Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations

渡邉, 直暉

(Degree)
博士（工学）

(Date of Degree)
2023-03-25

(Date of Publication)
2024-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第8647号

(URL)
https://hdl.handle.net/20.500.14094/0100482395

DOCTORAL DISSERTATION

Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations

酵素アノテーションの網羅的予測のための機械学習モデルの開発

January, 2023

GRADUATE SCHOOL OF ENGINEERING

KOBE UNIVERSITY

WATANABE Naoki

渡邉　直暉

Contents

CHAPTER I

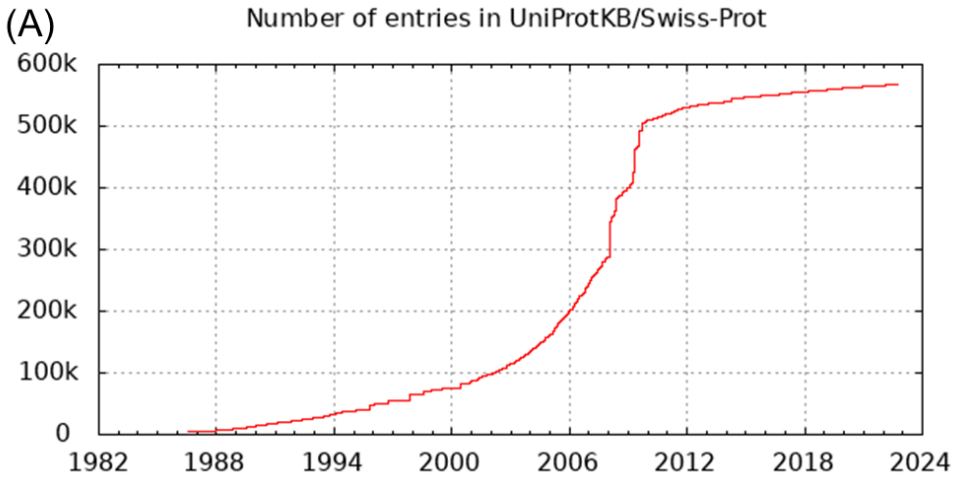
Introduction

I.1. Metabolic Engineering and Enzyme Sequences

Enzymes catalyze biochemical reactions and synthesize the molecules called as products by acting with the molecules called as substrates. All organisms utilize enzymes to metabolize and absorb nutrient sources, to synthesize energy, and to degrade harmful substances for the survival of life. Enzymes are not only essential proteins for each organism, but are used in various industrial fields.

Enzymes and metabolic engineering are closely linked, and the various enzymes are applied to microorganisms in order to biosynthesize a wide range of industrial chemicals, pharmaceuticals, antibiotics, and food additives[1,2]. However, microbial metabolic pathways and enzymes are not necessarily optimal. In particular, natural enzymes tend to be difficult to produce on a large scale, and even on a low scale, the productions of target substances are low[3,4]. Therefore, the syntheses of target compounds have been improved by designing and engineering optimized metabolic pathways, and utilizing enzymes derived from exogenous genes with high activity[5–7]. Moreover, the scope of target compounds that can be covered by the rapidly developing fields of genetic engineering, synthetic biology and metabolic engineering has recently increased and more complicated compounds have been synthesized[8–10]. Thus, enzyme modification and metabolic pathway design play an important role in the production of industrial useful substances.

Protein sequence information has been registered in various biological databases like that of the National Center for Biotechnology Information (NCBI)[11] and the protein database UniProt[12]. UniProt contains Swiss-Prot[15] and TrEMBL[16] databases. Protein sequences manually annotated from experiments are registered in the Swiss-Prot database, while sequences annotated via computational methods and unannotated sequences are registered in the TrEMBL database. In particular, the Swiss-Prot contains over 250,000 enzymes within about 560,000 protein sequences (Figure 1A). On the other hand, the number of the sequences is explosively increasing and is over 220 million (Figure 1B). Numerous hypothetical and uncharacterized enzyme functions are accelerating because of the development of genome sequencing technology[13,14]. Therefore, the number of available enzyme sequences could potentially increase by explosively increasing the number of unannotated sequences.



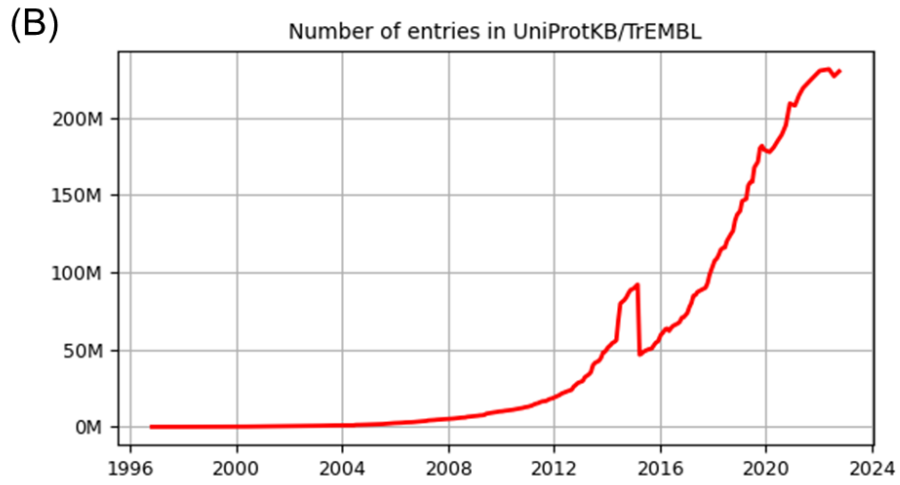


Figure 1. The number of protein sequences registered in (A) Swiss-Prot[15] and (B) TrEMBL[16].

The various protein annotations with the sequence information are registered in these databases. For example, the Enzyme Commission (EC) number system is used to classify enzymes using 4 digits based on the reaction type. Enzymes are also classified in Gene Ontology (GO) which is used to annotate proteins[17]. The first digit of EC numbers represents one of 7 main enzymatic reactions (Oxidoreductase, Transferase, Hydrolase, Lyase, Isomerase, Ligase and Translocase), and accordingly there are 7 first digit EC number classes referred to as EC 1 to EC 7. Translocase has been added to the first digits since 2018 and is related to the movement of ions or molecules across cell membranes. The second and third digits classify more details including the type of bonds, functional groups and cofactors involved in the catalyzed reaction. The fourth digit is a serial number to identify each specific enzyme (Figure 2). Enzymes or enzymatic reactions are classified as above and the EC number can be used to search for the candidate enzyme sequences that can synthesize the specific target compounds.

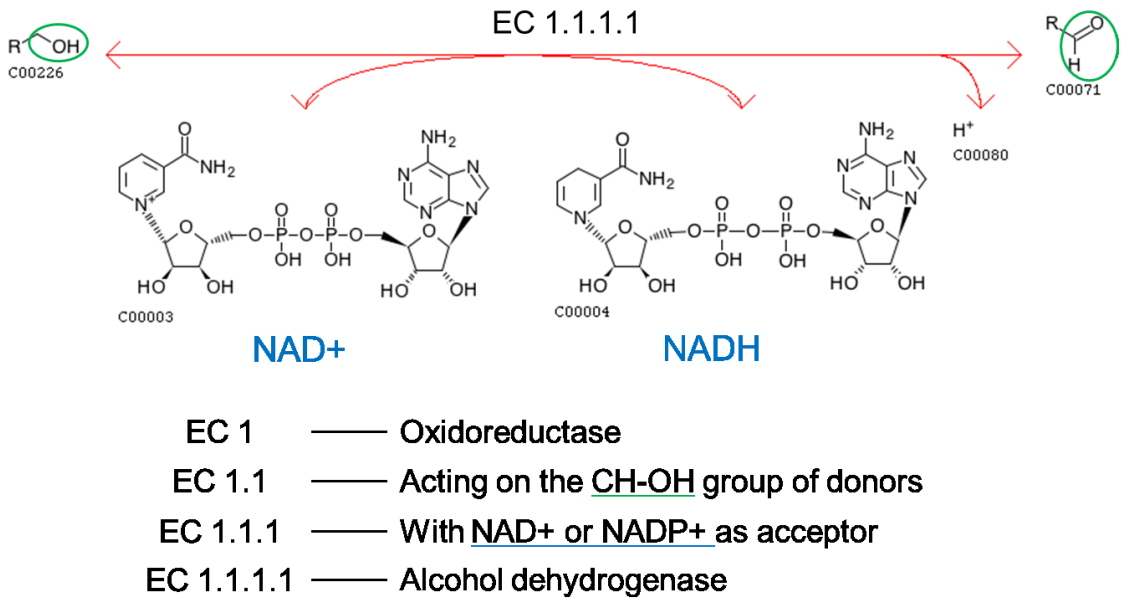


Figure 2. EC number system. Chemical equation[18] of EC 1.1.1.1 reaction catalyzed by alcohol dehydrogenase is shown as an example.

Some existing enzymes may react with unknown substrates to form newly characterized products in addition to known natural reactions. More importantly, newly discovered enzymes might catalyze novel reactions. Therefore, novel enzyme discovery is required to increase the production of target compounds and to expand the applications of metabolic engineering[8,9]. While experimental methods can improve the activity of an enzyme, it is difficult to discover new enzymes that have not been previously reported. Although new enzymes may exist within unannotated proteins that are increasing, experimental methods are limited due to high costs and time constraints. Therefore, a valid computational method to predict enzyme functions from sequence information is needed to discovery novel enzymes within a huge number of unannotated sequences.

I.2. Enzyme Annotation Prediction Based on Sequence Similarity

The most basic solution in computational methods is to use Basic Local Alignment Search Tool (BLAST) algorithm in which highly similar enzyme sequences to input sequences are searched from protein sequence databases and their functions are inferred based on the most similar annotated enzymes[19–21]. BLAST connects some protein sequence databases such as Non-redundant protein sequence database, and Nucleotide collection database in NCBI[11]. The latter database is used to search similar nucleotides to input nucleotide sequences. About 1.1 billion protein sequences are registered in Non-redundant protein sequence database and about 0.55 billion nucleotide sequences are registered in Nucleotide collection database on December 2022. By inputting enzyme sequences which can synthesize target compounds, some candidate enzymes are predicted if the similar sequences to input enzyme sequences are registered in databases. For example, BLAST has been used to annotate metagenomic contigs obtained from cow rumens and to identify protease inhibitor peptides[22,23].

Enzyme sequence based individual measures of quantifying the relationship between a protein sequence and a functional class, hierarchical clustering algorithms and Hidden Markov Models have been used to improve BLAST with some success[24–30]. These studies have reported enzyme annotation predictions. However, the sequence similarity based methods cannot predict the function of uncharacterized enzymes with low similarity to annotated enzymes. The annotations of the sequences regarded as hypothetical proteins by BLAST are not predicted. This is why the results using the methods are more highly depended on known information. Furthermore, it has recently been reported that machine learning predictions performed better than BLAST[31–35]. For

one reason, BLAST is not effective in predicting homologs with different activities. Therefore, several studies have recently reported machine learning methods for predicting various biological annotations as described in the next section.

I.3 Progress of Biological Annotation Prediction using Machine Learning

Machine learning can process vast amounts of available enzyme sequences and is suitable for the mass prediction of various biological functions. To build prediction models, protein sequence information with biological annotation should be transformed to the feature vectors and target classes which want to predict the annotations should be built. The feature vectors and classes are learned by a machine learning method and then a prediction model can be built. The feature vectors derived from unannotated sequences are inputted to the model to predict annotations (Figure 3). Several methods have been developed to extract various features from protein sequences and automatically transform them to feature vectors[36–39]. Moreover, the other methods have also been developed to automatically build and evaluate prediction models in addition to sequence feature extractions[40,41]. As with the annotation predictions based on sequence similarity methods, it is not easy to classify the sequences with high similarity and different functions, however, machine learning models can improve the prediction accuracy by optimizing feature extractions and training datasets. No results are outputted if BLAST does not find similar sequences to an input sequence, while machine learning models necessarily output some results.

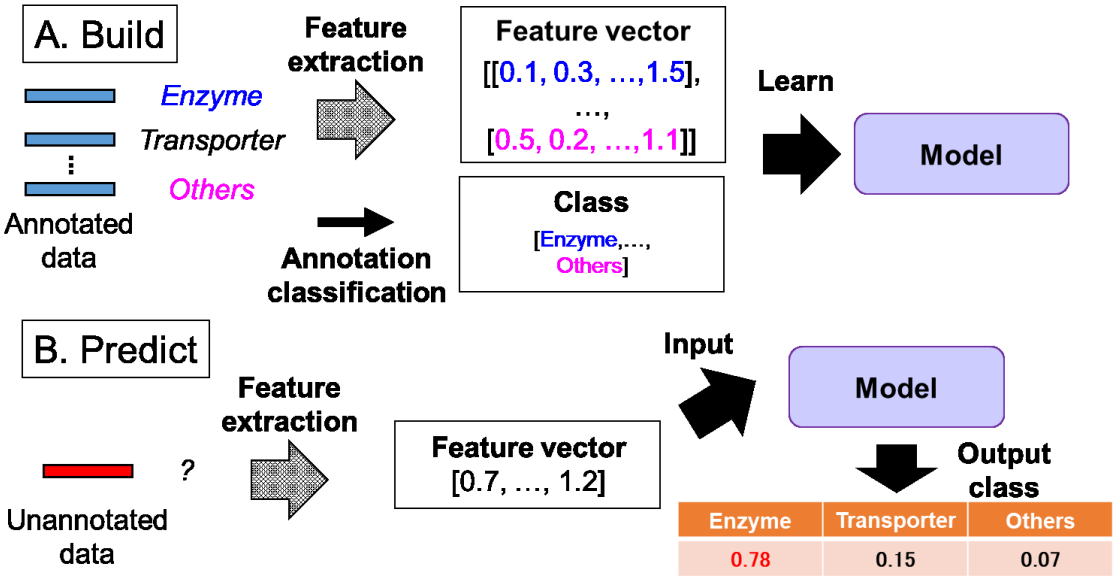


Figure 3. Classification task using machine learning.

The simplest feature extractions from protein sequences, amino acid composition and dipeptide composition, are calculated as the frequency of each amino acid and dipeptide contained within the protein sequence. Namely, the number of dimensions of the methods are 20 and 400, respectively. The methods were used for nuclear receptor and target enzyme classifications[42,43]. Due to the simple method, some accuracy can be achieved for simple predictions, however, complex classifications do not provide significant benefits such as EC number prediction[44–53], protein function prediction[50,54–57], and GO prediction[32,33,50,58,59]. Therefore, combining the methods with several physicochemical properties of amino acids included in protein sequences (hydrophobicity, polarizability, solvent accessibility and secondary structure), more detailed features could be extracted and the prediction results have been improved[54,55,60–65]. However, the above methods completely lose the information regarding the order of the amino acids in the sequences. As the number of data used in

machine learning models increases, the diversity of protein sequences increases and the features for each sequence are more important. Therefore, it is necessary to extract the sequence-like features of the 20^N(N: The length of amino acids) variations of sequences without losing them.

To overcome this problem, feature extractions such as one-hot encoding[44–47,66–68] and Word2Vec [32,33,47,69–73] are used. In one-hot encoding, each sequence is converted to a matrix with 20 rows and L columns (L: The number of residues). The matrix is included only 0 or only 1. For example, alanine is converted to [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] and cysteine is converted to [0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]. Next, Word2vec approach was used in the natural language processing study for the first time[74]. The approach learns words as high dimensional embedding and the feature vectors of the similar words end up near in feature vector space. Thus, a single amino acid sequence is regarded as a single sentence, and the raw sequence features are expressed as feature vectors. However, it may be possible to make highly accurate predictions using the feature extractions that lose sequence information when the number of data is not so large or when making simple predictions such as binary classification[75]. Therefore, some strategies aim to build highly accurate prediction models by using both sequence-like and no-sequence-like information to generate feature vectors with a wider range of information[44,45,48,49,65,76–78]. It is important to select feature extractions that match the scale of the training data and the feature information and the prediction annotations.

There are also various types of machine learning algorithms used for predictions. About 10 years ago, the most popular algorithms were the Support Vector Machine (SVM), Decision Trees, Random Forests (RF), and k-nearest neighbor (kNN), which were used to predict various annotations[35,54,79–84]. In some reports, multiple machine learning models were built and compared in order to achieve higher prediction accuracy. However, with the increase in available data, these classical machine learning methods have made it difficult to make highly accurate predictions. In particular, SVM has the critical shortcoming that the run time increases more explosively as the number of training data increases in comparison to the other classical machine learnings.

Annotation prediction using machine learning has evolved significantly in recent years because of the development of deep learning which was originally applied to natural language processing, voice recognition, and image recognition[85]. Deep learning can automatically learn input data while extracting the features and model performances are higher if training data are valid[86]. Several studies have reported deep learning methods for predicting protein functions[32,33,50,57], compound-protein interaction[60,72], protein structures[87,88], protein subcellular localization[66], enzyme commission numbers[46,48–50] and products in organic synthesis[89]. The advantages of deep learning enable to automatically extract features from input data, automatically discover complex patterns, and understand the features of data based on its original criteria. Accordingly, prediction models are often built by combining deep learning with one-hot encoding or Word2Vec. Overall, various machine learning algorithms have also been explored. Most importantly, various feature extractions and machine learnings are used because the optimal solutions for the methods depend on the annotation to be predicted.

I.4 Machine Learning for Prediction of Novel Enzymes and Metabolic Pathways

As described in the previous sections, conventional enzyme function prediction models built from enzyme sequences using machine learnings have recently been developed to discover novel enzymes. However, this strategy can only predict the enzyme information. In order to synthesize functional compounds using microorganisms, it is necessary to simultaneously predict enzymes and even substrates and products in enzymatic reactions. This prediction will not only discover new enzymes, but apply to the discovery of novel metabolic pathways.

In Chapter II, enzymatic reaction prediction models using multiple classical machine learning algorithms have been built by combining enzyme sequence and compound structural information, and the abilities of previously reported models for prediction of enzyme functions are expanded. The current combined models predict with higher accuracy than the models constructed with the same previous strategy. Therefore, the models are successfully built to provide the basis for predicting enzymatic reactions. However, these models cannot exclude unlikely enzyme-compound combinations, because they do not learn the combinations in which enzymatic reactions do not occur. Therefore, in Chapter III, while solving this problem, the models for more accurate predictions are developed by updating training datasets and feature extractions. As a result, the utilization of deep learning, which has explosively developed in recent years, has improved the prediction accuracy for extensive enzymatic reactions.

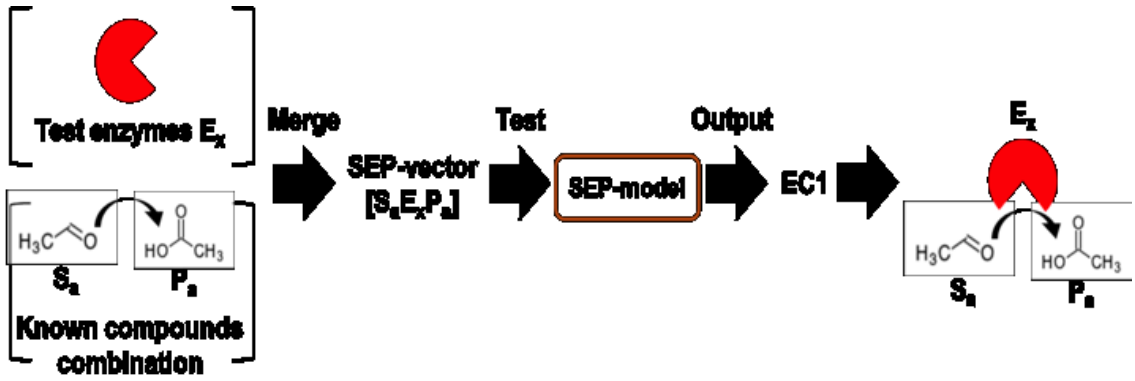
The current strategy has hypothesized that amino acid sequences used in the predictions are enzymes. However, the sequences may not be the enzyme when actually predicting unknown reactions. Thus, in Chapter IV, EC number prediction models are built so that proteins except for enzymes can be excluded from the candidate sequences for enzymatic reaction prediction. Incidentally, deep learning has enabled the prediction of the 3-dimensional structure of proteins from protein sequences with high accuracy[87,88]. The results indicate that deep learning is capable of capturing the more extensive enzyme features within a sequence. Moreover, several studies have reported models for predicting protein annotations from sequence information using Convolutional Neural Network[32,33,44–46,48,49,56,59,66,68,72,90–93], which is often used in image recognition. Therefore, EC number prediction models are built using ResNet, which contains the structures of multiple CNNs, and attempt to predict while capturing the structural features of enzyme sequences. As a result, more extensive EC numbers are predicted by the current model in comparison to previously reported models.

Combining the EC number prediction models with enzymatic reaction prediction models enables to predict candidate enzyme sequences with target functions and then predicts which compounds the enzymes will react with and which compounds the enzymes will able to synthesize. Therefore, the current system comprehensively predicts multiple annotations of enzymes. This system will help to select enzyme sequences and discover novel enzymatic reactions in metabolic pathways for the production of useful substances using microorganisms.

CHAPTER II

Exploration and Evaluation of Machine Learning Based Models for Predicting Enzymatic Reactions

Graphical Abstract



II.1. Introduction

Enzymes catalyze biochemical reactions and therefore are key targets for selection and modification within metabolic engineering applications as described in Chapter I. To expand the applications of metabolic engineering, it is essential to discover new enzymatic reactions and their related metabolic pathways. EC numbers are important enzyme classifiers used in annotations[94–96]. Moreover, enzyme sequence information is registered in various biological database[12,18,97]. Hundreds of genomes have recently been sequenced by way of next-generation sequencing[13,14]. With explosive increases in the number of available gene sequences, the number of unannotated sequences is increasing. Within the unannotated sequences, many novel enzyme functions may be

discovered. Therefore, enzymatic reaction prediction systems have been developed using various computational methods[98–100].

Of these methods, the most basic approaches involve the use of only amino acid sequence information to specifically predict EC numbers. This can include the use of BLAST in which highly similar sequences are found and their functions are inferred based on the most similar known enzymes[19–21]. Next, machine learning approaches which have the potential to acquire new knowledge from a large number of datasets have been pursued to build more accurate prediction models by increasing training data and feature extractions[48,49,52,53,55]. Li *et al*.[48] have evaluated and compared prediction results between multiple machine learning methods derived from enzyme sequence information. Li *et al*.[48], Zou *et al.*[49] and Che *et al*.[53] have also compared several feature extractions and machine learning methods. Yet, this strategy is still challenging because EC numbers are hierarchical and consist of thousands of classes. Therefore, prediction methods require a large number of multistep models. Several studies have reported machine learning methods in combination with BLAST. However, BLAST based methods cannot predict the function of uncharacterized enzymes with low similarity to annotated enzymes. Moreover, approaches using only enzyme sequences cannot predict detailed enzymatic reactions, which are dependent on information about substrates and products.

A second approach involves including substrate and product chemical structure information[82,101–108]. Yamanishi *et al*.[101], Kotera *et al*.[102], Matsuta *et al*.[103] and Moriya *et al*.[104] have proposed prediction methods based on structural changes of substrates to

products. These methods utilized substrates and products linked to EC numbers that are registered in KEGG and evaluated prediction using jackknife-type cross-validation. The methods of Yamanishi *et al.*[101] and Kotera *et al.*[102] have achieved practical coverage and prediction accuracy of EC numbers. However, the relationship between enzyme sequence and compound chemical structure cannot be inferred without including enzyme sequence information.

To overcome the above challenges in predicting novel enzymatic reactions, this Chapter II study further explores several classical machine learning algorithms with the capability to discover new enzyme functions. First, enzyme models (E models) are built using only amino acid sequences to predict the first digit of EC numbers. To expand the ability for prediction of enzymatic reactions in more detail, substrate-enzyme models (SE models) and substrate-enzyme-product models (SEP models) are developed to include substrate and product chemical structural information in addition to sequence information (Figure 4). Several SE and SEP models can directly predict enzymatic reactions and discover new metabolic pathways using both compound chemical structure and enzyme sequence information. After comparing each machine learning model, the addition of compound chemical structural information is found to be important for accurate prediction. A few other studies have evaluated enzyme prediction models in detail by comparing feature extractions and machine learning methods[48,51,53]. However, these studies use only enzyme sequence information and have not compared prediction results for each reaction type between machine learning methods. In the current SEP models, certain reaction types are more or less difficult to predict than others, depending on the machine learning method. For example, prediction of

oxidation/reduction reactions with hydrocarbons, alcohols, aldehydes and ketones is optimal using all SVM, RF, kNN and Multilayer Perceptron (MLP) based models. On the other hand, glycosyl transferases reactions are not predicted by SVM and MLP based models, but RF and kNN based models predict almost all of these reactions. Overall, the SEP-RF model is best, with an Average AUC score over 0.94 for prediction of enzymatic reactions in a single model and single step.

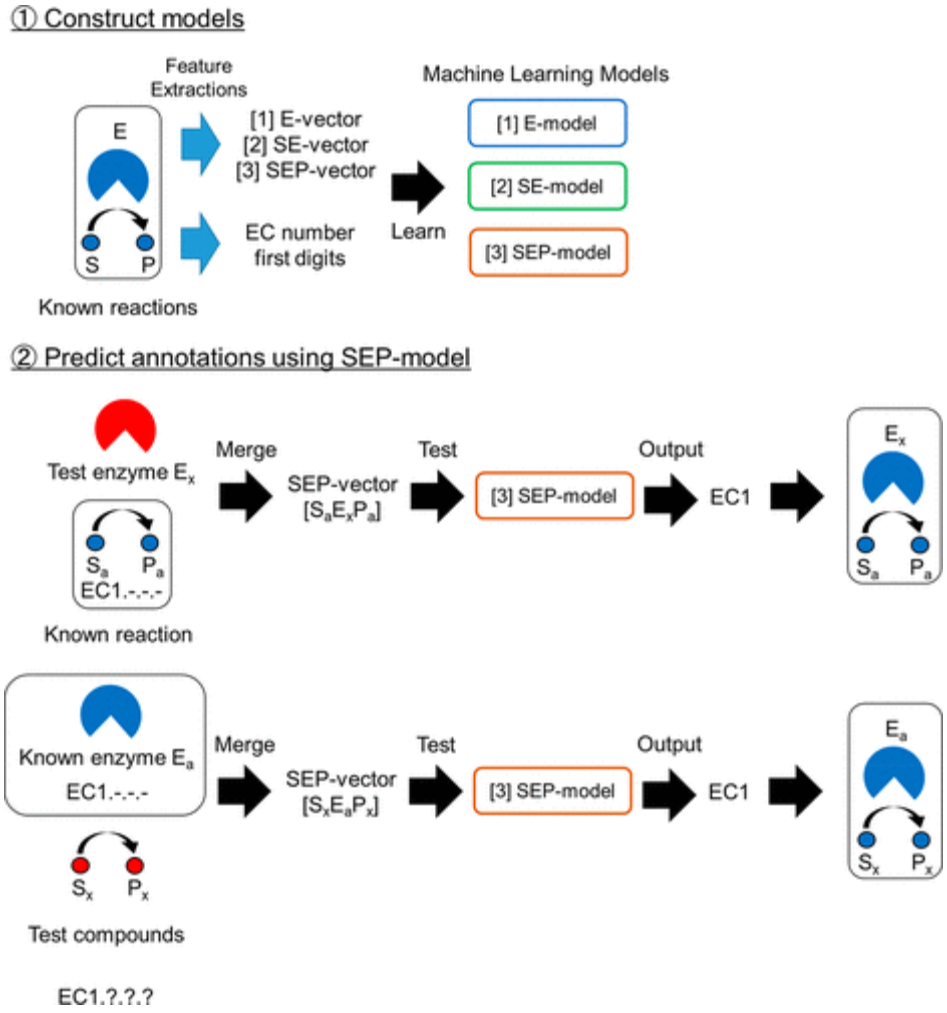


Figure 4. Scheme for enzyme prediction strategies used in Chapter II.

II.2. Materials and Methods

II.2.1. Data Collection

II.2.1-1. Training Datasets

EC numbers and reaction information for 38,320 enzymatic reactions were collected from BRENDA[97] and KEGG[18] to build 3 types of prediction models: E models, SE models, and SEP models. Some EC numbers, which correspond to substrate and products of protein, DNA or metal complexes, and those with very specific reactions, were all removed from training datasets. Enzyme sequences and simplified molecular-input line entry system (SMILES) strings for substrates and products were collected from Swiss-Prot and PubChem[109], respectively. For each gene, enzyme amino acid sequences from various species were aligned using MAFFT[110]. Consensus sequences were derived from alignments using EMBOSS[111] to decrease the size of training datasets. After removing duplicate information, 2882 enzyme, 25,320 substrate-enzyme, and 33,263 substrate-enzyme-product datasets were used to build E models, SE models, and SEP models, respectively.

II.2.1-2. Test Datasets

Sequences, annotations and EC numbers for 838 *Escherichia coli* K-12 enzymes were collected from Swiss-Prot[12] to train and evaluate the 3 models. For this chapter, only *E. coli* K-12 was selected based on availability of detailed annotations. Substrate and product datasets for all known EC numbers were collected from EC numbers according to KEGG. A total of 838 enzyme, 275 substrate-enzyme, and 299 substrate-enzyme-product datasets were used to test E models, SE models, and SEP models, respectively. A total of 210 enzyme sequences were included in 275 substrate-enzyme and 299

substrate-enzyme-product datasets. Vectors used in test datasets were not included as training dataset vectors. Table 1 shows the training and test datasets. Enzyme sequence similarity between training and test datasets is shown in Figure 5 and Figure S1 (Chapter II.6.4). Databases for comparing sequence similarity of test sequences with training sequences were built using BLAST+ 2.7.1[19–21]. BLAST results were also used to infer the function of 210 test sequences, in comparison to machine learning prediction using E, SE and SEP models.

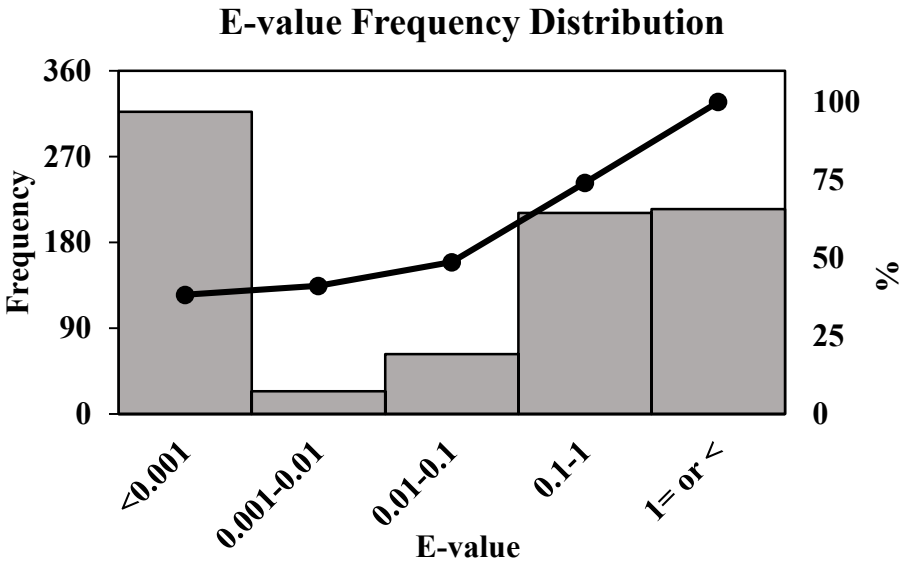


Figure 5. Enzyme sequence similarity between training and test datasets for E model evaluation. 8 of 838 test samples had low sequence identity to training sequences.

Table 1. List of Training and Test Datasets.

EC	Training datasets			Test datasets		
first digit	E	SE	SEP	E	SE	SEP
EC 1	681	7,928	8,760	157	53	56
EC 2	714	3,859	5,324	260	121	130
EC 3	964	11,794	17,173	291	24	25
EC 4	272	760	974	58	47	57
EC 5	156	291	302	45	16	17
EC 6	95	688	730	27	14	14
Sum	2,882	25,320	33,263	838	275	299

II.2.2. Feature Extractions

A total of 1437 dimensional E vectors were constructed from enzyme amino acid sequence features using PROFEAT[36,37] with 7 descriptors (Table 2A). PROFEAT is a reliable system that can extract various enzyme sequence features and select multiple descriptors. These descriptors have been established in many protein sequence analysis studies[38,39,53,55,60,64,112–114]. Similarly, 1,387 dimensional S and P vectors were derived from their respective chemical structural features using DRAGON (version 7.0.4)[115] with 13 descriptors (Table 2B). The DRAGON descriptors have been used to express various compound features by calculating quantitative structure-property relationships and quantitative structure-activity relationships[60,64,114,116]. The descriptors were applied to extract compound chemical structure in 2-dimensional (2D) spaces. Several studies have used both PROFEAT and DRAGON descriptors to predict drug-target interactions[60,64,114]. As illustrated in Figure 4, E vectors, SE vectors and SEP vectors

have 1,437, 2,824 and 4,211 dimensions, respectively. 3 types of vectors were extracted from individual enzymatic reactions. Test vectors were normalized based on training vectors. Moreover, in this chapter, these descriptors were evaluated for enzymatic reaction prediction.

Table 2. List of Descriptors: (A) PROFEAT, (B) DRAGON[115].

(A) PROFEAT Descriptors	Number of dimensions
Amino acid composition (AAC)	20
Dipeptide composition (DPC)	400
Autocorrelation descriptors (ACD)	270
Composition, transition and distribution (CTD)	504
Quasi-sequence-order descriptors (QSO)	160
Amphiphilic pseudo-amino acid composition (APAAC)	80
Total amino acid properties (TAAP)	3
Sum	1437

More detailed explanations of these descriptors are shown in Chapter II.6.1-1.

(B) DRAGON Descriptors	Number of dimension
Constitutional indices (CI_1)	47
Ring descriptors (RD)	32
Walk and path counts (WPC)	46
Connectivity indices (CI_2)	37
2D autocorrelations (2DA)	213
P_VSA-like descriptor (PV)	55
ETA indices (ETA)	23
Edge adjacency indices (EAI)	324
Functional group counts (FGC)	153
Atom-centred fragments (AF)	115
Atom-type E-state indices (AEI)	172
CATS 2D (C2D)	150
Molecular properties (MP)	20
Sum	1387

More detailed explanations of these descriptors are shown in Chapter II.6.1-2.

II.2.3. Machine Learning

Machine learning can rapidly learn various types of data including vectors derived from substrates, products and enzyme sequences. Multiple machine learning algorithms were employed to build enzymatic reaction prediction models for critical comparison[48,53,55,105]. Support Vector Machine (SVM), Random Forests (RF), k-Nearest Neighbor (kNN) and Multilayer Perceptron (MLP), which have been demonstrated in various biological annotation predictions, were used in this chapter. Explanations of

each method are given in the Chapter II.6.2. As illustrated in Figure 4, the E model, SE model and SEP model were built by learning E vectors, SE vectors and SEP vectors, respectively, in combination with corresponding EC number first digits. 6 types of SVM-OvR-models (e.g., EC 1 or Rest, EC 2 or Rest, …), and an SVM-Multi-model that merges 15 types of classifiers (e.g., EC 1 or EC 2, EC 1 or EC 3, …, EC 2 or EC 3, …, EC 5 or EC 6), were built to predict 6 types of EC number first digits. RF, kNN, MLP are normal Multi-models (e.g., EC 1 or EC 2 or …, EC 6). In OvR-models, posterior probability for test samples was calculated by Platt's method[117,118], which is based on the distance from the decision boundary. OvR test sample prediction classes were determined via probability thresholds. On the other hand, in Multi-models, prediction classes were determined as the class with the highest score. All machine learning models were evaluated using an *E. coli* K-12 test. Cross-validation was also included for SVM based models using One-versus-Rest (OvR) and Multiclass One-versus-One (Multi) methods.

Each model was optimized by tuning hyper-parameters. In E-SVM models, the hyper-parameters with the highest Accuracy were used for cross-validation, and in the SE, and SEP models, the same parameters were used because Accuracy was consistent throughout all cross-validation results. On the other hand, in Multi-models, the hyper-parameters with the highest Macro F_1 scores were selected for the *E. coli* K-12 test. All hyper-parameters are shown in Chapter II.6.4 (Table S1-S4). Each machine learning algorithm was used in the scikit-learn library[119]. Various parameters were used for cross-validation and the *E. coli* K-12 test, because Accuracy is not always the best metric for prediction when training and test datasets are imbalanced (Chapter II.6.3). Macro

Precision, Recall, F_1 score and AUC were used as metrics because the datasets are imbalanced in 6 EC number first digits.

II.2.4. Principal Component Analysis (PCA)

PCA was used to dimensionally compress and extract features for the SE and SEP models. PCA orthogonally projects data onto a lower dimensional linear space known as the principal subspace, resulting in maximization of projected data variance[120,121]. SE and SEP vector dimensions were compressed using PCA to decrease model building time, especially for the SVM-OvR models. Furthermore, PCA was used to identify important features of training vectors because the number of dimensions increases when adding the substrate and product information. SE- and SEP vectors were compressed into 6 types and 7 types of dimensions, respectively, and the resulting models were then compared. In addition, 30 dimensions were determined in descending order of factor loadings, up to the 10th principal component. The effect of origin vector dimensions on each principal component dimension was evaluated. Important variables for enzymatic reaction prediction were evaluated by removing descriptor dimensions throughout the 30 dimensions and 10 principal components followed by comparing prediction results with all other machine learning SEP models, not including SVM-Multi. PCA from the scikit-learn library[119] was used.

II.3. Results

II.3.1. EC Number Prediction from Amino Acid Sequence Information

Cross-validation and *E. coli* K-12 test results for E models built using only enzyme information are shown in Figure 6, and results for each EC number first digit are shown

in Tables 3 and 4. Cross-validation of SVM models indicates that the Accuracy and AUC of first-digit EC number prediction are slightly higher for EC 4, 5 and 6, which consist of smaller training datasets. In contrast, Macro Precision, Macro Recall and Macro F_1 scores are lower than Accuracy and AUC. Regarding test results of SVM prediction, although Average Accuracies are over 0.75, other statistics are lower than those from cross-validation. The Accuracy of first digit prediction is still slightly high for EC 4, 5 and 6, and much lower for EC 2 and 3. Macro Precision, Macro Recall and Macro F_1 scores of SVM-multi are lower than those from SVM-OvR. Other machine learning results are similar to SVM results (Figure 6C). Table S5 shows correct prediction rates using BLAST and E models with the *E. coli* K-12 test. BLAST prediction is regarded as correct if the EC number first digit of a training sample matches the corresponding test sample. Overall, correct prediction was lower with the E models relative to that of the BLAST method, although the E model can predict some functions that cannot be inferred using BLAST alone.

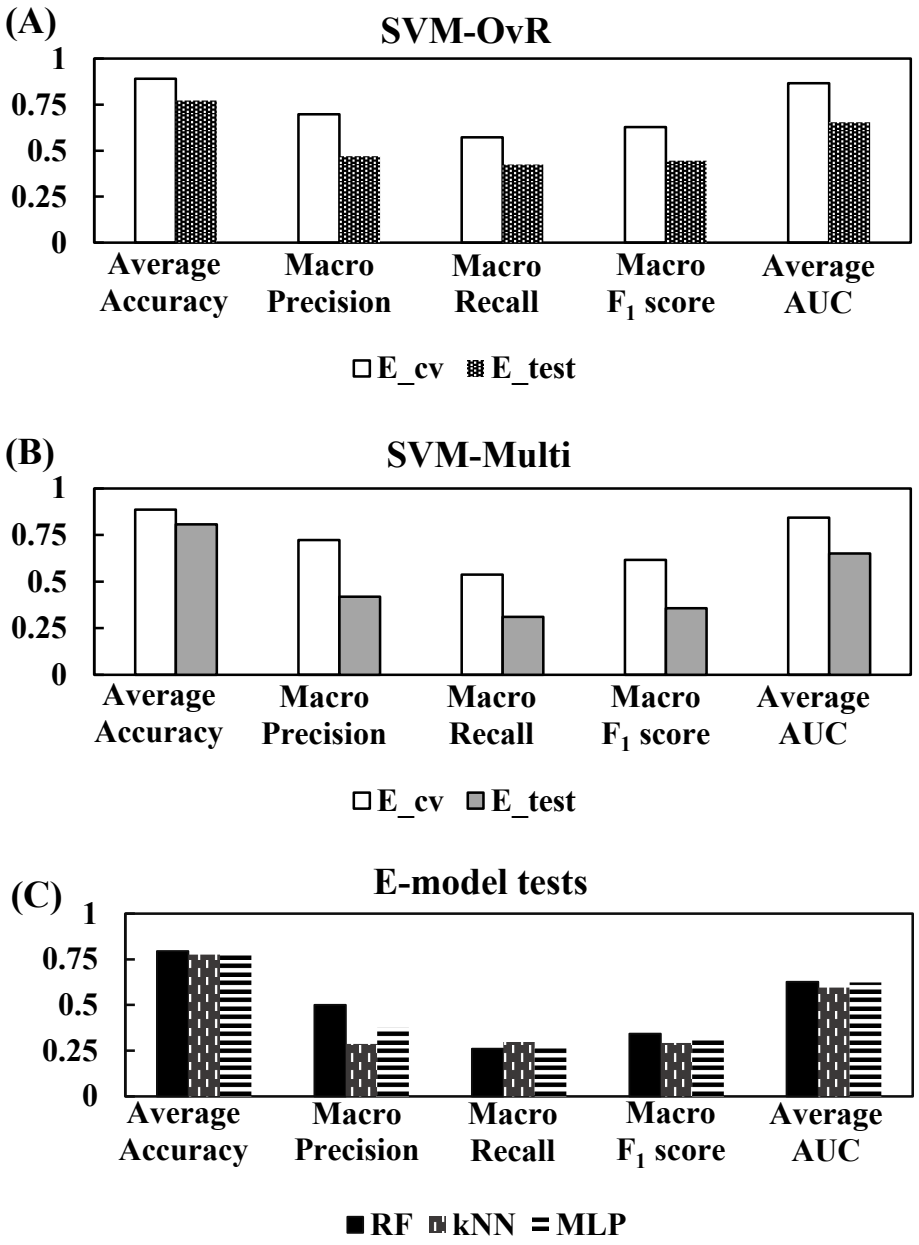


Figure 6. Cross-validation (cv) and *E. coli* K-12 test results for E models: (A) SVM-OvR, (B) SVM-Multi, and (C) RF, kNN, and MLP tests.

Table 3. Cross-Validation (CV) and *E. coli* K-12 Test (Test) Results for Each EC Number First Digit in E-SVM Models: (A) OvR and (B) Multi.

(A)	E-SVM-OvR model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.853	0.858	0.758	0.552	0.639
EC 2	0.792	0.851	0.560	0.755	0.643
EC 3	0.847	0.904	0.768	0.779	0.773
EC 4	0.914	0.846	0.555	0.445	0.494
EC 5	0.955	0.837	0.625	0.417	0.500
EC 6	0.982	0.903	0.920	0.484	0.634

(A)	E-SVM-OvR model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.741	0.676	0.360	0.490	0.415
EC 2	0.481	0.576	0.346	0.758	0.475
EC 3	0.563	0.672	0.433	0.832	0.569
EC 4	0.942	0.682	0.846	0.190	0.310
EC 5	0.946	0.613	0.500	0.0890	0.151
EC 6	0.962	0.702	0.333	0.185	0.238

(B)	E-SVM-Multi model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.833	0.835	0.645	0.649	0.647
EC 2	0.808	0.832	0.600	0.674	0.635
EC 3	0.818	0.884	0.689	0.829	0.753
EC 4	0.918	0.813	0.639	0.313	0.420
EC 5	0.961	0.787	0.906	0.308	0.459
EC 6	0.980	0.908	0.860	0.453	0.593

(B)	E-SVM-Multi model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.759	0.661	0.380	0.452	0.413
EC 2	0.621	0.602	0.389	0.392	0.391
EC 3	0.636	0.648	0.479	0.540	0.507
EC 4	0.913	0.674	0.326	0.241	0.277
EC 5	0.943	0.643	0.364	0.0890	0.143
EC 6	0.969	0.678	0.571	0.148	0.235

Table 4. *E. coli* K-12 Test Results for Each EC Number First Digit in the Machine Learning E, SE, and SEP Models. (A) RF, (B) kNN and (C) MLP.

