IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 May 2023, accepted 4 June 2023, date of publication 9 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284464

RESEARCH ARTICLE

RPPSP: A Robust and Precise Protein Solubility Predictor by Utilizing Novel Protein Sequence Encoder

FAIZA MEHMOOD, SHAZIA ARSHAD, AND MUHAMMAD SHOAIB
Department of Computer Science, University of Engineering and Technology Lahore, Lahore 54000, Pakistan
Corresponding author: Faiza Mehmood (faiza.mehmood@kics.edu.pk)

ABSTRACT Protein solubility prediction is essential to understand diverse types of biological processes and to explore the impact of different factors (ionic strength, temperature, PH of medium and electrostatic repulsion) on the productivity of proteins. It also plays an important role in disease analysis and drug development processes. Protein solubility prediction through experimental approaches is time-consuming, labour intensive and error-prone. To empower the process of protein solubility prediction and facilitate large scale analysis, 16 different computational predictors have been proposed. However, these predictors have low predictive performance mainly due to extraction of less semantic and discriminative features from raw protein sequences. Existing predictors either extract sequence order information or positional information, while both types of information are important to discriminate soluble and insoluble proteins. This paper presents a novel encoder CTAPAAC capable of generating statistical representations of protein sequences by extracting 4 different types of information correlation, distribution, composition and transition. Over 4 benchmark datasets a comprehensive intrinsic and extrinsic performance analysis of proposed and 14 most widely used existing protein sequence encoders reveals that proposed encoder has more potential in transforming soluble and insoluble protein sequences into statistical vectors having discriminative patterns among soluble and insoluble classes. Proposed encoder along with random forest classifier outperforms existing best performing protein solubility predictors with a significant margin of 6%, 7%, 25% and 10% over PSI:Biology, E.coli, price and Esol datasets in terms of accuracy. Source code of proposed predictor is publicly available at https://github.com/Faiza-Mehmood/RPPSP.

INDEX TERMS Solubility prediction, machine learning, statistical representation method, novel sequence encoder, composition transition distribution, physicochemical properties.

I. INTRODUCTION

Proteins are complex nitrogenous organic compounds, that are comprised of long chains of amino acids joined through peptide bonds [37]. These molecules are essential for humans and other living organisms as they perform diverse types of biological processes like metabolic reactions, PH maintenance, oxygen transformation and message transmission in cells [37]. They support immune system to enhance its defensive role and are responsible to control different types of physical processes such as speaking, listening, breathing, and walking and also control growth of different organs such as hair, nails, bones and skin [69]. To make sure proper working of biological and physical processes, healthy food maintains the quality and quantity of proteins in cells [79]. Based on molecular structure and biological roles, proteins are broadly categorized into three major categories namely: simple, conjugated and derived proteins [23]. A large number of amino acids form simple proteins, while conjugated proteins are formed by the combination of simple proteins and non-protein stuff in body [23]. Derived proteins are produced through simple and conjugated proteins e.g., peptides, proteoses, recombinant and denatured proteins [44]. Among different types of proteins, recombinant proteins a subtype

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

of derived proteins are gaining more attention where biologists are keenly interested to deeply explore their diverse types of roles in cellular processes [47]. Commonly, these proteins are considered more important as they control the expression of genes and translation of mRNA [77]. Apart from their involvement in biological processes, they are being utilized for the development of diverse types of medicines such as recombinant human insulin, recombinant hormones, interleukins, tumour necrosis factors, growth factors, blood clotting factors, interferons and thrombolytic drugs [48], [90].

Medicine industries produce recombinant proteins through genetic engineering processes by using multiple types of bacteria such as Escherichia coli (E.coli), and yeast [15]. Approximately 25-57% of recombinant proteins are soluble and 33-35% are insoluble [28], [70]. Protein solubility degree depends upon several factors such as ionic strength, temperature, PH of medium, type of solvent and balance between hydrophobic interaction and electrostatic repulsion in protein molecules [31], [93]. Ionic strength can increase or decrease protein solubility such as at isoelectric (PI) point net-charge becomes zero [93], which makes attractive forces predominant and protein becomes insoluble, while at lower and higher values of (PI) protein holds a negative or positive charge that increases the solubility of proteins [93]. Mainly pharmaceutical industries only utilize soluble proteins because insoluble proteins create complications such as in prokaryotes they cause overexpression of heterologous proteins that create inclusion bodies [27], [43]. These inclusion bodies cause three main diseases hepatitis, hepatorenal syndrome and hepato pulmonary syndrome [16]. Solubility level of recombinant proteins has significant importance in the development of medicines, novel therapies and antibodies [31].

A vast quantity of soluble proteins is essential for developing fine dispersed colloidal systems [72], [92], [93] that control the flow of blood and avoid blood clotting and ascites [32], [92]. The shortfall of soluble proteins may damage the colloidal system that affects diverse biological processes, such as in the case of weak colloidal system water enters interstitial tissues and can cause complication of multiple diseases such as chronic kidney disease that may lead to ascites, decompensated chronic liver disease (DCLD) that may also lead to ascites and hypovolemic shock [1], [32]. To perform diverse types of disease analysis, it is important to ensure the required quantity of soluble proteins in different cells. As different factors affect the solubility of proteins, so for the task of solubility prediction it is important to deeply explore the impact of these factors that can be utilized to maintain the solubility level in different cells. Protein solubility prediction is also important to diagnose colloidal system complications and drug development.

Traditional experimental approaches predict protein solubility levels by utilizing equilibrium concentration of proteins, solvent contributions and conformational entropy that is capable of accurately calculating the flexibility and sensitivity of structural behaviour [3], [86]. Protein solubility level prediction through experimental approaches is expensive, error-prone and time-consuming [86]. Hence, the aforementioned challenges analyze protein sequences impossible at large scale.

Following the success of machine and deep learning approaches in different application areas such as energy forecasting [46], [60], automation of finance departmental tasks [14], forgery analysis [45], [59] and bioinformatics [26], [49]. Researchers have developed 16 AI-based predictors that are capable of predicting solubility level of recombinant proteins from raw protein sequences [12], [63]. Working paradigm of existing predictors can be categorized into two main phases. Firstly, raw protein sequences are transformed into statistical vectors as machine and deep learning classifiers cannot directly operate on raw sequences due to their inherent dependency over statistical vectors [6], [8]. In the second phase, transformed vectors are utilized to train machine and deep learning based classifiers.

In the marathon of developing more powerful protein solubility predictor, to transform raw protein sequences into statistical vectors [5], [7], [8],researchers have utilized 19 different encoders. However, these encoders do not capture sequence order information such as interdependencies between amino acids and distributional, compositional as well as transitional information of amino acids [20], [34], [50], [51].

In second stage, for discriminating statistical vectors into soluble and insoluble classes, researchers have utilized 7 different machine learning classifiers [18], [64], [68], [75], [76] and 8 deep learning based classifiers. Among deep learning classifiers, 3 classifiers are convolutional neural network based [41], [56], [88] and 2 classifiers are long short term memory (LSTM) based [36], [67]. One classifier is graph based [18] and one method is based on language modelling [85].

Performance of classifiers mainly relies on the quality of generated statistical vectors, as protein sequences are made up of repetitive patterns of 20 unique amino acids [7]. While performing classification based on raw protein sequences, it is considered that distribution of unique amino acids at different positions is almost identical in the protein sequences of same class; while their positional distribution in the sequences of different classes slightly varies [9], [80]. The prime objective of protein sequence encoding methods is to extract position aware distribution of amino acids in protein sequences and encode such information into statistical vectors. It is widely accepted, that simple classifiers can produce better performance when they are fed with comprehensive features that contain discriminative patterns among different classes, while sophisticated classifiers may not perform better when they are fed with feature vectors that contain non-discriminative patterns. According to our best knowledge, there does not exist any encoding method that generates statistical representations of protein sequences by capturing amino acids occurrence and their

compositional and transitional information along with correlational information.

To empower the process of discriminating soluble and insoluble proteins using only raw protein sequences, contributions of this paper are manifold: **(I)** It presents CTAPAAC a novel sequence encoding method that transforms raw protein sequences into statistical vectors by capturing amino acid's correlational, distributional, compositional and transitional information from raw sequences. **(II)** To precisely capture amino acids correlational information, proposed encoder deeply explores the potential of different physicochemical properties. In order to reap the benefits of different physicochemical properties, it combines statistical representations of top-performing properties by utilizing a phenomenon similar to forward feature selection method. **(III)** Over 4 public benchmark datasets, it performs an intrinsic performance analysis of proposed CTAPAAC and 14 existing most widely used protein sequence encoding methods. **(IV)** Over 4 public benchmark datasets, using 7 different machine learning classifiers, it performs extrinsic performance analysis of proposed CTAPAAC and 14 existing most widely used protein sequence encoding methods. **(V)** Comprehensive performance comparison of proposed predictor with 19 existing protein solubility predictors over 4 public benchmark datasets.

II. LITERATURE SURVEY

In the marathon of developing AI based robust and precise end to end pipelines for the prediction of protein solubility, researchers have utilized diverse types of feature encoding methods and machine/deep learning based predictors that are summarized in this section.

Smialowski et al. [76] proposed a machine learning based approach namely PROSO that uses SVM and Naïve Bayes classifiers in series. For statistical vectors representation they used the frequency based encoder Amino acid composition (AAC). They carried out experimentation on E.coli species dataset collected from targetDB database,[1] and achieved performance values of 0.434 and 0.72 in terms of Matthews Correlation Coefficient (MCC) and accuracy respectively. Magnan et al. [57] proposed a machine learning model namely SolPro that used two stage SVM model and employed information of seven group based encoders including Natural-20, Hydropho-5, ConfSimi-7, BlosumSM-8, ClustEM-14, ClustEM-17, PhysChem-7 by computing frequencies of unimer, bimer and trimer of a sequence to generate the feature representation. For experimentation they prepared the balanced protein solubility dataset from different databases namely PDB, targetDB, and SwissProt and produced 0.5938 accuracy.

Huang et al. [40] proposed a scoring card method SCM that generates the dipeptide score of a sequence to predict its solubility. They trained intelligent genetic method as a classifier and performed experimentation using one benchmark dataset SolProDB and other self prepared dataset Sd957. Authors performed 10-fold cross validation and achieved accuracy of 0.8429 and 0.539 over Sd957 and SoLproDB respectively. Smialowski et al. [75] produced PROSO II extension of PROSO by replacing the first layer classifier with the combination of parzen window model to capture the similarity information of sequence Furthermore, they employed AAC features to logistic regression classifier followed by logistic regression classifier at second layer, and reported accuracy and MCC score of 0.754 and 0.39 respectively. Agostini et al. [2] developed a webserver namely ccSOL for the prediction of solubility in E.coli gene expressions (heterologous and endogenous). They feed hydrophobicity, hydrophilicity, disorder, α-helix and β-sheet information of sequence to ccSOL and outperformed Smialowski et al. [75] with 0.829 AUROC, Magnan et al. [57] with 0.857 AUROC and Niwa et al. [63] with 0.933 AUROC.

Rawi et al. [68] explored the combined effect of direct and structural feature information of sequences by feeding as input to gradient boosting classifier. To evaluate the proposed model namely PaRSnIP they used Smialowski et al. [75] dataset for training and Chang et al. [17] for testing. Further they achieved 0.70 accuracy and 0.48 MCC. Khurana et al. [41] proposed a deep learning based method namely DeepSol by feeding direct and structural information of sequences. They performed experimentation by using Smialowski et al. [75] as a train set and Chang et al. [17] as a test set and produced 0.77 accuracy. Following the success of generative adversarial approaches in diverse domains, Han et al. [33] utilised it for data augmentation. Further they utilised multi layer perceptron model for classification of sequences into soluble and insoluble classes. Authors performed experimentation over benchmark dataset [63] related to ecoli species and proposed predictor managed to produce performance values of 0.45 R^2 score and 0.05 MSE error.

Considering the B-factor normalization by computing the flexibility and dynamics of protein sequence structure [82], [87], Bhandari et al. [12] proposed a Solubility Weight Index "SWI" approach that computes length independent composition based weights to predict the level of solubility in protein sequence. To evaluate the SWI approach they performed experimentation over two benchmark datasets namely PSI:Biology [12] and esol [63] proteins of 196 different species that are expressed in E.coli and it produced performance figures of 0.71 AUC and 0.50 R^2 respectively. Furthermore, they have provided the web interface to predict the solubility level and maximize the protein expressions. Hou et al. [39] developed the structure based solubility prediction method namely SOLart that employs sequence basic, secondary struct, statistical potential and composition information by introducing ten new feature descriptors. They performed experimentation on 4 different benchmark datasets D_{Ecoli}, $D_{Scerevisiae}$, M_{Ecoli}, $M_{Scerevisiae}$ with 5 fold cross validation and managed to produced performance 25%

[1] http://targetdb.pdb.org/

23% 28% and 24% in terms of RMSE error respectively. They also have provided a web-interface for protein solubility prediction.

