OXFORD

Briefings in Bioinformatics, 2022, **23**(3), 1–16
https://doi.org/10.1093/bib/bbac071
Problem Solving Protocol

deepNEC: a novel alignment-free tool for the identification and classification of nitrogen biochemical network-related enzymes using deep learning

Naveen Duhan, Jeanette M. Norton and Rakesh Kaundal

Corresponding author. Rakesh Kaundal, Department of Plants, Soils, and Climate, College of Agriculture and Applied Sciences; Bioinformatics Facility, Center for Integrated BioSystems; Department of Computer Science, College of Science; Utah State University, Logan, UT 84322 USA. Tel.: +1 (435) 797-4117; Fax: +1 (435) 797-2766; E-mail: rkaundal@usu.edu

Abstract

Nitrogen is essential for life and its transformations are an important part of the global biogeochemical cycle. Being an essential nutrient, nitrogen exists in a range of oxidation states from +5 (nitrate) to −3 (ammonium and amino-nitrogen), and its oxidation and reduction reactions catalyzed by microbial enzymes determine its environmental fate. The functional annotation of the genes encoding the core nitrogen network enzymes has a broad range of applications in metagenomics, agriculture, wastewater treatment and industrial biotechnology. This study developed an alignment-free computational approach to determine the predicted nitrogen biochemical network-related enzymes from the sequence itself. We propose deepNEC, a novel end-to-end feature selection and classification model training approach for nitrogen biochemical network-related enzyme prediction. The algorithm was developed using Deep Learning, a class of machine learning algorithms that uses multiple layers to extract higher-level features from the raw input data. The derived protein sequence is used as an input, extracting sequential and convolutional features from raw encoded protein sequences based on classification rather than traditional alignment-based methods for enzyme prediction. Two large datasets of protein sequences, enzymes and non-enzymes were used to train the models with protein sequence features like amino acid composition, dipeptide composition (DPC), conformation transition and distribution, normalized Moreau–Broto (NMBroto), conjoint and quasi order, etc. The *k*-fold cross-validation and independent testing were performed to validate our model training. deepNEC uses a four-tier approach for prediction; in the first phase, it will predict a query sequence as enzyme or non-enzyme; in the second phase, it will further predict and classify enzymes into nitrogen biochemical network-related enzymes or non-nitrogen metabolism enzymes; in the third phase, it classifies predicted enzymes into nine nitrogen metabolism classes; and in the fourth phase, it predicts the enzyme commission number out of 20 classes for nitrogen metabolism. Among all, the DPC + NMBroto hybrid feature gave the best prediction performance (accuracy of 96.15% in *k*-fold training and 93.43% in independent testing) with an Matthews correlation coefficient (0.92 training and 0.87 independent testing) in phase I; phase II (accuracy of 99.71% in *k*-fold training and 98.30% in independent testing); phase III (overall accuracy of 99.03% in *k*-fold training and 98.98% in independent testing); phase IV (overall accuracy of 99.05% in *k*-fold training and 98.18% in independent testing), the DPC feature gave the best prediction performance. We have also implemented a homology-based method to remove false negatives. All the models have been implemented on a web server (prediction tool), which is freely available at http://bioinfo.usu.edu/deepNEC/.

Keywords: CNN, computational modeling, deep learning, enzyme classification, metagenomics, N-biochemical network, neural networks, prediction, nitrogen cycle, nitrification, denitrification

Introduction

Nitrogen (N) is an essential element of all life and is cycled throughout ecosystems through a number of interconnected biochemical reactions catalyzed by microbial enzymes. The biogeochemical processes which transform nitrogen between its many chemical forms, both living and non-living, are collectively termed as the nitrogen cycle [1, 2]. There are a number of microorganisms involved in the nitrogen cycle, and nitrogen is present in a range of oxidation states, from +5 in nitrate to −3 in ammonia and amino acids. There are four main nitrogen reducing processes: nitrogen fixation, assimilatory nitrate reduction, dissimilatory nitrate reduction and denitrification and two main oxidation pathways: nitrification and anaerobic ammonia oxidation in the N-biochemical network [3, 4]. The nitrogen cycle is a particularly complicated biochemical network as one reaction's end product is frequently the substrate

Naveen Duhan is a PhD student in the Center for Integrated BioSystems/Department of Plants, Soils, and Climate, College of Agriculture and Applied Sciences, Utah State University, USA.
Jeanette M. Norton is professor of Soil Microbiology in the Department of Plants, Soils and Climate, an Ecology Center associate and an adjunct professor in Biology at Utah State University.
Rakesh Kaundal is an assistant professor of Bioinformatics in the Department of Plants, Soils, and Climate, College of Agriculture and Applied Sciences, and director of the Bioinformatics Facility, Center for Integrated BioSystems, Utah State University, USA. He also has an adjunct faculty appointment in the Department of Computer Science, USU.

Received: November 12, 2021. **Revised:** January 25, 2022. **Accepted:** February 10, 2022

for another reaction [5]. Enzymes from microbes engaged in the N-cycle also have various regulatory effects, including either feedback or feed-forward controls, which is a fascinating aspect of the cycle [6, 7]. To better understand the nitrogen cycle in natural environments, understanding these feedback/feed-forward regulatory mechanisms, as well as their evolutionary dynamics, is critical. Recent advances in microbial genomics have revealed information on the whole range of genes encoding these enzymes involved in the nitrogen cycle [8–10].

With the advent of next-generation sequencing technologies, many environmental metagenomes are being sequenced. The functional annotation of these metagenomes plays a vital role in deciphering the functions of the microbial community and its associated enzymes. The classical approach uses experimental methods such as protein purification and enzymatic assays [11]. Conducting these focused biochemical experiments, however, requires significant time and expertise that cannot cope with the rapid increase in the metagenomic data volume. Advancements in computational methods may assist biologists in characterizing the diversity and complexity of the metagenomic data and give further insights prior to experimental validation.

According to Swiss-Prot [12, 13] latest release, a total of 258 733 enzymes are present among all 565 254 manually annotated proteins. These enzymes are classified into seven classes using the Enzyme Commission (EC) number [14], the most accepted numerical enzyme classification scheme.

Many computational methods have already been proposed to classify enzymes based on their EC numbers. As the first report addresses this problem with machine learning and sequence information [15], this problem has been investigated in three different ways. Researchers [16–20] generally predicted enzyme function by predicting the enzyme structure first because generally the function of protein is determined by its structure. Upon structure prediction, a database or library was searched for EC numbers and assigned to the template structure. Structure prediction is still relatively time-consuming, however. Another common assumption is that similar protein sequences tend to have similar functions. Therefore, several studies have been reported that utilize sequence similarities [21–25]. Homology-based methods are widely used to decipher the function of an enzyme, but they fail when there is no significant similarity found. Alignment-based methods have been widely used to determine the role of a protein sequence in an N-biochemical network. Currently, the most widely used methods are to extract features from protein sequences and then train and classify using machine learning approaches [26–44]. Although machine learning methods have been applied for two decades to address this problem with several web servers and software available, no methods have been developed that focus specifically on predicting nitrogen biochemical network-related enzymes. In this study, we present an alignment-free computational tool using deep learning, a class of machine learning, for the prediction of nitrogen biochemical network-related enzymes from raw protein sequences. We propose a web server deepNEC for that purpose.

Materials and methods

We have implemented deepNEC in four phases: Phase I identifies a query protein sequence as enzyme or non-enzyme; Phase II classifies the predicted enzyme sequence into nitrogen metabolism or non-nitrogen metabolism enzymes; Phase III classifies enzymes into nine nitrogen metabolism enzyme classes (four reduction, one oxidation and four hybrid) and Phase IV predicts the EC number associated with the nitrogen biochemical network class predicted in Phase III (Figure 1).

Preparation of deepNEC datasets

The phase I dataset was prepared by processing 563 972 enzyme protein sequences from the Swiss-Prot dataset available in February 2020 [12]. The dataset was separated into enzymes and non-enzymes based on their annotation downloaded from UniProt. Enzyme sequences of fragments with less than 50 amino acids annotated were omitted to prevent fragment data. We have used CD-HIT [45] with a 40% similarity threshold to sift the raw dataset to eradicate redundancy biases resulting in 28 287 enzyme sequences with low homology. For the non-enzyme part, 28 287 non-enzyme sequences were randomly extracted from Swiss-Prot non-enzyme sequences using a custom python script. These datasets (56 574 protein sequences) were separated into a training dataset (26 787 enzyme and 26 787 non-enzyme sequences) and an independent dataset (1500 enzyme and 1500 non-enzyme sequences) to test the generalization accuracy of the method proposed.

The phase II dataset was prepared by separating the nitrogen biochemical network-related enzymes and non-nitrogen biochemical network-related protein sequences. The sequences with >80% similarity were removed using CD-HIT [45] to eradicate redundancy bias resulting in 24 996 nitrogen biochemical network-related enzyme sequences with low homology and 27 884 non-nitrogen metabolism sequences. Finally, a dataset with these sequences was prepared by separating them into training and independent testing sets.

