Pathway-Specific Protein Domains (PSPD) Discrimination by Using a Hybrid Feature Space Based on Deep Neural Network(DNN)

Ali Ghulam
Shaanxi Normal University
XiuJuan Lei (xjlei@snnu.edu.cn)
Shaanxi Normal University https://orcid.org/0000-0002-9901-1732
Yuchen Zhang
Shaanxi Normal University
Zhenqiang Wu
Shaanxi Normal University

Research

Keywords: Proteins sequence, pathway-specific proteins, AAC, DPC, AAindex, PACC

Posted Date: September 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-70425/v1

RESEARCH ARTICLE

Pathway-specific protein domains (PSPD) discrimination by using a hybrid feature space based on deep neural networks (DNN)

Ali Ghulam[a], Xiujuan Lei[a], Yuchen Zhang[a], and Zhenqiang Wu[a]
School of Computer Science, Shaanxi Normal University, Xian, China.

Full list of author information is available at the end of the article

Abstract

The Pathway-specific protein domains (PSPDs) are important tools in examining drug growth as they provide a fast, reliable, and inexpensive way of estimating complex new molecular targets in specific diseases. The protein architecture prevents the formation of a direct correlation between signal transduction behavior and cellular structure. Accordingly, protein–tissue factor pathway inhibitor 2 isotypes 1 precursors have been used to encode peptide sequence information into specific feature structures. The measurable structure-activity classification model obtained by machine learning technology can predict pathway-specific protein interactions and new signaling peptides. We introduce deep neural network (DNN)-based PSPDs, abbreviated as DNNPSPDs, as the first pathway-specific protein domain that is built based on five extant models, namely, the AAindex, pseudo-amino acid composition, amino acid composition, composition mood of pseudoamino acids, and dipeptide composition. A total of 900 proteins with undetermined roles collected from the PDB data base are tested to evaluate the predictive power of this model. Various combinations of the available feature selection technologies are also combined to process a hybrid function space. DNNPSPDs predicts PSPDs by using features that are automatically learned from primary protein sequences. The sequences of pathway-associated proteins are sequentially fed into and decoded in neural network layers. Several classifications are also employed. DNNPSPDs achieves a prediction accuracy of 0.957 at a Matthew's correlation coefficient (MCC) of 91.86%, with DPC, and 2nd achieve high prediction score 0.936 at Matthew's correlation coefficient (MCC) of 88.02%, accuracy which is probably better. In terms of ROC–AUC, DNNPSPDs achieves a ROC–AUC curve of 0.982, which is larger than that of the other machine learning classifiers. A study using an alternative dataset reveals that our primary pathways, as pathway-specific protein domains, have accurate and reliable associations, thereby proving the viability of the proposed DNNPSPDs.

Keywords: Proteins sequence, pathway-specific proteins, AAC, DPC, AAindex, PACC

[a] *School of Computer Science, Shaanxi Normal University, Xi'an, China*
*Corresponding Author: Xiujuan Lei**
e-mail: xjlei@snnu.edu.cn,

Introduction

Machine learning techniques are dominant statistical strategies for predicting pathway-specific protein domains (PSPDs). Creating a realistic feature set and selecting matching machine learning algorithms are two main stages in deep neural network (DNN)-based and long- and short-term memory machine learning predictions. This study provides detailed information on each cell in the human pathway that is associated with many proteins, with each protein serving one or more specific functions. A cell can receive numerous signals at the same time and use these data to build an integrated action plan. PSPDs have proteins as their functional units. Many of these proteins are involved in various biological processes, although some of them are connected to certain pathways. The recent advances in experimental methods have improved the pathway prediction capabilities of theoretical methods, such as those based on homology and protein pathway analysis. In this review, we propose a novel method for predicting protein pathways in cancer. Genetically complex diseases have recently attracted much research attention, and studying the protein–pathway association has become a key process in predicting disease pathways. The disease pathway has the basic intention of disease pathogenesis and characteristics[1]. Many studies on proteins and association pathways have also begun to consider those components that are associated with cells, but the exact mechanical effects of these pathways remain unclear. Other studies show that some machine learning methods outperform in pathological prediction by relying on biological disease pathway has the basic intention of disease pathogenesis and characteristics[1]. Many studies on proteins and association pathways have also begun to consider those components that are associated with cells, but the exact mechanical effects of these pathways remain unclear. Other studies show that some machine learning methods outperform in pathological prediction by relying on biological networks. In order to limit the combinatorial explosion, the prediction of human cancer pathways is aimed at identifying signaling pathways in PSPD prediction[2]. However, previous studies have largely ignored the role of PSPDs in identifying the roles of proteins in the complex interactions and biological mechanisms that drive cellular procedures. While the function of proteins needs to be verified by hand in a wet laboratory, scientists require a suggestion before they can even attempt to determine the probable function of a protein. Biologists can use computers to make these gene-function assumptions. Studies show that genome sequencing has become a routine practice in determining the functions of proteins. However, the application of gene sequencing in laboratories remains controversial, thereby increasing the importance of computational gene purpose prediction. Computational approaches are deemed suitable for function prediction as they generate inferences from experimental data that identify the similarities between a gene and its known proteins. These approaches include sequence similarity tools, such as the basic local alignment search tool that searches for all previously recorded sequences and generates a list of possible roles for these sequences.

Computational biology urgently requires new methods that can accurately reflect the nature of biological procedures. Previous studies have attempted to identify such nature based on hierarchical multiple protein features. This approach takes evolutionary relationships into account unlike traditional sequence-based methods and has a better allocation function compared with the simple backbone of amino acids. Machine learning methods have also been used to predict whether a protein has a dual role. Genetically complex diseases have recently attracted much research attention, and studying PSPD associations plays an important role in predicting disease pathways. This study attempts to establish a connection between pathways and proteins, provide detailed information about the pathogenesis of a complex disease and its features, and highlight the role of signaling pathways in predicting protein–pathway associations[3]. Computational biology is a growing discipline that combines research methods with system biology to explore various biological phenomena. Cellular classifications, such as PSPDs, allow us to examine the structures of cells. Nevertheless, one major challenge in system biology is fully integrating genomic and proteomic knowledge into other data sources, such as printed literature, which can translate the original data into useful information and contribute to the present knowledge of biology[4]. The recognized proteins and the abundant differences in protein-tissue factor pathway inhibitor 2 isoform 1 precursor, Approximately, PSPD includes a simple insert that serves as a signal to retention cells. A standard alveolar septum structure during embryogenesis, normal gastrointestinal tract growth, normal Leydig cell development, and spermatogenesis are needed. The normal production of oligodendrocytes and normal formation of myelin are also needed in the spinal cord and cerebellum given their important roles in wound healing. PSPD associated with a pathway (R-HSA-3000171). This study examines the relationship between protein group coherence and pathway assignment based on a functional association. To this end, we use 15 proteins and 4 protein–pathway associations(NP_001230133.1.NP_057665.2.NP_001135936.1.NP_001135937.1) as the name of Evolutionarily conserved signaling intermediate in Toll pathway, mitochondrial isoform 1 precursor which is the type of (Homo sapiens). These above 15 proteins reviewed by (Swiss-port)- manually interpreted.

The UniProt Knowledgebase (UniProtKB) is considered the most source of operational data related to proteins given its complete, consistent, and powerful annotations. Each entry in UniProtKB (i.e., amino acid sequence, protein name or description, classification data, and reference information) requires as much annotation information as possible. The use of mining rules in predicting human pathways associated with prokaryotic UniProtKB data has received much attention in recent years[5]. Many researchers have also proposed

[a] *School of Computer Science, Shaanxi Normal University, Xi'an, China*
*Corresponding Author: Xiujuan Lei**
e-mail: xjlei@snnu.edu.cn,

innovative computational methods for predicting protein function and amino acid sequences. Compared with the mono-functional approach of a single protein, these methods can effectively predict the relationships of perfect protein sets with specific biological processes. Many studies have also applied mining to examine the human involvement in protein pathways with multiple independent functions. Post-translation modifications, such as phosphorylation and ubiquitization, can significantly affect protein functions and have often been used as control devices in signal transduction pathways[6]. Given that proteins communicate with one another, the interaction between proteins and related binding sites should be identified to facilitate hypothesis-driven research and explorations of regulatory networks[7]. The precise subcellular localization of proteins and their tissue distribution in vivo are also important in identifying protein function[8]. Checking if any proteins are affected can also help indicate a pathway[9]. The National Biotechnological Information Center Reference Sequence Database is among the main repositories for DNA and protein sequences and features[10]. Swiss-Prot is another common source of protein information[11]. However, these databases do not contain information on many protein characteristics and functions that are too complex to understand.

In addition to visualizing protein interaction networks, this diagram can also describe the roles of new molecules in large signal networks. These networks have a large number of interactive proteins that can highlight the patterns of certain classes of molecules, pathways, or cellular processes. Motivated by previous studies that have largely focused on sequence characteristics, including amino acid composition (AAC), chain transfer distribution, dipeptides composition (DPC), and pseudo-amino acid composition (PAAC), and have employed various feature selection strategies, such as correlation-, variance-analysis-, minimum-redundancy-, and maximum-correlation-based feature selection, we develop a novel domain called DNNPSPDs for protein function prediction[12-19]. Protein structures have been predicted in the literature by using several machine learning methods, such as artificial neural networks (ANNs). DNNs, as a subgroup of ANN, have several hidden layers. These networks take low-level features as inputs and create highly advanced features at each subsequent layer. DNN-based approaches have been widely applied in the fields of computer vision and natural language processing. The recent improvements in computational capacity have also allowed the scientific community to apply DNN-based methods across various domains, including biomedical data analysis, where DNN algorithms are shown to outperform the conventional predictive methods used in bioinformatics and cheminformatics[20]. DNNs can be divided into two classes. Multi-task DNNs classify input instances into multiple predefined classes/tasks, whereas single-task DNNs aim to produce a binary prediction. DNNs have also been categorized into several classes based on their designs and features. The most common architectures include feed-forward DNNs (i.e., multi-layered perceptron), recurrent neural networks, restricted Boltzmann systems, and deep belief networks[21]. In this study, we introduce DNNPSPDs and demonstrate its excellent performance in PSPDs prediction. Apart from tuning its parameters, we also compare the prediction performance of this model with that of other machine learning classification classifiers, such as AdaBoost (ABC), KNN classifier, SVC classifier, linear discriminant analysis (LDA), gradient boosting classifier (GBC), random forest classifier (RFC), Gaussian naïve Bayes classifier (GNB), decision tree classifier (DTC), multi-layer perceptron classifier (MLPC), and extra trees classifier (ETC). We also adopt feature extraction protocols that have been successfully used in solving various biological problems to determine the best feature extraction method. We hypothesize that AAindex, AAC, DPC, PseAAC, and PAAC are the best feature extraction methods. We also propose a method based on the aforementioned machine learning classifiers.

Material and method

Proposed model

By using a minimum-redundancy dataset, we construct DNNPSPDs as a novel machine learning model for predicting protein–pathway associations. Prior to the construction of this model, we investigate AAindex, AAC, DPC, PAAC and PseAAC, implement a two-step function selection protocol, and define the correct optimal feature selection protocol to remove the irrelevant functionalities. We then compare the prediction performance of five feature encoding models with that of DNNPSPDs and use the results as inputs to 10 machine learning classifiers. We also generate various feature space combinations to create hybrid paces. K-tenfold cross-validation tests are also conducted to evaluate the performance of these classifiers. Figure 1 illustrates the structure of the proposed model.

Datasets

We employ a quantitative approach that involves the use of a dataset mostly enhances the success rate simplification when used in machine learning models. We collect our data from the SMPDB, UniPortKB, and Swiss-Port databases. A total of 900 protein sequences are collected, among which 115 are PSPDs positive and 121 are non-PSPDs as negative. These pathway proteins usage of prediction of subcellular localizations[22]. We have downloaded the dataset from the above database. We preprocess the collected datasets according to the protein–pathway and protein–non-pathway relationships. We saved our dataset in CSV format and then set the parameters of the proposed model.

