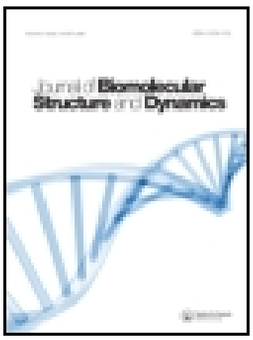


Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

DeEPn: A deep neural network based tool for enzyme functional annotation

Rahul Semwal, Imlimaong Aier, Pankaj Tyagi & Pritish Kumar Varadwaj

To cite this article: Rahul Semwal, Imlimaong Aier, Pankaj Tyagi & Pritish Kumar Varadwaj (2020): DeEPn: A deep neural network based tool for enzyme functional annotation, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1754292

To link to this article: https://doi.org/10.1080/07391102.2020.1754292

Accepted author version posted online: 10 Apr 2020.

Submit your article to this journal

Article views: 1

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tbsd20

Check for updates

DeEPn: A deep neural network based tool for enzyme functional annotation

Rahul Semwal[a], Imlimaong Aier[b], Pankaj Tyagi[a] and Pritish Kumar Varadwaj[b]

[a]Department of Information Technology (Bioinformatics), Indian Institute of Information Technology Allahabad, India

[b]Department of Bioinformatics and Applied Science, Indian Institute of Information Technology, Allahabad

*Corresponding author: pritish@iiita.ac.in

Accepted Manuscript

Abstract

With the advancement of high throughput techniques, the discovery rate of enzyme sequences has increased significantly in the recent past. All of these raw sequences are required to be precisely mapped to their respective functional attributes, which helps in deciphering their biological role. In the recent past, various prediction models have been proposed to predict the enzyme functional class; however, all of these models were able to quantify at most six functional enzyme classes (*EC1* to *EC6*) out of existing seven functional classes, making these approaches inappropriate for handling enzymes corresponding to the seventh functional class (*EC7*). In this study, a Deep Neural Network-based approach, *DeEPn*, has been proposed, which can quantify enzymes corresponding to all seven functional classes with high precision and accuracy. The proposed model was compared with two recently developed tools, *ECPred* and *SVM-Prot*. The result demonstrated that *DeEPn* outperformed *ECPred* and *SVM-Prot* in terms of predictive quality. The *DeEPn* tool has been hosted as a web-based tool at https://bioserver.iiita.ac.in/DeEPn/.

Keywords: Enzyme, Machine learning, Enzyme classification, Deep learning

Appendix

ACC: Accuracy

CNN: Convolutional neural network

CTD: Composition, Transition, and Distribution

EC: Enzyme Commission

FN: False negative

FP: False positive

FPR: False positive rate

IUBMB: International Union of Biochemistry and Molecular Biology

MCC: Matthew's correlation coefficient

PPV: Precision

ROC: Receiver operating characteristics

SEN: Sensitivity

TN: True negative

TP: True positive

TPR: True positive rate

Accepted Manuscript

1 Introduction

Enzymes are essential proteins that catalyze chemical reactions in *vivo* and play an important role in synchronizing biological processes. Enzymes perform variety of function, such as signal transduction, transport of nutrients, metabolism, and energy generation, which are essential for any living organism (Huang, Lin, Chang & Tang 2012). Most of the biological processes in the cell require enzymes for their survival (Wang, Wang, Yang & Deng 2011). Experimental evidence reports that about half of all proteins have enzymatic activities, making enzymes one of the largest and most important families of proteins (Rost 2002; Espadaler et al. 2008). Therefore, accurate enzyme function annotation is a crucial and essential step for analysis of metabolic fluxes and high-quality metabolic reconstruction. In fact, the malfunctioning or instability of few of these enzymes could lead to serious disorders, such as, the defalcation of myophosphorylase, which metabolize glucose in the cell, causing *McArdle's* disease. This leads to difficulty in swallowing, sucking, and respiratory failure (Fenichel 2007). Similarly, the deficiency of glucosylceramidase (*GCase*) enzyme activity causes hydrolysis of glucosylceramide into glucose and ceramide, leading to *Gaucher* disease which results in anaemia, enlargement of spleen and liver, and bone fractures (Pastores & Hughes 2018; Mahan, Bailey, Pilipenko & Prada 2019). To explore the root cause of such diseases, and to identify an essential method to cure them, the first and foremost step would be to perceive the function of related enzymes.

The traditional and most rigorous approach for performing enzyme function investigation is through enzyme assays (an experimental technique) which require extensive time and expertise for determining enzyme activity (Goddard & Reymond 2004; Li et al. 2018). However, with the emergence of new technologies and methods in proteomics (Jiang et al. 2017), the dimension of enzyme databases, as well as the discovery rate of new enzymes corresponding to existing or new families, have increased tremendously. This makes experimental methods inappropriate for handling

Accepted Manuscript

the rapidly growing rate of new enzymes (Godzik 2011). In such a case, a reliable automated process is required that can precisely predict the function of enzymes and allow exploration of significant properties of newly discovered enzymes according to their predicted family. However, automating functional characterization of enzymes with higher accuracy remains a challenging task in computational biology.

One great effort to annotate enzymes based on their functional characterization was performed by the International Union of Biochemistry and Molecular Biology (*IUBMB*) (Mahan et al. 2019). It is an international non-governmental organization that provides a nomenclature system known as Enzyme Commission number (*EC* number) to annotate enzymes. *EC* number classifies enzymes based on the chemical reactions they catalyze (Webb 1992; Moss 2010). It is a hierarchical ontological system composed of four-digit numbers separated by the period symbol (Cornish-Bowden 2014) . The first digit defines one of the main enzymatic functional class out of seven classes i) Oxidoreductases (EC1): These enzymes catalyze oxidoreduction reactions; for example, Aldehyde dehydrogenases (ALDH) removes toxic aldehydes through oxidoreduction from the body (Goedde et al. 1983) ii) Transferases (EC2): These enzymes catalyze the transfer of functional group from one molecule (donor) to other molecule (accepter); for example, kinase catalyzes the transfer of phosphate group from ATP and hence helps in regulating the biological process in living organism. (Manning, Whyte, Martinez, Hunter & Sudarsanam 2002) iii) Hydrolases (EC3): These enzymes catalyze the breakdown of larger molecules into smaller molecules with the help of water; for example, lipase breaks down dietary fats in living organism and helps in maintaining the digestive regulation (Rotticci, Rotticci-Mulder, Denman, Norin & Hult 2001) iv) Lyases (EC4): These enzymes reduce large complexes to small molecules by means of elimination reaction; for example, ornithine decarboxylase is a lyases family enzyme which produces the polymerase by means of elimination reaction, and thus provides stabilization to DNA and regulates cell growth. (Kern, Oliveira, Coffino & Hackert 1999) v) Isomerases (EC5): These are general classes of enzymes that synthesis isomers

Accepted Manuscript

of a molecule; for example, bisphosphoglycerate mutase is an enzyme essential for glycolysis and is integral part of red blood cell (Pritlove, Gu, Boyd, Randeva & Vatish 2006) vi) Ligases (EC6): These enzymes join two large molecules by synthesizing a chemical bond between them; for example, DNA ligase helps in DNA synthesis by joining the DNA strands (Pascal, O'Brien, Tomkinson & Ellenberger 2004) vii) Translocases (EC7): These are a general class of protein involved in transporting the molecule across the membrane; for example, transporter proteins are involved in moving ions across the membrane, and in maintaining the cellular environment (Medicine 2000). The next two digits define the more specific function of enzymes based on the chemical reaction they catalyze. In contrast, the last (fourth) digit defines the substrate specificity of the catalyzed reaction. Currently, the *EC* number system is the universally accepted commission system for annotating enzymes in biological databases.

