2024 OXCAM Programme – AI+ Course
Group Project Report

Programme	**Biotechnology Engineering & Healthcare Technology**		
Course Group	**BIOA-G3**	**Group Name**	**Stray Birds**
Group Members	**Dong Shuyi , Wang Xinyi , Yang Keyi , Qin Hanqing , Zheng Xi , Zhou Yangyang**		

Analysis of phase separation proteins based on machine learning to help assess neurodegeneration

1 Project Summary

Neurodegenerative diseases pose a growing burden on healthcare systems worldwide, such as Alzheimer's disease(AD), amyotrophic lateral sclerosis(ALS) □etc. Most neurodegenerative diseases are associated with aberrant phase separation(PS) proteins, leading researchers to study those proteins through multiple ways.
Here, we propose the prediction of aberrant phase separation proteins by using machine learning models based on decision tree. We demonstrate statistical analysis of protein characteristics in database to select seven biomarkers for data collection. Then, we prepare reliable datasets and input them to different models. By evaluating and visualizing the results ,we achieve the XGboost model with 70.88% accuracy. Ultimately, we obtain a diagnostic tool to help assess neurodegenerative diseases clinically, thus providing a new insight for researchers to explore the mechanism and promote the detection of rare neurodegenerative diseases globally.(Shen et al., 2021)(Pancsa, Vranken, & Mészáros, 2021)

2 Project Results

2.1 Introduction

Liquid-liquid phase separation (LLPS) is a reversible process of a homogeneous fluid de-mixing into two distinct liquid phases : one condensed phase and one dilute phase. Many biological processes have been revealed to be regulated by LLPS, while in some situations the liquid condensates can transform into solid aggregates aberrantly, which are implicated in related to neurodegenerative diseases.(Ding, Xu, Pei, & Li, 2023)(Gao, Li, Li, & Lin, 2022)Previous work has been conducted to screen out phase separation proteins , further challenge lies in understanding the conditions under which a protein changes from normal PS to abnormal PS, and what modifications (such as PTMs) occur in these conditions.(Chen et al., 2022)(Q. Li et al., 2020)(You et al., 2020) Our project demonstrates the prediction of aberrant phase separation proteins under different conditions and post-translational modification(PTMs)(Powell et al., 2024). Through different machine learning models, we finally achieve the results and performances of different models, we choose the model with best accuracy of 70.88% as an ideal predictor.

2.2 Methodology

XGBoost classification model.

(1)Choose bio-markers:We combine statistical analysis of protein characteristics with literature screening of bio-markers in database to select appropriate features.Since current phase separation predictors are mostly based on sequence-dependent features of proteins and lack of PTMs(post-translational modifications), features other than sequence could provide crucial information for identifying aberrant phase separation proteins.We propose PTMs (post-translational modifications) with specific types, partner of LLPS and IF (immunofluorescence) images as promising features to be incorporated into phase separation prediction .

(2)Data collection: By searching literatures, we found three comprehensive databanks for PTMs: LLPSDB, PhaSepDB and PhasePro and the Human Protein Atlas as a databank of immunofluorescence (IF) images. After screening and processing those databanks, 787 proteins data of 7 bio-markers were collected, including material-state of phase separation,partner, PTMs and specific types:phosphorylation(Phos), acetylation(Ac), methylation(Me), sumoylation(SUMO). Meanwhile, we select 30 immunofluorescence (IF) images of both normal and abnormal phase separation proteins. Finally we prepared the training datasets and grouped out 20%of them as testing datasets.

(Boyko & Surewicz, 2022)(X. Li, Du, Chen, Liu, et al., 2023)

Figure 1: Data pre-processing. Seven protein features are used in the dataset for model training: material-state, partner, PTMs, Phos (phosphorylation), Ac (acetylation), Me (methylation) and SUMO(SUMO). Among them, the last four features are different types of PTMs. We treat different feature data as digital forms of 0 and 1, and the corresponding contents of different features are shown in the following table.

