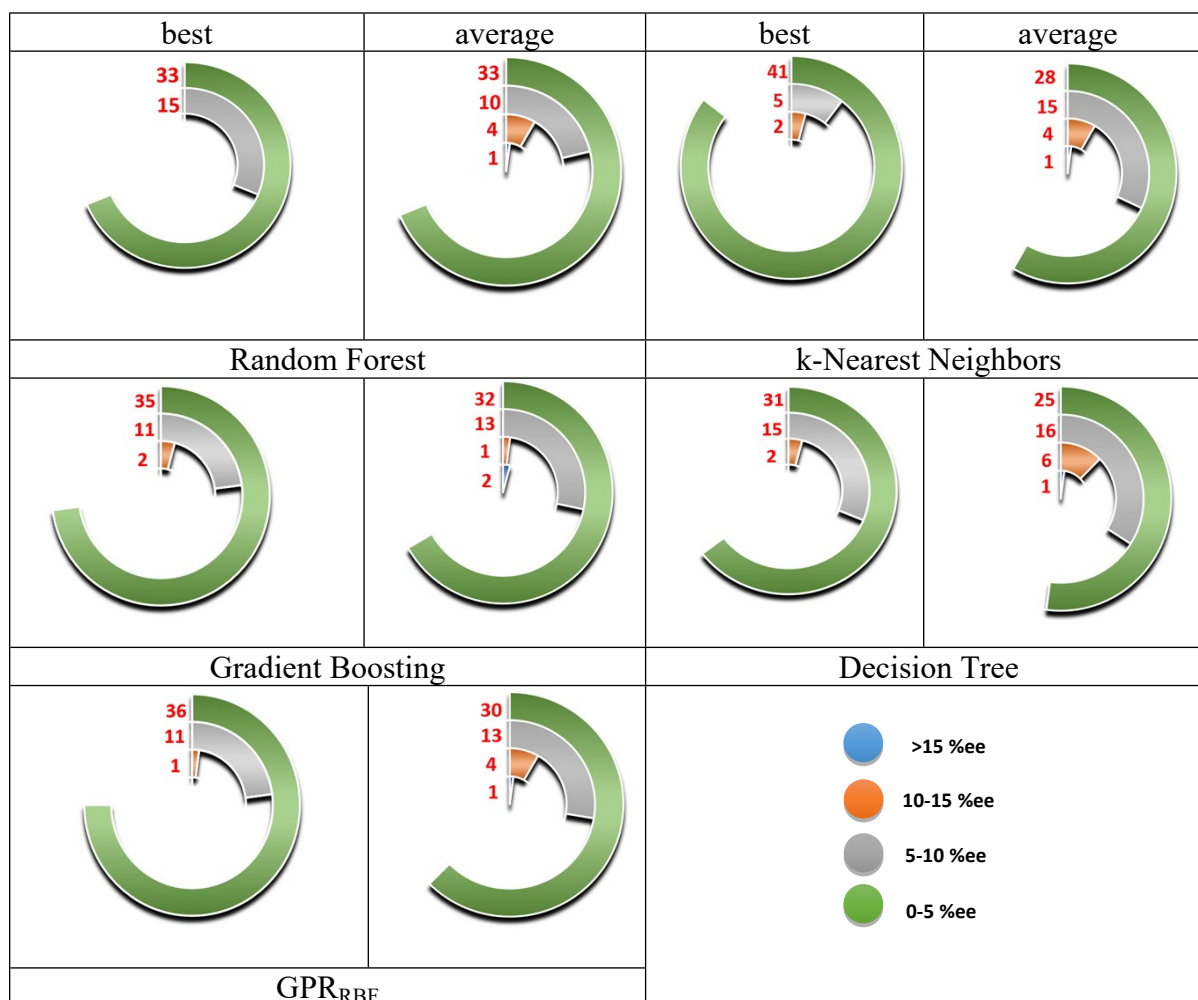


Fig S7. Performance of different ML algorithms in terms of the absolute error calculated as the difference between the predicted and actual %*ee* for 48 test samples in the best run and a typical run with an RMSE closest to the average RMSE over 100 runs. The pie charts show the number of samples exhibiting different ranges of the absolute error. 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 number of samples.

(9) Effect of train-test splitting

Table S24. Train and Test RMSEs for Different train-test Splits Obtained using the DNN Algorithm

train-test split ratio	train	test
90:10	4.50±0.40	6.14±1.31
80:20	4.48±0.46	6.32±0.90
70:30	4.36±0.54	6.39±0.83
60:40	4.27±0.42	6.59±0.70
50:50	4.18±0.57	6.75±0.67

(10) Assessment of feature importance

(10.1) Randomization of features

The outputs are shuffled at random among the rows (samples) such that no descriptor is associated to its true output value. This modified dataset is then augmented with synthetic data using SMOTE (SVM) technique. With these randomized feature values, we re-trained the model using DNN algorithm. The test and train RMSEs were found to be 17.72 ± 3.05 and 10.44 ± 1.53 respectively.

(10.2) Normally distributed set of random numbers

By replacing the original features with random numbers, with a mean of zero and a standard deviation of one, we created a new dataset. It is important to highlight that none of these generated values resemble or are related to the actual chemical descriptors. We trained our model on this new dataset augmented with synthetic data using the DNN algorithm. Inferior performances compared to that obtained with the original chemical descriptors are evident from the test and train RMSEs of 18.16 ± 4.14 and 0.75 ± 0.17 respectively.

(10.3) One-hot encoding

One-hot encoding can be used as a baseline model to examine the validity an ML model. In this experiment, each reaction component was encoded as a "one-hot" vector i.e., a binary vector 1 or 0 that only indicates the presence or absence of that component. With this one-hot encoded data, we trained our DNN algorithm. The test and train RMSEs were respectively found to be 14.98 ± 2.31 and 1.83 ± 0.27 .

(11) Correlation analysis

We used correlation analysis to examine the interdependencies between the features. Measure of co-linearity expressed as correlation coefficient in the range from +1 to -1. A positive correlation indicates that both variables move in the same direction, whereas a negative correlation indicates that as the value of one variable increases, the value of the other variable decreases. Value of zero indicates no co-linearity between the features. We performed the correlation analysis on the full feature matrix (240x153), the result is shown in [Fig S8](#).

For feature selection, we have chosen one of the features among the features with a correlation coefficient of 0.9 or higher (Table S25). The full feature matrix reduced to 240x114 from the original 240x153 dimension. We have used these 114 features to build the ML models with the DNN algorithm (Table S26). The test and train RMSEs were found to be inferior as compared to the performances with the original feature matrix (Table S27). Similarly, by setting the correlation coefficient to 0.8, the number of features is reduced to 83 (Table S28 S29). The DNN model with these 83 features resulted much poorer performance compared to the original feature list (Table S27).

