Designing effective predictors of protein post-translational modifications using iLearnPlus

Zhen Chen[1], Fuyi Li[2,3], Xiaoyu Wang[2,3], Yanan Wang[2,3], Lukasz Kurgan[5,*], Jiangning Song[2,3,4,*]

[1]Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou 450046, China;

[2]Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia;

[3]Monash Data Futures Institute, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia;

[4]Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Melbourne, Parkville, Victoria 3010, Australia;

[5]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA.

*To whom the correspondence should be addressed.

Jiangning Song, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. Tel: +61-3-9902-9304; Email: Jiangning.Song@monash.edu.

Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA Tel: +1-804-827-3986; Email: lkurgan@vcu.edu.

Abstract

Posttranslational modifications (PTMs) have vital roles in a myriad of biological processes, such as metabolism, DNA damage response, transcriptional regulation, protein-protein interactions, cell death, immune response, signaling pathways and aging. Identification of PTM sites is a crucial first step for biochemical, pathological and pharmaceutical studies associated with the functional characterization of proteins. However, experimental approaches for identifying PTM sites are relatively expensive, labor-intensive and time-consuming, partly due to the dynamics and reversibility of PTMs. In this context, computational methods that accurately predict PTMs serve as a useful alternative, especially when targeting large-scale whole-proteome annotations. We briefly summarize and review existing predictors of PTM sites in protein sequences. Moreover, we introduce the iLearnPlus platform that facilitates development of new predictive methods and apply it to generate a new PTM predictor. We elaborate a detailed procedure for the development of predictive models, particularly focusing on the deep learning (DL) techniques. We assess predictive performance of the developed DL model and demonstrate how to compare it against other machine learning algorithms. While we use iLearnPlus in the context of the PTM prediction, we emphasize that this platform can be used to design predictive systems for a broad spectrum of other related problems that cover prediction of structural and functional characteristics of proteins and nucleic acids.

1 Introduction

Posttranslational modifications refer to the reversible or irreversible chemical changes that some proteins undergo after translation [1]. PTMs play vital roles in a broad array of cellular processes, such as metabolism, signal transduction, stability, structural state, and localization of proteins [2-8]. For example, phosphorylation is implicated in orchestrating signal transduction, cytoskeleton rearrangement, and cell cycle progression [9, 10]. Moreover, ubiquitination mediates protein degradation by the Ub-proteasome system in eukaryotic cells [11] while malonylation plays vital roles for metabolic reprogramming in determining the function of immune cells [12]. To date, advances in experimental techniques have significantly assisted biologists in identifying various types of PTMs. Currently, over 680 types of PTMs have been characterized experimentally (see http://www.uniprot.org/docs/ptmlist.txt).

Given the prevalence and importance of PTMs, aberrant modifications are shown to be associated with various human diseases [4, 5, 13-16]. Systematic identification of different types of PTM substrates and PTM sites in proteomic data is becoming an urgent issue. To date, numerous efforts have been dedicated to the investigation of cellular mechanisms that underly PTMs, which is based on accurate identification of corresponding PTM substrates and sites. Advances in the PTM research benefit from computational studies that accurately predict PTM sites, significantly reducing the time and effort involved in the experimental identification. Compared with the labor-intensive and time-consuming experimental characterization of PTMs, computational prediction of PTMs in proteins provides a valuable and complementary approach to shortlist likely candidates for subsequent experimental validation. Thus, a variety of computational methods for PTM identification have been developed using various protein sequence features and state-of-the-art machine learning (ML) techniques [17-21]. These methods predict new PTM sites by learning features of the sequence context of experimentally verified PTM sites primarily using ML algorithms. We briefly overview existing computational predictors of PTM sites.

We describe an innovative and comprehensive platform for the development of new predictive methods, iLearnPlus [22], and we apply it to generate a new PTM predictor for lysine malonylation. We detail the procedure for development of predictive models based on iLearnPlus, focusing on the DL techniques. This includes benchmark dataset preparation, feature extraction, model construction and performance evaluation. In particular, we compare the results produced by the DL model against other ML algorithms.

While here we apply iLearnPlus for the lysine malonylation prediction, this software can be used to design, implement and comparatively validate predictive systems for many other related problems. These application areas broadly cover prediction of structural and functional characteristics of proteins and nucleic acids, such as secondary structure, intrinsic disorder, protein-ligand and nucleic acid-ligand binding, and many others.

2 Brief review of computational PTM site prediction

Recent years have witnessed the development and proliferation of computational approaches for the prediction of PTM and cleavage sites [17, 18, 20, 21, 23-39]. These methods differ in a variety of aspects, including the dataset collection and preprocessing, feature descriptors

and feature selection techniques employed, classification algorithms used, and performance evaluation strategies utilized.

Generally, the current models for the PTM sites prediction could be divided into three main categories based on the adopted techniques. The first category is based on peptide similarity. Methods in this group usually calculate a similarity score between the peptide that is being predicted and peptides with experimentally annotated PTM sites [38, 40]. The similarities are computed using a number of measures, such as the BLOcks SUbstitution Matrix (BLOSUM62) matrix [41] and position-specific scoring matrix (PSSM) [42]. Representative methods in this category include the Group-based Prediction System (GPS) series approaches for predicting phosphorylation [43], methylation [44], sumoylation [45], as well as the acetylation set enrichment-based (ASEB) approach [46] for the acetylation sites prediction [47].

