Heliyon 10 (2024) e36041

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

ProSol-multi: Protein solubility prediction via amino acids multi-level correlation and discriminative distribution

Hina Ghafoor [a,b,1], Muhammad Nabeel Asim [b,*,1], Muhammad Ali Ibrahim [a,b,1], Andreas Dengel [a,b]

[a] *Department of Computer Science, Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, 67663, Germany*
[b] *German Research Center for Artificial Intelligence GmbH, Kaiserslautern, 67663, Germany*

ARTICLE INFO

Keywords:
Protein solubility prediction
Machine learning
Physico-chemical properties
Amino acids multi-level correlation
Amino acids multi-level discriminative distribution
Protein sequence encoder
Protein sequence descriptor
Computational protein solubility analysis

ABSTRACT

Protein solubility prediction is useful for the careful selection of highly effective candidate proteins for drug development. In recombinant proteins synthesis, solubility prediction is valuable for optimizing key protein characteristics, including stability, functionality, and ease of purification. It contains valuable information about potential biomarkers or therapeutic targets and helps in early forecasting of neurodegenerative diseases, cancer, and cardiovascular disorders. Traditional wet-lab experimental protein solubility prediction approaches are error-prone, time-consuming, and costly. Researchers harnessed the competence of Artificial Intelligence approaches for replacing experimental approaches with computational predictors. These predictors inferred the solubility of proteins by analyzing amino acids distributions in raw protein sequences. There is still a lot of room for the development of robust computational predictors because existing predictors remain fail in extracting comprehensive discriminative distribution of amino acids. To more precisely discriminate soluble proteins from insoluble proteins, this paper presents ProSol-Multi predictor that makes use of a novel MLCDE encoder and Random Forest classifier. MLCDE encoder transforms protein sequences into informative statistical vectors by capturing amino acids multi-level correlation and discriminative distribution within raw protein sequences. The performance of proposed encoder is evaluated against 56 existing protein sequence encoding methods on a widely used protein solubility prediction benchmark dataset under two different experimental settings namely intrinsic and extrinsic. Intrinsic evaluation reveals that from all sequence encoders, proposed MLCDE encoder manages to generate non-overlapping clusters of soluble and insoluble classes. In extrinsic evaluation, 10 machine learning classifiers achieve better performance with proposed MLCDE encoder as compared to 56 existing protein sequence encoders. Moreover, across 4 public benchmark datasets, proposed ProSol-Multi predictor outshines 20 existing predictors by an average accuracy of 3%, MCC and AU-ROC of 2%. ProSol-Multi interactive web application is available at https://sds_genetic_analysis.opendfki.de/ProSol-Multi.

[*] Corresponding author.
E-mail address: Muhammad_Nabeel.Asim@dfki.de (M.N. Asim).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e36041
Received 16 May 2024; Received in revised form 1 August 2024; Accepted 8 August 2024

Available online 22 August 2024

0. Introduction

Proteins are key players in various biological processes, including pH maintenance, metabolic reactions, and message transmission across different cells [50]. They are essential for the proper functioning of the immune system and for the growth of organs like bones, skin, etc. [85]. Considering molecular structure as well as biological roles, proteins can be classified into three major classes namely: simple, conjugated, and derived [32]. Simple proteins are made up of a large number of amino acids, while conjugated proteins form when simple proteins combine with non-protein substances [32]. Derived proteins, are produced through the breakdown or combination of simple and conjugated proteins, such as proteoses, peptides, and denatured proteins [61]. From derived proteins, one sub-category of proteins that has gained significant attention is recombinant proteins, which are often produced through genetic engineering processes [63]. These proteins are used in the development of therapies, antibodies, and medicines like insulin, interleukins, thrombolytic and interferons drugs [65,113]. However, the solubility of recombinant proteins is a critical factor that determines their expression, stability, proper functionality [112] and overall effectiveness.

Soluble proteins efficiently perform their dedicated functionalities by readily dissolving in water or other aqueous solutions and maintaining their proper structure [112]. They can easily interact with other molecules and move across the cells to specific compartments through biological fluids. The motility characteristic enables them to distribute throughout the organism and facilitate enzymatic reactions, and signaling pathways [112]. Soluble proteins are essential for the development of finely dispersed colloidal systems [115,87,114] that control the flow of blood and prevent blood clotting and ascites [43,114]. On the other hand, insoluble proteins do not dissolve into water or other aqueous solutions [72,105]. These proteins form clumps or precipitates and lose their functional structure, leading to cellular dysfunction [72,105] and more than 40 different complications such as amyloids [25], Alzheimer's [26], and Parkinson's disease [57]. Insoluble proteins form aggregates, which can clog up the protein purification process, that affects recombinant protein synthesis pipelines [72,105]. Moreover, insoluble protein aggregates can reduce the bio-availability and efficacy of protein-based drugs by hindering their proper folding and function [91,47,80]. This can lead to decreased potency or even render the drugs inactive, thereby impeding the development of effective protein-based therapeutics [59,37].

Solubility does not solely depend on protein physicochemical properties [8] but also on the type of host and stringent cellular environment like pH, temperature, and ionic strength [41,35]. Up to date, only around a quarter of proteins are successfully expressed in soluble form.[2] Considering the importance of soluble proteins for functional research, drug development and insoluble proteins for the development of various diseases. Accurate prediction of protein solubility is an inevitable challenge in proteomics research and pharmaceutical industry. Protein solubility information helps in designing efficient bioprocessing strategies for large-scale protein production, reducing production costs, and improving overall process efficiency [10]. It aids in the rational design of protein variants with improved solubility, stability, and expression characteristics for various industrial and research applications [64]. Moreover, it helps in formulating biopharmaceutical products with improved stability, bioavailability, efficacy and shelf-life [10]. It assists in the identification of soluble enzymes for industrial applications and contribute to the development of more effective biocatalysts for diverse industrial processes [95]. In a nutshell, accurate protein solubility prediction can significantly accelerate the process of early detection of diverse diseases, development of potent therapeutic recombinant proteins, and precise selection of drug development candidates [95].

Trivial wet-lab experimental methods for predicting protein solubility involve measuring the concentration of proteins in a solution, determining the effects of the solvent, and evaluating the flexibility and sensitivity of the protein's structure [101,2]. These methods are expensive and time consuming [101]. Due to these challenges, analyzing the solubility of protein sequences on a large scale using experimental approaches becomes difficult [95]. Researchers have discovered that raw amino acid sequence is the most basic determinant of solubility of a protein [102,83]. Different studies have established a very strong correlation among protein solubility and various sequence based features like availability of the hydrophobic stretches, composition of unique amino acids types, length of raw sequence, and amino acid distribution patterns [102,83]. Following the huge success of Artificial Intelligence based computational approaches in the development of robust anti-inflammatory peptides predictors like AIPs-SnTCN [84], anti-viral peptides predictors like DeepAVP-TPPred [104], Deepstacked-AVPs [4], pAVP_PSSMDWT-EnC [3], anti-fungal peptides predictors like IAFPs-Mv-BiTCN [5] and other protein sequence analysis predictors [66,36], utilization of AI methods for development of protein solubility predictors have gained huge attention in protein engineering and design research community [93,49].

In recent times, according to our best knowledge, around 27 AI based predictors have been developed. On the basis of working paradigm, these predictors can be broadly classified into two categories, classification and regression. In the classification category, predictors analyze amino acids distribution patterns and utilize these patterns to categorize proteins into either soluble or insoluble classes [95,106]. Latest representative tool for this category is RPPSP [74] which was evaluated on most number of benchmark datasets in 2023. The unique selling point of the RPPSP predictor was its novel protein sequence encoder called CTAPAAC (Composition and Transition aware Amphiphilic Pseudo-Amino Acid Composition). This encoder transformed raw protein sequences into statistical vectors by extracting four different types of information: correlation, distribution, composition, and transition of amino acids. The fusion of different features helped the machine learning classifier namely Random Forest extract comprehensive discriminative patterns between soluble and insoluble proteins. RPPSP predictor outperformed state-of-the-art predictors with accuracy improvements of 6%, 7%, 25%, and 10% on PSI:Biology, E.coli, price, and Esol datasets respectively [74]. On the other hand, in regression category, predictors analyze amino acids distribution patterns and utilize these patterns to predict solubility values of protein sequences [45,78]. Latest representative tool of this category is HybridGCN [18] that was evaluated on 2 datasets in 2023. The unique selling

[2] http://targetdb.rcsb.org/metrics/.

point of HybridGCN was its novel approach of combining multiple types of protein features, including advanced deep learning features and classical biophysical features. This was done in a unique setting where protein sequence data was modeled using the graph structure, and a graph convolutional network constructed the mapping between protein sequences and the corresponding solubility values. Unlike existing predictors that typically relied on either handcrafted features or deep learning features alone, HybridGCN leveraged both through two key innovations: 1) It incorporated a zero-shot learning feature called ESM-1v, derived from a large protein language model, which captured comprehensive information about protein functions and structures. 2) It introduced an Adaptive Feature Re-weighting (AFR) module that dynamically adjusts the importance of different features, prioritizing the most informative ones for solubility prediction. This allowed HybridGCN to effectively integrate domain-specific knowledge from handcrafted features with discriminative insights from deep learning models and achieve state-of-the-art performance on 2 datasets [18].

Comparison of predictors framing protein solubility as a classification task or a regression task reveals that although prediction of exact solubility values is more useful as compared to classifying proteins directly into soluble and insoluble classes. However, considering rich patterns of 20 unique amino acids in diverse protein sequences make it extremely challenging to detect accurate continuous solubility values as compare to categorization of proteins into soluble and in-soluble classes. Overall objective of both types of predictors is to discriminate soluble proteins from non-soluble proteins. In regression predictors, error rate of continues values detection propagates to second level where based on a particular solubility threshold, proteins are categorized into soluble and insoluble classes [93,49].

To facilitate the readers, most recent 27 computational predictors are briefly summarized in Supplementary File-1. A closer look at Supplementary File-1 reveals that, from 27 computational predictors, more than 20 predictors fall under the category of protein solubility classification. These predictors pipelines comprise of two different stages. At first stage, protein sequences are transformed into statistical feature space. At second stage, statistical feature space is utilized to train machine or deep learning based classifiers. Existing predictors have used more than 25 unique sequence encoding methods to generate statistical representations of protein sequences. These sequence encoding methods can be broadly categorized into six different classes: amino acid composition [9,86,117,14,15], grouped amino acid composition [12,13,33,34,44], structural features [55,52], physico-chemical properties [103,68,70], substitution matrix [67], and scoring matrix [107]. Amino acid composition and grouped amino acid composition sequence encoders do not capture the order of amino acids. Furthermore, grouped amino acids composition encoders usually oversimplify protein sequences and overlook key functional characteristics of proteins. Also, these encoders immensely rely on the specific criteria used to group the amino acids which may not work well for different benchmark datasets. Physico-chemical properties based encoders utilize pre-computed values, which limits their potential to extract comprehensive interactions and relationships among amino acids within protein sequences. Substitution matrix based encoders lack to capture diversity of protein sequences, scoring matrix based encoders lack to capture local and global contextual information and show biaseness towards more abundant group of amino acids. Structure features based encoders lack to fully capture the structural complexity of proteins.

Considering the strong association of physico-chemical properties like charge, polarity, molecular weight, transmembrane tendency, hydrophobicity, hydrophilicity, etc. with solubility of proteins [8], prime focus of every other study has been to integrate different types of encoders with physico-chemical properties based encoders. The motive behind this was to generate informative and discriminative statistical representations of protein sequences that can assist even a simple machine learning classifier to accurately discriminate soluble proteins from insoluble proteins. However, these approaches only focus on the concatenation of diverse amino acids information such as occurrence frequency, transition, physico-chemical properties, etc. These approaches fully neglect the changes in relatedness and distribution of amino acids over difference distances in protein sequences. Hence, these approaches lack to generate physico-chemical properties aware discriminative distribution based statistical representations of protein sequences. This is why even sophisticated machine learning as well as deep learning classifiers fail to distinguish soluble proteins from insoluble proteins across different datasets, indicating a lot of room for the development of novel encoders and more effective predictors. Following the need of a precise and robust computational protein solubility predictor, contribution of this paper are manifold:

(I) It presents ProSol-Multi predictor that utilizes a potent sequence encoder MLCDE and Random Forest classifier to effectively discriminate soluble proteins from insoluble proteins using only raw protein sequence data **(II)** Proposed MLCDE encoder converts protein sequences into informative statistical vectors by extracting physico-chemical properties aware multi-level correlation and distribution information of amino acids **(III)** It thoroughly compares the potential of proposed MLCDE encoder with 56 existing protein sequence encoders under the hood of extrinsic and intrinsic evaluation. The core objective of an intrinsic evaluation is to analyze which of the sequence encoder captures highly discriminative distribution of amino acids and generates non-overlapping clusters for both soluble and insoluble protein classes. On the other hand, focus of an extrinsic evaluation is to analyze the impact of statistical vectors generated by proposed MLCDE and existing protein sequence encoders on the predictive performance of 10 distinct machine learning classifiers **(IV)** With an aim to showcase the predictive performance as well as generalizeability of proposed ProSol-Multi predictor, it performs a fair performance comparison of ProSol-Multi predictor with 20 most recent protein solubility predictors **(V)** To accelerate the process of discriminating soluble proteins from insoluble proteins, it develops an interactive web application available at (https://sds_genetic_analysis.opendfki.de/ProSol-Multi).

1. Results

This section illustrates amino acids discriminative distribution analysis in soluble and insoluble protein sequences. It performs a thorough extrinsic performance comparison of proposed MLCDE encoder with 56 existing protein sequence encoders using 10 distinct machine learning classifiers. It also performs an intrinsic performance comparison of proposed MLCDE encoder with existing encoders. Finally, it compares the effectiveness of proposed ProSol-Multi predictor with 20 existing predictors.

1.1. Amino acids distribution analysis in soluble and insoluble classes

We employ sequence logo library [98] to examine which amino acids have more than 70% position-wise occurrence probability in both soluble and insoluble protein sequences. This analysis highlights discriminative distribution patterns of amino acids in soluble and insoluble protein sequence classes. Fig. 14 illustrates the varying lengths of sequences, ranging from 20 to 1750 amino acids. Visualizing the amino acids discriminative distribution in longer sequences, especially in sequences having 175 or more amino acids, is difficult. To overcome this, we focus on the 1/4 part of the sequences based on only 40 amino acids from start and eradicate the amino acids that occur after 40^{th} position, assuming discriminative distribution of amino acids remain somewhat similar in other parts of the sequences. Additionally, in this analysis, we remove sequences from the datasets that have a length of less than 40 amino acids. Fig. 1 showcases the distribution of amino acids in soluble and insoluble classes across four benchmark datasets.

