Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07868-0

ORIGINAL ARTICLE

Energy Profile Bayes and Thompson Optimized Convolutional Neural Network protein structure prediction

Varanavasi Nallasamy[1] · Malarvizhi Seshiah[2]

Received: 6 August 2021 / Accepted: 21 September 2022

Abstract
In living organisms, proteins are considered as the executants of biological functions. Owing to its pivotal role played in protein folding patterns, comprehension of protein structure is a challenging issue. Moreover, owing to numerous protein sequence exploration in protein data banks and complication of protein structures, experimental methods are found to be inadequate for protein structural class prediction. Hence, it is very much advantageous to design a reliable computational method to predict protein structural classes from protein sequences. In the recent few years there has been an elevated interest in using deep learning to assist protein structure prediction as protein structure prediction models can be utilized to screen a large number of novel sequences. In this regard, we propose a model employing Energy Profile for atom pairs in conjunction with the Legion-Class Bayes function called Energy Profile Legion-Class Bayes Protein Structure Identification model. Followed by this, we use a Thompson Optimized convolutional neural network to extract features between amino acids and then the Thompson Optimized SoftMax function is employed to extract associations between protein sequences for predicting secondary protein structure. The proposed Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method tested distinct unique protein data and was compared to the state-of-the-art methods, the Template-Based Modeling, Protein Design using Deep Graph Neural Networks, a deep learning-based S-glutathionylation sites prediction tool called a Computational Framework, the Deep Learning and a distance-based protein structure prediction using deep learning. The results obtained when applied with the Biopython tool with respect to protein structure prediction time, protein structure prediction accuracy, specificity, recall, F-measure, and precision, respectively, are measured. The proposed EPB-OCNN method outperformed the state-of-the-art methods, thereby corroborating the objective.

Keywords Energy profile · Legion-Class Bayes · Protein structure identification · Thompson optimization · Convolutional neural network · Secondary structure prediction

1 Introduction

Proteins are the macromolecules that are almost universally in charge of bestowing out the numerous functionalities requisite to endure life, cell structural underpinning, immune safeguarding, enzymatic catalysis, cell signal transduction, and translation control. These numerous functionalities are made feasible by the distinctive three-dimensional structures applied by distinct protein molecules. The objective of protein structure prediction methods is to make use of computational representation to govern the spatial location of every atom in a protein molecule beginning from only its sequences of amino acid. Based on the homologous structures present in the Protein Data Bank (PDB), numerous protein structure models have been generally classified as template-based modeling (TBM) or template-free modeling (FM) approaches. Template-based modeling (TBM) was proposed in [1] via deep learning techniques, therefore increasing the precision in a significant manner. Even though a significant amount of precision

✉ Varanavasi Nallasamy
varanavasi.nallasamy@cognizant.com

[1] Cognizant Technology Solutions Pvt. Ltd, CHIL SEZ IT Park, Keeranatham, Saravanam Patti, Coimbatore, Tamil Nadu 641035, India

[2] Department of Computer Science, Thiruvalluvar Government Arts College, Rasipuram, Namakkal, Tamil Nadu, India

Published online: 07 October 2022

was said to be attained, there is still an issue with the cost of accuracy. Nevertheless, the swift improvement observed over the past few years alone bestows certain types of assistance in holistic protein structure prediction issue that may be addressed by employing deep learning, where predictions may continually attain accuracies that even may go on par with the experimental methods.

The functioning and structuring of protein are deliberated by the positioning of the linear arrangement of amino acids in 3D space. Protein design using deep graph neural networks (PD-DGNN) was proposed in [2] by utilizing energy-based scores and molecular dynamics. Followed by this, as validation for proof-of-principle, ProteinSolver was also utilized to bring about sequences that counterpart the formation of serum albumin, then synthesize the top-scoring representation and substantiate it in vitro utilizing circular dichroism, therefore contributing to accuracy. Even though accuracy was found to be improved, the specificity and recall measure were not included. One limitation of PD-DGNN methods for protein design is the observation and interpretation of steep learning curve and the immense magnitude of domain competence that is required to provide rational and logical predictions.

A deep learning-based S-glutathionylation sites prediction tool called a computational framework (CF) was proposed in [3] to significantly identify the species-specific S-glutathionylation sites. In this study, species-specific S-glutathionylation sites prediction was made on the basis of the deep learning and particle swarm optimization algorithms. With this, the prediction results were said to be significantly improved. Despite improvement observed in the prediction accuracy, the time involved in prediction was not focused. Though better performance is being achieved by means of DeepGSH tools, there are numerous characteristics to be included, i.e., the inclusion of additional features like information concerning evolutionary aspect, interactions between proteins, secondary structures and so on, which may result in an accurate performance.

With the potentiality of deep learning (DL) [4], it was featured in considering numerous magnitudes of data structure, dispensing with noisy data, acquiring raw features without the requirement for feature engineering, and incorporating in a sensible manner to fabricate sensible predictions for data not utilized in training. Moreover, the ideal objective of bioinformatics not only remained in ensuring prediction accuracy but also remained in thorough comprehension of the fundamental biological procedures at work. Each member of the protein structure family may possess a moderately distinct form or shape from every other member, and hence, this creates an intrinsic accuracy constraint to deep learning-based modeling. This feasibly highlights the increased significance of structure cleansing to the succeeding protein structure prediction.

A distance-based protein structure prediction using deep learning, called DPSP method, was proposed in [5]. The method utilized a distance geometry algorithm with the purpose of enhancing the threading of protein without the presence of good templates in a protein data bank. With this, a significant amount of accuracy was said to be achieved in addition to the concentration of errors. Although the current protocol for predicting via predicted distance distribution went well, it was not found to be optimal for constructing 3D models.

1.1 Contributions

Motivated by the above state-of-the-art methods for secondary protein structure prediction, in our work, an Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method is proposed. The major contributory factors of EPB-OCNN method are given below.

- A novel secondary protein structure prediction method is proposed based on Energy Profile Bayes and Thompson Optimized Convolutional Neural Network model, which can offer maximum accuracy and precision rate with minimum time and therefore contributing to overall prediction accuracy.
- A separate algorithm is designed for the settings of Energy Profile Bayes and Thompson Optimized Convolutional Neural Network, respectively, therefore addressing protein structure prediction time, accuracy, precision, specificity, recall, and MCC.
- An integrated theoretical, qualitative analysis and experimental results are given, which validate the proposed method.
- The performance was evaluated through extensive simulations based on Protein Data Bank dataset. In comparison with TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], our EPB-OCNN method is superior in terms of protein structure prediction time, accuracy, precision, specificity, and recall.

1.2 Contribution explanation

The explanation about the Energy Profile Legion-Class Bayes Protein Structure Identification algorithm and Thompson Optimized CNN Protein Secondary Structure Prediction algorithm is detailed below.

1.2.1 Energy Profile Legion-Class Bayes Protein Structure Identification algorithm

Energy Profile Legion-Class Bayes Protein Structure Identification algorithm differentiates itself from others

where it possesses energy profiles for two atoms of given types based on the protein sequence profile context of the atom. Next, Legion Class Bayes function via Location-Specific Resultant Matrix reveals the patterns with maximum precision.

1.2.2 Thompson Optimized CNN Protein Secondary Structure Prediction algorithm

Thompson Optimized CNN Protein Secondary Structure Prediction algorithm is that the optimization function in the pooling layer via nonlinear down-sampling aids in minimizing the dimensionality of the features and parameters for obtaining relevant and precise protein structure prediction. Also, the learning rate forming hyperparameter controls how much to change in response to the estimated error each time during the prediction process.

1.3 Organization of the paper

The fundamental materials and methods connected with this study and the related works are discussed in brief in Sect. 2. The pseudo-codes representation with the aid of block diagram of the proposed Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) methods is outlined in Sect. 3. The experimental setup along with the results is summarized and contextualized in Sect. 4. Finally, the study is concluded in Sect. 5.

2 Related works

Secondary protein structure prediction is a paramount issue as far as structural biology and structural bioinformatics are concerned. Though enormous progression in structural bioinformatics has been seen in the recent few years, specifically accuracy of predicted structures inclines to differ in an extensive manner based on supplementary information availability and frequency of homologous structures and sequences in databases. Soluble and membrane protein design was proposed in [6] for fine tuning protein structures of low resolution, modeling drug binding sites in an accurate manner and modeling solvent-mediated protein catalysis. However, it was found to be a laborious and cumbersome process which remained difficult to distinguish at high resolution. A spherical graph convolutional network denoted as a molecular graph was designed in [7] for accurate structure prediction via angular information. Moreover, the spherical convolution technique can also be integrated with other methods for assessment of the corresponding protein model quality, and hence can also pave the way for additional and supplementary input features. In this manner, it will be probable to attain even better prediction results by including biological and chemical information in the corresponding input graphs.

Protein loop modeling was presented in [8] by means of deep learning. Owing to the drawbacks of computing potentialities, simulation experiments were only performed in limited settings of distinct network configurations. However, additional network configurations can still more assist in enhancing the overall loop modeling performance to a greater extent. The objective of protein structure prediction is to employ computational models to estimate spatial location of every atom in protein initiating from its amino acid sequence. Despite stagnancy observed in recent past two decades, contemporary application of deep neural networks to spatial prediction and end-to-end model has extensively enhanced the protein structure prediction accuracy.

Incorporation of deep learning techniques into numerous steps of protein folding and design was proposed in [9]. Nonetheless, while there are unquestionably many issues in the domain, the advancement noticed over the past few years bestows expectation that one of the most laborious and significant biological issues, i.e., predicting protein structures at their stability state of affairs initiating from the amino acid sequences unattended could be addressed via the employment of deep learning within the destined future. However, due to a lack of sufficient solved structures, high-throughput deep transfer learning model facilitating drug discovery was designed in [10]. One of the crucial disadvantages of the deep learning technique is that it is laborious and cumbersome to interpret the resultant technique. Even though machine learning technique has been giving solutions toward this issue, extracting simple rules from the deep learning technique to explain why particular pair of residues is predicted to form a contact is yet to be hard to incept, or else a simpler model of contact prediction can also be developed.

Neural network model was presented in [11] for analysis of dynamics and function prediction. Neural networks protein structure and function prediction was performed for dynamic analysis comprising of prediction based on amino acid sequences using multiple sequence alignment, utilization of assay data for interaction between protein and compound prediction and application of molecular dynamic simulation for protein detection. Though study here only presented a single approach, room for improvement is still said to persist. Machine learning techniques were applied in [12] for AlphaFold2 protein structure prediction involving foundation reconfiguration for biomolecular modeling. Though the design and analysis of AlphaFold2 is said to be unquestionably a milestone in the prolonged history of protein structure prediction, achieving accuracy for single domain prediction is only possible.

However, predictions empathetic to insignificant transposes that result in crucial structure altercations are said to be uneven and untested.

In [13], deep learning methods like convolutional neural networks and recurrent networks were applied to enhance prediction accuracy involved in protein structure prediction. To be more specific, though the convolutional neural network and recurrent network heavily depend on certain types of sensitive models to detect similarity from dissimilar structures and sequences, it is not clear whether the predictions accurately denote low energy arrangements except for their correctness. However, further enhancements with the inclusion of additional resources required for computation and a high volume of data are certainly necessitated. Different machine learning methods using a support vector machine and decision tree were applied in [14] for enhancing stability for protein structure forecasting. This proposed support vector machine and decision tree model was found to be advantageous in protein analysis for the sequences provided in an anonymous structure. However, these machine learning methods can be applied to even other protein datasets to ensure an efficient, aggressive analysis of the protein structure prediction.