Chen et al. [18] explored the potential of 5 different encoding methods namely BLOSUM62, AAPHY7, PSSM, HMM and SPIDER3 to transform protein sequences into statistical vectors. They developed GraphSol named protein solubility predictor by using graph neural network to extract discriminative features and attention mechanism with an aim to feed classifier with more comprehensive weights of discriminative features. GraphSol predictor was evaluated on two different species (S.cerevisiae, E.coli) benchmark datasets by taking different combinations of statistical vectors, encoded through 5 different encoders. Graphsol managed to produce performance values of 0.782 accuracy, 0.873 AUC, and 0.501 R^2 score over benchmark public dataset namely esol independent test set and 0.37 R^2 score over S.cerevisiae test set. Wang et al. [88] proposed DDcCNN predictor which makes use of convolutional layers to extract discriminative features of sequences into soluble and insoluble classes. DDcCNN was fed with statistical vectors generated from raw sequences by capturing gap based local and global features. DDcCNN was evaluated on Smialowski et al [75] benchmark dataset and it produced performance figures 0.7782, 0.7613, 0.7932 and 0.57 in terms of accuracy, sensitivity, specificity and MCC respectively. Madani et al. [56] proposed DSResSol predictor relies on multiple Resnet blocks which make DSResSol training process smooth by providing different paths for gradient back flow. Further, they fed DSResSol model with statistical vectors that were generated by extracting diverse types of information from raw sequences such as sequence length, molecular weights, sequence instability and chemical properties of sequence including atomicity, gravy, hydrophobicity, charge etc. DSResSol performance was analysed by experimentation over two different independent test sets related to E.coli species benchmark datasets provided by [17] and [65]. Over both datasets DSResSol managed to produce performance values of 0.796 and 0.629 in terms of accuracy of chang et al. [17] and NESG [65] respectively.

Table 1 summarizes existing protein solubility predictors in terms of used sequence encoding methods, datasets and classifiers along with their performance values.

III. MATERIALS AND METHODS

This section briefly describes different modules of proposed RPPSP predictor graphical illustration which is shown in Figure 1. In Figure 1 data preparation module describes benchmark datasets that are used for experimentation, a comprehensive detail of this module is provided in subsection III-C. Second module illustrates proposed encoding method which is briefly described in subsection III-A. Third module represents different machine learning classifiers and evaluation measures, details of which are described in section III-D and III-E, respectively.

A. PROPOSED COMPOSITION AND TRANSITION AWARE AMPHIPHILIC PSEUDO-AMINO ACID COMPOSITION ENCODER (CTAPAAC)

Chou et al. [21] proposed a protein sequence encoding method named pseudo amino acid composition that transforms raw protein sequences into statistical vectors by capturing distributional information of 20 unique amino acids. To generate more comprehensive statistical vectors of protein sequences, Chou et al. [22] introduced modified version of pseudo amino acid composition named Amphiphilic Pseudo Amino Acid Composition (APAAC) [22]. APAAC encoder generates statistical vectors by capturing amino acid's distribution and their correlation [24], [30], [53] information from protein sequences. Although APAAC encoder captures amino acid's distributions and correlation information, however, it remains fail to extract their composition and transition information from protein sequences.

Considering the need of a powerful encoding method that generates statistical vectors by extracting diverse types of information, we introduce Composition and Transition aware Amphiphilic Pseudo-Amino Acid Composition (CTAPAAC) encoder capable of capturing 4 different types of information namely: amino acids correlation, distribution, composition and transition.

In order to capture amino acids correlation information, following chou et al., we utilize 3 physicochemical properties namely: hydrophobicity, hydrophilicity and side chain mass. These properties facilitate to determine the characteristics of protein sequences by computing diverse types of information such as ionic strength, interactive and catalytic mechanisms. Such information helps predictor to discriminate protein sequences into different classes. Table 2 illustrates physicochemical values of 20 amino acids for 3 properties.

For each property, mean and standard deviation of amino acids property values are computed using Equations 1 and 2. Utilizing mean and standard deviation, normalized property values are computed using Equation 3.

$$Mean[P_n] = \frac{\sum_{i=1}^{20} P_n[AA_i]}{20} \tag{1}$$

$$SD[P_n] = \sqrt{\frac{\sum_{i=1}^{20}(P_n[AA_i] - Mean[P_n])^2}{20}} \tag{2}$$

$$Properties[P_n] = \frac{P_n[AA_i] - Mean[P_n]}{SD[P_n]} \tag{3}$$

In Equations 1, 2 and 3 P_n represents physicochemical properties where:

$$\therefore n \in (hydrophobicity, hydrophilicity, sidechainmass)$$

'AA_i' denotes amino acids of protein sequence, i.e.

$$\therefore AA_i \epsilon [A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V]$$

TABLE 1. A comprehensive survey of existing protein solubility predictors.

Publication	Dataset	Encoding Methods	Approach	Performance
Smialowski et al. (2006) [76]	E.coli species dataset	Amino Acid Composition **(AAC)**	PROSO (SVM-NB)	Accuracy: 0.72
Magnan et al. (2009) [57]	SOLP	Natural-20, Hydropho-5, ConfSimi-7, BlosumSM-8, ClustEM-14, ClustEM-17, PhysChem-7 **(Grouped AAC)**	SolPro	Accuracy: 0.59
Huang et al. (2012) [40]	SolProDB SD97	score of dipeptide **(Grouped AAC)**	SCM	Accuracy: 0.53 Accuracy: 0.84
Smialowski et al. (2012) [75]	E.coli species dataset	Amino Acid Composition, Similarity **(AAC)**	PROSO II	Accuracy: 0.75
Agostini et al. (2014) [2]	E.coli species dataset SOLP Niwa et al. [63]	α-helix β-sheet **(physicochemical)**	ccSOL	AUROC: 0.83 AUROC: 0.85 AUROC: 0.93
Rawi et al. (2017) [68]	E.coli species dataset (train) Chang et al. [17] (test)	Direct and structural feature information **(Structural)**	PaRSnIP	Accuracy: 0.7
Khurana et al. (2018) [41]	E.coli species dataset (train) Chang et al. [17] (test)	Direct and structural feature information **(Structural)**	DeepSol	Accuracy: 0.77
Han et al. (2019) [33]	niwa et al. [63]	Amino Acid Composition **(AAC)**	MLP	R2 score: 0.45
Bhandari et al. (2020) [12]	PSI:Biology Esol	B-factor normalization**(AAC)**	SWI	AUC: 0.71 R2: 0.50
Hou et al. (2020) [39]	D-eColi D-cerevisiae M-eColi M-cerevisiae	sequence basic, secondary struct, statistical potential and composition information **(Structural)**	SOLart	RMSE: 25 RMSE: 23 RMSE: 28 RMSE: 24
Chen et al. (2021) [18]	esol S.cerevisiae	BLOSUM62 (Blocks substitution matrix), AAPHY7 (Physicochemical), PSSM (scoring matrix), HMM (matrix) and SPIDER3 **(Structural)**	GraphSol	Accuracy: 0.78 R2 score: 0.37
Madani et al. (2021) [56]	Chang et al. [17] NESG	sequence length & instability, molecular weights, chemical properties (atomicity, gravy, hydrophobicity, charge) **(Physicochemical)**	DSResSol	Accuracy: 0.79 Accuracy: 0.62
Wang et al. (2021) [88]	E.coli species dataset	Gap based local and global method **(Grouped AAC)**	DDcCNN	Accuracy: 0.77

TABLE 2. Physicochemical values of 20 amino acids for 3 properties namely hydrophobicity, hydrophilicity and side chain mass.

Properties	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Hydrophobicity	0.62	-2.53	-0.78	-0.9	0.29	-0.85	-0.74	0.48	-0.4	1.38	1.06	-1.5	0.64	1.19	0.12	-0.18	-0.05	0.81	0.26	1.08
Hydrophilicity	-0.5	3	0.2	3	-1	0.2	3	0	-0.5	-1.8	-1.8	3	-1.3	-2.5	0	0.3	-0.4	-3.4	-2.3	-1.5
Side Chain Mass	15	101	58	59	47	72	73	1	82	57	57	73	75	91	42	31	45	130	107	43

In order to understand the working paradigm of proposed encoder to generate statistical vectors by utilizing processed physicochemical values and raw protein sequences, let's take a hypothetical protein sequence.

$$A_1, A_2, A_3, A_4, A_5, A_6, \ldots\ldots\ldots\ldots\ldots, A_{L-1}, A_L \tag{4}$$

In hypothetical protein sequence, $A_1, A_2, \ldots A_L$ denotes different amino acids. To capture amino acids correlation information at different levels, following chou et al., we generate bi-mers with different lag values. Equation 5, illustrates bi-mers generated with different lag values.

$$\begin{cases} A_1A_2, A_2A_3, A_3A_4, \ldots\ldots A_{L-1}A_L & \text{with lag 1} \\ A_1A_3, A_2A_4, A_3A_5, \ldots\ldots A_{L-2}A_L & \text{with lag 2} \\ A_1A_4, A_2A_5, A_3A_6, \ldots\ldots A_{L-3}A_L & \text{with lag 3} \\ A_1A_{l+1}, A_2A_{2+l}, A_3A_{3+l}, \ldots\ldots A_{L-l}A_L & \text{with lag } l \end{cases} \tag{5}$$

Equation 6 describes the process of physicochemical properties information incorporation into generated bi-mers. Utilizing experimentally pre-computed physicochemical values of amino acids (provided in Table 2), generated bi-mers computes correlation information of amino acids at different gap

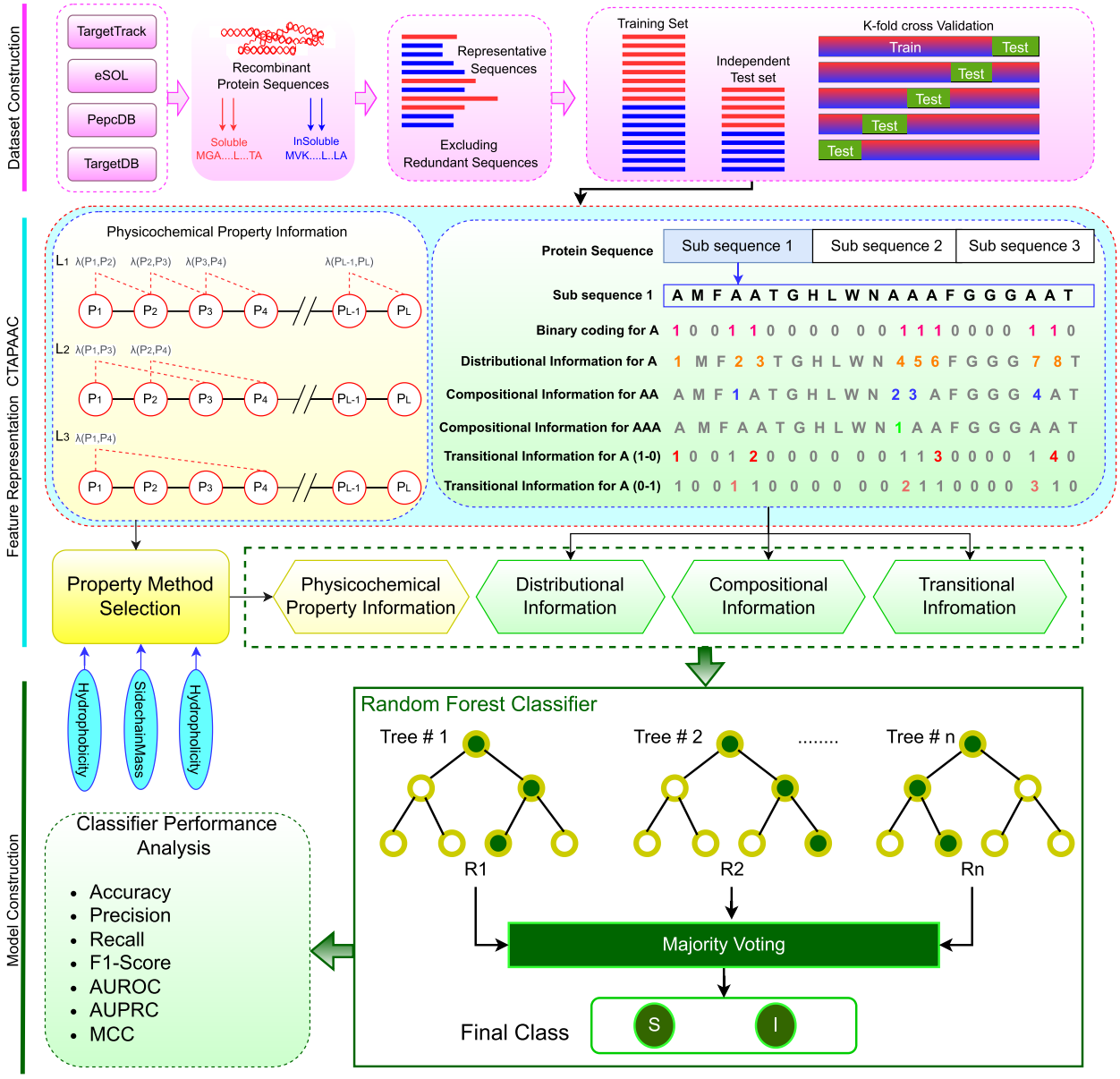


FIGURE 1. **Three basic modules of proposed predictor: (A) Dataset Construction: Public benchmark datasets are prepared by collecting sequences from different databases like TargetTrack, eSol, PecpDB and target DB (B) Feature Representation CTAPAAC: Generates numerical representations of protein sequences by using proposed encoder CTAPAAC (C) Model Construcion: Performance evaluation of 7 different machine learning classifiers.**

levels. s described in Equation 6.

$$Bi-mers\_Correlation\_information = A_1 \times A_{l+1} \therefore A_1 : Property\_value\_of\_First\_AminoAcid \therefore A_{l+1} : Property\_value\_of\_Second\_AminoAcid \quad (6)$$

$$Correlation[P_n][lag_k] = \frac{(\sum_{j=1}^{len(seq)-l} Property\_value\_of\_bimer[j] - Mean[P_n])/20}{\sqrt{\frac{\sum_{j=1}^{len(seq)-l}(Property\_value\_of\_bimer[j]-Mean[P_n])^2}{20}}} \quad (7)$$

Furthermore, correlation of bi-mers for each property at different lag values is computed using Equation 7, where lag_k represents range of lag values at which bi-mers with different gaps of amino acids in the sequence are generated and P_n denotes n^{th} property. In this Equation property value of each bi-mer is computed using Equation 6 and mean of each property is computed using Equation 1. Furthermore, to calculate overall correlation of bi-mers at different lag values, Equation 8 normalizes each correlation value at particular lag with bi-mers of the sequence and sum normalized values using all lag values. Equation 9 scales computed correlation

values at different scales in the range of 0.1 to 1.

$$Encoding[P_n] = \sum_{k=1}^{lag} \frac{Correlation[P_n][lag_k]}{seq(len) - lag_k}. \quad (8)$$

$$Encoding[p_n][lag_k] = \frac{w \times Correlation[P_n][lag_k]}{w \times Encoding[p_n] + 1}. \quad (9)$$

In Equation 9, w is weight parameter that contains scalar values.