In the past, various computational approaches were proposed based on *EC* numbers to annotate the function of enzymes. These approaches can be classified as follows:

1. **Homology based methods:**

   These methods (Tian, Arakaki & Skolnick 2004; Arakaki, Huang & Skolnick 2009; Yu, Zavaljevski, Desai & Reifman 2009; Quester & Schomburg 2011; Kumar & Skolnick 2012) are based on the common assumption that enzymes with high sequence similarity tend to have the same function. However, this direct inference only works when the enzymes under consideration are very similar to known enzymes; moreover, it was proved that under 60% sequence similarity, the accuracy of this method is significantly low (Tian & Skolnick 2003; Wang et al. 2011). Further, the technique is inappropriate when encountering a sequence with low homology in the current databases (Li et al. 2018).

2. **Structure-based methods:**

Accepted Manuscript

These methods (Dobson & Doig 2005; Roy, Yang & Zhang 2012; Nagao, Nagano & Mizuguchi 2014; Yang et al. 2015; Zhang, Freddolino & Zhang 2017) assume that enzymes with similar structures tend to have similar functions. To predict the function of enzymes under consideration, these methods first predict the structure of unknown enzymes, and then perform a structural comparison with known enzymes (validated and *EC* number assigned) available in databases or libraries. After structural comparison, the *EC* number or function of known enzymes which are most similar in structure, are assigned to unknown enzymes. However, the structure prediction methods are time-consuming and more prone to error (Correa & Dorn 2018). The propagation of error from sequence to structure, and structure to *EC* number prediction would produce a more accumulated error in the final result.

3. Sequence based methods:

These methods (Cai, Han, Ji, Chen & Chen 2003; Chou & Elrod 2003; Cai, Han, Ji & Chen 2004; Cai & Chou 2006; Huang, Chen, Hwang & Ho 2007; Lu, Qian, Cai & Li 2007; Shen & Chou 2007; Zhou, Chen, Li & Zou 2007; Lee, Lee & Ryu 2008; Nasibov & Kandemir-Cavas 2009; Qiu, Luo, Huang & Liang 2009; Wang, Wang, Yang & Deng 2010; Wang et al. 2011; De Ferrari, Aitken, van Hemert & Goryanin 2012; Kumar & Choudhary 2012; Volpato, Adelfio & Pollastri 2013; Sharif, Thrwat, Amin, Ella & Hefeny 2015; Li et al. 2016; Zou & Xiao 2016; Li et al. 2018) are the most popular, and they make use of either sequence or domain, or both as features to determine the function of enzymes. These features are either predefined or close to the suboptimal feature space, making feature extraction and classification task a separate phase of enzyme annotation. However, due to diversity associated with enzyme families, these limited set of features may not be able to capture all the important features crucial for enzyme functional annotation (Li et al. 2016).

Accepted Manuscript

In this study, to overcome the problems stated above, a deep convolution neural network approach, *DeEPn*, was proposed to predict the function of enzymes. *DeEPn* uses various sequence related enzyme features, and automatically weight the important ones for enzyme classification, which can be achieved through convolution neural network filters (Masci, Meier, Cireşan & Schmidhuber 2011; Zeng, Liu, Lai, Zhou & Zhao 2014). The approaches described above can only classify enzymes corresponding to six functional classes (*EC1* to *EC6*). Moreover, for the Translocase (*EC7*) functional class, which was introduced only recently by the *IUBMB* [45], the methods stated above do not provide any classification schemes; however, *DeEPn* was developed to accommodate all functional classes, including the updated one (*EC1* to *EC7*). *DeEPn* was compared with two recently developed tools, *ECPred*, and *SVM-Prot*, and the resulting comparison indicated that *DeEPn* outperformed the existing approaches. The *DeEPn* tool was implemented and hosted as a web prediction tool and is available on the server interface https://bioserver.iiita. ac.in/DeEPn/.

2 Materials and Method

2.1 Dataset collection and pre-processing

The prediction reliability and performance of any model depends on the dataset used for training to a large extent. In our study, to train the *DeEPn* model, the dataset (protein sequences) was taken from *Uniprot Knowledgebase* (release 2019_03), a freely available and high-quality web based resource of protein sequence and functional information (Apweiler et al. 2004). There are two types of the segment in *Uniprot Knowledgebase*: i) *UniProtKB/TrEMBL*, which contains unreviewed and automatically annotated protein entries ii) *UniProtKB/ Swiss-Prot*, which holds reviewed and manually annotated protein entries. To retrieve the reliable protein (enzyme and non-enzyme) entries the following filtering criteria were applied to sequences: not fragments (C-terminal or N-terminal should be absent), does not contain non-amino acid character, longer than 30 amino acids, manually annotated and reviewed. Furthermore, the sequences belong to more than one functional class were removed for further analysis, since these create dubious situations for the classifier. After the

Accepted Manuscript

filtering process, a total of 4,40,963 protein (enzymes and non-enzymes) sequences with the unique functional class were obtained. The number of protein sequences and corresponding functional categories was summarized in Table 1.

Algorithm 1: Update cluster algorithm

```
Input: RClust, EClust, NEClust
Result: Return updated RClust
    for each cluster clst_i in RClust do
        UpdateList_i ←∅
        for each sequence seq_i in clust_i do
            clust_j ←∅
            if EClust.is_representative(seq_i) then
                clust_j←EClust:get_cluster(seq_i )
            else if NEClust.is_representative(seq_i)
            then
                clustj←NEClust:get_cluster(seq_i )
            if clust_j ≠∅ then
                UpdateList_i.append(clust_j)
        end
        if UpdateList ≠∅ then
            clust_i←RClust.update(clust_i, UpdateList_i)
    end
```

After the filtering process, a stringent homology partition was performed on the filtered dataset, which ensured that *DeEPn* generalizes on the new dataset. To do this, *PSI-CD-HIT* (Li & Godzik 2006; Huang, Niu, Gao, Fu & Li 2010) tool was used, which cluster homologous sequences based on explicit constraints. The constraints used in this study were summarized as follows: if sequences are at least 30% identical, and if alignment covers at least 80% of the shorter sequence, then sequences are mapped onto the same cluster, otherwise on different clusters. To reduce the running time overhead, the three passes of clustering was employed: in the first pass enzyme clusters, *EClust*, were formed in which all enzyme sequences were clustered; resulting in 62,160 groups. In the second pass non-enzyme clusters, *NEClust*, were created in which all non-enzyme sequences were clustered; leading in 95,543 groups, and in the third pass representative clusters, *RClust*, were formed

Accepted Manuscript

in which the representative sequences (According to CD-HIT, for each group, the longest sequence within a cluster or group is known as representative sequence) of the previous two groups (*EClust*, *NEClust*) were clustered; resulting in 89,437 clusters. To make a complete dataset for *DeEPn* model, the Algorithm 1 was applied on *RClust*. In Algorithm 1, to increase the number of elements in the corresponding groups of *RClust*, each sequence s_i from *RClust* groups were checked against representative sequences of *EClust* and *NEClust*. The sequences belonging to *EClust* or *NEClust* were added to *RClust* if it contained representative sequence s_i. The resulting groups of *RClust* contains sequences from enzyme and non-enzyme family. Finally, the sequences from *RClust* clusters were mapped to one of five folds in such a way that all folds contain distinct sets of sequences. Four folds out of five folds were used for training and validation and one set was used for testing.

