

Machine Learning-Based Yield Prediction for First-Row Transition Metal Catalyzed Cross-Coupling Reactions

C. Rajalakshmi ^a, Vivek Vijay ^a, Abhirami Vijayakumar ^a, Parvathi Santhoshkumar ^a, John B Kottooran ^a,
Ann Miriam Abraham ^a, G Krishnaveni ^a, C S Anjanakutty ^a, Binuja Varghese ^a, Vibin Ipe Thomas ^{*a}
^a *Department of Chemistry, CMS College Kottayam (Autonomous), Kottayam, Kerala, India-686001.*

Supplementary Information

Table of Contents

Table S1: Regression results of our dataset employing DRFP, RDkitFP, and RXNFP as featurization methods.	2
Table S2: Regression results of our dataset employing DFT as a featurization method.	12
Table S3: Higher correlation results of all subsets made by considering the count of each element.	15
Table S4: Classification results of our dataset employing DRFP, RDkitFP, and RXNFP as featurization methods.	16
Table S5: Classification results of our dataset employing DFT as a featurization method.	20
Table S6: Higher classification results of all subsets made by considering count of each element.	21

Regression results:

Table S1: Regression results of our Dataset, employing DRFP, RDkitFP, RxnFP as featurisation methods.

Model type	Featurisation method	epoch	learning rate	test size	iterations	Stratification type	RMSE	MAE	Correlation
Random Forest	DRFP			0.2	10	mechanism	24.06	18.66	0.45
Random Forest	DRFP			0.2	10	Substrate class	24.16	18.86	0.46
Random Forest	DRFP			0.2	10	Coupling partner class	26.55	20.32	0.3
Random Forest	DRFP			0.2	10	No stratification	24.28	18.74	0.4
Random Forest	RDkitFP			0.2	10	mechanism	24.73	19.26	0.42
Random Forest	RDkitFP			0.2	10	Substrate class	25.79	19.81	0.38
Random Forest	RDkitFP			0.2	10	Coupling partner class	26.34	20.12	0.31
Random Forest	RDkitFP			0.2	10	No stratification	24.54	18.6	0.39
Random Forest	RxnFP			0.2	10	mechanism	26.08	21.26	0.36
Random Forest	RxnFP			0.2	10	Substrate class	27.76	23.43	0.29
Random Forest	RxnFP			0.2	10	Coupling partner class	27.53	22.16	0.25
Random Forest	RxnFP			0.2	10	No stratification	26.99	21.38	0.26
Random Forest HPT	DRFP			0.2	10	mechanism	25.03	19.46	0.39
Random Forest HPT	DRFP			0.2	10	Substrate class	24.87	19.22	0.39
Random Forest HPT	DRFP			0.2	10	Coupling partner class	24.76	19.14	0.4
Random Forest HPT	DRFP			0.2	10	No stratification	25.25	19.52	0.37
Random Forest HPT	RDkitFP			0.2	10	mechanism	25.13	19.74	0.38
Random Forest HPT	RDkitFP			0.2	10	Substrate class	24.72	19.19	0.39
Random Forest HPT	RDkitFP			0.2	10	Coupling partner class	24.58	19.25	0.41

Random Forest HPT	RDkitFP			0.2	10	No stratification	25.44	19.75	0.36
Random Forest HPT	RxnFP			0.2	10	mechanism	27.57	22.54	0.26
Random Forest HPT	RxnFP			0.2	10	Substrate class	26.96	21.88	0.28
Random Forest HPT	RxnFP			0.2	10	Coupling partner class	27.2	22.09	0.28
Random Forest HPT	RxnFP			0.2	10	No stratification	28	22.65	0.23
Nural Network	DRFP	800	0.0001	0.2	10	mechanism	28.8	24.22	0.19
Nural Network	DRFP	800	0.0001	0.2	10	mechanism	27.81	23.01	0.24
Nural Network	DRFP	1000	0.0001	0.2	10	Substrate class	28.33	23.58	0.2
Nural Network	DRFP	1000	0.0001	0.2	10	Substrate class	27.37	22.39	0.26
Nural Network	DRFP	800	0.0001	0.2	10	Coupling partner class	28.67	24.05	0.2
Nural Network	DRFP	800	0.0001	0.2	10	Coupling partner class	27.63	22.71	0.26
Nural Network	DRFP	1000	0.0001	0.2	10	No stratification	28.78	24.09	0.19
Nural Network	DRFP	1000	0.0001	0.2	10	No stratification	28.01	23.09	0.23
Nural Network	RDkitFP	800	0.0001	0.2	10	mechanism	29.56	24.82	0.15
Nural Network	RDkitFP	800	0.0001	0.2	10	mechanism	29.76	24.85	0.13
Nural Network	RDkitFP	1000	0.0001	0.2	10	Substrate class	29.15	24.3	0.16
Nural Network	RDkitFP	1000	0.0001	0.2	10	Substrate class	29.42	24.39	0.14
Nural Network	RDkitFP	800	0.0001	0.2	10	Coupling partner class	28.73	24.03	0.2
Nural Network	RDkitFP	800	0.0001	0.2	10	Coupling partner class	28.96	24.07	0.18
Nural Network	RDkitFP	1000	0.0001	0.2	10	No stratification	29.5	24.78	0.15
Nural Network	RDkitFP	1000	0.0001	0.2	10	No stratification	29.66	24.77	0.14
Nural Network	RxnFP	800	0.0001	0.2	10	mechanism	31.28	27.21	0.04
Nural Network	RxnFP	800	0.0001	0.2	10	mechanism	30.77	26.71	0.07
Nural Network	RxnFP	1000	0.0001	0.2	10	Substrate class	31.04	26.89	0.04

