# PScalpel: A Machine Learning-based Guider for Protein Phase-Separating Behaviour Alteration

**Jia Wang[1], Liyan Zhu[1], Zhe Wang[1], Chenqiu Zhang[2], Yaoxing Wu[3], Jun Cui[2], Jianqiang Li[4]***

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]MOE Key Laboratory of Gene Function and Regulation, Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
[3]Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, The First Affiliated Hospital of Sun Yat-sen University, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
[4]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
{jia.wang, lijq}@szu.edu.cn, zhuliyan2023@email.szu.edu.cn, zwanghs@foxmail.com, zhangchq26@mail2.sysu.edu.cn, {wuyaox5, cuij5}@mail.sysu.edu.cn

## Abstract

Missense mutations could affect the Liquid-Liquid Phase Separation (LLPS) propensity of proteins and lead to aberrant phase-separating behaviours, which are recently found to be associated with many diseases including Alzheimer's and cancer. However, the regulatory role of mutations in LLPS remains unclear due to challenges in accurately characterizing the LLPS ability of mutants, including the high similarity in features, lack of labeled data, and vast amounts of data involved. To bridge this gap and facilitate the discovery of therapeutic strategies, we propose the first machine learning-based guider for protein phase-separating behaviour alteration, PScalpel. PScalpel leverages both structural information and an auxiliary tasks-based graph contrastive learning framework to distinguish the mutants' LLPS ability, and incorporates a genetic algorithms-based recommendation method to identify mutants with desired LLPS properties. Comprehensive computational and biological experiments validate the effectiveness of PScalpel as a versatile tool for guiding alterations in protein phase separation behavior.

## Introduction

Liquid-Liquid Phase Separation (LLPS) has emerged as a pivotal and intricate mechanism underlying the maintenance of cellular homeostasis(Alberti, Gladfelter, and Mittag 2019). It orchestrates the condensation of biological macromolecules, including proteins and nucleic acids, into membrane-less, liquid-like condensates. Given the intimate connection between LLPS and the proteome, the occurrence of missense mutations can perturb the phase separation and transition behavior of proteins(Hopf et al. 2017; Sundaram et al. 2018; Schuster et al. 2020). Recent investigations have substantiated a causal relationship between aberrant phase-separating behaviors and the manifestation of disease pathologies, encompassing neurodegenerative disorders(Tsang et al. 2020) and cancer(Zou et al. 2022). Consequently, the elucidation of key sequence perturbations that exert a significant influence on LLPS ability assumes paramount importance in comprehending the multifaceted role of phase separation in molecular biology and harnessing this knowledge for applications in disease therapy.

Intuitively, the goal could be achieved in an evaluate-then-recommend fashion. However, the intricate and time-intensive nature of biochemical techniques employed in evaluating the LLPS ability hinders the systematic testing of all possible mutant combinations for a given protein. As the molecular principles underlying phase separation is being continuously uncovered, several protein features, such as disordered region, arginine rich domain and low-complexity aromatic-rich kinked segments, have been identified as the driving force thus the determinants of phase separation, which resulted in the development of the so-called first generation Phase Separation Proteins (PSPs) prediction tools.

More recently, observing that the molecular grammar of phase separation could be learned from its linear amino acid sequence, multiple machine learning based PSPs prediction models have also been proposed (van Mierlo et al. 2021; Sun et al. 2019; Saar et al. 2021).

Although these knowledge-based and sequence-based models achieved great success in LLPS propensity prediction, none of them is suitable to be used to distinguish LLPS ability of mutant proteins since they're all designed for natural proteins which differ significantly in protein sequences.

For mutant proteins, it remains a very challenging problem to precisely discriminate their phase separation likehood and screen out the ones that best meet the actual demand, mainly due to the following facts: **(1) Requirement of high model sensitivity**. As the mutants are extremely similar in sequences, the prediction model should be of high sensitivity to its input. Hence, new features should be explored for differentiating LLPS propensity of mutant proteins. **(2) Lack of labeled data**. Although the number of reported proteins mediating phase separation is growing rapidly, well labled data that can be used for evaluating mutant proteins' LLPS ability is extremely limited, which means the utilization of other related domain data should also be considered in the learning approach for better performance. **(3) Requirement**

*Corresponding author

**of high model efficiency**. Considering the diversities and complexities of variation characteristics, high efficiency is required for both the mutants' evaluation model and recommendation model.

In light of all these design goals, in this paper, we present the first efficient computational tool, **PScalpel**, for biologists to trace the decisive sequence perturbations associated with LLPS and thus alter protein phase-separating behaviors in a predictable way. As illustrated in Figure 1, the PScalpel framework consists three parts, each of which addresses at least one of the above mentioned design challenges:

(1) **Efficient and Accurate Extraction of Protein Structural Features.** Protein function is largely determined by structural characteristics. To enhance LLPS propensity evaluation among mutant proteins, we introduce structural features into the model. However, the vast number of potential mutants makes existing residue contact map prediction methods which lack the capacity to support the design of an efficient and accurate evaluation and recommendation tool. To address this, we develop **BetaFold**, a distilled model based on Alphafold2, designed to provide an efficient and accurate tool for extracting protein structural information.