Feature extraction techniques

Selecting the appropriate feature extraction techniques is a complex process that can greatly facilitate the exploration of biological features. These techniques also require tuning and fine adjustment [23]. The number of elements ***(n)*** in the function vector varies along with the function forms. Generally, the protein sequence vector ***(V)*** of the order index ***(i)*** can be interpreted as:

$$V_i = \left[f_1, f_2, f_3, \ldots\ldots f_n \right] \quad (1)$$

where ***fj*** is an element value of element ***j.*** The function dimension is also known as a component, vector, or column, and these terms are used interchangeably. The characteristics of proteins are usually derived from various sequences. If ***N***

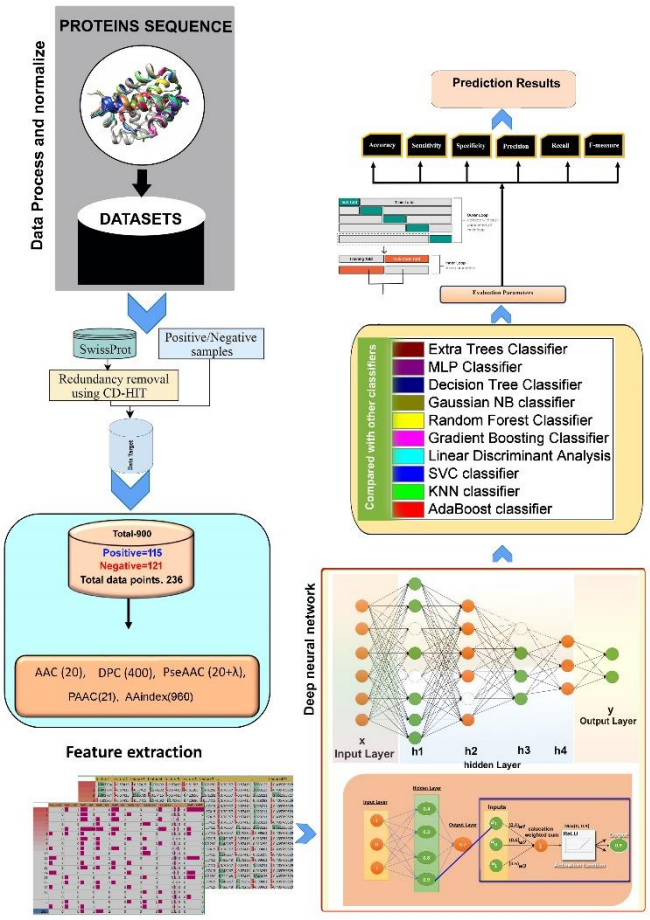

Figure.1 Proposed frame-work model

denotes the number of sequences and ***i=1*** to ***N*** is the order index of a sequence, then the characteristics of a group of protein sequences can be derived as:

$$\begin{bmatrix} f_1,1 & f_2 2 & f_3,3.... & f_1,N \\ f_2,1 & f_2,2 & f_2,3..... & f_2,N \\ f_3,1 & f_3,2 & f_3,3..... & f_3,N \\ f_N,1 & f_N,2 & f_N,3..... & f_N,n \end{bmatrix} \quad (2)$$

The above matrix can also be viewed as a table with ***N*** rows and ***n*** columns. While ***N*** is calculated by a number of sequences, ***n*** greatly depends on the applied feature extraction methods. These features are generally classified into seven groups, namely, AAC, autocorrelation, transformation and distribution of composition, quasi-sequence order, and PseAAC. AAC is computed as the percentage of amino acids in a peptide. Among the 20 produced vectors, only **1** matches an amino acid. Amino acids are organic compounds that mix together to form proteins, and both amino acids and proteins are fundamental components of any living being. These acids have also been described as building blocks of peptides and proteins. Each amino acid is taken from an amino group and a tetrahedral fuel-bound carboxyl group. The carbon is referred to as α-carbon (alpha-carbon). Amino acids, which originally takes the form of a molecular chain of 20 amino acids, vary from each other in comparison to their side chains. Each amino acid is unique in terms of its hydrophilicity, hydrophobicity, polarity, and charge. The features of amino acids in this study are extracted by using AAindex, PAAC, **AAC**, DPC, and PseACC, all of which involve a PSPD sequence formation that reliably classifies protein pathways. Machine learning techniques have been extensively used with organic difficulties in direction to predict protein complicated biological functions.

Composition of amino acids (AAindex)

We compute the binary profiles by using AAindex. When users choose 10 AAindex to construct an amino acid with 10 values, each value reflects **1** AAindex value. This function gives input AA Indices binary profile. If the normalized AAIndex score of a residue is negative, then this residue is assigned a value of **0**; otherwise, a value of **1** is assigned. Collected 236 amino acid indexes for each amino acid index. We choose the complete AAindex that not only accurately represents the physicochemical properties around the acetylation site but also generates redundancy and noise. Twelve physicochemical properties are eventually selected.

Amphiphilic PAAC (APAAC)

APAAC is an improved version of PAAC (Chou, 2005)[24]. We improve the SVM classifier by using different descriptors and find that using APAAC can slightly improve its prediction accuracy. In this case, we use APAAC to extract details on hydrophobicity, hydrophilicity, and amino acid sequences.

Composition of amino acids (AAC)

We determine AAC by calculating the present frequency of each amino acid. AAC is a feature sequence that has been commonly used to measure the occurrence frequency of 20 amino acids within a given sequence fragment. AAC can be expressed as where ***n*** denotes **1** of the 20 types of native amino acids given a sequence ***w*** (hence, ***w1,2,3, 4,…20***), and ***T*** denotes the size of the protein sequence. Many studies from different fields, such as bioinformatics, have developed AAC construction techniques to distinguish different protein structure categories[25], membrane protein types, and protein contact numbers. We obtain predictive data with 21 characteristics from the proposed model. We set the target variable and calculate AAC as given the absence of any missing value in our data, we do not check for null values.

$$\boldsymbol{AAC(w) = \frac{Total\ number\ of\ \ amino\ n(w)}{Total\ number\ of\ all\ possible\ amino\ acids,T} \times\ 100} \quad (3)$$

Dipeptide composition (DPC)

We calculate the DPC of each amino acid in a given peptide sequence length. Every single peptide/protein... Rhythms are classified based on DPC, which compares pairs of residues in a sequence (e.g., AA, AC, and AD). Various PSPD prediction and composition-based algorithms proposed in the literature are using DPC to classify protein sequences[26]. DPC is the composition of measurements for each of the 400 possible dipeptides produced by 20 amino acids. Similar AAC, DPC provides additional local arrangement information in a peptide/protein, as it is the pair of amino acids positioned adjacently [27] We calculate DPC as

$$DPC(i) = \frac{Total\ number\ of\ \ dipeptides\ n(i)}{Total\ number\ of\ all\ possible\ dipeptides,T} \times 100 \quad (4)$$

where $i¼\ 1, 2, 3 \ldots 400$, N denotes the number of dipeptides (represented by amino acid types i), and T represents the size of 400 dipeptides possibly molded by 20 amino acids.

Feature-based on Pseudo-Amino acid composition (Pse_PAAC)

Empirical evidence seems simple and has been used in the existing literature in the fields of bioinformatics and biomedical. By introducing the use of PseAAC[28] in preparing protein sequences, the problem is confused. PseAAC uses the values associated with the factors that represent sequential data[29], such as the subcellular localization of mycobacterium proteins and the superfamily and family classifications of snail toxins. Sub-cells are used to determine the quaternary structure of proteins[30, 31]. IFeature employs a comprehensive protein-related pathway sequence encoding scheme that covers 53 types of feature descriptors. This tool also allows users to choose specific amino acid characteristics from the AAindex database and is equipped with a Python package and web server for selecting features from the PseAAC of protein and peptide sequences by using the following equation:

$$\mathbf{F} = [f_1, \ldots . f_{20}. f_{20} + 1, \ldots . f_{20+\gamma}]^T \quad (5)$$

Chou introduced the concept of PseAAC for estimating cellular protein attributes and proposed a set of discrete numbers based on traditional AAC to determine the potential patterns of sequence order. PseAAC has been effectively used in solving many biological problems[32]. Where T denoted transposing sets, as such $f1,\ldots 20$ is the fraction of remaining **20** distinct amino acids are amino acid association variables determined on the basis of charge, hydrophilicity, and hydrophobicity. PseAAC has also been used in preparing RNA/DNA sequences[33].

Hybrid features PSPDs

For the analysis, we build a hybrid PSPD model with a pathway-specific domain based on AAC, DPC, and PseAAC and then check for the presence of peptides in the training dataset of pathway-specific and non-pathway-specific protein domain motifs. Where the epitope included the patched protein domain motives, AAindex, PAAC, AAC, DPC or PseAAC weight of + **1** was applied based on various classifiers for the machine learning performance. Similarly, the weight was applied to the same if the epitope is positive for non-pathway protein motives as **−1.**

Correlation Matrix

Figure. 2 presents a function correlation matrix constructed by Pyplot and visualized to 12×8 by using rcParams. Using xticks and yticks, the matrix of correlation was introduced with names. None of the features is significant to our target, whereas some features show either negative or positive correlations with the target value.

PSPDs binary classifications

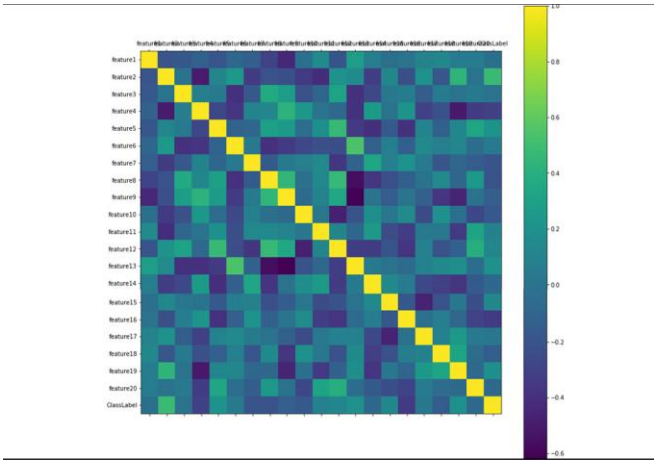

Figure. 2 High Correlation with our target value

We now describe those types of graphics that require only one command to visualize and provide a large amount of information. Figure 3 shows how different sets of features and marks are distributed, which further reinforces the need for scaling. Each bar chart in the figure denotes a specific category of variables that needs to be examined before the implementation of machine learning.

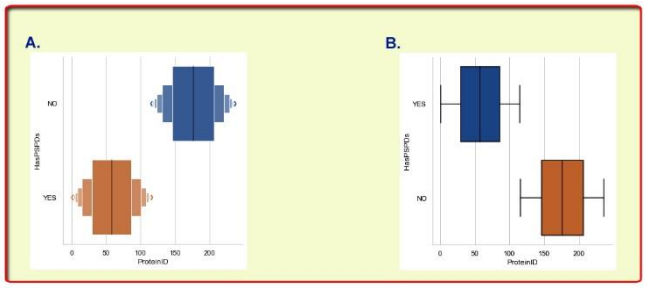

Figure 3. Binary classifications of pathway-specific protein domains

Proposed Deep Neural Networks, Pathway-specific protein domain's Model (DNNPSPDs)

Our proposed model comprises an input layer, several hidden layers, and an output layer. Any neural network with two or more hidden layers is considered a DNN[34]. The layers in a DNN are completely connected, and the secret, hidden, or output layer units are connected to all previous layer units (Figure 4-A). The output values are measured sequentially along with the network layer (Figure 4-B) and are transformed in a non-linear manner until the final output is determined.

The rules of **ReLU** present another problem[35]. Along with Sigmoid and Tanh, **ReLU** is an activation feature widely used in the literature that we also adopt in this study. We perform our optimization analyses and experiments by using **Adam** given its simple implementation, computational effectiveness, and low memory requirements. Accordingly, this software does not affect the gradient sparse when updating its parameters and is optimized for sparse gradients or high noise rates. We employ the cross-entropy method to

prevent very late weight updates. A cross-entropy is a non-negative function, and a smaller loss function corresponds to a better model performance. This is a part of our projected cost model. Thus, it is the objective function to prefer cross-entropy costs. Several scholars have also used **Adam**[36] to achieve an optimized cross-entropy loss detection function at a **dropout** rate of 0.5[37]. The **softmax** function[38] function is seen as a class probability.

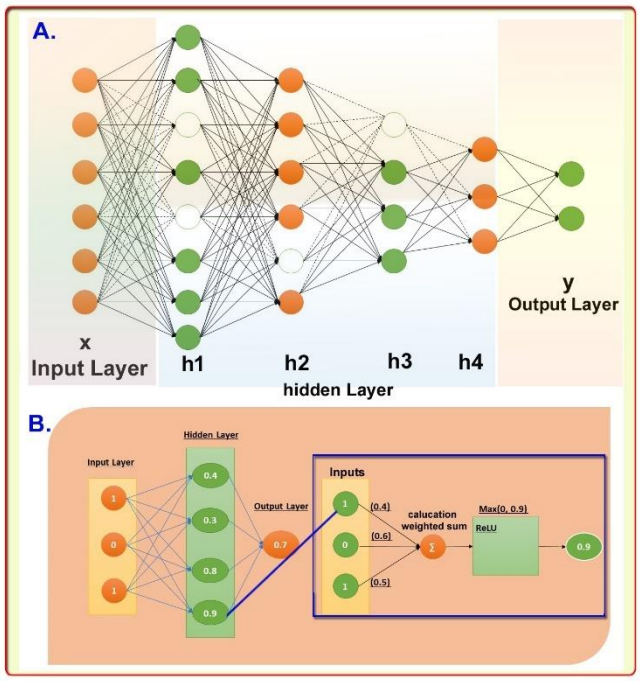

Figure 4. Proposed model

Figure 4. Architecture of the proposed DNNPSPDs model and preparation of the training dataset. **(A)** The DNN network structure, which comprises an input layer unit, four hidden layer units, and an output layer unit. **(B)** The functions of each hidden layer and a non-linear activation function that measures the output value.

$$ReLU(x)\begin{cases}0 & for\ \ x < 0\\ x & for\ \ x \geq 0\end{cases} \qquad (6)$$

derivate (with respective to x)

$$ReLU(x)\begin{cases}0 & for\ \ x < 0\\ 1 & for\ \ x \geq 0\end{cases} \qquad (7)$$

Activation function

DNNs have been widely studied due to neurons. A neuron generally accepted that a number that comes from the final branches of the neuron (pathway-specific proteins). What is happening to the neural network layer, we increased the input shape into a protein by the weight of that pathway-specific proteins and summarized all the neurons.