Chou et al. incorporated correlational information into distributional information by counting the occurrence frequency of each amino acid in a sequence, that is further normalized with overall correlation values of amino acids. Equations 8 and 10, describe mathematical expressions to compute correlation and distribution information of amino acids. Hence, a final 20-dimensional vector is generated, where each value represents normalized frequency of an unique amino acid.

$$Dist[AA] = \frac{frequency\ of\ AA\ in\ sequence}{w \times Encoding[P_n] + 1} \quad (10)$$

Rather than extracting and incorporating only distribution information of amino acids into physicochemical properties based correlation information of bi-mers, we propose to also extract and include, transition and composition information of amino acids.

1) COMPOSITION INFORMATION

Compositional information denotes consecutive two and three times occurrence frequency of an amino acid in a whole sequence and it can be computed using Equations 11 and 12, as shown at the bottom of the page. Figure 2 (b) presents a toy example to illustrates the process of computing amino acids composition information in a protein sequence at two different levels: consecutive two times and consecutive three times occurrence frequencies of amino acids. Hence, proposed encoder extracts and encodes compositional information of raw sequences into 40-dimensional vector in which first 20-dimensions represent consecutive two times occurrence and remaining 20-dimensions represent consecutive three times occurrence of an unique amino acid.

2) TRANSITION INFORMATION

Transition information describes how often different amino acids occur in a sequence. Figure 2 (c) shows the transitional information extraction process of amino acid 'A' with respect to other 19 amino acids. In the sequence, count of 'A' amino acid when post-amino acid is from other 19 amino acids is 4 as described in 1-0 transition and similarly count of 'A' amino acid when pre-amino acid is from other 19 amino acids is 4 as described in 0-1 transition.

$$Trans[!AA][AA] = \frac{Count\ of\ (0-1)\ Transition}{w \times Encoding[P_n] + 1} \quad (13)$$

$$Trans[AA][!AA] = \frac{Count\ of\ (1-0)\ Transition}{w \times Encoding[P_n] + 1} \quad (14)$$

Equations 13 and 14 describe mathematical expressions to compute 1-0 and 0-1 transition information of amino acids, respectively. Proposed encoder extracts and encodes transitional information into 40-dimensions vector where first 20-dimensions represent 1-0 transitions and last 20-dimensions represent 0-1 transitional frequencies of 20 unique amino acids. Equation 15 describes the concatenation process of all 4 different types of information.

$$CTAPAAC = \begin{cases} Encoding[p_n][lag_k] \bigoplus \\ Dist[AA] \bigoplus \\ Comp[AA^2] \bigoplus \\ Comp[AA^3] \bigoplus \\ Trans[!AA][AA] \bigoplus \\ Trans[AA][!AA] \end{cases} \quad (15)$$

Furthermore, to analyze whether more comprehensive information about distribution, transition and composition of amino acids can be captured at global level by taking full sequence or at local level from subsequences, we generate statistical vectors in two different settings. In first setting, we take full sequence of protein as explained above. In second setting, we first generate equal length subsequences represented by 'l'. To explain this setting more briefly let's we have a hypothetical sequence $A_1, A_2, A_3, \ldots A_L$, considering '$l = 3$' that means given sequence will be divided into 3 equal parts, each subsequence will be consider independent and a statistical vector will be generated separately. In this way for distributional information rather than 20 dimensional vector, $20 \times l$ dimensional vector will be generated. Similarly, for compositional $40 \times l$ dimensional, $40 \times l$ dimensional for transitional information will be produced.

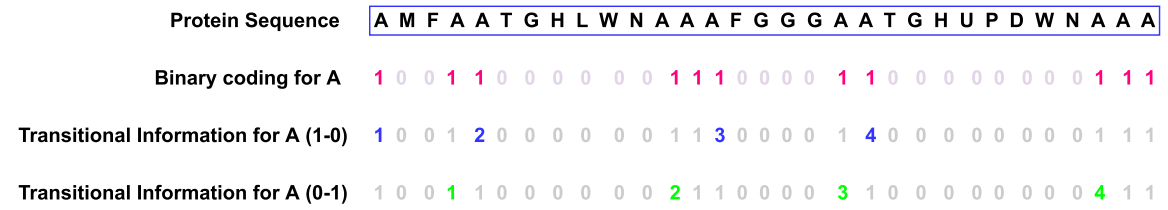
Finally, concatenate the all above captured information (correlational 9, distributional 10, compositional 11, 12 and transitional 13, 14) of protein sequence as shown in Equation 15. Thus, final statistical vector will be of dimension $[(20 \times l) \times 5 + lag] \times n$ where 20 are the unique amino acids, 'l' is subsequences, 5 is sequence order information captured through 5 different ways: distributional, 2 compositional (consecutive two, consecutive three) and two transitional (1-0, 0-1), while n is for number of physicochemical properties (hydrophobicity, hydrophilicity and side chain mass). To summarize the process of statistical vector generation of protein sequences through proposed encoder a pseudo code is given here.

$$Comp[AA^2] = \frac{Consecutive\ Two\ times\ occurrence\ of\ same\ AminoAcids}{w \times Encoding[P_n] + 1} \quad (11)$$

$$Comp[AA^3] = \frac{Consecutive\ Three\ times\ occurrence\ of\ same\ AminoAcids}{w \times Encoding[P_n] + 1} \quad (12)$$

(a) A toy example for computing distributional information of Amino acid 'A' that occur 11 times. So distributional information of amino acid A is 11 and similarly other amino acids distributional information can be captured

(b) A toy example for computing compositional information of Amino acid 'A' that 6 times occur in consecutive two patterns and 2 times occur in consecutive 3 patterns. So, compositional information of amino acid 'A' in two consecutive occurrences is 6 and in consecutive 3 occurrences is 2. Similarly other amino acids compositional information can be captured



(c) A toy example for computing transitional information of Amino acid 'A' that has 4 transitions from 1 to 0 and 4 transitions from 0 to 1. So 1-0 transitional information of amino acid 'A' is 4 and its 0-1 transitional information is also 4. Similarly, other amino acids transitional information can be captured

FIGURE 2. **The process of computing amino acids distribution composition and transition information.**

B. EXISTING PROTEIN SEQUENCE ENCODERS

This section briefly summarizes 14 existing most widely used protein sequence encoders. These encoders are briefly described in several manuscripts, so here we only provide short description, interested readers can see more detail in referred articles. Considering working paradigm for transforming protein sequences into statistical vectors, all 14 encoders can be put into 4 different categories namely; physicochemical properties, block substitution, amino acids distribution and group-based.

Among existing encoders the most simple encoding method is block substitution method [19] which directly maps amino acids to experimentally pre-computed values. AESNN3 [52] is block substitution based method in which each amino acid is replaced with 3-dimensional vector of aligned learning, hence generated sequence encoding vector has ('length of sequence' $\times$ 3) dimensions. Dipeptide Composition (DPC) [13], Distancepair [54], Dipeptide deviation from expected mean (DDE) [71], and Adaptive skip dipeptide composition (ASDC) [89] encoders belongs to amino acids distribution category. DPC [13] transforms protein sequences into statistical vectors by computing distribution of amino acids in the protein sequences. Distancepair [54] encoder generates statistical representation by computing amino acids bimers distribution at different gaps. For example, at distance $= 4$, it will calculate frequency of each amino acid at uni-mer level and compute frequency of bimers generated with gap 0, 1, and 2. DDE [71] also generates statistical representation of protein sequences by computing distribution of bi-mer codons. ASDC [89] is also amino acids distribution computation method, which first computes amino acids bi-mers occurrence frequency and normalize them with sequence length.

In order to capture information of amino acids in a comprehensive manner researchers have proposed group based encoding methods namely; Grouped Amino Acid Composition (GAAC) [94], Grouped Di-peptide Composition (GDPC) [94] and Grouped Tri-peptide Composition (GTPC) [94]. Group based encoder working paradigm can be summarised into 3 different phases. First, make different groups of amino acids based on their chemical structure, molecular formulas and structural behavior. Second, generate k-mers of raw sequences e.g. uni-mer, bi-mer, tri-mer. Third, calculate k-mers occurrence frequencies and normalize with k-mer sequence length. However, these encoders lack to capture inter-dependencies of

Algorithm 1 Variational Amphiphilic Pseudo-Amino Acid Composition Statistical Representation Method for Protein Sequences

Input:
Lag size **k**
number of subsequences l
number of properties n
Amino Acid **AA**
1. Normalize Property values of Amino Acids **Properties**$[P_n]$
Compute Mean of each Property **Mean**$[P_n]$
Compute Standard Deviation of each Property **SD**$[P_n]$
2. Compute correlation for each property with lag size
correlation$[P_n][l_{agk}]$
Generate bi-mers of sequence with lag size
for l = 1 to l **do**
  3. Capture Distributional information **Dist**[**AA**]
  Compute normalized frequency information of each AA
  4. Capture Compositional information **Comp**[**AA**]
  Compute normalized frequency information of two consecutive same AA occurrence **Comp**$[AA^2]$
  Compute normalized frequency information of three consecutive same AA occurrence **Comp**$[AA^3]$
  5. Capture Transitional information **Trans**[**AA**]
  Compute normalized frequency information of 0-1 transaction **Trans**[**!AA**][**AA**]
  Compute normalized frequency information of 1-0 transaction **Trans**[**AA**][**!AA**]
6. Generate the Statistical representation of Protein sequence **Encoding**
Combine the steps 2,3,4,5 captured information **Statistical Representation**

amino acids and sequence order information of protein sequences.

CTDC(Composition) [25], CTDT(Transition) [25] and CTDD(Distribution) [29] are physicochemical property based encoders that has 3 groups and each group has further 13 properties. CTDC(Composition) [25] encoder computes the frequency of amino acids with respect to groups and their respective physicochemical properties. CTDT(Transition) [25] encoder generates the bi-mer of raw sequences and then compute the frequency of bi-mers with respect to groups and their respective physicochemical properties. CTDD(Distribution) [29] encoder computes frequency by dealing with five different distributions 1%, 25%, 50%, 75%, and 100% of amino acids for each property in all 3 groups.

Quasi Sequence order (QSOrder) [73] encoder transforms protein sequences into statistical vectors using two different physicochemical properties schneider and grantham. Similarly, Pseudo-amino acid composition (PAAC) [21] encoder generates statistical representation oof raw sequences using three physicochemical properties Hydrophobicity, hydrophilicity and side chain mass.

C. BENCHMARK DATASETS

This section briefly describes the details of 4 benchmark datasets that are used to verify the practical significance of proposed predictor.

Since the last decade, protein structure exploration has been an active area of research. A well-known project named protein structure initiative (PSI) was started in 2000 and successfully completed in 2017. This project generated a large amount of experimentally verified data that is available in Targettrack database [11]. Participants registered it separately and provide different kinds of information for protein structures like solubility level and explicit expressions [74]. Bhandari et al. [12] utilized Targettrack database to develop soluble/insoluble proteins benchmark dataset PSI:biology that contains experimentally verified 3726 soluble and 7500 insoluble protein sequences, Distribution os dataset is shown in Figure 3. Over this dataset, performance of 10 different AI-based protein solubility predictors are evaluated under 5 folds cross-validation based experimentation.

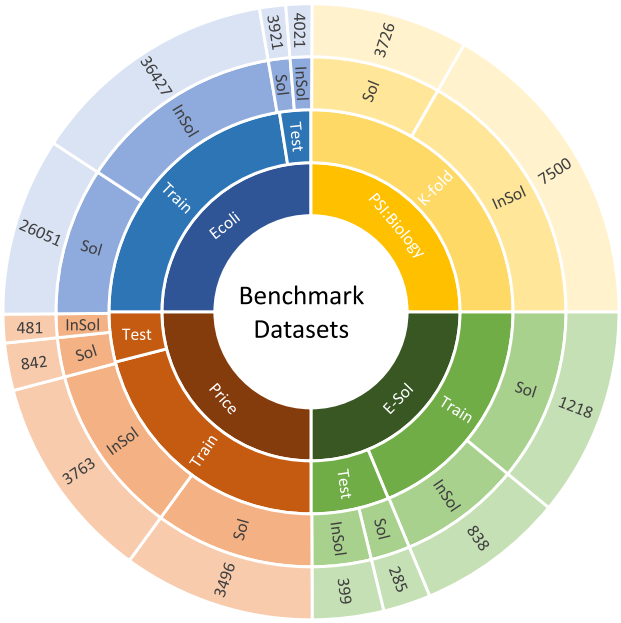


FIGURE 3. Statistics of 4 benchmark datasets in terms of samples distribution in soluble and insoluble classes.