(A)	Accuracy			AUC		
	E	SE	SEP	E	SE	SEP
EC 1	0.690	0.789	0.886	0.644	0.838	0.936
EC 2	0.600	0.724	0.843	0.566	0.800	0.941
EC 3	0.626	0.804	0.906	0.620	0.885	0.900
EC 4	0.940	0.898	0.930	0.650	0.865	0.945
EC 5	0.945	0.956	0.963	0.558	0.908	0.958
EC 6	0.969	0.956	0.970	0.718	0.956	0.976

(A)	Precision			Recall			F_1 score		
	E	SE	SEP	E	SE	SEP	E	SE	SEP
E C1	0.290	0.470	0.662	0.452	0.736	0.804	0.353	0.574	0.726
EC 2	0.349	0.747	0.812	0.335	0.562	0.831	0.342	0.642	0.821
EC 3	0.466	0.292	0.463	0.526	0.875	0.760	0.494	0.438	0.576
EC 4	0.900	0.913	0.950	0.155	0.447	0.667	0.265	0.600	0.784
EC 5	0.333	1.00	0.800	0.0220	0.250	0.471	0.042	0.400	0.593
EC 6	0.667	1.00	0.857	0.074	0.143	0.429	0.133	0.250	0.571

(B)	Accuracy			AUC		
	E	SE	SEP	E	SE	SEP
EC 1	0.696	0.836	0.906	0.665	0.784	0.887
EC 2	0.626	0.731	0.836	0.526	0.721	0.828
EC 3	0.650	0.913	0.916	0.578	0.801	0.791
EC 4	0.827	0.869	0.893	0.626	0.769	0.800
EC 5	0.905	0.920	0.946	0.541	0.635	0.778
EC 6	0.954	0.960	0.980	0.638	0.675	0.888

(B)	Precision			Recall			F_1 score		
	E	SE	SEP	E	SE	SEP	E	SE	SEP
EC 1	0.298	0.561	0.706	0.459	0.698	0.857	0.361	0.622	0.774
EC 2	0.385	0.720	0.846	0.342	0.636	0.762	0.363	0.675	0.802
EC 3	0.494	0.500	0.500	0.282	0.667	0.640	0.359	0.571	0.561
EC 4	0.163	0.617	0.755	0.362	0.617	0.649	0.225	0.617	0.698
EC 5	0.211	0.313	0.526	0.156	0.313	0.588	0.149	0.313	0.556
EC 6	0.238	0.714	0.786	0.185	0.357	0.786	0.208	0.476	0.786

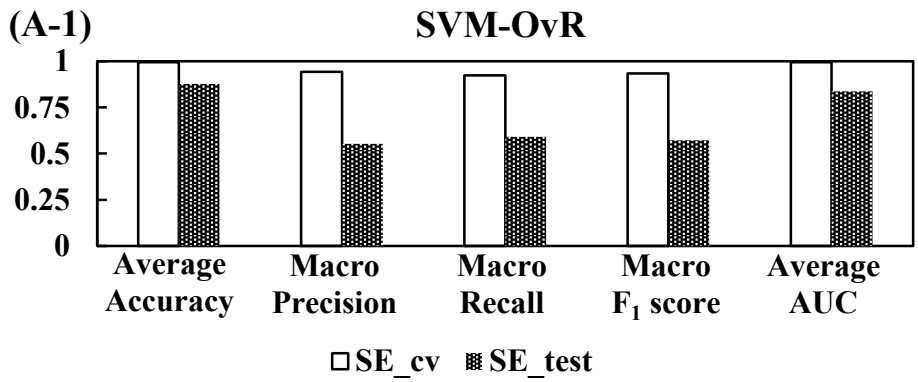
(C)	Accuracy			AUC		
	E	SE	SEP	E	SE	SEP
EC 1	0.732	0.775	0.880	0.624	0.814	0.936
EC 2	0.582	0.709	0.793	0.565	0.728	0.854
EC 3	0.636	0.909	0.910	0.598	0.829	0.896
EC 4	0.878	0.836	0.870	0.619	0.820	0.900
EC 5	0.938	0.920	0.940	0.638	0.788	0.814
EC 6	0.969	0.956	0.967	0.691	0.747	0.910

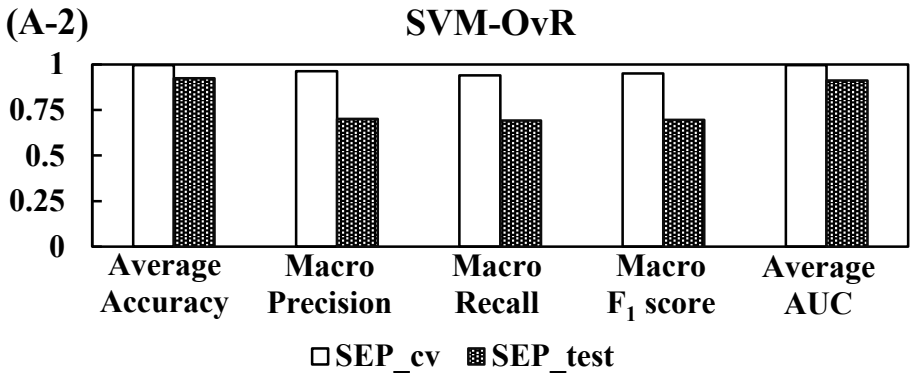
(C)	Precision			Recall			F_1 score		
	E	SE	SEP	E	SE	SEP	E	SE	SEP
EC 1	0.311	0.444	0.639	0.357	0.679	0.821	0.332	0.537	0.719
EC 2	0.372	0.711	0.793	0.504	0.570	0.708	0.428	0.633	0.748
EC 3	0.468	0.484	0.476	0.351	0.625	0.800	0.401	0.545	0.597
EC 4	0.194	0.521	0.673	0.241	0.532	0.614	0.215	0.526	0.642
EC 5	0.231	0.313	0.462	0.0670	0.313	0.353	0.103	0.313	0.400
EC 6	0.667	1.00	1.00	0.0740	0.143	0.286	0.133	0.250	0.444

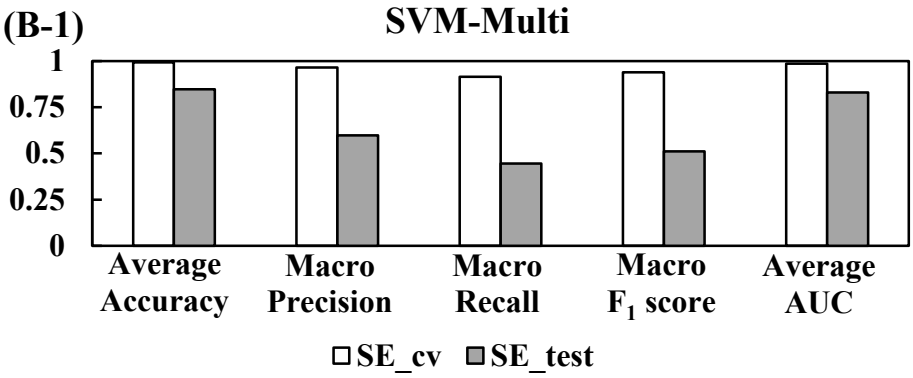
II.3.2. Enzymatic Reaction Prediction Using Substrate and Product Chemical Structural Information in Addition to Sequence Information

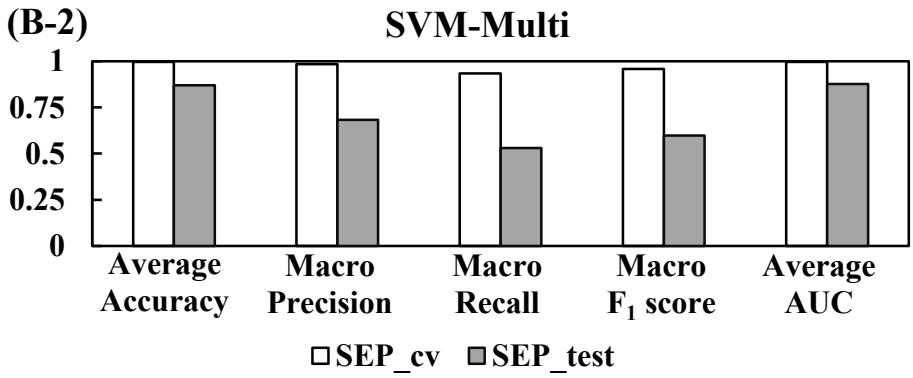
Cross-validation and *E. coli* K-12 test results for all SE and SEP models built using enzyme and compound information are illustrated in Figure 7. Prediction of EC number first digits from enzyme and compound information is shown in Table 4 and Table S6. Cross-validation of the SVM based SE and SEP models results in near-perfect scores,

with higher values for SEP models compared to SE models. Test results for all machine learning models are best for SEP models, second best for SE models, with lower prediction results for the E models. Overall, these SE and SEP results are also lower than the corresponding cross-validation metrics. Here, SVM-Multi test results for Macro Precision, Macro Recall, and Macro F_1 score decrease more than that of SVM-OvR. SEP model results for all machine learning methods are shown in Figure 8. For SEP-RF models, Average Accuracy, Macro F_1 score, and Average AUC are slightly higher than those of other methods. In comparison to BLAST, correct prediction was higher for SE models with the exception of SVM-OvR and MLP, and also higher for all SEP models (Figure 9 and Table S5). The SEP models correctly predict test samples with high E-values as well as those that cannot be determined using BLAST.




(A-2)
SVM-OvR
1
0.75
0.5
0.25
0
Average Accuracy
Macro Precision
Macro Recall
Macro F_1 score
Average AUC
SEP_cv
SEP_test


(B-1)
SVM-Multi
1
0.75
0.5
0.25
0
Average Accuracy
Macro Precision
Macro Recall
Macro F_1 score
Average AUC
SE_cv
SE_test


(B-2)
SVM-Multi
1
0.75
0.5
0.25
0
Average Accuracy
Macro Precision
Macro Recall
Macro F_1 score
Average AUC
SEP_cv
SEP_test

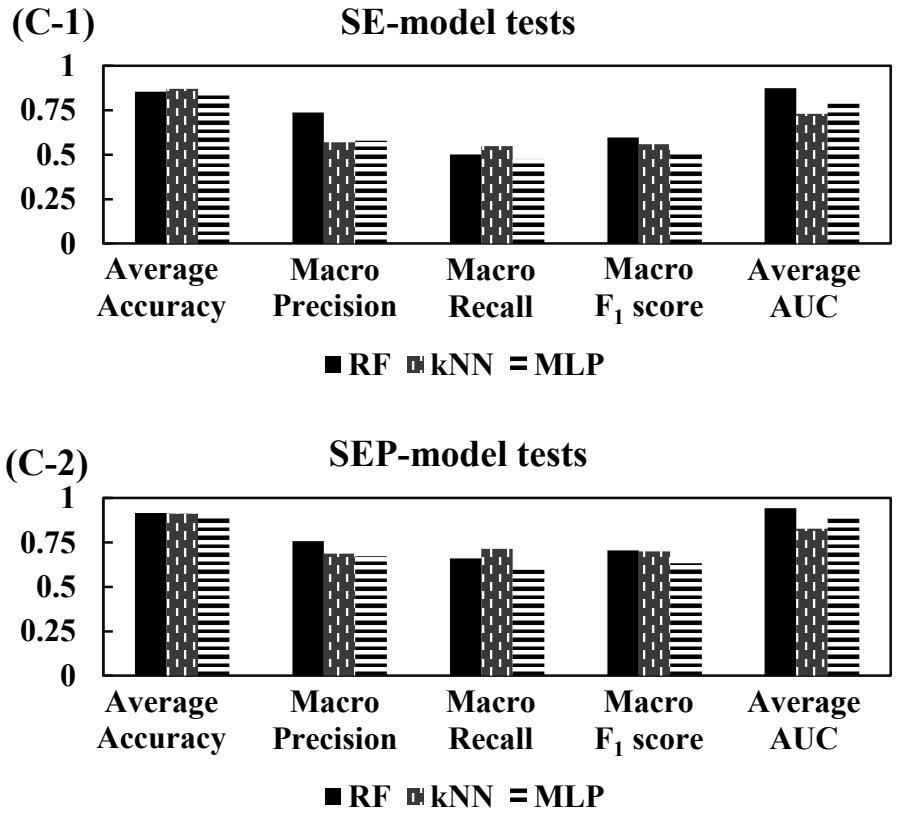


Figure 7. Cross-validation and *E. coli* K-12 test results for SE and SEP models: (A) SE-SVM-OvR (1) and SEP-SVM-OvR models (2), (B) SE-SVM-Multi (1) and SEP-SVM-Multi models (2), and (C) SE (1) and SEP models (2) using RF, kNN and MLP tests.

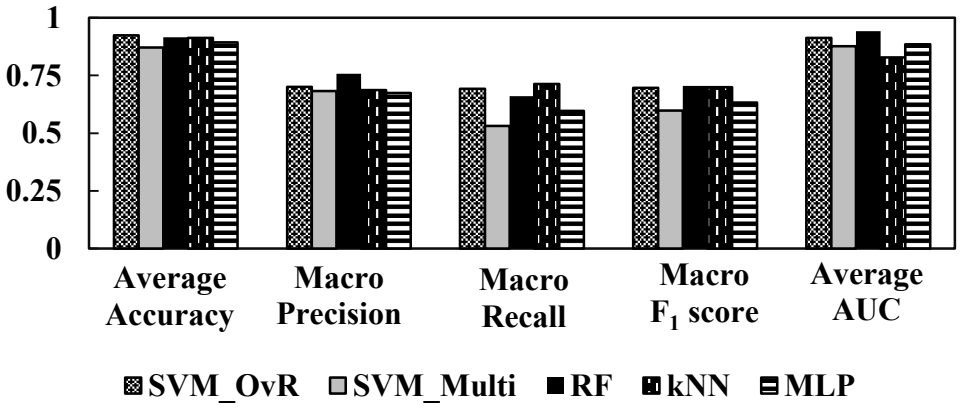


Figure 8. *E. coli* K-12 test results for all SEP models.

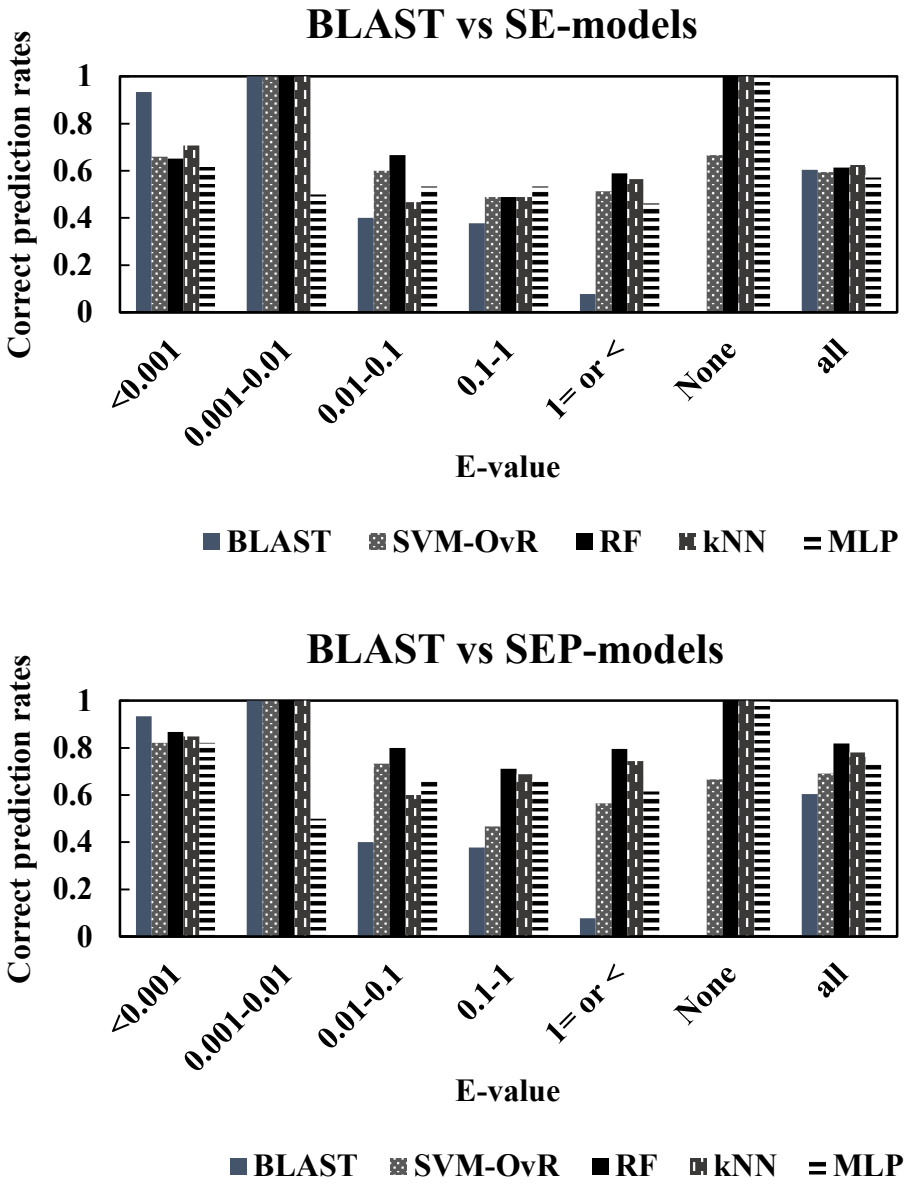


Figure 9. Performance comparison of BLAST with SE and SEP models with the *E. coli* K-12 test: (A) SE models and (B) SEP models. Abbreviations: None, test samples not predicted by BLAST; all, all test samples.

II.3.3. Evaluation of Important Factors for Training Datasets and Prediction

PCA is performed to shorten building times for SE-SVM-OvR and SEP-SVM-OvR models and to identify important factors for training datasets. Models are first built

using compressed vectors followed by evaluation of test results based on the number of dimensions. Macro F_1 score and Average AUC are calculated from the *E. coli* K-12 test as shown in Figure 10. Here, prediction is stable at 100 dimensions and higher, but accuracy decreases in 10 dimension models. Furthermore, prediction using origin SE (2,824 dimension) and SEP models (4,211 dimension) is slightly lower than that of PCA analysis.

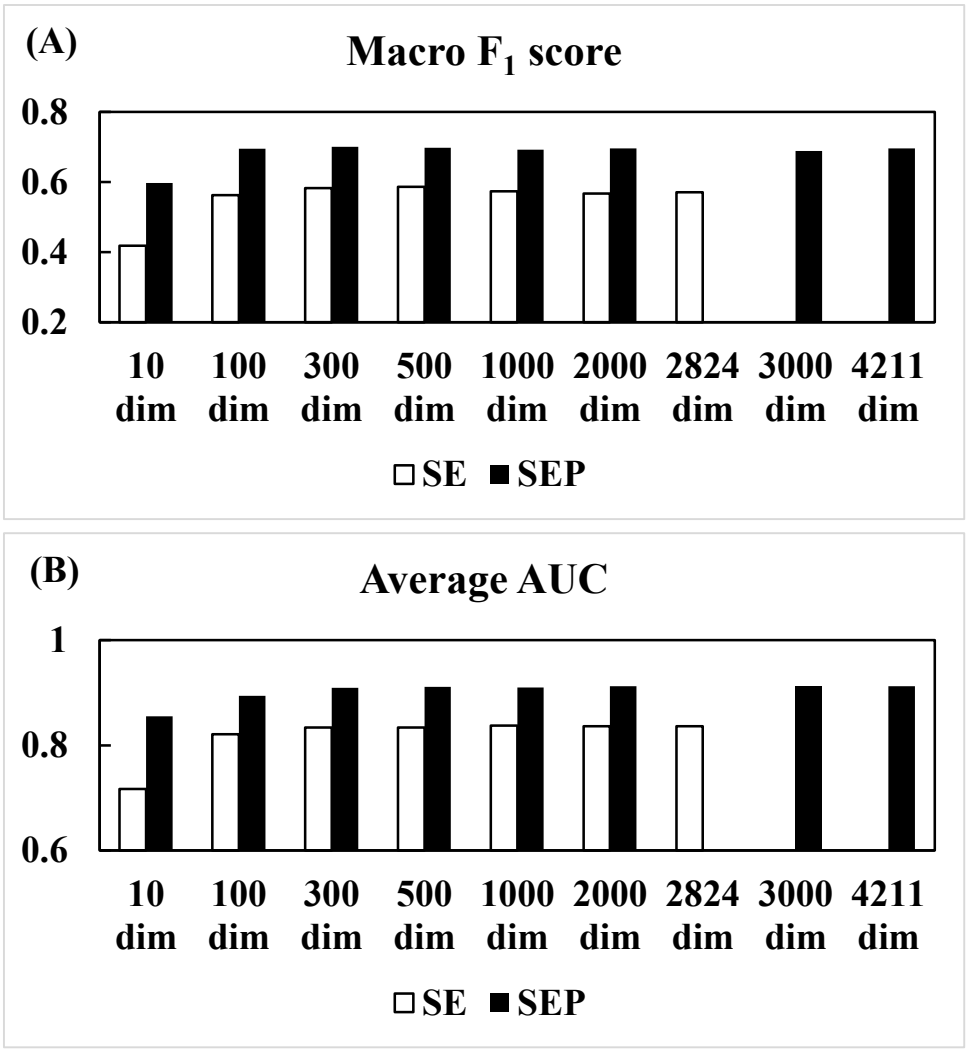


Figure 10. *E. coli* K-12 test of the compressed SEP-SVM-OvR model: (A) Macro F_1 score and (B) Average AUC.

Next, important training factors are examined. Proportions of variance up to 20 principal components and 30 dimensions across 10 principal components in SEP vectors are shown in Figure 11. Proportions of variance above 20 principal components are not used due to a gradual decrease around the 10th principal component up to the 20th principal component. For similar reasons, factor loadings are also evaluated in this range. Variations in the dimensions of each principal component during a high factor loading are observed. Factor loading in substrate and product dimensions is high with the exception of the 9th principal component, while enzyme factor loadings are high in the 9th and 10th principal components. Despite factor loading differences, all dimensions include the same descriptor ranges. Information in unique principal components 1 and 9 are shown in Figure 12. A higher compound factor loading is observed in the dimensions of Edge adjacency indices, Walk and Path counts, and 2D autocorrelation descriptors in the first few components (Table 2A). Moreover, in only the 2nd and 3rd components, Ring descriptor dimensions contained a higher factor loading. In contrast, a higher enzyme factor loading is observed in the C and T dimensions of Composition, Transition, and Distribution (CTD) descriptors in the 9th and 10th components (Table 2B). Here, details for these descriptors are omitted.

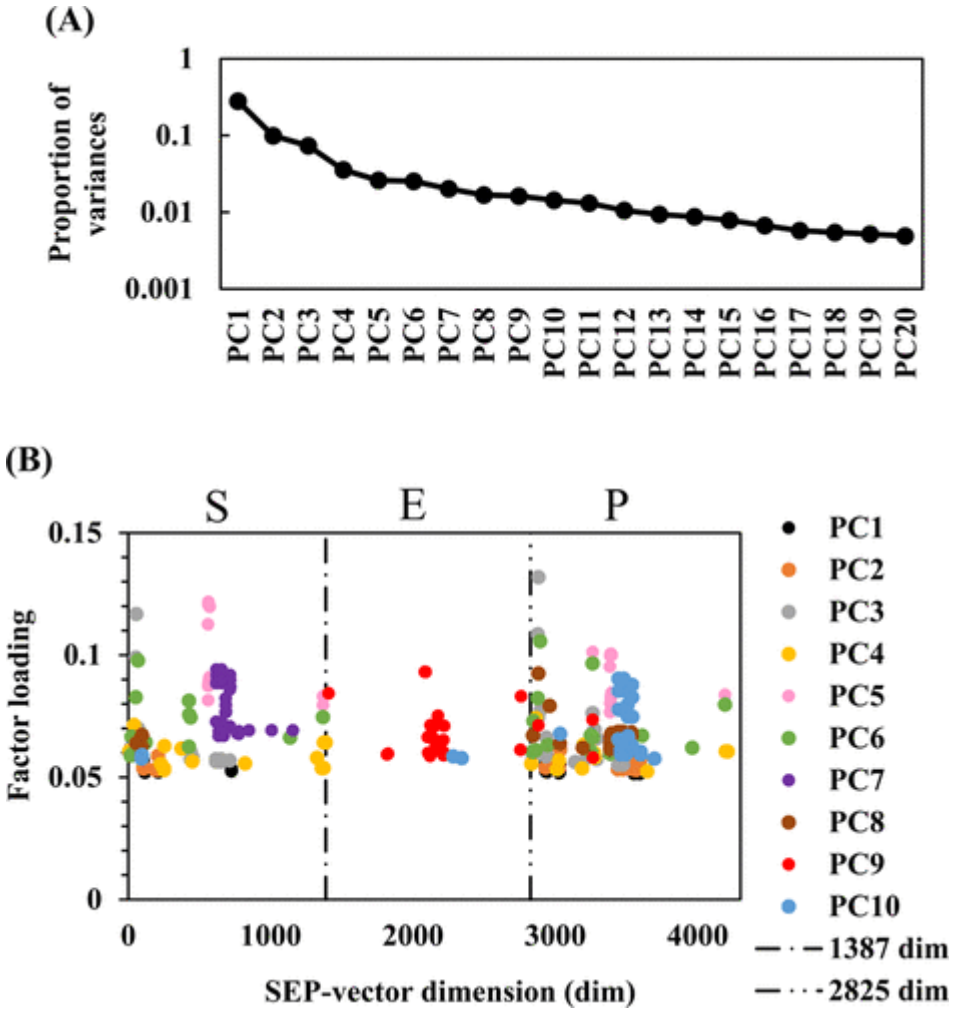


Figure 11. (A) Proportion of variances up to the 20th principal component. (B) Absolute values of loading up to the 10th principal component. 30 dimensions in descending order of absolute values for each principal component are shown.

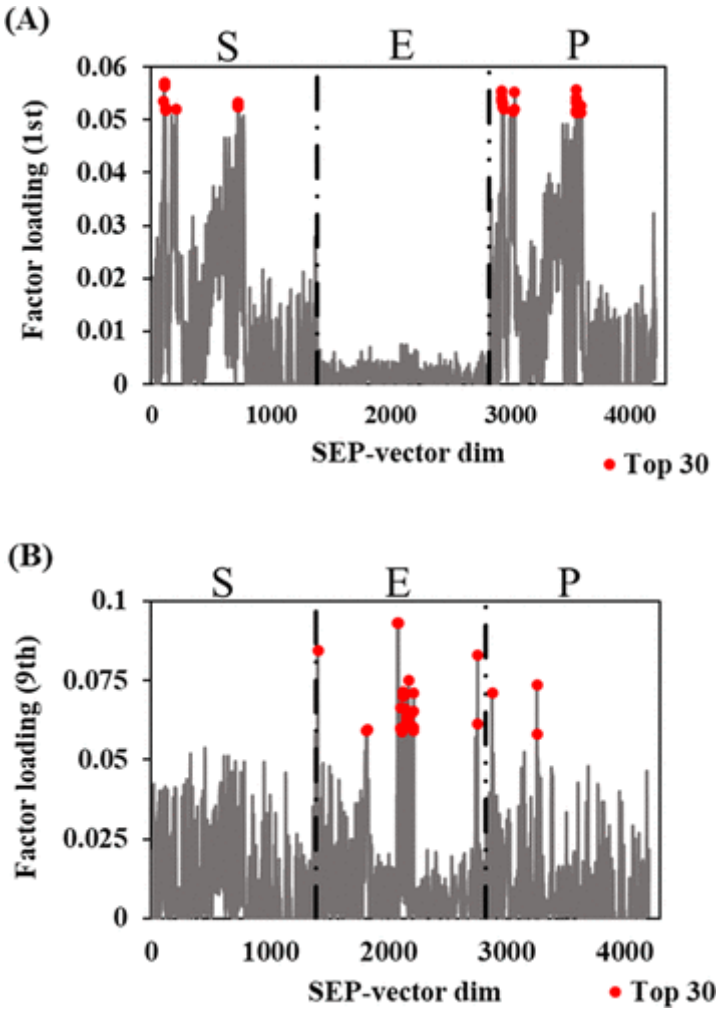


Figure 12. Absolute values for factor loadings in the (A) 1st principal component and (B) 9th principal component.

Several dimensions of SEP vectors are then reduced in the range of these 4 descriptors to examine important factors for enzymatic reaction prediction. 6 combinations of SEP models are used with all machine learning methods except SVM-Multi. 10 DRAGON and 6 PROFEAT descriptors were used to build all SEP models (Table 2). As shown in Figure 13, prediction varies across each method with the exception of RF.

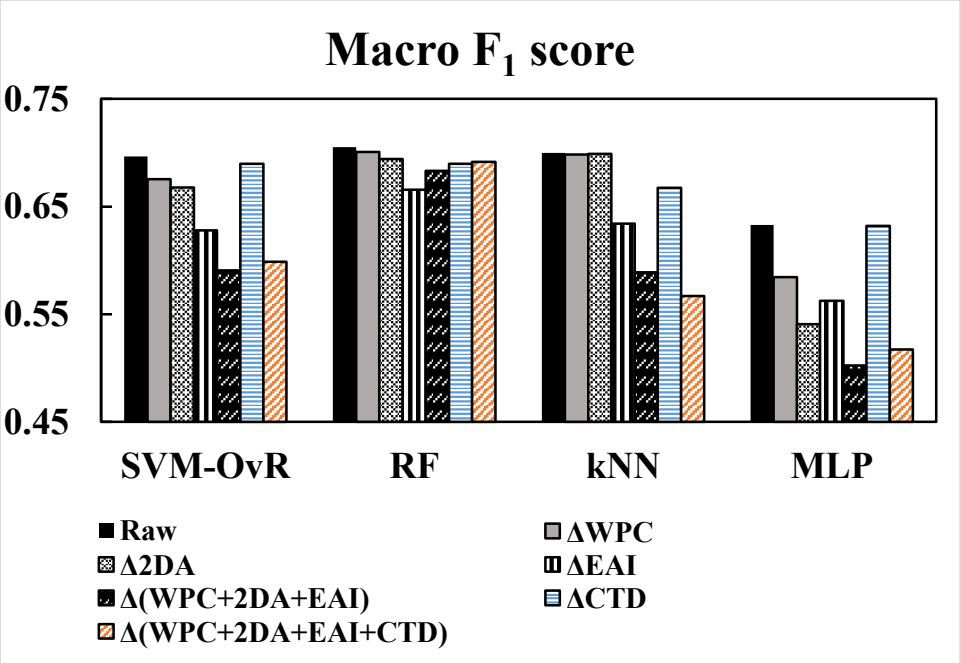


Figure 13. *E. coli* K-12 test Macro F_1 scores for dimensionally reduced SEP models. Abbreviations: raw, 4211 dimensions SEP vector; Δ, deleted; WPC, DRAGON Walk and Path counts descriptor;2DA, DRAGON 2D Autocorrelation descriptor; EAI, DRAGON Edge adjacency indices descriptor; CTD, PROFEAT composition, transition, and distribution descriptor.

II.3.4. SEP Model Prediction of Specific Reaction Types in the *E. coli* K-12 Test

When comparing SEP model prediction for specific reaction types, results for the SVM-OvR models appear less accurate (Figure 14 and Figure S2). Test predictions are regarded as incorrect if a reaction scores high probability in an incorrect EC first digit group, even if the correct group is also included. Misjudged reactions occurred least with the RF model followed closely by the kNN model. For 26 oxidation/reduction reactions with hydrocarbons, alcohols, aldehydes, and ketones (EC 1.1.X.X, EC 1.2.X.X, EC 1.3.X.X), prediction is optimal with all methods. For SVM-OvR and RF prediction of 3 EC 1.1.1.42 known reactions, only the reaction from isocitrate to

oxalosuccinate is not predicted. However, production of 2-oxoglutarate from isocitrate or oxalosuccinate is correctly predicted using all machine learning methods. Oxidation/reduction reactions of simple alcohols, aldehydes, and ketones, including those producing a single proton, tend to be correctly predicted. Next, 55 transfer reactions with acyl- and phosphorus-containing groups (EC 2.3.X.X and EC 2.7.1.X) are also correctly predicted by all methods. For example, EC 2.3.1.241, EC 2.3.1.242, and EC 2.3.1.243 reactions are predicted by all models, even though the substrate and product structures are large and complex. Test results for EC 2.7.1.202 reactions varied among the 4 methods, where these reactions are catalyzed by 8 enzymes in test datasets. RF and kNN methods can predict all EC 2.7.1.202 reactions, while SVM-OvR can correctly predict 5 of 8 sequences. Among the 3 enzyme sequences in this group that could not be predicted by SVM-OvR, there is high sequence identity. Moreover, reactions of 16 glycosyl transferases (EC 2.4.1.X and EC 2.4.2.X) are not predicted by SVM and MLP models, but RF and kNN predict about 80% of these reactions.

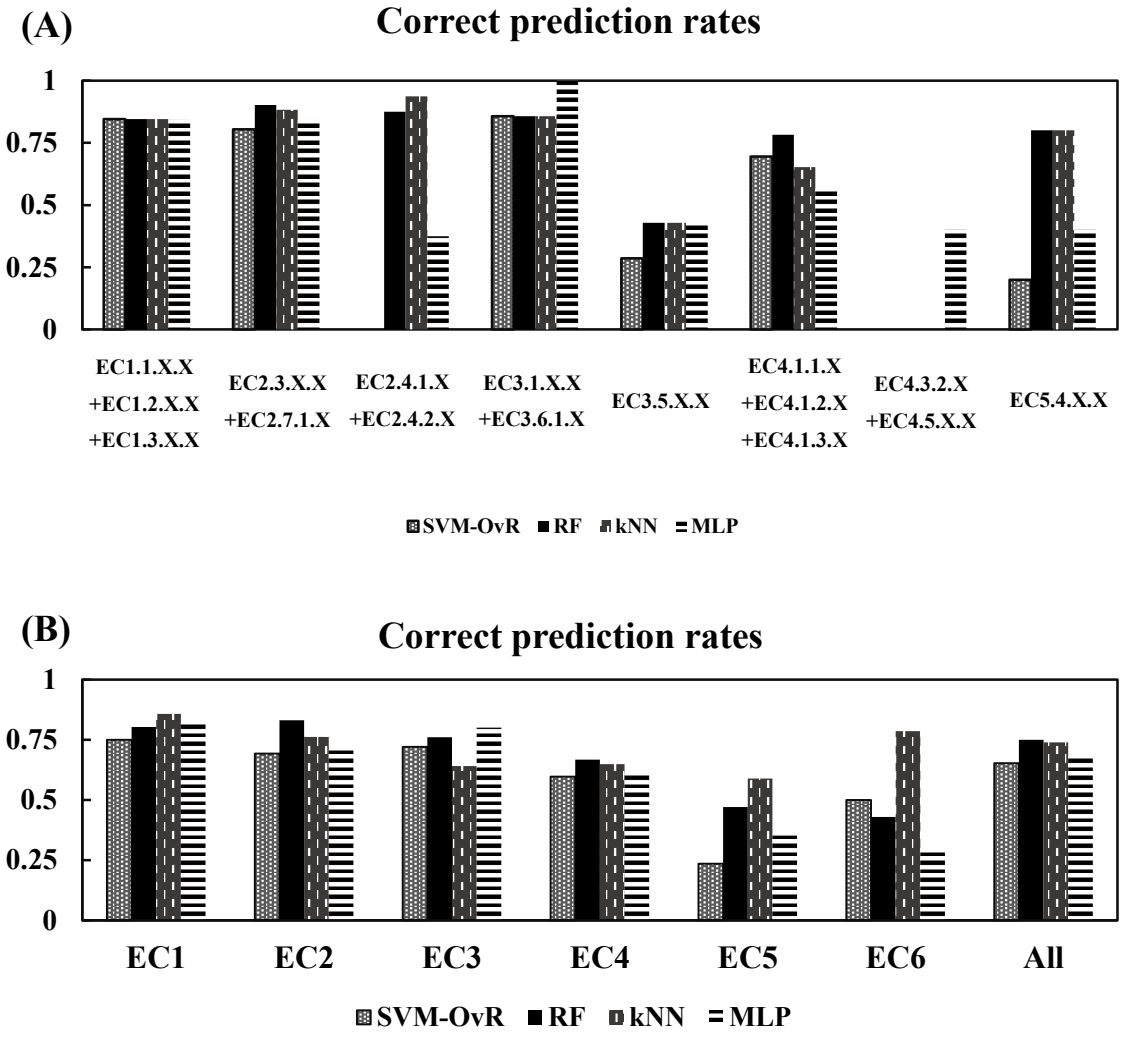


Figure 14. Performance comparison of (A) EC 2–3 digit groups and (B) EC first digits of 299 reactions tests using the SEP models.

Although 14 hydrolysis reactions with carboxylic acid esters and acid anhydrides (EC 3.1.X.X and EC 3.6.X.X) are well predicted, 7 reactions with carbon–nitrogen bonds, with the exception of peptide bonds (EC 3.5.X.X), are not predicted by the current models. For example, 5 EC 3.1.3.X reactions that involve hydrolysis of phosphate monoesters are predicted by all methods. However, the EC 3.1.4.52 hydrolysis of phosphate diester is only predicted using MLP. Next, 23 elimination–addition reactions

with carboxy, aldehyde and oxo-acid groups (EC 4.1.1.X, EC 4.1.2.X and EC 4.1.3.X) are correctly predicted. In contrast, amidine and carbon-halide lyase reactions (EC 4.3.2.X and EC 4.5.X.X) are unable to be predicted. In EC 4, most reactions in which large structures are eliminated from substrates tend not to be predicted. 17 relatively simple isomerization reactions are incorrectly predicted most of the time. For 5 intramolecular transfer reactions (EC 5.4.X.X), 80% are correctly predicted by RF and kNN models, whereas other models fail. Complex EC 6 reactions including cyclization are commonly misjudged. The kNN model is best for EC 5 and EC 6 reactions, in which the number of test datasets was small, while the RF model predicted the most test reactions overall.

II.4. Discussion

II.4.1. Model Evaluations

II.4.1-1. EC Number Prediction Using Amino Acid Sequence Information

Cross-validation of E models results in the best Average AUC and Average Accuracy when using SVM methods. The Accuracy for smaller EC first digit groups 4 to 6 is better than that of the larger EC first digit groups 1 to 3. Positive prediction is worse than negative prediction for EC 4 to 6 because Macro Precision, Macro Recall and Macro F_1 score are low (Figure 6, Table 3). Therefore, a higher number of negative samples in EC groups 4 to 6 resulted in high negative prediction and better overall Accuracy. *E. coli* K-12 tests with all machine learning E models resulted in much lower scores relative to that of cross-validation, especially regarding Macro Precision, Macro Recall and Macro F_1 score. Test results further indicate that the models can classify negative samples much better than positive samples. Moreover, with enzyme sequence

information alone, all machine learning results were less effective than BLAST results. Therefore, E models require further improvements to better match the EC number prediction of Li *et al.*[48], Zou *et al.*[49], Shen *et al.*[55]. However, with different training and test data across each study, it is difficult to compare results from various reports. Moreover, annotation of more EC 4 to 6 examples is needed to improve positive prediction of these enzymes.

II.4.1-2. Enzyme Prediction Using Chemical Structural Information

Cross-validation of SE-SVM and SEP-SVM models produces near-perfect performance parameters. Although *E. coli* K-12 test results are expectedly lower than that of cross-validation, addition of chemical structure information improves prediction relative to models with only enzyme information. However, the cross-validation results indicate overfitting to training datasets. Therefore, improvements in training dataset size, selection of chemical structural information, and regularization are needed in order to prevent overfitting. Furthermore, SVM-OvR methods should be better than SVM-Multi methods for enzymatic reaction prediction as indicated by all parameters except Average Accuracy and Macro Precision. Other machine learning methods improved when including chemical structure information to make SE and SEP models. This further emphasizes the importance of increasing dataset sizes and information for optimizing machine learning prediction. Of all the methods tested, the SEP-RF model was the most accurate in this chapter. Because RF is an ensemble method and learns with feature selection, RF models resulted in less overfitting compared to the other machine learning methods[114]. Regarding the comparison of machine learning prediction with BLAST results, the SE and SEP models can correctly predict more reactions for

enzyme sequences with low similarity or no similarity to database results. However, BLAST can nearly predict more reactions for enzymes with many highly similar and well-annotated sequencers.

II.4.2. Important Influencers in Training Datasets and Prediction

Consistent prediction accuracy in SVM-OvR SE and SEP models after dimensionality compression enables decreased building time. However, unimportant factors for reaction predictions may exist in origin vectors, leading to suboptimal feature extraction and lower prediction accuracy. In addition, dimensionality compression can minimize overfitting in SE- and SEP models because both models performed better in compressed tests of 300 and 500 dimensions (Figure 10). The evaluation of key training factors and factor loading in each dimension suggests that chemical structural information is more important than sequence information. This conclusion is consistent with the lower prediction of E models.

Results for up to 10 principal components indicate that the most important factors for feature extraction are 4 descriptors: DRAGON 2D autocorrelations descriptors, DRAGON Walk and Path counts descriptor, DRAGON Edge adjacency indices descriptor in first a few components and PROFEAT CTD descriptor in the 9th and 10th components (Table 2, Figure 11B, Figure 12). Moreover, Ring descriptors are somewhat important, but less critical than the best 4 descriptors. CTD descriptors were used in Li *et al.*[55] and should be important for protein sequence annotation prediction. However, reduction of these descriptors in the RF SEP model, including removal of 4 descriptors, results in no loss of accuracy (Figure 13). RF builds prediction models using Bootstrap

Sampling and random feature extraction to minimize cross-entropy error. Accordingly, the reduction of RF model dimensions has no significant influence on the consistency of results. Building machine learning models using RF is also very effective due to less strict feature extraction. While accurate SVM-OvR prediction requires the combination of 6 models; RF, kNN and MLP prediction can be directly determined.

The Walk and Path counts (WPC), 2D autocorrelations (2DA) and Edge adjacency indices (EAI) descriptors related to chemical structure in SEP vectors are important factors for accuracy of SVM-OvR, kNN and MLP models (Figure 13). However, for these 3 models of this Chapter II, which include all vector features, the CTD descriptor is not important for enzyme feature extraction. It is therefore implied that important feature extractions for enzyme prediction depend on the specific training dataset.

In deep learning models by Li *et al.*[48], Zou *et al.*[49], prediction results depended on the type of descriptors. These papers reported high prediction accuracy when including sequence motif information from Pfam[122]. Therefore, the current models in Chapter II may also benefit from addition of motif information. However, there might be no single most important shared factor for reaction prediction between various machine learning methods (Figure 13). Accordingly, results vary among machine learning models depending on reaction type, and multiple models should be carefully compared.

II.4.3. SEP Model Prediction of Specific Reaction Types in the *E. coli* K-12 Test

E. coli K-12 test results demonstrate that the SEP-RF model can often predict more reactions than the other models, although it is not good at predicting EC 5 and 6

reactions, which provide less training data. The SEP-kNN model can predict more EC 5 and 6 reactions relative to the RF model. These results emphasize that certain reaction types are more difficult to predict than others, depending on the machine learning method, especially for EC 2.4.1.X, EC 2.4.2.X, EC 4.3.2.X, EC 4.5.X.X and EC 5.4.X.X (Figure 14 A). Therefore, differences in substrate, product, or enzyme sequence within the same EC number group can influence prediction, and comprehensive methods should be developed to cover all reactions. Additional tests covering enzyme sequences from other model species are also needed to further develop prediction methods. Yet, 11 enzymes with multiple catalytic functions are predicted with above 50% accuracy by SEP models. These methods therefore demonstrate potential to predict multiple reactions for a single enzyme sequence.

II.4.4. Comparison with Other Studies

Li *et al.*[48] and Shen *et al.*[55] have proposed EC number prediction methods that can predict a single EC number for a single enzyme. The multiple models from these studies utilize stepwise prediction for second and third EC digits. This strategy may be challenging because the number of examples decreases greatly in higher level EC digits leading to imbalance in datasets. The SE and SEP models are advantageous because they predict detailed reactions in addition to EC numbers. Furthermore, SEP-RF model predicts reactions with an Average AUC score of over 0.94 using a single model in a single step.

II.5. Conclusion

In summary, this chapter demonstrates classical machine learning models that integrate structural information of substrates and products, in addition to enzyme sequence information, are effective predictors of enzymatic reactions. To improve the models further, more extensive testing using enzyme and isozyme sequences from additional species is needed because the models are evaluated using only *E. coli* K-12 test. Furthermore, SE and SEP models can be improved by optimizing feature extractions and increasing the variations of training datasets. This Chapter II study is expected to result in the discovery of new enzymes with novel functions, existing enzymes that may react with new substrates and unknown combinations of substrates-enzymes-products that can expand current metabolic pathways in the future.

II.6. Supplementary Information

II.6.1. Feature Extraction

II.6.1-1. PROFEAT Descriptors[36,37]

- Amino acid composition

The amino acid composition is the fraction of each amino acid type within a protein.

- Dipeptide composition

The dipeptide composition is the fraction of each dipeptide type within a protein.

- Autocorrelation descriptors

The descriptors describe how a considered property is distributed along an amino acid[123–125]. The amino acid properties used here are various types of amino acids index[126,127].

- Composition, transition and distribution[128,129]

The amino acids are divided in 3 classes according to its properties and each amino acid is encoded by one of the indices 1, 2 and 3 according to which class it belonged. The properties include hydrophobicity, normalized van der Waals volume, polarity, and polarizability.

First, composition descriptor is the global percent for each encoded class (1, 2, 3) in the sequence. Next, transition descriptor from class 1 to 2 is the percent frequency with which 1 is followed by 2 or 2 is followed by 1 in the encoded sequence. The descriptors from class 1 to 3 and class 2 to 3 are calculated using the same way. Last, distribution descriptor describes the distribution of each attribute in the sequence. There are 5 distribution descriptors for each attribute and they are the position percent in the whole sequence for the first residue, 25% residues, 50% residues, 75% residues and 100% residues, respectively, for a specified encoded class. For example, a sequence is encoded as “32132223311311222222” according to a property. The positions for the first residue “2”, the 2nd residue “2” (25%*10=2), the 5th “2” residue (50%*10=5), the 7th “2” (75%*10=7) and the 10th residue “2” (100%*10) in the encoded sequence are 2, 5, 15, 17,20 respectively, so the distribution descriptors for “2” are: 10.0 (2/20*100), 25.0 (5/20*100), 75.0 (15/20*100), 85.0 (17/20*100), 100.0 (20/20*100), respectively.

- Quasi-sequence-order descriptors

The quasi-sequence-order descriptors are proposed by Chou[130]. They are derived from the distance matrix between the 20 amino acids. The descriptors enable to consider sequence order effects in addition to each amino acid composition. More detailed explanations of the calculation using descriptors are shown in Chou[130].

- Amphiphilic pseudo-amino acid composition

This descriptor called type 2 pseudo-amino acid composition and is proposed by Chou[131]. The descriptor reflects the sequence-order correlations between all the contiguous residues along a protein chain through hydrophobicity and hydrophilicity in addition to each amino acid composition. More detailed explanations of the calculation using descriptors are shown in Chou[131].

- Total amino acid properties

The descriptor (TAAP) for a property P is defined as follows[113]:

$$TAAP(P) = \sum_{i=1}^{n} P_i$$

where P_i is the property of ith amino acid and n is the number of the amino acid in a sequence.

II.6.1-2. DRAGON Descriptors[115]

- Constitutional indices

Dragon calculates 47 constitutional descriptors, many of them are well explained by their definition such as the molecular weight, number of atoms and number of bonds.

- Ring descriptors

The 32 descriptors are numerical quantities encoding information about the presence of rings in a molecule.

- Walk and path counts

Walk and path counts are topological indices based on the counting of paths, walks and self-returning walks in an H-depleted molecular graph. Topological indices are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity, and can also encode chemical information concerning atom types and bond multiplicity[132]. A walk in a molecular graph is a sequence of pairwise adjacent edges leading from one vertex to another one; any edge can be traversed several times. A path is a walk without any repeated vertex or edge. The walk or path length is the number of edges traversed by the walk or path.

- Connectivity indices

Connectivity indices[133] are among the most popular topological indices. They are calculated from the H-depleted molecular graph where each vertex (non-hydrogen atom) is weighted by the vertex degree, that is, the number of connected non-hydrogen atoms.

- 2D autocorrelations

2D autocorrelations are molecular descriptors which describe how a considered property is distributed along a topological molecular structure[123–125,134]. The atomic properties used to weight molecular graphs are as follows:

▪Carbon-scaled atomic mass

▪Carbon-scaled atomic van der Waals volume

▪Carbon-scaled atomic Sanderson electronegativity

▪Carbon-scaled atomic polarizability

▪Carbon-scaled atomic ionization potential

▪Intrinsic state

- P_VSA-like descriptor

These are molecular descriptors defined as the amount of van der Waals surface area having a property P in a certain range[135]. The properties are as follows:

▪LogP

▪Molar refractivity

▪Mass

▪Van der Waals volume

▪Sanderson electronegativity

▪Polarizability

▪Ionization potential

▪Intrinsic state

- ETA indices

ETA indices are topological indices derived from the H-depleted molecular graph where a vertex is considered to be comprised of a core and a valence electronic environment. More detailed explanations of the calculation using descriptors are shown in Roy *et al* [136].

- Edge adjacency indices

These are topological indices calculated from the edge adjacency matrix of a molecule. The edge adjacency matrix is derived from the H-depleted molecular graph and encodes information about connectivity between graph edges. It is a square symmetric matrix of dimension $nBO \times nBO$, where nBO is the number of bonds between non-hydrogen atom pairs. The entries of the matrix equal one if the considered bonds are adjacent and zero otherwise. More detailed explanations of the calculation using descriptors are shown in Estrada[137].

- Functional group counts

These are simple molecular descriptors defined as the number of specific functional groups in a molecule. They are calculated on the basis of molecular composition and atom connectivity such as the number of terminal primary carbon, number of aldehydes and number of nitriles.

- Atom-centred fragments

These are simple molecular descriptors defined as the number of specific atom types in a molecule. They are calculated on the basis of molecular composition and atom

connectivity. DRAGON calculates 115 atom-centred fragments which are those defined by Ghose and Crippen[138].

- Atom-type E-state indices

Atom-type E-state indices are molecular descriptors that combine structural information about the electron accessibility associated with each atom-type, an indication of the presence or absence of a given atom-type and a count of the number of atoms of a given atom-type[139,140].

- CATS 2D

CATS 2D descriptors encode 2D features of molecules as an array of values[141,142]. They consist of bins, each bin being a substructure descriptor associated with a specific molecular feature. Atom Pairs are defined in terms of any pair of atoms and bond types connecting them. An atom pair is composed of 2 non-hydrogen atoms and an interatomic separation.

5 potential pharmacophore points (PPPs) are used: hydrogen-bond donor (D), hydrogen-bond acceptor (A), positively charged (P), negatively charged (N), and lipophilic (L). If an atom does not belong to any of the 5 PPP types, it is not considered. Moreover, an atom is allowed to be assigned to one or 2 PPP types. For each molecule, the number of occurrences of all 15 possible pharmacophore point pairs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) is determined and then associated with the number of intervening bonds between the 2 considered points, whereby the shortest path length is used. Topological distances of 0 – 9 bonds are considered to lead to a 150-

dimensional autocorrelation vector. More detailed explanations of the calculation using descriptors are shown in Fechner *et al and* Schneider *et al* [141,142].

- Molecular properties

This block includes a set of heterogeneous molecular descriptors describing physico-chemical and biological properties as well as some molecular characteristics obtained by literature models[138,143–148].