Protein biological functions are fundamentally connected with its structure. Hence, for the last few years, protein structure identification has been considered as hot research issue in the field of bioinformatics. Accurate protein structure identification may help the research communities in evaluating numerous protein functions. The primary structure of a protein involves a polymer of 20 amino acids, which are heavily responsible for numerous functions and these functions are said to be largely based on their corresponding structures. Hence, protein structure information bestows indicators in secondary and tertiary protein structure prediction.

A framework called protein distance net was proposed in [15] for protein structure prediction for training and testing real-valued distances. Protein distance can also be analyzed by testing the importance by including certain other features via covariance and precision matrix. Also, an in-depth analysis can also be made by concentrating on the loss aspect specifically for distance prediction. Yet another deep structural inference for proteins that integrated both deep learning and template based structural model was designed in [16] therefore solving protein structure prediction problem. Also, tertiary structure prediction in a large-scale manner was performed for over 1200 single-domain proteins. Moreover, it predicted the tertiary structure in a successful manner four times that was predicted previously. In [17], clustering recurrent neural network was proposed for predicting distance matrices, torsion angles and secondary structures. However, the method was even found to be highly expensive upon comparison to the shallow learning method. Owing to this reason, measures to speed up the overall learning process in addition to the accuracy maintenance of the prediction system was not focused.

Despite improvement observed inaccuracy, it was not found to be computationally efficient. Yet another novel computational method using deep learning was investigated in [18], therefore, achieving secondary structure prediction accuracy. Moreover, a learning strategy performed based on the multi-task model was also utilized in predicting secondary structures and the trans-membrane helixes. The novel computation method was elaborately trained and tested by employing an independent dataset that was found to be non-redundant in nature. As a result, the secondary structure prediction accuracy was found to be 78% as far as the non-transmembrane region was concerned and found to be 90% for the transmembrane region. A linear predictive coding model using position-specific score matrices was proposed in [19] for predicting protein structural class. To sum up, the method was found to be satisfactory upon comparison with other methods on a single type of features. Owing to this reason, cost-effective mechanism for predicting protein structural class was ensured.

A secondary structure prediction based on the position-specific scoring value using matrix representation was proposed in [20]. Although numerous and extensive machine learning algorithms have been provided for predicting secondary structures in every short period, the enhancements rate were found to be comparatively minimum. This is owing to the reason that the amino was already introduced in and around 2000. However, there have been only elementary shifts in the feature set prediction. However, to improve the rate of accuracy, the foundation must be enhanced. In [21], an integration of physical, chemical, statistical, and biological characteristics of protein were employed as the features with which a novel mechanism was presented with the purpose of predicting protein's post-translational modification sites, therefore contributing to accuracy. Despite improvement observed in accuracy, numerous types of protein post-translational modification must be elaborated in detail as far as the domain of biology is concerned. Also, specialized formation of a structure must be utilized as the means for feature prediction.

Yet another deep learning approach based on position specific scoring matrix employing deep network architecture was investigated in [22]. By taking into consideration the enormous endeavor essential for researchers to bring about small enhancements, the realistic objective would remain in concentrating on the enhancement in the overall prediction of protein. A review of deep learning involving protein structural modeling was investigated in [23].

Finally, with the objective of acquiring a greater purview into the biomolecule fundamental science, a specific requirement to associate artificial intelligence (AI) with the biochemical and biophysical properties would arise. Also, a holistic approach to the underlying strategies and hidden patterns that result in the development of therapeutics is also the need of the hour**.** A state-of-the-art machine learning method was proposed in [24] by building on the alpha fold model, therefore, ensuring high quality predictions. Key features on alpha-fold two were concentrated in [25] using attention mechanism, therefore, ensuring computational capability.

Protein model quality assessments were made in [26] by employing spherical convolutions via rotation-equivariant spherical filters, therefore ensuring critical assessment of structure prediction benchmarks. Despite critical assessment, the precision with which the benchmark arrived was not discussed. Gaining a thorough insight into the protein structure is contemplated as the laborious process toward the design and development of new types of drugs. Therefore, acquiring a thorough knowledge and understanding of its functionalities would provide a good understanding into life machinery and organization, therefore the paving way for great social influence.

In [27], protein engineering was performed by means of deep diving with only a small proportion of protein sequence descriptors. With this, protein redesign issues in the pharmaceutical industry were addressed in a precise manner. Even though an extensive protein sequence selection along with the state-of-the-art machine learning techniques have been provided in detail, still further enhancement would cause an overall improvement. To name an objective may be to enhance the learning rate polity that in turn would result in minimizing the training time and hence the overall performance improvement. Yet another critical assessment of protein structure prediction using a simple gradient descent algorithm was proposed in [28] with increased accuracy. With the analysis results, it can be inferred that the overall process can be optimized by means of simple gradient descent algorithm that, in turn, would acquire structures without complicated sampling procedures**.**

An ensemble of deep convolutional neural networks for protein function prediction was investigated in [29] to address the time analysis. To sum up, the proposed method can, in turn, bestow swift prediction of the protein functioning, therefore making room for pertinent applications, to name a few being identification of target concerning pharmacological application. Also, a thorough analysis can be made for hierarchical function prediction and enzyme annotation toward the enzyme classification system. Yet another protein structure prediction that was performed automatically using I-TASSER was proposed in [30]. Here, from the target proteins of the respective amino acid sequence, the I-TASSER initially generated a comprehensive lengthy atomic structure format based on multiple threading alignments. Followed by this, an assembly simulation performed in an iterative fashion based on the atomic-level structure refinement was also proposed. Finally, protein biological functions comprising of ligand-binding sites and commission number of the respective enzyme were then acquired from protein function databases based on the sequence and comparative structure profile. A prediction model of protein consisting of Kunitz-type trypsin inhibitor from the respective seeds of Acacia Nilotica (L) based on the antimicrobial and insecticidal activity was proposed in [31]. Here, two generation progenies were studied, therefore reducing the mean percent mortality.

Protein family identification and classification are one of the most paramount issues as far as bioinformatics and protein studies are concerned. In these cases, it becomes necessary to mention the protein family as it finds a place chiefly in smart drug therapies, functioning of protein and so on. However, determining these families with sequencing yet consumes an enormous time. A novel protein mapping method was designed in [32] based on the Fibonacci numbers and hashing table called (FIBHASH). The Fibonacci number was assigned to each amino acid code based on integer representations. Followed by these amino acid codes were inserted into hashing table for further classification using recurrent neural networks, therefore improving protein mapping accuracy to a greater extent. As of now, the novel coronavirus (COVID-19) is a swiftly proliferating disease with a high rate of mortality.

In [33], interactions between specific flavonols such as 2019-nCoV receptor binding domain (RBD) and cathepsins (CatB and CatL) were medically analyzed. Based on the Relative Binding Capacity Index (RBCI) value estimated based on the free energy of binding and calculated inhibition constants, results were analyzed that robinin (ROB) and gossypetin (GOS) were determined to be the most significant flavonols among all the targets. Biological organism sequence data like nucleotide and amino acid are stored in databases that comprise billions of records. With the objective of processing enormous data in a comparatively lesser amount of time, high-performance analysis models were designed.

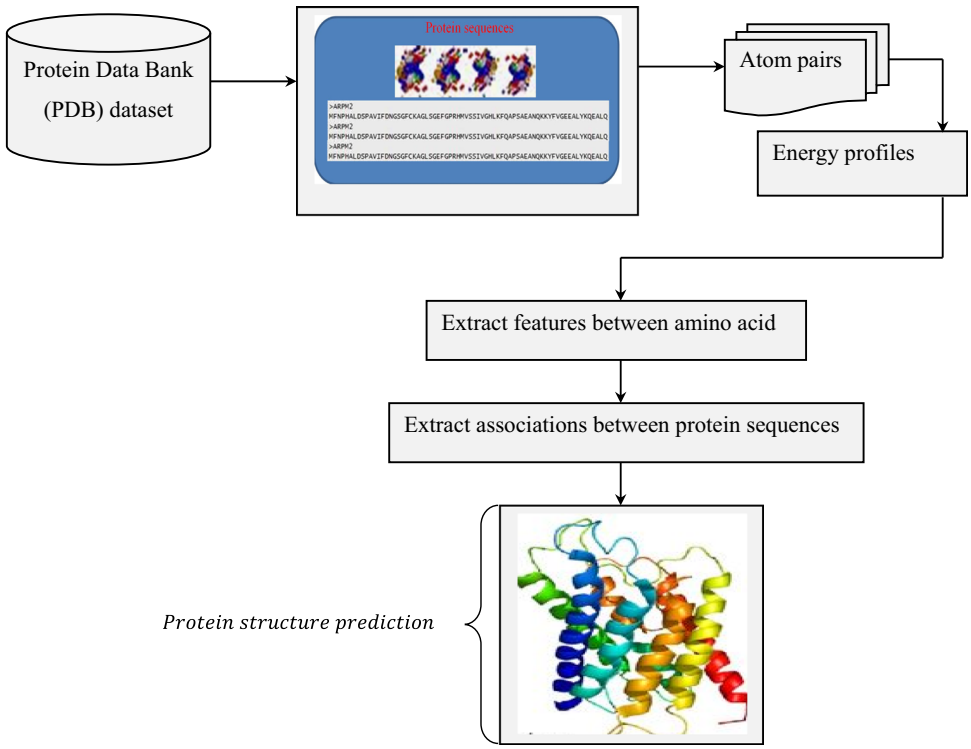

Fig. 1 Block diagram of Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method

In [34], pairwise and multiple sequence alignment operations were proposed to perform sequence alignment concerning bioinformatics in minimal time. As far as uncharacterized protein sequences are concerned, prediction of the functioning of protein in an automated fashion is said to be a critical issue to be handled. Over the past few years, deep learning-based algorithms were said to outperform the existing methods owing to the issues concerning overfitting and significance involved in training. A DEEPred was proposed in [35] that involved multitasking deep neural networks in a feed forward manner with the structure of hierarchical stack-based protein function prediction. With this organization, protein function prediction was found to be good.

From the above hypothesis, present-day research works explored above are mandatorily recapitulating necessity for novel secondary protein structure prediction method. Thus, in this work, it is concentrated on deliberately proposing an Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method that provides significant results for precise and accurate Protein Secondary Structure Prediction method with minimum time and maximum improvement in specificity, recall and F-measure.

3 Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) protein secondary structure prediction

This section mainly deals with the proposed Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method of protein secondary structure prediction from a broader point of view. The EPB-OCNN method is divided into three parts. The first part consists of the problem definition, whereas the second portion contains an extensive protein structure identification using Energy Profile Legion-Class Bayes Protein Structure Identification. The third subsection includes a detailed analysis of protein secondary structure prediction by employing the Thompson Optimized Convolutional Neural Network model. Figure 1 shows the block diagram of the EPB-OCNN method.