Number	material-state	partner	PTMs	Phos	Ac	Me	SUMO
0	normal LLPS	no partner	no PTMs	no Phos	no Ac	no Me	no SUMO
1	abnormal LLPS	with partner	with PTMs	with Phos	with Ac	with Me	with SUMO

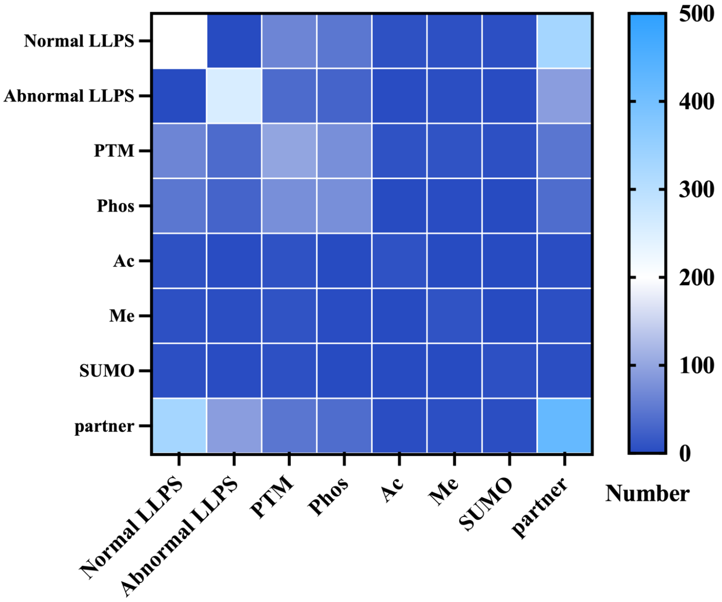


Figure 2: Heat map of data set used in model training. Described the number of proteins corresponding to different aspects.

(3) Model training:We use different machine learning models to make prediction, including decision tree(Zhou, V. (2022). Decision Tree - ID3.py [Python script]. In Zhouxiaonnan/machine-learning-notesandcode (GitHub repository). Retrieved from https://github.com/Zhouxiaonnan/machine-learning-notesandcode/Decision%20Tree/%E5%A3%B0%E6%%98 %20ID3.py),random forest(guofei9987. (2017, October 20). Random Forest Theory and Implementation [Web log post]. Retrieved from https://www.guofei.site/2017/10/20/randomforest.html) , XGBoost(Oliwia, S. (2021). xgboost-AutoTune [GitHub repository]. GitHub. https://github.com/SylwiaOliwia2/xgboost-AutoTune), and Convolutional Neural Network(CNN)(Doe, J. (2023). introgression/train.neural.net.introgression.CNN.PYTHON3.py [Python script],linear(Xu, H. (Ralph), F. M. (2022). Chapter 4: Multicollinearity. In Econometrics Regression Analysis [Text file]. Retrieved from http://www.example.com/path/to/Econometrics-Regression-Analysis/chapter4.md). Retrieved from http://www.example.com/path/to/script). We searched for our models' code on GitHub, then downloaded those codes and corresponding packages to create the required environment. By using Python based on Visual Studio Code, we input the prepared datasets and train the models . Then achieving prediction accuracy and results.By using desicion tree, random forest and XGBoost to predict aberrant phase separation through datasets of partner and PTMs,we get different prediction results of those models.We also use CNN based on immunofluorescence (IF) images dataset to predict aberrant phase separation and achieve results.