Another protein solubility prediction dataset namely eSOL was developed by Niwa et al. [63], they collected 4132 proteins from eSOL database. In order to train classifier more comprehensively by avoiding model over and under fitting in training process, they removed 1395 samples that had sequence similarity greater than 25 and E-value less than $1e^{-6}$. Authors also developed an independent test set that contains 285 soluble and 399 insoluble proteins as shown in Figure 3. To date, 6 different AI-based predictors have been evaluated over eSol dataset. Over, train set researchers performed 5 fold cross-validation to report their performance values and also reported their performance over an independent test set.

Smialowski et al. [75] prepared e-coli species dataset by collecting 58689 soluble and 70954 insoluble proteins from pepcDB database[2] [68]. In order to reduce the redundancy in training data they used CD-hit tool and removed

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

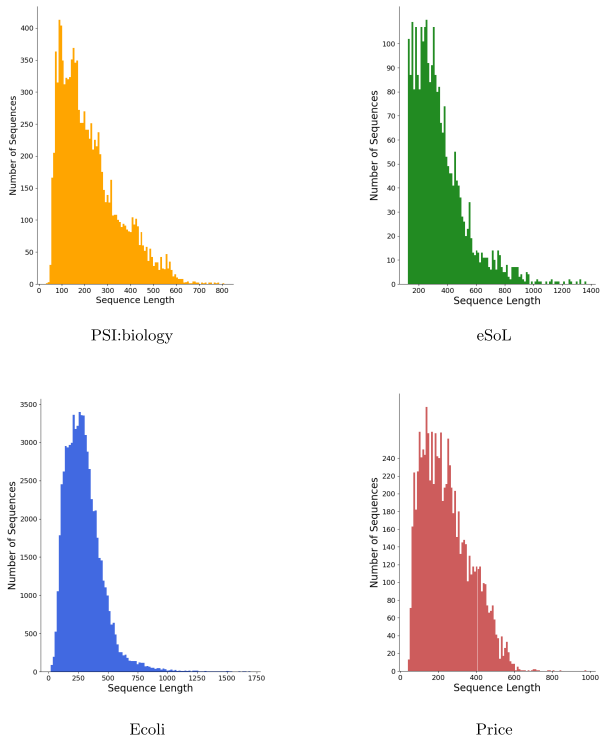


FIGURE 4. Sequence length distribution analysis in 4 benchmark datasets.

60223 sequences that have a similarity greater than 90%. Researchers utilized different types of experimentation on this dataset. However, recently Wang et al. [88] provided standard train test split of dataset with 90% training data and 10% test data. E-coli species dataset has been utilized to evaluate the performance of 6 AI-based predictors.

Price et al. [65] extracted 9644 E.coli protein sequences from one of the PSI centers namely North East Structural Genomics (NESG) to prepare protein solubility prediction dataset ''Price''. They processed dataset according to usability value where usability is the product of proteins expression value (E) and solubility level (S). They discard 2385 proteins which have a usability value of less than 4. They also provided independent test set having 842 soluble proteins and 481 insoluble proteins sequences. They performed experimentation in 2 ways: 5 fold cross-validation and independent test set. Price dataset has been utilized to evaluate the performance of 9 AI-based predictors.

Figure 4, illustrates sequence length distribution of 4 benchmark datasets. PSI:biology dataset sequences length varies from 50 to 800 amino acids. eSol dataset sequences length varies from 100 to 1300 amino acids. Length of sequences in the Price dataset varies from 50 to 950 amino acids. Among all 4 benchmark datasets, Ecoli dataset sequences have the highest length variability with a span of 20 to 1700 amino acids.

D. PROTEIN SOLUBILITY PREDICTORS

This section summarizes the details of 7 different classifiers that are used to perform an extrinsic performance analysis of proposed CTAPAAC and 14 existing encoders for the task of protein solubility prediction.

1) DECISION TREE CLASSIFIER (DT)

is a tree-based structure comprising of decision nodes and leaf nodes, the former decides the feature to split the data and later defines class of data point. The splitting criterion is based on maximum information gain for each node. To optimally split the data, model traverses' nodes by comparing every possible split and searches the best features which maximize entropy gain with minimal value of Gini index. It recursively splits the data points until reaches the class label.

2) RANDOM FOREST CLASSIFIER (RF)

is a collection of multiple random decision trees with low sensitivity to training data in contrast to decision tree. A set of datasets is generated from original dataset having random samples with replacement of data instances called bootstrapping. Decision trees are trained on boot strapped data with a subset of features. To make a prediction an average result from random decision trees is computed which is called aggregation.

3) EXTRA TREE CLASSIFIER (ET)

is an ensemble of bagging and random forest. In contrast to random forest ETC utilizes complete data to build a decision tree and randomly chooses data split at each node.

4) ADAPTIVE BOOSTING (AB)

operates sequentially to train a strong classifier from several weak classifiers. To determine overall error, each classifier is first fitted on the original data. The classifiers are trained using a modified dataset that assigns more weight to inaccurately labeled observations. An effective classifier is produced after N iterations of this technique.

5) LOGISTIC REGRESSION (LR)

model estimates probability of categorical data, by learning the relationship between independent and dependent variables. Mean squared error serves as cost function for linear regression. It will result in a non-convex function of variables if this is utilized for logistic regression (theta).

6) GRADIENT BOOSTING (GB)

builds sequential models with a focus to reduce error of previous model. This is achieved by training new model on errors in the previous model's prediction. It scales the previous model by adding prediction of new model. Hence, it determines error patterns that are not captured in previous model. Throughout the iterative process of learning it updates three parameters target, prediction and error until the predicted values are closer to the actual value.

Unlike gradient boost, **Extreme Gradient Boosting (XGB)** parallelly builds model and offers regularization parameters which improve model generalization over unseen data.

7) EXTENDED K NEAREST NEIGHBORS

ExtKNN [4] is an advanced version of traditional KNN algorithm. First, it creates sampled data set using bootstrap strategies and then calculates distance between data and target. Based on calculated distance, it locates the shortest distance, records predicted value, upgrades the target and repeats K times. Lastly, voting method is used to determine final prediction.

E. EVALUATION MEASURES

A fair performance comparison of proposed RPPSP approach with existing computational approaches [18], [41], [88] over 4 benchmark datasets is performed using 9 most commonly used evaluation metrics including accuracy (ACC), F_1-score (F_1), Precision (Pre), Recall (Rec), Specificity (SP), Sensitivity (SN), Matthews Correlation Coefficient (MCC), Area Under Receiver Operating Characteristics (AUROC) [42] and Area Under Precision Recall Curve (AUPRC) [55].

$$f(x) = \begin{cases} Accuracy = \frac{TP+FP}{TP+FP+TN+FN} \\ Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \\ Specificity = \frac{TN}{TN+FP} \\ Sensitivity = \frac{TP}{TP+FN} \\ F_1 = \frac{TP}{TP+\frac{(FP+FN)}{2}} \\ MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TN+FN)*(TN+FP)*(TP+FN)*(TP+FP)}} \\ AUROC = \frac{1}{2}(\frac{TN}{TN+FP} + \frac{TP}{TP+FN}) \\ AUPRC = \frac{1}{2}(\frac{TP}{TP+FN} * \frac{TP}{TP+FP}) \end{cases} \tag{16}$$

In Equation 16, true positive (TP) represents the number of protein sequences where solubility level is correctly predicted, true negative (TN) represents the number of protein sequences where insoluble proteins are correctly predicted. Whereas, false positive (FP) refers to number of protein sequences where soluble proteins are wrongly predicted and false negatives (FN) refer to number of protein sequences that are wrongly predicted in insoluble class. Furthermore, to ensure that the performance of proposed RPPSP predictor is not biased towards the magnitude of corpus classes, MCC, AUROC and AUPRC are used. While MCC computes overall model performance by taking all 4 performance parameters true positive, false positive, true negative and false negative into account along with size of positive and negative classes. AUROC assists to analyze trade-off among false positive rate and true positive rate through equivalently caring about true positive and true negatives. Whereas, AUPRC analyzes trade-off among true positive rate and positive predicted value, paying more attention on how efficiently model can predict soluble proteins from all proteins sequences.

IV. EXPERIMENTAL SETUP

Proposed predictor is implemented in python language by utilizing three main APIs namely Biopython,[3] scikit-learn[4] and iLearn Plus.[5] Optimal hyperparameters selection has significant impact on the performance of machine learning classifiers. Following the success of grid search approach in previous studies [61], [66], [81], we performed classifiers hyperparameters optimization through grid search. Initially, select the subsequence split value 'L' from 1 to 4 for statistical encoder. Performance of Boosting classifiers such as AdaBoost [91], decision tree [91], gradient boost [68] and random forest [18] is highly dependent upon number of estimators, learning rate, maximum depth and split criterion. ElasticNet [10], Extratree [10], logistic regressor [75] and SGD [58] classifier's performance can be improved by selecting appropriate values of alpha, l_1, tol, penalty, maximum iterations and selection criterion. Table 3 summarizes the ranges of different hyperparameters for 9 different classifiers and optimal values of hyperparameters for each dataset selected by grid search.

V. RESULTS

This section describes discriminative distribution of amino acids in soluble and insoluble protein sequences. It illustrates the performance of random forest classifier at different settings of proposed encoder. Furthermore, it summarizes intrinsic performance comparison of proposed encoder and traditional APAAC encoder. It also illustrates extrinsic performance comparison of proposed and traditional APAAC encoders using 7 different machine learning classifiers. Finally, over 4 benchmark datasets under the hood of 2 different experimental settings 5 folds cross-validation and independent test sets, it compares the performance of proposed encoder and RF classifier based predictor with 19 existing protein solubility predictors [2], [12], [18], [33], [35], [36], [38], [41], [57], [67], [68], [75], [78], [84], [88] in terms of 8 different evaluation measures.

A. AMINO ACIDS DISTRIBUTION ANALYSIS

We utilize sequence logo library [83] to analyze distribution of amino acids in both soluble and insoluble classes. As illustrated in Figure 4, length of sequences varies from 20 to 1750 amino acids. It is difficult to graphically visualize position aware distribution of amino acids in longer sequences such as sequences having length of 175 amino acids. Hence, to visualize position aware discriminative potential of amino acids in both classes, we take 40 amino acids from start of each sequence and discarded amino acids from the sequences that lie after 40^{th} position. Here, we assumed that distribution of amino acids in other part of sequences will be almost similar to their distribution in first 40 amino acids of the sequences. Furthermore, we discarded sequences from the

[3] https://biopython.org/
[4] https://scikit-learn.org/stable/
[5] https://github.com/Superzchen/iLearnPlus

TABLE 3. Grid search ranges of classifiers hyperparameters and their optimal values.

Reg.	Grid-Search Parameters	PSI:Biology	E-Coli species	Price (Independent)	Price (Kfold)	e.Sol (Independent)	e.Sol (Kfold)
Adaboost	lr=[1.0, 0.1, 0.01, 0.001, 0.0001], n_{be}=[10,20,30,40,50,60, 80,100,150,200,300,400,500] Algorithm=[SAMME, SAMME.r]	(0.01, 50, SAMME)	(0.01, 40, SAMME)	(0.01, 60, SAMME)	(0.01, 40, SAMME)	-	-
DT	n_{be}=[10,20,30,40,50,60, 80,100,150,200,300,400,500], criterion=[Entropy,Gini,mse, mae,friedman-mse]	(30, Entropy)	(30, Entropy)	(50, Entropy)	(60, Entropy)	(auto, mse)	(auto, mse)
ExtraTree	alpha=[1.0,0.1,0.001,0.0001], l_1=[1.0,0.75,0.5, 0.25, 0.1], tol=[$1e^{-5}$,$1e^{-4}$, $1e^{-3}$, $1e^{-2}$], max_{it}=[100,200,500,1000,1500,2000], selection=[Cyclic,Random]	(1.0, 0.5, $1e^{-4}$ 1000, Cyclic)	(1.0, 0.5, $1e^{-4}$ 1000, Cyclic)	(1.0, 0.5, $1e^{-4}$ 1000, Cyclic)	(1.0, 0.25, $1e^{-3}$ 500, Cyclic)	-	-
Gradient Boost	n_{be}=[10,20,30,40,50,60, 80,100,150,200,300,400,500], lr=[1.0,0.1,0.01,0.001,0.0001], c=[mse,friedman-mse,mae]	(100, 0.01, friedman-mse)	(100, 0.01, friedman-mse)	(100, 0.01, friedman-mse)	(100, 0.01, friedman-mse)	-	-
LR	Penalty=[l_1, l_2], C=[1,10,100,200,500], tol=[$1e^{-5}$,$1e^{-4}$,$1e^{-3}$,$1e^{-2}$], mx_{it}=[50,100,200,300,400,500]	(l_2, 1, $1e^{-4}$, 100)	(l_2, 1, $1e^{-4}$, 100)	(l_2, 1, $1e^{-4}$, 200)	(l_2, 1, $1e^{-4}$, 200)	(auto)	(auto)
SGD	Penalty= [l_1,l_2] alpha=[1.0,0.1,0.001,0.0001], l_1=[1.0,0.75,0.5, 0.15, 0.1], $val_{fraction}$=[1.0,0.1,0.001,0.001], max_{it}=[100,200,500,1000,1500,2000], n_{it}=[1,5,10,15,20]	-	-	-	-	(l_2, 0.0001, 0.15, 0.1, 1000, 5)	(l_2, 0.0001, 0.15, 0.1, 1000, 5)
EN	alpha=[1.0,0.1,0.001,0.0001], l_1=[1.0,0.75,0.5, 0.15, 0.1], tol=[$1e^{-5}$,$1e^{-4}$,$1e^{-3}$,$1e^{-2}$], Selection=[Cyclic, Random]	-	-	-	-	(1.0, 0.5, $1e^{-4}$, cyclic)	(1.0, 0.5, $1e^{-4}$, cyclic)
ExtKNN	k=[5,6,7,8,9,10], B=[5,6,7,8,9,10]	(10, 10)	(8, 10)	(10, 10)	(10, 9)	(10, 10)	(7, 7)
RF	n_{be}=[50,100,150,200,250,300,350, 400,450,500,600,700,800,900,1000], c=[gini,entropy,mse,mae] max_{depth}=[50,100,200,300]	(350, entropy, Null)	(300, entropy, Null)	(500, entropy, Null)	(500, entropy, Null)	(450, mse, 50)	(450, mse, 50)
XGB	lr=[1.0,0.1,0.01,0.001,0.0001] subsample=[0,0.2,0.4,0.6,0.8,1], lambda=[$1e^{-5}$,$1e^{-4}$,$1e^{-4}$,$1e^{-3}$,$1e^{-2}$], cols=[0,0.2,0.4,0.6,0.8,1] Obj=[Bin:Logistic,multi:softprob]	(Bin:Logistic)	(Bin:Logistic)	(Bin:Logistic)	(Bin:Logistic)	(0.1, 0.8, $1e^{-5}$, 0.8)	(0.1, 0.8, $1e^{-5}$, 0.8)

datasets which have length less than 40 amino acids. Figure 5 shows position aware distributions of amino acids in soluble and insoluble classes using 4 benchmark datasets. To make visual analysis simple, we drop rare amino acids with position aware occurrence probability less than 0.07 in the sequences of whole dataset. Such type of rare patterns do not contribute in learning discriminative patterns, so while performing discriminative analysis we just drop them.