Nural Network	RxnFP	1000	0.0001	0.2	10	Substrate class	30.5	26.31	0.08
Nural Network	RxnFP	800	0.0001	0.2	10	Coupling partner class	31.38	27.4	0.04
Nural Network	RxnFP	800	0.0001	0.2	10	Coupling partner class	30.97	26.99	0.07
Nural Network	RxnFP	1000	0.0001	0.2	10	No stratification	31.28	27.07	0.04
Nural Network	RxnFP	1000	0.0001	0.2	10	No stratification	30.81	26.54	0.07
Random Forest	DRFP			0.2	5	mechanism	24.13	18.79	0.45
Random Forest	DRFP			0.2	5	Substrate class	24.1	18.98	0.46
Random Forest	DRFP			0.2	5	Coupling partner class	26.42	20.29	0.31
Random Forest	DRFP			0.2	5	No stratification	24.37	18.86	0.4
Random Forest	RDkitFP			0.2	5	mechanism	24.64	19.05	0.43
Random Forest	RDkitFP			0.2	5	Substrate class	25.66	19.62	0.39
Random Forest	RDkitFP			0.2	5	Coupling partner class	26.37	20.29	0.31
Random Forest	RDkitFP			0.2	5	No stratification	24.61	18.79	0.39
Random Forest	RxnFP			0.2	5	mechanism	26.4	21.51	0.34
Random Forest	RxnFP			0.2	5	Substrate class	27.91	23.62	0.28
Random Forest	RxnFP			0.2	5	Coupling partner class	27.83	22.36	0.23
Random Forest	RxnFP			0.2	5	No stratification	27.16	21.52	0.25
Random Forest HPT	DRFP			0.2	5	mechanism	25.25	19.83	0.37
Random Forest HPT	DRFP			0.2	5	Substrate class	24.14	18.81	0.43
Random Forest HPT	DRFP			0.2	5	Coupling partner class	25.13	19.56	0.4
Random Forest HPT	DRFP			0.2	5	No stratification	25.69	19.69	0.35
Random Forest HPT	RDkitFP			0.2	5	mechanism	24.84	19.51	0.39
Random Forest HPT	RDkitFP			0.2	5	Substrate class	24.55	19.13	0.41

Random Forest HPT	RDkitFP			0.2	5	Coupling partner class	24.59	19.26	0.43
Random Forest HPT	RDkitFP			0.2	5	No stratification	25.47	19.67	0.36
Random Forest HPT	RxnFP			0.2	5	mechanism	27.36	22.41	0.26
Random Forest HPT	RxnFP			0.2	5	Substrate class	27.28	22.35	0.27
Random Forest HPT	RxnFP			0.2	5	Coupling partner class	27.35	22	0.29
Random Forest HPT	RxnFP			0.2	5	No stratification	27.82	22.37	0.24
Nural Network	DRFP	800	0.0001	0.2	5	mechanism	28.3	23.84	0.21
Nural Network	DRFP	800	0.0001	0.2	5	mechanism	27.64	22.92	0.25
Nural Network	DRFP	1000	0.0001	0.2	5	Substrate class	28.42	23.85	0.21
Nural Network	DRFP	1000	0.0001	0.2	5	Substrate class	27.17	22.27	0.28
Nural Network	DRFP	800	0.0001	0.2	5	Coupling partner class	28.54	23.9	0.23
Nural Network	DRFP	800	0.0001	0.2	5	Coupling partner class	27.62	22.79	0.28
Nural Network	DRFP	1000	0.0001	0.2	5	No stratification	29.06	24.16	0.17
Nural Network	DRFP	1000	0.0001	0.2	5	No stratification	28.46	23.3	0.2
Nural Network	RDkitFP	800	0.0001	0.2	5	mechanism	29.36	24.68	0.15
Nural Network	RDkitFP	800	0.0001	0.2	5	mechanism	29.81	24.89	0.12
Nural Network	RDkitFP	1000	0.0001	0.2	5	Substrate class	29.08	24.41	0.17
Nural Network	RDkitFP	1000	0.0001	0.2	5	Substrate class	29.51	24.62	0.15
Nural Network	RDkitFP	800	0.0001	0.2	5	Coupling partner class	28.6	23.89	0.22
Nural Network	RDkitFP	800	0.0001	0.2	5	Coupling partner class	28.84	24.05	0.21
Nural Network	RDkitFP	1000	0.0001	0.2	5	No stratification	29.91	25.06	0.12
Nural Network	RDkitFP	1000	0.0001	0.2	5	No stratification	30.01	25.12	0.12
Nural Network	RxnFP	800	0.0001	0.2	5	mechanism	31.13	27.15	0.05
Nural Network	RxnFP	800	0.0001	0.2	5	mechanism	30.68	26.72	0.07