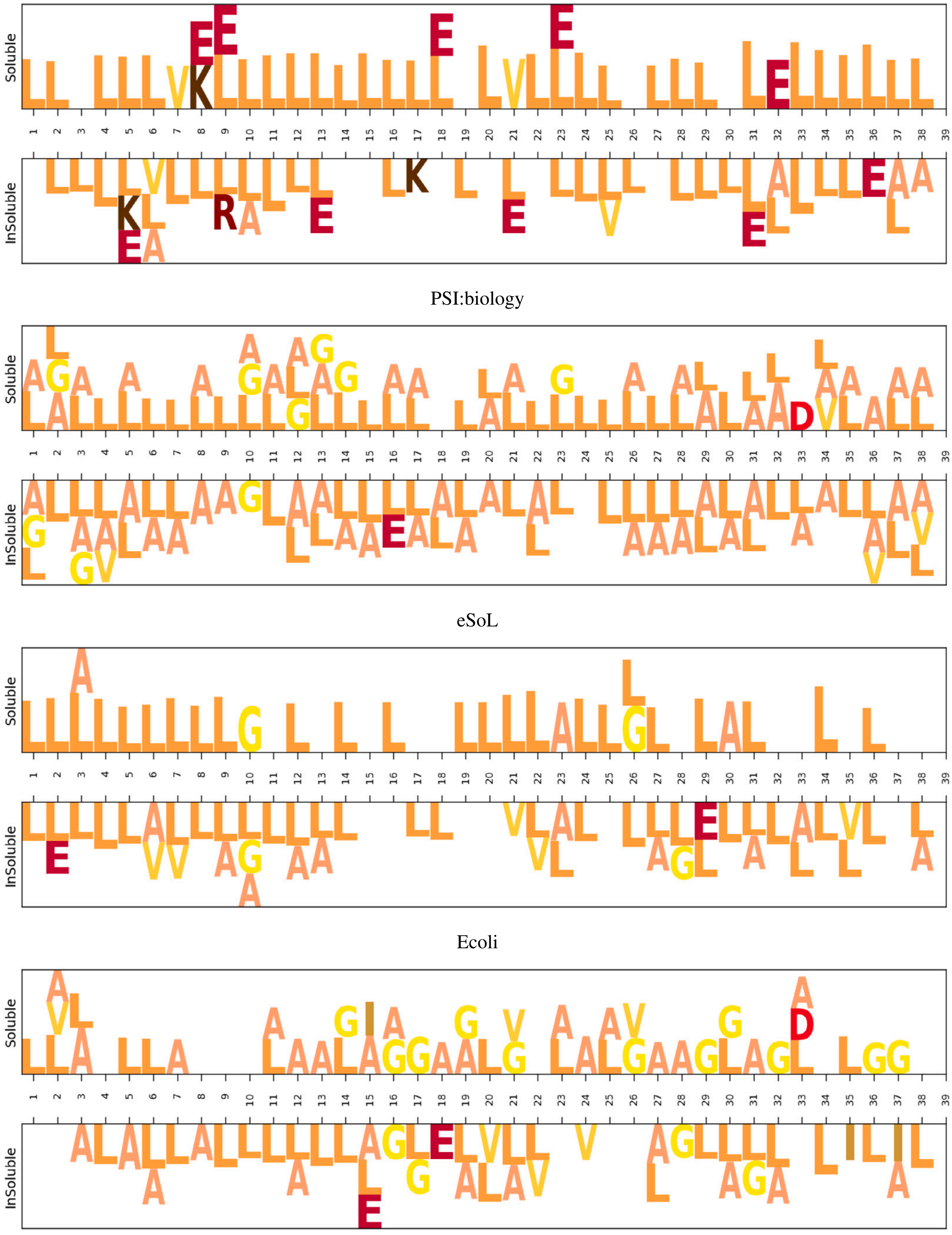
It is evident in Fig. 1 that, across 3 benchmark datasets including PSI:Biology, eSol and Ecoli, distribution of individual amino acids is quite similar in soluble and insoluble classes, such as frequent occurrence of 'L:leucine' amino acid at most unique positions. Whereas, in benchmark Price dataset, along with 'L:leucine', 'A:alanine' amino acid occur frequently at same positions in both classes sequences. This precise distributional analysis indicates that these sequences have useful amino acids patterns that can be used to distinguish soluble and insoluble protein sequences. Also, existing encoding methods are bound to generate very similar statistical vectors for soluble and insoluble protein sequences as they usually focus on correlation, distribution, composition, and transition of individual or group of amino acids. Less discriminative statistical vectors make the discrimination of soluble protein sequences from insoluble protein sequences, extremely difficult for machine learning classifiers. It can be inferred from Fig. 1 that, across all 4 benchmark datasets, group of amino acids distribution vary at most positions across soluble and insoluble classes, such as occurrence of 'L:leucine' and 'V:Valine' in soluble class, 'L:leucine' and 'K:Lysine' in insoluble class of benchmark PSI:Biology dataset. Likewise, 'L:leucine' followed by 'G:glycine' in soluble class and 'G:glycine' followed by 'L:leucine' in insoluble class of benchmark eSol dataset, and similar patterns are evident for other two datasets as well. Unlike existing traditional sequence encoders which lacks to capture relatedness and distribution of amino acids at different distances in protein sequences. Proposed MLCDE encoder captures physico-chemical properties aware multi-level correlation and discriminative distribution of amino acids. This functional paradigm helps to encode informative yet discriminative patterns that assist the machine learning classifier to accurately distinguish soluble protein sequences from insoluble protein sequences.

1.2. Proposed MLCDE encoder extrinsic performance comparison with existing encoders

To showcase the effectiveness of proposed MLCDE encoder, we perform a detailed performance comparison of proposed MLCDE encoder with 56 existing encoders of 14 distinct categories using benchmark PSI:biology dataset and 10 distinct classifiers. The results produced by proposed MLCDE encoder and 56 existing encoders across 10 classifiers in terms of 9 different evaluation metrics are given in Supplementary File-2, Table-3. Analysis of Supplementary File-2, Table-3 indicates that from amino acid distribution, amino acid groups distribution, and gap-based amino acid distribution categories, DDE, KSCTriad, and CKSNAP achieve top performance. From autocorrelation, co-variance, and binary encoding categories, NMBroto, auto-covariance, and binary 5bit-type-2 achieve better performance. From local-global context-aware, sequence order, and traditional network categories, WSRC-local-global, CTAPAAC, and enhanced-complex-network achieve best performance. From pre-trained neural network, physico-chemical properties, optimized physico-chemical properties, Fourier transformation, and substitution matrix based encoders categories, AESNN3, AAINDEX, Z-Scale, MappingClass eiip fourier, and BLOSUM62 achieve best performance. To better facilitate the readers, accuracy achieved by 14 top performing existing encoders and proposed MLCDE encoder across 10 different classifiers on benchmark PSI:Biology dataset are given in Supplementary File-2, Table-4 and graphically shown here in Fig. 2.

Fig. 2 illustrates that from existing 14 different encoders, sequence order based encoder CTAPAAC and amino acids distribution based encoder DDE achieve the best average accuracy of 72% due to their capability to handle variations in length of sequences by extracting relative abundance of different amino acids inside sub-sequences. This performance figure is followed by 71% achieved by CKSNAP and WSRC-local-global and 70% achieved by KSCTriad. Then, 4 different encoders namely ZScale, AAINDEX, BLOSUM62, and NMBroto achieve the average accuracy around 68% followed by 67% achieved by binary 5bit-type-2, MappingClass eiip fourier, auto covariance, and AESNN3 encoders. Among all 14 different encoders, enhanced complex network achieves the lowest average accuracy of 66% because it lacks to capture diverse kinds of amino acids distribution patterns. Unlike existing encoders that only manage to capture correlation and distribution patterns at limited level, proposed MLCDE encoder captures amino acids multi-level correlation and discriminative distribution. This helps MLDE encoder to outperform all 14 different encoders by producing an average accuracy of 73%. MLCDE encoder enhances the performance of all 10 different classifiers and it achieves best performance with AdaBoost classifier that outperforms existing sequence order based encoder CTAPAAC and amino acids distribution based encoder DDE by 1%. This trend of better performance is even more evident in Supplementary File-2, Table-3 across other evaluation measures.

To facilitate readers, Fig. 3 performs AU-ROC and AU-PRC performance comparison of proposed MLCDE encoder with 14 top-performing existing encoders using best performing machine learning classifier namely Random Forest on benchmark PSI:Biology dataset. It is evident in the Figure that 2 existing encoders namely CKSNAP, and CTAPAAC achieve AU-ROC of 81% followed by 80% achieved by 3 existing encoders namely DDE, WSRC-local-global, and KSCTriad. Among all existing encoders, 3 encoders namely binary-5-bit-type-2, mappingclass-EIIP-fourier, and enhanced complex network achieve low AU-ROC scores. Overall, from existing encoders, composition and transition aware encoders, and gap-based amino acid distribution encoders achieve better AU-ROC scores. However, among all encoders, proposed MLCDE encoder achieves best AU-ROC score with Random Forest classifier as it outperforms best performing existing encoders by a slight margin. Similar performance trends are shown by proposed and existing encoders in



PSI:biology

eSoL

Ecoli

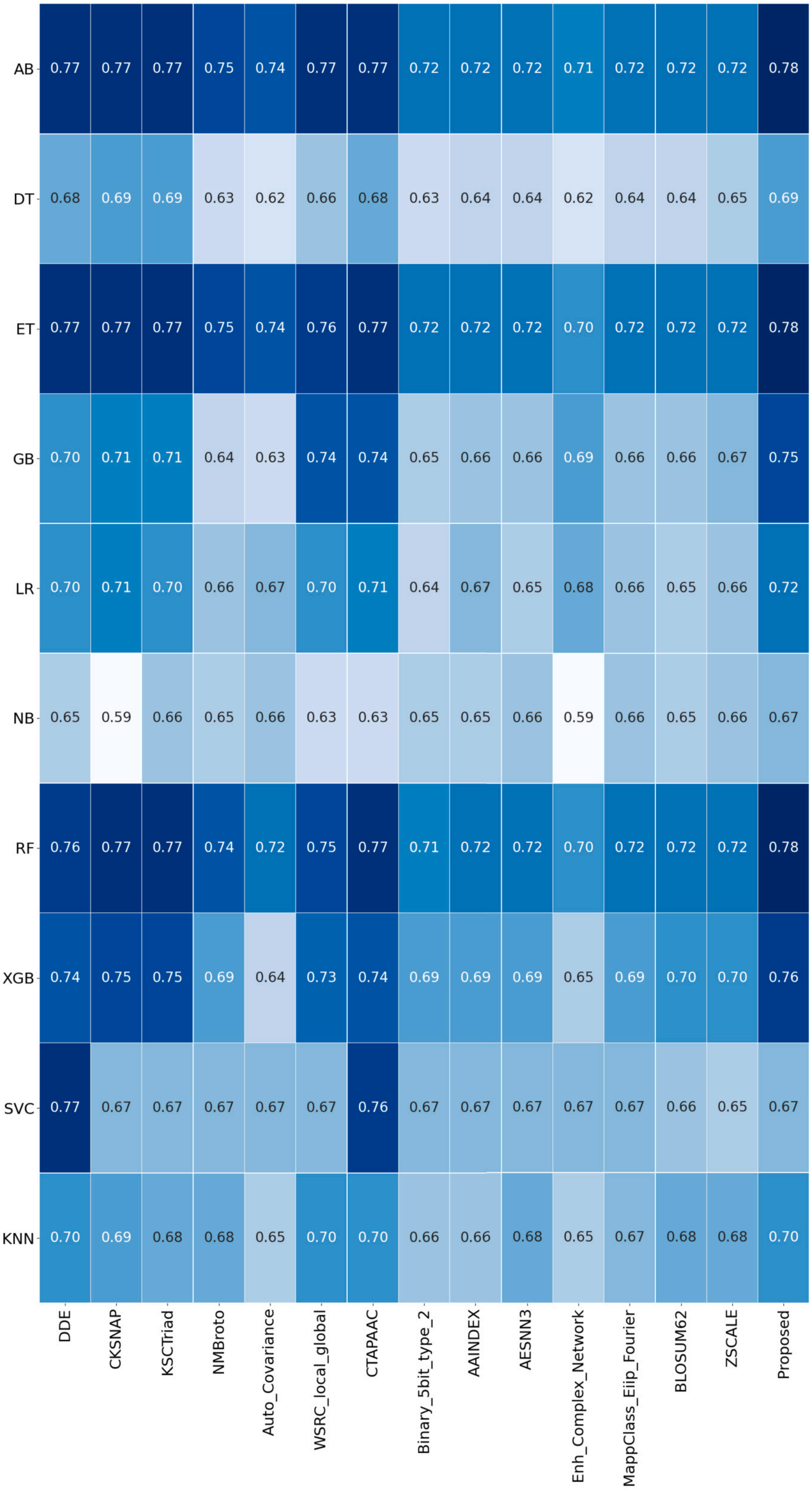
Price

Fig. 1. Distribution of amino acids across unique positions in soluble and insoluble protein sequences of 4 different benchmark datasets.

terms of AU-PRC. However, overall proposed encoder achieves peak performance of 82% in terms of AU-ROC whereas AU-PRC performance falls around 70%.

1.3. Proposed MLCDE encoder intrinsic performance comparison with existing top performing encoders

This section conducts an intrinsic performance comparison between the proposed encoder and 14 top performing encoders from existing 56 encoders. The goal is to assess the quality of the statistical vectors generated by both types of encoders. To perform an intrinsic performance analysis, we reduce the statistical vectors of proposed and existing encoders up to 20% dimensions through



	DDE	CKSNAP	KSCTriad	NMBroto	Auto_Covariance	WSRC_local_global	CTAPAAC	Binary_5bit_type_2	AAINDEX	AESNN3	Enh_Complex_Network	MappClass_Eiip_Fourier	BLOSUM62	ZSCALE	Proposed
AB	0.77	0.77	0.77	0.75	0.74	0.77	0.77	0.72	0.72	0.72	0.71	0.72	0.72	0.72	0.78
DT	0.68	0.69	0.69	0.63	0.62	0.66	0.68	0.63	0.64	0.64	0.62	0.64	0.64	0.65	0.69
ET	0.77	0.77	0.77	0.75	0.74	0.76	0.77	0.72	0.72	0.72	0.70	0.72	0.72	0.72	0.78
GB	0.70	0.71	0.71	0.64	0.63	0.74	0.74	0.65	0.66	0.66	0.69	0.66	0.66	0.67	0.75
LR	0.70	0.71	0.70	0.66	0.67	0.70	0.71	0.64	0.67	0.65	0.68	0.66	0.65	0.66	0.72
NB	0.65	0.59	0.66	0.65	0.66	0.63	0.63	0.65	0.65	0.66	0.59	0.66	0.65	0.66	0.67
RF	0.76	0.77	0.77	0.74	0.72	0.75	0.77	0.71	0.72	0.72	0.70	0.72	0.72	0.72	0.78
XGB	0.74	0.75	0.75	0.69	0.64	0.73	0.74	0.69	0.69	0.69	0.65	0.69	0.70	0.70	0.76
SVC	0.77	0.67	0.67	0.67	0.67	0.67	0.76	0.67	0.67	0.67	0.67	0.67	0.66	0.65	0.67
KNN	0.70	0.69	0.68	0.68	0.65	0.70	0.70	0.66	0.66	0.68	0.65	0.67	0.68	0.68	0.70

Fig. 2. Extrinsic performance comparison of proposed MLCDE encoder with 14 top-performing existing encoders over benchmark PSI:Biology dataset.

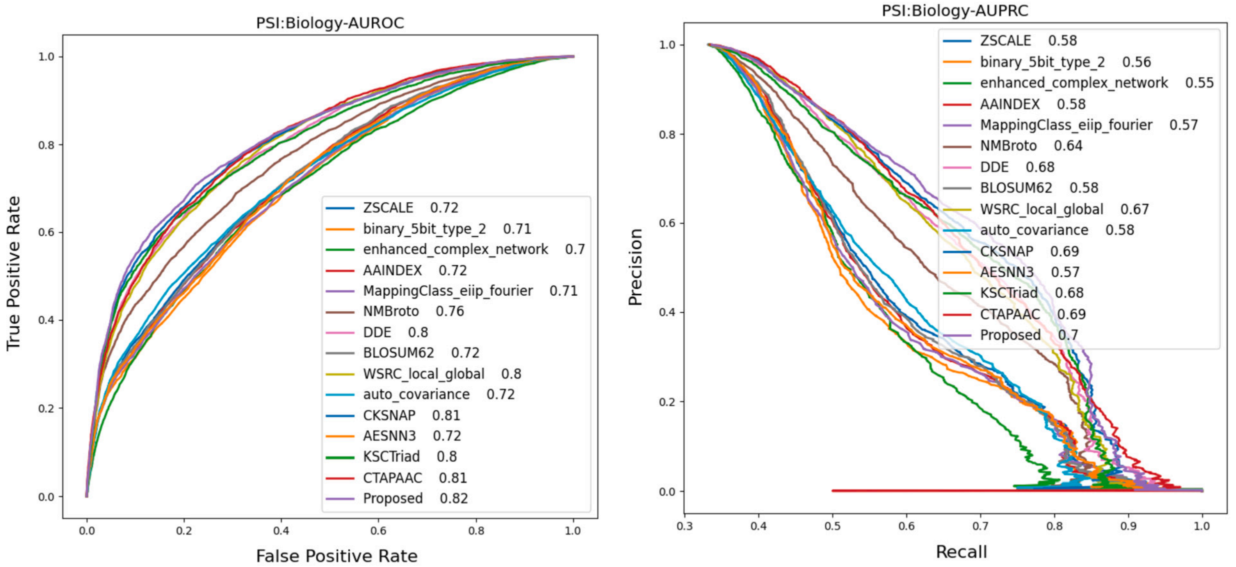


Fig. 3. Comparison of AU-ROC and AU-PRC produced by proposed MLCDE encoder and 14 top performing existing encoders using best performing random forest classifier on PSI:Biology dataset.

