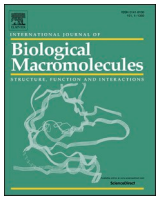
International Journal of Biological Macromolecules 278 (2024) 134601

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac



AlphaFold 2-based stacking model for protein solubility prediction and its transferability on seed storage proteins

Hyukjin Kwon, Zhenjiao Du, Yonghui Li *

Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA

ARTICLE INFO

Keywords:
Deep learning
AlphaFold
Protein solubility prediction
Alternative food protein
Seed proteins

ABSTRACT

Accurate protein solubility prediction is crucial in screening suitable candidates for food application. Existing models often rely only on sequences, overlooking important structural details. In this study, a regression model for protein solubility was developed using both the sequences and predicted structures of 2983 *E. coli* proteins. The sequence and structural level properties of the proteins were bioinformatically extracted and subjected to multilayer perceptron (MLP). Moreover, residue level features and contact maps were utilized to construct a graph convolutional network (GCN). The out-of-fold predictions of the two models were combined and fed into multiple meta-regressors to create a stacking model. The stacking model with support vector regressor (SVR) achieved R^2 of 0.502 and 0.468 on test and external validation datasets, respectively, displaying higher performance compared to existing regression models. Based on the improved performance compared to its based models, the stacking model effectively captured the strength of its base models as well as the significance of the different features used. Furthermore, the model's transferability was indirectly validated on a dataset of seed storage proteins using Osborne definition as well as on a case study using molecular dynamic simulation, showing potential for application beyond microbial proteins to food and agriculture-related ones.

1. Introduction

The food industry's concern about the sustainability of animal protein has grown significantly in the past decade, given the 58 % increase in meat demand over 20 years [1]. Recognizing the need for a more sustainable food system, researchers are exploring alternative protein sources, including plants, insects, and transgenic organisms [2]. However, selecting food proteins requires considering specific techno-functional properties. Aqueous solubility, in particular, is vital as it directly impacts various protein functionalities and key food characteristics like mouthfeel and digestibility [3]. Consequently, there is a growing demand for computational tools that can effectively predict or screen protein solubility, which would aid in selecting the optimal protein for a desired food product [4].

Protein solubility is a concept that researchers often approach with different definitions. While various methods for measuring solubility exist, such as induced precipitation using ammonium sulfate [5] or concentration via ultrafiltration [6], protein solubility can be broadly categories into relative and maximum solubility. Relative solubility refers to the fraction of the protein that remains in the supernatant after centrifugation [7]. On the other hand, maximum solubility pertains to the amount of a protein that can be dissolved in water without observable precipitation [8]. While maximum solubility aligns with the thermodynamic definition of solubility, it poses practical challenges in measurements, especially in the case of food proteins where multiple protein types are present. Therefore, a significant portion of research conducted in the fields of food and agriculture science relies on the relative definition of solubility [8].

With the advent of deep learning, researchers have adapted various model architectures to effectively predict the physicochemical properties of proteins. Initially, the basic multilayer perceptron (MLP) was used to predict protease cleavage sites [9] and to classify enzymes into six families [10]. Improvements in MLP, such as dropout [11], Xavier initialization [12], and diverse activation functions, have further expanded its application in modeling the behaviors of proteins. Examples include classification of seed storage proteins by Osborne fractionation [13], prediction of DNA and RNA binding sites [14], and identification of amyloid protein [15].

While MLP enables modeling complex, non-linear phenomena, the architecture struggles to handle the spatial arrangements of amino acids

* Corresponding author.
E-mail address: yonghui@ksu.edu (Y. Li).

https://doi.org/10.1016/j.ijbiomac.2024.134601
Received 7 May 2024; Received in revised form 29 July 2024; Accepted 7 August 2024
Available online 11 August 2024

in proteins. To address this, spatial models like graph convolutional network (GCN) have emerged as powerful tools. Originally proposed for node classification tasks [16], GCN learns on graph-structured data by aggregating feature information from neighboring nodes. This allows the architecture to capture both local graph topology and node attributes and makes GCN effective for representing protein structures, where nodes represent amino acids and edges represent their interactions. The advancement of attention mechanism [17] further enhanced GCN models by allowing them to dynamically emphasize on specific parts of the input data, thereby assigning levels of importance to different elements. This selective focus improved the learning of node representations, enhancing the model's performance and robustness on proteins. For instance, Baranwal et al. proposed a structure-based prediction model for protein-protein interaction sites using a graph attention network [18]. Similarly, Cheng et al. utilized a multi-head, self-attention graph network to predict drug-target protein interactions [19].

Another drawback of MLP is their inability to effectively handle context-aware data, which could be crucial for modeling protein sequences. The biological meaning of specific amino acids and motifs can heavily depend on their context within the sequence, which MLP struggles to capture without complex and impractical modifications. Recently, sequential models, particularly transformer-based protein language models, have been increasingly applied in the field. Pre-trained language models (PLM) like evolutionary scale modeling (ESM) [20] and ProtBert [21] have gained attention for their computational efficiency and exceptional performance in representing protein sequences [22]. Such PLMs provide embedding per each amino acid in a given protein sequence. These representations capture the contextual information of each amino acid within the sequence, similar to how embeddings in natural language processing capture the context of words in sentences [23]. The embeddings generated from pre-trained language models have been applied to predict allergenic proteins [24] and to identify alkaliphilic proteins [25].

Extensive research and development have been devoted to protein solubility prediction models due to its critical importance across various fields. Given the complex nature of protein solubility, various machine learning and deep learning architectures have been widely employed in these models [26–28]. Such tools can be classified based on several criteria; one of which is whether they perform binary classification or regression. In binary classification models, proteins are categorized as either soluble or insoluble, and the majority of the developed models fall into this category, including PROSO [29], PaRSnIP [30], and PLM_Sol [31]. The prevalence of binary classifiers can be attributed to the fact that the most available datasets on protein solubility, including the widely used TargetTrack database [32], define solubility as either soluble or insoluble. On the other hand, regression models provide continuous values for protein solubility, typically ranging from 0 to 1. GraphSol [33], and SoLart [34] are examples of regression models. Although fewer in number compared to binary counterparts, regression models provides significant engineering advantages as they provide insights into the predicted values' magnitude and offer easier outlier detection.

Another criterion for categorizing solubility prediction tools is the type of features they utilize. Sequence-based models exclusively rely on features derived from the protein sequence itself, including sequence length, amino acid composition, and isoelectric point, etc. [34]. For example, SVM-based CCSol combines six different sequence information such as secondary structure propensity, hydrophobicity, and hydrophilicity to define solubility parameters for each region by a sliding window of 21 amino acids [35]. Moreover, self-defined features based on domain knowledge are also employed. Another sequence-based model ProteinSol [36] utilizes different combinations of the ratio of amino acid residues, such as K-R or D-E, employing 35 sequence features in total. On the other hand, structure-based or structure-aware models leverage features extracted from the 3D structure of the proteins. These structural features can include the fraction of exposed residues or surface hydrophobicity. For instance, SoLart employs solubility-dependent statistical potentials derived from the type of residue, interatomic distance and angles as well as other structural features like solvent accessibility or the fraction of buried residues [34]. While incorporating 3D information into the model has the potential to enhance its performance, the availability of crystal structures for proteins is not always guaranteed. As a result, most existing models such as SOLpro [28], and Protein-Sol [36] were developed based solely on protein sequence information. However, the emergence of modern ab-initio structure prediction tools like AlphaFold (AF) offers promise for addressing this limited structure availability.

Recently, PLMs have been increasingly integrated into solubility prediction tools. PLMs typically produce embeddings for each residue, allowing graph-based models like GCN to be applied with predicted protein structures. For instance, Wang et al. used embeddings generated from ProtTrans PLM as node features [37]. Using AF-predicted structures, they constructed the binary GCN classifier DeepMutSol, which outperformed all state-of-the-art (SOTA) classifiers. Moreover, PLM-generated embeddings can be combined with other residue-based features. For example, Chen et al. utilized combined features from ESM-1v embeddings and evolutionary features like position-specific scoring matrix (PSSM) or hidden Markov model (HMM) [38]. With the predicted protein contact map from SPIDER3, they developed a solubility regressor HybridGCN, achieving ~10 % higher coefficient of determination compared to the previous SOTA model GraphSol on an external validation dataset. More recently, Li and Ming [39] combined the blosum62 matrix with ESM-1b embeddings to construct a GCN solubility regressor, GATSol, which is the current SOTA regressor based on AF-predicted structures.

In this study, regression models for predicting protein solubility were trained by employing two distinct physicochemical feature sets generated by both protein sequences and AF computed structures. The first set encompassed features derived from the protein sequences and structures, including hydrophobic index (GRAVY) [40] and hydrophobic patch area [41]. The second set comprised protein contact maps and residue-level features, such as seven physicochemical properties of amino acids (AAPHY7) [42]. The sequence + structural feature set was processed using MLP, while the residual feature + contact map set was managed through GCN. Furthermore, to harness the information provided by both approaches, a stacking method was employed, integrating the outputs of the MLP and GCN models. The performance of the stacking model was compared with existing regression models, and two regressors trained with ESM-2 pre-trained model. The relative importance of each features used in the base models MLP and GCN was computed from their saliency map. Additionally, the stacking model's transferability to plant proteins was indirectly validated using the Osborne definition of seed storage proteins. A case study involving coarse-grained molecular dynamic (CGMD) simulation was also conducted on two proteins with low and high predicted solubility.

2. Materials and methods

All datasets and the codes utilized to train and validate the models in this study are available at: https://github.com/john94kwon/Stacking-model-for-solubility-prediciton

2.1. Protein solubility definition

In this study, relative solubility of *E. coli* proteins was utilized to develop prediction models. Solubility was defined as the proportion of supernatant obtained after centrifugation divided by the initial quantity of overexpressed protein [43].

2.2. Datasets

2.2.1. E. coli dataset

A dataset consisting of solubility of proteins from the K-12 *E. coli*

strain was utilized for model development. The dataset encompassed solubility measurements for 3147 proteins that were expressed in a cell-free system [43]. Using the provided entry information in the data, the corresponding protein structures were retrieved from the AF protein structure database (https://alphafold.ebi.ac.uk/) as pdb format. In order to maintain structural reliability, structures with a pLDDT score below 70 for at least half of the total residues were discarded [44]. The pLDDT score represents the confidence in the location of each residue [45]. Furthermore, to mitigate biases arising from evolutionarily similar proteins, HHblits [46] was utilized to filter out those with <25 % sequence similarity [47]. To maintain consistency, only structures computed by AF were collected even for the sequences with experimentally determined structures. A total of 2983 protein structure files remained for analysis. This data was further split into training and testing sets, with 75 % (2237) allocated for training and 25 % (746) reserved for testing purposes. 5-fold cross validation along with grid search was employed for hyperparamter tuning.

2.2.2. *S. cerevisiae* dataset

For external validation of trained models, another solubility dataset consisting of *S. cerevisiae* (brewer's yeast) proteins was selected. This dataset included the solubility measurements of 108 yeast proteins that were overexpressed in the same system as *E. coli* proteins above [48]. We specifically chose the *S. cerevisiae* dataset to compare the performance of our models to existing ones, as several prediction models have already undergone external validation with the dataset [33,34].