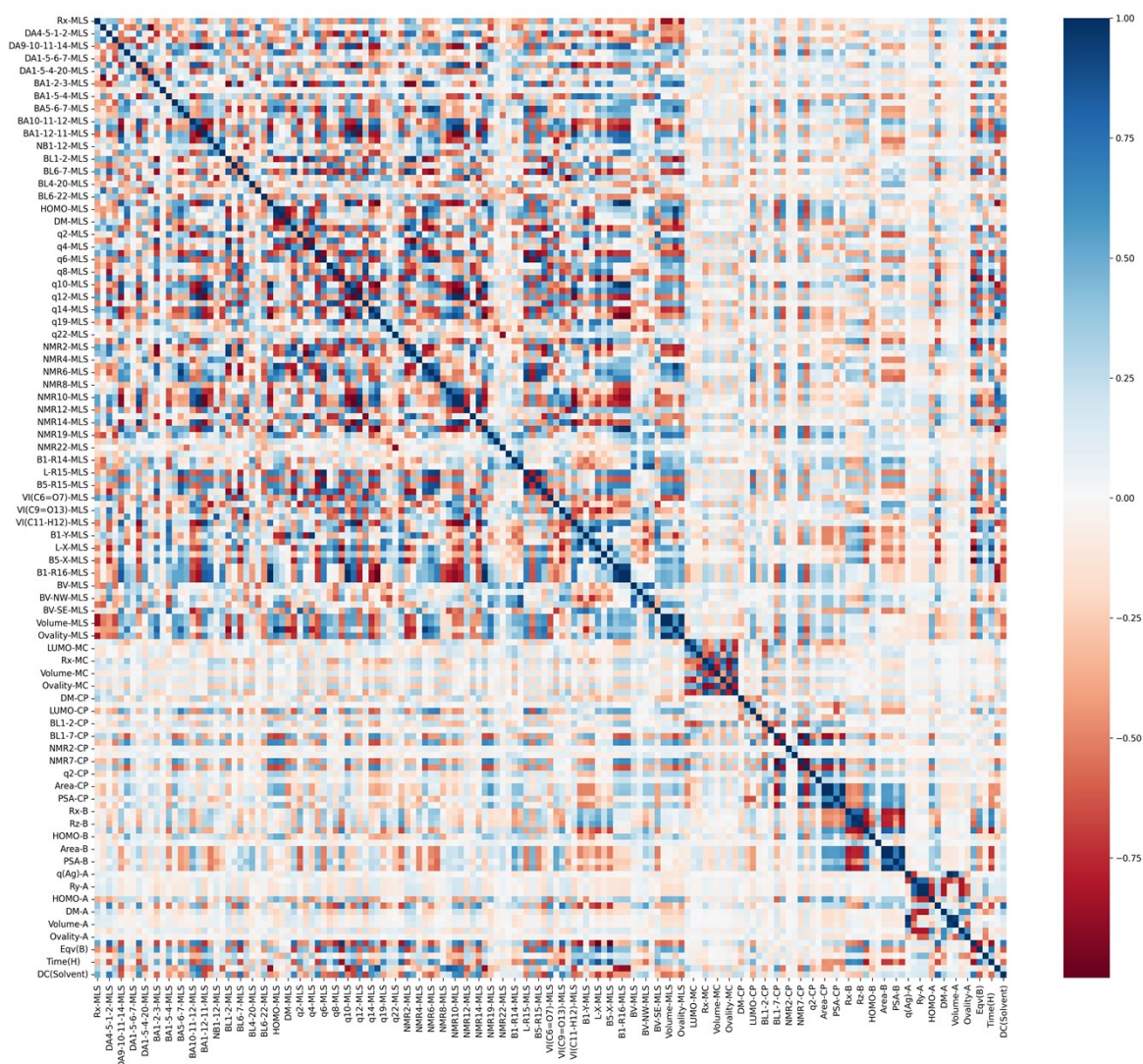


Fig S8. Pearson correlation matrix for full feature matrix

Table S25. List of Correlated Features Obtained after the Removal of the Correlated Features with Correlation Coefficient > 0.9. See Table S4 for Details of Feature

S. No.	parameters	correlated parameters
1	DA9-10-11-14-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS
2	DA1-12-11-14-MLS	DA1-12-11-14-MLS, NMR11-MLS, NMR14-MLS
3	BA5-6-7-MLS	BA5-6-7-MLS, NMR6-MLS
4	BA10-11-14-MLS	BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS
5	BA1-12-11-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, q11-MLS, VI(C11-H12)-MLS
6	BL1-2-MLS	BL1-2-MLS, NMR2-MLS, NMR3-MLS
7	BL6-7-MLS	BL6-7-MLS, q1-MLS, NMR2-MLS
8	BL10-15-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-

		R16-MLS, B5-R16-MLS
9	HOMO-MLS	HOMO-MLS, LUMO-MLS, NMR5-MLS, PSA-MLS
10	LUMO-MLS	HOMO-MLS, LUMO-MLS
11	DM-MLS	DM-MLS, q4-MLS, B1-Y-MLS
12	q1-MLS	BL6-7-MLS, q1-MLS, NMR2-MLS
13	q4-MLS	DM-MLS, q4-MLS, NMR3-MLS, B1-Y-MLS
14	q6-MLS	q6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
15	q9-MLS	q9-MLS, Eqv(L)
16	q10-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
17	q11-MLS	BA1-12-11-MLS, q11-MLS, VI(C11-H12)-MLS
18	q12-MLS	BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
19	NMR2-MLS	BL1-2-MLS, BL6-7-MLS, q1-MLS, NMR2-MLS
20	NMR3-MLS	BL1-2-MLS, q4-MLS, NMR3-MLS
21	NMR5-MLS	HOMO-MLS, NMR5-MLS, PSA-MLS
22	NMR6-MLS	BA5-6-7-MLS, NMR6-MLS, L-R15-MLS
23	NMR7-MLS	q6-MLS, NMR7-MLS, L-R15-MLS
24	NMR10-MLS	BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
25	NMR11-MLS	DA1-12-11-14-MLS, BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
26	NMR14-MLS	DA1-12-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
27	NMR15-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
28	L-R15-MLS	q6-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
29	B5-R15-MLS	q6-MLS, L-R15-MLS, B5-R15-MLS
30	VI(C11-H12)-MLS	BA1-12-11-MLS, q11-MLS, VI(C11-H12)-MLS
31	L-Y-MLS	L-Y-MLS, B5-Y-MLS
32	B1-Y-MLS	DM-MLS, q4-MLS, B1-Y-MLS
33	B5-Y-MLS	L-Y-MLS, B5-Y-MLS
34	L-X-MLS	L-X-MLS, B5-X-MLS
35	B5-X-MLS	L-X-MLS, B5-X-MLS
36	L-R16-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
37	B1-R16-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
38	B5-R16-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
39	Area-MLS	Area-MLS, Volume-MLS, Ovality-MLS
40	Volume-MLS	Area-MLS, Volume-MLS, Ovality-MLS
41	PSA-MLS	HOMO-MLS, NMR5-MLS, PSA-MLS
42	Ovality-MLS	Area-MLS, Volume-MLS, Ovality-MLS
43	PSA-MC	PSA-MC, q(Pd)-MC
44	q(Pd)-MC	PSA-MC, q(Pd)-MC
45	BL1-7-CP	BL1-7-CP, q1-CP
46	q1-CP	BL1-7-CP, q1-CP

47	Area-CP	Area-CP, Volume-CP, Ovality-CP
48	Volume-CP	Area-CP, Volume-CP, Ovality-CP
49	Ovality-CP	Area-CP, Volume-CP, Ovality-CP
50	Ry-B	Ry-B, Rz-B
51	Rz-B	Ry-B, Rz-B
52	Area-B	Area-B, Volume-B, Ovality-B
53	Volume-B	Area-B, Volume-B, Ovality-B
54	Ovality-B	Area-B, Volume-B, Ovality-B
55	q(Ag)-A	q(Ag)-A, Area-A, Volume-A
56	Ry-A	Ry-A, Rz-A
57	Rz-A	Ry-A, Rz-A
58	Area-A	q(Ag)-A, Area-A, Volume-A
59	Volume-A	q(Ag)-A, Area-A, Volume-A
60	Eqv(L)	q9-MLS, Eqv(L)

Table S26. List of 115 Features used for Building DNN Model

S. No.	parameters	S. No.	parameters	S. No.	parameters
1	Rx-MLS	40	q14-MLS	79	Ovality-MC
2	DA3-4-5-6-MLS	41	q15-MLS	80	DM-CP
3	DA4-5-1-2-MLS	42	q19-MLS	81	HOMO-CP
4	DA4-5-6-7-MLS	43	q20-MLS	82	LUMO-CP
5	DA9-10-11-14-MLS	44	q22-MLS	83	Rx-CP
6	DA9-10-11-12-MLS	45	NMR1-MLS	84	BL1-2-CP
7	DA1-5-6-7-MLS	46	NMR4-MLS	85	BL1-6-CP
8	DA1-12-11-14-MLS	47	NMR8-MLS	86	BL1-7-CP
9	DA1-5-4-20-MLS	48	NMR9-MLS	87	NMR1-CP
10	DA6-5-4-20-MLS	49	NMR12-MLS	88	NMR2-CP
11	BA1-2-3-MLS	50	NMR13-MLS	89	NMR6-CP
12	BA3-4-20-MLS	51	NMR19-MLS	90	NMR7-CP
13	BA1-5-4-MLS	52	NMR20-MLS	91	q2-CP
14	BA5-4-20-MLS	53	NMR22-MLS	92	q6-CP
15	BA5-6-7-MLS	54	L-R14-MLS	93	Area-CP
16	BA7-6-22-MLS	55	B1-R14-MLS	94	PSA-CP
17	BA10-11-12-MLS	56	B5-R14-MLS	95	Rx-B
18	BA10-11-14-MLS	57	B1-R15-MLS	96	Ry-B
19	BA5-1-12-MLS	58	VF(C6=O7)-MLS	97	DM-B
20	NB1-12-MLS	59	VI(C6=O7)-MLS	98	HOMO-B
21	NB7-12-MLS	60	VF(C9=O13)-MLS	99	LUMO-B
22	BL1-2-MLS	61	VI(C9=O13)-MLS	100	Area-B
23	BL5-6-MLS	62	VF(C11-H12)-MLS	101	PSA-B
24	BL6-7-MLS	63	L-Y-MLS	102	q(Ag)-A
25	BL11-12-MLS	64	L-X-MLS	103	Rx-A
26	BL4-20-MLS	65	B1-X-MLS	104	Ry-A