Nural Network	RxnFP	1000	0.0001	0.2	5	Substrate class	31.2	27	0.05
Nural Network	RxnFP	1000	0.0001	0.2	5	Substrate class	30.44	26.18	0.09
Nural Network	RxnFP	800	0.0001	0.2	5	Coupling partner class	31.6	27.61	0.05
Nural Network	RxnFP	800	0.0001	0.2	5	Coupling partner class	31.19	27.12	0.08
Nural Network	RxnFP	1000	0.0001	0.2	5	No stratification	31.18	27.03	0.04
Nural Network	RxnFP	1000	0.0001	0.2	5	No stratification	30.78	26.52	0.07
Random Forest	DRFP			0.3	10	mechanism	25.01	19.01	0.38
Random Forest	DRFP			0.3	10	Substrate class	24.67	18.94	0.41
Random Forest	DRFP			0.3	10	Coupling partner class	26.11	20.57	0.35
Random Forest	DRFP			0.3	10	No stratification	24.53	19.05	0.4
Random Forest	RDkitFP			0.3	10	mechanism	25.53	19.62	0.35
Random Forest	RDkitFP			0.3	10	Substrate class	25.34	19.37	0.38
Random Forest	RDkitFP			0.3	10	Coupling partner class	26.42	20.51	0.33
Random Forest	RDkitFP			0.3	10	No stratification	25.15	19.47	0.37
Random Forest	RxnFP			0.3	10	mechanism	27.5	22.45	0.25
Random Forest	RxnFP			0.3	10	Substrate class	28.11	23.43	0.24
Random Forest	RxnFP			0.3	10	Coupling partner class	27.76	22.79	0.26
Random Forest	RxnFP			0.3	10	No stratification	26.97	21.67	0.28
Random Forest HPT	DRFP			0.3	10	mechanism	25.59	19.96	0.36
Random Forest HPT	DRFP			0.3	10	Substrate class	25.33	19.57	0.37
Random Forest HPT	DRFP			0.3	10	Coupling partner class	25.09	19.58	0.38
Random Forest HPT	DRFP			0.3	10	No stratification	25.22	19.63	0.37
Random Forest HPT	RDkitFP			0.3	10	mechanism	25.65	20.19	0.36
Random Forest HPT	RDkitFP			0.3	10	Substrate class	25.02	19.56	0.39

Random Forest HPT	RDkitFP			0.3	10	Coupling partner class	25.11	19.68	0.38
Random Forest HPT	RDkitFP			0.3	10	No stratification	25.32	19.85	0.37
Random Forest HPT	RxnFP			0.3	10	mechanism	28.21	23.13	0.22
Random Forest HPT	RxnFP			0.3	10	Substrate class	27.26	22.28	0.27
Random Forest HPT	RxnFP			0.3	10	Coupling partner class	27.55	22.53	0.26
Random Forest HPT	RxnFP			0.3	10	No stratification	27.75	22.48	0.24
Nural Network	DRFP	800	0.0001	0.3	10	mechanism	28.86	24.31	0.19
Nural Network	DRFP	800	0.0001	0.3	10	mechanism	28.18	23.47	0.22
Nural Network	DRFP	1000	0.0001	0.3	10	Substrate class	28.47	23.81	0.21
Nural Network	DRFP	1000	0.0001	0.3	10	Substrate class	27.56	22.67	0.26
Nural Network	DRFP	800	0.0001	0.3	10	Coupling partner class	28.49	23.8	0.21
Nural Network	DRFP	800	0.0001	0.3	10	Coupling partner class	27.71	22.83	0.25
Nural Network	DRFP	1000	0.0001	0.3	10	No stratification	28.9	24.17	0.18
Nural Network	DRFP	1000	0.0001	0.3	10	No stratification	28.03	23.08	0.23
Nural Network	RDkitFP	800	0.0001	0.3	10	mechanism	29.5	24.77	0.15
Nural Network	RDkitFP	800	0.0001	0.3	10	mechanism	29.75	24.87	0.13
Nural Network	RDkitFP	1000	0.0001	0.3	10	Substrate class	29.29	24.43	0.16
Nural Network	RDkitFP	1000	0.0001	0.3	10	Substrate class	29.56	24.58	0.15
Nural Network	RDkitFP	800	0.0001	0.3	10	Coupling partner class	28.88	24.17	0.18
Nural Network	RDkitFP	800	0.0001	0.3	10	Coupling partner class	29.1	24.25	0.17
Nural Network	RDkitFP	1000	0.0001	0.3	10	No stratification	29.42	24.58	0.15
Nural Network	RDkitFP	1000	0.0001	0.3	10	No stratification	29.58	24.59	0.14
Nural Network	RxnFP	800	0.0001	0.3	10	mechanism	31.48	27.34	0.03
Nural Network	RxnFP	800	0.0001	0.3	10	mechanism	31.09	26.96	0.05