2.2.3. Seed storage protein dataset

To investigate the model's relevance in the solubility of plant proteins, a dataset comprising AF structures of a range of seed storage proteins was gathered. A dataset of 200 structures was compiled, covering 50 distinct structures per each Osborne classifications: albumin, globulin, prolamin, and glutelin. Specifically, the UniProt database was searched using the tag "seed storage protein" along with the Osborne classification (e.g., "seed storage protein glutelin"). The protein structures and sequences reviewed by UniProt were exclusively used for albumin, globulin, and prolamin. In the case of glutelin class, due to limited data availability, both reviewed and automatically annotated proteins were utilized. The number of source organisms for each class was 28 for albumin, 29 for globulin, 13 for prolamin, and 28 for glutelin (Supplementary Document 1, Table S1). The specific names, entries, and sequence information for all 200 storage proteins were provided as a separate .csv file (Supplementary Document 2). Moreover, the properties of five seed storage proteins from single sources, each belonging to a different Osborne class (2S albumin from mouse-ear cress, 11S legumin from pea, zein from maize, and glutelin from rice), were compared with those of human reference proteins. The names and entry of the compared proteins were also listed in Supplementary Document 2. The statistical significance between the selected proteins was computed using analysis of variance (ANOVA) with the R statistical package (ver 3.5.0), followed by Duncan's multiple range test with a confidence level of 95 %.

2.3. Feature extraction for MLP model

To train the MLP model, a feature set comprising a range of protein descriptors utilized in the field of bioinformatics was employed [49]. The features were extracted with the biopython module ver 1.81 [50], and included molecular weight, aromaticity, instability index, hydrophobic index, aliphatic index, absolute charge per residue, and hydrophilic index. In addition to these sequence features, an assortment of structural characteristics was extracted from the protein 3D structures computed from AF structures. Specifically, the QUILT software ver 1.3 [51] was employed to calculate the hydrophobic patch area, hydrophilic surface area, and the ratio of hydrophobic surface to total surface area. QUILT calculates the solvent-accessible surface area (SASA) of a protein by rolling a probe that emulates the van der Waals radius of a solvent molecule. The surface is defined by the path of the probe's center and the number of discrete sampling points. The software then identifies contiguous apolar surface areas, namely around sulfur and carbon atoms. In this study, a probe radius of 1.4 Å and 252 sampling points were used, following the suggested parameters from the author of the software (https://github.com/plijnzaad/quilt). The hydrophilic surface area was determined by subtracting the hydrophobic patch area from the total SASA, while the ratio of hydrophobic patches was calculated by dividing the total hydrophobic patch area by the total SASA.

Furthermore, with the ChimeraX software ver 1.6.1, a range of additional structural features was generated. These features encompassed mean lipophilicity potential, mean coulombic potential, surface area, volume, solvent-accessible surface area, number of hydrogen bonding, helix propensity, coil propensity, strand propensity, and number of favorable contacts. Prior to the feature extraction, all pdb structures files were converted into pqr format files using the pbd2pqr plugin ver 3.6.1 [52] with pH 7.0 under CHARMM force field. This conversion step was essential to account for the protonation states of the proteins, as they may significantly impact the coulombic potential or number of hydrogen bonding. Similar to QUILT's calculation of SASA, ChimeraX also determines the molecular surface of a protein, but it defines the surface by the trajectory of the probe sphere's outer edge instead of its center. This molecular surface is used to calculate the mean coulombic potential, mean lipophilicity potential, and molecular volume. The mean coulombic potential is calculated as the average electrostatic potential using atomic partial charges, coordinates, and distances from the surface. The mean lipophilicity potential is computed using atomic hydrophobicity values, coordinates, and distances from the surface. Molecular volume is determined by the volume enclosed by the molecular surface. The number of hydrogen bonds was calculated using default geometric criteria in ChimeraX [53]. Favorable contacts were determined by subtracting the number of clashes from the total contacts. In ChimeraX, contacts include all direct interactions (polar and nonpolar), while clashes are defined as unfavorable interactions where atoms are too close, considering their van der Waals radii. All feature extractions in ChimeraX were computed using default parameters.

In sum, a total of 7 sequence features and 13 structural features were gathered. The descriptions, example codes and results of each features using an example protein (B1-hordein from Barely) were provided in Table S2 and S3. No exclusive feature selection was conducted, as the utilization of common feature selection methods such as Pearson correlation, K-best selection, and recursive feature elimination resulted in inferior performance (Table S4).

2.4. Feature extraction for GCN model

In a graph neural network, a graph is represented as a collection of interconnected vertices. These vertices, commonly referred to as nodes, establish the foundation of the graph, while the connections between them are represented as edges [54]. When considering proteins, which are composed of linked residues, a similar graph-based representation can be applied. In this context, the amino acid residues of a protein are the nodes, and their interconnections—whether they involve physical contacts or actual peptide bonds—serve as the edges.

2.4.1. Node features for GCN model

In contrast to the features employed in the MLP model, where each feature denoted a characteristic of the entire protein, the features utilized in the GCN model represented the attributes of individual residues. Three distinct node features, including AAPHY7, Blosum62, and the coordinates of the alpha carbon were utilized to capture the characteristics of each amino acid residue. Briefly, AAPHY7 is a list of seven physicochemical properties of each amino acids, which includes steric parameter, residue hydrophobicity, residue volume, polarizability, isoelectric point, helix propensity, and sheet propensity [55]. Blosum62 is a

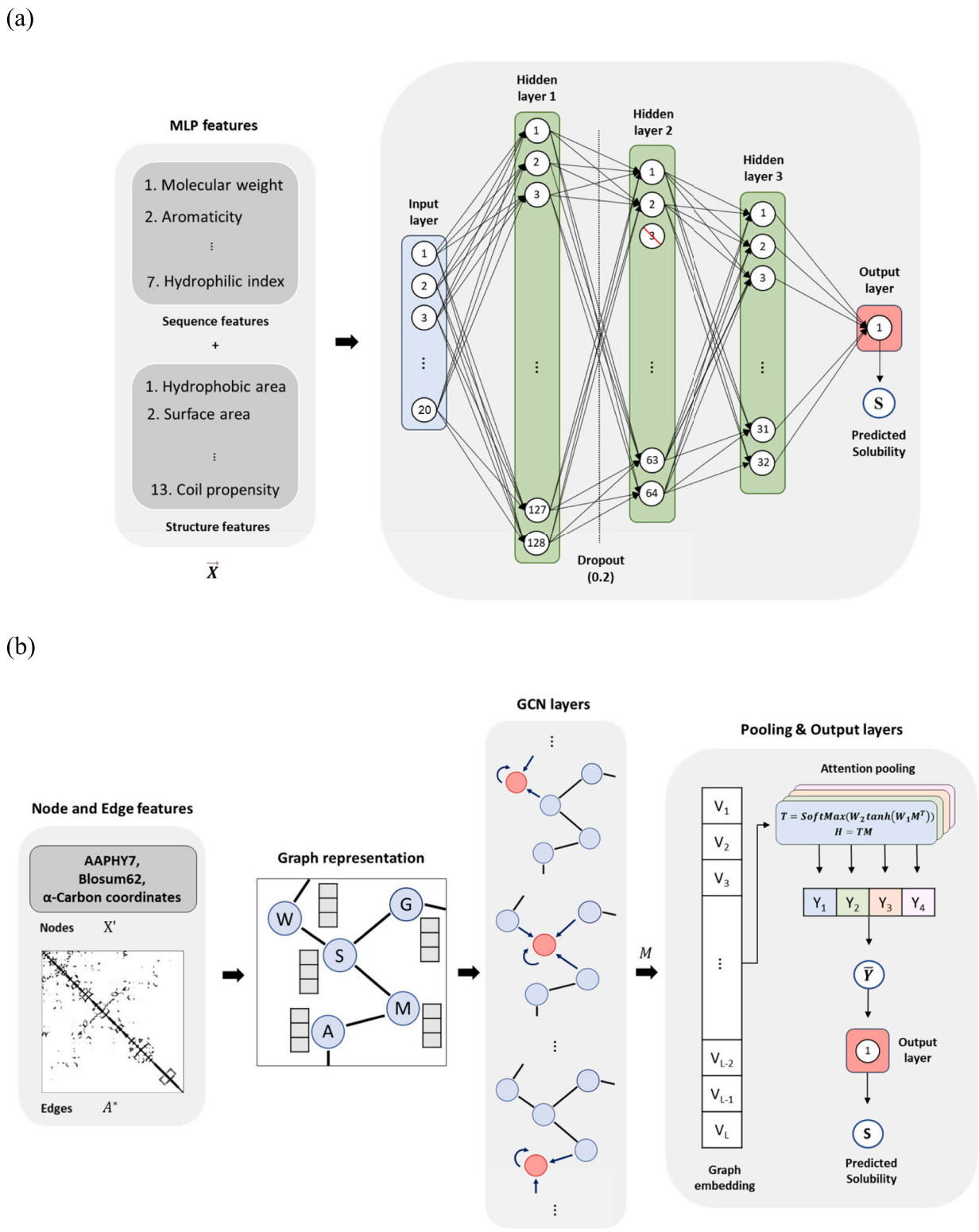


Fig. 1. The structure of the MLP (a), and GCN (b) model constructed.

substitution matrix derived from conserved regions of protein families. The higher score in the blosum matrix indicates favorable conservation, and physicochemical similarity between two residues [56]. Detailed descriptions on the acquisition of these residue-level features used were listed in Table S2 and S3.

2.4.2. *Protein contact map (edge for GCN model)*

In graph neural networks, the edge of a graph are often represented by its adjacency matrix A [57], which is a [n x n] matrix defined by Eq. (1). When dealing with proteins, their contact maps can be used to produce the adjacency matrix of the protein. The generation of contact maps involves calculating the distances between the alpha carbon residues and binarily mapping these distances based on a specified

threshold. Such contact maps, by definition, are consistent with the adjacency matrix used in graph neural networks and can be employed as edge matrix [58]. It should be noted that contact maps are equivalent of $\widetilde{A}$, the adjacency matrix added with identity matrix (Eq. (2)). Typically, the threshold distance for protein contact maps falls within the range of 6 to 12 Å [59]. In the context of this research, multiple contact maps were created using python script, and after evaluation, the contact map generated with a 12 Å threshold exhibited the highest performance and was therefore selected as the optimal choice.

$$[A]_{ij} = 1 \; if \; node \; i \; and \; j \; are \; connected, = 0 \; otherwise \tag{1}$$

$$\widetilde{A} = A + I \tag{2}$$

2.5. MLP model structure

The structure and parameters of the constructed MLP model were displayed in Fig. 1(a). The input to the model was a feature vector $\overrightarrow{X}$, which consisted of 20 sequence and structure features. Initially, $\overrightarrow{X}$ was fed into the input layer, and underwent a series of transformations as it passed through the hidden layers. These transformations were determined by the adjustable layer weights and activation functions, as defined by Eq. (3). In the equation, H^i represents the transformed feature vector at the i^{th} hidden layer, W^i denotes the trainable weight of the layer, σ denotes the activation function, and b is the bias term. At $i = 0$, the feature vector X is H^0. Moving toward the right side of the figure, the resulting vector from the third hidden layer was further processed at the output layer, where it was transformed into the predicted solubility, denoted as S. The hidden layers used in the MLP model consisted of 128, 64, and 32 weights, respectively. The LeakyReLU activation function was applied to these hidden layers. Moreover, a dropout layer with the rate 0.2 was employed at the first hidden layer to alleviate overfitting, and batch normalization was utilized in all three hidden layers [60]. To scale the predicted solubility within the range of 0 to 1, the sigmoid activation function was used at the output layer. The root mean square error (RMSE) was and Adamax optimizer were used as the loss function and the optimizer, respectively.