The phase III dataset was generated by categorizing nitrogen biochemical network enzymes into nine groups using KEGG pathways map00910 (https://www.genome.jp/pathway/map00910). The N-biochemical network has six sub-pathways (nitrogen fixation, assimilatory nitrate reduction, dissimilatory nitrate reduction, denitrification, nitrification and anammox). We downloaded the enzymes associated with these sub-pathways from NCBI (https://www.ncbi.nlm.nih.gov/) and UniProt

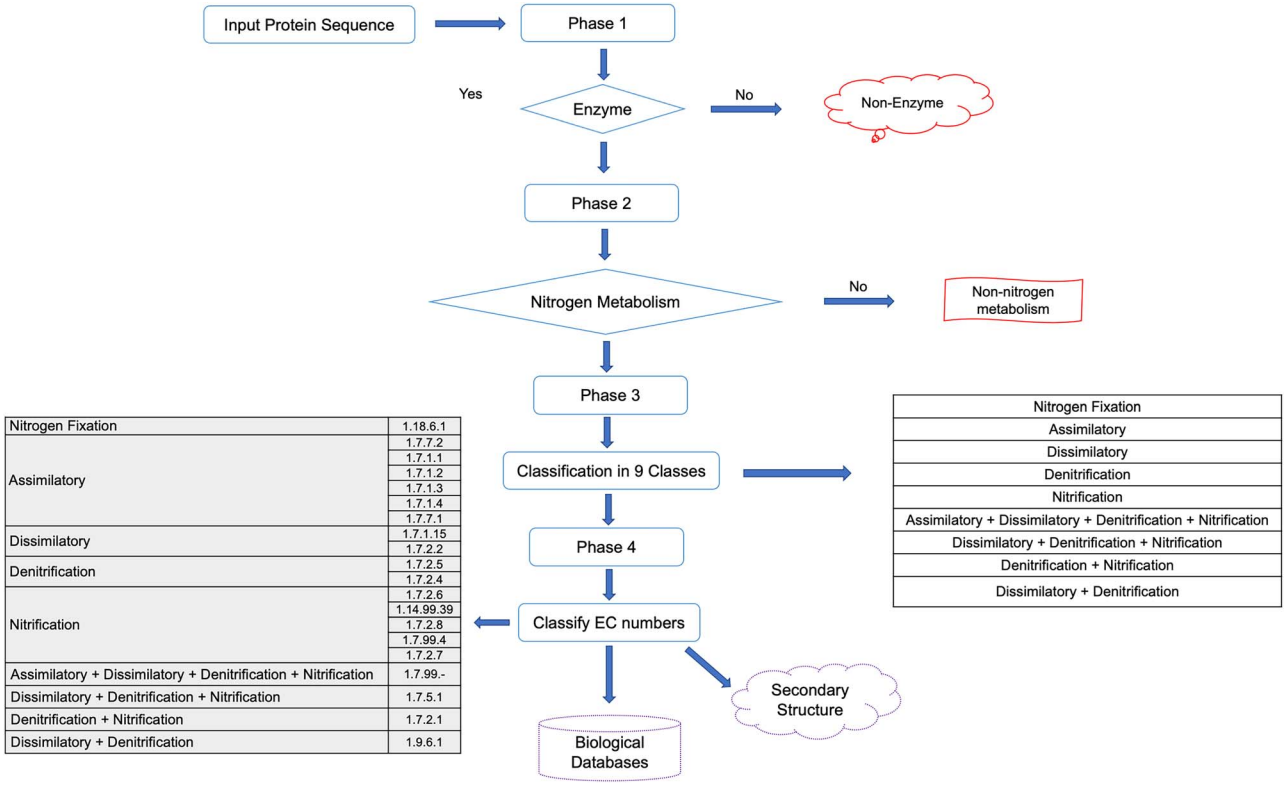

Figure 1. An overall workflow design of deepNEC.

Table 1. Nitrogen metabolism-related nine classes used in the study

Class	Train	Test
Nitrogen fixation	16 305	200
Assimilatory	24 466	200
Dissimilatory	15 253	200
Denitrification	9518	200
Nitrification	26 344	200
Assimilatory + dissimilatory + denitrification + nitrification	12 494	200
Dissimilatory + denitrification + nitrification	14 423	200
Denitrification + nitrification	24 895	200
Dissimilatory + denitrification	4323	200

(https://www.uniprot.org/) and classified them accordingly. We established four hybrid classes for enzymes that are engaged in several sub-pathways. The dataset was then divided into training and independent testing datasets. These nine nitrogen biochemical network classes are presented in Table 1.

The phase IV dataset was prepared by separating nitrogen biochemical network enzymes with their EC number. That resulted in 20 classes with EC numbers. Then these sequences were separated into training and independent testing datasets. These 20 ECs are presented in Table 2.

Sequence representation in deepNEC

In general, input data with a fixed size are accepted by the machine learning methods. The deep learning system eliminates the need for uniformity of manual dimensioning and manually built features which are impossible to support with the increasing quantity and sophistication of data by simultaneously carrying out feature reconstruction and classifier training. We have tried multiple sequence representations for training models for all phases. Best sequence representation is discussed below, and the training statistics of other sequence representations tried are presented in Supplementary File 1.

In Phase I, we have converted the protein sequence into a 640-length vector by combining two protein features [dipeptide composition (DPC) and normalized Moreau–Broto (NMBroto)]. The DPC has been determined to encapsulate the worldwide details on each protein sequence using the sequence order effects. This number, which has a fixed pattern of 400 (20 × 20), provides information on the structure of the amino acid and the local amino acid order. The following equation was used to measure the fraction of each dipeptide:

$$f(r,s) = \frac{N_{rs}}{N-1} r, s = 1, 2, \ldots, 20$$

where N_{rs} is the number of dipeptides represented by the amino acid type r and type s, N is the length of the sequence. NMBroto is an autocorrelation-based feature that describes the degree of association between two objects (protein or peptide sequences) in terms of their particular structural or physicochemical properties based on amino acid properties distributed in a sequence. The normalized Moreau–Broto autocorrelation

Table 2. Twenty classes of enyzmes related to nitrogen metabolism

Nitrogen metabolism type	EC number	Train	Test
Nitrogen fixation (nitrogenase)	1.18.6.1	16 305	200
Assimilatory nitrate reduction	1.7.7.2	5425	200
Assimilatory nitrate reduction	1.7.1.1	2095	200
Assimilatory nitrate reduction	1.7.1.2	5893	200
Assimilatory nitrate reduction	1.7.1.3	1522	200
Assimilatory nitrate reduction	1.7.1.4	7290	200
Assimilatory nitrate reduction	1.7.7.1	1250	200
Dissimilatory nitrate reduction	1.7.1.15	2036	200
Dissimilatory nitrate reduction	1.7.2.2	13 017	200
Denitrification (nitric oxide reductase)	1.7.2.5	4750	200
Denitrification (nitrous oxide reductase)	1.7.2.4	4568	200
Nitrification (hydroxylamine dehydrogenase)	1.7.2.6	1924	200
Nitrification (ammonia monooxygenase)	1.14.99.39	7374	200
Nitrification (hydrazine dehydrogenase)	1.7.2.8	2062	200
Nitrification	1.7.99.4	11 572	200
Nitrification (hydrazine synthase)	1.7.2.7	2612	200
Assimilatory + dissimilatory + denitrification + nitrification	1.7.9.9.	12 494	200
Dissimilatory + denitrification + nitrification (nitrate reductase)	1.7.5.1	14 423	200
Denitrification + nitrification [nitrite (oxido) reductase]	1.7.2.1	24 895	200
Dissimilatory + denitrification (periplasmic nitrate reductase)	1.9.6.1	4323	200

is defined as

$$I(d) = \frac{\sum_{i=1}^{N-d} (P_i * P_{i+d})}{N-d} d = 1, 2, \ldots, 30$$

where d is the lag of the autocorrelation, P_i is the value of ith amino acid in a property entry of AAindex.

In Phases II–IV, we have converted the protein sequence into a 400-length vector by the DPC feature as discussed above.

deepNEC training model architecture

deepNEC comprises two separate convolution neural networks (CNNs) that conduct two distinct classification tasks with a single protein sequence as input data. The CNN for Phases I and II consists of two 2D convolution layers, two max-pooling layers, three dropout layers, two batch normalization, one flattening layer and three fully connected layers and finally to an output (Figure 2A). The first layer in a CNN is always a convolutional layer. In particular, the first layer of our CNN includes a 640-size vector on which we have implemented our 2D convolutional operations with some default parameters, including $n \times n$ kernel size, f filters, 1×1 steps, and 1×1 zero-padding and then a convolutional operation was used to filter essential features of the motif. We learned about the network by modifying the hyper-parameters described above to find the appropriate choices. In addition, to reduce the size of matrix calculation, eliminate nonmaximal values and control overfitting, we used a 2D max-pooling layer with a 1×1 stride. Before three fully connected layers, we added a flattening layer to flatten the input. The first fully connected layer consists of 512 hidden nodes followed by a second fully connected layer of 256 hidden nodes and finally a fully connected layer of 2 hidden nodes for binary classification of deepNEC Phase I model (Figure 2A).