Nural Network	RxnFP	1000	0.0001	0.3	10	Substrate class	31.42	27.2	0.03
Nural Network	RxnFP	1000	0.0001	0.3	10	Substrate class	30.62	26.4	0.08
Nural Network	RxnFP	800	0.0001	0.3	10	Coupling partner class	31.34	27.28	0.04
Nural Network	RxnFP	800	0.0001	0.3	10	Coupling partner class	30.87	26.77	0.07
Nural Network	RxnFP	1000	0.0001	0.3	10	No stratification	31.38	27.14	0.03
Nural Network	RxnFP	1000	0.0001	0.3	10	No stratification	30.88	26.55	0.06
Random Forest	DRFP			0.3	5	mechanism	25.08	19.04	0.38
Random Forest	DRFP			0.3	5	Substrate class	24.7	18.93	0.41
Random Forest	DRFP			0.3	5	Coupling partner class	26.15	20.51	0.35
Random Forest	DRFP			0.3	5	No stratification	24.38	19.01	0.41
Random Forest	RDkitFP			0.3	5	mechanism	25.85	20.01	0.34
Random Forest	RDkitFP			0.3	5	Substrate class	25.29	19.33	0.38
Random Forest	RDkitFP			0.3	5	Coupling partner class	26.49	20.54	0.33
Random Forest	RDkitFP			0.3	5	No stratification	25.22	19.68	0.37
Random Forest	RxnFP			0.3	5	mechanism	27.24	22.43	0.27
Random Forest	RxnFP			0.3	5	Substrate class	27.99	23.2	0.24
Random Forest	RxnFP			0.3	5	Coupling partner class	27.73	22.7	0.26
Random Forest	RxnFP			0.3	5	No stratification	27.37	22.12	0.26
Random Forest HPT	DRFP			0.3	5	mechanism	25.75	20.13	0.34
Random Forest HPT	DRFP			0.3	5	Substrate class	24.57	19.11	0.41
Random Forest HPT	DRFP			0.3	5	Coupling partner class	25.73	20.12	0.36
Random Forest HPT	DRFP			0.3	5	No stratification	25.17	19.61	0.38
Random Forest HPT	RDkitFP			0.3	5	mechanism	25.57	20.01	0.35
Random Forest HPT	RDkitFP			0.3	5	Substrate class	24.64	19.27	0.41

Random Forest HPT	RDkitFP			0.3	5	Coupling partner class	25.27	19.83	0.39
Random Forest HPT	RDkitFP			0.3	5	No stratification	25.39	19.76	0.36
Random Forest HPT	RxnFP			0.3	5	mechanism	28.07	22.92	0.22
Random Forest HPT	RxnFP			0.3	5	Substrate class	27.24	22.37	0.28
Random Forest HPT	RxnFP			0.3	5	Coupling partner class	27.74	22.66	0.26
Random Forest HPT	RxnFP			0.3	5	No stratification	27.85	22.58	0.24
Nural Network	DRFP	800	0.0001	0.3	5	mechanism	28.63	24.13	0.19
Nural Network	DRFP	800	0.0001	0.3	5	mechanism	28.11	23.37	0.22
Nural Network	DRFP	1000	0.0001	0.3	5	Substrate class	28.59	24.19	0.2
Nural Network	DRFP	1000	0.0001	0.3	5	Substrate class	27.25	22.45	0.27
Nural Network	DRFP	800	0.0001	0.3	5	Coupling partner class	28.76	24.22	0.2
Nural Network	DRFP	800	0.0001	0.3	5	Coupling partner class	27.87	23	0.25
Nural Network	DRFP	1000	0.0001	0.3	5	No stratification	29.18	24.43	0.16
Nural Network	DRFP	1000	0.0001	0.3	5	No stratification	28.14	23.01	0.22
Nural Network	RDkitFP	800	0.0001	0.3	5	mechanism	29.29	24.59	0.15
Nural Network	RDkitFP	800	0.0001	0.3	5	mechanism	29.43	24.62	0.14
Nural Network	RDkitFP	1000	0.0001	0.3	5	Substrate class	29.05	24.3	0.18
Nural Network	RDkitFP	1000	0.0001	0.3	5	Substrate class	29.47	24.57	0.15
Nural Network	RDkitFP	800	0.0001	0.3	5	Coupling partner class	28.98	24.3	0.19
Nural Network	RDkitFP	800	0.0001	0.3	5	Coupling partner class	29.01	24.2	0.19
Nural Network	RDkitFP	1000	0.0001	0.3	5	No stratification	29.88	24.94	0.12
Nural Network	RDkitFP	1000	0.0001	0.3	5	No stratification	30.21	25.11	0.1
Nural Network	RxnFP	800	0.0001	0.3	5	mechanism	31.23	27.12	0.03
Nural Network	RxnFP	800	0.0001	0.3	5	mechanism	30.75	26.65	0.06