$$H^{[i+1]} = \sigma\left(W^i H^i + b\right) \tag{3}$$

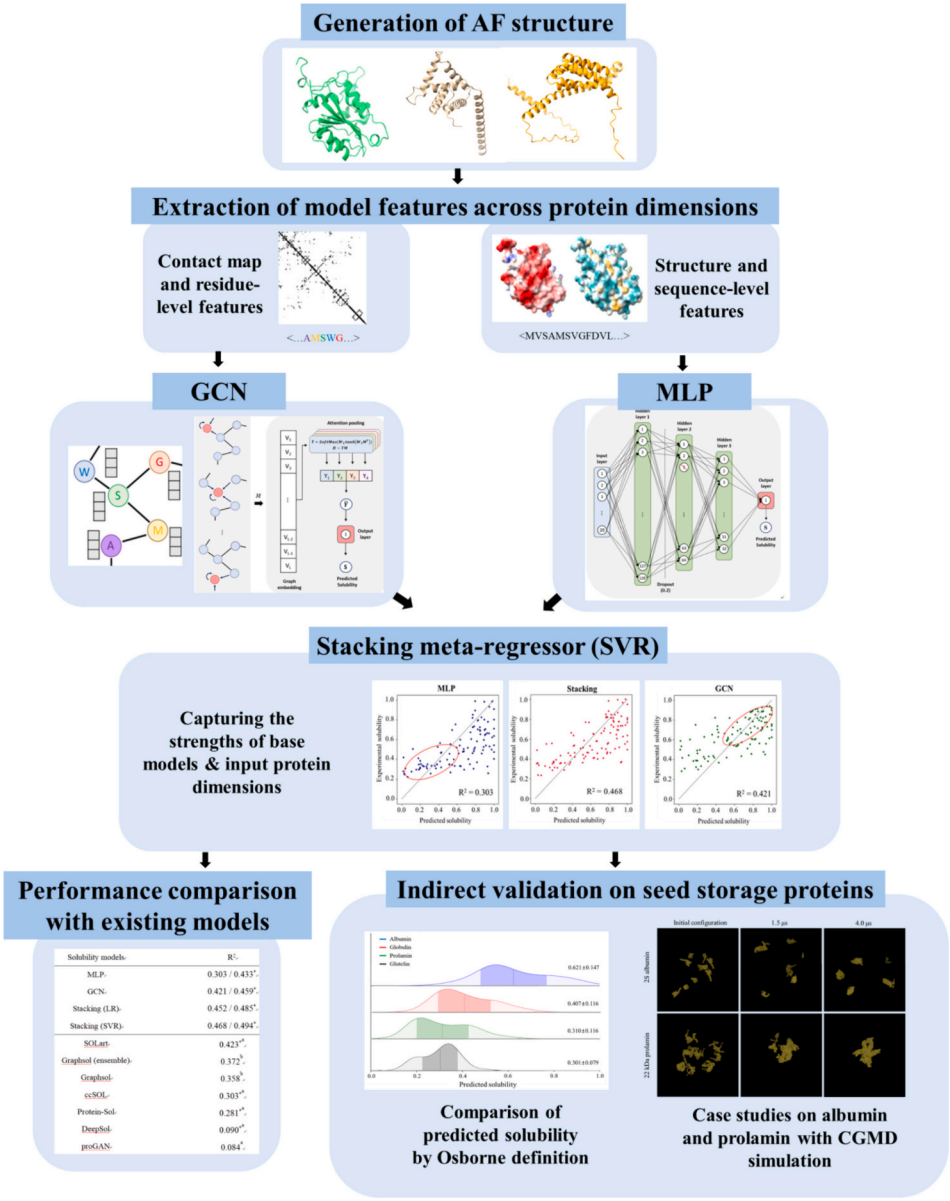


Fig. 2. Overall workflow for the construction and validation of the stacking solubility model.

2.6. GCN model structure

The basic structure of the GCN model used in this experiment was adapted from the work of Chen et al. [33]. The model architecture and its process were displayed in Fig. 1(b). The protein graph was represented by the combination of the node feature vector X' and the generated contact map $\tilde{A}$. These two vectors were then fed into the GCN model, which consists of GCN layers and a pooling & output layer.

2.6.1. GCN layers

In the GCN layer, the node features of individual residues were updated based on the features of their neighboring residues as well as their own. This process is illustrated in the "GCN layer" of Fig. 1b and can be mathematically represented by Eq. (4) [16]. In a similar manner to MLP model, GCN model involved the feature vector at the ith layer denoted as H^i, the trainable weights of the ith layer was denoted as W^i, and the activation function was represented by σ. The key difference between MLP and GCN stems from the additional element A*, which corresponded to $\tilde{A}$ normalized by the degree matrix D (Eq. (5)). The normalized adjacency matrix A* encoded the connectivity between the nodes, and the dot product of the normalized adjacency matrix and node feature matrix, A·X' represented sum of neighboring node features. The weight matrix W, was then updated to synchronize the impact of neighboring nodes to accurately represent a protein graph, and eventually its solubility. Thus, unlike MLP, where each feature was treated independently, the nodes in GCN were considered in connection with other neighboring nodes based on the contact map provided [61]. The GCN model employed two hidden layers with LeakyReLU activation functions, consisting of 256 and 64 weights, respectively. Batch normalization was applied to normalize the layers to mitigate overfitting.

$$H^{[i+1]} = \sigma\left(A^* H^i W^i + b\right) \tag{4}$$

$$A^* = D^{-\frac{1}{2}} \tilde{A} D^{\frac{1}{2}} \tag{5}$$

2.6.2. Pooling & output layer

One problem with the GCN layer presented would be the variation in protein length. If a protein's length was L and had f number of node features, the size of A^* and X' would be [L x L] and [L x f], respectively. This would make their dot product an [L x f] matrix. If the size in the first and second GCN layers were set to be [f x p] and [p x q], respectively, then the resulting matrix M from passing two GCN layers would have the dimension of [L x q]. Consequently, this matrix M would then assume a variable size dependent on L. To address this variability in protein size, the multi-head self-attention pooling mechanism proposed by Lin et al. [62] was employed. In Eq. (6), T represents the attention scoring matrix, which determines the significance of the relationships between each residue and all the other ones [63]. W1 and W2 are learnable parameters. If the shapes of W1 and W2 are [m x L] and [n x m], respectively, then the final dimension of T would be [n x L]. By taking the dot product of T and M, the resulting matrix Y assumes a fixed size of [n x q], regardless of the protein length L (Eq. (7)). The number of attention heads, or the number of attention mechanisms applied in parallel, was set to four. The average of these four vectors was passed through an output layer to produce the predicted solubility. A sigmoid activation function was employed at the output layer.

$$T = SoftMax\left(W_2 tanh\left(W_1 M^T\right)\right) \tag{6}$$

$$Y = TM \tag{7}$$

2.7. Benchmark models: MLP and GCN with ESM-2 embedding

To compare the performance of models derived from the features in Sections 2.3 and 2.4, another set of parameters was obtained using the pre-trained protein language model ESM-2. Developed by Facebook researchers, ESM-2 is the latest version of ESM and has outperformed many existing models in various structure forecasting tasks [20]. Given the relatively small size of the training dataset and the number of features in Section 2.3 (20 features), the esm2_t6_8M_UR50D model, which generates the lowest dimension embeddings (320 features), was selected. The created 320 features per residues were averaged by the sequence length to ensure equal feature dimension (1 × 320). The generated sequence embeddings were then fed into the MLP structure described in Section 2.5. Moreover, the embeddings were directly fed into the GCN model as node features without taking their average. In the case of the GCN model, the number of epochs was reduced from 10 to 5 to prevent model overfitting.

2.8. Stacking model

The two feature sets used for MLP and GCN might offer different perspectives on the specific protein. In the MLP model, features derived from the protein sequences and unmodified structures were utilized. Conversely, the GCN model trained on graph structures reconstructed from residual features and contact maps. To leverage the information provided by both approaches, different meta-regressors including linear regression (LR), support vector regressor (SVR), decision tree (DT), and k-nearest neighbor (KNN) were employed to combine the outcomes of the MLP and GCN models (Fig. 2). Specifically, the meta-regressors were trained using the out-of-fold cross validation results from each 5-fold cross-validation from individual base models to prevent data leakage and model overfitting.

2.9. Model hyperparameters and evaluation metrics

To ensure optimal performance, all model hyperparameters were tuned using a 5-fold cross-validation technique on the training dataset. The specific hyperparameters, along with their respective values, can be found in Table S5. In order to assess the predictive power of the models constructed, coefficient of determination (R^2) and root mean square error (RMSE) were utilized. In an extension to the prediction of continuous solubility values, a binary classification was also performed with threshold solubility value of 0.5. Proteins with solubility values above 0.5 were classified as soluble, while those below were classified as insoluble. The performance of the classification was assessed by accuracy, F1 score, AUC (area under curve), and MCC (Matthews correlation coefficient). For the external validation using the *S.cerevisiae* dataset, both the coefficient of determination and the square of Pearson's correlation were utilized.

2.10. Identification of influential features/residues

To assess the importance of features in the base models MLP and GCN, saliency maps of the trained models were generated with respect to the input features. Training datasets were fed into the trained models with PyTorch's requires_grad function enabled, allowing the measurement of gradients during backpropagation. The models were set to evaluation mode to negate the effects of dropout and batch normalization. The magnitude of the saliency for each feature was averaged across individual training data to determine the overall influence of each feature. In the case study on seed storage proteins, the relative importance of each residue for the GCN model was identified by summing the magnitudes of all features per residue. The most important regions of residues were selected with thresholds of 0.20 for 2S albumin and 0.15 for 22 kDa prolamin. The electrostatic potential maps for the two proteins were generated using the adaptive Poisson-Boltzmann solver (APBS) web server [64] with the CHARMM force field and default parameters.

2.11. Case study: molecular dynamics simulation

The aggregation behavior of seed storage proteins was studied using CGMD simulation. AF predicted structures of 2S-albumin from soybean (Uniprot entry: P19594) and 22 kDa alpha zein prolamin from maize (Uniprot entry: P04700) were downloaded to prepare the initial coordinates and topologies of the proteins. Ten molecules of each protein structure were randomly inserted into water boxes with dimensions of $39.43 \times 39.43 \times 39.43\ nm^3$ (albumin) and $45.54 \times 45.54 \times 45.54\ nm^3$ (prolamin) using the gmx insert-molecules function in the GROMACS package. The box dimensions were chosen to replicate approximately 5 % protein concentration. The prepared system was converted into a coarse-grained system using the Martinize 2 script [65] under the Martini 3 force field. The INSANE (INSert membrANE) program (https://github.com/Tsjerk/Insane) was employed to solvate the proteins, and ions were added to neutralize the system. With the simulated system, 5000 steps of energy minimization were followed by 500 ns NVT equilibrium and 4 μs NPT production runs. The timestep used for equilibrium and production runs was 20 fs. The pressure and temperature were maintained at 1 atm and 300 K using the Parrinello-Rahman barostat and velocity-rescaling coupling thermostat, respectively [66]. All simulations were conducted using GROMACS (version 2022.5), and snapshots of the trajectories were generated using VMD software (version 1.9.3) [67].