As shown in the above figure, the protein sequences are obtained from Protein Data Bank (PDB) [20] dataset. Next, protein structure identification called Energy Profile Legion-Class Bayes is designed based on the energy profiles of corresponding atom pairs to describe protein energy and accordingly rank different conformations based on energy. Next, the Thompson Optimized Convolutional

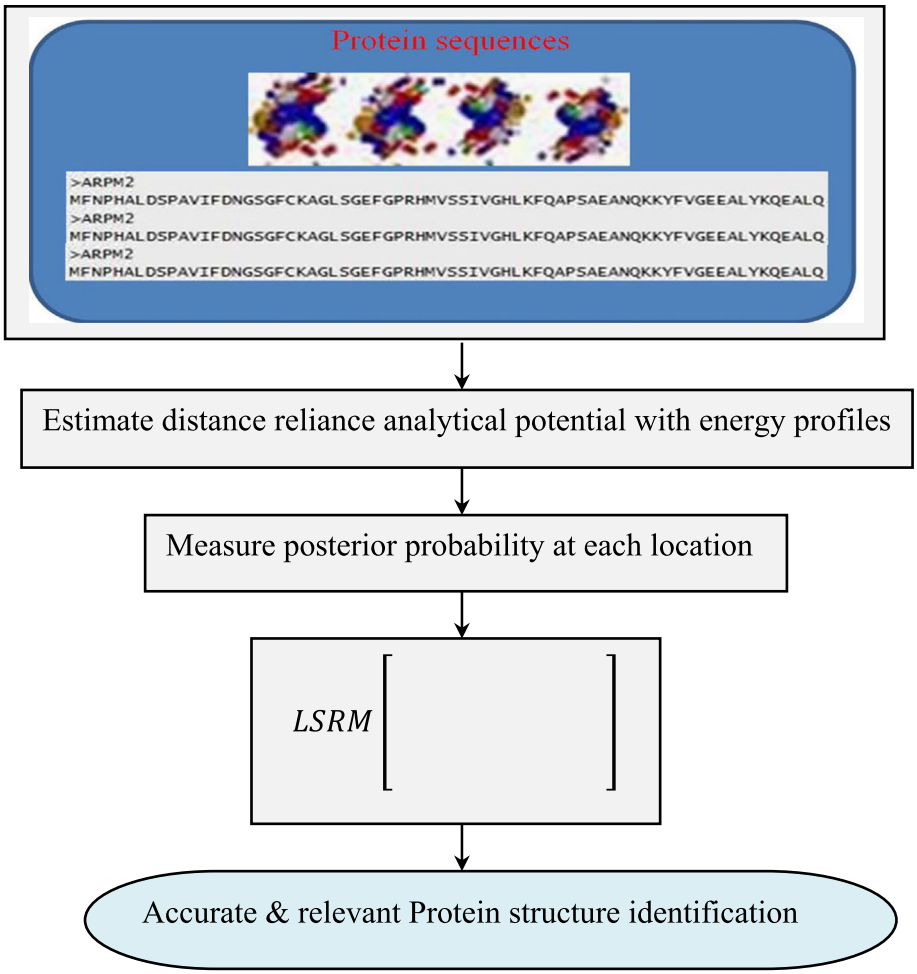

Fig. 2 Block diagram of Energy Profile Legion-Class Bayes Protein Structure identification

Neural Network model is designed by extracting features between amino acid and extracting the association between protein sequence, optimizing parameters involved in predicting protein structure via CNN.

3.1 Problem definition

Protein structure prediction from amino acid sequence has been a striking confront for decagons. Therefore, atomic-level structures of proteins are frequently the initiating locality to perceive protein structure identification and engineer them. Naturally occurring proteins denote only a minuscule subset of probable amino acid sequences designated by evolutionary process to carry out a biological function. The state-of-the-art methodology for protein structure prediction is based on the thermodynamic hypothesis. The thermodynamics concerning protein structure prediction states that the indigenous protein structure must possess the lowest free energy. Identifying the lowest-energy state is demanding owing to the fact as it occupies an enormous space of possible conformations obtainable to a protein. Considering these challenges, in our work, Legion-Shape Protein Structure Identification is proposed. Next, with the identified secondary protein structure and optimized parameter learning, protein structure prediction employing Thompson function is made for robust and accurate prediction.

3.2 Energy Profile Legion-Class Bayes Protein Structure Identification

To start within this section, protein structure identification using Energy Profile Legion-Class Bayes is designed. For this, first, protein and location based analytical Evolutionary Duo Distance Reliance Potential (EDDRP) is obtained. We configure the perceived probability in EDDRP by the

evolutionary information in addition to atom types. EDDRP differentiates itself from others in that it possesses numerous energy profiles for two atoms of given types, according to protein taken into consideration and sequence profile atom context. With this EDDRP, Energy Profile Legion-Class Bayes Protein Structure Identification is made. Figure 2 shows the block diagram of Energy Profile Legion-Class Bayes Protein Structure Identification.

As shown in the above figure, with protein sequences obtained as input, initially, distance reliance analytical potential with energy profiles is measured. Second, the posterior probability at each location is measured to form Location-Specific Resultant Matrix. Finally, with the obtained Location-Specific Resultant Matrix, accurate and relevant protein structures are identified. To start with, distance reliance analytical potential and energy profiles for atom pairs are mathematically represented as given below.

$$U\left(\mathrm{Dis}|A_i, A_j, \mathrm{Area}_i, \mathrm{Area}_j, \mathrm{Rad_G}\right) = T\mathrm{Log}\left(\frac{\mathrm{Prob}\left(\mathrm{Dis}|A_i, A_j, \mathrm{Area}_i, \mathrm{Area}_j, \mathrm{Rad_G}\right)}{\mathrm{Ref}(\mathrm{Dis}|\mathrm{Rad_G})}\right) \quad (1)$$

From Eq. (1), 'T' is the temperature factor corresponding to each amino acid, '$\mathrm{Ref}(\mathrm{Dis}|\mathrm{Rad_G})$' denotes the reference state 'Ref' with respect to distance 'Dis' and gyration radius '$\mathrm{Rad_G}$' of protein considered for simulation. In addition, '$\mathrm{Prob}\left(\mathrm{Dis}|A_i, A_j, \mathrm{Area}_i, \mathrm{Area}_j, \mathrm{Rad_G}\right)$' denotes the discovered probability of two atoms 'A_i,' 'A_j' linked within a distance 'Dis' and gyration radius 'Rad_G.'

With the utilization of the above distance reliance analytical potential model, accurate energy functions are evolved for describing protein physics and sampling protein sequence. With the concept of Legion Class Bayes employed in our work and energy profiles for pair of atoms, let us consider, '$P = \{P_1, P_2, \ldots, P_n P_{n+1}[D_1], P_{n+2}, \ldots, P_{2n}[D_2],\ P_{m+1}, P_{m+2}, \ldots, P_{mn}[D_n]\}$,' where '$P_i = \{1, 2, \ldots .m\}$' with '$i$' denoting amino acid position in protein sequence. In the case of Legion-class problem, the trial sample 'S' comprises of '$(\mathrm{CL}_1, \mathrm{CL}_2, \ldots, \mathrm{CL}_n)$' where '$\mathrm{CL}_1$' is the first class, '$\mathrm{CL}_2$' is the second class, and 'CL_n' denotes the last class, respectively. Then, the posterior probability for each class is estimated employing the probability of amino acid at each location in trial samples. With this, the Legion Class Bayes is mathematically formulated as given below.

$$\mathrm{Prob}(\mathrm{CL}_i|S) = \frac{\mathrm{Prob}(S|\mathrm{CL}_i)\mathrm{Prob}(CL_i)}{\mathrm{Prob}(S)} \quad (2)$$

From Eq. (2), '$\mathrm{Prob}(\mathrm{CL}_i|\mathrm{S})$' represents posterior probabilities of amino acid at each location with respect to each class 'CL_i' and samples 'S,' '$\mathrm{Prob}(\mathrm{CL}_i)$' denotes the prior probabilities of amino acid at each location for the corresponding class '(CL_i),' '$\mathrm{Prob}(S|\mathrm{CL}_i)$' represents the likelihood of amino acid at each location and '$\mathrm{Prob}(S)$' denotes the probability of the overall trial sample taken into consideration for simulation. The likelihood of each sample 'S' with respect to distance 'D_i' is mathematically formulated as given below.

$$\mathrm{Prob}(S|D_i) = \prod_{j=1}^{m} \mathrm{Prob}\left(S_j|D_i\right) \quad (3)$$

With the resultant Legion Class Bayes value Location-Specific Resultant Matrix (LSRM), an evolutionary energy profiles for pair of atoms is estimated. The LSRM is then mathematically formulated for a protein sequence 'P' possessing area 'Area' as given below

.

$$P_{\mathrm{LSRM}} = \begin{bmatrix} R(1 \to 1) & R(1 \to 2) & \ldots & R(1 \to i) & \ldots & R(1 \to 20) \\ R(2 \to 1) & R(2 \to 2) & \ldots & R(2 \to i) & \ldots & R(2 \to 20) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ R(j \to 1) & R(j \to 2) & \ldots & R(j \to i) & \ldots & R(j \to 20) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ R(\mathrm{Area} \to 1) & R(\mathrm{Area} \to 2) & \ldots & R(\mathrm{Area} \to i) & \ldots & R(\mathrm{Area} \to 20) \end{bmatrix} \quad (4)$$

From Eq. (4), '$R(i \to j)$' corresponds to the resultant outcome of 'ith' amino acid location, that was swapped by 'j' amino acid in protein sequence during the process of evolution. The 'P_{LSRM}' is produced for multiple sequence with protein sequence 'P,' possessing '$A * 20$' resultant outcomes. Finally, the probability of evolution 'Evol' (i.e., protein structure identification) from 'p-th' to 'q-th' amino acid 'Evol_{pq}' is mathematically formulated as given below.

$$\mathrm{Evol}_{pq} = \mathrm{Prob}_{ip}\mathrm{Prob}_{jq}[1 \le p \le 20;\ \le q \le 20] \quad (5)$$

From Eq. (5), the probability of evolution from 'p-th' to 'q-th' amino acid 'Evol_{pq}' is obtained for 20 resultant outcomes (i.e., sliding window ranging between 13 and 19). The pseudo-code representation of Energy Profile Legion-Class Bayes Protein Structure Identification is given below.

Input: Protein Data Bank Dataset 'PDB', Features '$F = F_1, F_2, \ldots, F_n$'

Output: Accurate and reliable relevant protein data '$PD = pd_1, pd_2, .., pd_n$'

1: **Initialize** trial sample 'S', temperature factors 'T', protein data size 'n'

2: **Begin**

3: **For** each Protein Data Bank Dataset 'PDB' and Features 'F'

4: **Estimate** distance reliance analytical potential with energy profiles as in equation (1) based on the temperature corresponding to each amino acid, reference state with respect to distance and gyration radius of protein considered for simulation

5: **Measure** posterior probability with the probability of amino acid at each location as in equation (2) with respect to each class and samples, prior probabilities of amino acid at each location for the corresponding class, likelihoods of amino acid at each location and probability of the overall trial sample considered for simulation

6: **Estimate** Location Specific Resultant Matrix as in equation (4) for the resultant outcome of 'ith' amino acid location being swapped by 'j' amino acid in a protein sequence.

7: **Measure** probability of evolution (i.e., protein structure identification) as in equation (5) from '$p - th$' to '$q - th$' amino acid evolution obtained for 20 resultant outcomes

8: **Return** (protein data)

9: **End for**

10: **End**

Algorithm 1 Energy Profile Legion-Class Bayes Protein Structure Identification

As given in the above Energy Profile Legion-Class Bayes Protein Structure Identification algorithm, the objective remains in obtaining accurate and reliable protein data utilizing the Legion-Class Bayes function in addition to the higher-resolution energy profiles. Initially with PDB dataset provided as input, for each amino acid, distance reliance analytical potential with energy profiles is measured. With this function, optimal molecules are arrived at with minimum time. Next, to the energy profile values, the Legion Class Bayes function is applied to obtain Location-Specific Resultant Matrix, therefore revealing patterns with maximum precision. Finally, utilizing the probability of evolution, protein structure identification is made with improved true positive rate.

Initially, with the PDB dataset provided as input for each amino acid, the first distance relies on analytical potential with energy profiles using the features, amino acid name/residue name, residue number, X, Y, Z coordinates and occupancy. Followed by this, the posterior probability with the probability of amino acid at each location was measured employing record type, atom number, atom name, amino acid name, chain name, residue number, X, Y, Z coordinates, occupancy, temperature factors and element symbols, respectively. Next, Location-Specific Resultant Matrix was estimated based on the record type, atom number, atom name, residue number, chain name, residue name, X, Y, Z coordinates, occupancy, temperature factors and element symbols, respectively. Finally, the probability of evolution is obtained to acquire the final features, namely amino acid name/residue name, residue number, X, Y, Z coordinates and occupancy, respectively.