27	BL4-19-MLS	66	BV-MLS	105	HOMO-A
28	BL6-22-MLS	67	BV-SW-MLS	106	LUMO-A
29	BL10-15-MLS	68	BV-NW-MLS	107	DM-A
30	HOMO-MLS	69	BV-NE-MLS	108	PSA-A
31	DM-MLS	70	BV-SE-MLS	109	Ovality-A
32	q2-MLS	71	Area-MLS	110	Eqv(B)
33	q3-MLS	72	HOMO-MC	111	Eqv(A)
34	q5-MLS	73	LUMO-MC	112	Time(H)
35	q6-MLS	74	DM-MC	113	T (?C)
36	q7-MLS	75	Rx-MC	114	DC(Solvent)
37	q8-MLS	76	Ry-MC		
38	q9-MLS	77	Volume-MC		
39	q13-MLS	78	PSA-MC		

Table S27. Performance comparison after reduction of features based on correlation matrix

number of features	train	test
114	6.58±1.34	8.57±1.78
83	3.77±0.43	8.25±1.39

Table S28. List of Correlated Features Obtained after Removing the Correlated Features with Correlation Coefficient > 0.8

S. No.	parameters	correlated parameters
1	DA4-5-1-2-MLS	DA4-5-1-2-MLS, BL6-7-MLS, q9-MLS
2	DA9-10-11-14-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, q11-MLS, NMR15-MLS, VI(C11-H12)-MLS, B1-R16-MLS, B5-R16-MLS
3	DA1-12-11-14-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
4	BA1-2-3-MLS	BA1-2-3-MLS, q3-MLS, NMR5-MLS, NMR6-MLS, VF(C6=O7)-MLS, PSA-MLS
5	BA5-6-7-MLS	BA5-6-7-MLS, HOMO-MLS, q6-MLS, NMR5-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS, VF(C6=O7)-MLS, PSA-MLS
6	BA10-11-12-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, NMR9-MLS, NMR10-MLS, NMR11-MLS
7	BA10-11-14-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, q14-MLS, q15-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
8	BA1-12-11-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, q11-MLS, NMR15-MLS, VI(C11-H12)-MLS, B1-R16-MLS, B5-R16-MLS
9	NB1-12-MLS	NB1-12-MLS, VF(C11-H12)-MLS
10	BL1-2-MLS	BL1-2-MLS, BL6-7-MLS, q1-MLS, q4-MLS, NMR2-MLS, NMR3-MLS
11	BL6-7-MLS	DA4-5-1-2-MLS, BL1-2-MLS, BL6-7-MLS, q1-MLS, q5-MLS, q9-MLS, NMR2-MLS, Eqv(L)

12	BL11-12-MLS	BL11-12-MLS, NMR12-MLS
13	BL10-15-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
14	HOMO-MLS	BA5-6-7-MLS, HOMO-MLS, LUMO-MLS, NMR5-MLS, PSA-MLS
15	LUMO-MLS	HOMO-MLS, LUMO-MLS, NMR5-MLS, PSA-MLS
16	DM-MLS	DM-MLS, q4-MLS, B1-Y-MLS
17	q1-MLS	BL1-2-MLS, BL6-7-MLS, q1-MLS, q5-MLS, q9-MLS, NMR2-MLS
18	q2-MLS	q2-MLS, BV-SE-MLS
19	q3-MLS	BA1-2-3-MLS, q3-MLS
20	q4-MLS	BL1-2-MLS, DM-MLS, q4-MLS, NMR3-MLS, B1-Y-MLS
21	q5-MLS	BL6-7-MLS, q1-MLS, q5-MLS, q9-MLS, q14-MLS, B1-R15-MLS
22	q6-MLS	BA5-6-7-MLS, q6-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
23	q7-MLS	q7-MLS, q13-MLS
24	q8-MLS	q8-MLS, q13-MLS
25	q9-MLS	DA4-5-1-2-MLS, BL6-7-MLS, q1-MLS, q5-MLS, q9-MLS, NMR2-MLS, Eqv(L)
26	q10-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, q11-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
27	q11-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, q10-MLS, q11-MLS, NMR12-MLS, VI(C11-H12)-MLS
28	q12-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, q15-MLS, NMR9-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
29	q13-MLS	q7-MLS, q8-MLS, q13-MLS, VF(C6=O7)-MLS
30	q14-MLS	BA10-11-14-MLS, q5-MLS, q14-MLS, q15-MLS
31	q15-MLS	BA10-11-14-MLS, q12-MLS, q14-MLS, q15-MLS, NMR10-MLS, T (?C)
32	NMR1-MLS	NMR1-MLS, L-X-MLS, B5-X-MLS, Time(H)
33	NMR2-MLS	BL1-2-MLS, BL6-7-MLS, q1-MLS, q9-MLS, NMR2-MLS, NMR3-MLS
34	NMR3-MLS	BL1-2-MLS, q4-MLS, NMR2-MLS, NMR3-MLS, NMR13-MLS
35	NMR5-MLS	BA1-2-3-MLS, BA5-6-7-MLS, HOMO-MLS, LUMO-MLS, NMR5-MLS, NMR6-MLS, VF(C6=O7)-MLS, PSA-MLS
36	NMR6-MLS	BA1-2-3-MLS, BA5-6-7-MLS, q6-MLS, NMR5-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS, VF(C6=O7)-MLS
37	NMR7-MLS	BA5-6-7-MLS, q6-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
38	NMR9-MLS	BA10-11-12-MLS, q12-MLS, NMR9-MLS, VF(C9=O13)-MLS, T (?C)
39	NMR10-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, q15-MLS, NMR10-MLS, NMR11-MLS, NMR14-

		MLS
40	NMR11-MLS	DA1-12-11-14-MLS, BA10-11-12-MLS, BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
41	NMR12-MLS	BL11-12-MLS, q11-MLS, NMR12-MLS, VI(C11-H12)-MLS
42	NMR13-MLS	NMR3-MLS, NMR13-MLS
43	NMR14-MLS	DA1-12-11-14-MLS, BA10-11-14-MLS, q12-MLS, NMR10-MLS, NMR11-MLS, NMR14-MLS
44	NMR15-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
45	B5-R14-MLS	B5-R14-MLS, BV-NW-MLS, BV-NE-MLS
46	L-R15-MLS	BA5-6-7-MLS, q6-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
47	B1-R15-MLS	q5-MLS, B1-R15-MLS
48	B5-R15-MLS	BA5-6-7-MLS, q6-MLS, NMR6-MLS, NMR7-MLS, L-R15-MLS, B5-R15-MLS
49	VF(C6=O7)-MLS	BA1-2-3-MLS, BA5-6-7-MLS, q13-MLS, NMR5-MLS, NMR6-MLS, VF(C6=O7)-MLS
50	VF(C9=O13)-MLS	NMR9-MLS, VF(C9=O13)-MLS
51	VF(C11-H12)-MLS	NB1-12-MLS, VF(C11-H12)-MLS
52	VI(C11-H12)-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, q11-MLS, NMR12-MLS, VI(C11-H12)-MLS
53	L-Y-MLS	L-Y-MLS, B5-Y-MLS, L-X-MLS
54	B1-Y-MLS	DM-MLS, q4-MLS, B1-Y-MLS
55	B5-Y-MLS	L-Y-MLS, B5-Y-MLS
56	L-X-MLS	NMR1-MLS, L-Y-MLS, L-X-MLS, B5-X-MLS, Eqv(B), Time(H)
57	B5-X-MLS	NMR1-MLS, L-X-MLS, B5-X-MLS, Eqv(B), Time(H)
58	L-R16-MLS	BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
59	B1-R16-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
60	B5-R16-MLS	DA9-10-11-14-MLS, BA1-12-11-MLS, BL10-15-MLS, q10-MLS, NMR15-MLS, L-R16-MLS, B1-R16-MLS, B5-R16-MLS
61	BV-MLS	BV-MLS, BV-NW-MLS
62	BV-NW-MLS	B5-R14-MLS, BV-MLS, BV-NW-MLS
63	BV-NE-MLS	B5-R14-MLS, BV-NE-MLS
64	BV-SE-MLS	q2-MLS, BV-SE-MLS
65	Area-MLS	Area-MLS, Volume-MLS, Ovality-MLS
66	Volume-MLS	Area-MLS, Volume-MLS, Ovality-MLS
67	PSA-MLS	BA1-2-3-MLS, BA5-6-7-MLS, HOMO-MLS, LUMO-MLS, NMR5-MLS, PSA-MLS
68	Ovality-MLS	Area-MLS, Volume-MLS, Ovality-MLS
69	Rx-MC	Rx-MC, Ry-MC, q(Pd)-MC