The CNN for Phases III and IV consists of two 2D convolution layers, two max-pooling layers, two dropout layers, two batch normalizations, one flattening layer and two fully connected layers, and finally to an output (Figure 1B). The first layer in a CNN is always a convolutional layer. In particular, the first layer of our CNN includes a 400-size vector on which we have implemented 2D convolutional operations with some default parameters, including $n \times n$ kernel size, f filters, 1×1 steps and 1×1 zero-padding. We used a 2D max-pooling layer with a 1×1 stride. Before three fully connected layers, we added a flattening layer to flatten the input. The first fully connected layer consists of 512 hidden nodes followed by a second fully connected layer of 14 hidden nodes with softmax activation for binary classification of deepNEC Phase-II model (Figure 2B).

deepNEC CNN training

In deepNEC, weights were initialized randomly using a uniform distribution. Rectified linear unit (ReLU) was used as an activation function for the convolution layer and hidden layers of the fully connected layers (i.e. a final stage of each CNN within deepNEC).

$$f(x) = \max(0, x)$$

The softmax function was used as an activation function for an output layer of the fully connected layer.

$$\sigma(Z)_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_i}}$$

To minimize the internal-covariate-shift by normalizing the input distribution of each layer to regular

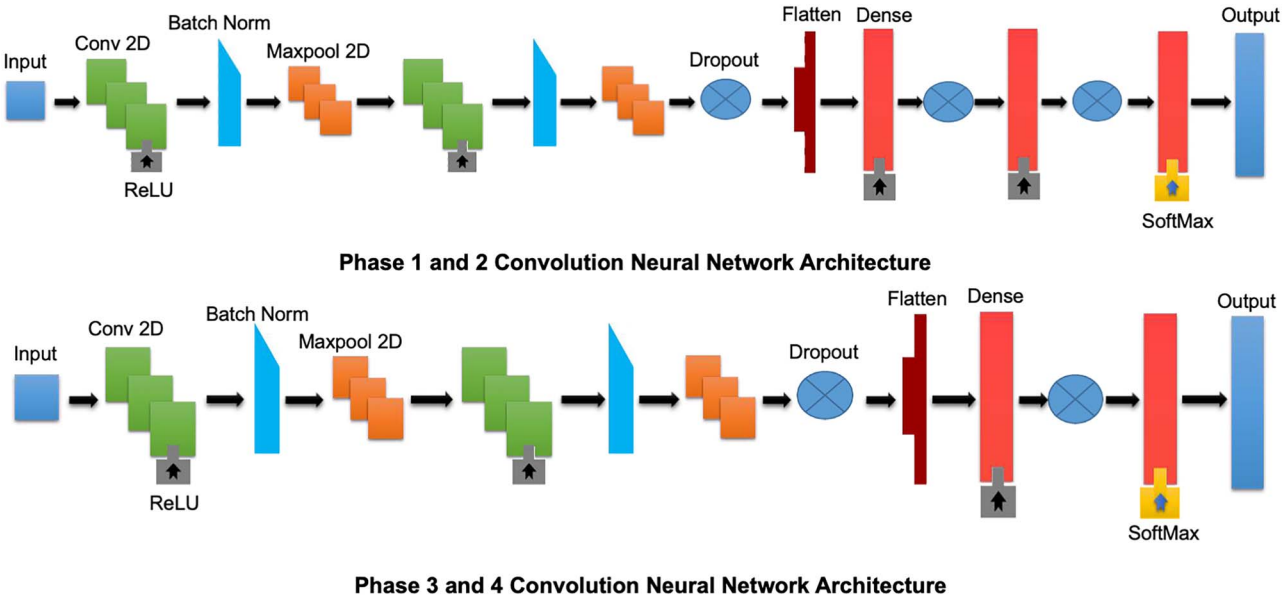

Figure 2. CNN architecture of deepNEC. The different terminology used in the figure are explained here: Input: This represents the input sequence vector for training; Conv 2D is a 2 dimensional convolution kernel that, when convolved with the layer input, yields a tensor of outputs; Batch Norm: Batch normalization is a training strategy for very deep neural networks that standardizes the inputs to each mini-batch; ReLU: Rectified linear unit; Maxpool2D: Max pooling operation for 2D spatial data; Dropout: Dropout refers to the practice of randomly 'dropping out,' or omitting, units during the neural network training process; Flatten: a flattened layer to process data into one-dimensional array. Dense: a deep neural network layer which each neuron receives input from all the neurons; Softmax: Softmax is a mathematical function that transforms a number vector into a probability vector; Output: output prediction probabilities.

Table 3. Enzyme versus non-enzyme (phase I) training and independent testing metrics results.

Metrics	Training average 10-fold	Independent testing
Sensitivity (%)	95.76	94.47
Specificity (%)	95.64	92.40
Precision (%)	95.64	92.55
Accuracy (%)	95.70	93.43
*F*1-score (%)	95.70	93.50
MCC	0.914	0.868

Table 4. Nitrogen metabolism versus non-nitrogen metabolism (phase II) training and independent testing metrics results.

Metrics	Training average 10-fold	Independent testing
Sensitivity (%)	99.51	96.60
Specificity (%)	99.88	100
Precision (%)	99.87	100
Accuracy (%)	99.71	98.30
*F*1-score (%)	99.69	98.27
MCC	0.9941	0.9665

Gaussian distribution, batch normalization was used. Batch normalization also acts as a regularizer, which helps prevent the overfitting of a deep learning algorithm. Each neuron learns the feature representation of the input signal across the period depending on their learning abilities, resulting in varying learning rates for each neuron of the network to maximize the objective function. Therefore, the Stochastic gradient descent (SGD) optimizer was used as the optimizer feature in this work. The categorical entropy of loss function was used to train the neural network. CNN in Phases I and II was trained up to 100 epochs.

Evaluation criteria

For binary classification problem that classifies an enzyme against a non-enzyme sequence, we defined enzymes as the positive data and non-enzymes as negative data. Similar approaches were applied for the nitrogen-related enzymes also. We used 10-fold cross-validation techniques to develop our model and evaluated the training process, and the independent dataset was used to assess the ability of our model. Different statistical approaches were used to evaluate the performance of each classifier in both phases. In particular, each query point from the test set has its correct class label in a traditional supervised binary classification problem. During the evaluation process, however, the classifier maps the question points to one of the following categories: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). To accomplish these categories for each class, the multiclass classification uses a one-on-one approach to rest. In this method, the query point belongs to a class, considered as a positive or a negative point. On this basis, TP, TN, FP and FN are determined for each class, and the following statistical approaches are used to test the output of the classifier for each class as was done in many other studies [38, 46–48].

Recall/sensitivity can be defined as the ability of a classifier to predict all relevant data correctly. It can

Table 5. Nitrogen metabolism nine classes (phase III) training metrics results

Metrics	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)	MCC
Nitrogen fixation	99.92	99.98	99.86	99.98	99.89	0.9988
Assimilatory	96.14	99.08	95.39	98.59	95.76	0.9492
Dissimilatory	96.21	99.46	95.35	99.13	95.78	0.9529
Denitrification	99.8	99.93	98.93	99.92	99.36	0.9932
Nitrification	85.96	99.75	98.7	97.3	91.89	0.9059
Assimilatory + dissimilatory + denitrification + nitrification	97.94	99.69	96.64	99.54	97.28	0.9704
Denitrification + nitrification	99.9	99.89	99.46	99.89	99.68	0.9962
Dissimilatory + denitrification + nitrification	99.47	99.66	89.76	99.65	94.36	0.9432
Dissimilatory + denitrification	98.94	98.21	85.67	98.28	91.83	0.9117
Overall	96.14	99.59	95.96	99.03	95.91	0.9544

Table 6. Nitrogen metabolism nine classes (phase III) Independent testing metrics results

Metrics	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)	MCC
Nitrogen fixation	99.5	99.94	99.5	99.89	99.5	0.9944
Assimilatory nitrate reduction	95.5	98.56	89.25	98.22	92.27	0.9133
Dissimilatory nitrate reduction	92.5	99.44	95.36	98.67	93.91	0.9317
Denitrification	99	99.94	99.5	99.83	99.25	0.9915
Nitrification	80	99.31	93.57	97.17	86.25	0.8501
Assimilatory + dissimilatory + denitrification + nitrification	98	99.75	98	99.56	98	0.9775
Denitrification + nitrification	100	99.88	99.01	99.89	99.5	0.9944
Dissimilatory + denitrification + nitrification	98.5	99.69	97.52	99.56	98.01	0.9776
Dissimilatory + denitrification	96	98.38	88.07	98.11	91.87	0.9091
Overall	95.44	99.43	95.53	98.98	95.39	0.9488

be calculated as the ratio between predicted TP to all positive observations (TP + FN).

$$\text{Sensitivity} = \frac{(\text{TP})}{(\text{TP} + \text{FN})}$$

Specificity can be defined as the percentage of negatively labeled instances that were predicted as negative; this can be calculated as the ratio between predicted TNs to all negative observations TN/(TN + FP).

$$\text{Specificity} = \frac{(\text{TN})}{(\text{TN} + \text{FP})}$$

Precision can be defined as the ability of a classifier to predict only relevant data correctly and is calculated as the ratio between predicted TP to all predicted positive observations (TP + FP).

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

Accuracy is the measure of correct prediction out of the total forecast.

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

*F*1 score: The *F*1 score uses only three categories (TP, FP andFN) to evaluate the performance of the classifier. It is the weighted average of precision and recall and takes values between 0 and 1, where zero value represents the worst classifier, and the value one represents the best classifier.

$$F1 - \text{score} = 2 * \frac{(\text{Precision} * \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})}$$

Matthews correlation coefficient (MCC) can be defined as the correlation between the observed and predicted values. The reason behind calculating MCC is that the accuracy and specificity sometimes overestimate the performance of the classifier. The MCC value of +1 represents the best prediction, 0 represents random prediction, and −1 represents the disagreement between the correct class and the predicted class.