The secondary category relies on conventional ML algorithms using sequence-derived features. Here, ML algorithms are used to derive predictive models from experimentally annotated training data. Authors of these methods manually develop an approach to transform the input sequences into a fixed-size numeric feature set that is subsequently input into the ML model. The fixed-size feature vector is required by these types of models. Fortunately, development of these feature sets from biological sequences (i.e., protein and nucleic acid sequences) is supported by a variety of convenient tools, such as ProtrWeb [48], iFeature [49], BioSeq-Analysis [50], iLearn [51] and iLearnPlus [22]. Conventional ML algorithms (i.e., ML algorithms that exclude deep learners) are employed to use the extracted features to build an accurate predictive model. Conventional ML algorithms that are commonly used to predict PTMs include support vector machine [18, 52, 53], random forest [54-56], shallow artificial neural network [57, 58], k-nearest neighbors [59, 60], logistic regression [32], and their ensembles [31].

The third category covers end-to-end approaches that rely on deep learning techniques. For the end-to-end approaches, the protein sequence is not encoded into a feature vector but rather used directly as an input to a deep neural network. These deep learners extract latent features by themselves. Many of the recent PTM predictors belong to this category. Examples include DeepNitro [34], CapsNet_PTM [30], DeepPTM [28], DeepSuccinylSite [35], MusiteDeep [29], DeepPPSite [36], MultiLyGAN [37], and nhKcr [39].

As described above, dozens of computational methods have been developed for the prediction of various types of PTM sites. However, dedicated predictors are missing for numerous PTM types, and the rapid accumulation of the experimental data motivate the need to develop many more predictors in a near future. We use the iLearnPlus platform to demonstrate how easy and convenient it is to design, develop and validate a new PTM predictor, focusing on the recently popular deep learning/end-to-end methods.

3 Design of novel predictive methods using iLearnPlus

3.1 iLearnPlus

iLearnPlus is an advanced software package that provides a comprehensive platform to analyze various structural and functional characteristics of the DNA, RNA and protein sequences and to efficiently conceptualize, design, implement and comparatively evaluate ML-based solutions for prediction of these characteristics [22]. This platform includes four modules:

- iLearnPlus-Basic for analysis and prediction using feature-based representation of protein/RNA/DNA sequences and a selected ML classifier;
- iLearnPlus-Estimator that facilitates comprehensive feature extraction from protein and nucleic acid sequences;
- iLearnPlus-AutoML that provides automated benchmarking and optimization of predictive accuracy by considering different ML algorithms and features; and
- iLearnPlus-LoadModel that enables uploading, deploying and testing models on user's own data.

Altogether, iLearnPlus supports a broad spectrum of activities including feature extraction and analysis, rational design of ML models, training and empirical assessment of ML classifiers, comparative statistical analysis of classifiers, and visualisation of data and predictive results. As a highlight, iLearnPlus covers 21 ML algorithms including 7 types of popular and modern deep learners.

iLearnPlus can be utilized by users with limited bioinformatics expertise, such as biologists and biochemists, who can take advantage of the easy and convenient to use webserver version at http://ilearnplus.erc.monash.edu/. All activities, including generation of models and testing are performed on the server side. More experienced bioinformaticians should use the command line and/or GUI (Graphical User Interface) versions that can be downloaded from https://github.com/Superzchen/iLearnPlus/.

Here, we illustrate how to use this platform to conceptualize, design and test a deep-learning based predictor of protein lysine malonylation sites.

3.2 Data collection and preprocessing

Lysine malonylation (Kmal) is a recently discovered PTM type [61] that is associated with several important cellular processes [62-65]. Only a few methods can predict the Kmal sites [31, 66-69]. Experimental data on the lysine-malonylated proteins are retrieved from mice and humans in two proteomic assays [70, 71]. Based on work in [69], the following data preprocessing steps are used to derive datasets needed for the development of predictive models:

1) Kmal-containing proteins are retrieved from the UniProt database [72], and protein sequences with sequence identities greater than 30% are removed using the CD-HIT tool [73];
2) The annotated Kmal sites are considered as positive samples, and the remaining lysine residues on the same proteins are considered as negative samples;
3) 31-residue long peptides (-15 to + 15) with the lysine site in the center are extracted for each sample. If the positive peptides are identical to the negative peptides then the negative peptides are removed;
4) All the samples are randomly divided into two parts. About 80% of all samples are subjected to five-fold cross-validation, and the remaining are used as an independent test dataset (i.e., dataset excluded from the classifier training procedure). The finalized version of the training dataset contains 4,242 positive peptides and 71,809 negative peptides, while the independent test dataset has 1,046 positive peptides and 16,827 negative peptides.

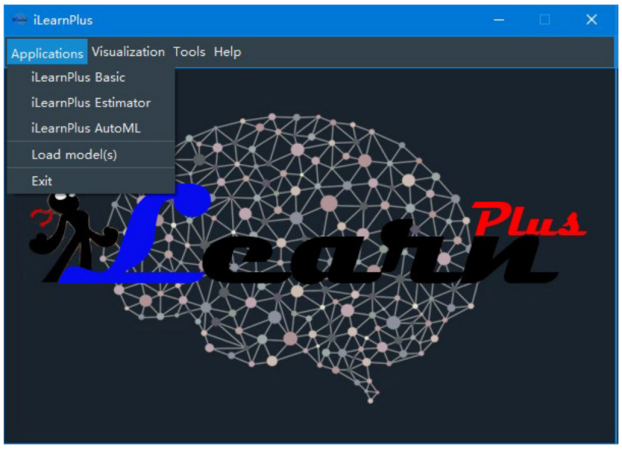

Figure 1. The main interface of iLearnPlus.