70	Ry-MC	Rx-MC, Ry-MC
71	Volume-MC	Volume-MC, Ovality-MC
72	PSA-MC	PSA-MC, q(Pd)-MC
73	Ovality-MC	Volume-MC, Ovality-MC
74	q(Pd)-MC	Rx-MC, PSA-MC, q(Pd)-MC
75	BL1-7-CP	BL1-7-CP, q1-CP
76	NMR1-CP	NMR1-CP, NMR7-CP
77	NMR7-CP	NMR1-CP, NMR7-CP, Area-CP, Volume-CP, Ovality-CP
78	q1-CP	BL1-7-CP, q1-CP
79	Area-CP	NMR7-CP, Area-CP, Volume-CP, Ovality-CP
80	Volume-CP	NMR7-CP, Area-CP, Volume-CP, Ovality-CP
81	Ovality-CP	NMR7-CP, Area-CP, Volume-CP, Ovality-CP
82	Ry-B	Ry-B, Rz-B
83	Rz-B	Ry-B, Rz-B
84	DM-B	DM-B, Area-B, Volume-B, Ovality-B
85	Area-B	DM-B, Area-B, Volume-B, Ovality-B
86	Volume-B	DM-B, Area-B, Volume-B, Ovality-B
87	PSA-B	PSA-B, Ovality-B
88	Ovality-B	DM-B, Area-B, Volume-B, PSA-B, Ovality-B
89	q(Ag)-A	q(Ag)-A, Area-A, Volume-A
90	Rx-A	Rx-A, Rz-A
91	Ry-A	Ry-A, Rz-A
92	Rz-A	Rx-A, Ry-A, Rz-A
93	Area-A	q(Ag)-A, Area-A, Volume-A
94	Volume-A	q(Ag)-A, Area-A, Volume-A
95	Eqv(L)	BL6-7-MLS, q9-MLS, Eqv(L)
96	Eqv(B)	L-X-MLS, B5-X-MLS, Eqv(B)
97	Time(H)	NMR1-MLS, L-X-MLS, B5-X-MLS, Time(H)
98	T (?C)	q15-MLS, NMR9-MLS, T (?C)

Table S29. List of 115 Features used for Building DNN Model

S. No.	parameters	S. No.	parameters	S. No.	parameters
1	Rx-MLS	29	q8-MLS	57	LUMO-CP
2	DA3-4-5-6-MLS	30	q19-MLS	58	Rx-CP
3	DA4-5-1-2-MLS	31	q20-MLS	59	BL1-2-CP
4	DA4-5-6-7-MLS	32	q22-MLS	60	BL1-6-CP
5	DA9-10-11-14-MLS	33	NMR1-MLS	61	BL1-7-CP
6	DA9-10-11-12-MLS	34	NMR4-MLS	62	NMR1-CP
7	DA1-5-6-7-MLS	35	NMR8-MLS	63	NMR2-CP
8	DA1-12-11-14-MLS	36	NMR19-MLS	64	NMR6-CP
9	DA1-5-4-20-MLS	37	NMR20-MLS	65	q2-CP
10	DA6-5-4-20-MLS	38	NMR22-MLS	66	q6-CP
11	BA1-2-3-MLS	39	L-R14-MLS	67	PSA-CP
12	BA3-4-20-MLS	40	B1-R14-MLS	68	Rx-B
13	BA1-5-4-MLS	41	B5-R14-MLS	69	Ry-B
14	BA5-4-20-MLS	42	VI(C6=O7)-MLS	70	DM-B
15	BA5-6-7-MLS	43	VI(C9=O13)-MLS	71	HOMO-B

16	BA7-6-22-MLS	44	L-Y-MLS	72	LUMO-B
17	BA5-1-12-MLS	45	B1-X-MLS	73	PSA-B
18	NB1-12-MLS	46	BV-MLS	74	q(Ag)-A
19	NB7-12-MLS	47	BV-SW-MLS	75	Rx-A
20	BL1-2-MLS	48	Area-MLS	76	Ry-A
21	BL5-6-MLS	49	HOMO-MC	77	HOMO-A
22	BL11-12-MLS	50	LUMO-MC	78	LUMO-A
23	BL4-20-MLS	51	DM-MC	79	DM-A
24	BL4-19-MLS	52	Rx-MC	80	PSA-A
25	BL6-22-MLS	53	Volume-MC	81	Ovality-A
26	DM-MLS	54	PSA-MC	82	Eqv(A)
27	q2-MLS	55	DM-CP	83	DC(Solvent)
28	q7-MLS	56	HOMO-CP		

(12) Details of parameters and performance of different ML algorithm for the

Unbound/Free ligand model

We considered a simpler model in which each of the reacting components, i.e., ligand, catalyst, and coupling partner, was treated as a free entity in its native state. Features were collected from the optimized geometries (following the procedure as described in section 4) of all these individual molecules. Tables S30 and S31 lists the parameters of these components.

Table S30. Parameter Details of Various Reacting Components for the Unbound Model

ligand (L)			
local parameters			
bond length (BL)	1-2, 2-3, 3-4, 5-6, 4-9, 5-10, 1-11, 3-8, 3-7	bond angle (BA)	1-2-3, 2-3-4, 3-4-5, 3-4-9, 8-3-4, 8-3-2, 8-3-7, 4-5-10, 2-1-11
dihedral angle (DA)	1-2-3-4, 2-3-4-5, 2-3-4-9, 3-4-5-6, 1-2-3-8, 1-2-3-7, 7-3-4-5		
charge (q)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	NMR shift (NMR)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
vibrational frequency	5-6, 4-9	sterimol(L, B1, B5)	R ¹⁴ , R ¹⁵ , Y

(VF) & intensity (VI)			
global parameters			
HOMO energy, LUMO energy, Dipole Moment (DM), Rotational Constants (R_x , R_y), Area, Volume, PSA, Ovality			
substrate (S)			
local parameters			
bond length (BL)	1-2, 4-5, 4-7, 3-8, 1-12	bond angle (BA)	1-2-3, 3-4-5, 4-3-8, 2-1-12, 3-2-6
dihedral angle (DA)	2-3-4-5, 12-1-2-6	NMR shift (NMR)	1, 2, 3, 4, 5, 6, 7, 8, 12
charge (q)	1, 2, 3, 4, 6	sterimol(L, B1, B5)	R^{16} , X
global parameters			
HOMO energy, LUMO energy, Dipole Moment (DM), rotational constant (R_x), Volume, PSA			
coupling partner (CP)			
local parameters			
bond length (BL)	bond length (BL)	bond length (BL)	
global parameters			
HOMO energy, LUMO energy, Dipole moment (DM), Rotational constant (R_x), Area, PSA			
metal-catalyst precursor (MC)			
local parameter			
q(Pd)			
global parameters			
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y), Volume, PSA, Ovality			
base (B)			
global parameters			
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y , R_z), Area, Volume, PSA, Ovality			
additive (A)			
local parameter			
q(Ag)			
global parameters			
HOMO energy, LUMO energy, Dipole moment (DM), Rotational Constants (R_x , R_y , R_z), Area, Volume, PSA, Ovality			
secondary parameters			
dielectric constant of solvent (DC)			

experimental parameters

amount of ligand - Eqv(L), amount of base - Eqv(B), amount of additive - Eqv(A), reaction time and reaction temperature.

Table S31. Full List of Features for the Unbound Model

P1	Rx-L	P81	BA1-2-3-S
P2	Ry-L	P82	BA3-4-5-S
P3	DA1-2-3-4-L	P83	BA4-3-8-S
P4	DA2-3-4-5-L	P84	BA2-1-12-S
P5	DA2-3-4-9-L	P85	BA3-2-6-S
P6	DA3-4-5-6-L	P86	DA2-3-4-5-S
P7	DA1-2-3-8-L	P87	DA12-1-2-6-S
P8	DA1-2-3-7-L	P88	HOMO-S
P9	DA7-3-4-5-L	P89	LUMO-S
P10	BA1-2-3-L	P90	Rx-S
P11	BA2-3-4-L	P91	NMR1-S
P12	BA3-4-5-L	P92	NMR2-S
P13	BA3-4-9-L	P93	NMR3-S
P14	BA8-3-4-L	P94	NMR4-S
P15	BA8-3-2-L	P95	NMR5-S
P16	BA8-3-7-L	P96	NMR6-S
P17	BA4-5-10-L	P97	NMR7-S
P18	BA2-1-11-L	P98	NMR8-S
P19	BL1-2-L	P99	NMR12-S
P20	BL2-3-L	P100	q1-S
P21	BL3-4-L	P101	q2-S
P22	BL5-6-L	P102	q3-S
P23	BL4-9-L	P103	q4-S
P24	BL5-10-L	P104	q6-S
P25	BL1-11-L	P105	Volume-S
P26	BL3-8-L	P106	PSA-S
P27	BL3-7-L	P107	HOMO-MC
P28	HOMO-L	P108	LUMO-MC
P29	LUMO-L	P109	DM-MC
P30	DM-L	P110	Rx-MC
P31	q1-L	P111	Ry-MC
P32	q2-L	P112	Volume-MC
P33	q3-L	P113	PSA-MC
P34	q4-L	P114	Ovality-MC
P35	q5-L	P115	q(Pd)-MC
P36	q6-L	P116	DM-CP
P37	q7-L	P117	HOMO-CP
P38	q8-L	P118	LUMO-CP
P39	q9-L	P119	Rx-CP
P40	q10-L	P120	BL1-2-CP
P41	q11-L	P121	BL1-6-CP
P42	NMR1-L	P122	BL1-7-CP
P43	NMR2-L	P123	NMR1-CP