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

Web server development

The deepNEC web server was developed using PHP 7.4, JavaScript, Jquery. In all four phases, the predicted

Table 7. Nitrogen metabolism-related 20 EC number (phase IV) training results statistics

Metrics	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)	MCC
Nitrogen fixation						
1.18.6.1	99.92	99.98	99.86	99.98	99.89	0.9988
1.7.7.2	99.12	99.86	99.55	99.69	99.34	0.9914
Assimilatory nitrate reduction						
1.7.1.1	61.71	98.96	85.38	95.64	71.49	0.7036
1.7.1.2	86.73	91.11	76.61	90.01	81.33	0.7484
1.7.1.3	69.10	97.46	65.71	95.61	67.13	0.6495
1.7.1.4	92.93	98.41	96.35	96.71	94.61	0.9227
1.7.7.1	97.77	99.88	97.85	99.77	97.80	0.9769
Dissimilatory nitrate reduction						
1.7.1.15	99.95	99.96	99.76	99.96	99.85	0.9983
1.7.2.2	99.96	99.95	99.99	99.96	99.97	0.9983
Denitrification						
1.7.2.5	99.87	99.65	99.66	99.76	99.76	0.9952
1.7.2.4	99.65	99.87	99.86	99.76	99.75	0.9952
Nitrification						
1.7.2.6	95.88	96.64	95.60	99.35	95.72	0.9538
1.7.2.7	98.44	99.51	95.88	99.40	97.13	0.9682
1.7.2.8	94.47	99.66	96.06	99.24	95.25	0.9485
1.14.99.39	100	99.98	99.95	99.98	99.97	0.9997
1.7.99.4	99.10	99.56	99.47	99.35	99.29	0.9870
Assimilatory + dissimilatory + denitrification + nitrification						
1.7.99.-	97.94	99.69	96.64	99.54	97.28	0.9704
Denitrification + nitrification						
1.7.5.1	99.9	99.89	99.46	99.89	99.68	0.9962
Dissimilatory + denitrification + nitrification						
1.7.2.1	99.47	99.66	89.76	99.65	94.36	0.9432
Dissimilatory + denitrification						
1.9.6.1	98.94	98.21	85.67	98.28	91.83	0.9117
Overall	97.10	99.20	93.92	99.05	95.35	0.9488

enzymes are further linked to different biological databases like NCBI [49], UniProt [13], BRENDA [50], KEGG [51, 52] and JGI IMG/M [53] to provide more comprehensive information about their annotations. It also provides a secondary structure prediction option using PsiPred [54, 55]. Along with the deep learning-based prediction, we have also included a similarity-based method using BLAST and diamondBLAST in the web server. Users can choose from three different prediction type options (DNN, Homology and Combined). It allows users to compare the machine learning results with similarity-based methods. Users can give an *e*-value, coverage and percent identity parameter for filtering the similarity-based results.

Development environment

CNN modeling was implemented using the Keras (version 2.4.0) Python package (https://keras.io/) with TensorFlow backend (version 2.2.0) [56]. Python module scikit-learn [57] was used to create a confusion matrix.

Results and discussion

Development of deepNEC

In deepNEC development, CNN was used out of various deep learning approaches due to its demonstrated success in the identification of functional areas (e.g. motifs and domains, enzyme/non-enzyme classification) in a biological sequence (e.g. protein sequence).

deepNEC result analysis

In addition to the evaluation criteria stated above, receiver operating characteristics (ROC) and Precision–Recall graphical parameters have been used to demonstrate promising efficiency for our deepNEC models for functional annotation of enzymes. The deepNEC phase-I evaluation metrics for enzyme versus non-enzyme classes are depicted in Table 3. The phase-I model's classification performance was demonstrated by the average of 10-fold training/testing procedure: precision (95.64%), recall (95.76%), accuracy (95.70%) and MCC (0.914). On independent benchmark (*testing of the models on completely unknown data not used in the training/testing procedure*), we got a precision (92.55%), recall (94.46%), accuracy (93.43%) and MCC (0.868), on testing of 3000 protein sequences. The deepNEC phase-II evaluation metrics pertaining to nitrogen biochemical network-related versus non-nitrogen biochemical network-related classes are reported in Table 4. The phase-II model's classification performance was demonstrated by the average 10-fold training precision (99.87%), recall (99.51%), accuracy (99.71%) and MCC (0.9941) parameters; and the precision (100%), recall (96.60%), accuracy (98.30%) and MCC (0.9665) performance on

Table 8. Nitrogen metabolism-related 20 EC number (phase IV) independent testing results statistics

Metrics	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)	MCC
Nitrogen fixation						
1.18.6.1	99.5	99.94	99.5	99.89	99.5	0.9944
Assimilatory nitrate reduction						
1.7.7.2	100	99.80	99.01	99.83	99.50	0.9940
1.7.1.1	68.00	98.60	90.67	93.50	77.71	0.7505
1.7.1.2	86.50	88.60	60.28	88.25	71.05	0.6561
1.7.1.3	76.00	97.70	86.86	94.08	81.07	0.7782
1.7.1.4	88.00	98.20	90.72	96.50	89.34	0.8726
1.7.7.1	95.50	99.90	99.48	99.17	97.45	0.9698
Dissimilatory nitrate reduction						
1.7.1.15	100	100	100	100	100	1
1.7.2.2	100	100	100	100	100	1
Denitrification						
1.7.2.5	99.50	99.50	99.50	99.50	99.50	0.99
1.7.2.4	99.50	99.50	99.50	99.50	99.50	0.99
Nitrification						
1.7.2.6	97.00	99.25	97.00	98.80	97.00	0.9625
1.7.2.7	99.00	99.25	97.05	99.20	98.01	0.9752
1.7.2.8	96.00	99.37	97.46	98.70	96.72	0.9591
1.14.99.39	100	100	100	100	100	1
1.7.99.4	99.00	99.87	99.49	99.70	99.24	0.9906
Assimilatory + dissimilatory + denitrification + nitrification						
1.7.99.-	98	99.75	98	99.56	98	0.9775
Denitrification + nitrification						
1.7.5.1	100	99.88	99.01	99.89	99.5	0.9944
Dissimilatory + denitrification + nitrification						
1.7.2.1	98.5	99.69	97.52	99.56	98.01	0.9776
Dissimilatory + denitrification						
1.9.6.1	96	98.38	88.07	98.11	91.87	0.9091
Overall	94.80	98.85	94.95	98.18	94.64	0.9370

independent testing of 2000 protein sequences. Similarly, deepNEC phase-III evaluation criteria for nine nitrogen biochemical network classes are described in Tables 5 and 6. The phase-III model's classification performance was demonstrated by the overall average 10-fold training precision (95.96%), recall (96.14%), accuracy (99.03%) and MCC (0.9544) parameters of nine classes; and the precision (95.53%), recall (95.44%), accuracy (95.39%) and MCC (0.9488) performance on independent testing of 1800 protein sequences. The deepNEC phase-IV evaluation criteria for 20 nitrogen biochemical network-related enzyme classes are described in Tables 7 and 8. The phase-IV model's classification performance was demonstrated by an overall average of 10-fold training procedure which gave a precision (93.92%), recall (97.10%), accuracy (99.05%) and MCC (0.9488) parameters of 20 classes; and the precision (94.95%), recall (94.80%), accuracy (98.18%) and MCC (0.9370) performance on independent testing of 4000 protein sequences. According to [58, 59], however, accuracy alone is not a reasonable indicator of a model classification performance owing to the accuracy paradox; thus, the *F*1-score was also calculated for the corresponding deepNEC phases. Out of both the positive and negative examples, the *F*1-score is the harmonic description of precision and recall, showing the classifier's ability to detect the real true sample, and is thus viewed as a more appropriate output measure relative to accuracy alone. While the *F*1-score is a more accurate parameter relative to accuracy to measure the model performance, the *F*1-score often overestimates the classification performance in the case of unbalanced test data [60]. The *F*1-scores for deepNEC phase I (Table 3), for deepNEC phase II (Table 4), for the overall performance of deepNEC phase III (Tables 5 and 6) and for the overall performance of deepNEC phase IV (Tables 7 and 8) indicate a high performance of the models in all classes [60]. A balanced success metric MCC score (>0.90) was determined for all the classes in each phase, guaranteeing the higher predictive efficiency of our deepNEC models.

ROC curve analysis

For assessing the efficiency of the classification model, ROC is a graphical metric. This illustrates the relationship between true positive rate or sensitivity (TPR) and false positive rate or 1-specificity (FPR) at different thresholds, where FPR is usually plotted against the *x*-axis and TPR against the *y*-axis [61–63]. Out of the total negative cases, the FPR of the classification model is the determination of false-positive prediction. In all positive cases, at the same time, the TPR defines the TP forecast. ROC curves are depicted in Figures 3 and 4. The top-left corner, where sensitivity and accuracy are 100%, reveals the optimal classification condition of the ROC curve.