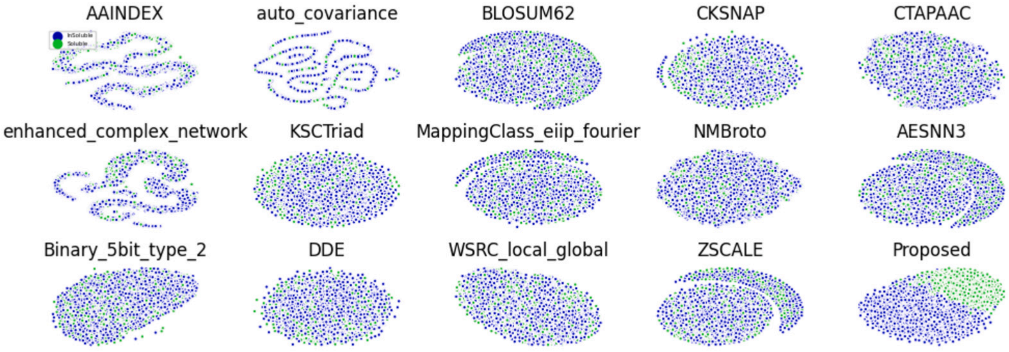


Fig. 4. Intrinsic performance comparison of proposed MLCDE encoder with 14 top-performing existing encoders over benchmark PSI:Biology dataset.

principal component analysis and up to 2 dimensions through t-distributed stochastic neighbor embedding method. Fig. 4 showcases the clusters of soluble and insoluble classes produced by proposed and existing traditional encoders on benchmark PSI:Biology dataset.

A first look at Fig. 4 reveals that existing encoders fail to generate highly disjoint clusters for soluble and insoluble classes due to lack of discriminative features. Contrarily, proposed MLCDE encoder generates highly non-overlapping clusters for soluble and insoluble classes due to the abundance of discriminative features. This is primarily due to the functional paradigm of proposed MLCDE encoder. Instead of only focusing on simple correlation, distribution, composition, and transition information, proposed MLCDE encoder focuses on the extraction physico-chemical properties aware multi-level correlation and discriminative distribution of amino acids.

1.4. Proposed ProSol-Multi predictor performance comparison with existing protein solubility predictors

This section compares the performance of proposed ProSol-Multi predictor with 20 existing predictors including SoluProt [52] CamSol [95], ParsnIP [83], DeepSol [58], SWI [8], ProteinSol [48], Prot5-P [99], ESM-MSA-P [99], CCSol [1], ESM1b-F [99], ESM12-F [99], PROSO II [92], SolPro [71], DDcCNN [108], ProGan [46], NetSolP [99], GraphSol [17], TAPE [82], SeqVec [49], and RPPSP [74]. From these 20 predictors, 11 predictors have been evaluated on benchmark PSI:Biology dataset, 6 predictors have been evaluated on benchmark PRICE dataset, 7 predictors have been evaluated on benchmark E.coli dataset, and 1 predictor has been evaluated on E.sol dataset. Similarly, 10 predictors have been evaluated on Price independent test set and 7 predictors have been evaluated on E.sol independent test set. We have taken these predictors performance values on respective datasets from their research papers and compared it with our proposed ProSol-Multi predictor performance below.

Fig. 5 illustrates the 5-fold accuracy, AU-ROC and MCC produced by proposed ProSol-Multi predictor and 10 different existing predictors on benchmark PSI:Biology dataset. It is apparent from Fig. 5 that among the existing predictors, correlation, distribution, composition and transition information aware RPPSP predictor achieves the best accuracy of 77%, AU-ROC of 81%, and MCC of

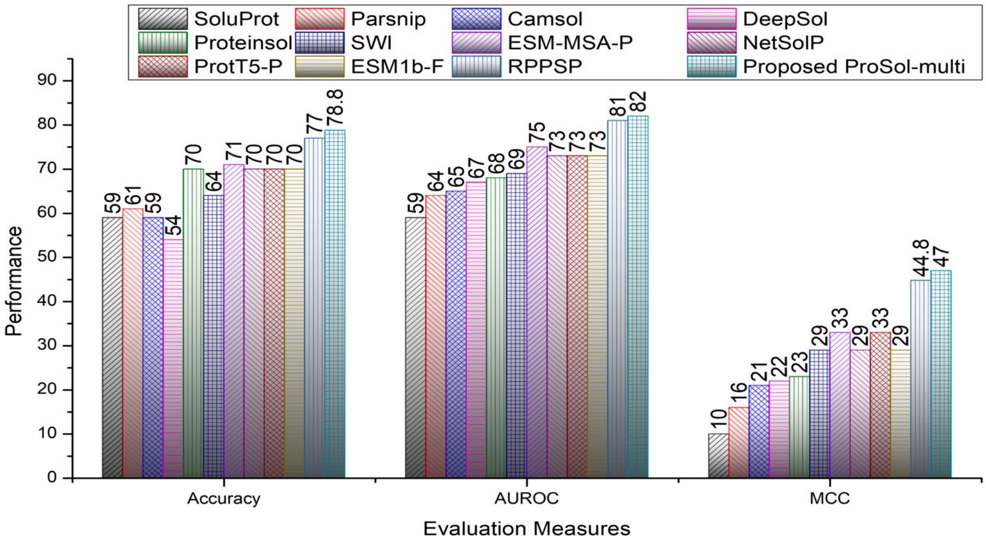


Fig. 5. Performance comparison of proposed ProSol-Multi predictor with 10 existing predictors over benchmark PSI:biology dataset.

45%. Second and third best performances across most evaluation metrics are achieved by language modeling based predictors. More deep predictors like ESM-MSA [99] based on 33-layer language model show better performance than less deep language model based predictors such as ProT5-P [99], NetSolP [99], and ESM1b-F [99] based on 24-layer language models. Fourth best performance is marked by amino acid composition aware predictors like SWI [8] and ProteinSol [48], except SoluProt [52] that marks limited performance and overall lowest MCC anf AU-ROC. Physico-chemical properties aware predictor CAMSOL and structural information aware predictors PARSNIP and DeepSol [58] show decent performance across most evaluation metrics.

Overall, proposed ProSol-Multi predictor outperforms all 10 existing predictors by a decent margin. Precisely, ProSol-Multi predictors outshines structural information and physico-chemical properties aware predictors by an accuracy margin of 18%, MCC margin of 25%, and AU-ROC margin of 15%. Furthermore, it outperforms state-of-the-art RPPSP predictor [74] on benchmark PSI:Biology dataset by accuracy, precision, recall, F1, specificity, and MCC of 2%, and AU-ROC of 1%. The prime reason behind the dominant performance of proposed ProSol-Multi predictor is the functional paradigm proposed MLCDE encoder. Unlike existing encoders who focus to extract diverse kinds of informative features, it focuses on the changes of physico-chemical properties aware correlation and distribution at different distances in protein sequences. This functional paradigm helps to encode highly discriminative distribution that helps even a simple Random Forest classifier to accurately distinguish soluble proteins from insoluble proteins.

Furthermore, Fig. 6 demonstrates the accuracy, precision, MCC, and AU-ROC achieved by proposed ProSol-Multi and 6 existing predictors on benchmark Price dataset. Like benchmark PSI:Biology dataset, here again, correlation, distribution, composition, and transition information based predictor (RPPSP) and language model based predictors (NetsolP [99], ESM1b-F [99] and ESM12-F [99] perform better followed by amino acid composition oriented predictors (SoluProt, SWI). However, proposed ProSol-Multi predictor surpasses language modeling oriented predictors accuracy, precision, as well as MCC by 3%, AU-ROC by 1%, amino acid composition oriented predictors accuracy by 2%, precision by 3%, AU-ROC by 2% and MCC by 3%. ProSol-Multi predictors also outshines state-of-the-art RPPSP predictor by accuracy of 2%, precision, MCC, and AU-ROC of 1%.

Moreover, Fig. 7 compares the performance of proposed ProSol-Multi predictor with 7 existing predictors on benchmark E.coli dataset in terms of 4 distinct evaluation measures. In Fig. 7, it is evident that amino acid correlation, distribution, composition, and transition information based predictor RPPSP achieves best performance among existing predictors. Among the three predictors based on amino acid composition, DDcCNN performs better than PROSOII [92] and SolPro [71]. DDcCNN achieves the highest accuracy, specificity, and MCC compared to all other predictors. However, when it comes to sensitivity, the amino acid structural information-based predictor DeepSol [58] performs better. Overall, DeepSol is third-best performer followed by amino acid structural information aware predictor called Parnsip. On the other hand, the physicochemical properties-based predictor CCSol achieves the lowest scores in terms of accuracy, specificity, sensitivity, and MCC.

Among all predictors, proposed ProSol-Multi predictor surpasses all existing predictors by a significant margin in terms of all 4 evaluation metrics. Specifically, it beats amino acid composition information based predictors by 19%, 17%, 10%, and 8% in terms of MCC, sensitivity, accuracy, and specificity. It outshines amino acid structural information based predictors by a mean performance of 11%, and physico-chemical properties aware predictors by a mean performance of 41%. It outperforms state-of-the-art RPPSP predictor by an accuracy of 2%, sensitivity of 7%, specificity of 1%, and MCC of 2%. Furthermore, Fig. 8 reveals the performance of existing RPPSP and proposed ProSol-Multi predictor on benchmark Esol dataset. It is evident in Fig. 8 that proposed ProSol-Multi predictor achieves a significant AU-ROC increment of 7%, accuracy, precision, recall, and F1 increment of 1%. This remarkable improvement can be attributed primarily to the exceptional effectiveness of the innovative sequence encoder employed by the proposed ProSol-Multi predictor.

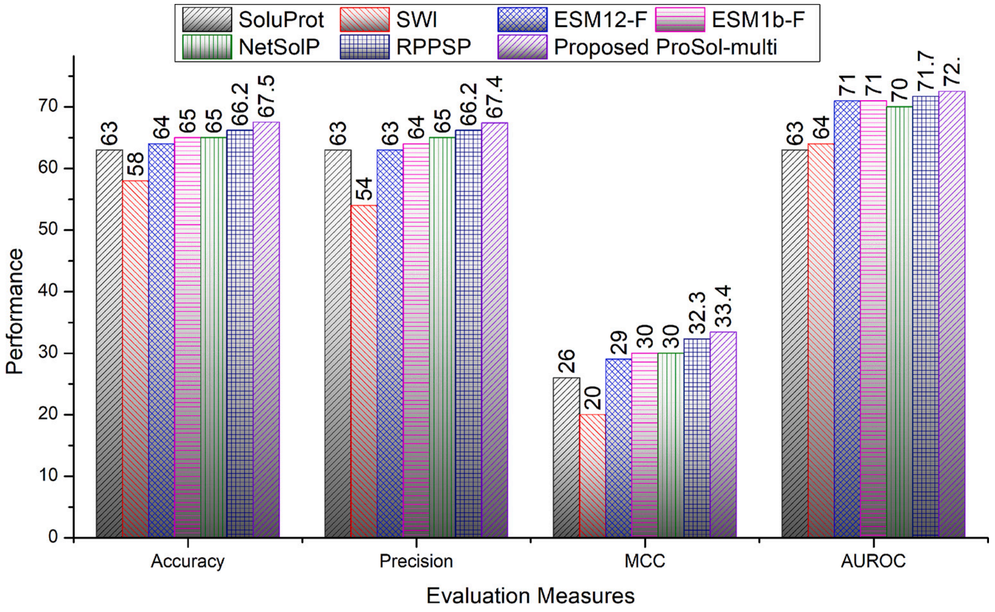


Fig. 6. Performance comparison of proposed ProSol-Multi predictor with 6 existing predictors over benchmark price dataset.

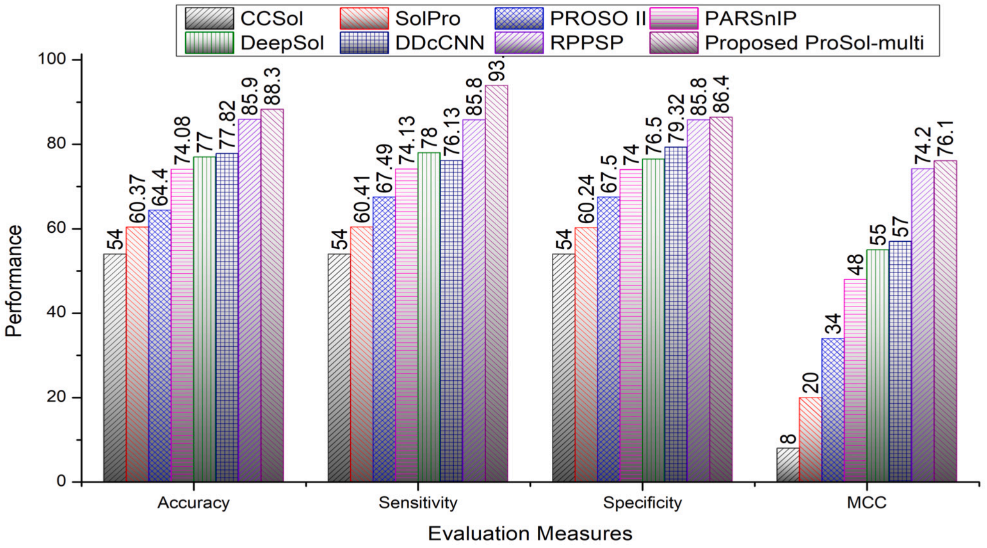


Fig. 7. Performance comparison of proposed ProSol-Multi predictor with 7 existing predictors over E.coli benchmark dataset.

1.4.1. *Proposed ProSol-Multi predictor performance comparison with existing predictors using independent test sets*

To better illustrate the predictive power of proposed ProSol-Multi predictor, we compare its performance with existing predictors on two independent test sets after training them on different core datasets. Fig. 9 illustrates the performance achieved by ProSol-Multi and 10 existing predictors on Price independent test set after training them on Price core dataset. In Fig. 9, a thorough performance analysis in terms of 4 different evaluation metrics indicates that, in contrast to the core datasets, a variety of predictors, including those based on amino acid structure information, amino acid composition, and physicochemical properties, exhibit similar performance on Price independent test set. However, once again, amino acid correlation, distribution, composition, and transition information based RPPSP predictor achieves the best performance followed by language modeling-based predictors. Among all the existing predictors, predictors based on amino acid composition information achieve the lowest performance across most evaluation metrics. Similar to the core datasets, the proposed ProSol-Multi predictor attains optimal predictive performance on the Price independent test set, marking the performance of almost 100% and outperforming state-of-the-art RPPSP predictor with an accuracy, precision and MCC margin of 1%.