P44	NMR3-L	P124	NMR2-CP
P45	NMR4-L	P125	NMR6-CP
P46	NMR5-L	P126	NMR7-CP
P47	NMR6-L	P127	q1-CP
P48	NMR7-L	P128	q2-CP
P49	NMR8-L	P129	q6-CP
P50	NMR9-L	P130	Area-CP
P51	NMR10-L	P131	Volume-CP
P52	NMR11-L	P132	PSA-CP
P53	L-R14-L	P133	Ovality-CP
P54	B1-R14-L	P134	Rx-B
P55	B5-R14-L	P135	Ry-B
P56	L-R15-L	P136	Rz-B
P57	B1-R15-L	P137	DM-B
P58	B5-R15-L	P138	HOMO-B
P59	L-Y-L	P139	LUMO-B
P60	B1-Y-L	P140	Area-B
P61	B5-Y-L	P141	Volume-B
P62	Area-L	P142	PSA-B
P63	Volume-L	P143	Ovality-B
P64	PSA-L	P144	q(Ag)-A
P65	Ovality-L	P145	Rx-A
P66	VF(C5=O6)-L	P146	Ry-A
P67	VI(C5=O6)-L	P147	Rz-A
P68	VF(N4-H9)-L	P148	HOMO-A
P69	VI(N4-H9)-L	P149	LUMO-A
P70	L-X-S	P150	DM-A
P71	B1-X-S	P151	Area-A
P72	B5-X-S	P152	Volume-A
P73	L-R16-S	P153	PSA-A
P74	B1-R16-S	P154	Ovality-A
P75	B5-R16-S	P155	Eqv(L)
P76	BL1-2-S	P156	Eqv(B)
P77	BL4-5-S	P157	Eqv(A)
P78	BL4-7-S	P158	Time(H)
P79	BL3-8-S	P159	T (°C)
P80	BL1-12-S	P160	DC(Solvent)

(12.1) Performance with different ML algorithms for different subsets

Table S32. Train and Test RMSEs Obtained using the RF Algorithm for Various Subsets

set	train	test
L_A	2.29±0.12	4.68±1.47
L_B	2.55±0.2	6.17±1.7
L_C	3.13±0.24	6.92±1.92
L_D	2.86±0.26	8.04±3.42
L_A-L_B	4.28±0.19	8.11±2.35
$L_A-L_B-L_C$	5.35±0.18	8.34±2.05

$L_A-L_B-L_C-L_D$	5.22±0.11	7.52±1.05
-------------------	-----------	-----------

Table S33. Train and Test RMSEs Obtained using the kNN Algorithm for Various Subsets

set	train	test
L_A	2.65±0.21	4.65±1.24
L_B	4.66±1.57	8.09±2.36
L_C	6.07±1.53	10.98±6.55
L_D	4.71±1.09	8.73±2.48
L_A-L_B	3.78±0.3	7.59±2.54
$L_A-L_B-L_C$	4.58±0.89	8.07±3.02
$L_A-L_B-L_C-L_D$	4.02±0.43	7.86±2.47

Table S34. Train and Test RMSEs Obtained using the GB Algorithm for Various Subsetsⁱ

set	train	test
L_A	1.95±0.46	5.00±1.27
L_B	1.77±0.57	7.77±2.10
L_C	2.34±0.86	7.44±3.08
L_D	1.39±0.33	7.74±2.92
L_A-L_B	3.52±0.84	8.67±2.05
$L_A-L_B-L_C$	3.82±0.66	8.47±1.41
$L_A-L_B-L_C-L_D$	4.09±0.25	8.31±1.32

ⁱLigand subsets with very small number of samples (L_B , number of samples = 56) and (L_D , number of samples = 36) results in significant over-fitting.

Table S35. Train and Test RMSEs Obtained using the DT Algorithm for Various Subsetsⁱ

set	train	test
L_A	2.73±0.12	5.72±1.65
L_B	1.57±0.19	7.03±2.33
L_C	2.95±0.19	7.9±2.49
L_D	1.16±0.15	7.92±4.43
L_A-L_B	4.34±0.28	8.73±3.33
$L_A-L_B-L_C$	6.66±0.35	10.08±3.19
$L_A-L_B-L_C-L_D$	6.27±0.32	9.97±2.02

ⁱLigand subsets with very small number of samples (L_B , number of samples = 56) and (L_D , number of samples = 36) results in significant over-fitting.

Table S36. Train and Test RMSEs Obtained using the DNN Algorithm for Various Subsets

set	train	test
L_A	2.96±0.40	5.32±1.40
L_B	4.95±1.13	7.96±2.26
L_C	7.44±2.11	10.17±4.16
L_D	5.79±0.34	7.99±1.84
L_A-L_B	5.45±2.01	7.51±2.18
$L_A-L_B-L_C$	6.70±1.50	8.55±2.67
$L_A-L_B-L_C-L_D$	7.91±1.33	8.54±2.00

Table S37. Train and Test RMSEs Obtained using the GPR_{RBF} Algorithm for VariousSubsetsⁱ

set	train	test
L_A	1.6±0.39	4.7±1.51
L_B	6.81±1.19	7.84±2.35
L_C	0.98±1.3	17.15±10.96
L_D	6.91±0.37	8.65±2.17
L_A-L_B	3.75±0.77	6.39±1.82
L_A-L_B-L_C	1.94±0.38	10.37±4.33
L_A-L_B-L_C-L_D	3.01±0.32	8.97±2.94

ⁱLigand subsets with relatively smaller number of samples (**L_A**, number of samples = 69) and (**L_C**, number of samples = 79) results in significant over-fitting.

(13) Performance of the DNN algorithm for MLS model with different binary and ternary combinations of samples

Table S38. Train and Test RMSEs Obtained using the DNN Algorithm for Various Subsets

set	train	test
L_A-L_D	4.29±0.38	6.50±1.29
L_A-L_C	3.16±0.54	5.80±1.11
L_B-L_C	4.59±0.49	6.49±1.20
L_B-L_D	2.76±0.47	4.99±0.89
L_C-L_D	4.24±0.70	7.60±3.42
L_A-L_B-L_D	4.91±0.45	6.99±1.19
L_A-L_C-L_D	4.11±0.54	7.07±2.60
L_B-L_C-L_D	4.53±0.43	7.79±2.43

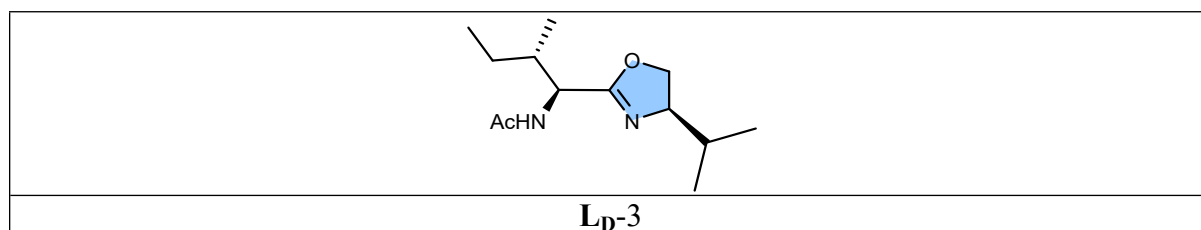
(14) Details of out-of-bag sets

External validation is essential for determining the quality of a machine learning model. An out-of-sample test, which involves testing the model performance on samples that are not present in the training set, is a better way to assess the model generalizability. We have considered three different sets of out-of-bag samples as listed in the following section.