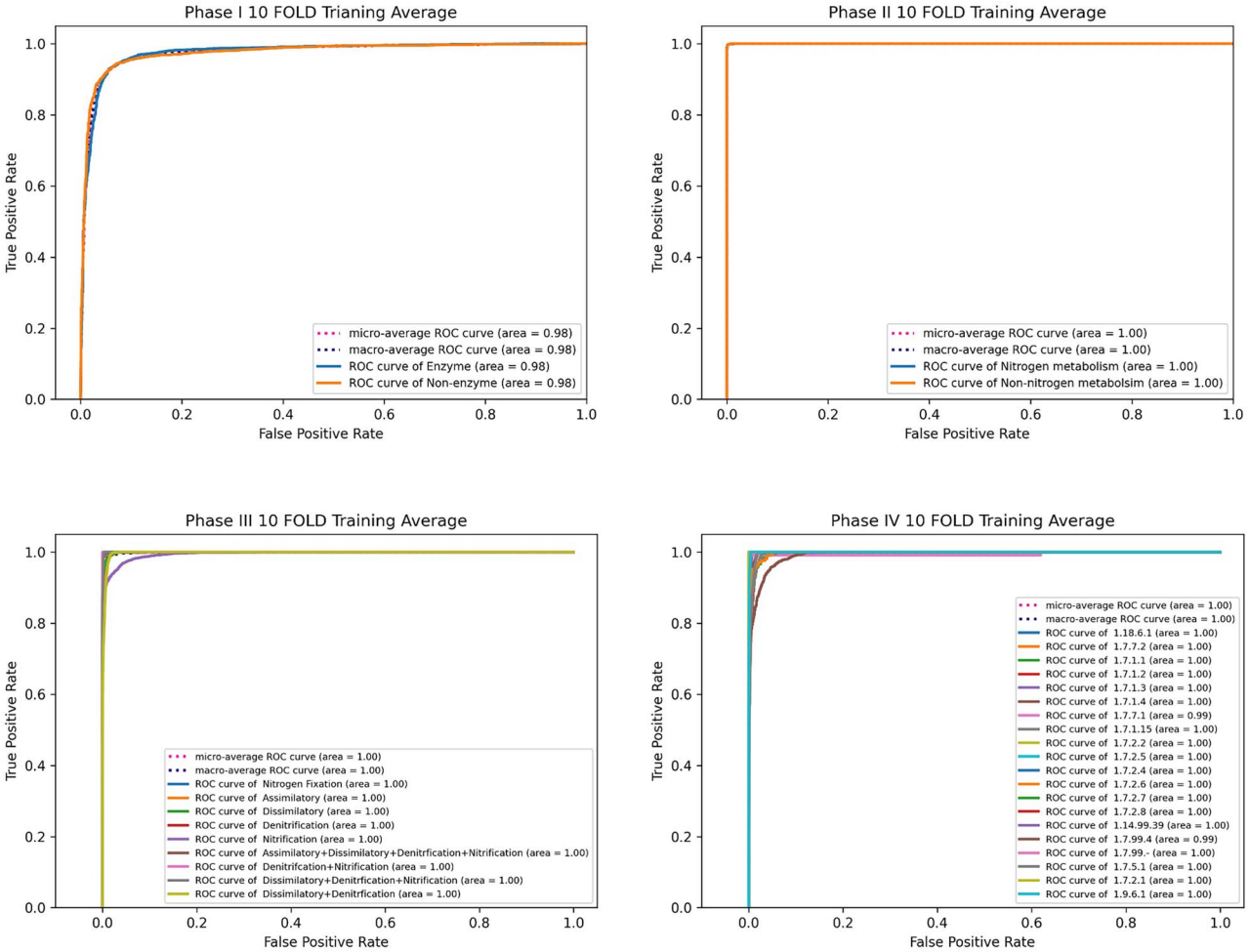

Figure 3. 10-FOLD average training ROC curves for deepNEC, all four phases.

The diagonal line [coordinate (0, 0) to (0, 1)] indicates the random output of the classification. Therefore, the ROC of the model must be over the diagonal line to get a good classification pattern, i.e. the more the area below the curve, the higher the performance of the classifier. The model's micro average ROC has also been plotted along with the fourteen classes to demonstrate the deepNEC model's average ROC efficiency. The deepNEC model's micro average ROC metric was plotted against micro average FPR (FPRμ) and micro average TPR (TPRμ), where FPRμ and TPRμ reflected the contribution of all fourteen classes and defined as

$$\mathrm{TPR}_{\mu} = \frac{\sum_{i=1}^{k} \mathrm{TP}_i}{\sum_{i=1}^{k} \mathrm{TP}_i + \mathrm{FN}_i}$$

$$\mathrm{FPR}\mu = \frac{\sum_{i=1}^{k} \mathrm{FN}_i}{\sum_{i=1}^{k} TP_i + \mathrm{FN}_i}$$

Precision–recall analysis

The precision–recall curve, plotted at different thresholds against recall and precision on the *x*-axis and *y*-axis respectively, is another graphical metric to determine the classifier's efficiency. The precision–recall curve helps to evaluate the classifier's optimistic predictive efficiency as it does not include real adverse cases in its estimation (Figures 5 and 6).

The optimal classification condition for the precision–recall curve is the upper right corner with a region under the curve of 1 square unit, where the precision and recall of the classifier is 100%, guaranteeing the classifier's ability to predict TP without any FP predictions. There is a fair trade-off between accuracy and recall, i.e. the region below the precision–recall curve must close to 1 square unit to provide a good classification model.

Benchmarking of deepNEC with other tools

Since deepNEC is the first reported tool specifically for nitrogen-related enzymes, we compared the performance of deepNEC with the existing tools for enzyme classification like ECPred [64], DeepEC [65], DeEPn [66] and DEEPre [44]. Due to the unavailability of the DeEPn and DEEPre servers, we were not able to include them for further analysis. ECPred uses multiple binary classifiers, while DeepEC is a combination of deep learning and homology methods. We downloaded 50 sequences from UniProt which were not part of our original training/testing. The comparison statistics are

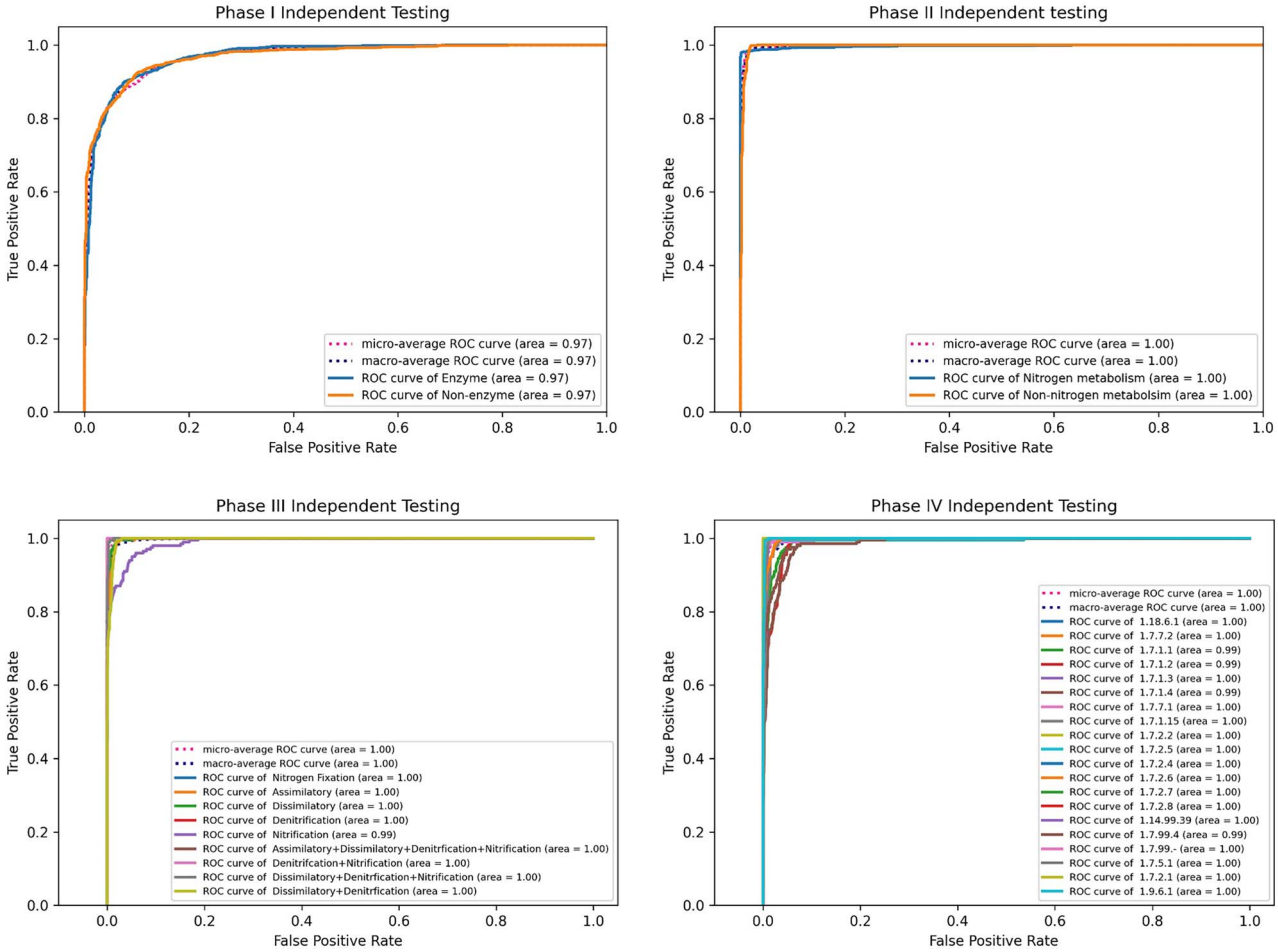

Figure 4. Independent testing ROC curves for deepNEC, all four phases.

represented in Figure 7 and Table 9. The accuracy of deepNEC was 92%, while the accuracy of ECPred and DeepEC were 76 and 74%, respectively. The *F*1-score of deepNEC was much higher (92.31) than ECPred (80) and DeepEC (74.51), validating the fact that our classifier had a better chance of separating the true samples from the negative samples. As expected, the MCC of deepNEC was much higher (+0.843) than ECPred (+0.567) and DeepEC (+0.480), showing the ability of the deepNEC classifier to make a more precise prediction.

