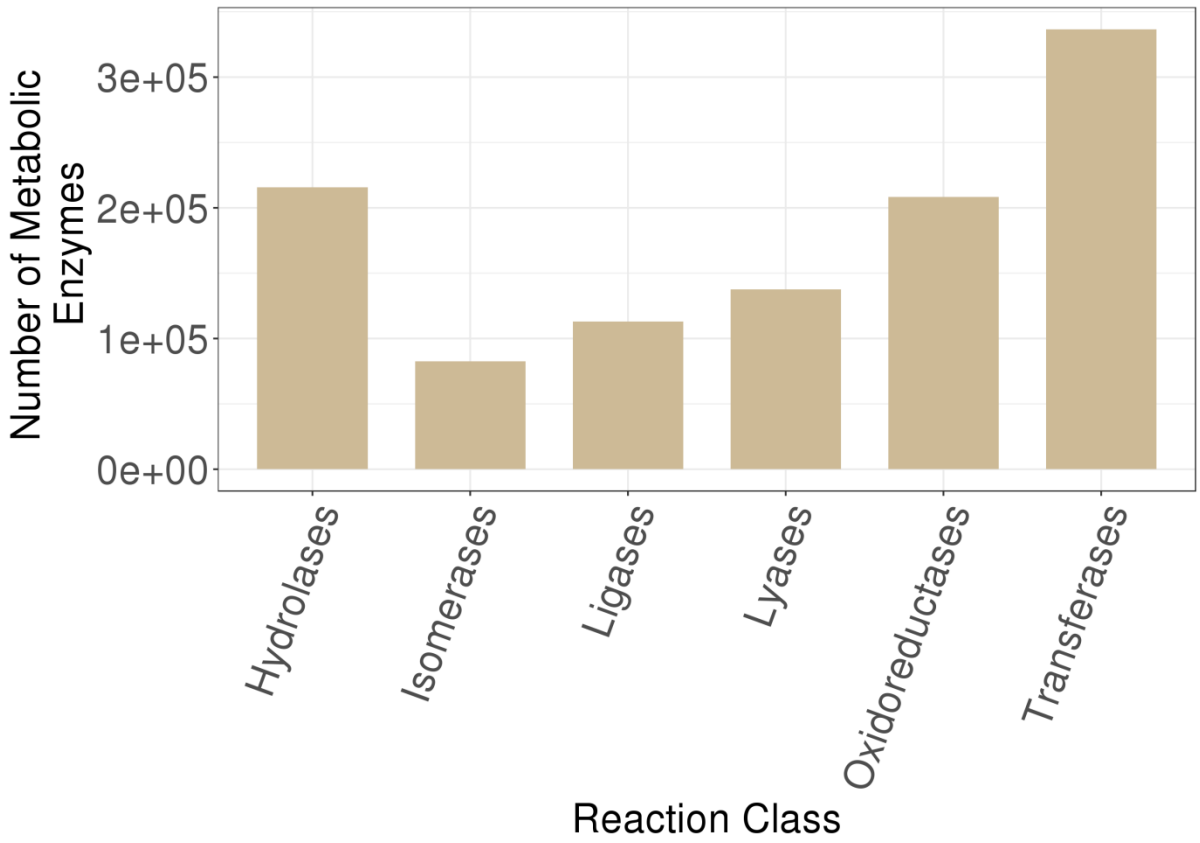
iScience, Volume *24*

Supplemental Information

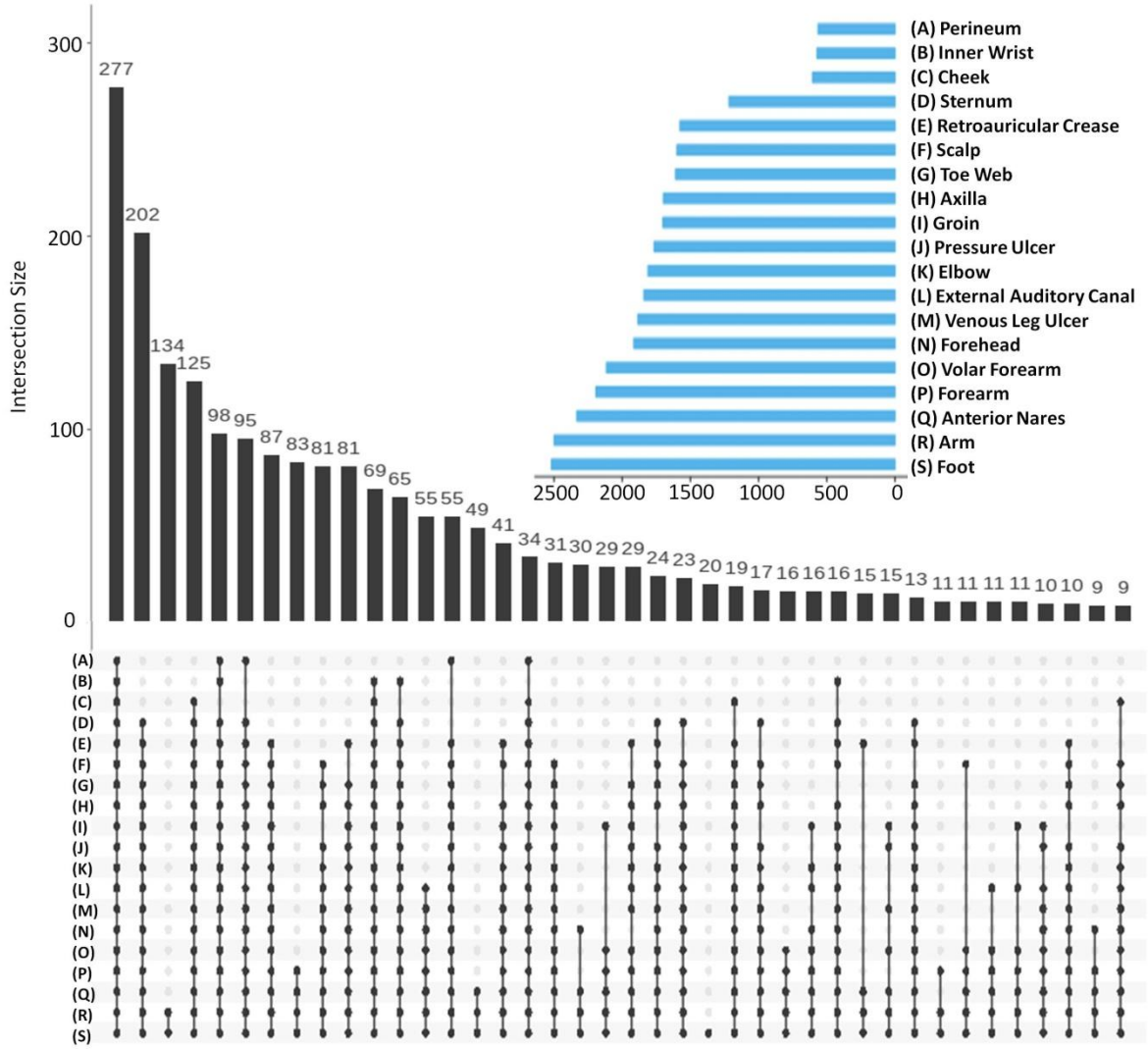
SkinBug: an artificial intelligence approach to predict human skin microbiome-mediated metabolism of biotics and xenobiotics

Shubham K. Jaiswal, Shitij Manojkumar Agarwal, Parikshit Thodum, and Vineet K. Sharma