(2) **LLPS ability distinguishing model for mutants.** Drawing inspiration from the successful application of Contrastive Learning(CL)(You et al. 2020; Oord, Li, and Vinyals 2018; He et al. 2020) in handling large, high-dimensional unlabeled datasets, we explore the feasibility of GCL in this context. Traditional GCL methods often heavily rely on data augmentation to generate positive and negative pairs, which may not suit models requiring high sensitivity. To overcome this, we propose Twin Towers Graph Contrastive Learning ($\mathrm{T}^3\mathrm{GCL}$). $\mathrm{T}^3\mathrm{GCL}$ simultaneously trains two related auxiliary tasks, namely disease-related mutant classification and normal LLPS propensity prediction, using GCL. By leveraging the inherent consistency information of the target task, $\mathrm{T}^3\mathrm{GCL}$ aims to enhance model's performance and sensitivity. Additionally, we introduce a model-based transfer learning approach (TLPSDM) to enhance the Phase Separation Distinguishing Model (PSDM). By freezing all layers except the final one, we exploit the feature extraction capabilities of the pre-trained model and adapt it to our specific single protein mutation sequence dataset.

(3) **GA-based efficient recommendation method.** We construct a Phase Separation ability Distinguishing model for Mutants (PSDM) using the integration of **BetaFold** and $\mathrm{T}^3\mathrm{GCL}$. Then the PSDM model is used to serve as the fitness function within the Genetic algorithm (GA) framework, facilitating an accelerated search for optimal alterations.

## The Proposed Method

In this paper, we present PScalpel aimed at assisting biologists in identifying mutation sites within the vast space of mutation data that could significantly impact the phase-separation ability of proteins. Given the limitations of using knowledge-based or sequence-based features alone to distinguish the phase separation ability of mutant proteins, we propose the integration of tertiary structure information and introduce a novel contrastive learning method called $\mathrm{T}^3\mathrm{GCL}$ to enhance the feature extraction capabilities of our evaluation model, PSDM. Finally, we employ a genetic algorithm to recommend mutants that best align with the desired alteration in protein phase-separating behavior, thereby addressing the practical needs of protein modification.

**Efficient Prediction of Protein Graph Structure** In this section, we propose an efficient protein graph structure prediction model called BetaFold. The framework diagram of BetaFold is depicted in Figure 2 (a). To ensure computational efficiency, Multiple Sequence Alignment (MSA) is not utilized to match similarities among amino acid fragments from known three-dimensional protein sequences, as done in traditional protein structure prediction models. Instead, to achieve satisfactory performance solely through machine learning training, the predicted results of the AlphaFold2 model are approximated to actual three-dimensional protein structure data and used as training data for our model.

Furthermore, the model is enhanced by simplifying the complex task of predicting the absolute position of all atoms within a protein. In contrast to AlphaFold2, which aims to determine absolute position information for all atoms, BetaFold focuses solely on relative position information between the $C_\alpha$ atoms of individual amino acids and those of neighboring amino acids. This simplification transforms the task into a binary classification problem: predicting whether a pair of amino acids are adjacent nodes. Specifically, in this work, the distance between $C_\alpha$ atoms of neighboring amino acids is defined as less than $10\mathring{A}$.

**The Twin Towers Graph Contrastive Learning** Due to the limited availability of well-labeled data and recognizing the effectiveness of contrastive learning in unsupervised learning domains, we propose the utilization of a contrastive learning (CL)-based model for predicting phase separation in mutants (Xu et al. 2021; You et al. 2020). However, the conventional approach of employing random data augmentation to generate positive pairs in CL deviates from the model's primary objective of achieving high sensitivity, rendering these techniques unsuitable for direct application. To address this aforementioned challenge, we introduce the Twin Towers Graph Contrastive Learning ($\mathrm{T}^3\mathrm{GCL}$) framework. As depicted in Figure 2 (b), the dataset pertaining to the target task is denoted as A, while the two auxiliary task datasets are represented as B and C, respectively. It is assumed that the auxiliary datasets B and C contain ample well-labeled data, with dataset A being closely associated with the intersection of the two auxiliary datasets. Consequently, despite the absence of labels specifically related to the target task in dataset A, its intrinsically consistent information can still be jointly learned by concurrently training two auxiliary tasks on datasets B and C. To achieve the aforementioned objective, it is imperative to appropriately balance the gradients of the two auxiliary tasks. This is because the learning complexities of the two tasks may signif-