Furthermore, Fig. 10 illustrates the performance achieved by ProSol-Multi and 6 existing predictors on Esol independent test set after performing training on Esol core dataset. Performance analysis in terms of 5 distinct evaluation metrics reveals that like Price

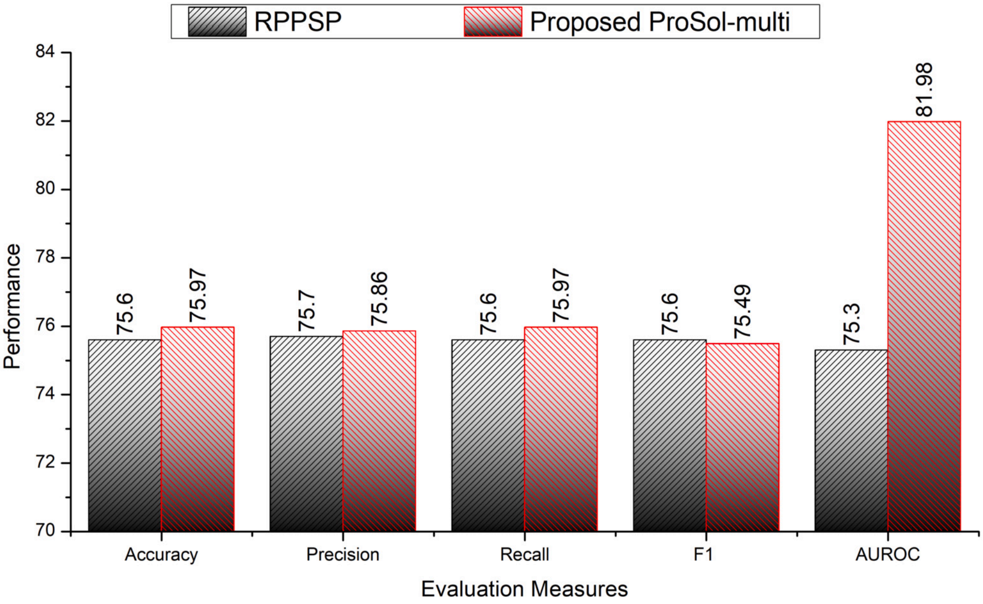


Fig. 8. Performance comparison of proposed ProSol-Multi predictor with existing predictors over E.Sol dataset.

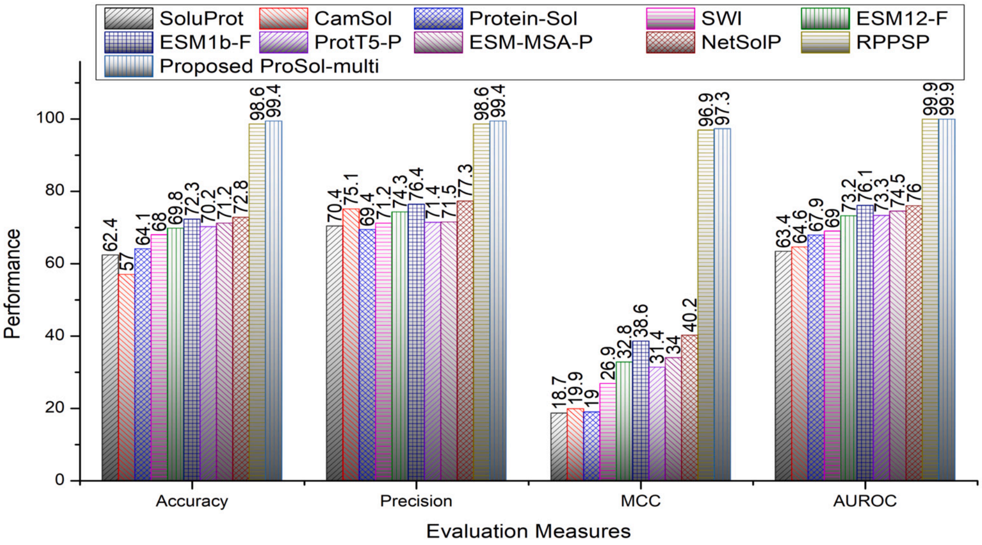


Fig. 9. Performance comparison of proposed ProSol-Multi predictor with 10 existing predictors over price independent test set.

independent test sets, existing predictors exhibit comparable performance trends, except for the amino acid composition information based predictor called ProteinSol. The accuracy, precision, recall, F1-score, and AUROC of most existing predictors fall around 78%, 79%, 71%, 74%, and 87% respectively. State-of-the-art RPPSP predictor achieves a performance of over 88% across all evaluation metrics. However, proposed ProSol-Multi predictor significantly outperforms state-of-the-art RPPSP predictor, achieving an accuracy, precision, recall increment of 7%, F1-score increment of 4% and AU-ROC increment of 6%.

1.5. Feature importance evaluation using SHAP analysis

Shapley additive explanations (SHAP) analysis is a method to understand how each feature in a model contributes to its predictions. It is based on game theory and provides a fair way to distribute the importance of a prediction among all features. Specifically, for each prediction, SHAP calculates the contribution of each feature by comparing the prediction with and without each feature present. This process is repeated for all predictions in the dataset. The results are then aggregated to give overall feature importance. Fig. 11 illustrates importance of top features on test set of benchmark PSI:Biology dataset. These features are ranked from top to bottom based on their overall importance in predicting solubility. Considering the size of features and similarity in importance trends of top features across all sequences of test set, here we have shown results on only 50 randomly selected sequences of test set. To illustrate better,

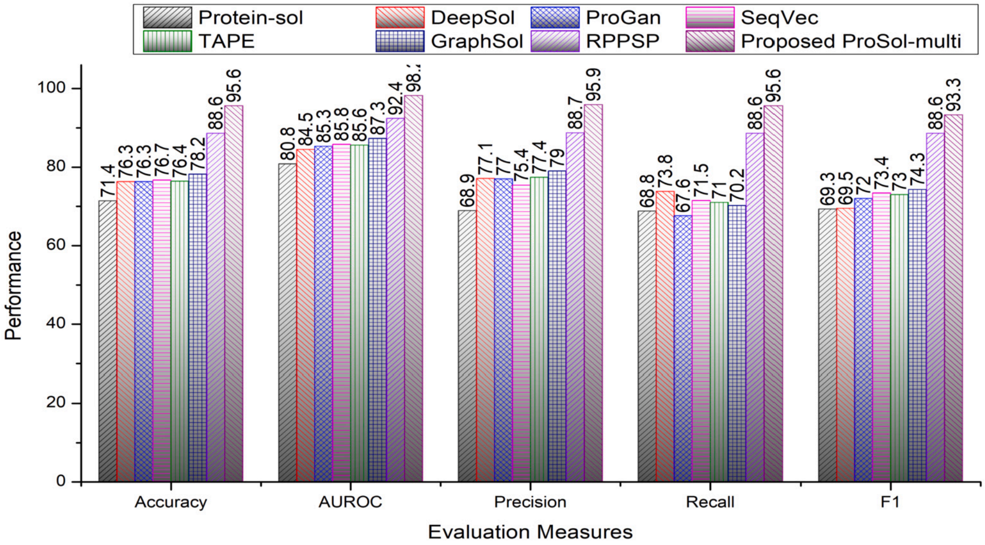


Fig. 10. Performance comparison of proposed ProSol-Multi predictor with 7 existing predictors over E.Sol independent test set.

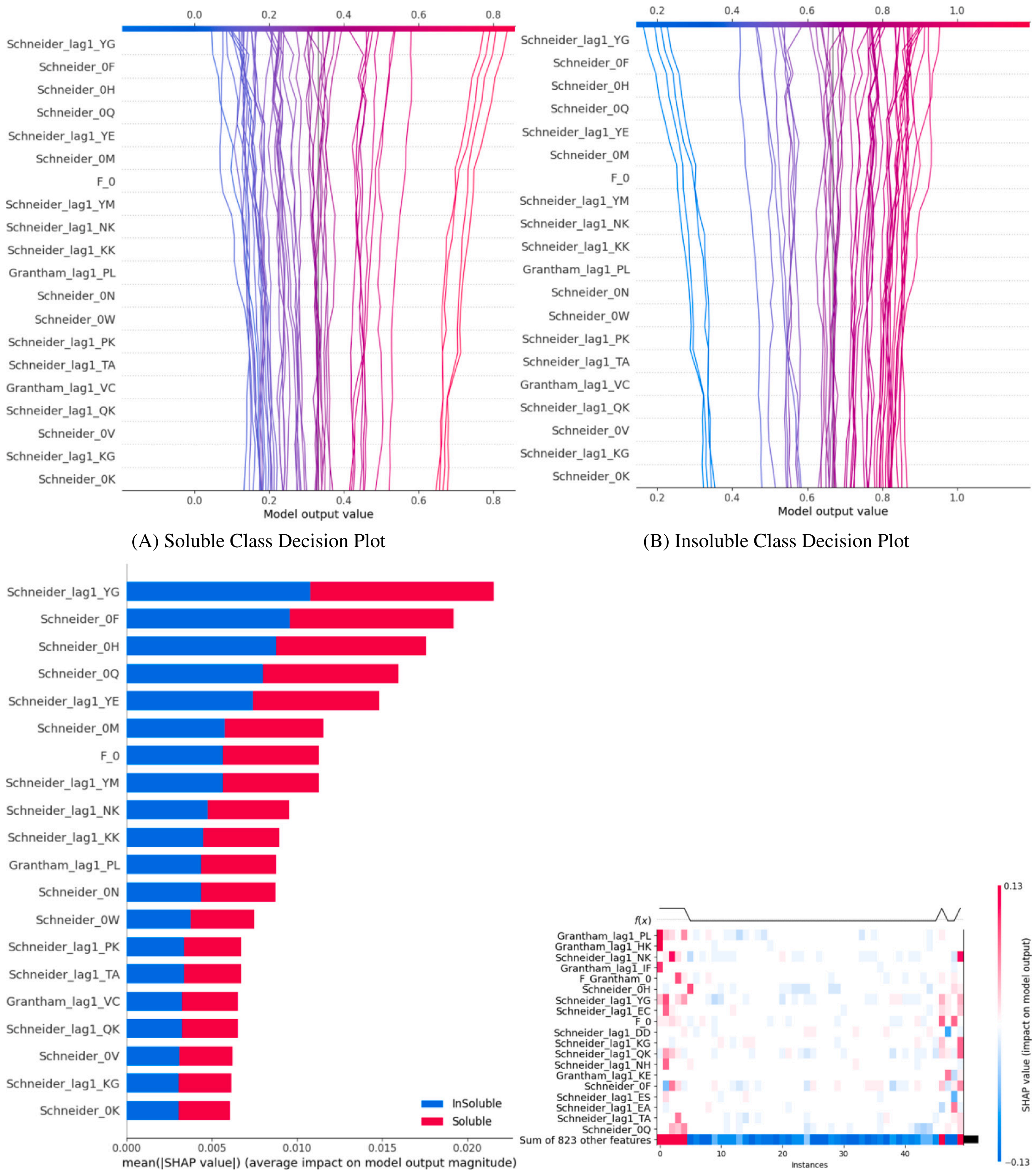
3 different types of plots are shown in Fig. 11 including decision plot, summary plot, and heatmap. Proposed MLCDE encoder makes use of two pre-computed amino acid distance matrices (Schneider and Grantham) which average the values of 4 physico-chemical properties (hydrophobicity, hydrophilicity, polarity, and side chain volume) to find bi-mer multi-level correlation information. Also, it captures individual amino acid distribution and bi-mer multi-level distribution information. Across all 3 kinds of plots, y-axis shows top contributing individual amino acids as well as bi-mers with specific lag values along with amino acid physico-chemical properties aware distance matrix. Whereas, X-axis shows model output value in decision plots, average impact on model output magnitude in summary plot and number of instances or sequences in heatmap.

A critical analysis of the feature importance for both soluble and insoluble protein classes reveals that, overall in multi-level feature extraction setting, features extracted using lag-0 and lag-1 prove most discriminative. For soluble proteins, the most influential features appear to be Schneider_lag1_YG, Schneider_0F, and Schneider_0H. These features consistently push predictions towards higher solubility values in the soluble decision plot shown in Fig. 11-A. The summary plot shown in Fig. 11-C confirms this by showing these features have the largest impact on model output for soluble proteins. Schneider_lag1_YG, in particular, stands out as the most important feature for solubility prediction, with the longest bar in the summary plot and the most dramatic shifts in the decision plot. For insoluble proteins, while the same top features (Schneider_lag1_YG, Schneider_0F, Schneider_0H) are important, their effect is reversed. In the insoluble decision plot shown in Fig. 11-B, these features often push predictions towards lower values, indicating insolubility. The summary plot (Fig. 11-C) shows that these features have significant impact for both classes, but their influence is slightly stronger for soluble proteins. Interestingly, features like Schneider_0Q and Schneider_lag1_YE show more varied effects in both decision plots, suggesting they might be important for distinguishing edge cases or specific subgroups within each class. The heatmap shown in Fig. 11-D supports this by showing these features have both positive and negative impacts across different instances.

Lower-ranked features like Grantham_lag1_PL, Schneider_0N, and Schneider_0W have smaller but still noticeable effects. They appear to fine-tune predictions after the top features have established the general trend towards solubility or insolubility. It is worth noting that many features related to amino acid pairings or bi-mers (lag1 features) are important for both classes, suggesting that local sequence patterns play a crucial role in determining protein solubility. The presence of both Schneider and Grantham indices in the top features indicates that both physico-chemical properties aware correlation information of bi-mers as well as individual amino acids distribution and bi-mers distribution information extracted at lag 0 and lag 1 significantly contribute to solubility prediction over PSI-Biology test set. Overall, while many features are important for both classes, their impact is often reversed between soluble and insoluble proteins. The model appears to rely on a complex interplay of these features, with the top features establishing a strong initial prediction and lower-ranked features providing nuanced adjustments.

2. Discussion

A thorough comparison of the proposed ProSol-Multi predictor with existing protein solubility predictors using benchmark core datasets as well as independent test sets demonstrates the superiority of the proposed predictor. The proposed predictor exhibits a decent improvement in performance on the benchmark datasets and substantial improvements on independent test sets, which indicates its strong ability to generalize well across datasets of different species In scenarios requiring analysis of large protein datasets with high sequence variability, ProSol-Multi's ability to capture comprehensive informative yet discriminative patterns provides more robust predictions compared to methods relying on simpler sequence representations. Its strong performance across different datasets also indicates potential utility in comparative proteomics studies or when working with proteins from less-studied organisms. The



(A) Soluble Class Decision Plot

(B) Insoluble Class Decision Plot

(C) Soluble and Insoluble Classes Summary Plot

(D) HeatMap

Fig. 11. Feature importance evaluation for soluble and insoluble classes using 3 different plots namely decision plots, summary plot and heatmap generated by performing SHAP analysis on test set of benchmark PSI:Biology dataset.

main factor contributing to the exceptional predictive performance as well as generalizability of the proposed ProSol-Multi predictor is the utilization of a more efficient MLCDE sequence encoder.

MLCDE encoder greatly improves upon conventional protein sequence encoding methods by capturing amino acids physico-chemical properties aware multi-level correlation and discriminative distribution in highly variable length soluble and insoluble protein sequences. Unlike simpler encoders that focus on single-level features, MLCDE extracts bi-mer correlations at multiple levels by varying lag values from 1 to 20 which allows it to capture both local and global sequence patterns that may influence solubility. MLCDE incorporates physicochemical properties of amino acids by utilizing two pre-calculated distance matrices (Grantham and Schneider), which account for characteristics such as hydrophobicity, hydrophilicity, polarity, and side chain volume. This integration allows the encoder to implicitly consider factors that affect solubility beyond simple amino acid composition. The encoder also captures distribution information at multiple levels which complements the correlation patterns and provides a richer set of discriminative features. Comprehensive features extracted by MLCDE encoder are best leveraged by the Random Forest classifier due to its ability to handle high-dimensional data and extract non-linear relationships crucial for accurate protein solubility prediction. Proposed ProSol-Multi predictor holds considerable implications for protein engineering and drug development. Its ability to accurately predict solubility across diverse datasets suggests that, it could serve as a valuable resource in the preliminary stages of protein-based drug design, and can potentially minimize the time and financial resources needed for experimental validation of candidate proteins. It can also optimize protein production workflow to generate a kind of proteins that are easy to purify at scale. It can help to prioritize enzymes with respect to their efficiency of use as detergents, food processors in different industrial applications. Furthermore, the success of the MLCDE encoder in capturing relevant solubility-related features implies that similar approaches could be employed in other protein property prediction tasks. This opens up new avenues for research in areas such as protein-protein interaction prediction, protein function annotation, and structural stability prediction. However, it is important to note that while MLCDE encoder implicitly accounts for some physicochemical properties, it may not fully capture all factors that influence protein solubility in complex biological environments. Also, MLCDE does not explicitly consider protein secondary structure, which could be a limitation in some scenarios where structural information significantly impacts solubility.

