AlphaSolu: AlphaFold-aware Protein Solubility Prediction Using Spatial Structures

Jinzhi Ouyang 23020231154218*,artificial intelligence class
Ze Lin 23020231154207*,information school class
Zhekai Lin 31520231154301*,information school class
Yihang Liu 31520231145280*,information school class
Oucheng Huang 31520231154291*,information school class

*School of Informatics, Xiamen University
{ouyangjinzhi, linze, linzhekai, liuyihang}@stu.xmu.edu.cn, huangoucheng@gmail.com

Abstract

With the development of heterologous expression of recombinant protein applied in many fields, the solubility of protein attracts people's attention, as poorly soluble proteins may form insoluble aggregates which can reduce the yield of expressed target proteins. Therefore, we propose AlphaSolu, a AlphaFold-aware framework to predict protein solubility with high performance. Unlike most approaches in the past, AlphaSolu explores the information in 3D structure instead of one-dimensional sequence based on graph neural network, where amino acid sequence representation and contact map reconstruct a graph as the input. It catches spatial information which may contain abundant solubility information about protein. Experiments indicates the framework we proposed achieves higher accuracy and Matthew's correlation coefficient outperforming the state-of-the-art approaches, that enables mass production of heterologous expressions on an industrial scale. It is worth mentioning that we wrap it into an end-to-end protein solubility prediction framework and users only need raw sequences to gain the results.

Introduction

It is widely known that genetic engineering has an enormous impact on all aspects of human life, not only biomedicine but also agriculture, environmental conservation. A vital but challenging step in the procedure of genetic engineering is heterologous protein expression. Heterologous protein expression indicates transferring a gene from one organism to another that we call host organism, typically a microorganism like bacteria, yeast, or mammalian cells. The foreign gene is inserted into the host organism's DNA, enabling it to produce the desired protein. This process is crucial in biotechnology and research for producing specific proteins.

However, solubility is a crucial property in heterologous protein expression, significantly impacting the success of protein expression, purification, and downstream applications. Poorly soluble proteins often form insoluble aggregates or inclusion bodies which reduces target protein yield. Therefore, they requires more purification steps to achieve the desired purity and yield. In contrast, soluble proteins can be more efficiently purified using simpler approaches like affinity chromatography, thus resulting in higher protein purity. If an approach predicting protein solubility based on amino acid sequences is available, we can directly select the soluble proteins for research. It leads to higher yields and purity in heterologous protein expression, facilitating large-scale industrial screening and production.

Previous work for predicting protein solubility were based on statistical properties of amino acid sequences. Some work(Wilkinson and Harrison 1991; Davis et al. 1999) relied on the average charge, the relative quantities of four types of residues (Asp, Glu, Lys, and Arg), and their corresponding turn-forming content. Besides, experimental observations revealed significant differences between soluble and insoluble proteins in terms of hydrophobic segments, glutamine content, negatively charged residues, and the percentage of aromatic amino acids(Christendat et al. 2000). And the accuracy of predicting solubility using these statistical features was approximately 65%. However, such approaches have significant limitations. They rely on a deep understanding of protein biology and observations of soluble and insoluble proteins, and need to mannally extract discriminative features, which is relatively inefficient.

How to explore the solubility information in proteins and efficiently extract the corresponding features to achieve high accuracy is still a challenging question. On one hand, past researchers(Smialowski et al. 2007; Rawi et al. 2018; Khurana et al. 2018; Wu and Yu 2021) preferred to use sequence-based features to predict solubility which is proved limited. More and more recent work(Chen et al. 2021; Yuan et al. 2022; Bryant, Pozzati, and Elofsson 2022; Song et al. 2022) indicate proteins of high-dimensional structure may contain a wealth of additional solubility information. For example, 3D structure includes the positions of individual atoms and the intricate folding patterns. On the other hand, with the rise of deep learning techniques, automated feature extraction becomes possible.