II.6.2. Machine Learning Algorithms

II.6.2-1. Support Vector Machine (SVM)

SVM separates negative and positive samples via a decision boundary with maximized margins[117,121,149]. The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points. Soft margin SVM allows some of the training points to be misclassified because training datasets are not always linearly separable in practice. The gaussian kernel is the common SVM kernel function given by:

$$k(x, x_n) = e^{-\gamma||x-x_n||^2}.$$

II.6.2-2. Random Forests (RF)

RF is an ensemble method which generates many decision trees and determines prediction classes based on a majority vote[150,151]. Decision trees are built using a Bootstrap Sampling method of all training datasets in which some datasets for each tree are randomly selected while allowing multiple samples. RF avoids overfitting which is

common in other machine learning methods and RF learning occurs via random feature selection from origin vectors.

II.6.2-3. *k*-Nearest Neighbor (*k*NN)

*k*NN learns feature vectors and corresponding classes. For testing, prediction classes are determined as the classes that occur most frequently among the *k* training samples nearest to test samples vectors[121,152,153].

II.6.2-4. Multilayer Perceptron (MLP)

MLP is a type of feed-forward neural network that is built from a network with 3 types of layers: input, hidden and output[121,154,155]. MLP has recently been applied to pattern recognition. *D* dimensional vectors $(x_1, x_2, \ldots, x_D)$ are inputted. Within the hidden layer, *M* linear combinations of inputs are constructed and then transformed using a differentiable nonlinear activation function. *K* dimensional vectors define the prediction probability of each class and are outputted using the same linear combinations used in other layers. For training, weighted vectors of each layer are applied to minimize a cross-entropy error function. MLP updates and adjusts the weights iteratively using error back-propagation to determine the minimum of the error function[156].

II.6.3. Performance Evaluation Parameters for Cross-Validation and the *E. coli* K-12 Test

To evaluate prediction model performance, the following values are calculated, given by:

$$Accuracy(Acc) = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision(Pr) = \frac{TP}{TP + FP}$$

$$Recall(Re) = \frac{TP}{TP + FN}$$

$$F_1\ score(F_1) = \frac{2 \cdot Pr \cdot Re}{Pr + Re}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives. Moreover, the values below were also calculated which are given by:

$$Average\ Accuracy = \frac{1}{L}\sum_{i=1}^{L} Accuracy_i$$

$$Macro\ Precision\ (Pr_M) = \frac{1}{L}\sum_{i=1}^{L} Precision_i$$

$$Macro\ Recall\ (Re_M) = \frac{1}{L}\sum_{i=1}^{L} Recall_i$$

$$Macro\ F_1\ score\ (F_{1M}) = \frac{2 \cdot Pr_M \cdot Re_M}{Pr_M + Re_M}$$

where L represents the number of prediction classes (the number of EC number first digits)[157,158]. In this chapter, various parameters were utilized because prediction results cannot be evaluated using only Accuracy when test datasets are imbalanced. A receiver operating characteristic (ROC) curve was produced for each model using posterior probabilities of SVM-OvR test samples, distances from the SVM-Multi decision boundary and prediction probabilities from the other machine learning methods. The area under ROC curve (AUC) was then calculated as a benchmark of prediction ability. The vertical line of ROC curves represents true positive rates and the horizontal line represents false positive rates, as given by:

$$TPr(True\ positive\ rate) = \frac{TP}{TP + FN}$$

$$FPr(False\ positive\ rate) = \frac{FP}{FP + TN}$$

AUC was calculated using the scikit-learn library[119]. Average AUC for each class was

also calculated.

II.6.4. Supplementary Figures and Tables

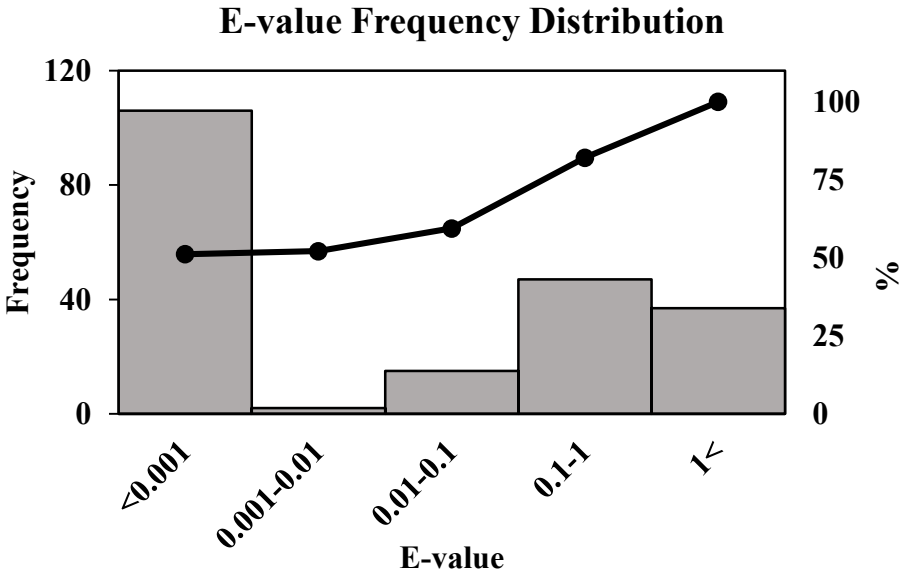
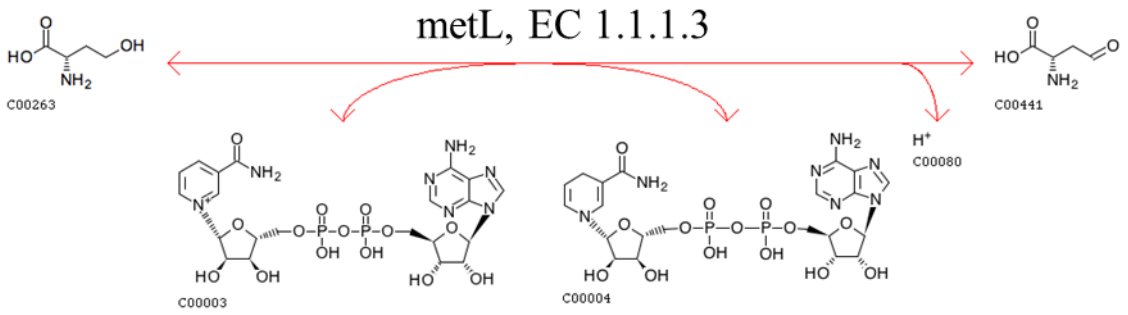
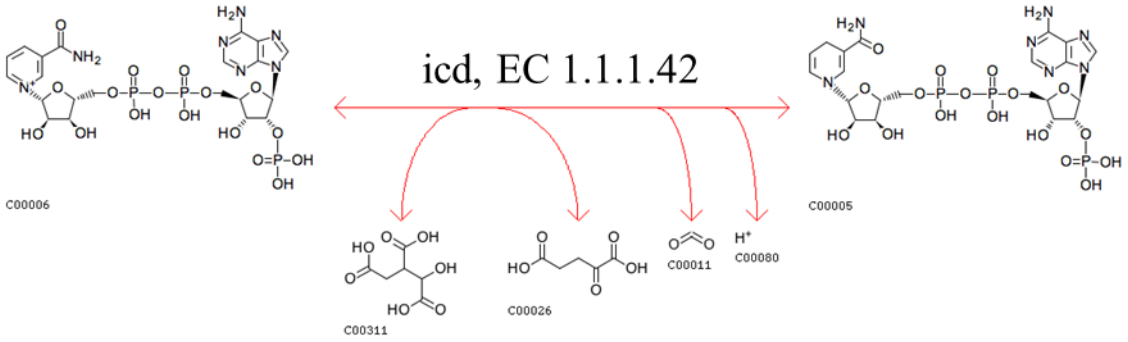


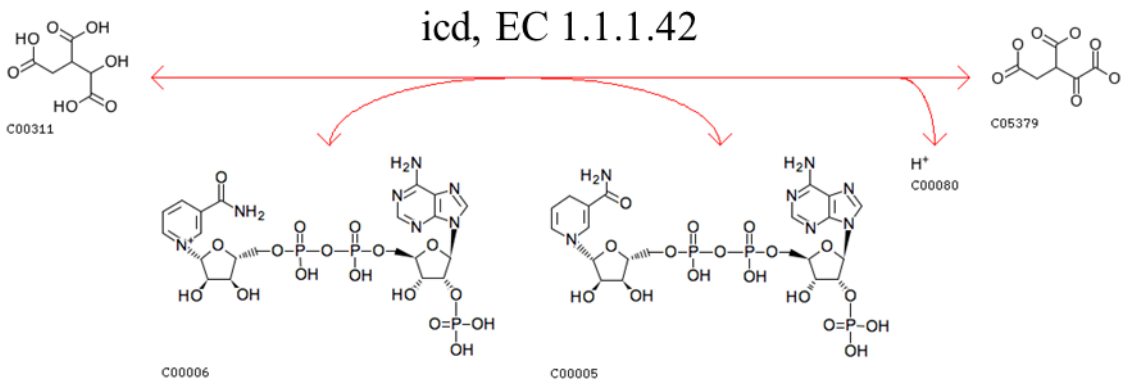
Figure S1. Enzyme sequences based similarity between training and test datasets in SE and SEP model evaluations. 3 of 210 test samples had low sequence identity to training sequences.



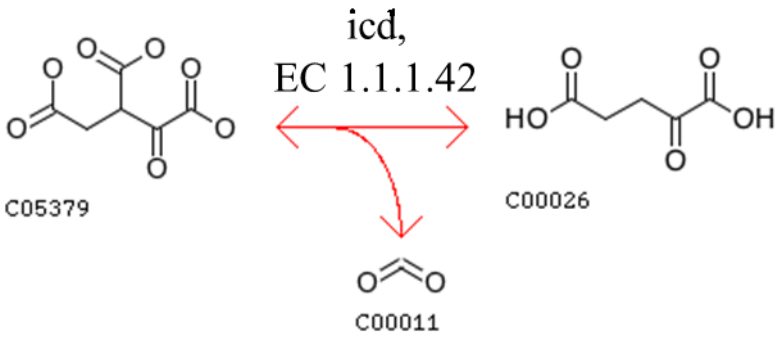
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)

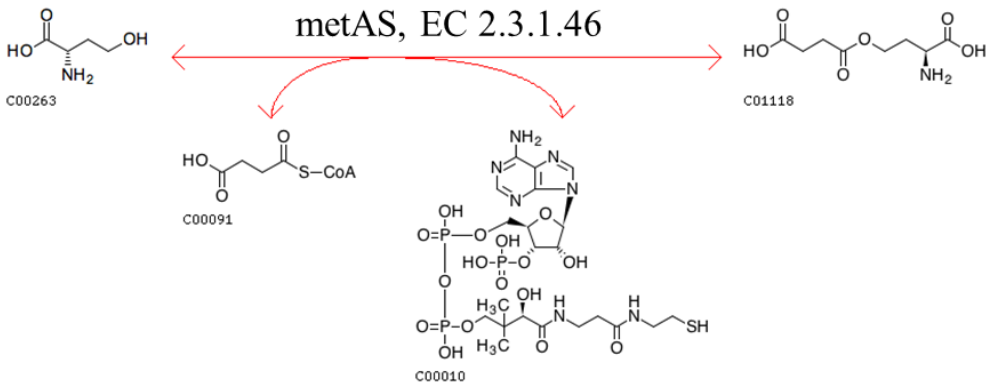


(SVM-OvR, RF, kNN, MLP)=(×,×,O,O)



(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)

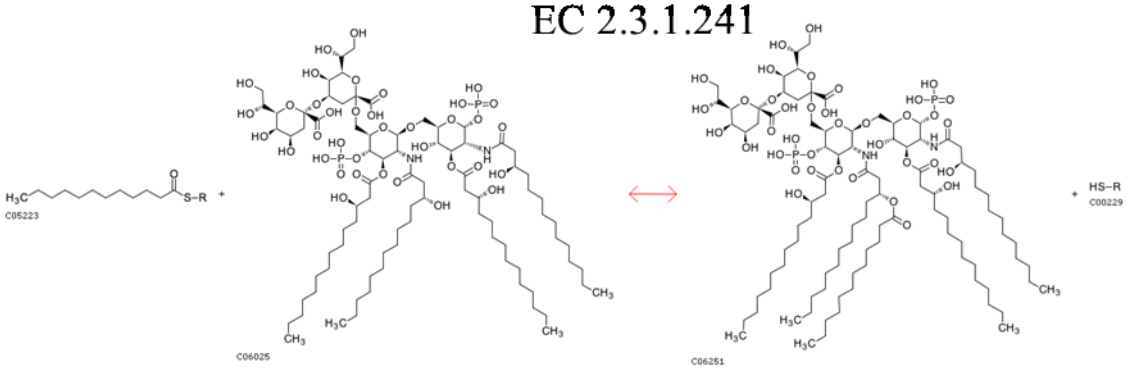
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)

lpxL,

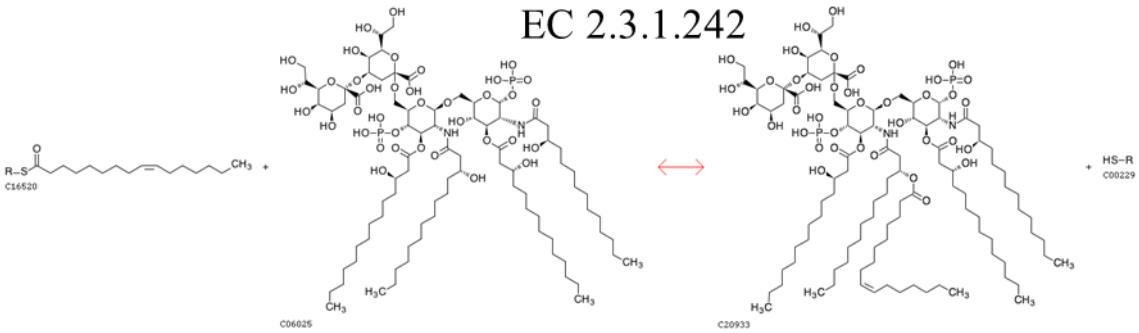
EC 2.3.1.241



(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)

lpxP,

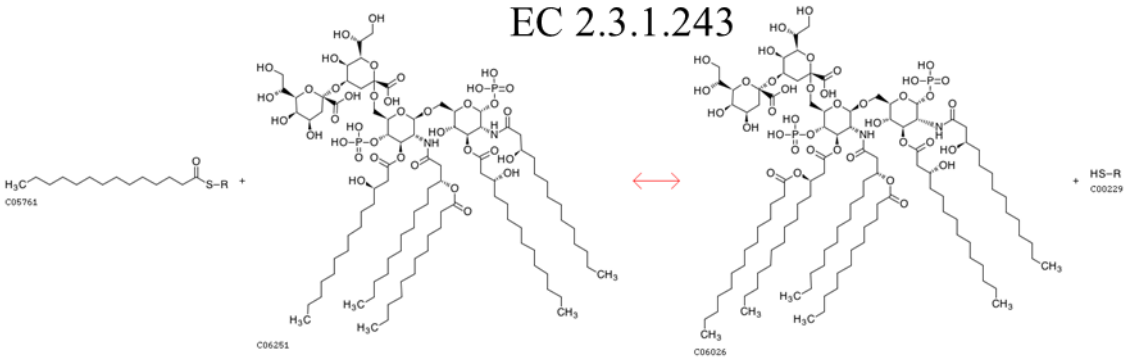
EC 2.3.1.242



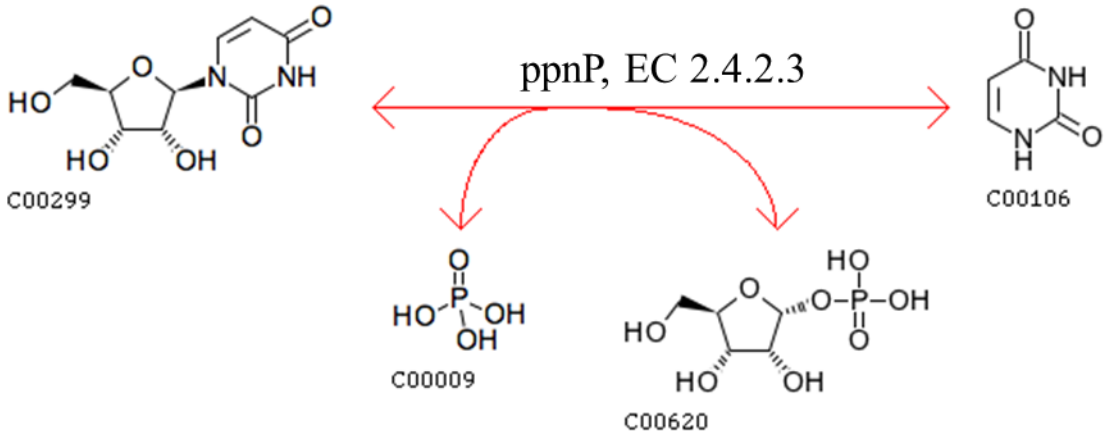
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)

lpxM,

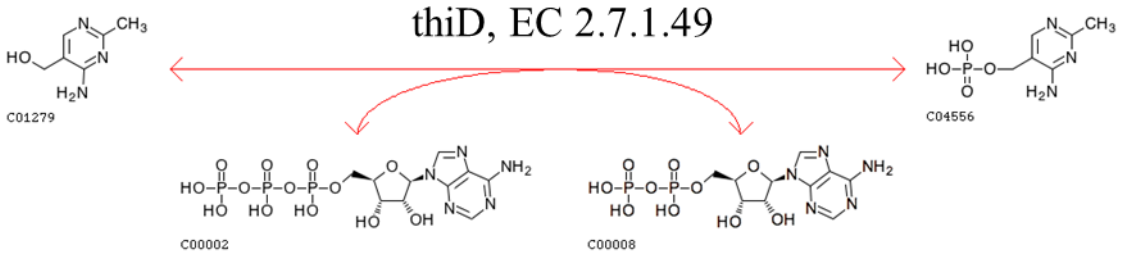
EC 2.3.1.243



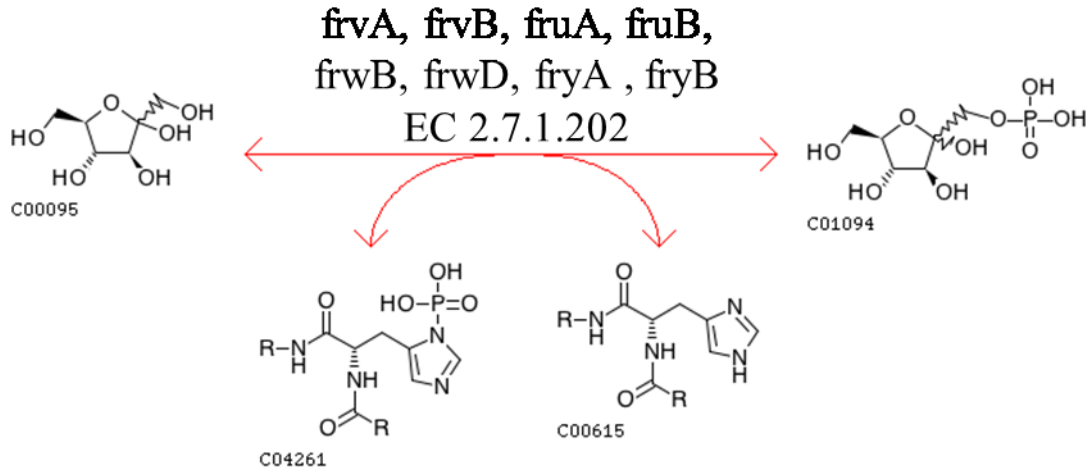
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



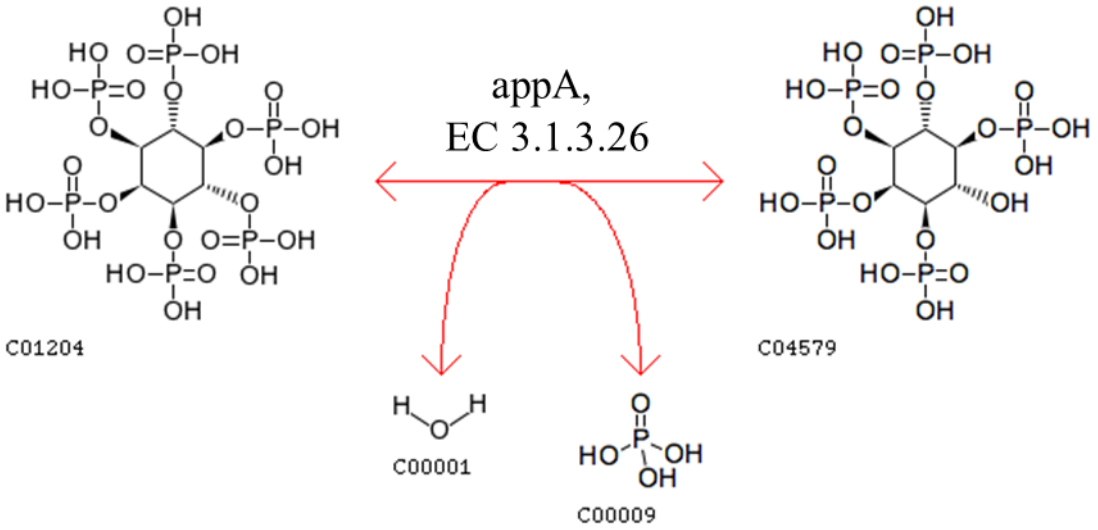
(SVM-OvR, RF, kNN, MLP)=(×,○,○,×)



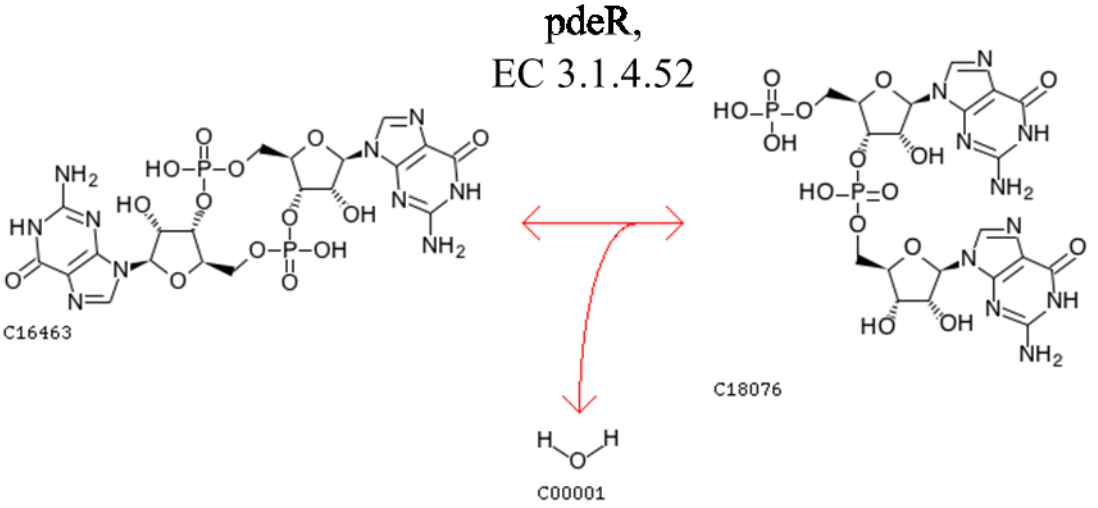
(SVM-OvR, RF, kNN, MLP)=(○,○,○,○)



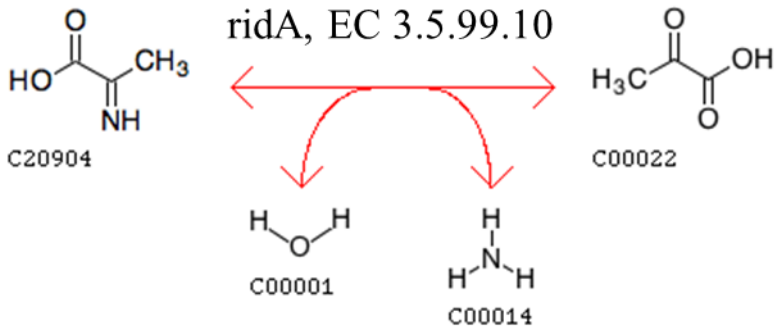
(SVM-OvR, RF, kNN, MLP)=(○:5,○:8,○:8,○:5)



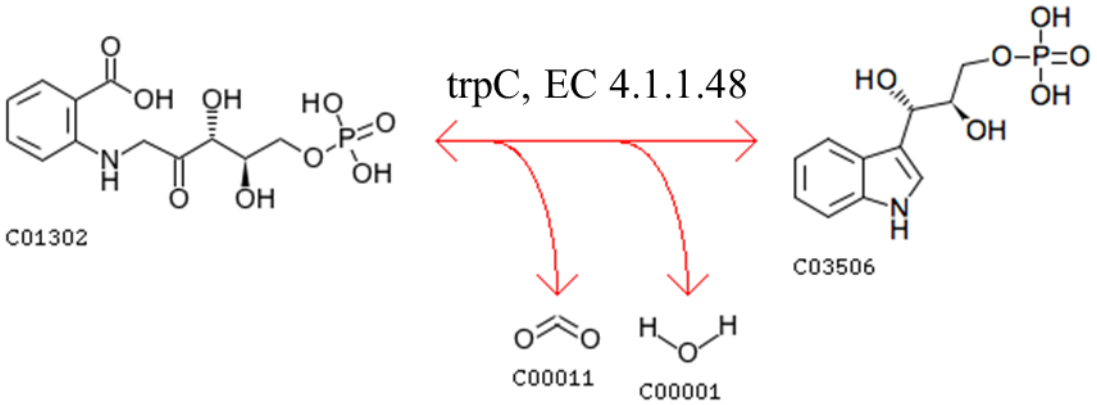
(SVM-OvR, RF, kNN, MLP)=(○,○,○,○)



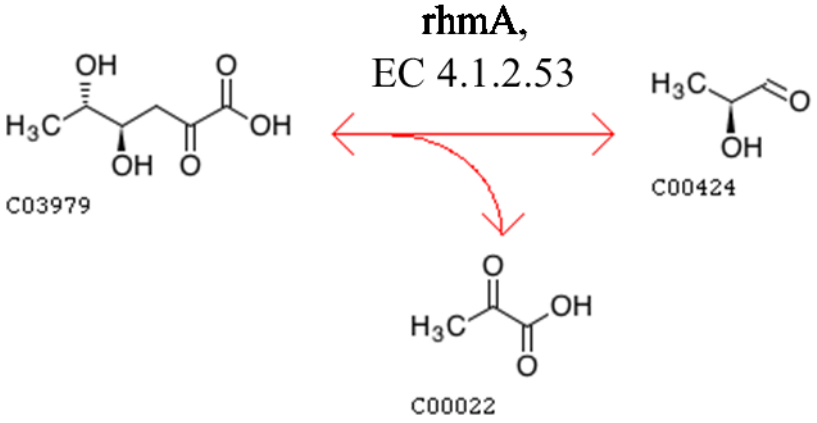
(SVM-OvR, RF, kNN, MLP)=(×,×,×,○)



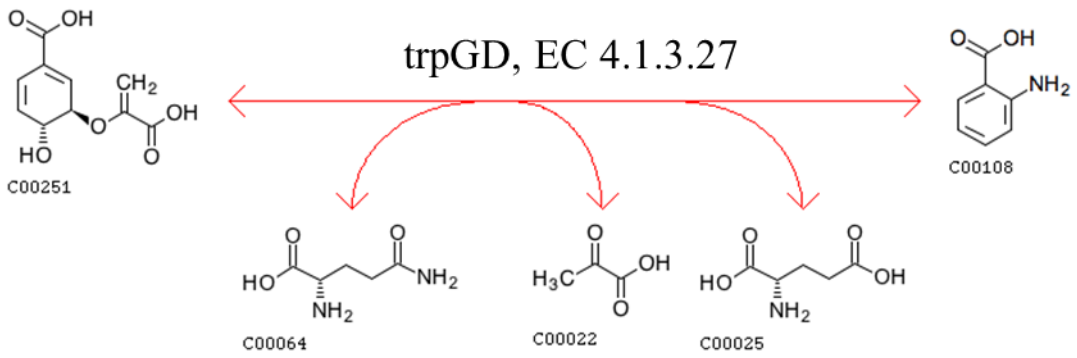
(SVM-OvR, RF, kNN, MLP)=(×,×,×,×)



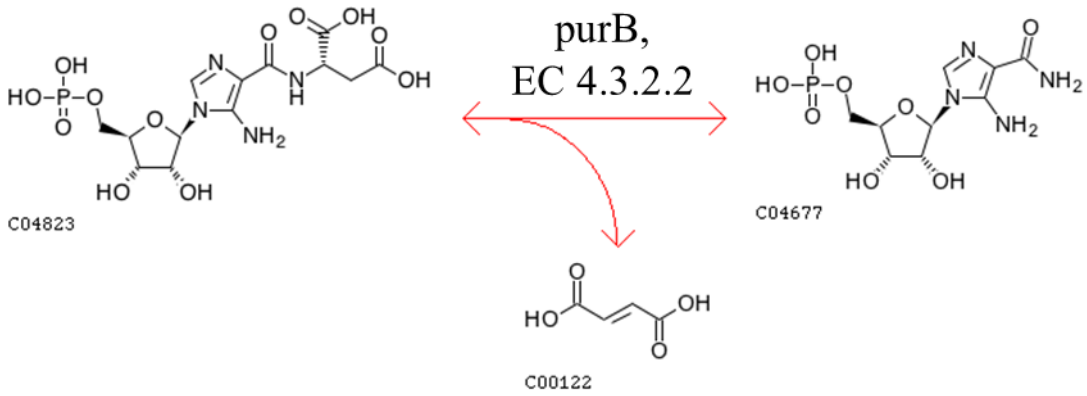
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



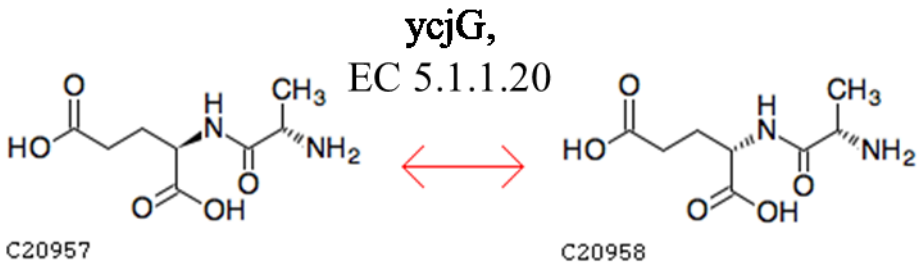
(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



(SVM-OvR, RF, kNN, MLP)=(O,O,O,O)



(SVM-OvR, RF, kNN, MLP)=(×,×,×,×)



(SVM-OvR, RF, kNN, MLP)=(×,×,×,×)

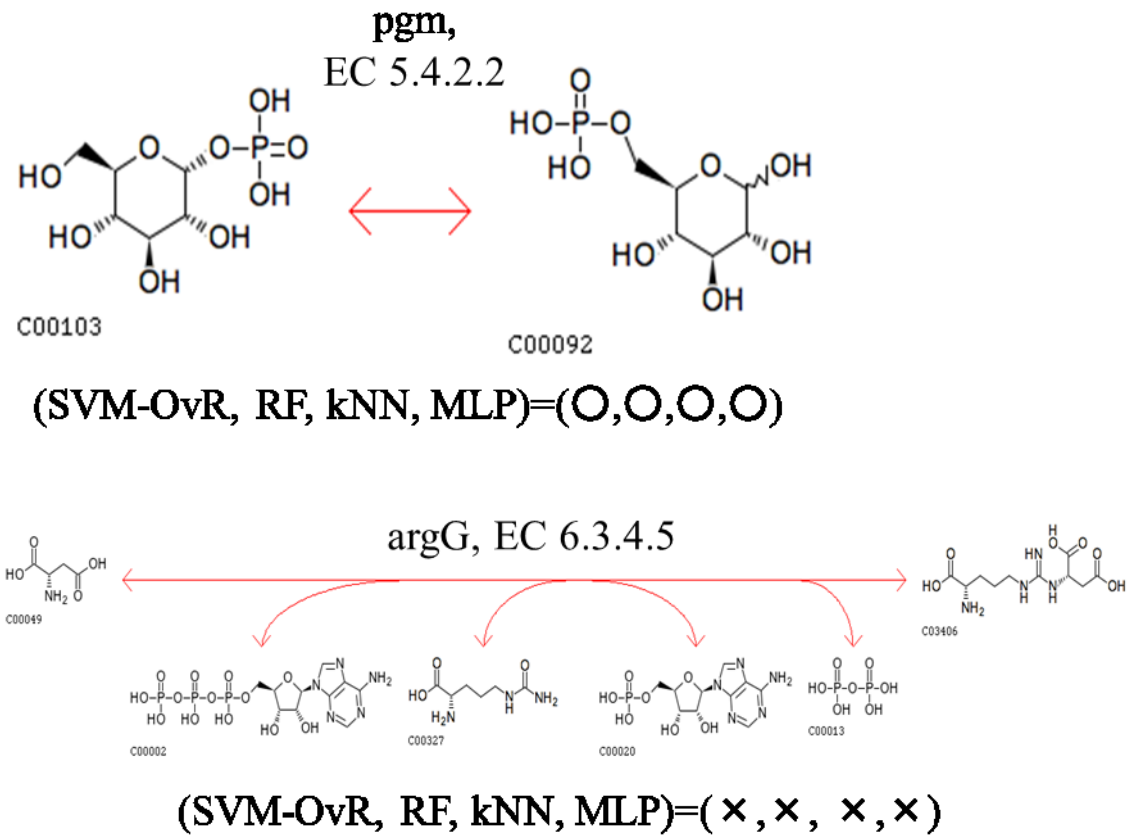


Figure S2. Enzyme chemical equations[18] and prediction results. ○ shows correct prediction, and × shows incorrect prediction.

Table S1. List of Hyper-parameters in SVM Models: (A) OvR, (B) Multi.

(A)	E-SVM-OvR model		SE-SVM-OvR model		SEP-SVM-OvR model	
	C	*gamma*	*C*	*gamma*	*C*	*gamma*
EC 1	2	0.015625	2	0.015625	2	0.015625
EC 2	8	0.015625	8	0.015625	8	0.015625
EC 3	4	0.015625	4	0.015625	4	0.015625
EC 4	2	0.015625	2	0.015625	2	0.015625
EC 5	1	0.015625	1	0.015625	1	0.015625
EC 6	16	0.0078125	16	0.0078125	16	0.0078125

(B)	E, SE, SEP-SVM- Multi model	
	C	*gamma*
-	4	0.015625

Regularization parameter *C*, Gaussian kernel coefficient *gamma* and the other parameters were defalut[119].

Table S2. List of Hyper-parameters in RF Models.

E-RF model		SE-RF model		SEP-RF model	
NT	*criterion*	*NT*	*criterion*	*NT*	*criterion*
50	Gini	400	Gini	550	Entropy

The number of trees (*NT*), *criterion* for the information gain and the other parameters were defalut[119].

Table S3. List of Hyper-parameters in kNN Models.

E-kNN model				SE-kNN model			
NN	*Algorithm*	*P*	*W*	*NN*	*Algorithm*	*P*	*W*
5	Ball tree	Manhattan	Distance	1	Ball tree	Euclidean	Uniform

SEP-kNN model			
NN	*Algorithm*	*P*	*W*
1	Ball tree	Euclidean	Uniform

The number of neighbors (NN), Algorithm used to compute the nearest neighbors, Power parameter (P) for the Minkowski metric, weight function (W) and the other parameters were defalut[119].

Table S4. List of Hyper-parameters in MLP Models.

E-MLP model				SE-MLP model			
Activation	*solver*	*L2*	*batch size*	*Activation*	*solver*	*L2*	*batch size*
Logistic	Adam	0.01	50	Tanh	Lbfgs	0.0001	None

SEP-MLP model			
Activation	*solver*	*L2*	*batch size*
Logistic	Adam	0.001	50

Activation function for the hidden layer (Activation), solver for weight optimization, L2 penalty parameter (L2), batch size of size of mini-batches for stochastic optimizers and the other parameters were defalut[119].

Table S5. Performance Comparison of BLAST and Machine Learning Methods in the *E. coli* K-12 Test.

	BLAST	E model				SE model				SEP model				
E-value	-	SVM-OvR	RF	kNN	MLP	SVM-OvR	RF	kNN	MLP	SVM-OvR	RF	kNN	MLP	Number of samples
<0.001	0.934	0.509	0.434	0.519	0.406	0.66	0.651	0.708	0.623	0.821	0.868	0.849	0.821	106
0.001-0.01	1	0	1	0.5	0	1	1	1	0.5	1	1	1	0.5	2
0.01-0.1	0.4	0.133	0.467	0.267	0.2	0.6	0.667	0.467	0.533	0.733	0.8	0.6	0.667	15
0.1-1	0.378	0.178	0.156	0.178	0.378	0.489	0.489	0.489	0.533	0.467	0.711	0.689	0.667	45
1= or <	0.077	0.256	0.231	0.385	0.359	0.513	0.59	0.564	0.462	0.564	0.795	0.744	0.615	39
None	0	0.667	0.667	0.667	1	0.667	1	1	1	0.667	1	1	1	3
All	0.605	0.362	0.348	0.405	0.381	0.595	0.614	0.624	0.571	0.69	0.819	0.781	0.738	210

Table S6. Cross-Validation (CV) and *E. coli* K-12 Test (Test) Results for Each EC Number First digit in SE-SVM and SEP-SVM Models: (A) OvR, (B) Multi.

(A)	SE-SVM-OvR model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.993	0.999	0.992	0.986	0.989
EC 2	0.992	0.999	0.979	0.971	0.975
EC 3	0.995	1.00	0.995	0.994	0.995
EC 4	0.993	0.993	0.891	0.878	0.884
EC 5	0.995	0.981	0.816	0.749	0.781
EC 6	0.999	0.998	0.988	0.968	0.978

(A)	SE-SVM-OvR model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.862	0.867	0.632	0.679	0.655
EC 2	0.756	0.857	0.669	0.884	0.762
EC 3	0.927	0.884	0.591	0.542	0.565
EC 4	0.887	0.804	0.750	0.511	0.608
EC 5	0.895	0.812	0.276	0.500	0.356
EC 6	0.938	0.794	0.400	0.429	0.414

(A)	SEP-SVM-OvR model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.995	1.00	0.989	0.992	0.991
EC 2	0.995	0.999	0.987	0.984	0.986
EC 3	0.997	1.00	0.996	0.998	0.997
EC 4	0.997	0.996	0.954	0.926	0.940
EC 5	0.997	0.982	0.867	0.755	0.807
EC 6	0.999	0.999	0.980	0.982	0.981

(A)	SEP-SVM-OvR model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.910	0.910	0.754	0.768	0.761
EC 2	0.876	0.936	0.808	0.938	0.868
EC 3	0.940	0.926	0.640	0.640	0.640
EC 4	0.910	0.899	0.875	0.614	0.722
EC 5	0.936	0.881	0.438	0.412	0.424
EC 6	0.973	0.921	0.688	0.786	0.733

(B)	SE-SVM-Multi model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.991	0.996	0.987	0.985	0.986
EC 2	0.989	0.997	0.956	0.975	0.965
EC 3	0.992	0.999	0.987	0.996	0.991
EC 4	0.994	0.980	0.934	0.805	0.890
EC 5	0.997	0.950	0.929	0.763	0.838
EC 6	0.998	0.993	0.995	0.926	0.959

(B)	SE-SVM-OvR model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.862	0.867	0.632	0.679	0.655
EC 2	0.756	0.857	0.669	0.884	0.762
EC 3	0.927	0.884	0.591	0.542	0.565
EC 4	0.887	0.804	0.750	0.511	0.608
EC 5	0.895	0.812	0.276	0.500	0.356
EC 6	0.938	0.794	0.400	0.429	0.414

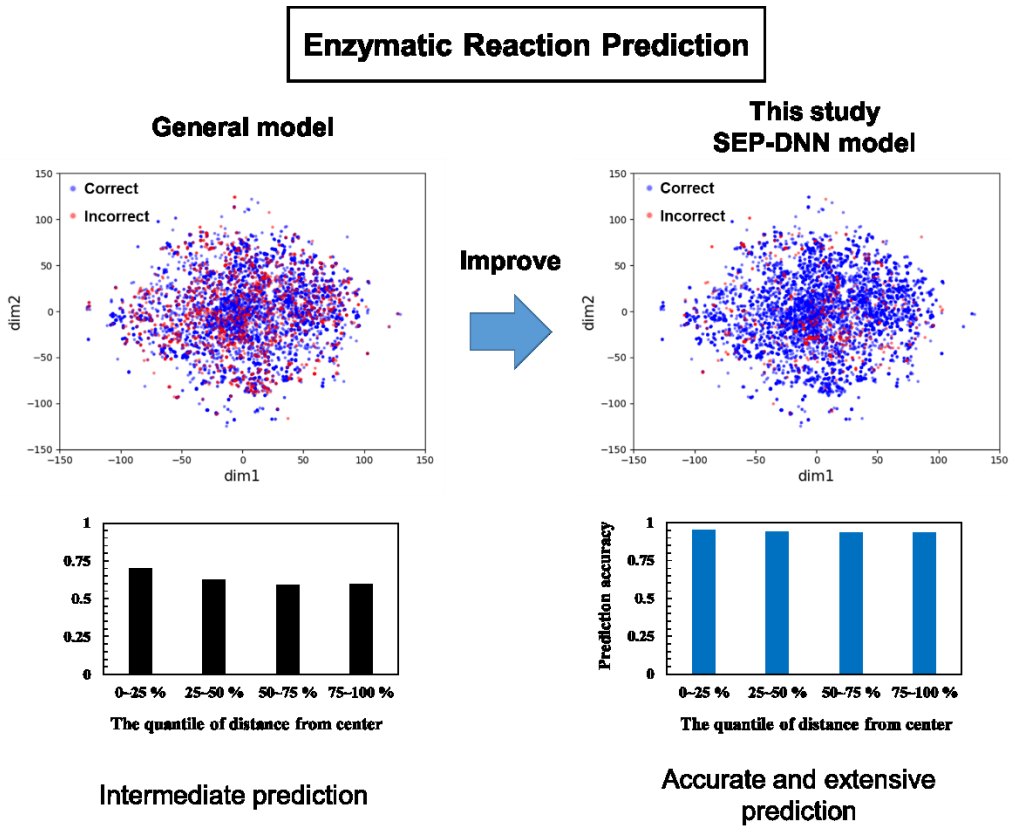
(B)	SEP-SVM-Multi model CV				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.996	1.00	0.991	0.992	0.992
EC 2	0.995	0.999	0.982	0.987	0.984
EC 3	0.994	1.00	0.990	0.998	0.994
EC 4	0.996	0.998	0.985	0.885	0.932
EC 5	0.998	0.981	0.963	0.781	0.863
EC 6	0.999	1.00	0.992	0.966	0.978

(B)	SEP-SVM-Multi model Test				
	Accuracy	AUC	Precision	Recall	F_1 score
EC 1	0.900	0.882	0.717	0.768	0.741
EC 2	0.769	0.919	0.780	0.654	0.711
EC 3	0.759	0.909	0.236	0.840	0.368
EC 4	0.896	0.870	0.964	0.474	0.635
EC 5	0.936	0.816	0.400	0.235	0.296
EC 6	0.963	0.867	1.00	0.214	0.353

CHAPTER III

Comprehensive Machine Learning Prediction of Extensive Enzymatic Reactions

Graphical Abstract



III.1. Introduction

Discovery and engineering of novel enzymes are necessary to increase the scope of target compounds which are difficult or expensive to biosynthesize with conventional methods. The number of unannotated sequences is explosively increasing and the numerous hypothetical and uncharacterized enzyme functions are accelerating.

Moreover, experimental verification of all available unknown protein sequences cannot be achieved due to high costs and time limitations. Therefore, machine learning, which can process vast amounts of available enzyme sequences, is suitable for the mass prediction of protein biological functions[32,33,61,65,78,159–162].

Several studies have reported machine learning methods for predicting novel enzymes and reactions[163–165]. To discover new enzymes that can biosynthesize a target compound, deep neural network (DNN) based models for EC number prediction have been developed[44–46,48–51]. The development of deep learning has made it possible to deal with larger data sets in comparison to classical machine learning. Here, an EC number is assigned using protein feature matrices derived from amino acid sequences. Second and subsequent digits of EC numbers, which represent the type of bonds or functional groups involved in the catalytic reaction, are also predicted to infer detailed enzyme function. In addition, models have been developed to predict the functions of multifunctional enzymes. However, the prediction of reactions that are not included in the training data has not been discussed in previous reports. In order to develop a valid model for comprehensive enzymatic reaction prediction, the model must be evaluated against unknown reactions, including novel enzymatic reactions that are not included in databases.

In Chapter II, enzymatic reaction prediction models using multiple classical machine learning algorithms are built by combining enzyme and compound information. However, these models cannot completely exclude unlikely enzyme-compound combinations from candidates. In this chapter, the models are updated to predict

whether or not enzymatic reactions will occur by including unlikely enzyme–compound combinations as negative training data. Moreover, several machine learning models are developed to predict more extensive and comprehensive enzymatic reactions. Large-scale datasets and feature extractions are updated from Chapter II and the improved models are evaluated using test data from 2 distinct databases. The 2 test datasets include the enzyme sequences derived from various species to evaluate the ability of extensive predictions for each model. E, SP (substrate–product), SE, and SEP models are built using DNN, kNN, MLP, and RF, with matching datasets and feature extractions. As a result, the prediction performances of 10 out of 16 models (not E-MLP, E-RF, and all SP models) improve over SEP-RF model which is the most accurate model as described in chapter II. These results suggest that the updated datasets and feature extractions are suitable for comprehensive enzyme prediction.

In addition, SE and SEP models do not require rigorous optimization of datasets and feature extractions when comparing the process of building E models. However, even the E-DNN and E-kNN models, which show lower accuracy than the current SE and SEP models, are still more accurate than the previously reported model[46] regardless of the test datasets. Finally, the SEP-DNN model shows the highest prediction accuracy of Chapter III with Macro F_1 scores up to 0.966, covering extensive enzymatic reactions for substrate–product combinations that are not included in the training data.

III.2. Materials and Methods

III.2.1. Data Collection

Enzymatic reaction data for E, SP, SE, and SEP models are collected from the KEGG[18] database. Positive datasets for SP, SE, and SEP models include the reactions of EC 1 to EC 6 enzymes, and the respective models are built from the corresponding combinations of substrate–product, substrate–enzyme, and substrate–enzyme–product. EC 7 reactions are not included in any of the datasets because too few reactions are registered in KEGG. This is because significant imbalance in the number of datasets for each class predicted by machine learning has a significant negative impact on model performances. EC 7 reactions are not so related to the enzymatic reactions for the production of functional substances using microorganisms. Negative datasets are randomly generated from combinations of enzymes, substrates, and products that are not expected to participate in reactions together. E model data consist of only enzyme information for EC 1 to EC 6.

To build positive datasets, the EC number, substrate, and product for each enzymatic reaction are collected from KEGG ENZYME, REACTION, and RCLASS. Corresponding enzyme sequences for each EC number group are collected from KEGG GENES. Simplified molecular-input line entry system (SMILES) strings for substrates and products are built using RDkit[166]. Mol files for compounds registered in KEGG are transformed into SMILES strings using RDkit. Isomeric compound structural information is transformed to isomeric SMILES, while non-isomeric compound information is transformed to canonical SMILES. Enzyme sequences that are duplicated or include non-canonical amino acids are removed. To keep datasets balanced, highly

similar enzyme sequences are omitted by clustering at 90% identity using CD-HIT[167] and then, only a single enzyme sequence from each cluster is included. In addition, enzyme sequences that cluster with enzyme sequences in independent test datasets are removed from training data. Positive datasets for SP, SE, and E models are built from the SEP dataset by removing unnecessary data and resulting duplicate data. For the E model, enzymes with multiple EC numbers are removed.