Furthermore, the physicochemical similarities among certain amino acid residues have led to the development of reduced amino acid alphabets. Reduced amino acid alphabets group similar amino acids based on their physicochemical properties, such as hydrophobicity, size, or charge, and can potentially simplify protein sequence analysis without significant loss of information. For protein solubility prediction, a reduced alphabet that emphasizes properties known to affect solubility could be particularly effective. For example, reduced alphabets that distinguish between hydrophobic, hydrophilic, charged, and neutral amino acids might capture the key factors influencing a protein's interaction with water. The composition preference of these reduced groups in soluble versus insoluble proteins could provide a simpler yet informative feature set. By applying the multi-level correlation and distribution analysis to such reduced alphabets, proposed MLCDE encoder might capture more robust patterns related to solubility. This approach could also improve the interpretability of the model, as the features would be directly linked to broader physicochemical categories rather than individual amino acids. Furthermore, analyzing the composition preference of reduced amino acid alphabets in soluble and insoluble proteins could reveal important trends. For instance, it might highlight the optimal balance of hydrophobic and hydrophilic residues for solubility, or the impact of charged residue distribution on solubility. These insights could not only improve ProSol-Multi prediction accuracy but also guide protein engineering efforts aimed at enhancing solubility. However, it's important to note that while reduced alphabets can offer advantages, they may also lose some fine-grained information. The optimal level of reduction requires precise balance between simplification and information retention for solubility prediction.

3. A user-friendly and interactive ProSol-Multi web server

We have deployed proposed ProSol-Multi predictor as an interactive and easy to use web server at https://sds_genetic_analysis.opendfki.de/ProSol-Multi. Scientific community can utilize this web server to predict the solubility of novel protein sequences belonging to different species. Researchers and practitioners can also leverage this web server for validating experimentally verified soluble proteins, training as well as optimizing predictive model from scratch, and performing prediction on novel protein sequences of new or existing species.

4. Conclusion

The paper in hand develops a robust protein solubility predictor called ProSol-Multi. ProSol-Multi makes use of a powerful encoder MLCDE to extract physico-chemical properties aware amino acids multi-level correlation and discriminative distribution in protein sequences. Prime reason behind the limited performance of existing protein solubility predictors is the use of traditional sequence encoders. These encoders only focus on individual or group of amino acids structural, physico-chemical, or occurrence characteristics. However, proposed MLCDE encoder focuses on the physico-chemical properties aware changes in the relatedness and distribution of amino acids at different distances in protein sequences. Consequently, MLCDE encoder manages to generate informative yet discriminative statistical vectors for soluble and insoluble classes, that assist ProSol-Multi predictor to achieve state-of-the-art performance on 4 different benchmark datasets. Specifically, it achieves an average accuracy increment of 3%, MCC and AU-ROC of 2%. To facilitate scientific community, an interactive ProSol-Multi web application is developed. We anticipate that ProSol-Multi predictor would prove complementary for wet-lab experiments experts, proteomics researchers, and practitioners involved in the marathon of acquiring deeper understanding of protein function and develop more effective drugs. A compelling future line of current work would be to comprehensively assess the potential of proteomics language models for learning rich and discriminative representations of protein sequences for accurate solubility prediction task. Furthermore, considering the strong influence of amino acid physicochemical properties and structural dynamics on protein solubility, the idea of incorporating reduced amino acid alphabets as well as structural information within MLCDE encoder can also be explored.

5. Limitations of the study

A limitation of this study is that it focuses solely on determining whether a protein is soluble or insoluble, without providing exact solubility values. Considering solubility values impact proteins bioavailability, stability, therapeutic efficacy, folding, and interactions, it is crucial to include the prediction of precise solubility values. Identification of soluble proteins along with exact solubility levels can guide the setup of more targeted and efficient wet-lab drug discovery experiments.

STAR * METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCE TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Data Sources
  - Model Architecture
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Model Training Optimization
  - Model Evaluation Criterion

6. Transparent methods

STAR * METHODS

KEY RESOURCE TABLE

RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
PSI:Biology Dataset	[8]	https://data.mendeley.com/preview/3ftg38trjc
E.Sol Dataset	[77]	https://data.mendeley.com/preview/3ftg38trjc
E.coli	[92]	https://data.mendeley.com/preview/3ftg38trjc
Price	[81]	https://data.mendeley.com/preview/3ftg38trjc
Software and Algorithms		
MLCDE Encoder	This Study	https://sds_genetic_analysis.opendfki.de/ProSol-Multi/
Scikit-Learn	Open Source	https://scikit-learn.org/
Python	Open Source	https://www.python.org/
Pytorch	Open Source	https://pytorch.org/
Grid Search	[6]	http://tinyurl.com/54phrd33
CD-HIT	[40]	http://weizhong-lab.ucsd.edu/cd-hit/

RESOURCE AVAILABILITY

Lead Contact For more detailed information and requests for code as well as materials must be directed to and will be completed by Lead Contact, Dr. Muhammad Nabeel Asim (Muhammad_Nabeel.Asim@dfki.de).

Materials Availability No new materials are generated in this study.

Data and Code Availability

- This study performs protein solubility prediction over existing benchmark core datasets and independent test sets provided by different researchers [8,77,92,81]. All benchmark core datasets and independent test sets used in this study are available at https://sds_genetic_analysis.opendfki.de/ProSol-Multi.
- A user-friendly web application that enables the researchers to train robust solubility predictor from scratch and use pre-trained predictor can be accessed at https://sds_genetic_analysis.opendfki.de/ProSol-Multi.
- Full source code of proposed ProSol-Multi predictor is give at https://sds_genetic_analysis.opendfki.de/ProSol-Multi.
- Further details required for re-analysis of the data presented in the current study can be obtained from the corresponding author upon request.

METHOD DETAILS

Proposed ProSol-Multi predictor workflow can be divided into 3 distinct modules, graphical illustration of which is given in Fig. 12. First module explains benchmark protein solubility prediction datasets used for sake of experimentation, essential description of

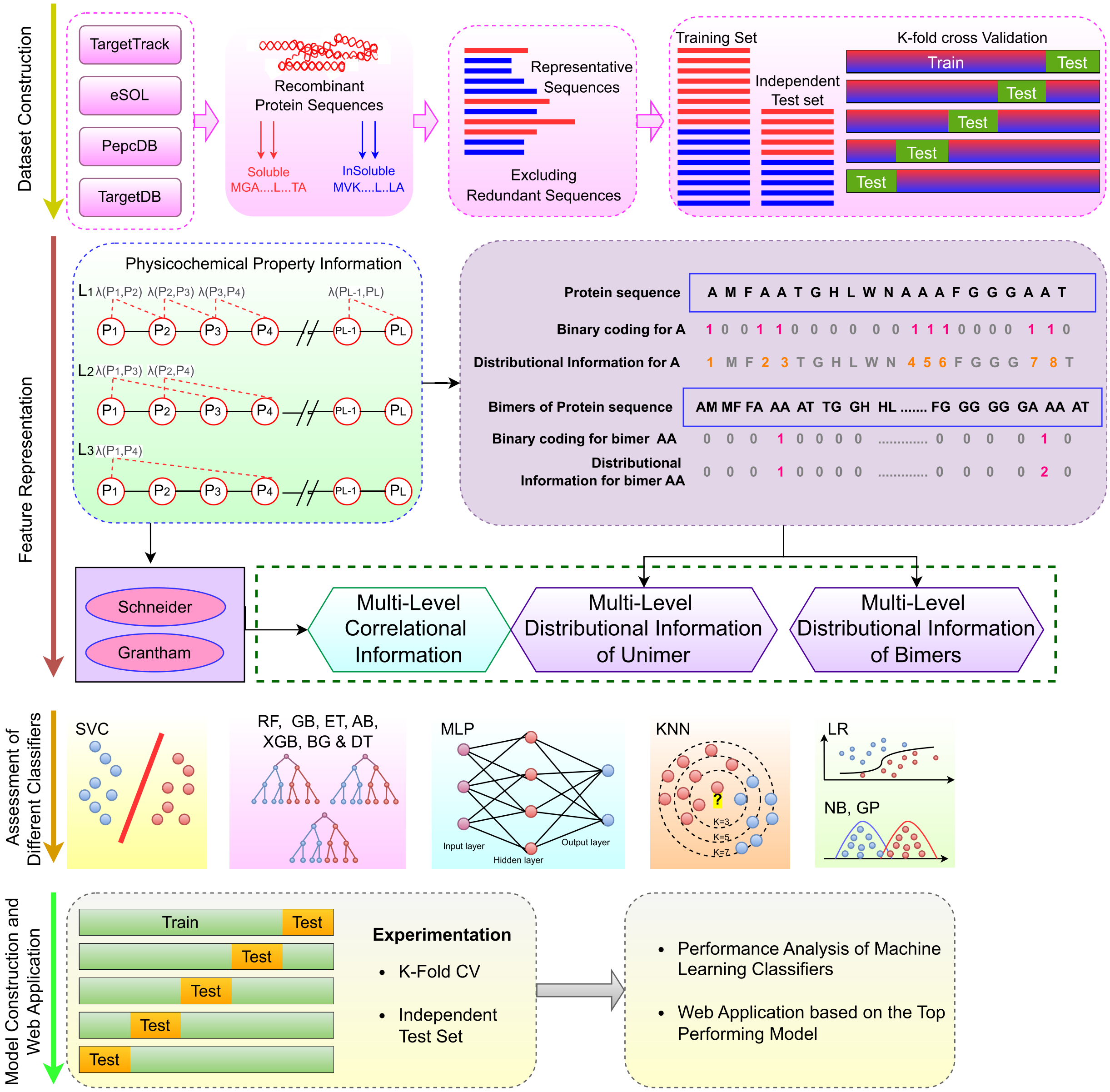

Dataset Construction
TargetTrack
eSOL
PepcDB
TargetDB
Recombinant Protein Sequences
Soluble MGA....L...TA
InSoluble MVK....L..LA
Representative Sequences
Excluding Redundant Sequences
Training Set
Independent Test set
K-fold cross Validation
Train
Test
Feature Representation
Physicochemical Property Information
L1 λ(P1,P2) λ(P2,P3) λ(P3,P4) λ(PL-1,PL)
L2 λ(P1,P3) λ(P2,P4)
L3 λ(P1,P4)
P1 P2 P3 P4 PL-1 PL
Protein sequence A M F A A T G H L W N A A A F G G G A A T
Binary coding for A 1 0 0 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0
Distributional Information for A 1 M F 2 3 T G H L W N 4 5 6 F G G G 7 8 T
Bimers of Protein sequence AM MF FA AA AT TG GH HL .... FG GG GG GA AA AT
Binary coding for bimer AA 0 0 0 1 0 0 0 0 .... 0 0 0 0 1 0
Distributional Information for bimer AA 0 0 0 1 0 0 0 0 .... 0 0 0 0 2 0
Schneider
Grantham
Multi-Level Correlational Information
Multi-Level Distributional Information of Unimer
Multi-Level Distributional Information of Bimers
Assessment of Different Classifiers
SVC
RF, GB, ET, AB, XGB, BG & DT
MLP
Input layer
Hidden layer
Output layer
KNN
K=3
K=5
K=7
LR
NB, GP
Model Construction and Web Application
Train
Test
Experimentation
K-Fold CV
Independent Test Set
Performance Analysis of Machine Learning Classifiers
Web Application based on the Top Performing Model

Fig. 12. Three fundamental modules of proposed ProSol-Multi predictor: **(A) Dataset construction:** Public benchmark datasets are prepared by collecting sequences from different databases like TargetTrack, eSol, PecpDB and target DB **(B) Feature representation MLCDE:** Generates statistical representations of protein sequences using proposed MLCDE encoder **(C) Model construction:** Performance Evaluation of 10 different machine learning classifiers.

which is given in data sources section. Second module explains proposed protein sequence encoder, essential description of which is provided in model architecture section. Third module explains classifiers along with evaluation measures, essential descriptions of which are provided in model architecture and model evaluation criterion sections respectively.

Data Sources

Benchmark Datasets

Development of robust protein solubility predictor significantly depends on quality of annotated soluble and insoluble protein sequences available in datasets. This section discusses available benchmark datasets and sheds lights on annotation quality of their inherent sequences. Up to date, researchers have developed around 23 different protein solubility prediction datasets namely rWH, TargetDB, SolPro, PROSO II, ccSOL, SolProDB, TargetTrack (3 versions), ESPRESSO, CamSol, Protein-Sol, D-eColi, D-cerevisiae, M-eColi, M-cerevisiae, PSI:Biology, eSol, Price, Camsol Mutation, E.coli, PKAD, and S.cerevisiae.

A close analysis of Supplementary File-1 indicates that, among 23 datasets, only 9 datasets are publicly available and 14 datasets namely rWH [112], TargetDB, CCSol [1], SolProDB, TargetTrack (sequences = 36,990), ESPRESSO [51], Camsol [95], Protein-Sol

[48], D-eColi, D-cerevisiae, M-eColi, M-cerevisiae, PKAD, and S.cerevisiae are not publicly available. From 9 publicly available datasets, we have skipped both TargetTrack (40,448/28,972, 1000/1001) and TargetTrack (5718/5718) datasets. Because, multiple studies have established that there exist a significant level of bias in the training and test sets of these datasets, hence these datasets are not reliable to develop protein solubility predictors [100,53]. Specifically, Hon et al. [53] compared the class labels of protein sequences of TargetTrack dataset with Price [81] dataset and found that almost 19% of class labels were totally different despite with 100% same protein sequences.

Furthermore, Supplementary File-1 reveals that publicly available protein solubility experimental datasets have been developed from 3 different databases namely TargetDB [20], PepcDB [62], and TargetTrack. For example, SolPro dataset was developed using TargetDB, and PROSO II was developed using PepcDB [7]. It is important to mention that TargetTrack merges both TargetDB and PepcDB. Considering the presence of biaseness in TargetTrack database and the fact that both SolPro and PROSO II datasets were developed more than 10 years ago. Researchers have skipped these datasets in recent studies [100]. Following the footsteps of researchers, we have also skipped these datasets due to non-reliability and mis-match of dataset statistics. For example, researchers claimed that PROSO II had 82,999 proteins, however, we only found 22,614 proteins at a different web server called ccSol omics.[3] Apart from this, camsol-mutation independent test set has been skipped by recent studies. This is because previous researchers [100] have not clearly illustrated which benchmark dataset they used for training the model before evaluating on camsol-mutation test set.

This study performs comprehensive evaluation of proposed ProSol-Multi predictor on 4 more reliable datasets namely PSI:Biology, eSol, E.coli, and Price. Considering many centers around the world have recorded the explicit expression as well as solubility labels of target proteins in TargetTrack database. With an aim to prepare a highly reliable benchmark protein solubility prediction dataset from TargetTrack database, in 2020, Bhandari et al. [8] carefully extracted a subset of 11,226 proteins from TargetTrack database with explicit labels to create PSI:Biology dataset. These labels were obtained by expressing the proteins in E.coli species. Although new techniques may change the soluble status of proteins, however, experimentally verified labels and the fact of having 66% soluble proteins make this dataset highly reliable for the development of protein solubility predictors. To date, more than 10 predictors have been evaluated on PSI:Biology. To evaluate our proposed ProSol-Multi predictor, we have used this dataset due to reliability of experimentally verified labels.