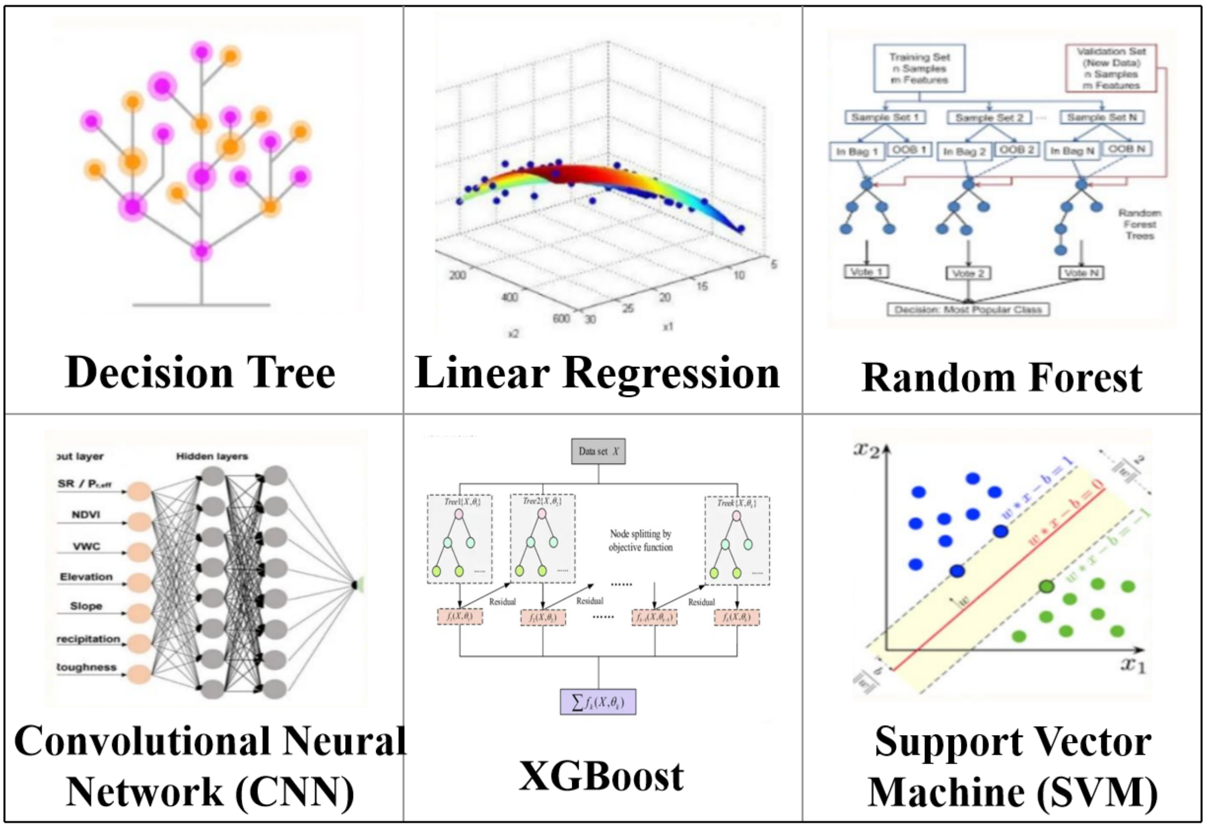


Figure 3: Different models
https://img2020.cnblogs.com/blog/1504684/202003/1504684-20200316075712580-403570783.png
https://www.edrawsoft.cn/wp/wp-content/uploads/2020/09/suijisenlinhfa.png
https://github.com/fengdu78/Coursera-ML-AndrewNg-Notes/raw/master/images/01105c3afd1315acf0577f8493137dcc.png
https://i-blog.csdnimg.cn/direct/4317da33d83c41f0b5a6f8598b676e89.png
https://blog.csdn.net/weixin$_4$0493805$/article/details/$121927256
$https://www.mdpi.com/$2624 − 7402/5/4/109