3. Results and discussion

3.1. Comparison of extracted features among datasets

Due to the different source organisms, significant variance among the proteins in each dataset could exist. To gauge the variation in the extracted feature values within each dataset, the average and standard deviation of each MLP feature were presented in Table S6. As shown in the table, most values of the extracted MLP features were fairly consistent across the three datasets. The two most prominent differences were in the instability index (38.011 ± 10.269 for *E.coli*, 36.929 ± 7.648 for *S.cerevisiae*, and 61.186 ± 24.673 for seed storage) and SASA (16,151.322 ± 8318.966 for *E.coli*, 16,763.76 ± 6839.81 for *S.cerevisiae*, and 23,930.924 ± 10,064.226 for seed storage). However, even for these features, high variability within each set was observed, as indicated by their high standard deviations. For the GCN features, a direct comparison of proteins was not possible due to the nature of the graph-based model. However, the relative proportion of each amino acid was presented in the Table S7. Similar to the MLP features, no apparent difference among the datasets was observed. While structural differences in proteins across source organisms, particularly in conserved domains or motifs, likely exist, the proteins used in this study did not show strong deviations in the features employed to train the models.

Although the three datasets displayed similar range of feature values on average, it could be worthwhile to compare the characteristics of seed storage proteins to some reference proteins, especially of the four Osborne classes. Therefore, five AF predicted structures of three different human proteins, namely keratin, hemoglobin, and lysozyme were collected as reference proteins. The proteins were selected to represent a broad range of properties, as they are of different origins: connective tissues, blood, and saliva, respectively. Moreover, five AF structures of albumin, globulin, prolamin, and glutelin with respective single source were compared (Table S8). As shown in the table, the four Osborne groups exhibited significantly different physicochemical properties from each other and the reference proteins. For example, maize zein had the highest ratio of hydrophobic surface (0.707 ± 0.008), while 11S pea globulin exhibited the lowest (0.544 ± 0.012). Rice glutelin had the highest ratio of coiled secondary structures (53.208 ± 1.706 %), whereas zein had the lowest (26.086 ± 4.081 %). The differences among the proteins were also highlighted by the relative amino acid fractions, such as the highest Leu (18.838 ± 0.870 %) and Cys (5.790 ± 0.534 %) contents in zein and 2S albumin, respectively. This comparison suggested that, despite being categorized as seed storage proteins, the four Osborne fractions exhibit different physicochemical properties. Furthermore, the statistically significant differences among the proteins indicated a deviation of each Osborne class from the reference human proteins.

Table 1
Prediction performances of the MLP and benchmark models (cross-validation / testing).

Models[a]	Regression metrics		Binary classification metrics			
	R^2	RMSE	Accuracy	F1	AUC	MCC
LR	0.350 /0.368	0.256 /0.260	0.743 /0.742	0.728 /0.728	0.822 /0.831	0.486 /0.488
SGD	0.353 /0.365	0.255 /0.260	0.740 /0.736	0.724 /0.721	0.824 /0.830	0.481 /0.475
SVR	0.472 /0.443	0.231 /0.244	0.776 /0.772	0.748 /0.745	0.854 /0.849	0.546 /0.548
RF	0.431 /0.436	0.240 /0.245	0.765 /0.784	0.732 /0.753	0.838 /0.848	0.525 /0.561
GB	0.436 /0.441	0.239 /0.244	0.761 /0.785	0.732 /0.758	0.838 /0.849	0.517 /0.565
MLP	0.455 /0.469	0.234 /0.238	0.768 /0.795	0.737 /0.764	0.845 /0.858	0.530 /0.583

[a] Linear regression (LR), stochastic gradient descent (SGD), support vector regression (SVR), random forest (RF), Gradient boosting (GB), multilayer perceptron (MLP).

3.2. Linear correlation analysis between structure/sequence features with solubility

Prior to MLP model construction, a linear correlation analysis was conducted between each sequence/structure descriptor and solubility within *E.coli* dataset (Table S9). As discussed in Section 3.1, significant deviations in the descriptors compared to their mean values were observed, indicating the diversity of proteins in the dataset. The highest linear correlation with solubility ($r = 0.361$) was exhibited by the hydrophilic index, followed by absolute charge per residue ($r = 0.357$) and aliphatic index ($r = -0.335$). These correlations are in line with the expected behavior, as the presence of hydrophilic or charged residues are known to facilitate hydrogen bond and electrostatic attractions between the protein and surrounding water [68]. These interactions effectively prevent aggregation resulting from hydrophobic interactions and contribute to the promotion of protein solubility. In contrast, certain parameters, such as the instability index ($r = -0.042$) or strand propensity ($r = 0.097$), showed very weak or no linear influence.

3.3. Solubility prediction using MLP model

The predictive performances of the MLP model along with those of benchmark models were displayed in Table 1. On the test *E. coli* dataset, the MLP model exhibited testing R^2 of 0.469, followed closely by gradient boosting (GB) method at 0.441 and random forest (RF) model at 0.436. SVR, which has been extensively used in non-deep learning solubility prediction models, achieved R^2 of 0.443. The MLP model's highest performance seemed to stem from its capability in capturing intricate relationships as a deep-learning model [69]. Conversely, the stochastic gradient descent (SGD) regression and LR models demonstrated inferior performances, recording R^2 values of 0.365 and 0.368, respectively. The disparity in performance between models with higher R^2 (>0.43) and lower R^2 (<0.37) likely arose from the non-linear nature of protein solubility, as SGD in scikitlearn and LR are linear models [70]. The four non-linear models (MLP, RF, GB, SVR) also exhibited better accuracy (>0.77) in binary classification compared to (<0.73) of linear models (SGD, LR), with MLP displaying the highest accuracy of 0.795. Therefore, in both regression and binary classification, MLP model outperformed other benchmark architectures.

Table 2
Prediction performances of stacking models with different meta-regressors and benchmark/base models (cross-validation / testing).

Models	Regression metrics			Binary classification metrics		
	R^2	RMSE	Accuracy	F1	AUC	MCC
ESM-2 (MLP)	0.372 /0.364	0.251 /0.261	0.740 /0.750	0.691 /0.752	0.825 /0.841	0.481 /0.522
ESM-2 (GCN)	0.439 /0.453	0.239 /0.241	0.762 /0.773	0.723 /0.726	0.839 /0.859	0.519 /0.538
GCN	0.414 /0.435	0.262 /0.246	0.747 /0.787	0.762 /0.774	0.838 /0.857	0.558/ 0.577
MLP	0.455 /0.469	0.234 /0.238	0.768 /0.795	0.737 /0.764	0.845 /0.858	0.530 /0.583
GraphSol	0.469 /0.476	0.236 /0.237	0.779 /0.786	0.752 /0.759	0.860 /0.869	0.545 /0.557
Stacking (LR)	0.488 /0.498	0.227 /0.231	0.784 /0.801	0.755 /0.782	0.856 /0.874	0.563 /0.600
Stacking (DT)	0.455 /0.458	0.235 /0.240	0.766 /0.792	0.729 /0.749	0.839 /0.848	0.525 /0.577
Stacking (KNN)	0.473 /0.481	0.231 /0.235	0.775 /0.803	0.745 /0.781	0.851 /0.867	0.545 /0.602
Stacking (SVR)	0.487 /0.502	0.227 /0.230	0.783 /0.804	0.751 /0.783	0.856 /0.876	0.560 /0.604

3.4. Solubility prediction using GCN model

The GCN model was constructed and trained using the generated node features and contact maps from AF structures. Along with the GCN model, the latest regression model, GraphSol was trained as a reference using the same training set. While both the GCN and GraphSol employed the same solubility dataset for training, GraphSol incorporated a greater number of node features (91 features compared to the GCN model's 30 features). Furthermore, GraphSol utilized adjacency matrices generated from SPIDER3 instead of AF. Table 2 displayed the performances of the two models; the GCN model exhibited inferior predictive power compared to GraphSol in regression task (R^2 0.435 vs 0.476) but comparable classification accuracy (0.786 vs 0.787). The better performance of the GraphSol was likely due to the higher number of features, which would promote a more finely-trained model. Unlike the GCN model, Graphsol incorporated evolutionary features, namely hidden markov matrix (HMM) and position-specific scoring metrics (PSSM). Both HMM and PSSM are matrices derived from multiple sequence alignment (MSA) of a specific protein sequence, where related protein sequences are aligned based on similarities [71]. Using the aligned sequences, HMM and PSSM summarize the characteristics of a set of related protein. Specifically, they assign probabilities, or scores to residues based on the observed alignment. These scores convey evolutionary information and can be utilized to predict protein functionality and structure [72]. However, as the goal of this study is to create a model applicable to food and agricultural proteins, evolutionary information was excluded for training our GCN model, even with the inferior performance in the *E. coli* test dataset. Since MSA generated using a protein of particular group or lineage represents the sequences similar to the group, it was likely that HMM or PSSM produced solely from *E. coli* dataset would over-represent the group of protein related to *E. coli* family [73,74].

3.5. Performance of the stacking model

Stacking model, also known as ensemble model, is a powerful approach that combines the prediction results of multiple models to create a more accurate and generalized model [75]. As indicated in Table 2, the stacking model with SVR regressor constructed in this research achieved R^2 of 0.502, outperforming MLP and GCN using ESM-2 embeddings ($R^2 = 0.364$ and 0.453, respectively), its base models MLP ($R^2 = 0.469$), GCN ($R^2 = 0.435$) as well as GraphSol ($R^2 = 0.476$) in the regression task. Among the meta-regressors used, SVR displayed the highest performance ($R^2 = 0.502$) followed by LR ($R^2 = 0.498$), KNN ($R^2 = 0.481$) and DT ($R^2 = 0.458$). The better performance of the stacking model compared to its base models suggested that the meta-regressors successfully captured crucial aspects from the respective base models. To visually depict the prediction performance of the models, plots were generated to compare the actual solubility with the predicted solubility (Fig. 3a). From the figure, it was observable that the predicted solubility were more closely aligned with the diagonal line y = x compared to the LR model or the base models, indicating higher linearity and a higher R^2 value. This linearity was particularly evident in the region where predicted solubility falls between 0.2 and 0.4. Moreover, the stacking model outperformed the other models in binary classification task as well (Fig. 3b).

3.6. Characterization of important features

In order to determine the most influential features used in the base models MLP and GCN, the saliency maps of each model was generated. Saliency map assigns the significance scores of each features used by computing the gradient of the score function in respect to each inputs [76]. As shown in Fig. 4a, the most influential features of the MLP models were the hydrophilicity index and mean lipophilicity potential, followed by absolute charge per residue, number of hydrogen bonds, molecular weight, and the ratio of hydrophobic area. These results are consistent with existing reports on the strong influence of hydrophilic, hydrophobic, charged nature of proteins, as well as molecular weight, on solubility [5,77]. Based on the magnitudes of the computed saliency, sequence-level and structure-level features appeared to demonstrate a similar level of importance in the MLP model.

In the case of the GCN model, the blosum62 mapping displayed higher level of influence compared to AAPHY7 and coordinates features (Fig. 4b). Notably, the blosum62 mapping for the charged residues, including R, D, E, and K showed more significant saliencies compared to other residues. Derived from the observation of substitutions in conserved blocks of protein sequences, blosum62 provides the similarity of a residue compared to another residue [56]. Therefore, GCN seemed to put strong emphasis on the similarity of each residue to the charged ones. Among the AAPHY7 features used, the residual isoelectric point displayed the highest impact, which could correlate with the degree of charged residues. From the high saliency of the residual charge-related features, it was conceivable that the GCN model puts the most emphasis on the charge states of the protein residues.