Algorithm 1: Update cluster algorithm

Input: *RClust, EClust, NEClust*
Result: Return updated *RClust*

```
for each cluster clst_i in RClust do
    UpdateList_i ←∅
    for each sequence seq_i in clust_i do
        clust_j ←∅
        if EClust.is_representative(seq_i) then
            clust_j←EClust:get_cluster(seq_i )
        else if NEClust.is_representative(seq_i )
        then
            clustj←NEClust:get_cluster(seq_i )
        if clust_j ≠∅ then
            UpdateList_i.append(clust_j)
    end
    if UpdateList ≠∅ then
        clust_i←RClust.update(clust_i,UpdateList_i)
end
```

Accepted Manuscript

2.2 Feature Calculation

Protein sequences were converted into numerical feature vectors by *protr* (Xiao, Cao, Zhu & Xu 2015), a library package, from *R* software environment (Team 2013). The calculated features include

conjoint triad, pseudo amino acid composition, CTD (Composition, Transition, and Distribution), quasi sequence order features, auto correlation, and amino acid composition. The description related to features were summarized in Table 2.

2.3 DeEPn

For enzyme classification, *DeEPn* uses 1D convolutional neural network (*CNN*). The 1D *CNN* is akin to the traditional fully connected multilayer neural network; both of these maintains hierarchical architecture, performs dot product with trainable units (weights) w_i and some input signal x_i, adding a bias b, and applying some activation or non-linear function θ. The output of a particular neuron can be defined as:

$$\alpha = \theta(\sum_i (w_i x_i + b)) \quad (1)$$

Where x_i are the outputs of previous layers, and w_i are the strength of connections (weight) between the preceding and current layer. However, unlike traditional fully connected multilayer neural networks, the 1D *CNN* does not maintain full connectivity in all levels of hierarchy. Instead of this, through weight sharing, they maintain local connections between the previous and current layer. For more detail, readers are encouraged to refer to the article (Goodfellow, Bengio & Courville 2016). Following are the two most common layers used in *CNN*, which makes it different from the fully connected multilayer neural *network*:

i. **Convolutional layers:** These layers contain neurons corresponding to the weights (convolutional filter), which performs the dot product between weights and preceding layer outputs through which the filters are locally connected. These layers help the *CNN* to learn essential features (feature map/feature representation) for classification. Apart from this, one can control various hyperparameters, such as filter size, number of filter, padding, stride, and activation function of the output.

Accepted Manuscript

ii. **Pooling layers:** The convolutional layers use filters to learn the most robust feature representations of the input signal, in which the low level filter learns the simplex features, while the high-level filter learns complex ones. However, the limitation with the filter is that it learns position dependent feature mapping from the input signal; which results in different feature representation of the input if there is a small variation in the input signal. To avoid such situations, polling layers were introduced in *CNN*. The pooling layer downsamples the input signal and provides a position independent feature representation of the input signal. Due to this, filters can learn more robust features for classification.

Figure 1, depicts the detailed architecture of the *DeEPn* model. The 1D *CNN* for enzyme classification was inspired from text classification (Shu, Xu & Liu 2017).

In text classification, texts were represented into 1D feature vector, which was applied as an input to the convolutional layer to learn the robust features for classification. A similar strategy was applied in the *DeEPn* model for enzyme classification. The input to the *DeEPn* model were enzyme features of size [9920 X 1]. The *CNN* then convolve with input to extracts important features by applying 46 filters, each of size [3 X 1], resulting in [9914 X 20] feature map. The resultant feature map from the convolutional layers was applied as input to the dropout, batch normalization, and max pooling layers which resulted in [4957 X 20] feature map. The dropout layer avoids the problem due to coadaptation, ensuring that the neurons in the network learn stable features on their own without overfitting. The batch normalization layer reduces the problem of internal co-variant shift, ensuring that the neurons in the succeeding layer do not receive the different distribution of the input signal, and the max pooling layer performs the down sampling of feature map ensuring that the network learns reliable features irrespective of the position of that feature in the input signal. Finally, the feature map of size [4957 X 20] was applied as an input to fully connected layers, which are a combination of dense, dropout, and batch normalization layer, as shown in the Figure 1. The output layer contains eight neurons corresponding to 8 classes (7 enzyme class (*EC1* to *EC7*) and 1 non-

Accepted Manuscript

enzyme class). The activation function used at the output layer was "*softmax*" while the other layers either use "*relu*" or "*sigmoid*" activation function as shown in Figure 1. Over the course of time, each neuron learns the feature representation of the input signal based on their learning capability, resulting in each neuron of the network having different learning rates for optimizing the objective function. Hence, in the proposed work, *adam* optimizer was used as the optimizer function. The loss function used to train the neural network was "*categorical loss entropy*" function. The traces of training-validation loss and training validation accuracy was shown in Figure 2. The train and test accuracy of model are 99.87%, and 99.86% respectively.

Accepted Manuscript

2.4 Evaluating Criteria

To evaluate the performance of each classifier, different statistical scores were used. In fact, in a typical supervised binary classification problem, each query point from the test sets have their own true class label. However, during the evaluation process, the classifier maps the query points onto one of following categories: true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) (Semwal, Mondal & Nandi 2017). To achieve such categories for each class, the multiclass classification uses one versus rest approach. In this approach, the query point belongs to a particular class, considered as a positive or negative point. Based on this, *TP*, *TN*, *FP*, and *FN* is calculated for each class, then the following statistical scores are used to evaluate the performance of classifier correspond to each class:

Accuracy (*ACC*): It is the measure of correct prediction out of the total forecast.

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (2)$$

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

$$PPV = \frac{(TP)}{(TP+FP)} \quad (3)$$

Recall/Sensitivity (*SEN*): It can be defined as the ability of a classifier to predict all relevant data correctly. It can be calculated as the ratio between predicted true positive (*TP*) to all positive observations (*TP*+*FN*).

$$SEN = \frac{(TP)}{(TP+FN)} \quad (4)$$

***F1-score*:** The *F1-score* uses only three categories (*TP, FP,* and *FN*) 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.

$$F1 - score = 2 * \frac{(PPV*SEN)}{(PPV+SEN)} \quad (5)$$

Matthew's correlation coefficient (*MCC*): It can be defined as the correlation between the observed and predicted values. The reason behind calculating *MCC* is that the accuracy and *F1-scores* sometimes overestimate the performance of the classifier (Abma 2009). The *MCC* value +1 represents the best prediction, 0 represents random prediction, and -1 represents the disagreement between correct class and predicted class.