Use of deepNEC for annotation of bacterial genomes

Understanding the role played by enzymes and biota in the functioning of agro-ecosystems is important for improving capabilities for soil health management. Several microbiome databases are generated every day with the advent of NGS technology. Quantitative data for nitrogen metabolism and enzyme activity will help us better understand the role of enzymes in soils. The development of deepNEC is one step in this direction to assist experimental biologists in the correct identification/classification of different types of enzymes from genomes and environmental metagenomes. For the performance evaluation of deepNEC on genome sequence data, we obtained 20 bacterial complete genomes from the NCBI (Table 10). Then we used deepNEC to predict all of the proteomes. Results of all the four phases for 20 genomes, with full results are available in Supplementary File 2. The highest percentage number of sequences annotated as enzymes was found in *Frankia alni* (66.84%) followed by *Rhodobacter capsulatus* (63.36%), *Cupriavidus taiwanesis* (61.08%) and *Candidatus Nitrosocosmicus franklandus* (35.78%) has the least percentage of sequences annotated as enzymes. From these enzymes, the highest percentage of sequences in *Nitrospina gracilis* (7.56%) were annotated as nitrogen biochemical network-related followed by *Nitrospira moscoviensis* (7.08%), *F. alni* (6.66%), *R. capsulatus* (6.36%), while the least percentage of nitrogen biochemical network-related enzymes were annotated in *Candidatus Nitrosocosmicus franklandus* (2.50%). In phase III, these nitrogen biochemical network-related enzymes were further assigned to nine different classes. For nitrogen fixation, the highest percentage of sequences were annotated in *Croscosphaera watsonii* (26.53%), followed by *Anabaena cylindrica* (23.80%) whereas the least number was in *Nitrosomonas oligotropha* (1.96%). Similarly, for the assimilatory class, the most percentage of sequences were present in *Nitrobacter winogradskyi* (41.30%) followed

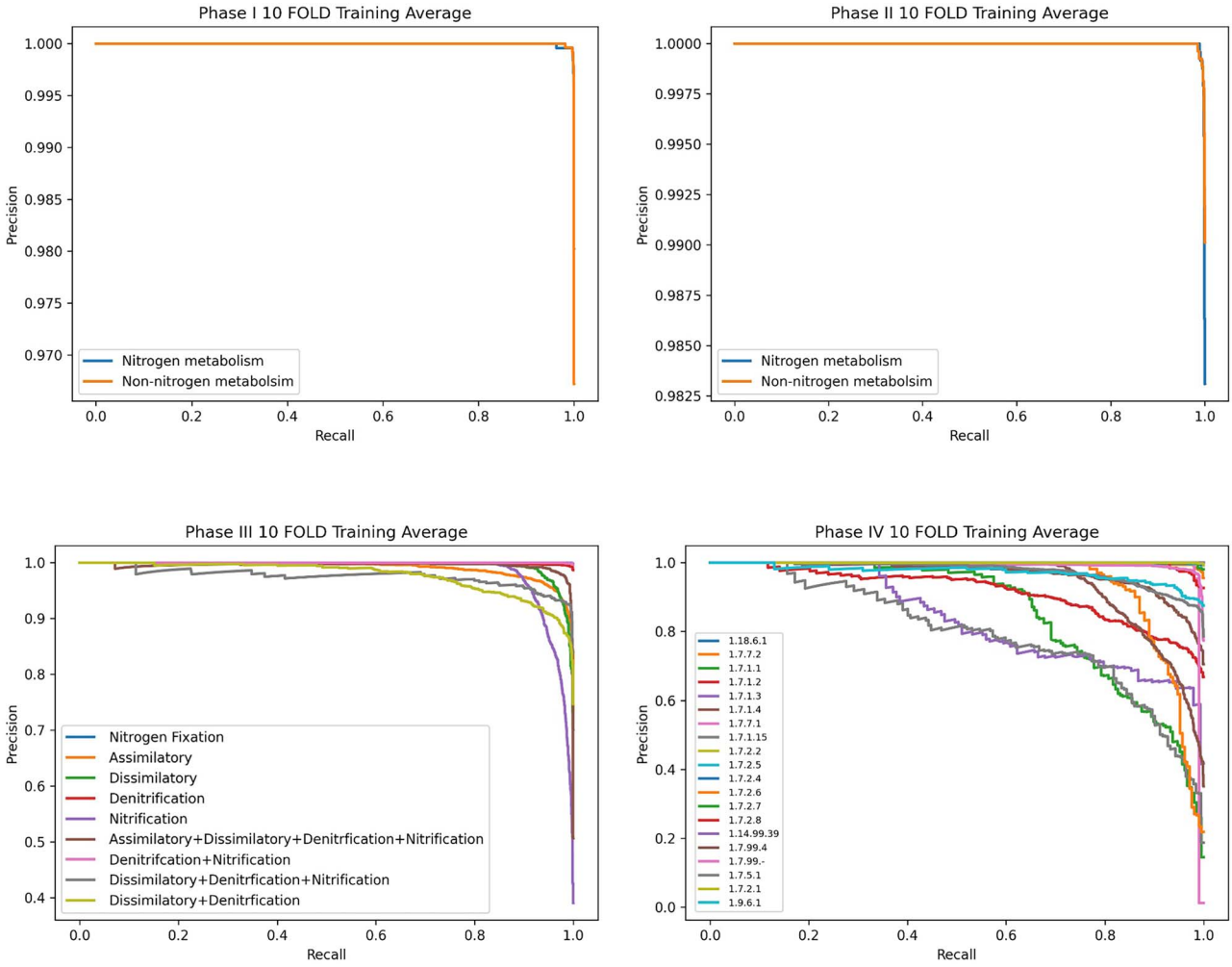

Figure 5. 10-FOLD average training PRC curves for deepNEC, all four phases.

by *F. alni* (33.98%) and the least percentage was in *Nitrosomonas eutropha* (13.04%). In the dissimilatory class, *Nitrososphaera viennensis* (10.20%) has the highest enzymes predicted and *C. watsonii, N. winogradskyi* and *A. cylindrica* have no enzymes predicted for dissimilatory nitrate reduction. *N. europaea* (27.02%) has the highest number of predicted enzymes for denitrification class, while *Candidatus Jettenia caeni* (5.43%) has the lowest number. For the nitrification class, highest number of enzymes were predicted in *Nitrosomonas halophila* (44.70%) followed by *Candidatus Brocadia sinica* (42.57%), while *Candidatus Nitrosocosmicus franklandus* (16.66%) has the least percentage of enzymes for nitrification.

These prediction results are in line with the reported annotations and functions of these bacterial genomes [67–85] and indicate that deepNEC can be used as a tool for annotations of new genomes or metagenomes. A negative control dataset (e.g. altered enzyme protein sequences with their functionalities gone) would also be highly beneficial in enhancing deepNEC's prediction capabilities for mutations in domains and binding site residues of protein sequences. Additional training of deepNEC with the negative control dataset as well as new nitrogen biochemical network-related enzymes would allow for more accurate detection of changes in enzymatic function caused by mutation or rearrangements. This feature is particularly valuable for examining homologous enzymes, such as those derived from new (meta)genome data that contain alterations with previously unknown impacts on enzyme functioning. Above all, it will be critical to utilize deepNEC in a variety of situations, either independently or as part of a third-party software package, and to gather input from biochemists, enzymologists and biotechnologists to more rigorously validate deepNEC predictions and to guide future improvements.

Comparison with similarity-based method

To compare the prediction performance of our alignment free method with the alignment-based methods such as BLAST, we downloaded one proteome, e.g. *Nitrosospira multiformis* proteome. We applied our deep learning based approach as well homology based approach to this dataset. deepNEC outperformed similarity-based methods like BLAST/diamond BLAST in the prediction of novel nitrogen metabolism-related enzymes. Out of the total 2739 proteins in *N. multiformis*; 1488 as enzymes and 1251 non-enzymes were predicted with deepNEC, whereas

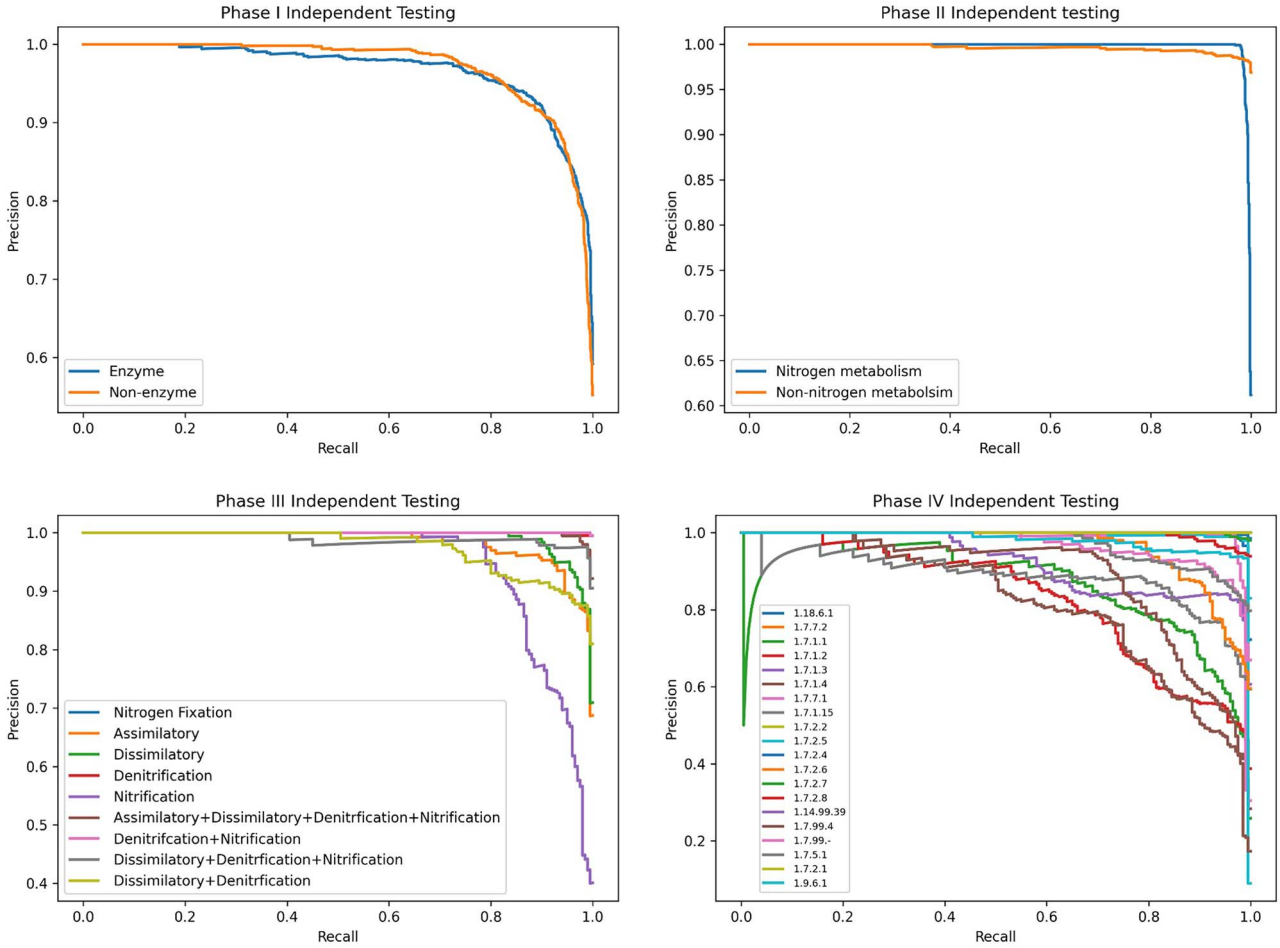

Figure 6. Independent testing PRC curves for deepNEC, all four phases.