The role of ReLU Activation Function

Many studies have highlighted **ReLU** as the most used activation function in the world. Accordingly, the application of **ReLU** in neural networks and deep learning has been intensively examined. We apply the following **ReLU** function as the activation function of all neurons:

$$R(z) = max(0,\ z) \qquad (8)$$

The conducted to determine ReLU, as shown in figure 5, is half corrected (from the bottom), as you can see. ***R(z)*** is equal to ***0*** if z is below ***0*** and takes a value of z if z is equal to or above 0.Range: [***0*** to endlessness]

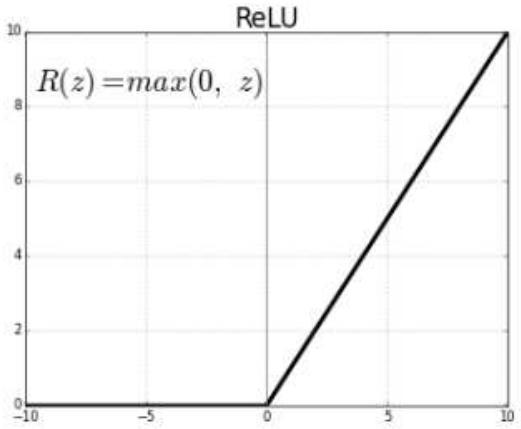

Figure 5. ReLU activation function

However, any negative value automatically becomes ***0***, thereby reducing the ability of the proposed model to suit or train correctly from the data. In other words, any negative input given to the **ReLU** function automatically transforms the value into ***0*** in the ***graph*** and, in turn, does not map the negative data.

For instance, if the weights are ***w1, w2, w3*** (Figure 4), an input layer, a***1, a2, a3,*** and ***wN*** inputs... We provide a summary of ***w1*a1 + w2*a2 + w3*a3*** .... Small * Small.

where ***R*** represents the activation function, ***w*** represents the connected weight matrix, $\boldsymbol{a^l}$ represents the input layer values for the ***z-th*** layer neuron output indicating class. In recent years the Models of Neural Networks (**NNs**) state-of-the-art performance language modeling efficiency and are now undergoing adoption on biological issues.

$$R(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}\ \ for\ j = 1, \ldots K. \qquad (9)$$

Mathematics is accompanied by softmax where z is the reference vector for output layer and ***j*** indexes ***1,2, 3*** ....*K*. Similar to a sigmoid function, the softmax function keeps the output value of each variable between **0** and **1** yet splits each output to **1** (Figure 4). In this case, the total output number is **1**. The output of the softmax function is equal to the categorical distribution of probabilities, which shows that each class is valid. The softmax function is mathematically expressed below, where z is a vector of output inputs (if we have 10 output units, then **10** elements will appear in z, which denotes the total number of output neurons in the softmax layer). The majority of studies in this area have used Keras and TensorFlow to enforce their models. The parameters used in the experiment are listed in Table 1.

Table 1. Parameters of the DNN model settings.

Parameter Name	P.Range	Recommendation
Learning rate	1, 0.1, 0.01, 0.001, 0.0001	0.01
Weight initialization	uniform, normal, glorot_normal, glorot_uniform,lecun_uniform, he_normal, he_uniform	glorot_normal
Per-parameter adaptive learning rate methods	SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam	Adam
Activation function	relu, tanh, sigmoid, softmax, softplus, softsign, hard_sigmoid	relu
Dropout rate	0.1, 0.2, 0.5, 0.8	0.5

In our experiment, cost function is a key aspect of lost function that denotes the distance between the expected and real values. We use cross-entropy as our cost function. Sigmoid functions are mostly used in shallow neural networks and require a low initialization power. Tanh functions are mostly used to address symmetry problems with two classifications, whereas **ReLU** is often used in deep learning. A sparse activation of neurons in a neural network is caused by a unilateral **ReLU** inhibition. We control the extraneous variables by using the **ReLU** activation function to ensure an easy operation and good learning ability.

Evaluation of method croos-validation

We utilize various cross-validation approaches to analyze the effects of our statistical parameter forecasts. Specifically, we adopt independent dataset testing, k-10-fold cross-valuation, and k-folding test. The jackknife test has been used progressively in previous research to check the accuracy of various predictors and to evaluate the forecasts of classifiers. In our analysis, we divide our dataset into 10 parts and subject them to a cross-validation experiment.

Computational tools for experiments

We use qualitative/quantitative approaches along with some machine learning tools, such as MATLAB R2018a, to evaluate and build partial algorithms. Specifically, we utilize MATLAB to select digital descriptors from a protein sequence and Weka for the classification. We then examine the output of different classifiers. The MATLAB method is used as a programming language in the fourth generation.

Result and performance evaluation

Performance evaluation of classifiers

Assume that $\boldsymbol{M}$ is a dataset that includes **N** samples, $\boldsymbol{Xi}$ is the feature space, and $\boldsymbol{Yi}$ denotes the settings of the target set, where $\boldsymbol{I \epsilon M}$**.** We measure precision as

$$\boldsymbol{Accuracy} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\boldsymbol{X_i} \cap \boldsymbol{Y_i}| \,/ (\boldsymbol{X_i} \cup \boldsymbol{Y_i}) \tag{10}$$

$$\boldsymbol{Sensitivity} = \boldsymbol{TP/TP + FN} \tag{11}$$

$$\boldsymbol{Specificity} = \boldsymbol{TN/FP + TN} \tag{12}$$

$$\boldsymbol{F - measure} = \mathbf{2} \times (\boldsymbol{Recall} \times \boldsymbol{Precision/Recall} + \boldsymbol{Precision}) \tag{13}$$

$$\boldsymbol{Precision} = \boldsymbol{TP/TP + FP} \tag{14}$$

$$\boldsymbol{Re\ call} = \boldsymbol{TP/TP + FN} \tag{15}$$

A real positive $\boldsymbol{TP}$ is an event in which the model accurately predicts the positive class, a true negative $\boldsymbol{TN}$ is an outcome where the model accurately predicts the negative class, and a false positive $\boldsymbol{FP}$ is an outcome where the model incorrectly predicts the positive class. False negatives, $\boldsymbol{FN}$, represent cases where the forecast is negative and the actual category is positive. The four parameters shown in Eqs. (6) to (10) are then determined as follows. Not more comfortable to know, especially the correlation coefficient of Mathew, and quiet.

$$\boldsymbol{Sn} = \mathbf{1} - \frac{\boldsymbol{N_-^+}}{\boldsymbol{N^-}} \tag{16}$$

$$\boldsymbol{Sp} = \mathbf{1} - \frac{\boldsymbol{N_+^-}}{\boldsymbol{N^-}} \tag{17}$$

$$\boldsymbol{Acc} = \mathbf{1} - \frac{\boldsymbol{N_-^+}}{\boldsymbol{N^+}} + \frac{\boldsymbol{N_+^-}}{\boldsymbol{N^-}} \tag{18}$$

$$\mathbf{Mcc} = \frac{\mathbf{1} - \left[\frac{\mathbf{N_-^+}}{\mathbf{N^+}} + \frac{\mathbf{N_+^-}}{\mathbf{N^-}}\right]}{\sqrt{\left[\mathbf{1} + \frac{\mathbf{N_+^-} + \mathbf{N_-^+}}{\mathbf{N^+}}\right]\left[\mathbf{1} + \frac{\mathbf{N_+^-} - \mathbf{N_-^+}}{\mathbf{N^-}}\right]}} \tag{19}$$

where Eqs. (16) to (19) measure flexibility, Matthew's similitude, accuracy, and characteristics. We calculate the correlation among these parameters to evaluate the Matthew's similitude. The number of true PSPD $\boldsymbol{N_-^+}$ proteins that have failed to identify as non-PSPD proteins, the amount of total estimated PSPD $\boldsymbol{N^+}$ protéins, $\boldsymbol{N - th}$ total number of checked nonPSPD proteins and the a mount of failed PSPD proteins, $\boldsymbol{N_+^-}$ And the number of non-PSPD proteins incorrectly identified as PSPD proteins.

Impact of extraction algorithm

The proposed DNNPSPDs model demonstrates an excellent performance in formulating the protein sequences PseAAC DPC ,AAindex, AAC, and PAAC. We use datasets with the same data points in various classification tasks and perform principal component analysis to condense the hybrid feature space. This approach demonstrates the usefulness of the individual and hybrid feature spaces of different classifications. We use DNN in all analyses, while PseAAC DPC ,AAindex, AAC, and PAAC are used to determine the optimal parameters that have critical effects on the model development. We also conduct a content analysis to calculate the values of various parameters, and we use PSPDs predictability as a metric to determine those parameters with

the correct features. Given that the testing dataset samples include the synthetic amino acid ***O***, we set the value of ***l*** to 1, 2, 3, 4, 5, and 6. The optimum parameter values of DNNPSPDs and the other models are shown in Table 2.

Table 2. DNN Identifying optimum parameter for various models

Proposed model	ACC	Precision	npv	Sensitivity	Specificity	MCC	F1	ROC-Auc
Pse_AAC	93.60%	95.46%	93.49%	92.12%	94.99%	88.02%	93.26%	0.982%
DPC	95.72%	96.11%	96.25%	95.53%	95.83%	91.86%	95.57%	0.981%
AAindex	87.13%	86.35%	89.97%	91.36%	83.18%	75.35%	88.19%	0.965%
AAC	81.82%	85.51%	86.87%	82.95%	80.83%	67.65%	80.45%	0.942%
PAAC	69.36%	663.6%	77.48%	71.21%	67.65%	41.58%	66.29%	0.814%

Comparison of predicted ROC-AUC score

Although multi-information fusion improves the prediction efficiency of a model to some degree, this approach also provides redundant feature information, thereby affecting the model classification accuracy and reducing the calculation speed. We compare the average prediction, precision, and measurement performance of DDNSPSDs with those of other models. Figure 6 shows the ROC–AUC curves that correspond to various PSPDs datasets. DDNSPSDs achieves the highest ROC–AUC score (0.983), followed by DPC (0.982), PseAAC (0.983), AAindex (0.965), AAC (0.943), and PAAC (0.815). The ROC–AUC curves are shown in Figure 6.

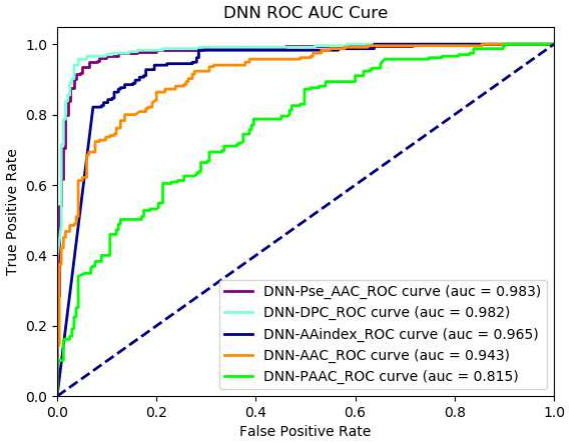

Figure. 6. Proposed model ROC-auc score result.

Identifying a comparison of combined features extraction approaches

In specific, the ROC Auc curves corresponding to the DNNPSPDs datasets, which contain the ROC cure corresponding to the 5 five feature extraction approaches, than AUC cure values are , AAC(auc=0.91), AAindex(auc=0.93) DPC(acu=0.98), PAAC(auc=0.77) and Pse_AAC(auc=0.98). According to these ROC-auc performances of DNN (Pse_AAC, DPC) better than other approaches are shown in figure 7 in supplymentry material (Appendix. A)

Comparison of proposed method with other ML approaches.

Figure 8 demonstrates that intuitively. The ROC-AUC curve used by the DNNPSPD process includes ROC curves relating to other methods and AUC values are accuracy achieved,9360%, the precision achieved 95.46%, npv achieved score 93.49, Sensitivity,92.12%, specificity,94.99, MCC achieved score 88.02%, F1-measure 93.26% and ROC AUC. 98.28% as shown in figure 8 in supplementary mater-ail (Appendix. A). The results for our proposed model DNN and with combining Ten 10 classification systems such as, KNN classifier (KNN)[39-42], SVC classifier (SVC)[43,44], Decision Tree Classifier (DTC)[45], Random Forest Classifier (RFC)[45,46,47], AdaBoost classifier (ABC)[48], Linear Discriminant Analysis (LDA)[49], Gradient Boosting Classifier (GBC)[50], Gaussian NB classifier (GNB)[51], MLP Classifier (MLPC)[52], and Extra Trees Classifier (ETC)[53], with AAC spaces function, shown figure 8 in supplementary mater- ail (Appendix. A).