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FN)*(TP+FP)*(TN+FP)*(TN+FN)}} \quad (6)$$

Accepted Manuscript

3 Results and Discussion

3.1 *DeEPn* result analysis

In order to evaluate the performance of the *DeEPn* model, apart from the above specified evaluation criteria, two graphical performance measuring parameters *ROC* and *Precision-Recall* curve were also used, which demonstrated promising performance of our *DeEPn* model for enzyme functional annotation. Table 3 showed the evaluation metric of *DeEPn* corresponding to seven enzyme functional classes and one non-enzyme class. The accuracy (more than 99%) parameter showed that *DeEPn* gave an excellent classification performance for all categories. However, according to Abma et al. (Abma 2009), and Valverde et al. (Valverde-Albacete, Carrillo-de-Albornoz & Peláez-Moreno 2013), due to the accuracy paradox, the accuracy alone is not a good indicator for classification performance; thus the *F1-score* of *DeEPn* was also calculated corresponding to all eight classes. The *F1-score* is the harmonic summation of precision and recall, indicating the ability of the classifier to detect the actual true sample, out of all positive and negative examples, and is hence considered as a more robust performance metric for classification compared to the accuracy. Table 3, depicted that the *F1-score* of *DeEPn* model was highly significant (≥ 99%), corresponding to all classes. To evaluate the model performance, *F1-score* is a more robust metric compared to accuracy; however, in the case of unbalanced test cases, the *F1-score* sometimes overestimate the classification performance (Boughorbel, Jarray & El-Anbari 2017). To avoid such a situation, a balanced performance metric *MCC* score (≥ +0.98) was calculated for all classes ensuring the higher predictive quality of our *DeEPn* model.

Accepted Manuscript

3.2 *ROC* analysis (receiver operating characteristics)

ROC is a graphical measure for evaluating the performance of the classification model. It shows the relationship between *TPR* (true positive rate or sensitivity) and *FPR* (false positive rate or 1-

specificity) at various thresholds, where in general *FPR* is plotted against x-axis and *TPR* is plotted against y-axis (Semwal, Aier, Raj & Varadwaj 2017). The *FPR* of the classification model is the determination of false positive prediction out of total negative cases. At the same time, the *TPR* is the determination of true positive prediction out of all positive cases and is defined in Eq 4. In our study, the *ROC* curve was generated using the *sklearn* (Pedregosa et al. 2011) package "*roc_auc*" (Fawcett 2006) and Figure 3 was used to depict the *ROC* curve corresponding to all eight classes plotted against different thresholds.

The perfect classification situation of *ROC* curve was depicted by the top left corner, where the sensitivity and specificity are 100%. The random performance of classification was depicted by the diagonal line (coordinate (0, 0) to (0, 1)) (Semwal et al. 2017). Hence, to have a good classification model, the *ROC* of the model must be above the diagonal line, i.e. more the area under the curve, better a classifier is. In our study, the micro average (Van Asch 2013) *ROC* of the model has also plotted along with the eight classes to demonstrate the average *ROC* performance of the *DeEPn* model. Micro average *ROC* metric of *DeEPn* model was plotted against micro average *FPR* (*FPRμ*) and micro average *TPR* (*TPRμ*), where micro average *FPR* and *TPR* represented the contribution of all eight classes and was defined in Eq 7 and Eq 8.

$$TPR_{\mu} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FN_i} \quad (7)$$

$$FPR_{\mu} = \frac{\sum_{i=1}^{k} FN_i}{\sum_{i=1}^{k} TP_i + FN_i} \quad (8)$$

Where *k* was used to denote the number of classes. The area under the curve for each class: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases, Translocases, and Non-Enzyme was 0.99962 square unit, 0.99962 square unit,0.99921 square unit, 0.99998 square unit,0.99994 square unit, 0.99999 square unit, 0.99956 square unit, and 0.99953 respectively.

Accepted Manuscript

3.3 *Precision-Recall* analysis

Another graphical measure to evaluate the performance of the classifier is the *Precision-Recall* curve, plotted at various thresholds against recall and precision on x-axis and y-axis respectively. The *Precision-Recall* curve helps in analyzing the positive predictive quality of classifier as it does not use actual negative cases in its calculation. Figure 4 represented the *Precision-Recall* curve of *DeEPn* model corresponding to all eight classes, plotted with different thresholds. The perfect classification condition for *Precision-Recall* curve is the top right corner with an area under the curve of 1 square unit, where precision and recall of classifier are 100% ensuring the ability of the classifier to predict actual true positive without any false positive prediction. Therefore, to have a good classification model, there must be a good trade-off between the precision and recall, i.e. area under the *Precision-Recall* curve must be close to 1 square unit. In our study, the micro average (Van Asch 2013) *Precision-Recall* curve of *DeEPn* has also plotted along with eight classes to demonstrate the average *Precision-Recall performance* of *DeEPn* model. In Figure 4, the area under the curve of each classes: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases, Translocases, and Non-Enzyme was 0.99 849 square unit, 0.99937 square unit,0.99738 square unit, 0.99964 square unit,0.99906 square unit, 0.99999 square unit, 0.99844 square unit, and 0.99922 square unit respectively.

3.4 Comparison with other existing Tools

To perform an unbiased and data independent comparison among the *DeEPn* and other existing tools (*ECPred* (Dalkiran et al. 2018), *SVM-Prot* (Cai et al. 2003; Li et al. 2016)), the dataset of *ECPred*, *SVM-Prot*, and *DEEPre* (Li et al. 2018) were combined. Due to the unavailability of the *DEEPre* server, we exclude the tool for further comparative analysis. *ECPred* uses multiple binary classifiers, while *SVM-Prot* uses the Support Vector Machine to predict enzyme functional classes. Figure 5

Accepted Manuscript

represented individual class comparison results, while Table 4 described the average score corresponding to the evaluation criteria. In all four evaluation metric criteria, *DeEPn* outperformed all existing tools. The micro average accuracy of *DeEPn* model was 99%, while the accuracy of *ECPred* and *SVM-Prot* were 92% and 56%, respectively. The *F1-score* of *DeEPn* was much higher (0.99) than *ECPred* (0.95) and *SVM-Prot* (0.58), validating the fact that the classifier had a higher chance of making a clear distinction between the true sample and negative samples. Again as expected, the *MCC* score of *DeEPn* was much higher (+0.9929) than *ECPred* (+0.9078) and *SVM-Prot* (+0.4640), showing the ability of the classifier to make a more accurate prediction even if distribution among classes were non uniform.