Nural Network	RxnFP	1000	0.0001	0.3	5	Substrate class	31.04	26.84	0.06
Nural Network	RxnFP	1000	0.0001	0.3	5	Substrate class	30.67	26.43	0.08
Nural Network	RxnFP	800	0.0001	0.3	5	Coupling partner class	31.51	27.34	0.04
Nural Network	RxnFP	800	0.0001	0.3	5	Coupling partner class	31.14	26.94	0.07
Nural Network	RxnFP	1000	0.0001	0.3	5	No stratification	31.25	27.06	0.04
Nural Network	RxnFP	1000	0.0001	0.3	5	No stratification	30.94	26.73	0.06

Here epoch, learning rate are different parameters of Neural Network, R^2 is the coefficient of Correlation, RMSE (Root Mean Square Error), MAE (mean absolute error).

Table S2: Regression results of our dataset employing DFT as a featurization method.

model type	featurisation method	epoch	learning rate	test size	iterations	Stratification type	RMSE	MAE	Correlation
Random Forest	DFT			0.2	10	mechanism	24.07	18.66	0.45
Random Forest	DFT			0.2	10	Substrate class	26.12	20.11	0.37
Random Forest	DFT			0.2	10	Coupling partner class	27.82	20.69	0.24
Random Forest	DFT			0.2	10	No stratification	23.54	18.55	0.44
Random Forest HPT	DFT			0.2	10	mechanism	24.72	19.27	0.4
Random Forest HPT	DFT			0.2	10	Substrate class	24.61	19.09	0.4
Random Forest HPT	DFT			0.2	10	Coupling partner class	24.84	19.26	0.4
Random Forest HPT	DFT			0.2	10	No stratification	24.92	19.24	0.39
Nural Network	DFT	800	0.0001	0.2	10	mechanism	30.93	26.46	0.06
Nural Network	DFT	800	0.0001	0.2	10	mechanism	30.6	26.01	0.08
Nural Network	DFT	1000	0.0001	0.2	10	Substrate class	30.3	25.72	0.09
Nural Network	DFT	1000	0.0001	0.2	10	Substrate class	30.03	25.43	0.1
Nural Network	DFT	800	0.0001	0.2	10	Coupling partner class	31.13	26.5	0.06
Nural Network	DFT	800	0.0001	0.2	10	Coupling partner class	31.21	26.44	0.05

Nural Network	DFT	1000	0.0001	0.2	10	No stratification	31.39	26.38	0.03
Nural Network	DFT	1000	0.0001	0.2	10	No stratification	31.16	26.12	0.05
Random Forest	DFT			0.2	5	mechanism	24.05	18.81	0.45
Random Forest	DFT			0.2	5	Substrate class	26.15	20.12	0.37
Random Forest	DFT			0.2	5	Coupling partner class	27.55	20.71	0.25
Random Forest	DFT			0.2	5	No stratification	23.66	18.55	0.43
Random Forest HPT	DFT			0.2	5	mechanism	24.88	19.43	0.39
Random Forest HPT	DFT			0.2	5	Substrate class	24.48	19.08	0.41
Random Forest HPT	DFT			0.2	5	Coupling partner class	25.25	19.52	0.4
Random Forest HPT	DFT			0.2	5	No stratification	24.58	18.98	0.41
Nural Network	DFT	800	0.0001	0.2	5	mechanism	30.6	26.09	0.08
Nural Network	DFT	800	0.0001	0.2	5	mechanism	30.47	25.96	0.09
Nural Network	DFT	1000	0.0001	0.2	5	Substrate class	30.69	26.11	0.08
Nural Network	DFT	1000	0.0001	0.2	5	Substrate class	30.19	25.63	0.11
Nural Network	DFT	800	0.0001	0.2	5	Coupling partner class	31.46	26.74	0.06
Nural Network	DFT	800	0.0001	0.2	5	Coupling partner class	30.91	26.17	0.09
Nural Network	DFT	1000	0.0001	0.2	5	No stratification	31.85	26.87	0
Nural Network	DFT	1000	0.0001	0.2	5	No stratification	31.33	26.46	0.04
Random Forest	DFT			0.3	10	mechanism	24.33	18.59	0.41
Random Forest	DFT			0.3	10	Substrate class	25.34	19.31	0.38
Random Forest	DFT			0.3	10	Coupling partner class	28.16	21.23	0.24
Random Forest	DFT			0.3	10	No stratification	23.69	18.54	0.44
Random Forest HPT	DFT			0.3	10	mechanism	25.32	19.93	0.37
Random Forest HPT	DFT			0.3	10	Substrate class	25.25	19.78	0.38