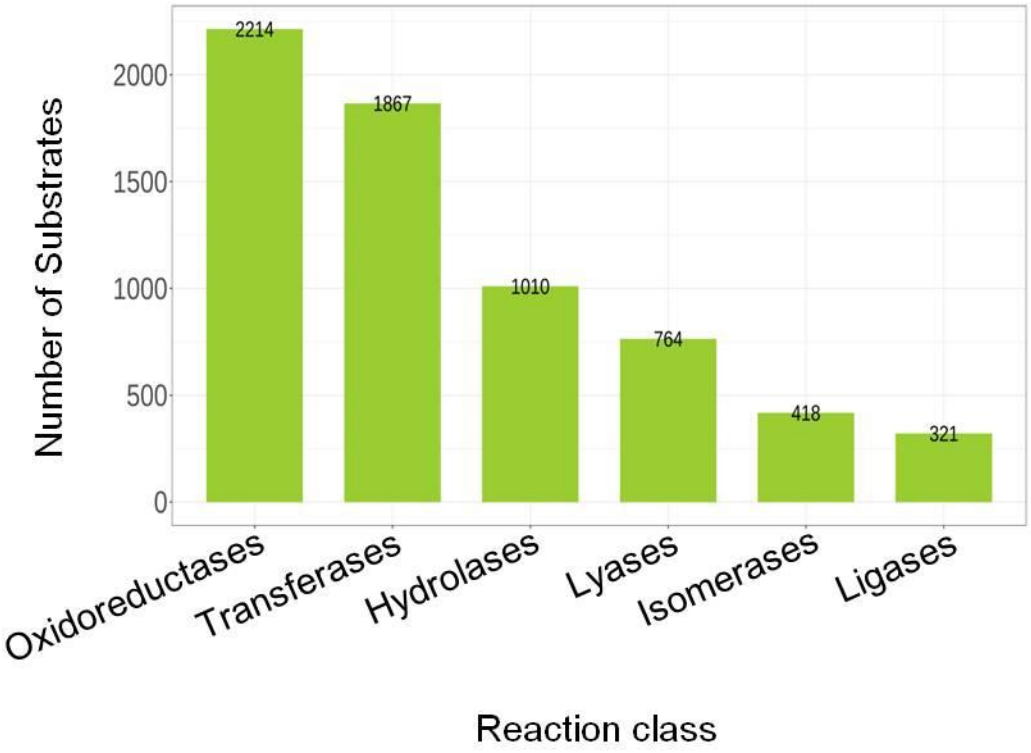
SUPPLEMENTARY FIGURES



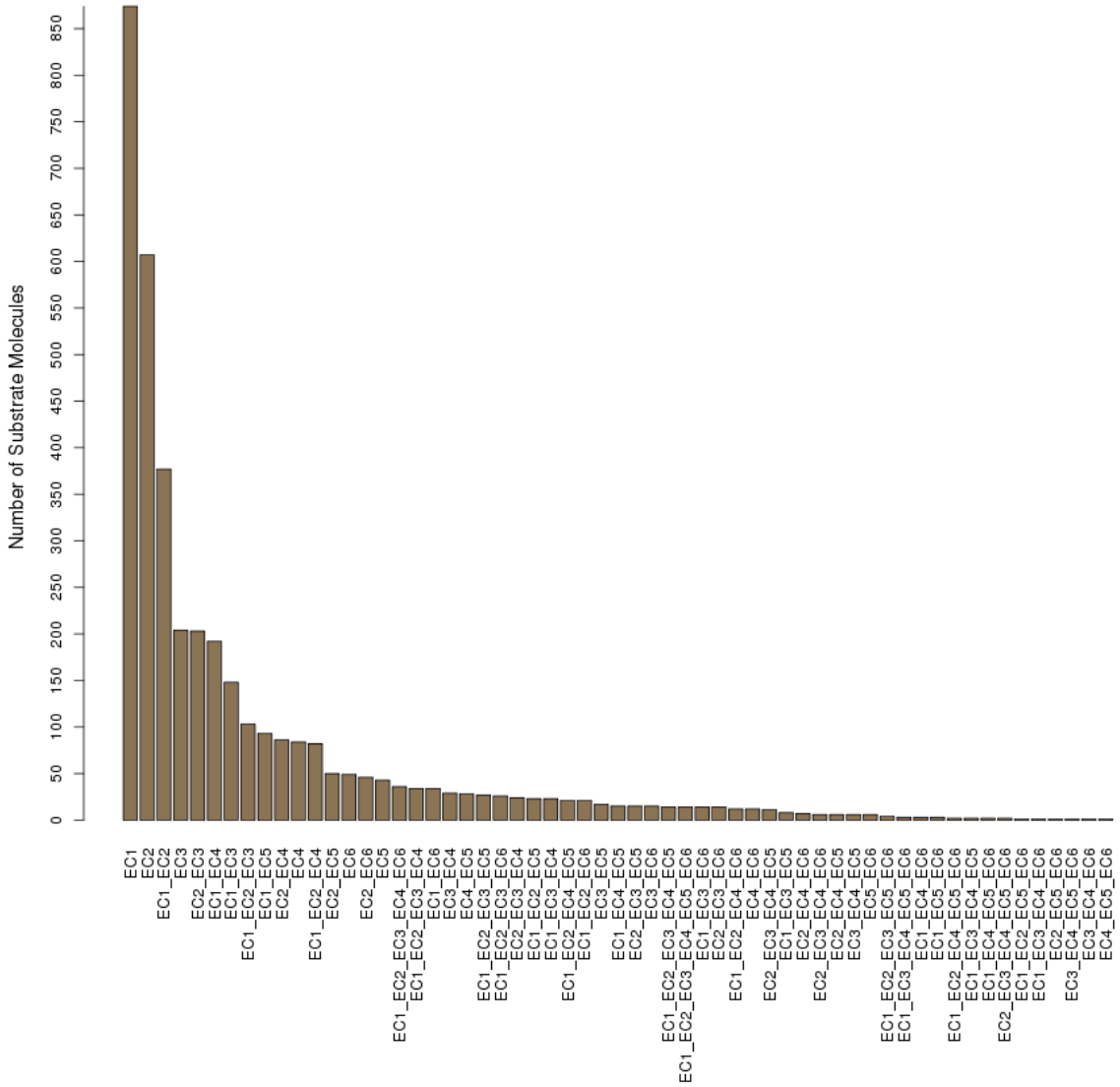
Supplementary Figure S1: Bar plot of number of metabolic enzymes from different reaction classes present in the human skin microbiome (Related to Figure 1)



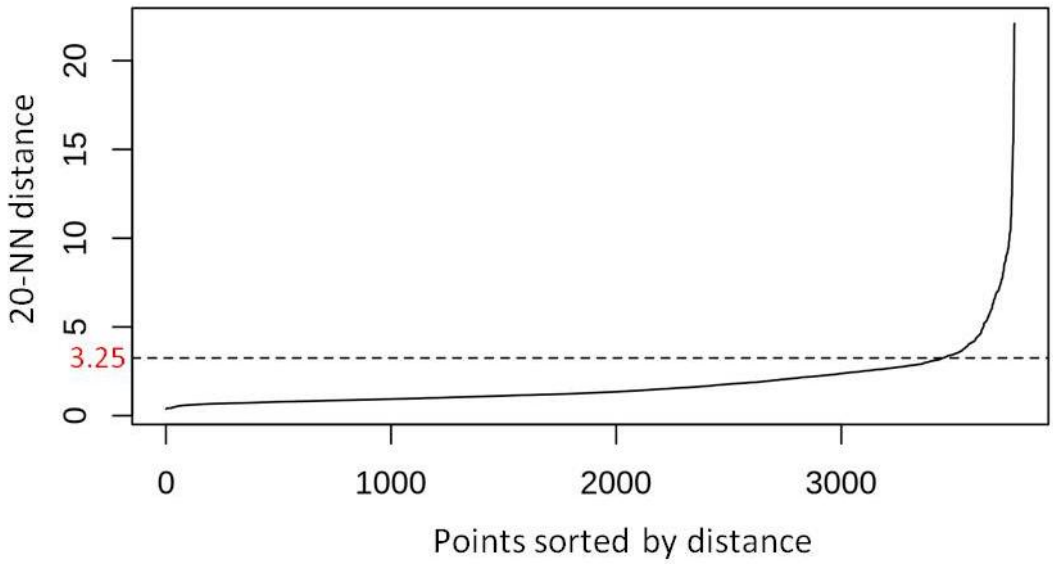
Supplementary Figure S2: Matrix layout for the intersection of unique enzymatic reactions from 19 different skin sites. The blue horizontal bar depicts the absolute number of unique enzymatic reactions that can occur on the each skin site. The vertical bar plot represents the number of unique enzymatic reactions (top of bars) shared by the different skin sites. The sites that share a particular number are shown as the intersection of filled ellipsoids at x-axis. The ellipsoids are placeholders for individual skin site. (Related to Figure 1)



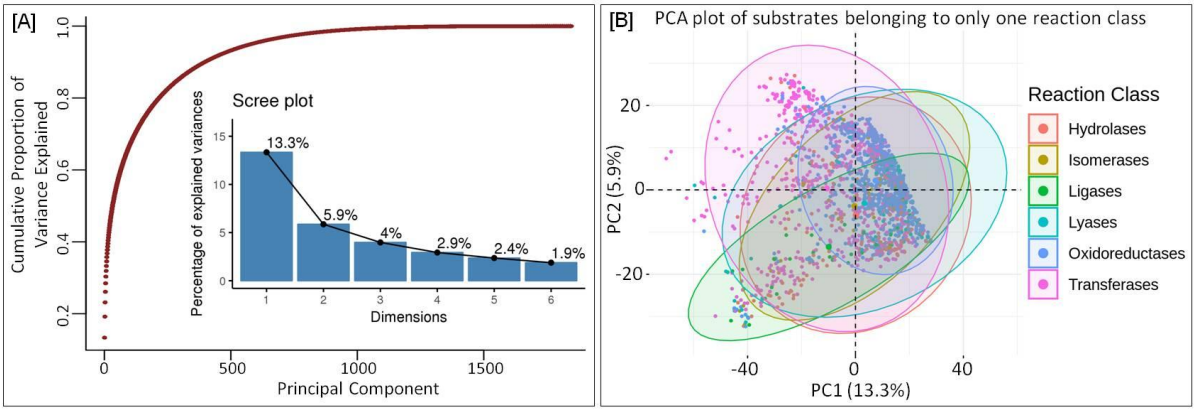
Supplementary Figure S3: Distribution of substrate molecules across different reactions classes (annotated as one-digit EC number). Here all inclusive approach has been used where if one molecule can undergo reactions from different reaction claseses the count is incremented in all of them. (Related to Figure 1)



Supplementary Figure S4: Distribution of 3,769 substrate molecules across different reactions classes (annotated as one-digit EC number) unique combinations. (Related to Figure 1)



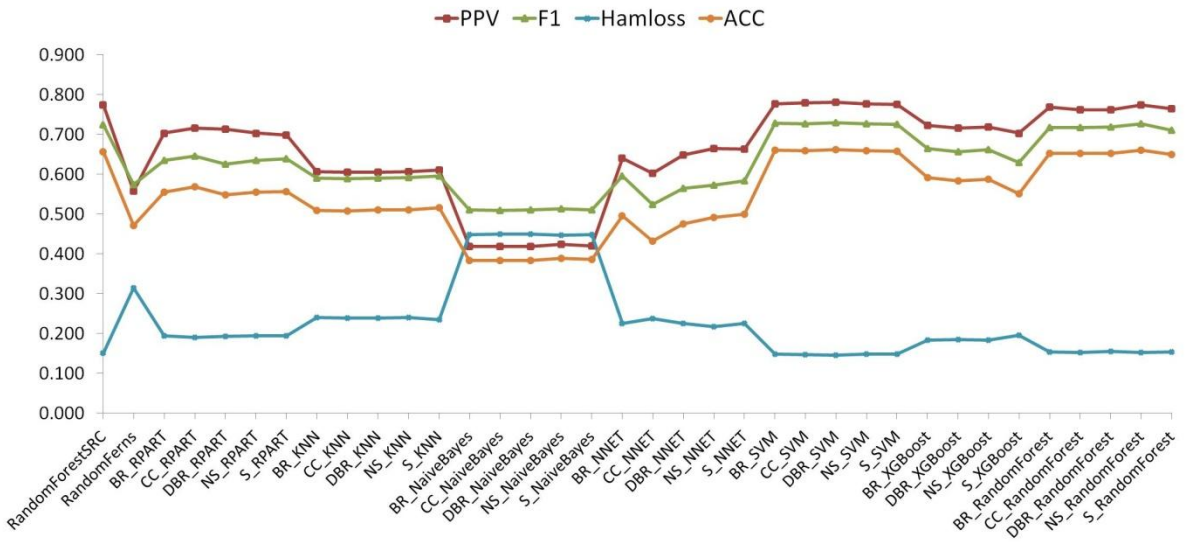
Supplementary Figure S5: The distribution of 20-nearest neighbours distances across the dataset of 3,769 substrate molecules based on the selected 2,322 variables. The knee point at 3.25 is the most suitable epsilon value for the density based clustering using DBSCAN. (Related to Figure 2)