Classification performance with AAC feature space.

The results for our proposed model DNN model and with combining Ten 10 classification systems such as AdaBoost classifier (ABC), KNN classifier (KNN), SVC classifier (SVC), Linear Discriminant Analysis (LDA), Gradient Boosting Classifier (GBC), Random Forest Classifier (RFC), Gaussian NB classifier (GNB), Decision Tree Classifier (DTC), MLP Classifier (MLPC), and Extra Trees Classifier (ETC), with AAC spaces function, shown in table 3. in supplementary materail (Appendix. B). Figure. 9. Performance of various classifiers in predicting pathway-specific protein domain based on **AAC feature encoding method**. Areas under the Precision-Recall curves (AUPRC) indicate that AdaBoost classifier(ABC) outperformed KNN classifier(KNN), SVC classifier(SVC), Linear Discriminant Analysis(LDA),Gradient Boosting Classifier(GBC),Random Forest Classifier(RFC),Gaussian NB classifier(GNB), Decision Tree Classifier(DTC), Multiple Layer Perceptron (MLP) and (ETC) classifiers.

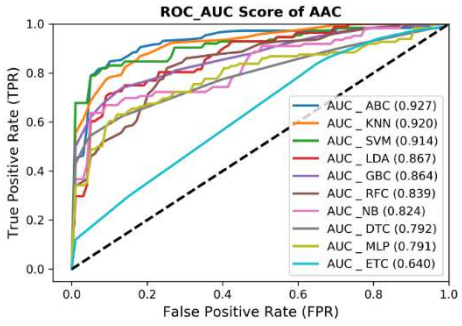

Figure 9. Combine compared ACC model predicted data.

Classification performance of DPC model

In the Table. 4 in supplementary material (Appendix. B). compared DPC with the 10 classifiers in terms of several parameters, including AUC, accuracy, precision and score of 85.22%, and then we presented ROC-AUC achieved score with DPC as shown in figure 10. Figure 10 clearly shows that DPC outperforms all classifiers with a ROC–AUC score of 0.939.

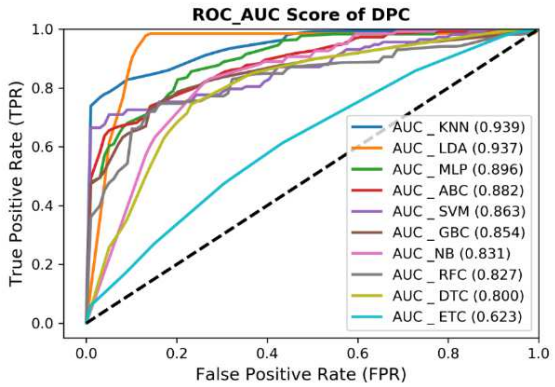

Figure 10. Combine compared DPC model predicted data.

Performance of classifier using (AAC+PseAAC)

PseAAC involves AAC principles and sequence correlation variables. Table 5 in supplementary material (Appendix. B). presents the expected effects of those classifiers that use ABC and PseAAC hybrid feature spaces. ABC, LDA, KNN, and NB achieve accuracies of 0.915% 0.879% (86.82%), 85.58% (78.38%), and 84.31% (54.26%) when using AAC (PseAAC), respectively. Although we investigate 10 classifiers in this work, we only focus on the accuracies of the 3 aforementioned classifiers, all of which have achieved higher accuracies by using ABC and LDA instead of PseAAC as shown in Figure 11. reports the results of classifiers that use PseAAC extraction feature spaces. DNN achieves the best accuracy of 93.60%, followed by ABC.

Classification performance of PseAAC, ROC-AUC

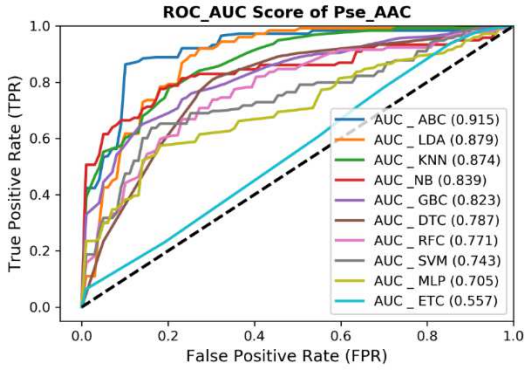

Figure 11. Combine compared Pse_ACC model ROC-AUC

Performance of classifer using (AAC+DPC)

In Table 6 in supplementary material (Appendix. B). presents the accuracies of those classifiers that use AAC and DPC. DNN, ABC, KNN, SVC, and LDA achieve accuracies of 87.3% (95.72%), 87.30% (85.99%), 84.31% (69.89%), and 77.98% (90.66%) when using AAC (DPC), respectively. Most of these classifiers achieve higher accuracy when using DPC instead of AAC.

Performance of classifiers using (DPC+PseAAC)

As shown in Table 6, in supplementary material (Appendix. B). PseAAC yields excellent classification results for ABC, which achieves 86.82%, 80.00%, 86.36%, 86.96%, 80.00%, and 86.74% accuracy, sensitivity, specificity, precision, recall, and F1-measure, respectively. Meanwhile, LDA achieves an accuracy, sensitivity, specificity, precision, recall, and F1-measure of 82.62%, 96.00%, 63.64%, 75.00%, 96.00%, and 83.27%, respectively. However, when using DPC, ABC only achieves an accuracy of 81.37%, which is lower than that achieved by using PseAAC.

Quality of classifiers selected for minimum-redundancy-maximum significance (mRMR)

K-5-fold cross-validation is one of the most popular methods for model evaluation and model selection in the field of machine learning. The central concept of cross-validation is that any finding is checked in our dataset. K-5-fold cross-validation is a unique cross-validation case where k iterations are performed over a dataset. In each round, the dataset is divided into k parts, where one part is used for validation and the other k−1 parts are fused into a model assessment training subset. In K-10-fold cross-validation, 9 sets are used to prepare the training sets whereas the 1 remaining set is used for practice or testing. This process is repeated 10 times. The manipulation precision and Matthew's correlation coefficient (MCC) can be used to test the output of different modules. We also assess the efficiency of classifiers in compressing the function (feature) space. Table 4 shows the performance of these classifiers in a compressed function space. When DPC and PseAAC are used as extraction models, DNN achieves the best accuracies of 95.72% and 93.60%, respectively. Meanwhile, ABC achieves the best accuracy (87.30%), sensitivity (96.00%), specificity (81.82%), precision (85.71%), recall (96.00%), and F1-measure (86.77%) when using AAC. Similarly, ABC achieves the best accuracy of (86.82%), sensitivity (76.00%), specificity, (81.82%), precision (82.61%), recall (76.00%), and F1-measure (85.22%) when using PseAAC. All used classifiers output on individual and mixed-function spaces after empirical evaluation. In sum, using the AAC and PseAAC hybrid feature yields promising results for ABC. We also reduce the function space by implementing the mRMR selection technology, which negatively affects the performance of classifiers by removing some essential features of the space.

Specifically, these classifiers show better outputs in the original feature space than in the compressed feature space after an empirical evaluation. It was therefore decided that

the achievement was due to the PC and PseAAC mixed feature area, as well as AdaBoost capacity for discrimination.

Performance comparison of our model with existing models.

Table 7 compares the performance of DNNSPSDs with some extant classification methods. Jung et al.[54] proposed the ECMPP method[54], where five pathway–protein characteristics are used for the classification. This approach yields a precision of 85.71%, sensitivity of 96.00%, and specificity of 81.82%[54]. Meanwhile, the PSPD model introduced in[55], which bases its predictions on 10 classifiers with PSSM spaces, achieves a precision and sensitivity of 96.00% and 81.82%, respectively. Yang et al. proposed the IECMP model[56], which predicts PSPD proteins with a mixed-characteristic set classification and achieves precision, sensitivity, and specificity of 86.40%, 87.80%, and 88.67%, respectively. The proposed DNNSPSDs model outperforms the IECP model by 10.36% in terms of precision[56]. In addition, the computer model described in this paper will be provided with a web server.

Table 7. Performance comparison of different prediction models.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
ECMPP [54]	77.80	56.30	95.60	
IECMP [56]	86.40	87.80	84.90	
DPP-PseAAC [57]	77.42	83.87	70.97	55.30
Proposed method	95.72	**95.53**	95.83	91.86

Performance comparison of different classifiers

Among the examined classifiers, DNN achieves the best accuracies of 95.72% and 93.61% when using DPC and PseAAC, respectively. ABC shows the second-highest accuracy of 85.99% when using DPC, followed by KNN (85.99%). Overall, ABC emerges as the best classifier, followed by KNN and LDA (with 90.66% accuracy) when using DPC as shown in Figure. 12 in supplimentery material (Appendix A).

Performance comparison of (AAC, DPC and PseAAC model)

We adopt qualitative/quantitative techniques to analyze the extraction of PSPD functional annotation features. Feature engineering is an important step in the application of machine learning methods. Given that AAC is the most popular feature of PSPDs, we use PAAC, AAIndex, AAC, DPC, and PseAAC to determine which model performs best in terms of features construction as shown in Figure.13 in supplimentery material (Appendix A).

Case Study

Qualitative/quantitative techniques are also used to analyze 31 proteins–WNT1-inducible-signaling pathway protein 2 isoform 1 precursor, which contains was [Homo sapiens]. We analyze the relationship between the proteins and pathways by using the three aforementioned feature extraction models and then use the 10 classifiers to test the accuracy of these models. The analysis is conducted based on the 4 proteins–CCN6_human cellular communication network factor, 9-G
protein pathway suppressor 2, 2-proteins TIP41, TOR signaling pathway regulator-like, 10-epidermal growth factor receptor pathway substrate 8, and 6-disease-pathway association 6 proteins as the name of tissue factor pathway inhibitor isoform a precursor as shown figure 14.

WNT1 inducible signalling pathway protein sequence

We present a protein-based pathway specificity prediction for protein domains to be used for classifying domain-specific pathways. AAH74841.1 protein Wnt / Frizzled signaling pathway downstream regulator[58]. We presented the top 10 protein WNT1 inducible signaling pathway protein sequence and database entry id evidence, as shown in table 8 and Cell survival linked. Attenuates apoptosis p53-mediated by the activation of AKT kinase in response to damage to DNA. The anti-protein Bcl-) is upregulated. Numerous cancers, including breast and colon tumors, demonstrate overexpression. In addition to fundamentally modifying cells, malignant cells show soluble microenvironmental signals, including WNT1 (WISP1), which is a secreted matricellular protein that increases the number of cancers and has been associated with reduced survival rates[59]. We collect these proteins from the Signor 2.0 database name of (The signaling network open resource and evidence of pathway map taken from https://signor.uniroma2.it/. As shown in figure 14.

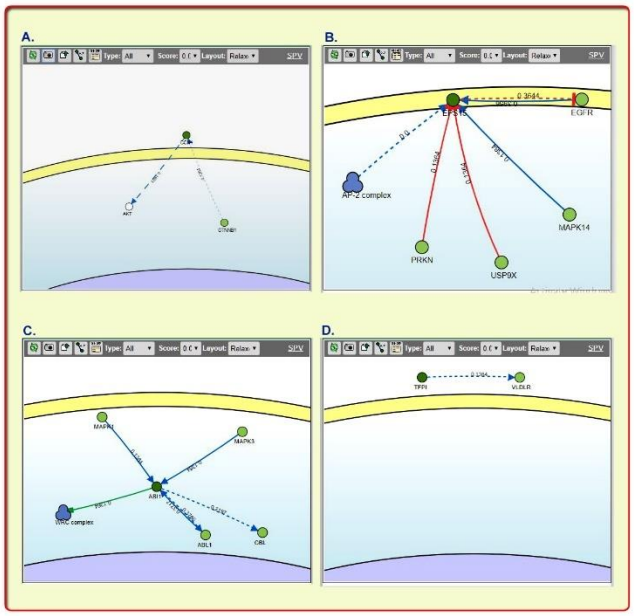

Figure 14. Proteins functions involve in pathways

Figure 14. (A) WNT1/frizzled-signaling pathway downstream regulator that is linked to cell survival, attenuates p53-mediated apoptosis due to DNA damage by

activating AKT kinase, upregulates the Bcl-X(L) anti-apoptotic protein, supports skin and melanoma fibroblasts, and facilitates the binding of proteoglycans, decorin, and bigly to skin fibroblasts in vitro. (B) Controlling cell growth, regulating mitogenic signals, managing cell proliferation, and internalizing receptor tyrosine kinase (RTK)-type ligand-inducible receptors, especially EGFR, which plays a role in clathrin-coated pit (CCP) assembly, acts as a clathrin adapter for post-Golgi trafficking, and is involved in the maturation, invagination, or budding of CCPs, endocytosis of integrin beta-1 (ITGB1) and transferrin receptor (TFR), and internalization of ITGB1 as a DAB2-dependent cargo (which, in turn, requires DAB2). (C) Interaction with non-receptor tyrosine kinases ABL1 and/or ABL2 in the negative regulation of cell growth and transformation, EGF-induced pathway activation regulation, cytoskeletal reorganization, and signaling of EGFR. Together with EPS8, these functions participate in the transfer of signals from Ras to Rac. The ABI1, EPS8, and SOS1 trimeric complexes exhibit a Rac-specific guanine nucleotide exchange factor (GEF) activity in vitro, and ABI1 tends to be an adapter in the group. These functions also include ENAH-ABL1/c-Abl-mediated phosphorylation, recruitment of WASF1 to lamellipodia, controlling the WASF1 protein level, controlling dendritic outgrowth and branching in the brain, and determining the form and amount of synaptic neuron contacts. (D) Direct inhibition of factor X (X(a)) and VIIa/tissue factor activity, possibly by forming a quaternary Xa/LACI/VIIa/TF complex, which has an antithrombotic function and potential to interact with plasma lipoproteins.

