# Moonlighting protein prediction using physico-chemical and evolutional properties via machine learning methods

**Farshid Shirafkan**
University of Tehran
**Sajjad Gharaghani** ( s.gharaghani@ut.ac.ir )
University of Tehran
**Karim Rahimian**
Tarbiat Modares University
**Reza Sajedi**
Tarbiat Modares University
**Javad Zahiri**
Tarbiat Modares University

**Research Article**

**Keywords:** Moonlighting proteins, Multitasking proteins, Physico-Chemical Properties, Evolutionary Properties, PSSM, Pseudo-amino acid composition, Random forests, SVM, Predictive Bioinformatics

**Posted Date:** December 18th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-126672/v1

RESEARCH

# Moonlighting protein prediction using physico-chemical and evolutional properties via machine learning methods

Farshid Shirafkan[1], Sajjad Gharaghani[1*], Karim Rahimian[2], Reza Hasan Sajedi[3] and Javad Zahiri[2*]

[*]Correspondence:
s.gharaghani@ut.ac.ir;
zahiri@modares.ac.ir
[1] Laboratory of Bioinformatics and Drug Design, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran, Tehran, Iran
[2] Bioinformatics and Computational Omics Lab (BioCOOL), Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran, Tehran, Iran
Full list of author information is available at the end of the article

## Abstract

**Background:** Moonlighting proteins (MPs) are a subclass of multifunctional proteins in which more than one independent or usually distinct function occurs in a single polypeptide chain. Identification of unknown cellular processes, understanding novel protein mechanisms, improving the prediction of protein functions, and gaining information about protein evolution are the main reasons to study MPs. They also play an important role in disease pathways and drug-target discovery. Since detecting MPs experimentally is quite a challenge, most of them were detected randomly. Therefore, introducing an appropriate computational approach seems to be rational.

**Results:** In this study, we would like to represent a competent model for detecting moonlighting and non-moonlighting proteins by extracted features from protein sequences. Then, we will represent a scheme for detecting outlier proteins. To do so, 15 distinct feature vectors were used to study each one's effect on detecting MPs. Furthermore, 8 different classification methods were assessed to find the best performance. To detect outliers, each one of the classifications was implemented 100 times by 10 fold cross-validation on feature vectors, then proteins which miss classified 80 times or more, were grouped. This process was applied to every single feature vector and in the end, the intersection of these groups was determined as the outlier proteins. The results of 10 fold cross-validation on a dataset of 351 samples (containing 215 moonlighting and 136 non-moonlighting proteins) show that the decision tree method on all feature vectors has the highest performance among all methods in this research and also in other available methods. Besides, the study of outliers shows that 57 of 351 proteins in the dataset could be an appropriate candidate for the outlier. Among the outlier proteins, there are non-moonlighting proteins (such as P69797) that have been misclassified by 8 different classification methods with 16 different feature types. Because these moonlighting proteins have been obtained by computational methods, the results of this study could reduce the likelihood of hypothesizing that, these proteins are non-moonlighting.

**Conclusions:** Moonlighting proteins are difficult to identify by experiments. Our method enables identification of novel moonlighting proteins using distinct feature vectors. It also indicates that a number of non-moonlight proteins are likely to be moonlight.

**Keywords:** Moonlighting proteins; Multitasking proteins; Physico-Chemical Properties; Evolutionary Properties; PSSM; Pseudo-amino acid composition; Random forests; SVM, Predictive Bioinformatics

## Introduction

These days most of the researches at the cellular level, study protein functions. Protein function and its mechanism are contemporary topics in biology [2]. The reason for studying this issue is due to the latent importance of this vital macromolecule in the metabolism of organisms and pathogens. Most of the proteins are unifunctional, but a considerable number of them that have been discovered so far are multifunctional. Moonlighting proteins (MPs) are a class of multifunctional proteins. In precise, the word moonlight is applied to proteins with at least two different unrelated functions [23] in a way that this multifunctionality is not due to gene fusion, multiple domains, multiple splice variants, proteolytic fragments, families of homologous and pleitropic effect [2]. Considering this definition, it should be noted that the moonlighting proteins are different from multifunctional proteins because multifunctional proteins are generated by gene fusion and/ or alternative splicing [21, 22].

Another important feature of moonlighting proteins is the independence of their functions in a way that inactivation of one function doesn't affect the other protein functions [20]. Moreover, moonlighting is distinct from pleiotropisms which have a constant role although they participate in multicellular functions and can be present in different pathways. However, multifunctional moonlighting proteins have several different mechanisms [19].

The first example of moonlighting proteins reported in the late 1980s by Piatigorsky and Wistow [25]. They noticed that crystallin which is a structural protein in the eye lens has an enzymatic role as well. Hitherto these proteins have been discovered in mammals, yeast, worms, bacteria, plants, viruses, archaea, and many other organisms. To record the data related to these proteins several online databases were established. MoonProt [9] and MultitaskProtDB-II [15] have reported 400 and 694 proteins respectively, in their last update.

Moonlighting proteins contain various sub-types: 1) different sites for different functions in the same domain 2) different sites for different domains in different domains 3) implementing the same residue for different functions 4) implementing different residues of the same site for different functions 5) implementing structural composition or different folding for different functions [28].

Although there have been several studies on moonlighting proteins in recent decades, a lot about these proteins such as the number of these proteins still have remained unknown. Detection of protein functions, how to target a function without affecting other functions, and discovering the expression patterns changes to find a novel function is from the major questions in biology that giving them answers were so vital and important [22].

In addition to the mentioned matters, detecting unknown cellular processes, perception of new protein mechanisms, improving protein function prediction, a significant role in disease pathways, obtaining information on protein evolution, and drug discovery are the reasons that make MPs studies more appealing. According to previous studies, 78% of moonlighting proteins are involved in human disease pathways and 48% of MPs are the targets of active medicines. For example, phosphoglucose isomerase is an enzyme in glycolysis and also is a cytokine (autocrine motility factor) which has a significant role in breast cancer metastasis [10]. To study more on this issue there are several other examples in [21].

The reasons mentioned above on one side and the challenging laboratory and experimentally methods in detecting these proteins on the other side have made computational methods so remarkable. To date, several computational methods have been used to detect moonlighting proteins.

Since moonlighting proteins, tend to interact with other proteins with different functions or in different pathways, they can be detected by protein-protein interaction (PPI) [7]. Cheng developed an algorithm called MoonFinder that used function

annotations of proteins that had interaction with long non-coding RNAs, to detect moonlight long non-coding RNAs. Gómez et al. implemented sequence similarity to other protein families with different functions to detect moonlighting proteins.