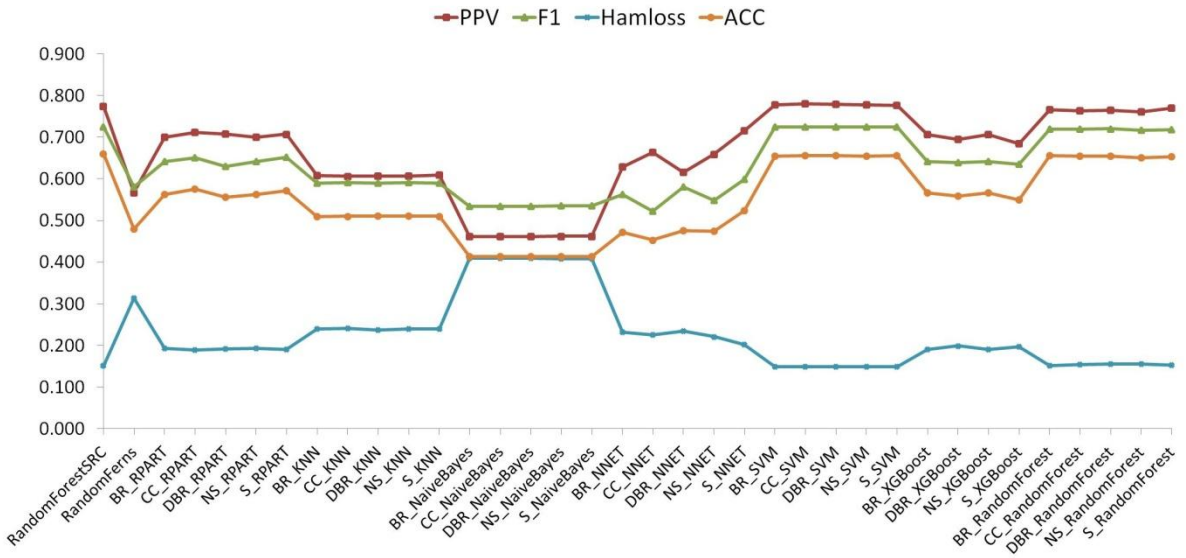
Supplementary Figure S6: Evaluating the diversity and complexity of a subset dataset where one substrate can undergo only one type of reaction (Related to Figure 2)

[A] The cumulative scree plot and a normal scree plot from the PCA analysis of all the substrate molecules from the dataset of substrates that can undergo only one type of reaction class. All the selected 2,322 features were used for performing the PCA analysis. The x-axis is the principal component number, y-axis for the dot plot is the cumulative variance explained by the individual principal components, and the y-axis for the bar plot is the percentage of variance explained by the individual principal components.

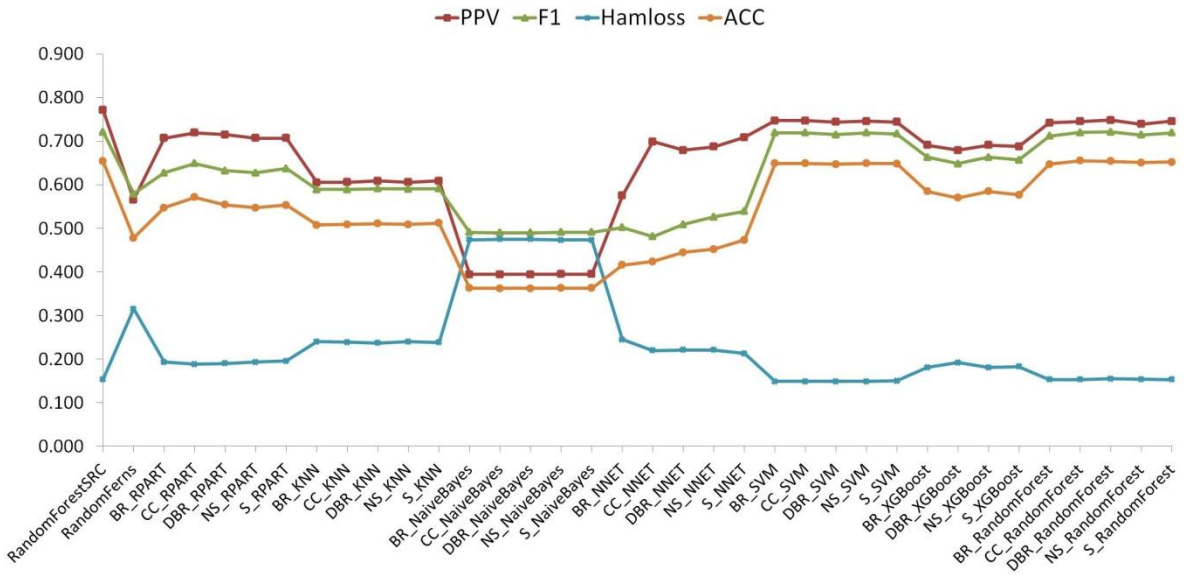
[B] The PCA plot of the substrate molecules that can undergo only one type of reaction class using the principal component PC-1 and PC-2 from the PCA analysis. Different reaction classes are coloured differently and the ellipsoids are drawn for each reaction class based on the distribution of substrate molecules in the plot.



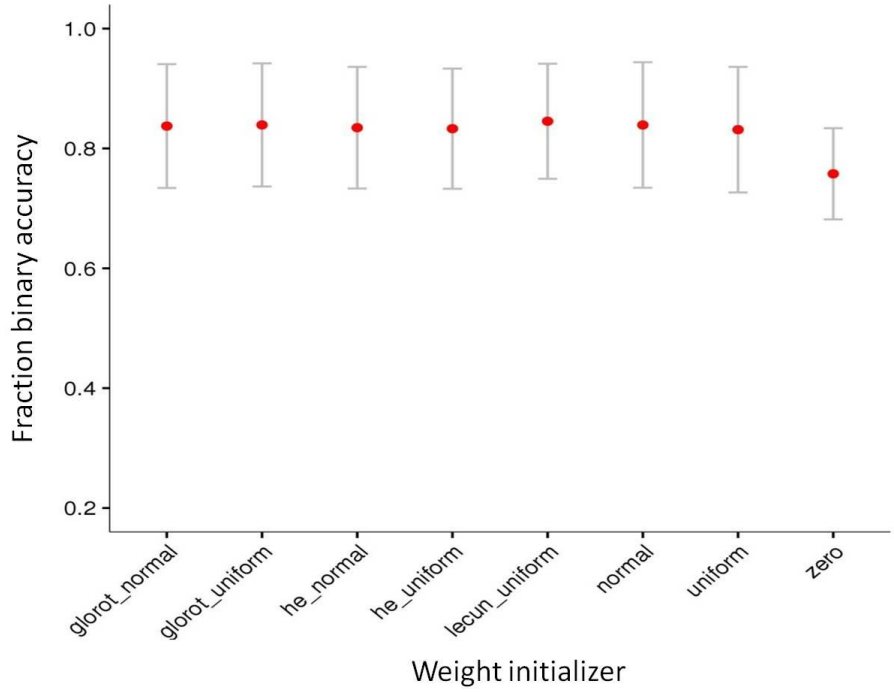
Supplementary Figure S7: The performance of different multiclass multilabel classification algorithms on the multilabel dataset with ECFP fingerprints, FCFP fingerprints, boruta selected descriptors, and boruta selected fingerprints. ACC – Multilabel accuracy, PPV – Multilabel precision or Multilabel positive predicted value, Hamloss – Hamming loss, F1 – Multilabel F1 score. Binary relevance – BR, Classifier chains – CC, Nested stacking – NS, Dependent binary relevance – DBR, Stacking – S. (Related to Figure 4)



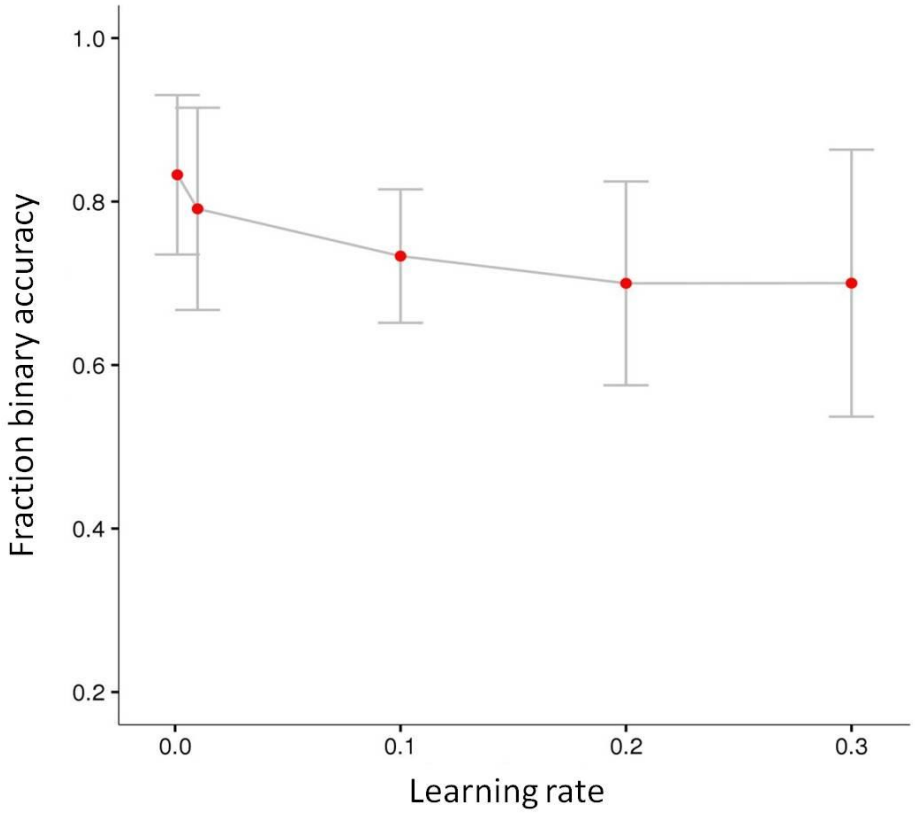
Supplementary Figure S8: The performance of different multiclass multilabel classification algorithms on the multilabel dataset with ECFP fingerprints, boruta selected descriptors, and boruta selected fingerprints. ACC – Multilabel accuracy, PPV – Multilabel precision or Multilabel positive predicted value, Hamloss – Hamming loss, F1 – Multilabel F1 score. Binary relevance – BR, Classifier chains – CC, Nested stacking – NS, Dependent binary relevance – DBR, Stacking – S. (Related to Figure 4)



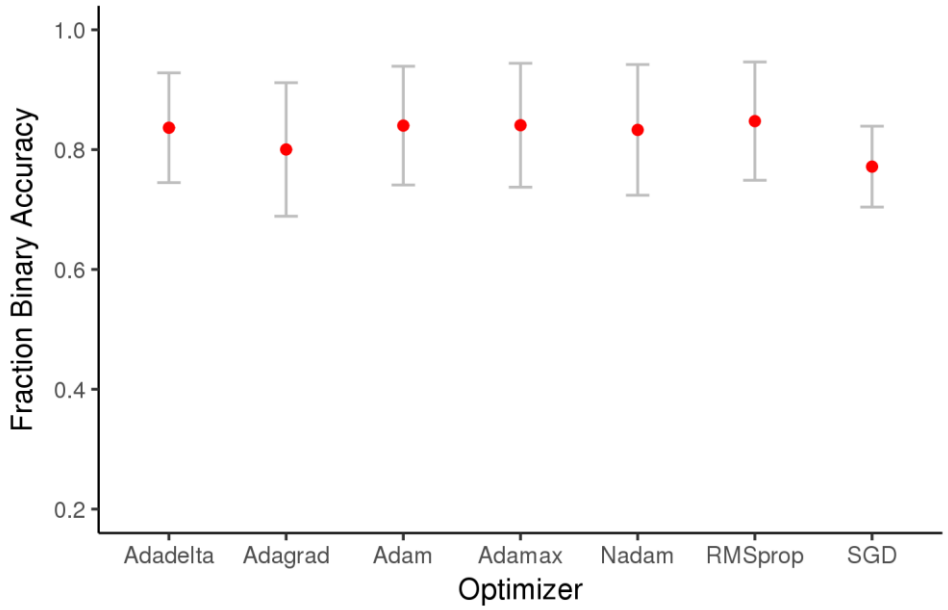
Supplementary Figure S9: The performance of different multiclass multilabel classification algorithms on the multilabel dataset with FCFP fingerprints, boruta selected descriptors, and boruta selected fingerprints. ACC – Multilabel accuracy, PPV – Multilabel precision or Multilabel positive predicted value, Hamloss – Hamming loss, F1 – Multilabel F1 score. Binary relevance – BR, Classifier chains – CC, Nested stacking – NS, Dependent binary relevance – DBR, Stacking – S. (Related to Figure 4)