3.3 Model construction and performance evaluation

The iLearnPlus platform is used to construct the model and assess the model performance. Figure 1 shows the main GIU interface of iLearnPlus. The predictive model is built using the following sequence of nine steps:

1) Transform the positive and negative peptides into FASTA format. The FASTA header consists of three parts: part 1, part 2 and part 3, which are separated by the "|" symbol (**Figure 2**). Part 1 is the sequence name. Part 2 is the sample category information, which can be filled with any integer. For instance, users may use "1" to indicate the positive samples and "0" to represent the negative samples for a binary classification task, or use "0, 1, 2, …" to represent different classes in a multiclass classification task. Part 3 indicates the role of the sample, where for instance "training" would indicate that the corresponding sequence would be used as part of the training set in the *k*-fold cross-validation test, and "testing" that the sequence would be used as part of the independent test dataset;
2) Click "iLearnPlus Basic" (**Figure 1)** to launch the iLearnPlus-Basic module that is shown in **Figure 2**; This module facilitates generation of features from the sequences.

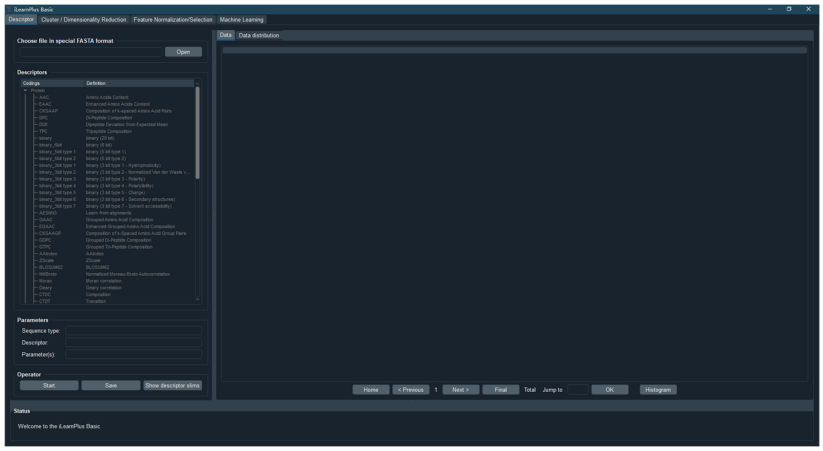

Figure 2. An example of extracting feature descriptors using the iLearnPlus-Basic module.

3) Click the “Open” button in the “Descriptor” panel, and select the file with the protein sequences. The biological sequence type (i.e. DNA, RNA or protein) is automatically detected based on the input sequences. Click the “Save” button to save the feature descriptors to a file named “binary.csv”;
4) Click the “binary” descriptor. We use the default parameters here;
5) Click the “Start” button to calculate the descriptor. This initiates a procedure to compute numeric features from the input peptide sequences, which is needed to subsequently use the ML algorithm. The feature encoding and graphical presentation are displayed in the “Data” and “Data distribution” areas, respectively;
6) Switch to the “Machine Learning” panel and load data through the “Select” button (**Figure 3**). Select “Descriptor data” in the data selection dialog box and click the “OK” button; This step moves the process to the production of the ML model from the already prepared feature-based dataset.

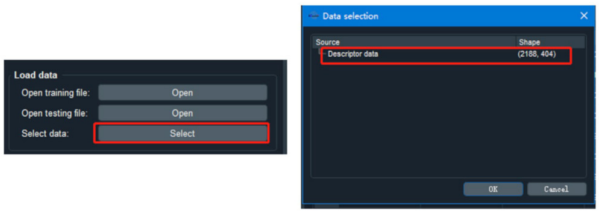

Figure 3. Load data using the data selection explorer.

7) Select “Net_1_CNN” and set the “Input channels” as 20 (**Figure 4**). This denotes that we select a modern deep convolutional neural network (CNN) model. The default values are used for the remaining parameters;

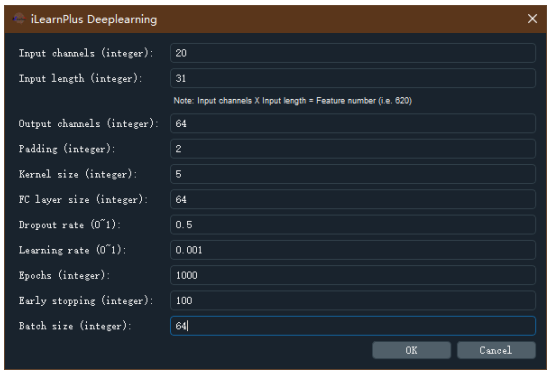

Figure 4. An example of parameter setting for the deep convolutional neural network (CNN) algorithm.