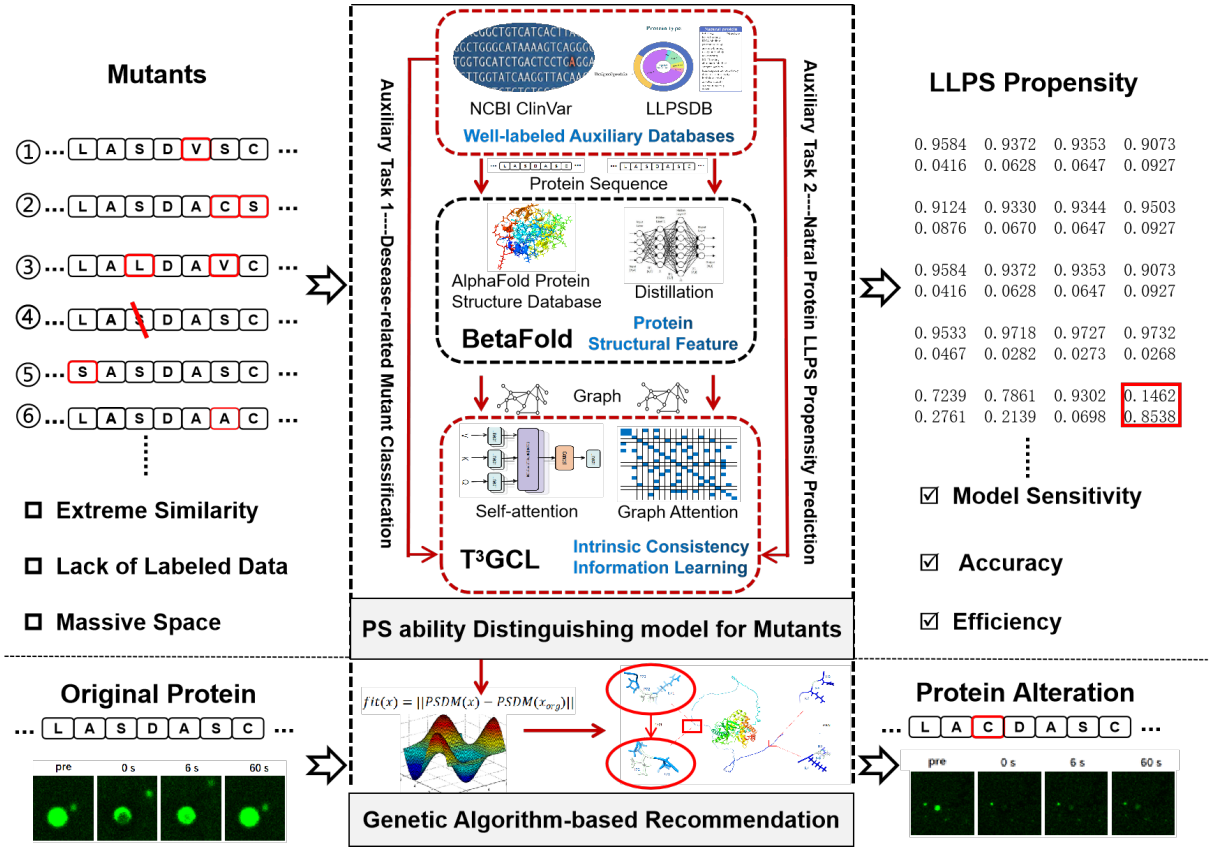


Figure 1: **Architecture of the proposed PScalpel scheme.** In PScalpel, a protein graph structure prediction model, BetaFold, and the Twin Towers Graph Comparative Learning ($T^3$GCL) framework are proposed and combined to construct a Phase Separation ability Distinguishing Model (PSDM) for mutants. Genetic algorithm (GA) was then adopted to complete the sequence perturbation recommendation task.

icantly differ, resulting in the model primarily capturing the internal consistency features of one task while disregarding the other.

**PS Ability Distinguishing Model for Mutants** In this section, the proposed $T^3$GCL framework is incorporated with the BetaFold method to construct an efficient and effective PS ability distinguishing model for mutants (PSDM). Specifically, the target dataset of phase-separated mutant protein is denoted as $A$. We use the the clinical mutation dataset NCBI ClinVar as auxiliary dataset $B$ and the phase-separating proteins dataset LLPSDB for $C$. The appropriateness of this selection hinges upon the existing understanding that anomalous phase separations and transitions bear a causal relationship with various human diseases, including neurodegenerative disorders and cancer (Schuster et al. 2020).

It is noteworthy that in the case of the clinical mutation dataset $B$, the original contrastive learning loss function exhibits limited discriminatory ability in distinguishing positive and negative examples. Therefore, we modified the loss function as follows. Denoting the similarity function as $sim(x, y) = 1 - arccos(x^T y/||x||||y||)/\pi$, the loss function is defined as:

$$\ell = -\log \frac{exp(sim(x, x^+))}{exp(sim(x, x^+)) + exp(sim(x, x^-))}, \quad (3)$$

where $x$ is the original protein sequence, $x^+$ is the protein sequence with benign mutation, and $x^-$ is the protein sequence with pathogenic mutation. Regarding the auxiliary task conducted on dataset $C$, it is anticipated that the model would acquire the capability to learn the invariance associated with protein phase separation ability.Hence in Formula (3), $x$ and $x^+$ represent the sequences of proteins mediating phase separation, and $x^-$ denotes the protein sequence without phase separation ability.

To enhance learning performance, adjustments are made to the learning rates of the two auxiliary tasks. Specifically, for the gradients of auxiliary task $C$, denoted as $grad_C$, a modification is applied using the equation:

$$grad_C = grad_C \frac{||grad_B||}{||grad_C||}\tau, \quad (4)$$

where $\tau$ is a proportional constant and is set as $0.05$ in this work. Furthermore, various feature extraction techniques are explored to achieve optimal performance, including self-attention (Vaswani et al. 2017), Graph Isomorphism (GIN)

