International Journal of
Molecular Sciences

MDPI

Article

DSResSol: A Sequence-Based Solubility Predictor Created with Dilated Squeeze Excitation Residual Networks

Mohammad Madani [1,2], Kaixiang Lin [2] and Anna Tarakanova [1,3,*]

[1] Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA; mohammad.madani@uconn.edu
[2] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269, USA; kaixiang.lin@uconn.edu
[3] Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269, USA
* Correspondence: anna.tarakanova@uconn.edu

Abstract: Protein solubility is an important thermodynamic parameter that is critical for the characterization of a protein's function, and a key determinant for the production yield of a protein in both the research setting and within industrial (e.g., pharmaceutical) applications. Experimental approaches to predict protein solubility are costly, time-consuming, and frequently offer only low success rates. To reduce cost and expedite the development of therapeutic and industrially relevant proteins, a highly accurate computational tool for predicting protein solubility from protein sequence is sought. While a number of in silico prediction tools exist, they suffer from relatively low prediction accuracy, bias toward the soluble proteins, and limited applicability for various classes of proteins. In this study, we developed a novel deep learning sequence-based solubility predictor, DSResSol, that takes advantage of the integration of squeeze excitation residual networks with dilated convolutional neural networks and outperforms all existing protein solubility prediction models. This model captures the frequently occurring amino acid k-mers and their local and global interactions and highlights the importance of identifying long-range interaction information between amino acid k-mers to achieve improved accuracy, using only protein sequence as input. DSResSol outperforms all available sequence-based solubility predictors by at least 5% in terms of accuracy when evaluated by two different independent test sets. Compared to existing predictors, DSResSol not only reduces prediction bias for insoluble proteins but also predicts soluble proteins within the test sets with an accuracy that is at least 13% higher than existing models. We derive the key amino acids, dipeptides, and tripeptides contributing to protein solubility, identifying glutamic acid and serine as critical amino acids for protein solubility prediction. Overall, DSResSol can be used for the fast, reliable, and inexpensive prediction of a protein's solubility to guide experimental design.

Keywords: deep learning; dilated convolutional neural network; DSResSol; protein solubility; squeeze excitation residual network

Citation: Madani, M.; Lin, K.; Tarakanova, A. DSResSol: A Sequence-Based Solubility Predictor Created with Dilated Squeeze Excitation Residual Networks. *Int. J. Mol. Sci.* **2021**, *22*, 13555. https://doi.org/10.3390/ijms222413555

Academic Editors: Jung Hun Oh and Mingon Kang

Received: 29 October 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

Solubility is a fundamental protein property, that can give useful insights into the protein's function or potential usability, for example, in foams, emulsions, and gels [1], and therapeutics applications such as drug delivery [2,3]. In practice, the analysis of protein solubility is the most important determinant of success (i.e., high yields) in therapeutic protein and protein-based drug production [4,5]. In the research setting, producing a soluble recombinant protein is essential for investigating the functional and structural properties of the molecule [6]. To improve yields experimentally, there exist certain refolding methods that utilize weak promoters and fusion proteins or optimize expression conditions, e.g., by using low temperatures [4,5]. However, these methods cannot ensure the production of soluble proteins from a relatively small trial batch size as they are limited by production cost and time. Given these concerns, reliable computational approaches for discovering

potentially soluble protein targets for experimental testing can help to avoid expensive experimental trial and error approaches.

A protein's structure and sequence features such as the isoelectric point, polarity, hydrophobicity, turn-forming amino acids, etc., are crucial intrinsic factors in protein solubility determination [7–9]. On this basis, several in silico approaches have been developed to predict protein solubility by using the protein sequence and its features. The majority of these tools use traditional machine learning models such as support vector machines (SVM) [10] and gradient boosting machines [11], employing pre-extracted features (i.e., features that are extracted from the protein sequences via other bioinformatics tools before being fed into machine learning models) as input for these models. For example, SOLpro employs two-stage SVM models for training 23 extracted features from the protein sequences [5]. PROSO II utilizes a two-layered structure, including the Parzen window [12] and first level logistic regression models as the first layer and a second-level logistic regression model as the second layer [13]. In more recent models such as PaRSnIP [14], gradient boosting machine models are used. This predictor utilizes the frequency of mono-, di-, and tripeptides from the protein sequence in addition to other biological features such as secondary structure, and the fraction of exposed residues in different solvent accessibility cutoffs as training features. SoluProt is the newest solubility predictor using a gradient boosting machine for training [15]. To evaluate the performance of this tool, a new independent test set was utilized. Notably, the frequency of important dimers extracted from the protein sequences was used as an input feature of the SoluProt model [15]. All the aforementioned models are two-stage models, with a first stage set up for extracting and selecting features and a second stage employed for classification. Deep learning (DL) models circumvent the need for a two-stage model. DeepSol is the first deep learning-based solubility predictor proposed by Khurana and coworkers [16] built as a single stage predictor through the use of parallel convolutional neural network layers [17] with different filter sizes to extract high dimensional structures encoding frequent amino acid k-mers and their non-linear local interactions from the protein sequence as distinguishable features for protein solubility classification [16].

In this study, a novel deep learning architecture and framework is proposed to create a sequence-based solubility predictor that outperforms all currently available state-of-the-art predictors: Dilated Squeeze excitation Residual network Solubility predictor (DSResSol). Specifically, we employ parallel Squeeze-and-Excitation residual network blocks that include dilated convolutional neural network layers (D-CNNs) [18], residual networks blocks (ResNet) [19], and Squeeze-and-Excitation (SE) neural network blocks [20] to capture not only extracted high dimensional amino acid k-mers from the input protein sequence but also both local and long-range interactions between amino acid k-mers, thereby increasing the information extracted by the model from the protein sequence. This framework can capture independent, non-linear interactions between amino acid residues without increasing the training parameters and run time and offers a significant improvement in performance compared to other models. Our work is inspired by recent studies using dilated convolutional neural networks and SE-ResNet for protein sequence and text classification [21,22].

The traditional method to capture long-range interactions in data and to solve vanishing gradient problems is to use Bidirectional Long Short-Term (BLSTM) memory networks [23,24]. However, BLSTM implementation significantly increases the number of parameters in the model. Thus, we use D-CNNs instead of BLSTM because D-CNNs perform simple CNN operations, but over every nth element in the protein sequence, resulting in captured long-range interactions between amino acid k-mers. ResNet is a recent advance in neural networks that utilizes skip connections to jump over network layers to avoid the vanishing gradient problem and gradually learns the feature vectors with many fewer parameters [19]. The Squeeze-and-Excitation (SE) neural network explicitly blocks model interdependencies between channels, thereby directing higher importance to specific channels within the feature maps over others. Thus, by designing a novel architecture that

combines these three neural networks in a specific manner together with parallel CNNs layers, a highly accurate solubility predictor is built. In the first model instance, DSResSol (1), we use only protein sequence as input and protein solubility as output. The second model instance, DSResSol (2), includes pre-extracted biological features added to the model as a hidden layer to improve the model's performance. DSResSol is evaluated on two different independent test sets and is the first protein solubility predictor outperforming all existing tools on two distinct test sets, confirming its useability for different classes of proteins expressed in various host cells. By contrast, all existing tools have been built to work on a single test set for a specific class of proteins. DSResSol shows an improvement in accuracy of up to 13% compared with existing tools. Additionally, DSResSol reduces bias within the insoluble protein class compared to existing tools. We further investigate the most important single amino acids, dipeptides, and tripeptides contributing to protein solubility, which are directly extracted from feature maps in layers of the DSResSol architecture, with close alignment to experimental findings. We find that DSResSol is a reliable predictive tool that can be used for possible soluble and insoluble protein targets, achieving high accuracy, with improved relevance for guiding experimental studies.

2. Results

2.1. Model Performance

A 10-fold cross-validation is performed for the training process. In each cross-validation step, the training set is divided into ten parts where nine parts are used for training and one part is used for validation. The performance of the DSResSol model is reported by using the ten models. To evaluate the stability in performance results, we use four different metrics: accuracy, precision, recall, and F1-score. Figure 1 represents the box plots of these metrics for all ten models obtained through 10-fold cross-validation for both DSResSol (1) and DSResSol (2) on both independent test sets. Notably, the variance in box plots corresponding to each metric for both DSResSol (1) and DSResSol (2) is very small, highlighting the outstanding stability in the performance of the DSResSol predictor. For example, for DSResSol (2), among 10 models, the best model has an accuracy of 79.2% and the weakest model has an accuracy of 78.4%, with a variance of 0.8% (Tables S1 and S2).

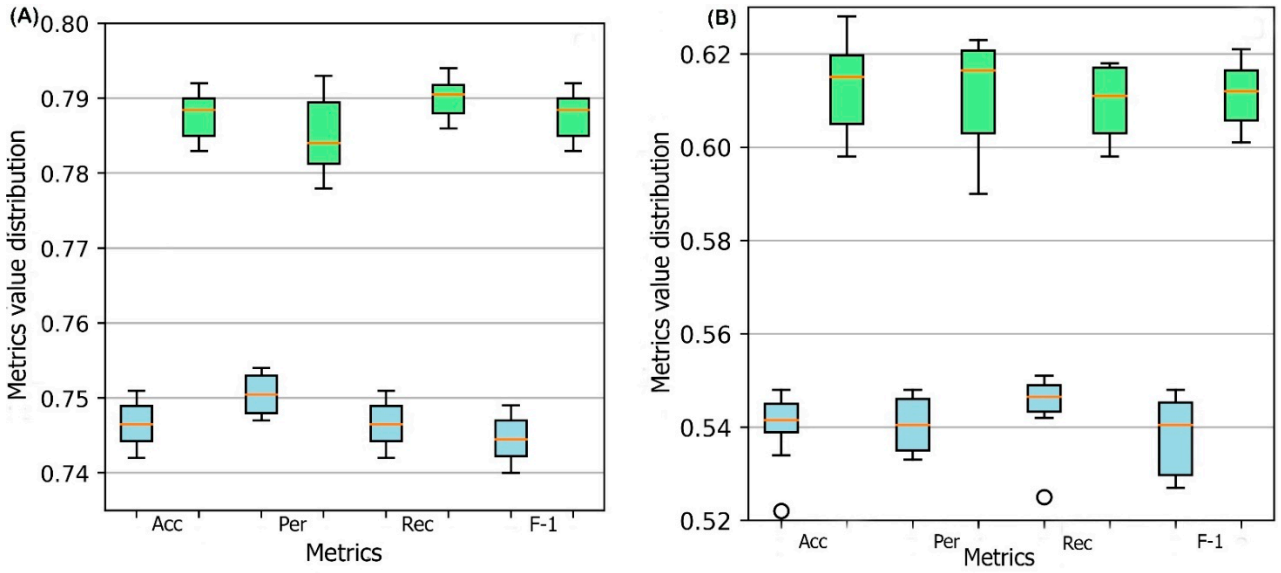

Figure 1. Box plot for 10 models obtained from 10-fold cross-validation for both DSResSol (1) and DSResSol (2) considering four metrics: ACC (accuracy), Per (precision), Rec (recall), and F-1 (f-1 score) for (**A**) Chang et al. test set [25], (**B**) NESG test set. Note: blue and green box plots represent the score distribution for DSResSol (1) and DSResSol (2), respectively.