E.Sol dataset was compiled by Niwa et al. [77] by collecting 4132 proteins from the eSOL database in 2009 [77]. To ensure a more comprehensive training process and avoid overfitting or underfitting the model, they excluded 1395 sequences having a similarity value greater than 25 along with an E-value lower than $1e^{-6}$. Additionally, the authors established an independent test set consisting of 285 soluble proteins and 399 insoluble proteins. To date, around 7 predictors have been evaluated on this dataset.

E. coli was compiled by Smialowski et al. [92] by gathering 58,689 soluble proteins and 70,954 insoluble proteins from public pepcDB database.[4] To eliminate redundancy in training corpus, they employed the CD-hit tool and eliminated 60,223 sequences with a similarity exceeding 90%. Researchers have performed evaluation on this dataset using different experimental settings. However, standard split containing 90% data in training and 10% data in test set has been commonly used by most researchers [108]. To date, around 7 predictors have been evaluated on this dataset.

Price dataset was prepared by Price et al. [81] by gathering a total of 9644 protein sequences from North East Structural Genomics (NESG) center. The dataset was processed based on its usability value, which is computed by multiplying protein expression value (E) with protein solubility level (S). In order to ensure the compilation of a high-quality dataset, they removed 2,385 proteins with a usability value lower than 4. Additionally, they facilitated an independent test set consisting of 842 soluble and 481 insoluble proteins raw sequences. This dataset has around 64% soluble proteins. To date, around 10 predictors have been evaluated on this dataset. Statistics of all four benchmark datasets are given in Fig. 13 and distribution of sequence lengths across all four benchmark datasets are demonstrated in Fig. 14. It is evident in Fig. 14 that, in PSI:biology dataset, protein sequences exhibit sequence lengths ranging from 50 to 800 amino acids. In the eSol dataset, sequence lengths vary between 100 and 1300 amino acids. In Price dataset, protein sequences lengths ranges from 50 to 950 amino acids. From all four benchmark protein solubility prediction datasets, the Ecoli dataset demonstrates the highest variability in sequence lengths, spanning from 20 to 1700 amino acids.

Model Architecture

Predictive pipeline of every other protein solubility predictor involves the generation of statistical representations of raw protein sequences on the basis of amino acids distributions followed by the use of classifiers for the extraction of useful patterns. Existing predictors remain fail in extracting comprehensive discriminative distribution of amino acids due to the use of less effective sequence encoders. Proposed ProSol-Multi predictor makes use of a novel MLCDE encoder and Random Forest classifier for accurate protein solubility prediction. Novel MLCDE encoder transforms protein sequences into informative statistical vectors by capturing amino acids multi-level correlation and discriminative distribution within raw protein sequences. To better explain the novelty and worth of MLCDE encoder, we first classify and discuss the functional paradigms of 56 existing protein sequence encoders followed by description of MLCDE encoder and machine learning classifiers.

A Bird's Eye View on Existing Protein Sequence Encoders

To the best of our knowledge, 56 distinct protein sequence encoders are extensively being utilized for converting amino acid sequences into statistical vectors. These encoders can be classified into 14 different categories based on the diverse information they acquire from raw protein sequences. These categories include amino acid distribution, amino acid groups distribution, gap-

[3] http://s.tartaglialab.com/static_files/shared/tutorial_ccsol_omics.html#4.

[4] http://pepcdb.sbkb.org/.

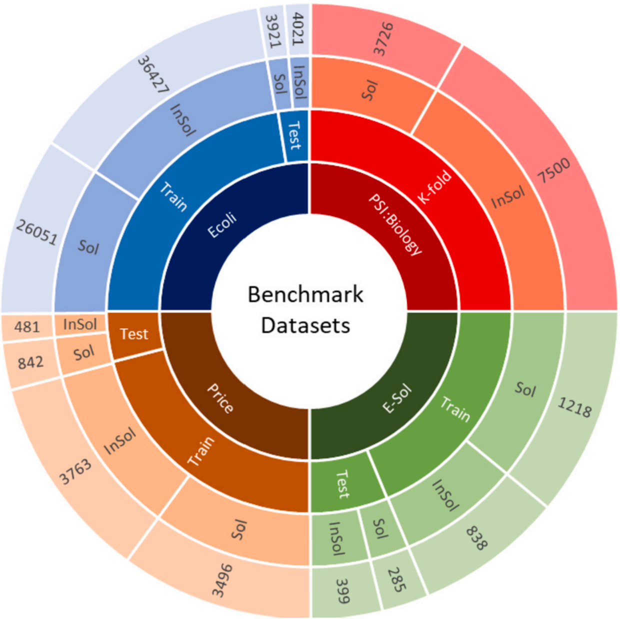


Fig. 13. Number of soluble and insoluble protein sequences in 4 different benchmark datasets.

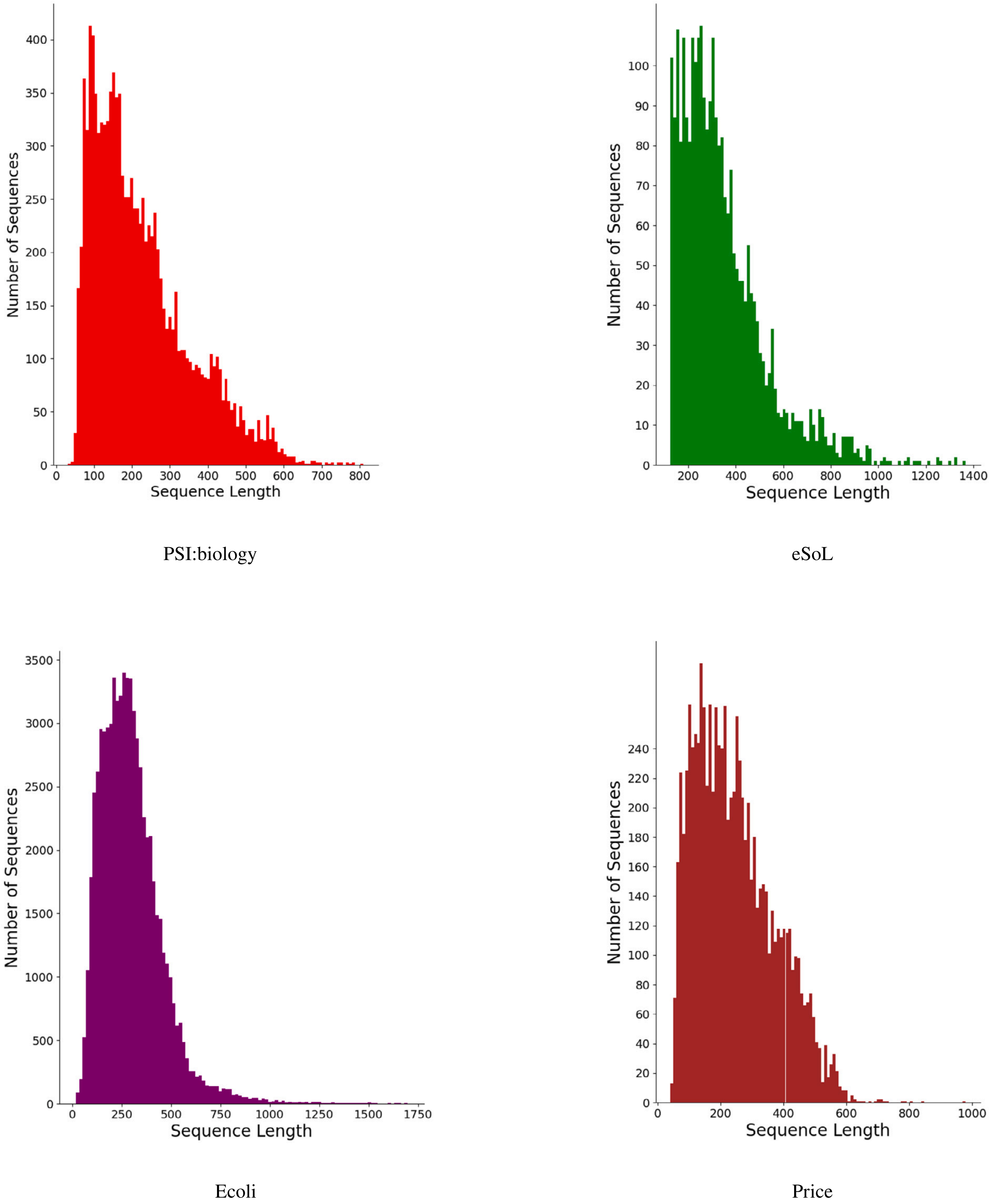
based amino acid distribution, co-variance, autocorrelation, local-global context-awareness, binary encoding, sequence order, physico-chemical properties, optimized physico-chemical properties, traditional networks, Fourier transformation-based, and substitution matrix based encoders.

Amino acid distribution encoders including Kmer [9,86], TPC [9,86], DPC [9,86], EAAC [117,14], ANF [15], DDE [86], EGAAC are most common protein sequence encoders. These encoders generate statistical vectors on the basis of frequency of individual or combination of amino acids called k-mers within protein sequences. They provide information about the overall composition of individual amino acids or k-mers within protein sequences, revealing the relative abundance or scarcity of certain amino acids or k-mers. Amino acid group distribution encoders like CTDC [12,13,33,34,44], CTDT [12,13,33,34,44], CTDD [12,13,33,34,44], GDPC [116,14], GAAC [116,14], GTPC [116,14], CTriad [90], and KSCTriad [116,14] classify the unique amino acids into distinct groups based on their particular physico-chemical characteristics including hydrophobicity, polarity, or charge. These encoders analyze the distribution of different amino acid groups within the protein sequence, offering valuable information about the overall sequence physicochemical characteristics. By organizing the amino acids in this manner, these encoders provide a comprehensive understanding of the specific groups of amino acids that significantly contribute to proteins functions.

Some encoders that utilize gap-based amino acid distribution include Adaptive skip Dipeptide composition (ASDC) [110], CKSNAP [19,21], and CKSAAGP [116,14]. These encoders divide the protein sequences into bi-mers with unique gap values and learn the distribution of distinct bi-mers. The gap value decides the distance among amino acids that form a pair, and significantly impact the kind of interactions focused by the encoders. Precisely, a smaller value of gap focuses on the extraction of short range interactions, whereas a larger value of gap focuses on the extraction of long-range interactions present within protein sequences. Autocorrelation encoders like Moran [38,69], Geary [94], and NMBroto [54] extract the similarity among amino acids or k-mers in protein sequences. These encoders calculate the correlation coefficients by considering the physico-chemical properties of the amino acids or k-mers, such as charge or hydrophobicity. By doing so, they provide valuable insights into the pairwise interactions as well as dependencies among the amino acids in protein sequences, enabling the detection of certain functional motifs. Covariance encoders provide insights into the relationship between two amino acids or k-mers by indicating how they vary together. Unlike autocorrelation encoders that estimate both the strength as well as direction of the relationship among two amino acids or k-mers, covariance encoders solely indicate the direction of the relationship without specifying its strength.

Protein sequences encoders that have a local-global context awareness include WSRC-global [75], WSRC-local [75], and WSRC-local-global [75], provide valuable information about the distribution and changes of amino acids in various segments of protein sequences. These encoders take into account the composition and transition of amino acids, offering insights into the distribution and arrangement of these building blocks within the protein sequence. Another category of encoders, called sequence order encoders include APAAC [30], PAAC [29], SOCNumber [89,27,31], CTAPAAC [74], and QSOrder [89,27,31]. These encoders consider not only the distribution but also the order or arrangement of amino acids in protein sequences based on different distances. Different distances reflect various levels of global or local interactions among certainly arranged unique amino acids. A different approach is taken by binary encoders [68,70,110,16,23,111]. These encoders convert protein sequences into statistical vectors having 1s and 0s on the basis of presence and absence of individual amino acids or k-mers.

Physico-chemical characteristics as well as pre-trained neural network based encoders like AAIndex [103] as well as AESNN3 [68,70] respectively, replace amino acids with pre-calculated floating-point values. Optimized physico-chemical characteristics based



PSI:biology

eSoL

Ecoli

Price

Fig. 14. Analysis of sequence lengths across 4 different benchmark datasets.

encoders like ZScale [24], uses physico-chemical characteristics to describe the amino acids. It employs various techniques like Partial least squares (PLS), principal component analysis (PCA), and Multiple Linear regression to eradicate less informative characteristics and retain only the most informative ones. On the other hand, traditional network-based encoders like complex networks or enhanced complex networks represent the protein sequences as graphs where the nodes correspond to amino acids and the edges represent interactions or relationships among them. Furthermore, Fourier transformation based encoders like MappingClass-eiip-fourier or

MappingClass-integer-fourier make use of Electron-Ion Interaction Potential floating point values or integer values to substitute the unique amino acids. These encoders try to improve the encoding of comprehensive hidden patterns as well as trends such as common components inside protein sequences by applying Fourier transformation. Whereas, substitution matrix-based encoders such as BLOSUM62 [67] develop matrices that primarily assign different scores to amino acids on the basis of their frequencies in different yet related protein sequences. This scoring paradigm facilitates a way of computing the similarity among distinct amino acids.

Proposed Protein Sequence Encoder

Proposed Multi-Level Correlation and Discriminative Distribution Encoder (MLCDE) extends the functional paradigm of Quasi Sequence Order (QSO) [28] encoder. QSO encoder converts protein sequences into statistical vectors by extracting two kinds of information: correlation and distribution. QSO encoders first generates bi-mers based on different Lags, where every Lag value allows the QSO encoder to focus on specific amino acid. For example, with Lag 1, immediate amino acids are combined for bi-mers generation. With Lag 2, amino acid available at every second position with respect to the position of current amino acid is used for bi-mers generation, and so on. QSO encoder extracts correlations on the basis of average physico-chemical properties based distance between amino acids involved in unique bi-mers generated using different Lags. It captures distribution information on the basis of occurrence frequency of individual amino acids inside protein sequences. Although QSO encoder extracts bi-mers correlation information at different levels. However, it does not capture distribution information at different levels. This limitation over-simplify both soluble and insoluble protein sequences, and makes their discrimination difficult even for sophisticated machine or deep learning classifiers. Hence, an extensive set of informative and discriminative patterns which enable even simple machine learning classifier to more accurately distinguish soluble proteins from insoluble proteins can not be acquired by traditionally extracting distribution information at single level. Therefore, to encode comprehensive discriminative patterns that can distinguish soluble protein sequences from insoluble protein sequences, proposed MLCDE encoder thoroughly captures bi-mers multi-level correlation and distribution information.