3.3 Thompson Optimized Convolutional Neural Network Protein Secondary Structure Prediction model

The accurate protein secondary structure prediction not only warrants us to realize complicated association between protein sequence and protein structure, but also assists in analyzing the functioning of the protein. In this work, a deep learning algorithm based on convolutional neural network, called Thompson Optimized Convolutional Neural Network Protein Secondary Structure Prediction model has been applied to protein secondary structure prediction. The objective of designing this model remains in optimizing network parameters and speed up overall process. Structure of Thompson Optimized Convolutional Neural Network Protein Secondary Structure Prediction model is shown in Fig. 3.

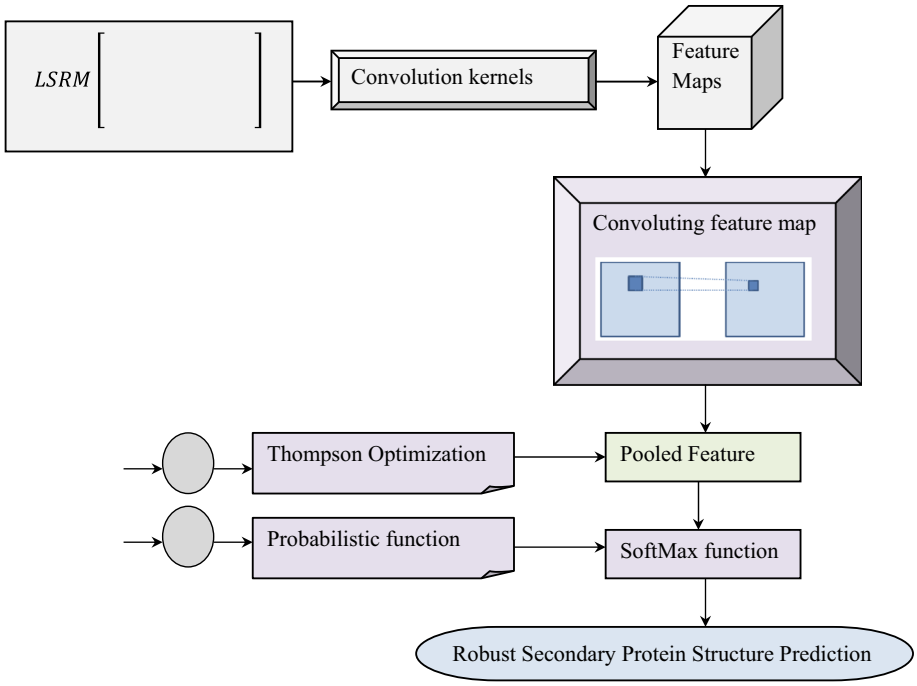

Fig. 3 Structure of Thompson Optimized Convolutional Neural Network protein secondary structure prediction

The Thompson Optimized Convolutional Neural Network extracts features between amino acid. The input of each neuron in the convolutional layer comes from Location-Specific Resultant Matrix (LSRM) in a definite area. In addition, the size of this definite area is obtained by means of a convolution kernel. The feature map initially 'FM^i' is formulated as given below.

$$\mathrm{FM}^i = \left(\mathrm{CK}_1^i, \mathrm{CK}_2^i, \mathrm{CK}_3^i, \ldots, \mathrm{CK}_n^i\right) \tag{6}$$

From Eq. (6), 'CK_n^i' corresponds to the convolution kernel of the 'i-th' layer with 'n' representing a number of convolution kernels. The functioning of convolution is to perceive convolution operation via a feature extraction filter for input matrix LSRM. Each region, in turn, is obtained by multiplying the input 'LSRM' matrix and weights and then added with the offset constant 'Off' to produce the feature map.

$$\mathrm{FM}_l^i = \mathrm{fun}\left(\sum_l \mathrm{LSRM}\left[\mathrm{CK}_l^i\right] * W_l^i + \mathrm{off}\right) \tag{7}$$

From Eq. (7), '$\mathrm{LSRM}\left[\mathrm{CK}_l^i\right]$' denotes the feature map obtained by the convolution kernel 'CK' of the input 'LSRM' matrix data with the weight of the 'i-th' convolution kernel represented by 'W_l^i' and offset value denoted by 'off' for 'l' amount of convolution kernels, respectively. Next, the pooling layer known as nonlinear down-sampling aids in minimizing dimensionality features and parameter to minimize the frequency of calculation. To adjust weights during training or learning process, our work uses a Thompson Function. Finally, the output layer of our forms fully connected layer and SoftMax layer. The SoftMax function layer in our protein secondary structure prediction employs activation function as fined below.

$$\mathrm{Prob}\left(\underset{i}{\mathrm{Pr}}, \mathrm{CL}\right) = \frac{\mathrm{Prob}(\mathrm{CL}/\mathrm{Pr}_i)\mathrm{Prob}\ (\mathrm{Pr}_i)}{\sum_{j=1}^{\mathrm{CL}} \mathrm{Prob}\ \left(\mathrm{CL}/\mathrm{Pr}_j\right)\mathrm{Prob}(\mathrm{Pr}_i)} \tag{8}$$

From Eq. (8), '$\mathrm{Prob}(\mathrm{CL}/\mathrm{Pr}_i)$' refers to the probability of given class sample (i.e., from four different classes) and '$\mathrm{Prob}(\mathrm{Pr}_i)$' denotes the prior probability of the protein secondary structure class. Owing to the increasing number of protein sequences present in a protein data bank, a considerable amount of time is said to be consumed while modeling the CNN to the protein secondary structure prediction. This is because of the reason that it takes a significant amount of time to modify the hyperparameters of CNN. This work employs a Thompson optimization algorithm to optimize the hyperparameters of CNN, employing learning rate (i.e., 0.01), impulse and regularization factor, respectively. First, let us assume that the Gaussian kernel function is selected as an acquisition function to obtain consecutive sampling points. Thompson optimization of hyperparameters is Gaussian prior modeling of the loss function '$f(pd)$' by hyperparameters of corresponding protein data 'pd_i.'

$$L\left(\mathrm{OP}_i, V_j\right) = \mathrm{CL}[\mathrm{OP}_i(pd_i)Y_i] \tag{9}$$

As the observations on the secondary protein structure prediction involves a considerable amount of noise, Gaussian noise 'α' is added to each observation sample '$L\left(\mathrm{OP}_i, V_j\right)$' for the objective function '$\mathrm{CL}[\mathrm{OP}_i(pd_i)Y_i]$.'

$$f(pd) = L\left(\mathrm{OP}_i, V_j\right) + \alpha \tag{10}$$

Let us further consider input hyperparameters '$pd = (pd_1, pd_2, pd_3, \ldots, pd_n)$' obtains the output '$Y = \mathrm{CL}\left(pd_i, V_j\right)$.' Then, the Thompson optimization for hyperparameters 'f_{pd}' is mathematically formulated as given below.

$$f_{pd} = \int \mathrm{IF}[\mathrm{Exp}(r|pd, \alpha) = \max\ \mathrm{Exp}(r|pd, \alpha)]\mathrm{Prob}(\alpha|D)d\alpha \tag{11}$$

The expected improvement based on the above Thompson optimization is given below.

$$(pd|D) = \mathrm{Exp}\left[\max\left(\alpha, f_{pd} - f_{\mathrm{best}}\right)\right] \tag{12}$$

From Eq. (12), 'f_{best}' forms optimal solution for hyperparameters, learning rate, impulse, and regularization factor, respectively. The pseudo-code representation of Thompson Optimized CNN Protein Secondary Structure Prediction is given below.

As given in the above Thompson Optimized CNN Protein Secondary Structure Prediction algorithm, the objective remains in predicting secondary protein structure sequence with maximum precision and accuracy, therefore, contributing to robustness. To achieve this, the Location-Specific Resultant Matrix is employed as input to the CNN. Followed by which, for each protein data, convolved feature map is evolved by means of kernel weight and offset parameter. Next, Thompson optimized learning rate, impulse and regularization factor is estimated and applied to pooled layer, therefore obtaining robust protein structure prediction sequence. Also, in our work, first, the loss function is evaluated via Thompson optimization of hyperparameters. Second, if the detected loss is starting to increase, the weights are reset based on the Gaussian noise 'α' back to where the minimum occurred. This ensures that the proposed method using Thompson Optimized CNN Protein Secondary Structure Prediction algorithm won't continue to learn noise and overfit the data. In this manner, Thompson Optimized CNN Protein Secondary Structure Prediction algorithm addresses overfitting, hence guaranteeing accurate results.

Input: Protein Data Bank dataset 'DS', Features '$F = F_1, F_2, \ldots, F_n$'

Output: Relevant and precise protein structure prediction

1: **Initialize** relevant protein data '$PD = pd_1, pd_2, .., pd_n$', amount of convolution kernels 'n',

2: **Begin**

3: **For** each relevant protein data 'PD'

4: **Formulate** a feature map as in equation (6) with the corresponding convolution kernel of '$i-th$' layer for 'n' convolution kernels

5: **Evaluate** feature map for corresponding input '$LSRM$' matrix data as in equation (7) using feature map by convolution kernel of the input matrix data with the weight of the '$i-th$' convolution kernel denoted by 'W_l^i' and offset represented by 'off' for 'l' convolution kernels

6: **Estimate** the SoftMax function layer as in equation (8) by means of the probability of given class sample and the prior probability of protein secondary structure class

//Optimization

7: **Model** loss function as in equation (9)

8: **Estimate** Thompson optimization for hyperparameters as in equation (11) via input hyperparameters

9: **Obtain** optimal solution for the hyperparameters as in equation (12)

10: **Return** (protein structure prediction sequence '$f_1, f_2, \ldots, f_n$')

11: **End for**

12: **End**

Algorithm 2 Thompson Optimized CNN Protein Secondary Structure Prediction

4 Experimental setting and qualitative analysis

In this section, initially, dataset details are provided. Subsequent subsections contain the performance metrics analysis and discussion, comparison with the state-of-the-art methods, and the statistical analysis, respectively.

4.1 Dataset details

The Protein Data Bank (PDB) [21] used in our work corresponds to a database containing three-dimensional protein structure hitherto determined by nuclear magnetic resonance. The PDB dataset consists of the protein structural domains that have been categorized based on the structure similarities and amino acid sequences provide a detailed and comprehensive description of the structural and evolutionary relationships between proteins**.** However, for practical applications, only four classes '$\text{all}-\alpha$,' '$\text{all}-\beta$,' '$\alpha+\beta$' and 'α/β' are considered.With an overall '1673' protein sequences, they are classified as '$443(\text{all}-\alpha)$,' '$443(\text{all}-\beta)$,' '$441(\alpha+\beta)$' and '$346(\alpha/\beta)$,' respectively (as given in the Table 1), with a typical 80% training data size and 20% testing data size. The dataset is a list of protein sequences that are an arrangement of amino acids having 1673 as the sample size. Out of it, 80% remains the training data size (i.e., 1339) and 20% (i.e., 334.6 = 335) is the testing data size.

Table 1 Typical PDB dataset with four categories utilized for benchmarking

Dataset	$\text{all}-\alpha$	$\text{all}-\beta$	$\alpha+\beta$	α/β	Total
Protein Data Bank (PDB)	443	443	441	346	1673

The number of proteins in each category and the total number of proteins in PDB dataset

With 335 amino acids arrangements in the list of protein sequences, protein data ranging between 500 and 5000 are considered for simulation.