3.7. Comparison with other methods using external S. cerevisiae test set

An external validation was conducted to assess the robustness of the stacking model in comparison to previous models using *S. cerevisiae* dataset. Table 3 displayed the prediction performance of the base models, the stacking model, and reference models in terms of the square of Pearson's correlation (r^2) and coefficient of determination (R^2). As shown in Table 3, the stacking model with SVR regressor outperformed all the existing competitors in both metrics ($r^2 = 0.494$ compared to 0.423 of SOLart, and $R^2 = 0.468$ compared to 0.424 of GATSol). The stacking model with LR regressor displayed slightly lower performance compared to using SVR regressor ($r^2 = 0.485$ and $R^2 = 0.452$). Notably, the stacking models, and the top competitors SOLart, GraphSol, HybridGCN, and GATSol were all structure-based models, while the underperforming models were sequence-based. The better performances of structure-based models indicated the crucial role of structural information in predicting protein solubility. Interestingly, our stacking model, derived from AF-predicted structures, yielded higher performance than SOLart, which was constructed upon experimentally determined structures, by approximately 17 %. This improvement could be attributed to the larger number of protein structures in the training set (2237 compared to 406). Furthermore, it was reported that the accuracy of AlphaFold predictions is comparable to experimental structures [78]. Compared to GraphSol, our stacking model demonstrated a

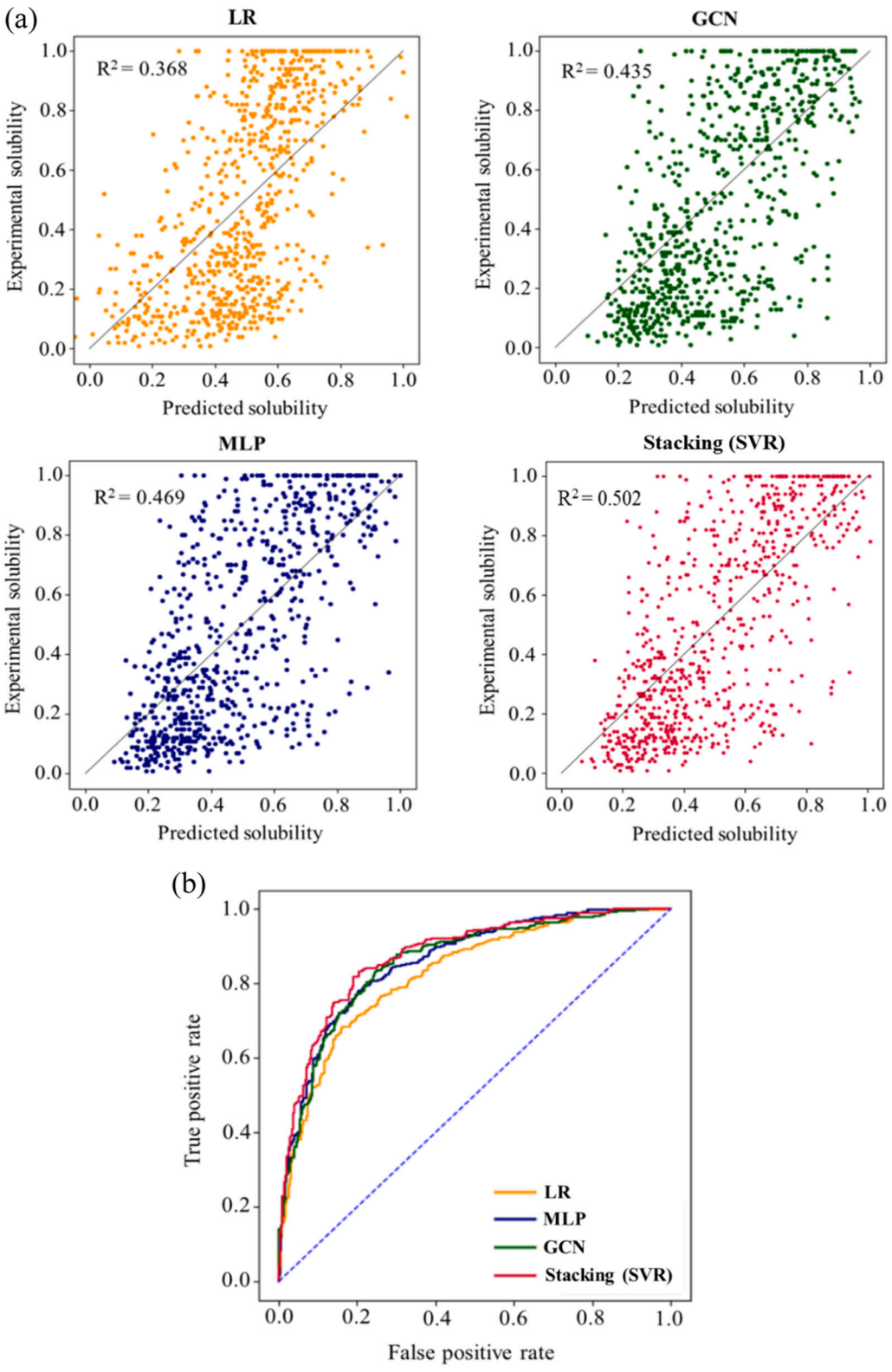
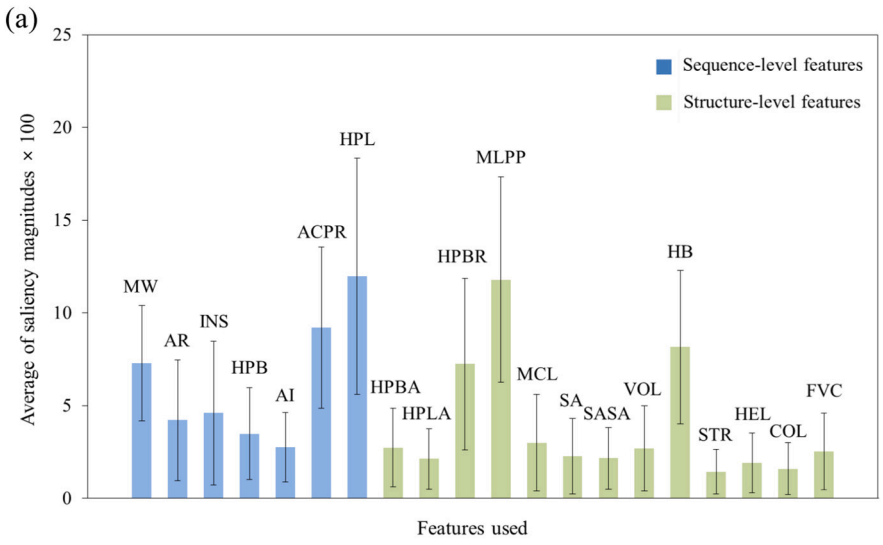
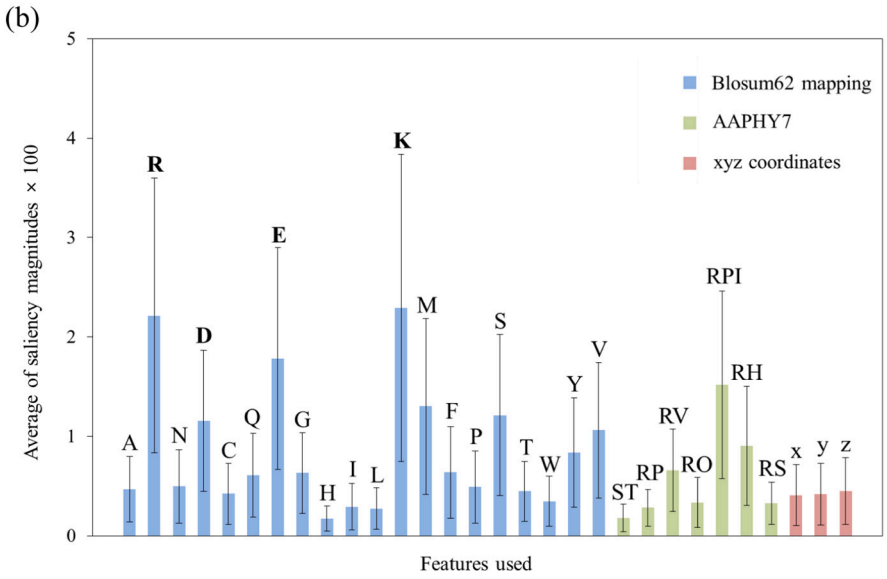


Fig. 3. The scatter plots of predicted and experimental solubility (a) and ROC curve (b) for each model on *E. coli* test dataset.

performance improvement of around 30 %. This difference was more pronounced in the *S. cerevisiae* test set compared to the *E. coli* test set (5.46 %). The larger performance gap was believed to stem from the differences in the node features used. As mentioned earlier, GraphSol incorporated evolutionary information represented by HMM and PSSM. Excluding these evolutionary features was thought to help avoid overfitting the stacking model to the proteins similar to *E. coli* family. This notion was further supported by the fact that the performance of the base model GCN was lower than that of GraphSol in the *E. coli* dataset but higher in the *S. cerevisiae* dataset. Moreover, when compared with PLM-based prediction tools, the stacking model (SVR) outperformed HybridGCN and GATSol by 23.80 % and 10.37 %, respectively. Derived



MW: molecular weight, AR: aromaticity, INS: instability index, HPB: hydrophobic index, AI: aliphatic index, ACPR: absolute charge per residue, HPL: hydrophilic index, HPBA: hydrophobic area, HPLA: hydrophilic area, HPBR: ratio of hydrophobic area, MLPP: mean lipophilic potential, MCL: mean Coulombic potential, SA: surface area, SASA: solvent accessible surface area, VOL: volume, HB: number of hydrogen bonds, STR: strand propensity, HEL: helix propensity, COL: coil propensity, FVC: favorable interaction



ST: steric parameter, RP: residue polarizability, RV: residue volume, RO: residue hydrophobicity, RPI: residue isoelectric point, RH: residue helix propensity, RS: residue strand propensity

Fig. 4. The relative importance of features determined from saliency maps of the trained MLP (a) and GCN (b). The blosum62 mappings for charged residues are bolded in (b).

from the authors of Graphsol, HybridGCN employs ESM-1b language model as well as all the node features used in Graphsol. On the other hand, GATSol utilizes ESM-1v (an updated version of ESM-1b) and blosum 62 mapping as its features. While direct comparison was not feasible, as the stacking model does not employ language model embedding, a possible explanation for its better performance could be the high feature dimensions of the two transformers used (1280 per residue).

To illustrate the improvement of performance achieved by stacking the MLP and GCN models, scatter plots were generated for the predictions made in the *S. cerevisiae* set (Fig. 5). As seen from the scatter plots of the base models, the major deviations in GCN model results were at the region where the predicted solubility is between 0.0 and 0.4. On the other hand, the MLP model exhibited the most deviation in the predicted solubility ranging from 0.6 to 1.0. In the case of the stacking model, it was evident that the predictions in the regions where each base model failed were successfully alleviated. This improvement suggested the stacking model's ability to leverage the strength of its base models. While the base models MLP and GCN were both constructed from predicted structures of the same proteins, they utilized distinct-level information of the proteins. Specifically, the MLP model leveraged information from the protein's structure and sequence, and GCN utilized a recreated graph structure based on connectivity and node features. MLP approach might overlook the intricate interactions and spatial

Table 3
Prediction performance of the stacking models and existing prediction models on *S. cerevisiae* dataset.

Solubility models	R^2
ESM-2 (GCN)	0.280 / 0.314*
MLP	0.303 / 0.433*
GCN	0.421 / 0.459*
Stacking (LR)	0.452 / 0.485*
Stacking (SVR)	0.468 / 0.494*
GATSol	0.424[a]
HybridGCN (ensemble)	0.390[b]
HybridGCN	0.378[b]
SOLart	0.423*[c]
Graphsol (ensemble)	0.372[c]
Graphsol	0.358[c]
ccSOL	0.303*[d]
Protein-Sol	0.281*[d]
DeepSol	0.090*[d]
proGAN	0.084[d]

* Square of Pearson's correlation.
[a] Generated by Li et al. [39].
[b] Generated by Chen et al. [38].
[c] Generated by Chen et al. [33].
[d] Generated by Hou et al. [34].

relationships between amino acids. Conversely, GCN could underestimate long-range interactions and lacked a comprehensive viewpoint on protein structure [79]. Hence, the stacking model seemed to be able to achieve a more holistic understanding of proteins compared to its base models.