Chapple et al. used a protein-protein interaction network to extract features that enable them to identify extreme multifunctional proteins. These types of proteins belong to several functional modules that are engaged in different functions that MPs are one of their subclasses. First, they detected the overlapping cluster of a PPI network. These clusters contained highly interconnected proteins that tend to get involved in the identical cellular process. In the next phase of their study clusters were annotated by the common function of most of the clusters' proteins. Proteins that were found in the intersection of the clusters were then selected as the candidates. The candidate proteins had more than one function and their first and second functions were not identical. They observed that the number, degree, and the relationship of domains with the disease in candidate proteins were more than the ones in proteins that were in the intersection of clusters but had not been selected due to the identical first and second functions. Also, the average degree in the candidate proteins is higher than the hubs but candidates are less disorder than the hubs.

Aashish Jain et al. developed a new method by text mining to detect moonlighting protein using various information sources [2]. In this method moonlighting proteins were detected by analyzing database entries, literature, and big data omics utilizing the DextMP algorithm. This research was implemented and applied to the genome proteins of Arabidopsis thaliana, Caenorhabditis elegans, and Drosophila melanogaster.

In another study, done by khan et al., the functional features of MPs were identified by using a computational framework from various proteomics aspects. They created a model for prediction of moonlighting protein based on gene ontology

(GO), PPI, gene expression, phylogenetic profiles, genetic interactions, network-based graph properties, and the number and length of intrinsically disordered regions. The prediction accuracy of this method by applying the random forest classification algorithm was 72%.

In this study, we want to propose a prediction model for moonlighting protein detection by extracting features from protein sequence and machine learning classification methods. Different feature vectors have been used to study the effects of each feature vector on MPs prediction. Besides, 8 classification methods that have a good impression in bioinformatics studies were used.

## Material and Methods

### Dataset

In this study, a dataset of 351 proteins was utilized that contained 136 nonmoonlighting and 215 moonlighting proteins. This dataset contains proteins derived from different organisms. Table 1 represents the percentage of proteins based on each organism for each class. To collect moonlighting proteins, the MoonProt database (http://www.moonlightingproteins.org/) and for non-moonlighting proteins, Khan et al. (2014) method based on function annotation, were used [13]. To date, the moonlight database contains 400 moonlighting proteins and the set of proteins that were introduced by the khan method are 150 samples, since data redundancy can lead to bias, CD-hit was utilized to remove the redundant or similar protein. The sequence identity cut-off was considered 40. Finally, a set of 351 proteins was obtained.

### Feature extraction

15 feature vectors used in this study have been shown in table 2. The name of each set of features, length of the feature vector, and a brief definition were described in table 2.

All feature vectors from 1 to 14 were extracted by the ftrCool library [1]. The IF Set of features, which has been shown in table 2 under number 15, consists of several features that each one has been extracted with an appropriate tool. These features are as follows; length, molecular mass, isoelectric point, charge, hydrophobicity, aliphatic index, instability index, GC-content, Hydrogen binding, number of hydrogen bond in alpha-helix (h-AlphaHelix), number of hydrogen bond in beta-sheet (h-BetaSheet), Kidera factor features, MS-WHIM score, post-translational modification, disorder, Amino Acid Composition, Pseudo Amino Acid Composition, and position-specific scoring matrix (PSSM). The PSSM set of features was extracted by the bioinformatics tool POSSUM [12].

## Machine Learning Methods

In this study following classification methods were used; Support Vector Machine (SVM), Random Forest (RF), Multi-Layer perceptron (MLP), Logistic Regression (LR), Decision Tree (DT), Ada Boost (ADA), Naïve Bayes (NB), K nearest neighbor (KNN).

SVM is one of the most applicable methods of machine learning that utilizes an optimized hyperplane to distinguish classes [5]. One of the advantages of this method is high accuracy and also high performance. This method is only applicable to two-class or binary problems and therefore it can be considered as its disadvantage [11, 14]. This method is used for hot spot detection in proteins.

KNN is one of the simplest algorithms of machine learning [6]. In this method the distance of an object to the k nearest neighbors is calculated, and then the object adapts the label that has the most numbers between k nearest objects. Prediction of the hot spot in proteins and PPI are the applications of this method [17, 27].

NB this classification method is based on Bayes' theorem and independence assumptions between the data. This assumption can highly reduce the complexity of the classification. Simplicity and low computational costs are the advantages of this

method while the independence assumption and normalization of data distribution could have a negative impact on the accuracy and precision of the algorithm. Several successful applications of this method have been reported for PPI, so far [4,18,24].

DT one of the most useful classification methods that can visualize the relation between classes and feature vector [26]. Each leaf of the decision tree represents a class. Branches perform as ways to classes based on the contents of the feature vectors. Although this method is simple, it can increase the classification error. This method is successful in PPI prediction [8].

MLP artificial neural networks are made based on human contemplation that can process simultaneously [30]. Multilayer neural network is a type of artificial neural network that consists of at least three layers, input, hidden, and output layer. Each layer contains several nodes (neurons) that exhibit a specific output in the network. Edges connect the nodes and each edge contains a value called weight. The edges transfer output of a neuron to another. The last layer is the output layer and the result could be extracted from this layer. This network demonstrates good results in human virus PPI prediction [29].

Ada boost is one of the ensemble methods that the compositions of classifiers are used for better and more accurate prediction. In this method, weaker classifiers called week learners are utilized. Each week learner creates an output (a class) for each sample. Then the linear sum of these week learners is selected so that the classifier error is minimized. RF is a collection of decision trees. Every decision tree, is trained by a subset of features. The selection of this subset is done randomly. Ultimately, to calculate the predicted value, the majority of votes are used. This method has been successful in predicting PPI that was presented by Akbaripour et al. [16].

## Implementing the Classifiers

To implement the machine learning models, scikit-learn library was used. Each model has parameters in which to get the best results, they ought to be adjusted. The best value of k in KNN is 5 and the Euclidean distance function was set. The kernel function in the SVM method was set equal to RBF and marginal parameter C was set to 1. In the decision tree, the Gini criteria was used, and max-depth, and min-sample is set to 3 and 5 respectively. The number of weak learners, in the Ada boost method, considered 200, and the number of decision trees in RF method was set equal to 50. Ultimately, to perform and execute MLP, two layers of 20 and 3 neurons in the hidden layer with the maximum iteration of 150 were used. Out of all available activation function for this method, sgd exhibited better results. Figure1 shows the implementation of classification models.