It is worth noting that the reason for significant differences between the performance of the model on the first and second test sets is not due to overfitting or overtraining. The

difference between accuracy in training and validation for both DSResSol 1 and DSResSol (2) is less than 1.5% (Tables S1 and S2), suggesting that neither model has an overfitting problem. Therefore, we conclude that the difference between the performance in different test sets originates from the nature of the test sets. Specifically, the protein sequences within the second test set are expressed in *E. coli*. while, for the training process, we used a training set that includes mixed proteins (expressed in *E. coli* or other host cells) to achieve a more comprehensive model (Table 1).

Table 1. Performance of DSResSol in comparison with known existing models on first independent test set [25]. Note: Best performing method is in bold.

Model	ACC	MMC	Selectivity (Soluble)	Selectivity (Insoluble)	Sensitivity (Soluble)	Sensitivity (Insoluble)	Gain (Soluble)	Gain (Insoluble)
DSResSol (2)	**0.796**	**0.589**	0.817	**0.782**	**0.769**	0.823	**1.634**	**1.564**
DSResSol (1)	0.751	0.508	0.786	0.722	0.691	0.813	1.572	1.445
SoluProt	0.682	0.382	0.701	0.670	0.722	0.643	1.403	1.342
DeepSol S2	0.762	0.546	**0.821**	0.721	0.681	**0.843**	1.642	1.442
DeepSol S3	0.760	0.543	0.801	0.725	0.707	0.822	1.602	1.451
PaRSnIP	0.720	0.472	0.761	0.723	0.698	0.743	1.522	1.446
DeepSol S1	0.720	0.471	0.752	0.706	0.691	0.749	1.504	1.412
PROSSO II	0.638	0.345	0.671	0.682	0.693	0.662	1.342	1.365
SCM	0.600	0.214	0.650	0.572	0.422	0.773	1.301	1.145
PROSO	0.581	0.161	0.582	0.575	0.541	0.622	1.164	1.151
CCSOL	0.543	0.083	0.543	0.539	0.514	0.572	1.087	1.081
RPSP	0.520	0.032	0.522	0.517	0.447	0.588	1.044	1.035

To compare the performance of DSResSol with the best existing prediction tools, two different testing sets are employed, the first one proposed by Chang et al. [25] and the second one, NESG dataset, proposed by Price et al. [26] and refined by Hon et al. [15]. Tables 1 and 2 display the performance of eight solubility predictors on both test sets. DSResSol (2) outperforms all available sequence-based predictor tools when the performance is assessed by accuracy, MCC, sensitivity for soluble proteins, selectivity for insoluble proteins, and gain for insoluble proteins.

Table 2. Performance of DSResSol in comparison with known existing models on NESG test set [15]. Note: Best performing method is in bold.

Method	ACC	MCC	Selectivity (Soluble)	Selectivity (Insoluble)	Sensitivity (Soluble)	Sensitivity (Insoluble)	Gain (Soluble)	Gain (Insoluble)
DSResSol (2)	**0.629**	**0.273**	0.606	**0.663**	**0.73**	0.53	1.212	**1.326**
DSResSol (1)	0.557	0.169	0.558	0.555	0.54	0.58	1.117	1.11
SoluProt	0.578	0.189	0.575	0.581	0.6	0.56	1.15	1.162
PROSSO II	0.565	0.143	0.578	0.555	0.48	0.66	1.157	1.11
SWI	0.558	0.142	0.545	0.58	0.7	0.42	1.09	1.16
CamSol	0.535	0.115	0.548	0.527	0.39	0.68	1.097	1.054
ESPRESSO	0.493	0.093	0.493	0.492	0.55	0.44	0.987	0.984
rWH	0.519	0.133	0.532	0.513	0.31	0.73	1.065	1.026
DeepSol S2	0.546	0.132	**0.894**	0.56	0.22	**0.88**	**1.788**	1.12
SOLpro	0.5	0.089	0.5	0.5	0.48	0.52	1	1

For the first test set, we find that only the sensitivity value for the insoluble class, and the selectivity value for the soluble class were slightly inferior to the close competitor, DeepSol S2 [16]. The accuracy and MCC of DSResSol (2) are higher compared to DeepSol by at least 4% and 7%, respectively (Table 1). In addition, the sensitivity of DSResSol (2) for both soluble and insoluble proteins is close in value, suggesting that the DSResSol (2) model can predict both soluble and insoluble protein sequences with high accuracy and minimal bias. This consistency is missing in DeepSol S2 and DeepSol S3, the most accurate predictors to date. The sensitivity of DSResSol (2) for insoluble protein is about 83% which is comparable to the current best predictor (DeepSol S2 = 85%). On the other hand, DSResSol (2) can identify soluble proteins with higher predictive accuracy (77%) than all existing models, including DeepSol S2 (68%), DeepSol S3 (70%), and PaRSnIP (70%). There is a 16.2% and 12.7% difference in sensitivity between soluble and insoluble classes, for DeepSol S2 and DeepSol S3, respectively, where the insoluble class is predicted with higher accuracy. By contrast, the difference for the DSResSol (2) model is less than 6%, representing the outstanding capability of DSResSol (2) for identifying both soluble and insoluble classes and thereby reducing prediction bias.

The DSResSol (1) model performs comparably to DeepSol S2 and DeepSol S3 models and outperforms other models such as PaRSnIP and DeepSol S1. The performance of DSResSol (1) is competitive with DeepSol S2 and DeepSol S3; notably, however, in contrast to DeepSol S2 and DeepSol S3, DSResSol (1) obtains a similar performance without additional biological features as complimentary information in the training process. The accuracy of DSResSol (1) is only 1% lower than DeepSol S2, and higher by at least 4% in accuracy and 7% in MCC than DeepSol S1, suggesting that our proposed model architecture can capture more meaningful information from the protein sequence than DeepSol S1 by using only protein sequence as input for the training process.

Figure 2A,B show the Receiver Operating Characteristic (ROC) curve and the recall vs. precision curve for seven different solubility predictors, using the first independent test set [25]. The area under curve (AUC) and area under precision recall curve (AUPR) for DSResSol (2) are 0.871 and 0.872, respectively, which is at least 2% higher than other models confirming that the DSResSol (2) model outperforms other state-of-the-art available predictors. Figure 2C shows the accuracy of models in different probability threshold cutoffs. The highest accuracy for DSResSol (2) and DSResSol (1) is achieved at the probability thresholds equal to 0.5. This achieved result is due to using a balanced training and testing set.

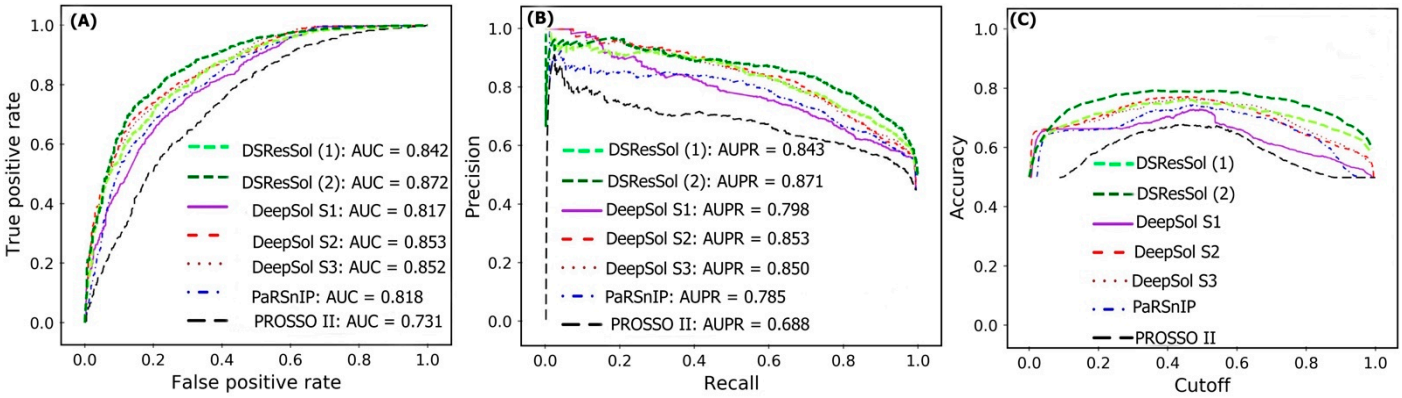

Figure 2. Comparison of the performance of DSResSol models with DeepSol [16] and PaRSnIP [14] models. (**A**) Receiver operating curve (ROC), (**B**) recall-precision curve, (**C**) accuracy-threshold cutoffs curve. The cutoff threshold discriminates between the soluble and the insoluble proteins. The curve for PROSSO II is obtained with permission from Bioinformatics [16].