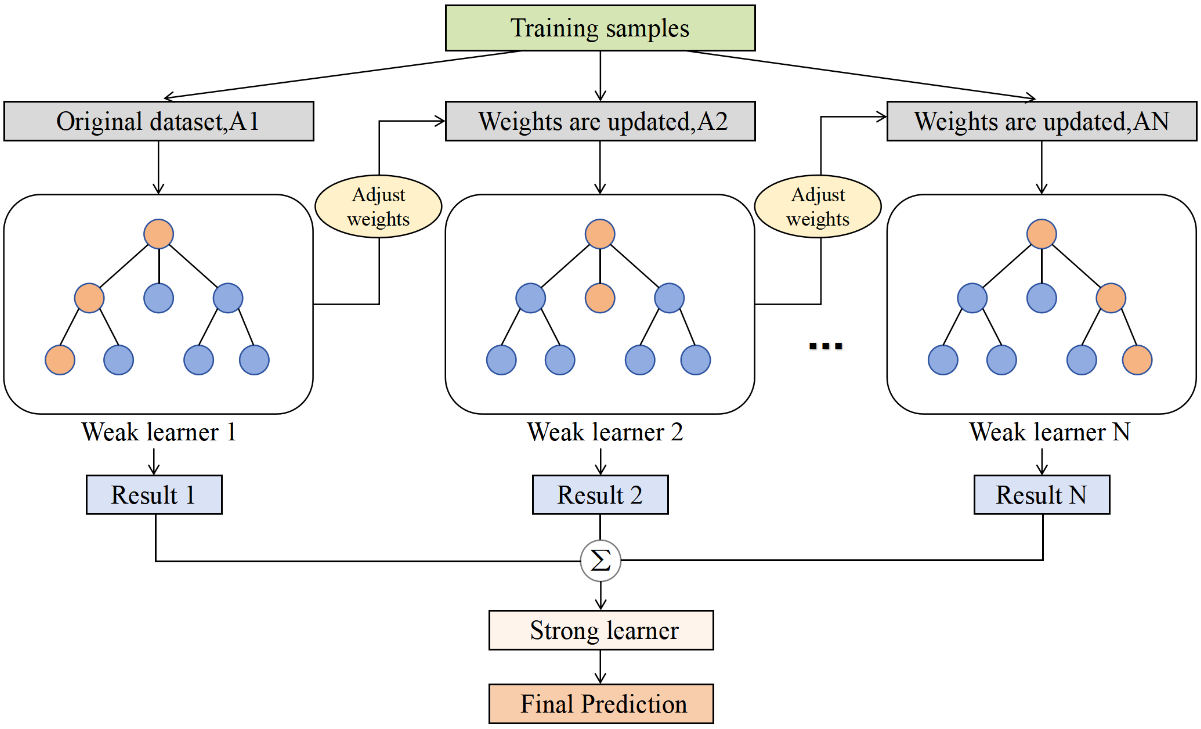


Figure 4: model principle
The XG Boost model. We combine several weak learners into one strong learner by a certain method. That is, multiple trees make decisions together, and the result of each tree is the difference between the target value and the predicted result of all the previous trees, all the results are added together to get the final result, so as to achieve the improvement of the effect of the whole model.

(4) Model evaluation :We input the testing datasets of experimental data to evaluate the performance of model prediction and compare prediction accuracy and performance of different models.Multiple methods, such as heat maps and pie charts, are used to visualize the results.

2.3 Results Our results are mainly divided into two parts:dataset results and model results.
2.3.1 Dataset results Through results analysis, we achieve different weights of seven features,and the relationship between them.Among seven features,partner accounts for the highest one, meaning the importance of partner for phase separation.The weights of phosphorylation(Phos),acetylation(Ac), methylation(Me) and sumoylation(SUMO) demonstrate the potential impact of PTMs for aberrant phase separation.

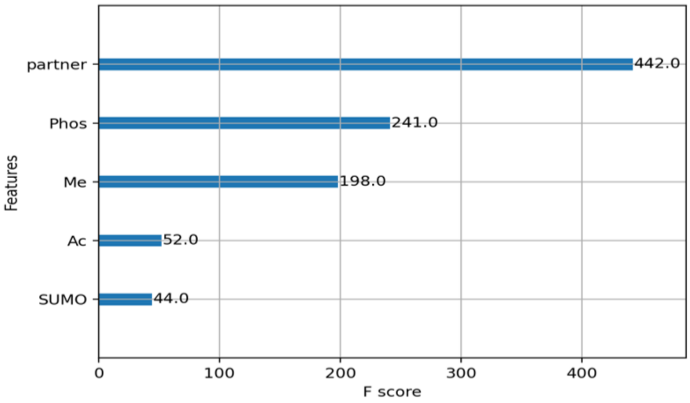


Figure 5: This is the weights of corresponding features and the potential relationship of them for aberrant phase separation.

2.3.2 Model results We get the code of various models from diversified open source websites for our comparison and selection, and then modify the codes according to our subject. We conducted multiple training sessions to improve the learning ability of the machine (fig.2). We tried five models of machine learning, which included the decision tree, random forest, convolutional neural network, support vector machine, XGBoost, and linear regression.(fig.3) As shown in the figure is the machine learning accuracy of various models. Finally, we selected the model XGBoost and expanded it through one 0 and 1 decisions, and finally determined the weight of each feature.

矩形截图 Ctrl + Alt + A

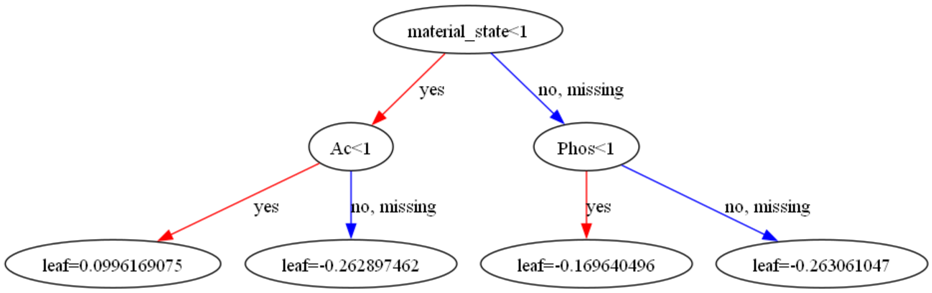


Figure 6: XGBoost-Tree
This is the result of XGBoost ,if material state is hydrogel or solid we define as yes, if it is liquid we define as no.Then based on 0 or 1 numbers we classify different types of proteins to conduct this decision tree