Random Forest HPT	DFT			0.3	10	Coupling partner class	25.52	19.99	0.36
Random Forest HPT	DFT			0.3	10	No stratification	25.16	19.56	0.38
Nural Network	DFT	800	0.0001	0.3	10	mechanism	31.09	26.49	0.05
Nural Network	DFT	800	0.0001	0.3	10	mechanism	30.97	26.39	0.06
Nural Network	DFT	1000	0.0001	0.3	10	Substrate class	30.65	26	0.08
Nural Network	DFT	1000	0.0001	0.3	10	Substrate class	30.41	25.76	0.1
Nural Network	DFT	800	0.0001	0.3	10	Coupling partner class	31.14	26.42	0.05
Nural Network	DFT	800	0.0001	0.3	10	Coupling partner class	31.13	26.32	0.05
Nural Network	DFT	1000	0.0001	0.3	10	No stratification	31.37	26.38	0.03
Nural Network	DFT	1000	0.0001	0.3	10	No stratification	31	25.98	0.05
Random Forest	DFT			0.3	5	mechanism	23.99	18.48	0.43
Random Forest	DFT			0.3	5	Substrate class	25.34	19.35	0.38
Random Forest	DFT			0.3	5	Coupling partner class	28.13	21.2	0.24
Random Forest	DFT			0.3	5	No stratification	23.99	18.64	0.43
Random Forest HPT	DFT			0.3	5	mechanism	25.25	19.7	0.37
Random Forest HPT	DFT			0.3	5	Substrate class	24.75	19.47	0.4
Random Forest HPT	DFT			0.3	5	Coupling partner class	26.02	20.31	0.35
Random Forest HPT	DFT			0.3	5	No stratification	24.63	19.13	0.4
Nural Network	DFT	800	0.0001	0.3	5	mechanism	31.36	26.78	0.03
Nural Network	DFT	800	0.0001	0.3	5	mechanism	30.81	26.27	0.06
Nural Network	DFT	1000	0.0001	0.3	5	Substrate class	30.64	25.92	0.08
Nural Network	DFT	1000	0.0001	0.3	5	Substrate class	30.32	25.56	0.1
Nural Network	DFT	800	0.0001	0.3	5	Coupling partner class	31.1	26.4	0.07
Nural Network	DFT	800	0.0001	0.3	5	Coupling partner class	31.12	26.31	0.07

Nural Network	DFT	1000	0.0001	0.3	5	No stratification	31.43	26.54	0.03
Nural Network	DFT	1000	0.0001	0.3	5	No stratification	31.47	26.51	0.02

Table S3: Higher correlation results of all subsets made by considering the count of each element.

Dataset	model type	featurisation	test size	iterations	Stratification type	RMSE	MAE	Correlation
Dataset Subset count 1	Random forest	RDkitFP	0.2	10	No stratification	25.55	19.95	0.37
Dataset Subset count 2	Random forest	DRFP	0.2	10	No stratification	27.89	23.00	0.31
Dataset Subset count 3	Random forest	DFT	0.3	10	No stratification	26.41	20.87	0.39
Dataset Subset count 5	Random forest	DFT	0.2	10	No stratification	25.31	20.20	0.40
Dataset Subset count 7	Random forest	DFT	0.2	10	No stratification	27.10	21.45	0.31

Classification results:

Below, you will find the classification results for these specific subsets, along with their respective accuracy and precession values:

Table S4: Classification results of our dataset employing DRFP, RDkitFP, and RXNFP as featurization methods.

Model type	Featurisation method	Test size	Iterations	Stratification type	Accuracy	Precession
KNN classification	DRFP	0.2	10	mechanism	0.45	0.43
KNN classification	DRFP	0.2	10	Substrate class	0.45	0.49
KNN classification	DRFP	0.2	10	Coupling partner class	0.49	0.47
KNN classification	DRFP	0.2	10	No stratification	0.39	0.34
KNN classification	RDkitFP	0.2	10	mechanism	0.4	0.39
KNN classification	RDkitFP	0.2	10	Substrate class	0.43	0.45
KNN classification	RDkitFP	0.2	10	Coupling partner class	0.44	0.41
KNN classification	RDkitFP	0.2	10	No stratification	0.46	0.43
KNN classification	RxnFP	0.2	10	mechanism	0.42	0.39
KNN classification	RxnFP	0.2	10	Substrate class	0.38	0.43
KNN classification	RxnFP	0.2	10	Coupling partner class	0.39	0.38
KNN classification	RxnFP	0.2	10	No stratification	0.38	0.37
KNN classification HPT	DRFP	0.2	10	mechanism	0.48	0.46
KNN classification HPT	DRFP	0.2	10	Substrate class	0.45	0.46
KNN classification HPT	DRFP	0.2	10	Coupling partner class	0.5	0.47