(14.1) Set-1

Ligand diversity

L_D (N-acyl-protected amino oxazoline (APAO))
--



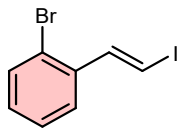
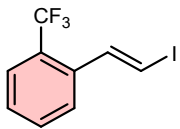
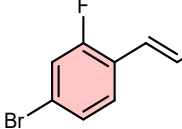
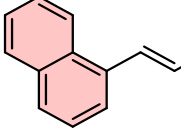
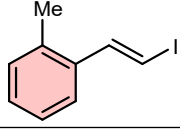
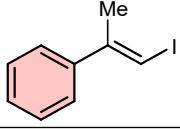
Coupling partner diversity

Table S39. Identities and Notations of the Coupling Partners used for Pd(II)-Catalyzed Enantioselective Arylation of the Cyclobutyl Carboxylic Amide

CP2	CP24	CP25	CP26	CP27	CP28
CP29	CP30	CP35	CP36	CP37	CP38
CP39	CP40	CP41	CP56	CP55	CP89
CP90	CP91				

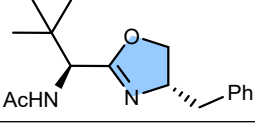
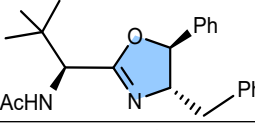
Table S40. Identities and Notations of the Coupling Partners used for Pd(II)-Catalyzed Enantioselective Alkenylation of the Cyclobutyl Carboxylic Amide

CP68	CP69	CP72	CP73

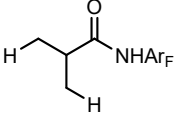
CP74	CP76	CP78	CP79
			
CP81	CP82	CP83	CP85
			
CP87	CP88		

(14.2) Set-2

Ligand diversity

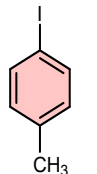
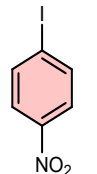
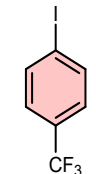
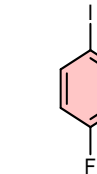
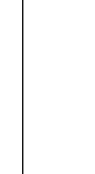

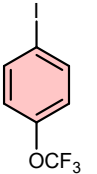
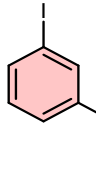
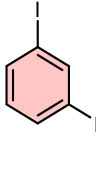
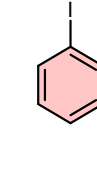
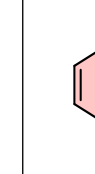
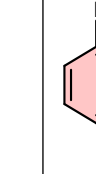
L_D (N-acyl-protected amino oxazoline [APAO])	
	
LD-1	LD-4

Substrate diversity

acyclic substrate

Ar = 4-(CF ₃)-C ₆ F ₄
S6

Coupling partner diversity

Table S41. Identities and Notations of the Coupling Partners used for Pd(II)-Catalyzed Enantioselective C–H Arylation of Isobutyric Acid

					
CP2	CP25	CP26	CP28	CP29	CP30
					
CP32	CP33	CP36	CP37	CP38	CP39


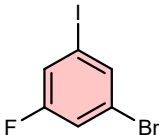
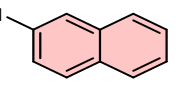
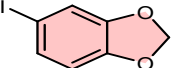
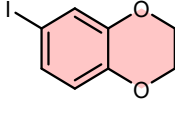
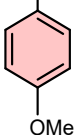
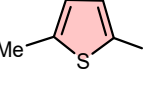
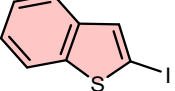


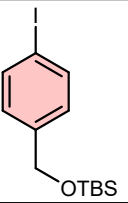
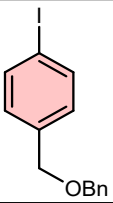
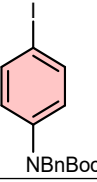
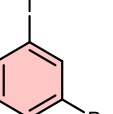
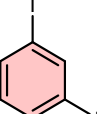
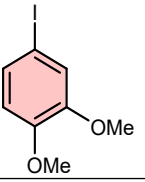
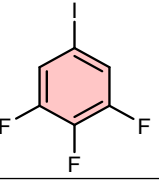
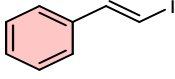
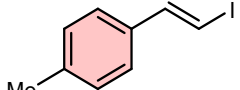
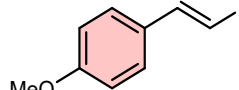
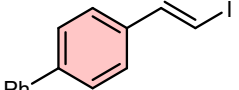
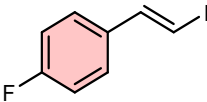
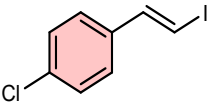
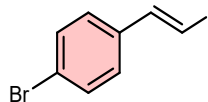
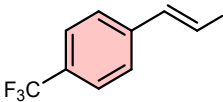
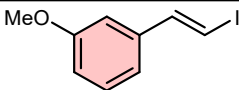
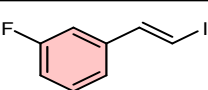
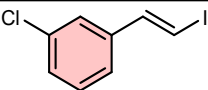
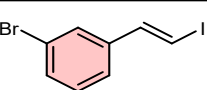
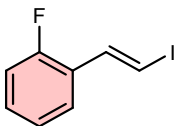
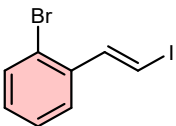
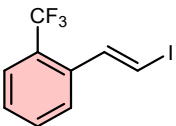
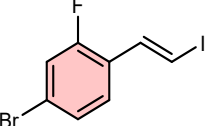
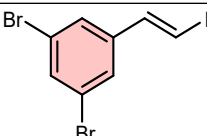
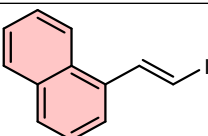
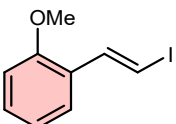
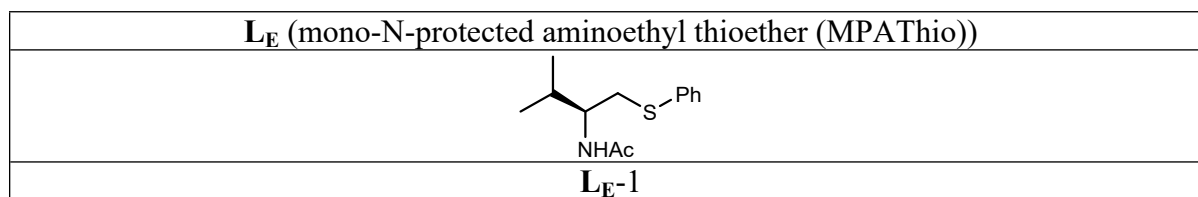
					
CP40	CP52	CP53	CP54	CP55	CP56
					
CP57	CP58	CP59	CP60	CP61	CP62
					
CP63	CP64	CP65	CP66	CP67	

Table S42. Identities and Notations of the Coupling Partners used for Pd(II)-Catalyzed Enantioselective Vinylation of Isobutyric Acid

			
CP68	CP69	CP70	CP71
			
CP72	CP73	CP74	CP75
			
CP76	CP77	CP78	CP79
			
CP80	CP81	CP82	CP83
			
CP84	CP85	CP86	

(14.3) Set-3

Ligand diversity



Substrate diversity

Table S43. Identities and Notations of the Substrates used for γ -C(sp³)-H Arylation of Free Cyclopropylmethylamine

S7	S8	S9	S10	S11
S12	S13	S14	S15	S16
S17	S18	S19	S20	S21
S22	S23			

Table S44. Variance Analysis to Understand the Effect of Inclusion of Randomized Samples in the Training Set for Predictions on the Out-of-Bag Samples

feature	variance before addition of samples	variance after inclusion of four samples	variance difference
VF(C9=O13)-MLS	420.4970688	30027.13541	29606.64
VI(C9=O13)-MLS	63731.9411	72633.71554	8901.774
NMR2-MLS	4274.198402	5441.466174	1167.268
NMR1-MLS	1343.426868	1598.388449	254.9616
NMR9-MLS	5.667467416	247.620762	241.9533
NMR8-MLS	23.01744518	237.9591261	214.9417
Area-MLS	18066.93177	17854.36202	212.5698
Volume-MLS	21162.45422	20972.27192	190.1823
VF(C11-H12)-MLS	13800.39711	13960.94809	160.551
NMR13-MLS	103.0361718	221.2909603	118.2548
VI(C6=O7)-MLS	22074.19406	22165.08087	90.88682

Area-CP	2748.941174	2821.361494	72.42032
NMR3-MLS	2612.476455	2684.487823	72.01137
Volume-CP	2754.571195	2823.184739	68.61354
BA10-11-14-MLS	6.373564591	37.65649288	31.28293
NMR7-MLS	1866.966003	1837.195294	29.77071
Area-B	518.8457345	547.5890209	28.74329
Volume-MC	1917.789869	1891.176001	26.61387
Volume-B	334.09499	359.0887874	24.9938
VF(C6=O7)-MLS	1062.719007	1085.650668	22.93166
NMR15-MLS	3665.106113	3683.207868	18.10176
Area-A	3.537894889	21.41916759	17.88127
Volume-A	4.08883604	19.34769639	15.25886
PSA-B	249.7082519	262.9220656	13.21381
VI(C11-H12)-MLS	27727.06607	27714.79066	12.27541

To examine the unexpectedly higher performance of the trained DNN on the Set-3 OOB samples, we have compared the feature variance in our datasets (a) before and (b) after the inclusion of the four randomized samples to the training set. Interestingly, a features exhibited, as described in Fig S9, are found to exhibit large differences in the variance upon adding four OOB samples. The vibrational frequency and intensity of (C9–X13, where X = O or H) are the top two features with the largest difference in variance. It may further be noted that the samples in the training set contains a carbonyl group (C=O) whereas and in Set-3, it is a C–H bond. The 25 features with the high variance difference are listed in the Table S44. With these characteristics, the migration of four samples seems to improve the learning capability of the DNN to be deployed for predicting samples from Set-3 out-of-bag.