Table 8. WNT1 inducible signaling pathway protein sequence and database entry id evidence

Protein ID	Protein Sequence Name	UniProtKB ID	Length
AAI07739.1	G protein pathway suppressor 2	Q13227	327
AAI03904.1	G protein pathway suppressor 2	Q13227	327
AAI03902.1	G protein pathway suppressor 2	Q13227	327
AAH13652.1	G protein pathway suppressor 2	Q13227	327
AAH00155.3	G protein pathway suppressor 1	Q13098	491
AAH64503.1	G protein pathway suppressor 1	Q13098	491
NP_004480.1	G protein pathway suppressor 2	Q13227	327
hsa:2873	GPS1; G protein pathway suppressor 1	Q13098	491
hsa:2874	GPS2; G protein pathway suppressor 2	Q13227	327
sp|Q13227.3|	GPS2: G protein pathway suppressor 2		

GPS2 protein pathway suppressor 2[human]

The GPS G-protein pathway protein 2 is encoded by gene GPS2 in humans. Table 9 lists the top 10 GPS2 G-proteins pathway suppressor 2 in humans. GPS2 codes a protein that participates in the cascade signaling of G protein-mitogen-activated protein kinase (MAPK), may effectively suppress a signal mediated by RAS and MAPK when over-expressed in mammalian cells, and interfere with JNK activity, which suggests that signal replacement may be a function of this gene[60]. GPS2 also functions as a B-cell production regulator by inhibiting UBE2N/Ubc13, thereby reducing activation by related (B) pathways for signaling Toll-like (TLR) and B-cell antigen receptors (BCRs). Action as the main mediator for mitochondrial stress reaction relocates to the nucleus following desumoylation and promotes specifically the expression of nuclear-encrypted mitochondrial genes in response to depolarization[61].

Tissue factor pathway inhibitor protein sequence

Plasmin-mediated matrix reshaping control may prevent the formation of trypsin, plasmin, factor VIIa, and tissue factor Xa and have no effects on thrombin. Table 10 presents the top 10 matrix protein-tissue factor pathway inhibitor 2 isoform 1 precursors[62]. Serine proteinase inhibitors play an essential function in the combination of tissue turnovers. In this analysis, trypsin/elastase/plasmin inhibitors of the extracellular matrix of the human skin-

Table 9. GPS2 protein sequence and database entry IDs

Protein ID	Protein Sequence Name	UniProtKB ID	Length
AAI07739.1	G protein pathway suppressor 2	Q13227	327
AAI03904.1	G protein pathway suppressor 2	Q13227	327
AAI03902.1	G protein pathway suppressor 2	Q13227	327
AAH13652.1	G protein pathway suppressor 2	Q13227	327
AAH00155.3	G protein pathway suppressor 1	Q13098	491
AAH64503.1	G protein pathway suppressor 1	Q13098	491
NP_004480.1	G protein pathway suppressor 2	Q13227	327
hsa:2873	GPS1; G protein pathway suppressor 1	Q13098	491
hsa:2874	GPS2; G protein pathway suppressor 2	Q13227	327
sp|Q13227.3|	GPS2: G protein pathway suppressor 2		

transformed fibroblasts are isolated and determined in the partially amino-terminal amino acid sequence. Substrate reverse zymography tracks the antitrypsin activity of these inhibitors. The amino acid sequence homology of the 31-kDa inhibitor has been proven to be novel by a computer. Meanwhile, the 33-kDa inhibitor sequence is 70% to 90% similar to an amino-terminal sequence known as the 32-kDa inhibitor of the tissues factor or tissue factor pathway inhibitor-2.

Epidermal growth factor receptor protein sequence

Adapter to regulate actin cytoskeleton dynamics and architecture controlling many cellular protrusions. Different processes may be controlled depending on their relationship with other signal transducers[63]. These processes include the axonal production of philopodia, stereocidal volume, dendritic migration of cells, and migration and invasion of cancer cells.

Table 10. Tissue factor pathway inhibitor protein sequence and database entry IDs

Protein ID	Protein Sequence Name	Length
AAH05330.1	Tissue factor pathway inhibitor	235
AAH15514.1	Tissue factor pathway inhibitor	304
NP_006278.1	tissue factor pathway inhibitor	304
NP_001316168.1	tissue factor pathway inhibitor	304
NP_001316170.1	tissue factor pathway inhibitor	304
NP_001316169.1	tissue factor pathway inhibitor	304
NP_001305870.1	tissue factor pathway inhibitor	304
NP_001027452.1	tissue factor pathway inhibitor	304
hsa:7035	tissue factor pathway inhibitor	304
hsa:7980	tissue factor pathway inhibitor	235

We analyze the top 10 epidermal growth factor receptor pathways shown in Table 11. Component of a WHRN and MYO15A complex located at stereo-types and needed to elongate the stereo-actin center. Cell cycle degradation is required during the G2 phase to sustain the changes in cell structure[64]. With its active barbed finish activity and ability to modulate Rac activity, Eps8 is involved in actin dynamics. In addition, IRSp53 is bound to Eps8. Here's a preview of Eps8's novel actin interconnect.

Table 11. Factor receptor protein sequence and database entry IDs

Protein ID	Protein Sequence Name	UniProtKB ID	Length
AAH30010.1	epidermal growth factor receptor pathway	Q12929	822
EAW84545.1	epidermal growth factor receptor pathway		
EAW84544.1	epidermal growth factor receptor pathway		
hsa:58513	epidermal growth factor receptor pathway	Q9UBC2	864
hsa:2060	epidermal growth factor receptor pathway	P42566	896
hsa:2059	epidermal growth factor receptor pathway	Q12929	822
sp|Q8TE67.2|	epidermal growth factor receptor pathway		
sp|Q9H6S3.2|	epidermal growth factor receptor kinase		
sp|Q8TE68.3|	epidermal growth factor receptor kinase		
sp|Q9H6S3.2|	epidermal growth factor receptor kinase		

Future Direction

Machine learning is a valuable tool for modeling the interaction between protein structures and features that are derived from human pathways or multiple biological data sources. The algorithmic approaches for feasibility concern relevant to single task networks will be useful for potential innovations. We may build and evaluate a unique task protein feature prediction DNN-based method based on these solutions. Although DNNPSPDs can improve pathway-specific protein prediction accuracy and precision to some extent, its predictability and algorithm efficiency require further improvements. We shall also attempt to build our fundamental knowledge on proteins to produce more successful hidden features, to take biological meaning into account, and to incorporate specific effective algorithms, such as convolutional neural networks[65], capsule networks[66], and generative opponent networks. The versatility of our work contributes to the advancements in protein function analysis and association predictions with human pathways based on convolutional neural networks and several specific biological data sources, such as the graph embedding features or pathway involvement of protein–protein networks. Future studies should attempt to identify the best way of building a concept for our web-based software in addition to utilizing expanded vocabulary and different datasets.

Discussion

To the best of our understanding, the application of deep learning algorithms in predicting functional large-scale protein pipelines has not extensively examined in the literature. Moreover, previous experiments have only focused on small protein sets and functional groups. In these experiments, the application of DNNs has been extended to predict the protein functions of various forms, such as amino acid sequences[67], 3D structural properties, non-protein networks, molecular and functional aspects, and specific DNN computational feed-forward architectures (i.e., single- or multi-task feed-forward DNNs, recurrent neural networks, deep auto-encoder neural networks, and profound re-task)[68].

The technical complex research methods that restrict the size of the input data and number of integrable functional classes present a key barrier in designing realistic DNN prediction devices. Given this limitation, previous studies have only focused on few protein families. Therefore, new analytical methods with high efficiency and applicability in real scenarios should be proposed to facilitate in vitro research on protein-pathway recognition.

In this work, we propose DNNPSPDs as a novel hierarchical multitasking deep learning approach for predicting protein term interactions from protein sequence records. We also conduct a robust DNN-based predictive model characteristics analysis. This work is among the first to use DNNs in predicting sequence-based pathway specific protein functions and contributes to the literature by creating a broad-based deep learning predictive framework with a 1 multi-task stack and 101 feed-forward DNNs that can predict thousands of functional concepts based on a protein or gene. We also examine the forecasts of pathway-specific training instances, which present a major problem in the field of automated protein function prediction, by proposing an approach that enhances the quality of automated functional predictions that have been previously developed through machine testing.

Conclusion

PSPD functional annotation is an important challenge in the genomic era. The most widely used feature selection technologies, such as AAindex, AAC, DPC, PAAC, and PseAAC, are used for feature extraction models to express protein pathway sequences. We adopt several analytical methods to predict the function of novel pathway–proteins associations. However, PSPDs profiles are complex models with many free parameters. The difficulty of controlling external parameters lies in setting position-specific residue scores and combining a structure with

multiple sequence information. Our proposed method shows a higher accuracy compared with the other extant methods proposed in the literature. This study processes the data for protein–pathway association datasets and then trains and tests 10 machine learning models. Our proposed model achieves the highest prediction accuracy of 0.957 at an MCC of 91.86%, followed by DPC (0.936 at an MCC of 88.02%). In addition, DNNPSPDs achieves an ROC–AUC score of 0.982 by using PseAAC and 0.981 by using DPC. This model also has higher accuracy than other classifiers, including LDA, GBC, RFC, GNBC, DTC, MLP, ETC, and DNN. All methods have randomly predicted 115 pathway–protein associations for carbon metabolism, lipid, energy, and non-standard amino acids. These pathway–protein associations are preserved by strong co-inheritance patterns in genetic information processing and may be linked to one another via physical cell interactions.

Author Contributions

Xiujuan Lei, Zhenqiang Wu, Ali Ghulam, and Yuchen Zhang, jointly contributed to the design of the study. Xiujuan Lei conceptualized the review and finalized the manuscript. Ali Ghulam wrote the initial manuscript. Yuchen Zhang helped to design method and draft the manuscript. Zhenqiang Wu revised the manuscript and polished the expression of English. All of the authors have read and approved the final manuscript.

Acknowledgements

This work was supported by the funding from National Nat ural Science Foundation of China (61972451, 61672334, 6 1902230) and the Fundamental Research Funds for the Cen tral Universities (No. GK201901010).

Availability of data and materils

In regard to data availability, the following information was provided: This code is based on python sklearn packages, which implement the majority of tasks and workflows mentioned in this contribution. In the public PDB database you can find the raw data: www.rcsb.org/pdb/. We provide a remove similarity of proteins sequence based on third-party tools have CD-HIT web tools been used for this purpose. The simulated evolutionary scenarios used throughout this URL is available as ((http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit), (CD-HIT web tools.) The used all datasets can be found here: NCBI database (https://www.ncbi.nlm.nih.gov/), and https://www.ncbi.nlm.nih.gov/protein/ database and then also verified and compared with (https://www.uniprot.org/uniprot/) Uniport database. Further, the code used to support the findings of this study are available from the corresponding author upon request.