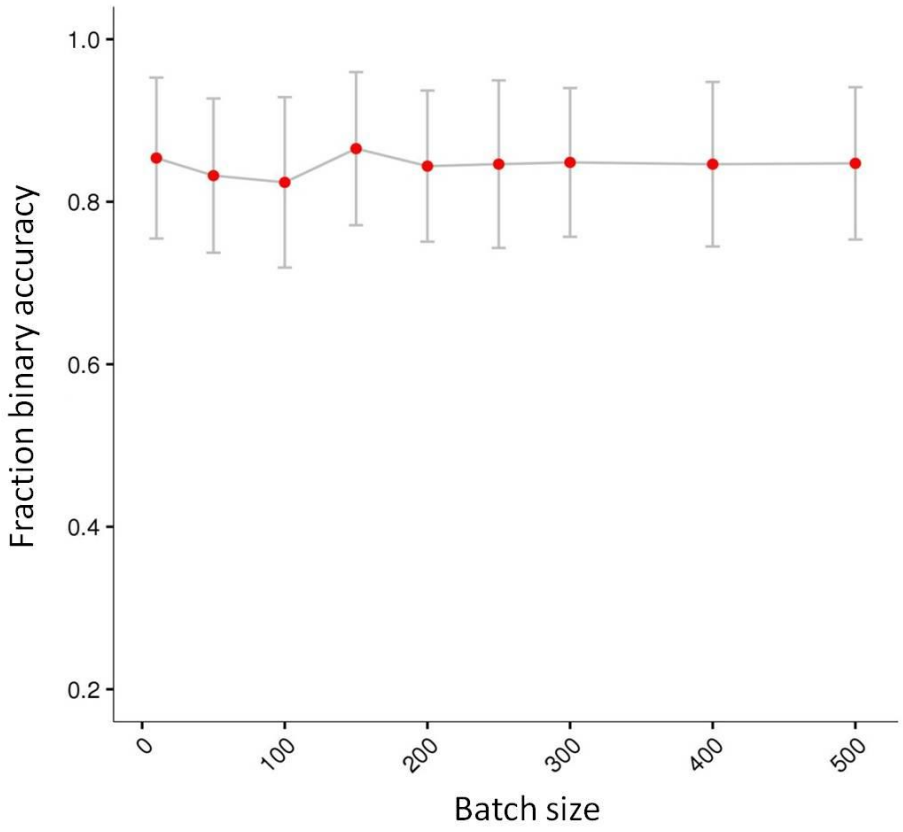
Supplementary Figure S10: Selecting the most optimum weight initializer for ANN models using grid search method using 5-fold cross validation. Different weight initializers are plotted against their respective fraction binary accuracies. The most optimum weight initializer was “lecun_uniform”. (Related to Figure 5)



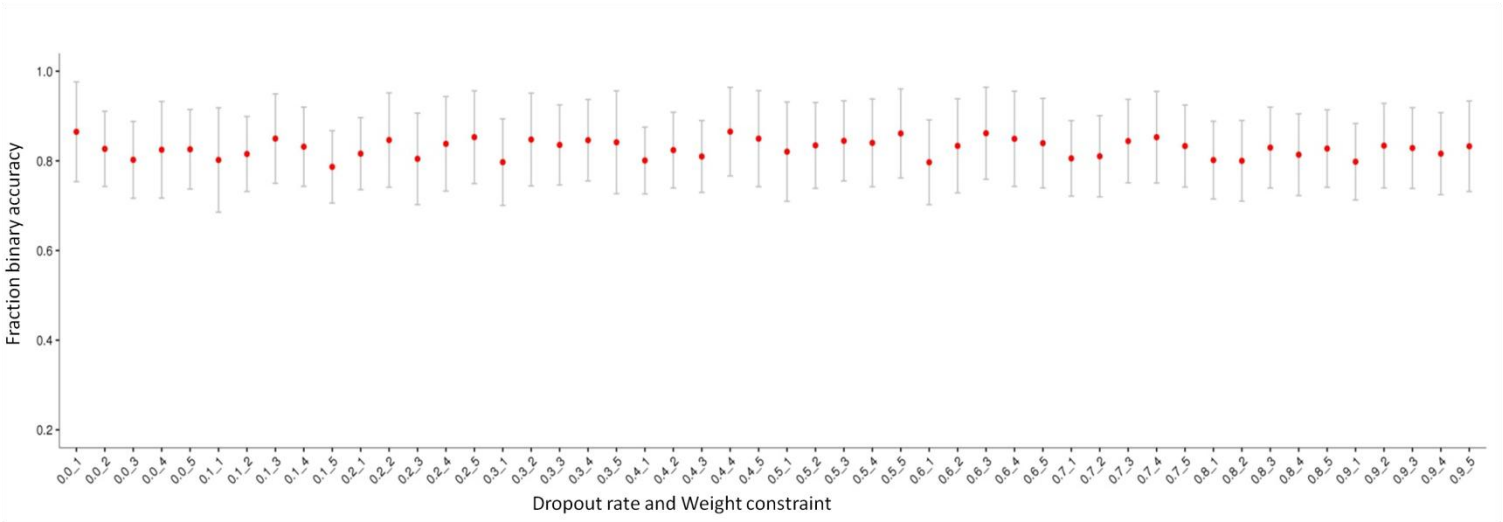
Supplementary Figure S11: Selecting the most optimum value of learning rate for ANN models using grid search method using 5-fold cross validation. Different values of learning rates are plotted against their respective fraction binary accuracies. The most optimum value for the learning rate was "0.001". (Related to Figure 5)



Supplementary Figure S12: Selecting the best performing optimizer for ANN models using grid search method using 5-fold cross validation. Different optimizers are plotted against their respective fraction binary accuracies. The best performing optimizer was “RMSprop”. (Related to Figure 5)



Supplementary Figure S13: Selecting the most optimum value of batch size for ANN models using grid search method using 5-fold cross validation. Different values of batch size are plotted against their respective fraction binary accuracies. The most optimum value for the batch size was “150”. (Related to Figure 5)



Supplementary Figure S14: Selecting the most optimum value of dropout rate and weight constraint for ANN models for performing dropout regularization using grid search method using 5-fold cross validation. Different combinations of values of dropout rate and weight constraint are plotted against their respective fraction binary accuracies. The most optimum value for the dropout rate was “0.4” and most optimum value of weight constraint was “4”. (Related to Figure 5)

SUPPLEMETARY TABLES

Table S1: The binary performance of multiclass multilabel model for predicting the reaction class on 5-fold cross validation testing. (Related to Figure 4)

Reaction Class	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1 Score	FDR	GPR
Oxidoreductases	0.855	0.207	0.111	0.345	0.793	0.568	0.806	0.786	0.834	0.214	0.836
Transferases	0.841	0.235	0.225	0.244	0.765	0.531	0.774	0.757	0.766	0.243	0.766
Hydrolases	0.824	0.210	0.538	0.089	0.790	0.422	0.821	0.656	0.543	0.344	0.551
Lyases	0.844	0.164	0.620	0.047	0.836	0.420	0.857	0.673	0.486	0.327	0.506
Isomerases	0.900	0.082	0.571	0.020	0.918	0.522	0.931	0.734	0.541	0.266	0.561
Ligases	0.886	0.082	0.761	0.017	0.918	0.336	0.931	0.577	0.338	0.423	0.371

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

Table S2: The binary performance of multiclass multi-label model for predicting the reaction class on blind set testing. (Related to Figure 4)

Reaction Class	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1 Score	FDR	GPR
Oxidoreductases	0.871	0.174	0.081	0.309	0.826	0.638	0.855	0.813	0.863	0.188	0.864
Transferases	0.884	0.222	0.217	0.226	0.778	0.557	0.783	0.774	0.778	0.226	0.778
Hydrolases	0.810	0.186	0.487	0.094	0.814	0.450	0.859	0.625	0.563	0.375	0.566
Lyases	0.896	0.120	0.536	0.036	0.880	0.516	0.899	0.722	0.565	0.278	0.579
Isomerases	0.943	0.060	0.417	0.032	0.940	0.551	0.968	0.583	0.583	0.417	0.583
Ligases	0.878	0.030	0.429	0.013	0.970	0.602	0.981	0.667	0.615	0.333	0.617