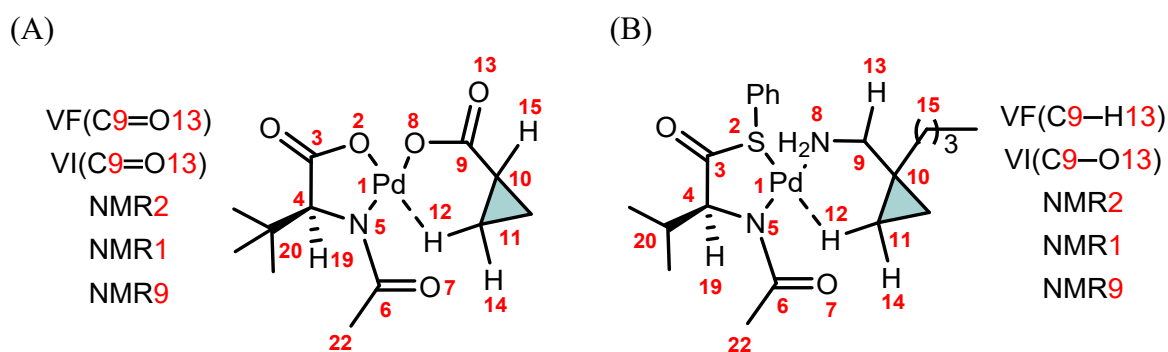


Fig S9. A representative example of the MLS model employed for feature extraction for a (A) typical sample in the general training set, and (B) one from Set-3 out-of-bag sample

exhibiting larger differences in the feature variance. Only relevant features with highest differences in variance between the test and out-of-bag sets are shown here. See Table S5 for additional details of features.

15. Selective Reduction of Features

We have handpicked a set of features to assess the importance of features in the learning of the DNN model (Table S45). As can be seen, the removal of the mechanistically important features, resulted in a test RMSE of 9.92 ± 1.79 %*ee*.

Table S45. Train and Test RMSEs Obtained using the DNN Model with Reduced Number of Features

	features removed = 62	train	test
	type of features: NPA charges, NMR, sterimol, vibrational frequency and vibrational intensity, buried volume	7.97 ± 1.31	13.39 ± 3.08
	features retained = 91		
features	'Rx-MLS', 'DA3-4-5-6-MLS', 'DA4-5-1-2-MLS', 'DA4-5-6-7-MLS', 'DA9-10-11-14-MLS', 'DA9-10-11-12-MLS', 'DA1-5-6-7-MLS', 'DA1-12-11-14-MLS', 'DA1-5-4-20-MLS', 'DA6-5-4-20-MLS', 'BA1-2-3-MLS', 'BA3-4-20-MLS', 'BA1-5-4-MLS', 'BA5-4-20-MLS', 'BA5-6-7-MLS', 'BA7-6-22-MLS', 'BA10-11-12-MLS', 'BA10-11-14-MLS', 'BA1-12-11-MLS', 'BA5-1-12-MLS', 'NB1-12-MLS', 'NB7-12-MLS', 'BL1-2-MLS', 'BL5-6-MLS', 'BL6-7-MLS', 'BL11-12-MLS', 'BL4-20-MLS', 'BL4-19-MLS', 'BL6-22-MLS', 'BL10-15-MLS', 'HOMO-MLS', 'LUMO-MLS', 'DM-MLS', 'Area-MLS', 'Volume-MLS', 'PSA-MLS', 'Ovality-MLS', 'HOMO-MC', 'LUMO-MC', 'DM-MC', 'Rx-MC', 'Ry-MC', 'Volume-MC', 'PSA-MC', 'Ovality-MC', 'q(Pd)-MC', 'DM-CP', 'HOMO-CP', 'LUMO-CP', 'Rx-CP', 'BL1-2-CP', 'BL1-6-CP', 'BL1-7-CP', 'NMR1-CP', 'NMR2-CP', 'NMR6-CP', 'NMR7-CP', 'q1-CP', 'q2-CP', 'q6-CP', 'Area-CP', 'Volume-CP', 'PSA-CP', 'Ovality-CP', 'Rx-B', 'Ry-B', 'Rz-B', 'DM-B', 'HOMO-B', 'LUMO-B', 'Area-B', 'Volume-B', 'PSA-B', 'Ovality-B', 'q(Ag)-A', 'Rx-A', 'Ry-A', 'Rz-A', 'HOMO-A', 'LUMO-A', 'DM-A', 'Area-A', 'Volume-A', 'PSA-A', 'Ovality-A', 'Eqv(L)', 'Eqv(B)', 'Eqv(A)', 'Time(H)', 'T (°C)', 'DC(Solvent)'		

16. ML Model Interpretability using SHAP (Shapley Additive Explainability)

Interpretability is important to gather additional trust of ML models, which in turn, can help in making informed decisions about how to improve them. The interpretation of complex DNN model findings adds a layer of model validation, converting it to a grey box rather than

a black box model. In this work, we used the Shapley additive explanations (SHAP) method, which is used to explain the ML model predictions.¹⁴ The SHAP method is based on the Shapley values that quantify the contributions of individual participants to a collaborative game in the context of game theory.¹⁵ This idea is extended to feature attributions by treating a team performance as an output and each player contribution as the feature significance. The best test run (RMSE of 4.1 %) of the trained DNN model based on molecular features derived from the MLS entity was selected. The trained model was then used in the feature attribution procedure. Herein, we have employed the DeepExplainer to approximate SHAP values for DNN, which is an enhanced version of the DeepLIFT algorithm (Deep SHAP).¹⁶

17. Conformational Sampling using CREST

The CREST sampler is used for carrying out additional conformational analyses. A representative set of eight different MLS complexes, from three subsets (L_A , L_C , L_D) was subjected to the CREST sampling first. The single point energy calculations were performed on all the CREST derived conformers using the DFT(B3LYP-D3) method. The two lowest energy conformers (denoted as crest conformer set 1 and crest conformer set 2) were chosen for further geometry optimization using same level of theory, and the energies were then compared with the original geometries (Table S46). Ten out of the sixteen CREST conformers were found to be of higher energies than the original conformers of different ligand sets. The six lower energy conformers, identified by the sampler showed only a modest difference (~ 0.5 kcal/mol lower) as compared to the original conformer. It may also be noted that our initial geometry optimization was performed by considering various possible noncovalent interactions (lone pair- π , π - π stacking, C-H $\cdots\pi$) and C-H \cdots Pd agostic interaction that are likely in the metal-ligand-substrate (MLS) complex. These findings indicate that the geometries of the MLS entity and the molecular descriptors collected from them are adequate for the present study.

Table S46. Comparison of energies (in kcal/mol) between the conformers obtained through the CREST sampling and the original conformer of the MLS entity

MLS entity	original (kcal/mol)	crest conformer set 1 (kcal/mol)	crest conformer set 2 (kcal/mol)
L_A-2-s2	0	0.4	0.4
L_A-3-s2	0	-0.7	-0.5
L_A-4-s2	0	0.2	0.2
L_A-5-s2	0	-0.4	0.2
L_D-1-s3	0	3.5	3.5
L_D-2-s4	0	0.3	0.3
L_C-1-s2	0	-0.5	-0.2
L_C-2-s2	0	-0.02	0.6

Additionally, we have extracted descriptors from the CREST sampled MLS conformers. To examine the variation in descriptor value, we have calculated % change in each of the descriptor values as, $\Delta D = ((D_{\text{org}} - D_{\text{CREST}})/D_{\text{org}})*100$, where D_{org} and D_{CREST} respectively denote the descriptor values of our original MLS entity and that of the CREST conformer. For each of the 16 conformers, the % change was then plotted in the form of a heat map in Fig. S10. The larger number of cells spanned by the green color indicates that majority of descriptors did not change with respect to the descriptors obtained from the original conformer. Some of the parameters (e.g., dihedral angles DA4-5-1-2-MLS and DA1-5-6-7-MLS) showed a large variation as these are for different conformers. For the higher energy conformers of the **L_D-1-s3** MLS system (Table S46), the vibrational intensity and NMR values (VI(C6=O7)-MLS, VI(C11-H12)-MLS, and NMR7) were found to be different than the original conformer.

Using the descriptors obtained from the CREST conformers, two new datasets were formed by replacing original descriptors. Test and train RMSE obtained by using our DNN algorithm on these new datasets are tabulated in Table S47. An increase in the test RMSE for both the new conformer sets further demonstrates that the molecular descriptor derived from original MLS conformers are more suitable for the system examined in this study.