8) Set the *K* number as 5; This sets up the test procedure as the 5-fold cross validation.
9) Click the “Start” button to start the modelling. This starts the process of generating the CNN model using the training dataset in the 5-fold cross-validation setting. The resulting prediction score, performance evaluation metrics for the cross-validation and independent test and the Receiver Operating Characteristic (ROC) curves are displayed in **Figure 5**;

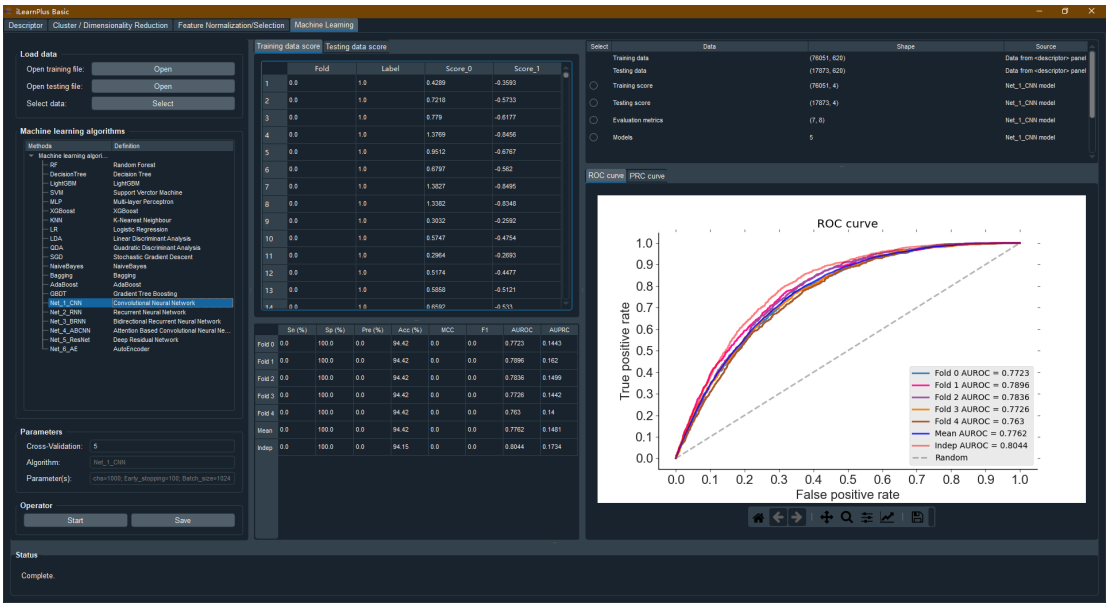

Figure 5. An example of prediction model construction using the CNN algorithm.

We observe that iLearnPlus produces a relatively accurate predictor of the Kmal sites, with the average AUROC (area under the ROC curve) over the 5 cross-validation folds of about 0.78 (**Figure 5)**. This model achieves a similar AUROC of 0.80 on the independent test dataset, which suggests that the trained model is robust and did not overfit the cross-validation experiment.

3.4 Comparison with other ML algorithms

So far, we designed and evaluated a predictive model using the CNN algorithm. Now, we use the iLearnPlus-AutoML module to compare the predictive performance between different ML algorithms. This requires executing the following sequence of five steps:

1) Open the "data.csv" file;
2) Select ML algorithms. Here, we consider five ML algorithms including "RF" (random forest), "DecisionTree" (decision three), "LightGBM" (gradient boosted forest), "LR" (logistic regression), and "CNN" (deep convolutional neural network).
3) Set the *K* number as 5; This defines the test procedure to be the 5-fold cross validation.
4) Click the "Start" button to train the five models. For each of the selected ML algorithms, the program will build the predictive model automatically, one by one, and correspondingly test them via the cross-validation protocol.
5) The results are displayed using four panels:
   - The evaluation metrics for the five classifiers are displayed in the table widget (**Figure 6**);
   - The correlation matrix of the five classifiers is displayed in the form of a heatmap (**Figure 7**); This heatmap quantifies the degree of similarity between the results/predictions produced by the five models.
   - Boxplots for the evaluation metrics (**Figure 8**); These plots quantify the spread of the values of the metrics over the five cross-validation folds.
   - ROC and precision-recall curve (PRC) curves (**Figure 9**). These curves quantify the quality of the putative scores produced by the five predictors.