For the second test set, (the newest test set to date), the performances of both models are evaluated and compared with eight different available sequence-based tools. Table 2 represents the performance of the DSResSol models on the NESG test set. Evaluation metrics include both threshold-dependent metrics such as accuracy and MCC as well as

threshold-independent metrics such as area under ROC curve value. The results represent that the accuracy and MCC of DSResSol (2) is at least 5% and 50% more than SoluProt tools, respectively. Furthermore, DSResSol (2) achieves the highest AUC value (0.68) among other tested solubility predictors on the second independent test set. The DSResSol (1) model, using only protein sequences for training, achieves comparable results with other tools such as SoluProt [15] and PROSSO II [13]. The accuracy value for DSResSol (1) only is 2% lower than the best competitor (SoluProt). This demonstrates an outstanding performance of the DSResSol (1) model which does not take advantage of using additional biological features for training, confirming that DSResSol (1) indeed captures the most meaningful features from the protein sequence to distinguish soluble proteins from insoluble ones. Furthermore, the sensitivity of DSResSol (2) for soluble proteins is 73% which is significantly higher than SoluProt (at 13%). Figure 3A, B displays the threshold-independent evaluation metrics to show the performance of our models in comparison to three different existing models for the second independent test set (NESG) [26]. The area under ROC curve (AUC) and area under precision-accuracy curve (AUPR) are 0.683 and 0.678, respectively, which is at least 8% higher than the best existing competitor, SoluProt. Figure 3C shows the accuracy of the tested models at different solubility thresholds. The highest accuracy (62%) for DSResSol models is obtained at the solubility threshold equal to 0.5.

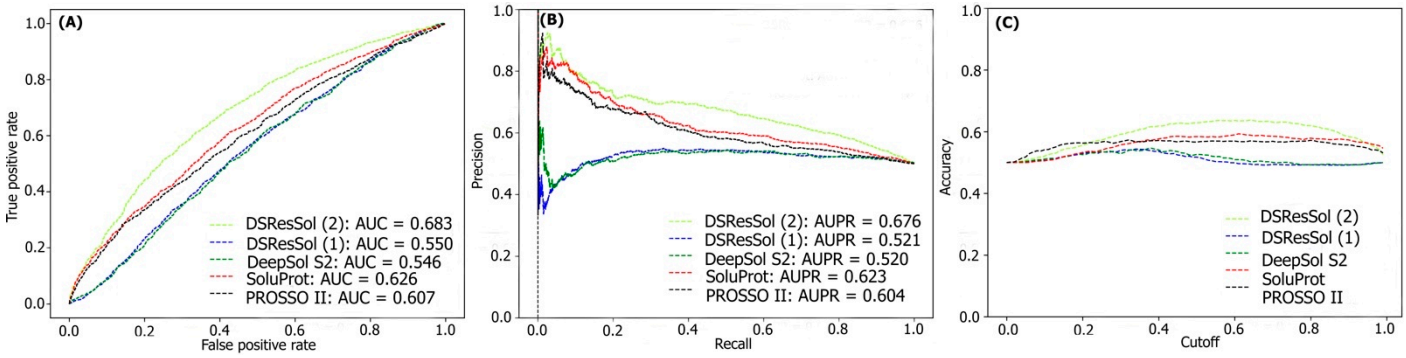

Figure 3. Comparison of the performance of DSResSol models with DeepSol [16], PROSSO II [13], and SoluProt [15] models. (**A**) Receiver operating curve (ROC), (**B**) recall-precision curve, (**C**) accuracy-threshold cutoffs curve. The cutoff threshold discriminates between the soluble and the insoluble proteins.

We also use the probability score distribution to evaluate DSResSol on both test sets. For the first test set, we consider the probability score distribution of DSResSol and close competitors (PaRSnIP [14] and DeepSol [16]) in violin plots (Figure 4) for both soluble and insoluble proteins. The distribution of scores for the four models shown in Figure 4 does not follow a normal distribution. For soluble and insoluble proteins, the score distribution plot shows that although DeepSol S2 like DSResSol (1) and DSResSol (2) gives more than a 99% level of confidence for the solubility prediction, the density of scores for DSResSol (2) in soluble proteins (values near score = 1) and for insoluble proteins (values near score = 0) is much higher than for DeepSol S2, confirming the better performance of DSResSol over DeepSol S2. In contrast to DSResSol and DeepSol S2, the score distribution for the PaRSnIP model does not reach a score = 1 for soluble and 0 for insoluble proteins, suggesting poor performance for PaRSnIP. To compare the score distribution of DSResSol (1) and DeepSol S2, we can see that near the probability score cutoff = 0.5, DSResSol (1) has greater density scores than DeepSol S2, suggesting lower accuracy in comparison to DeepSol S2. The mean score for each model is also computed. For the insoluble class, the mean score of DSResSol (2) (0.12) is significantly lower than DeepSol S2 (0.26) and PaRSnIP (0.37). For the soluble class, the mean score for DSResSol (2) (0.81) is much higher than DeepSol S2 (0.62) and PaRSnIP (0.61), suggesting the improved performance of DSResSol (2) over close competitors. In other words, a lower mean score value for the insoluble class and a higher mean score value for the soluble class represent higher confidence in the model. Finally, we note that the density of scores in the DSResSol (1) model beyond

a score of 0.5 is higher and lower for the insoluble and soluble classes, respectively, than DSResSol (2), suggesting lower accuracy (Figure 4). This result suggests that DSResSol (1) wrongly predicts more solubility values than DSResSol (2). This result can be understood considering that DSResSol (2) takes advantage of 85 additional biological features to establish a more accurate predictive model. For the second test set, a similar analysis is performed.

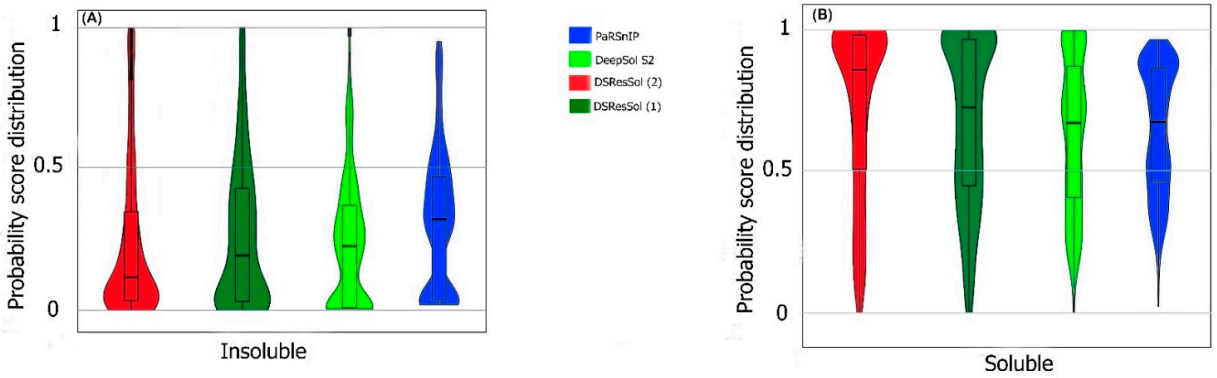

Figure 4. Violin plots represent the probability score distribution of DSResSol (1) and (2), DeepSol S2 [16], and PaRSnIP [14] for (**A**) insoluble and (**B**) soluble classes in the first test set [25].

For the second test set, the probability score distribution with two competitors, SoluProt [15] and DeepSol [16], is compared. Figure 5 shows this analysis for the second test set in two violin plots. The density score distribution near the probability value 0 for the DeepSol S2 model is higher than DSResSol (2) and DSResSol (1), indicating that DeepSol S2 works better than DSResSol (2) for insoluble protein prediction. Furthermore, the density of score distribution near the value = 0.5 for DSResSol is higher than DeepSol S2, confirming the slightly better performance of DeepSol S2 on insoluble protein prediction in comparison to DSResSol (2). Furthermore, the mean score distribution of the SoluProt model is 0.63 for insoluble proteins, representing its relatively poor performance on insoluble proteins. However, based on Figure 5B, DSResSol (2) outperforms DeepSol S2 and SoluProt for the soluble class. The mean of scores distribution on soluble proteins is 0.72 for DSResSol (2) while this value for DeepSol S2 and SoluProt is 0.42 and 0.66, respectively, confirming DSResSol (2) is the best candidate tool for soluble proteins prediction. Furthermore, the density of scores distribution for the soluble class in DSResSol (2) is higher than both DeepSol S2 and SoluProt (e.g., close to value = 1 in the violin plot), which further validates the outstanding performance of DSResSol (2) on the soluble class.

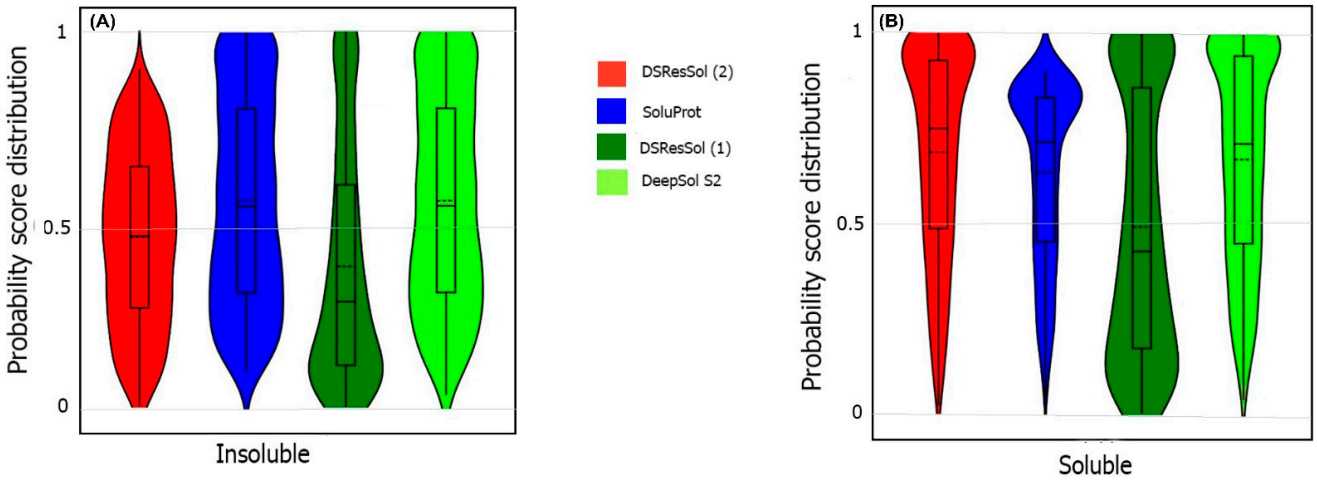

Figure 5. Violin plots representing the probability score distribution of DSResSol (1) and (2), DeepSol S2 [16], and SoluProt [15] for (**A**) insoluble and (**B**) soluble classes in NESG test set.