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

Table S3: The binary performance of multiclass multilabel model for predicting the subclasses of "Oxidoreductases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Acting on the CH-OH group of donors	0.944	0.115	0.237	0.056	0.885	0.733	0.892	0.868	0.812	0.132	0.814
Acting on the aldehyde or oxo group of donors	0.921	0.079	0.611	0.013	0.921	0.520	0.929	0.791	0.521	0.209	0.555
Acting on the CH-CH group of donors	0.914	0.099	0.503	0.022	0.901	0.586	0.911	0.812	0.617	0.188	0.635
Acting on the CH-NH2 group of donors	0.945	0.046	0.686	0.008	0.954	0.449	0.961	0.698	0.433	0.302	0.468
Acting on the CH-NH group of donors	0.936	0.036	0.723	0.004	0.964	0.440	0.967	0.743	0.403	0.257	0.453
Acting on NADH or NADPH	0.899	0.007	0.714	0.000	0.993	0.533	0.993	1.000	0.444	0.000	0.535
Acting on other nitrogenous compounds as donors	0.952	0.017	0.800	0.000	0.983	0.420	0.983	0.900	0.327	0.100	0.424
Acting on a sulfur group of donors	0.920	0.026	0.742	0.002	0.974	0.437	0.977	0.773	0.386	0.227	0.446
Acting on a heme group of donors	0.744	0.001	1.000	0.000	0.999	0.000	0.999	0.000	0.000	NA	NA
Acting on diphenols and related substances as donors	0.881	0.014	1.000	0.000	0.986	0.000	0.986	0.000	0.000	NA	NA
Acting on a peroxide as acceptor	0.927	0.018	0.739	0.002	0.982	0.436	0.984	0.750	0.387	0.250	0.442
Acting on hydrogen as donor	0.877	0.004	0.692	0.000	0.996	0.554	0.996	1.000	0.471	0.000	0.555
Acting on single donors with incorporation of molecular oxygen (oxygenases)	0.890	0.071	0.659	0.008	0.929	0.503	0.933	0.824	0.483	0.176	0.530
Acting on paired donors, with incorporation or reduction of molecular oxygen	0.894	0.197	0.223	0.175	0.803	0.602	0.819	0.784	0.780	0.216	0.780
Oxidizing metal ions	0.949	0.002	0.357	0.000	0.998	0.801	0.998	1.000	0.783	0.000	0.802
Acting on CH or CH2 groups	0.930	0.026	0.653	0.002	0.974	0.530	0.976	0.839	0.491	0.161	0.539
Acting on iron-sulfur proteins as donors	0.997	0.003	0.700	0.000	0.997	0.547	0.997	1.000	0.462	0.000	0.548
Acting on reduced flavodoxin as donor	0.993	0.002	1.000	0.000	0.998	0.000	0.998	0.000	0.000	NA	NA
Acting on phosphorus or arsenic in donors	0.880	0.003	1.000	0.000	0.997	0.000	0.997	0.000	0.000	NA	NA
Catalysing the reaction X-H + Y-H = X-Y	0.916	0.017	0.814	0.001	0.983	0.381	0.983	0.800	0.302	0.200	0.386
Reducing C-O-C group as acceptor	0.813	0.004	1.000	0.000	0.996	0.000	0.996	0.000	0.000	NA	NA
Other oxidoreductases	0.798	0.007	0.737	0.000	0.993	0.466	0.993	0.833	0.400	0.167	0.468

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary

negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S4: The binary performance of the multiclass multi-label model for predicting the subclasses of "Transferases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Transferring one-carbon groups	0.896	0.150	0.360	0.069	0.850	0.610	0.870	0.781	0.704	0.219	0.707
Transferring aldehyde or ketonic groups	0.939	0.017	1.000	0.001	0.983	-0.004	0.984	0.000	0.000	1.000	0.000
Acyltransferases	0.853	0.169	0.502	0.057	0.831	0.512	0.848	0.747	0.597	0.253	0.610
Glycosyltransferases	0.890	0.144	0.384	0.059	0.856	0.608	0.873	0.790	0.692	0.210	0.698
Transferring alkyl or aryl groups, other than methyl groups	0.855	0.090	0.754	0.012	0.910	0.383	0.918	0.708	0.365	0.292	0.417
Transferring nitrogenous groups	0.956	0.062	0.465	0.011	0.938	0.650	0.944	0.863	0.660	0.137	0.679
Transferring phosphorus-containing groups	0.943	0.106	0.278	0.053	0.894	0.697	0.917	0.809	0.763	0.191	0.764
Transferring sulfur-containing groups	0.891	0.073	0.745	0.014	0.927	0.366	0.937	0.617	0.361	0.383	0.397
Transferring selenium-containing groups	0.412	0.001	1.000	0.000	0.999	0.000	0.999	0.000	0.000	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S5: The binary performance of the multiclass multi-label model for predicting the subclasses of "Hydrolases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Acting on ester bonds	0.878	0.204	0.237	0.177	0.796	0.586	0.814	0.774	0.768	0.226	0.768
Acting on sulfur-nitrogen bonds	0.336	0.002	1.000	0.000	0.998	0.000	0.998	0.000	0.000	NA	NA
Acting on carbon-phosphorus bonds	0.844	0.005	1.000	0.000	0.995	0.000	0.995	0.000	0.000	NA	NA
Acting on sulfur-sulfur bonds	0.426	0.001	1.000	0.000	0.999	0.000	0.999	0.000	0.000	NA	NA
Acting on carbon-sulfur bonds	0.766	0.009	1.000	0.000	0.991	0.000	0.991	0.000	0.000	NA	NA
Glycosylases	0.890	0.092	0.452	0.023	0.908	0.622	0.919	0.817	0.656	0.183	0.669
Acting on ether bonds	0.944	0.032	0.475	0.002	0.968	0.690	0.969	0.941	0.674	0.059	0.703
Acting on peptide bonds (peptidases)	0.906	0.032	1.000	0.000	0.968	0.000	0.968	1.000	0.000	NA	NA
Acting on carbon-nitrogen bonds, other than peptide bonds	0.903	0.156	0.283	0.094	0.844	0.638	0.870	0.784	0.749	0.216	0.750
Acting on acid anhydrides	0.952	0.058	0.383	0.023	0.942	0.646	0.960	0.744	0.674	0.256	0.677
Acting on carbon-carbon bonds	0.928	0.060	0.907	0.010	0.940	0.159	0.949	0.357	0.147	0.643	0.182
Acting on halide bonds	0.990	0.021	0.606	0.000	0.979	0.621	0.979	1.000	0.565	0.000	0.628
Acting on phosphorus-nitrogen bonds	0.355	0.003	1.000	0.000	0.997	0.000	0.997	0.000	0.000	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S6: The binary performance of the multiclass multi-label model for predicting the subclasses of "Lyases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Carbon-carbon lyases	0.851	0.224	0.259	0.201	0.776	0.537	0.823	0.710	0.725	0.290	0.726
Carbon-oxygen lyases	0.839	0.223	0.231	0.214	0.777	0.554	0.742	0.810	0.789	0.190	0.789
Carbon-nitrogen lyases	0.781	0.109	0.961	0.011	0.891	0.076	0.899	0.300	0.070	0.700	0.109
Carbon-sulfur lyases	0.922	0.053	0.691	0.001	0.947	0.524	0.947	0.944	0.466	0.056	0.540
Carbon-halide lyases	0.834	0.011	0.583	0.001	0.989	0.585	0.990	0.833	0.556	0.167	0.589
Phosphorus-oxygen lyases	0.968	0.023	0.941	0.001	0.977	0.166	0.978	0.500	0.105	0.500	0.171
carbon-phosphorus lyases	0.760	0.004	1.000	0.000	0.996	0.000	0.996	0.000	0.000	NA	NA
Other lyases	0.953	0.026	0.833	0.006	0.974	0.256	0.979	0.429	0.240	0.571	0.267

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S7: The binary performance of the multiclass multi-label model for predicting the subclasses of "Isomerases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Racemases and epimerases	0.946	0.108	0.133	0.095	0.892	0.765	0.926	0.832	0.849	0.168	0.849
cis-trans-Isomerases	0.822	0.027	0.529	0.005	0.973	0.601	0.977	0.800	0.593	0.200	0.614
Intramolecular oxidoreductases	0.924	0.167	0.356	0.077	0.833	0.605	0.843	0.802	0.714	0.198	0.719
Intramolecular transferases	0.827	0.165	0.646	0.049	0.835	0.389	0.859	0.636	0.455	0.364	0.475
Intramolecular lyases	0.932	0.059	0.292	0.015	0.941	0.767	0.946	0.902	0.793	0.098	0.799
Other isomerases	0.277	0.005	1.000	0.000	0.995	0.000	0.995	0.000	0.000	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S8: The binary performance of the multiclass multi-label model for predicting the subclasses of "Ligases" class on 5-fold cross validation testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Forming carbon-oxygen bonds	0.902	0.086	0.463	0.029	0.914	0.581	0.933	0.733	0.620	0.267	0.627
Forming carbon-sulfur bonds	0.939	0.134	0.167	0.112	0.866	0.722	0.888	0.833	0.833	0.167	0.833
Forming carbon-nitrogen bonds	0.933	0.124	0.099	0.151	0.876	0.752	0.890	0.864	0.882	0.136	0.882
Forming carbon-carbon bonds	0.892	0.067	0.864	0.007	0.933	0.264	0.939	0.600	0.222	0.400	0.286
Forming phosphoric-ester bonds	0.981	0.019	0.500	0.007	0.981	0.568	0.987	0.667	0.571	0.333	0.577
Forming nitrogen-D-metal bonds	0.969	0.019	1.000	0.000	0.981	0.000	0.981	0.000	0.000	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S9: The binary performance of the multiclass multi-label model for predicting the subclasses of "Oxidoreductases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Acting on the CH-OH group of donors	0.973	0.099	0.158	0.070	0.901	0.776	0.922	0.857	0.850	0.143	0.850
Acting on the aldehyde or oxo group of donors	0.975	0.047	0.462	0.013	0.953	0.624	0.963	0.778	0.636	0.222	0.647
Acting on the CH-CH group of donors	0.932	0.058	0.320	0.014	0.942	0.749	0.948	0.895	0.773	0.105	0.780
Acting on the CH-NH2 group of donors	0.964	0.012	0.500	0.000	0.988	0.703	0.988	1.000	0.667	0.000	0.707
Acting on the CH-NH group of donors	0.996	0.012	0.333	0.006	0.988	0.661	0.994	0.667	0.667	0.333	0.667
Acting on NADH or NADPH	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on other nitrogenous compounds as donors	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
Acting on a sulfur group of donors	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
Acting on a heme group of donors	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on diphenols and related substances as donors	0.994	0.006	1.000	0.000	0.994	0.000	0.994	0.000	0.000	NA	NA
Acting on a peroxide as acceptor	0.994	0.006	1.000	0.000	0.994	0.000	0.994	0.000	0.000	NA	NA
Acting on hydrogen as donor	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on single donors with incorporation of molecular oxygen (oxygenases)	0.854	0.052	0.583	0.013	0.948	0.521	0.958	0.714	0.526	0.286	0.546
Acting on paired donors, with incorporation or reduction of molecular oxygen	0.941	0.122	0.125	0.120	0.878	0.755	0.890	0.864	0.870	0.136	0.870
Oxidizing metal ions	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on CH or CH2 groups	0.994	0.006	0.500	0.000	0.994	0.705	0.994	1.000	0.667	0.000	0.707
Acting on iron-sulfur proteins as donors	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on reduced flavodoxin as donor	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on phosphorus or arsenic in donors	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Catalysing the reaction X-H + Y-H = X-Y	1.000	0.006	1.000	0.000	0.994	0.000	0.994	1.000	0.000	NA	NA
Reducing C-O-C group as acceptor	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Other oxidoreductases	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary

negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S10: The binary performance of the multiclass multi-label model for predicting the subclasses of "Transferases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Transferring one-carbon groups	0.924	0.133	0.302	0.070	0.867	0.654	0.892	0.789	0.741	0.211	0.742
Transferring aldehyde or ketonic groups	NA	0.006	NA	0.006	0.994	0.000	1.000	0.000	0.000	1.000	NA
Acyltransferases	0.862	0.165	0.541	0.050	0.835	0.492	0.852	0.739	0.567	0.261	0.583
Glycosyltransferases	0.961	0.089	0.220	0.043	0.911	0.764	0.926	0.865	0.821	0.135	0.822
Transferring alkyl or aryl groups, other than methyl groups	0.945	0.057	0.538	0.014	0.943	0.561	0.953	0.750	0.571	0.250	0.588
Transferring nitrogenous groups	0.968	0.051	0.500	0.007	0.949	0.639	0.953	0.875	0.636	0.125	0.661
Transferring phosphorus-containing groups	0.937	0.089	0.286	0.033	0.911	0.731	0.922	0.862	0.781	0.138	0.785
Transferring sulfur-containing groups	0.967	0.063	0.667	0.027	0.937	0.345	0.960	0.429	0.375	0.571	0.378
Transferring selenium-containing groups	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S11: The binary performance of the multiclass multi-label model for predicting the subclasses of "Hydrolases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Acting on ester bonds	0.895	0.202	0.216	0.191	0.798	0.591	0.826	0.763	0.773	0.237	0.773
Acting on sulfur-nitrogen bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on carbon-phosphorus bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on sulfur-sulfur bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Acting on carbon-sulfur bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Glycosylases	0.968	0.083	0.250	0.056	0.917	0.672	0.958	0.692	0.720	0.308	0.721
Acting on ether bonds	0.997	0.024	0.400	0.000	0.976	0.765	0.975	1.000	0.750	0.000	0.775
Acting on peptide bonds (peptidases)	0.964	0.012	1.000	0.000	0.988	0.000	0.988	0.000	0.000	NA	NA
Acting on carbon-nitrogen bonds, other than peptide bonds	0.931	0.119	0.200	0.085	0.881	0.715	0.915	0.800	0.800	0.200	0.800
Acting on acid anhydrides	0.963	0.048	0.143	0.039	0.952	0.731	0.987	0.667	0.750	0.333	0.756
Acting on carbon-carbon bonds	0.959	0.036	1.000	0.000	0.964	0.000	0.964	0.000	0.000	NA	NA
Acting on halide bonds	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
Acting on phosphorus-nitrogen bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S12: The binary performance of the multiclass multi-label model for predicting the subclasses of "Lyases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Carbon-carbon lyases	0.895	0.169	0.192	0.154	0.831	0.650	0.868	0.778	0.792	0.222	0.793
Carbon-oxygen lyases	0.907	0.169	0.194	0.138	0.831	0.664	0.781	0.879	0.841	0.121	0.841
Carbon-nitrogen lyases	0.850	0.077	1.000	0.000	0.923	0.000	0.923	1.000	0.000	NA	NA
Carbon-sulfur lyases	0.968	0.031	0.333	0.016	0.969	0.651	0.984	0.667	0.667	0.333	0.667
Carbon-halide lyases	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Phosphorus-oxygen lyases	1.000	0.015	1.000	0.000	0.985	0.000	0.985	1.000	0.000	NA	NA
carbon-phosphorus lyases	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Other lyases	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S13: The binary performance of the multiclass multi-label model for predicting the subclasses of "Isomerases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Racemases and epimerases	0.973	0.083	0.077	0.087	0.917	0.824	0.955	0.857	0.889	0.143	0.889
cis-trans-Isomerases	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000
Intramolecular oxidoreductases	0.931	0.083	0.273	0.000	0.917	0.806	0.893	1.000	0.842	0.000	0.853
Intramolecular transferases	0.842	0.139	0.714	0.000	0.861	0.494	0.853	1.000	0.444	0.000	0.535
Intramolecular lyases	0.961	0.056	0.200	0.032	0.944	0.768	0.968	0.800	0.800	0.200	0.800
Other isomerases	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

Table S14: The binary performance of the multiclass multi-label model for predicting the subclasses of "Ligases" class on stratified random sampling split testing. (Related to Table 1 and 2)

Reaction subclass	AUC	MMCE	FNR	FPR	ACC	MCC	NPV	PPV	F1	FDR	GPR
Forming carbon-oxygen bonds	0.957	0.040	0.500	0.000	0.960	0.692	0.958	1.000	0.667	0.000	0.707
Forming carbon-sulfur bonds	0.887	0.200	0.200	0.200	0.800	0.592	0.857	0.727	0.762	0.273	0.763
Forming carbon-nitrogen bonds	0.910	0.240	0.231	0.250	0.760	0.519	0.750	0.769	0.769	0.231	0.769
Forming carbon-carbon bonds	0.958	0.080	1.000	0.042	0.920	-0.042	0.958	0.000	0.000	1.000	0.000
Forming phosphoric-ester bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA
Forming nitrogen-D-metal bonds	NA	0.000	NA	0.000	1.000	0.000	1.000	0.000	NA	NA	NA

AUC = Area under the curve, MMCE = Binary mean misclassification error, FNR = Binary false negative rate, FPR = Binary false positive rate, ACC = Binary accuracy, MCC = Matthews correlation coefficient, NPV = Binary negative predictive value, PPV = Binary positive predicted value, F1 = Binary F1 score, FDR = Binary false discovery rate, GPR = Geometric mean of binary precision and binary recall

NA - Although, stratified random sampling was used to split the six datasets into the training and testing datasets, yet, some of the reaction subclasses had no representation in the respective test datasets, thus few of the binary matrices could not be calculated for these reaction subclasses. These are represented as NA values

TRANSPARENT METHODS

Construction of microbial species database

The data and text mining of the available literature was performed to construct a manually curated database of bacterial species present at different skin sites. The protein sequences of genomes of the bacterial species from the constructed database were retrieved from NCBI Reference Sequence Database (RefSeq) (O'Leary et al., 2015). Only complete genome assemblies were used from the RefSeq database. The genomes of all the available strains of a species were used to construct the pangenome of that species which helped to compile the metabolic potential of all the strains for a particular species. The pangenome of a species includes all the genes from all the different strains of that species. The pangenomes were constructed for all the bacterial species that are experimentally known to be a part of skin microbiome, and for which the complete genomes were available on the NCBI RefSeq database. The information on different bacterial species for which the pangenomes were constructed, and the skin sites harbouring these species is provided in **Supplementary Data Sheet 1**. The information on 19 different sites primarily including the sebaceous, moist, and dry skin niches was retrieved from literature, manually curated, and was used for further analysis. For the construction of pangenome, the protein sequences of all genomes of a species were merged and clustered at 100% identity using CD-HIT v4.6 (Li and Godzik, 2006).

Construction of skin microbiome specific metabolic information database

The ExPASy enzyme database was used to find the Uniprot/SwissProt IDs of all the annotated enzymes that belong a particular metabolic reaction annotated as four-digit EC number (Gasteiger et al., 2003). The protein sequences for these enzymes were downloaded from the Uniprot database (Consortium, 2014). The homology search of these enzyme sequences was performed against each pangenome to identify all the metabolic enzymes present in that pangenome using the NCBI BLASTP program (Altschul et al., 1990). The hits were filtered using the cut-off criteria of identity >50%, bit-score >100, query coverage >50%, subject coverage >50%, E-value $<10^{-10}$, mismatch percentage <50%, and gap percentage <50%. Finally, a database of complete reactions annotated as four-digit EC number and corresponding metabolic enzymes from all the pangenomes was constructed. Each of the metabolic enzymes was tagged with the bacterial species pangenome containing the enzyme. Further, the metabolic enzymes were also tagged with the skin sites that harbour the bacterial species with those enzymes.

Construction of reaction, RDM pattern, and substrate database

All the enzymatic reactions and their corresponding reactions IDs were retrieved from KEGG database (Kanehisa and Goto, 2000). For each reaction ID, the corresponding reactions pairs and respective RDM patterns were also retrieved from the KEGG database (Kanehisa and

Goto, 2000). From this data the databases of reactions, reaction pairs, and RDM patterns were constructed. From the reactions, the primary substrates were identified and a database of primary substrates and their respective reactions annotated as four-digit EC number was constructed.

Calculation of molecular features of substrates

The structural and chemical features were calculated for each of the substrate molecule in the substrate database. Thus, the molecular information of substrates was translated into machine-readable features that include chemical properties parameters, linear structural fingerprints, and circular molecular connectivity information. The chemical features were calculated using the PaDEL software (Yap, 2011). These chemical features included different types of chemical descriptions such as acidic atom count, aromatic atom count, aromatic bonds count, carbon types, molecular distance edge etc. encoded into 240 different values. Two types of structural fingerprints were calculated: linear and circular. The linear fingerprints were calculated using the PaDEL software (Yap, 2011). A total of 12 different types of linear fingerprints (Fingerprinter, Pubchem, MACCS, Atom pairs 2D, KlekotaRoth etc.) were calculated that were represented as 10,208 bits (values either 0 or 1). The two types of circular/topological fingerprints, Morgan FCFP - 512 bits and Morgan ECFP - 512 bits, were calculated using RDkit software (Landrum, 2016).

Feature selection

The Boruta algorithm implemented in R as the “Boruta” package was used to extract the important features among all the above calculated molecular features (Kursa and Rudnicki, 2010). Boruta is a wrapper algorithm for feature selection that uses “Random Forest” algorithm, and scores each feature and marks them as important, unimportant or tentative. The tentative features were then finalized as important or unimportant using “TentativeRoughFix” function of Boruta package in R. The variable importance was calculated for each EC reaction (EC1 to EC6) class separately. Finally, the important features for each EC were merged and unique sorted to obtain the final set of important features.

Principal component and cluster analysis

Principal component analysis was performed using the “prcomp” function from “stats” package in R v3.4.4. This function performs the principal component analysis (PCA) by performing the singular value decomposition of the input data (Mankin, 2003). This method is the preferred method for better numeric accuracy. The PCA and scree plots were generated using the “factoextra” and “ggfortify” package in R v3.4.4 (Kassambara and Mundt, 2017). The density-based clustering was performed using the “fpc” and “dbscan” package in R v3.4.4. The kNN distance plot was generated using the “kNNdistplot” function

from “dbscan” package in R v3.4.4 (Tran et al., 2013). The density cluster plot was generated using the “factoextra” package in R v3.4.4 (Kassambara and Mundt, 2017).

Hierarchical clustering

The hierarchical clustering was performed using the ‘hclust’ function of ‘stats’ package in R v3.4.4. The approximate unbiased p-values (AUp) and the bootstrap probability (BP) values for each branch/cluster were calculated using multiscale bootstrap resampling and using normal bootstrap resampling, respectively. The optimum number of clusters was identified to be two based on the average silhouette method .

Construction of machine learning models

Dataset construction

The dataset of 3,769 substrate molecules was randomly split into a working and blind dataset with a ratio of 95:5. Thus, the working dataset had 3,602 molecules and the blind dataset had 167 molecules. The working dataset was utilized for the training and statistical evaluation of the machine learning model, and the blind dataset was used for the independent evaluation of the model. The dataset was highly skewed with a higher number of substrate molecules for “Oxidoreductases” and “Transferases” in comparison to other reaction classes. Also the abundances of substrate molecules belonging to different combinations of reaction classes were also highly variable. Thus, a modified strategy of stratified random sampling approach was used to divide the working dataset into the training and testing dataset for modeling. The details of the dataset construction are mentioned in **Supplementary Text S1**.

