BOLTZMAN LENNARD POTENTIAL AND KULLBACK LEIBLER DEEP NEURAL NETWORK FOR PROTEIN STRUCTURE PREDICTION)

Mr.VARANAVASI NALLASAMY
Senior Associate,
Cognizant Technology Solutions Pvt. Ltd,
Coimbatore, Tamil Nadu, India

Ms.MALARVIZHI SESHIAH
Assistant Professor
Department of Computer Science,
Thiruvalluvar Government Arts College,
*Rasipuram,*Namakkal,
Tamil Nadu, India.

***Abstract*—Proteins are indispensable to life, and comprehending their structure can smooth an impersonal perception of their consequence. Procuring an explicit protein structure representation is an essential step toward understanding the strengthening of biology. Despite recent advancements in experimental methods have profusely improved the potentialities to experimentally ascertain protein structures, the gap between the protein sequences number and well-known protein structures is persistently growing. A deeper insight into protein structure prediction is one of the methods to fill this gap. Recently, the protein structure prediction field has noticed a great deal of fosters due to Deep Learning (DL)-based methods as evidenced in the most recent protein structure prediction.The proposed Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD) method for protein structure prediction consists of three layers. They are input layer, hidden layers and output layer. The amino acid sequences acquired from the protein data bank dataset are provided in the input layer. Then, in the first hidden layer, Boltzman and Lennard Jones Potential-based protein structure identification is performed. Followed by which, in the second hidden layer, KullbackLeibler Relative Entropy-based protein structured prediction is made. Finally, by means of Affine map activation function the actual predicted protein structure is provided in the output layer. Simulations are conducted using Biopython tool to measure the efficiency of the proposed method using different metrics like, protein structure prediction time, protein structure prediction accuracy, precision, specificity, recall, F-measure and ROC curve.**

Keywords—*Deep Learning, Protein Structure Prediction, Boltzman, Lennard Potential, KullbackLeibler Divergence*.

I. INTRODUCTION

With the evolution of computational methods for predicting three-dimensional (3D) protein structures from the protein sequence has gone a long way into two distinct

complementary paths that concentrate on either physical interactions or evolutionary history. On the other hand, protein intrinsic disorder is due to the absence of the inclination of a protein to fold into a well investigated and rigid formation. Owing to this reason the dynamic protein structures can be inferred in an experimentally manner because their backbone angles differ according to time due to their innate flexibility.

Despite recent advancements the existing methods for protein structure prediction fall far short of atomic accuracy, specifically with the absence of the availability of the homologous structure. AlphaFolda neural network-based method was proposed in [1] where initially the computational process was described with the purpose of predicting the protein structures in a regular fashion. With this design atomic accuracy was found to be improved even in cases where no similar structure was known.

An ensemble of deep Squeeze-and-Excitation residual inception and long short-term memory (LSTM) networks called, SPOT-Disorder2 was proposed in [2] with the objective of predicting protein intrinsic disorder. Here, the input was obtained from information acquired through evolutionary structure for predicting single dimensional structural properties. As a result, the SPOT-Disorder2 method was found to be advantages over the previous method based on LSTM networks alone.

Occurrence of proteins in a natural fashion denotes only an infinitesimal subset of all probable amino acid sequences obtained by the evolutionary process to carry out a discrete biological function. However, proteins with more robustness involving exorbitant thermal firmness, resistance to degradation, tighter binding might not been explored. To address this issue in this work, Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD) method for protein structure prediction is proposed in this work.

The contributions of BLP-KLD method comprises of the following:

- To propose a Deep Neural Network based method called, Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD)for protein structure identification and prediction.
- To design Boltzman and Lennard Jones Potential-based protein structure identification based on the interaction between atoms by means of Orientation Discrete Quasi Chemical Approximation that in turn returns or predicts computationally efficient protein structures.
- To present KullbackLeibler Relative Entropy-based protein structure prediction via pair-wise distance by employing intermolecular amino acid pair potential, therefore contributing to precision and recall.
- To evaluate the performance through extensive simulations based on protein data bank dataset to measure the metrics, protein structure prediction

time, accuracy, precision, specificity, recall, F-measure.

The organization of the paper is as given below: Section 2 introduces the related works of protein structure identification and prediction proposed by different research persons, Section 3 introduces the data and methods used in experiments, Section 4 introduces the experimental results and comparison with other methods in performance, and Section 5 introduces the conclusion of this paper.