Overall in all 4 datasets, distribution of amino acids in both classes is almost similar such as in 3 datasets (PSI:biology, eSol and Ecoli) both soluble and insoluble classes amino acid 'L:leucine' occurs frequently. In price dataset along with L two other amino acids 'A:alanine' and 'G:glycine' also occur and as compared to other datasets in this dataset distribution is more discriminative. This distributional analysis reveals that encoding methods which consider only amino acids distributional information to generate statistical vectors will degrade the performance of classifiers. Because, statistical vectors generated through such encoding method, will not contain discriminative patterns and classifiers performance relies on the discriminative patterns present in input vectors. In a nutshell, it can be concluded that statistical vectors generated through the extraction of 4 different types of information (distributional, correlational, transitional and compositional) may facilitate classifiers to more precisely discriminate between soluble and insoluble classes.

B. PERFORMANCE ANALYSIS OF RF CLASSIFIER AT DIFFERENT SETTINGS OF PROPOSED ENCODER

The proposed encoder transforms raw protein sequences into statistical vectors by capturing four different types of information namely: physicochemical, distributional, compositional and transitional. With an aim to analyze whether correlational, distributional, transitional and compositional information can be more comprehensively captured at the global level from the full protein sequence or at the local level where first we divide the sequence into different-length subsequences and then from each subsequence, we extract

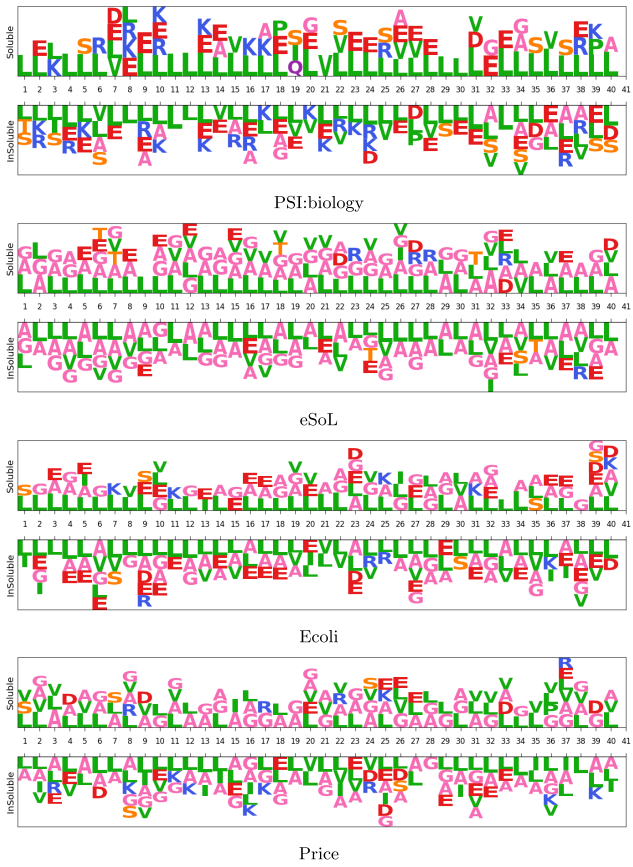


FIGURE 5. Amino acids distribution analysis among soluble and insoluble proteins using 4 benchmarkÂ datasets.

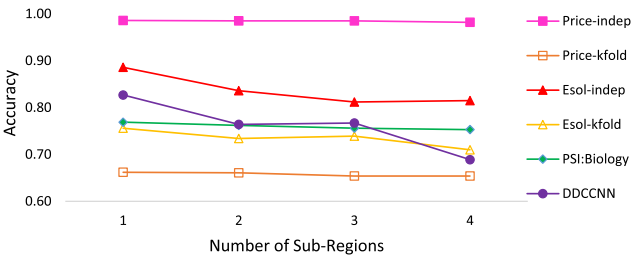


FIGURE 6. Over different benchmark datasets, Random Forest classifier performance analysis using subregions of sequences.

all four types of information and finally concatenate information of all subsequences. Furthermore, at the local level, we generate subsequences in 3 different settings. In 1^{st} setting we divide full sequences into 2 equal length subsequences, in 2^{nd} and 3^{rd} settings we generate 3 and 4 equal length subsequences, respectively. Figure 6 illustrates the performance values produced by the Random Forest classifier by taking statistical vectors generated through the proposed encoder by capturing sequence information at global and local levels. Over 3 datasets (price-indep, price k-fold and PSI biology) Random Forest classifier has produced a similar performance for both types of vectors generated at the global and local levels. Among the other 3 datasets, the Esol-indep dataset classifier produces better performance at the global level. Furthermore, at the local level among 3 different settings, the classifier produces better performance at settings 1.

In a nutshell, it can be concluded that, better statistical representations can be generated by capturing physicochemical, compositional, transitional and invariance information from full sequence, rather than taking subsequences. However, in particular task maximum length of sequence is around 1650. The idea of capturing information from subsequences may works better for tasks where sequence length is in several thousand.

C. INTRINSIC PERFORMANCE COMPARISON OF TRADITIONAL AND PROPOSED SEQUENCE ENCODERS

This section performs intrinsic performance comparison of proposed and 14 traditional encoders, where aim is to analyze the quality of statistical vectors generated through both proposed and traditional encoders. It is widely, accepted that if input samples contain discriminative features between different classes then a simple classifier can produced better performance, conversely, in case of non discriminative a sophisticated classifier may also remains fail to produce better performance [62]. This intrinsic analysis graphically visualize statistical vectors generated through proposed and traditional encoders, where prime objective is to analyze whether statistical vectors of soluble and insoluble classes are more separable for traditional encoders or proposed encoder.

Figure 7 illustrates the clusters of positive and negative classes across E.coli benchmark dataset, visualized by reducing generated statistical vectors to 20% dimensions using principal component analysis and 2 dimensions using t-distributed stochastic neighbor embedding. Graphical analysis of soluble and insoluble classes clusters across all traditional encoders indicates that existing encoder's statistical vectors remain fails to generate highly disjoint clusters. Unlike existing encoders, proposed encoder CTAPAAC generated statistical vectors produce disjoint clusters for both classes. Existing encoders remain fail to produce disjoint clusters because these encoders generate statistical vectors by utilizing physicochemical properties or extracting amino acids distribution information and does not take compositional and transitional invariances of amino acids into account which leads to extraction of limited inherent relationships. On the other hand, proposed CTAPAAC encoder extracts and fuses 4 different types of information that helps to generate statistical vectors having discriminative features in soluble and insoluble classes of proteins.

Supplementary file (I) contains graphical representation of positive and negative classes clusters for proposed and existing 14 encoders across 3 benchmark datasets namely E.sol, PSI:Biology and Price. Similar to E.coli dataset, over other 3 datasets existing encoders generate highly overlapping clusters while proposed encoder produce disjoint clusters. This intrinsic analysis across 4 different benchmark datasets, reveals that proposed encoder is generic and has potential for generating discriminative vectors among different classes.

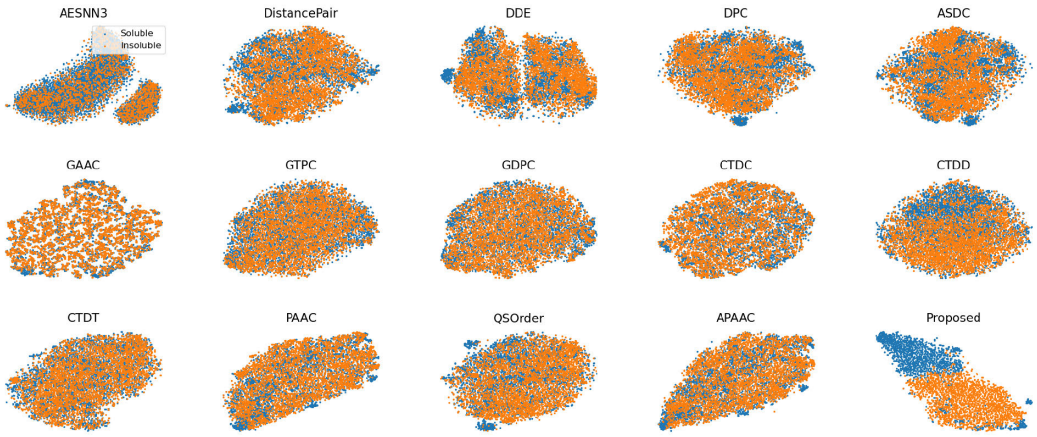


FIGURE 7. Intrinsic performance analysis of proposed CTAPAAC and existing sequence encoders over benchmark E.coli dataset.

D. EXTRINSIC PERFORMANCE COMPARISON OF TRADITIONAL AND PROPOSED SEQUENCE ENCODERS

This section performs a comprehensive performance comparison of the proposed protein sequence encoder with 14 most widely used protein sequence encoders using 7 different classifiers across four benchmark datasets.

Table 4 illustrates the accuracies of 7 different classifiers produced using statistical vectors generated through proposed encoder and 14 existing encoders that belongs to 4 different categories: 1 block substitution, 3 group amino acid distribution, 4 amino acid distribution, and 6 physicochemical properties based encoders. A critical analysis of Table 4 indicates that, on E.coli dataset, from existing 4 different categories sequence encoders, amino acid distribution based sequence encoders mark better performance across most classifiers followed by physicochemical properties, block substitution, and group based amino acid distribution based sequence encoders. From 4 amino acid distribution based sequence encoders, ASDC [89] sequence encoder achieves peak performance of 80.26% using RF classifier. From 6 physicochemical properties based sequence encoders, CTDD [29] achieves best performance of 75.27%, and block substitution based AESNN3 sequence encoder [52] marks the best performance of 72% using ET classifier. From 3 group based amino acid distribution based sequence encoders, GTPC [94] achieves the best performance of 63.03% using XGB classifier. Overall all sequence encoders achieve better performance with tree based classifiers. Among all sequence encoders, proposed sequence encoder CTAPAAC significantly outperforms existing encoders using RF classifier, achieving the peak performance of 86%.

On E.Sol dataset, physicochemical properties based sequence encoders perform best followed by amino acid distribution class encoders, group based amino acid distribution, and block substitution based sequence encoders, respectively. From physicochemical properties based encoders, QSOrder achieves the best performance of 84.67% using LR classifier and from amino acid distribution based encoders, DistancePair [54] sequence encoder achieves best performance of 82.63% using RF classifier. From group based encoders, GDPC [94] sequence encoder achieves the best performance of 78.25% using XGB classifier, block subsitution based encoder AESNN3 [52] achieves best performance of 71.53% using RF classifier. Overall, on E.Sol dataset, once again tree based classifiers achieve better performance across most sequence encoders. Proposed sequence encoder CTAPAAC outperforms all existing sequence encoders by a decent margin, achieving a peak performance of 89% using RF classifier.

On PSI-Biology dataset, amino acid distribution based sequence encoders perform better as compared to group based amino acid distribution, physicochemical properties based, and block substitution based sequence encoders. From these classes, DistancePair [54], GTPC [94], APAAC [22], and AESNN3 [52] achieve the performances of 77.04%, 76.28%, 76.03%, and 71.77% respectively using ET and RF classifiers. Among all, proposed sequence encoder CTAPAAC achieves the best performance of 77% using RF classifier.

On Price-Independent dataset, physicochemical properties based sequence encoders perform better than group based, standalone amino acid distribution, and block substitution based sequence encoders. From these classes, APAAC [22], GDPC [94], DistancePair [54], AESNN3 [52] achieve good performance around 98.5%, 98.41%, 98.34%, and 98.11% respectively using ET and RF classifiers. Among all, proposed sequence encoder CTAPAAC achieves the best performance of 99% using RF classifier.