3.8. Transferability of the stacking model to seed storage proteins

While the stacking model demonstrated a promising performance on microbial proteins, its applicability to food-related proteins remained unknown. Validating the transferability of the model would ideally involve comparing predicted values with experimental data. However, in the field of food and agricultural science, the solubility of individual protein is not commonly reported, making it nearly impossible to construct a solubility dataset. To address this challenge, the Osborne classification of plant storage proteins was employed. This classification system categorizes proteins based on their solubility in various solvents, dividing them into albumin (water soluble), globulin (salt soluble), prolamin (alcohol soluble), and glutelin (alkaline soluble) fractions [80]. These four fractions are relevant for predicting aqueous solubility, as each displays distinct solubility behaviors in water. Specifically, albumins are highly soluble in water, while globulins are partially soluble in water [81]. In contrast, glutelin and prolamin, which require more harsh conditions to solubilize, are known to be insoluble in water [82]. The process of obtaining Osborne fractions first involves solubilizing a sample in water, centrifuging it, and collecting the supernatant as albumin. The remaining residue is then subjected to saline water, and the supernatant is collected as globulin. This process continues with prolamin using an alcohol solution and with glutelin using an alkaline solution [83]. The major proteins in each supernatant are then classified into the respective Osborne classes. Therefore, although numerical quantification does not follow, the process of obtaining the albumin fraction is analogous to the relative solubility definition used in *E.coli* dataset. In this sense, by comparing the predicted solubilities of proteins falling under each classification, it would be possible to indirectly demonstrate the applicability of the model to food-related protein solubility. Based on the documented solubility of each Osborne fractions, it was anticipated that albumin would exhibit the highest solubility, followed by globulin with lower solubility, and finally, prolamin and glutelin would be insoluble.

Thus, the stacking model was applied to a collected dataset of 200 seed storage proteins, and their predicted solubility was presented in Fig. 6. From the figure, it was evident that the stacking model successfully ranked the dataset according to the expected trend. The mean predicted solubility for each classification was 0.621 ± 0.147 for albumin, 0.407 ± 0.116 for globulin, 0.310 ± 0.116 for prolamin, and 0.301 ± 0.079 for glutelin. Notably, the model predicted some level of solubility for glutelin and prolamin, which differed from the expected behavior. This discrepancy likely stemmed from the inherent uncertainty in the model (R^2 of 0.502 in the *E. coli* dataset). Nevertheless, it was clear that the model was capable of ranking the relative solubility of plant proteins, extending beyond the microbial proteins it was trained on. Moreover, the existing models were tested on the seed storage protein dataset (Table 4). Evaluations of Graphsol and HybridGCN were

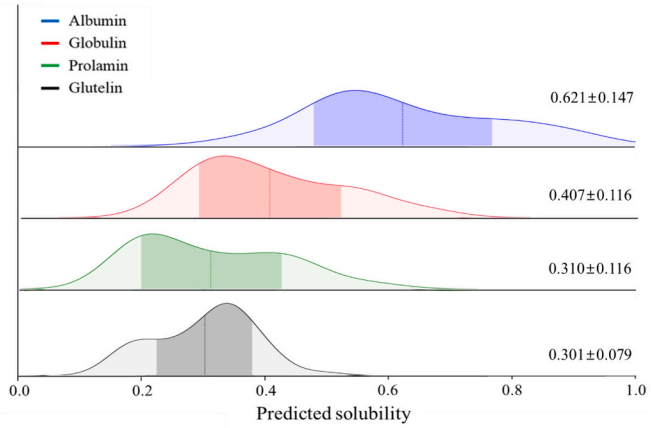


Fig. 6. Stacking model (SVR) predictions for the solubility of seed storage proteins by their Osborne classifications.

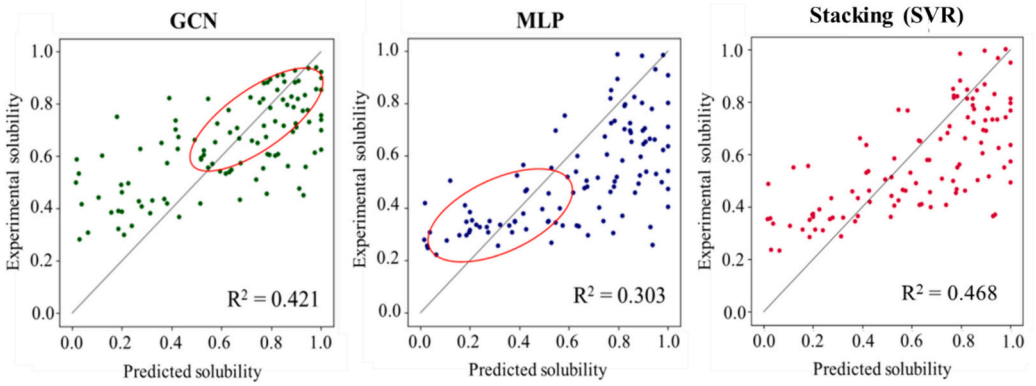


Fig. 5. The scatter plots of predicted and experimental solubility on *S. cerevisiae* validation dataset.

Table 4
Predictions of existing models on seed storage protein dataset.

	Albumin (soluble)	Globulin (partially soluble)	Prolamin (insoluble)	Glutelin (insoluble)
DeepSol	0.492 ±0.137	0.464 ±0.078	0.315 ±0.167	0.456 ±0.096
Protein-Sol	0.599 ±0.120	0.373 ±0.101	0.528 ±0.070	0.325 ±0.091
SoLart	0.659 ±0.067	0.500 ±0.065	0.565 ±0.032	0.521 ±0.109
GATSol	0.578 ±0.119	0.356 ±0.066	0.496 ±0.105	0.336 ±0.117
Stacking (SVR)	0.621 ±0.147	0.407 ±0.116	0.310 ±0.116	0.301 ±0.079

excluded due to the heavy computational demands of generating PSSM and HMM. As shown in the table, while albumin was generally predicted to be the most soluble across all models, discrepancies were observed for other Osborne fractions. For example, Protein-Sol, SoLart, and GATSol, on average, predicted prolamin fractions to be more soluble in water than globulin, which did not align with the reported behaviors of globulin (partially soluble) and prolamin (insoluble). Although the stacking model did not perfectly reproduce the insolubility of prolamin and glutelin, it displayed the largest solubility difference between albumin (water soluble) and prolamin and glutelin. Furthermore, the stacking model was the only one that predicted the expected solubility ranks of seed storage proteins—namely, albumin > globulin > prolamin ≈ glutelin—among the tested models.

3.8.1. *Case study on seed storage proteins: identification of important residues*

To determine the significance of specific residues in predicting solubility using the GCN model, the saliency values generated for 2S albumin from soybean and 22 kDa zein (prolamin) from maize were analyzed. These proteins were selected as they displayed the highest and lowest predicted solubility: 0.898 for albumin and 0.142 for prolamin in the stacking model, with GCN model predictions of 0.825 and 0.169, respectively. The saliency magnitudes for each residue were presented in Fig. S1, showing that certain residues displayed a higher influence on the GCN model. Specifically, residues 8–31 (region 1), 65–83 (region 2), and 139–143 (region 3) of 2S albumin were found to have the greatest impact. For 22 kDa prolamin, residues 31–135 (region 1), 160–168 (region 2), and 171–175 (region 3) were identified as the most influential (Fig. 7a). Notably, these regions exhibited fewer charged residues per length compared to the overall protein. The selected regions of albumin had charged residue densities of 0.0869, 0.278, and 0.000, respectively, compared to 0.367 for the entire protein. For prolamin, the regions had charged group densities of 0.019, 0.000, and 0.000, respectively, compared to 0.023 for the whole protein. This charge disparity was further highlighted by the electrostatic potential data from the APBS web server (Fig. 7b).

3.8.2. *Case study on seed storage proteins: CGMD simulation*

To further investigate the dissolution behaviors of storage proteins predicted by the stacking model, 4 μs CGMD simulations were conducted with 2S albumin and 22 kDa prolamin (Fig. 8a). The important residues from the saliency map of the GCN models were colored in red. As shown in the figure, both 2S albumin and 22 kDa prolamin exhibited

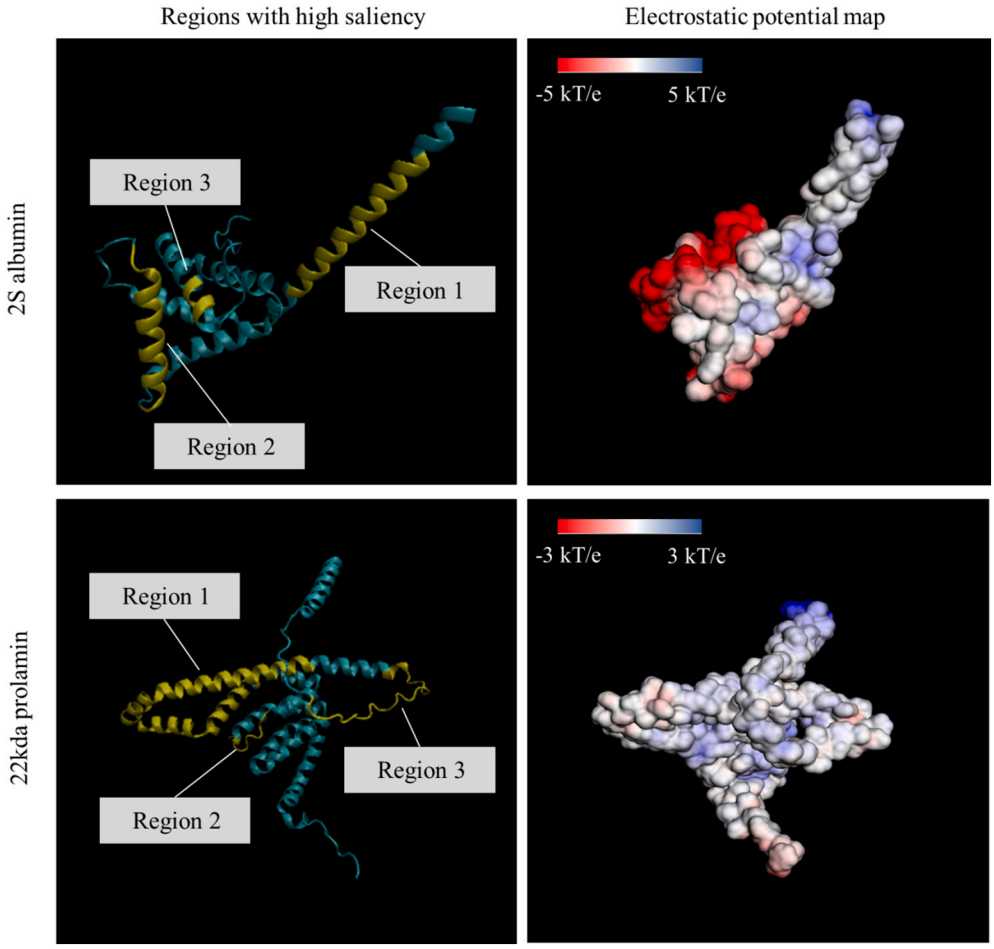


Fig. 7. Selected regions with high saliency and the electrostatic potential maps of soybean 2S albumin and maize 22 kDa prolamin.