II. RELATED WORKS

The study of the protein structure prediction method is specifically split into two distinct parts. One is the development of protein amino acid sequence identification strategies, and the other is the application of prediction algorithms. This section briefly reviews related research.Significant advancement has of late been coerced by exploiting genetic information. It is probable hence to deduce which amino acid residues are in proximity by examining skewness in homologous sequences that in turn assists in protein structure prediction of. A neural network was trained in [3] for making precise predictions between pairs of residues, that provided supplementary information regarding protein structure. With this information a mean potential force was enumerated that precise determined protein shape.

The swift growth in the protein numbers and the heterogeneity of these activities take exception to computational methods for automated function prediction. A graph convolutional network called, DeepFRI was proposed in [4] with the purpose of predicting protein functions by capitalizing sequence features acquired from protein structures, therefore ensuring diverse confident function predictions. However, high resolution structure prediction still remains a main challenge to be addressed. Deep learning technique was implemented in [5] to prediction high resolution sequences. Yet another deep learning method using remodeling [6] was performed for missing sequences that remained a long challenging issue as far as bioinformatics was concerned.

Molecular attribute predictions are an elementary chore in the area of drug discovery. Precise prediction using computational methods would paramount quicken the comprehensive procedure of identifying recommended drug candidates in an accelerated and reasonable manner.

A holistic review of prediction of molecular property using graph neural networks was investigated in [7]. An elaborative review of algorithms and the consequence of evolution and co-evolution to attain improved predictions by means of the prevailing part of Deep Learning techniques to predict protein structures along with the challenges and opportunities were proposed in [8].

Residue-residue contact information is indispensable for comprehending the implementation of protein folding, and hasbeen applied in de novoprotein structure prediction. In [9], deep residue contact predictor, called, DeepConPred2, was proposed that revealed considerably enhanced execution and adequately minimized running time. A deep transfer learning method was designed in [10] with the purpose of predicting membrane protein contact map by learning relationship between sequence structure, which addressed the issue of solved membrane protein structures.

Illustrating the influence of amino acid mutations on protein-protein inter-linkage takes part in the decisive part in designing of drug. In [11], a novel deep learning method based on structure called, GeoPPI was designed with the purpose of predicting transpose of binding affinity upon mutations. Also on the basis of the three-dimensional protein structure, GeoPPI initially learnt geometric characterization that also encoded protein structure topology features by means of self-supervised learning model. These representations were then utilized as features to train gradient boosting with the purpose of predicting protein-protein binding changes. Moreover, meaningful features were also learnt that in turn distinguished interactions between atoms in protein structures.

Protein contacts possess crucial information for the discernment of protein structure and hence, predicting contact from sequence is predominant issue. Over the past few years, electrifying advancement has been made on this issue, however, the predicted protein contacts without numerous sequence homologs is still of low quality and hence not found its use for de novo structure prediction. A novel deep learning method was proposed in [12] that predicted by means of combining evolutionary coupling (EC) and information acquired via progression preservation via an ultra-deep neural network utilizing dual deep residual neural networks, therefore ensuring accurate contact assisted folding.

Protein secondary structure prediction (SSP) includes different types of applications. Nevertheless, there has been moderately restricted enhancement in accuracy for years. Input features were operated in Secondary Structure Element based Position Specific Scoring Matrix (SSE-PSSM) [13], based on which a state-of-the-art coordinate of machine learning features were found to be established. With this the overall accuracy was said to be improved. On the contrary, a substantial fragment of prevailing protein interactions were not known and also experimental evaluation was found to be resource consuming. To this objective, a Graph Signal Processing based method was applied in [14] for designing protein-protein interaction (PPI) network as a graph. Moreover, a Markovian model of the signal on graph was designed that in turn validated the characterization of congruence node to deduce graph edges, therefore ensuring accuracy.

Comprehending the sequence-to-structure association is pivotal for the successful structure prediction. Several methods provided in the previous works included single shallow learning model however those models possessed the disadvantages of lacking sequence-to-structure relationship. To further increase the protein structure prediction performance, a novel Clustering Recurrent Neural Network (CRNN) was proposed in [15], therefore providing prediction accuracy. Yet another fast and flexible protein design was performed in [16]

employing Deep Graph Neural Networks, therefore improving accuracy on missing residues with high improved score.

Protein structure prediction has been considerably boosted by deep learning however, most endeavors are dedicated to template free modeling. Nevertheless a small number of deep learning techniques are designed for TBM (template-based modeling), a sough after mechanism for protein structure prediction. A new and novel method called, New Deep-learning Threader (NDthreader) was proposed in [17] to address the issues concerning TBM. Yet another ThreaderAI was designed in [18] with the purpose of predicting residue-residue aligning probability matrix by combining sequence profile, structural features, and residue-residue contacts. With these structural formulation templates query alignment was constructed by appertaining dynamic programming algorithm on probability matrix, therefore improving alignment accuracy in a significant manner.