References

1. Bebek, G., Yang, J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics 8, 335 (2007). https://doi.org/10.1186/1471-2105-8-335 .
2. Boudellioua I, Saidi R, Hoehndorf R, Martin MJ, Solovyev V . *PLOS ONE* 11(7): e0158896. https://doi.org/10.1371/journal.pone.0158896-002
3. Valente GT, Acencio ML, Martins C, Lemke N (2013) The Development of a Universal In Silico Predictor of Protein-Protein Interactions. PLOS ONE 8(5): e65587. https://doi.org/10.1371/journal.pone.0065587 .
4. Suraj Peri. et al. genome.cshlp.org on January 4, 2020 - *Published by Cold Spring Harbor Laboratory Press*. -0004
5. Boudellioua I, et at. PLoS ONE 11(7): e0158896. http://doi:10.1371/journal.pone.0158896-0005
6. Karin, M. and Ben-Neriah, Y. 2000. Annu. Rev. Immunol. 18: 621-663. -6 https://doi.org/10.1146/annurev.immunol.18.1.621 PMid:10837071
7. Pawson, T. and Nash, P. 2003.Science 300: 445-452. -7 https://doi.org/10.1126/science.1083653 PMid:12702867
8. Nakai, K. 2000. Protein Chem. 54: 277-344.-8 https://doi.org/10.1016/S0065-3233(00)54009-1
9. Hanash, S. 2003. Nature 422: 226-232. -9 https://doi.org/10.1038/nature01514 PMid:12634796
10. Pruitt, Kim D. , et al. Trends in Genetics 16.1(2000):44-47. -10 https://doi.org/10.1016/S0168-9525(99)01882-X
11. Brigitte Boeckmann 等. *C R Biol* 328.10-11:0-899. -11
12. Wang, H., Feng, L., Zhang, Z., Webb, G. I., Lin, D., and Song, J. (2016). Sci. Rep. 6:21383. doi: 10.1038/srep21383 -12 https://doi.org/10.1038/srep21383 PMid:26906024 PMCid:PMC4764925
13. Wang, M., Zhao, X. M., Takemoto, K., Xu, H., Li, Y., Akutsu, T., et al. (2012). PLoS ONE 7: e43847. doi: 10.1371/journal.pone.0043847 -13 https://doi.org/10.1371/journal.pone.0043847 PMid:22937107 PMCid:PMC3427247
14. Lin, H., Liu, W. X., He, J., Liu, X. H., Ding, H., and Chen, W. (2015). Sci Rep. 5:16964. doi: 10.1038/srep16964 -14 https://doi.org/10.1038/srep16964 PMid:26648527 PMCid:PMC4673586
15. Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016). *Oncotarget* 7, 44310–44321. doi: 10.18632/oncotarget.10027. -15
16. Tang, H., Su, Z. D., Wei, H. H., Chen, W., and Lin, H. (2016). Biophys. *Res. Commun.* 477, 150–154. doi: 10.1016/j.bbrc.2016.06.035 -16
17. Gupta, S.,Mittal, P.,Madhu,M. K., and Sharma, V. K. (2017). *Front. Immunol*. 8:1430. doi: 10.3389/fimmu.2017.01430 -17
18. Manavalan, B., and Lee, J. (2017). *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222. -18
19. Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, *B., et al.* (2017). Sci. Rep. 7:6862. http://doi: 10.1038/s41598- 017-07199-4-19
20. Hinton, G. et al. IEEE Signal Process. Mag. 82–97, https://doi.org/10.1109/MSP.2012.2205597 (2012).
21. Min, S., Lee, B. & Yoon, *S.Brief. Bioinform.* 18, 851–869 (2016).
22. Kandaswamy, K.K., Pugalenthi,., Kalies,K.U.,Hartmann,E.,Martinetz,T.,2013.J.Theor. Biol. 317,377-383. https://doi.org/10.1016/j.jtbi.2012.10.015 PMid:23123454.
23. Li, Z.R., et al., Nucleic Acids Res, 2006. 34(Web Server issue): p. W32-7. https://doi.org/10.1093/nar/gkl305 PMid:16845018 PMCid:PMC1538821.
24. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M. et al. Sci Rep 9, 7344 (2019). https://doi.org/10.1038/s41598-019-43708-3
25. Thusberg, J., Olatubosun, A., Vihinen, M., 2011.Hum. Mutat.32,358-368. https://doi.org/10.1002/humu.21445 PMid:21412949.
26. Kalita, M.K., Nandal, U.K., Pattnaik, A., Sivalingam, A., Ramasamy, G., Kumar, M., Raghava, G.P.,Gupta,D.,2008.PLoS One 3,e2605_1-e2605_12. https://doi.org/10.1371/journal.pone.0002605 PMid:18596929 PMCid:PMC2435623.

27. Gupta, S., Ansari, H.R., Gautam, A., Raghava, G.,2013. Biol.Direct8,27. https://doi.org/10.1186/1745-6150-8-27 PMid:24168386 PMCid:PMC3831251.
28. Chou, K.-C., 2011.J. Theor.Biol.273,236-247. https://doi.org/10.1016/j.jtbi.2010.12.024 PMid:21168420 PMCid:PMC7125570.
29. Ali, F., Hayat, M., 2015.J. Theor.Biol.384,78-83. https://doi.org/10.1016/j.jtbi.2015.07.034 PMid:26297889 .
30. Sun,X.Y.,Shi,S.P.,Qiu,J.D.,Suo,S.B.,Huang,S.Y.,Liang,R.P., 2012. Mol.Biosyst.8,3178-3184. https://doi.org/10.1039/c2mb25280e PMid:22990717.
31. Mei, S.,2012.J. Theor.Biol.310,80-87. https://doi.org/10.1016/j.jtbi.2012.06.028 PMid:22750634 .
32. Du, P., Wang,X.,Xu,C.,Gao,Y.,2012. Anal.Biochem.425,117-119. https://doi.org/10.1016/j.ab.2012.03.015 PMid:22459120.
33. Chan, J.F.,Lau,S.K.,To,K.K.,Cheng,V.C.,Woo,P.C.,Yuen,K.-Y.,2015. Microbiol. Rev.28,465-522. https://doi.org/10.1128/CMR.00102-14 PMid:25810418 PMCid:PMC4402954.
34. Gui, Y., Wang, R., Wei, Y., & Wang, X. (2019). Journal Of Biological Systems, 1–18. http://doi:10.1142/S0218339019500013 Url To Share This Paper: Sci-Hub.Tw/10.1142/S0218339019500013
35. Glorot X, Bordes A, Bengio Y. International conference on artificial intelligence & statistics, 2011; 15: 315-23.- ReLU-34
36. Wang X, Wu YJ, Wang RJ, et al. Plos one. 2016; 14(6): e0217312-adam-35
37. Hinton GE, Srivastava N, A Krizhevsky, I Sutskever, RR Salakhutdinov. Computer science. 2012; 3 (4):212-23.- dropout-36
38. Xie, Zengyan, Xiaoya Deng, and Kunxian Shu. International Journal of Molecular Sciences 21.2 (2020): 467.-37
39. Akkus, A.,Guvenir,H.A.,1995. Proc.ICML96,12–19.
40. Hayat, M.,Khan,A.,2011.J.Theor.Biol. 271,10-17. https://doi.org/10.1016/j.jtbi.2010.11.017 PMid:21110985 .
41. Zahoor, J.,Abrar,M.,Hussain,D.,2008.Springer-Verlag,BerlinHeidelberg,pp. 40–51.
42. Hayat, M.,Khan,A.,2012.IETCommun.6, 3257-3264 https://doi.org/10.1049/iet-com.2011.0170
43. Bhasin, M., Raghava,G.,2004. Nucleic Acids Res.32,W414–W419
44. Mandle, A.K.,Jain,P.,Shrivastava, S.K.,2012. Int.J.Soft Comput.3,67–78.
45. Duda, R.O., Hart,P.E.,Stork,D.G.,2012.John Wiley & Sons, California.
46. Liaw A, Wiener M. R News 2001;23.
47. Fern N-DM, Cernadas E, et al. J Mach Learn Res 2014;15:3133–81.
48. Meir, R., Rätsch, G., 2003.Springer, pp.118–183.
49. Nour, Majid, and Kemal Polat. Mathematical Problems in Engineering 2020 (2020).
50. S. Raschka and V. Mirjalili, 2nd Edition, 2nd ed. Packt Publishing Ltd, 2017.
51. Soni, J., Ansari,U.,Sharma,D.,Soni,S.,2011.Int.J.Comput.Appl.17, 43–48.
52. B. Çarklı Yavuz, N. Yurtay and O. Ozkan, IEEE Access, vol. 6, pp. 45256-45261, 2018.
53. Yonasi, S., & Singh, Y. (2018). Predicting Cellular Protein localization Sites on Ecoli's Minimal Dataset using a Comparison of Machine Learning Techniques.
54. Jung, J., Ryu, T., Hwang, Y., Lee, E., Lee, D., 2010.J. Comput. Biol. 17,97–105.
55. Anitha, J., Rejimoan, R., Sivakumar,K.C.,Sathish,M.,2012. IJCA Spec.Issue Adv.Comput. Commun. Technol. HPC Appl.1,7–11.
56. Yang, R., Zhang,C.,Gao,R.,Zhang,L.,2015. PLoSOne 10,1–21.
57. Rahman, M Saifur , et al. Journal of Theoretical Biology 452(2018).
58. Xu L, Corcoran RB, Welsh JW, Pennica D, Levine AJ. Genes Dev. 2000;14(5):585–595.
59. Su F, Overholtzer M, Besser D, Levine AJ.Genes Dev. 2002;16(1):46–57. http://doi:10.1101/gad.942902
60. Spain BH, Bowdish KS, Pacal AR, Staub SF, Koo D, Chang CY, Xie W, Colicelli J (Dec 1996). Molecular and Cellular Biology. 16 (12): 6698–706. http://doi:10.1128/mcb.16.12.6698 PMC 231672. PMID 8943324.
61. Cardamone MD, Tanasa B, Cederquist CT, et al. Mol Cell. 2018;69(5):757–772.e7. http://doi:10.1016/j.molcel.2018.01.037
62. Rao CN, Liu YY, Peavey CL, Woodley DT. Arch Biochem Biophys. 1995;317(1):311–314. http://doi:10.1006/abbi.1995.1168
63. Disanza A, Carlier MF, Stradal TE, et al. Nat Cell Biol. 2004;6(12):1180–1188. http://doi:10.1038/ncb1199
64. Disanza A, Mantoani S, Hertzog M, et al. Nat Cell Biol. 2006;8(12):1337–1347. http://doi:10.1038/ncb1502
65. Cui, Y., Dong, Q., Hong, D. et al.BMC Bioinformatics 20, 93 (2019). https://doi.org/10.1186/s12859-019-2672-1
66. Nguyen, B.P., Nguyen, Q.H., Doan-Ngoc, G. et al. BMC Bioinformatics 20, 634 (2019). https://doi.org/10.1186/s12859-019-3295-2
67. Liu, X. L. arXiv 1–38 (2017).
68. Tavanaei, A. et al. IEEE Int. Conf. Bioinforma. Biomed. 145–149, https://doi.org/10.1109/BIBM.2016.7822509 (2016).