KNN classification HPT	DRFP	0.2	10	No stratification	0.47	0.43
KNN classification HPT	RDkitFP	0.2	10	mechanism	0.41	0.39
KNN classification HPT	RDkitFP	0.2	10	Substrate class	0.46	0.49
KNN classification HPT	RDkitFP	0.2	10	Coupling partner class	0.43	0.4
KNN classification HPT	RDkitFP	0.2	10	No stratification	0.41	0.35
KNN classification HPT	RxnFP	0.2	10	mechanism	0.36	0.35
KNN classification HPT	RxnFP	0.2	10	Substrate class	0.39	0.42
KNN classification HPT	RxnFP	0.2	10	Coupling partner class	0.38	0.34
KNN classification HPT	RxnFP	0.2	10	No stratification	0.41	0.39
KNN classification	DRFP	0.2	5	mechanism	0.48	0.5
KNN classification	DRFP	0.2	5	Substrate class	0.41	0.41
KNN classification	DRFP	0.2	5	Coupling partner class	0.47	0.45
KNN classification	DRFP	0.2	5	No stratification	0.51	0.5
KNN classification	RDkitFP	0.2	5	mechanism	0.43	0.41
KNN classification	RDkitFP	0.2	5	Substrate class	0.4	0.43
KNN classification	RDkitFP	0.2	5	Coupling partner class	0.36	0.38
KNN classification	RDkitFP	0.2	5	No stratification	0.44	0.43
KNN classification	RxnFP	0.2	5	mechanism	0.37	0.35
KNN classification	RxnFP	0.2	5	Substrate class	0.39	0.41
KNN classification	RxnFP	0.2	5	Coupling partner class	0.34	0.34
KNN classification	RxnFP	0.2	5	No stratification	0.4	0.38
KNN classification HPT	DRFP	0.2	5	mechanism	0.49	0.5
KNN classification HPT	DRFP	0.2	5	Substrate class	0.42	0.44
KNN classification HPT	DRFP	0.2	5	Coupling partner class	0.48	0.44
KNN classification HPT	DRFP	0.2	5	No stratification	0.54	0.53
KNN classification HPT	RDkitFP	0.2	5	mechanism	0.47	0.44
KNN classification HPT	RDkitFP	0.2	5	Substrate class	0.34	0.36
KNN classification HPT	RDkitFP	0.2	5	Coupling partner class	0.42	0.38
KNN classification HPT	RDkitFP	0.2	5	No stratification	0.44	0.42
KNN classification HPT	RxnFP	0.2	5	mechanism	0.42	0.42

KNN classification HPT	RxnFP	0.2	5	Substrate class	0.44	0.44
KNN classification HPT	RxnFP	0.2	5	Coupling partner class	0.4	0.37
KNN classification HPT	RxnFP	0.2	5	No stratification	0.41	0.39
KNN classification	DRFP	0.3	10	mechanism	0.44	0.44
KNN classification	DRFP	0.3	10	Substrate class	0.44	0.45
KNN classification	DRFP	0.3	10	Coupling partner class	0.46	0.45
KNN classification	DRFP	0.3	10	No stratification	0.43	0.4
KNN classification	RDkitFP	0.3	10	mechanism	0.42	0.41
KNN classification	RDkitFP	0.3	10	Substrate class	0.4	0.4
KNN classification	RDkitFP	0.3	10	Coupling partner class	0.41	0.4
KNN classification	RDkitFP	0.3	10	No stratification	0.43	0.41
KNN classification	RxnFP	0.3	10	mechanism	0.39	0.39
KNN classification	RxnFP	0.3	10	Substrate class	0.36	0.34
KNN classification	RxnFP	0.3	10	Coupling partner class	0.38	0.37
KNN classification	RxnFP	0.3	10	No stratification	0.39	0.35
KNN classification HPT	DRFP	0.3	10	mechanism	0.47	0.47
KNN classification HPT	DRFP	0.3	10	Substrate class	0.46	0.46
KNN classification HPT	DRFP	0.3	10	Coupling partner class	0.45	0.45
KNN classification HPT	DRFP	0.3	10	No stratification	0.46	0.42
KNN classification HPT	RDkitFP	0.3	10	mechanism	0.41	0.41
KNN classification HPT	RDkitFP	0.3	10	Substrate class	0.44	0.44
KNN classification HPT	RDkitFP	0.3	10	Coupling partner class	0.43	0.41
KNN classification HPT	RDkitFP	0.3	10	No stratification	0.44	0.43
KNN classification HPT	RxnFP	0.3	10	mechanism	0.4	0.39
KNN classification HPT	RxnFP	0.3	10	Substrate class	0.4	0.38
KNN classification HPT	RxnFP	0.3	10	Coupling partner class	0.38	0.38
KNN classification HPT	RxnFP	0.3	10	No stratification	0.36	0.36
KNN classification	DRFP	0.3	5	mechanism	0.44	0.44
KNN classification	DRFP	0.3	5	Substrate class	0.46	0.47
KNN classification	DRFP	0.3	5	Coupling partner class	0.48	0.44
KNN classification	DRFP	0.3	5	No stratification	0.48	0.48
KNN classification	RDkitFP	0.3	5	mechanism	0.43	0.41