Accepted Manuscript

4 Conclusion

DeEPn is a machine learning prediction model for predicting the function of enzymes, based on the *EC* number system proposed by the International Union of Biochemistry and Molecular Biology (*IUBMB*). Not only can *DeEPn* quantify enzymes corresponding to six functional class (*EC1* to *EC6*), but also quantify enzyme corresponding to the seventh functional class (*EC7*), which was not available in earlier *EC* number prediction tools. To predict the enzymatic function of amino acid sequences, *DeEPn* converts these amino-acid sequences into various sequence-related features, such as *AAC*, *CTD*, *PseAAC*, etc. The training and validation dataset of *DeEPn* model was retrieved from *UniProtKB/Swiss-Prot* segment of *Uniprot* Knowledgebase, which contains reviewed and manually annotated entries related to protein sequences. Independent dataset comparison of *DeEPn* model was further performed against other states of art *EC* number prediction tools (*ECPred* and *SVM-Prot*) to analyze the predictive power of our proposed model. The result showed that *DeEPn* outperformed all other existing mechanisms in terms of its average accuracy (99.42%), precision (0.99), recall (0.99),

and Matthew's correlation coefficient (+0.9929). In conclusion, *DeEPn* was able to outperform several alternative tools for correctly annotating the enzyme function with higher accuracy and precision.

Acknowledgements

The authors acknowledge the Department of Bioinformatics & Applied Sciences, Central Computing Facility, Indian Institute of Information Technology-Allahabad for providing computing facility.

Conflict of interest

The authors have no Conflict of Interest.

References

Abma, B. (2009). Evaluation of requirements management tools with support for traceability-based change impact analysis. *Master's thesis, University of Twente, Enschede*.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. & Magrane, M. (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research 32*: D115-D119.

Arakaki, A. K., Huang, Y. & Skolnick, J. (2009). EFICAz 2: enzyme function inference by a combined approach enhanced by machine learning. *BMC bioinformatics 10*: 107.

Boughorbel, S., Jarray, F. & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one 12*.

Cai, C., Han, L., Ji, Z. & Chen, Y. (2004). Enzyme family classification by support vector machines. *Proteins: Structure, Function, and Bioinformatics 55*: 66-76.

Accepted Manuscript

Cai, C., Han, L., Ji, Z. L., Chen, X. & Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic acids research 31*: 3692-3697.

Cai, Y.-D. & Chou, K.-C. (2006). Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. *Journal of theoretical biology 238*: 395-400.

Chou, K.-C. & Elrod, D. W. (2003). Prediction of enzyme family classes. *Journal of Proteome Research 2*: 183-190.

Cornish-Bowden, A. (2014). Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspectives in Science 1*: 74-87.

Correa, L. D. L. & Dorn, M. (2018). *A knowledge-based artificial bee colony algorithm for the 3-D protein structure prediction problem*. 2018 IEEE Congress on Evolutionary Computation (CEC), IEEE.

Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V. & Doğan, T. (2018). ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC bioinformatics 19*: 1-13.

De Ferrari, L., Aitken, S., van Hemert, J. & Goryanin, I. (2012). EnzML: multi-label prediction of enzyme classes using InterPro signatures. *BMC bioinformatics 13*: 61.

Dobson, P. D. & Doig, A. J. (2005). Predicting enzyme class from protein structure without alignments. *Journal of molecular biology 345*: 187-199.

Espadaler, J., Eswar, N., Querol, E., Avilés, F. X., Sali, A., Marti-Renom, M. A. & Oliva, B. (2008). Prediction of enzyme function by combining sequence similarity and protein interactions. *BMC bioinformatics 9*: 249.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters 27*: 861-874.

Fenichel, G. M. (2007). Hypotonia, Arthrogryposis, and rigidity. *Neonatal neurology. 4th ed. Churchill Livingston*: 37-68.

Accepted Manuscript

Goddard, J.-P. & Reymond, J.-L. (2004). Enzyme assays for high-throughput screening. *Current opinion in biotechnology 15*: 314-322.

Godzik, A. (2011). Metagenomics and the protein universe. *Current opinion in structural biology 21*: 398-403.

Goedde, H., Agarwal, D., Harada, S., Meier-Tackmann, D., Ruofu, D., Bienzle, U., Kroeger, A. & Hussein, L. (1983). Population genetic studies on aldehyde dehydrogenase isozyme deficiency and alcohol sensitivity. *American journal of human genetics 35*: 769.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*, MIT press.

Huang, C.-C., Lin, C.-Y., Chang, C.-W. & Tang, C. Y. (2012). *Automatic Prediction of Enzyme Functions from Domain Compositions Using Enzyme Reaction Prediction Scheme*. 2012 International Conference on Biomedical Engineering and Biotechnology, IEEE.

Huang, W.-L., Chen, H.-M., Hwang, S.-F. & Ho, S.-Y. (2007). Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. *Biosystems 90*: 405-413.

Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics 26*: 680-682.

Jiang, Z., Kumar, M., Padula, M. P., Pernice, M., Kahlke, T., Kim, M. & Ralph, P. J. (2017). Development of an efficient protein extraction method compatible with LC-MS/MS for proteome mapping in two Australian seagrasses Zostera muelleri and Posidonia australis. *Frontiers in plant science 8*: 1416.

Kern, A. D., Oliveira, M. A., Coffino, P. & Hackert, M. L. (1999). Structure of mammalian ornithine decarboxylase at 1.6 Å resolution: stereochemical implications of PLP-dependent amino acid decarboxylases. *Structure 7*: 567-581.

Kumar, C. & Choudhary, A. (2012). A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP Journal on Bioinformatics and Systems Biology 2012*: 1.

Accepted Manuscript

Kumar, N. & Skolnick, J. (2012). EFICAz2. 5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics 28*: 2687-2688.

Lee, B. J., Lee, H. G. & Ryu, K. H. (2008). *Design of a novel protein feature and enzyme function classification*. 2008 IEEE 8th International Conference on Computer and Information Technology Workshops, IEEE.

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics 22*: 1658-1659.

Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L. & Gao, X. (2018). DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics 34*: 760-769.

Li, Y. H., Xu, J. Y., Tao, L., Li, X. F., Li, S., Zeng, X., Chen, S. Y., Zhang, P., Qin, C. & Zhang, C. (2016). SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PloS one 11*.

Lu, L., Qian, Z., Cai, Y.-D. & Li, Y. (2007). ECS: an automatic enzyme classifier based on functional domain composition. *Computational biology and chemistry 31*: 226-232.

Mahan, F. R., Bailey, L., Pilipenko, V. & Prada, C. (2019). Pain and fatigue associated with generalized joint hypermobility in Gaucher disease. *Molecular Genetics and Metabolism 126*: S97.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science 298*: 1912-1934.

Masci, J., Meier, U., Cireşan, D. & Schmidhuber, J. (2011). *Stacked convolutional auto-encoders for hierarchical feature extraction*. International conference on artificial neural networks, Springer.

Medicine, N. L. o. (2000). *Medical subject headings*, US Department of Health and Human Services, Public Health Service, National ….

Moss, G. P. (2010). Enzyme nomenclature. *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse*.

Accepted Manuscript

Nagao, C., Nagano, N. & Mizuguchi, K. (2014). Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *PloS one 9*.

Nasibov, E. & Kandemir-Cavas, C. (2009). Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. *Computational biology and chemistry 33*: 461-464.

Pascal, J. M., O'Brien, P. J., Tomkinson, A. E. & Ellenberger, T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature 432*: 473-478.