Table S47. Comparison of the model performance obtained using the original conformer of the MLS entity and the additional conformers as obtained through the CREST sampling

Dataset	train	test
original MLS dataset	4.48±0.46	6.32±0.90
CREST conformation set 1	4.53±0.41	6.45±0.95
CREST conformation set 2	4.57±0.53	6.42±1.02

18. Analyses of Performance of the DNN in the Low %*ee* Region

In the case of Pd-catalyzed enantioselective β -C(sp³)-H functionalization using the MPAA ligand family only very few examples are experimentally known with low %*ee*. Such a distribution of %*ee* is therefore the ground truth pertaining to this catalytic asymmetric transformation.

Additional analyses were performed as follows,

(A) The test RMSEs of reactions belonging to different class intervals were evaluated, which are found to be >80 (5.73), <80 (8.08), <60 (6.37). Good performance noted in the <60%*ee* class suggests the model generalizability even for low %*ee* reactions.

(B) Another out-of-bag (OOB) set from the original 240 reactions was also created. Four samples with <80 %*ee* were chosen at random as the new OOB set, leaving 236 reactions for the training set. This new OOB set was evaluated using the DNN model trained on 236 reactions. This process was repeated using five such OOB sets, with each run carrying 4 new randomly chosen samples in the OOB set. Similarly, another set of OOB consisting of 6 randomly chosen reactions were also considered, results of both these are provided in Table

S47. These findings demonstrate that the MLS model is effective in predicting low %*ee* reactions as well.

Table S47. Test RMSEs Obtained using the DNN Model with Different OOB sets with samples with <80 %*ee*

randomly chosen 4 reactions		randomly chosen 6 reactions	
runs	RMSE	runs	RMSE
1	7.42	1	8.89
2	9.17	2	7.41
3	8.53	3	8.54
4	8.09	4	7.84
5	5.87	5	6.54
avg.±std. dev.	7.82±1.13	avg.±std. dev.	7.84±0.83

19. Workflow for New Experiment Planning

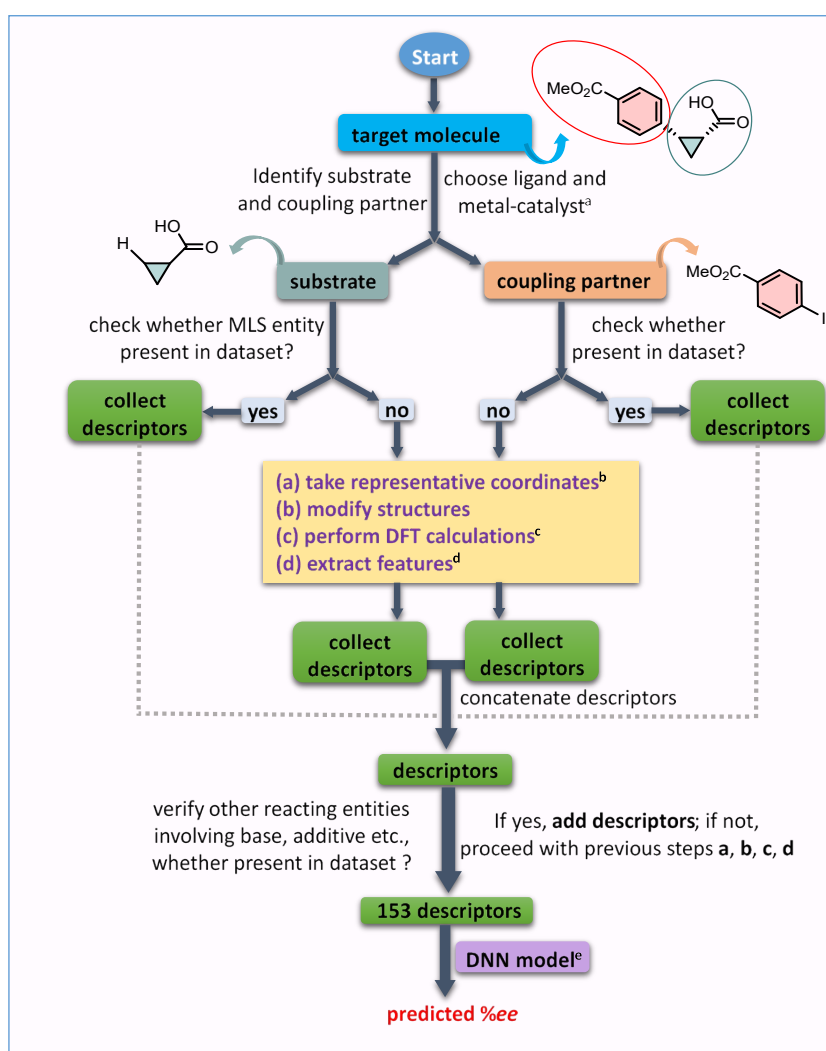


Fig S11. A step-by-step workflow is provided for doing new experiments. ^a Select the catalyst and ligand when designing a new reaction. ^b The Github repository contains a representative set of Cartesian coordinates for all reacting components. ^{c,d} Use the methods described in Sections 3 and 4 for quantum chemical calculations and feature extraction. ^e Use the *DNN-new-exp.py* python file from the Github repository for out-of-bag samples prediction.

References

-
- (a) Shen, P.-X.; Hu, L.; Shao, Q.; Hong, K.; Yu, J.-Q. *J. Am. Chem. Soc.* **2018**, *140*, 6545–6549. (b) Hu, L.; Shen, P.-X.; Shao, Q.; Hong, K.; Qiao, J. X.; Yu, J.-Q. *Angew.Chem.Int. Ed.* **2019**, *58*, 2134–2138. (c) Xiao, K.-J.; Lin, D.W.; Miura, M.; Zhu, R.-Y.;Gong, W.; Wasa, M.; Yu, J.-Q. *J. Am. Chem. Soc.* **2014**, *136*, 8138–8142. (d) Wu, Q.-F.; Wang, X.-B.; Shen, P.-X.; Yu, J.-Q. *ACS Catal.* **2018**, *8*, 2577–2581.
 - (a) Liu, B.; Romine, A. M.; Rubel, C. Z.; Engle, K. M.; Shi, B.-F. *Chem. Rev.* **2021**, *121*, 14957–15074. (b) Hao, W.; Bay, K. L.; Harris, C. F.; King, D. S.; Guzei, I. A.; Aristov, M. M.; Zhuang, Z.; Plata, R. E.; Hill, D. E.; Houk, K. N.; Berry, J. F.; Yu, J.-Q.; Blackmond, D. G. *ACS Catal.* **2021**, *11*, 11040–11048.
 - (a) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. *Sci. Rep.* **2017**, *7*, 3582. (b) Mitchell, J. B. *OWiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481. (c) Reid, J. P.; Sigman, M. S. *Nature* **2019**, *571*, 343–348.
 - The experimental reaction conditions are crucial to chemical reactivity. Changes in ligand/Pd-catalyst loading can as well affect the enantioselectivity. This is study, we have incorporated reaction conditions, even as the reactants were the same. Inclusion of reaction conditions as parameters provide improves the number of samples in then dataset.

-
5. Gaussian 09, Revision D.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W. Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2013.
6. (a) Slater, J. C. *Quantum Theory of Molecules and Solids, Vol. 4: The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974. (b) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211. (c) Becke, A. D. *Phys. Rev. A Gen. Phys.* **1988**, *38*, 3098–3100. (d) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B Condens. Matter* **1988**, *37*, 785–789. (e) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (f) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104. (g) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.
7. (a) Häussermann, U.; Dolg, M.; Stoll, H.; Preuss, H.; Schwerdtfeger, P.; Pitzer, R. M. *Mol. Phys.* **1993**, *78*, 1211–1224. (b) Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. *Theo. Chim. Acta* **1990**, *77*, 123–141.
8. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.
9. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. *ACS Catal.* **2019**, *9*, 2313–2323.
10. Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo,

L. Nat. Chem. **2019**, *11*, 872–879.

11. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

12. Scikit-learn: Machine Learning in Python, Pedregosa et al., *JMLR* **2011**, *12*, 2825–2830.

13. Hackeling, G. *Mastering Machine Learning with Scikit-Learn: Apply Effective Learning Algorithms to Real-World Problems using Scikit-Learn*. Packt Publishing Ltd, 2014.

14. (a) Lundberg, S.; Lee, S.-I. *Adv. Neural Information Processing*; Curran Associates, **2017**, 4765–4774. (b) Sundararajan, M.; Najmi, *AarXiv [cs.AI]*, 2019. <https://arxiv.org/abs/1908.08474>. (c) *arXiv [stat.ML]*, 2019. <https://arxiv.org/abs/1910.13413>.

15. Shapley, L. S. *Contrib. Teor. Games* **1953**, *2*, 307–317.

16. Shrikumar, A.; Greenside, P.; Kundaje, A. *arXiv [cs.CV]* **2017**. <https://doi.org/10.48550/ARXIV.1704.02685>.