4.2 Performance metrics analysis and Discussion

In this section, performance analysis of metrics such as prediction time, accuracy, ROC curve, precision, specificity, recall, F-measure, Mathew Correlation Coefficient and precision–recall curves are discussed.

Table 2 lists the hyperparameters and their description employed in our proposed method.

4.2.1 Performance analysis of protein structure prediction time

In this section, a detailed analysis of protein structure prediction time is made. During prediction of secondary protein structure, a considerable amount of time is said to be consumed. This is mathematically expressed as given below.

$$\text{PSP}_{\text{time}} = \sum_{i=1}^{n} P_i * \text{Time}\left[\text{Prob}\left(\underset{i}{\text{Pr}}, \text{CL}\right)\right] \tag{13}$$

From Eq. (13), the protein structure prediction time 'PSP_{time}' is measured based on the number of protein data considered for simulation 'P_i' and time consumed in analyzing protein secondary structure prediction using Soft-Max function 'Time[Prob(Pr_i, CL)].' It is measured in terms of milliseconds (ms). Table 3 provides the protein structure prediction time of the proposed method EPB-OCNN method and the state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyperparameter with a sliding window of 13.

Figure 4 shows a graphical representation of the secondary protein structure prediction analysis of the proposed

Table 2 Hyperparameters and description

S. no	Hyperparameters	Description
1	Number of hidden layers used	Two hidden layers are used (the first hidden layer from the convolution and the second hidden layer from the pooling)
2	Activation function used in hidden layers	Nonlinear down-sampling function (i.e., linear activation function) is used in hidden layer
3	Activation function used in output layer	Sigmoid activation function
4	Learning rate	The value of the learning rate used in our work is 0.01
5	The momentum set	The momentum is set of 0.9
6	Batch size	Batch size in our work refers to the samples from the training dataset. In our work, the batch size is 5000 as samples are considered for simulation
7	Number of epochs	The number of epochs in our work is 10

Table 3 Tabulation for protein structure prediction time

Number of protein data	Protein structure prediction time (ms)					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	55	60	82.5	92.5	97.5	100
1000	68.20	90.05	110.05	140.15	170.20	200.25
1500	80.09	120.25	135.55	160.25	185.15	220.05
2000	90.09	135.05	180.05	200.05	220.35	280.35
2500	120.25	170.05	200.05	240.45	280.25	300.05
3000	135.15	200.15	230.25	270.25	320.05	350.15
3500	160.55	215.05	275.05	320.05	380.25	400.15
4000	195.05	235.25	300.05	380.25	419.45	480.05
4500	220.25	275.05	330.15	420.05	480.05	520.25
5000	265.05	290.25	375.15	480.43	530.05	590.05

method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. As shown in the above figure, with the x axis representing a number of protein data and the y axis representing protein structure prediction time, increasing protein data result in an increase in secondary protein structure time also. This is because with a large number of protein data sequence considered for simulation, protein structure modeling and designing also increase. This, in turn, causes an increase in the corresponding time. But simulations show that the proposed method EPB-OCNN achieved betterment in comparison with the five state-of-the-art methods. Single protein structure prediction time of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was observed to be 0.110 ms, 0.120 ms, 0.165 ms, 0.185 ms, 0.195 ms, and 0.200 ms, respectively. With this, the overall prediction time of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was observed to be 55 ms, 60 ms, 82.5 ms, 92.5 ms, 97.5 ms, and 100 ms, respectively, for 500 number of protein data, therefore reducing time using EPB-OCNN method. The reason was due to the application of distance reliance on analytical potential precise energy for describing protein and protein sequence sampling. As a result, the protein structure prediction time of the proposed method EPB-OCNN in comparison with the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was observed to be improved by 23%, 38%, 49%, 55%, and 60%, respectively.

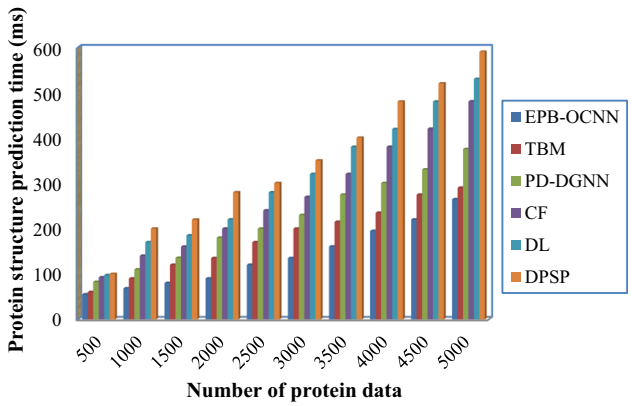

Fig. 4 Protein structure prediction time analyses

4.2.2 Performance analysis of protein structure prediction accuracy

The second parameter of significance is the accuracy involved during the prediction of secondary protein structure. This is mathematically stated as given below.

$$\mathrm{PSP}_{\mathrm{acc}} = \sum_{i=1}^{n} \frac{\mathrm{PS}_{\mathrm{CP}}[\mathrm{Prob}(\mathrm{Pr}_i, \mathrm{CL})]}{P_i} * 100 \tag{14}$$

From Eq. (14), protein structure prediction accuracy '$\mathrm{PSP}_{\mathrm{acc}}$' is measured based on protein structure correctly predicted using Softmax function '$\mathrm{PS}_{\mathrm{CP}}[\mathrm{Prob}(\mathrm{Pr}_i, \mathrm{CL})]$' and the overall number of protein data considered for simulation 'P_i.' It is measured in terms of percentage (%). Table 4 provides the protein structure prediction accuracy values of proposed method EPB-OCNN, and the state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4] and DPSP [5], respectively, for hyperparameters with a sliding window of 13.

Figure 5 shows a graphical representation of the secondary protein structure prediction accuracy analysis of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. As shown in the above figure, protein structure prediction accuracy is found to be inversely proportional to number of protein data considered for simulation. This is because an increasing number of protein data, a large number of protein sequence to be analyzed are kept in a stack and this, in turn, results in making a small

Table 4 Tabulation for protein structure prediction accuracy

Number of protein data	Protein structure prediction accuracy (%)					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	98.8	98	97	96	95	94
1000	98.80	96.25	92.45	92.10	91.10	90.45
1500	98.60	94.25	90.55	89.45	89.10	88.25
2000	98.30	93.45	87.65	86.25	85.35	85.10
2500	98.25	90.55	85.75	84.35	83.25	82.10
3000	97.15	88.25	82.45	81.10	80.65	79.10
3500	97.10	85.35	81.55	80.25	79.35	79.10
4000	97.10	84.25	81.25	80.10	79.10	78.65
4500	96.95	81.35	80.35	78.45	78.25	78.10
5000	96.65	81.10	76.45	75.31	74.25	74.10

amount of the wrong predictions. However, simulation analysis made with 500 protein data shows that the accuracy of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was observed to be 98.8%, 98%, 97%, 96%, 95%, and 94%, respectively. With this analysis, secondary structure prediction accuracy of the proposed method EPB-OCNN was found to be better than that of other five state-of-the-art methods. The betterment was due to the application of Energy Profile Legion-Class Bayes Protein Structure Identification algorithm. Legion-Class Bayes function was applied along with the higher-resolution energy profiles to model secondary protein structure prediction. As a result, the secondary protein structure prediction accuracy of the proposed method EPB-OCNN in comparison with the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was found to be improved by 10%, 15%, 16%, 18%, and 21%, respectively.

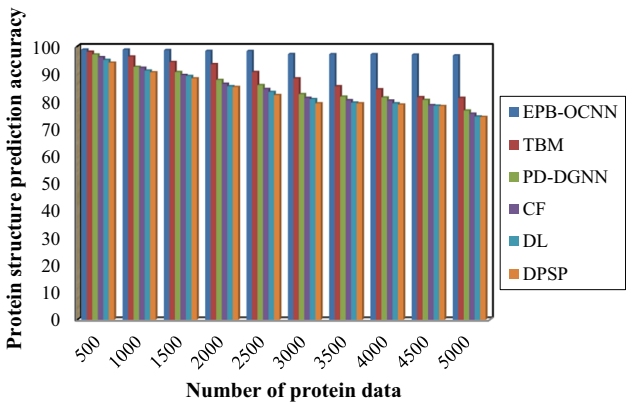

Fig. 5 Protein structure prediction accuracy analyses

4.2.3 Performance analysis of ROC curve

In this section, receiver operating characteristic (ROC) curves are analyzed to measure the prediction rate performance. With the assistance of ROC curve, binary classification is made based on either protein structure correctly predicted or wrongly predicted. The ROC curve is measured based on the true positive, false positive, true negative and false negative. With these four probable outcomes, receiver operating characteristic (ROC) curve is made where false positive rate is plotted on the x axis, and the true positive rate is plotted on the y axis. Table 5 provides the ROC curve analysis values of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyperparameters with a sliding window of 19.

Figure 6 illustrates the graphical representation of ROC curve analysis of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. In the above graphical representation, the x axis denotes the false positive rate, whereas the y axis denotes the true positive rate. A diagonal line from (0, 0) in the lower left-hand corner to (1, 1) in the upper right-hand corner is drawn. This diagonal line displays the protein structure prediction test results. Also, the ROC makes an analysis of the protein structure prediction based on the true positive rate and false positive rate for each possible cut point value of the test. From the above figure, it is illustrative that the roc curve of EPB-OCNN method is identified to be comparatively better than that of the five state-of-the-art methods, therefore corroborating the secondary protein structure prediction rate.

Table 5 Tabulation for ROC curve

False positive rate	True positive rate					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
0.1	0	0	0	0	0	0
0.2	0.33	0.23	0.18	0.17	0.16	0.15
0.3	0.48	0.33	0.23	0.21	0.18	0.17
0.4	0.63	0.48	0.36	0.28	0.23	0.21
0.5	0.71	0.54	0.48	0.40	0.34	0.31
0.6	0.76	0.63	0.51	0.44	0.40	0.34
0.7	0.83	0.74	0.63	0.55	0.43	0.41
0.8	0.95	0.85	0.75	0.68	0.58	0.53
0.9	0.98	0.88	0.83	0.76	0.68	0.63
1.0	1	0.93	0.87	0.83	0.75	0.68

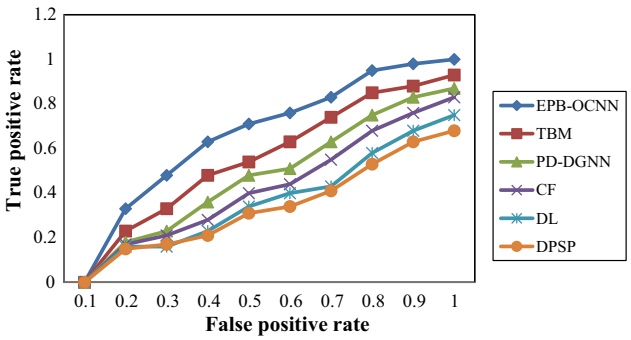

Fig. 6 ROC curve analyses

4.2.4 Performance analysis of precision

The next parameter of significance is precision. While predicting secondary protein structure, precision measurement has to be evolved. This is mathematically expressed as given below.

$$P = \left[\frac{t_p}{t_p + f_p}\right] \tag{15}$$

From Eq. (15), precision 'P' is measured based on the true positive rate 't_p' (i.e., protein structure correctly predicted as it is) and the false positive rate 'f_p' (i.e., protein structure incorrectly predicted). It is measured in terms of percentage (%). Table 6 gives the analysis of the precision of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyperparameters with a sliding window of 19.