A novel computational method, Trans Membra Protein Secondary Structure (TMPSS) was presented in [19], to predict the secondary structures in non-transmembrane parts, therefore contributing to accuracy. Recurrent neural network was applied in [20] for accurate prediction via torsion angles. Protein structure prediction and its influence on deep learning techniques were investigated in [21]. A review of deep learning techniques for modeling and designing protein structure was designed in [22].

A novel computational method using stack sparsed auto encoder was proposed in [23] to implement protein-protein interaction. Here with the application of Legendre moment feature extraction model predictive accuracy was said to be enhanced significantly. Yet another method for protein-protein interaction utilizing backward and forward recurrent neural network was proposed in [24]. In [25], a deep learning neural network model called, DeepHiFam producing high accuracy to classify proteins in a hierarchical fashion across numerous levels simultaneously was proposed.

Based on the aforementioned materials and methods, in this work, Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD) method is proposed for protein structure prediction.

III. METHODOLOGY

In conventional methods, protein structure predictions from data are performed by means of physical equations and modeling. However, deep learning puts forward a distinct criterion in which algorithms deduce in an automatic fashion or learn a relationship between inputs and outputs from a set of postulates. The proposed Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD) method consists of three layers. They are one input layer, two hidden layers and one output layer. Initially, the amino acid sequences acquired from the protein data bank dataset forms as input and provided in the input layer. Then, in the first hidden layer, protein structure identification is made based on the interaction between atoms by means of Orientation Discrete Quasi Chemical Approximation via Boltzman and Lennard Jones Potential function. Next, in the second hidden layer, protein structure prediction is done by means of KullbackLeibler Relative Entropy. Finally, the predicted protein structure is arrived at the output layer using affine map activation function. The structure of BLP-KLD method is given below

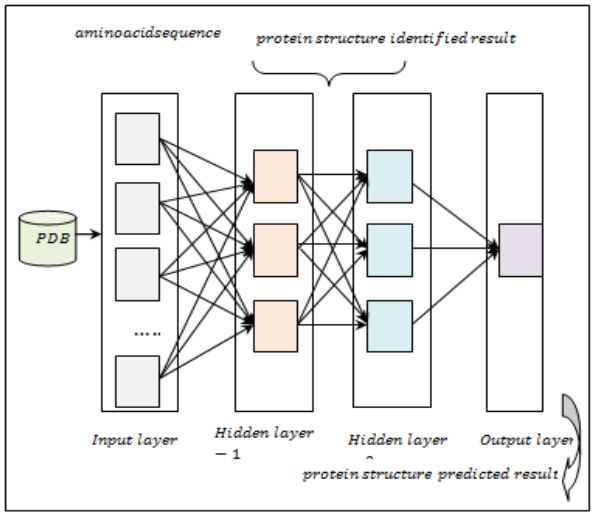

Fig. 1. Structure of BLP-KLD method

As shown in the above figure, let us consider a group of 'n' training samples consisting of features '$F = F_1, F_2, \dots, F_n$' in an input space 'X' (i.e., amino acid sequences), and equivalent labels 'y' in certain output space 'Y' where '$(x_i, y_i), i = 1,2,..n$' are sampled independently. Moreover, let us consider a function '$f: x \to y$' in function class 'Cl', and a loss function '$Loss: y * y$' that estimates how much 'f(x)' varies from the equivalent label 'y'. The objective of the supervised deep learning is to identify a function '$f \in Cl$' that reduces the expected loss '$Exp[Loss(f(x), y)]$' for '(x, y)' sampled from overall sample set.Then, the optimization form is represented as given below.

$$Opt = \min \frac{1}{n} \sum_{i=1}^{n} Loss[f(x_i), y_i] \quad (1)$$

From the above equation (1), the optimization function 'Opt' is evolved on the basis of the loss function 'Loss' with respect to the input '$f(x_i)$' and output space '$f(y_i)$' respectively. Then, the affine map for the above optimality function is represented for 'L' layers with activation function '$\sigma(.)$' as given below.

$$f(x) = M_L \sigma\left(M_{L-1}\sigma\left(M_{L-2}\sigma(M_{L-3}) \dots \dots \sigma(M_{L-N})\right)\right) \quad (2)$$

In the first hidden layer (i.e., hidden layer 1), Quasi Chemical Approximation for understanding molecular mechanism to evaluate interactions between atoms via Boltzman and Lennard Jones Potential function is designed. Then, to evaluate interactions between atoms via Inverse

Boltzman of two atom types 'p' and 'q' is mathematically expressed as given below.