Training and evaluation

The prediction of reaction class is a multiclass multilabel problem because one substrate molecule can undergo more than one type of reaction among the six types of reactions classes. In machine learning, there exists two methods to model the multiclass multi-label problem, one is problem transformation method where the multiclass multi-label problem is divided into several multiclass or binary problems, and another is algorithm adaptation method where the algorithms are adapted to perform the multiclass multi-label predictions. For the problem transformation method all the algorithms used for binary or multiclass classification can be used, whereas for algorithm adaptation method the algorithms need to be changed before using them for the multiclass multi-label classification. In the problem transformation method, a learner known as “wrapped multilabel learner” is employed on the “core learner”. The function of wrapped learner is to manage and combine several core learners so that they can work in synchronization to achieve the multilabel classification. The core learner is any traditional algorithm for binary or multiclass classification. We used five different wrapper methods: (1) binary relevance (BR) method, (2) classifier chains (CC)

method, (3) Nested stacking (NS) method, (4) Dependent binary relevance (DBR) method, and (5) Stacking method. We used the seven core learners for each of the above mentioned wrapper methods, these are: k-Nearest Neighbors (kNN), Recursive Partitioning (RPART), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost),Perceptive Neural Network (NNET), Naive Bayes, Random Forests (RF). In the algorithm adaptation method, we used two methods, randomForestSRC (RFSRC) and random ferns (RFerns).

The performance of the models was evaluated using two types of matrices, multilabel - to assess the capability of the model to perform the multilabel classification, and binary - to assess the capability of the model to perform the binary classification for each label. Five matrices were used in the multilabel case namely: Multilabel Accuracy, Multilabel Sensitivity or Recall or True Positive Rate, Multilabel Precision or Positive Prediction Value (PPV), Multilabel F1 measure (F1), and Hamming loss (Charte and Charte, 2015). The formulas for these matrices for multiclass multilabel classification are mentioned below (Charte and Charte, 2015):

$$Multilabel\ Accuracy = \frac{1}{|D|}\sum_{i=1}^{i=|D|}\frac{|Pi\ \cap\ Ti|}{|Pi\ \cup\ Ti|};\ Multilabel\ Precision\ = \frac{1}{|D|}\sum_{i=1}^{i=|D|}\frac{|Pi\ \cap\ Ti|}{|Ti|}$$

$$Multilabel\ Recall\ or\ Sensitivity\ = \frac{1}{|D|}\sum_{i=1}^{i=|D|}\frac{|Pi\ \cap\ Ti|}{|Pi|}$$

$$Multlabel\ F1\ measure = \ 2 * \frac{Precision * Recall}{Precision + Recall};\ Hamming\ loss = \frac{1}{|D|}\sum_{i=1}^{i=|D|}\frac{|Pi\ \Delta\ Ti|}{|C|}$$

Where, D is the total number of instances in the multiclass multilabel dataset, C is the complete set of labels present in the multiclass multilabel dataset, *Pi* is the predicted labels for the i^{th} instance, and Ti is the true labels for the i^{th} instance. The set operations used were: ∩ meaning intersection, ∪ meaning union, and *Δ* meaning symmetric difference.

We used eight matrices to evaluate the binary performance: Binary Accuracy, Mean Misclassification Error (MMCE), Matthews Correlation Coefficient (MCC), Binary Precision or Positive Predicted Value (PPV), Area under the curve (AUC), Binary False Negative Rate (FNR), Binary False Positive Rate (FPR), Binary Sensitivity or Recall or True Positive Rate (TPR), Binary Specificity or True Negative Rate (TNR), Binary Negative predictive value (NPV), Binary False discovery rate (FDR), and Binary Geometric Mean of binary precision and binary recall (GPR). The formulas for these binary performance matrices are mentioned in **Supplementary Text S2**. The final model for reaction class and subclass prediction was constructed with the method that showed the best multilabel and binary performance.

Construction of artificial neural network (ANN) models

Dataset construction

The aim of constructing the ANN model was to improve upon the learning about the class-specific patterns, thus, only the substrates where the molecule could undergo the reactions of only one type of reaction class were extracted from the working dataset (as mentioned above) and were used for the construction of ANN models. This dataset had a total of 1,758 substrate molecules with the distribution of molecules across different reactions classes: “Oxidoreductases” - 832, “Transferases” - 573, “Hydrolases” - 195, “Lyases” - 79, “Isomerases” – 41, and “Ligases”- 36. It is evident from the distribution that the dataset is very biased and imbalanced, thus, the stratified random sampling was performed to split this dataset into training and testing dataset. For stratified random sampling, this dataset was first divided into six parts, one for each reaction class, and then each reaction class dataset was splitted separately into training and testing dataset using random sampling with the split ration of approximately 90:10. Now all the six training sets were merged to create the final training dataset and all the six testing datasets were merged to create the final testing dataset.

Training and evaluation

The ANN network was constructed in Python using libraries tensorflow v1.4.1 and keras v2.2.4. Based on the nature of the problem, the best suited multilayer perceptron model that is based on the backpropagation method for training is used. In the backpropagation method, the error rate is provided as feedback to the whole neural network that is known as back propagating the error, which is then used by an optimizer algorithm to optimize the parameters of artificial neural network.

Three different matrices were used to evaluate the performance of the ANN model: categorical accuracy, binary accuracy, and log loss/binary cross entropy. Since it is a multiclass classification problem the target variable here is one hot encoded. The categorical accuracy checks if the maxima in the true values and the maxima in the predicted values have the same index, if yes, it is considered a true prediction, else it is considered a wrong prediction. This is performed on all test dataset instances, and the fraction of correct predictions out of total predictions on test dataset gives the categorical accuracy. In contrast, for calculating the binary accuracy, at first all the probabilities are converted into values with the threshold of 0.5 (if <0.5 means 0, and if >0.5 means 1), then all the true values of each instance are compared with the predicted values. If the true value is equal to the predicted value then it is considered as correct prediction, else it is considered a wrong prediction. This was also performed on all the values of each of the test dataset instance, and the fraction of correct predictions out of the total predictions gives the binary accuracy. The formula to calculate the log loss/binary cross entropy is:

$$-\sum_{c=1}^{N} Y(i,c)\log\left(P(i,c)\right)$$

Where, N is the number of different classes present in the dataset, log is the natural logarithm, Y(i,c) is the indicator if the classification is correct (1 if yes and 0 if no) for i^{th} observation for c class, and P(i,c) is the probability predicted by the ANN model for i^{th} observation for c class

The hypermeters of the ANN models were also optimized based on the three evaluation matrices mentioned above to obtain the best performance from the ANN model. To calculate the optimum number of neurons in the hidden layer, the values close to the average of the size of input and output layers were tried and the best value was selected while keeping the number of hidden layer as one. Different number of hidden layers were tried to select the best performing ANN model with the most optimum number of hidden layers. A range of epoch values from 1 to 4000 were tried and based on the plateau in the performance an optimum value was selected. The other parameters of the ANN models were optimized using the grid search method with 5-fold cross validation the details are mentioned in **Supplementary Text S3**. The parameters optimized were: Weight initializer, Learning rate, Optimizer, Batch size, Dropout rate, and Weight constraint.

Statistical evaluation of the machine learning and ANN models

We used three methods to statistically evaluate the performance of the machine learning and ANN models. These three methods are split testing, cross validation, and blind set testing. The details of these methods are mentioned in **Supplementary Text S4**.

Molecular similarity search

The open source chemoinformatics tool Open Babel v2.3.2 was used for performing the molecular similarity search using the inbuilt default fingerprint FP2 which is a path-based fingerprint. The complete substrate molecule database was divided into several reaction subclass specific databases, depending on the type of reaction subclass a particular substrate can undergo. Once the reaction class and subclass are predicted by the machine learning and ANN models, the molecular similarity search against the predicted reactions subclass specific database is performed and Tanimoto Coefficient or Jaccard Index was calculated. The formula for calculating the Tanimoto Coefficient or Jaccard Index is:

$$Tanimotto\ coefficient\ or\ jaccard\ index\ T(a,b) = \frac{Nc}{Na + Nb - Nc}$$

Where, T(a,b) is the tanimoto coefficient for molecule a and b, Na is number of bits that are 1 in the fingerprints of molecule a, Nb is number of bits that are 1 in the fingerprints of

molecule b, and Nc is the number of bits that are 1 in the intersection of fingerprints of molecule a and b.

K-nearest neighbour (KNN) model construction or lazy learning

KNN is a preferred method for the identification of structurally and chemically similar molecules to the input molecule in the search against a heterogeneous database (Soucy and Mineau, 2001). The KNN algorithm was implemented using the R package "FNN" (Beygelzimer et al., 2015). The k-nearest neighbours for any given molecule were extracted using the function "get.knnx" from the "FNN" package that uses "Euclidean distance" as the measure of similarity between molecules.

Identification of reaction center

The reaction centers were identified by using the RDM pattern information that is associated with each of the substrate-product pair of an enzyme catalyzed reaction in KEGG database (Kanehisa, 2002). In the RDM pattern database constructed in this study, all the complete metabolic reactions are associated with the respective Reaction Class (RC) pairs, and all the RC pairs were tagged with corresponding RDM patterns. For a given biochemical reaction available in KEGG, the KEGG-defined RDM (Reaction center, Difference region, Matching region) patterns contain the information on the KEGG atom type changes at the reaction center, matched region of the molecule, and the difference region of the molecule (Kotera et al., 2013). Here a reaction center is the atom where the reaction occurs, a matched region is the region common between substrate and product that remained unchanged after the reaction, and a difference region is the part of molecule that changed after the reaction. The RDM patterns are derived from the structural alignments of the substrates and products which identifies the reaction center, matched and difference regions (Yamanishi et al., 2009). To identify the reaction center in a molecule for each of the predicted metabolic reaction, all the RC pairs and corresponding RDM patterns were extracted. Using these RDM patterns, the reaction centers were identified by in-house python scripts. Thus, this computational approach is similar to the biochemical approach in which the primary substrate and product are compared to identify the reaction center where the biochemical reaction has occurred in the enzyme active site.

SUPPLEMENTARY TEXT

Supplementary Text S1: (Related to Figure 4 and 5)

The dataset of 3,769 substrate molecules was randomly split into a working and blind dataset with a ratio of 95:5, the working dataset had 3602 molecules and the blind dataset had 167 molecules. The working dataset was utilized further for the training and statistical evaluation of the machine learning model and the blind dataset was used for the independent evaluation of the model. Since the dataset was much skewed with a very higher number of substrate molecules for “Oxidoreductases” and “Transferases” in comparison to other reaction classes and also abundance of substrate molecules with different combinations of reaction classes was very variable, thus, a modified strategy of stratified random sampling approach was used to divide the working dataset into the training and testing dataset for modeling.