As mentioned above, our goal is to design a protein solubility prediction with high accuracy using high-dimensional features of proteins, especially three-dimensional ones. The introduction of AlphaFold(Jumper et al. 2021), which is a model developed by DeepMind for predicting the three-dimensional structure of proteins, has injected new vitality into research related to structural proteins. Therefore, we propose a deep framework AlphaSolu, using 3D structure

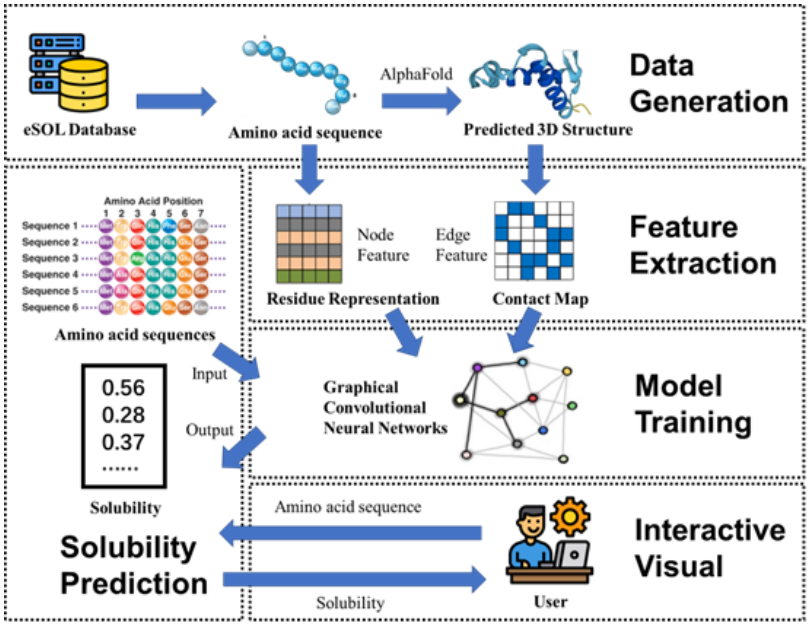

Figure 1: The Framework of AlphaSolu.

predicted by AlphaFold based on graph neural network, as shown in figure 1. In a nutshell, our contributions are threefold.

- We propose AlphaSolu, a AlphaFold-aware framework to predict protein solubility catching spatial information about protein, different from many previous sequenced-based approaches.
- We use graph neural network to automatically extract spartial features, where amino acid sequence representation and contact map reconstruct a graph as the input.
- By implementing AlphaSolu and other benchmarks, results demonstrate ours outperforms others and achieves higher accuracy under robustness and reliability.

Related Work

A number of work use machine learning techniques to predict protein solubility. Smialowski et al. (2007) introduced a machine learning-based solubility prediction system, PROSO, consisting of a two-level layers with a Support Vector Machine (SVM) and a Naive Bayesian classifier, and it could achieve 72% accuracy with the data in TargetDB. Subsequently, Smialowski et al. (2012) improved their previous work and proposed PROSO II achieving accuracy up to 75.4%. Han, Wang, and Zhou (2019) applied support vector machine techniques to solubility prediction, classifying solubility prediction as a classification problem instead of a regression one. PaRSnIP(Rawi et al. 2018) and SoluProt(Hon et al. 2021) employed gradient boosting machine to predict protein solubility. The above work based on machine learning techniques need manually extract features of amino acid sequences.

Due to the ability of neural networks to automatically learn information embedded in amino acid sequences, many subsequent works have employed neural networks to predict protein solubility. DeepSOL(Khurana et al. 2018) used two-dimensional information as inputs to a convolutional neural network, with one dimension representing sequences and the other representing structures. Sequence information included one-dimensional amino acid characteristics, such as sequence length, molecular weight, net charge, aliphatic index (AIs), grand average of hydropathy (GRAVY), and others. Structural one was derived from high-dimensional protein structure features like SS3 and SS8 predicted by SCRATCH(Cheng et al. 2005) using amino acid sequences. This approach achieved accuracy up to 77%, marking a significant improvement over previous approaches. EPSOL(Wu and Yu 2021) also utilized a convolutional neural network for feature extraction, but compared to DeepSOL, it included a broader range of feature dimensions. Additionally, it used a sliding window model for extracting raw amino acid sequences, resulting an accuracy value of 79% and a Matthews correlation coefficient (MCC) value of 0.58.