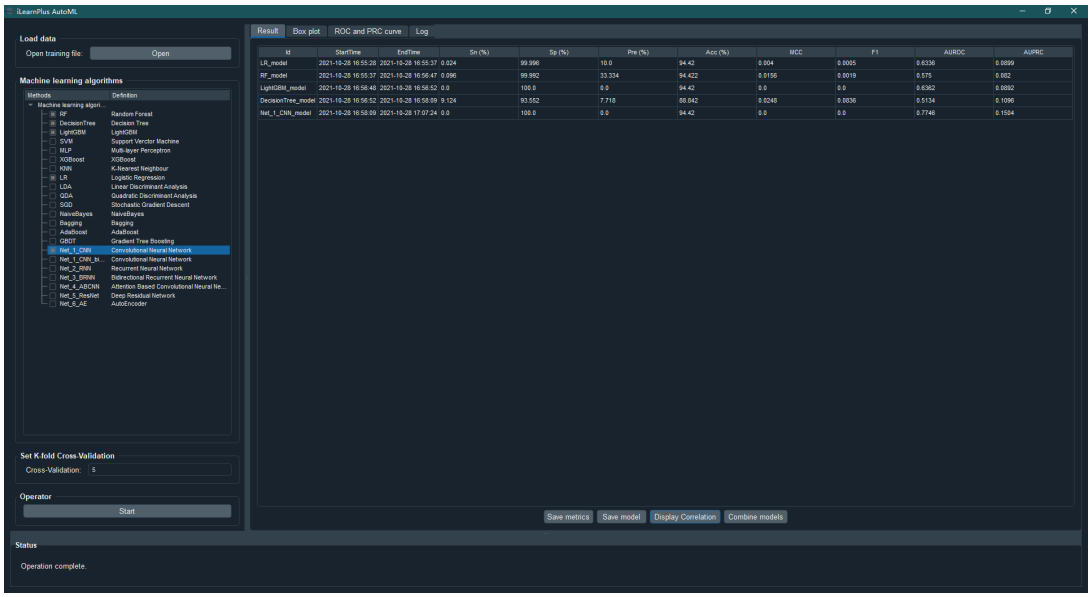

Figure 6. Performance evaluation metrics for the five models. The metrics include sensitivity (Sn), specificity (Sp), precision (Pre), accuracy (Acc), Matthew's correlation coefficient (MCC), F1, AUROC, and AUPRC (area under the precision-recall curve).

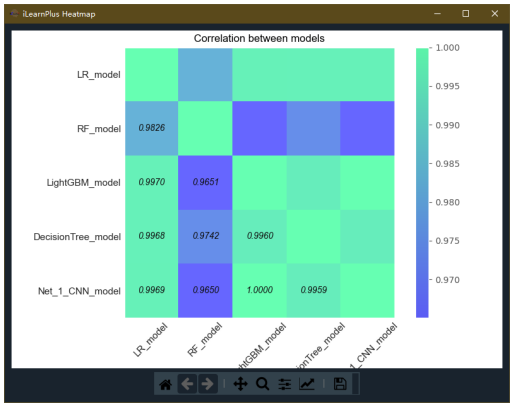

Figure 7. The correlation matrix generated by the iLearnPlus-AutoML module.

Figure 8. The boxplots generated by the iLearnPlus-AutoML module.

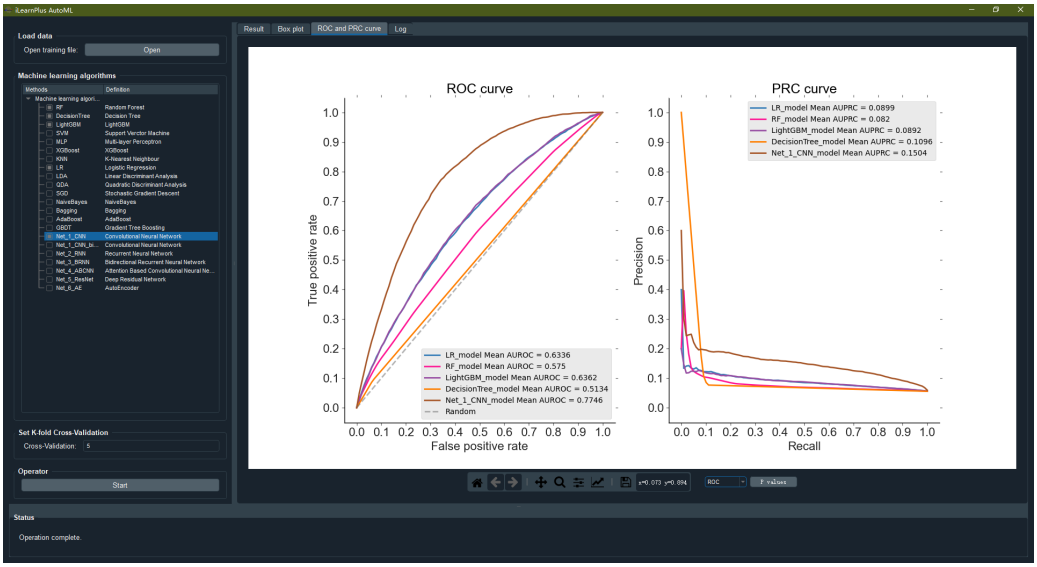

Figure 9. ROC and PR curves generated by the iLearnPlus-AutoML module to evaluate the predictive performance of the five models.

The five-step process automates numerous activities that include loading of the data, selection of ML algorithms, setting up test protocol, running training and test experiments across the five selected algorithms, and generation of a wide range of helpful metrics and plots that summarize and compare the corresponding results of the five ML algorithms. From **Figure 7**, we learn that the results generated by the various ML algorithms are highly correlated, with Pearson correlation coefficients ranging between 0.98 and 1. This is not surprising since these models solve the same problem using the same training dataset. While the five results are correlated, **Figure 6** and **Figure 9** reveal that the corresponding predictive performance is substantially different. The most accurate solution that relies on the CNN model achieves AUROC of 0.77, which agrees with the results in **Figure 5**. The other ML algorithms are not as accurate as the CNN model, with the second-best algorithm (i.e. the

gradient boosted forest) securing AUROC of 0.64 and the simplest decision tree obtaining AUROC of 0.51. This type of analysis allows the users to easily and conveniently compare different solutions, in this case by relying on different ML algorithms, to select the one that generates favorable levels of predictive quality.