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editori

Figures

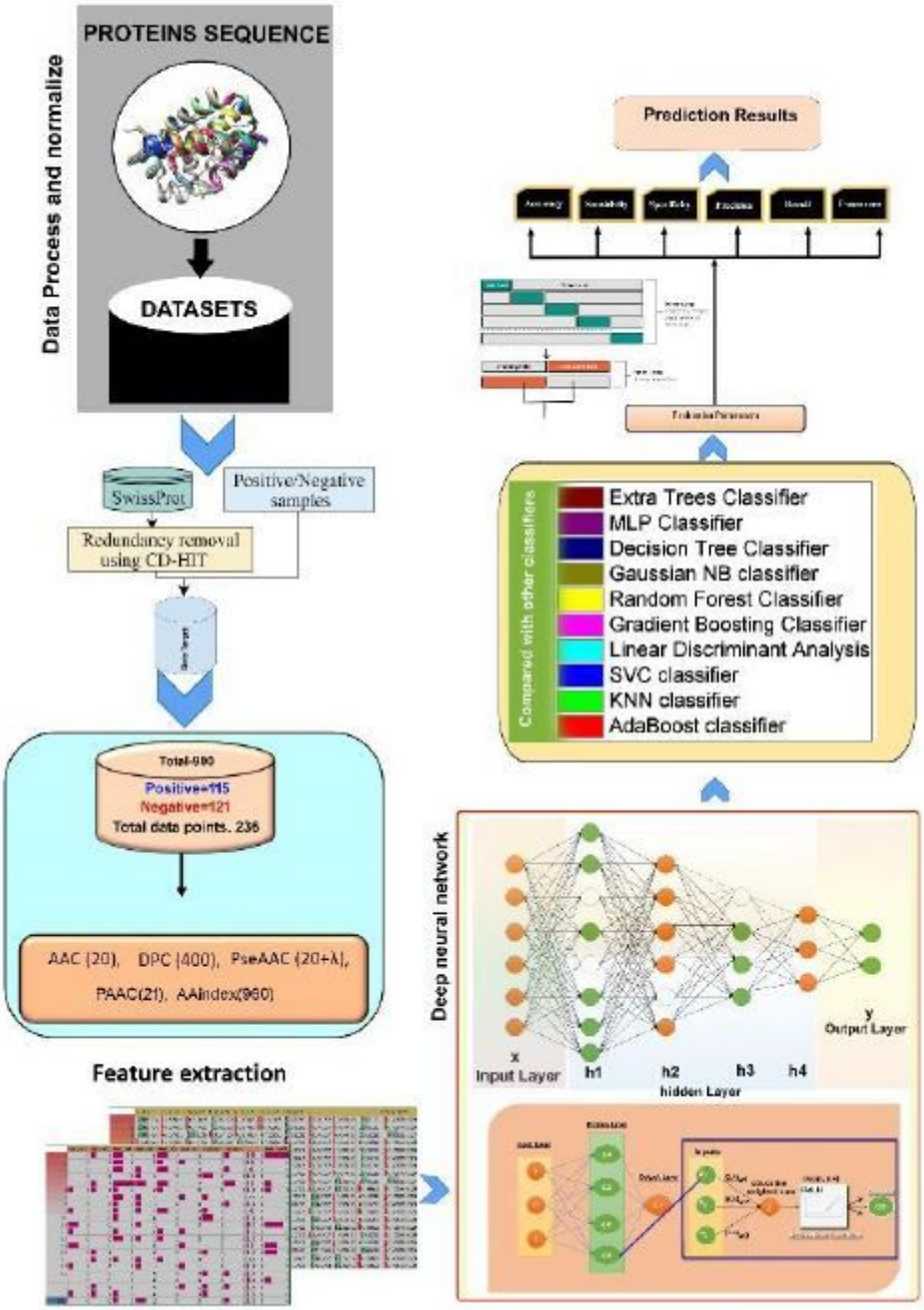

Figure 1

Proposed frame-work model

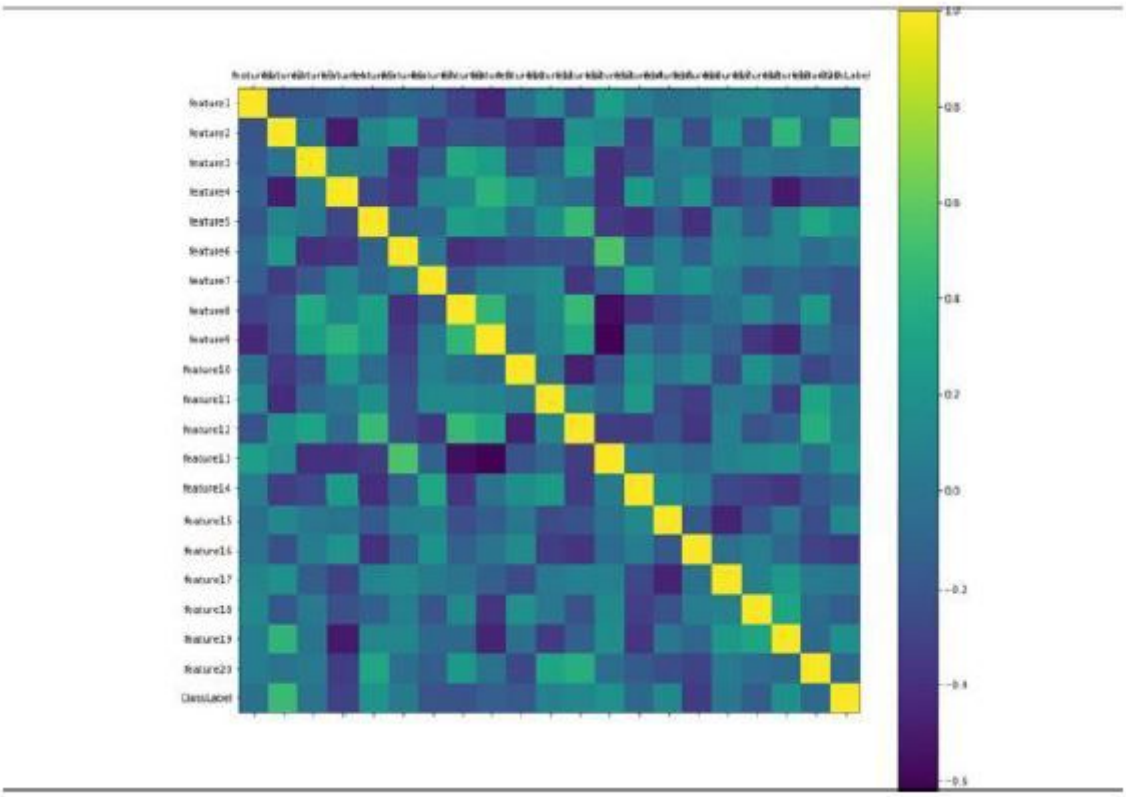

Figure 2

High Correlation with our target value

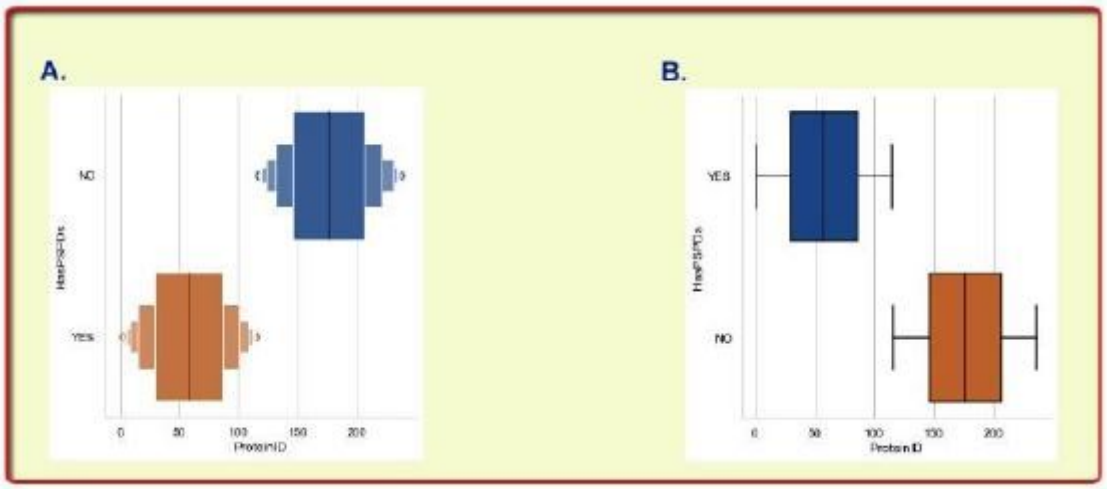

Figure 3

Binary classifications of pathway-specific protein domains

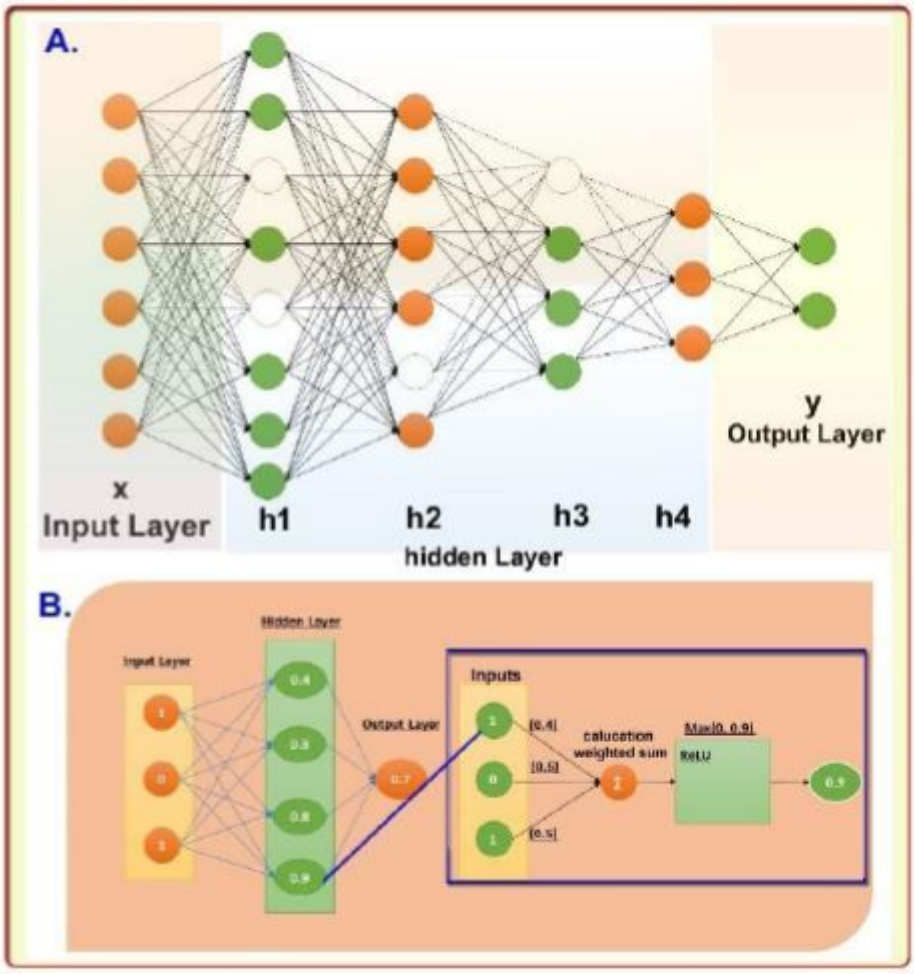

Figure 4

Proposed model

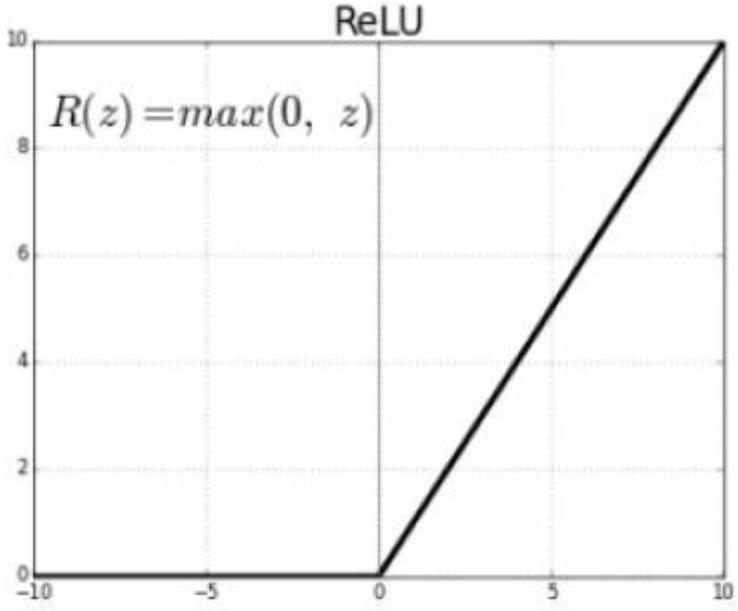

Figure 5

ReLU activation function

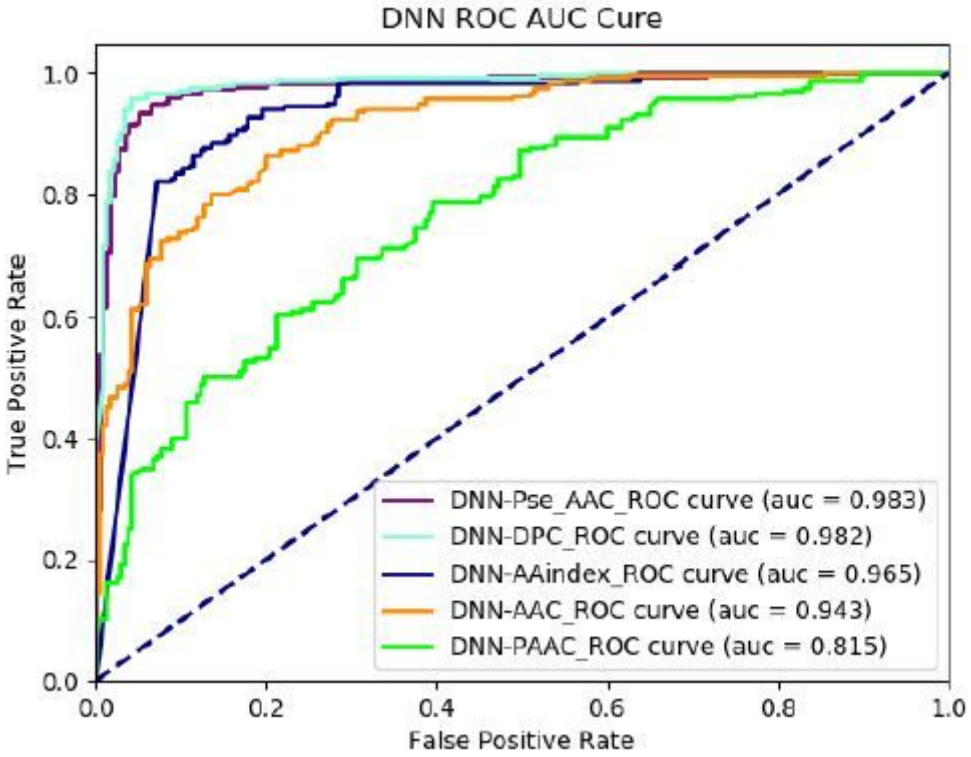

Figure 6

Proposed model ROC-auc score result.