First, desirable features (table2) were extracted from protein sequences (MPs and non-MPs). Each of the features in table2 was saved in separate vectors with different dimensions. Then each of the feature vectors was used for classification models training. The trained models were compared to one another and the best feature vector and also the best model were selected. Finally, the best model was used for protein classification.

## Performance evaluation

16 feature vectors and 8 classification methods were studied in this research that contained a total of 128 different results. To achieve the best result, 10 fold cross-validation was used and the results' assessment was done using F-measure, Precision, Recall, Accuracy (acc), and MCC.

$$acc = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$precision = (TP + TN)/(TP + FP + TN + FN) \quad (2)$$

$$recall = TP/(TP + FN) \quad (3)$$

$$F - Measure = 2 * (precision * recall)/(precision + recall) \quad (4)$$

$$MCC = (TP * TN - FP * FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (5)$$

In these equations, TP represents the number of true positives, FP, TN, and FN show the number of false positives, true negatives, and false negatives, respectively. For further information refer to [3]. Besides what mentioned area under the curve of ROC (AUC) was also used.

## Results and Discussion

### Results of model performance

10 fold cross-validation was used to get the results. All the columns that correlated greater than 0.4 were removed from the dataset. The summary of the results was exhibited as a heatmap in figure2. In this figure accuracy, Matt, Recall, Precision, F-measure, AUC were shown for each feature and every classification method. SVM with the feature vector of SAAC and MLP with the CDT feature vector had the highest and lowest precision, respectively. This performance was confirmed with other assessments such as AUC, F1 score, Precision, and Recall.

Figure2 indicates that in order of precedence, SVM, RF, and NB have much higher accuracy compare to other models. By this comparison, based on feature vectors, SAAC, QSO, and SGAAC showed higher accuracy than other features.

Although the implementation of all features could produce a relative improvement in model accuracy, using all the features generally did not lead to the best results. For instance, in the SVM method, using every feature resulted in an accuracy of 0.741 which was higher than other models except for QSO and SAAC. This is applicable to other methods and features.

## Outlier Detection

There are many methods available to detect outliers. In this study, the number of proteins that were misclassified was used. In this method, every classifier was implemented 100 times by 10 fold cross-validation on data. Then each protein that was misclassified 80 times would take into account and results for each model and every feature were obtained (figure 3). In this study, the number of times a protein misclassified was called the misclassification rate (MR). Proteins with a 0.8 misclassify rate were considered as Candidate Outlier Proteins (COP). For example in the MLP method by using the feature vector SOCN near 75 proteins considered COP.

It is clear that the number of COP for each model and each feature vector differs greatly. For example, in the NB method by using SOCN, the number of COPs was more than 135 that represented 38% of proteins, while in the DT method using AllData, this value was 28 that was 8% of all data.

Table 3 indicates the average number of COPs for each classification model. To determine the average in a classifier, the sum of the number of COPs for all feature vectors was used. As it was shown, NB, SVM, LR, KNN, RF, DT, ADA, and MLP in order of precedence have the highest to the lowest number of COPs.

Furthermore, the means of MR are demonstrated in table 3. For instance, the ADA method contains an average of 62 proteins that their average MR is 93.60%. This value was then called the Average Misclassify Rate (AMR). As it is obvious, not only the number of misclassified proteins is high, but also they have high MR value. For example, in the NB method, there is an average of 97 proteins that have

AMR greater than 98.27%. This indicates that each protein in this set has a high MR value.

Figure4, shows the AMR for every feature vector and each classification model. As it is shown in this figure, the MR value, for none of the classification model, is not less than 90.

To improve the results, COPs were eliminated and classification models with 10 fold cross-validation were implemented on remaining proteins. The obtained results, in figure5, indicate a noticeable improvement. As it is exhibited in figure 5, the SVM method using the feature vector of CTDC and CTDT has a better performance compared to the other feature vectors. It should be noted that this raise in accuracy is achieved by losing approximately 27% of proteins. This also applied to the NB method in which using AAAC and SOCN lead to 0.97 and 0.96 accuracies, respectively, but near 26% and 38% of the proteins lost considering outliers. To eliminate this bias, results with less number of outliers could be used. Therefore in this study, 6 models with fewer outliers were selected.

Table4 demonstrates the performance of the selected models before and after the elimination of the outliers. As it is shown, the DT method with the feature vector of AllData could obtain an accuracy of 0.84 by removing only 8% proteins. This result indicates the 0.24 improvement compared to the performance before removing the outliers.

Other assessments confirm the performance improvement. The comparison of the introduced models, in table4, indicates that eliminating more proteins does not necessarily lead to higher accuracy and DDE feature vector ( with 45 eliminated proteins and an accuracy of 0.77) and AllData (with 28 eliminated proteins and an accuracy of 0.84) using the DT method confirm that. This also applied to the MLP using CTDC, QSO, SGAAC features. The ADA method using the IF dataset has the most accuracy between the selected sets. In this method by detecting 45

proteins and eliminating them, the model reached an accuracy of 0.88 which was much higher compared to the DT method using the DDE feature vector.

### Intersection of COPs

To obtain outlier proteins, the intersection of COPs was used. In this method, a set of the obtained COPs was made, from every feature, using a classifier. Then all the obtained sets were intersected and the achieved result was considered as outliers. Figure 6 exhibits the results obtained from the first method in which 57 proteins were considered as outliers. In this set, 3 proteins, O75821, P69786, P69797 in 8 methods, these proteins, P43274, P47176, Q8WWY3, P25043, P36160, in 7 methods, these three proteins, P47813, P03004, P11655, in 5 methods, and finally these 5 proteins, A2A5z6, P38624, P32368, P32332 Q13325, in 4 methods, were considered as COPs. Other proteins were repeated in 3 methods or less. Figure7 shows each of the proteins. 26.5% of proteins were repeated in 2 or fewer methods. It is notable that 19% of proteins in figure7 are moonlighting proteins and 81% are non-moonlighting proteins.

## Conclusion

Moonlighting proteins are important molecules in cell cycles. They have a significant role in regulatory activities and disease-related pathways. Experimental methods have their complications in detecting moonlighting proteins, therefore using computational methods has attracted much attention in detecting moonlighting proteins. Many computational methods have been used to detect these proteins, so far.

However, studies that have used machine learning methods along with feature extraction are rare. In this study, 8 classification methods and 16 different feature vectors were used to detect moonlighting proteins. Moreover, to detect outliers the method of counting misclassified proteins was utilized. In this method, sets of proteins that were misclassified more than 80 times for each classifier and every

feature vectors were considered as candidate outlier protein. Then the intersection of these sets was represented as outlier proteins.