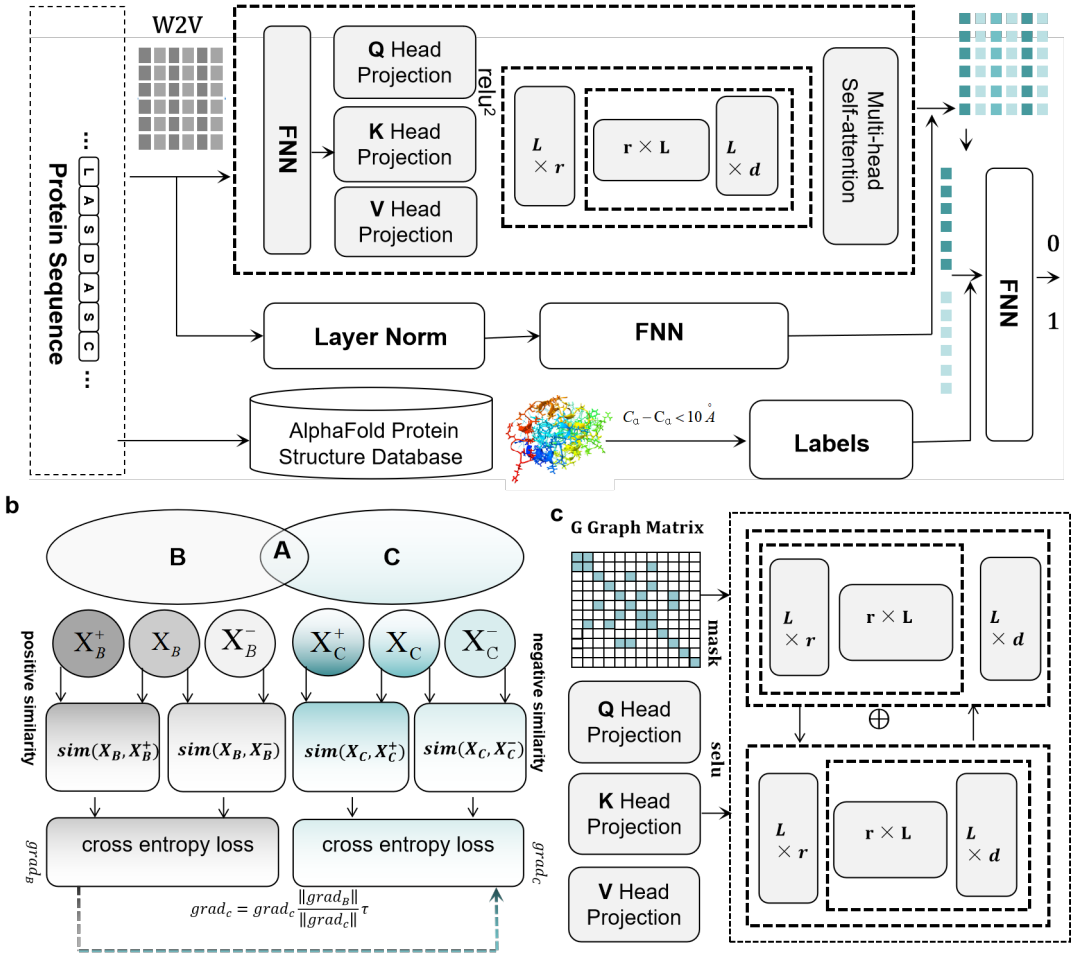


Figure 2: **A diagram illustrating the PScalpel framework. a**, Network architecture of BetaFold. It shows the network architecture of the proposed BetaFold model. **b**, The workflow of $T^3GCL$. **c**, The workflow of self-attention+GAT. It illustrates the workflow of self-attention+GAT (Graph Attention Network) in the PScalpel framework.

(Xu et al. 2018), and Graph Attention (GAT) (Veličković et al. 2017). Experimental results demonstrate that the model employing self-attention and GAT yields the most favorable outcomes. Implementation details are shown in Figure 2 (c). The feature vectors obtained by $T^3GCL$ are finally mapped to the corresponding categories to complete the PS ability evaluation task of the mutant proteins.

**Transfer Learning with PSDM** To enhance LLPS prediction ability of PSDM in single protein mutation sequence, we developed TLPSDM by integrating transfer learning with PSDM. Transfer learning(Iman, Arabnia, and Rasheed 2023), which leverages knowledge from other domains to enhance target domain performance. Specifically, we employ a model-based transfer learning method. By freezing the initial layers of a pre-trained model, we retain its ability to capture broad protein features. By training only the final layer, we fine-tune the model for our specific prediction task, leading to improved predictive accuracy.The structure of the TLPSDM model is shown in Figure 3.

**Guidance for Protein Modification** To expedite the search for the optimal solution within the vast mutation space, a genetic algorithm is employed. The fitness function for this algorithm is defined as follows:

$$fit(x) = ||PSDM(x) - PSDM(x_{org})||, \qquad (5)$$

where $x$ represents the mutated protein sequence and $x_{org}$ denotes the original protein sequence. The fitness function quantifies the difference between the phase separation distance matrices (PSDMs) of the mutated sequence and the original sequence. The recommendation algorithm follows the subsequent steps. Firstly, a batch of random mutants for the target protein is generated and utilized as the initial population. Subsequently, a selection process is employed to identify and retain a certain number of superior specimens based on the fitness function $fit(x)$. The retained specimens then undergo crossbreeding and mutation to generate a new population. This process is iteratively repeated for a specified number of iterations. Finally, the optimal specimens within the population are identified as the optimal solution to the problem. In order to align with the scenario of natu-

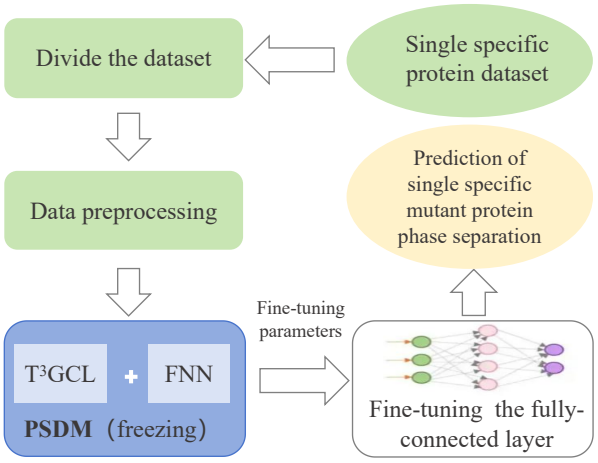


Figure 3: **Diagram of the TLPSDM model structure.** The figure illustrates the integration of transfer learning with PSDM to create the TLPSDM model, showing steps like loading pre-trained models, freezing layers, and fine-tuning the final layer with single protein mutation sequence data.

ral mutations, where modifications typically occur at a few specific sites, it is crucial to limit the simultaneous modification of numerous sites. Consequently, this work restricts the modification of only one or a few sites. In terms of the experimental setup, it is ensured that the number of mutation sites in each sample does not exceed a small positive integer denoted as $k$. For this particular study, $k$ is set to 2.