In a nutshell, proposed sequence encoder CTAPAAC significantly outperforms all existing sequence encoders, by achieving an increment of 6%, 4%, 1%, and 1% on E.coli, E.Sol, PSI-Biology, and Price-Independent datasets using RF classifier. This is mainly due to the competence of the proposed sequence encoder for capturing four different types of information from protein sequences,

TABLE 4. Extrinsic performance comparison of 7 different classifiers by utilizing proposed and existing 14 encoders in terms of accuracy.

Dataset	Classifier	Block Substitution	Amino acids distribution				Group based			Physicochemical Properties based						Proposed CTAPAAC
		AESNN3 [52]	DistancePair [54]	ASDC [89]	DPC [13]	DDE [71]	GAAC [94]	GTPC [94]	GDPC [94]	CTDC [25]	CTDT [25]	CTDD [29]	QSOrder [73]	PAAC [21]	APAAC [22]	
Ecoli	AB	67.12	72.21	78.09	68.27	68.45	55.19	58.02	57.62	61.82	57.06	74.55	67.33	68.04	67.4	82.7
	DT	65.70	64.86	63.45	63.69	63.84	52.64	56.94	52.87	55.88	53.79	65.50	57.18	56.95	57.2	65.7
	ET	72.17	77.86	79.50	73.91	75.14	55.18	59.68	57.64	61.95	57.57	75.27	66.23	67.59	67.4	85.1
	GB	62.17	61.34	63.69	63.71	63.61	52.20	55.88	53.58	55.35	53.59	65.83	57.01	56.96	57.5	66.4
	LR	66.80	69.50	53.83	55.92	74.07	51.33	52.01	51.66	59.80	56.46	64.68	62.65	63.40	63.1	64.2
	XGB	70.86	76.59	77.79	74.89	75.45	55.10	63.03	58.16	62.97	60.05	73.86	67.16	67.55	65	83.2
	ExtKNN	45.49	43.62	43.62	44.91	50.31	49.45	52.93	51.65	52.31	52.37	52.46	52.65	52.46	52.65	52.93
	RF	69.00	75.81	80.26	69.20	69.47	54.96	58.32	57.61	62.35	57.92	74.72	67.07	67.94	68	**85.9**
Esol	DT	58.39	65.55	63.65	61.46	64.38	65.84	63.36	63.65	67.74	67.15	57.66	71.39	66.57	67.7	70.5
	LR	52.85	47.59	74.31	74.60	72.41	70.07	71.82	73.58	75.04	73.58	69.93	84.67	82.48	81.6	82
	SGD	60.58	59.42	57.52	57.52	69.78	57.66	57.52	57.52	70.95	68.47	42.77	74.01	55.04	42.5	45.4
	EN	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.52	57.5	57.5
	XGB	70.95	81.46	80.15	81.02	78.39	74.16	79.71	78.25	79.12	79.56	69.49	83.21	83.50	82.9	85.4
	ExtKNN	42.33	57.81	57.51	57.22	51.38	57.51	57.51	57.81	57.66	57.81	57.37	57.51	57.66	57.37	57.66
	RF	71.53	82.63	80.00	81.02	78.54	73.72	79.27	77.81	80.15	79.27	70.95	84.38	84.82	83.8	**88.6**
PSI:Biology	AB	71.75	76.83	76.15	76.78	76.02	67.83	75.66	74.22	74.27	74.18	70.18	76.15	76.11	76.1	**77**
	DT	63.53	68.81	66.95	69.07	67.53	60.71	66.79	65.05	65.42	65.30	61.21	67.83	68.17	66.7	67.7
	ET	71.63	77.04	76.42	77.05	76.88	67.42	76.28	74.55	75.09	75.17	71.65	76.19	75.92	75.3	77
	GB	64.90	69.74	67.56	69.62	69.58	61.89	67.62	66.03	65.99	66.00	62.00	68.48	69.27	69	69.1
	LR	64.58	68.22	66.81	66.81	70.17	66.65	66.79	66.66	67.49	66.54	66.69	68.94	69.77	69.8	70.1
	XGB	70.60	75.98	75.15	75.56	75.42	65.46	74.11	70.45	71.58	71.72	68.82	73.09	73.62	74	75
	ExtKNN	46.83	47.51	46.83	51.70	47.51	51.24	60.12	55.58	58.22	55.58	51.64	51.64	46.83	55.58	66.73
	RF	71.77	76.84	76.02	76.54	75.64	67.95	75.63	73.70	74.27	74.42	70.00	76.05	75.88	76.3	**77**
Price:Indep	AB	98.11	98.19	98.03	98.11	97.96	97.73	98.34	98.26	97.58	98.03	97.88	98.19	98.26	98.4	98.4
	DT	97.73	97.88	97.73	97.88	97.43	97.73	97.51	97.96	97.88	97.43	97.81	97.73	97.96	98	97.9
	ET	98.03	98.19	98.26	98.19	98.11	97.96	98.19	98.41	97.81	98.19	97.58	98.19	98.41	98.5	98.3
	GB	97.96	97.73	97.20	97.81	97.88	98.03	97.88	97.58	97.35	97.66	97.58	97.81	97.81	97.9	97.7
	LR	69.01	63.87	53.14	54.65	64.78	56.24	56.76	57.14	60.39	60.39	59.64	62.89	65.91	65.1	64.9
	XGB	97.51	97.96	98.34	98.03	97.28	87.23	97.58	97.51	96.90	97.43	97.96	97.35	97.05	96.5	98.1
	ExtKNN	63.64	63.64	63.71	63.79	63.71	63.71	63.41	63.71	63.49	63.64	63.79	62.96	63.64	63.64	63.49
	RF	98.11	98.34	98.34	97.96	98.26	97.88	98.11	98.19	98.03	98.41	97.81	98.26	98.41	96.6	**98.6**

specifically correlation, composition, distribution, and transition information of amino acids important for accurately predicting protein solubility. Overall, proposed encoder produce best performance across all datasets and from existing encoders from existing sequence encoders, amino acids distribution based sequence encoders achieve peak performance on two datasets (E.coli, PSI-Biology), and on the remaining two datasets (E.Sol, Price-Independent), physicochemical properties based sequence encoders achieve the peak performance. While block substitution-based sequence encoder lacks to capture the comprehensive discriminative distribution of amino acids present in sequences due to their reliance on experimentally identified values. Group based amino acid distribution based sequence encoders pay more attention to higher order residues dependencies, neglect the inter-relatedness of amino acids, their positions, and their context in the sequences. Furthermore, the prime reason behind the dominant performance of tree-based machine learning classifiers across all sequence encoders is that these classifiers make use of ensemble learning paradigm which extracts comprehensive non-linear and complex relationships of features and combines the predictions of multiple estimators to deduce the final predictions.

Supplementary file (II) contains 9 different evaluation measures based performance of 7 classifiers using statistical vectors generated through 14 existing and proposed encoders for 4 different benchmark datasets, accuracy measure, and other 8 measures in comparison to existing encoders proposed encoder produce better performance. Based on performance analysis across different evaluation measures and benchmark datasets, it can be concluded that proposed CTAPAAC encoder, generates more comprehensive statistical vectors of protein sequences by compositional, transitional, distributional and correlational information, from raw protein sequences.

Furthermore, to evaluate the generalization of proposed encoder over 4 benchmark datasets, we generate statistical vectors through proposed encoder and feed them to 7 different machine learning classifiers to perform statistical testing. Here we compute probability value (p-value) that represents the fraction of randomized datasets. Primarily, this significance test makes sure if distribution of amino acids in raw sequences slightly changes then whether proposed encoder will manage to generate same quality statistical vectors and further whether classifiers will be able to produce similar performance values. To add slight noise in the datasets,

sequences of the datasets are permuted. Over 4 benchmark datasets, all classifiers produce p-value less than 0.1 with 10 permutations and almost 0 value with 1000 permutations. The lowest p-value reveals that proposed encoder has potential to generate comprehensive and discriminative statistical representation for soluble and insoluble protein sequences even when raw sequences are slightly noisy.

E. PERFORMANCE COMPARISON OF PROPOSED CTAPAAC PREDICTOR WITH EXISTING PREDICTORS OVER CORE DATASETS

This section performs a comprehensive extrinsic performance comparison of proposed predictor with existing 19 predictors (SoluProt [38], ParsnIP [68], CamSol [78], DeepSol [41], ProteinSol [35], SWI [12], ESM-MSA-P [84], Prot5-P [84], ESM1b-F [84], CCSol [2], SolPro [57], PROSO II [75], PARSnIP [68], DDcCNN [88], NetSolP [84], ProGan [33], SeqVec [36], TAPE [67] and GraphSol [18]) over 4 benchmark datasets. To perform a fair performance comparison of proposed predictor with existing predictors [2], [12], [18], [33], [35], [36], [38], [41], [57], [67], [68], [75], [78], [84], [88], following experimental settings of existing predictors for PSI:biology [12] and price datasets, we computed performance of proposed predictor in 5 folds cross-validation settings. Furthermore, for E.coli dataset, following evaluation criteria of existing predictors [2], [12], [18], [33], [35], [36], [38], [41], [57], [67], [68], [75], [78], [84], [88], proposed predictor is evaluated on standard split where 90% sequence samples are used for predictor training and 10% sequences are used to test the predictor.

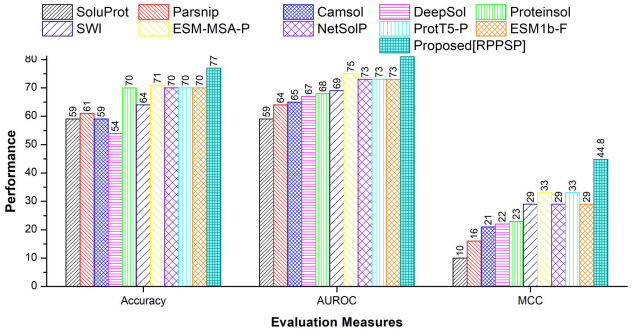


FIGURE 8. Performance comparison of proposed RPPSP and 10 existing predictors using benchmark PSI:biology dataset.

Figure 8 compares the performance of proposed predictor and 10 existing predictors [12], [35], [38], [41], [68], [78], [84] in terms of accuracy, AUROC and MCC. It is evident in the Figure 8 that from existing predictors, ESM-MSA-P predictor [84] that makes use of 33-layer evolutionary scale language model and a linear classifier achieves the best performance of 71%, 75% and 33% in terms of accuracy, AUROC and MCC. Other language modeling based approaches including ProT5-P [84], ESM1b-F [84] and NetSolP [84] utilize 24-layers and linear classifiers to achieve second best predictive performance among all existing predictors. Third best performance across most evaluation metrics is achieved by amino acid composition based predictors such as ProteinSol [35] and SWI [12]. Among all amino acid composition based predictors, SoluProt [38] achieves lower performance as well as overall lowest AUROC and MCC. From existing two amino acid structural information based predictors, DeepSol [41] marks better AUROC and MCC, whereas Parsnip achieves better accuracy. The only existing amino acid physicochemical properties based predictor achieves similar performance figures as achieved by amino acids secondary structural information based predictors.

Among all predictors, proposed predictor outperforms language modeling based predictors by the accuracy and AUROC of 6% and MCC of 12%. It beats amino acids physicochemical properties and structural information based predictors by 7%, 12%, 16% in terms of accuracy, AUROC and MCC, respectively. This is primarily due to the use of an optimal sequence encoder which unlike existing sequence encoders fuses composition and transition information to capture heterogeneous short and long relations of amino acids which are important to accurately predict protein solubility.

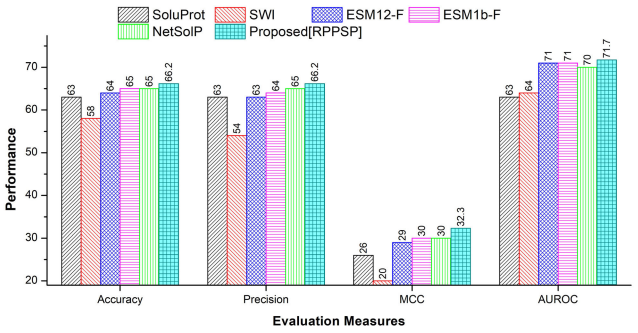


FIGURE 9. Performance comparison of proposed RPPSP and 10 existing predictors using benchmark Price dataset.

Figure 9 compares performance of proposed and 5 existing predictors [12], [38], [84] over Price dataset in terms of four different evaluation metrics. As indicated by the Figure 9, from existing predictors, deep evolutionary scale language modeling based predictors such as NetsolP [84], ESM1b-F [84] and ESM12-F [84] achieve better performance followed by amino acid composition based predictors namely SoluProt and SWI. Among all existing predictors, SWI [12] achieves the lowest accuracy, precision and MCC whereas SoluProt marks the lowest AUROC.

Proposed protein solubility predictor outperforms language modeling based predictors by accuracy and precision of 1%, MCC and AUROC of 2% and amino acid composition based predictors by accuracy and precision of 3%, MCC of 6% and AUROC of 8% due to effective characterization performed using novel sequence encoder.

Figure 10 performs performance comparison of proposed predictor with 6 existing predictors [2], [41], [57], [68], [75], [88] over E.coli dataset in terms of accuracy, sensitivity,

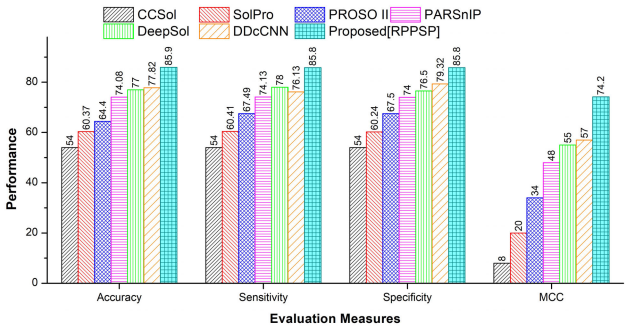


FIGURE 10. Performance comparison of proposed RPPSP and 6 existing predictors using E.coli benchmark dataset.