In this approach, to account for the differences in substrates belonging to different combinations of reaction classes the working dataset was divided into pure (contains substrates that can undergo only one type of reaction among different reaction classes) and mixed datasets (contains substrates that can undergo multiple reactions among different reaction classes). The pure dataset which had 1756 substrate was statistically down-sampled to randomly select the same number (lowest in the sample = 36) of substrates for each reaction class. Thus, the down-sampled pure had a total of 216 substrates (36 of each reaction class). The mixed dataset had 1,846 substrate molecules which was split into two datasets large and small with the ratio of 95:5, the large part had 1,774 substrates, whereas the small part had 72 substrate molecules. The training dataset was constructed by merging the down-sampled pure dataset (216 substrates) and the large part of mixed dataset (1,774 substrates), and had a total of 1,990 substrate molecules. The testing dataset was constructed by merging the remaining of pure dataset after down-sampling (1,540 substrates) and the small part of the mixed dataset (72 substrates), and has a total of 1,612 substrate molecules. These final training and testing datasets corresponded to an approximate ratio of 55:45 of the working dataset of 3602 substrate molecules.

Similarly, for the training of machine learning models for reaction subclass prediction the working dataset was divided into six parts, one for each reaction class. The same substrate could belong to multiple parts if it can undergo reactions from multiple reaction classes. The numbers of reaction sub-classes in each reaction class were: “Oxidoreductases” - 22, “Transferases” - 9, “Hydrolases” -13, “Lyases” -8, “Isomerases” – 6, and “Ligases”- 6. For each dataset the stratified random sampling was performed to split the input dataset into training and testing dataset with the split ration of 90:10.

Supplementary Text S2: (Related to Figure 4)

To evaluate the binary performance we used eight matrices, Binary Accuracy, Mean Misclassification Error (MMCE), Matthews Correlation Coefficient (MCC), Binary Precision or Positive Predicted Value (PPV), Area under the curve (AUC), Binary False Negative Rate (FNR), Binary False Positive Rate (FPR), Binary Sensitivity or Recall or True Positive Rate (TPR), Binary Specificity or True Negative Rate (TNR), Binary Negative predictive value (NPV), Binary False discovery rate (FDR), and Binary Geometric Mean of binary precision and binary recall (GPR). The formulas for these binary performance matrices are mentioned below:

$$Binary\ Accuracy = \frac{TP + TN}{P + N}$$

$$MMCE = \frac{FP + FN}{P + N};\ MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Binary\ Precision = \frac{TP}{FP + TP};\ Binary\ FNR = \frac{FN}{TP + FN};\ Binary\ FPR = \frac{FP}{TN + FP}$$

$$Binary\ Sensitivity\ or\ Recall = \frac{TP}{FN + TP};\ Binary\ Specificity\ = \frac{TN}{FP + TN}$$

$$Binary\ NPV = \frac{FP}{TN + FP};\ Binary\ FDR = \frac{FP}{TP + FP};\ Binary\ GPR = \sqrt{Precision * Recall}$$

Where, TP is true positives, FP is false positives, TN is true negatives, FN is false negatives, P is the total number of positives, and N is the total number of negatives in the input dataset.

Supplementary Text S3: (Related to Figure 5)

The other parameters of the ANN models were optimized using the grid search method with 5-fold cross validation, the parameters along with the values tested are: (1) Weight initializer – Zero, Normal, Uniform, Glorot normal, Glorot uniform, He normal, He uniform, and Lecun uniform (2) Learning rate – 0.0, 0.02, 0.1, 0.2, and 0.3 (3) Optimizer – Adadelta, Adagrad, Adam, Adamax, Nadam, RMSprop, and SGD (4) Batch size – 0, 50, 100, 150, 200, 250, 300, 400, and 500 (5) Dropout rate and Weight constraint – [0.0, 1], [0.0, 2], [0.0, 3], [0.0, 4], [0.0, 5], [0.1, 1], [0.1, 2], [0.1, 3], [0.1, 4], [0.1, 5], [0.2, 1], [0.2, 2], [0.2, 3], [0.2, 4], [0.2, 5], [0.3, 1], [0.3, 2], [0.3, 3], [0.3, 4], [0.3, 5], [0.4, 1], [0.4, 2], [0.4, 3], [0.4, 4], [0.4, 5], [0.5, 1], [0.5, 2], [0.5, 3], [0.5, 4], [0.5, 5], [0.6, 1], [0.6, 2], [0.6, 3], [0.6, 4], [0.6, 5], [0.7, 1], [0.7, 2], [0.7, 3], [0.7, 4], [0.7, 5], [0.8, 1], [0.8, 2], [0.8, 3], [0.8, 4], [0.8, 5], [0.9, 1], [0.9, 2], [0.9, 3], [0.9, 4], and [0.9, 5]. The final model was constructed using the most optimum parameters selected based on the gird search method.

Supplementary Text S4: (Related to Figure 4 and 5)

a) *Split testing*: As mentioned in the dataset construction part the complete working dataset was divided into training and testing dataset using a specific splitting approach. The models were trained on the training dataset and evaluated in the test dataset.
b) *Cross validation*: In this study, we used 5-fold cross validation for machine learning models and ANN models. In this method, during the process of training the dataset was randomly divided into five equal parts and five iterations of training and testing are performed. In each of the iteration four parts are used for training and the rest one part is used for the testing. This way in five iterations each of the training instances is used for testing the model and thus, avoiding any bias in the evaluation of the performance matrices. Finally, the mean/median and standard deviation value of performance matrices across five iterations is used to evaluate any bias in the model such as over-fitting or under-fitting.
c) *Blind set testing*: Approximately 5% of the randomly selected instances are kept aside before starting the training and testing process of model and, the model never sees these instances at any stage of its training and testing, hence called a blind dataset to model. Therefore, the performance of the model on this blind dataset is considered to be a real or unbiased performance of the model.

Supplementary Text S5: (Related to Figure 1)

To further evaluate the variability in skin sites in terms of enzymatic reactions it is critical to know the reactions that are common to the different sites. To identify the number of reactions that are common to different sites the matrix layout analysis performed using the 'UpSetR' package in R (Conway et al., 2017; Lex and Gehlenborg, 2014). It generates a matrix layout diagram for visualizing the set intersections.

Supplementary Text S6: (Related to Figure 1)

A skin microbiome specific metabolic enzyme database of four-digit EC number and corresponding metabolic enzymes from all the pangenomes was constructed. Each metabolic enzyme in this database is tagged with the bacterial species if their pangenome harbors this enzyme. Also the metabolic enzymes were tagged with the skin sites based on the presence and absence of the bacterial species harboring the enzyme on that particular skin site. All the well-annotated enzymatic reactions were extracted from KEGG database

and corresponding reaction, primary substrates, RC pair and RDM pattern databases were constructed. A total of 10,629 reactions, 3,769 primary substrates, and 2,592 RC pairs and RDM patterns were extracted from the KEGG database. Also each of the primary substrate in the database was tagged with the reaction class (EC-one digit), reaction subclass (EC-two digit), and complete reaction category (EC-four digit). The four types of molecular features were calculated for each of the primary substrate: chemical descriptors, linear fingerprints, Morgan ECFP fingerprints, and Morgan FCFP fingerprints. All this data was used for the training of the prediction models and for making the final metabolism predictions.

Supplementary Text S7: (Related to Figure 4 and 5)

Feature selection was performed using the Boruta package on chemical descriptor and linear fingerprints (Kursa and Rudnicki, 2010). Boruta selected 194 features out of 240 chemical descriptors and 1,104 features out of 10,208 linear fingerprints across different reaction classes. All features of Morgan FCFP and Morgan ECFP were included and were not subjected to feature selection as these features are elementary and all the bits are needed to adequately describe a substrate molecule. The total number of features used for further analysis was 2,322: 194 from chemical descriptors, 1104 linear fingerprints, 512 Morgan ECFP fingerprints, and 512 Morgan FCFP fingerprints.

Supplementary Text S8: (Related to Figure 6)

A previous study have reported that the adjustment in the threshold of multiclass multilabel classification model could significantly improve on their performance (Al-Otaibi et al., 2014; Fan and Lin, 2007). Thus, although we evaluated the performance of our models using the threshold value of 0.5 for all the machine learning and ANN models so that we do not overestimate the sensitivity of our models, the prediction threshold of the models deployed in the web server was lowered for the reaction subclass prediction models of “Oxidoreductases”, “Transferases”, “Hydrolases”, and “Lyases” classes from 0.5 to 0.2 because they had a range of 8 to 22 different subclasses and a high threshold could lead to miss out on some possible reaction subclasses.

REFERENCES

Al-Otaibi, R., Flach, P., and Kull, M. (2014). Multi-label classification: A comparative study on threshold selection methods. Paper presented at: First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD 2014.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.
Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., Li, S., and Li, M.S. (2015). Package 'FNN'. Accessed June *1*.
Charte, F., and Charte, D. (2015). Working with Multilabel Datasets in R: The mldr Package. R Journal *7*.
Consortium, U. (2014). UniProt: a hub for protein information. Nucleic acids research *43*, D204-D212.
Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938-2940.
Fan, R.-E., and Lin, C.-J. (2007). A study on threshold selection for multi-label classification. Department of Computer Science, National Taiwan University, 1-23.
Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic acids research *31*, 3784-3788.
Kanehisa, M. (2002). The KEGG database. Paper presented at: Novartis Foundation Symposium (Wiley Online Library).
Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research *28*, 27-30.
Kassambara, A., and Mundt, F. (2017). Package 'factoextra'. Extract and visualize the results of multivariate data analyses *76*.
Kotera, M., Tabei, Y., Yamanishi, Y., Moriya, Y., Tokimatsu, T., Kanehisa, M., and Goto, S. (2013). KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. BMC systems biology *7*, S2.
Kursa, M.B., and Rudnicki, W.R. (2010). Feature selection with the Boruta package. J Stat Softw *36*, 1-13.
Landrum, G. (2016). RDKit: open-source cheminformatics software.
Lex, A., and Gehlenborg, N. (2014). Points of view: Sets and intersections. Nature Methods *11*, 779-779.
Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.
Mankin, E. (2003). Principal Components Analysis: A How-To Manual for R. Desde http://psych colorado edu/wiki/lib/exe/fetch php.
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., and Ako-Adjei, D. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research *44*, D733-D745.
Soucy, P., and Mineau, G.W. (2001). A simple KNN algorithm for text categorization. Paper presented at: Proceedings 2001 IEEE International Conference on Data Mining (IEEE).
Tran, T.N., Drab, K., and Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometrics and Intelligent Laboratory Systems *120*, 92-96.
Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., and Kanehisa, M. (2009). E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics *25*, i179-i186.
Yap, C.W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry *32*, 1466-1474.