## Experimental Results

### Datasets

To train BetaFold, the prediction results of human protein structures of Alphafold2 on the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) are utilized. A 0-1 matrix is then generated by calculating and determining whether the Euclidean distance between the $C_\alpha$ of any two amino acids on the protein is less than $10\mathring{A}$. Totally 23,391 human protein sequences and their corresponding 0-1 matrices are obtained, among which 21,051 are used as the training set and 2,340 as the validation set. The test dataset is composed of native protein structures from the RCSB-PDB database (released on 07/20/2022) (Berman et al. 2000). And to ensure that there's no data overlap between the test dataset and the validation dataset, only those native protein structures updated after February 2, 2022 in RCSB-PDB are used. Sequences whose lengths are less than 100 are also deleted from the test dataset. Hence, finally, we obtain a test dataset containing 731 proteins. In addition, in order to compare Betafold with other residue contact prediction models, we obtained a test data set containing 34 proteins from the 14th Community Wide Experience on the Critical Assessment of Techniques for Protein Structure Prediction about the CASP14.

As to $\mathrm{T^3GCL}$, two auxiliary datasets, the clinical mutation data and the phase separation data, are downloaded from the ClinVar website (https://www.ncbi.nlm.nih.gov/clinvar/) and LLPSDB (Li et al. 2020; Ning et al. 2020) and PDB (Berman et al. 2000) respectively. For ClinVar, we screened data that related to cancer and neurodegenerative diseases and removed early termination and missense parts of the data. And only gene data with both benign and pathogenic mutations are remained to construct the auxiliary dataset. Thus, we obtained a dataset of 1,590 benign mutations and 1,371 pathogenic mutations. In order to expand the dataset, after translating the gene into protein sequence, a Cartesian product of the benign mutation dataset and the pathogenic mutation dataset of the same original protein sequence is applied to get a triple array (benign mutant protein, original protein, pathogenic mutant protein), which corresponds to ($x^+$, $x$, $x^-$) in Formula (3). For LLPSDB, we screened protein sequences $PS^+$ (493) with naturally occurring sequence constructs and phase separation capability. And for PDB, we screened protein sequences $PS^-$ (1556) that did not contain any disordered residues which means they're highly unlikely to phase separate (Saar et al. 2021). We then divided the phase separation data into two independent parts, $70\%$ of which is used as the training set for $\mathrm{T^3GCL}$ and $30\%$ as the validation set. Similarly, the training dataset is constructed as ($PS^+$, $PS^+$, $PS^-$) corresponding to ($x^+$, $x$, $x^-$) in Formula (3).

To evaluate the performance of PSDM, we screened the mutated single proteins from LLPSDB. And after deletion of those repetitive sequences and empty sequences, 283 phase-separated mutant proteins are left. It's worth to mention that due to the small amount of data, these 283 sequences are used only for testing rather than training nor validation.

To train TLPSDM and evaluate its performance in predicting the phase separation capabilities of single proteins, we utilized mutation amino acid sequences from cGAS and TDP43 proteins as distinct single protein datasets, performing five-fold cross-validation on each. The cGAS dataset comprises 85 amino acid mutation sequences with LLPS capabilities and 45 sequences without LLPS capabilities. The TDP43 dataset includes 69 sequences with LLPS capabilities and 47 sequences without LLPS capabilities. For the five-fold cross-validation, each dataset was partitioned into five subsets, with four subsets used for training and one subset used for testing in each iteration.

### Input Features

In the proposed PSDM model, ProtVec is used to generate the word vectors for amino acids (Asgari and Mofrad 2015) processing. The amino acid sequence in character type is then encoded and converted into a numerical matrix of dimension $100 \times L$. To further improve the model performance, low-complexity regions scores (van Mierlo et al. 2021) are also introduced as an additional feature in PSDM.

### Network Architecture and Hyper-parameters

The proposed BetaFold architecture consists of two parts, the feature extraction layer and the application layer. The feature extraction layer uses the multi-head self-attention mechanism to extract the relationship features between amino acids, while the feature extraction layer contains $N = 6$ identical blocks. Each block is a residual structural unit (He et al. 2016) whose dimension $d_{model} = 128$. It is