2.2. Effect of Sequence Length on Solubility Prediction

To illustrate the effect of protein sequence length on protein solubility prediction, we divide both test sets into five separate sets of different sequence length, in the range: {[0, 100], [100, 200], [200, 300], [300, 400], [400, ∞]}. To evaluate how sequence length affects the solubility score, the score distribution is shown for proteins predicted to be soluble and insoluble in five different sequence length ranges (Figure 6). True Positive (TP) and True Negative (TN) predictions correspond to the soluble and insoluble classes predicted correctly. False Positive (FP) and False Negative (FN) predictions correspond to the insoluble and soluble classes predicted incorrectly. Figure 6A shows that the median decreases monotonically as sequence length increases, suggesting that longer sequence length results in reduced solubility. In other words, the shorter protein sequences are more soluble as proposed by Kramer et al. [27]. The median scores for TP sets for five sequence length ranges are 0.93, 0.92, 0.90, 0.86, and 0.83, respectively. These values highlight the outstanding performance of DSResSol on the soluble class (a value of 1 corresponds to soluble protein). From Figure 6B, we observe that an increase in sequence length in the TN sets yields a decrease in the score distribution for insoluble proteins, suggesting that DSResSol can more easily predict insoluble proteins having longer than shorter sequences. The median of TN sets for five sequence length ranges is 0.21, 0.18, 0.16, 0.08, and 0.05, respectively, showing good performance of the DSResSol predictor (a value of 0 corresponds to insoluble protein). Furthermore, the difference between median TN and FN (Median (FN) -Median (TN)) for the insoluble class as well as the difference between the median of TP and FP (Median (TP)—Median (FP)) is calculated for the soluble class (Table 3). Proteins in the soluble class in the sequence length range (0, 100) and proteins in the insoluble class in the sequence length range (400 < L < ∞) have maximum values (0.29 for soluble and 0.34 for insoluble), confirming that the DSResSol model can predict proteins in these sequence length ranges with higher relative confidence than proteins with other sequence length ranges.

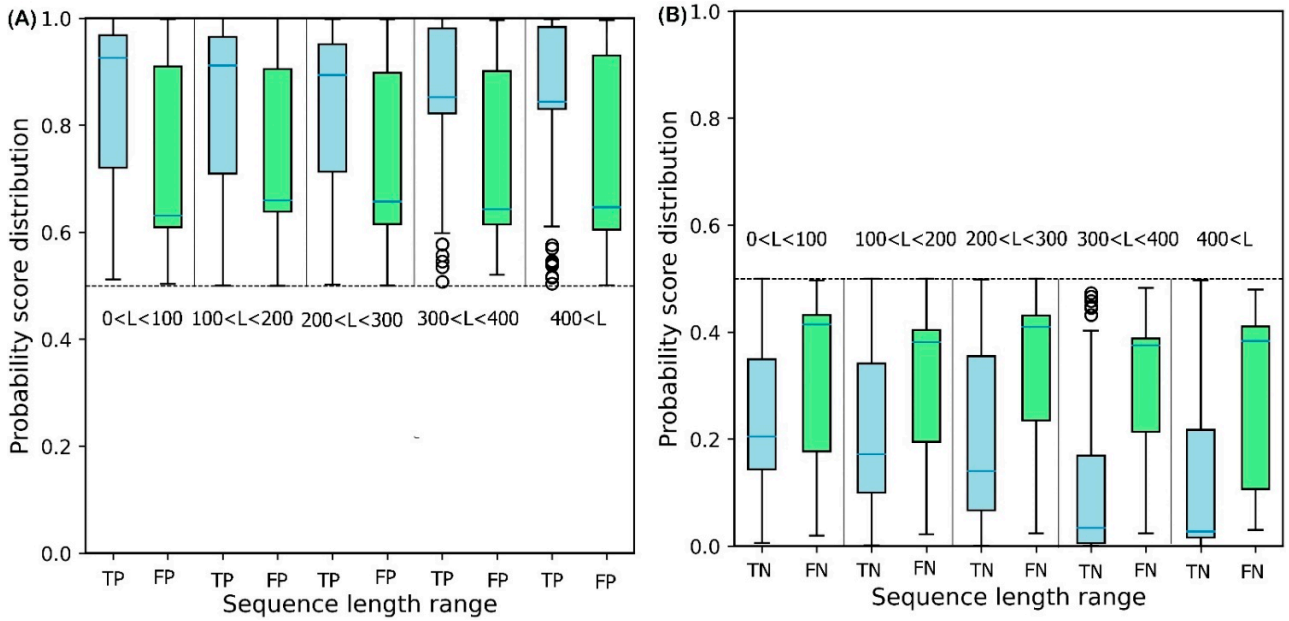

Figure 6. Probability score distributions for proteins predicted in both test sets to be (**A**) soluble and (**B**) insoluble for 5 different sequence length ranges: 0 < L < 100, 100 < L < 200, 200 < L < 300, 300 < L< 400, and 400 < L < ∞. TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative. Blue horizontal line in box plot of each set shows the median of the score distribution for that set.

Table 3. Difference between the median of True Positive and False Positive (TP, FP) soluble proteins as well as False Negative and True Negative (FN, TN) for insoluble proteins for different sequence length ranges. M = Median. Median of each category showed as horizontal blue line in box plots in Figure 6.

Sequence Length Range	M(TP)—M(FP)	M(FN)—M(TN)
(0, 100)	0.29	0.23
(100, 200)	0.24	0.22
(200, 300)	0.24	0.26
(300, 400)	0.22	0.31
(400, ∞)	0.22	0.34

2.3. Key Amino Acids, Dipeptides, and Tripeptides for Protein Solubility

To investigate the most important amino acids and di- and tripeptides contributing to protein solubility, these are directly extracted from the DSResSol model. As discussed, nine initial CNNs in DSResSol are responsible for capturing amino acid k-mers from k = 1 to 9. The feature maps obtained from each initial CNN, having dimensions 1200 × 32, are associated with amino acid k-mers for the corresponding protein sequence. To extract key amino acids associated with protein solubility, the feature vector, called activation vector, is needed for each protein sequence. These feature vectors for each protein sequence in our training set are extracted as follows. First, we pass the feature maps, which we receive from the CNN layer having a filter size of 1, through a reshape layer to assign features maps with dimension 32 × 1200. Then, these feature vectors are fed to a Global Average Pooling layer to obtain the feature vectors of length 1200 for each protein sequence, which represents the activation vector for that protein sequence. Each value in the activation vector, called activation value, is associated with a corresponding amino acid within the original protein sequence. Hence, higher activation values suggest a larger contribution to the classification results and protein solubility. The amino acids corresponding to the top 20 activation values for each protein sequence in the training dataset are counted. The total number of each amino acid corresponding to the top 20 activation values for all protein sequences in the training dataset represents the importance of that amino acid in protein solubility classification. The same process is applied for feature maps obtained from initial CNN layers with a filter size of 2 and 3 and the total number of pairs and triplets are counted, corresponding to the top 20 activation values across all protein sequences, to gain insight into the contribution of di- and tripeptides in protein solubility prediction. Figure 7 depicts the most important amino acids, dipeptides, and tripeptides contributing to protein solubility. We found that glutamic acid, serine, aspartic acid, asparagine, histidine, and glutamine are key amino acids contributing to protein solubility. Glutamic acid, aspartic acid, and histidine are amino acids with electrically charged side chains, while serine, asparagine, and glutamine have polar uncharged side chains. Interestingly, in one experimental study reported by Trevino et al., glutamine, glutamic acid, serine, and aspartic acid contribute most favorably to protein solubility [28]. Figure 7B,C shows that two and three consecutive glutamine amino acids (EE and EEE) are the most important dipeptides and tripeptides contributing to protein solubility. These results are consistent with experimental data proposed by Islam et al. [29]. Additionally, polar residues and residues that have negatively charged side chains such as glutamic acid and aspartic acid are, in general, more likely to be solvent-exposed than other residues [28] and can bind water better than other residues [30]. These observations are well-correlated with protein solubility and consistent with our analysis that identifies these as key amino acids for protein solubility prediction. In another investigation, Chan et al. demonstrated that positively charged amino acids are correlated with protein insolubility [31], consistent with our findings that histidine in the single, dipeptide, and tripeptide state strongly impacts protein solubility prediction. Moreover, Nguyen et al. found negatively charged

fusion tags as another way to improve protein solubility [32,33], consistent with our findings.