Classification methods were implemented on selected datasets twice, with and without the elimination of COPs. Results show that the decision tree method with the elimination of only 28 proteins, had a satisfactory performance (assessing by F1, Matt, Recall, Precision, Accuracy, and AUC) compared to other available methods. Furthermore, results indicate that 57 proteins out of 351 are outliers. These are proteins that have been repeated at least one time in each classifier for each feature. The average misclassification rate for these proteins is greater than 90% that is a high value. Therefore, they could be good outlier candidates. These proteins belong to different organisms. Statically, 19% of outlier proteins are moonlighting proteins, and the rest of them are non-moonlighting proteins.

By studying non-moonlighting proteins that were considered COPs, it is specified that their characteristics are more like moonlighting proteins and it is better to derive them out of the non-moonlighting proteins category, because it may later become clear that they were Moonlighting proteins.

# Appendix

**Acknowledgements**

Not applicable.

**Funding**

Not applicable.

**Abbreviations**

Not applicable.

**Availability of data and materials**

All data generated or analysed during this study are included in this published article and its supplementary information files.

**Ethics approval and consent to participate**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable

**Authors' contributions**

Shirafkan and Rahimian designed and performed experiments, analysed data and co-wrote the paper. Gharaghani and Zahiri supervised the research. All authors read and approved the final manuscript

**Authors' information**

Laboratory of Bioinformatics and Drug Design, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Sajjad Gharaghani, Farshid Shirafkan

Bioinformatics and Computational Omics Lab (BioCOOL), Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran

Javad Zahiri, Karim Rahimian

Department of Biochemistry, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran

Reza H Sajedi

**Author details**

[1] Laboratory of Bioinformatics and Drug Design, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran, Tehran, Iran. [2] Bioinformatics and Computational Omics Lab (BioCOOL), Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran, Tehran, Iran. [3] Department of Biochemistry, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran, Tehran, Iran.

**References**

1. Sare Amerifar. ftrcool: Feature extraction from biological sequences, 2020.
2. H. Gali A. Jain and D. Kihara. Identification of moonlighting proteins in genomes using text mining techniques. *Proteomics*, 18(21–22), Nov 2018.
3. M. Bramer. Measuring the performance of a classifier. *Springer, London*, page 175–187, 2013.
4. D. Nguyen C. D. Nguyen, K. J. Gardiner and K. J. Cios. Prediction of protein functions from protein interaction networks: A naive bayes approach. *in Lecture Notes in Computer Science*, 5351:788–798, 2008.
5. C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn*, 20(3):273–297, Sep 1995.
6. T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Trans*, 13(1):21–27, 1967.
7. A. Gomez et al. Do protein-protein interaction databases identify moonlighting proteins? *Mol. Biosyst*, 7(8):2379–2382, Aug 2011.
8. A. Sikandar et al. Decision tree based approaches for detecting protein complex in protein protein interaction network (ppi) via link and sequence analysis. *IEEE Access*, 6:22108–22120, Feb 2018.
9. C. Chen et al. Moonprot 2.0: An expansion and update of the moonlighting proteins database. *Nucleic Acids Res*, 46(D1):D640–D644, Jan 2018.
10. H. Watanabe et al. Purification of human tumor cell autocrine motility factor and molecular cloning of its receptor. *J. Biol. Chem*, 266(20):13442–13448, 1991.
11. J. Xia et al. Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features. *Oncotarget*, 7(14):18065–18075, 2016.
12. Javad Zahiri et al. Ppievo: Protein–protein interaction prediction from pssm based evolutionary information. *Genomics*, 102(4):237–242, 2013.
13. Khan et al. Genome-scale identification and characterization of moonlighting proteins. *Biology Direct*, 2014.

14. L. Deng et al. Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. *J. Comput. Biol*, 20(11):878–891, Nov 2013.
15. L. Franco-Serrano et al. Multitaskprotdb-ii: An update of a database of multitasking/moonlighting proteins. *Nucleic Acids Res*, 46(D1):D645–D648, Jan 2018.
16. M. Akbaripour-Elahabad et al. rpicool: A tool for in silico rna-protein interaction detection using random forest. *Theor. Biol*, 402:1–8, 2016.
17. M. R. Guarracino and A. Nebbia. Predicting protein-protein interactions with k-nearest neighbors classification algorithm. *Lect. Notes Comput. Sci*, 6160:139–150, 2010.
18. X. Lin H. Geng, T. Lu and F. Yan. Prediction of protein-protein interaction sites based on naive bayes classifier. *Biochem*, 2015:788–798, 2008.
19. D. H. E. W. Huberts and I. J. van der Klei. Moonlighting proteins: An intriguing mode of multitasking. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1803(4. Elsevier):520–525, Apr 2010.
20. C. J. Jeffery. Moonlighting proteins. *Trends in Biochemical Sciences*, 24(1. Elsevier Ltd):8–11, Jan 1999.
21. C. J. Jeffery. Why study moonlighting proteins? *Frontiers in Genetics*, 6(JUN), 2015.
22. C. J. Jeffery. Protein moonlighting: What is it, and why is it important? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1738), Jan 2018.
23. M. Krantz and E. Klipp. Moonlighting proteins - an approach to systematize the concept. *In Silico Biol*, 14(1–2):1–13, Jan 2020.
24. Y. Murakami and K. Mizuguchi. Applying the naive bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics*, 26(15):1841–1848, Jun 2010.
25. J. Piatigorsky and G. J. Wistow. Enzyme/crystallins: Gene sharing as an evolutionary strategy. *Cell*, 57(2):197–199, Apr 1989.
26. J. R. Quinlan. Induction of decision trees. *Springer*, 1(1):81–106, Mar 1986.
27. B. Wang S. S. Hu, P. Chen and J. Li. Protein binding hot spots prediction from sequence only by a new ensemble learning method. *Amino Acids*, 49(10):1773–1785, Oct 2017.
28. Daisuke Kihara Sayoni Das, Ishita Khan and Christine Orengo. Exploring structure function relationships in moonlighting proteins. In Brian Henderson, editor, *Moonlighting Proteins: Novel Virulence Factors in Bacterial Infections*, page 53. Wiley blackwell, University College London, 1nd edition, 2017.
29. Q. Li X. Yang, S. Yang and Z. Zhang. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput. Struct. Biotechnol. J*, 18:153–161, Jan 2020.
30. X. Yao. Evolving artificial neural networks. *Proc. IEEE*, 87(9):1423–1447, 1999.