4 Summary

Prediction of the PTM sites from the protein sequences is an active research area that requires the development of novel methods that would provide results for the many PTM types that lack predictors and that would take advantage of the newly released experimental data to improve over the current solutions. Generation of the predictive systems is a relatively complex process that involves collection of training and test data, various data conversions that include feature encoding, extraction and selection, modelling that covers setup and generation of predictive models using various ML algorithms, and comparative analysis to select the best solution. The execution of this entire process can be automated and facilitated with modern software platforms, such as iLearnPlus [22]. We use the example of the prediction of the lysine malonylation sites to demonstrate how to use iLearnPlus to develop accurate models and to perform comparative analysis. We find that predictors that rely on deep neural networks outperform more classical ML algorithms for this predictive task.

Importantly, we highlight the fact that iLearnPlus can be utilized to conceptualize, design, test, and deploy predictive solutions for many other related problems that extend beyond the PTM predictions. These problems cover prediction of functional and structural annotations from the proteins and nucleic acid sequences. Examples include prediction of the protein secondary structure [74-78] and other structural features of proteins [79], RNA secondary structure [80, 81], protein-nucleic acids interactions [82-84], protein-protein interactions [85-88], intrinsic disorder and its functions [89-97], cleavage sites [98], and many other annotations.

References

1. Uversky, V.N., *Posttranslational Modification*, in *Brenner's Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Editors. 2013, Academic Press: San Diego. p. 425-430.
2. Mann, M. and O.N. Jensen, *Proteomic analysis of post-translational modifications.* Nat Biotechnol, 2003. **21**(3): p. 255-61.
3. Hendriks, I.A., et al., *Site-specific mapping of the human SUMO proteome reveals co-modification with phosphorylation.* Nat Struct Mol Biol, 2017. **24**(3): p. 325-336.
4. Xu, H., et al., *PTMD: A Database of Human Disease-associated Post-translational Modifications.* Genomics Proteomics Bioinformatics, 2018. **16**(4): p. 244-251.
5. Li, F., et al., *PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact.* Brief Bioinform, 2020. **21**(3): p. 1069-1079.
6. Duan, G. and D. Walther, *The roles of post-translational modifications in the context of protein interaction networks.* PLoS computational biology, 2015. **11**(2): p. e1004049-e1004049.
7. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
8. Zhou, J.H., S.W. Zhao, and A.K. Dunker, *Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation.* Journal of Molecular Biology, 2018. **430**(16): p. 2342-2359.

9. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
10. Kuntz, E.M., et al., *Targeting mitochondrial oxidative phosphorylation eradicates therapy-resistant chronic myeloid leukemia stem cells.* Nat Med, 2017. **23**(10): p. 1234-1240.
11. Hagai, T. and Y. Levy, *Ubiquitin not only serves as a tag but also assists degradation by inducing protein unfolding.* Proc Natl Acad Sci U S A, 2010. **107**(5): p. 2001-6.
12. Galvan-Pena, S., et al., *Malonylation of GAPDH is an inflammatory signal in macrophages.* Nat Commun, 2019. **10**(1): p. 338.
13. D'Amore, C. and M. Salvi, *Editorial of Special Issue "Protein Post-Translational Modifications in Signal Transduction and Diseases".* Int J Mol Sci, 2021. **22**(5).
14. Li, F., et al., *GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome.* Bioinformatics, 2015. **31**(9): p. 1411-9.
15. Li, F., et al., *GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features.* Sci Rep, 2016. **6**: p. 34595.
16. Li, F., et al., *Positive-unlabelled learning of glycosylation sites in the human proteome.* BMC Bioinformatics, 2019. **20**(1): p. 112.
17. Gianazza, E., et al., *In silico prediction and characterization of protein post-translational modifications.* J Proteomics, 2016. **134**: p. 65-75.
18. Wang, B., M. Wang, and A. Li, *Prediction of post-translational modification sites using multiple kernel support vector machine.* PeerJ, 2017. **5**: p. e3261.
19. Zhou, F., et al., *A general user interface for prediction servers of proteins' post-translational modification sites.* Nat Protoc, 2006. **1**(3): p. 1318-21.
20. Audagnotto, M. and M. Dal Peraro, *Protein post-translational modifications: In silico prediction tools and molecular modeling.* Comput Struct Biotechnol J, 2017. **15**: p. 307-319.
21. He, W., L. Wei, and Q. Zou, *Research progress in protein posttranslational modification site prediction.* Brief Funct Genomics, 2018. **18**(4): p. 220-229.
22. Chen, Z., et al., *iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization.* Nucleic Acids Res, 2021. **49**(10): p. e60.
23. Li, F., et al., *Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome.* Bioinformatics, 2018. **34**(24): p. 4223-4231.
24. Song, J., et al., *PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy.* Bioinformatics, 2018. **34**(4): p. 684-687.
25. Li, F., et al., *DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites.* Bioinformatics, 2020. **36**(4): p. 1057-1065.
26. Li, F., et al., *Procleave: Predicting Protease-specific Substrate Cleavage Sites by Combining Sequence and Structural Information.* Genomics Proteomics Bioinformatics, 2020. **18**(1): p. 52-64.
27. Ahmed, S., et al., *DeepPPSite: A deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information.* Anal Biochem, 2020. **612**: p. 113955.
28. Baisya, D.R. and S. Lonardi, *Prediction Of Histone Post-Translational Modifications Using Deep Learning.* Bioinformatics, 2020.
29. Wang, D., et al., *MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization.* Nucleic Acids Res, 2020. **48**(W1): p. W140-W146.
30. Wang, D., Y. Liang, and D. Xu, *Capsule network for protein post-translational modification site prediction.* Bioinformatics, 2019. **35**(14): p. 2386-2394.
31. Zhang, Y., et al., *Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework.* Brief Bioinform, 2019. **20**(6): p. 2185-2199.