$$U(IaDis|p,q) = -kT\ln \frac{Prob(IaDis|p,q)}{RS(IaDis)} \quad (3)$$

From the above equation (3), interactions between atom types 'p' and 'q' is obtained by employing inter atom distance 'IaDis', a Boltzman constant 'k', corresponding temperature 'T' to its reference state 'RS' of the inter atom distance respectively. Followed by which the summation given below refers to all amino acid pairs in the protein data bank (PDB). This is mathematically stated as given below.

$$\sum U(IaDis|p,q) = -kT\sum \ln \frac{Prob(IaDis|p,q)}{RS(IaDis)} \quad (4)$$

Followed by which the intermolecular amino acid pair potential is estimated for obtaining intermolecular amino acid pair interactions. This is formulated using the Lennar Jones potential as given below.

$$P = Pot_{LJ}(IaDis|p,q)4\varepsilon\left[\left(\frac{\sigma}{(IaDis|p,q)}\right)^{12} - \left(\frac{\sigma}{(IaDis|p,q)}\right)^{6}\right] \quad (5)$$

From the above equation (5), the Lennar Jones potential 'Pot_{LJ}' is estimated based on the dispersion energy of intermolecular amino acid pair potential 'ε', distance between two interacting particles or amino acid pair '$IaDis|p,q$' and 'σ' representing the distance at which the potential 'Pot' is zero. Followed by the protein structure identification based on the interaction between atoms, protein structure prediction are said to be employed in determining the protein three dimensional shape from its amino acid sequence. However, it is probable to deduce which amino acid residues are in contact by determining the correlation in analogous sequences that in turn assists in protein structure prediction. In our work, the neural network is trained based on the intermolecular amino acid pair potential that in turn would accurately make the predictions given its identified sequence. The neural network predictions in our work include pair-wise distances using KullbackLeibler Relative Entropy. The entropy score is first written as given below.

$$H = KD(IF_i[P]||\ BF_i[P]) \quad (6)$$

From the above equation (6), the entropy score 'H' is obtained based on the Kullback Divergence function 'KD' with respect to interposed frequency 'IF_i' and background frequency 'BF_i' of the corresponding amino acid identified protein sequence 'P'. Followed by which the protein structure prediction made via pair-wise distance using KullbackLeibler Relative Entropy in turn would provide specific information about the structure (i.e., therefore enhancing the accuracy rate) is mathematically formulated as given below.

$$RE = \sum_i IF_i \log 2\left(\frac{IF_i[P]}{BF_i[P]}\right) \quad (7)$$

From the above equations (7), the relative entropy 'RE' for each sequence is measured. Finally, the predicted results 'PR_i' of the corresponding amino acid identified protein sequence is obtained as given below.

$$PR_i = \sum_{i=1}^{20} RE_i \quad (8)$$

The pseudo code representation of Boltzman Lennard Potential and KullbackLeibler Divergence-based Deep Neural Network is given below.

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

Output: Accurate, precise and timely protein structure prediction

Step 1: Initialize sample 'x', temperature 'T', protein data size 'n', layers 'L'
Step 2: Initialize interposed frequency 'IF_i' and background frequency 'BF_i'
Step 3: Begin
//Input layer
Step 4: Obtain features 'F' in an input space 'X' (i.e., amino acid sequences),
Step 5: Formulate optimization function as in equation (1)
Step 6: Obtain affine map as in equation (2)
//Hidden layer – 1 [protein structure identification]
Step 7: For two atom types 'p' and 'q'
Step 8: Evaluate interactions between atoms via Inverse Boltzman as in equation (3)
Step 9: Obtain summation to all amino acid pairs in protein data bank as in equation (4)
Step 10: Evaluate intermolecular amino acid pair potentialas in equation (5)
Step 11: End for
//Hidden layer – 2 [protein structure prediction]
Step 12: For each intermolecular amino acid pair potential
Step 13: Estimate entropy score as in equation (6)
Step 14: Evaluate KullbackLeibler Relative Entropy as in equation (7)
Step 15: Evaluate predicted results as in equation (8)
Step 16: End for
//Output layer [predicted results]
Step 17: For each intermolecular amino acid pair potential
Step 18: If '$PR_i \geq 0$'
Step 19: Correct protein structure prediction
Step 20: End if
Step 21: If '$PR_i < 0$'
Step 22: Incorrect protein structure prediction
Step 23: End if
Step 24: End for
Step 25: End

Algorithm Boltzman Lennard Potential and KullbackLeibler Divergence Deep Neural Network-based protein structure prediction

As given in the above Boltzman Lennard Potential and KullbackLeibler Divergence Deep Neural Network algorithm for protein structure identification and prediction, the overall step is divided into four distinct stages via deep learning. In the first stage, the amino acid sequences are acquired from the protein data bank dataset and provided as input in the input layer. Next, in the second stage, or in the first hidden layer, identification of protein structure is made by means of Boltzman and Lennard Potential function. With this function, time and accuracy with which the protein structure is identified are said to be ensured. In the third

stage, or in the second hidden layer, the protein structure prediction is made by employing KullbackLeibler Divergence function, therefore improving precision and recall. Finally, in the fourth stage or the output layer, using affine map activation function, the predicted results are provided.