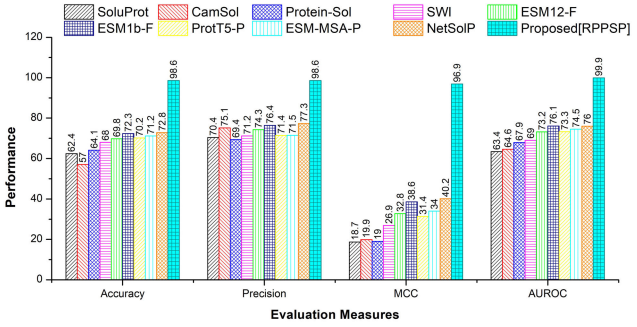


FIGURE 11. Over Price Independent test set, performance comparison of proposed RPPSP and 9 existing predictors.

specificity and MCC. It can be seen in Figure 10 that among three amino acid composition based predictors, DDcCNN achieves better performance followed by PROSOII [75] and SolPro [57]. DDcCNN achieves peak performance among all existing predictors in terms of accuracy, specificity and MCC, however, better sensitivity is achieved by amino acid structural information based predictor namely DeepSol [41]. Overall, second best performance from existing predictors is also achieved by DeepSol followed by another amino acid structural information based predictor namely Parnsip. Overall, physicochemical properties based predictor CCSol achieves the lowest accuracy, specificity, sensitivity and MCC.

Among all approaches, proposed predictor outperforms all existing predictors across all four evaluation metrics. It achieves the accuracy increment of 8%, sensitivity increment of 10%, specificity increment of 7% and MCC increment of 17% as compared to amino acid composition based predictors. It outperforms amino acid structure based predictors by an average of 11% and physicochemical properties based predictors by an average of 41% mainly due to the supreme effectiveness of novel sequence encoder.

F. PERFORMANCE COMPARISON OF PROPOSED CTAPAAC PREDICTOR WITH EXISTING PREDICTORS OVER TWO INDEPENDENT TEST SETS

Following the evaluation criteria of existing studies related to protein solubility prediction, we evaluate proposed predictor on two independent test sets. We train the proposed predictor on full core Price dataset to test the predictor on Price independent test set. Similarly, proposed predictor is trained on full E.Sol dataset and tested on E.Sol independent test set.

Figure 11 indicates the performance produced by proposed predictor and 9 existing predictors [12], [35], [38], [78], [84], [85] in terms of accuracy, precision, MCC and AUROC. Unlike core datasets, here, a variety of predictors including amino acid structure information based predictors, physicochemical properties based predictors and amino acid composition based predictors mark quite similar performance. However, once again peak performance across all four evaluation metrics is achieved by evolutionary scale language modeling based predictors, achieving the best accuracy of 73%, precision of 77%, MCC of 40.2% and AUROC of 76%. From all existing predictors, amino acid composition based predictor achieves the lowest performance across most evaluation metrics. Like all core datasets, on Price independent dataset, proposed predictor achieves optimal predictive performance across all four evaluation metrics, achieving 99% accuracy, precision and AUROC and 97% MCC.

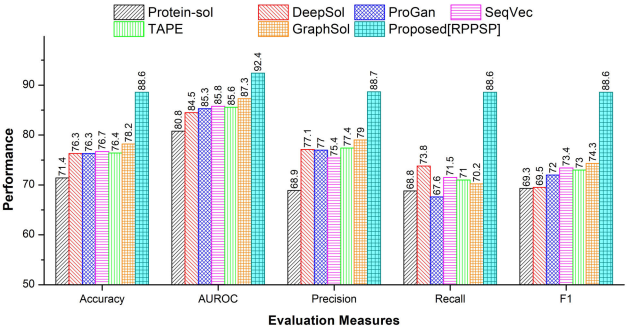


FIGURE 12. Over E.Sol Independent test set, performance comparison of proposed RPPSP and 6 existing predictors.

Furthermore, Figure 12 illustrates the performance of proposed and six existing predictors produced over E.Sol independent dataset in terms of five different evaluation metrics. Like Price independent datasets, here, once again, all existing predictors show quite similar performance trends except amino acid composition based ProteinSol. Performance of most existing approaches falls around 78%, 87%, 79%, 71% and 74% in terms of accuracy, AUROC, precision, recall and F1-score. Proposed protein solubility predictor once again outperforms all existing predictors achieves more than 88% performance across all distinct evaluation metrics. Overall, it achieves the increment of 11%, 5%, 10%, 19% and 15% in terms of accuracy, AUROC, precision, recall and F1-score.

In a nutshell, a comprehensive performance comparison of proposed protein solubility predictor with a variety of existing predictors using different benchmark core and independent datasets proves the dominance of proposed approach. It achieves decent performance increment on core datasets and huge performance increments on independent datasets, indicating great generalizability. Prime reason of supreme

predictive performance and generalizability of proposed predictor is the use of more effective sequence encoder that unlike existing sequence encoders manages to capture discriminative distribution of amino acids in highly variable length protein sequences that are important to accurately predict protein solubility across different species.

VI. LIMITATIONS OF PROPOSED ENCODER

Proposed encoder transforms raw protein sequences into statistical vectors by extracting 4 different types of information namely correlation, distribution, composition and transition. However, while extracting 4 different types of information generated vectors may contain some redundant information. This redundant information may hinder the predictive performance of machine learning classifiers. To fully utilize the potential of proposed encoder, we believe in overall predictive pipeline induction of appropriate feature selection method may further improve the performance of proposed solubility predictor.

VII. CONCLUSION

Protein solubility prediction plays important role in understanding diverse types of intracellular and pathological processes and pave way for the development of novel therapies and drugs. Considering the need of statistical encoder that discretizes protein sequences, this paper presents a novel sequence encoder that transforms raw protein sequences into statistical vectors by extracting 4 different types of information namely correlation, distribution, transition and composition. A comprehensive experimentation, over 4 benchmark datasets reveals that proposed encoder significantly improves the performance of 7 different machine learning classifiers. Overall, proposed encoder along with random forest classifier outperforms existing predictors over 4 benchmark datasets namely: PSI:Biology, price (5fold), e.coli, price (independent), esol (independent) with a significant margin of 6%, 1%, 7%, 25% and 10% in terms of accuracy, respectively. As proposed encoder has potential to extract diverse types of information from raw protein sequences while generating statistical representation, so we believe proposed encoder can be utilized to perform other protein sequence analysis tasks such as protein family classification and protein subcellular location prediction. To enhance the future capabilities of the proposed predictor, an important direction is to involve incorporating an appropriate feature selection method. This selection method would help in effectively filtering out redundant information from the generated vectors. By removing unnecessary or redundant features, the proposed predictor can focus on the most relevant and informative ones, which would ultimately lead to improved performance and accuracy.

COMPETING INTERESTS

Authors declare that there is no known competing financial interest or personal relationships which could have influenced this article.

DATA AVAILABILITY

Code and benchmark datasets are available at: https://github.com/Faiza-Mehmood/RPPSP.git

REFERENCES

[1] Y. Adachi, "Dynamic aspects of coagulation and flocculation," *Adv. Colloid Interface Sci.*, vol. 56, pp. 1–31, Mar. 1995.
[2] F. Agostini, D. Cirillo, C. M. Livi, R. D. Ponti, and G. G. Tartaglia, "*Cc* SOL *omics*: A webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*," *Bioinformatics*, vol. 30, no. 20, pp. 2975–2977, Oct. 2014.
[3] F. Agostini, M. Vendruscolo, and G. G. Tartaglia, "Sequence-based prediction of protein solubility," *J. Mol. Biol.*, vol. 421, nos. 2–3, pp. 237–241, Aug. 2012.
[4] A. Ali, M. Hamraz, N. Gul, D. M. Khan, S. Aldahmani, and Z. Khan, "A *k* nearest neighbour ensemble via extended neighbourhood rule and feature subsets," *Pattern Recognit.*, vol. 142, Oct. 2023, Art. no. 109641.
[5] M. N. Asim, A. Fazeel, M. A. Ibrahim, A. Dengel, and S. Ahmed, "MP-VHPPI: Meta predictor for viral host protein-protein interaction prediction in multiple hosts and viruses," *Frontiers Med.*, vol. 9, Nov. 2022, Art. no. 1025887.
[6] M. N. Asim, M. Ali Ibrahim, M. I. Malik, A. Dengel, and S. Ahmed, "ChrSLoc-Net: Machine learning-based prediction of channelrhodopsins proteins within plasma membrane," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.
[7] M. N. Asim, M. A. Ibrahim, M. I. Malik, A. Dengel, and S. Ahmed, "ADH-PPI: An attention-based deep hybrid model for protein-protein interaction prediction," *iScience*, vol. 25, no. 10, Oct. 2022, Art. no. 105169.
[8] M. N. Asim, M. A. Ibrahim, M. I. Malik, A. Dengel, and S. Ahmed, "LGCA-VHPPI: A local-global residue context aware viral-host protein-protein interaction predictor," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0270275.
[9] M. N. Asim, M. I. Malik, C. Zehe, J. Trygg, A. Dengel, and S. Ahmed, "A robust and precise ConvNet for small non-coding RNA classification (RPC-snRC)," *IEEE Access*, vol. 9, pp. 19379–19390, 2021.
[10] A. Banga, R. Ahuja, and S. C. Sharma, "Stacking regression algorithms to predict $PM_{2.5}$ in the smart city using Internet of Things," *Recent Adv. Comput. Sci. Commun.*, vol. 15, no. 1, pp. 60–76, Feb. 2022.
[11] H. M. Berman, C. L. Lawson, B. Vallat, and M. J. Gabanyi, "Anticipating innovations in structural biology," *Quart. Rev. Biophys.*, vol. 51, p.e8, Jul. 2018.
[12] B. K. Bhandari, P. P. Gardner, and C. S. Lim, "Solubility-weighted index: Fast and accurate prediction of protein solubility," *Bioinformatics*, vol. 36, no. 18, pp. 4691–4698, Sep. 2020.
[13] M. Bhasin and G. P. S. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *J. Biol. Chem.*, vol. 279, no. 22, pp. 23262–23266, May 2004.
[14] S. Carta, A. Corriga, A. Ferreira, D. R. Recupero, and R. Saia, "A holistic auto-configurable ensemble machine learning strategy for financial trading," *Computation*, vol. 7, no. 4, p. 67, Nov. 2019.
[15] W.-C. Chan, P.-H. Liang, Y.-P. Shih, U.-C. Yang, W.-C. Lin, and C.-N. Hsu, "Learning to predict expression efficacy of vectors in recombinant protein production," *BMC Bioinf.*, vol. 11, no. S1, pp. 1–12, Jan. 2010.
[16] R. Chandra, S. Shukla, and M. Kumar, "The hydropericardium syndrome and inclusion body hepatitis in domestic fowl," *Tropical Animal Health Prod.*, vol. 32, no. 2, pp. 99–111, 2000.
[17] C. C. H. Chang, J. Song, B. T. Tey, and R. N. Ramanan, "Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: Protein solubility prediction," *Briefings Bioinf.*, vol. 15, no. 6, pp. 953–962, Nov. 2014.
[18] J. Chen, S. Zheng, H. Zhao, and Y. Yang, "Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map," *J. Cheminformatics*, vol. 13, no. 1, pp. 1–10, Dec. 2021.
[19] Y.-Z. Chen, Z. Chen, Y.-A. Gong, and G. Ying, "SUMOhydro: A novel method for the prediction of sumoylation sites based on hydrophobic properties," *PLoS ONE*, vol. 7, no. 6, Jun. 2012, Art. no. e39195.
[20] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y.-Z. Chen, T. Akutsu, R. J. Daly, G. I. Webb, Q. Zhao, L. Kurgan, and J. Song, "ILearnPlus: A comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization," *Nucleic Acids Res.*, vol. 49, no. 10, p. e60, Jun. 2021.

[21] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins, Struct., Function, Genet.*, vol. 43, no. 3, pp. 246–255, 2001.

[22] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005.

[23] Z. Dische, "A new specific color reaction of hexuronic acids," *J. Biol. Chem.*, vol. 167, no. 1, pp. 189–198, Jan. 1947.

[24] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, Oct. 2009.

[25] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence.," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[26] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, p. 505, 2017.

[27] B. Fahnert, H. Lilie, and P. Neubauer, "Inclusion bodies: Formation and utilisation," in *Physiological Stress Responses in Bioprocesses*. Sweden: Springer, 2004, pp. 93–142.

[28] Y. Fang and J. Fang, "Discrimination of soluble and aggregation-prone proteins based on sequence information," *Mol. BioSyst.*, vol. 9, no. 4, pp. 806–811, 2013.

[29] G. Govindan and A. S. Nair, "Composition, transition and distribution (CTD)—A dynamic feature for predictions based on hierarchical structure of cellular sorting," in *Proc. Annu. IEEE India Conf.*, Dec. 2011, pp. 1–6.

[30] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Res.*, vol. 36, no. 9, pp. 3025–3030, May 2008.

[31] N. Habibi, S. Z. M. Hashim, A. Norouzi, and M. R. Samian, "A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in *Escherichia coli*," *BMC Bioinf.*, vol. 15, no. 1, pp. 1–16, Dec. 2014.