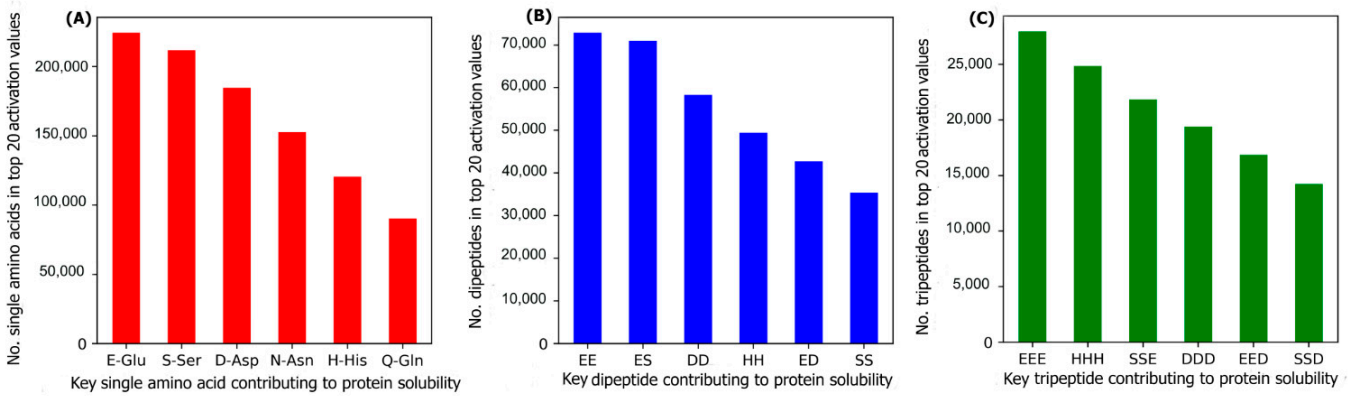

Figure 7. The number of (**A**) amino acids, (**B**) dipeptides, and (**C**) tripeptides corresponding to the top 20 activation values, obtained from the initial CNNs in DSResSol model, across all protein sequences within the training dataset.

2.4. Effect of Additional Biological Features on DSResSol Performance

To evaluate the effect of each additional biological feature group on DSResSol performance, we consider each feature group independently in the DSResSol (1) model (Tables 4 and 5). When only solvent accessibility-related features are added to DSResSol (1), the accuracy of the model on the first test set increases from 0.751 to 0.782, and the accuracy of the model on the second test improves from 0.557 to 0.618. Adding secondary structure-related features to DSResSol (1) improves the accuracy for the first test set from 0.751 to 0.763 and for the second test set from 0.557 to 0.582. We also consider the fraction of exposed residues and secondary structure content for soluble and insoluble proteins in the training data. We identify that the soluble protein class has 61.2% helix and beta strand content. In total, 68.7% of the residues are exposed residues with relative solvent accessibility cutoff higher than 65%. On the other hand, for the insoluble proteins in the training set, 81% of the secondary structure content is random coils. Further, 78% of residues are buried with relative solvent accessibility less than 35%, suggesting that the proteins having highly ordered structure and solvent-exposed residues with larger relative solvent accessibility cutoffs have a greater tendency to be soluble. By contrast, proteins with a higher degree of disordered secondary structure, such as random coil, and buried residues are predominantly insoluble. These results represent the influence on protein solubility propensity by solvent accessibility and secondary structure and are supported by experimental data. Kramer et al. have previously demonstrated the correlation between solvent accessibility and secondary structure content with protein solubility. They proposed that soluble proteins have larger negatively-charged surface area and are thus amenable to bind water [27]. Furthermore, Tan et al. identified the significant relationship between protein solubility and ordered secondary structure content such as helix and beta sheets [34]. They found large helix and beta sheet content within the most soluble proteins [34]. Thus, our results, which suggest that ordered secondary structures such as helix and beta sheets, as well as a larger fraction of solvent-exposed residues with higher relative solvent accessibility cutoffs, contribute to protein solubility, correlate well with experimental findings.

Table 4. Performance of the DSResSol model after adding each biological feature group to the DSResSol (1) model for the first test set. The accuracy of DSResSol (1) without biological features is 0.751.

Model	ACC DSResSol (1) after Adding the Additional Biological Features	ACC Improvement
DSResSol (1) + Solvent accessibility related features	0.787	3.7%
DSResSol (1) + Secondary structure related features	0.762	1.1%
DSResSol (1) + order/disorder related features	0.757	0.6%
DSResSol (1) + global sequence features	0.756	0.5%

Table 5. Performance of the DSResSol model after adding each biological feature group to the DSResSol (1) model for the second test set. The accuracy of DSResSol (1) without biological features is 0.557.

Model	ACC DSResSol (1) after Adding the Additional Biological Features	ACC Improvement
DSResSol (1) + Solvent accessibility related features	0.618	6.1%
DSResSol (1) + Secondary structure related features	0.582	2.5%
DSResSol (1) + order/disorder related features	0.564	0.7%
DSResSol (1) + global sequence features	0.561	0.4%

2.5. Effect of Sequence Identity Cutoff on DSResSol Performance

To develop input datasets, we removed the redundant protein sequences in the training set with sequence identity over 25%. Moreover, the protein sequences having more than 15% sequence similarity with both test sets have been eliminated in the training set. To analyze the impact of identity cutoff on DSResSol performance, we considered different cutoffs to train DSResSol (Table 6 and Table 7). The results indicate that by changing the sequence identity cutoffs, the performance of the DSResSol predictor improves to 75.1% for the first test set and to 55.7% for the second test set, suggesting that the optimal identity cutoff is 25% [16]. In fact, these results show that the existence of similar sequences within the training set leads to overfitting or overtraining of the model, which results in a decrease in model performance on the test sets.

Table 6. Performance comparison for DSResSol (1) on the first independent test set for different cutoff sequence identity. Note: Best performing method is in bold.

Model	ACC	MMC	Sensitivity (Soluble)	Sensitivity (Insoluble)
DSResSol (1) Cutoff 25%	**0.751**	**0.508**	**0.691**	**0.813**
DSResSol (1) Cutoff 15%	0.744	0.491	0.686	0.805
DSResSol (1) Cutoff 20%	0.743	0.488	0.701	0.795
DSResSol (1) Cutoff 30%	0.744	0.492	0.688	0.801

Table 7. Performance comparison for DSResSol (1) on the second independent test set for different cutoff sequence identity. Note: Best performing method in bold.

Model	ACC	MCC	Sensitivity (Soluble)	Sensitivity (Insoluble)
DSResSol (1) Cutoff 25%	**0.557**	**0.166**	**0.545**	**0.568**
DSResSol (1) Cutoff 15%	0.553	0.157	0.542	0.567
DSResSol (1) Cutoff 20%	0.555	0.164	0.547	0.563
DSResSol (1) Cutoff 30%	0.552	0.159	0.541	0.567

3. Discussion

In this study, we propose a novel sequence-based solubility predictor that uses a SE-ResNet neural network. In the first model, DSResSol (1), only raw protein sequences are used as input to distinguish soluble proteins from insoluble proteins. In the second model, DSResSol (2), to improve the performance of the first proposed model, 85 pre-extracted biological features are added as input. We observe that the performance of DSResSol (2) is superior to existing state-of-the-art tools when the model performance is evaluated on two distinct independent test sets. In particular, for the first test set, the accuracy of DSResSol (2) is at least 3% higher over the best performing model to date, DeepSol S2 [16]. For the second test set, the accuracy of DSResSol is more than 5% higher than SoluProt [15], the top-performing existing tool on this test set.

The main reason for the improved performance of the DSResSol predictor in comparison to other existing models originates from the SE-ResNet architecture. DeepSol [16], a close competitor, used only some parallel CNNs to extract feature maps from the input protein sequence. In fact, the DeepSol model could only capture contextual features, and amino acid k-mers of different lengths and their local non-linear interactions. By contrast, the DSResSol model not only extracts amino acid k-mers and their local interactions but also captures long-range interactions between amino acid k-mers with different lengths. This is because DSResSol greatly benefits from the specific SE-ResNet architecture, including dilated CNNs. SE-ResNet blocks in the DSResSol model are responsible for capturing frequently occurring amino acid k-mers where k = {1, 2, . . . , 9}, and their local and global interactions. Extracting these high order k-mers and their interactions gives valuable structural information about features such as protein folds [35], which are discriminative and important features for protein solubility prediction [14]. In the SE-ResNet block, the dilated CNN efficiently extracts long-range interactions among k-mers, while preventing over-fitting using dropout on the weights, which leads to good generalization performance. Furthermore, SE-ResNet captures more information from the input feature maps related to amino acid k-mers because it not only reduces gradient vanishing owing to feature reusability but also highlights the most important information from feature maps, which results in the capture of complex sequence–contact relationships while using fewer parameters than other methods [36]. In addition, by adding 85 biological features to the DSResSol model, the performance of the predictor is significantly improved. This suggests that these pre-extracted features are complementary to contextual features obtained from the SE-ResNet model.

We employed the DSResSol model to identify a mechanistic understanding of the relationship between sequence length and solubility propensity. For both insoluble and soluble classes, we observed monotonically decreasing score distributions when sequence length increases, suggesting that proteins with longer sequence length have a higher tendency to be insoluble. Furthermore, by analyzing the DSResSol model results, we found that glutamine, serine, and aspartic acid are key amino acids that favorably contribute to

protein solubility. Interestingly, this result correlates with experimental studies reported by Islam et al. and Trevino et al. [28,29]. We also found that secondary structure and relative solvent accessibility features are determinative in protein solubility prediction. We demonstrated that soluble proteins include a large number of exposed residues at relative solvent accessibility cutoffs of more than 65% and residues having ordered secondary structure content such as helix and beta sheets. On the other hand, for insoluble proteins, a large number of residues is buried and disordered. These results are supported by experimental findings proposed by Kramer et al. [27] and Tan et al. [34].

Overall, DSResSol presents an interpretable, predictive tool that effectively learns key structural and biological features of protein sequences for predicting protein solubility, with mechanistic implications that are highly correlated with experimental findings.

4. Materials and Methods

4.1. Data Preparation and Feature Engineering

To create the training set, we used the Target Track database [37]. Based on the methods proposed in previous studies [5,13], the solubility value for each protein sequence within the training set was inferred. A protein is labeled as insoluble if it cannot be expressed or purified experimentally. On the other hand, a protein is considered soluble if it is realized as soluble, purified, crystallized, etc., e.g., an experimental state requiring the protein to be soluble. To maintain the generality of our training set, in contrast to previous studies such as SoluProt [15], we did not impose a limitation on the expression system for selecting the proteins included in our training set. To reduce the noise and redundancy from our training set, the following tasks were performed: (1) removing the transmembrane proteins based on the annotations from the Target Track database; (2) removing the proteins considered as insoluble but associated with a PDB structure; (3) eliminating protein sequences from the training set with a sequence identity of more than 25% via CD-HIT [38] to avoid any bias because of homologous sequences within the training and testing sets. Finally, we used a fairly balanced training set that included approximately the same number of proteins within the soluble and insoluble classes. Thus, in total, our final training set contained 40,317 protein sequences including 19,718 soluble and 20,599 insoluble protein sequences.