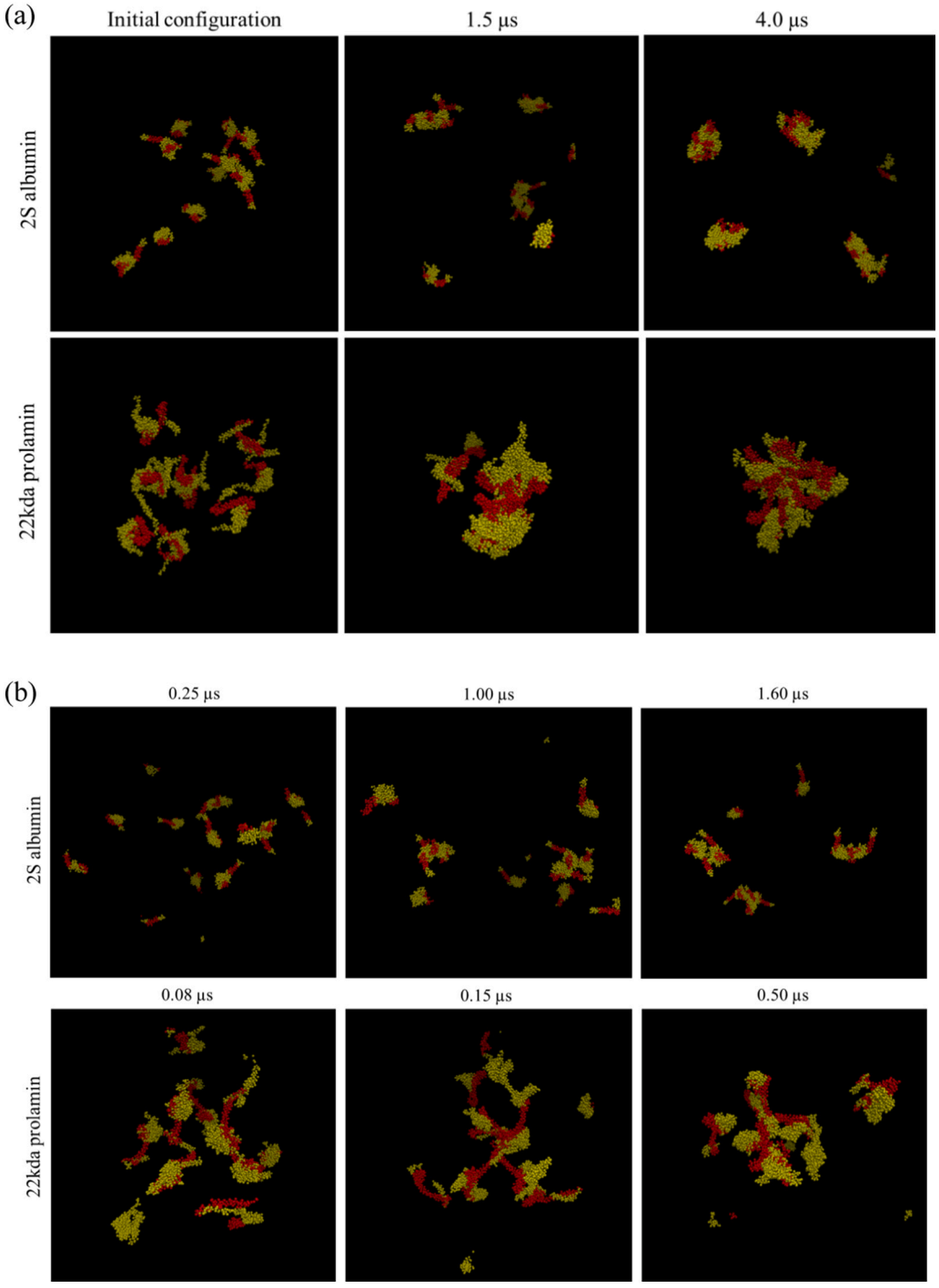


Fig. 8. Snapshots of CGMD simulations for soybean 2S albumin and maize 22 kDa prolamin at different intervals (a) and initial aggregation stage (b).

aggregating behaviors over the simulation time. However, while 2S albumin tended to aggregate into small clusters of two or three monomers, larger aggregates involving multiple 22 kDa prolamins were observed. The more aggregation-prone behavior of prolamin compared to albumin was also reflected in their mean square displacement (MSD) over time (Fig. S2). MSD measures the average squared distance that particles move over a certain time interval. Compared to 2S albumin, 22 kDa prolamin displayed faster and more severe aggregation, reducing overall mobility and MSD. The diffusion coefficient, calculated from the slope of the MSD plot, was 0.027910 ± 0.0302 nm^2/s for 2S albumin and 0.005749 ± 0.0195 nm^2/s for 22 kDa prolamin. Given that the formation of large aggregates leads to increased amount of protein precipitation, which significantly decreases solubility [84], the simulation results seemed to support the solubility predictions made by the stacking model.

While Fig. 8a successfully demonstrated the predicted behaviors of the two proteins, the role of the important regions determined in Section 3.8.1 in protein aggregation remained unclear. Therefore, the trajectories of the simulations during the early stage of aggregation were investigated. From the snapshots of the trajectories, it was observed that the initial aggregation began with the contact between proteins and the influential regions (Fig. 6b), whether in the large aggregate of prolamin or the smaller ones of albumin. This was more prominent in the prolamin simulation at 0.15 μs, where the binding of multiple selected regions

to a prolamin monomer was observed. The binding of these regions further promoted overall aggregation, as both selected and unselected regions came into contact and aggregated (0.50 μs in Fig. 8b and 1.5 μs in Fig. 8a). The snapshots of the early stage of aggregation suggested that the selected regions were more aggregation-prone compared to the unselected regions, potentially inducing further aggregation. This point was further supported by the fewer charged groups in the identified regions (Fig. 4), which would result in less favorable interaction of these regions with water compared to highly charged ones. The importance of charge-related features, the deficiency of charged groups within the selected residues, and the more aggregation-prone nature of the selected regions from the simulation therefore collectively suggested that the GCN model predicts solubility by identifying residues likely to aggregate, emphasizing the charged states of residues.

4. Conclusion

In this study, a novel solubility prediction model was developed using the protein structures generated from AlphaFold 2. Based on the predicted structures, two distinctive feature sets on their sequence+-structural information and on residual feature+contact map were subjected to MLP and GCN model, respectively. In the case of GCN model, the use of evolutionary features related to E.coli proteins was avoided during training, presumably providing more robust prediction applicable to food related proteins. The resulting out-of-fold predictions from the two base models were combined to create stacking models using various meta-regressors. The stacking model with SVR achieved the best performance among existing models in an external validation dataset. The higher performance of the stacking model compared to its base models suggested that stacking approach effectively leveraged the strengths of base models and the information from the distinct protein features. The stacking model (SVR) was further validated using a dataset consisting of seed storage proteins, generating expected solubility trend for seed storage proteins and exhibiting the potential transferability from microbial proteins to food and agricultural proteins. The limitation of this study is the lack of wet experimental validation data, due to the difficulty in obtaining isolated seed storage proteins. Additionally, although the extracted physicochemical features for training the model were fairly consistent across each dataset, the adequacy of using *E.coli* proteins to model seed storage proteins needs further validation. Moreover, the lack of multimeric structure data might impede the model performance, which could be improved with multimer prediction tools. Lastly, a further comparison with existing regressors on different dataset outside the one used in the research would make the performance evaluation more rationale. Yet, this research marks the first study to develop a solubility model for food proteins, and the framework adopted here could serve as a pioneering approach for future research in this area.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2024.134601.

CRediT authorship contribution statement

Hyukjin Kwon: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhenjiao Du:** Writing – review & editing, Methodology, Investigation. **Yonghui Li:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that there is no known conflict of interest.

Acknowledgements

This is contribution No. 24-083-J from the Kansas Agricultural Experimental Station. This research was supported by the USDA National Institute of Food and Agriculture Hatch project (7003330).

References

[1] T. Whitnall, N. Pitts, Global trends in meat consumption, Agric. Commod. 9 (1) (2019) 96–99.
[2] J. Westerhout, T. Krone, A. Snippe, L. Babe, S. McClain, G.S. Ladics, G.F. Houblen, K.C. Verhoeckx, Allergenicity prediction of novel and modified proteins: not a mission impossible!, Development of a random forest allergenicity prediction model, Regulatory Toxicology and Pharmacology 107 (2019) 104422.
[3] R. Deng, M. Mars, R.G. Van Der Sman, P.A. Smeets, A.E. Janssen, The importance of swelling for in vitro gastric digestion of whey protein gels, Food Chem. 330 (2020) 127182.
[4] P. Wood, M. Tavan, A review of the alternative protein industry, Curr. Opin. Food Sci. 47 (2022) 100869.
[5] S. Trevino, M. Scholtz, C. Pace, Measuring and increasing protein solubility, J. Pharm. Sci. 97 (10) (2008) 4155–4166.
[6] P. Evans, K. Wyatt, G.J. Wistow, O.A. Bateman, B.A. Wallace, C. Slingsby, The P23T cataract mutation causes loss of solubility of folded γD-crystallin, J. Mol. Biol. 343 (2) (2004) 435–444.
[7] M. Schnepf, Protein-water interactions, biochemistry of food, Proteins (1992) 1–33.
[8] L. Grossmann, D.J. McClements, Current insights into protein solubility: a review of its importance for alternative proteins, Food Hydrocoll. 137 (2023) 108416.
[9] Z. Yang, R. Thomson, Bio-basis function neural network for prediction of protease cleavage sites in proteins, IEEE Trans. Neural Netw. 16 (2005) 263–274.
[10] W. Weinert, H. Lopes, Neural networks for protein classification, Appl. Bioinforma. 3 (2004) 41–48.
[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, the, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[12] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR workshop and conference proceedings 13 (2010) 249–256.
[13] V. Radhika, V. Rao, Computational approaches for the classification of seed storage proteins, J. Food Sci. Technol. 52 (2015) 4246–4255.
[14] O. Arican, O. Gumus, PredDRBP-MLP: prediction of DNA-binding proteins and RNA-binding proteins by multilayer perceptron, Comput. Biol. Med. 164 (2023) 107317.
[15] Y. Li, Z. Zhang, Z. Teng, X. Liu, PredAmyl-MLP: prediction of amyloid proteins using multilayer perceptron, Comput. Math. Methods Med. 2020 (2020) 8845133.
[16] T. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint. arXiv:1609.02907, 2016.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).
[18] M. Baranwal, A. Magner, J. Saldinger, E. Turali-Emre, P. Elvati, S. Kozarekar, J. VanEpps, N. Kotov, A. Violi, A. Hero, Struct2Graph: a graph attention network for structure-based predictions of protein–protein interactions, BMC Bioinform. 23 (2022) 370.
[19] Z. Cheng, C. Yan, F. Wu, J. Wang, Drug-target interaction prediction using multi-head self-attention and graph attention network, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (2021) 2208–2218.
[20] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130.
[21] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (2022) 2102–2110.
[22] F. Zhang, J. Li, Z. Wen, C. Fang, FusPB-ESM2: fusion model of ProtBERT and ESM-2 for cell-penetrating peptide prediction, Comput. Biol. Chem. 2024 (2024) 108098.
[23] C. Tran, S. Khadkikar, A. Porollo, Survey of protein sequence embedding models, Int. J. Mol. Sci. 24 (4) (2023) 3775.
[24] Z. Du, Y. Xu, C. Liu, Y. Li, pLM4Alg: protein language model-based predictors for allergenic proteins and peptides, J. Agric. Food Chem. 72 (2023) 752–760.
[25] M. Susanty, M. Mursalim, R. Hertadi, A. Purwarianti, T. Rajab, Classifying alkaliphilic proteins using embeddings from protein language model, Comput. Biol. Med. 173 (2024) 108385.
[26] R. Rawi, R. Mall, K. Kunji, C.-H. Shen, P.D. Kwong, G.-Y. Chuang, PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine, Bioinformatics 34 (7) (2018) 1092–1098.
[27] S. Khurana, R. Rawi, K. Kunji, G.-Y. Chuang, H. Bensmail, R. Mall, DeepSol: a deep learning framework for sequence-based protein solubility prediction, Bioinformatics 34 (15) (2018) 2605–2613.
[28] M. Hebditch, M.A. Carballo-Amador, S. Charonis, R. Curtis, J. Warwicker, Protein–sol: a web tool for predicting protein solubility from sequence, Bioinformatics 33 (19) (2017) 3098–3100.
[29] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann, D. Frishman, PROSO II–a new method for protein solubility prediction, FEBS J. 279 (12) (2012) 2192–2200.