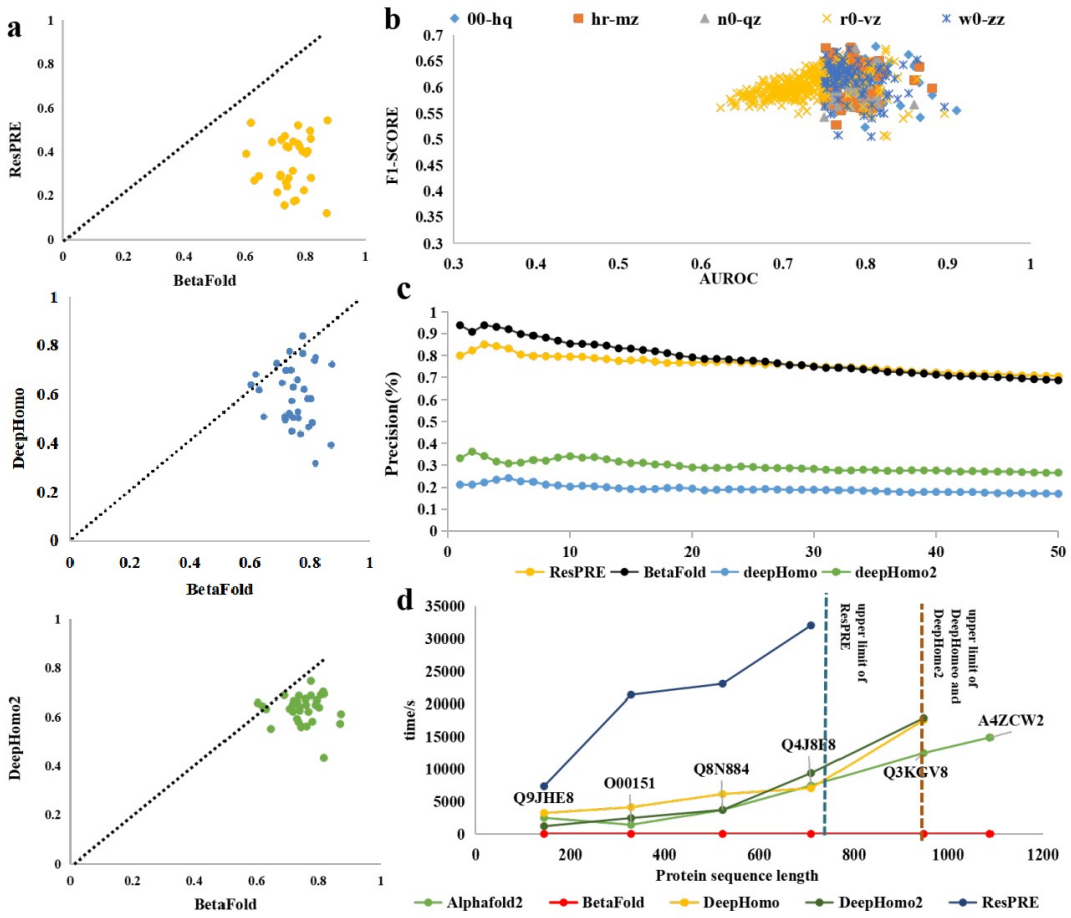


Figure 4: **Results of BetaFold. a**, Comparison with RRCP. Each point on the coordinate system represents the AUCROC value of a predicted protein on the BetaFold model and the corresponding baseline model. **b**, BetaFold on PDB subsets. Considering the potential errors inherent in AlphFold2, the three-dimensional structural data of natural proteins obtained from experimental calculations were downloaded from the RCSB-PDB database to construct the test dataset. **c**, Performance on CASP14. This section presents the performance of BetaFold, ResPRE, DeepHomo, and DeepHomo2 on the CASP14 dataset, which consists of 34 realistic targets with experimental structures. The precision (%) is depicted as a function of the number of predicted contacts. **d**, Efficiency Comparison.

formed by the parallel connection of a self-attention layer and a fully connected feed-forward network. For the self-attention layer, we set $h = 8$. And for each head, the dimension is set as $d_k = d_v = d_{model}/h = 16$. For the seek of high efficiency, the kernel method (Choromanski et al. 2020) as well as the Squared ReLU activation (So et al. 2021) are also adopted in the self-attention layer. As to the application layer, the fully connected feed-forward network is used to linearly transform the feature matrix of the amino acid sequence output from the feature extraction layer, where each column corresponds to the feature vector of an amino acid in the protein sequence. Then the feature vectors are concatenated in pairwise to obtain new feature vectors that representing the relationship between amino acids. Finally, a fully-connected feed-forward network is used to classify the relational feature vectors, predicting whether the Euclidean distance between any specific pair of amino acids is less than $10\mathring{A}$. Since the complexity of predicting the relational feature vector is $O(L^2)$, only a hidden layer of dimension 64 is added to the feed-forward neural network, and Squared ReLU activation (So et al. 2021) is used to reduce the load of calculation. The learning rate and dropout rate are set to 0.001 and 0.5 respectively, and we use Adam (Kingma and Ba 2014) as the gradient descent algorithm.

The training of the PSDM model is performed in two steps. Firstly, we use the proposed $\mathrm{T}^3\mathrm{GCL}$ method to extract the approximated features from the phase-separated data of the mutant proteins. This part is composed of a $100 \times 128$ linear transformation matrix and a self-attention+GAT layer of dimension 128. We propose to represent the features of protein by summing all node features in one iteration. The dropout rate is set to 0 and the SeLU activation (Klambauer et al. 2017) is used to guarantee the high sensitivity of the model to site changes in protein sequences. Then a fully

connected feed-forward network is utilized to functionally map the generated feature vectors to the corresponding categories. Two hidden layers, with dimensions 64 and 16 respectively, are added to the feed-forward neural network.

## Comparison Experiment

**Effectiveness of Betafold** A performance comparison was conducted between Betafold and the most advanced Residue-Residue Contact Prediction (RRCP) model using the CASP14 dataset. Figure 4(a) shows the AUROC predictions by BetaFold, ResPre(Li et al. 2019), DeepHome(Yan and Huang 2021), and DeepHome2(Lin, Yan, and Huang 2022) for each protein. To evaluate the performance of the models, we utilized the trained Betafold model and the residue-residue contact prediction model as baseline methods to predict 34 protein sequences from the CASP14 dataset. The predicted results were compared with the actual values, and the AUROC indicators were calculated for each model on the test dataset. Subsequently, the performance differences between the Betafold model and the other baseline models in predicting different proteins were analyzed. The findings demonstrate that the Betafold method outperforms the baseline methods.