For model evaluation, two different independent test sets were utilized. Both test sets include proteins that have been expressed in *E. coli*:

1. The first test dataset was proposed by Chang et al. [25]. This dataset includes 2001 protein sequences and their corresponding solubility values;
2. The second test dataset, first proposed by Hon et al. [15], has been constructed from a dataset generated by the North East Structural Consortium (NESG) [26] and includes 9644 proteins expressed in *E. coli*. The original dataset consists of integer values in the range from 0 to 5 for levels of expression and soluble fraction recovery. We maintained consistency between the procedure for constructing the training set and the test set. Finally, similar to the SoluProt test set [15], the solubility level of each protein within the NESG test set was transferred to a binary value. In the original NESG dataset, the solubility values for protein sequences are an integer value in the range from 0 to 4. To transfer these values to binary values, similar to the SoluProt method, we considered protein sequences having a solubility of value 0, 1, 2, as insoluble and those sequences have solubility values of 3 or 4 as soluble proteins.

To decrease the overlap between sequences within both test sets and the training set, all protein sequences in the training set which have a sequence identity of more than 15% with protein sequences in both test sets were eliminated. This significantly reduces redundant sequences in our training set. Table 8 catalogues the dataset construction/reduction process for both the training set and both test sets in detail.

Table 8. Construction steps for dataset preparation and number of sequences retained in each dataset construction step. Note: final amount of data within training and testing sets after pre-processing are in bold.

Construction Step	Training Set	Soluble	Insoluble	Test Set 1	Soluble	Insoluble	Test Set 2	Soluble	Insoluble
Input	129,593	-	-	2001	1000	1001	9703	-	-
Pre-processing and solubility assignment	109,648	-	-	2001	1000	1001	-	-	-
Redundancy removal	87,969	40,905	14,064	2001	1000	1001	9423	5718	3705
Removal of short sequences and sequences with unknown residues	82,902	50,004	32,898	2001	1000	1001	9420	5715	3705
Removal of transmembrane proteins	76,274	45,603	30,671	2001	1000	1001	8769	5421	3348
Removal of insoluble sequences with available PDB structure	72,756	42,530	30,226	2001	1000	1001	8754	5421	3333
Clustering to 25% identity	49,369	26,422	22,947	2001	1000	1001	3945	2078	1867
Overlap removal with test sets 15% identity	46,028	24,920	21,108	2001	1000	1001	3945	2078	1867
Class and length balancing	**40,317**	19,718	20,599	**2001**	1000	1001	**3729**	1864	1865

4.2. Model Architecture

Protein solubility prediction is a binary classification problem. Within the datasets, each protein sequence is assigned a solubility value equal to 0 or 1. Thus, the solubility propensity for each sequence evaluated by our model is assigned a score in the range from 0 to 1. The DSResSol (1) model includes 5 architectural units, including a single embedding layer, nine parallel initial CNNs with different filter sizes, nine parallel SE-ResNet blocks, three parallel CNNs, and fully connected layers, sequentially (Figure 8). The architecture of DSResSol (2) is equivalent to DSResSol (1). DSResSol (2) has an additional input layer to receive 85 additional biological features.

To use protein sequences as inputs of the DSResSol model, we employed two major preprocessing approaches. First, protein sequences were parameterized to the vectors $X = \{x_0, x_1, x_2, \ldots, x_L\}$ where $x_i \in \{0, 1, 2, \ldots, 20\}$. The numbers from 1 to 20 represent amino acid residues, and 0 indicates a gap [39]. Second, each sequence was padded to the fixed-length vector having length L = 1200 to generate same-sized vectors. Input features were converted to embedded vectors via the embedding layer, which is a lookup table for mapping and transforming the discrete input into continuous fixed-sized vectors. Thus, during the training process, a continuous feature was learned from each amino acid. The embedding layer transformed the input sequence vector $x \in R^{1200\times21}$ to a dense continuous feature representation via the embedding weight matrix $W_e \in R^{50\times21}$. The output of the embedding layer, i.e., the feature map, was $E = x \times W_e$. The embedding dimension was 50. Note that training the W_e happens along the whole network. After this layer, the embedded vectors were fed to nine CNN layers with different filter sizes k from 1 to 9: $k \in \{0, 1, 2, \ldots, 9\}$. (Figure 8A, Block (1)). Different filter sizes in the CNNs were employed to extract amino acid k-mers, i.e., "biological words," with different sizes between one (monopeptide) to nine (nonapeptide) from the input sequences. This component of the model was inspired by DeepSol [16]. In fact, the filter size in the CNN is equal to the convolutional window size along the characters of the sequence. Produced feature maps from each CNN layer were received by a Squeeze-and-Excitation residual network (SE-ResNet) block consisting of two main parts: a residual network part and a Squeeze-and-Excitation block, linked via a residual connection (Figure 8A, Block (2)).

Residual neural networks (ResNet) [19] are an outstanding new discovery for neural networks making neural networks deeper, by resolving the gradient vanishing problem with fewer parameters than traditional neural networks such as CNNs. In ResNet, the gradients can flow directly through the skip connections backwards from later layers to

initial filters [19,36]. ResNets are utilized in a wide range of applications including natural language processing [40] and image classification [19].

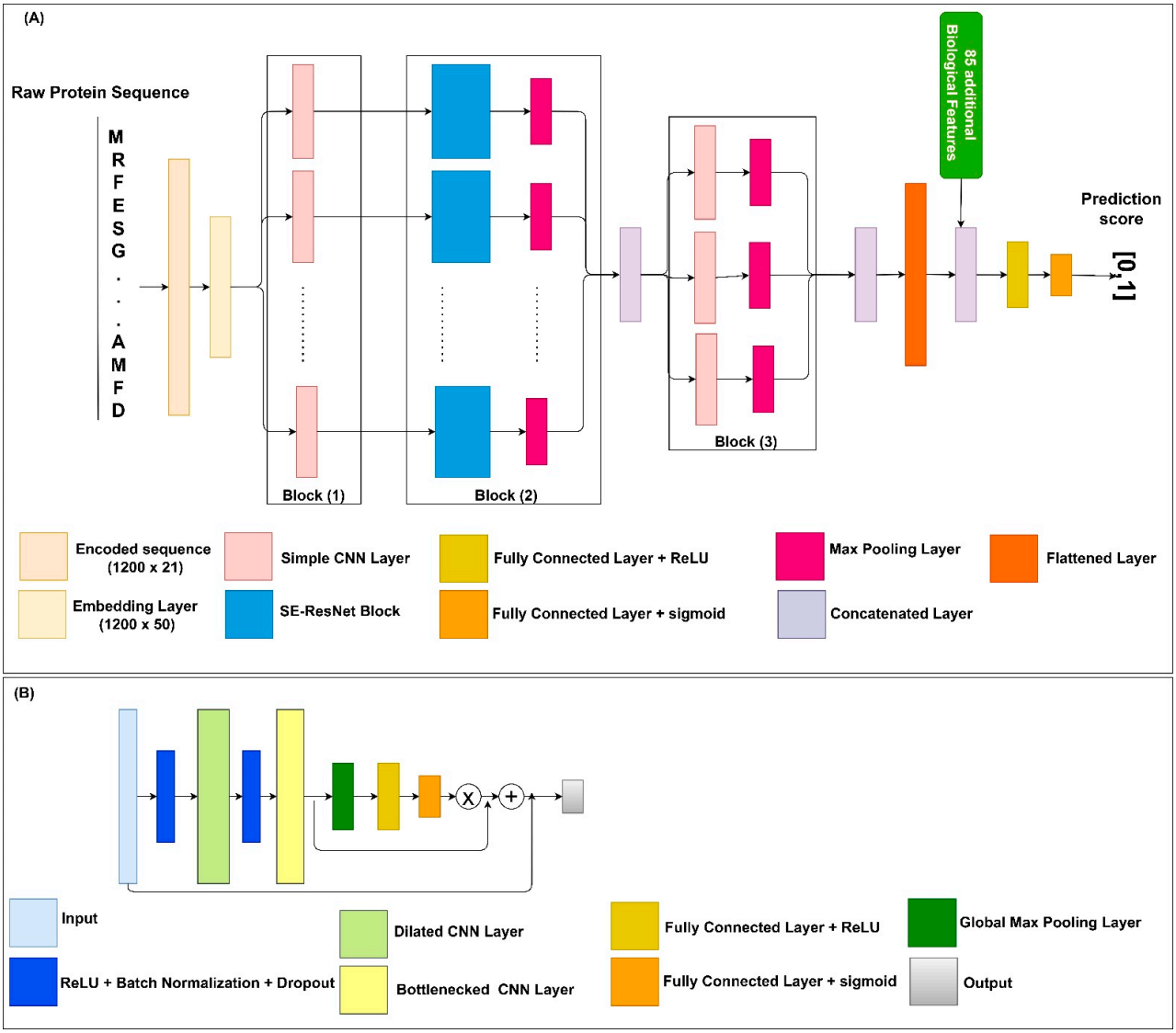

Figure 8. (**A**) Schematic of the DSResSol model architecture. DSResSol Block 1 has 9 initial CNNs with 32 filters within each CNN and filter sizes from k = 1 to 9; Block 2 includes 9 SE-ResNet blocks; Block 3 has another 3 CNNs with 32 filters with filter sizes {11, 13, 15}. A total of 85 additional biological features are used in DSResSol (2) only. (**B**) Schematic of the SE-ResNet block architecture. Each SE-ResNet block has a single dilated CNN with dilation rate of 2 and filter size of 3, and single bottlenecked CNN with filter size of 1. All Max Pooling layers have 1 stride and window size of 3.