[32] S. H. A. Hamid, F. Lananan, H. Khatoon, A. Jusoh, and A. Endut, "A study of coagulating protein of moringa oleifera in microalgae bio-flocculation," *Int. Biodeterioration Biodegradation*, vol. 113, pp. 310–317, Sep. 2016.

[33] X. Han, L. Zhang, K. Zhou, and X. Wang, "ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework," *Comput. Chem. Eng.*, vol. 131, Dec. 2019, Art. no. 106533.

[34] W. He, C. Jia, and Q. Zou, "4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction," *Bioinformatics*, vol. 35, no. 4, pp. 593–601, Feb. 2019.

[35] M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis, and J. Warwicker, "Protein–Sol: A web tool for predicting protein solubility from sequence," *Bioinformatics*, vol. 33, no. 19, pp. 3098–3100, Oct. 2017.

[36] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–17, Dec. 2019.

[37] J. R. Hepler and A. G. Gilman, "G proteins," *Trends Biochem. Sci.*, vol. 17, no. 10, pp. 383–387, 1992.

[38] J. Hon, M. Marusiak, T. Martinek, A. Kunka, J. Zendulka, D. Bednar, and J. Damborsky, "SoluProt: Prediction of soluble protein expression in *Escherichia coli*," *Bioinformatics*, vol. 37, no. 1, pp. 23–28, Apr. 2021.

[39] Q. Hou, J. M. Kwasigroch, M. Rooman, and F. Pucci, "SOLart: A structure-based method to predict protein solubility and aggregation," *Bioinformatics*, vol. 36, no. 5, pp. 1445–1452, 2020.

[40] H.-L. Huang, P. Charoenkwan, T.-F. Kao, H.-C. Lee, F.-L. Chang, W.-L. Huang, S.-J. Ho, L.-S. Shu, W.-L. Chen, and S.-Y. Ho, "Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition," *BMC Bioinf.*, vol. 13, no. S17, pp. 1–14, Dec. 2012.

[41] S. Khurana, R. Rawi, K. Kunji, G.-Y. Chuang, H. Bensmail, and R. Mall, "DeepSol: A deep learning framework for sequence-based protein solubility prediction," *Bioinformatics*, vol. 34, no. 15, pp. 2605–2613, Aug. 2018.

[42] M. Koklu and K. Tutuncu, "Tree based classification methods for occupancy detection," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 675. Bristol, U.K.: IOP Publishing, 2019, Art. no. 012032.

[43] R. R. Kopito, "Aggresomes, inclusion bodies and protein aggregation," *Trends Cell Biol.*, vol. 10, no. 12, pp. 524–530, Dec. 2000.

[44] H. Korhonen, A. Pihlanto-Leppäla, P. Rantamäki, and T. Tupasela, "Impact of processing on bioactive proteins and peptides," *Trends Food Sci. Technol.*, vol. 9, nos. 8–9, pp. 307–319, 1998.

[45] B. S. Kumar, R. Cristin, K. Karthick, and T. Daniya, "Study of shadow and reflection based image forgery detection," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2019, pp. 1–5.

[46] P. Kumari and D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," *J. Cleaner Prod.*, vol. 318, Oct. 2021, Art. no. 128566.

[47] R. Kunert and D. Reinhart, "Advances in recombinant antibody manufacturing," *Appl. Microbiology Biotechnol.*, vol. 100, no. 8, pp. 3451–3461, Apr. 2016.

[48] M. R. Ladisch and K. L. Kohlmann, "Recombinant human insulin," *Biotechnol. Prog.*, vol. 8, no. 6, pp. 469–478, 1992.

[49] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, 2006.

[50] D. Lee, R. Karchin, and M. A. Beer, "Discriminative prediction of mammalian enhancers from DNA sequence," *Genome Res.*, vol. 21, no. 12, pp. 2167–2180, Dec. 2011.

[51] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, and Y.-Y. Ou, "Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites," *PLoS ONE*, vol. 6, no. 3, Mar. 2011, Art. no. e17331.

[52] B. Liu, X. Gao, and H. 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.*, vol. 47, no. 20, p. e127, Nov. 2019.

[53] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "RepDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, Apr. 2015.

[54] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, and K.-C. Chou, "IDNA-Prot|dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Sep. 2014, Art. no. e106691.

[55] Y.-X. Liu, X. Liu, C. Cen, X. Li, J.-M. Liu, Z.-Y. Ming, S.-F. Yu, X.-F. Tang, L. Zhou, J. Yu, K.-J. Huang, and S.-S. Zheng, "Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study," *Hepatobiliary Pancreatic Diseases Int.*, vol. 20, no. 5, pp. 409–415, Oct. 2021.

[56] M. Madani, K. Lin, and A. Tarakanova, "DSResSol: A sequence-based solubility predictor created with dilated squeeze excitation residual networks," *Int. J. Mol. Sci.*, vol. 22, no. 24, p. 13555, Dec. 2021.

[57] C. N. Magnan, A. Randall, and P. Baldi, "SOLpro: Accurate sequence-based prediction of protein solubility," *Bioinformatics*, vol. 25, no. 17, pp. 2200–2207, Sep. 2009.

[58] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," 2017, *arXiv:1704.04289*.

[59] F. Marra, G. Poggi, F. Roli, C. Sansone, and L. Verdoliva, "Counter-forensics in machine learning based forgery detection," *Proc. SPIE*, vol. 9409, pp. 181–191, Mar. 2015.

[60] F. Mehmood, M. U. Ghani, M. N. Asim, R. Shahzadi, A. Mehmood, and W. Mahmood, "MPF-Net: A computational multi-regional solar power forecasting framework," *Renew. Sustain. Energy Rev.*, vol. 151, Nov. 2021, Art. no. 111559.

[61] M. A. Muslim, "Support vector machine (SVM) optimization using grid search and unigram to improve e-commerce review accuracy," *J. Soft Comput. Exp.*, vol. 1, no. 1, pp. 8–15, 2020.

[62] M. Nabeel Asim, M. Ali Ibrahim, A. Fazeel, A. Dengel, and S. Ahmed, "DNA-MP: A generalized DNA modifications predictor for multiple species based on powerful sequence encoding method," *Briefings Bioinf.*, vol. 24, no. 1, Jan. 2023, Art. no. bbac546.

[63] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi, "Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 11, pp. 4201–4206, Mar. 2009.

[64] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[65] W. N. Price, S. K. Handelman, J. K. Everett, S. N. Tong, A. Bracic, J. D. Luff, V. Naumov, T. Acton, P. Manor, R. Xiao, B. Rost, G. T. Montelione, and J. F. Hunt, "Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. Coli*," *Microbial Informat. Experimentation*, vol. 1, no. 1, pp. 1–20, Dec. 2011.

[66] I. Priyadarshini and C. Cotton, "A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis," *J. Supercomput.*, vol. 77, no. 12, pp. 13911–13932, Dec. 2021.

[67] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, "Evaluating protein transfer learning with tape," in *Proc. Adv. Neural Inf. Process. Syst.*, 32, 2019, pp. 1–11.

[68] R. Rawi, R. Mall, K. Kunji, C.-H. Shen, P. D. Kwong, and G.-Y. Chuang, "PaRSnIP: Sequence-based protein solubility prediction using gradient boosting machine," *Bioinformatics*, vol. 34, no. 7, pp. 1092–1098, Apr. 2018.

[69] P. S. Regojo, "Burn care basics: How to extinguish problems," *Nursing*, vol. 33, no. 3, pp. 50–53, Mar. 2003.

[70] T. Samak, D. Gunter, and Z. Wang, "Prediction of protein solubility in *E. coli*," in *Proc. IEEE 8th Int. Conf. E-Sci.*, Oct. 2012, pp. 1–8.

[71] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor," *OMICS, A J. Integrative Biol.*, vol. 19, no. 10, pp. 648–658, Oct. 2015.

[72] C. H. Schein, "Solubility and secretability," *Current Opinion Biotechnol.*, vol. 4, no. 4, pp. 456–461, Aug. 1993.

[73] G. Schneider and P. 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," *Biophysical J.*, vol. 66, no. 2, pp. 335–344, Feb. 1994.

[74] C. Y. Seiler, J. G. Park, A. Sharma, P. Hunter, P. Surapaneni, C. Sedillo, J. Field, R. Algar, A. Price, J. Steel, A. Throop, M. Fiacco, and J. LaBaer, "DNASU plasmid and PSI: Biology-materials repositories: Resources to accelerate biological research," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1253–D1260, Jan. 2014.

[75] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann, and D. Frishman, "PROSO II—A new method for protein solubility prediction," *FEBS J.*, vol. 279, no. 12, pp. 2192–2200, Jun. 2012.

[76] P. Smialowski, A. J. Martin-Galiano, A. Mikolajka, T. Girschick, T. A. Holak, and D. Frishman, "Protein solubility: Sequence based prediction and experimental verification," *Bioinformatics*, vol. 23, no. 19, pp. 2536–2542, Oct. 2007.

[77] H. P. Sørensen and K. K. Mortensen, "Advanced genetic strategies for recombinant protein expression in *Escherichia coli*," *J. Biotechnol.*, vol. 115, no. 2, pp. 113–128, Jan. 2005.

[78] P. Sormanni, F. A. Aprile, and M. Vendruscolo, "The CamSol method of rational design of protein mutants with enhanced solubility," *J. Mol. Biol.*, vol. 427, no. 2, pp. 478–490, Jan. 2015.

[79] W. R. Springer and D. E. Koshland, "Identification of a protein methyltransferase as the cheR gene product in the bacterial sensing system.," *Proc. Nat. Acad. Sci. USA*, vol. 74, no. 2, pp. 533–537, Feb. 1977.

[80] M. Stricker, M. N. Asim, A. Dengel, and S. Ahmed, "CircNet: An encoder–decoder-based convolution neural network (CNN) for circular rna identification," *Neural Comput. Appl.*, vol. 34, pp. 1–12, Jan. 2021.

[81] B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 1–6, 2020.

[82] G. Taherzadeh, M. Campbell, and Y. Zhou, "Computational prediction of N- and O-linked glycosylation sites for human and mouse proteins," in *Computational Methods for Predicting Post-Translational Modification Sites*. Cham, Switzerland: Springer, 2022, pp. 177–186.

[83] A. Tareen and J. B. Kinney, "Logomaker: Beautiful sequence logos in Python," *Bioinformatics*, vol. 36, no. 7, pp. 2272–2274, Apr. 2020.

[84] V. Thumuluri, H.-M. Martiny, J. J. A. Armenteros, J. Salomon, H. Nielsen, and A. R. Johansen, "NetSolP: Predicting protein solubility in *Escherichia coli* using language models," *Bioinformatics*, vol. 38, no. 4, pp. 941–946, Jan. 2022.

[85] V. Thumuluri, H.-M. Martiny, J. J. A. Armenteros, J. Salomon, H. Nielsen, and A. Johansen, "NetSolP: predicting protein solubility in *E. coli* using language models," *Bioinformatics*, vol. 38, no. 4, pp. 941–946, 2022.

[86] H. Tjong and H.-X. Zhou, "Prediction of protein solubility from calculation of transfer free energy," *Biophysical J.*, vol. 95, no. 6, pp. 2601–2609, Sep. 2008.

[87] M. Vihinen, E. Torkkila, and P. Riikonen, "Accuracy of protein flexibility predictions," *Proteins, Struct., Function, Genet.*, vol. 19, no. 2, pp. 141–149, Jun. 1994.

[88] X. Wang, Y. Liu, Z. Du, M. Zhu, A. C. Kaushik, X. Jiang, and D. Wei, "Prediction of protein solubility based on sequence feature fusion and DDc-CNN," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 13, no. 4, pp. 703–716, Dec. 2021.

[89] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, Dec. 2018.

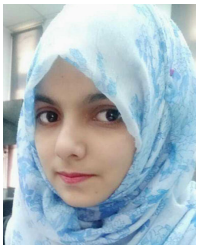
[90] P. T. Wingfield, "Overview of the purification of recombinant proteins," *Current Protocols Protein Sci.*, vol. 80, no. 1, pp. 1–6, Apr. 2015.

[91] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping," *CATENA*, vol. 187, Apr. 2020, Art. no. 104396.

[92] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[93] J. F. Zayas, "Solubility of proteins," in *Functionality of Proteins in Food*. Cham, Switzerland: Springer, 1997, pp. 6–75.

[94] C. Zhou, C. Wang, H. Liu, Q. Zhou, Q. Liu, Y. Guo, T. Peng, J. Song, J. Zhang, L. Chen, Y. Zhao, Z. Zeng, and D.-X. Zhou, "Identification and analysis of adenine N^6-methylation sites in the rice genome," *Nature Plants*, vol. 4, no. 8, pp. 554–563, Jul. 2018.



FAIZA MEHMOOD is currently pursuing the Ph.D. degree in computer science (bio-informatics) with the University of Engineering and Technology at Lahore, Lahore, Pakistan. Since 2019, she has been a Senior Researcher (Team Lead) with the Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology at Lahore. She has completed several natural language processing, bio-informatics, and energy projects. She has published various articles in well-reputed journals.

SHAZIA ARSHAD is currently a Professor with the Department of Computer Science, University of Engineering and Technology at Lahore, Lahore, Pakistan. Her research interests include software engineering and bio-informatics.

MUHAMMAD SHOAIB is currently the Dean and a Professor with the Department of Computer Science, University of Engineering and Technology at Lahore, Lahore, Pakistan. His research interests include information retrieval, software metrics, and bio-informatics.