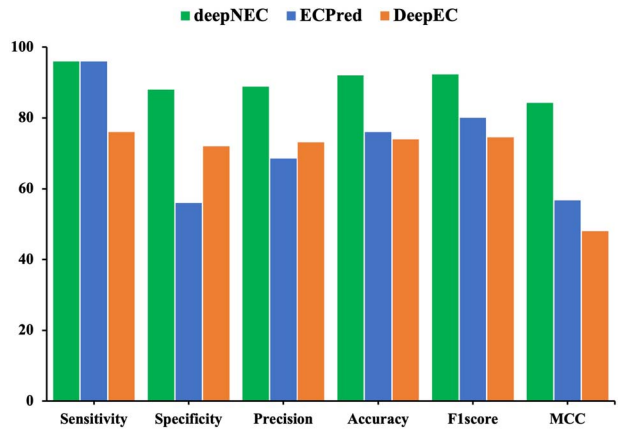

Figure 7. deepNEC comparison with existing tools.

820 enzymes and 1919 non-enzymes were predicted using diamondBLAST in phase I. In phase II, 83 nitrogen metabolism-related and 1405 non-nitrogen metabolism-related enzymes were predicted using deepNEC out of the total 1488 enzymes from phase I, and 17 nitrogen metabolism-related and 803 non-nitrogen metabolism-related were predicted using diamondBLAST out of the 820 enzymes predicted in phase I. In phase III, deepNEC classified 83 nitrogen metabolism-related enzymes into 6 classes while diamondBLAST classified 16 of 17 nitrogen

Table 9. Comparison of deepNEC with other tools for Enzyme versus non-enzyme classification

Metrics	deepNEC	ECPred	DeepEC
Sensitivity (%)	96	96	76
Specificity (%)	88	56	72
Precision (%)	88.89	68.57	73.08
Accuracy (%)	92	76	74
F1-score (%)	92.31	80	74.51
MCC	0.843	0.567	0.480

metabolism-related enzymes in 6 different classes. In phase IV, we have selected nitrification class for the EC number prediction. In this phase IV EC prediction, 31 nitrification-related enzymes were classified in 5 EC numbers while with diamondBLAST only 3 EC numbers were predicted for 4 nitrification-related enzymes. Thus, based on this comparison, we can say that deepNEC predicted more enzymes than similarity-based methods using the same dataset.

Conclusion

DeepNEC is a machine learning prediction model for predicting nitrogen biochemical network-related enzymes. It is implemented in four phases, where phase I classifies a protein sequence in enzyme/non-enzyme;

Table 10. Annotation of 20 microbial genomes with deepNEC

Genus and species	Strain	Lineage/cluster	Accession	References
Ammonia oxidizing bacteria (*Beta-Proteobacteria*)				
Nitrosospira briensis	C-128	cluster 3	GCF_000619905.2	[67]
N. multiformis	ATCC 25196	cluster 3	GCA_000196355.1	[68]
N. oligotropha	Nm45	cluster 6A	GCF_009833085.1	[69]
N. europaea	ATCC 19718	cluster 7	GCF_000009145.1	[70]
N. eutropha	C-91	cluster 7	GCF_000014765.1	[71]
N. halophila	Nm1		GCF_900107165.1	[72]
Ammonia oxidizing bacteria (*Gamma-Proteobacteria*)				
Candidatus Nitrosolglobus terrae	TAO100		GCF_002356115.1	[86]
Ammonia oxidizing archaea (*Thaumarchaeota*)				
N. viennensis	EN-76^T	*Nitrososphaera*	GCF_000698785.1	[74]
Ca. Nitrosocosmicus franklandus	C13 NFRAN1	*Nitrososphaera* sister	GCF_900696045.1	[75]
Nitrite oxidizing bacteria (*Proteobacteria*, *Nitrospirae*)				
N. winogradskyi	NB-255	Alpha-proteobacteria	GCA_000012725.1	[76]
N. moscoviensis	NSP-S1	lineage II *Nitrospira*	GCF_001273775.1	[87]
N. gracilis	3/211		GCF_000341545.2	[77]
Comammox bacteria (*Nitrospirae*)				
Nitrospira inopinata	ENR4	lineage II (clade A.1)	GCF_001458695.1	[78]
Anammox (*Plantomycetes*)				
Ca. Jettenia caeni	KSU-1		GCA_000296795.1	[79]
Ca. Brocadia sinica	JPN1		GCF_000949635.1	[80]
Nitrogen-fixing bacteria				
C. taiwanesis	LMG 19424		GCF_000069785.1	[81]
A cylindrica	PCC 7122		GCF_000317695.1	[82]
F. alni	ACN14A		GCF_000058485.1	[83]
C. watsonii	WH 8501		GCF_000167195.1	[84]
R. capsulatus	A12		GCF_014622665.1	[85]

phase II further classifies the predicted enzyme into nitrogen metabolism enzyme/non-nitrogen metabolism enzyme; phase III classifies the predicted nitrogen biochemical network-related enzyme in nine classes; phase IV then provides an EC number for the predicted nitrogen metabolism class. For the prediction of nitrogen biochemical network-related enzymes, deepNEC converts protein sequence into DPC and NMBroto features. The training and validation dataset of the deepNEC model was retrieved from *UniProtKB/Swiss-Prot*. Independent dataset comparison of the deepNEC model was further performed against other enzyme classification tools (DeepEC, ECPred) to analyze the performance of our model. deepNEC was able to outperform these alternatives. Researchers will be able to use deepNEC to predict and analyze a variety of new nitrogen biochemical network-related enzymes from genomes and metagenomes.

Key Points

- Nitrogen biochemical network is an important pathway in the global biogeochemical cycle.
- deepNEC is a deep learning-based tool for the identification and classification of N-biochemical network-related enzymes.
- It is a four-phase prediction tool and further provides annotations of the predicted enzymes to external databases. It can also predict secondary and tertiary structures.
- Independent validation of deepNEC on independent test data shows an efficient and accurate prediction. Also performed annotations on 20 novel bacterial genomes.
- The resource will be useful for the functional annotation of microbial genomes and metagenomes.

Acknowledgment

Special thanks are due to Shelby McCowan (Bioinformatics Linux Systems Administrator), who helped with the installation of the backend software that deepNEC uses through SLURM jobs and with the configuration of the head node of the High-Performance Computing (HPC) cluster to implement the web server successfully. The authors sincerely thank the anonymous referees for all the suggestions and help in improving the research article.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The prediction software developed from this study is available freely at http://bioinfo.usu.edu/deepNEC/. The standalone version of deepNEC can be downloaded from https://navduhan@bitbucket.org/navduhan/deepnec.git.

All the training/testing data and the independent testing data used in this study are available as well on the web server.

Funding

This work was funded by the Office of Research, USU (Research Catalyst grant # A45112 to R.K.). The funding body had no involvement in the design of this study, data collection, analysis, interpretation or article preparation.

Conflict of Interest

The author declare no conflict interest.