To provide a more detailed performance comparison between the BetaFold model and other RRCP models, we contructed a precision (%) graph on the CASP14 dataset, as shown in Figure 4(c). This graph illustrates the percentage of correctly predicted edges among the top n predicted edges with confidence in the model's prediction results. BetaFold maintains higher accuracy compared to other RRCP models, even as the number of predicted edges increases.

The performance evaluation of the BetaFold model encompasses multiple subsets of the PDB dataset is illustrated in Figure 4(b), with an AUROC value of approximately 0.87, indicating minimal bias and reliable predictions close to the actual 3D protein structures. Additionally, the model's performance was assessed across various subdatasets, revealing a consistent AUROC range interval.

Regarding efficiency, we conducted a comparative analysis of the processing time between BetaFold, AlphaFold2, and other residue-residue contact prediction models for protein sequences of varying lengths. As depicted in Figure 4(d)), the results clearly demonstrate the superior time performance exhibited by the BetaFold model in comparison to other models.

**Effectiveness of PSDM** The experimental results of PSDM are demonstrated in Table1 that existing PS protein prediction tools perform remarkably well on natural proteins; however, none of them exhibit the ability to differentiate the PS propensity of mutant proteins. With AUROC values ranging from 0.54 to 0.57 for methods like PSPredictor, DeePhase, and FNN. These results suggest random predictions when categorizing mutant protein phase-separated samples. Although PSAP achieves a higher AUROC of 0.6445, its F1-score(0.5080) indicates that it is still unsuitable for integration into the protein alteration recommendation system. As the first LLPS propensity evaluation model specifically designed for mutant proteins, PSDM demonstrates the best performance in this particular task.

| Method | Natural Proteins | | Mutant Proteins | |
|---|---|---|---|---|
| | AUROC | F1-score | AUROC | F1-score |
| PSPredictor | 0.9136 | **0.9324** | 0.5403 | 0.5288 |
| PSAP | 0.9433 | 0.8915 | 0.6445 | 0.5080 |
| DeePhase | 0.9482 | 0.8854 | 0.5707 | 0.5310 |
| FNN | **0.9675** | 0.8587 | 0.5752 | 0.5227 |
| PSDM | 0.8933 | 0.7746 | **0.7226** | **0.6474** |

Table 1: The results of PS prediction using PSDM and other methods on both natural and mutant proteins are presented.

**Effectiveness of TLPSDM** The experimental results of TLPSDM are demonstrated in Table 2 that presents the AUROC results of various predictive methods on the cGAS and TDP-43 datasets. The results indicate that existing methods such as PSPPI, PSAP, FNN, and PSDM perform reasonably well. However, the proposed TLPSDM model achieves the highest AUROC values on both datasets, with 0.776 for cGAS and 0.667 for TDP-43. This significant improvement demonstrates that TLPSDM is highly effective in leveraging transfer learning to enhance the prediction of LLPS capabilities in single protein mutation sequences.

| DataSet | PSPPI | PSAP | FNN | PSDM | TLPSDM |
|---|---|---|---|---|---|
| cGAS | 0.489 | 0.743 | 0.757 | 0.756 | **0.776** |
| TDP-43 | 0.500 | 0.516 | 0.551 | 0.612 | **0.667** |

Table 2: Comparison of Predictive Methods on cGAS and TDP-43 (AUROC Evaluation).

**Ablation Study** Ablation studies have been conducted to evaluate the efficacy of each component within the proposed PSDM framework. Figure 5 illustrates the performance comparison of PSDM implementations with GAT and self-attention mechanisms. The results demonstrate that the inclusion of the protein graph structure predicted by the BetaFold model enhances the performance of PSDM, as evidenced by the superior performance achieved with GAT. Furthermore, the incorporation of $\mathrm{T^3GCL}$ notably enhances the feature extraction capabilities of the model, resulting in significant performance improvements.

## Discussion and Conclusion

In this study, we introduce PScalpel which is the first machine learning tool tailored for biologists to track and analyze sequence perturbations in liquid-liquid phase separation (LLPS) phenomena. PScalpel includes the Protein Phase-Separating Mutants Distinguishing Model (PSDM) and a Genetic Algorithm (GA)-based method for suggesting alterations. This tool enables predicting and manipulating protein phase-separating behaviors reliably. We also make two noteworthy contributions as interim achievements. Firstly,we present an efficient and accurate residue-residue contact prediction model called BetaFold, which extracts vital graph structure features from proteins and has the potential to enhance performance in other protein structure-based tasks. Secondly, we introduce a novel Graph Convolutional