Pastores, G. M. & Hughes, D. A. (2018). Gaucher disease. *GeneReviews®[Internet]*, University of Washington, Seattle.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research 12*: 2825-2830.

Pritlove, D., Gu, M., Boyd, C., Randeva, H. & Vatish, M. (2006). Novel placental expression of 2, 3-bisphosphoglycerate mutase. *Placenta 27*: 924-927.

Qiu, J.-D., Luo, S.-H., Huang, J.-H. & Liang, R.-P. (2009). Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform. *Journal of theoretical biology 256*: 625-631.

Quester, S. & Schomburg, D. (2011). EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC bioinformatics 12*: 376.

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of molecular biology 318*: 595-608.

Rotticci, D., Rotticci-Mulder, J. C., Denman, S., Norin, T. & Hult, K. (2001). Improved enantioselectivity of a lipase by rational protein engineering. *ChemBioChem 2*: 766-770.

Roy, A., Yang, J. & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research 40*: W471-W477.

Accepted Manuscript

Semwal, R., Aier, I., Raj, U. & Varadwaj, P. K. (2017). Pharmadoop: a tool for pharmacophore searching using Hadoop framework. *Network Modeling Analysis in Health Informatics and Bioinformatics 6*: 20.

Semwal, V. B., Mondal, K. & Nandi, G. C. (2017). Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach. *Neural Computing and Applications 28*: 565-574.

Sharif, M. M., Thrwat, A., Amin, I. I., Ella, A. & Hefeny, H. A. (2015). Enzyme function classification based on sequence alignment. *Information Systems Design and Intelligent Applications*, Springer**:** 409-418.

Shen, H.-B. & Chou, K.-C. (2007). EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. *Biochemical and biophysical research communications 364*: 53-59.

Shu, L., Xu, H. & Liu, B. (2017). Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tian, W., Arakaki, A. K. & Skolnick, J. (2004). EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic acids research 32*: 6226-6239.

Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of molecular biology 333*: 863-882.

Valverde-Albacete, F. J., Carrillo-de-Albornoz, J. & Peláez-Moreno, C. (2013). *A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks*. International Conference of the Cross-Language Evaluation Forum for European Languages, Springer.

Van Asch, V. (2013). Macro-and micro-averaged evaluation measures [[basic draft]]. *Belgium: CLiPS 49*.

Volpato, V., Adelfio, A. & Pollastri, G. (2013). Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. *BMC bioinformatics 14*: S11.

Accepted Manuscript

Wang, Y.-C., Wang, X.-B., Yang, Z.-X. & Deng, N.-Y. (2010). Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein and Peptide Letters 17*: 1441-1449.

Wang, Y.-C., Wang, Y., Yang, Z.-X. & Deng, N.-Y. (2011). Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. *BMC systems biology 5*: S6.

Webb, E. C. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press.

Xiao, N., Cao, D.-S., Zhu, M.-F. & Xu, Q.-S. (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics 31*: 1857-1859.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods 12*: 7.

Yu, C., Zavaljevski, N., Desai, V. & Reifman, J. (2009). Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases. *Proteins: Structure, Function, and Bioinformatics 74*: 449-460.

Zeng, D., Liu, K., Lai, S., Zhou, G. & Zhao, J. (2014). Relation classification via convolutional deep neural network.

Zhang, C., Freddolino, P. L. & Zhang, Y. (2017). COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research 45*: W291-W299.

Zhou, X.-B., Chen, C., Li, Z.-C. & Zou, X.-Y. (2007). Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of theoretical biology 248*: 546-551.

Accepted Manuscript

Zou, H.-L. & Xiao, X. (2016). Classifying multifunctional enzymes by incorporating three different models into chou's general pseudo amino acid composition. *The Journal of membrane biology 249*: 551-557.

Accepted Manuscript

Table 1. Summary of Enzyme and Non-Enzyme Dataset.

SI.No.	Class Description	Number of proteins
1	Oxidoreductases (EC1)	25,378
2	Transferases (EC2)	70,756
3	Hydrolases (EC3)	45,703
4	Lyases (EC4)	17,237
5	Isomerases (EC5)	10,375
6	Ligases (EC6)	22,238
7	Translocases (EC7)	8,160
8	Non-Enzyme	2,41,116

Accepted Manuscript

Table 2. List of protein sequence features calculated using 'protr'.

SI.No.	Feature Group	Feature Name	Dimension
1	Conjoint Triad	Conjoint Triad	343
2	Pseudo-Amino Acid Composition	Pseudo-Amino Acid Composition	50
		Amphiphilic Pseudo-Amino Acid Composition	80
3	CTD	Composition	21
		Transition	21
		Distribution	105
4	Quasi-Sequence-Order	Sequence-Order-Coupling Number	60
		Quasi-Sequence-Order Descriptors	100
5	Autocorrelation	Moran Autocorrelation	240
		Geary Autocorrelation	240
		Normalized Moreau-Broto Autocorrelation	240
6	Amino Acid Composition	Amino Acid Composition	20
		Dipeptide Composition	400
		Tripeptide Composition	8000

Accepted Manuscript

Table 3. *DeEPn* statistical scores corresponds to each classes.

Performance Measures Classes	ACC	PPV	SEN	F1-score	MCC
Oxidoreductases	0.9987	0.99	0.99	0.99	+0.9884
Transferases	0.9978	0.99	0.99	0.99	+0.9920
Hydrolases	0.9977	0.99	0.99	0.99	+0.9876
Lyase	0.9994	1.00	0.99	0.99	+0.9931
Isomerases	0.9997	1.00	0.99	0.99	+0.9935
Ligases	0.9998	1.00	1.00	1.00	+0.9983
Translocases	0.9997	1.00	0.99	0.99	+0.9943
Non-Enzyme	0.9960	0.99	1.00	1.00	+0.9920

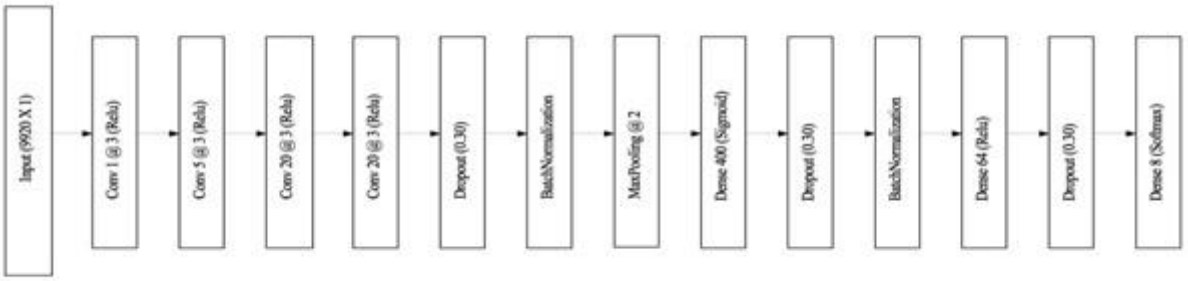
Accepted Manuscript

Table 4. *DeEPn*, *ECPred*, and *SVM-Prot* average statistical scores corresponds to each classes.