Figure 7 illustrates a graphical representation of precision analyses of the proposed method EPB-OCNN, and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. To analyze the precision factor, protein sequence data in the range of 500 to 5000 are taken into consideration. From the above figurative representation, it is inferred that the precision of the proposed method EPB-OCNN is found to be comparatively higher than that of the five state-of-the-art methods. The reason behind the improvement was due to the application of Legion Class Bayes function employed for obtaining energy profile resultant values via Location-Specific

Table 6 Tabulation for precision

Number of protein data	Precision					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	0.99	0.93	0.77	0.75	0.72	0.70
1000	0.94	0.88	0.78	0.77	0.75	0.80
1500	0.94	0.87	0.79	0.80	0.77	0.76
2000	0.99	0.87	0.80	0.73	0.73	0.72
2500	0.99	0.90	0.90	0.79	0.76	0.75
3000	0.96	0.89	0.91	0.77	0.75	0.73
3500	0.94	0.90	0.88	0.78	0.77	0.75
4000	0.96	0.89	0.89	0.78	0.76	0.75
4500	0.98	0.89	0.89	0.68	0.66	0.65
5000	0.94	0.89	0.89	0.79	0.77	0.75

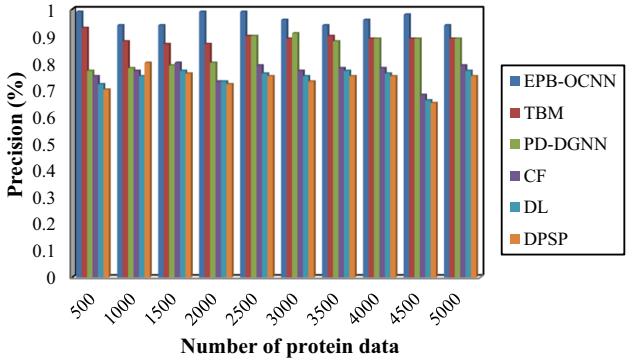

Fig. 7 Precision analyses

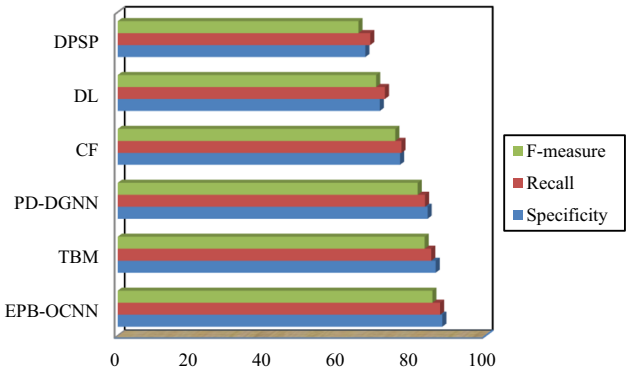

Fig. 8 Specificity, recall, and F-measure analyses

Resultant Matrix. With this pattern, maximum precision is said to be revealed. The precision of the proposed method EPB-OCNN method in comparison with the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] is improved by 8%, 14%, 26%, 30%, and 31%, respectively.

4.2.5 Performance analysis of specificity, recall and F-measure

Specificity denotes the percentage ratio of negatives that are correctly identified as with not possessing the actual secondary protein structure. On the other hand, recall denotes the percentage ratio of relevant protein sequences instances that were retrieved as it is. Finally, F-measure represents the harmonic mean of both the precision and recall. Table 7 provides the specificity, recall and F-measure analyses of the proposed method EPB-OCNN and five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4] and DPSP [5], respectively, for hyperparameters with a sliding window of 13.

Figure 8 shows the graphical representation of specificity, recall and F-measure analysis of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. From the above figure, the specificity rate of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] is inferred to be 88.25%, 86.45%, 84.25%, 76.75%, 71.25%, and 67.32%, respectively. In addition, the recall rate of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was found to be 87.65%, 85.25%, 83.55%, 77.15%, 72.65%, and 68.65%, respectively. From the specificity and recall results, it is identified that specificity, recall, and F-measure of the proposed method EPB-OCNN were found to be better than the five state-of-the-art methods. The reason behind improvement was due to the application of Thompson optimized function to obtain learning rate, impulse, and regularization factor. With this, improvement was observed at a true positive rate in turn reducing true negative rate, therefore contributing to specificity, recall and F-measure.

4.2.6 Performance analysis of precision–recall curve

In this section, the precision–recall curve is analyzed. The precision–recall curve shows graphical representation of secondary protein structure prediction at different threshold values. This ROC curve plot two different parameters, namely precision and recall. Table 8 makes a detailed analysis of the precision–recall curve of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyperparameters with a sliding window of 13.

Figure 9 illustrates the graphical representation of precision–recall analysis of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. In this study, simulation was performed for all the six methods with unique recall values ranging between 0.1 and 1. For

Table 7 Tabulation for specificity, recall and F-measure

Metrics	Methods					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
Specificity	88.25	86.45	84.25	76.75	71.25	67.32
Recall	87.65	85.25	83.55	77.15	72.65	68.65
F-measure	85.55	83.45	81.55	75.45	70.25	65.45

Table 8 Tabulation for precision–recall curve

Recall	Precision					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
0.1	0.26	0.23	0.21	0.18	0.17	0.15
0.2	0.29	0.27	0.26	0.22	0.21	0.19
0.3	0.38	0.37	0.35	0.32	0.30	0.26
0.4	0.43	0.40	0.39	0.36	0.32	0.29
0.5	0.44	0.42	0.39	0.36	0.34	0.33
0.6	0.45	0.43	0.41	0.35	0.36	0.33
0.7	0.46	0.44	0.43	0.39	0.37	0.35
0.8	0.48	0.46	0.45	0.40	0.38	0.36
0.9	0.48	0.48	0.46	0.40	0.38	0.38
1	0.52	0.49	0.47	0.42	0.39	0.39

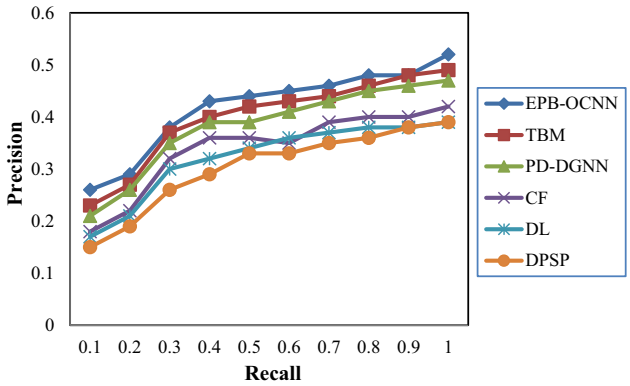

Fig. 9 Precision–recall analyses

these simulation values, the precision–recall value of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was observed to be 0.26, 0.23, 0.21, 0.18, 0.17, and 0.15, respectively. From the simulation results, optimality between precision and recall was found by the comparison of the precision–recall value of the proposed method with that of the five other existing methods. The precision–recall improvement was observed owing to the application of Thompson Optimized Convolutional Neural Network Protein Secondary Structure Prediction model. With this model, optimizing the network parameters, in turn, resulted in speeding up of overall process and hence causing an improvement in the precision–recall curve of the proposed method EPB-OCNN over the five other existing methods.

4.2.7 Performance analysis of MCC coefficient

Matthews Correlation Coefficient considers true positive, true negative, false positive and false negatives and hence considered as a balanced measure even when it is used with classes of distinct sizes. Therefore, MCC refers to a correlation coefficient between observed and predicted binary classifications, returning a value between -1 and $+1$. The coefficient of $+1$ denotes a perfect prediction made, 0 is not better than the random prediction made, and finally, -1 denotes total disagreement between prediction and observation. This is mathematically formulated as given below.

$$\mathrm{MCC} = \frac{(\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{16}$$

From Eq. (16), the Matthews Correlation Coefficient 'MCC' is measured using the true positive rate 'TP,' true negative rate 'TN,' false positive rate 'FP' and the false negative rate 'FN,' respectively. Table 9 provides the MCC resultant values of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively.

Table 9 Tabulation for MCC

Number of protein data	MCC					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	0.93	0.88	0.82	0.81	0.78	0.77
1000	0.92	0.87	0.81	0.80	0.77	0.74
1500	0.91	0.85	0.80	0.79	0.75	0.73
2000	0.90	0.84	0.78	0.76	0.73	0.71
2500	0.90	0.84	0.77	0.76	0.72	0.69
3000	0.88	0.83	0.77	0.73	0.70	0.66
3500	0.88	0.83	0.76	0.71	0.68	0.64
4000	0.85	0.81	0.75	0.70	0.66	0.63
4500	0.84	0.80	0.73	0.68	0.63	0.62
5000	0.83	0.78	0.72	0.66	0.62	0.57

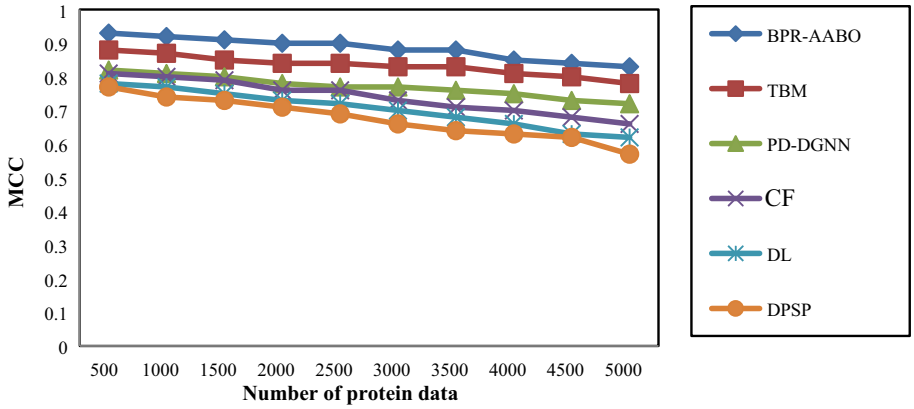

Fig. 10 MCC analyses

Finally, Fig. 10 shows a graphical representation of MCC analysis of the proposed method EPB-OCNN and the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. As illustrated in the above figure, with the increase in the number of protein data, a significant amount of decrease in the MCC is noted. This is because increasing the number of protein data sequences considered for simulation compromises the results of classifications being made for secondary protein structure prediction. This, in turn, reduces the entire MCC values for all the six methods. However, simulation analysis showed betterment of the proposed method EPB-OCNN method over the existing five state-of-the-art methods. The reason behind the improvement was the incorporation of Thompson Optimized CNN Protein Secondary Structure Prediction algorithm. By applying this algorithm, secondary protein structure prediction was made using Location-Specific Resultant Matrix as input to the CNN. Next, the convolved feature map was obtained using kernel weight and offset parameter for each protein data based on the Thompson optimization function, therefore corroborating the result.

4.2.8 Comparison of algorithms

The objective of the proposed Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method remains in utilizing the protein structure sequence features and focusing on both the loss and the precision function so that the secondary protein structure prediction can be made in an accurate and precise manner. A comprehensive comparative analysis is provided in Table 10.

With the purpose of improving the precision involved during protein structure prediction, spatial locations of every atom in a protein molecule were analyzed in [1]. Though precision was analyzed, error or loss function involved was not focused. Deep graph neural networks were employed in [2], where the analysis of both loss and accuracy was made while designing protein. However, the true and false positive rate was not concentrated while doing validation for proof-of-principle. Deep learning and particle swarm optimization algorithms were integrated into [3] to identify the species-specific S-glutathionylation with maximum accuracy. However, the precision involved in optimization was not included. Along with the accuracy, the noise factor was analyzed in [4] by means of feature engineering. Distance-based protein structure prediction via deep learning was made in [5], therefore contributing to the accuracy factor. Table 10 lists the comparison of algorithms made with different methods.