The architecture of ResNet in the DSResSol model contains two CNNs, a dilated CNN [41] and a bottlenecked CNN [22], followed by batch normalization and a rectified linear unit as the activation function [42]. Figure 8B shows the ResNet block within the SE-ResNet module. Using a bottlenecked convolution layer speeds up the computation and increases the ResNet block's depth by using fewer parameters and a thinner ResNet block [43]. An n-dilated convolution captures local and global information about amino acid k-mers without significantly increasing the model parameters because it behaves like a simple convolution operation over every nth element in a sequence [22]. In fact, this type of convolution enables the model to capture long-range interactions across the sequence. The dilated convolution utilizes kernels that have holes. In this way, not only is the overall receptive field of convolution wider but also the complexity and number of parameters are

reduced [18,22]. The utilization of n-dilated convolution was inspired by a recent study for protein family classification from protein sequence [22].

In addition to ResNet, each SE-ResNet block consists of a single Squeeze-and-Excitation (SE) block. The SE block focuses on more important channels within the feature maps. In other words, the Squeeze-and-Excitation blocks can recalibrate the channels in the learned feature maps, which results in stimulating more important channels and hindering weak channels within the feature maps [44,45]. The feature input of the SE block was passed through the Global Max Pooling layer. This layer reduces each channel in the input feature map to a single value, which is the maximum value within each channel. Suppose the input tensor of the SE block has the shape of L × C, where C is the number of channels within the feature map and L is the feature dimension. After passing this tensor through the Global Max Pooling operator, the shape of the output will be reduced to C × 1. To map adaptive scaling weights for the output of the Global Max Pooling layer, we employed two fully connected (FC) layers. In the first FC layer, the number of units was set to C/8, and the activation function was the rectified linear unit (ReLU) [42]. In the second FC layer, the number of units was set to C to project back the first FC layer's output to the same dimensional space as the input, returning to C neurons. In summary, the C × 1 tensor input was passed through the first FC layer. Next, a weighted tensor of the same shape was obtained from the second FC layer as output. The sigmoid was utilized as the second FC activation function to scale the value to a range from 0 to 1. Using a simple broadcasted element-wise multiplication, the second FC layer's output was applied to the SE block's initial input [44]. To complete the SE-ResNet block, the SE block's output was concatenated with the ResNet block's input (Figure 8A). These processes are the same for each SE-ResNet block. The feature maps derived from each SE-ResNet block were fed to the Max Pooling layer to accumulate feature maps by taking maximum values over the sub-region along with the feature map. The output Max Pooling layers were merged to generate feature maps for the next stage of the model which included three convolution layers with a filter size of 11, 13, and 15, respectively, followed by three Max Pooling layers (Figure 8A, Block (3)). This stage is responsible for extracting more contextual features from the merged outputs of the SE-ResNet module. Finally, all three feature maps obtained from this stage were concatenated. The output of the previous stage was flattened to a 1D array, then fed into a single FC layer with hidden neurons of size 128 and ReLU as the activation function. The final FC layer with sigmoid as the activation function generated the probability score for solubility propensity.

4.3. Additional Features

Similar to PaRSnIP [14], secondary structure, solvent accessibility, structural order/disorder, and global sequence features were employed as additional biological features. Secondary structure (SS) and relative solvent accessibility (RSA) features were obtained through the SCRATCH webserver [46], order/disorder (O/D) information from the ESpritz webserver [47], and global sequence features from the python package modLAMP [48]. To calculate features from the SS and RSA, we employed the PaRSnIP procedure [14]. The fraction content of SS for 3- and 8-state SS was calculated. In addition, the fraction of exposed residues at different cutoffs (FER-RSA) from 0% to 95% with 5% intervals was obtained. The FER-RSA values were multiplied by the hydrophobicity of exposed residues to extract another RSA-based feature group. We introduced additional O/D-related features to our biological features set, augmenting the original features used in PaRSnIP. For O/D-related features, the number of disordered regions limited to fewer than 5 amino acids, sized between 5 and 10 amino acids, and sized larger than 10 amino acids, as well as the frequency of each amino acid in disordered regions was calculated. In total, 85 biological features were extracted from the protein sequence. A summary of all biological features used in our model is presented in Figure 9. These 85 pre-extracted biological features were concatenated to the feature maps derived from the flattening layer before fully connected

layers in DSResSol (2). To feed these additional features to DSResSol (2) as input features, we normalized them to values between 0 and 1.

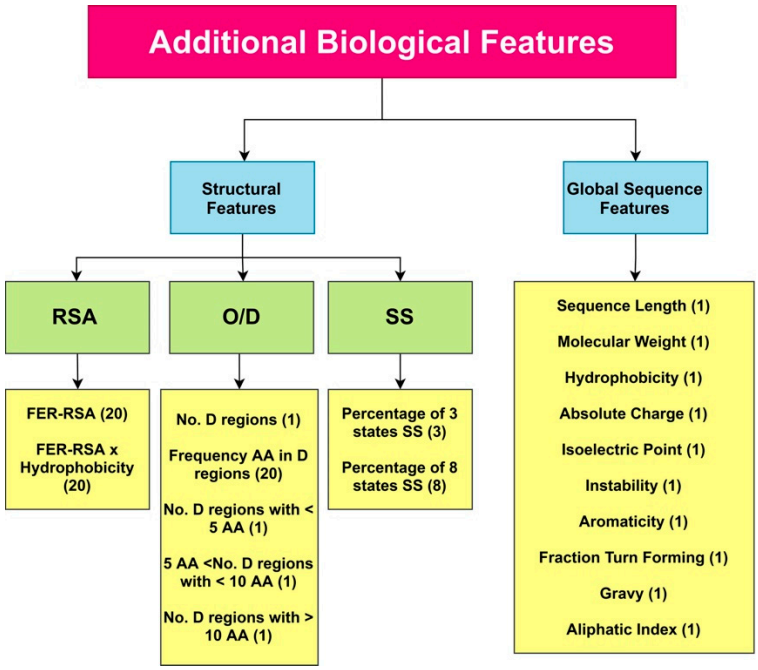

Figure 9. A total of 85 additional pre-extracted biological features are captured from the protein sequence and used for training the DSResSol (2) model. RSA: Relative Solvent Accessibility, SS: Secondary Structure, O/D: Order/ Disorder, FER-RSA: Fraction Exposed Residue at Relative Solvent Accessibility Cutoffs. The number of features for each type is shown in parentheses.

4.4. Training and Hyperparameter Tuning

The DSResSol model utilizes a binary cross-entropy [49] objective function to classify the protein sequence into two classes. Both models DSResSol (1) and (2) are fit for a different number of training epochs with the Adam optimizer [50]. Performance of the models depends on different hyperparameters such as: learning rate—the step size in which the optimizer receives the parameter space and updates the parameters; number of epochs—the number of iterations for training the model; batch size—the number of training examples that should be received before updating the parameters; size and number of filters in each convolution layer; embedding dimension; the number of units in FC layers, etc. We have performed hyperparameter tuning by employing a grid search on 10-fold cross-validation similar to DeepSol [16]. Table 9 represents the tuned hyperparameters for both DSResSol (1) and DSResSol (2) models. After identifying optimal hyperparameters, 10-fold cross-validation was performed to train the models. We also used the early stopping approach during the training to avoid overfitting [16].

4.5. Evaluation Metrics

To evaluate the performance of the DSResSol predictor in comparison with other predictors, we used accuracy, Matthew's correlation coefficient (MCC), sensitivity, selectivity, and gain metrics as described in PaRSnIP [14]. The sensitivity for a class is the ratio of correctly classified instances of that class to the total number of instances of that class, in a dataset. Selectivity for a class is the ratio of accurately classified instances of a class to the total number of instances predicted to be in a class. The gain value for a class represents the ratio of selectivity of a class to the proportion of those instances in the complete dataset.

Table 9. Tested hyperparameters in network layers and optimal values generated via the Grid Search Method [51].

Layers	Number of Tested Units or Filters	Optimal Value	Filter Size	Parameters	Tested Values	Optimal Value
Embedding layer	(50, 100, 150)	50	-	Epochs	50	50
Initial CNNs	(32, 64, 128, 256)	32	{1, 2, . . . , 9}	Learning rate	(0.005, 0.008, 0.01, 0.02)	0.008
Dilated CNN	(32, 64, 128, 256)	32	3	Batch size	(32, 64, 128, 256)	64
Bottlenecked CNN	(32, 64, 128, 256)	32	1	Decay rate	$(10^{-7}, 10^{-8}, 10^{-9})$	10^{-7}
Final CNNs	(32, 64, 128, 256)	32	{11, 13, 15}	Early stopping value	(3, 4, 5, 6)	5
FC	(64, 128, 256, 512)	128	-	-	-	-
MaxPooling	(2, 3, 5, 7)	3	-	-	-	-

5. Conclusions

In this study, we introduce the sequence-based protein solubility predictor, Dilated Squeeze excitation Residual network Solubility predictor (DSResSol), that outperforms all available bioinformatic tools for solubility prediction when the performance is assessed by different evaluation metrics such as accuracy and MCC. DSResSol improves accuracy for protein solubility prediction up to 5% for all proteins and up to 13% for soluble proteins in comparison with close competitors. In contrast to other existing models, DSResSol accurately identifies both soluble and insoluble proteins (assessed by close sensitivity values for soluble and insoluble classes), suggesting that the superior accuracy of DSResSol originates from good and balanced performance on both classes. The parallel SE-ResNet blocks with dilated CNNs comprehensively captures long-range non-linear interactions between amino acid k-mers in addition to local interactions, which facilitates the extraction of more meaningful features from protein sequences to improve model accuracy. The model's robustness originates not only from its novel deep learning architecture but also from its comprehensive training dataset. We have employed a novel training set that is cleaned from the noisy Target Track data via multiple steps for removing redundant protein sequences and evaluated the performance of the DSResSol tool with two independent test sets. Further analysis of DSResSol feature maps suggests that glutamine, serine, and aspartic acid are key amino acids that favorably contribute to protein solubility, a result correlated with experimental studies. Notably, the framework for generating comprehensive feature maps developed as part of this tool can be utilized independently as a feature extraction tool to prepare meaningful input features for other predictors, as well as for the prediction of other protein properties. In its present form, DSResSol can be effectively used as a predictive tool for evaluating solubility for designed proteins or for detecting changes in solubility and degradation upon mutation. Overall, the model can help avoid time-consuming trial-and-error approaches in experimental designs to facilitate the identification and prioritization of possible soluble targets.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms222413555/s1.