Performance Measures Classes	*ACC*	*PPV*	*SEN*	*F1-score*	*MCC*
DeEPn	0.9942	0.99	0.99	0.99	+0.9929
ECPred	0.9232	0.92	0.97	0.95	+0.9078
SVM-Prot	0.5629	0.56	0.59	0.58	+0.4640

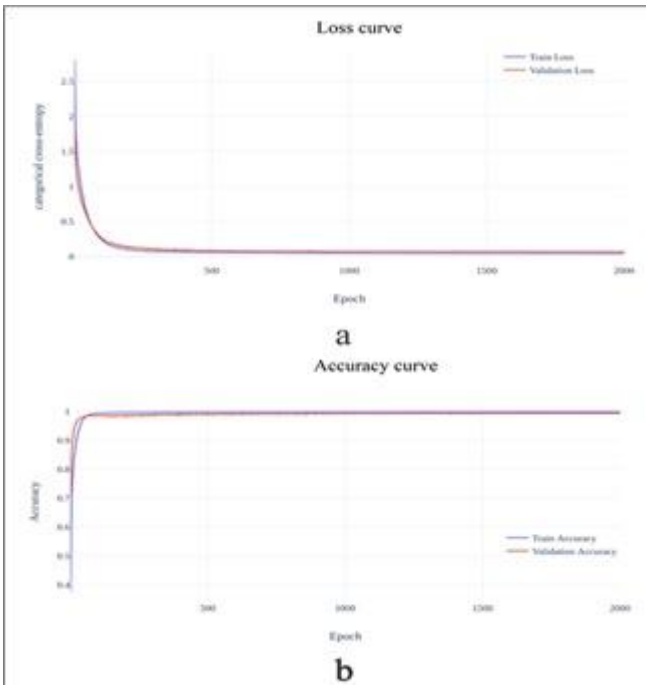
Accepted Manuscript

Figure 1. Architectural view of DeEPn model. The vertical rectangular bars represent the layer of DeEPn model and arrow lines between two layers represent data flow between layers.



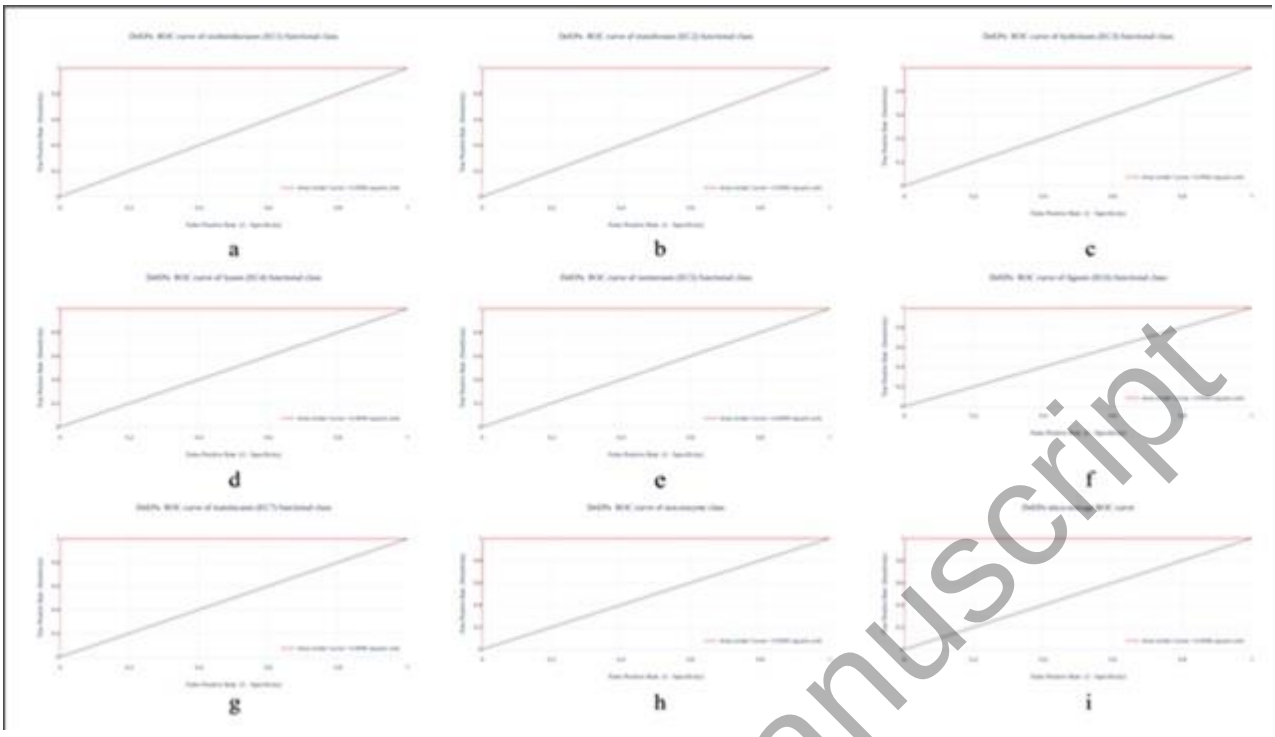
Accepted Manuscript

Figure 2. a) Tracing of training and validation loss b) Tracing of training and validation accuracy during training of DeEPn.