Negative datasets for the SE and SEP models are derived from combinations of compounds and enzymes that are not expected to occur in nature. The SP negative dataset is randomly generated. This enables direct prediction of whether or not an enzymatic reaction may happen, in addition to basic EC number estimation. Negative combinations of substrate–enzyme or substrate–enzyme–product are built by randomly shuffling substrate and product combinations from one EC number group with enzymes from another EC number group. Using this approach, it is possible to prevent the models from relying on compound information alone to judge correct predictions. The number of negative substrate–enzyme combinations for SE models is less than the number of negative substrate–enzyme–product combinations for SEP models, but within each model, the amount of negative data is the same throughout all EC number groups. The amount of negative data is determined by dividing the average number of positive data combinations for each of the 6 EC number groups by the total number of EC numbers. The distributions of quantitative similarity between positive and negative samples of SE and SEP models are shown in Figure S3 (Chapter III.5.4). More than half of the negative samples are highly similar to positive samples, with Tanimoto coefficient values over 0.8. Moreover, the quantitative similarity distributions of

training and independent test datasets for SE and SEP models are shown in Figure S4. Independent test data for SE and SEP models include more enzymatic reactions with a lower similarity to training data than that of the E models.

Total amounts of positive and negative data for each EC number first digit in E, SP, SE, and SEP models are shown in Table 5A. Data derived from the KEGG database were randomly split into training, validation, and test data, at an approximate ratio of 8:1:1 (Table 5B). Training data are used for building models; validation data are used for evaluating models when using DNN and MLP models; and test data are used for evaluating all models after training. The amount of training data is greatly expanded over Chapter II to cover a wider range of enzymatic reactions. Independent test data are included to evaluate model performance using reactions that are not included in training data. Independent test data are collected from 5595 enzymes registered in Swiss-Prot and the corresponding substrates and products (Table 5B). Moreover, additional 7 negative datasets are built as shown in Table 6 to evaluate the effect of negative dataset size on SP, SE, and SEP prediction.

Table 5. Dataset Sizes of E, SP, SE, and SEP Models: (A) Total Number of Samples in Training, Validation, and Test Data for Each EC Number Class and (B) Number of Samples in Training, Validation, Test, and Independent Test Data at an Approximate Ratio of 8:1:1.

(A)	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	Negative
E	176,673	217,934	106,463	87,046	46,786	32,936	-
SP	1,808	1,431	647	616	270	179	4,900
SE	443,251	467,678	225,712	156,861	75,067	69,210	237,160
SEP	477,480	528,938	281,795	179,772	79,978	75,812	270,450

(B)	Training	Validation	Test	Independent Test
E	535,955	65,949	65,934	5,595
SP	7,991	930	930	640
SE	1,347,430	163,762	163,747	5,600
SEP	1,518,882	187,647	187,669	5,597

Negative represents negative datasets for enzymatic reactions.

Table 6. Negative Training Dataset Variations for SP, SE and SEP Models.

Model / No.	1	2	3	4	5	6	7	8
SP	240	880	1600	2160	2720	3280	3920	4800
SE	3,080	6,160	18,480	30,800	52,360	98,560	194,040	286,640
SEP	3,005	6,010	21,035	33,055	54,090	111,185	216,360	327,545

The 7th negative dataset is the original dataset for the first models.

III.2.2. Feature Extraction

1024 dimensional enzyme feature vectors (E vectors) are transformed from enzyme amino acid sequences using ProtVec[73] which has been used in biological function predictions[71,168] for the first DNN, kNN, MLP, and RF models as shown in Figure S5. The ProtVec models are built from several millions of enzymes in KEGG. Enzymes with duplicated sequences and non-canonical amino acids [B (aspartic acid or asparagine), U (selenocysteine), and X (unknown)] are removed from the ProtVec training data. Moreover, the number of enzymes in the data is reduced using CD-HIT with a 90% identity threshold. As a result, 2,746,981 unique enzyme sequences are used to build the ProtVec models.

1024 dimensional substrate and product feature vectors (S and P vectors) are generated using SMILES using SMILESVec[69] for the first DNN, kNN, MLP, and RF models as shown in Figure S5. The SMILESVec models are built from the SMILES strings of several millions of compounds from KEGG and ChEMBL[169]. Mol files from KEGG and CHEMBL are transformed to SMILES strings using RDkit. Duplicate compounds are removed from the SMILESVec training data. The length of SMILES strings is limited from 10 to 250 characters to omit overly long SMILES strings and to build 4-gram models. As a result, 1,934,543 compounds are finally used to build the SMILESVec models. Moreover, 256, 512 and 2,048 dimensional S, E, and P vectors are generated to evaluate the relationship between model accuracy and the number of ProtVec and SMILESVec dimensions.

III.2.3. Machine Learning

Several machine learning algorithms are compared to evaluate tendencies for comprehensive enzymatic reaction prediction. E, SP, SE, and SEP models are built using DNN, kNN, MLP, and RF, resulting in 16 independent prediction models: E-DNN, E-kNN, E-MLP, E-RF, SP-DNN, SP-kNN, SP-MLP, SP-RF, SE-DNN, SE-kNN, SE-MLP, SE-RF, SEP-DNN, SEP-kNN, SEP-MLP, and SEP-RF. These machine learning algorithms have been demonstrated in various biological function predictions. SVM methods are not used because the time of building SVM models is much longer in comparison to other machine learning methods even when the data of the previous chapter whose number is much smaller than that of Chapter III is used. Next, all models are optimized by evaluating the effects of dataset size and feature vector dimensions on prediction accuracy. E models only perform positive prediction of first digit EC number groups, while the SP, SE, and SEP models predict whether or not data belongs to a first digit EC number group. In the SE and SEP models, first digits of EC numbers are still used because second and subsequent digits are inferred by substrate and product information. The SEP-DNN model structure is shown in Figure S6A. In the Covert Feature layers, substrate, enzyme, and product features are compressed and extracted, and then the resulting 3 vectors are merged. From EC detector layers, the score for each class is then outputted. SEP-MLP model is built as shown in Figure S6B. MLP models are built using a model structure that is similar to that of the DNN models (Figure S6B). A single dense layer is used as a hidden layer in MLP models, while 4 dense layers are used in DNN models. DNN and MLP models are built using Tensorflow[170], while kNN and RF models are built using scikit-learn[119]. Hyper-parameters for DNN and MLP models are as follows: Batch size = 128, activation function = ReLU, optimizer =

Adam, learning rate = 0.001, dropout rate = 0.3, and epoch = 100. The DNN and MLP models are built using 90 epochs, where the Macro F_1 scores in the test are highest. A categorical cross-entropy loss function is used to train DNN and MLP models, and trainable parameters are updated for each batch. Loss values, including those shown in Figure S7, are calculated as the average of all batches. Hyper-parameters for kNN and RF models are shown in Tables S7 and S8 (Chapter III.5.4). Multiple evaluation parameters (Chapter III.5.1) are used to compare the performance of each model.

III.2.4. Calculation of Variation in Correct Enzymatic Reaction Prediction

To estimate the variation of correct prediction, the sum of the euclidean distance between correct sample vectors and the center of the gravity vector for all samples, including test and independent test data, is calculated for each model. When a feature vector of a correct sample is represented as $X_k = (x_1, x_2, \dots, x_m)$, the distance between the sample and center of gravity in all samples is calculated as:

$$d_k = \sqrt{\sum_{i=1}^{m} \left(X_{k_i} - X_{C_i}\right)^2}$$

where m is the number of dimensions in the feature vector and the center of gravity vector is represented by the following equation:

$$X_c = \left(\frac{1}{n}\sum_{i=1}^{n} x_{1_i}, \frac{1}{n}\sum_{i=1}^{n} x_{2_i}, \ \dots, \frac{1}{n}\sum_{i=1}^{n} x_{m_i}\right)$$

where n is the number of all samples. The sum of distances between all samples and the center of gravity is calculated as:

$$V = \sum_{k=1}^{n} d_k$$

When the V value is higher, the model correctly predicts more extensive enzymatic reactions. The V values for the current SE, SEP models and DeepEC by Ryu *et al.*[46] are calculated and compared. Prediction accuracy for each quantile of distance between a feature vector and the center of gravity vector is calculated for each model.

III.3. Results and Discussion

III.3.1. Initial Model Evaluation

16 independent enzymatic reaction prediction models are evaluated in this chapter based on 4 machine learning algorithms with 4 combinations of enzyme, substrate, and product information: E-DNN, E-kNN, E-MLP, E-RF, SP-DNN, SP-kNN, SP-MLP, SP-RF, SE-DNN, SE-kNN, SE-MLP, SE-RF, SEP-DNN, SEP-kNN, SEP-MLP, and SEP-RF. All models are evaluated using 2 types of tests: (1) test using data derived from the KEGG database and (2) independent test using data derived from the Swiss-Prot database. The independent test is used to evaluate the prediction performance for enzymatic reactions of enzymes that are not included in the training data. Evaluation results are shown in Figure 15 and Tables S9 to S11. All DNN (Figure S7) and MLP models are optimized using validation datasets. Test results are best in epoch 90 according to Macro F_1 scores. When comparing the performance of each model based on all evaluation parameters, the SP models score the lowest, followed by the E models in both tests. On the other hand, the combination of enzyme and compound information improves the relative prediction accuracy of the SE and SEP models. Furthermore, correct prediction with the SEP models is the most accurate, with the exception of the

RF model (Figure 15 and Tables S9 to S11). SE and SEP models are both highly valid as indicated by the high prediction accuracy in the test and the independent test. These results suggest that the updated SE and SEP models predict enzymatic reactions with high accuracy regardless of which machine learning algorithm is used. Moreover, all E, SE, and SEP models, with the exception of the E-RF model, are able to predict reactions for enzyme classes with large and small dataset sizes with similar accuracy. These results indicate that the prediction models of this chapter are not biased toward majority classes.

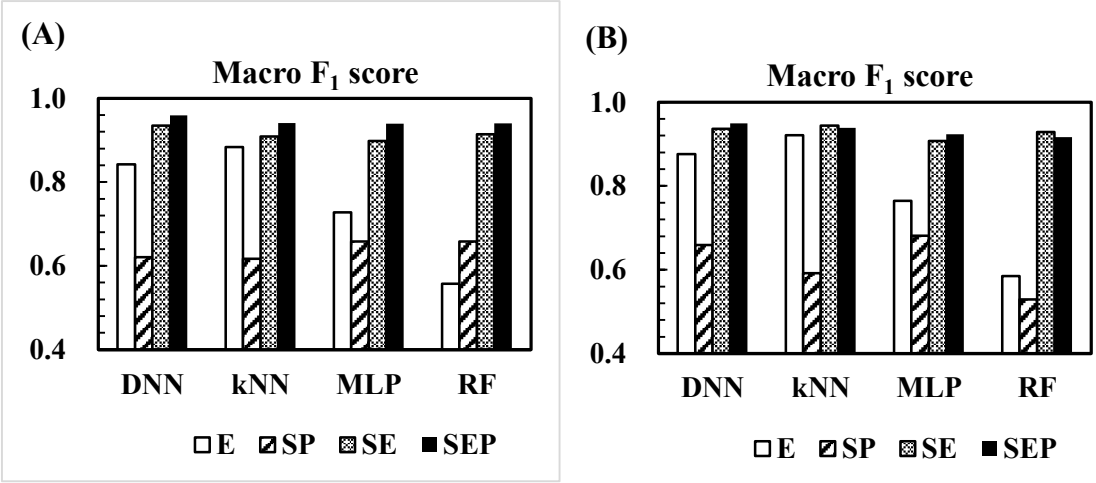


Figure 15. Macro F_1 score results in the (A) test and the (B) independent test for the first E, SP, SE, and SEP models built using 4 machine learning algorithms.

Addition of substrate and product information improves the prediction accuracy of E and SE models. On the other hand, reaction prediction with SP models, which are derived from only compound information, is less accurate than that of E and SE models. The weaker SP prediction is a result of the lower amount of substrates and products in the training data compared to the number of enzymes. The test also shows that negative

prediction by SE and SEP models is more inaccurate than the corresponding positive prediction. Here, the negative test samples are more difficult to discriminate than positive samples because the negative datasets are randomly constructed. On the other hand, SP models predict negative samples with higher accuracy than positive samples because the negative training data of SP models include mixed combinations of substrates and products from various EC number groups.

The E-DNN model misjudges many non-EC 3 enzymes as EC 3 according to lower Precision scores, while the corresponding SE and SEP models greatly improve EC 3 reaction prediction. Therefore, EC 3 reaction prediction using E-DNN derived from only enzyme sequence information is more difficult, and the inclusion of compound information into the model results in significant prediction improvement. Unlike the E-DNN model, the E-kNN model predicts all EC classes with equally high accuracy. The E-kNN model is optimal for relatively simple predictions such as EC number first digit prediction. On the other hand, DNN is more suitable for complicated predictions such as predicting second and subsequent digits or predicting multiple functions[44–46,48,49]. Since kNN is used by Dalkiran *et al*.[51] for EC number prediction, kNN may also be utilized for complicated prediction targets with improvements in datasets and feature extractions. The E-MLP model predicts all samples with low accuracy, especially with EC 3 and EC 5 enzymes that are not discriminated without addition of compound information. Similarly, SE- and SEP-MLP models are weaker than corresponding DNN models. This is because the MLP models do not follow the complete DNN model structure where features of substrates, enzymes, and products are extracted before being combined.

The E-RF model shows particularly low accuracy. Here, the enzymes from EC 3 to EC 6 tend to be predicted incorrectly because the amount of training data for these enzymes is low in comparison to that of EC 1 and EC 2 enzymes. However, when using the RF algorithm, the addition of compound information in SE and SEP models results in a greater improvement in prediction compared to that of other machine learning models. In the SE and SEP models with the RF algorithm, substrate and product dimensions are the most important factors for prediction (Figure S8), further emphasizing the importance of including compound information. Overall, the SEP-DNN model is the most accurate of all the first models when considering evaluations from both tests.

In this chapter, training data and feature extractions are updated from the previous Chapter II and several models for comprehensive enzymatic reaction prediction are subsequently built. Figure 16 shows a comparison of the prediction results of all the models from the current chapter and Chapter II with matching test data. Here, the current kNN, MLP, and RF models improve in prediction accuracy, with E models improving over 1.7-fold. The current E, SE, and SEP models built with DNN and kNN exhibit higher prediction accuracy than that of the previous SEP-RF model, the strongest model of Chapter II. This indicates that the combination of updated datasets and feature extractions results in a significant improvement in prediction. In other words, it is concluded that the present method is effective in predicting enzymatic reactions.

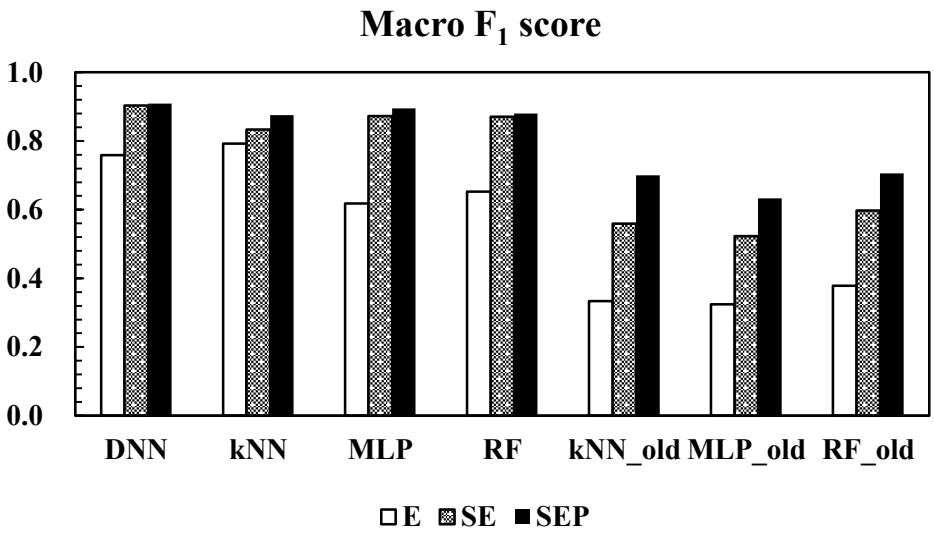


Figure 16. Macro F_1 score results for the additional test when using the first E, SE, and SEP models from the current chapter and Chapter II.

III.3.2. Relationship between Model Accuracy and the Amount of Negative Training Data

In order to optimize the current models, the effect of negative training dataset size on the prediction accuracy of SP, SE, and SEP models is next evaluated. Figure S9 shows the results when using test data from Table 5B. All models improve as the negative dataset size increases up to about 200,000 for all models except for the SP models, which improves up to about 2000 (Figure S9). After approaching maximum prediction, the addition of randomly generated negative data does not significantly improve overall accuracy. Therefore, models with stable prediction accuracy can be built by simply adding a sufficient amount of negative reaction data.

III.3.3. Relationship between Model Accuracy and Number of Dimensions for Each Feature Vector

The effect of the number of SMILESVec dimensions and ProtVec dimensions on prediction accuracy of all E, SP, SE, and SEP models is next evaluated. Figure 17 and Table S12 show the test and independent test results. For example, the 256 dimension SE model is built using 256 SMILESVec dimensions for substrate vectors and 256 ProtVec dimensions for enzyme vectors. The SP models show almost constant prediction accuracy even when the number of dimensions increases in all machine learning models. E, SE, and SEP models in DNN and MLP tend to improve as the number of feature vector dimensions increases. Prediction accuracy in the SE and SEP models increases up to 1024 dimensions but does not increase with 2048 dimensions. The E models continue to improve up to 2048 dimensions. This suggests that a large amount of enzyme sequence information is required to improve accurate prediction using DNN and MLP. MLP models are relatively similar to DNN models in terms of model structure, and accordingly, the results are similar. On the other hand, kNN and RF models do not significantly improve as the vector dimensions increase. Especially when using RF, prediction accuracy actually decreases as the dimensions increase. Thus, while all machine learning results support that combining enzyme and compound information improves enzymatic reaction prediction, a higher number of vector dimensions does not necessarily improve prediction. Therefore, it is necessary to examine the specific conditions of each prediction model on a case-by-case basis to improve model accuracy.

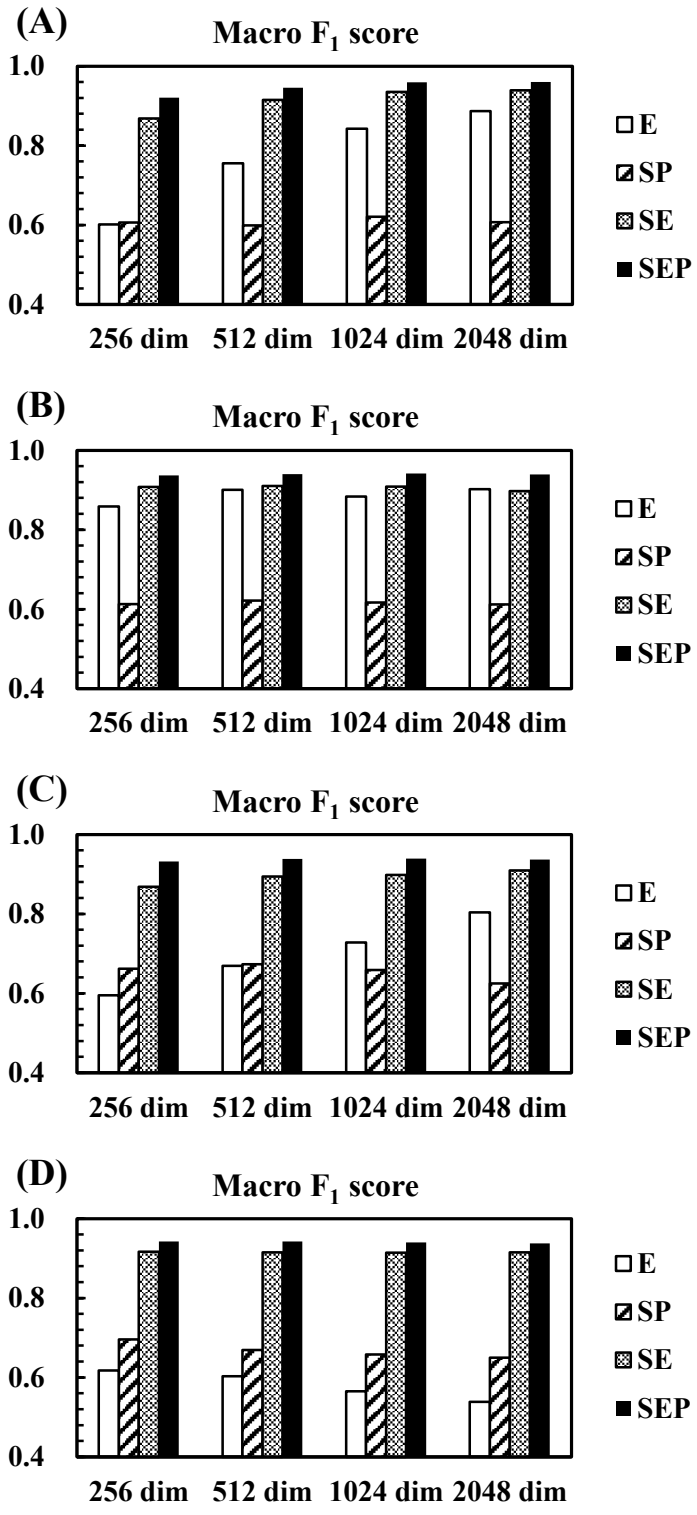


Figure 17. Macro F_1 scores of test datasets for (A) DNN, (B) kNN, (C) MLP, and (D) RF algorithms, running E, SP, SE, and SEP models derived from SMILESVec and ProtVec

vectors with 4 different numbers of dimensions.

The dimensional analysis also supports that enzymatic reaction prediction improves with the addition of compound information and that accuracy can be maintained when using low dimensional vectors. Furthermore, SE and SEP models are more versatile than the corresponding E models because the number of SE and SEP dimensions does not need to be strictly determined. In particular, for the SEP model, the information on substrate, enzyme, and product is already sufficient for accurate prediction regardless of the number of features when compared to the SE model.

III.3.4. Model Evaluation for Optimized Models

E, SP, SE, and SEP models built with the 4 machine learning algorithms are next optimized based on the case-by-case trends for negative training dataset size and feature vector dimensions. Optimized models with the highest tested prediction accuracies are then adopted. Table S13 shows the parameters for each optimized model, and Figure 18 and Tables S14 to S16 show the results. All SP models except the SP-DNN model show improved test and independent test results, and the overall SP prediction accuracy remained low. The consistently low prediction accuracy of all SP models indicates that SP models are not suitable for enzymatic reaction prediction. On the other hand, all optimized E models show improved prediction, while optimized SE and SEP models maintain an already high prediction accuracy without significant further improvement. Therefore, SE and SEP models that combine enzyme and compound features do not require extensive fine-tuning of training data and feature vectors. Furthermore, all optimized E, SE, and SEP models maintain higher prediction accuracy than the previous

models of Chapter II (Table S16). After optimization, the SEP-DNN model exhibits the highest prediction accuracy in this chapter. Moreover, linear discriminant analysis (LDA)-SE and LDA-SEP models, built and evaluated using matching datasets and feature extractions, result in a decrease of over 0.1 in Macro F_1 scores in comparison to the other SE and SEP models.

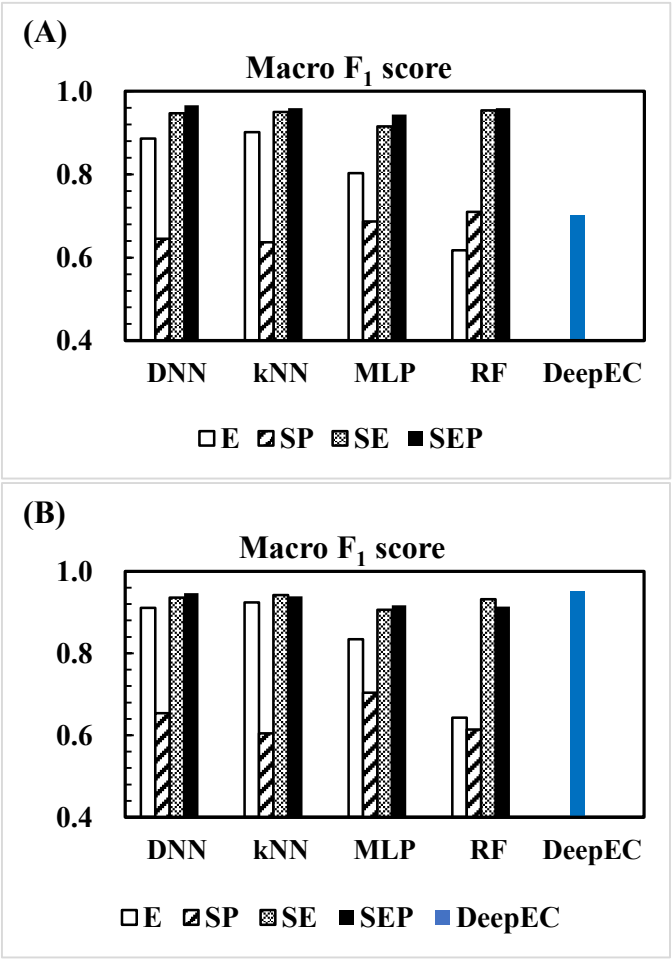


Figure 18. Macro F_1 score results for the (A) test and the (B) independent test, when using optimized E, SP, SE, and SEP models built from 4 machine learning algorithms, in comparison to DeepEC[46]. DeepEC is trained with enzyme sequences from UniProt (Swiss-Prot and TrEMBL), while the models of the Chapter III study are trained with

enzyme information from KEGG, which does not include reactions in the independent test.

SEP models, which predict enzymatic reactions with the highest accuracy, correctly predict over 80% of the samples from the independent test data using all machine learning models. Some enzymatic reactions are easy to predict and others are more challenging depending on the particular machine learning model (Figure S10). For each EC number except for EC 5, SEP-DNN and SEP-kNN models predict enzymatic reactions with similar high accuracy. For EC 5 prediction, all machine learning models show high accuracy. Reactions that MLP and RF predict correctly are also almost always correct in DNN or kNN models. Accordingly, the number of reactions that are successfully predicted by the MLP and RF models is smaller than that of the DNN and kNN models. This suggests that SEP-MLP and SEP-RF models exhibit relatively lower prediction performance.

Next, all SE and SEP models are evaluated using challenging reactions from independent test data of which the substrates or substrate–product combinations are not included in the training data (Figures 19 and 20). DNN models predict these challenging reactions with higher accuracy compared to other machine learning models, with SEP-DNN showing the highest prediction accuracy. SEP-DNN correctly predicts almost all EC 1.1.1.X and EC 2.7.2.X reactions, even though none of the corresponding reactions are used for training. For example, 2 steroids and a sterol comprise the substrates of the 3 reactions which group together at the bottom of the EC 1.1.1.X 2-dimensional plots (Figure 20A); only SEP-DNN can correctly predict all 3 reactions. For EC 2.7.2.X

reactions, SEP-RF and SEP-kNN cannot predict the reactions which act on acetic acid and propionic acid (Figure 20B). These results suggest that DNN models are more robust for prediction of unknown enzymatic reactions in comparison to other machine learning models and that the SEP-DNN model can be utilized for prediction of novel enzymatic reactions that are not present in the training data. Since prediction accuracies of DNN models are higher than that of MLP models due to a more complex model structure (Figure S6), future improvements in DNN based prediction are expected to result from additional optimization of model structure.

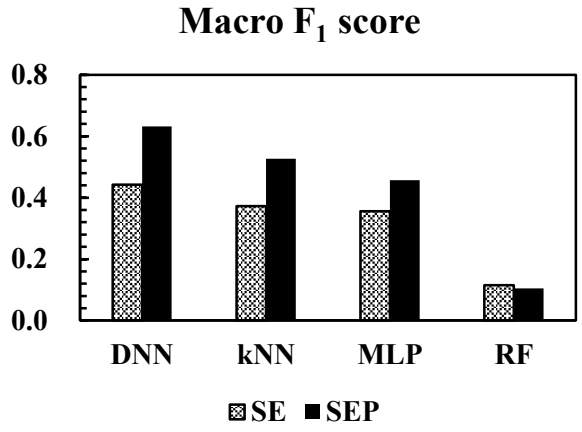


Figure 19. Macro F_1 scores for SE and SEP models built from selected independent test data, in which the substrates and substrate–product combinations are not included in the training data.

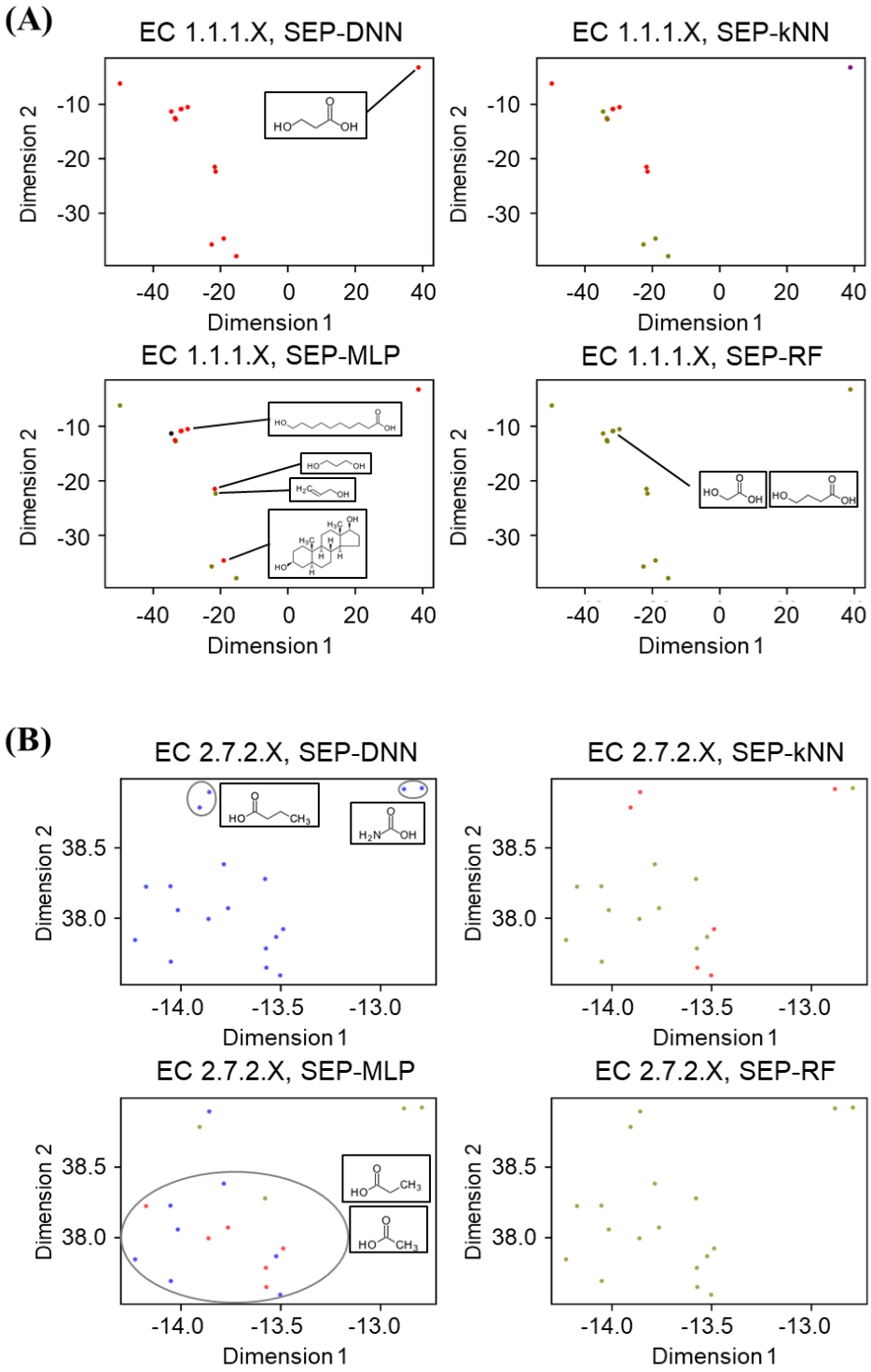


Figure 20. Visualization of independent test results for (A) EC 1.1.1.X and (B) EC 2.7.2.X reactions with test compounds that are not included in the training data. Results for SEP-DNN (upper left), SEP-kNN (upper right), SEP-MLP (lower left), and SEP-RF

(lower right) models are shown. The SEP-DNN model correctly predicts (A) 16 of 24 EC 1.1.1.X reactions and (B) 18 of 19 EC 2.7.2.X reactions. For EC 2.7.2.X reactions (B) substrates and products of the 2 reactions in the upper right are carbamic acid and carbamoyl phosphate, respectively; substrates of the 2 reactions in the upper left are butyric acid; substrates of remaining EC 2.7.2.X reactions are all acetic acid and propionic acid. The 2-dimensional plots are built from hidden layer vectors for each reaction of the SEP-DNN model in epoch 90 using t-SNE[171] (t-distributed stochastic neighbor embedding). Chemical structures represent substrates of each reaction. Red points represent reactions predicted as EC 1, blue points represent reactions predicted as EC 2, black points represent reactions predicted as EC 4, the purple point represents the reaction predicted as EC 5, and olive points represent reactions predicted as negative.

The above evaluations of challenging reactions help to ensure that models are valid for comprehensive enzymatic reaction prediction. Similar evaluations have been performed in other deep learning studies[172]. In the report by Dalkiran *et al*.[51], 30 test enzymes, which were not included in the training data, were used to evaluate the model. However, it is not clear whether or not enzymes that share high sequence identity with test enzymes were used as training data. On the other hand, in the Chapter III study, enzymes that share 90% or higher sequence identity with independent test enzymes are removed from the training data.

III.3.5. Comparative Evaluation of Enzymatic Reaction Prediction

As a benchmark, the current models are compared with DeepEC by Ryu *et al.*[46], which predicts EC number fourth digits from only enzyme sequences. In the current

evaluation, DeepEC predictions are regarded as correct if predicted EC numbers match that of true EC number first digit groups. Here, test and independent test data of E models are also used to evaluate DeepEC prediction. Because the independent test data are derived from Swiss-Prot and DeepEC is trained with data from UniProt (Swiss-Prot and TrEMBL), while the models of the Chapter III study are trained with data from KEGG, DeepEC performance is slightly better in the independent test (Figure 18B). However, when using the test data derived from KEGG, most of the current models exhibit significantly higher test prediction accuracy than that of DeepEC (Figure 18A). Although DeepEC is built from only enzyme sequence information like that of the current E models, even E-DNN and E-kNN show stronger prediction results in the test.

Furthermore, Table 7 shows the variations in correct enzymatic reaction prediction (Materials and Methods Section 4). Here, combined test and independent test data are used to evaluate the current SE and SEP models in comparison to DeepEC. As a result, the V values of SE and SEP models are much higher than that of DeepEC. For each model, prediction accuracy in each quantile of distance between a feature vector and the center of gravity vector is shown in Figure 21. The current models consistently predict samples not only near but also far from the center of gravity of all test samples with high accuracy. These results indicate that the current models predict more extensive enzymatic reactions in comparison to DeepEC. kNN models show slightly more variation in correct prediction than that of DNN models; however, run time of the SEP-kNN model is longer than that of other machine learning SEP models when predicting more than 500 samples (Figure S12). This longer time required for prediction is a significant shortcoming. Therefore, it is concluded that the SEP-DNN model is

currently the best choice for quick and extensive selection of enzymes to biosynthesize target compounds.

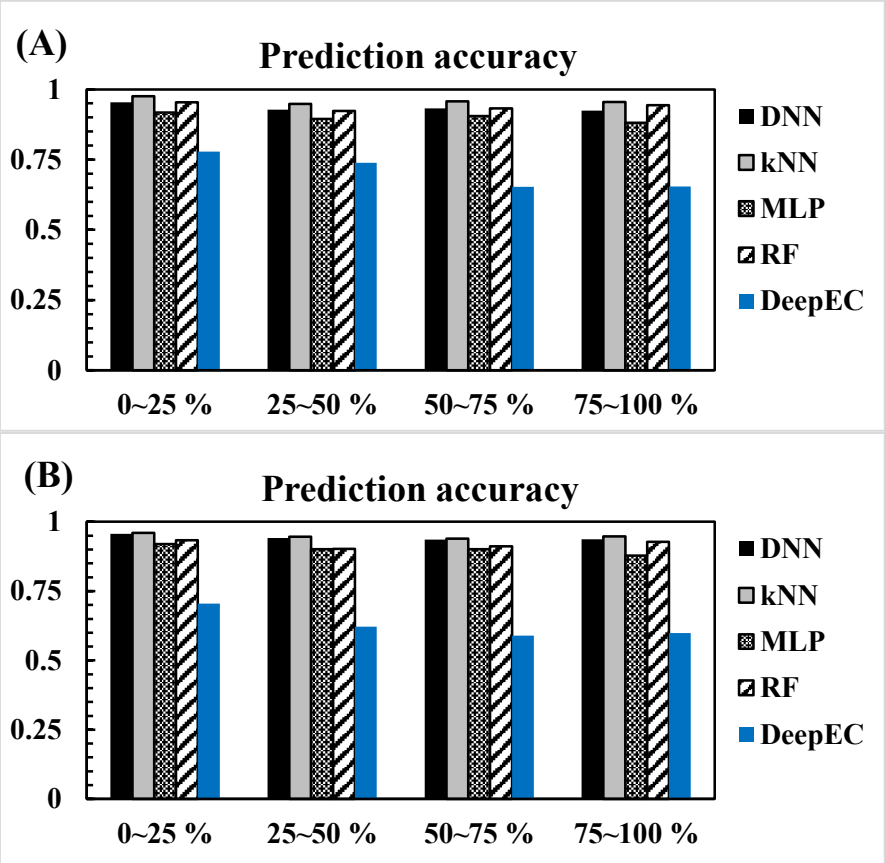


Figure 21. Prediction accuracy in each quantile of distance between a feature vector and the center of gravity vector of all (A) SE and (B) SEP test samples. (A) 4 SE models and DeepEC and (B) 4 SEP models and DeepEC.

Table 7. Evaluation of the Variation in Correct Enzymatic Reaction Prediction (*V*).

V (10^6)	DNN	kNN	MLP	RF	DeepEC
SE	0.877	0.900	0.874	0.848	0.651
SEP	1.416	1.424	1.337	1.381	0.933

SE, the number of samples is 20,493. SEP, the number of samples is 25,329.

III.3.6. Limitation

The number of available compounds in the training data is much smaller than the number of available enzyme sequences, and prediction results greatly depend upon the included compound information. For example, SE and SEP models show lower prediction accuracy for independent test reactions with compounds that do not exist in the training data (Figure 19) and for reactions that have low Tanimoto coefficient values based on low similarity with training data. Accordingly, SEP models have difficulty in predicting reactions for first digit EC number groups that share substrates and products with other first digit EC number groups (Chapter III.5.3). Therefore, the reaction prediction models can be further improved by optimizing compound feature extractions. It is also necessary to consider reducing the number of dimensions for feature vectors.

Negative training datasets for SE and SEP models consist of random SE and SEP combinations in order to prevent the models from relying only on compound feature information; however, it is possible that some of the random SE and SEP combinations might occur in nature. Furthermore, the incorrect predictions of the current models are most often misjudged as negative. These false negative predictions are due to negative samples that are similar to positive samples, with high Tanimoto coefficient values (Figure S3). Therefore, improved methods to build negative training data are also needed.

III.4. Conclusion

In this chapter, training data and feature extractions from Chapter II are updated to develop machine learning models for comprehensive enzymatic reaction prediction. 10

out of 16 of the current models show higher accuracy than the Chapter II models. The current SE and SEP models that combine enzyme and compound features can exclude unlikely enzyme-compound combinations. These models show improved prediction and require less optimization in comparison to the E models. Overall, the SEP-DNN model most quickly and accurately predicts extensive enzymatic reactions and is robust for prediction of reactions that are not included in training data. This model will more greatly help to select enzyme sequences and discover novel enzymatic reactions in metabolic pathways for the production of useful substances than the other Chapter III models and the Chapter II models

III.5. Supplementary Information

III.5.1. Performance Evaluation Parameters for Test Data

To evaluate prediction model performance, the following values are calculated, given by:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1\ \ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively. TP and TN are the number of samples that are correctly predicted, while FP and FN are the number of samples that are incorrectly predicted. The values below are also calculated as given by:

$$Macro\ \ Precision = \frac{1}{L}\sum_{i=1}^{L} Precision_i$$

$$Macro\ \ Recall = \frac{1}{L}\sum_{i=1}^{L} Recall_i$$

$$Macro\ \ F_1\ \ score = \frac{2 \cdot Macro\ \ Precision \cdot Macro\ \ Recall}{Macro\ \ Precision + Macro\ \ Recall}$$

where L represents the number of prediction classes[157,158]. Receiver operating characteristic (ROC) curves are produced for each model using prediction scores from each machine learning method. The area under ROC curve (AUC) is then calculated as a benchmark of prediction ability. The vertical lines of ROC curves represent true positive rates and the horizontal line represents false positive rates, as given by:

$$TPr(True\ \ positive\ \ rate) = \frac{TP}{TP + FN}$$

$$FPr(False\ \ positive\ \ rate) = \frac{FP}{FP + TN}$$

AUC is calculated using the scikit-learn library[119]. Average AUC for each prediction class is also calculated.

III.5.2. DNN and MLP Model Training

The 7th dataset of Table 6 with 1,024 dimensional SMILESVec and ProtVec vectors are used to train the first models. Figure S7 shows loss function curves for training and validation in DNN models. A categorical cross-entropy loss function is used to train DNN and MLP models, and trainable parameters are updated for each batch. Loss values, including those shown in Figure S7, are calculated as the average of all batches.

For E-DNN, SE-DNN, and SEP-DNN models, validation loss decreases as training proceeds, indicating that these models are optimized. However, the SP-DNN validation loss stops decreasing after reaching epoch 40, indicating that the SP model is not optimized. Moreover, validation loss in the SEP-DNN model, which is derived from the largest amount of feature information, is the lowest, while loss in the E-DNN model is the highest of E-DNN, SE-DNN and SEP-DNN models. In the second half of tested epochs (epochs 50-100) for SE-DNN and SEP-DNN models, the training loss values are almost the same as the validation loss values. The matching loss values suggest that overfitting is not a problem for SE-DNN and SEP-DNN models. MLP models are built in a similar way to the DNN models, and MLP validation results are similar to DNN results.

III.5.3. Prediction Results for Independent Test of SEP Models

Figure S10 shows representative prediction results from the independent test of SEP-DNN, SEP-kNN, SEP-MLP and SEP-RF models, which predict enzymatic reactions with the highest accuracy among optimized models, and Figure S11 shows prediction results visualized in 2 dimensions. 292 of 362 EC 1.1.1.X reactions, which are oxidation/reduction reactions with hydrocarbons where NAD+ or NADP+ acts as the acceptor, are correctly predicted by all models. As shown in Figure S11A, most of the correctly predicted EC 1.1.1.X reactions are similarly judged by all 4 machine learning algorithms, with some slight variations depending on the specific algorithm. There are only 8 EC 1.1.1.X reactions that all models do not correctly predict, which indicates that the prediction models are accurate. On the other hand, for EC 1.10.3.X, which acts on diphenol-related compounds with oxygen as an acceptor, 13 of 15 reactions are

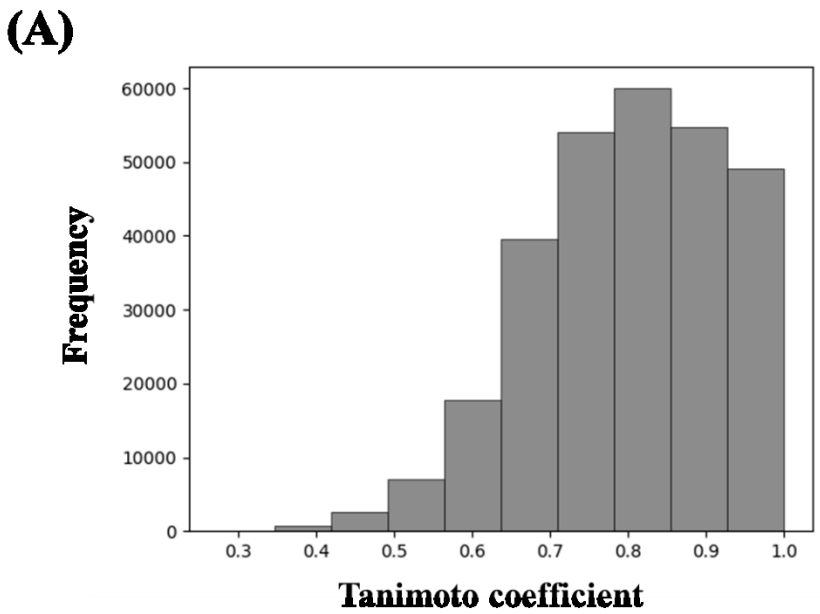
misclassified by all models. This difficulty in prediction may be due to the lack of related compounds in training datasets. Prediction results vary significantly for 4 of 66 reactions (EC 1.13.11.12, EC 1.13.11.58, EC 1.13.11.60, EC 1.13.11.62) that incorporate 2 oxygen atoms into linoleate. Only the latter 2 reactions, which incorporate oxygen into C8 and C10, respectively, are incorrectly predicted as negative by most of the SEP models. This indicates that slight differences in enzymes and products of SEP combinations can significantly influence prediction results. Out of 35 monooxygenase reactions of EC 1.14.13.X, 19 are correctly predicted, while the remaining reactions are predicted with varying results depending on the model (Figure S11C). The SEP-kNN model predicts the most reactions that are misjudged by other models.

L-glutamine is converted to L-glutamate by both EC 3.5.1.2 and EC 4.1.3.27, although the mechanisms of each class are different. However, EC 3.5.1.2 reactions are predicted as EC 4 by all models. Since the models classify reactions based on EC number, it is more difficult to predict reactions for multiple EC numbers that include the same compounds. Moreover, for 186 hexosyltransferases in EC 2.4.1.129 and EC 2.4.1.227 that recognize peptidoglycan substrate, all reactions are incorrectly predicted. This may be because the training data does not contain similar compounds.

For 73 phosphotransferases in EC 2.7.2.X that use a carboxy group as an acceptor, the SEP-DNN model exhibits the highest prediction accuracy, while SEP-kNN and SEP-RF models misjudge the same reactions (Figure S11D). Here, the predictions are correct for all models in 3 of the 4 clusters (Figure S11D). EC 3.4.11.X aminopeptidase reactions are often misclassified in 1 of the 3 clusters by SEP-MLP and SEP-RF models (Figure

S11E). The low accuracies of SEP-RF model for EC 4.3.2.X prediction, and of all machine learning algorithms for EC 3.1.2.X, EC 6.2.1.X, and EC 6.3.2.X prediction, are attributed to the lack of related compounds in the training data. These overall results show that SEP models with various machine learning algorithms exhibit similar prediction trends for most reactions, while some reactions are easier to predict than others depending on the specific machine learning algorithm. In particular, enzymatic reaction prediction involving compounds that are not included in the training data is generally more difficult, although some of these types of reactions are still correctly predicted.