IV. RESULTS AND DISCUSSION

In this project, we propose a weighted rule-mining technique (say, RANWAR or Rank-based Weighted Association Rule-mining) to rank the rules using two novel rule-interestingness measures, Rank-based Weighted Condensed Support (WCS) and Weighted Condensed Confidence (WCC) measures to bypass the pro blem. These measures are basically depended on the rank of items (genes). Using the rank, we assign weight to each item. RANWAR generates much smaller number of frequent item sets than the state-of-the-art association rule mining algorithms. Thus, it saves time of execution of the algorithm.

We employed the following evaluation metrics to have an extensive and comprehensive perception of the deep neural learning process of the method and to study the performance of the methods. We use extensively applied used measures, protein structure prediction accuracy, time, precision, recall, specificity, F1-measure, and receiver operating characteristics (ROC) which are further discussed in the following section.

4.1.1. Protein structure prediction accuracy: Thisis a significant performance measurement of a protein structure prediction method that estimates the percentage ratio of correct prediction amount to the total amount of protein data samples as given below.

$$PSP_{acc} = \sum_{i=1}^{n} \frac{PS_{CP}[aaseq]}{P_i} * 100 \qquad (9)$$

From the above equation (9), the protein structure prediction accuracy 'PSP_{acc}' is evaluated on the basis of protein structure correctly predicted 'PS_{CP}' for the respective amino acid sequence '[aaseq]' and the protein data samples 'P_i'. It is measured in terms of percentage (%).

4.1.2 Protein structure prediction time: The second important measure for protein structure prediction with the given amino acid sequence is the time it consumes in performing the process. This is mathematically formulated as given below.

$$PSP_{time} = \sum_{i=1}^{n} P_i * \text{Time}\,[PR_i] \qquad (10)$$

From the above equation (10), the protein structure prediction time 'PSP_{time}' is measured on the basis of the protein data samples 'P_i' and the time consumed in prediction 'Time $[PR_i]$'. It is measured in terms of milliseconds (ms).

4.1.3 Precision: Precision assists in providing an evaluation of the protein structure prediction method for imbalanced dataset. Precision is mathematically formulated as given below.

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \qquad (11)$$

From the above equation (11), the precision rate 'Precision' is measured based on the ratio of number of true-positives 'TP' and the number of true and false-positive predictions 'TP + FP' of the method.

4.1.4 Recall: Recall is the metric that dispenses an estimation of the ratio of protein data samples from a class that is correctly predicted by the method.

$$\text{Recall} = \frac{(TP)}{(TP+FN)} \qquad (12)$$

From the above equation (12), the recall rate 'Recall' is measured based on the ratio of number of true-positives 'TP' and the number of true and false negative predictions 'TP + FN' of the method.

4.1.5 F-measure: F-measure as given below provides an estimation to take the overall idea of method performance when both the recall and precision are significant. Protein structure prediction is a domain area where concentration has to be paid not only to the number of correct predictions but also to the incorrect prediction of the method. The f-measure is mathematically formulated as given below.

$$F - \text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (13)$$

4.1.6 Specificity: Specificity refers to the ratio of those who do not have the condition (judged negative by the 'Gold Standard') that received a negative result on this test. Specificity refers to the tests potentiality to correctly reject predicted protein structures without a condition. Specificity of a test is the proportion of those predicted protein structures that do not posses as it is and test negative for the condition.

$$\text{Specificity} = \frac{(TN)}{(TN+FP)} \qquad (14)$$

From the above equation (14), the specificity rate 'Specificity' is measured based on the true negative 'TN' and false positive 'FP' rate respectively.

4.1 Discussion

4.1.1 Performance measure of protein structure prediction accuracy

We tested the BLP-KLD method with two different protein structure prediction methods for 5000 distinct numbers of protein data as shown in table 1. The results in table 1 show that BLP-KLD method has the higher accurate rate than the other two considered methods, AlphaFold [1] and SPOT-Disorder2 [2] respectively.