[30] R. Rawi, R. Mall, K. Kunji, C.-H. Shen, P.D. Kwong, G.-Y. Chuang, PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine, Bioinformatics 34 (7) (2018) 1092–1098.
[31] X. Zhang, X. Hu, T. Zhang, Y. Ling, C. Liu, N. Xu, H. Wang, W. Sun, PLM_Sol: predicting protein solubility by benchmarking multiple protein language models with the updated *Escherichia coli* protein solubility dataset, bioRxiv (2024), https://doi.org/10.1101/2024.04.22.590218, 2024-04.
[32] H.M. Berman, M.J. Gabanyi, A. Kouranov, D. Micallef, J. Westbrook, Protein structure initiative–targettrack 2000-2017–all data files, Zenodo (2017), https://doi.org/10.5281/zenodo.1234567.
[33] J. Chen, S. Zheng, H. Zhao, Y. Yang, Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map, J. Chem. 13 (1) (2021) 1–10.
[34] Q. Hou, J.M. Kwasigroch, M. Rooman, F. Pucci, SOLart: a structure-based method to predict protein solubility and aggregation, Bioinformatics 36 (5) (2020) 1445–1452.
[35] F. Agostini, D. Cirillo, C.M. Livi, R.D. Ponti, G. G., Tartaglia, ccSOL omics: a webserver for large-scale prediction of endogenous and heterologous solubility in E. Coli, Bioinformatics 30 (20) (2014) 2975–2977.
[36] C.N. Magnan, A. Randall, P. Baldi, SOLpro: accurate sequence-based prediction of protein solubility, Bioinformatics 25 (17) (2009) 2200–2207.
[37] J. Wang, S. Chen, Q. Yuan, J. Chen, D. Li, L. Wang, Y. Yang, Predicting the effects of mutations on protein solubility using graph convolution network and protein language model representation, J. Comput. Chem. 45 (8) (2024) 436–445.
[38] L. Chen, R. Wu, F. Zhou, H. Zhang, J.K. Liu, HybridGCN for protein solubility prediction with adaptive weighting of multiple features, J. Chem. 15 (1) (2023) 118.
[39] B. Li, D. Ming, GATSol, an enhanced predictor of protein solubility through the synergy of 3D structure graph and large language modeling, BMC Bioinform. 25 (1) (2024) 204.
[40] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (1982) 105–132.
[41] P. Lijnzaad, H.J. Berendsen, P. Argos, Hydrophobic patches on the surfaces of protein structures, proteins: structure, Function, and Bioinformatics 25 (3) (1996) 389–397.
[42] V. Gligorijević, P.D. Renfrew, T. Kosciolek, J.K. Leman, D. Berenberg, T. Vatanen, et al., Structure-based protein function prediction using graph convolutional networks, Nat. Commun. 12 (1) (2021) 3168.
[43] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, et al., Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins, Proc. Natl. Acad. Sci. 106 (11) (2009) 4201–4206.
[44] A. Ros-Lucas, N. Martinez-Peinado, J. Bastida, J. Gascón, J. Alonso-Padilla, The use of alphafold for in silico exploration of drug targets in the parasite Trypanosoma cruzi, Front. Cell. Infect. Microbiol. 12 (2022) 944748.
[45] M.A. Pak, K.A. Markhieva, M.S. Novikova, D.S. Petrov, I.S. Vorobyev, E. S. Maksimova, et al., Using AlphaFold to predict the impact of single mutations on protein stability and function, PLoS One 18 (3) (2023) e0282689.
[46] M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods 9 (2) (2012) 173–175.
[47] M. Shapovalov, R.L. Dunbrack Jr., S. Vucetic, Multifaceted analysis of training and testing convolutional neural networks for protein secondary structure prediction, PLoS One 15 (5) (2020) e0232528.
[48] E. Uemura, T. Niwa, S. Minami, K. Takemoto, S. Fukuchi, K. Machida, et al., Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding, Sci. Rep. 8 (1) (2018) 678.
[49] H. Helmick, H. Turasan, M. Yildirim, A. Bhunia, A. Liceaga, J.L. Kokini, Cold denaturation of proteins: where bioinformatics meets thermodynamics to offer a mechanistic understanding: pea protein as a case study, J. Agric. Food Chem. 69 (22) (2021) 6339–6350.
[50] P.J. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, et al., Biopython: freely available python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (11) (2009) 1422.
[51] P. Lijnzaad, H.J. Berendsen, P. Argos, A method for detecting hydrophobic patches on protein surfaces, proteins: structure, Function, and Bioinformatics 26 (2) (1996) 192–203.
[52] T.J. Dolinsky, J.E. Nielsen, J.A. McCammon, N.A. Baker, PDB2PQR: An automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations, Nucleic Acids Res. 32 (suppl_2) (2004). W665-W7.
[53] J.E. Mills, P.M. Dean, Three-dimensional hydrogen-bond geometry and probability information from a crystal survey, J. Comput. Aided Mol. Des. 10 (1996) 607–622.
[54] J. Jia, A.R. Benson, Residual correlation in graph neural network regression, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 588–598.
[55] J. Meiler, M. Müller, A. Zeidler, F. Schmäschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Mol. Model. Ann. 7 (9) (2001) 360–369.
[56] W.R. Rudnicki, T. Mroczek, P. Cudek, Amino acid properties conserved in molecular evolution, PLoS One 9 (6) (2014) e98983.
[57] H. Yuan, J. Tang, X. Hu, S. Ji, Xgnn, Towards model-level explanations of graph neural networks, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 430–438.
[58] K. Guo, M.J. Buehler, Rapid prediction of protein natural frequencies using graph neural networks, Dig. Dis. 1 (3) (2022) 277–285.
[59] G. Pollastri, P. Baldi, Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners, Bioinformatics 18 (suppl_1) (2002) S62–S70.
[60] C. Garbin, X. Zhu, O. Marques, Dropout vs. batch normalization: an empirical study of their impact to deep learning, Multimed. Tools Appl. 79 (2020) 12777–12815.
[61] S. Zhang, H. Tong, J. Xu, R. Maciejewski, Graph convolutional networks: a comprehensive review, Comput. Soc. Netw. 6 (1) (2019) 1–23.
[62] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, arXiv preprint arXiv:170303130, 2017.
[63] S. Pittala, C. Bailey-Kellogg, Learning context-aware structural representations to predict antigen and antibody binding interfaces, Bioinformatics 36 (13) (2020) 3996–4003.
[64] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L.E. Felberg, D.H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, Improvements to the APBS biomolecular solvation software suite, Protein Sci. 27 (1) (2018) 112–128.
[65] P. Kroon, F. Grunewald, J. Barnoud, M. Tilburg, P. Souza, T. Wassenaar, S. Marrink, Martinize2 and vermouth: unified framework for topology generation, Elife 12 (2023) RP90627.
[66] O. Nnyigide, K. Hyun, Charge-induced low-temperature gelation of mixed proteins and the effect of pH on the gelation: a spectroscopic, rheological and coarse-grained molecular dynamics study, Colloids Surf. B Biointerfaces 230 (2023) 113527.
[67] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38.
[68] X. Han, W. Ning, X. Ma, X. Wang, K. Zhou, Improving protein solubility and activity by introducing small peptide tags designed with machine learning models, Metab. Eng. Commun. 11 (2020) e00138.
[69] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (3) (2021) 685–695.
[70] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, API Design for Machine Learning Software: Experiences from the Scikit-Learn Project, arXiv Preprint arXiv:13090238, 2013.
[71] D.P. Ismi, R. Pulungan, Deep learning for protein secondary structure prediction: pre and post-AlphaFold, Comput. Struct. Biotechnol. J. 20 (2022) 6271–6286.
[72] M. Rashid, S. Saha, G.P. Raghava, Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs, BMC Bioinform. 8 (2007) 1–9.
[73] M. Delorenzi, T. Speed, An HMM model for coiled-coil domains and a comparison with PSSM-based predictions, Bioinformatics 18 (4) (2002) 617–625.
[74] C.J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, P. Bucher, PROSITE: a documented database using patterns and profiles as motif descriptors, Brief. Bioinform. 3 (3) (2002) 265–274.
[75] O. Sagi, L. Rokach, Ensemble learning: a survey, Wiley interdisciplinary reviews, Data Min. Knowl. Disc. 8 (4) (2018) e1249.
[76] A. Ismail, H. Bravo, S. Feizi, Improving deep learning interpretability by saliency guided training, Adv. Neural Inf. Proces. Syst. 34 (2021) 26726–26739.
[77] C. Van Oss, Hydrophobicity and hydrophilicity of biosurfaces, Curr. Opin. Colloid Interface Sci. 2 (5) (1997) 503–512.
[78] M. Akdel, D.E. Pires, E.P. Pardo, J. Jänes, A.O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R.A. Laskowski, G. Pozzati, A. Shenoy, A structural biology community assessment of AlphaFold2 applications, Nat. Struct. Mol. Biol. 29 (11) (2022) 1056–1067.
[79] L. Rampášek, G. Wolf, Hierarchical graph neural nets can capture long-range interactions, 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2021, pp. 1–6.
[80] A. Kumar, R. Nayak, S.R. Purohit, P.S. Rao, Impact of UV-C irradiation on solubility of Osborne protein fractions in wheat flour, Food Hydrocoll. 110 (2021) 105845.
[81] J. Yang, R. Kornet, E. Ntone, M.G. Meijers, I.A. van den Hoek, L.M. Sagis, P. Venema, M.B. Meinders, C.C. Berton-Carabin, C.V. Nikiforidis, E.B. Hinderink, Plant protein aggregates induced by extraction and fractionation processes: impact on techno-functional properties, Food Hydrocoll. 110223 (2024).
[82] S.K. Sathe, V.D. Zaffran, S. Gupta, T. Li, Protein solubilization, J. Am. Oil Chem. Soc. 95 (8) (2018) 883–901.
[83] W.H. van der Walt, L. Schussler, W.H. van der Walt, Fractionation of proteins from low-tannin sorghum grain, J. Agric. Food Chem. 32 (1) (1984) 149–154.
[84] A.P. Golovanov, G.M. Hautbergue, S.A. Wilson, L.Y. Lian, A simple method for improving protein solubility and long-term stability, J. Am. Chem. Soc. 126 (2004) 8933–8939.