**Figures**

**Figure 1** schematic view of Pipeline for moonlight proteins detections. A) collecting proteins B) Extracting Features. C)Training models (MLP, Knn, LR, Ada Boost, SVM, RF, DT, NB). D) performance evaluation.

**Figure 2** models performance before remove outlier proteins. Using heat map to compare performance (Matt, Recall, Precision, F-Measure, AUC) of models. Red and Green color indicates worse and better results respectively.

**Figure 3** Misclassification Frequency. X Axis and Y axis indicate Features vector and Frequency Respectivly. Each line in chart indicate a model.

**Figure 4** Radar Plot showing the average misclassify rate for every feature vector and for each classification model.

**Figure 5** models performance after remove outlier proteins. Using a heat map to compare performance (Matt, Recall, Precision, F-Measure, AUC) of models. Red and Green color indicates worse and better results respectively.

**Figure 6** Word cloud of Cop proteins names. Bigger and smaller words indicate the higher and lower misclassication rate, respectively.

**Figure 7** Misclassification rate for outlier proteins. 19% of proteins are moonlighting proteins and 81% are non-moonlighting proteins.

#### Tables

**Table 1** The number of moonlighting and non moonlighting proteins.(moonlight exist in different organism).

| Organism | Moonlight | NonMoonlight |
|---|---|---|
| Mus Musclus | 11 | 39 |
| Human | 57 | 48 |
| E.Coli | 24 | 16 |
| Yeast | 23 | 33 |
| Rat | 5 | 0 |
| Drome | 8 | 0 |
| Arath | 5 | 0 |
| Other | 82 | 0 |
| Total | 215 | 136 |

**Table 2** Candidate extracted features. CKSAAp with length of 401 and SOCN with length of 20 are the longest and shortest length respectively.

| Row# | Feature Name | Description | No. of Feature# |
|---|---|---|---|
| 1 | DDE | Dipeptide Deviation from Expected Mean value | 400 |
| 2 | CTD | Composition Transition Distribution | 273 |
| 3 | AAAC | Amino Acid Auto Correlation | 180 |
| 4 | CT | Conjoint Triad | 343 |
| 5 | SAAC | Split Amino Acid Composition | 60 |
| 6 | AAKPC | Amino Acid to K part, Composition | 60 |
| 7 | SOCN | Sequence Order Coupling Number | 20 |
| 8 | SGAAC | Split Group Amino Acid Composition | 40 |
| 9 | QSO | Quasi Sequence Order | 50 |
| 10 | GrpDDE | Group Dipeptide Deviation from Expected Mean | 64 |
| 11 | CkSAAp | Composition of k-spaced Amino Acids pairs | 401 |
| 12 | CTDC | Composition TransitionDistribution | 39 |
| 13 | EVAA | Expected Value for each Amino Acid | 400 |
| 14 | CTDT | CompositionTransition Distribution | 39 |
| 15 | IF | Some interest feature | 106 |

**Table 3** Average number of COPs for each classification model.

| name | ADA | KNN | NB | DT | LR | SVM | RF | MLP |
|---|---|---|---|---|---|---|---|---|
| AVERAGE OF MR | 93.60 | 97.93 | 98.27 | 93.30 | 97.34 | 98.46 | 97.07 | 92.94 |
| FREQUENCY AVERAGE | 62.94 | 85.63 | 97.06 | 63.69 | 87.69 | 88.06 | 77.63 | 52.38 |

**Table 4** Comparison of models performance with fewer outlier proteins before and after removing outlier.

| After remove outlier | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Model | Mr(%) | MF | ACC | AUC | Recall | F1 | Matt | Precision |
| All Data | DT | 8 | 28 | 0.85 | 0.85 | 0.83 | 0.68 | 0.85 | 0.85 |
| SGAAC | MLP | 10 | 38 | 0.82 | 0.79 | 0.79 | 0.61 | 0.82 | 0.82 |
| QSO | MLP | 11.3 | 40 | 0.86 | 0.85 | 0.84 | 0.70 | 0.86 | 0.86 |
| CTDC | MLP | 11.9 | 42 | 0.80 | 0.79 | 0.77 | 0.58 | 0.80 | 0.80 |
| DDE | DT | 12.8 | 45 | 0.78 | 0.76 | 0.74 | 0.51 | 0.78 | 0.78 |
| IF | ADA | 12.8 | 45 | 0.89 | 0.87 | 0.87 | 0.75 | 0.89 | 0.89 |
| Before remove outlier | | | | | | | | | |
| All Data | DT | | | 0.60 | 0.57 | 0.55 | 0.13 | 0.60 | 0.60 |
| GAAC | MLP | | | 0.62 | 0.60 | 0.60 | 0.19 | 0.62 | 0.62 |
| QSO | MLP | | | 0.73 | 0.72 | 0.71 | 0.44 | 0.73 | 0.73 |
| CTDC | MLP | | | 0.60 | 0.50 | 0.40 | 0.01 | 0.60 | 0.60 |
| DDE | DT | | | 0.66 | 0.63 | 0.63 | 0.27 | 066 | 066 |
| IF | ADA | | | 0.70 | 0.67 | 0.67 | 0.35 | 0.70 | 0.70 |

**Additional Files**

Additional file 1 — Extracted feature data-set1.

Includes calculated values for features CTDT, CTDC, CT, DDE, EVAA, GrpDDE, QSO, SAAC, SGAAC, SOCN.

Additional file 2 — Extracted feature data-set2.

Includes calculated values for features IF, AAAC, AAKPC, CKSAAP, CTD.