III.5.4. Supplementary Figures and Tables



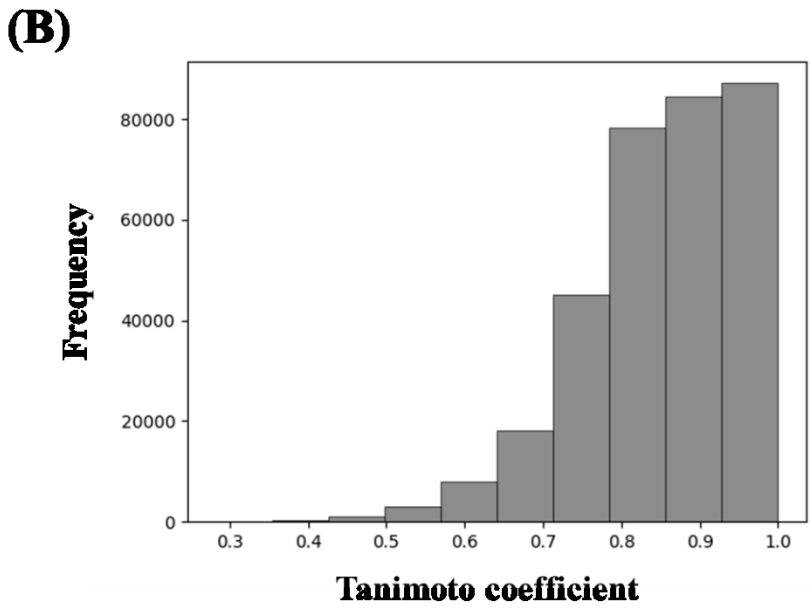
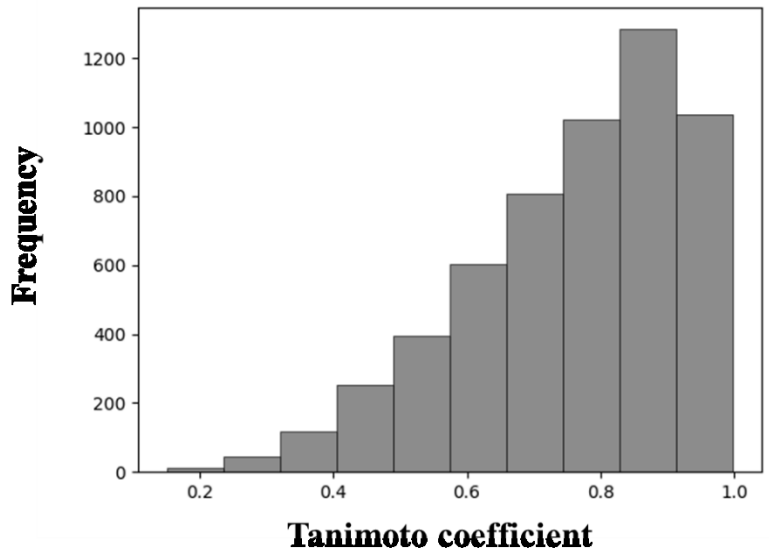
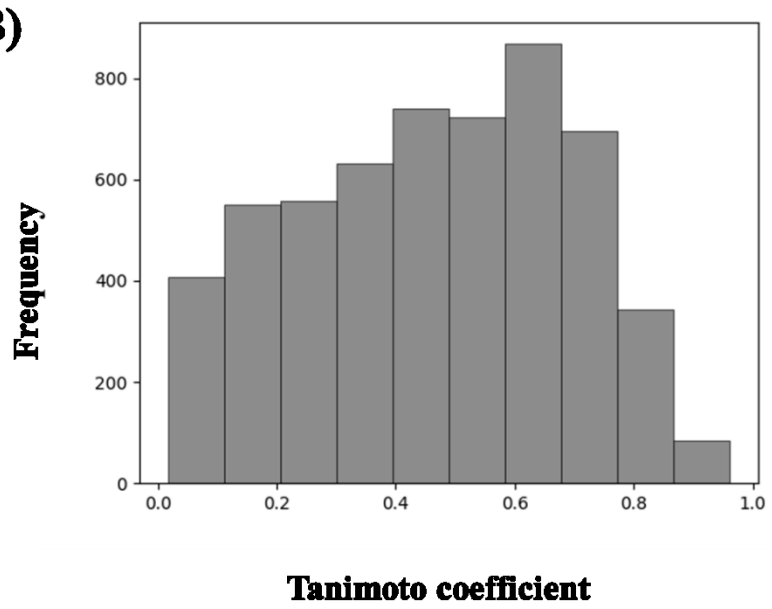


Figure S3. Quantitative similarity distributions of the largest negative training datasets (Table 6 dataset 8), for (A) SE and (B) SEP models. Tanimoto coefficients are calculated for each negative sample paired with the corresponding positive sample of maximum similarity.

(A)


Frequency
1200
1000
800
600
400
200
0
0.2
0.4
0.6
0.8
1.0
Tanimoto coefficient

(B)


Frequency
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Tanimoto coefficient

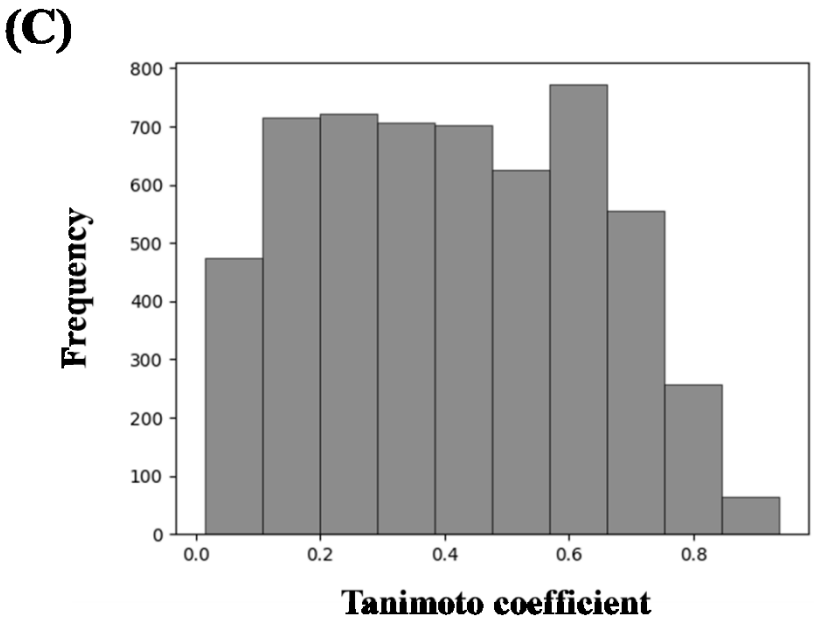


Figure S4. Quantitative similarity distributions of independent test datasets for (A) E, (B) SE and (C) SEP models. Tanimoto coefficients are calculated for each independent test sample paired with the corresponding training sample of maximum similarity.

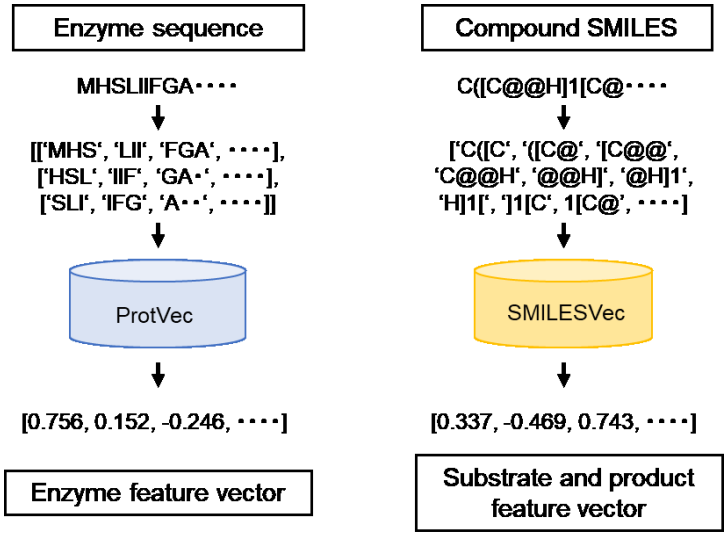
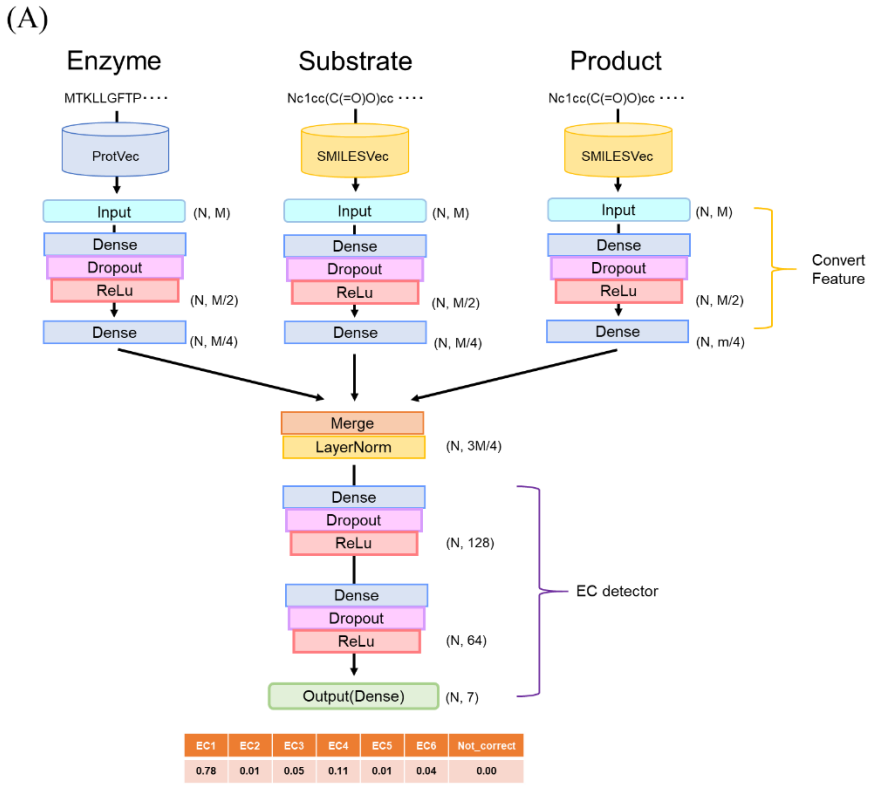


Figure S5. Flow chart for building enzyme feature vectors and substrate/product feature vectors.



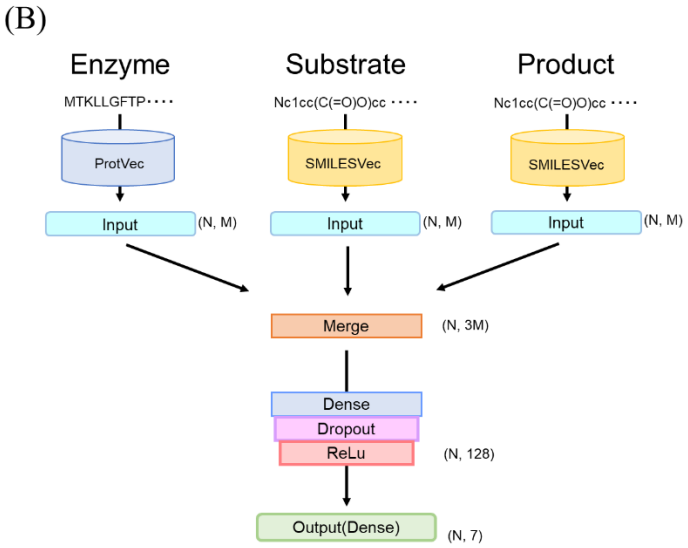


Figure S6. (A) SEP-DNN model and (B) SEP-MLP model structure derived from enzyme ProtVec feature vectors and compound SMILESVec feature vectors. N

represents the amount of data and M represents the number of dimensions for feature vectors (M = 256, 512, 1,024, 2,048).

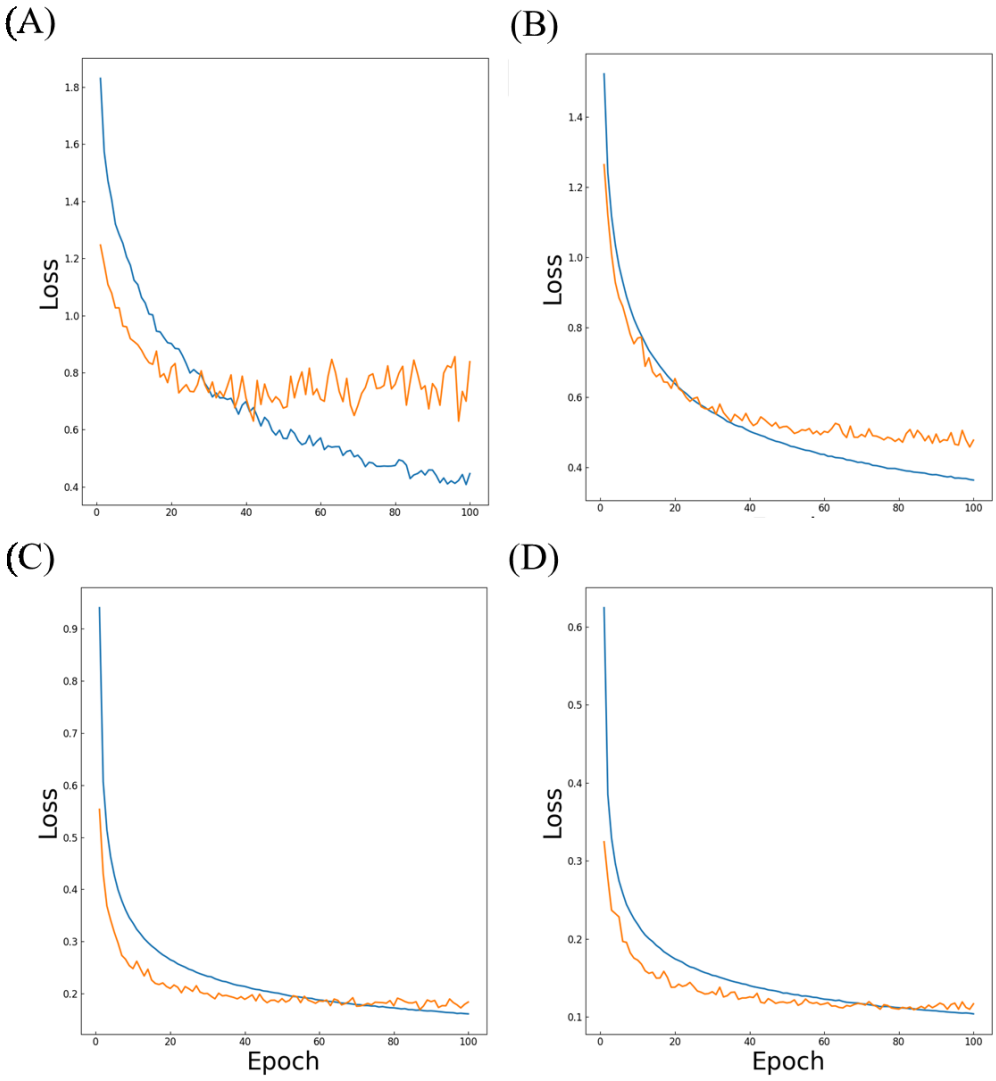


Figure S7. Training and validation loss curves (blue line: training, orange line: validation) of DNN models for 100 epochs. Curves for (A) SP, (B) E, (C) SE, and (D) SEP models are shown.

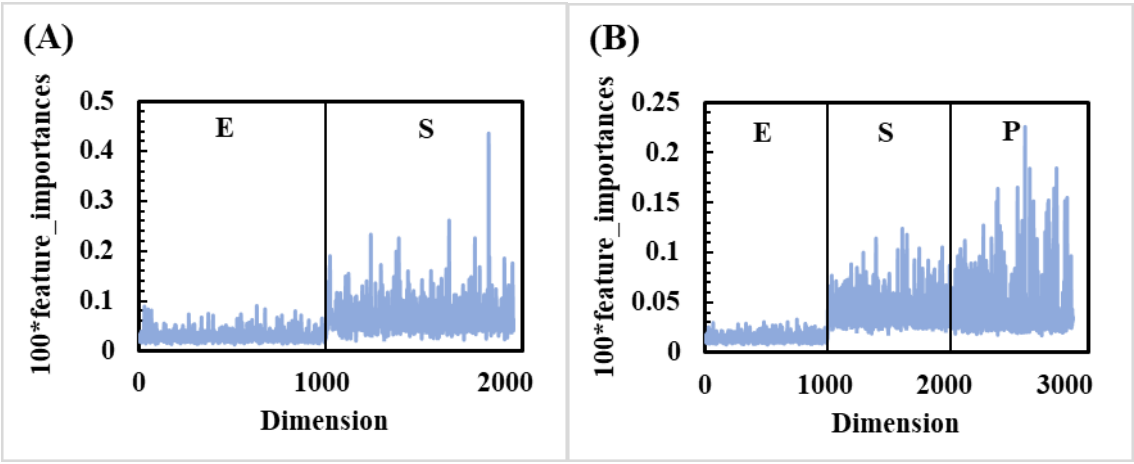
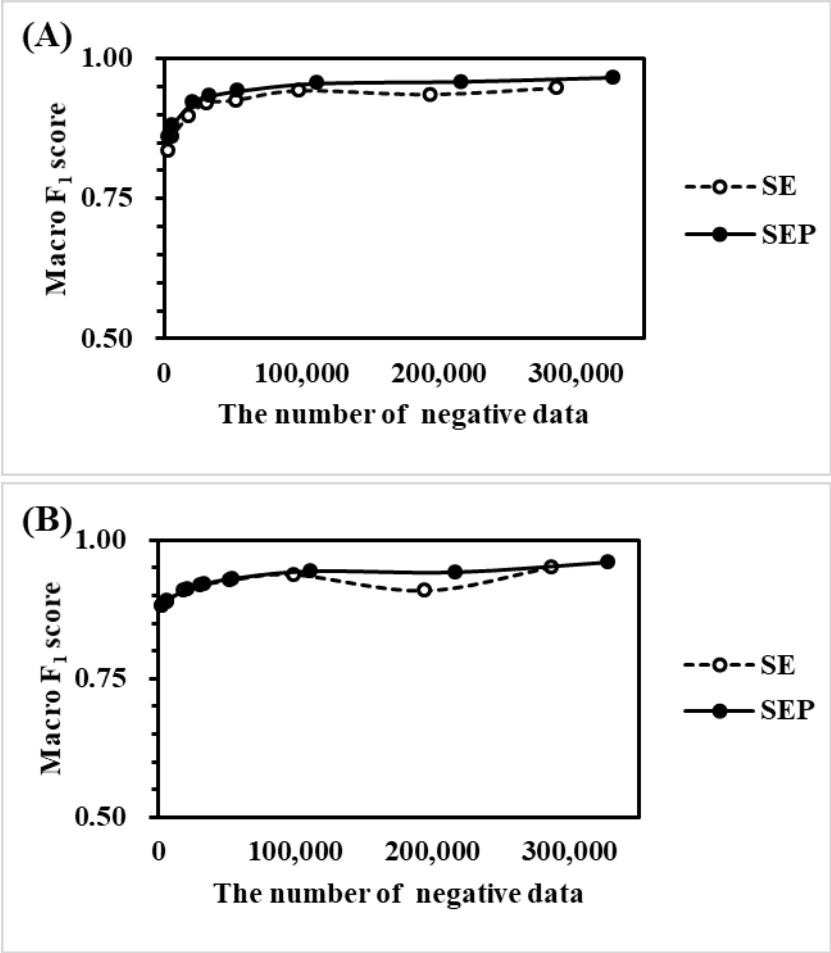


Figure S8 Impurity-based feature importance values in (A) SE-RF, and (B) SEP-RF models. The feature importance values are multiplied by 100.



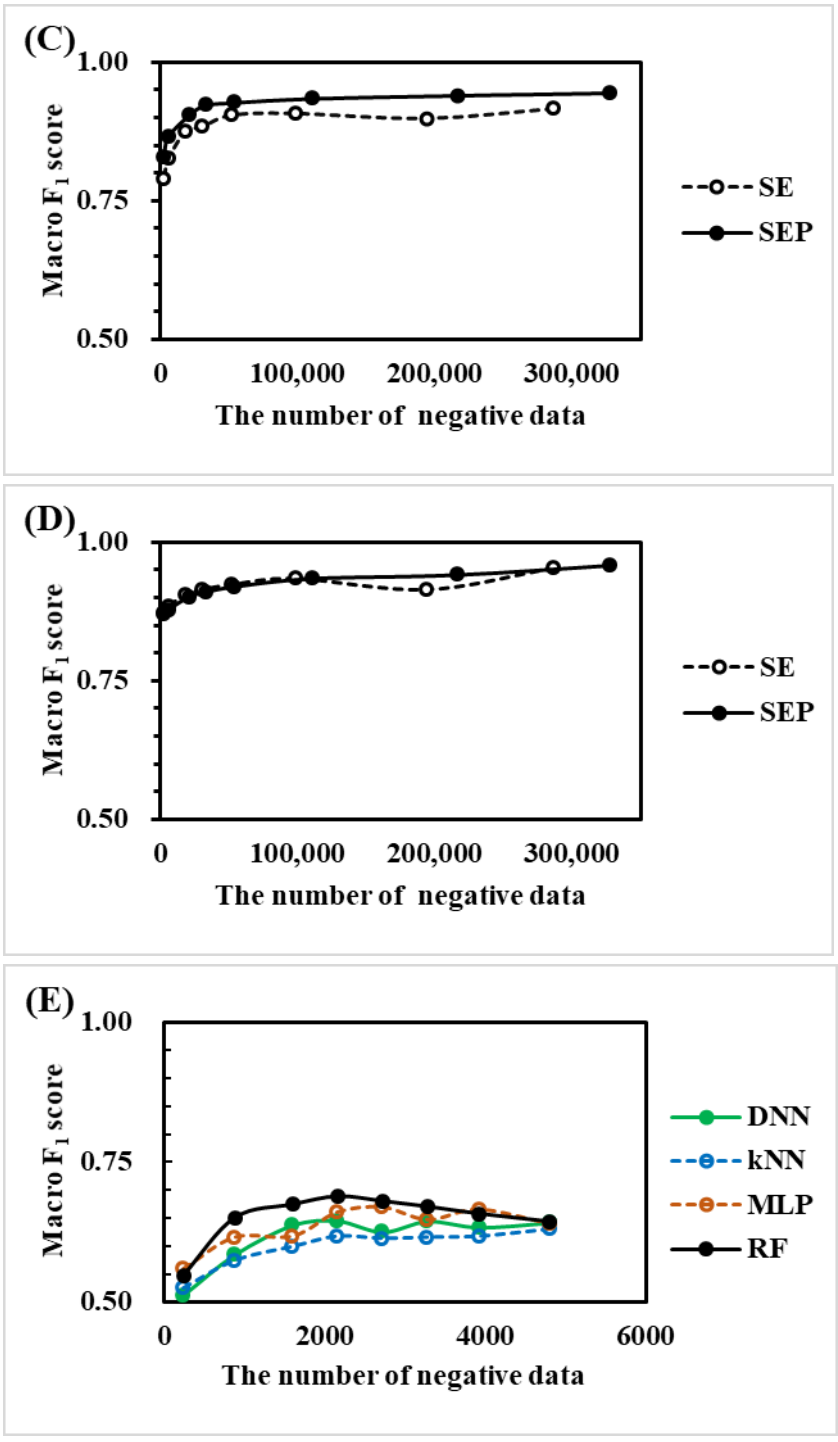
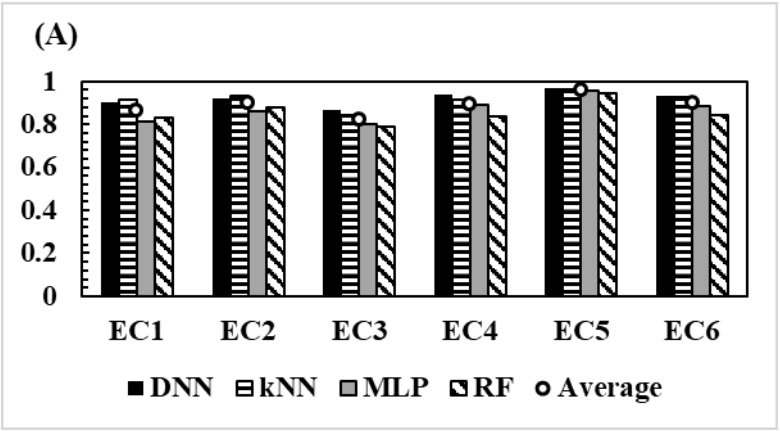
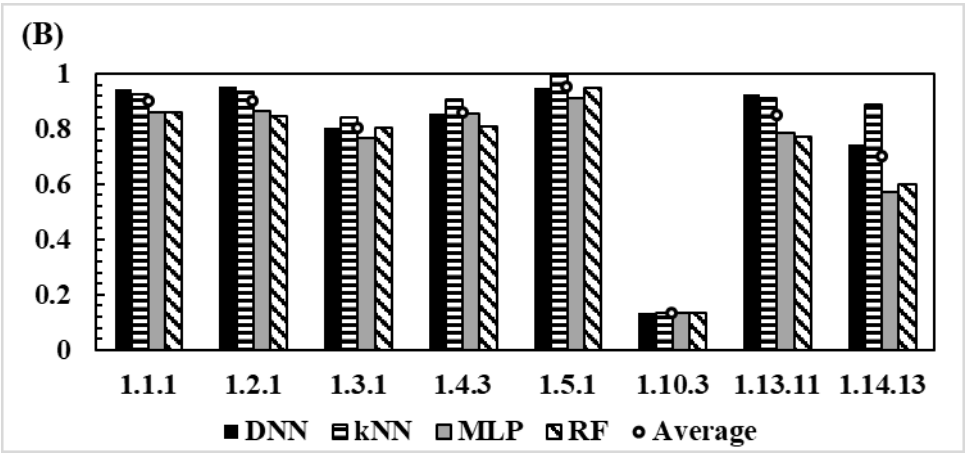
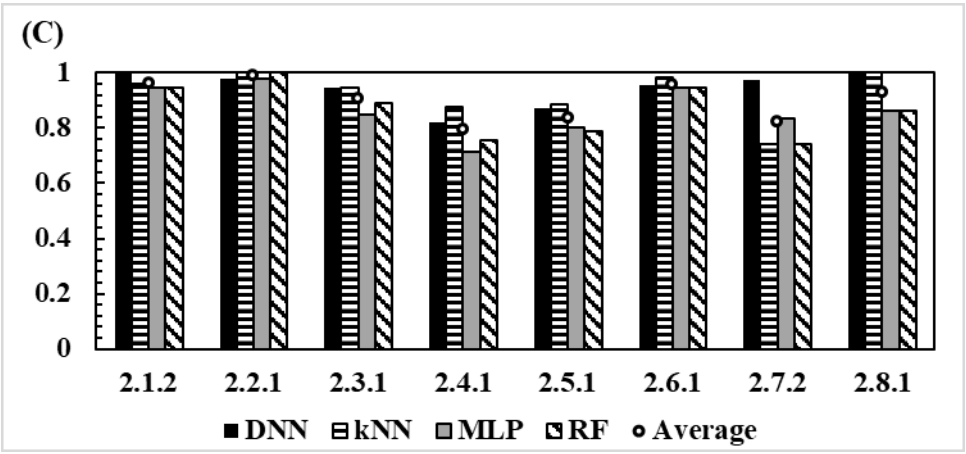


Figure S9. Macro F_1 score of test results with varying negative dataset sizes: (A) SE-DNN and SEP-DNN, (B) SE-kNN and SEP-kNN, (C) SE-MLP and SEP-MLP, (D) SE-RF and SEP-RF, and (E) SP models with 4 machine learning algorithms.


(A)
1
0.8
0.6
0.4
0.2
0
EC1
EC2
EC3
EC4
EC5
EC6
DNN
kNN
MLP
RF
Average


(B)
1
0.8
0.6
0.4
0.2
0
1.1.1
1.2.1
1.3.1
1.4.3
1.5.1
1.10.3
1.13.11
1.14.13
DNN
kNN
MLP
RF
Average


(C)
1
0.8
0.6
0.4
0.2
0
2.1.2
2.2.1
2.3.1
2.4.1
2.5.1
2.6.1
2.7.2
2.8.1
DNN
kNN
MLP
RF
Average

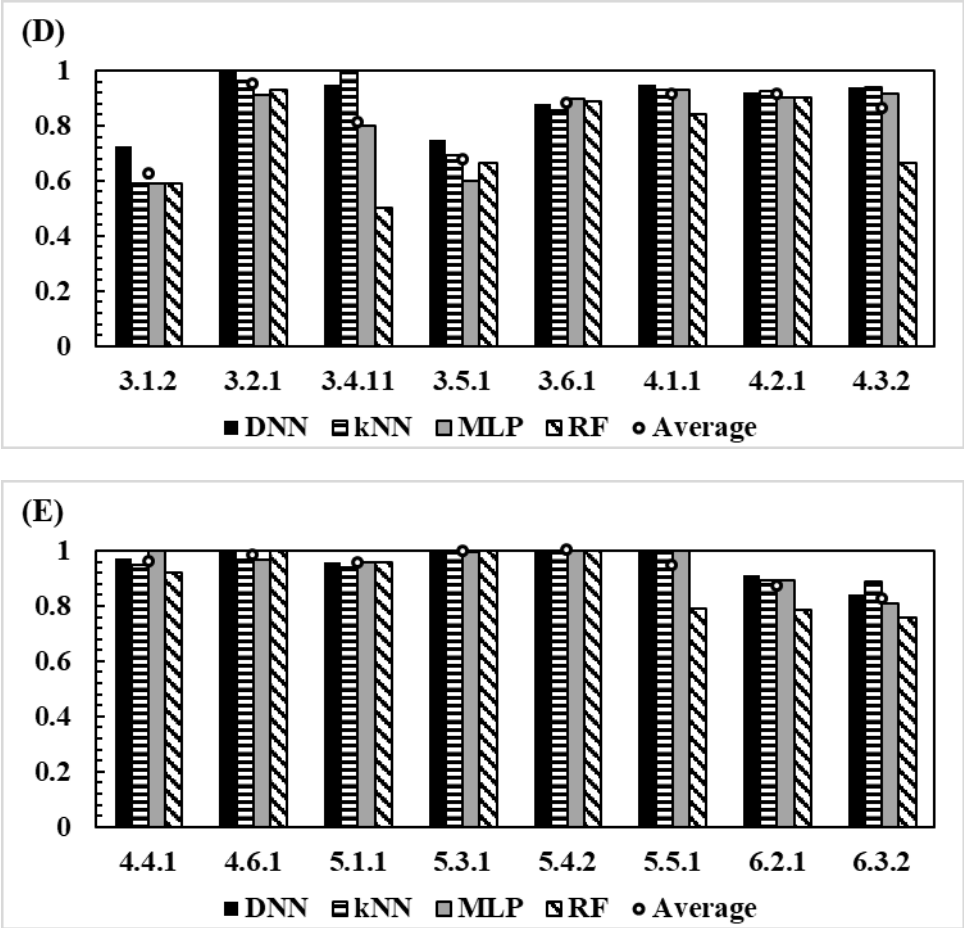
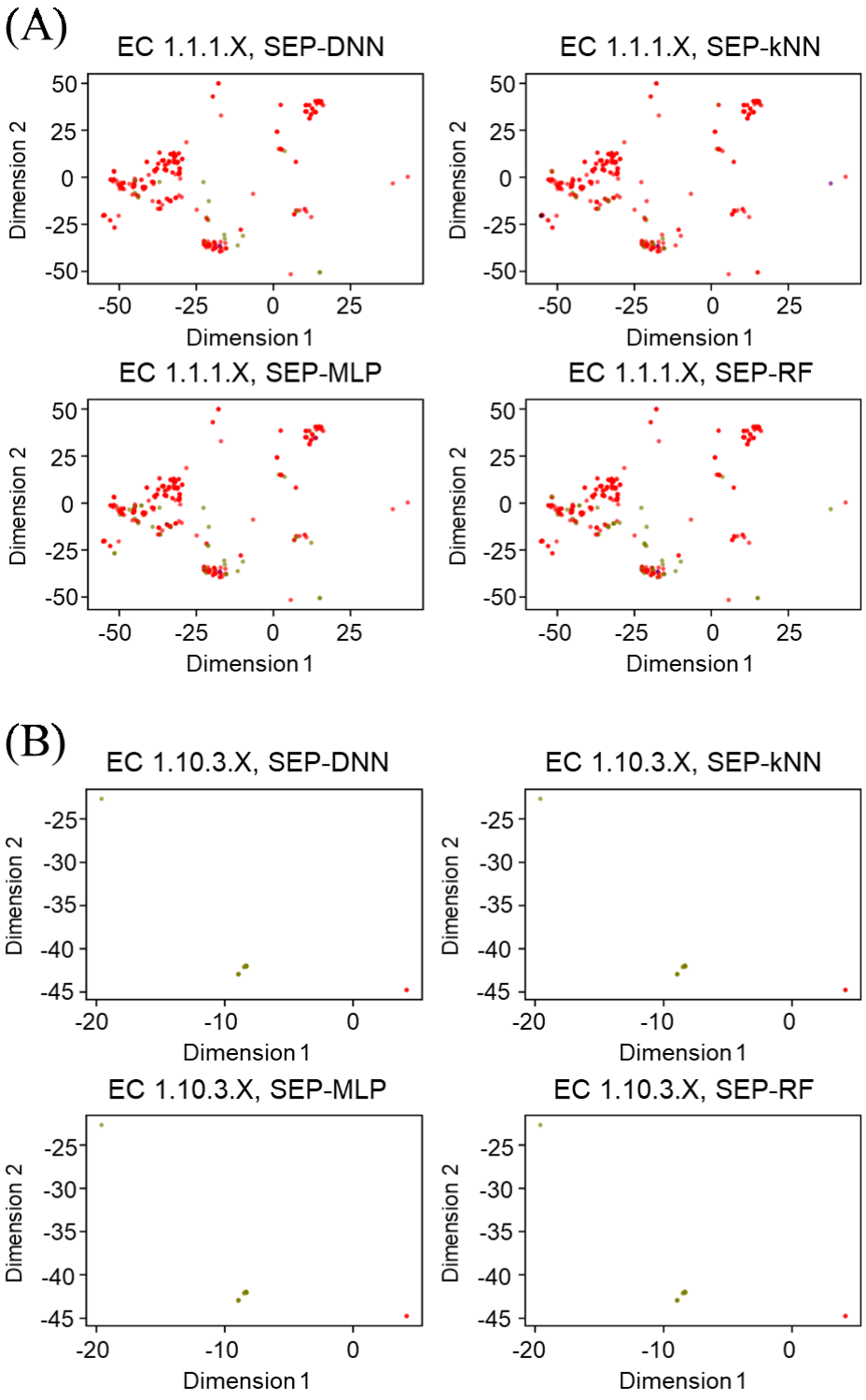
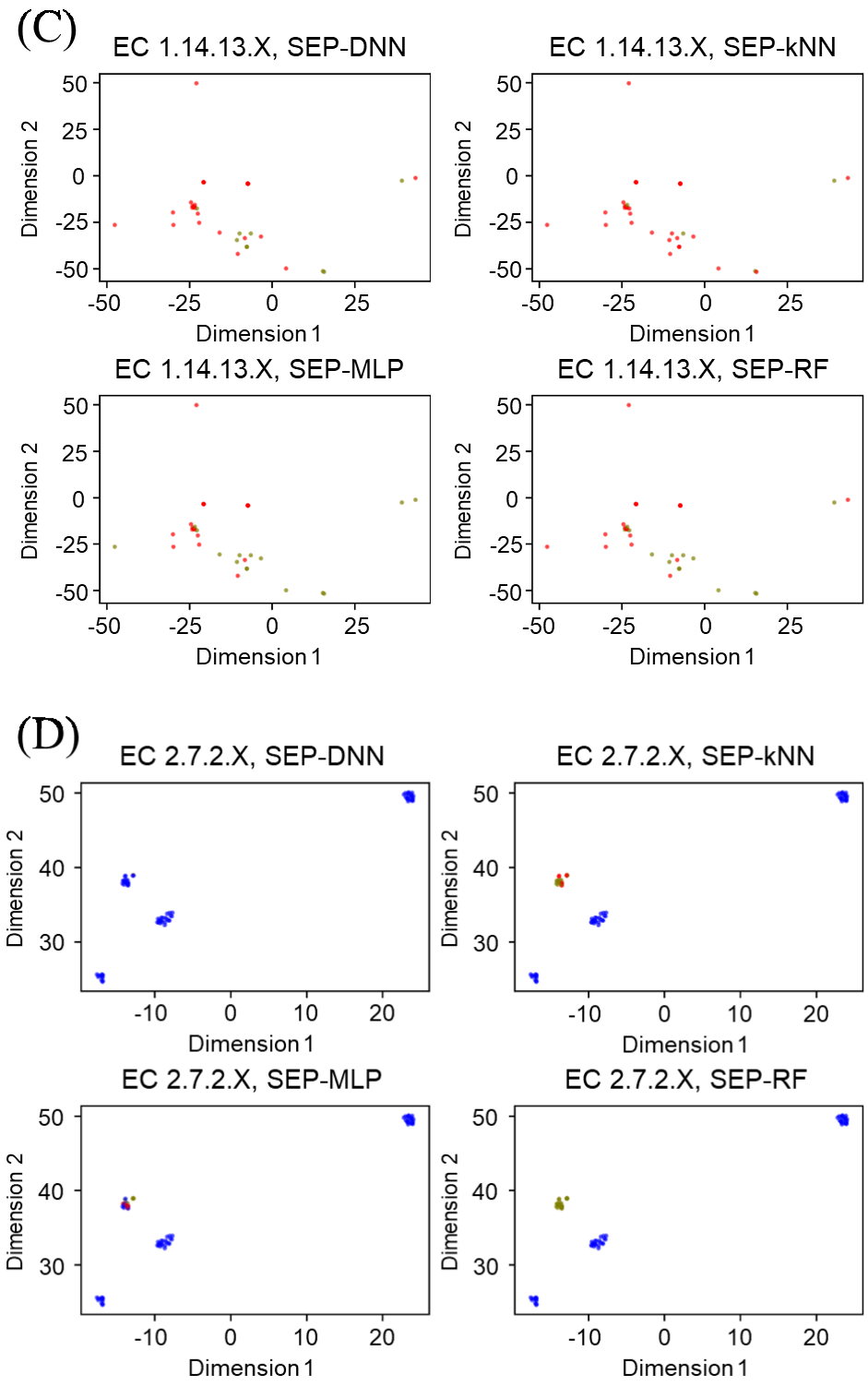


Figure S10. Independent test Recall for EC number first digit groups and third digit groups using SEP models with each machine learning algorithm: (A) All EC number first digits, (B) ~ (E) selected EC 1 ~ EC 6 third digit groups.


(A)
EC 1.1.1.X, SEP-DNN
EC 1.1.1.X, SEP-kNN
EC 1.1.1.X, SEP-MLP
EC 1.1.1.X, SEP-RF
Dimension 1
Dimension 2
(B)
EC 1.10.3.X, SEP-DNN
EC 1.10.3.X, SEP-kNN
EC 1.10.3.X, SEP-MLP
EC 1.10.3.X, SEP-RF
Dimension 1
Dimension 2


(C)
EC 1.14.13.X, SEP-DNN
EC 1.14.13.X, SEP-kNN
EC 1.14.13.X, SEP-MLP
EC 1.14.13.X, SEP-RF
Dimension 1
Dimension 2
(D)
EC 2.7.2.X, SEP-DNN
EC 2.7.2.X, SEP-kNN
EC 2.7.2.X, SEP-MLP
EC 2.7.2.X, SEP-RF
Dimension 1
Dimension 2

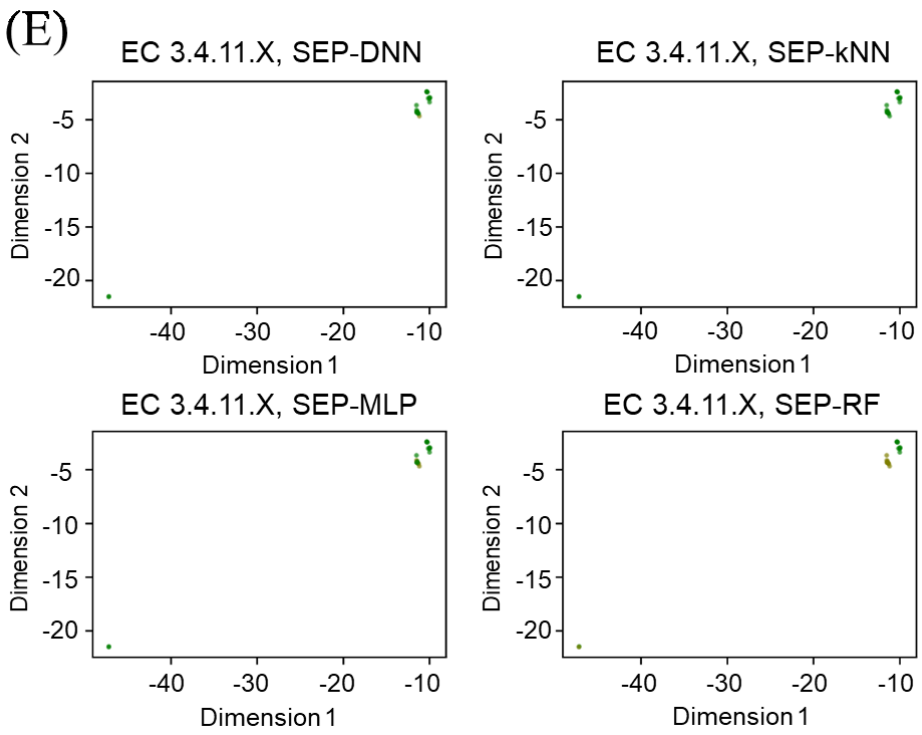


Figure S11. Visualization of independent test results of SEP-DNN (upper left), SEP-kNN (upper right), SEP-MLP (lower left), and SEP-RF (lower right): (A) EC 1.1.1.X, (B) 1.10.3.X, (C) 1.14.13.X, (D) 2.7.2.X, and (E) 3.4.11.X. Hidden layer vectors are derived from the SEP-DNN model in epoch 90 using t-SNE[171]. Red points are predicted as EC 1, blue points are predicted as EC 2, green points are predicted as EC 3, orange points are predicted as EC 6 and olive points are predicted as negative.

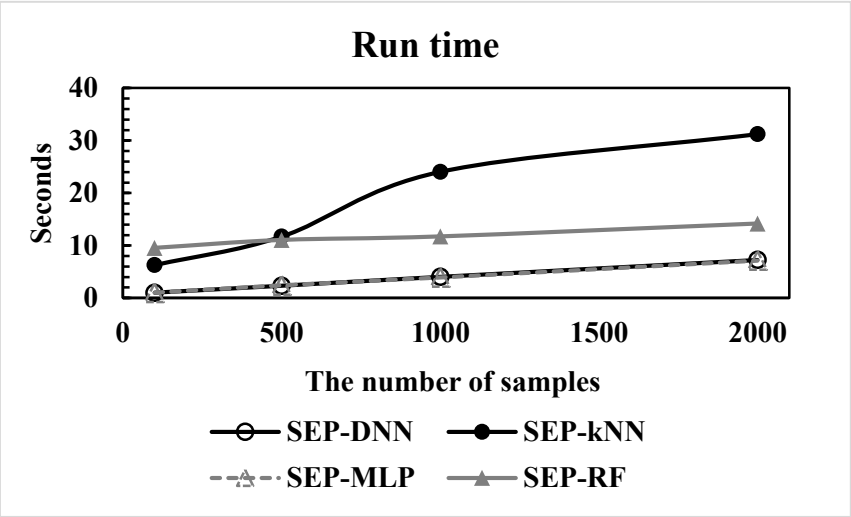


Figure S12. Run time of SEP models

Table S7. The number of neighbors (*NN*) and Power parameter (*P*) for the Minkowski Metric in k-Nearest Neighbor (kNN) Models.

E, SP, SE, SEP	
NN	*P*
1	Euclidean

Other hyper-parameters are used as default[119]. The same parameters are used in all kNN models.

Table S8. The number of Trees (*NT*), *criterion* for the Information Gain in Random Forests (RF) Models.

E		SP		SE		SEP	
NT	*criterion*	*NT*	*criterion*	*NT*	*criterion*	*NT*	*criterion*
250	Entropy	450	Entropy	50	Entropy	650	Entropy

E		SP		SE		SEP	
NT	*criterion*	*NT*	*criterion*	*NT*	*criterion*	*NT*	*criterion*
500	Gini	300	Entropy	500	Entropy	500	Entropy

The other hyper-parameters are default[119]. The upper table shows the parameters in the first models, while the lower table shows in the optimized models.

Table S9. Initial Model Test Evaluations for E, SP, SE and SEP Models Built Using 4 Machine Learning Algorithms.

DNN Test Results									
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**	**Negative**
E	**F_1 score**	**0.842**	**0.873**	**0.845**	**0.781**	**0.831**	**0.865**	**0.858**	-
	Recall	**0.845**	**0.849**	**0.824**	**0.874**	**0.823**	**0.828**	**0.870**	-
	Precision	**0.844**	**0.899**	**0.867**	**0.706**	**0.839**	**0.905**	**0.847**	-
	AUC	**0.976**	**0.975**	**0.965**	**0.972**	**0.976**	**0.984**	**0.986**	-
SP	**F_1 score**	**0.620**	**0.747**	**0.633**	**0.671**	**0.574**	**0.429**	**0.453**	**0.835**
	Recall	**0.759**	**0.771**	**0.623**	**0.825**	**0.744**	**0.857**	**0.750**	**0.745**
	Precision	**0.566**	**0.725**	**0.644**	**0.566**	**0.468**	**0.286**	**0.324**	**0.951**
	AUC	**0.943**	**0.947**	**0.913**	**0.957**	**0.938**	**0.948**	**0.953**	**0.941**
SE	**F_1 score**	**0.935**	**0.955**	**0.958**	**0.960**	**0.930**	**0.943**	**0.948**	**0.853**
	Recall	**0.944**	**0.942**	**0.961**	**0.965**	**0.961**	**0.971**	**0.982**	**0.828**
	Precision	**0.927**	**0.969**	**0.955**	**0.955**	**0.900**	**0.917**	**0.916**	**0.879**
	AUC	**0.994**	**0.996**	**0.996**	**0.998**	**0.997**	**0.999**	**1.000**	**0.973**
SEP	**F_1 score**	**0.959**	**0.977**	**0.977**	**0.974**	**0.976**	**0.974**	**0.975**	**0.863**
	Recall	**0.962**	**0.985**	**0.987**	**0.991**	**0.989**	**0.991**	**0.994**	**0.801**
	Precision	**0.958**	**0.968**	**0.967**	**0.958**	**0.964**	**0.957**	**0.956**	**0.935**
	AUC	**0.997**	**0.999**	**0.999**	**0.999**	**1.000**	**1.000**	**1.000**	**0.980**

kNN Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.884	0.916	0.904	0.867	0.886	0.865	0.861	–
	Recall	0.883	0.918	0.907	0.849	0.894	0.890	0.841	–
	Precision	0.884	0.914	0.902	0.887	0.880	0.840	0.883	–
	AUC	0.930	0.943	0.929	0.914	0.938	0.939	0.917	–
SP	F_1 score	0.617	0.762	0.611	0.544	0.506	0.533	0.500	0.841
	Recall	0.644	0.818	0.689	0.596	0.564	0.571	0.500	0.771
	Precision	0.592	0.714	0.549	0.500	0.458	0.500	0.500	0.924
	AUC	0.799	0.866	0.802	0.779	0.767	0.781	0.746	0.850
SE	F_1 score	0.909	0.940	0.952	0.952	0.913	0.929	0.919	0.753
	Recall	0.909	0.944	0.962	0.960	0.923	0.944	0.920	0.713
	Precision	0.908	0.937	0.941	0.944	0.905	0.914	0.919	0.797
	AUC	0.947	0.960	0.969	0.976	0.956	0.970	0.958	0.843
SEP	F_1 score	0.941	0.965	0.970	0.967	0.955	0.952	0.959	0.804
	Recall	0.942	0.986	0.988	0.987	0.972	0.974	0.976	0.709
	Precision	0.941	0.944	0.953	0.949	0.938	0.931	0.943	0.928
	AUC	0.966	0.983	0.985	0.989	0.983	0.986	0.987	0.850

MLP Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.728	0.791	0.729	0.704	0.720	0.658	0.765	–
	Recall	0.775	0.752	0.652	0.752	0.772	0.865	0.857	–
	Precision	0.704	0.836	0.828	0.663	0.675	0.531	0.691	–
	AUC	0.947	0.939	0.914	0.935	0.946	0.969	0.976	–
SP	F_1 score	0.658	0.785	0.675	0.638	0.602	0.514	0.500	0.894
	Recall	0.714	0.781	0.656	0.789	0.641	0.643	0.625	0.859
	Precision	0.624	0.789	0.696	0.536	0.568	0.429	0.417	0.931
	AUC	0.942	0.957	0.932	0.964	0.950	0.955	0.883	0.951
SE	F_1 score	0.898	0.937	0.924	0.934	0.911	0.906	0.914	0.760
	Recall	0.917	0.929	0.906	0.966	0.944	0.975	0.977	0.724
	Precision	0.883	0.946	0.944	0.904	0.881	0.847	0.859	0.798
	AUC	0.988	0.993	0.991	0.997	0.997	0.999	0.999	0.941
SEP	F_1 score	0.940	0.965	0.962	0.967	0.961	0.962	0.962	0.797
	Recall	0.947	0.963	0.965	0.988	0.977	0.988	0.991	0.756
	Precision	0.933	0.967	0.959	0.946	0.946	0.938	0.935	0.842
	AUC	0.992	0.998	0.996	0.999	0.999	0.999	0.999	0.952

RF Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.565	0.631	0.625	0.408	0.410	0.427	0.429	–
	Recall	0.436	0.662	0.874	0.273	0.261	0.274	0.274	–
	Precision	0.803	0.603	0.486	0.806	0.962	0.970	0.990	–
	AUC	0.867	0.844	0.838	0.848	0.881	0.904	0.887	–
SP	F_1 score	0.658	0.766	0.584	0.581	0.580	0.571	0.560	0.869
	Recall	0.580	0.766	0.500	0.474	0.513	0.429	0.438	0.939
	Precision	0.761	0.766	0.701	0.750	0.667	0.857	0.778	0.808
	AUC	0.932	0.944	0.892	0.944	0.923	0.922	0.957	0.943
SE	F_1 score	0.914	0.943	0.941	0.942	0.919	0.928	0.919	0.790
	Recall	0.910	0.958	0.964	0.960	0.928	0.938	0.930	0.695
	Precision	0.918	0.927	0.919	0.924	0.911	0.919	0.909	0.916
	AUC	0.992	0.993	0.994	0.998	0.993	0.997	0.999	0.969
SEP	F_1 score	0.940	0.965	0.966	0.964	0.964	0.965	0.955	0.782
	Recall	0.941	0.985	0.987	0.989	0.984	0.983	0.983	0.678
	Precision	0.939	0.945	0.947	0.941	0.945	0.948	0.928	0.922
	AUC	0.996	0.998	0.998	0.999	0.999	1.000	1.000	0.978

Table S10. Initial Model Independent Test Evaluations for E, SP, SE and SEP Models Built Using 4 Machine Learning Algorithms.