Table 1. Protein structure prediction accuracy comparison with BLP-KLD, AlphaFold [1] and SPOT-Disorder2 [2]

Number of protein data	Protein structure prediction accuracy (%)		
	BLP-KLD	AlphaFold	SPOT-Disorder2
500	99	98.4	97.4
1000	98.90	97.55	93.75

1500	98.80	95.55	91.85
2000	98.50	94.75	88.95
2500	98.45	91.85	86.75
3000	98.35	89.55	83.75
3500	98.20	86.65	82.85
4000	98.10	85.55	82.55
4500	97.35	82.65	81.65
5000	97.15	82.40	77.75

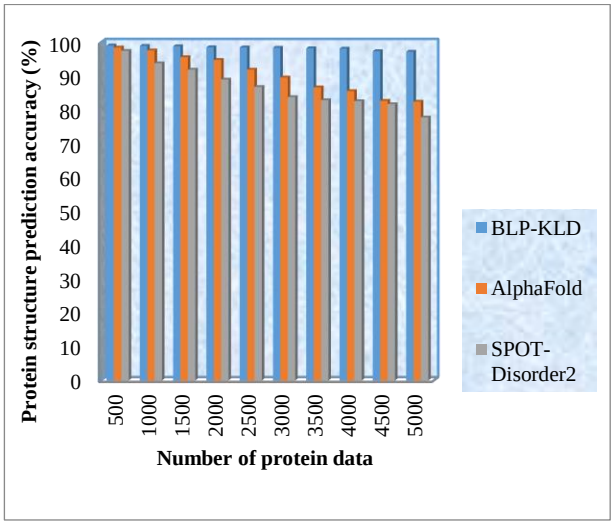

Fig. 2. Graphical representation of protein structure prediction accuracy

Figure 2 given above shows the protein structure prediction accuracy using three different methods, BLP-KLD, AlphaFold [1] and SPOT-Disorder2 [2] respectively. From the figure it is inferred that the accuracy rate comparatively better than [1] and [2]. The reason behind the improvement was owing to the application of Boltzman and Lennard Potential function. By applying this function, accurate identification are made, hence improving the accuracy using BLP-KLD by 9% compared to [1] and 14% compared to [2].

4.1.2 Performance measure of protein structure prediction time

In this section, the protein structure prediction time is tested for the proposed BLP-KLD method with two different protein structure prediction methods [1] and [2] for 5000 protein samples data as shown in table 2. The results in table 2 show that BLP-KLD method has the lesser prediction time than the other two state-of-the-art methods, AlphaFold [1] and SPOT-Disorder2 [2] respectively.

Table 2. Protein structure prediction time of BLP-KLD, AlphaFold [1], SPOT-Disorder2 [2]

Number of	Protein structure prediction time (ms)		
protein data	BLP-KLD	AlphaFold	SPOT-Disorder2
500	52.5	57.5	67.5
1000	58.20	80.15	100.15
1500	70.09	110.35	125.65
2000	80.09	125.15	170.15
2500	110.25	160.15	190.15
3000	122.15	190.25	220.35
3500	150.55	205.15	265.15
4000	183.05	225.35	290.15
4500	210.25	265.15	320.25
5000	253.05	280.35	365.25

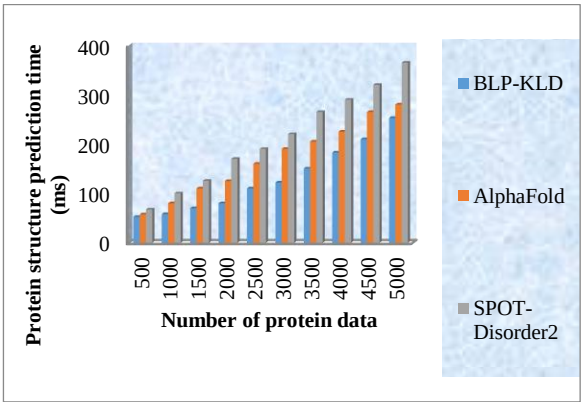

Fig. 3. Graphical representation of protein structure prediction time

Figure 3 given above illustrates the protein structure prediction time using proposed BLP-KLD and two different existing methods, AlphaFold [1] and SPOT-Disorder2 [2] respectively. From the figure the time is said to be reduced than [1] and [2]. The reason behind the minimization of time is due to the application of Lennard Potential function that identifies the proteins based on the interaction between atoms, therefore minimizing the prediction time using BLP-KLD by 25% compared to [1] and 39% compared to [2].

4.1.3 Performance measure of precision

The precision rate is measured in this section for the proposed BLP-KLD method with two different protein structure prediction methods [1] and [2] with respect to 5000 distinct protein samples data as shown in table 3. From the results in table 3 is inferred that BLP-KLD method has the greater precision rate than the other two state-of-the-art methods, AlphaFold [1] and SPOT-Disorder2 [2] respectively.