References

1. Fowler D, Coyle M, Skiba U, *et al.* The global nitrogen cycle in the Twentyfirst century. *Philos Trans R Soc B Biol Sci* 2013;**368**:20130164.
2. Galloway JN, Dentener FJ, Capone DG, *et al.* Nitrogen cycles: past, present, and future. *Biogeochemistry* 2004;**70**:153–226.
3. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive earth's biogeochemical cycles. *Science* 2008;**320**:1034–9.
4. Gruber N, Galloway JN. An earth-system perspective of the global nitrogen cycle. *Nature* 2008;**451**:293–6.
5. Reed DC, Algar CK, Huber JA, *et al.* Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc Natl Acad Sci* 2014;**111**:1879–84.
6. Landolfi A, Dietze H, Koeve W, *et al.* Overlooked runaway feedback in the marine nitrogen cycle: the vicious cycle. *Biogeosciences* 2013;**10**:1351–63.
7. Vitousek PM, Howarth RW. Nitrogen limitation on land and in the sea: how can it occur? *Biogeochem* 1991;**13**:87–115.
8. Ye RW, Thomas SM. Microbial nitrogen cycles: physiology, genomics and applications. *Curr Opin Microbiol* 2001;**4**:307–12.
9. Jetten MSM. The microbial nitrogen cycle. *Environ Microbiol* 2008;**10**:2903–9.
10. Kuypers MMM, Marchant HK, Kartal B. The microbial nitrogen-cycling network. *Nat Rev Microbiol* 2018;**16**:263–76.
11. Goddard JP, Reymond JL. Enzyme assays for high-throughput screening. *Curr Opin Biotechnol* 2004;**15**:314–22.
12. Bairoch A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acid Res* 2000;**28**:45–8.
13. Consortium TU, Bateman A, Martin M-J, *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acid Res* 2021;**49**:D480–9.
14. Cornish-Bowden A. Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspect Sci* 2014;**1**:74–87.
15. des Jardins M, Karp PD, Krummenacker M, *et al.* Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**:92–9.
16. Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. *J Mol Biol* 2005;**345**:187–99.
17. Nagao C, Nagano N, Mizuguchi K. Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *PLoS One* 2014;**9**:e84623.
18. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acid Res* 2012;**40**.
19. Yang J, Yan R, Roy A, *et al.* The I-TASSER suite: protein structure and function prediction. *Nat Method* 2014;**12**:7–8.
20. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acid Res* 2017;**45**:W291–9.
21. Arakaki AK, Huang Y, Skolnick J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinform* 2009;**10**:107.
22. Kumar N, Skolnick J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics* 2012;**28**:2687–8.
23. Quester S, Schomburg D. EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinform* 2011;**12**:376.
24. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acid Res* 2004;**32**:6226–39.
25. Yu C, Zavaljevski N, Desai V, *et al.* Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. *Proteins Struct Funct Bioinform* 2009;**74**:449–60.
26. Cai CZ, Han LY, Ji ZL, *et al.* SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acid Res* 2003;**31**:3692–7.
27. Cai CZ, Han LY, Ji ZL, *et al.* Enzyme family classification by support vector machines. *Protein Struct Funct Genet* 2004;**55**:66–76.
28. Cai YD, Chou KC. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J Proteome Res* 2005;**4**:967–71.
29. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**:10–9.
30. Chou KC, Elrod DW. Prediction of enzyme family classes. *J Proteome Res* 2003;**2**:183–90.
31. De Ferrari L, Aitken S, van Hemert J, *et al.* EnzML: multi-label prediction of enzyme classes using InterPro signatures. *BMC Bioinform* 2012;**13**:61.
32. Huang WL, Chen HM, Hwang SF, *et al.* Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. *Biosystems* 2007;**90**:405–13.
33. Kumar C, Choudhary A. A top-down approach to classify enzyme functional classes and sub-classes using random forest. *Eurasip J Bioinform Syst Biol* 2012;**2012**:1.
34. Li YH, Xu JY, Tao L, *et al.* SVM-prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One* 2016;**11**:e0155290.
35. Lu L, Qian Z, Cai YD, *et al.* ECS: an automatic enzyme classifier based on functional domain composition. *Comput Biol Chem* 2007;**31**:226–32.
36. Nasibov E, Kandemir-Cavas C. Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. *Comput Biol Chem* 2009;**33**:461–4.
37. Qiu JD, Luo SH, Huang JH, *et al.* Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform. *J Theor Biol* 2009;**256**:625–31.
38. Bin SH, Chou KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Commun* 2007;**364**:53–9.
39. Volpato V, Adelfio A, Pollastri G. Accurate prediction of protein enzymatic class by N-to-1 neural networks. *BMC Bioinform* 2013;**14**:S11.
40. Claesson MJ, Wang Q, O'Sullivan O, *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acid Res* 2010;**38**:e200–0.

41. Wang YC, Wang Y, Yang ZX, *et al.* Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. *BMC Syst Biol* 2011;**5**:S6.
42. Wang Y-C, Wang X-B, Yang Z-X, *et al.* Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept Lett* 2012;**17**:1441–9.
43. Bin ZX, Chen C, Li ZC, *et al.* Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J Theor Biol* 2007;**248**:546–51.
44. Li Y, Wang S, Umarov R, *et al.* DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* 2018;**34**:760–9.
45. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
46. Feng PM, Chen W, Lin H, *et al.* IHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem* 2013;**442**:118–25.
47. Liu B, Liu F, Wang X, *et al.* Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acid Res* 2015;**43**:W65–71.
48. Kaundal R, Saini R, Zhao PX. Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis. *Plant Physiol* 2010;**154**:36–54.
49. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.
50. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. *Nucleic Acid Res* 2002;**30**:47.
51. Ogata H, Goto S, Sato K, *et al.* KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acid Res* 1999;**27**:29–34.
52. Kanehisa M, Furumichi M, Tanabe M, *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acid Res* 2017;**45**:D353–61.
53. Chen I-MA, Chu K, Palaniappan K, *et al.* The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acid Res* 2021;**49**:D751–63.
54. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;**16**:404–5.
55. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acid Res* 2019;**47**:W402–7.
56. Abadi M, Barham P, Chen J, *et al.* Tensorflow: a system for large-scale machine learning. *12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI 16)* 2016;265–83
57. Pedregosa F, Michel V, Grisel Oliviergrisel O, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**. https://www.jmlr.org/papers/v12/pedregosa11a.html.
58. Abma BJM, Kurtev I. *Evaluation of requirements management tools with support for traceability-based change impact analysis*, Master's thesis in Software Engineering. University of Twente, 2009.
59. Valverde-Albacete FJ, Carrillo-de-Albornoz J, Peláez-Moreno C. A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. *Lect Note Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2013;**8138 LNCS**:41–52.
60. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLoS One* 2017;**12**:e0177678.
61. Semwal R, Aier I, Raj U, *et al.* Pharmadoop: a tool for pharmacophore searching using Hadoop framework. *Netw Model Anal Heal Inform Bioinform* 2017;**6**:1–9.
62. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;**39**:561–77.
63. Swets JA. Measuring the accuracy of diagnostic systems. *Sci Sci* 1988;**240**:1285–93.
64. Dalkiran A, Rifaioglu AS, Martin MJ, *et al.* ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinform* 2018;**19**:334.
65. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci* 2019;**116**:13996–4001.
66. Semwal R, Aier I, Tyagi P, *et al.* DeEPn: a deep neural network based tool for enzyme functional annotation. *J Biomol Struct Dyn* 2020;**39**:2733–2743.
67. Rice MC, Norton JM, Valois F, *et al.* Complete genome of *Nitrosospira briensis* C-128, an ammonia-oxidizing bacterium from agricultural soil. *Stand Genomic Sci* 2016;**11**:1–8.
68. Norton JM, Klotz MG, Stein LY, *et al.* Complete genome sequence of *Nitrosospira multiformis*, an ammonia-oxidizing bacterium from the soil environment. *Appl Environ Microbiol* 2008;**74**:3559.
69. Sedlacek CJ, McGowan B, Suwa Y, *et al.* A Physiological and genomic comparison of Nitrosomonas cluster 6a and 7 ammonia-oxidizing bacteria. *Microb Ecol* 2019; **78**:985–94
70. Chain P, Lamerdin J, Larimer F, *et al.* Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J Bacteriol* 2003;**185**:2759–73.
71. Stein LY, Arp DJ, Berube PM, *et al.* Whole-genome analysis of the ammonia-oxidizing bacterium, *Nitrosomonas eutropha* C91: implications for niche adaptation. *Environ Microbiol* 2007;**9**:2993–3007.
72. IMG-taxon 2675903041 annotated assembly - Genome - Assembly - NCBI. Nitrosomonas halophila proteome. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/107/165/GCA_900107165.1_IMG-taxon_2675903041_annotated_assembly/.
73. Hayatsu M, Tago K, Uchiyama I, *et al.* An acid-tolerant ammonia-oxidizing γ-proteobacterium from soil. *ISME J* 2017;**11**:1130–41.
74. Stieglmeier M, Klingl A, Alves RJE, *et al. Nitrososphaera viennensis* gen. Nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. *Int J Syst Evol Microbiol* 2014;**64**(Pt 8):2738–52.
75. Lehtovirta-Morley LE, Ross J, Hink L, *et al.* Isolation of '*Candidatus Nitrosocosmicus franklandus*', a novel ureolytic soil archaeal ammonia oxidiser with tolerance to high ammonia concentration. *FEMS Microbiol Ecol* 2016;**92**(5):fiw057.
76. Starkenburg SR, Chain PSG, Sayavedra-Soto LA, *et al.* Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Appl Environ Microbiol* 2006;**72**:2050–63.
77. Lücker S, Nowka B, Rattei T, *et al.* The genome of *Nitrospina gracilis* illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front Microbiol* 2013;**4**:27.
78. Daims H, Lebedeva EV, Pjevac P, *et al.* Complete nitrification by Nitrospira bacteria. *Nature* 2015;**528**:504–9.
79. Ali M, Oshiki M, Awata T, *et al.* Physiological characterization of anaerobic ammonium oxidizing bacterium '*Candidatus Jettenia caeni*'. *Environ Microbiol* 2015;**17**:2172–89.
80. Oshiki M, Ali M, Shinyako-Hata K, *et al.* Hydroxylamine-dependent anaerobic ammonium oxidation (anammox) by "*Candidatus Brocadia sinica*". *Environ Microbiol* 2016;**18**:3133–43.
81. Amadou C, Pascal G, Mangenot S, *et al.* Genome sequence of the beta-rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia. *Genome Res* 2008;**18**:1472–83.

82. ASM31769v1 - Genome - Assembly - NCBI. Anabaena cylindrica refernce genome NCBI. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/317/695/GCA_000317695.1_ASM31769v1/.
83. Normand P, Lapierre P, Tisa LS, *et al.* Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res* 2007;**17**:7–15.
84. ASM16719v1 - Genome - Assembly - NCBI. Crocosphaera watsonii refernece genom NCBI. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/167/195/GCA_000167195.1_ASM16719v1/.
85. ASM1462266v1 - Genome - Assembly - NCBI. Rhodobacter capsulatus NCBI Genome. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/014/622/665/GCA_014622665.1_ASM1462266v1/.
86. Hayatsu M, Tago K, Uchiyama I, *et al.* An acid-tolerant ammonia-oxidizing & gamma-proteobacterium from soil. *ISME J* 2017;**11**:1130–41.
87. Koch H, Lücker S, Albertsen M, *et al.* Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. *Proc Natl Acad Sci* 2015;**112**:11371–6.