DNN Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.876	0.889	0.894	0.772	0.864	0.907	0.931
	Recall	0.884	0.886	0.870	0.835	0.858	0.905	0.952
	Precision	0.870	0.892	0.919	0.717	0.870	0.910	0.911
	AUC	0.980	0.984	0.979	0.958	0.975	0.986	0.996
SP	F_1 score	0.659	0.705	0.614	0.685	0.692	0.457	0.629
	Recall	0.694	0.692	0.492	0.630	0.741	0.762	0.848
	Precision	0.627	0.719	0.815	0.750	0.649	0.327	0.500
	AUC	0.904	0.884	0.858	0.913	0.902	0.910	0.957
SE	F_1 score	0.936	0.916	0.941	0.913	0.947	0.967	0.927
	Recall	0.920	0.863	0.905	0.877	0.945	0.974	0.955
	Precision	0.953	0.976	0.980	0.952	0.949	0.961	0.901
	AUC	0.992	0.993	0.992	0.988	0.991	0.995	0.992
SEP	F_1 score	0.950	0.942	0.954	0.936	0.949	0.981	0.932
	Recall	0.931	0.905	0.934	0.892	0.929	0.971	0.953
	Precision	0.969	0.982	0.975	0.984	0.970	0.992	0.912
	AUC	0.996	0.997	0.993	0.995	0.995	0.999	0.999

kNN Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.921	0.929	0.940	0.874	0.918	0.921	0.941
	Recall	0.922	0.923	0.946	0.835	0.934	0.945	0.947
	Precision	0.921	0.935	0.935	0.918	0.902	0.899	0.935
	AUC	0.953	0.952	0.954	0.912	0.958	0.968	0.971
SP	F_1 score	0.591	0.692	0.648	0.519	0.455	0.565	0.607
	Recall	0.548	0.746	0.594	0.403	0.412	0.619	0.515
	Precision	0.642	0.645	0.713	0.727	0.507	0.520	0.739
	AUC	0.741	0.789	0.744	0.684	0.675	0.800	0.753
SE	F_1 score	0.944	0.943	0.956	0.925	0.947	0.978	0.913
	Recall	0.928	0.925	0.938	0.883	0.949	0.976	0.898
	Precision	0.960	0.961	0.975	0.972	0.945	0.979	0.928
	AUC	0.960	0.957	0.962	0.940	0.970	0.987	0.946
SEP	F_1 score	0.938	0.939	0.946	0.905	0.939	0.967	0.932
	Recall	0.919	0.920	0.930	0.850	0.913	0.966	0.933
	Precision	0.959	0.959	0.963	0.968	0.968	0.968	0.930
	AUC	0.955	0.954	0.955	0.923	0.954	0.982	0.964

MLP Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.764	0.799	0.774	0.685	0.791	0.684	0.854
	Recall	0.811	0.788	0.693	0.713	0.830	0.908	0.932
	Precision	0.740	0.809	0.877	0.660	0.755	0.549	0.788
	AUC	0.953	0.952	0.935	0.923	0.952	0.975	0.984
SP	F_1 score	0.681	0.725	0.673	0.676	0.699	0.571	0.683
	Recall	0.669	0.741	0.563	0.622	0.671	0.571	0.848
	Precision	0.693	0.710	0.835	0.740	0.731	0.571	0.571
	AUC	0.898	0.900	0.843	0.879	0.882	0.932	0.952
SE	F_1 score	0.907	0.882	0.903	0.894	0.934	0.947	0.877
	Recall	0.888	0.822	0.850	0.874	0.927	0.968	0.885
	Precision	0.928	0.951	0.963	0.915	0.941	0.927	0.868
	AUC	0.980	0.971	0.982	0.980	0.980	0.992	0.973
SEP	F_1 score	0.923	0.900	0.930	0.892	0.935	0.972	0.905
	Recall	0.877	0.827	0.873	0.822	0.894	0.953	0.895
	Precision	0.974	0.988	0.995	0.976	0.979	0.992	0.916
	AUC	0.985	0.978	0.979	0.983	0.989	0.994	0.989

RF Independent Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.590	0.634	0.681	0.400	0.472	0.412	0.491
	Recall	0.463	0.726	0.888	0.260	0.311	0.264	0.327
	Precision	0.814	0.563	0.552	0.862	0.978	0.935	0.992
	AUC	0.905	0.888	0.876	0.867	0.918	0.931	0.948
SP	F_1 score	0.529	0.655	0.507	0.425	0.504	0.444	0.500
	Recall	0.399	0.681	0.371	0.286	0.376	0.286	0.394
	Precision	0.785	0.630	0.802	0.829	0.762	1.000	0.684
	AUC	0.899	0.885	0.877	0.892	0.896	0.896	0.945
SE	F_1 score	0.929	0.910	0.934	0.920	0.944	0.967	0.898
	Recall	0.892	0.853	0.905	0.871	0.915	0.953	0.855
	Precision	0.969	0.975	0.965	0.974	0.975	0.981	0.945
	AUC	0.986	0.989	0.986	0.981	0.988	0.992	0.978
SEP	F_1 score	0.917	0.913	0.935	0.884	0.911	0.970	0.881
	Recall	0.861	0.844	0.881	0.797	0.847	0.950	0.848
	Precision	0.980	0.995	0.997	0.993	0.986	0.992	0.916
	AUC	0.997	0.997	0.997	0.993	0.997	0.998	0.998

Table S11. Initial Model Additional Test Evaluations for E, SE and SEP Models Built Using 4 Machine Learning Algorithms to Compare to the Chapter II.

DNN Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.759	0.667	0.800	0.571	0.814	0.900	0.800
	Recall	0.781	0.581	0.761	0.875	0.923	0.818	0.727
	Precision	0.778	0.781	0.843	0.424	0.727	1.000	0.889
	AUC	0.926	0.911	0.900	0.969	0.978	0.973	0.826
SE	F_1 score	0.904	0.744	0.891	0.936	0.882	0.968	0.966
	Recall	0.862	0.604	0.843	0.917	0.872	0.938	1.000
	Precision	0.949	0.970	0.944	0.957	0.891	1.000	0.933
	AUC	0.965	0.888	0.975	0.995	0.968	0.967	1.000
SEP	F_1 score	0.909	0.737	0.919	0.941	0.964	0.875	1.000
	Recall	0.868	0.625	0.869	0.960	0.930	0.824	1.000
	Precision	0.955	0.897	0.974	0.923	1.000	0.933	1.000
	AUC	0.990	0.966	0.990	0.999	0.996	0.985	1.000

kNN Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.793	0.727	0.804	0.649	0.794	0.870	0.842
	Recall	0.797	0.651	0.783	0.750	0.962	0.909	0.727
	Precision	0.789	0.824	0.828	0.571	0.676	0.833	1.000
	AUC	0.873	0.806	0.821	0.850	0.946	0.949	0.864
SE	F_1 score	0.833	0.733	0.827	0.704	0.833	0.903	0.963
	Recall	0.794	0.623	0.762	0.792	0.833	0.824	0.929
	Precision	0.877	0.892	0.903	0.633	0.833	1.000	1.000
	AUC	0.884	0.802	0.849	0.874	0.899	0.912	0.964
SEP	F_1 score	0.877	0.800	0.837	0.750	0.906	0.938	1.000
	Recall	0.835	0.714	0.731	0.840	0.842	0.882	1.000
	Precision	0.924	0.909	0.979	0.677	0.980	1.000	1.000
	AUC	0.912	0.849	0.859	0.902	0.919	0.941	1.000

MLP Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.618	0.627	0.671	0.545	0.611	0.526	0.727
	Recall	0.714	0.488	0.565	0.750	0.846	0.909	0.727
	Precision	0.617	0.875	0.825	0.429	0.478	0.370	0.727
	AUC	0.907	0.855	0.836	0.941	0.950	0.942	0.921
SE	F_1 score	0.872	0.659	0.841	0.880	0.872	0.968	0.966
	Recall	0.833	0.528	0.744	0.917	0.872	0.938	1.000
	Precision	0.916	0.875	0.968	0.846	0.872	1.000	0.933
	AUC	0.948	0.819	0.969	0.998	0.964	0.941	0.999
SEP	F_1 score	0.895	0.681	0.844	0.941	0.926	0.938	1.000
	Recall	0.834	0.554	0.731	0.960	0.877	0.882	1.000
	Precision	0.965	0.886	1.000	0.923	0.980	1.000	1.000
	AUC	0.947	0.851	0.898	0.996	0.981	0.957	1.000

RF Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.652	0.575	0.779	0.571	0.700	0.778	0.167
	Recall	0.542	0.581	0.902	0.500	0.538	0.636	0.091
	Precision	0.820	0.568	0.686	0.667	1.000	1.000	1.000
	AUC	0.888	0.830	0.827	0.916	0.952	0.869	0.935
SE	F_1 score	0.867	0.690	0.851	0.889	0.837	0.970	0.929
	Recall	0.799	0.547	0.795	0.833	0.750	0.941	0.929
	Precision	0.946	0.935	0.915	0.952	0.947	1.000	0.929
	AUC	0.965	0.934	0.961	0.968	0.966	0.963	0.999
SEP	F_1 score	0.880	0.727	0.818	0.889	0.871	0.938	1.000
	Recall	0.786	0.571	0.692	0.800	0.772	0.882	1.000
	Precision	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AUC	0.992	0.988	0.982	0.996	0.990	0.993	1.000

Table S12. Macro F_1 Scores for DNN, kNN, MLP and RF Algorithms, Running E, SP, SE and SEP Models Derived from SMILESVec and ProtVec Vectors with 4 Different Numbers of Dimensions: (A) Test Results, (B) Independent Test Results.

(A)						(B)					
Algorithms	Model	256 dim	512 dim	1024 dim	2048 dim	Algorithms	Model	256 dim	512 dim	1024 dim	2048 dim
DNN	E	0.602	0.755	0.842	0.887	DNN	E	0.641	0.808	0.876	0.910
	SP	0.606	0.599	0.620	0.607		SP	0.654	0.658	0.659	0.648
	SE	0.868	0.915	0.935	0.939		SE	0.889	0.927	0.936	0.945
	SEP	0.920	0.946	0.959	0.960		SEP	0.938	0.947	0.950	0.954
kNN	E	0.858	0.901	0.884	0.902	kNN	E	0.858	0.901	0.921	0.924
	SP	0.613	0.621	0.617	0.612		SP	0.605	0.597	0.591	0.576
	SE	0.908	0.910	0.909	0.897		SE	0.941	0.944	0.944	0.947
	SEP	0.937	0.940	0.941	0.939		SEP	0.944	0.937	0.938	0.940
MLP	E	0.607	0.681	0.728	0.798	MLP	E	0.645	0.720	0.764	0.827
	SP	0.662	0.673	0.658	0.625		SP	0.688	0.678	0.681	0.664
	SE	0.868	0.894	0.898	0.909		SE	0.874	0.907	0.907	0.906
	SEP	0.932	0.938	0.940	0.937		SEP	0.917	0.916	0.923	0.928
RF	E	0.618	0.603	0.565	0.538	RF	E	0.643	0.629	0.590	0.551
	SP	0.696	0.669	0.658	0.650		SP	0.583	0.537	0.529	0.533
	SE	0.917	0.915	0.914	0.915		SE	0.935	0.934	0.929	0.932
	SEP	0.943	0.942	0.940	0.938		SEP	0.915	0.915	0.917	0.917

Table S13. Optimal parameters for Each Model.

	DNN				kNN			
	E	SP	SE	SEP	E	SP	SE	SEP
Negative data	-	No.4	No.8	No.8	-	No.8	No.8	No.8
Dimension	2,048	1,024	1,024	1,024	2,048	512	1,024	1,024

	MLP				RF			
	E	SP	SE	SEP	E	SP	SE	SEP
Negative data	-	No.4	No.8	No.8	-	No.5	No.8	No.8
Dimension	2,048	512	1,024	1,024	256	256	256	256

Abbreviation: MLP, Multilayer Perceptron.

Table S14. Optimized Model Test Evaluations for E, SP, SE and SEP Models Built Using 4 Machine Learning Algorithms.

DNN Test Results									
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.887	0.909	0.890	0.860	0.877	0.896	0.889	-
	Recall	0.890	0.894	0.881	0.898	0.882	0.888	0.896	-
	Precision	0.884	0.924	0.898	0.826	0.872	0.903	0.882	-
	AUC	0.986	0.943	0.929	0.914	0.938	0.939	0.917	-
SP	F_1 score	0.645	0.731	0.615	0.695	0.659	0.579	0.388	0.851
	Recall	0.762	0.745	0.648	0.860	0.718	0.786	0.813	0.765
	Precision	0.595	0.719	0.585	0.583	0.609	0.458	0.255	0.959
	AUC	0.947	0.941	0.910	0.960	0.961	0.947	0.962	0.947
SE	F_1 score	0.947	0.968	0.966	0.967	0.948	0.949	0.959	0.870
	Recall	0.955	0.972	0.966	0.989	0.982	0.979	0.992	0.803
	Precision	0.942	0.965	0.966	0.946	0.916	0.921	0.928	0.949
	AUC	0.996	0.998	0.998	0.999	0.999	1.000	1.000	0.979
SEP	F_1 score	0.966	0.982	0.981	0.982	0.979	0.977	0.978	0.887
	Recall	0.970	0.991	0.987	0.994	0.995	0.997	0.997	0.830
	Precision	0.964	0.972	0.975	0.969	0.962	0.959	0.960	0.954
	AUC	0.997	0.999	0.999	0.999	1.000	1.000	1.000	0.984

kNN Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.902	0.933	0.919	0.891	0.904	0.885	0.878	–
	Recall	0.902	0.929	0.925	0.874	0.911	0.907	0.866	–
	Precision	0.902	0.936	0.913	0.909	0.898	0.864	0.891	–
	AUC	0.941	0.953	0.941	0.929	0.948	0.948	0.930	–
SP	F_1 score	0.637	0.770	0.674	0.562	0.537	0.516	0.500	0.874
	Recall	0.646	0.802	0.713	0.596	0.564	0.571	0.438	0.839
	Precision	0.627	0.740	0.640	0.531	0.512	0.471	0.583	0.913
	AUC	0.798	0.851	0.807	0.770	0.769	0.780	0.747	0.862
SE	F_1 score	0.950	0.969	0.976	0.975	0.961	0.962	0.961	0.835
	Recall	0.949	0.982	0.990	0.991	0.979	0.978	0.974	0.752
	Precision	0.951	0.957	0.961	0.959	0.943	0.946	0.948	0.939
	AUC	0.971	0.983	0.987	0.992	0.986	0.988	0.986	0.872
SEP	F_1 score	0.959	0.974	0.979	0.977	0.974	0.970	0.974	0.850
	Recall	0.959	0.994	0.996	0.995	0.992	0.992	0.991	0.755
	Precision	0.959	0.955	0.962	0.960	0.956	0.950	0.957	0.973
	AUC	0.976	0.989	0.990	0.994	0.994	0.995	0.995	0.876

MLP Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.803	0.833	0.809	0.775	0.790	0.782	0.832	–
	Recall	0.821	0.791	0.796	0.799	0.819	0.865	0.854	–
	Precision	0.790	0.880	0.822	0.753	0.762	0.713	0.811	–
	AUC	0.965	0.960	0.947	0.959	0.967	0.979	0.980	–
SP	F_1 score	0.687	0.788	0.693	0.646	0.549	0.710	0.545	0.875
	Recall	0.762	0.823	0.713	0.737	0.718	0.786	0.750	0.804
	Precision	0.641	0.756	0.674	0.575	0.444	0.647	0.429	0.959
	AUC	0.938	0.954	0.926	0.964	0.952	0.960	0.866	0.945
SE	F_1 score	0.915	0.947	0.939	0.949	0.922	0.932	0.927	0.793
	Recall	0.931	0.940	0.936	0.973	0.967	0.983	0.986	0.728
	Precision	0.904	0.954	0.941	0.926	0.880	0.886	0.875	0.869
	AUC	0.991	0.995	0.993	0.998	0.998	0.999	0.999	0.954
SEP	F_1 score	0.944	0.966	0.962	0.971	0.969	0.968	0.961	0.812
	Recall	0.952	0.962	0.963	0.987	0.989	0.993	0.995	0.778
	Precision	0.937	0.971	0.961	0.955	0.950	0.944	0.930	0.848
	AUC	0.992	0.998	0.997	0.999	0.999	1.000	1.000	0.956

RF Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6	Negative
E	F_1 score	0.618	0.677	0.661	0.513	0.514	0.515	0.496	–
	Recall	0.501	0.702	0.886	0.379	0.354	0.353	0.333	–
	Precision	0.805	0.654	0.528	0.794	0.934	0.950	0.968	–
	AUC	0.891	0.876	0.870	0.879	0.902	0.917	0.900	–
SP	F_1 score	0.710	0.794	0.667	0.598	0.667	0.667	0.615	0.907
	Recall	0.659	0.875	0.623	0.509	0.615	0.571	0.500	0.916
	Precision	0.771	0.727	0.717	0.725	0.727	0.800	0.800	0.898
	AUC	0.930	0.947	0.900	0.941	0.918	0.910	0.939	0.954
SE	F_1 score	0.954	0.971	0.974	0.971	0.964	0.967	0.967	0.857
	Recall	0.954	0.980	0.988	0.988	0.979	0.984	0.981	0.780
	Precision	0.955	0.961	0.961	0.955	0.950	0.950	0.953	0.951
	AUC	0.998	0.999	0.999	1.000	0.999	1.000	1.000	0.991
SEP	F_1 score	0.960	0.976	0.978	0.976	0.976	0.977	0.972	0.853
	Recall	0.960	0.990	0.992	0.995	0.991	0.990	0.991	0.773
	Precision	0.959	0.961	0.964	0.959	0.962	0.965	0.954	0.951
	AUC	0.999	0.999	0.999	1.000	1.000	1.000	1.000	0.992

Table S15. Optimized Model Independent Test Evaluations for E, SP, SE and SEP Models Built Using 4 Machine Learning Algorithms.

DNN Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.910	0.919	0.921	0.845	0.900	0.927	0.950
	Recall	0.917	0.913	0.907	0.871	0.909	0.942	0.960
	Precision	0.904	0.925	0.936	0.820	0.891	0.913	0.941
	AUC	0.987	0.991	0.988	0.968	0.985	0.991	0.998
SP	F_1 score	0.654	0.653	0.652	0.706	0.686	0.520	0.624
	Recall	0.676	0.600	0.594	0.655	0.706	0.619	0.879
	Precision	0.634	0.716	0.722	0.765	0.667	0.448	0.483
	AUC	0.905	0.881	0.843	0.919	0.910	0.914	0.965
SE	F_1 score	0.936	0.921	0.940	0.916	0.947	0.963	0.922
	Recall	0.923	0.881	0.904	0.896	0.943	0.971	0.945
	Precision	0.948	0.964	0.980	0.937	0.950	0.956	0.900
	AUC	0.991	0.991	0.992	0.986	0.989	0.994	0.994
SEP	F_1 score	0.947	0.942	0.954	0.918	0.957	0.981	0.926
	Recall	0.923	0.902	0.923	0.866	0.941	0.971	0.933
	Precision	0.972	0.986	0.986	0.977	0.973	0.992	0.919
	AUC	0.996	0.997	0.994	0.991	0.995	0.999	0.998

kNN Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.924	0.926	0.948	0.876	0.925	0.929	0.936
	Recall	0.924	0.920	0.956	0.838	0.935	0.945	0.950
	Precision	0.924	0.933	0.940	0.918	0.915	0.913	0.922
	AUC	0.954	0.951	0.960	0.914	0.960	0.969	0.972
SP	F_1 score	0.605	0.706	0.657	0.538	0.486	0.565	0.593
	Recall	0.551	0.773	0.594	0.412	0.424	0.619	0.485
	Precision	0.669	0.650	0.736	0.778	0.571	0.520	0.762
	AUC	0.745	0.802	0.750	0.692	0.687	0.800	0.738
SE	F_1 score	0.942	0.943	0.955	0.925	0.947	0.979	0.905
	Recall	0.925	0.921	0.935	0.884	0.947	0.974	0.888
	Precision	0.961	0.965	0.975	0.970	0.947	0.984	0.922
	AUC	0.959	0.956	0.961	0.940	0.969	0.986	0.941
SEP	F_1 score	0.939	0.939	0.948	0.904	0.942	0.967	0.929
	Recall	0.915	0.915	0.931	0.840	0.916	0.963	0.925
	Precision	0.964	0.964	0.966	0.978	0.970	0.971	0.932
	AUC	0.954	0.953	0.956	0.919	0.956	0.981	0.960

MLP Independent Test Results								
Model	**Parameter**	**Macro/Average**	**EC1**	**EC2**	**EC3**	**EC4**	**EC5**	**EC6**
E	F_1 score	0.834	0.849	0.848	0.726	0.842	0.824	0.916
	Recall	0.852	0.826	0.822	0.731	0.882	0.910	0.940
	Precision	0.820	0.875	0.875	0.720	0.805	0.753	0.893
	AUC	0.966	0.968	0.960	0.937	0.969	0.981	0.985
SP	F_1 score	0.704	0.708	0.697	0.627	0.678	0.750	0.683
	Recall	0.688	0.768	0.584	0.529	0.682	0.714	0.848
	Precision	0.721	0.657	0.865	0.768	0.674	0.789	0.571
	AUC	0.903	0.880	0.845	0.896	0.881	0.956	0.959
SE	F_1 score	0.906	0.873	0.911	0.898	0.932	0.943	0.872
	Recall	0.887	0.820	0.864	0.858	0.929	0.958	0.895
	Precision	0.925	0.932	0.963	0.942	0.935	0.929	0.851
	AUC	0.978	0.968	0.968	0.981	0.982	0.995	0.974
SEP	F_1 score	0.917	0.884	0.919	0.885	0.932	0.975	0.901
	Recall	0.868	0.815	0.859	0.802	0.892	0.958	0.885
	Precision	0.971	0.966	0.988	0.988	0.975	0.992	0.917
	AUC	0.981	0.970	0.979	0.977	0.988	0.993	0.980

RF Independent Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.643	0.680	0.715	0.518	0.584	0.477	0.570
	Recall	0.530	0.766	0.896	0.374	0.422	0.322	0.399
	Precision	0.819	0.612	0.595	0.847	0.947	0.917	0.994
	AUC	0.926	0.912	0.901	0.899	0.941	0.948	0.954
SP	F_1 score	0.614	0.698	0.591	0.527	0.590	0.500	0.607
	Recall	0.498	0.811	0.477	0.370	0.482	0.333	0.515
	Precision	0.801	0.612	0.777	0.917	0.759	1.000	0.739
	AUC	0.907	0.889	0.874	0.892	0.907	0.914	0.964
SE	F_1 score	0.932	0.908	0.936	0.923	0.943	0.969	0.909
	Recall	0.892	0.844	0.906	0.870	0.912	0.955	0.868
	Precision	0.975	0.983	0.969	0.982	0.977	0.984	0.953
	AUC	0.990	0.992	0.989	0.983	0.993	0.993	0.991
SEP	F_1 score	0.914	0.906	0.933	0.881	0.907	0.969	0.882
	Recall	0.855	0.831	0.877	0.792	0.838	0.947	0.845
	Precision	0.981	0.996	0.997	0.993	0.989	0.992	0.921
	AUC	0.995	0.997	0.996	0.988	0.997	0.997	0.998

Table S16. Optimized Model Additional Test Evaluations for E, SP, SE and SEP Models Built Using 4 Machine Learning Algorithms to Compare to the Chapter II.

DNN Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.778	0.692	0.798	0.682	0.787	0.870	0.842
	Recall	0.812	0.628	0.750	0.938	0.923	0.909	0.727
	Precision	0.780	0.771	0.852	0.536	0.686	0.833	1.000
	AUC	0.941	0.924	0.924	0.987	0.984	0.940	0.886
SE	F_1 score	0.893	0.753	0.898	0.875	0.854	0.968	0.966
	Recall	0.854	0.604	0.835	0.875	0.872	0.938	1.000
	Precision	0.936	1.000	0.971	0.875	0.837	1.000	0.933
	AUC	0.961	0.891	0.982	0.986	0.959	0.948	1.000
SEP	F_1 score	0.926	0.800	0.927	0.939	0.964	0.938	0.966
	Recall	0.884	0.679	0.877	0.920	0.947	0.882	1.000
	Precision	0.972	0.974	0.983	0.958	0.982	1.000	0.933
	AUC	0.988	0.960	0.992	0.999	0.997	0.981	1.000

kNN Additional Test Results								
Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.768	0.693	0.826	0.611	0.828	0.800	0.800
	Recall	0.780	0.605	0.826	0.688	0.923	0.909	0.727
	Precision	0.757	0.813	0.826	0.550	0.750	0.714	0.889
	AUC	0.864	0.783	0.838	0.819	0.938	0.944	0.861
SE	F_1 score	0.848	0.756	0.830	0.679	0.860	0.968	0.963
	Recall	0.813	0.642	0.769	0.750	0.851	0.938	0.929
	Precision	0.885	0.919	0.903	0.621	0.870	1.000	1.000
	AUC	0.894	0.814	0.852	0.853	0.912	0.969	0.964
SEP	F_1 score	0.877	0.800	0.837	0.750	0.906	0.938	1.000
	Recall	0.835	0.714	0.731	0.840	0.842	0.882	1.000
	Precision	0.924	0.909	0.979	0.677	0.980	1.000	1.000
	AUC	0.912	0.849	0.859	0.902	0.919	0.941	1.000

MLP Additional Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.686	0.657	0.708	0.611	0.716	0.581	0.842
	Recall	0.725	0.512	0.685	0.688	0.923	0.818	0.727
	Precision	0.706	0.917	0.733	0.550	0.585	0.450	1.000
	AUC	0.917	0.881	0.854	0.940	0.969	0.934	0.926
SE	F_1 score	0.855	0.605	0.806	0.902	0.867	0.968	0.933
	Recall	0.823	0.491	0.719	0.958	0.830	0.938	1.000
	Precision	0.890	0.788	0.916	0.852	0.907	1.000	0.875
	AUC	0.935	0.820	0.850	0.996	0.967	0.977	0.998
SEP	F_1 score	0.886	0.698	0.818	0.870	0.944	0.938	1.000
	Recall	0.801	0.536	0.692	0.800	0.895	0.882	1.000
	Precision	0.992	1.000	1.000	0.952	1.000	1.000	1.000
	AUC	0.944	0.814	0.894	0.998	0.986	0.969	1.000

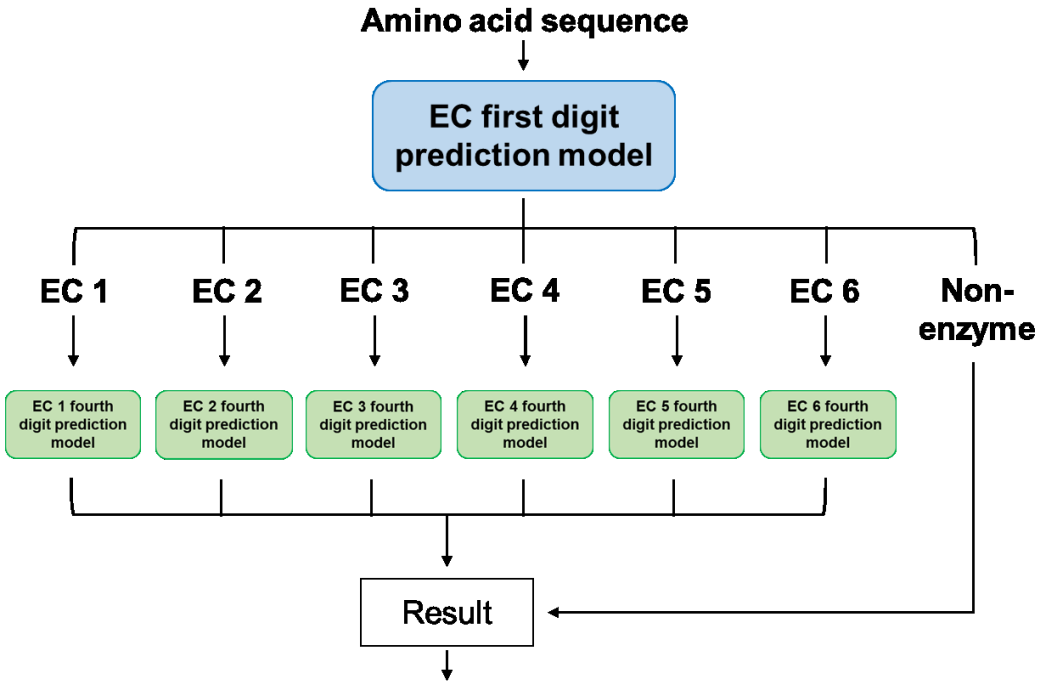
RF Additional Test Results

Model	Parameter	Macro/Average	EC1	EC2	EC3	EC4	EC5	EC6
E	F_1 score	0.627	0.602	0.755	0.429	0.711	0.706	0.308
	Recall	0.528	0.581	0.870	0.375	0.615	0.545	0.182
	Precision	0.772	0.625	0.667	0.500	0.842	1.000	1.000
	AUC	0.894	0.843	0.849	0.940	0.965	0.883	0.886
SE	F_1 score	0.847	0.667	0.852	0.889	0.815	0.933	0.897
	Recall	0.779	0.547	0.785	0.833	0.702	0.875	0.929
	Precision	0.929	0.853	0.931	0.952	0.971	1.000	0.867
	AUC	0.962	0.930	0.957	0.966	0.962	0.962	0.999
SEP	F_1 score	0.871	0.727	0.818	0.864	0.837	0.938	1.000
	Recall	0.771	0.571	0.692	0.760	0.719	0.882	1.000
	Precision	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AUC	0.985	0.986	0.983	0.996	0.983	0.963	1.000

CHAPTER IV

EnzymeNet: Residual Neural Networks model for Enzyme Commission numbers prediction

Graphical Abstract



Sequence	EC first digit	EC first digit score	EC fourth digit	EC fourth digit score
No. 1	EC 1	0.956	1.1.1.1	0.852
No. 2	Non-enzyme	0.912	-	-

IV.1. Introduction

Novel enzyme discovery is required to increase the production of target compounds[5,173]. The number of unannotated protein sequences is explosively increasing because of genome sequence technology[12]. Therefore, a valid computational method to predict enzyme functions from sequence information is needed to discover novel enzymes

within a huge number of unannotated sequences. Of these methods, one of the most basic approaches is machine learning which can learn various types of data. Machine learning algorithms have been applied to predict various protein annotations[32,33,66]. Then, several studies have been reported to predict EC numbers[27,46,51]. However, these studies have not discussed the evaluation of the sequences with numerous consecutive identical amino acids observed within unannotated sequences. The previously reported models for prediction of EC numbers sometimes misclassify such sequences as enzymes. Therefore, prediction models need to exclude the exceptional sequences from the candidate enzyme for comprehensive enzyme annotation prediction.

In the previous chapters, enzymatic reaction prediction models using multiple machine learning algorithms are built from enzyme and compound information and are evaluated for their ability to be utilized to predict comprehensive enzymatic reactions. However, the amino acid sequences used in the prediction may not be the enzyme when actually predicting unknown reactions. Therefore, in this chapter, EnzymeNet models based on Residual Neural Networks (ResNet)[174] are developed to predict EC numbers while removing proteins except for enzymes from the candidate sequences used in enzymatic reaction prediction. ResNet, which includes multiple CNN layers, has been demonstrated in protein structure and ligand-binding site predictions[175,176] and can address vanishing gradient problem occurring in deep learning models with deeper layers. Moreover, several CNN models built from the image-like features which were transformed to one-hot encoding from sequence information have been demonstrated in various enzyme annotation predictions as described in Chapter I. Enzyme sequence information is considered to consist of structural information because several reports

enable to predict protein structure from sequence information using deep learning methods. Therefore, EnzymeNet models are built using ResNet, which can be learned while capturing extensive enzyme features. The models predict EC numbers in 2 steps: 1) EC number first digit or negative and 2) complete EC number prediction. Moreover, the models exclude exceptional sequences with numerous consecutive identical amino acids in the first step. Therefore, the optimized condition of EnzymeNet models to remove such sequences are determined using several different datasets. The models were more accurate for extensive enzyme sequences with lower similarity to EnzymeNet training data than 4 previously reported models based on machine learning and sequence similarity methods[27,46,51,177]. EnzymeNet will help to determine enzyme annotations and to discover novel enzymes for useful substance production.

IV.2. Materials and Methods

IV.2.1. Data collection

IV.2.1-1. Data for Prediction of EC Number First Digits and Negative Sequences

To build positive data, enzyme sequences for each EC number class are collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) GENES[18] released on July 2019. There are 7 first digit EC number classes referred to as EC 1 to EC 7. EC 7 enzymes are not included in any of the data because too few enzymes are registered in KEGG. Enzyme sequences that are duplicated, with multiple EC numbers, or included non-canonical amino acids are removed and the length of amino acid residues is limited from 100 to 1000.

To keep data balanced, highly similar enzyme sequences are omitted by clustering at 90 % identity using CD-HI and then only a single enzyme sequence from each cluster is included. More than 80% of the EC numbers consist of fewer than 800 sequences. Therefore, similar sequences are removed by decreasing the identity until the number of sequences within each EC number is fewer than 800. As a result, 1,049,807 unique enzyme sequences are used to build and to evaluate EnzymeNet models.

To remove non-enzyme protein sequences and the exceptional sequences in the first prediction of EnzymeNet, negative data is built in 3 ways as follows: 1) Non-enzyme, 2) Random substitution and 3) Consecutive substitution. 3 random substitution and 3 consecutive substitution datasets are built to optimize the models for the prediction.

1) Non-enzyme

Proteins except for enzyme sequences are collected from Swiss-Prot[12] released on 2021. The sequences that are duplicated or included non-canonical amino acids were removed and the length of amino acid residues is limited from 100 to 1000. Only a single enzyme sequence from each cluster is used after clustering at 90 % identity to remove the sequence redundancy in the data. As a result, 142,378 non-enzyme sequences are used.

2) Random substitution

16,964 sequences are randomly extracted from the enzyme sequences included in the positive data. For each sequence, random 20 % of the amino acids of the sequence are substituted with the other amino acids (Figure 22A). The position and type of the substituted amino acids are randomly selected. This strategy is inspired by masked

language models, such as Bidirectional Encoder Representations from Transformers (BERT), which randomly masks some of the tokens from input, and the objective is to predict the original vocabulary id of the masked word based only on its context[178]. Therefore, in order to make EnzymeNet models understand original amino acid patterns of enzymes and the other sequence patterns, the artificial random substitution sequences are built. Moreover, 10 % and 40 % random substitution datasets are generated to evaluate the effect of the rate of substituted amino acids on this prediction.

3) Consecutive substitution

16,964 sequences are randomly extracted from positive data. For each sequence, 50 ~ 80% of the amino acids in the sequence are substituted with consecutive identical amino acids (Figure 22B). The position, type, and rate of the substituted amino acids are randomly selected. Previously reported models tend to predict such sequences as enzymes. Therefore, the current models enabled to remove the sequences, which are found within unannotated sequences. Moreover, 1 ~ 25 % and 26 ~ 49 % consecutive substitution datasets are generated to explore the relationship between prediction accuracy and the rate of substituted amino acids.

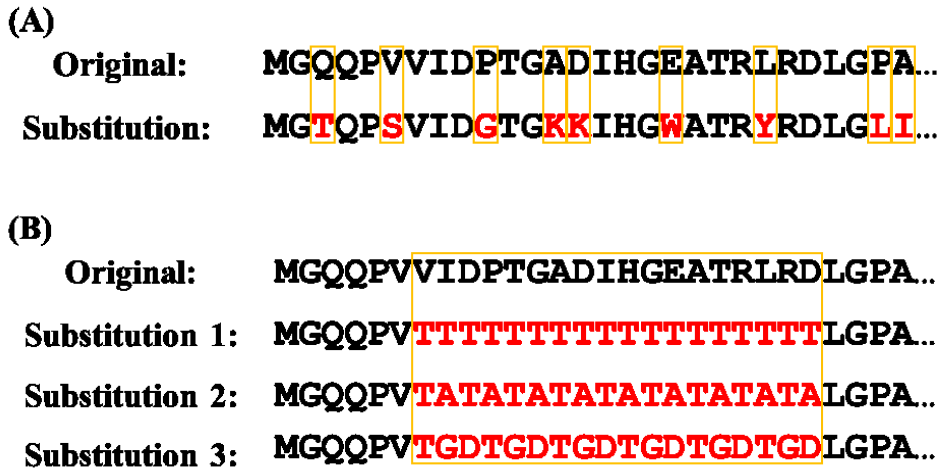


Figure 22. Examples of (A) Random substitution and (B) Consecutive substitution. The substituted positions of amino acids are surrounded by orange-bordered squares and substituted amino acids are marked in red color.

All positive and negative data are merged. All data are randomly split into training, validation, and test data, at an approximate ratio of 8 : 1 : 1 (Table 8A). Training, validation and test data are used for building models, evaluating all models in training and evaluating all models after training, respectively. The common test data (Table 8B) extracted from the 6 artificial negative test datasets is built to evaluate EnzymeNet models of 6 versions and to determine the optimal models in the first prediction. Moreover, EnzymeNet models are compared to 4 previously reported models using the common data.

IV.2.1-2. Data for Prediction of Complete EC Numbers

Positive data for EC first digit prediction is separated by each EC number fourth digit. Highly similar enzyme sequences are omitted by clustering at 90 % identity to decrease sequence redundancy. Moreover, the sequences with EC numbers that contained much

fewer sequences in each EC number fourth digit are removed. The data is randomly split into training, validation, and test data, at an approximate ratio of 8 : 1 : 1 (Table 8A). The test data is used in the evaluation of EnzymeNet models and previously reported models.

Table 8. Datasets Size of EnzymeNet. The Amounts of Training, Validation, and Test, at an Approximate Ratio of 8 : 1 : 1. (B) Common Test Data Size.

(A) Step	Type	EC	Training	Validation	Test
First prediction	Positive	EC 1	180,177	22,160	22,161
		EC 2	279,647	34,487	34,473
		EC 3	206,177	25,408	25,418
		EC 4	81,624	10,057	10,054
		EC 5	50,103	6,180	6,182
		EC 6	44,521	5,489	5,489
	Negative	Non-enzyme	113,881	14,225	14,272
		Random substitution	13,585	1,695	1,684
		Consecutive substitution	13,580	1,710	1,674
Second prediction	Positive	EC 1 fourth digit	122,858	14,961	14,961
		EC 2 fourth digit	284,539	35,139	35,139
		EC 3 fourth digit	222,466	27,511	27,511
		EC 4 fourth digit	658,52	8,065	8,065
		EC 5 fourth digit	55,187	6,825	6,825
		EC 6 fourth digit	107,511	13,380	13,380

(B)	Common test
EC1	22,161
EC2	34,473
EC3	25,418
EC4	10,054
EC5	6,182
EC6	5,489
Non-enzyme	14,272
Random substitution 10%	558
Random substitution 20%	558
Random substitution 40%	558
Consecutive substitution 1~25%	558
Consecutive substitution 26~49%	558
Consecutive substitution 50~80%	558

IV.2.2. Model Construction

EnzymeNet models which are built using ResNet50v2[174] consisted of 2 predictions; 1) prediction of EC number first digits and negative, 2) prediction of complete EC numbers. The model structure in the first prediction is shown in Figure 23. In Embedding Postprocessor layer[179], each amino acid included in each sequence is transformed into tokens which can be treated by deep learning and the tokens are transformed to (n, 1024, 128) feature maps. The positional information of each amino acid is added to the feature maps by Positional Embedding, and (n, 1024, 1024) feature maps are outputted. Next, in ConvertImg layer, feature maps are transformed to image-like (n, 256,256,3) feature maps, which are passed through ResNet50v2. Several studies have reported various biological prediction using CNN which has been often used in image recognition as described in Chapter I. ResNet can address vanishing gradient problem occurring in deep learning models with deeper layers. Therefore, ResNet which

is expanded CNN model are used to build EC number prediction models. From the final layer, the scores for 7 classes are then outputted. Moreover, 6 models referred to as EnzymeNet version 01 to 06 (v_01 to v_06) models are built from same positive and non-enzyme datasets, and from different artificial datasets obtained under different conditions of random and consecutive substitutions to explore the optimal condition in the first prediction (Table 9). EnzymeNet v_03 and v_05 models with higher accuracies using common test data are used for the following analyses.

The 2 EnzymeNet models in the second prediction are built by applying transfer learning for the EnzymeNet v_03 and v_05 models in the first step, respectively. For each model, 6 models for EC 1 to EC 6 are built. In the prediction, EC number first digits are predicted by first prediction model and then complete EC numbers are predicted by one of the 6 models selected from the results. When a result of the first prediction is negative, the second prediction is not performed. The all models in this study are built using Tensorflow[170]. A categorical cross-entropy loss function is used to train the models, and trainable parameters were updated for each batch.

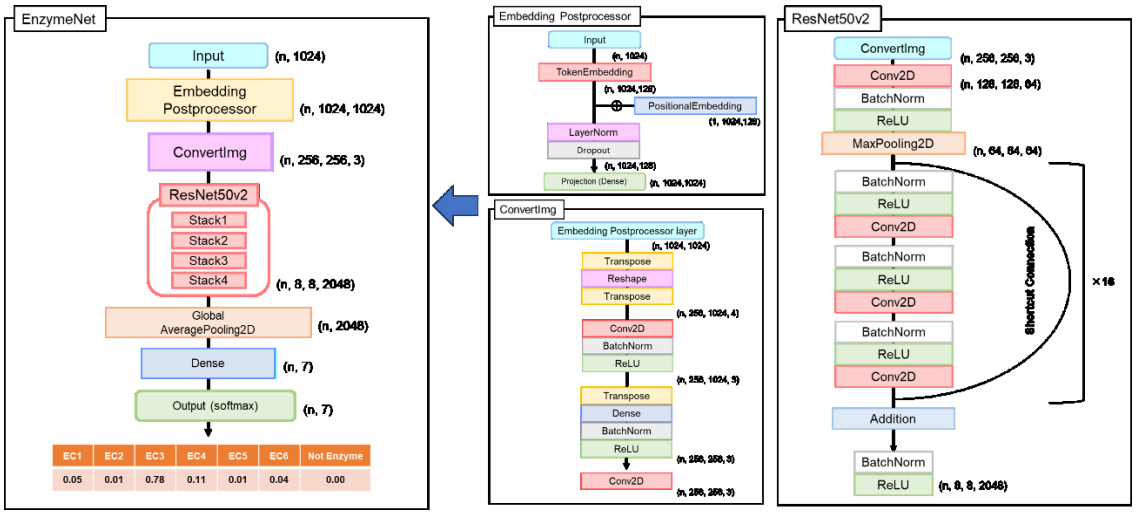


Figure 23. EnzymeNet structure in the first prediction using ResNet50v2.

Table 9. Type of Artificial Negative Datasets Used in 6 EnzymeNet Versions in the First Prediction.

Model	Random substitution	Consecutive substitution
EnzymeNet version_01 (v_01)	N/A	N/A
EnzymeNet version_02 (v_02)	10 %	50 ~ 80 %
EnzymeNet version_03 (v_03)	20 %	1 ~ 25 %
EnzymeNet version_04 (v_04)	20 %	26 ~ 49 %
EnzymeNet version_05 (v_05)	20 %	50 ~ 80 %
EnzymeNet version_06 (v_06)	40 %	50 ~ 80 %

Random and consecutive substitution datasets were not used as negative data in version_01.

IV.2.3. Model Evaluation

Accuracy, F_1 score, Precision, Recall, and Matthews correlation coefficient (MCC) are used for the evaluation of EnzymeNet models using test and common test data. Moreover, to evaluate the ability for both predictions, EC number first digits are predicted using test data for complete EC number prediction and then complete EC numbers are predicted. The incorrect test samples in the first prediction are not

predicted in the next step. The values are calculated using the scikit-learn library[119]. All EnzymeNet models are optimized using validation results. Then, the EnzymeNet models are compared with 4 EC number prediction models, 1) DeepEC[46], 2) DETECT v2[27], 3) ECPred[51], and 4) ProteInfer[177] using common test data for EC first digit prediction, and test data for complete EC prediction. DeepEC, ECPred and ProteInfer are built using machine learning methods while DETECT v2 is built using sequence similarity strategies[20]. The samples which DETECT v2 and ProteInfer do not predict are regarded as negative, because the models can predict only enzymes. The samples which DeepEC and ECPred do not predict are regarded as incorrect. Accuracy, Macro F_1 score, Macro Precision, Macro Recall are used in the model comparison. To compare the accuracy for prediction of all EC numbers in test data to all models, these values for all EC numbers are calculated using the number of each class of EC numbers. Moreover, the additional evaluations of these models are conducted using 2 test data, which enzyme sequences with high similarity to the training data are removed from, by lowering the sequence identity threshold using CD-HIT. One data is the enzyme and non-enzyme sequences extracted from common test data for prediction of EC number first digits and the other data is test data for prediction of complete EC numbers. To evaluate prediction model performance, the following values were calculated, given by:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN))}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives and false negatives, respectively. TP and TN are the number of samples that are correctly predicted, while FP and FN are the number of samples that are incorrectly predicted. The values below are also calculated as given by:

$$Macro\ \ Precision = \frac{1}{L}\sum_{i=1}^{L} Precision_i$$

$$Macro\ \ Recall = \frac{1}{L}\sum_{i=1}^{L} Recall_i$$

$$Macro\ \ F_1\ \ score = \frac{2 \cdot Macro\ \ Precision \cdot Macro\ \ Recall}{Macro\ \ Precision + Macro\ \ Recall}$$

IV.3. Results

IV.3.1. EC Number First Digit and Negative Predictions Using EnzymeNet Models

Figure 24 shows loss function curves for training and validation in the first prediction. The validation loss function decreases as epochs proceed. The results indicate that all EnzymeNet models for the prediction do not overfit. Test results are shown in Table 10. The model performances of all versions increase as epochs proceed. The models are built using 1,500, 1,300, 1,400, 1,500, 1,400 and 1,500 epochs, respectively, where the MCC is highest and the other values are relatively higher. Prediction accuracies show no significant differences among the models.