KNN classification	RDkitFP	0.3	5	Substrate class	0.39	0.38
KNN classification	RDkitFP	0.3	5	Coupling partner class	0.4	0.38
KNN classification	RDkitFP	0.3	5	No stratification	0.39	0.39
KNN classification	RxnFP	0.3	5	mechanism	0.38	0.37
KNN classification	RxnFP	0.3	5	Substrate class	0.38	0.36
KNN classification	RxnFP	0.3	5	Coupling partner class	0.34	0.34
KNN classification	RxnFP	0.3	5	No stratification	0.38	0.36
KNN classification HPT	DRFP	0.3	5	mechanism	0.48	0.51
KNN classification HPT	DRFP	0.3	5	Substrate class	0.47	0.48
KNN classification HPT	DRFP	0.3	5	Coupling partner class	0.5	0.49
KNN classification HPT	DRFP	0.3	5	No stratification	0.47	0.48
KNN classification HPT	RDkitFP	0.3	5	mechanism	0.39	0.37
KNN classification HPT	RDkitFP	0.3	5	Substrate class	0.38	0.37
KNN classification HPT	RDkitFP	0.3	5	Coupling partner class	0.49	0.46
KNN classification HPT	RDkitFP	0.3	5	No stratification	0.39	0.38
KNN classification HPT	RxnFP	0.3	5	mechanism	0.4	0.39
KNN classification HPT	RxnFP	0.3	5	Substrate class	0.41	0.4
KNN classification HPT	RxnFP	0.3	5	Coupling partner class	0.41	0.38
KNN classification HPT	RxnFP	0.3	5	No stratification	0.4	0.38

HPT stands for hyperparameters tuning using GridSearchCV.

Table S5: Classification results of our dataset employing DFT as a featurization method.

Model type	Featurisation method	Test size	iterations	Stratification type	Accuracy	Precession
KNN classification	DFT	0.2	10	mechanism	0.42	0.41
KNN classification	DFT	0.2	10	Substrate class	0.36	0.36
KNN classification	DFT	0.2	10	Coupling partner class	0.41	0.4
KNN classification	DFT	0.2	10	No stratification	0.4	0.36
KNN classification HPT	DFT	0.2	10	mechanism	0.47	0.45
KNN classification HPT	DFT	0.2	10	Substrate class	0.39	0.4
KNN classification HPT	DFT	0.2	10	Coupling partner class	0.42	0.42
KNN classification HPT	DFT	0.2	10	No stratification	0.43	0.39

KNN classification	DFT	0.2	5	mechanism	0.43	0.41
KNN classification	DFT	0.2	5	Substrate class	0.37	0.4
KNN classification	DFT	0.2	5	Coupling partner class	0.44	0.45
KNN classification	DFT	0.2	5	No stratification	0.43	0.43
KNN classification HPT	DFT	0.2	5	mechanism	0.43	0.41
KNN classification HPT	DFT	0.2	5	Substrate class	0.41	0.41
KNN classification HPT	DFT	0.2	5	Coupling partner class	0.48	0.43
KNN classification HPT	DFT	0.2	5	No stratification	0.45	0.43
KNN classification	DFT	0.3	10	mechanism	0.38	0.37
KNN classification	DFT	0.3	10	Substrate class	0.41	0.41
KNN classification	DFT	0.3	10	Coupling partner class	0.41	0.4
KNN classification	DFT	0.3	10	No stratification	0.41	0.38
KNN classification HPT	DFT	0.3	10	mechanism	0.42	0.39
KNN classification HPT	DFT	0.3	10	Substrate class	0.42	0.43
KNN classification HPT	DFT	0.3	10	Coupling partner class	0.44	0.42
KNN classification HPT	DFT	0.3	10	No stratification	0.39	0.38
KNN classification	DFT	0.3	5	mechanism	0.4	0.4
KNN classification	DFT	0.3	5	Substrate class	0.37	0.39
KNN classification	DFT	0.3	5	Coupling partner class	0.46	0.44
KNN classification	DFT	0.3	5	No stratification	0.38	0.38
KNN classification HPT	DFT	0.3	5	mechanism	0.46	0.46
KNN classification HPT	DFT	0.3	5	Substrate class	0.44	0.43
KNN classification HPT	DFT	0.3	5	Coupling partner class	0.45	0.4
KNN classification HPT	DFT	0.3	5	No stratification	0.42	0.4

HPT stands for hyperparameters tuning using GridSearchCV.

Table S6: Higher classification results of all subsets made by considering count of each element

Dataset	Model type	Featurisation	Test size	Iterations	Stratification type	Accuracy	Precession
Dataset subset count 1	KNN classification	RDkitFP	0.3	10	No stratification	0.43	0.42
	KNN	DRFP	0.3	10	No	0.44	0.45

Dataset subset count 2	Classification HPT				stratification		
	KNN Classification HPT	DFT	0.3	5	No stratification	0.44	0.43
Dataset subset count 3	KNN classification	DRFP	0.2	10	No stratification	0.46	0.46
Dataset subset count 5	KNN Classification HPT	RDkitFP	0.3	5	No stratification	0.46	0.44
	KNN Classification HPT	DFT	0.2	5	No stratification	0.46	0.45
Dataset subset count 7	KNN classification	RDkitFP	0.2	10	No stratification	0.49	0.51

(In the present study, the stratification criteria mechanism refers to the category of cross-coupling reactions (Suzuki coupling, Sonogashira coupling etc.) considered to collect the literature-mined data)