# Figures

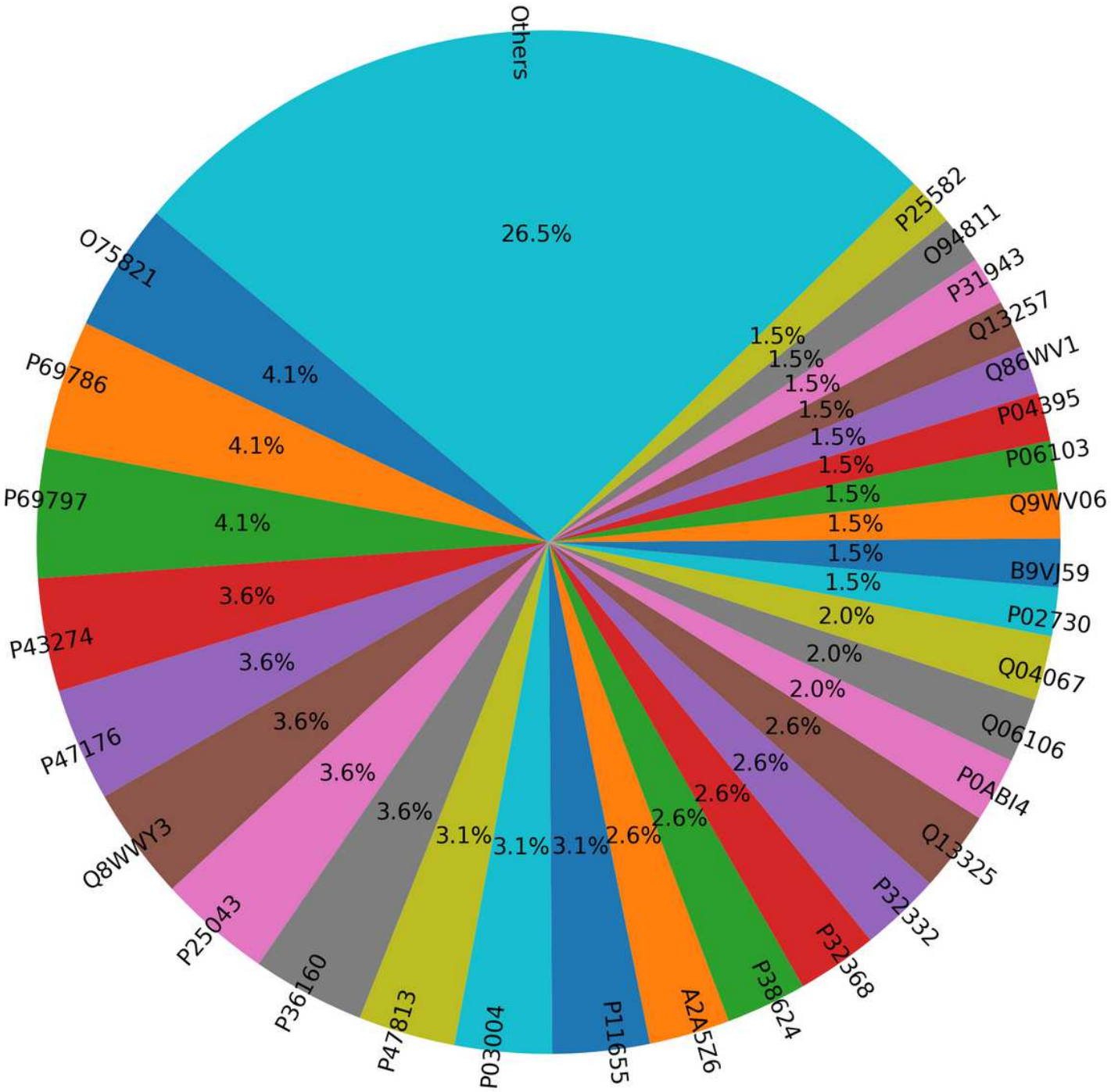


**Figure 1**

Misclassication rate for outlier proteins. 19% of proteins are moonlighting proteins and 81% are non-moonlighting proteins.

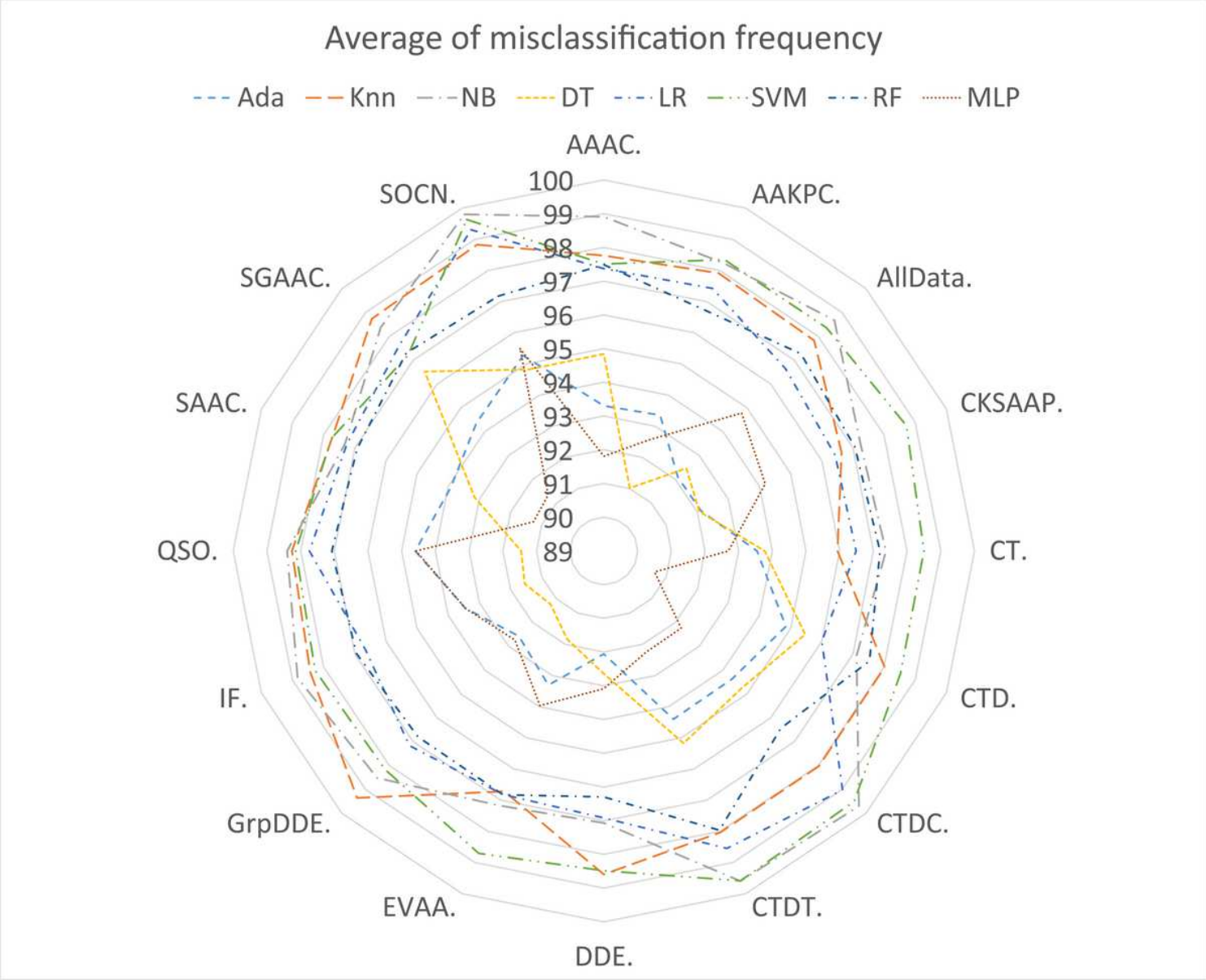


**Figure 1**

Radar Plot showing the average misclassify rate for every feature vector and for each classication model.