GraphSOL(Chen et al. 2021) stood as the first protein solubility prediction using graph convolutional neural network(GCN), the node features of which included amino acid encodings from Blosum62(Mount 2008), physical-chemical properties, evolutionary information, and predicted structural properties. Edge features were derived from contact maps predicted by SPOT-Contact(Hanson et al. 2018). Although graph convolutional neural networks excel in utilizing spatial information, GraphSOL has two potential negative effects. As it relies on high-dimensional structural features predicted by SPIDER3 and SPOT-Contact, they may not be entirely accurate. Additionally the extracted high-dimensional structural features are limited to two-dimensional structures, while higher-dimensional structures such as three-dimensional structures hold untapped solubil-

Name	Amino acid sequence	hydrolysis degree
aaeX	MSLFPVIVVFGLSFPPIFFELLLSLAIF	0.34
aas	MLFSFFRNLCRVLYRVRVTGDTQALKGERVLIT	0.07
aat	MRLVQLSRHSIAFPSPEGALREPNGLLALGGDLSP	0.08

Table 1: Original amino acid sequence data format.

ity information.

Proposed Solution

Data Preparation

Amino Acid Sequence The original training set containing amino acid sequences is derived from GraphSol(Chen et al. 2021), which is collected by the eSOL database(Niwa et al. 2009). This database contains a comprehensive set of Escherichia coli protein solubility data, corresponding to gene names in the NCBI database. A total of 3144 data were initially obtained, with a balanced distribution of positive and negative samples (1:1), ensuring there were no issues with class imbalance. The dataset were performed to preprocess to avoid the repeating sequences and finally we gained 3140 sequences with corresponding solubility. The format of amino acid sequence data is illustrated in Table 1.

Three-dimensional Structure AlphaFold(Jumper et al. 2021) developed by the DeepMind to predict the three-dimensional structure of proteins is currently open-source. Taking the 3140 amino acid sequences obtained above as input, it outputs the corresponding PDB files for each amino acid sequences. The PDB files contain comprehensive and rich three-dimensional structural information, which can help us predict the protein's solubility in subsequent steps.

Feature Extraction

Node Feature SeqVec(Heinzinger et al. 2019) is a pre-trained model used to generate protein/residue representations. It is trained on the UniRef50 dataset using the language model ELMo(Peters et al. 2018). Traditional natural language processing models do not focus on biological context information, while SeqVec utilizes a large amount of unlabeled data from the UniRef50 database to capture biophysical properties. The trained model can directly obtain representations with biophysical features based on amino acid sequences, which significantly enhances the performance of downstream tasks. Given the protein's raw sequences, we utilizes the SeqVec(Heinzinger et al. 2019) to output residue-level sequence representations. And each protein obtains an $m \times 1024$ dimensional representation, where m represents the number of residues in the protein. To ensure the consistency of training, we perform padding operations on the obtained representations, resulting in 3140 $L \times 1024$ dimensional features, where L is a hyperparameter.

Edge Feature Edge feature are represented as a contact map, which is also a matrix consisting of 0s and 1s. This definition is essentially similar to an adjacency matrix and reflects the connectivity between protein residues. If the distance between two residues is less than this distance, it indicates a connected relationship between them. Otherwise, if the distance is too large, they are considered unconnected. The contact map is defined as follows:

$$C_{p,q} = \begin{cases} 1, & \text{if } \delta_{p,q} > 8A \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Note that p and q represent two different residues and 8A is a commonly used distance in the contact map(Song et al. 2022). The coordinate data in the protein's three-dimensional structure PDB file precisely reflects spatial information. We calculate the Euclidean distance between residues to obtain a distance matrix namely as contact map. The contact map has dimensions of $m \times m$, where m represents the number of residues in the protein. After performing padding operations on the original contact maps, we get 3140 $L \times L$ dimensional features.