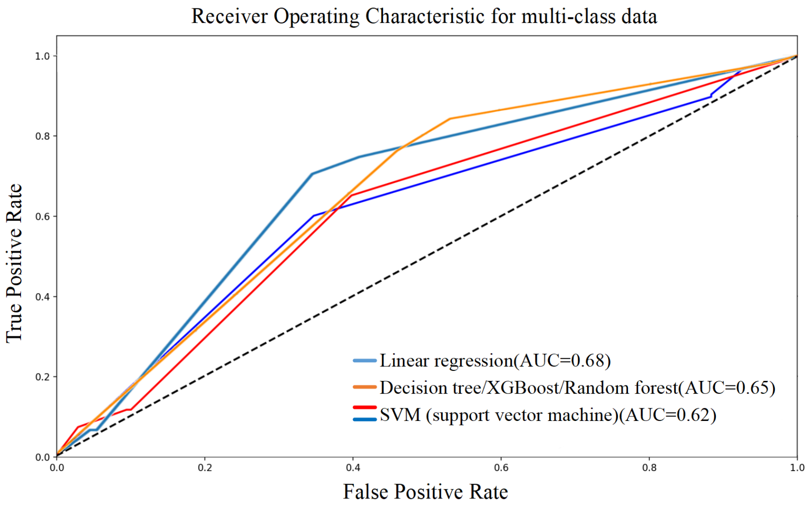
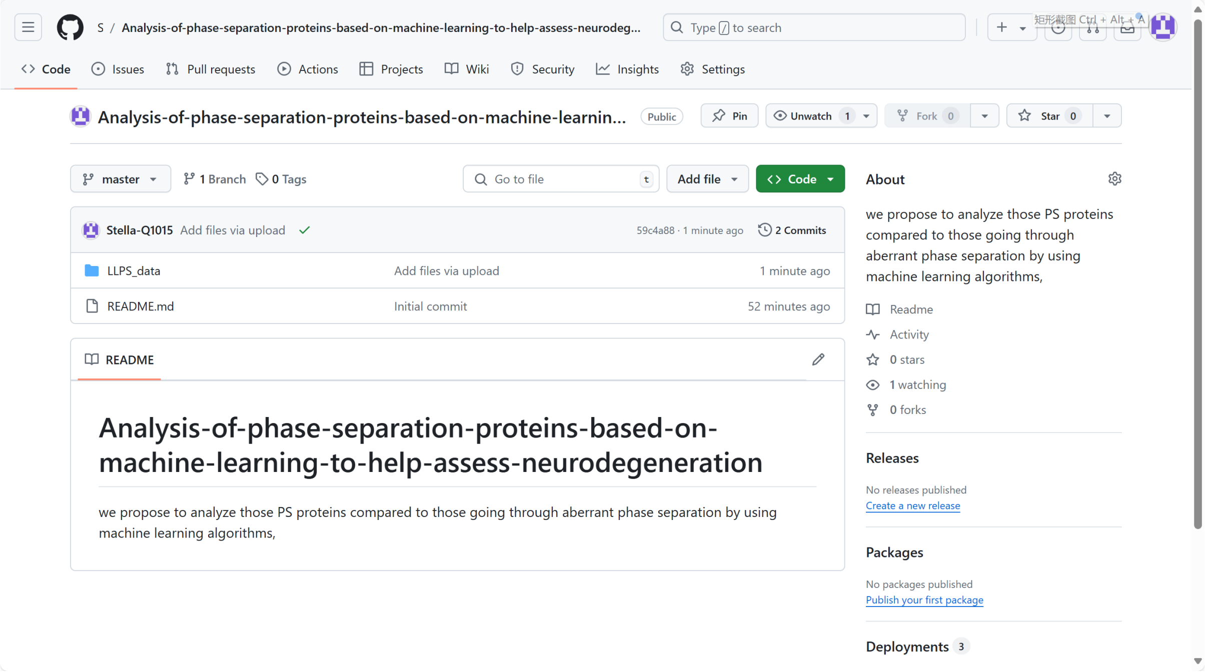


Figure 7: False Positive Rate
To measure the performance of those model prediction, we plotted the ROC curve for each predictor,showing the great prediction performance of XGBoost model.

2.3.3 Model code publication on GitHub

We finally publish our project result on GitHub and upload the files of our project as open source.



Analysis-of-phase-separation-proteins-based-on-machine-learnin...