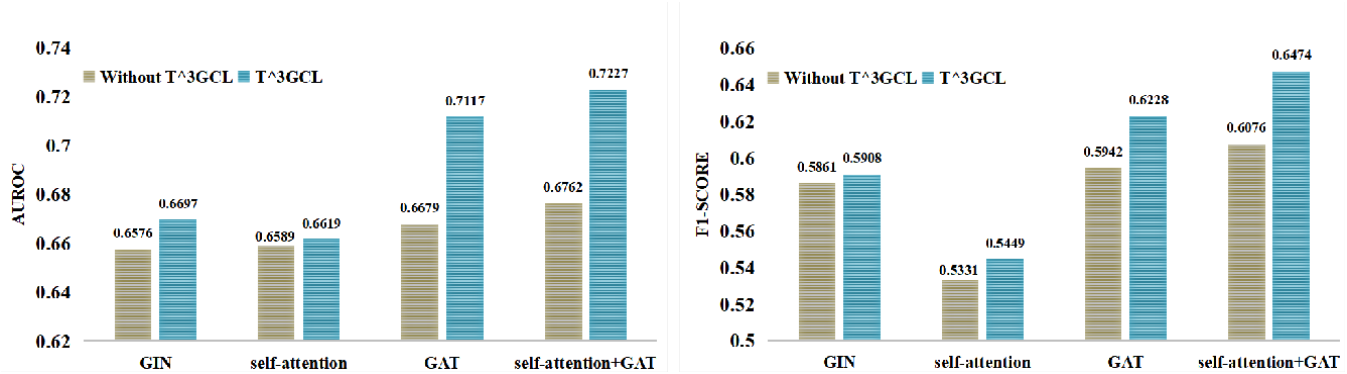


Figure 5: **ablation study of PSDM**. ablation study of PSDM. The performance of PSDM was evaluated on the phase separation dataset of mutant proteins using F1 Score and AUROC.

Network (GCN) Learning method, termed $T^3GCL$, for scenarios where data augmentation is not feasible.Additionally, we enhance PSDM by integrating transfer learning, resulting in TLPSDM, which leverages pre-trained knowledge to improve predictive accuracy and classification performance for LLPS capabilities in single protein mutation sequences.

Through extensive experimentation, we validate the efficacy of our approach in precisely identifying key perturbations related to protein phase separation. Additionally, our tool efficiently recommends optimal protein alterations tailored to specific application requirements.

**Code** — https://github.com/zly20020208/PScalpel

## Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2020YFA0908700, in part by the Natural Science Foundation of China under Grant 62476170, 62073225, 62203134, in part by the National Natural Science Funds for Distinguished Young Scholar under Grant 62325307, in part by the Natural Science Foundation of Guangdong Province under Grants 2023B1515120038, in part by Shenzhen Science and Technology Innovation Commission under Grants JCYJ20220531103401003, 20220809141216003, in part by the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, the Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235, in part by the Scientific Instrument Developing Project of Shenzhen University under Grant 2023YQ019.

## References

Alberti, S.; Gladfelter, A.; and Mittag, T. 2019. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*, 176(3): 419–434.

Asgari, E.; and Mofrad, M. R. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11): e0141287.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research*, 28(1): 235–242.

Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Schärfe, C. P.; Springer, M.; Sander, C.; and Marks, D. S. 2017. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2): 128–135.

Iman, M.; Arabnia, H. R.; and Rasheed, K. 2023. A review of deep transfer learning and recent advancements. *Technologies*, 11(2): 40.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-normalizing neural networks. *Advances in neural information processing systems*, 30.

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

Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; and Zhang, Y. 2019. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, 35(22): 4647–4655.

Lin, P.; Yan, Y.; and Huang, S.-Y. 2022. DeepHomo2. 0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning. *Briefings in Bioinformatics*.

Ning, W.; Guo, Y.; Lin, S.; Mei, B.; Wu, Y.; Jiang, P.; Tan, X.; Zhang, W.; Chen, G.; Peng, D.; et al. 2020. DrLLPS:

a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic acids research*, 48(D1): D288–D295.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Saar, K. L.; Morgunov, A. S.; Qi, R.; Arter, W. E.; Krainer, G.; Lee, A. A.; and Knowles, T. P. 2021. Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proceedings of the National Academy of Sciences*, 118(15): e2019053118.

Schuster, B. S.; Dignon, G. L.; Tang, W. S.; Kelley, F. M.; Ranganath, A. K.; Jahnke, C. N.; Simpkins, A. G.; Regy, R. M.; Hammer, D. A.; Good, M. C.; et al. 2020. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences*, 117(21): 11421–11431.

So, D.; Mańke, W.; Liu, H.; Dai, Z.; Shazeer, N.; and Le, Q. V. 2021. Searching for Efficient Transformers for Language Modeling. *Advances in Neural Information Processing Systems*, 34: 6010–6022.

Sun, T.; Li, Q.; Xu, Y.; Zhang, Z.; Lai, L.; and Pei, J. 2019. Prediction of liquid-liquid phase separation proteins using machine learning. *BioRxiv*, 842336.

Sundaram, L.; Gao, H.; Padigepati, S. R.; McRae, J. F.; Li, Y.; Kosmicki, J. A.; Fritzilas, N.; Hakenberg, J.; Dutta, A.; Shon, J.; et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8): 1161–1170.

Tsang, B.; Pritišanac, I.; Scherer, S. W.; Moses, A. M.; and Forman-Kay, J. D. 2020. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell*, 183(7): 1742–1756.

van Mierlo, G.; Jansen, J. R.; Wang, J.; Poser, I.; van Heeringen, S. J.; and Vermeulen, M. 2021. Predicting protein condensate formation using machine learning. *Cell reports*, 34(5): 108705.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I.; et al. 2017. Advances in neural information processing systems. *Attention is all you need*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, M.; Wang, H.; Ni, B.; Guo, H.; and Tang, J. 2021. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, 11548–11558. PMLR.

Yan, Y.; and Huang, S.-Y. 2021. Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes. *Briefings in bioinformatics*, 22(5): bbab038.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823.

Zou, H.; Pan, T.; Gao, Y.; Chen, R.; Li, S.; Guo, J.; Tian, Z.; Xu, G.; Xu, J.; Ma, Y.; et al. 2022. Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes. *Nucleic acids research*, 50(9): e49–e49.