32. Li, F., et al., *Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome.* Bioinformatics, 2018.
33. Lo Monte, M., et al., *ADPredict: ADP-ribosylation site prediction based on physicochemical and structural descriptors.* Bioinformatics, 2018. **34**(15): p. 2566-2574.
34. Xie, Y., et al., *DeepNitro: Prediction of Protein Nitration and Nitrosylation Sites by Deep Learning.* Genomics Proteomics Bioinformatics, 2018. **16**(4): p. 294-306.
35. Thapa, N., et al., *DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction.* BMC Bioinformatics, 2020. **21**(Suppl 3): p. 63.
36. Ahmed, S., et al., *DeepPPSite: A deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information.* Anal Biochem, 2021. **612**: p. 113955.
37. Yang, Y., et al., *Prediction and analysis of multiple protein lysine modified sites based on conditional wasserstein generative adversarial networks.* BMC Bioinformatics, 2021. **22**(1): p. 171.
38. Li, F., et al., *Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods.* Brief Bioinform, 2019. **20**(6): p. 2150-2166.
39. Chen, Y.Z., et al., *nhKcr: a new bioinformatics tool for predicting crotonylation sites on human nonhistone proteins based on deep learning.* Brief Bioinform, 2021. **22**(6).
40. Chen, Z., et al., *Large-scale comparative assessment of computational predictors for lysine post-translational modification sites.* Brief Bioinform, 2019. **20**(6): p. 2267-2290.
41. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
42. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
43. Xue, Y., et al., *GPS: a comprehensive www server for phosphorylation sites prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W184-7.
44. Deng, W., et al., *Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins.* Brief Bioinform, 2017. **18**(4): p. 647-658.
45. Zhao, Q., et al., *GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W325-30.
46. Li, T., et al., *Characterization and prediction of lysine (K)-acetyl-transferase specific acetylation sites.* Mol Cell Proteomics, 2012. **11**(1): p. M111 011080.
47. Liu, Z., et al., *GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins.* Mol Biosyst, 2011. **7**(10): p. 2737-40.
48. Xiao, N., et al., *protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences.* Bioinformatics, 2015. **31**(11): p. 1857-1859.
49. Chen, Z., et al., *iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences.* Bioinformatics, 2018. **34**(14): p. 2499-2502.
50. Liu, B., X. Gao, and H. Zhang, *BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches.* Nucleic Acids Res, 2019. **47**(20): p. e127.
51. Chen, Z., et al., *iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data.* Brief Bioinform, 2020. **21**(3): p. 1047-1057.
52. Cao, W., et al., *Prediction of N-myristoylation modification of proteins by SVM.* Bioinformation, 2011. **6**(5): p. 204-6.
53. Chen, X., et al., *Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity.* Bioinformatics, 2015. **31**(2): p. 194-200.
54. Chang, C.C., et al., *SUMOgo: Prediction of sumoylation sites on lysines by motif screening models and the effects of various post-translational modifications.* Sci Rep, 2018. **8**(1): p. 15512.
55. Hamby, S.E. and J.D. Hirst, *Prediction of glycosylation sites using random forests.* BMC Bioinformatics, 2008. **9**: p. 500.