Feature selection As we have applied the energy profiles for pair of atoms, protein data were found to be highly relevant for identification, and then prediction was made using the proposed EPB-OCNN method, and this was not performed in [3, 4] 5. However, local structure and structural features were utilized in [1] 2.

Linear/nonlinear/collinear data In the proposed EPB-OCNN method, energy profiles were applied to the raw PDB dataset, and hence, relevant protein structures were said to be identified. Hence, irrespective of the type of data, by applying energy profiles, further processing was ensured. In the case of [1, 2, 4], 5, only linear data was said to be applied. However, in [3], the data type was not mentioned. Hence, upon comparison with the other state-of-the-art methods, the proposed method EPB-OCNN performed well for both linear and nonlinear data.

Optimization algorithm Thompson optimization function is used to fine tune the parameters in the proposed EPB-OCNN method that in turn reduces the processing time and hence speeds up the entire process. In the case of torsion optimization applied in [1], any small error at local residue may result in big RMSE, therefore, compromising protein structure prediction accuracy. In the case of gradient descent applied in [3], it results in redundant computation, therefore, increasing protein structure prediction time. Though deep learning and deep neural networks applied in [4], 5 learnt the network parameters by fine

Table 10 Comparison of algorithms

S. no		Methods					
		EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
1	Feature selection	Energy profiles for pair of atoms	Local structural feature selection	Structural features	No feature selection algorithm is applied separately	No feature selection algorithm is applied separately	No feature selection algorithm is applied separately
2	Linear/ nonlinear/collinear data	Any protein data type	Linear type	Linear sequence of amino acids	Not applicable	Can be applied with linear data only	Can be applied with linear data only
3	Optimization algorithm	Thompson Optimization function (with momentum set to 0.9)	Torsion angle optimization	Not applied	Gradient descent	Gradient-based weight optimization	ResNet
4	Activation function	Sigmoid activation function	Per residue network activation	Not applicable	Linear function	Linear function	Linear function
5	Hyper parameters (regularization parameter, learning rate)	Hyper parameter is optimized via Thompson function and changes according to the amino acid sequence used for simulation	Not applicable	Not applicable	0.6	0.6	0.5
6	Neural network construction method	Convolutional model	Deep residual neural network	Deep graph neural network	Deep Neural Networks	Deep learning	deep convolutional residual neural network
7	Weight calculation of nodes	Optimization model	Markov random field model	Hidden Markov models	62	Not available	Not available
8	Error handling	Gaussian prior modeling	Not handled	Spearman's correlation coefficient	Not used	Static – zero training error	Absolute error calculation

tuning the parameters, it resulted in higher convergence rate.

Activation function In the proposed EPB-OCNN method, the sigmoid activation function is utilized and hence controls between exploitation and exploration during optimization. With residue network activation used in [1], premature convergence was said to occur. Also, with linear function utilized in [3–5], and the absence of both exploitation and exploration, it resulted in premature convergence.

Hyperparameters Optimization of the hyperparameters were made using the proposed EPB-OCNN method, whereas learning rate used in three existing methods [3–5] was found to be 0.6, 0.6, and 0.5, respectively. Also, with optimized learning rates used in our work, optimal protein structure identification with a minimum time of the proposed method EPB-OCNN is said to be achieved upon comparison with the state-of-the-art methods.

Neural network construction method In the proposed EPB-OCNN method, convolutional model was used that in turn updated the protein amino acid sequencing based on the optimization process, therefore discarding premature convergence. Though deep residual and deep graphs were utilized in [1–5], with the lack of precise optimization model, optimized results were not arrived at.

Weight calculation of nodes The score values are only obtained from the Thompson optimization model, therefore ensuring optimization. In the case of the other state-of-the-art methods, it was not available, therefore proposed EPB-OCNN method is contributing to the better precision–recall and ROC curve.

Error rate In the proposed EPB-OCNN method, Gaussian prior modeling was applied based on the LSRM matrix, which in turn minimized the error or loss rate. No such provision was included to address error aspect in [2],

4. The absolute calculation was done in [5]. Zero training error was seen in [4].

4.2.9 Statistical test/analysis

The statistical test for secondary protein structure prediction is performed using McNemar test (Table 11). This McNemar test is employed while we are identifying a change in ratio for the paired protein structure. To evaluate the McNemar test, the protein structure data is said to be placed into a 2×2 contingency table, with the cell frequencies equaling the number of pairs. The McNemar test formula is then measured as given below.

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{17}$$

Figure 11 shows the McNemar test (M-test) of the proposed method EPB-OCNN and the existing five state-of-the-art methods. From the figure, it is illustrative that by performing the M-test analysis for simulation ranging between 500 and 5000 numbers of protein data, an increasing trend was found. Despite this result, with simulations conducted for 500 protein data, a comparative improvement was observed in the proposed EPB-OCNN method upon comparison with the five state-of-the-art methods. The reason behind the improvement was due to the application of the optimization function for fine tuning the hyperparameters. With this, the M-test result of the proposed method EPB-OCNN in comparison with the five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] was said to be improved by 6%, 7%, 9%, 10%, and 12%, respectively.

4.2.10 L2 Loss function

L2 loss function is applied to reduce the error. L2 loss function is measured as the sum of all the squared difference between the true value and the predicted value. This is mathematically evaluated as given below.

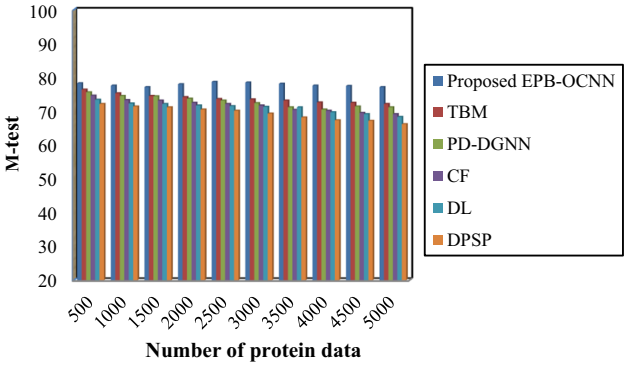

Fig. 11 M-test analysis

$$L2\,\text{Loss Function} = \sum_{i=1}^{n} \left(y_{\text{true}} - y_{\text{predicted}}\right)^2 \tag{18}$$

From Eq. (18), L2 loss function is evaluated. Table 12 provides the L2 loss function values of proposed method EPB-OCNN, and the state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyper parameters with a sliding window of 13.

Figure 12 demonstrates the L2 loss function of the proposed method EPB-OCNN and the existing five state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively. The number of protein data is taken in the horizontal direction, and the L2 loss function is observed at the vertical axis. The number of protein data is considered in the range of 500 and 5000 to conduct the simulation. The reason behind the improvement was application of the L2 loss function for fine tuning the hyperparameters with the aid of Thompson optimization algorithm. With this, the L2 loss function of the proposed method EPB-OCNN in comparison with the five state-of-the-art methods provides minimal loss. Let us considers 500 protein data for conducting the experiments in the first

Table 11 Tabulation for McNemar test

Number of protein data	McNemar test (M-test)					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	78.23	76.35	75.55	74.55	73.35	72.10
1000	77.55	75.25	74.45	73.25	72.25	71.35
1500	77.10	74.45	74.35	73.10	72.10	71.10
2000	77.95	74.10	73.65	72.45	71.65	70.45
2500	78.65	73.55	73.10	72.10	71.45	70.10
3000	78.45	73.45	72.35	71.65	71.25	69.25
3500	78.10	73.10	71.10	70.35	71.10	68.10
4000	77.55	72.55	70.45	70.10	69.65	67.25
4500	77.45	72.45	71.35	69.45	69.10	67.10
5000	77.10	72.10	71.10	69.10	68.25	66.10

Table 12 Tabulation for L2 loss function

Number of protein data	L2 loss function					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	36	100	225	400	625	900
1000	144	1406.25	5700.25	6241	7921	9120.25
1500	441	7439.06	20,093.1	25,043.1	26,732.3	31,064.1
2000	1156	17,161	61,009	75,625	85,849	88,804
2500	1914.06	55,814.1	126,914	153,077	175,352	200,256
3000	7310.25	124,256	277,202	321,489	336,980	393,129
3500	10,302.3	262,913	416,993	477,827	522,368	535,092
4000	13,456	396,900	562,500	633,616	698,896	729,316
4500	18,837.6	704,341	781,898	940,415	957,952	971,210
5000	28,056.3	893,025	1,386,506	1,523,990	1,657,656	1,677,025

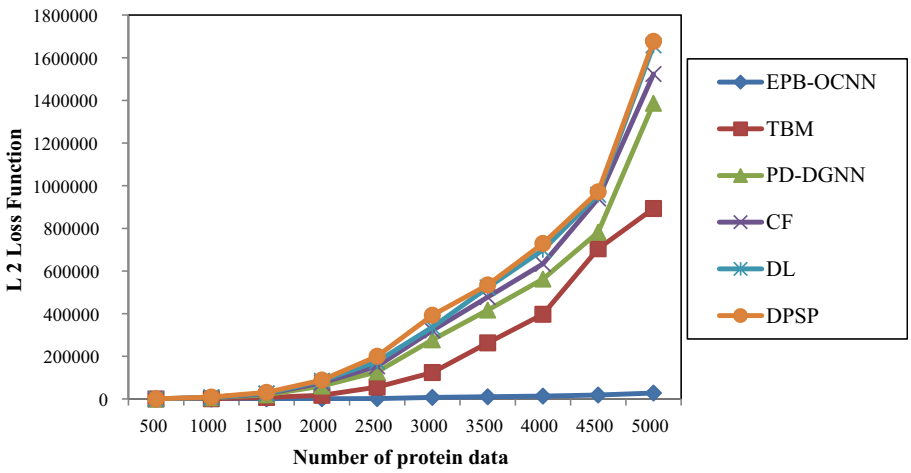

Fig. 12 L2 loss function

iteration. By applying the proposed EPB-OCNN, 494 data are correctly predicted and the L2 loss function is 36 whereas the L2 loss function of the existing TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] are 100, 225, 400, 625, and 900, respectively, followed which various performance results are observed for each method. For each method, ten different results are observed. The performance of the proposed EPB-OCNN has achieved a better result for L2 loss function than other existing methods.

4.2.11 Root mean square error

Root mean square error (RMSE) is measured by taking the square root of above mentioned L2 loss function. This is mathematically computed as given below.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \left(y_{\mathrm{true}} - y_{\mathrm{predicted}}\right)^2} \tag{19}$$

From Eq. (19), root mean square error '*RMSE*' is computed. Table 13 provides the RMSE values of proposed method EPB-OCNN, and the state-of-the-art methods, TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5], respectively, for hyperparameters with a sliding window of 13.

Figure 13 displays the root mean square error of the proposed method EPB-OCNN and the existing five state-of-the-art methods with respect to the number of protein data. The x-axis denotes the number of protein data, and the y-axis represents the root mean square error. In the experimentation process, the different number of protein data is taken as input in the ranges of 500 and 5000. From the observed results, the RMSE is minimized using the introduced EPB-OCNN method. This is because of the implementation of the loss function to optimize the hyperparameters by using the Thompson optimization algorithm. In the first iteration, 500 protein data is used to estimate the experiments. The RMSE of the proposed EPB-OCNN is 6, whereas the RMSE of the existing TBM [1], PD-DGNN [2], CF [3], DL [4], and DPSP [5] is 10, 15, 20, 25, and 30, respectively. The proposed EPB-OCNN obtained good results for RMSE compared to the state-of-the-art methods.