Author Contributions: M.M. and A.T. conceived and designed the research. M.M. developed the DSResSol models and performed all analysis. K.L. created the web server for DSResSol model. M.M. and A.T. wrote and edited the final manuscript with input from all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work utilized the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. XSEDE resources Stampede 2 and Ranch at the Texas Advanced Computing Center and Bridges at the Pittsburg Supercomputing Center through allocation TG-MCB180008 were used.

Data Availability Statement: The source code, datasets, and web server for this model are available at https://github.com/mahan-fcb/DSResSol (accessed on 30 July 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zayas, J.F. Solubility of Proteins. In *Functionality of Proteins in Food*; Zayas, J.F., Ed.; Springer: Berlin/Heidelberg, Germany, 1997; pp. 6–75, ISBN 978-3-642-59116-7.
2. Jain, A.; Jain, A.; Gulbake, A.; Shilpi, S.; Hurkat, P.; Jain, S.K. Peptide and protein delivery using new drug delivery systems. *Crit. Rev. Ther. Drug Carr. Syst.* **2013**, *30*, 293–329. [CrossRef] [PubMed]
3. Madani, M.; Tarakanova, A. Molecular Design of Soluble Zein Protein Sequences. *Biophys. J.* **2020**, *118*, 45a. [CrossRef]
4. Idicula-Thomas, S.; Balaji, P.V. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci. Publ. Protein Soc.* **2005**, *14*, 582–592. [CrossRef] [PubMed]
5. Magnan, C.N.; Randall, A.; Baldi, P. SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics* **2009**, *25*, 2200–2207. [CrossRef]
6. Chan, W.-C.; Liang, P.-H.; Shih, Y.-P.; Yang, U.-C.; Lin, W.; Hsu, C.-N. Learning to predict expression efficacy of vectors in recombinant protein production. *BMC Bioinform.* **2010**, *11*, S21. [CrossRef] [PubMed]
7. Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C.M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **2003**, *424*, 805–808. [CrossRef]
8. Bhandari, B.K.; Gardner, P.P.; Lim, C.S. Solubility-Weighted Index: Fast and accurate prediction of protein solubility. *Bioinformatics* **2020**, *36*, 4691–4698. [CrossRef]
9. Diaz, A.A.; Tomba, E.; Lennarson, R.; Richard, R.; Bagajewicz, M.J.; Harrison, R.G. Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnol. Bioeng.* **2010**, *105*, 374–383. [CrossRef] [PubMed]
10. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
11. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
12. Babich, G.A.; Camps, O.I. Weighted Parzen windows for pattern classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 567–570. [CrossRef]
13. Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II—A new method for protein solubility prediction. *FEBS J.* **2012**, *279*, 2192–2200. [CrossRef]
14. Rawi, R.; Mall, R.; Kunji, K.; Shen, C.-H.; Kwong, P.D.; Chuang, G.-Y. PaRSnIP: Sequence-based protein solubility prediction using gradient boosting machine. *Bioinforma. Oxf. Engl.* **2018**, *34*, 1092–1098. [CrossRef] [PubMed]
15. Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **2021**, *37*, 23–28. [CrossRef]
16. Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; Mall, R. DeepSol: A deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **2018**, *34*, 2605–2613. [CrossRef] [PubMed]
17. Lecun, Y.; Bengio, Y. Convolutional networks for images, speech, and time-series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995. Available online: https://nyuscholars.nyu.edu/en/publications/convolutional-networks-for-images-speech-and-time-series (accessed on 21 March 2021).
18. Yu, F.; Koltun, V. Multi-Scale Context Aggrgation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122. Available online: http://arxiv.org/abs/1511.07122 (accessed on 21 March 2021).
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. Available online: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html (accessed on 12 December 2020).
21. Yang, Z.; Hu, Z.; Salakhutdinov, R.; Berg-Kirkpatrick, T. Improved Variational Autoencoders for Text Modeling using Dilated Convolutions. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3881–3890. Available online: http://proceedings.mlr.press/v70/yang17d.html (accessed on 17 March 2021).
22. Bileschi, M.L.; Belanger, D.; Bryant, D.; Sanderson, T.; Carter, B.; Sculley, D.; DePristo, M.A.; Colwell, L.J. Using Deep Learning to Annotate the Protein Universe. *bioRxiv* **2019**, *20*, 626507. [CrossRef]
23. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
24. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin/Heidelberg, Germany, 7–12 August 2016; pp. 207–212.
25. Chang, C.C.H.; Song, J.; Tey, B.T.; Ramanan, R.N. Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: Protein solubility prediction. *Brief. Bioinform.* **2014**, *15*, 953–962. [CrossRef]
26. Price, W.N.; Handelman, S.K.; Everett, J.K.; Tong, S.N.; Bracic, A.; Luff, J.D.; Naumov, V.; Acton, T.; Manor, P.; Xiao, R.; et al. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microb. Inform. Exp.* **2011**, *1*, 6. [CrossRef]

27. Kramer, R.M.; Shende, V.R.; Motl, N.; Pace, C.N.; Scholtz, J.M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102*, 1907–1915. [CrossRef]
28. Trevino, S.R.; Scholtz, J.M.; Pace, C.N. Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J. Mol. Biol.* **2007**, *366*, 449–460. [CrossRef]
29. Islam, M.M.; Khan, M.A.; Kuroda, Y. Analysis of amino acid contributions to protein solubility using short peptide tags fused to a simplified BPTI variant. *Biochim. Biophys. Acta* **2012**, *1824*, 1144–1150. [CrossRef]
30. Kuntz, I.D. Hydration of macromolecules. III. Hydration of polypeptides. *J. Am. Chem. Soc.* **1971**, *93*, 514–516. [CrossRef]
31. Chan, P.; Curtis, R.A.; Warwicker, J. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep.* **2013**, *3*, 3333. [CrossRef] [PubMed]
32. Nguyen, T.K.M.; Ki, M.R.; Son, R.G.; Pack, S.P. The NT11, a novel fusion tag for enhancing protein expression in Escherichia coli. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 2205–2216. [CrossRef]
33. Zhang, Z.; Tang, R.; Zhu, D.; Wang, W.; Yi, L.; Ma, L. Non-peptide guided auto-secretion of recombinant proteins by super-folder green fluorescent protein in Escherichia coli. *Sci. Rep.* **2017**, *7*, 6990. [CrossRef] [PubMed]
34. Tan, L.; Hong, P.; Yang, P.; Zhou, C.; Xiao, D.; Zhong, T. Correlation Between the Water Solubility and Secondary Structure of Tilapia-Soybean Protein Co-Precipitates. *Molecules* **2019**, *24*, 4337. [CrossRef]
35. Hou, J.; Adhikari, B.; Cheng, J. DeepSF: Deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics* **2018**, *34*, 1295–1303. [CrossRef] [PubMed]
36. Li, Z.; Lin, Y.; Elofsson, A.; Yao, Y. Protein Contact Map Prediction Based on ResNet and DenseNet. *BioMed Res. Int.* **2020**, *2020*, e7584968. [CrossRef]
37. Berman, H.M.; Westbrook, J.D.; Gabanyi, M.J.; Tao, W.; Shah, R.; Kouranov, A.; Schwede, T.; Arnold, K.; Kiefer, F.; Bordoli, L.; et al. The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.* **2009**, *37*, D365–D368. [CrossRef] [PubMed]
38. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
39. Harris, D.; Harris, S. *Digital Design and Computer Architecture*; Morgan Kaufmann: Burlington, MA, USA, 2010; ISBN 978-0-08-054706-0.
40. Conneau, A.; Schwenk, H.; Barrault, L.; Lecun, Y. Very Deep Convolutional Networks for Natural Language Processing. *KI—Künstl. Intell.* **2016**, *26*, 1.
41. Chang, S.; Zhang, Y.; Han, W.; Yu, M.; Guo, X.; Tan, W.; Cui, X.; Witbrock, M.; Hasegawa-Johnson, M.; Huang, T.S. Dilated Recurrent Neural Networks. *arXiv* **2017**, arXiv:1710.02224. Available online: http://arxiv.org/abs/1710.02224 (accessed on 25 March 2021).
42. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* **2015**, arXiv:1505.00853.
43. Han, D.; Yun, S.; Heo, B.; Yoo, Y. ReXNet: Diminishing Representational Bottleneck on Convolutional Neural Network. *arXiv* **2020**, arXiv:2007.00992. Available online: http://arxiv.org/abs/2007.00992 (accessed on 25 March 2021).
44. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
45. Roy, A.G.; Navab, N.; Wachinger, C. Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks. *IEEE Trans. Med. Imaging* **2019**, *38*, 540–549. [CrossRef]
46. Cheng, J.; Randall, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33*, W72–W76. [CrossRef] [PubMed]
47. Walsh, I.; Martin, A.J.M.; Di Domenico, T.; Tosatto, S.C.E. ESpritz: Accurate and fast prediction of protein disorder. *Bioinformatics* **2012**, *28*, 503–509. [CrossRef]
48. Müller, A.T.; Gabernet, G.; Hiss, J.A.; Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **2017**, *33*, 2753–2755. [CrossRef] [PubMed]
49. Ramos, D.; Franco-Pedroso, J.; Lozano-Diez, A.; Gonzalez-Rodriguez, J. Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. *Entropy* **2018**, *20*, 208. [CrossRef]
50. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980. Available online: http://arxiv.org/abs/1412.6980 (accessed on 5 April 2021).
51. Li, W.; Xing, X.; Liu, F.; Zhang, Y. Application of Improved Grid Search Algorithm on SVM for Classification of Tumor Gene. *Int. J. Multimed. Ubiquitous Eng.* **2014**, *9*, 181–188. [CrossRef]