Table 3. Precision comparison with BLP-KLD, AlphaFold [1] and SPOT-Disorder2 [2]

Number of protein data	Precision		
	BLP-KLD	AlphaFold	SPOT-Disorder2

500	0.99	0.94	0.78
1000	0.95	0.89	0.79
1500	0.95	0.88	0.80
2000	0.99	0.88	0.81
2500	0.99	0.91	0.91
3000	0.97	0.90	0.92
3500	0.95	0.91	0.89
4000	0.97	0.90	0.89
4500	0.98	0.90	0.89
5000	0.95	0.90	0.89

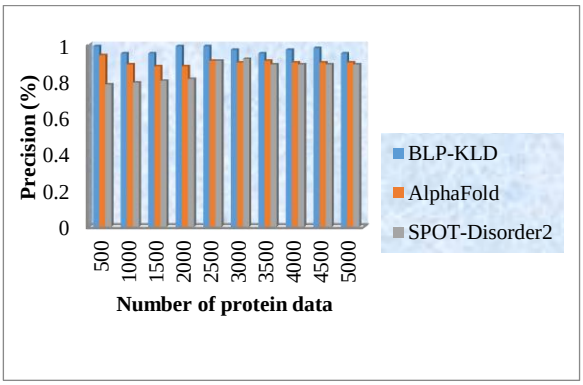

Fig. 4. Figure 4 Graphical representation of precision rate

Figure 4 given above shows the precision rate achieved using the proposed BLP-KLD and two distinct existing methods, AlphaFold [1] and SPOT-Disorder2 [2] with respect to 5000 different protein data samples. From the figure it is inferred that the precision rate is said to be improved using BLP-KLD when compared to [1] and [2]. The reason behind the improvement of precision was owing to the application of KullbackLeibler Divergence function that in turn made precise prediction using BLP-KLD and therefore improving by 8% compared to [1] and 13% compared to [2].

4.1.4 Performance measure of ROC curve

The ROC curve is provided in this section for the proposed BLP-KLD method with two state-of-the-art methods [1] and [2] with respect to 5000 distinct protein samples data is illustrated in table 4.

Table 4. ROC curve comparison with BLP-KLD, AlphaFold [1] and SPOT-Disorder2 [2]

False positive rate	True positive rate		
	BLP-KLD	AlphaFold	SPOT-Disorder2
0.1	0	0	0
0.2	0.44	0.33	0.29
0.3	0.59	0.43	0.34
0.4	0.74	0.58	0.47
0.5	0.82	0.64	0.59
0.6	0.87	0.73	0.62
0.7	0.94	0.84	0.74
0.8	0.96	0.91	0.86
0.9	0.98	0.92	0.88
1.0	1	0.93	0.90

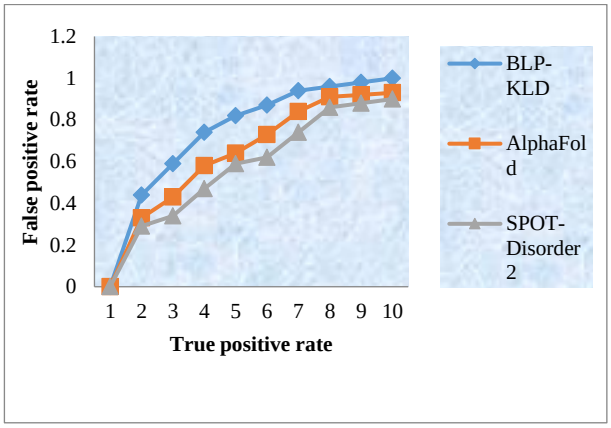

Fig. 5. Graphical representation of ROC curve

Figure 5 given above shows the illustration of ROC curve with the true positive rate of corresponding protein structure prediction in x axis and the false positive rate of the respective protein structure prediction in y axis respectively. From the above inferences shows better curve inclination by using proposed BLP-KLD method upon comparison with the state-of-the-art methods, [1] and [2]. The reason behind the improvement was that the correct prediction of protein structure using BLP-KLD was found to be better than [1] and [2] owing to the application of Boltzman Lennard Potential and KullbackLeibler Divergence Deep Neural Network algorithm. By applying this algorithm the falsification of prediction was eliminated during the first stage itself in the hidden layer 1 and was not proceed to the hidden layer 2. Hence, betterment was found to be observed using BLP-KLD method. The improvement was found to be 20% upon comparison with [1] and 36% upon comparison with [2].