Word Frequency Feature It is obvious that Word frequency information is one of the crucial features. An amino acid sequence consists of 20 letters representing 20 kinds of amino acids separately and the protein solubility is closely related to the kind of amino acids. Therefore, we can extract statistical word frequency features of the 20 letters in each amino acid sequence. To prevent numerical instability when inputting the features into a neural network, the word frequency feature is computed by dividing the count by the total length of the sequence, rather than just considering the raw count. The formula for calculating the frequency feature f_i of letter i is as follows:

$$f_i = \frac{O_i}{L} \tag{2}$$

where O_i is the number of occurrences of the letter and L is the total length of the sequence.

As a result, each amino acid sequence can obtain a 20-dimensional word frequency feature.

Deep Learning Framework

Graph Convolutional Network The node features and edge features extracted can be constructed into a graph convolutional neural network (GCN). The input data of the graph convolutional neural network is a graph structure, in which each node represents a residue in the protein structure using an implicit vector, and the edge represents the relationship or connection relationship between nodes, which is usually represented by the adjacency matrix as contact map.

Through the convolution operations of the contact graph, the residue representation of each node is convolved with the contact graph of its neighbor nodes, so as to update the node representation. A graph convolutional neural network is represented as follows:

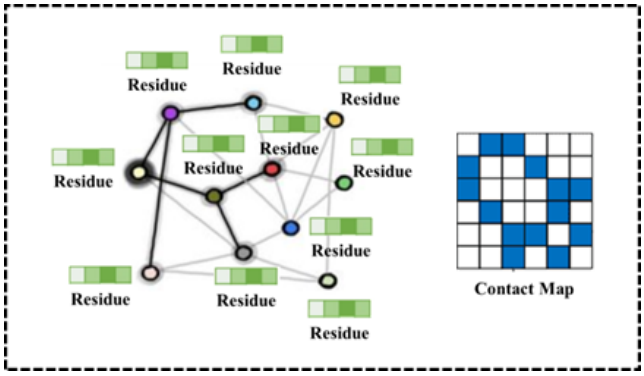

Figure 2: A graph convolutional network composed of residue representations and a contact map.

Protein Solubility Prediction Network Combined with graph convolutional network and fully connected network, we design an accurate protein solubility prediction network as 3shows. On one hand, the input of amino acid word frequency feature passes through the fully connected layer and changes from 20 dimensions to 1024 dimensions. On the other hand, the graph composed of one-dimensional sequence and three-dimensional structure information passes through three-layer graph convolutional neural network. In this process, each node constantly updates its own point features according to its relationship with its neighbor nodes. The three layers represent capturing the features of the neighbor nodes whose reference distance is less than or equal to 3. The obtained two 1024-dimensional features are concatenated together to form a 2048-dimensional feature, passing through the full connection layer for smoothing. Finally the Sigmoid activation function keeps the solubility result ranging from 0 to 1.

The input of the neural network is composed of the word frequency feature (20-dimensional) of an amino acid sequence, node feature and edge feature. And the output is the protein solubility corresponding to the input of amino acid sequence, the value of which ranges from 0 to 1. Due to most applications of protein solubility only need to know whether the protein is soluble or not, past work(Wu and Yu 2021; Chen et al. 2021; Khurana et al. 2018) have classified it as a classification problem. Therefore, we set the threshold p as 0.5 to classify whether it is soluble. If the output of solubility is greater than 0.5, the protein is considered soluble. Otherwise it is considered that the protein is insoluble.

Experiments

Environmental Setting

We implement a prototype of AlphaSolu by the deep learning framework PyTorch 1.12.1 with Python 3.8. Hardware support is shown as Table 2.

CPU	Intel(R) Xeon(R) Gold 5122 CPU @ 3.60GHz
GPU	NVIDIA GeForce RTX 2080 Ti
CUDA Version	10.2
Operating System	Ubuntu 18.04.6
Ubuntu 18.04.6	440.44

Table 2: Environmental setting.

Training and Testing

In order to improve the reliability of solubility prediction, we use K-fold approach for cross-validation, in which k can be used as a hyperparameter to adjust. In the current experiment, the value of k is set to 5, that is, the original data is divided into 5 subsets, one of which will be used as the test set and the other 4 subsets as the training set. The training set size is 2512 and the test set size is 628. According to the original data, the features are extracted and the three features described above are obtained as the input of neural network.