The working paradigm of proposed MLCDE encoder can be divided into 3 major steps: **I** Extraction of multi-level bi-mer correlation along with individual amino acid distribution, **II** Extraction of multi-level bi-mer distribution, **III** Concatenation of both kinds of features. To gain a more precise understanding of functional paradigm of proposed MLCDE encoder, let's take a generic sequence $S = AA_1, AA_2, ...AA_n$, where each AA_i denotes a specific amino acid from a set of 20 distinct amino acids. In the first step, proposed encoder iteratively skips 1-to-20 amino acids by varying Lag value from 1-to-20 to prepare 20 different chains of bi-mers for the given sequence S, shown in Equation (1). While chain of bi-mers based on smaller lag values (e.g., 1 or 2) enable the extraction of relatedness among adjacent or nearby amino acids, chain of bi-mers based on larger lag values (e.g., 10 or 20) enable the extraction of relatedness between faraway amino acids in the given sequence S.

$$\begin{cases} AA_1AA_2, AA_2AA_3, AA_3AA_4, \dots AA_{L-1}AA_L \text{ with Lag 1} \\ AA_1AA_3, AA_2AA_4, AA_3AA_5, \dots AA_{L-2}AA_L \text{ with Lag 2} \\ AA_1AA_4, AA_2AA_5, AA_3AA_6, \dots AA_{L-3}AA_L \text{ with Lag 3} \\ AA_1AA_5, AA_2AA_6, AA_3AA_7, \dots AA_{L-4}AA_L \text{ with Lag 4} \\ AA_1AA_6, AA_2AA_7, AA_3AA_8, \dots AA_{L-5}AA_L \text{ with Lag 5} \\ \vdots \; \vdots \\ AA_1AA_{21}, AA_2AA_{22}, AA_3AA_{23}, \dots AA_{L-20}AA_L \text{ with Lag 20} \end{cases} \tag{1}$$

Across each chain of bi-mers, coupling factor that precisely captures the distance among two amino acids is calculated for each bi-mer to acquire correlation information. Following the footsteps of Chou et al. [28], we make use of 2 pre-calculated amino acid distance matrices (dimensions of 20 × 20 = 400) facilitated by Grantham et al., [42] and Schneider et al., [88]. The values of four distinct physico-chemical properties, namely hydrophobicity, hydrophilicity, polarity, and side chain volume are averaged in these matrices using the Manhattan distance. The calculation of the coupling factor for each $bi - mer_q$ based on two amino acids AA_y and AA_z using Grantham and Schneider amino acid distance matrices can be mathematically written as:

$$\begin{gathered} Coupling\ Factor[bimer_q] = Distance\ Matrix_i\ (AA_y, AA_z) \\ Distance\ Matrix_i \in \{Schneider, Grantham\} \end{gathered} \tag{2}$$

Afterwards, for every bi-mers chain based on specific Lag value, the correlational values of inherent bi-mers are added together and then divided by sequence length in order to obtain normalized correlation features. Since the correlational values are obtained from two distinct amino acid distance matrices, a total of 20 normalized correlation values are obtained for 20 bi-mers chains using the distance matrix of Schneider et al. [88] and another 20 normalized correlation values are obtained using distance matrix of Grantham et al. [42] through Equation (3).

$$\begin{aligned} &\text{Normalized Correlation}[Distance\ Matrix_i][Lag_k] \\ &= \frac{\sum_{q=1}^{Sequence\ Length-1}(Coupling\ Factor[bimer]_q)}{Sequence\ Length - 1}. \end{aligned} \tag{3}$$

In above expression, Lag_k represents a particular bi-mers chain having a specific length. Subsequently, we calculate two distinct values by separately summing up 20 floating point values based on the Schneider matrix and 20 floating point values based on

Grantham matrix, through Equation (4). By doing so, we are able to measure the overall correlation between the bi-mers at 20 separate Lag values by employing two distinct average values of four different physico-chemical properties.

$$\text{Overall Correlation}[Distance\ Matrix_i] = \sum_{n=1}^{20} \text{Normalized Correlation}[Distance\ Matrix_i][Lag_k]. \tag{4}$$

Afterwards, we optimize the estimated normalized correlation values by applying a weight factor of 0.25 through Equation (5).

$$\text{Optimized Normalized Correlation}[Distance\ Matrix_i][Lag_k] = \frac{weight \times \text{Normalized Correlation}[Distance\ Matrix_i][Lag_k]}{1 + weight \times \text{Overall Correlation}[Distance\ Matrix_i]}. \tag{5}$$

Proposed MLCDE encoder combines multi-level bi-mers correlational features with distributional features of individual amino acids. In order to accomplish this, it computes the occurrence frequency at which each amino acid appears within a sequence using Equation (6), and then adjust the resulting value by normalizing it with the overall correlation values of the sequence, which are computed using Equation (4).

$$\text{Amino Acid Distribution}[AA_y] = \frac{AA_y \text{ occurrence frequency in sequence}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{6}$$

In second step, with an aim to capture comprehensive discriminative patterns in protein sequences, proposed MLCDE encoder captures bi-mers multi-level distribution to complement bi-mers multi-level correlation patterns. To accomplish this, proposed MLCDE encoder first computes occurrence frequency of bi-mers in every unique chain of bi-mers. Then, it separately divides the bi-mers occurrence frequency with overall correlation values of the given sequence S based on two different distance matrices to obtain 20 floating point values with respect to Schneider distance matrix and 20 floating point values with respect to Grantham distance matrix respectively using Equation (7).

$$\text{Multi-Level Bi-mers Distribution}[AA_y, AA_z] = \frac{AA_y, AA_z \text{ occurrence frequency in sequence } [Lag_k]}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{7}$$

Finally, in third step, it concatenates the multi-level bi-mers correlational, individual amino acid distributional, and bi-mers multi-level distributional features to generate discriminative statistical vectors for soluble and insoluble protein sequences using equation (8).

$$\text{MLCDE} = \begin{cases} \begin{matrix} Optimized Normalized \\ Correlation[DistanceMatrix_i][Lag_k] \end{matrix} \oplus \\ \text{Amino Acid Distribution}[AA_y] \oplus \\ \text{Multi-Level Bi-mers Distribution}[AA_y, AA_z] \end{cases} \tag{8}$$

Machine Learning Classifiers

We assess the performance impact of proposed protein sequence encoder MLCDE using 10 most widely used machine learning classifiers of 4 different categories [73]. Specifically, 6 tree-based, 1 generative, 2 discriminative, and 1 nearest neighbor based classifiers. From generative classifiers, we have used Naive Bayes (NB) [109]. NB assumes that all features of sequences are independent of each other. It uses Bayes theorem to compute probability of every class and conditional probability of every feature given the class during the training. During inference, it computes the probability of given sequence belonging to every class and predicts the class with highest probability.

From tree-based classifiers, we have used Decision tree (DT) [56,60], Random Forest (RF) [60], Extra tree(ET) [60], AdaBoost (AB) [60], Gradient Boost (GB) [60], and Extreme Gradient Boost (XGB) [60]. DT recursively splits the given sequences on the basis of selected features to build a tree that effectively separates the unique classes. To determine the optimal split, DT makes use of information gain or Gini impurity criteria [56,96]. RF classifier is an ensemble classifier based on bagging and random feature selection. It constructs a series of base classifiers by randomly sampling subsets of training sequences using copy with replacement method. Every base classifier is independently trained on random features selected from all features. During inference, RF predicts the class that receives majority votes. ET classifier goes one step further to RF classifier. It randomizes the creation of base classifiers by choosing arbitrary cut-points rather than choosing optimal cut-points to split the nodes. This randomization significantly enhances the diversity between base classifiers which may improve model's accuracy. During inference, ET classifier also considers majority voting to predict final class for a given sequence.

AB classifier [60] is also an ensemble classifier which combines several weak classifiers to construct a strong classifier. The underlay boosting algorithm operates iteratively where every subsequent classifier is trained to fix the errors of previous classifier. GB classifier also works similar to AB classifier, however the major difference is that GB makes use of gradient descent optimization to learn the optimal parameters of model at every iteration. XGB [60] classifier is also a boosting classifier and works similar to GB classifier. The key difference is that, unlike GB classifier, it makes use of L1 and L2 regularization strategies to avoid over-fitting, and uses tree-pruning to discard redundant branches of tree. K-Nearest Neighbors (KNN) [97] is the only nearest neighbor based classifier used in the study. KNN is a simple classifier and it does not require any training. It makes prediction by applying majority voting on the class of k nearest neighbors, computed through certain distance metric such as Euclidean distance. From discrimnative category, we have used Support Vector Machine (SVC or SVM) [11] and Logistic Regression (LR) [76] classifiers. SVM classifier tries to find the best hyperplane to separate sequences into different classes, LR classifier models the relationship among features and the likelihood of belonging to a specific class.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model Training Optimization

Proposed MLCDE encoder is implemented in Python and 56 existing protein sequence encoders are taken from iLearnPlus [22]. Machine learning classifiers implementations are taken from Scikit-Learn [79] library. Additionally, we have implemented the ProSol-Multi predictor web server using the Django framework [39]. To optimize the performance of both encoders and classifiers, we have experimented with different values of hyperparameters. Through hyperparameter optimization strategy called Grid search, we have searched for the best hyperparameter values within a specified range using the training data. To facilitate readers, existing encoders hyperparameter search space and optimal values across 10 different classifiers are mentioned in Supplementary File-2, Table-1. Moreover, initial and optimal hyperparameters values of 10 different classifiers for all 4 public benchmark datasets are given in Supplementary File-2, Table-2. To perform an unbiased performance comparison of proposed ProSol-Multi predictor with existing 20 predictors, following the evaluation criterion of existing predictors, we perform 5-fold cross validation on benchmark PSI:biology, and Price datasets. Whereas, E.Coli dataset is divided into training and test sets using standard 90/10 split and performance comparison is made using 10% test data. Furthermore, we compare the performance of proposed ProSol-Multi predictor with existing predictors on Price and eSol independent test sets.

Model Evaluation Criterion

To conduct a fair performance comparison of the proposed ProSol-Multi predictor with existing predictors, we use the same 7 evaluation measures that have been used extensively in previous studies [74,52]. These evaluation measures include Accuracy (ACC), Sensitivity (SEN) or Recall (REC), Precision (PRE), Specificity (SPE), Matthews correlation coefficient (MCC), F1-score, and area under the receiver operating characteristic (AU-ROC) [74]. Considering, these evaluation metrics have been comprehensively discussed in previous studies [74,52], here, we only facilitate their mathematical expressions.

$$
f(x) = \begin{cases}
\text{Accuracy (ACC)} = (T_P + T_N)/(T_P + T_N + F_P + F_N) \\
\text{Precision (PRE)} = T_P/(T_P + F_P) \\
\text{Recall (REC) or Sensitivity (SEN)} = T_P/(T_P + F_N) \\
\text{Specificity (SPE)} = T_N/(T_N + F_P) \\
\text{F1-Score} = 2 \times \dfrac{PRE \times REC}{PRE + REC} \\
\text{MCC} = T_P \times T_N - F_P \times F_N/Q \\
\text{Q} = \sqrt{(T_P + F_N)(T_P + F_P)(T_N + F_P)(T_N + F_N)}
\end{cases} \tag{9}
$$

In equation (9), variables TN and TP represent the count of correctly predicted insoluble and soluble protein sequences respectively. Conversely, FN and FP represent the count of incorrectly predicted insoluble and soluble protein sequences respectively. It is worth noting that the classifier's performance improves as the values of these seven evaluation metrics increase. It is worth noting that classifier's performance improves as the values of these seven evaluation metrics increase.

CRediT authorship contribution statement

Hina Ghafoor: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad Nabeel Asim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad Ali Ibrahim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andreas Dengel:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that no generative AI tool is used for scientific writing, editing, and reviewing.

Acknowledgements

This research has received no funding from any public or private organization.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36041.

References

[1] Federico Agostini, Davide Cirillo, Carmen Maria Livi, Riccardo Delli Ponti, Gian Gaetano Tartaglia, ccsol omics: a webserver for large-scale prediction of endogenous and heterologous solubility in E. coli, Bioinformatics 30 (20) (2014) 2975–2977.

[2] Federico Agostini, Michele Vendruscolo, Gian Gaetano Tartaglia, Sequence-based prediction of protein solubility, J. Mol. Biol. 421 (2–3) (2012) 237–241.

[3] Shahid Akbar, Farman Ali, Maqsood Hayat, Ashfaq Ahmad, Salman Khan, Sarah Gul, Prediction of antiviral peptides using transform evolutionary & shap analysis based descriptors by incorporation with ensemble learning strategy, Chemom. Intell. Lab. Syst. 230 (2022) 104682.

[4] Shahid Akbar, Ali Raza, Quan Zou, Deepstacked-avps: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model, BMC Bioinform. 25 (1) (2024) 102.

[5] Shahid Akbar, Quan Zou, Ali Raza, Fawaz Khaled Alarfaj, iafps-mv-bitcn: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks, Artif. Intell. Med. 151 (2024) 102860.

[6] James Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl, Algorithms for hyper-parameter optimization, Adv. Neural Inf. Process. Syst. 24 (2011).

[7] Helen M. Berman, John D. Westbrook, Margaret J. Gabanyi, Wendy Tao, Raship Shah, Andrei Kouranov, Torsten Schwede, Konstantin Arnold, Florian Kiefer, Lorenza Bordoli, et al., The protein structure initiative structural genomics knowledgebase, Nucleic Acids Res. 37 (suppl_1) (2009), D365–D368.

[8] Bikash K. Bhandari, Paul P. Gardner, Chun Shen Lim, Solubility-weighted index: fast and accurate prediction of protein solubility, Bioinformatics 36 (18) (2020) 4691–4698.

[9] Gajendra P.S. Raghava, Manoj Bhasin, Classification of nuclear receptors based on amino acid composition and dipeptide composition, J. Biol. Chem. 279 (22) (2004) 23262–23266.

[10] Jordan W. Bye, Lauren Platts, Robert J. Falconer, Biopharmaceutical liquid formulation: a review of the science of protein stability and solubility in aqueous environments, Biotechnol. Lett. 36 (2014) 869–875.

[11] Evgeny Byvatov, Gisbert Schneider, Support vector machine applications in bioinformatics, Appl. Bioinform. 2 (2) (2003) 67–77.

[12] Lianyi Han, Zhiliang Ji, Chen Xiang, Chen Yu Zong, C.Z. Cai, Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (13) (2003) 3692–3697.

[13] Lianyi Han, Zhiliang Ji, Chen Yu Zong, C.Z. Cai, Enzyme family classification by support vector machines, Proteins 55 (1) (2004) 66–76.

[14] Huaming Chen, Fuyi Li, Wang Lei, Yaochu Jin, Chi-Hung Chi, Lukasz Kurgan, Jiangning Song, Jun Shen, Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions, Brief. Bioinform. 22 (3) (2020), 1–NA.

[15] Hong Tran, Zhi-Yong Liang, Hao Lin, Liqing Zhang, Chen Wei, Identification and analysis of the n(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome, Sci. Rep. 5 (1) (2015) 13859.

[16] Xiang Chen, Jian-Ding Qiu, Shao-Ping Shi, Sheng Bao Suo, Shu-Yun Huang, Ru-Ping Liang, Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites, Bioinformatics (Oxford, England) 29 (13) (2013) 1614–1622.

[17] Jianwen Chen, Shuangjia Zheng, Huiying Zhao, Yuedong Yang, Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map, J. Cheminform. 13 (1) (2021) 1–10.

[18] Long Chen, Rining Wu, Feixiang Zhou, Huifeng Zhang, Jian K. Liu, Hybridgcn for protein solubility prediction with adaptive weighting of multiple features, J. Cheminform. 15 (1) (2023) 118.

[19] Ke Chen, Lukasz Kurgan, Jishou Ruan, Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs, BMC Struct. Biol. 7 (1) (2007) 25.

[20] Li Chen, Rose Oughtred, Helen M. Berman, John D. Westbrook, TargetDB: a target registration database for structural genomics projects, Bioinformatics (Oxford, England) 20 (16) (2004) 2860–2862.

[21] Ke Chen, Yingfu Jiang, Li Du, Lukasz Kurgan, Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs, J. Comput. Chem. 30 (1) (2008) 163–172.

[22] Zhen Chen, Pei Zhao, Chen Li, Fuyi Li, Dongxu Xiang, Yong-Zi Chen, Tatsuya Akutsu, Roger J. Daly, Geoffrey I. Webb, Quanzhi Zhao, et al., ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization, Nucleic Acids Res. 49 (10) (2021) e60.

[23] Zhen Chen, Yuan Zhou, Jiangning Song, Ziding Zhang, hcksaap_ubsite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties, Biochim. Biophys. Acta, Proteins Proteomics 1834 (8) (2013) 1461–1467.

[24] Yong-Zi Chen, Zhen Chen, Yu-Ai Gong, Ying Guoguang, SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties, PLoS ONE 7 (6) (2012) e39195.

[25] Fabrizio Chiti, Christopher M. Dobson, Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade, Annu. Rev. Biochem. 86 (1) (2017) 27–68.

[26] Minee L. Choi, Sonia Gandhi, Crucial role of protein oligomerization in the pathogenesis of Alzheimer's and Parkinson's diseases, FEBS J. 285 (19) (2018) 3631–3644.

[27] Kuo-Chen Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2) (2000) 477–483.

[28] Kuo-Chen Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2) (2000) 477–483.

[29] Kuo-Chen Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (3) (2001) 246–255.