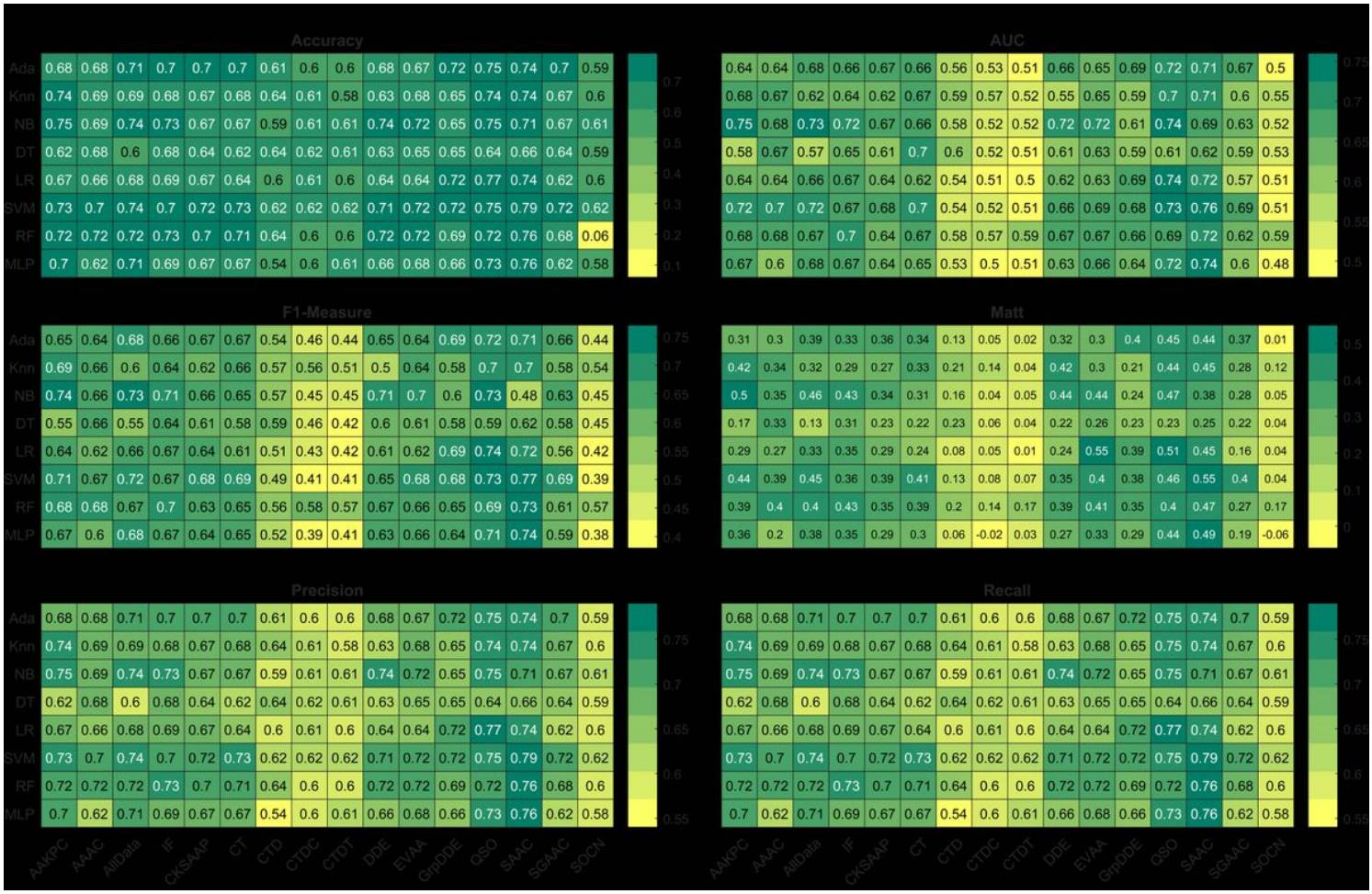


**Figure 1**

models performance before remove outlier proteins. Using heat map to compare performance (Matt, Recall, Precision, F-Measure, AUC) of models. Red and Green color indicates worse and better results respectively.


## Figure 1

models performance after remove outlier proteins. Using a heat map to compare performance (Matt, Recall, Precision, F-Measure, AUC) of models. Red and Green color indicates worse and better results respectively.


**Figure 1**

Word cloud of Cop proteins names. Bigger and smaller words indicate the higher and lower misclassication rate, respectively.