We set the batch size and learning rate to 16 and 1e-4, separately. The weight decay is 1e-5 and the total number of epochs is set to 30. However, we set early stopping whose patience is 10 to avoid overfitting. Cross entropy loss function is performed to evaluate the optimization.

Performance Evaluation

Evaluation Metrics We classify the protein solubility prediction into a classification problem. Therefore, for a classification problem, the most important thing is whether the predicted value is the same as the true label. Since this is a binary classification problem, there are only four relationships, that is, 0 is predicted to be 0, 0 is predicted to be 1, 1 is predicted to be 0, 1 is predicted to be 1, and these four relationships are recorded as TN, FP, FT, TP.

We use metrics including accuracy, precision, recall, F1-score, AUC, MCC (Matthew's correlation coefficient) to evaluate the performance of our framework that are widely used in past work(Wu and Yu 2021; Khurana et al. 2018; Chen et al. 2021).

Results To ensure the robustness of our framework, the experiment of protein solubility prediction is repeated several times and the accuracy averaged is 80.90%. In comparison to some state-of-the-art approaches in recent years for predicting protein solubility, such as GraphSol (Chen et al. 2021) with an accuracy of 78% and EPSOL (Khurana et al. 2018) with an accuracy of 79%, our framework AlphaSolu demonstrates a significant improvement in accuracy.

We not only consider accuracy as a crucial metric but also evaluates various other metrics mentioned above. Corresponding to 3, the F1-score reaches 0.797, and the AUC is 0.813. Both metrics reflect the robustness and stability of the model, indicating that the predictions of protein solubility using our framework are reliable. A comprehensive comparison of all metrics with advanced protein solubility prediction approaches is also presented in Table 3, clearly demonstrating the superior accuracy and reliability of the proposed framework.

Besides, we use some metrics such as prediction response time, ease of operation, and code maintainability to evalu-

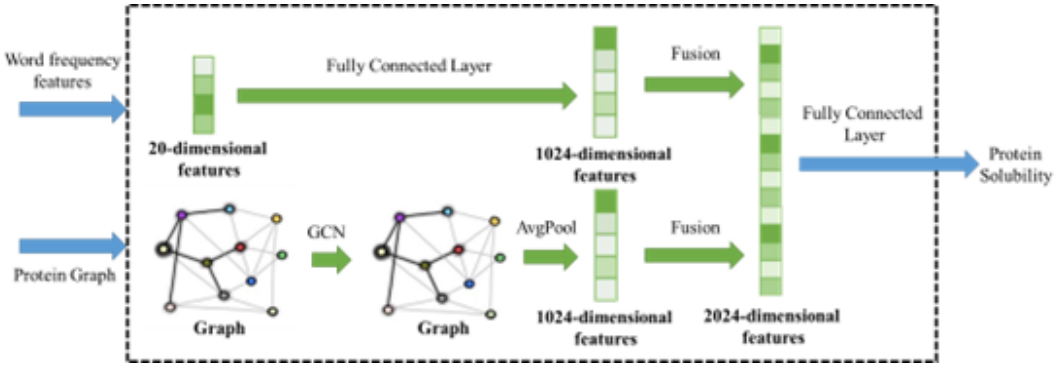

Figure 3: The protein solubility prediction neural network in AlphaSolu.

Approach	ACC	Precision	Recall	F1	AUC	MCC
XGboost(Chen and Guestrin 2016)	0.756	0.748	0.690	0.718	0.829	·
DeepSol(Khurana et al. 2018)	0.763	0.771	0.738	0.695	0.845	0.540
GraphSol(Chen et al. 2021)	0.782	**0.790**	0.702	0.743	**0.873**	·
EPSOL(Wu and Yu 2021)	0.790	0.787	0.787	0.787	·	0.580
Our	**0.809**	0.752	**0.849**	**0.797**	0.813	**0.622**

Table 3: The correctness compared with other protein solubility prediction approaches.

ate AlphaSolu, as illustrated in Table 4. By comparing with other protein solubility approaches presented in Table 4, AlphaSolu demonstrates superior performance in terms of the time required for predicting compared to EPSOL(Wu and Yu 2021) and GraphSol(Chen et al. 2021), which have a considerable amount of features, resulting in a longer processing time. Additionally, our framework give an visual interface leading a easy way to use.