Table 13 Tabulation for RMSE function

Number of protein data	Root mean square error					
	EPB-OCNN	TBM	PD-DGNN	CF	DL	DPSP
500	6	10	15	20	25	30
1000	12	37.5	75.5	79	89	95.5
1500	21	86.25	141.75	158.25	163.5	176.25
2000	34	131	247	275	293	298
2500	43.75	236.25	356.25	391.25	418.75	447.5
3000	85.5	352.5	526.5	567	580.5	627
3500	101.5	512.75	645.75	691.25	722.75	731.5
4000	116	630	750	796	836	854
4500	137.25	839.25	884.25	969.75	978.75	985.5
5000	167.5	945	1177.5	1234.5	1287.5	1295

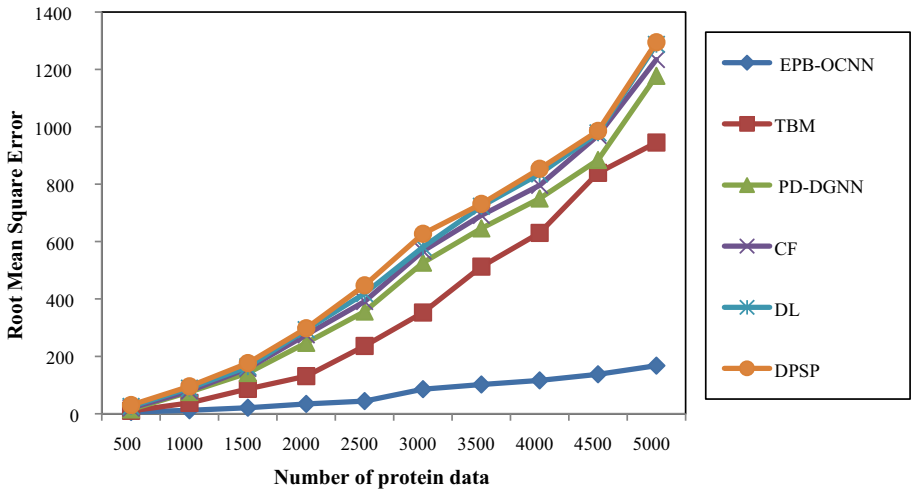

Fig. 13 Root mean square error loss function

5 Conclusion

In bioinformatics, secondary protein secondary structure prediction is a very significant task. To have better apprehension between the sequencing of proteins and their structural formations, we propose an Energy Profile Bayes and Thompson Optimized Convolutional Neural Network (EPB-OCNN) method. Secondary protein secondary structure prediction is a work of considerable importance in the area of bioinformatics. Hence it is mandatory to completely realize the purpose and protein structure. In this work, Energy Profile Legion-Class Bayes Protein Structure Identification and Thompson Optimized Convolutional Neural Network Protein Secondary Structure Prediction models are combined to predict secondary protein secondary structure. The Energy Profile Legion-Class Bayes first measures energy profiles and extracts protein sequence features for identifying secondary protein structures. Next, Thompson Optimized Convolutional Neural Network uses Location-Specific Resultant Matrix as input of convolutional neural network with optimization performed via Thompson optimization function. This is done to predict secondary protein secondary structure. Upon comparison with prediction results of state-of-the-art methods, protein structure prediction accuracy, time and precision of the proposed method EPB-OCNN method are relatively strong and can accomplish very consequential effects and possess good precision. Additional protein descriptors employing regularization techniques may also be explored. Inclusion of categorical variables to produce amino acid descriptors are also worth further investigation. Future versions of quantum computers, with their potential to simulate quantum-chemical systems, may also shed light on the protein structure prediction.

Author contributions VN contributed to the conceptualization; methodology; formal analysis and investigation; writing—original draft preparation; writing—review and editing; resources; MS was involved in the supervision.

Funding No funds, grants, or other support was received.

Declarations

Conflict of interest The authors have no financial or proprietary interests in any material discussed in this article.

Data availability Yes.

Code availability Yes.

References

1. Pearce R, Zhang Y (2021) Toward the solution of the protein structure prediction problem. J Biol Chem. https://doi.org/10.1016/j.jbc.2021.100870
2. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM (2020) Fast and flexible protein design using deep graph neural networks. Cell Syst 11(4):402–411. https://doi.org/10.1016/j.cels.2020.08.016
3. Lia S, Yub K, Wang D, Zhang Q, Liu ZX, Zhao L, Cheng H (2020) (2020) Deep learning based prediction of species-specific protein Sglutathionylation sites. Biochim Biophys Acta (BBA) Proteins Proteomics 1868(7):1–6. https://doi.org/10.1016/j.bbapap.2020.140422
4. Kandathil SM, Greener JG, Jones DT (2019) Recent developments in deep learning applied to protein structure prediction. Proteins Struct Funct Bioinform. https://doi.org/10.1002/prot.25824
5. Xu J, Wang S (2019) Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins Struct Funct Bioinform. https://doi.org/10.1002/prot.25810
6. Lai JK, Ambia J, Wang Y, Barth P (2017) Enhancing structure prediction and design of soluble and membrane proteins with explicit solvent-protein interactions. Structure 25(7):1758–1770. https://doi.org/10.1016/j.str.2017.09.002
7. Igashov I, Pavlichenko N, Grudinin S (2021) Spherical convolutions on molecular graphs for protein model quality assessment. Mach Learn Sci Technol. https://doi.org/10.1088/2632-2153/abf856
8. Nguyen SP, Li Z, Xu D, Shang Y (2017) New Deep Learning Methods for Protein Loop Modeling. IEEE Transactions on Computational Biology and Bioinformatics 16(2):596–606. https://doi.org/10.1109/TCBB.2017.2784434
9. Pearce R, Zhang Y (2021) Deep learning techniques have significantly impacted protein structure prediction and protein design. Struct Biol 68(68):104–207. https://doi.org/10.1016/j.sbi.2021.01.007
10. Wang S, Li Z, Yu Y, Xu J (2017) Folding membrane proteins by deep transfer learning. Cell Syst 5(3):202–211. https://doi.org/10.1016/j.cels.2017.09.001
11. Tsuchiya Y, Tomii K (2020) Neural networks for protein structure and function prediction and dynamic analysis. Biophys Rev 12(2):569–573. https://doi.org/10.1007/s12551-020-00685-6
12. AlQuraishi M (2021) Machine learning in protein structure prediction. Curr Opin Chem Biol Egypt J Med Hum Genet 65(65):1–8. https://doi.org/10.1016/j.cbpa.2021.04.005
13. Torrisi M, Pollastri G, Le Q (2020) Deep learning methods in protein structure prediction. Comput Struct Biotechnol J 18(18):1301–1310. https://doi.org/10.1016/j.csbj.2019.12.011
14. Afify HM, Abdelhalim MB, Mabrouk MS, Sayed AY (2021) Protein secondary structure prediction (PSSP) using different machine algorithms. Egypt J Med Hum Genet 22(1):1–10. https://doi.org/10.1186/s43042-021-00173-w
15. Adhikari B (2020) A fully open-source framework for deep learning protein real-valued distances. Sci Rep. https://doi.org/10.1038/s41598-020-70181-0]
16. Gao M, Zhou H, Skolnick J (2020) DESTINI: a deep-learning approachto contact-driven protein structureprediction. Sci Rep. https://doi.org/10.1038/s41598-019-40314-1
17. Zhong W, Gu F (2020) Predicting local protein 3D structures using clustering deep recurrent neural network. ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2020.3005972
18. Liu Z, Gong Y, Bao Y, Guo Y, Wang H, Lin GN (2021) TMPSS: a deep learning-based predictor for secondary structureand topology structure prediction of alpha-helical transmembrane proteins. Front Bioeng Biotechnol. https://doi.org/10.3389/fbioe.2020.629937
19. Yufang Q, Xiaoqi Z, Jun W, Ming C, Changjie Z (2015) Prediction of protein structural class based onLinear predictive coding of PSI-BLAST profiles. Open Life Sci 10:529–536. https://doi.org/10.1515/biol-2015-0055
20. Chen TR, Juan SH, Huang YW, Lin YC, Lo WC (2021) A secondary structure-based position-specific scoring matrix applied to the improvement in protein secondary structure prediction. PLoS ONE. https://doi.org/10.1371/journal.pone.0255076
21. Bao W, Yuan CA, Zhang Y, Han K, Nandi AK, Honig B, Huang DS (2017) Mutli-features prediction of protein translational modification sites. IEEE/ACM Trans Comput Biol Bioinform 15(5):1453–1460. https://doi.org/10.1109/TCBB.2017.2752703
22. Spencer M, Eickholt J, Cheng J (2014) A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinform 12(1):103–112. https://doi.org/10.1109/TCBB.2014.2343960
23. Gao W, Mahajan SP, Sulam J, Gray JJ (2020) Deep learning in protein structural modeling and design. Patterns. https://doi.org/10.1016/j.patter.2020.100142
24. Tunyasuvunakoo K, Adler J, Wu Z, Green T, Zielinski M (2021) Highly accurate protein structure predictionfor the human proteome. Nature. https://doi.org/10.1038/s41586-021-03828-1
25. Bouatta N, Sorger P, AlQuraishi M (2021) Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? Acta Crystallogr Sect D Struct Biol 77(8):982–991. https://doi.org/10.1107/S2059798321007531
26. Igashov I, Pavlichenko N, Grudinin S (2021) Spherical convolutions on molecular graphs for protein model quality assessment. Mach Learn Sci Technol 2(4):045005. https://doi.org/10.1088/2632-2153/abf856
27. Xu Y, Verma D, Sheridan RP, Liaw A, Ma J, Marshall NM, McIntosh J, Sherer EC, Svetnik V, Johnston JM (2020) Deep dive into machine learning models for protein engineering. J Chem Inf Model 3(60):2773–2790. https://doi.org/10.1021/acs.jcim.0c00073
28. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L (2020) Improved protein structure prediction usingpotentials from deep learning. Nature 8(428):706–710. https://doi.org/10.1038/s41586-019-1923-7
29. Evangelia IZ (2017) Prediction of protein function usinga deep convolutional neural networkensemble. Peer J Comput Sci. https://doi.org/10.7717/peerj-cs.124
30. Yang J, Zhang Y (2019) Protein structure and function prediction using I-TASSER. Curr Protocols Bioinform. https://doi.org/10.1002/0471250953.bi0508s52
31. Mehmood S, Imran M, Ali A, Munawar A, Khaliq B, Anwar F, Saeed Q, Buck F, Hussain S, Saeed A, Ashraf MY, Akrem A (2020) Model prediction of a Kunitz-type trypsin inhibitor protein from seeds of *Acacia nilotica* L. with strong antimicrobial and insecticidal activity. Turk J Biol. https://doi.org/10.3906/biy-2002-20

32. Alakuş TB, Türkoğlu İ (2021) A novel Fibonacci hash method for protein family identification by usingrecurrent neural networks. Turk J Electr Eng Comput Sci 29(1):370–386. https://doi.org/10.3906/elk-2003-116
33. Istifli ES, Tepe AŞ, Netz PA, Sarikürkcü C, Kilic IH, Tepe B (2021) Determination of the interaction between the receptor binding domain of 2019-nCoV spike protein, TMPRSS2, cathepsin B and cathepsin L and glycosidic and aglycon forms of some flavonols. Turk J Biol. https://doi.org/10.3906/biy-2104-51
34. Yilmaz C, Gok M (2021) System designs to perform bioinformatics sequence alignment. Turk J Electr Eng Comput Sci. https://doi.org/10.3906/elk-1105-22
35. Sureyya Rifaioglu A, Doğan T, Jesus Martin M, Cetin-Atalay R, Atalay V (2019) DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Sci Rep. https://doi.org/10.1038/s41598-019-43708-3