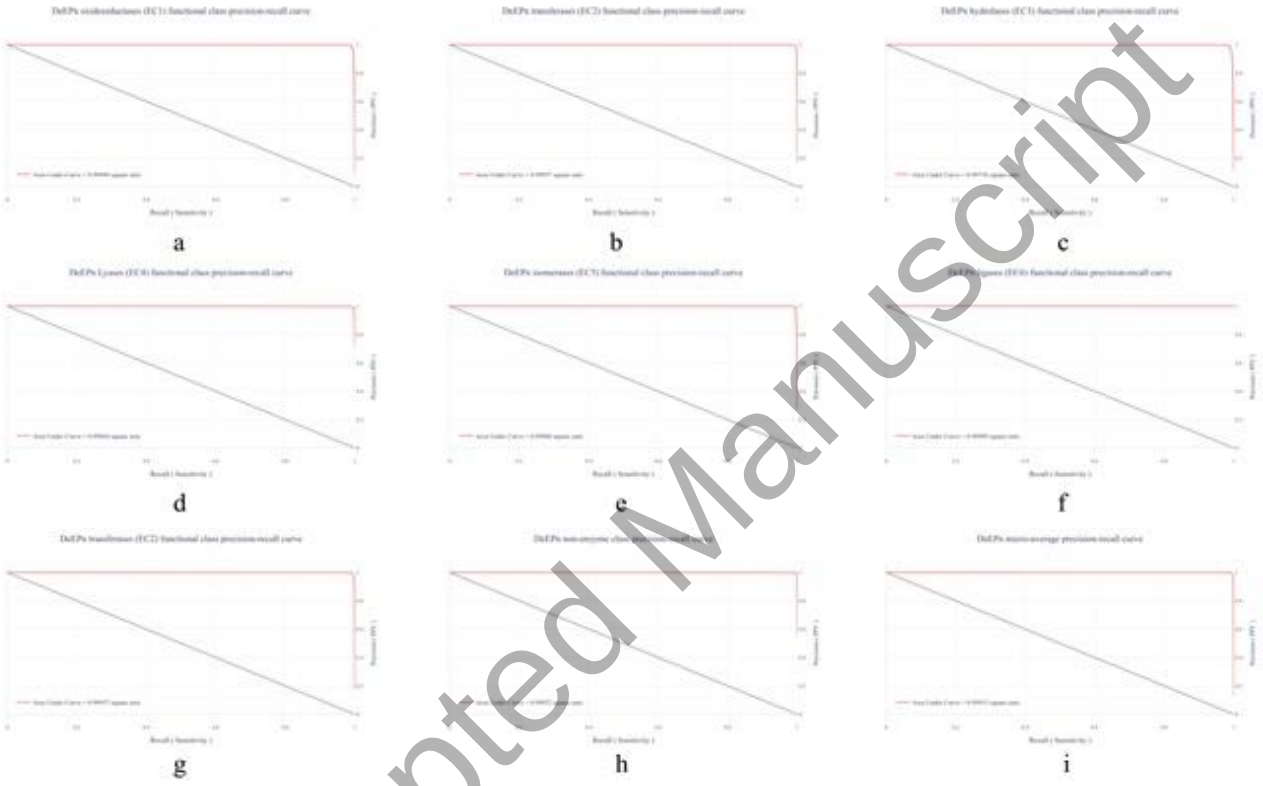
Accepted Manuscript

Figure 3. *ROC* performance analysis of *DeEPn* against eight classes. a) Oxidoreductases (*EC1*) class ROC curve. b) Transferases (*EC2*) CLASS *ROC* curve. c) Hydrolases (*EC3*) class *ROC* curve d) Lyases (*EC4*) class *ROC* curve e) Isomerases (*EC5*) class *ROC* curve. f) Ligases (*EC6*) class *ROC* curve. g) Translocases (*EC7*) class *ROC* curve. h) Non-Enzyme class *ROC* curve. i) Micro-Average *ROC* urve corresponds to eight classes.



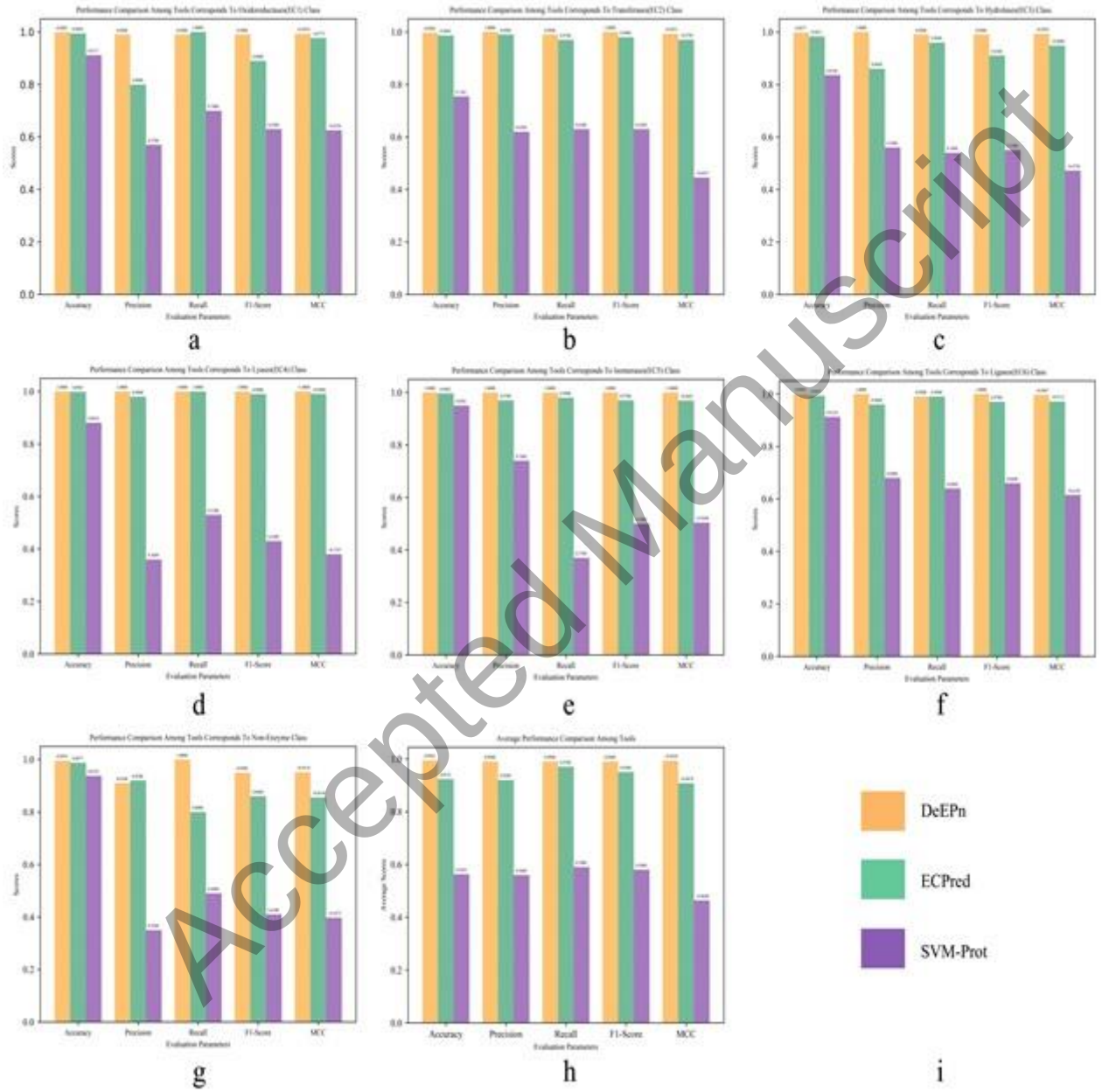
Accepted Manuscript

Figure 4. *Precision-Recall* performance analysis of *DeEPn* against eight classes. **a)** Oxidoreductases (*EC1*) class *precision-recall* curve. **b)** Transferases (*EC2*) class *precision-recall* curve. **c)** Hydrolases (*EC3*) class *precision-recall* curve **d)** Lyases (*EC4*) class *precision-recall* curve. **e)** Isomerases (*EC5*) class *precision-recall* curve. **f)** Ligases (*EC6*) class *precision-recall* curve. **g)** Translocases (*EC7*) class *precision-recall* curve. **h)** Non-Enzyme class *precision-recall* curve. **i)** Micro-Average *precision-recall* curve corresponds to eight classes.



Accepted Manuscript

Figure 5. *DeEPn*, *EC_Pred*, and *SVM-Prot* statistical scores comparison corresponds to **a)** Oxidoreductases (*EC1*) class, **b)** Transferases (*EC2*) class, **c)** Hydrolases (*EC3*) class, **d)** Lyases (*EC4*) class, **e)** Isomerases (*EC5*) class, **f)** Ligases (*EC6*) class, **g)** Non-Enzyme class. **h)** Average statistical score corresponds to eight classes. **i)** Bar plot colour indicator corresponds to *DeEPn*, *ECPred*, and *SVM-Prot*.



Accepted Manuscript