Stella-Q1015 Add files via upload ✓		59c4a88 · 1 minute ago 2 Commits
LLPS_data	Add files via upload	1 minute ago
README.md	Initial commit	52 minutes ago

About

we propose to analyze those PS proteins compared to those going through aberrant phase separation by using machine learning algorithms,

Analysis-of-phase-separation-proteins-based-on-machine-learning-to-help-assess-neurodegeneration

we propose to analyze those PS proteins compared to those going through aberrant phase separation by using machine learning algorithms,

Figure 8: Model code publication on GitHub

We publish the codes of trained model on GitHub and upload our programme in files to achieve the final results of our project.

https://github.com/Stella-Q1015/Analysis-of-phase-separation-proteins-based-on-machine-learning-to-help-assess-neurodegeneration

2.4 Discussions

- 2.4.1 Model improvement
  (1)Datasets optimization: Due to the incompleteness and incompatibility of databank,we could not find an appropriate way to score the level of post-translational modifications(PTMs) with specific numbers, the datasets we obtained still lack of detailed information and the volume need to be increased.We suggest to optimize the limitation of databank and integrate data from different databases to complement each other. This includes consolidating information on protein phase separation conditions. Specifically, focus on consolidating data related to biophysical conditions for proteins, including concentration, temperature, interacting partners, and post-translational modifications (PTMs).
  (2)Model comparison: By comparing the prediction accuracy of different models based on PTMs and partner database and identifying the factors influencing their performance□we present prediction accuracy through bar graph and model performance by Receiver Operating Characteristic(ROC) curve.
- 2.4.3 Application and significance
  System Application based on predictive model:
  (1)Patients provide biological samples.
  (2)Analyze samples' protein
  (3)AI prediction model assesses aberrant phase separation protein .
  (4)assist disease treatment and scientific research.
  Futural significance of predictive tool:
  (1)Disease Mechanism Understanding: Predicting how abnormal LLPS contributes to diseases can provide insights into the molecular mechanisms behind conditions such as neurodegenerative diseases and cancers.
  (2)Drug Development: Understanding the role of abnormal LLPS in disease can guide the development of drugs targeting these aberrant phase separation proteins.
  (3)tailored treatment: AI predictions help aid in creating personalized treatment plans by identifying individual-specific LLPS-related abnormalities, leading to tailored therapeutic strategies.
  (4)Diagnostic Tools: First, train machine learning models based on biological databases. After collecting biological samples from patients and analyzing protein characteristics, use the model to assess whether proteins exhibit abnormal phase separation for pathological research. Subsequently, provide point-to-point disease treatment.

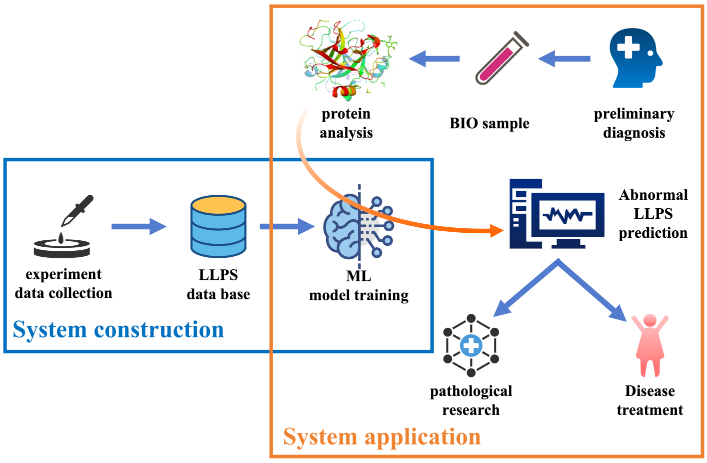


Figure 9: Healthcare application of phase separation proteins analysis system
A idea flowchart of possible directions for the application of the abnormal LLPS prediction project. It is divided into two parts: system construction and system application. System construction refers to the work done in this project, which involves collecting LLPS data, processing LLPS data, and conducting machine learning to provide a predictive model for abnormal phase separation. We believe that this model can have certain applications in pathological research and clinical treatment.

- 2.4.4 Ethics
  (1) Data Privacy and Protection: The protein data used in the research involve personal biological information, which is of great significance to set regulations to protect data.
  (2)Model Bias: The datasets based on well-studied experimental proteins could affect accuracy and fairness, leading to misdiagnosis or missed diagnoses for different certain groups.
  (3)Misuse of Technology: Research findings might be improperly used in non-medical contexts, such as insurance or employment discrimination, meaning that clear guidelines are needed to prevent misuse.
  (4)Diagnostic Accuracy: Although the model achieves a 70.88% accuracy rate, there is still room for improvement.Misdiagnosis or missed diagnoses could have serious consequences for patients, requiring extensive validation before clinical use.
  (5)Intellectual Property and Conflicts of Interest: Research outcomes may involve patent and intellectual property issues, requiring clarification of researchers' rights and potential conflicts of interest.