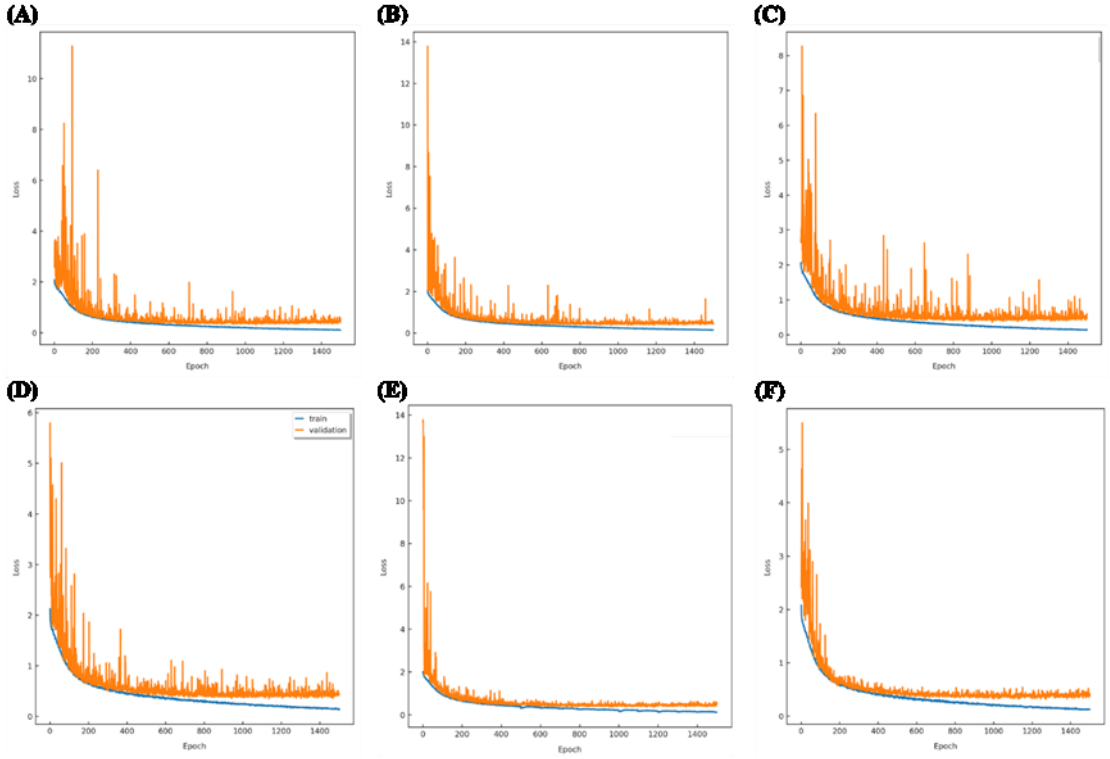


Figure 24. Training and validation loss curves (blue line: training, orange line: validation) of the EnzymeNet models in the first prediction using 6 different negative datasets for 1,500 epochs. (A) EnzymeNet v_01 model, (B) EnzymeNet v_02 model, (C) EnzymeNet v_03 model, (D EnzymeNet v_04 model, (E) EnzymeNet v_05 model, (F) EnzymeNet v_06 model as shown in Table 9.

Table 10. Test Results of EC Number First Digit Prediction Using 6 EnzymeNet Models for 1,500 Epochs.

EnzymeNet v_01 model					EnzymeNet v_02 model				
Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
100	0.601	0.571	0.634	0.494	100	0.596	0.639	0.558	0.438

200	0.776	0.779	0.773	0.691	200	0.733	0.735	0.732	0.641
300	0.735	0.714	0.758	0.656	300	0.761	0.731	0.794	0.743
400	0.802	0.783	0.822	0.751	400	0.843	0.847	0.839	0.768
500	0.817	0.792	0.843	0.765	500	0.841	0.840	0.841	0.771
600	0.860	0.856	0.865	0.815	600	0.841	0.837	0.845	0.800
700	0.860	0.841	0.881	0.830	700	0.853	0.855	0.852	0.788
800	0.874	0.871	0.876	0.838	800	0.841	0.834	0.849	0.827
900	0.882	0.884	0.879	0.846	900	0.857	0.853	0.861	0.822
1,000	0.886	0.883	0.888	0.850	1,000	0.868	0.876	0.860	0.807
1,100	0.887	0.892	0.882	0.849	1,100	0.875	0.880	0.870	0.805
1,200	0.886	0.879	0.894	0.853	1,200	0.869	0.867	0.870	0.829
1,300	0.857	0.832	0.884	0.834	1,300	0.878	0.872	0.884	0.834
1,400	0.889	0.881	0.897	0.862	1,400	0.860	0.854	0.866	0.833
1,500	0.898	0.897	0.899	0.868	1,500	0.895	0.903	0.888	0.842

EnzymeNet v_03 model					EnzymeNet v_04 model				
Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
100	0.597	0.626	0.571	0.448	100	0.612	0.638	0.587	0.470
200	0.736	0.733	0.739	0.653	200	0.729	0.700	0.761	0.660
300	0.800	0.799	0.801	0.746	300	0.765	0.738	0.794	0.706
400	0.788	0.764	0.814	0.732	400	0.769	0.737	0.804	0.710
500	0.825	0.825	0.825	0.777	500	0.790	0.759	0.823	0.737

600	0.828	0.803	0.854	0.785	600	0.820	0.798	0.843	0.777
700	0.860	0.867	0.852	0.808	700	0.842	0.825	0.860	0.796
800	0.864	0.864	0.864	0.818	800	0.807	0.785	0.831	0.757
900	0.863	0.867	0.860	0.813	900	0.841	0.818	0.865	0.802
1000	0.872	0.877	0.868	0.829	1000	0.860	0.848	0.872	0.825
1100	0.814	0.798	0.831	0.763	1100	0.827	0.797	0.859	0.797
1200	0.881	0.887	0.874	0.845	1200	0.869	0.861	0.877	0.834
1300	0.859	0.841	0.877	0.831	1300	0.830	0.815	0.845	0.781
1400	0.881	0.888	0.874	0.848	1400	0.863	0.849	0.877	0.828
1500	0.852	0.846	0.858	0.812	1500	0.870	0.857	0.883	0.843

EnzymeNet v_05 model					EnzymeNet v_06 model				
Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
100	0.645	0.652	0.638	0.511	100	0.645	0.652	0.638	0.511
200	0.754	0.75	0.759	0.662	200	0.754	0.75	0.759	0.662
300	0.771	0.75	0.793	0.699	300	0.771	0.75	0.793	0.699
400	0.810	0.793	0.828	0.761	400	0.810	0.793	0.828	0.761
500	0.833	0.823	0.844	0.793	500	0.833	0.823	0.844	0.793
600	0.792	0.773	0.813	0.734	600	0.792	0.773	0.813	0.734
700	0.857	0.87	0.844	0.795	700	0.857	0.87	0.844	0.795
800	0.872	0.881	0.864	0.835	800	0.872	0.881	0.864	0.835
900	0.869	0.872	0.866	0.818	900	0.869	0.872	0.866	0.818

1000	0.875	0.888	0.862	0.827	1000	0.875	0.888	0.862	0.827
1100	0.867	0.878	0.856	0.813	1100	0.867	0.878	0.856	0.813
1200	0.878	0.871	0.886	0.847	1200	0.878	0.871	0.886	0.847
1300	0.850	0.82	0.883	0.83	1300	0.850	0.82	0.883	0.83
1400	0.885	0.888	0.883	0.854	1400	0.885	0.888	0.883	0.854
1500	0.885	0.889	0.881	0.849	1500	0.885	0.889	0.881	0.849

Next, the models are evaluated using common test data (Tables 11 and 12). The results of the overall first step prediction maintain constant high accuracy among EnzymeNet models, while the prediction results of the negative samples vary. EnzymeNet v_03 model is more accurate for negative sequences than the other models. EnzymeNet v_01 model which learns only non-enzyme dataset as negative data predicts artificial negative samples with much lower accuracy. The EnzymeNet v_03 and v_05 models are regarded as optimized models in the first step prediction because the models more correctly predict both all test sequences and artificial sequences. Table 13 shows the common test results using the 2 models for each class. EnzymeNet v_06 model is not selected as optimized models because the model learns more different artificial sequences from original enzyme sequences and more easily classify the sequences than the other models. All models predict consecutive substitution samples with higher accuracy than random substitution samples.

Table 11. Common Test Results of the First Step Using 6 EnzymeNet Models.

Model	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
EnzymeNet v_01	1500	0.885	0.885	0.884	0.849
EnzymeNet v_02	1300	0.869	0.855	0.883	0.832
EnzymeNet v_03	1400	0.885	0.891	0.878	0.852
EnzymeNet v_04	1500	0.868	0.855	0.882	0.841
EnzymeNet v_05	1400	0.883	0.886	0.881	0.851
EnzymeNet v_06	1500	0.889	0.894	0.884	0.854

Model	Accuracy of random substitution samples	Accuracy of consecutive substitution samples	Accuracy of all artificial negative samples
EnzymeNet v_01	0.090	0.561	0.325
EnzymeNet v_02	0.443	0.829	0.636
EnzymeNet v_03	0.445	0.910	0.678
EnzymeNet v_04	0.250	0.746	0.498
EnzymeNet v_05	0.383	0.777	0.580
EnzymeNet v_06	0.398	0.784	0.591

Table 12. Common Test Results of Prediction of Artificial Negative Data.

Model	Random 10%	Random 20%	Random 40%	Consecutive 1 ~ 25%	Consecutive 26 ~ 49%	Consecutive 50 ~ 80%
EnzymeNet v_01	0.072	0.095	0.102	0.215	0.606	0.862
EnzymeNet v_02	0.188	0.376	0.765	0.527	0.964	0.995
EnzymeNet v_03	0.197	0.380	0.758	0.760	0.973	0.998
EnzymeNet v_04	0.109	0.229	0.412	0.351	0.891	0.995
EnzymeNet v_05	0.136	0.324	0.688	0.428	0.910	0.993
EnzymeNet v_06	0.111	0.297	0.785	0.430	0.925	0.996

Random: Random substitution, Consecutive: Consecutive substitution.

Table 13. Common Test Results of First Prediction for Each Class Using Optimized 2 EnzymeNet Models.

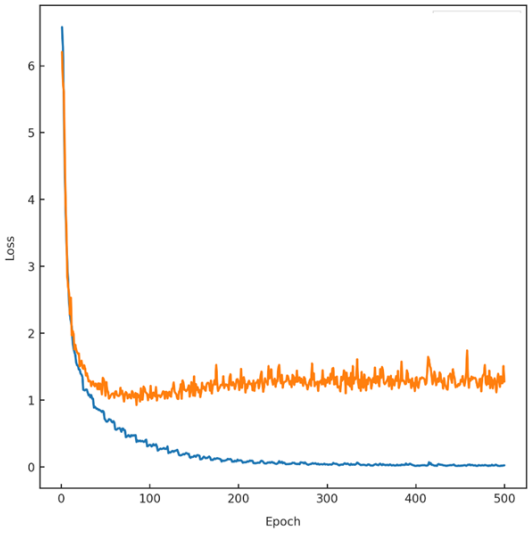
EnzymeNet v_03 model				EnzymeNet v_05 model			
Class	F_1 score	Precision	Recall	Class	F_1 score	Precision	Recall
EC1	0.909	0.892	0.928	EC1	0.917	0.951	0.885
EC2	0.898	0.909	0.887	EC2	0.890	0.861	0.921

EC3	0.881	0.886	0.876	EC3	0.881	0.870	0.893
EC4	0.904	0.904	0.903	EC4	0.900	0.950	0.855
EC5	0.924	0.929	0.919	EC5	0.913	0.900	0.926
EC6	0.890	0.963	0.828	EC6	0.894	0.860	0.929
Negative	0.780	0.758	0.804	Negative	0.780	0.807	0.754

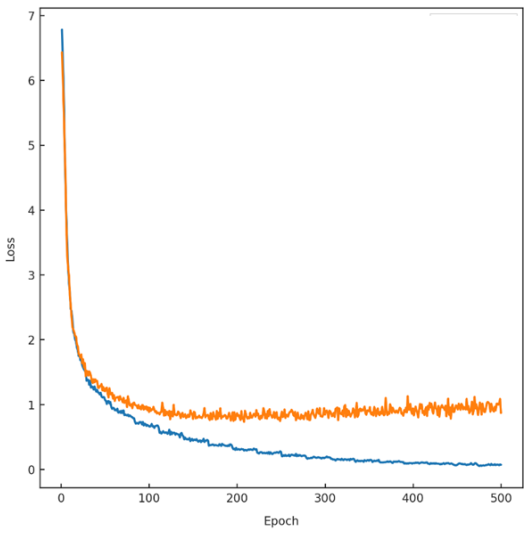
IV.3.2. Complete EC prediction using EnzymeNet models

Figure 25 shows loss function curves for training and validation in the second prediction using EnzymeNet v_05 models. The results of EnzymeNet v_03 models are similar to that of EnzymeNet v_05 models. Unlike the first prediction, the validation loss functions in 6 models for EC 1 to EC 6 insufficiently decrease in comparison to the training loss. However, all models are not regarded as overfitting, because all validation loss functions do not significantly increase. The EnzymeNet v_05 models for EC 1 to EC 6 are built using 400, 500, 400, 400, 90, and 300 epochs, respectively (Table 14). On the other hand, the EnzymeNet v_03 models are built using 500, 500, 400, 350, 90 and 450 epochs, respectively. Both models also predict test data with high accuracy in the second prediction although the accuracies are lower than that of EC first digit prediction.

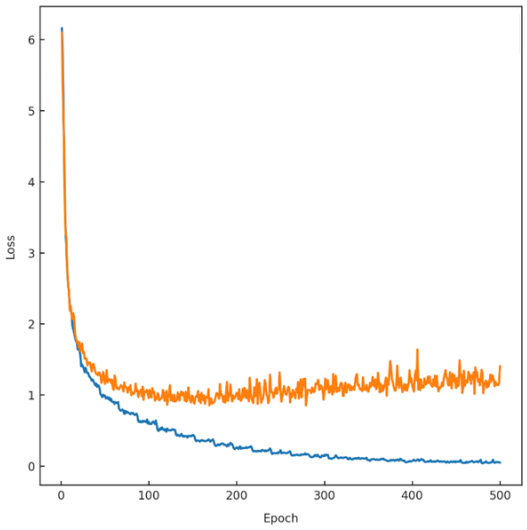
(A) EC 1.X.X.X



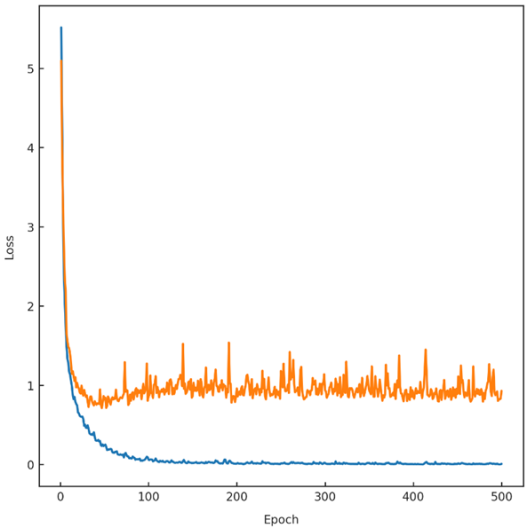
(B) EC 2.X.X.X



(C) EC 3.X.X.X



(D) EC 4.X.X.X



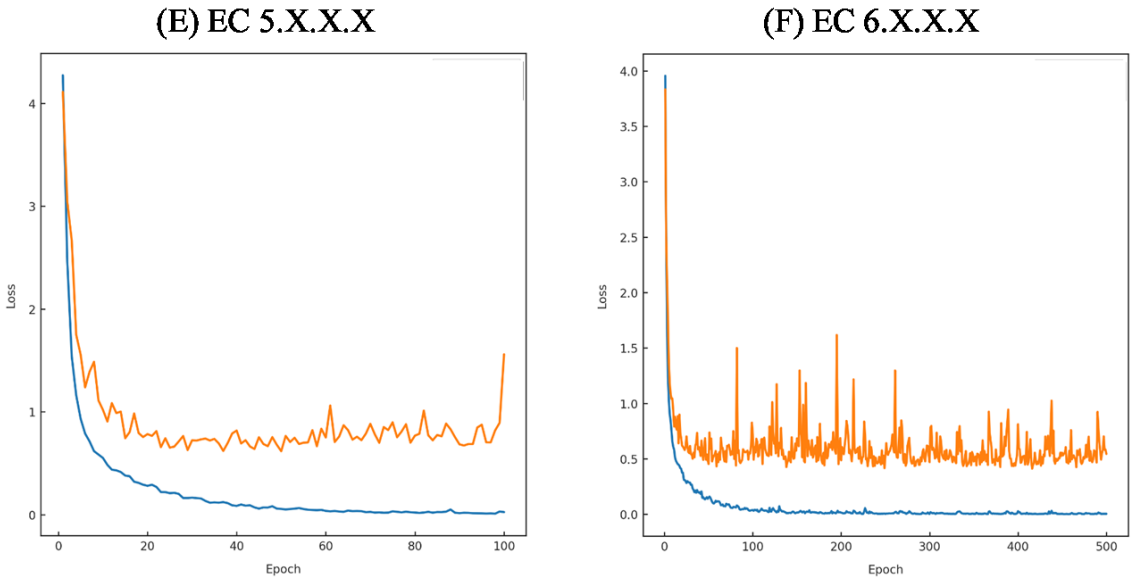


Figure 25. Training and validation loss curves (blue line: training, orange line: validation) of each model in the second prediction using EnzymeNet v_05 model. (A) EC 1, (B) EC 2, (C) EC 3, (D) EC 4, (E) EC 5 and (F) EC 6.

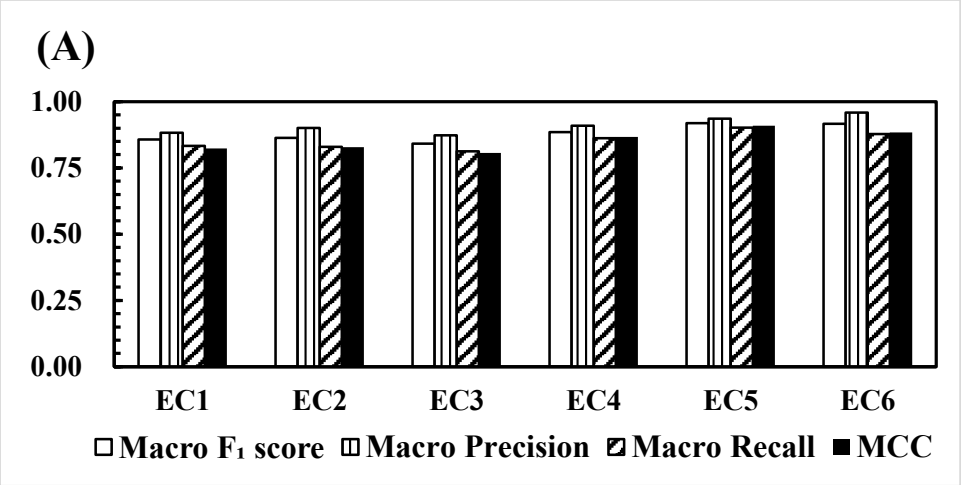
Table 14. Test Results of Complete EC Number Prediction. (A) EnzymeNet v_03 models and (B) EnzymeNet v_05 models.

(A) EnzymeNet v_03 models					
Class	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
EC1	500	0.860	0.872	0.849	0.837
EC2	500	0.873	0.884	0.863	0.861
EC3	400	0.852	0.857	0.847	0.838
EC4	350	0.891	0.900	0.881	0.885
EC5	90	0.927	0.934	0.921	0.925
EC6	450	0.954	0.961	0.947	0.955

(B) EnzymeNet v_05 models

Class	Epoch	Macro F_1 score	Macro Precision	Macro Recall	MCC
EC1	400	0.842	0.858	0.827	0.820
EC2	500	0.865	0.872	0.858	0.852
EC3	400	0.838	0.852	0.825	0.818
EC4	400	0.880	0.889	0.871	0.869
EC5	90	0.897	0.909	0.886	0.887
EC6	300	0.928	0.934	0.923	0.933

Continuous prediction results of EC first digits and complete ECs for the models are shown in Figure 26. The data of complete EC number prediction is used in this evaluation. The incorrect test samples in the first predictions are not performed in the next predictions. As a result, the prediction accuracies are slightly lower than in only complete EC prediction, but remain high. As with EC first digit prediction, EnzymeNet v_03 models are more accurate than EnzymeNet v_05 models.



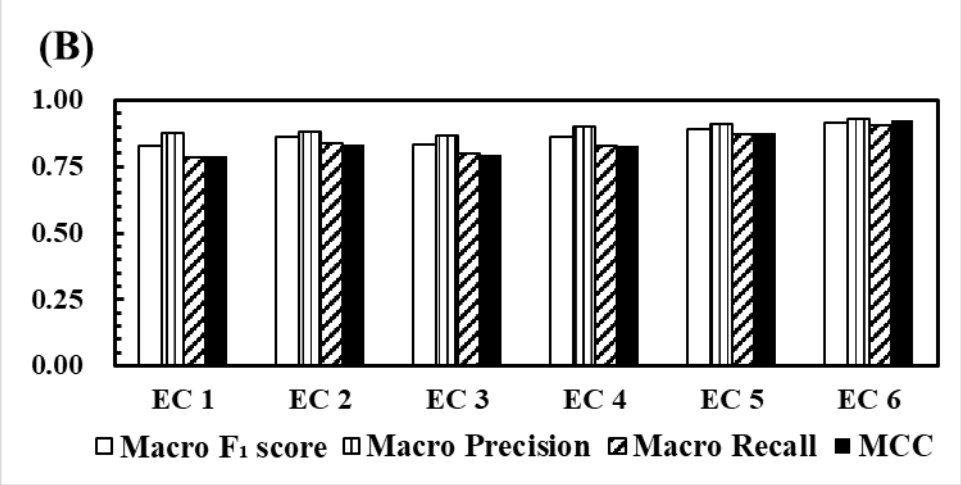


Figure 26. Continuous test results of (A) EnzymeNet v_03 models, (B) EnzymeNet v_05 models.

IV.3.3. Comparative evaluation of EC number prediction

As a benchmark, the EnzymeNet models are compared with DeepEC, DETECT v2, ECPred, and ProteInfer using common test data, and test data for prediction of complete ECs. Figure 27 and Table 15 show the comparative results of common test data. Both EnzymeNet models exhibit higher test prediction accuracy and higher both Macro Precision and Macro Recall. The accuracies of DeepEC and DETECT v2 are lower than

those of other models and the Macro Recalls are lower than the Macro Precisions. Moreover, the ability to classify non-enzyme and random substitution sequences using EnzymeNet models is lower than that of DETECT v2 and ProteInfer (Figure 27B and Table 15). Random substitution sequences tend to be more incorrectly predicted than consecutive substitution sequences. Both EnzymeNet models predict correctly more consecutive substitution sequences than the other models.

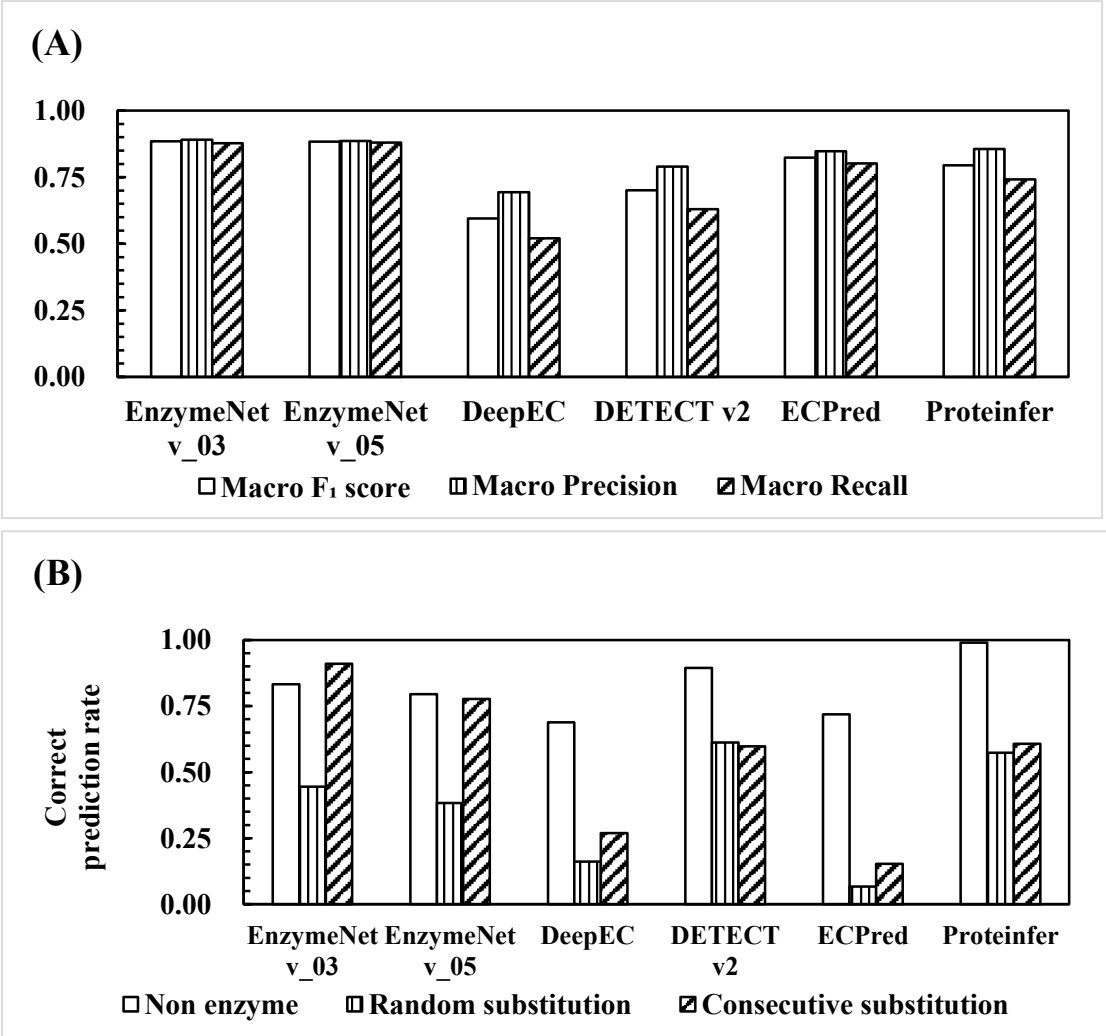


Figure 27. Comparative common test results of EC first digit prediction. (A) Model accuracy results of prediction of all EC first digits or negative, (B) Correct prediction

rate in only negative prediction. The number of non-enzyme test sequences was 14,272, the number of random substitution test sequences and consecutive substitution test sequences were 1,674, respectively.

Table 15. Comparative Common Test Results of Prediction of Artificial Negative Data.

Model	Random 10%	Random 20%	Random 40%	Consecutive 1~25%	Consecutive 26~49%	Consecutive 50~80%
EnzymeNet v_03	0.197	0.380	0.758	0.760	0.973	0.998
EnzymeNet v_05	0.136	0.324	0.688	0.428	0.910	0.993
DeepEC	0.165	0.152	0.167	0.199	0.294	0.638
DETECT v2	0.532	0.572	0.731	0.514	0.554	0.683
ECPred	0.050	0.054	0.097	0.100	0.194	0.425
ProteInfer	0.348	0.545	0.826	0.444	0.728	0.751

Next, the prediction results of complete EC prediction are shown in Figure 28. Both EnzymeNet models show higher prediction accuracy with Macro F_1 scores up to 0.850 than the other models. The conditions of negative artificial datasets in EnzymeNet v03 models are more suitable for EC prediction because of higher accuracies of all evaluations. On the other hand, the accuracies of the other models in the second prediction decrease much more than those of the first prediction. The test enzyme sequences include 2,591 ECs and some ECs are easy to predict using the previously

reported models. However, the F_1 scores of 1,877 of 2,591 ECs for EnzymeNet v_03 models are higher.

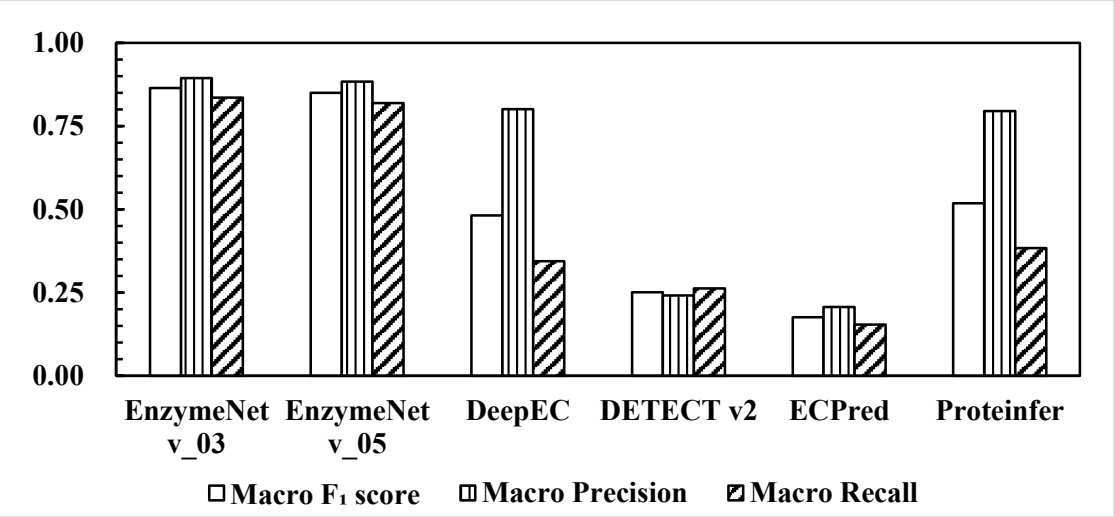


Figure 28. Comparative test results of complete EC prediction.

Figure 29 shows the results of the 2 datasets, which similar sequences to the training datasets are removed from, in EC first digit and complete EC predictions. The lower sequence identity threshold is, the more difficult the predictions are not depending on prediction models. In the first prediction, both EnzymeNet models predict more correctly in 70 and 80 sequence identity thresholds. However, ECPred is the most accuracy between all models. On the other hand, both EnzymeNet models are more accuracy in the complete EC prediction not depending on the value of sequence identity thresholds. All models in these evaluations show the decreases in prediction accuracy as more similarity sequences are removed.

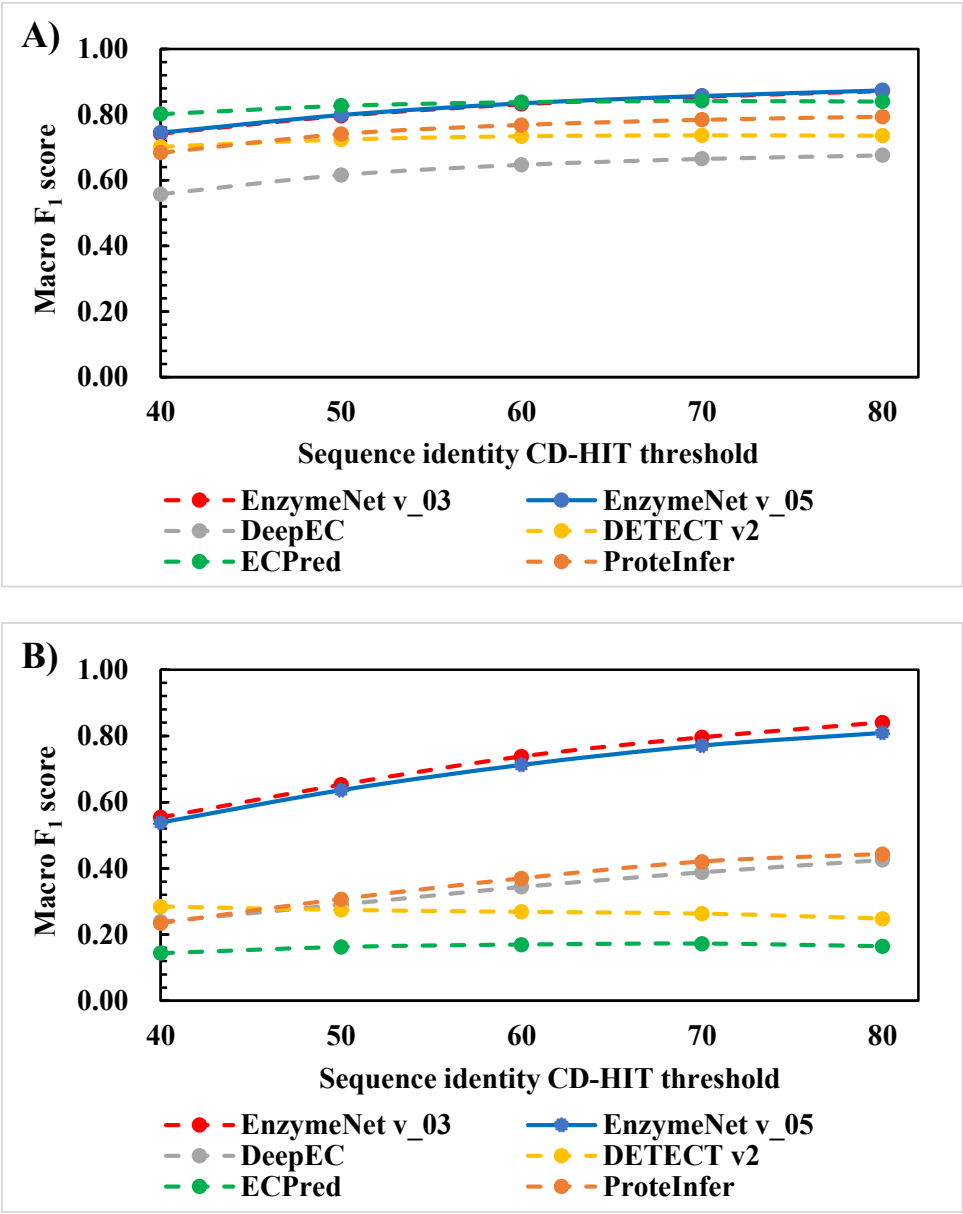


Figure 29. Macro F_1 scores of (A) EC first digit prediction and (B) complete EC number prediction using the common test and test sequences removing similar sequences to training enzyme sequences for each sequence identity threshold using CD-HIT.

IV.4. Discussion

EnzymeNet models are developed to predict complete EC number for each amino acid sequence in 2 step predictions while removing non-enzyme proteins and exceptional sequences. To discover novel enzymes within a vast number of unannotated protein sequences, enzyme prediction models for enzyme functions need to efficiently learn the patterns of amino acids for each enzyme sequence. Therefore, EnzymeNet models are built to enable to remove the sequences with numerous consecutive identical amino acids, which are found within unannotated sequences, as well as non-enzyme proteins. The conventional EC number prediction models have not considered such sequences. Moreover, EnzymeNet models deeply learn various patterns of amino acid sequences by adding the random substitution sequences, which are similar to the original enzymes, to the datasets.

First, the methods of generating artificial negative sequences in the first prediction are optimized. All EnzymeNet models in the evaluations of test and common test data maintain high prediction accuracy. However, the prediction results of artificial negative sequences using common test data are significantly different depending on the models. EnzymeNet v_01 model which does not learn the artificial sequences do not predict almost the sequences. This is because machine learning models generally have difficulty predicting the data which is so different from training data.

Considering the results of the positive and negative data, 2 complete EC number prediction models are built based on EnzymeNet v_03 and v_05 models, which exhibit higher prediction accuracy of the overall sequences. Moreover, the artificial negative

condition of EnzymeNet v_03 model is suitable for the prediction because the model predicts the consecutive substitution sequences constructed by all conditions with higher accuracy. The results of EnzymeNet v_03 models in the complete EC digit prediction and the continuous predictions are almost as accurate as those of EnzymeNet v_05 models. This indicates that the conditions of generating artificial negative samples do not have significant influence on the overall prediction accuracy in both predictions.

Next, the EnzymeNet models are compared with 4 previously reported EC number prediction models. In the prediction of common test data, EnzymeNet models exhibit higher prediction accuracy. Furthermore, the previously reported models cannot predict the sequences with consecutive identical amino acids which are apparently non-enzyme. The results and common test results of EnzymeNet v_01 model indicate that prediction models cannot predict the exceptional sequences without learning them. However, the EnzymeNet models cannot classify non-enzyme and random substitution sequences with the highest accuracy. Even though the number of non-enzyme sequences is clearly larger than that of enzyme sequences, EnzymeNet models learn more enzymes. This is why the accuracies of the models for non-enzyme sequences were lower. Moreover, EnzymeNet models have difficulty classifying the random substitution sequences because the models learn both random substitution sequences and pre-substituted enzyme sequences (original enzyme sequences before substituting) which are similar to each other. Since the other models except for DeepEC are built from fewer enzyme sequences than EnzymeNet models, it is assumed that the other models do not learn the pre-substituted enzymes and are able to predict them without confusion.

For complete EC number prediction, EnzymeNet models show much higher prediction accuracy than the other models in comparison to the results of the first prediction. On the other hand, DeepEC and ProteInfer correctly classify positive enzymes because the Macro Precisions are much higher than Macro Recalls. Therefore, the putative enzymes which are predicted as positive by these models can be assigned new annotations. The other prediction models[45,49] are built from various enzyme features while EnzymeNet models, DeepEC and ProteInfer need so simple features, namely, one-hot encoding, token and positional embedding. These simple feature extractions do not have a large effect on prediction results, which depend on only amino acid pattern information. Moreover, the amount of training data in DeepEC is almost as large as that of EnzymeNet models. This suggests that prediction accuracy does not necessarily rely on the amount of training data for each model. Building optimized model structure to match prediction target is required.

Finally, all models are evaluated using the datasets which similar sequences to the training datasets are removed from. EnzymeNet models also exhibit higher prediction accuracy for difficult enzyme sequences in EC number complete prediction even though the models do not show the highest accuracy in the first prediction. The results suggest EnzymeNet models can correctly predict EC numbers for more extensive sequences in comparison to reported models. However, EnzymeNet models cannot correctly predict some enzymes with lower similarity to training data. To improve the abilities of EnzymeNet models for the difficult positive, non-enzyme and random substitution predictions further, updated methods to build training data and model structure are needed. The common decreases (Figure 29) in prediction accuracy of all machine

learning models except for DETECT v2 as lowering the threshold indicate that the evaluation of the difficult enzymes in the predictions may be insufficient.

In summary, EnzymeNet models can exclude the exceptional sequences from the candidate sequence in addition to the EC number prediction, which are more accurate for extensive enzyme sequences than the reported models. Moreover, up to 4,000 sequences are predicted using EnzymeNet in about 10 minutes at one time. Therefore, EnzymeNet models enable to apply to find available enzymes from metagenomics registered in sequences databases[11,180]. For the putative enzyme sequences predicted using EnzymeNet models, the SEP models developed in the previous chapters can predict corresponding substrates and products, namely, detailed enzymatic reaction annotations. The robustness of EnzymeNet models will lead to predict enzyme annotations related to enzymatic reactions for mass unannotated protein sequences and to discover novel enzymes for biosynthesis of functional compounds using microorganisms.

IV.5. Usage of EnzymeNet

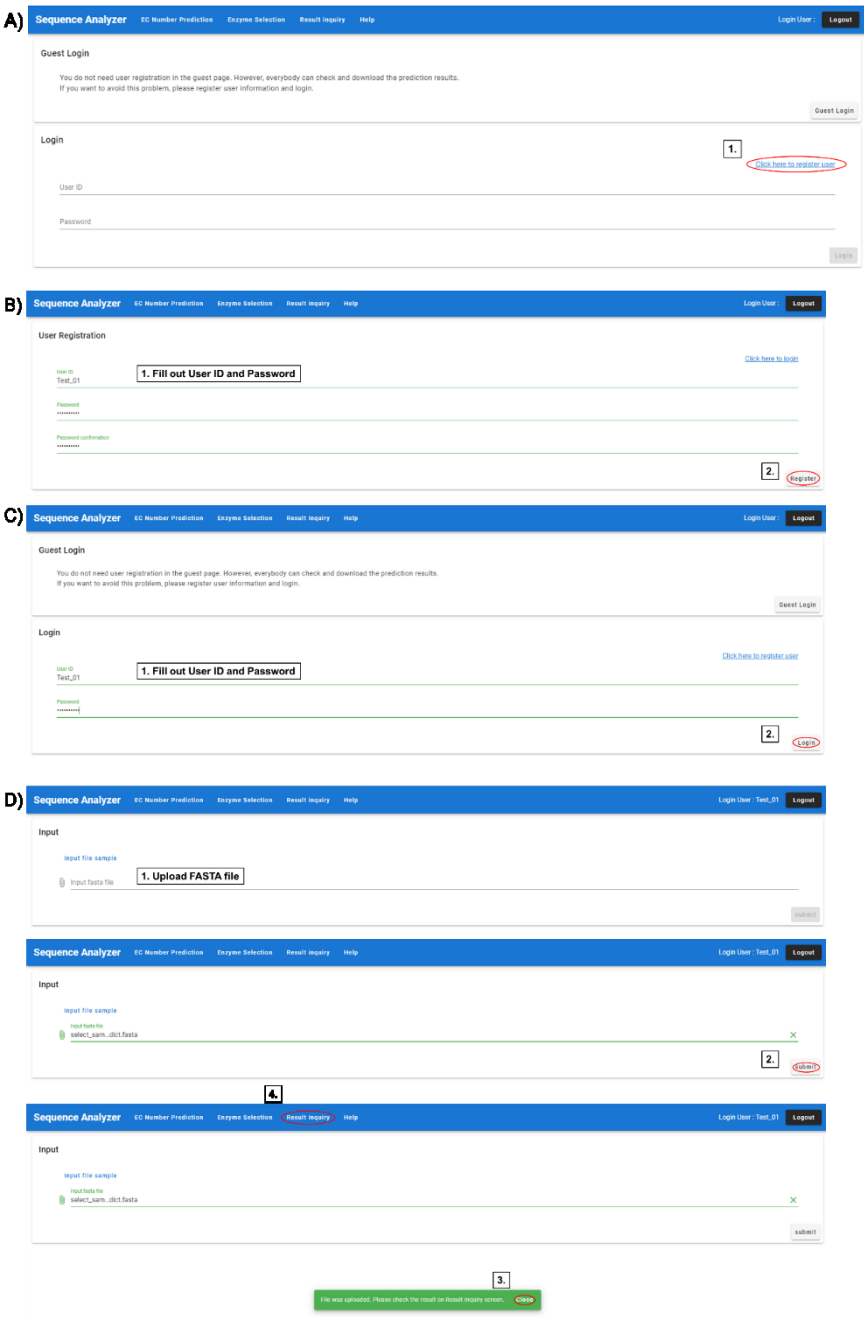
The usage of EnzymeNet webserver (https://m-ai.org/enzymenet/) is shown in Figure 30. The procedure for using this server consists of several steps: (1) register user information, (2) log in, (3) upload and submit the file including some number of amino acid sequences described in FASTA format, and (4) download the results named as input file name. EC number prediction page is automatically jumped after registering user information for the first time. The FASTA headers must be described as follows.

>*SequenceId ProteinInformation*

where *SequenceId* and *ProteinInformation* represent sequence id or protein identifier such as hsa:351 and explanation of protein, respectively. A space is needed between *SequenceId* and *ProteinInformation*. When no strings in *ProteinInformation* are described, the result is not outputted. EnzymeNet cannot predict the protein sequences whose length of amino acid residues is 1000 and more than. The results are removed from EnzymeNet server when the users finish downloading them or close download page.

As a result, the prediction labels and scores for first digit and complete EC number are outputted. When EnzymeNet predicts a sample as negative in the first prediction, the label and score for complete EC number are not outputted. Up to 4,000 sequences are predicted in about 10 minutes at one time.

Guest prediction page without the registration of the user information is prepared as shown in Figure 30A. However, in the guest mode, everybody can watch and download the results. If you avoid this, you should register user information.


A)
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User :
Logout
Guest Login
You do not need user registration in the guest page. However, everybody can check and download the prediction results.
If you want to avoid this problem, please register user information and login.
Guest Login
Login
1.
Click here to register user
User ID
Password
Login
B)
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User :
Logout
User Registration
Click here to login
User ID
Test_01
1. Fill out User ID and Password
Password
Password confirmation
2.
Register
C)
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User :
Logout
Guest Login
You do not need user registration in the guest page. However, everybody can check and download the prediction results.
If you want to avoid this problem, please register user information and login.
Guest Login
Login
Click here to register user
User ID
Test_01
1. Fill out User ID and Password
Password
2.
Login
D)
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User : Test_01
Logout
Input
Input file sample
Input fasta file
1. Upload FASTA file
submit
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User : Test_01
Logout
Input
Input file sample
Input fasta file
select_sam...dict.fasta
2.
submit
4.
Sequence Analyzer
EC Number Prediction
Enzyme Selection
Result inquiry
Help
Login User : Test_01
Logout
Input
Input file sample
Input fasta file
select_sam...dict.fasta
submit
3.
File was uploaded. Please check the result on Result inquiry screen.
Close

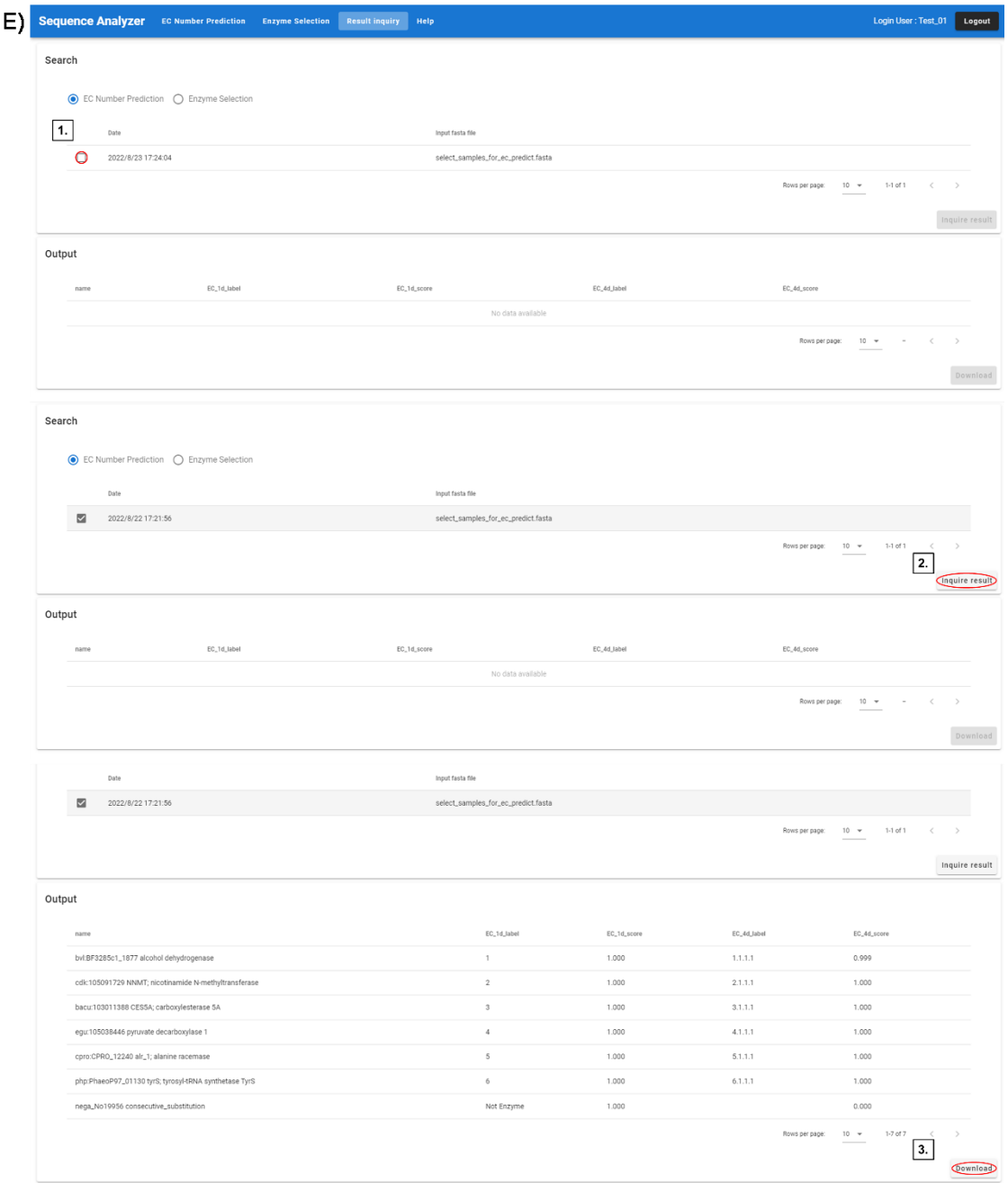


Figure 30. The usage of EnzymeNet. The numbers described in the figure represent the order of operation. A) Start page, B) user registration page, C) Log in page, D) upload and submission page and E) result page. User registration is required before utilizing EC number prediction without using guest mode. User ID and Password are filled out. Then, FASTA file describing one or arbitrary amino acid sequences is uploaded and

submitted after login. “Result Inquiry” is clicked on, the input file is selected, and “Inquire result” is clicked on. The results are shown if prediction of all test sequences is finished. All prediction results can be downloaded.