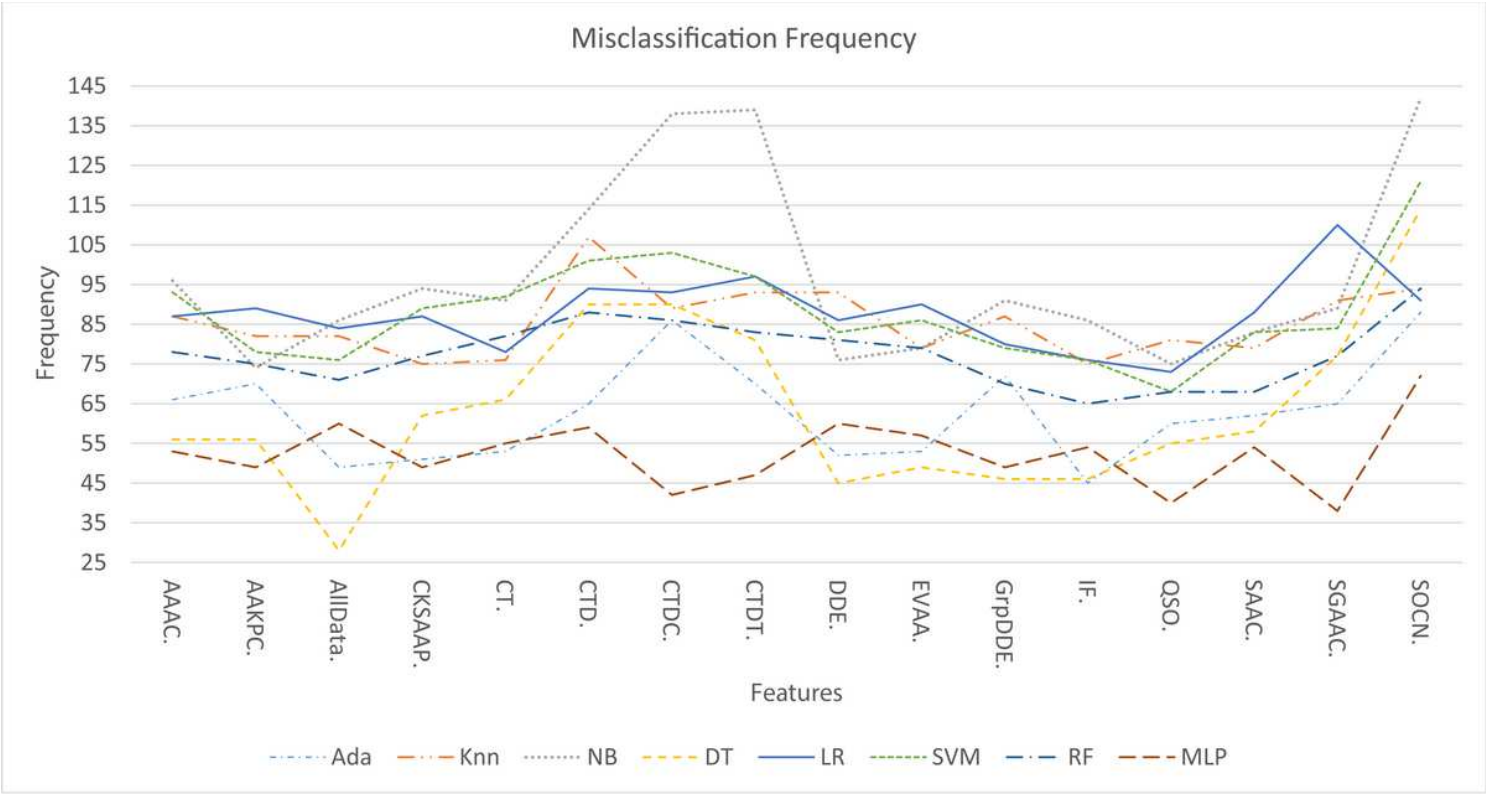


**Figure 1**

Misclassication Frequency. X Axis and Y axis indicate Features vector and Frequency Respectivly. Each line in chart indicate a model.

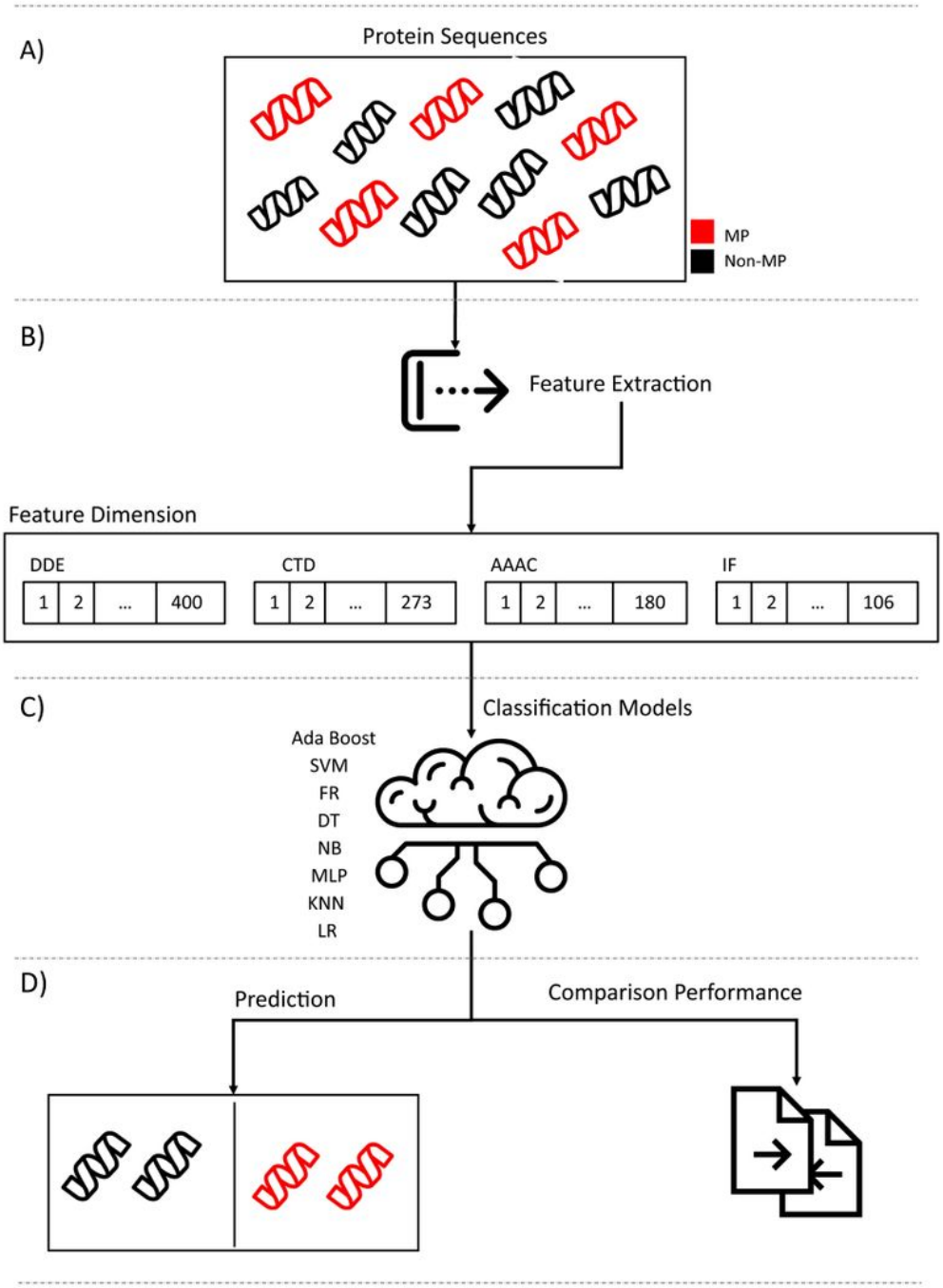


**Figure 1**

schematic view of Pipeline for moonlight proteins detections. A) collecting proteins B) Extracting Features. C)Training models (MLP, Knn, LR, Ada Boost, SVM, RF, DT, NB). D) performance evaluation.

# Supplementary Files

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

- Additionalfile2.xlsx
- Additionalfile1.xlsx