2.5 Limitations

- 2.5.1 Databank Incompatibility and Loss of Data
  We use LLPSDB, PhaSepDB, and PhasePro as chosen databank. However, each datasets has its limitations which complicate the process of integrating and utilizing the data effectively. LLPSDB and PhasePro are unable to determine material-state of proteins,the only databank we can find material-state is PhasepDB, which though lack data of conditions and level of post-translational modifications (PTMs). The absence of PTMs rating and conditions leading to a potential loss of valuable information. Each dataset provides valuable insights, but the gaps and limitations in the data make it challenging to build a cohesive understanding of protein phase separation.
- 2.5.2 Modelling Limitations
  Given incompleteness and sometimes incompatible nature of the data from these various databank, modeling and analysis of phase separation events are inherently limited. Predictive models may struggle to accurately assess the impact of specific PTMs or environmental conditions on phase separation because these factors are not uniformly represented across the datasets.The integration of these datasets into a single model requires sophisticated data processing and normalization techniques to mitigate the effects of these gaps.
- 2.5.3 Setting Thresholds for Accuracy and Reliability of Prediction Results
  It is crucial to set clear thresholds for accuracy and reliability in our predictive models. This involves defining acceptable levels for predictions based on the quality and completeness of the data used. For instance, when integrating data from LLPSDB, PhaSepDB, and PhasePro, thresholds should be established to determine the minimum quality of data required for a reliable prediction. By setting appropriate thresholds for accuracy, we can better manage the impact and enhance the overall reliability of predictions.

2.6 Conclusions

To sum up, we achieve a model with 70.88% accuracy and develop a diagnostic tool to predict phase separation proteins under different conditions and post-translational modification(PTMs).Futural regulations should be considered about intellectual property of our predictorand the protection of patients' data privacy.It should also be noted that models based on experimental evidences usually bias to well-studied proteins, which may impact the fairness of prediction. Ultimately we provide a new insight for researchers to explore the mechanism and promote the detection of rare neurodegenerative diseases globally.

3 Project skills

- 3.1 Team shared learning
  Our team focuses on collective learning and continuous improvement. Each team member utilizes their own strengths to master different sections and subsequently shares their knowledge and skills with peers through shared documents, group meetings, and team-building exercises. Throughout the project, we adhere strictly to the Gantt chart for scheduling. After completing each stage of progress, the team timely communicates and follows up to conduct a thorough summary. We compile datasets grounded in biological analysis and statistical science, input the datasets into our chosen model, and commence training. Finally, the team conducts a comprehensive analysis of the results and evaluates the performance of the predictive model. By incorporating AI and ML approaches, our team adopts their application to practical project achievements, enabling more precise and data-driven solutions and effectively harnessing AI and machine learning tools.
- 3.2 Team overcome difficulties
  In order to build an efficient team and ensure that all members achieve the project objectives, our team has overcome many challenges. Following timely communication within the team, we screen for suitable biomarkers within complex databases and continuously work to digitize text data into a format suitable for machine learning. After several rounds of code debugging, we have ultimately developed a model with high accuracy.
- 3.3 Team development and improvement
  Through previous online learning, our team has mastered the principles of AI and developed a solid understanding of machine learning algorithms, ensuring a solid foundation for practical applications. This golden opportunity allows our team to delve into the fascinating fields of AI and ML. The application of machine learning empowers us to solve complex challenges with innovative and efficient approaches, enabling us to further understand how to align AI capabilities with our project's strategic goals to drive innovation. It also allows us to explore complex interactions and non-linear effects using large data sets.