[30] Kuo-Chen Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics (Oxford, England) 21 (1) (2004) 10–19.

[31] Kuo-Chen Chou, Yu-Dong Cai, Prediction of protein subcellular locations by GO–FunD–ZPseAA predictor, Biochem. Biophys. Res. Commun. 320 (4) (2004) 1236–1239.

[32] Zacharias Dische, A new specific color reaction of hexuronic acids, J. Biol. Chem. 167 (1) (1947) 189–198.

[33] Inna Dubchak, Ilya Muchnik, Stephen R. Holbrook, Sung-Hou Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. USA 92 (19) (1995) 8700–8704.

[34] Inna Dubchak, Muchnik Ilya, ChristopherMayor, Igor Dralyuk, Sung-Hou Kim, Recognition of a protein fold in the context of the scop classification, Proteins 35 (4) (1999) 401–407.

[35] F.U. Ellis, R.J. Hartl, Principles of protein folding in the cellular environment, Curr. Opin. Struct. Biol. 9 (1) (1999) 102–110.

[36] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, Timothy L. Kline, Machine learning for medical imaging, Radiographics 37 (2) (2017) 505.

[37] Beatrix Fahnert, Hauke Lilie, Peter Neubauer, Inclusion bodies: formation and utilisation, Physiol. Stress Resp. Bioprocess. (2004) 93–142.

[38] Zhi-Ping Feng, Chun-Ting Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, J. Protein Chem. 19 (4) (2000) 269–275.

[39] Jeff Forcier, Paul Bissex, Wesley J. Chun, Python Web Development with Django, Addison-Wesley Professional, 2008.

[40] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, Weizhong Li, Cd-hit: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.

[41] Bertrand Garcia-Moreno, Adaptations of proteins to cellular and subcellular ph, J. Biol. 8 (11) (2009) 98.

[42] Richard Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (4154) (1974) 862–864.

[43] Siti Hajar Abdul Hamid, Fathurrahman Lananan, Helena Khatoon, Ahmad Jusoh, Azizah Endut, A study of coagulating protein of Moringa oleifera in microalgae bio-flocculation, Int. Biodeterior. Biodegrad. 113 (2016) 310–317.

[44] Lianyi Han, C.Z. Cai, Siew Lin Lo, Maxey C.M. Chung, Yu Zong Chen, Prediction of RNA-binding proteins from primary sequence by a support vector machine approach, RNA (New York, N. Y.) 10 (3) (2004) 355–368.

[45] Xi Han, Wenbo Ning, Xiaoqiang Ma, Xiaonan Wang, Kang Zhou, Improving protein solubility and activity by introducing small peptide tags designed with machine learning models, Metabolic Eng. Commun. 11 (2020) e00138.

[46] Xi Han, Liheng Zhang, Kang Zhou, Xiaonan Wang, Progan: protein solubility generative adversarial nets for data augmentation in dnn framework, Comput. Chem. Eng. 131 (2019) 106533.

[47] David J. Hauss, Oral Lipid-Based Formulations: Enhancing the Bioavailability of Poorly Water-Soluble Drugs, vol. 170, CRC Press, 2007.

[48] Max Hebditch, M. Alejandro Carballo-Amador, Spyros Charonis, Robin Curtis, Jim Warwicker, Protein–Sol: a web tool for predicting protein solubility from sequence, Bioinformatics 33 (19) (2017) 3098–3100.

[49] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, Burkhard Rost, Modeling aspects of the language of life through transfer-learning protein sequences, BMC Bioinform. 20 (1) (2019) 1–17.

[50] John R. Hepler, Alfred G. Gilman, G proteins, Trends Biochem. Sci. 17 (10) (1992) 383–387.

[51] Tamotsu Hirose, Shuichi Noguchi, Espresso: a system for estimating protein expression and solubility in protein expression systems, Proteomics 13 (9) (2013) 1444–1456.

[52] Jiri Hon, Martin Marusiak, Tomas Martinek, Antonin Kunka, Jaroslav Zendulka, David Bednar, Jiri Damborsky, Soluprot: prediction of soluble protein expression in Escherichia coli, Bioinformatics 37 (1) (2021) 23–28.

[53] Jiri Hon, Martin Marusiak, Tomask Martinek, Antonin Kunka, Jaroslav Zendulka, David Bednar, Jiri Damborsky, SoluProt: prediction of soluble protein expression in Escherichia coli, Bioinformatics (Oxford, England) 37 (1) (2021) 23–28.

[54] David S. Horne, Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities, Biopolymers 27 (3) (1988) 451–477.

[55] Qingzhen Hou, Jean Marc Kwasigroch, Marianne Rooman, Fabrizio Pucci, Solart: a structure-based method to predict protein solubility and aggregation, Bioinformatics 36 (5) (2020) 1445–1452.

[56] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning, vol. 112, Springer, 2013.

[57] Lorraine V. Kalia, Suneil K. Kalia, Pamela J. McLean, Andres M. Lozano, Anthony E. Lang, α-synuclein oligomers and clinical implications for Parkinson disease, Ann. Neurol. 73 (2) (2013) 155–169.

[58] Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, Raghvendra Mall, Deepsol: a deep learning framework for sequence-based protein solubility prediction, Bioinformatics 34 (15) (2018) 2605–2613.

[59] Ron R. Kopito, Aggresomes, inclusion bodies and protein aggregation, Trends Cell Biol. 10 (12) (2000) 524–530.

[60] Vandana Korde, C. Namrata Mahender, Text classification and classifiers: a survey, Int. J. Artif. Intell. Appl. 3 (2) (2012) 85.

[61] Hannu Korhonen, Anne Pihlanto-Leppäla, Pirjo Rantamäki, Tuomo Tupasela, Impact of processing on bioactive proteins and peptides, Trends Food Sci. Technol. 9 (8–9) (1998) 307–319.

[62] Andrei Kouranov, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E. Bourne, Helen M. Berman, The RCSB PDB information portal for structural genomics, Nucleic Acids Res. 34 (2006) (Database issue) D302–5.

[63] Renate Kunert, David Reinhart, Advances in recombinant antibody manufacturing, Appl. Microbiol. Biotechnol. 100 (8) (2016) 3451–3461.

[64] Aleksander Kuriata, Aleksandra E. Badaczewska-Dawid, Jordi Pujols, Salvador Ventura, Sebastian Kmiecik, Protocols for rational design of protein solubility and aggregation properties using aggrescan3d standalone, in: Computer Simulations of Aggregation of Proteins and Peptides, Springer, 2022, pp. 17–40.

[65] Michael R. Ladisch, Karen L. Kohlmann, Recombinant human insulin, Biotechnol. Prog. 8 (6) (1992) 469–478.

[66] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A. Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez, et al., Machine learning in bioinformatics, Brief. Bioinform. 7 (1) (2006) 86–112.

[67] Tzong-Yi Lee, Shu-An Chen, Hsin-Yi Hung, Yu-Yen Ou, Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites, PLoS ONE 6 (3) (2011) e17331.

[68] Kuang Lin, Alex C.W. May, William R. Taylor, Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types, J. Theor. Biol. 216 (3) (2002) 361–365.

[69] Zong Lin, Xian-Ming Pan, Accurate prediction of protein secondary structural content, J. Protein Chem. 20 (3) (2001) 217–220.

[70] Bin Liu, Xin Gao, Hanyu Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, Nucleic Acids Res. 47 (20) (2019) e127.

[71] Christophe N. Magnan, Arlo Randall, Pierre Baldi, SOLpro: accurate sequence-based prediction of protein solubility, Bioinformatics 25 (17) (2009) 2200–2207.

[72] Mark C. Manning, Danny K. Chou, Brian Murphy, Robert W. Payne, Derrick S. Katayama, Stability of protein pharmaceuticals: an update, Pharm. Res. 27 (4) (2010) 544–575.

[73] Faheem Masoodi, Mohammad Quasim, Syed Bukhari, Sarvottam Dixit, Shadab Alam, Applications of Machine Learning and Deep Learning on Biological Data, CRC Press, 2023.

[74] Faiza Mehmood, Shazia Arshad, Muhammad Shoaib, RPPSP: a robust and precise protein solubility predictor by utilizing novel protein sequence encoder, IEEE Access 11 (A) (2023) 59397–59416.

[75] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Ahtisham Fazeel, Andreas Dengel, Sheraz Ahmed, DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method, Brief. Bioinform. 24 (1) (2023), bbac546.

[76] Andrew Ng, Michael Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes, Adv. Neural Inf. Process. Syst. 14 (2001).

[77] Tatsuya Niwa, Bei-Wen Ying, Katsuyo Saito, WenZhen Jin, Shoji Takada, Takuya Ueda, Hideki Taguchi, 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.

[78] Marc Oeller, Ryan Kang, Rosie Bell, Hannes Ausserwöger, Pietro Sormanni, Michele Vendruscolo, Sequence-based prediction of ph-dependent protein solubility using camsol, Brief. Bioinform. 24 (2) (2023), bbad004.

[79] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[80] Marcus Pickhardt, Carmen Lawatscheck, Hans G. Börner, Eckhard Mandelkow, Inhibition of tau protein aggregation by rhodanine-based compounds solubilized via specific formulation additives to improve bioavailability and cell viability, Curr. Alzheimer Res. 14 (7) (2017) 742–752.

[81] W. Nicholson Price, Samuel K. Handelman, John K. Everett, Saichiu N. Tong, Ana Bracic, Jon D. Luff, Victor Naumov, Thomas Acton, Philip Manor, Rong Xiao, et al., Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in e. coli, Microbial Inf. Experiment. 1 (1) (2011) 1–20.

[82] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, Yun Song, Evaluating protein transfer learning with tape, Adv. Neural Inf. Process. Syst. 32 (2019).

[83] Reda Rawi, Raghvendra Mall, Khalid Kunji, Chen-Hsiang Shen, Peter D. Kwong, Gwo-Yu Chuang, PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine, Bioinformatics 34 (7) (2018) 1092–1098.

[84] Ali Raza, Jamal Uddin, Abdullah Almuhaimeed, Shahid Akbar, Quan Zou, Ashfaq Ahmad, Aips-sntcn: predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks, J. Chem. Inf. Model. 63 (21) (2023) 6537–6554.

[85] Patricia S. Regojo, Burn care basics: how to extinguish problems, Nursing2022 33 (3) (2003) 50–53.

[86] Vijayakumar Saravanan, Namasivayam Gautham, Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor, OMICS J. Integr. Biol. 19 (10) (2015) 648–658.

[87] Catherine H. Schein, Solubility and secretability, Curr. Opin. Biotechnol. 4 (4) (1993) 456–461.

[88] Gisbert Schneider, Paul Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (2) (1994) 335–344.

[89] Gisbert Schneider, Paul Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (2) (1994) 335–344.

[90] Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, Hualiang Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. USA 104 (11) (2007) 4337–4341.

[91] Tessa Sinnige, Karen Stroobants, Christopher M. Dobson, Michele Vendruscolo, Biophysical studies of protein misfolding and aggregation in in vivo models of Alzheimer's and Parkinson's diseases, Q. Rev. Biophys. 53 (2020) e10.

[92] Pawel Smialowski, Gero Doose, Phillipp Torkler, Stefanie Kaufmann, Dmitrij Frishman, Proso ii–a new method for protein solubility prediction, FEBS J. 279 (12) (2012) 2192–2200.

[93] Pawel Smialowski, Antonio J. Martin-Galiano, Aleksandra Mikolajka, Tobias Girschick, Tad A. Holak, Dmitrij Frishman, Protein solubility: sequence based prediction and experimental verification, Bioinformatics 23 (19) (2007) 2536–2542.

[94] Robert R. Sokal, Barbara A. Thomson, Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population, Am. J. Phys. Anthropol. 129 (1) (2005) 121–131.

[95] Pietro Sormanni, Francesco A. Aprile, Michele Vendruscolo, The camsol method of rational design of protein mutants with enhanced solubility, J. Mol. Biol. 427 (2) (2015) 478–490.

[96] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, Reading, MA, USA, 2005.

[97] Songbo Tan, An effective refinement strategy for knn text classifier, Expert Syst. Appl. 30 (2) (2006) 290–298.

[98] Ammar Tareen, Justin B. Kinney, Logomaker: beautiful sequence logos in python, Bioinformatics 36 (7) (2020) 2272–2274.

[99] Vineet Thumuluri, Hannah-Marie Martiny, Jose J. Almagro Armenteros, Jesper Salomon, Henrik Nielsen, Alexander Rosenberg Johansen, NetSolP: predicting protein solubility in Escherichia coli using language models, Bioinformatics 38 (4) (2022) 941–946.

[100] Vineet Thumuluri, Hannah-Marie Martiny, Jose J. Almagro Armenteros, Jesper Salomon, Henrik Nielsen, Alexander Johansen, Netsolp: predicting protein solubility in e. coli using language models, 2021, bioRxiv.

[101] Harianto Tjong, Huan-Xiang Zhou, Prediction of protein solubility from calculation of transfer free energy, Biophys. J. 95 (6) (2008) 2601–2609.

[102] Kyle Trainor, Aron Broom, Elizabeth M. Meiering, Exploring the relationships between protein sequence, structure and solubility, Curr. Opin. Struct. Biol. 42 (2017) 136–146.

[103] Chun Wei Tung, Shinn-Ying Ho, Computational identification of ubiquitylation sites from protein sequences, BMC Bioinform. 9 (1) (2008) 310.

[104] Matee Ullah, Shahid Akbar, Ali Raza, Quan Zou, Deepavp-tppred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm, Bioinformatics 40 (5) (2024), btae305.

[105] Salvador Ventura, Sequence determinants of protein aggregation: tools to increase protein solubility, Microb. Cell Fact. 4 (1) (2005) 11.

[106] Chao Wang, Quan Zou, Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with deepsolue, BMC Biol. 21 (1) (2023) 1–11.

[107] Lei Wang, Zhu-Hong You, Xing Chen, Jian-Qiang Li, Xin Yan, Wei Zhang, Yu-An Huang, An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences, Oncotarget 8 (3) (2017) 5149.

[108] Xianfang Wang, Yifeng Liu, Zhiyong Du, Mingdong Zhu, Aman Chandra Kaushik, Xue Jiang, Dongqing Wei, Prediction of protein solubility based on sequence feature fusion and DDcCNN, Interdiscip. Sci. Comput. Life Sci. 13 (4) (2021) 703–716.

[109] Geoffrey I. Webb, Eamonn Keogh, Risto Miikkulainen, Naïve bayes, in: Encyclopedia of Machine Learning, vol. 15, 2010, pp. 713–714.

[110] Leyi Wei, Chen Zhou, Huangrong Chen, Jiangning Song, Ran Su, ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides, Bioinformatics (Oxford, England) 34 (23) (2018) 4007–4016.

[111] Gilbert White, William Seffens, Using a neural network to backtranslate amino acid sequences, Electron. J. Biotechnol. 1 (2) (1998) 196–201.

[112] David L. Wilkinson, Roger G. Harrison, Predicting the solubility of recombinant proteins in Escherichia coli, Bio/technology (Nature Publishing Company) 9 (5) (1991) 443–448.

[113] Paul T. Wingfield, Overview of the purification of recombinant proteins, Current Protocols Protein Sci. 80 (1) (2015), 6–1.

[114] Bing Xu, Naiyan Wang, Tianqi Chen, Mu Li, Empirical evaluation of rectified activations in convolutional network, arXiv preprint, arXiv:1505.00853, 2015.

[115] Joseph F. Zayas, Solubility of proteins, in: Functionality of Proteins in Food, Springer, 1997, pp. 6–75.

[116] Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nat. Genet. 50 (8) (2018) 1171–1179.

[117] Chao Zhou, Changshi Wang, Hongbo Liu, Qiangwei Zhou, Qian Liu, Yan Guo, Ting Peng, Jiaming Song, Jianwei Zhang, Lingling Chen, Yu Zhao, Zhixiong Zeng, Dao-Xiu Zhou, Identification and analysis of adenine n6-methylation sites in the rice genome, Nature Plants 4 (8) (2018) 554–563.