Approaches	Time(s)	Visual Interface	Maintainability
Our	3.24(±1.23)	√	√
EPSOL	3.98(±2.47)	×	√
GraphSol	10.62(±1.59)	×	√

Table 4: The efficiency and maintainability compared with other protein solubility prediction approaches.

Conclusion

With the continuous advancement of genetic engineering, the expression of heterologous proteins has found applications in various fields. However, poorly soluble proteins may form insoluble aggregates or inclusion bodies, leading to a decrease in the yield of the target protein. Predicting protein solubility in advance can facilitate large-scale screening and production in industries, thereby enhancing efficiency. Therefore, our framework developed a protein solubility prediction that outperforms other state-of-the-art approaches.

We use amino acid sequences and corresponding solubility in the eSOL database as training and testing sets. Experiments indicate that the proposed three-dimensional structure-based protein solubility prediction framework has an excellent performance. It achieves higher accuracy, shorter prediction times, and provides users with a visual interface for facilitating user-friendly interactions.

However, in three-dimensional structure PDB files, besides atomic coordinates, there is still a wealth of information like bond angles, bond lengths, charges, and so on. Therefore, we plan to further explore and analyze these information through feature engineering to predict protein solubility.

References

Bryant, P.; Pozzati, G.; and Elofsson, A. 2022. Improved prediction of protein-protein interactions using AlphaFold2. *Nature communications*, 13(1): 1265.

Chen, J.; Zheng, S.; Zhao, H.; and Yang, Y. 2021. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *Journal of cheminformatics*, 13(1): 1–10.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; and Baldi, P. 2005. SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research*, 33(suppl_2): W72–W76.

Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J. R.; Booth, V.; Mackereth, C. D.; Saridakis, V.; Ekiel, I.; et al. 2000. Structural proteomics of an archaeon. *Nature structural biology*, 7(10): 903–909.

Davis, G. D.; Elisee, C.; Newham, D. M.; and Harrison, R. G. 1999. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnology and bioengineering*, 65(4): 382–388.

Han, X.; Wang, X.; and Zhou, K. 2019. Develop machine learning-based regression predictive models for engineering protein solubility. *Bioinformatics*, 35(22): 4640–4646.

Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; and Zhou, Y. 2018. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, 34(23): 4039–4045.

Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; and Rost, B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1): 1–17.

Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; and Damborsky, J. 2021. SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics*, 37(1): 23–28.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; and Mall, R. 2018. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15): 2605–2613.

Mount, D. W. 2008. Using BLOSUM in sequence alignments. *Cold Spring Harbor Protocols*, 2008(6): pdb–top39.

Niwa, T.; Ying, B.-W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; and Taguchi, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences*, 106(11): 4201–4206.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. arXiv:1802.05365.

Rawi, R.; Mall, R.; Kunji, K.; Shen, C.-H.; Kwong, P. D.; and Chuang, G.-Y. 2018. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, 34(7): 1092–1098.

Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; and Frishman, D. 2012. PROSO II–a new method for protein solubility prediction. *The FEBS journal*, 279(12): 2192–2200.

Smialowski, P.; Martin-Galiano, A. J.; Mikolajka, A.; Girschick, T.; Holak, T. A.; and Frishman, D. 2007. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, 23(19): 2536–2542.

Song, B.; Luo, X.; Luo, X.; Liu, Y.; Niu, Z.; and Zeng, X. 2022. Learning spatial structures of proteins improves protein–protein interaction prediction. *Briefings in bioinformatics*, 23(2): bbab558.

Wilkinson, D. L.; and Harrison, R. G. 1991. Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/technology*, 9(5): 443–448.

Wu, X.; and Yu, L. 2021. EPSOL: sequence-based protein solubility prediction using multidimensional embedding. *Bioinformatics*, 37(23): 4314–4320.

Yuan, Q.; Chen, S.; Rao, J.; Zheng, S.; Zhao, H.; and Yang, Y. 2022. AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Briefings in Bioinformatics*, 23(2): bbab564.