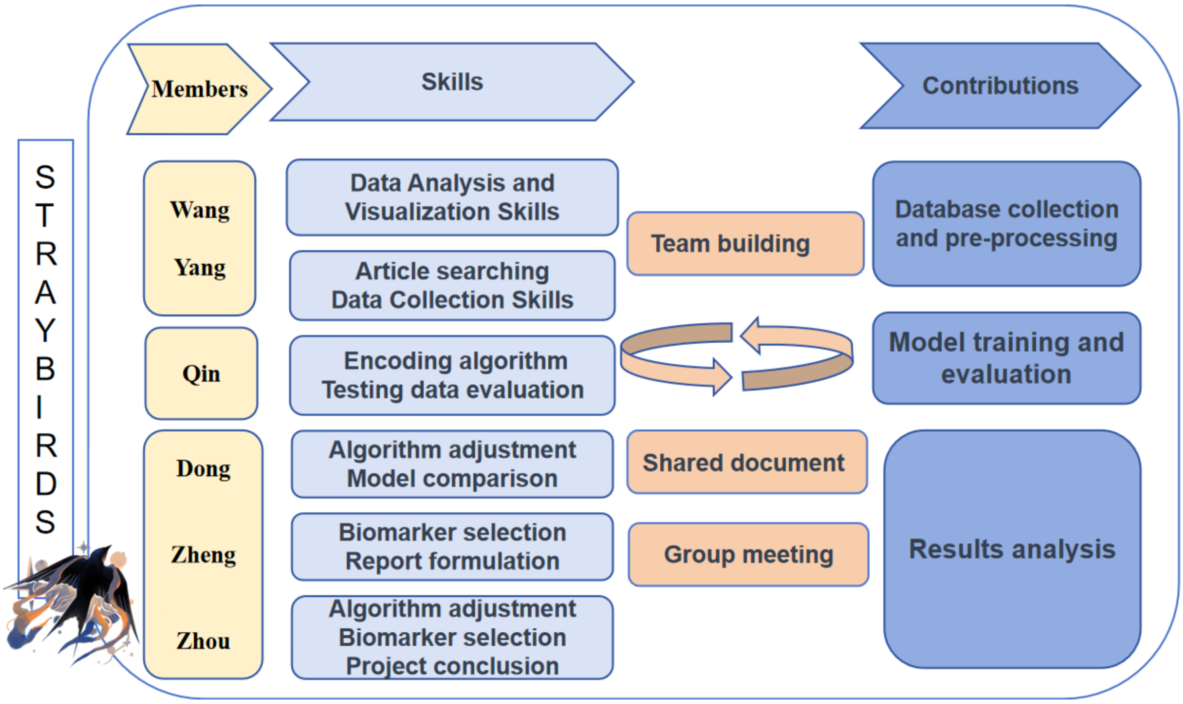

STRAYBIRDS
Members
Skills
Contributions
Wang
Yang
Qin
Dong
Zheng
Zhou
Data Analysis and Visualization Skills
Article searching Data Collection Skills
Encoding algorithm Testing data evaluation
Algorithm adjustment Model comparison
Biomarker selection Report formulation
Algorithm adjustment Biomarker selection Project conclusion
Team building
Shared document
Group meeting
Database collection and pre-processing
Model training and evaluation
Results analysis

Figure 10: Project skills

4 References

References

Boyko, S., & Surewicz, W. K. (2022). Tau liquid–liquid phase separation in neurodegenerative diseases. *Trends in cell biology*, *32*(7), 611–623.

Chen, Z., Hou, C., Wang, L., Yu, C., Chen, T., Shen, B., … Li, T. (2022). Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proceedings of the National Academy of Sciences*, *119*(24), e2115369119.

Ding, M., Xu, W., Pei, G., & Li, P. (2023). Long way up: rethink diseases in light of phase separation and phase transition. *Protein & Cell*, pwad057.

Gao, Y., Li, X., Li, P., & Lin, Y. (2022). A brief guideline for studies of phase-separated biomolecular condensates. *Nature Chemical Biology*, *18*(12), 1307–1318.

Li, Q., Peng, X., Li, Y., Tang, W., Zhu, J., Huang, J., … Zhang, Z. (2020). Llpsdb: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic acids research*, *48*(D1), D320–D327.

Li, X., Du, Y., Chen, X., Liu, C., et al. (2023). Emerging roles of o-glycosylation in regulating protein aggregation, phase separation, and functions. *Current Opinion in Chemical Biology*, *75*, 102314.

Pancsa, R., Vranken, W., & Mészáros, B. (2021). Computational resources for identifying and describing proteins driving liquid–liquid phase separation. *Briefings in Bioinformatics*, *22*(5), bbaa408.

Powell, W. C., Nahum, M., Pankratz, K., Herlory, M., Greenwood, J., Poliyenko, D., … others (2024). Post-translational modifications control phase transitions of tau. *bioRxiv*.

Shen, B., Chen, Z., Yu, C., Chen, T., Shi, M., & Li, T. (2021). Computational screening of phase-separating proteins. *Genomics, Proteomics and Bioinformatics*, *19*(1), 13–24.

You, K., Huang, Q., Yu, C., Shen, B., Sevilla, C., Shi, M., … Li, T. (2020). Phasepdb: a database of liquid–liquid phase separation related proteins. *Nucleic acids research*, *48*(D1), D354–D359.