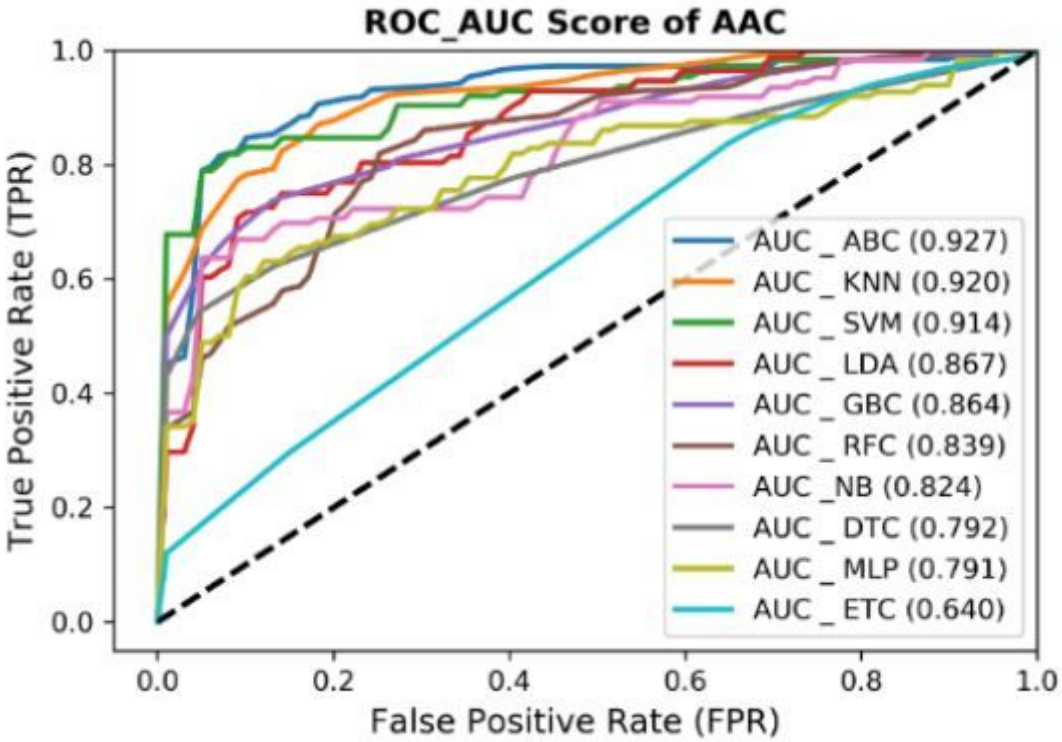

Figure 7

Combine compared ACC model predicted data.

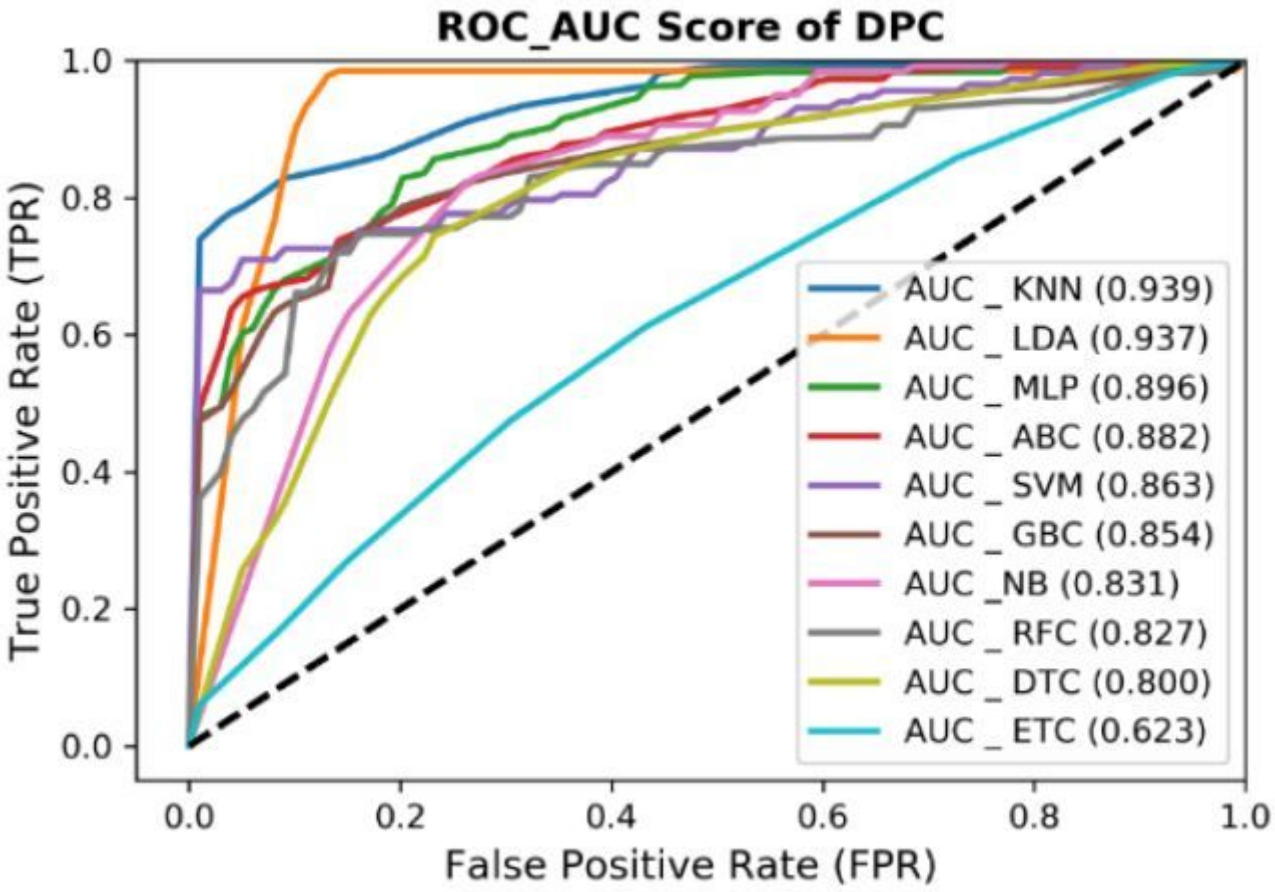

Figure 8

Combine compared DPC model predicted data.

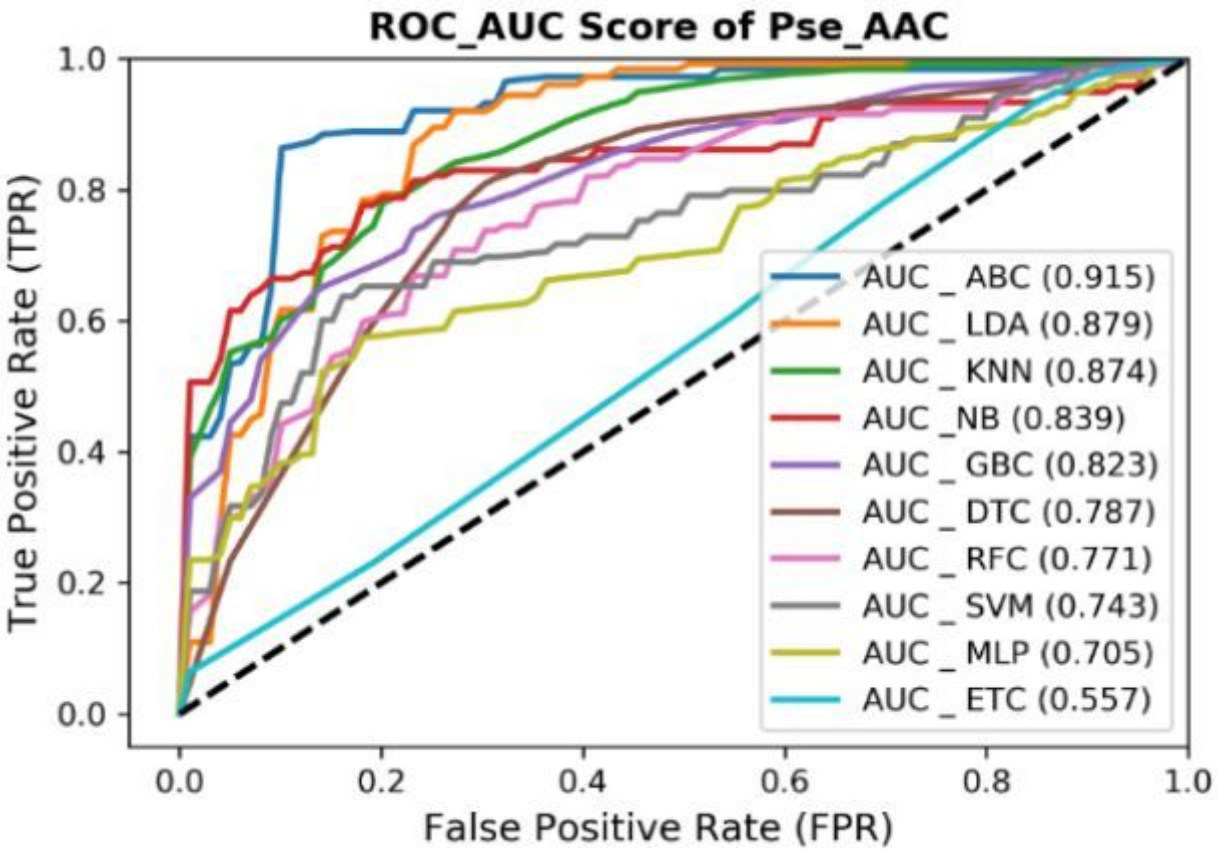

Figure 9

Combine compared Pse_ACC model ROC-AUC

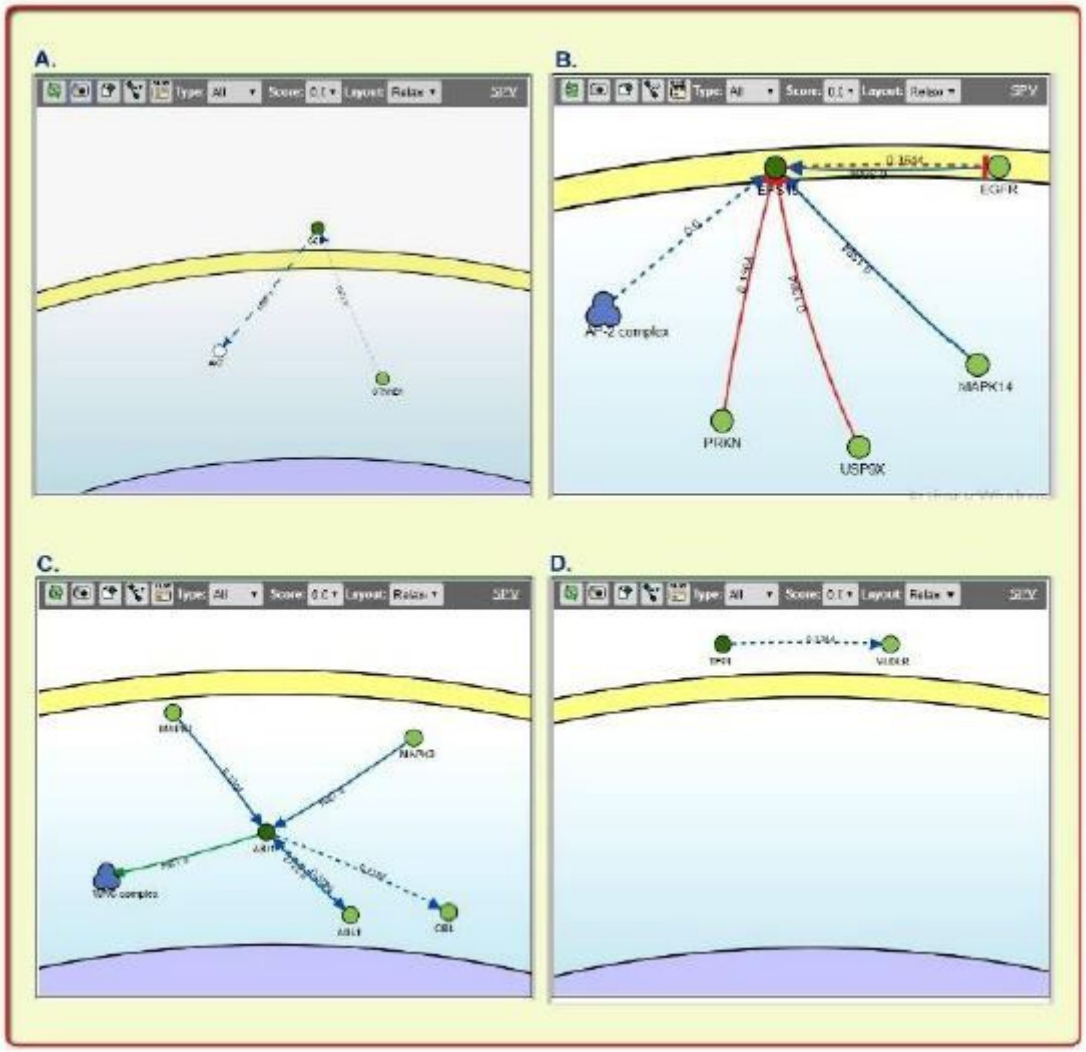

Figure 10

Proteins functions involve in pathways

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplimenteryMaterial.pdf