56. Wang, Y.G., et al., *Accurate prediction of species-specific 2-hydroxyisobutyrylation sites based on machine learning frameworks.* Anal Biochem, 2020. **602**: p. 113793.
57. Basu, S. and D. Plewczynski, *AMS 3.0: prediction of post-translational modifications.* BMC Bioinformatics, 2010. **11**: p. 210.
58. Dewhurst, H.M. and M.P. Torres, *Systematic analysis of non-structural protein features for the prediction of PTM function potential by artificial neural networks.* PLoS One, 2017. **12**(2): p. e0172572.
59. Feng, K.Y., et al., *Using WPNNA classifier in ubiquitination site prediction based on hybrid features.* Protein Pept Lett, 2013. **20**(3): p. 318-23.
60. Niu, S., et al., *Prediction of tyrosine sulfation with mRMR feature selection and analysis.* J Proteome Res, 2010. **9**(12): p. 6490-7.
61. Peng, C., et al., *The first identification of lysine malonylation substrates and its regulatory enzyme.* Mol Cell Proteomics, 2011. **10**(12): p. M111 012658.
62. Nie, L.B., et al., *Global Proteomic Analysis of Lysine Malonylation in Toxoplasma gondii.* Front Microbiol, 2020. **11**: p. 776.
63. Ma, Y., et al., *Malonylome Analysis Reveals the Involvement of Lysine Malonylation in Metabolism and Photosynthesis in Cyanobacteria.* J Proteome Res, 2017. **16**(5): p. 2030-2043.
64. Qian, L., et al., *Global Profiling of Protein Lysine Malonylation in Escherichia coli Reveals Its Role in Energy Metabolism.* J Proteome Res, 2016. **15**(6): p. 2060-71.
65. Hirschey, M.D. and Y. Zhao, *Metabolic Regulation by Lysine Malonylation, Succinylation, and Glutarylation.* Mol Cell Proteomics, 2015. **14**(9): p. 2308-15.
66. Chung, C.R., et al., *Incorporating hybrid models into lysine malonylation sites prediction on mammalian and plant proteins.* Sci Rep, 2020. **10**(1): p. 10541.
67. Ahmad, W., et al., *Mal-Light: Enhancing Lysine Malonylation Sites Prediction Problem Using Evolutionary-based Features.* IEEE Access, 2020. **8**: p. 77888-77902.
68. Xiang, Q., et al., *Prediction of Lysine Malonylation Sites Based on Pseudo Amino Acid.* Comb Chem High Throughput Screen, 2017. **20**(7): p. 622-628.
69. Chen, Z., et al., *Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting Malonylation Sites.* Genomics Proteomics Bioinformatics, 2018. **16**(6): p. 451-459.
70. Nishida, Y., et al., *SIRT5 Regulates both Cytosolic and Mitochondrial Protein Malonylation with Glycolysis as a Major Target.* Mol Cell, 2015. **59**(2): p. 321-32.
71. Colak, G., et al., *Proteomic and Biochemical Studies of Lysine Malonylation Suggest Its Malonic Aciduria-associated Regulatory Role in Mitochondrial Function and Fatty Acid Oxidation.* Mol Cell Proteomics, 2015. **14**(11): p. 3056-71.
72. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.
73. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-2.
74. Kashani-Amin, E., et al., *A systematic review on popularity, application and characteristics of protein secondary structure prediction tools.* Curr Drug Discov Technol, 2018. **16**(2): p. 159-172.
75. Zhang, H., et al., *Critical assessment of high-throughput standalone methods for secondary structure prediction.* Brief Bioinform, 2011. **12**(6): p. 672-88.
76. Oldfield, C.J., K. Chen, and L. Kurgan, *Computational Prediction of Secondary and Supersecondary Structures from Protein Sequences.* Methods Mol Biol, 2019. **1958**: p. 73-100.
77. Meng, F. and L. Kurgan, *Computational Prediction of Protein Secondary Structure from Sequence.* Curr Protoc Protein Sci, 2016. **86**: p. 2 3 1-2 3 10.
78. Jiang, Q., et al., *Protein secondary structure prediction: A survey of the state of the art.* J Mol Graph Model, 2017. **76**: p. 379-402.
79. Kurgan, L. and F.M. Disfani, *Structural protein descriptors in 1-dimension and their sequence-based predictions.* Curr Protein Pept Sci, 2011. **12**(6): p. 470-89.
80. Zhao, Q., et al., *Review of machine learning methods for RNA secondary structure prediction.* PLoS Comput Biol, 2021. **17**(8): p. e1009291.

81. Tahi, F., T.T.V. Du, and A. Boucheham, *In Silico Prediction of RNA Secondary Structure.* Methods Mol Biol, 2017. **1543**: p. 145-168.
82. Yan, J., S. Friedrich, and L. Kurgan, *A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues.* Brief Bioinform, 2016. **17**(1): p. 88-105.
83. Zhang, J., Z. Ma, and L. Kurgan, *Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains.* Brief Bioinform, 2019. **20**(4): p. 1250-1268.
84. Miao, Z. and E. Westhof, *A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs.* PLoS Comput Biol, 2015. **11**(12): p. e1004639.
85. Zhang, J., S. Ghadermarzi, and L. Kurgan, *Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins.* Bioinformatics, 2020. **36**(18): p. 4729-4738.
86. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.
87. Chen, H., et al., *Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions.* Brief Bioinform, 2020.
88. Fernandez-Recio, J., *Prediction of protein binding sites and hot spots.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2011. **1**(5): p. 680-698.
89. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.
90. He, B., et al., *Predicting intrinsic disorder in proteins: an overview.* Cell Res, 2009. **19**(8): p. 929-49.
91. Zhao, B. and L. Kurgan, *Surveying over 100 predictors of intrinsic disorder in proteins.* Expert Rev Proteomics, 2021: p. 1-11.
92. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Computational prediction of functions of intrinsically disordered regions.* Prog Mol Biol Transl Sci, 2019. **166**: p. 341-369.
93. Kurgan, L., M. Li, and Y. Li, *The Methods and Tools for Intrinsic Disorder Prediction and their Application to Systems Medicine*, in *Systems Medicine*, O. Wolkenhauer, Editor. 2021, Academic Press: Oxford. p. 159-169.
94. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins.* Brief Bioinform, 2010. **11**(2): p. 225-43.
95. Dosztányi, Z. and P. Tompa, *Bioinformatics Approaches to the Structure and Function of Intrinsically Disordered Proteins*, in *From Protein Structure to Function with Bioinformatics*, D. J. Rigden, Editor. 2017, Springer Netherlands: Dordrecht. p. 167-203.
96. Liu, Y., X. Wang, and B. Liu, *A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction.* Brief Bioinform, 2019. **20**(1): p. 330-346.
97. Atkins, J.D., et al., *Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies.* Int J Mol Sci, 2015. **16**(8): p. 19040-54.
98. Bao, Y., et al., *Toward more accurate prediction of caspase cleavage sites: a comprehensive review of current methods, tools and features.* Brief Bioinform, 2019. **20**(5): p. 1669-1684.