CHAPTER V

General Conclusion and Future work

Enzymes play an important role in the production of substances using microorganisms. The variety of useful substances that can be biosynthesized by enzymes has increased with the rapid development of various technologies such as genetic engineering, synthetic biology, and metabolic engineering. In order to access more useful compounds, novel enzyme discovery encourages the expansion of current metabolic pathways. Moreover, the number of proteins registered in protein sequence databases is increasing due to sequencing advances. Most of the increasing number of sequences are unannotated sequences and new enzyme sequences need to be discovered within them. The optimal method to achieve this is a computational approach.

In this study, enzyme reaction prediction models are developed to discover novel enzymes and enzymatic reactions using several machine learnings which have the potential to acquire new knowledge from a large number of datasets. First, E (Enzyme) models are built from enzyme sequence information using the same strategy as conventional enzyme function predictions. Next, SE (Substrate-Enzyme) and SEP (Substrate-Enzyme-Product) models combined enzyme sequence information with compound chemical structure information predict enzyme-compound combinations in enzymatic reactions. While accuracy of E models is not optimal, SE models and SEP models predict EC numbers and reactions with high accuracy using all tested machine learning-based methods. In comparison to BLAST, correct prediction is higher for most

of SE and SEP models. Here, SEP-RF model achieves the best performance using *E. coli K-12* test.

In order to improve prediction models, new E, SP, SE and SEP models are developed using several machine learning algorithms, including Deep Neural Network by updating training datasets and feature extractions. Moreover, these SE and SEP models can predict whether or not enzyme reactions will occur. Improvements in prediction performances over that of the previous SEP-RF result in the same test indicate that the updated methods are more effective for prediction of enzymatic reactions. The SEP-DNN model exhibits the highest prediction accuracy with Macro F_1 scores up to 0.966 using test enzymatic reactions involving a number of enzyme sequences derived from various species and with robust prediction of unknown enzymatic reactions that are not included in the training data. This model can predict more extensive enzymatic reactions in comparison to previously reported model. The SEP-type models can select enzyme sequences which may biosynthesize functional compounds. Moreover, the models can also apply to predict new substrates which a known enzyme may act with and new products which a known enzyme may synthesize because the models can predict enzymatic reactions for the substrate-enzyme-products combinations.

On the other hand, the enzymatic reaction prediction models need to be further improved in several points. First, it is also necessary to improve feature vectors because all SE and SEP models show lower prediction accuracy for the test reactions with compounds that do not exist in the training data and for reactions which are low similarity with training data. Second, the current models tend to misjudge some

reactions as negative because most of the negative samples are similar to positive samples. Thus, improved methods to build negative training data are needed. Moreover, the reactions involving compounds that are not uniquely determined, such as those indicated as “R” in the chemical structure, and the reactions involving polymer synthesis are not included in the datasets in this study. To predict more extensive enzymatic reactions, the models need to learn the enzymatic reactions built by fitting various compounds to these reactions.

Excluding non-enzyme proteins is necessary before predicting with the appropriate combinations of enzymes and compounds using enzymatic reaction prediction models. Therefore, EC number prediction models named EnzymeNet are developed to predict enzyme annotations for enzymatic reaction in addition to exclude non-enzyme and exceptional sequences. In order to reduce the number of steps in the prediction, the models roughly predict the first digits of the EC number and then determine the full EC numbers. As a result, EnzymeNet models predict EC numbers for extensive enzyme sequences and even the sequences, which are low similarity with training data, with higher accuracy than previously reported models. These results indicate that the enzymatic reaction prediction models in the previous chapter can potentially improve the model accuracy using ResNet.

Combining the EC number prediction models with enzymatic reaction prediction models enables to predict comprehensive enzyme annotations related to enzymatic reactions. First, the EC number prediction models select only enzyme from the amino acid sequences and roughly estimate a reaction catalyzed by the putative enzyme. Next,

based on the EC prediction results, the enzymatic reaction prediction models predict the substrate that is likely to react with the enzyme and the product that is likely to be synthesized. This system, which combines 2 prediction steps, is evaluated only on annotated data, and therefore must be optimized depending on the target to be predicted. The current system will help to select enzyme sequences and discover novel enzymatic reactions including missing links in metabolism and biosynthesis pathways for the production of useful substances using microorganisms.

Acknowledgment

First, the author would like to express the sincere gratitude to Prof. Michihiro Araki, his supervisor Prof. Chiaki Ogino and Prof. Akihiko Kondo for their guidance in all aspects of the author's research, including to conceive and design the research and their suggestions for ideas. The author would like to express the sincere gratitude to Dr. Yuki Kuriya, Masahiro Murata and Masaki Yamamoto for their useful advice and their continuous support of the research. The author is also particularly grateful to Assoc. Prof. Christopher John Vavricka Jr. for useful discussions of the author's research and collaboration of his research. The author would like to thank students and staffs in Kondo group, previous Araki and Hasunuma group researchers. Finally, the author with handicaps is very grateful to his parents for their daily support. The author could continue research for a long time thanks to their help.

List of Publications

Chapter II

Watanabe, N.; Murata, M.; Ogawa, T.; Vavricka, C. J.; Kondo, A.; Ogino, C.; Araki, M. Exploration and Evaluation of Machine Learning-Based Models for Predicting Enzymatic Reactions. *J. Chem. Inf. Model.* **2020**, 60, 1833–1843.

Chapter III

Watanabe, N.; Yamamoto, M.; Murata, M.; Vavricka, C. J.; Ogino, C.; Kondo, A.; Araki, M. Comprehensive Machine Learning Prediction of Extensive Enzymatic Reactions. *J. Phys. Chem. B*. **2022**, 126, 6762-6770.

Chapter IV

Watanabe, N.; Yamamoto, M.; Murata, M.; Kuriya, Y.; Araki, M. EnzymeNet: Residual Neural Networks model for Enzyme Commission numbers prediction (Under submission to Bioinformatics).

Coauthored Publications

Vavricka, CJ.; Takahashi, S.; **Watanabe, N**.; Takenaka, M.; Matsuda, M.; Yoshida, T.; Suzuki, R.; Kiyota, H.; Li, J.; Minami, H.; Ishii, J.; Tsuge, K.; Araki, M.; Kondo, A.; Hasunuma, T. Machine learning discovery of missing links that mediate alternative branches to plant alkaloids, Nature Communications, **2022**, 13, 1405.

Reference

(1) Choi, J. M.; Han, S. S.; Kim, H. S. Industrial Applications of Enzyme

Biocatalysis: Current Status and Future Aspects. *Biotechnol. Adv.* **2015**, *33* (7), 1443–1454. https://doi.org/10.1016/j.biotechadv.2015.02.014.

(2) Basso, A.; Serban, S. Industrial Applications of Immobilized Enzymes—A Review. *Mol. Catal.* **2019**, *479*, 110607. https://doi.org/10.1016/j.mcat.2019.110607.

(3) Na, D.; Kim, T. Y.; Lee, S. Y. Construction and Optimization of Synthetic Pathways in Metabolic Engineering. *Curr. Opin. Microbiol.* **2010**, *13* (3), 363–370. https://doi.org/10.1016/j.mib.2010.02.004.

(4) Böttcher, D.; Bornscheuer, U. T. Protein Engineering of Microbial Enzymes. *Curr. Opin. Microbiol.* **2010**, *13* (3), 274–282. https://doi.org/10.1016/j.mib.2010.01.010.

(5) Otte, K. B.; Hauer, B. Enzyme Engineering in the Context of Novel Pathways and Products. *Curr. Opin. Biotechnol.* **2015**, *35*, 16–22. https://doi.org/10.1016/j.copbio.2014.12.011.

(6) Lee, J. W.; Na, D.; Park, J. M.; Lee, J.; Choi, S.; Lee, S. Y. Systems Metabolic Engineering of Microorganisms for Natural and Non-Natural Chemicals. *Nat. Chem. Biol.* **2012**, *8* (6), 536–546. https://doi.org/10.1038/nchembio.970.

(7) Hammer, S. C.; Knight, A. M.; Arnold, F. H. Design and Evolution of Enzymes for Non-Natural Chemistry. *Curr. Opin. Green Sustain. Chem.* **2017**, *7*, 23–30. https://doi.org/10.1016/j.cogsc.2017.06.002.

(8) Yang, D.; Park, S. Y.; Park, Y. S.; Eun, H.; Lee, S. Y. Metabolic Engineering of Escherichia Coli for Natural Product Biosynthesis. *Trends Biotechnol.* **2020**, *38* (7), 745–765. https://doi.org/10.1016/j.tibtech.2019.11.007.

(9) Victorino da Silva Amatto, I.; Gonsales da Rosa-Garzon, N.; Antônio de Oliveira

Simões, F.; Santiago, F.; Pereira da Silva Leite, N.; Raspante Martins, J.; Cabral, H. Enzyme Engineering and Its Industrial Applications. *Biotechnol. Appl. Biochem.* **2021**, No. 2. https://doi.org/10.1002/bab.2117.

(10) Dasgupta, A.; Chowdhury, N.; De, R. K. Metabolic Pathway Engineering: Perspectives and Applications. *Comput. Methods Programs Biomed.* **2020**, *192*, 105436. https://doi.org/10.1016/j.cmpb.2020.105436.

(11) Agarwala, R.; Barrett, T.; Beck, J.; Benson, D. A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J. R.; Bryant, S. H.; Canese, K.; Cavanaugh, M.; Charowhas, C.; Clark, K.; Dondoshansky, I.; Feolo, M.; Fitzpatrick, L.; Funk, K.; Geer, L. Y.; Gorelenkov, V.; Graeff, A.; Hlavina, W.; Holmes, B.; Johnson, M.; Kattman, B.; Khotomlianski, V.; Kimchi, A.; Kimelman, M.; Kimura, M.; Kitts, P.; Klimke, W.; Kotliarov, A.; Krasnov, S.; Kuznetsov, A.; Landrum, M. J.; Landsman, D.; Lathrop, S.; Lee, J. M.; Leubsdorf, C.; Lu, Z.; Madden, T. L.; Marchler-Bauer, A.; Malheiro, A.; Meric, P.; Karsch-Mizrachi, I.; Mnev, A.; Murphy, T.; Orris, R.; Ostell, J.; O'Sullivan, C.; Palanigobu, V.; Panchenko, A. R.; Phan, L.; Pierov, B.; Pruitt, K. D.; Rodarmer, K.; Sayers, E. W.; Schneider, V.; Schoch, C. L.; Schuler, G. D.; Sherry, S. T.; Siyan, K.; Soboleva, A.; Soussov, V.; Starchenko, G.; Tatusova, T. A.; Thibaud-Nissen, F.; Todorov, K.; Trawick, B. W.; Vakatov, D.; Ward, M.; Yaschenko, E.; Zasypkin, A.; Zbicz, K. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46* (D1), D8–D13. https://doi.org/10.1093/nar/gkx1095.

(12) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Bye-AJee, H.; Cowley, A.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Castro, L. G.;

Figueira, L.; Garmiri, P.; Georghiou, G.; Gonzalez, D.; Hatton-Ellis, E.; Li, W.; Liu, W.; Lopez, R.; Luo, J.; Lussi, Y.; MacDougall, A.; Nightingale, A.; Palka, B.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G.; Rosanoff, S.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Speretta, E.; Turner, E.; Tyagi, N.; Volynkin, V.; Wardell, T.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; ArgoudPuy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; De Castro, E.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Nouspikel, N.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L. S.; Zhang, J. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45* (D1), D158–D169. https://doi.org/10.1093/nar/gkw1099.

(13) Thermes, C. Ten Years of Next-Generation Sequencing Technology. *Trends Genet.* **2014**, *30* (9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001.

(14) Mardis, E. R. The Impact of Next-Generation Sequencing Technology on Genetics. *Trends Genet.* **2008**, *24* (3), 133–141.

https://doi.org/10.1016/j.tig.2007.12.007.

(15) *UniProtKB/Swiss-Prot Release 2022_04 statistics. Expasy. 2022.* https://web.expasy.org/docs/relnotes/relstat.html (accessed October 19, 2022).

(16) *Current Release Statistics < Uniprot < EMBL-EBI. EMBL-EBI UniProt. 2022.* https://www.ebi.ac.uk/uniprot/TrEMBLstats (accessed October 19, 2022).

(17) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29. https://doi.org/10.2174/138161282466618052210520 2.

(18) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27–30. https://doi.org/10.1093/nar/28.1.27.

(19) Altschul, S. F.; Gish, W. Local Alignment Statistics. *Methods Enzymol.* **1996**, *266*, 460–480. https://doi.org/10.1007/978-1-84800-320-0_7.

(20) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. *BMC Bioinformatics* **2009**, *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

(21) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

(22) Stewart, R. D.; Auffret, M. D.; Warr, A.; Wiser, A. H.; Press, M. O.; Langford, K. W.; Liachko, I.; Snelling, T. J.; Dewhurst, R. J.; Walker, A. W.; Roehe, R.;

Watson, M. Assembly of 913 Microbial Genomes from Metagenomic Sequencing of the Cow Rumen. *Nat. Commun.* **2018**, *9* (1), 1–11. https://doi.org/10.1038/s41467-018-03317-6.

(23) Retzl, B.; Hellinger, R.; Muratspahić, E.; Pinto, M. E. F.; Bolzani, V. S.; Gruber, C. W. Discovery of a Beetroot Protease Inhibitor to Identify and Classify Plant-Derived Cystine Knot Peptides. *J. Nat. Prod.* **2020**, *83* (11), 3305–3314. https://doi.org/10.1021/acs.jnatprod.0c00648.

(24) Tian, W.; Arakaki, A. K.; Skolnick, J. EFICAz: A Comprehensive Approach for Accurate Genome-Scale Enzyme Function Inference. *Nucleic Acids Res.* **2004**, *32* (21), 6226–6239. https://doi.org/10.1093/nar/gkh956.

(25) Audit, B.; Levy, E. D.; Gilks, W. R.; Goldovsky, L.; Ouzounis, C. A. CORRIE: Enzyme Sequence Annotation with Confidence Estimates. *BMC Bioinformatics* **2007**, *8* (SUPPL. 4), 1–6. https://doi.org/10.1186/1471-2105-8-S4-S3.

(26) Yu, C.; Zavaljevski, N.; Desai, V.; Reifman, J. Genome-Wide Enzyme Annotation with Precision Control: Catalytic Families (CatFam) Databases. *Proteins Struct. Funct. Bioinforma.* **2009**, *74* (2), 449–460. https://doi.org/10.1002/prot.22167.

(27) Nursimulu, N.; Xu, L. L.; Wasmuth, J. D.; Krukov, I.; Parkinson, J. Improved Enzyme Annotation with EC-Specific Cutoffs Using DETECT V2. *Bioinformatics* **2018**, *34* (19), 3393–3395. https://doi.org/10.1093/bioinformatics/bty368.

(28) Piovesan, D.; Giollo, M.; Leonardi, E.; Ferrari, C.; Tosatto, S. C. E. INGA: Protein Function Prediction Combining Interaction Networks, Domain Assignments and Sequence Similarity. *Nucleic Acids Res.* **2015**, *43* (W1), W134–

W140. https://doi.org/10.1093/nar/gkv523.

(29) Quester, S.; Schomburg, D. EnzymeDetector: An Integrated Enzyme Function Prediction Tool and Database. *BMC Bioinformatics* **2011**, *12*. https://doi.org/10.1186/1471-2105-12-376.

(30) Tatusova, T.; Dicuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E. P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K. D.; Borodovsky, M.; Ostell, J. NCBI Prokaryotic Genome Annotation Pipeline. *Nucleic Acids Res.* **2016**, *44* (14), 6614–6624. https://doi.org/10.1093/nar/gkw569.

(31) Radivojac, P.; Clark, W.; Oron, T.; Schnoes, A. A Large-Scale Evaluation of Computational Protein Function Prediction. *Nature* **2013**, *10* (3), 221–227. https://doi.org/10.1038/nmeth.2340.A.

(32) Kulmanov, M.; Khan, M. A.; Hoehndorf, R. DeepGO: Predicting Protein Functions from Sequence and Interactions Using a Deep Ontology-Aware Classifier. *Bioinformatics* **2018**, *34* (4), 660–668. https://doi.org/10.1093/bioinformatics/btx624.

(33) Kulmanov, M.; Hoehndorf, R.; Cowen, L. DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics* **2020**, *36* (2), 422–429. https://doi.org/10.1093/bioinformatics/btz595.

(34) Meher, P. K.; Sahu, T. K.; Banchariya, A.; Rao, A. R. DIRProt: A Computational Approach for Discriminating Insecticide Resistant Proteins from Non-Resistant Proteins. *BMC Bioinformatics* **2017**, *18* (1), 1–14. https://doi.org/10.1186/s12859-017-1587-y.

(35) Arakaki, A. K.; Huang, Y.; Skolnick, J. EFICAz2: Enzyme Function Inference by a Combined Approach Enhanced by Machine Learning. *BMC Bioinformatics*

2009, *10*. https://doi.org/10.1186/1471-2105-10-107.

(36) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence. *Nucleic Acids Res.* **2006**, *34* (WEB. SERV. ISS.), 32–37. https://doi.org/10.1093/nar/gkl305.

(37) Rao, H. B.; Zhu, F.; Yang, G. B.; Li, Z. R.; Chen, Y. Z. Update of PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 2), 385–390. https://doi.org/10.1093/nar/gkr284.

(38) Cao, D. S.; Xu, Q. S.; Liang, Y. Z. Propy: A Tool to Generate Various Modes of Chou's PseAAC. *Bioinformatics*. 2013, pp 960–962. https://doi.org/10.1093/bioinformatics/btt072.

(39) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Wang, Y.; Webb, G. I.; Smith, A. I.; Daly, R. J.; Chou, K. C.; Song, J. IFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* **2018**, *34* (14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140.

(40) Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T. T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D. R.; Akutsu, T.; Webb, G. I.; Chou, K. C.; Smith, A. I.; Daly, R. J.; Li, J.; Song, J. ILearn: An Integrated Platform and Meta-Learner for Feature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data. *Brief. Bioinform.* **2020**, *21* (3), 1047–1057. https://doi.org/10.1093/bib/bbz041.

(41) Bonidia, R. P.; Santos, A. P. A.; De Almeida, B. L. S.; Stadler, P. F.; Da Rocha,

U. N.; Sanches, D. S.; De Carvalho, A. C. P. L. F. BioAutoML: Automated Feature Engineering and Metalearning to Predict Noncoding RNAs in Bacteria. *Brief. Bioinform.* **2022**, *23* (4), 1–13. https://doi.org/10.1093/bib/bbac218.

(42) Kumar, R.; Srivastava, A.; Kumari, B.; Kumar, M. Prediction of β-Lactamase and Its Class by Chou's Pseudo-Amino Acid Composition and Support Vector Machine. *J. Theor. Biol.* **2015**, *365*, 96–103. https://doi.org/10.1016/j.jtbi.2014.10.008.

(43) Bhasin, M.; Raghava, G. P. S. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* **2004**, *279* (22), 23262–23266. https://doi.org/10.1074/jbc.M401932200.

(44) Memon, S. A.; Khan, K. A.; Naveed, H. HECNet: A Hierarchical Approach to Enzyme Function Classification Using a Siamese Triplet Network. *Bioinformatics* **2020**, *36* (17), 4583–4589. https://doi.org/10.1093/bioinformatics/btaa536.

(45) Khan, K. A.; Memon, S. A.; Naveed, H. A Hierarchical Deep Learning Based Approach for Multi-Functional Enzyme Classification. *Protein Sci.* **2021**, *30* (9), 1935–1945. https://doi.org/10.1002/pro.4146.

(46) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep Learning Enables High-Quality and High-Throughput Prediction of Enzyme Commission Numbers. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (28), 13996–14001. https://doi.org/10.1073/pnas.1821905116.

(47) Nallapareddy, M. V.; Dwivedula, R. ABLE: Attention Based Learning for Enzyme Classification. *Comput. Biol. Chem.* **2021**, *94* (November 2020), 107558. https://doi.org/10.1016/j.compbiolchem.2021.107558.

(48) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34* (5), 760–769. https://doi.org/10.1093/bioinformatics/btx680.

(49) Zou, Z.; Tian, S.; Gao, X.; Li, Y. MlDEEPre: Multi-Functional Enzyme Function Prediction with Hierarchical Multi-Label Deep Learning. *Front. Genet.* **2019**, *9*, 714. https://doi.org/10.3389/fgene.2018.00714.

(50) Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal Deep Sequence Models for Protein Classification. *Bioinformatics* **2020**, *36* (8), 2401–2409. https://doi.org/10.1093/bioinformatics/btaa003.

(51) Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: A Tool for the Prediction of the Enzymatic Functions of Protein Sequences Based on the EC Nomenclature. *BMC Bioinformatics* **2018**, *19* (1), 1–13. https://doi.org/10.1186/s12859-018-2368-y.

(52) Shen, H. B.; Chou, K. C. EzyPred: A Top-down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364* (1), 53–59. https://doi.org/10.1016/j.bbrc.2007.09.098.

(53) Che, Y.; Ju, Y.; Xuan, P.; Long, R.; Xing, F. Identification of Multi-Functional Enzyme with Multi-Label Classifier. *PLoS One* **2016**, *11* (4), 1–13. https://doi.org/10.1371/journal.pone.0153503.

(54) Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z. SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence. *Nucleic Acids Res.* **2003**, *31* (13), 3692–3697. https://doi.org/10.1093/nar/gkg600.

(55) Li, Y. H.; Xu, J. Y.; Tao, L.; Li, X. F.; Li, S.; Zeng, X.; Chen, S. Y.; Zhang, P.; Qin, C.; Zhang, C.; Chen, Z.; Zhu, F.; Chen, Y. Z. SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity. *PLoS One* **2016**, *11* (8), 1–14. https://doi.org/10.1371/journal.pone.0155290.

(56) Seo, S.; Oh, M.; Park, Y.; Kim, S. DeepFam: Deep Learning Based Alignment-Free Method for Protein Family Modeling and Prediction. *Bioinformatics* **2018**, *34* (13), i254–i262. https://doi.org/10.1093/bioinformatics/bty275.

(57) Zhang, F.; Song, H.; Zeng, M.; Li, Y.; Kurgan, L.; Li, M. DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions. *Proteomics* **2019**, *19* (12). https://doi.org/10.1002/pmic.201900019.

(58) Sureyya Rifaioglu, A.; Doğan, T.; Jesus Martin, M.; Cetin-Atalay, R.; Atalay, V. DEEPred: Automated Protein Function Prediction with Multi-Task Feed-Forward Deep Neural Networks. *Sci. Rep.* **2019**, *9* (1), 1–16. https://doi.org/10.1038/s41598-019-43708-3.

(59) Hong, J.; Luo, Y.; Zhang, Y.; Ying, J.; Xue, W.; Xie, T.; Tao, L.; Zhu, F. Protein Functional Annotation of Simultaneously Improved Stability, Accuracy and False Discovery Rate Achieved by a Sequence-Based Deep Learning. *Brief. Bioinform.* **2019**, *00* (August), 1–11. https://doi.org/10.1093/bib/bbz081.

(60) Hamanaka, M.; Taneishi, K.; Iwata, H.; Ye, J.; Pei, J.; Hou, J.; Okuno, Y. CGBVS-DNN: Prediction of Compound-Protein Interactions Based on Deep Learning. *Mol. Inform.* **2017**, *36* (1), 1–10. https://doi.org/10.1002/minf.201600045.

(61) Li, G.; Rabe, K. S.; Nielsen, J.; Engqvist, M. K. M. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth. Biol.* **2019**, *8* (6), 1411–1420. https://doi.org/10.1021/acssynbio.9b00099.

(62) Wang, H.; Hu, X. Accurate Prediction of Nuclear Receptors with Conjoint Triad Feature. *BMC Bioinformatics* **2015**, *16* (1), 402. https://doi.org/10.1186/s12859-015-0828-1.

(63) Jamali, A. A.; Ferdousi, R.; Razzaghi, S.; Li, J.; Safdari, R.; Ebrahimie, E. DrugMiner: Comparative Analysis of Machine Learning Algorithms for Prediction of Potential Druggable Proteins. *Drug Discov. Today* **2016**, *21* (5), 718–724. https://doi.org/10.1016/j.drudis.2016.01.007.

(64) Yu, H.; Chen, J.; Xu, X.; Li, Y.; Zhao, H.; Fang, Y.; Li, X.; Zhou, W.; Wang, W.; Wang, Y. A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data. *PLoS One* **2012**, *7* (5). https://doi.org/10.1371/journal.pone.0037608.

(65) Manavalan, B.; Basith, S.; Shin, T. H.; Wei, L.; Lee, G. MAHTPred: A Sequence-Based Meta-Predictor for Improving the Prediction of Anti-Hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* **2019**, *35* (16), 2757–2765. https://doi.org/10.1093/bioinformatics/bty1047.

(66) Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning. *Bioinformatics* **2017**, *33* (21), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431.

(67) Karasuyama, M.; Inoue, K.; Nakamura, R.; Kandori, H.; Takeuchi, I.

Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach. *Sci. Rep.* **2018**, *8* (1), 1–4. https://doi.org/10.1038/s41598-018-33984-w.

(68) Xue, L.; Tang, B.; Chen, W.; Luo, J. DeepT3: Deep Convolutional Neural Networks Accurately Identify Gram-Negative Bacterial Type III Secreted Effectors Using the N-Terminal Sequence. *Bioinformatics* **2019**, *35* (12), 2051–2057. https://doi.org/10.1093/bioinformatics/bty931.

(69) Öztürk, H.; Ozkirimli, E.; Özgür, A. A Novel Methodology on Distributed Representations of Proteins Using Their Interacting Ligands. *Bioinformatics* **2018**, *34* (13), i295–i303. https://doi.org/10.1093/bioinformatics/bty287.

(70) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58* (1), 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

(71) Do, D. T.; Le, T. Q. T.; Le, N. Q. K. Using Deep Neural Networks and Biological Subwords to Detect Protein S-Sulfenylation Sites. *Brief. Bioinform.* **2021**, *22* (3), 1–11. https://doi.org/10.1093/bib/bbaa128.

(72) Mahdaddi, A.; Meshoul, S.; Belguidoum, M. EA-Based Hyperparameter Optimization of Hybrid Deep Learning Models for Effective Drug-Target Interactions Prediction. *Expert Syst. Appl.* **2021**, *185* (October 2020), 115525. https://doi.org/10.1016/j.eswa.2021.115525.

(73) Asgari, E.; Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* **2015**, *10* (11), 1–15. https://doi.org/10.1371/journal.pone.0141287.

(74) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word

Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*; 2013; pp 1–12.

(75) Vavricka, C. J.; Takahashi, S.; Watanabe, N.; Takenaka, M.; Matsuda, M.; Yoshida, T.; Suzuki, R.; Kiyota, H.; Li, J.; Minami, H.; Ishii, J.; Tsuge, K.; Araki, M.; Kondo, A.; Hasunuma, T. Machine Learning Discovery of Missing Links That Mediate Alternative Branches to Plant Alkaloids. *Nat. Commun.* **2022**, *13* (1). https://doi.org/10.1038/s41467-022-28883-8.

(76) Xie, Y.; Luo, X.; Li, Y.; Chen, L.; Ma, W.; Huang, J.; Cui, J.; Zhao, Y.; Xue, Y.; Zuo, Z.; Ren, J. DeepNitro: Prediction of Protein Nitration and Nitrosylation Sites by Deep Learning. *Genomics, Proteomics Bioinforma.* **2018**, *16* (4), 294–306. https://doi.org/10.1016/j.gpb.2018.04.007.

(77) Agrawal, P.; Kumar, S.; Singh, A.; Raghava, G. P. S.; Singh, I. K. NeuroPIpred: A Tool to Predict, Design and Scan Insect Neuropeptides. *Sci. Rep.* **2019**, *9* (1), 1–12. https://doi.org/10.1038/s41598-019-41538-x.

(78) Yu, K.; Zhang, Q.; Liu, Z.; Du, Y.; Gao, X.; Zhao, Q.; Cheng, H.; Li, X.; Liu, Z. X. Deep Learning Based Prediction of Reversible HAT/HDAC-Specific Lysine Acetylation. *Brief. Bioinform.* **2020**, *21* (5), 1798–1805. https://doi.org/10.1093/bib/bbz107.

(79) De Ferrari, L.; Aitken, S.; van Hemert, J.; Goryanin, I. EnzML: Multi-Label Prediction of Enzyme Classes Using InterPro Signatures. *BMC Bioinformatics* **2012**, *13* (1), 1–12. https://doi.org/10.1186/1471-2105-13-61.

(80) Dale, J. M.; Popescu, L.; Karp, P. D. Machine Learning Methods for Metabolic Pathway Prediction. *BMC Bioinformatics* **2010**, *11* (1), 15. https://doi.org/10.1186/1471-2105-11-15.

(81) Nath, N.; Mitchell, J. B. Is EC Class Predictable from Reaction Mechanism? *BMC Bioinformatics* **2012**, *13* (1), 60. https://doi.org/10.1186/1471-2105-13-60.

(82) Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S. Prediction of Metabolic Reactions Based on Atomic and Molecular Properties of Small-Molecule Compounds. *Bioinformatics* **2011**, *27* (11), 1537–1545. https://doi.org/10.1093/bioinformatics/btr177.

(83) Wang, Y.-C.; Wang, Y.; Yang, Z.-X.; Deng, N.-Y. Support Vector Machine Prediction of Enzyme Function with Conjoint Triad Feature and Hierarchical Context. *BMC Syst. Biol.* **2011**, *5 Suppl 1* (Suppl 1), S6. https://doi.org/10.1186/1752-0509-5-S1-S6.

(84) Kandaswamy, K. K.; Chou, K. C.; Martinetz, T.; Möller, S.; Suganthan, P. N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties. *J. Theor. Biol.* **2011**, *270* (1), 56–62. https://doi.org/10.1016/j.jtbi.2010.10.037.

(85) Dong, S.; Wang, P.; Abbas, K. A Survey on Deep Learning and Its Applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. https://doi.org/10.1016/j.cosrev.2021.100379.

(86) Sengupta, S.; Basak, S.; Saikia, P.; Paul, S.; Tsalavoutis, V.; Atiah, F.; Ravi, V.; Peters, A. A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends. *Knowledge-Based Syst.* **2020**, *194*, 105596. https://doi.org/10.1016/j.knosys.2020.105596.

(87) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.;

Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 590–596. https://doi.org/10.1038/s41586-021-03819-2.

(88) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Dustin Schaeffer, R.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Christopher Garcia, K.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(89) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

(90) Huang, K. Y.; Hsu, J. B. K.; Lee, T. Y. Characterization and Identification of Lysine Succinylation Sites Based on Deep Learning Method. *Sci. Rep.* **2019**, *9* (1), 1–15. https://doi.org/10.1038/s41598-019-52552-4.

(91) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug-Target Binding Affinity Prediction. *Bioinformatics* **2018**, *34* (17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593.

(92) Shao, D.; Huang, L.; Wang, Y.; He, K.; Cui, X.; Wang, Y.; Ma, Q.; Cui, J. DeepSec: A Deep Learning Framework for Secreted Protein Discovery in Human Body Fluids. *Bioinformatics* **2021**, No. August, 1–8. https://doi.org/10.1093/bioinformatics/btab545.

(93) Dao, F.-Y.; Lv, H.; Su, W.; Sun, Z.-J.; Huang, Q.-L.; Lin, H. IDHS-Deep: An Integrated Tool for Predicting DNase I Hypersensitive Sites by Deep Neural Network. *Brief. Bioinform.* **2021**, *00* (December 2020), 1–8. https://doi.org/10.1093/bib/bbab047.

(94) Tipton, K.; Boyce, S. History of the Enzyme Nomenclature System. *Bioinformatics* **2000**, *16* (1), 34–40. https://doi.org/10.1093/bioinformatics/16.1.34.

(95) Cornish-Bowden, A. Current IUBMB Recommendations on Enzyme Nomenclature and Kinetics. *Perspect. Sci.* **2014**, *1* (1–6), 74–87. https://doi.org/10.1016/j.pisc.2014.02.006.

(96) *Moss, G, P. Enzyme Nomenclature. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). 2022.*

(97) Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Res.* **2017**, *45* (D1), D380–D388. https://doi.org/10.1093/nar/gkw952.

(98) Friedberg, I. Automated Protein Function Prediction - The Genomic Challenge. *Brief. Bioinform.* **2006**, *7* (3), 225–242. https://doi.org/10.1093/bib/bbl004.

(99) Sleator, R. D.; Walsh, P. An Overview of in Silico Protein Function Prediction. *Arch. Microbiol.* **2010**, *192* (3), 151–155. https://doi.org/10.1007/s00203-010-

0549-9.

(100) Libbrecht, M. W.; Noble, W. S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16* (6), 321–332. https://doi.org/10.1038/nrg3920.

(101) Yamanishi, Y.; Hattori, M.; Kotera, M.; Goto, S.; Kanehisa, M. E-Zyme: Predicting Potential EC Numbers from the Chemical Transformation Pattern of Substrate-Product Pairs. *Bioinformatics* **2009**, *25* (12), 179–186. https://doi.org/10.1093/bioinformatics/btp223.

(102) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.* **2004**, *126* (50), 16487–16498. https://doi.org/10.1021/ja0466457.

(103) Matsuta, Y.; Ito, M.; Tohsato, Y. ECOH: An Enzyme Commission Number Predictor Using Mutual Information and a Support Vector Machine. *Bioinformatics* **2013**, *29* (3), 365–372. https://doi.org/10.1093/bioinformatics/bts700.

(104) Moriya, Y.; Shigemizu, D.; Hattori, M.; Tokimatsu, T.; Kotera, M.; Goto, S.; Kanehisa, M. PathPred: An Enzyme-Catalyzed Metabolic Pathway Prediction Server. *Nucleic Acids Res.* **2010**, *38* (SUPPL. 2), 138–143. https://doi.org/10.1093/nar/gkq318.

(105) Tian, S.; Djoumbou-Feunang, Y.; Greiner, R.; Wishart, D. S. CypReact: A Software Tool for in Silico Reactant Prediction for Human Cytochrome P450 Enzymes. *J. Chem. Inf. Model.* **2018**, *58* (6), 1282–1291. https://doi.org/10.1021/acs.jcim.8b00035.

(106) Mellor, J.; Grigoras, I.; Carbonell, P.; Faulon, J. L. Semisupervised Gaussian Process for Automated Enzyme Search. *ACS Synth. Biol.* **2016**, *5* (6), 518–528. https://doi.org/10.1021/acssynbio.5b00294.

(107) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J. L. RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers. *Metab. Eng.* **2018**, *45* (December 2017), 158–170. https://doi.org/10.1016/j.ymben.2017.12.002.

(108) Cai, Y.; Yang, H.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Multiclassification Prediction of Enzymatic Reactions for Oxidoreductases and Hydrolases Using Reaction Fingerprints and Machine Learning Methods. *J. Chem. Inf. Model.* **2018**, *58* (6), 1169–1181. https://doi.org/10.1021/acs.jcim.7b00656.

(109) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(110) Katoh, K.; Misawa, K.; Kuma, K. I.; Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30* (14), 3059–3066. https://doi.org/10.1093/nar/gkf436.

(111) Rice, P.; Longden, L.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16* (6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

(112) Liu, B. BioSeq-Analysis: A Platform for DNA, RNA and Protein Sequence Analysis Based on Machine Learning Approaches. *Briefings in Bioinformatics*. 2018, pp 1280–1294. https://doi.org/10.1093/bib/bbx165.

(113) Gromiha, M. M.; Suwa, M. Influence of Amino Acid Properties for

Discriminating Outer Membrane Proteins at Better Accuracy. *Biochim. Biophys. Acta - Proteins Proteomics* **2006**, *1764* (9), 1493–1497. https://doi.org/10.1016/j.bbapap.2006.07.005.

(114) Ezzat, A.; Wu, M.; Li, X. L.; Kwoh, C. K. Computational Prediction of Drug-Target Interactions Using Chemogenomic Approaches: An Empirical Survey. *Brief. Bioinform.* **2018**, *20* (4), 1337–1357. https://doi.org/10.1093/bib/bby002.

(115) *Dragon 7.0 – Kode Chemoinformatics. Kode Chemoinformatics. 2019.* https://chm.kode-solutions.net/pf/dragon-7-0/ (accessed October 19, 2022).

(116) Amos, R. I. J.; Haddad, P. R.; Szucs, R.; Dolan, J. W.; Pohl, C. A. Molecular Modeling and Prediction Accuracy in Quantitative Structure-Retention Relationship Calculations for Chromatography. *TrAC - Trends Anal. Chem.* **2018**, *105*, 352–359. https://doi.org/10.1016/j.trac.2018.05.019.

(117) Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–39. https://doi.org/10.1145/1961189.1961199.

(118) Wu, T. F.; Lin, C. J.; Weng, R. C. Probability Estimates for Multi-Class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **2004**, *5*, 975–1005.

(119) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(120) I. T. Jolliffe. *Principal Components Analysis*; Springer-Verlag: New York, 1989.

(121) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.

(122) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Res.* **2016**, *44* (D1), D279–D285. https://doi.org/10.1093/nar/gkv1344.

(123) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and Sar Studies. Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984**, *19* (1), 66–70.

(124) Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **1950**, *37* (1), 17–23.

(125) Geary, R. The Contiguity Ratio and Statistical Mapping. *Inc. Stat.* **1954**, *5* (3), 115–146.

(126) *AAindex. GenomeNet. 2017.* http://www.genome.ad.jp/dbget/aaindex.html (accessed October 13, 2022).

(127) Kawashima, S.; Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **2000**, *28* (1), 374. https://doi.org/10.1093/nar/28.1.374.

(128) Dubchak, I.; Muchnik, I.; Holbrook, S. R.; Kim, S. H. Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (19), 8700–8704. https://doi.org/10.1073/pnas.92.19.8700.

(129) Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S. H. Recognition of a Protein Fold in the Context of the SCOP Classification. *Proteins Struct. Funct. Genet.* **1999**, *35* (4), 401–407. https://doi.org/10.1002/(SICI)1097-0134(19990601)35:4<401::AID-PROT3>3.0.CO;2-K.

(130) Chou, K. C. Prediction of Protein Subcellar Locations by Incorporating Quasi-Sequence-Order Effect.Pdf. *Biochem. Biophys. Res. Commun.* **2000**, *19* (2), 477–483.

(131) Chou, K. C. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* **2005**, *21* (1), 10–19. https://doi.org/10.1093/bioinformatics/bth466.

(132) Gutman, I.; Rücker, C.; Rücker, G. On Walks in Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 739–745. https://doi.org/10.1021/ci000149u.

(133) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, Networks and Connectivity Indices. *Proteomics* **2008**, *8* (4), 750–778. https://doi.org/10.1002/pmic.200700638.

(134) Gálvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (3), 520–525. https://doi.org/10.1021/ci00019a008.

(135) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18* (4–5), 464–477. https://doi.org/10.1016/S1093-3263(00)00068-1.

(136) Roy, K.; Ghosh, G. Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies. *Internet Electron. J. Mol. Des.* **2003**, *2*, 599–620.

(137) Estrada, E. Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.

(138) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.*

1987, *27*, 21–35.

(139) Hall, L. H. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.

(140) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039–1045. https://doi.org/10.1021/ci00028a014.

(141) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *J. Comput. Aided. Mol. Des.* **2003**, *17* (10), 687–698. https://doi.org/10.1023/B:JCAM.0000017375.61558.ad.

(142) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chemie - Int. Ed.* **1999**, *38* (19), 2894–2896. https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F.

(143) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. 30-Modelling and Prediction by WHIM Descriptors. Part 8. Toxicity and Physico-Chemical Properties of Environmental Priority Chemicals by 2D-TI and 3D-WHIM Descriptors.Pdf. *SAR QSAR Environ. Res.* **1997**, *7*, 173–193.

(144) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. https://doi.org/10.1021/jm000942e.

(145) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40* (1), 127–130.

(146) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102* (21), 3762–3772. https://doi.org/10.1021/jp980230o.

(147) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Determination of McGowan Volumes for Ions and Correlation with van Der Waals Volumes. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1848–1854. https://doi.org/10.1021/ci0341114.

(148) Verhaar, H. J. M.; van Leeuwen, C. J.; Hermens, J. L. M. Classifying Environmental-Pollutants .1. Structure-Activity-Relationships for Prediction of Aquatic Toxicity. *Chemosphere* **1992**, *25* (4), 471–491. https://doi.org/10.1016/0045-6535(92)90280-5.

(149) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. https://doi.org/10.1023/A:1022627411411.

(150) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(151) Utgoff, P. E. Incremental Induction of Decision Trees. *Mach. Learn.* **1989**, *4* (2), 161–186. https://doi.org/10.1023/A:1022699900025.

(152) Cover, T. M.; Hart, P. E. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13* (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

(153) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185.

https://doi.org/10.1080/00031305.1992.10475879.

(154) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks* **1989**, *2* (5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

(155) Gardner, M. W.; Dorling, S. R. Artificial Neural Networks (the Multilayer Perceptron) - a Review of Applications in the Atmospheric Sciences. *Atmos. Environ.* **1998**, *32* (14–15), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

(156) Rumelhart, D. .; Hinton, G. .; Williams, R. . Learning Internal Representations By Error Propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*; MIT Press, 1986; pp 318–362.

(157) Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, 2008.

(158) Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34* (1), 1–47.

(159) Kerepesi, C.; Daróczy, B.; Sturm, Á.; Vellai, T.; Benczúr, A. Prediction and Characterization of Human Ageing-Related Proteins by Using Machine Learning. *Sci. Rep.* **2018**, *8*, 4094. https://doi.org/10.1038/s41598-018-22240-w.

(160) Zhou, Y.; Li, G.; Dong, J.; Xing, X. hui; Dai, J.; Zhang, C. MiYA, an Efficient Machine-Learning Workflow in Conjunction with the YeastFab Assembly Strategy for Combinatorial Optimization of Heterologous Metabolic Pathways in Saccharomyces Cerevisiae. *Metab. Eng.* **2018**, *47* (January), 294–302. https://doi.org/10.1016/j.ymben.2018.03.020.

(161) Zhang, Y.; Xie, R.; Wang, J.; Leier, A.; Marquez-Lago, T. T.; Akutsu, T.; Webb,

G. I.; Chou, K. C.; Song, J. Computational Analysis and Prediction of Lysine Malonylation Sites by Exploiting Informative Features in an Integrative Machine-Learning Framework. *Brief. Bioinform.* **2019**, *20* (6), 2185–2199. https://doi.org/10.1093/bib/bby079.

(162) Wang, Y.; Li, F.; Bharathwaj, M.; Rosas, N. C.; Leier, A.; Akutsu, T.; Webb, G. I.; Marquez-Lago, T. T.; Li, J.; Lithgow, T.; Song, J. DeepBL: A Deep Learning-Based Approach for in Silico Discovery of Beta-Lactamases. *Brief. Bioinform.* **2020**, *00* (September), 1–12. https://doi.org/10.1093/bib/bbaa301.

(163) Lawson, C. E.; Martí, J. M.; Radivojevic, T.; Jonnalagadda, S. V. R.; Gentz, R.; Hillson, N. J.; Peisert, S.; Kim, J.; Simmons, B. A.; Petzold, C. J.; Singer, S. W.; Mukhopadhyay, A.; Tanjore, D.; Dunn, J. G.; Garcia Martin, H. Machine Learning for Metabolic Engineering: A Review. *Metab. Eng.* **2021**, *63*, 34–60. https://doi.org/10.1016/j.ymben.2020.10.005.

(164) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10* (2), 1210–1223. https://doi.org/10.1021/acscatal.9b04321.

(165) Aromolaran, O.; Aromolaran, D.; Isewon, I.; Oyelade, J. Machine Learning Approach to Gene Essentiality Prediction: A Review. *Brief. Bioinform.* **2022**, *23* (1), bbab419. https://doi.org/10.1093/bib/bbab419.

(166) *RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed October 14, 2021)*.

(167) Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

(168) Charoenkwan, P.; Nantasenamat, C.; Hasan, M. M.; Manavalan, B.; Shoombuatong, W. BERT4Bitter: A Bidirectional Encoder Representations from Transformers (BERT)-Based Model for Improving the Prediction of Bitter Peptides. *Bioinformatics* **2021**, *37* (17), 2556–2562. https://doi.org/10.1093/bioinformatics/btab133.

(169) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(170) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. **2016**, arXiv preprint arXiv: 1603.04467v2.

(171) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2625.

(172) Zeng, J.; Giese, T. J.; Ekesan, Ş.; York, D. M. Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution. *J. Chem. Theory*

Comput. **2021**, *17* (11), 6993–7009. https://doi.org/10.1021/acs.jctc.1c00201.

(173) Ali, M.; Ishqi, H. M.; Husain, Q. Enzyme Engineering: Reshaping the Biocatalytic Functions. *Biotechnol. Bioeng.* **2020**, *117* (6), 1877–1894. https://doi.org/10.1002/bit.27329.

(174) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision, Springer, Cham*; 2016; pp 630–645. https://doi.org/10.1007/978-3-319-46493-0_38.

(175) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics* **2019**, *35* (14), 2403–2410. https://doi.org/10.1093/bioinformatics/bty1006.

(176) Shi, W.; Lemoine, J. M.; Shawky, A. E. M. A.; Singha, M.; Pu, L.; Yang, S.; Ramanujam, J.; Brylinski, M. BionoiNet: Ligand-Binding Site Classification with off-the-Shelf Deep Neural Network. *Bioinformatics* **2020**, *36* (10), 3077–3083. https://doi.org/10.1093/bioinformatics/btaa094.

(177) Sanderson, T.; Bileschi, M. L.; Belanger, D.; Colwell, L. J. ProteInfer: Deep Networks for Protein Functional Inference. *bioRxiv* **2021**, 2021.09.20.461077. https://doi.org/10.1101/2021.09.20.461077.

(178) Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2019; Vol. 1, pp 4171–4186.

(179) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. In *International Conference on Learning Representations*; 2020; pp 1–17.

(180) Mitchell, A. L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M. R.; Kale, V.; Potter, S. C.; Richardson, L. J.; Sakharova, E.; Scheremetjew, M.; Korobeynikov, A.; Shlemov, A.; Kunyavskaya, O.; Lapidus, A.; Finn, R. D. MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Res.* **2020**, *48* (D1), D570–D578. https://doi.org/10.1093/nar/gkz1035.

Doctoral Dissertation, Kobe University

“Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations”, 194 pages.

Submitted on January, 17, 2023.

When published on the Kobe University institutional repository /Kernel/, the publication date shall appear on the cover of the repository version.