V. CONCLUSION

The prediction of protein structures initiating from sequences of amino acid has prevailed an exceptional issue in biological research. Over the past few years, there requires a necessity to design high-accuracy protein structure prediction mechanisms, owing to the increasing interval between known protein sequences and the empirically persistent structures. In this work, Boltzman Lennard Potential and KullbackLeibler Divergence-based deep learning (BLP-KLD) for high accuracy empirically persistent structures towards protein structure prediction. First, protein structure identification was made by applying Boltzman Lennard Potential model. Second the actual protein structure prediction was performed using KullbackLeibler Divergence-based deep neural network

model. With this, precise and accurate prediction was made in a timely manner.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, “Highly accurate protein structure prediction with AlphaFold”, Nature, Springer, Aug 2021 [AlphaFold]

[2] [2] Jack Hanson, Kuldip K. Paliwal, Thomas Litfin, Yaoqi Zhou, “SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning”, Genomics Proteomics Bioinformatics, Elsevier, Feb 2019 [SPOT-Disorder2]

[3] [3] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, “Improved protein structure prediction usingpotentials from deep learning”, Nature, Springer, Jan 2020

[4] [4] Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman,Daniel Berenberg, TommiVatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis,Ramnik J. Xavier, Rob Knight, Kyunghyun Cho & Richard Bonneau, “Structure-based protein function prediction usinggraph convolutional networks”, Nature Communications, Oct 2021

[5] [5] Robin Pearce and Yang Zhang, “Toward the solution of the protein structure predictionproblem”, Journal of Biological Chemistry, Springer, Jun 2021

[6] [6] Son P. Nguyen, Zhaoyu Li, Dong Xu, Yi Shang, “New Deep Learning Methods for Protein Loop Modeling”, IEEE Transactions on Computational Biology and Bioinformatics, Oct 2017

[7] [7] Oliver Wieder, Stefan Kohlbachera, MélaineKuenemann, Arthur Garona, Pierre Ducrota, Thomas Seidela, Thierry Langera, “A compact review of molecular property prediction with graph neural networks”, Drug Discovery Today: Technologies, Elsevier, Oct 2020

[8] [8] Mirko Torrisi, Gianluca Pollastri, Quan Le, “Deep learning methods in protein structure prediction”, Computational and Structural Biotechnology Journal, Elsevier, Jan 2020

[9] [9] Wenze Ding, WenzhiMao, Di Shao,Wenxuan Zhang, Haipeng Gong, “DeepConPred2: An Improved Method for the Prediction of Protein Residue Contacts”, Computational and Structural Biotechnology Journal, Elsevier, Nov 2018

[10] [10] Sheng Wang, Zhen Li, Yizhou Yu, Jinbo Xu, “Folding Membrane Proteins by Deep Transfer Learning”, Cell Systems, Elsevier, Sep 2017

[11] [11] Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, Jian Peng, “Deep geometric representations for modelingeffects of mutations on protein-proteinbinding affinity”, PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1009284 August 4, 2021

[12] [12] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, Jinbo Xu, “Accurate De Novo Prediction of ProteinContact Map by Ultra-Deep Learning Model”, PLOS Computational Biology | DOI:10.1371/journal.pcbi.1005324 January 5, 2017

[13] [13] Teng-Ruei Chen, Sheng-Hung Juan, Yu-Wei Huang, Yen-Cheng Lin, Wei-Cheng Lo, “A secondary structure-based position-specificscoring matrix applied to the improvement inprotein secondary structure prediction”, PLOS ONE | https://doi.org/10.1371/journal.pone.0255076 July 28, 2021

[14] [14] Stefania Colonnese, Manuela Petti, Lorenzo Farina, Gaetano Scarano, Francesca Cuomo, “Protein-Protein Interaction Predictionvia Graph Signal Processing”, IEEE Access, Oct 2021

[15] [15] Wei Zhong, Feng Gu, “Predicting Local Protein 3D Structures Using Clustering Deep Recurrent Neural Network”, IEEE Transactions on Computational Biology and Bioinformatics, Jul 2020

[16] [16] Alexey Strokach, David Becerra,CarlesCorbi-Verge, Albert Perez-Riba,Philip M. Kim, “Fast and Flexible Protein Design Using Deep GraphNeural Networks”, Cell Systems, Elsevier, Oct 2020

[17] [17] Fandi Wu, Jinbo Xu, “Deep template-based protein structureprediction”, PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1008954 May 3, 2021

[18] [18] Haicang Zhang and Yufeng Shen, “Template-based prediction of proteinstructure with deep learning”, BMC Genomics, Aug 2020

[19] [19] Zhe Liu, Yingli Gong, Yihang Bao, Yuanzhao Guo, Han Wang and Guan Ning Lin, “TMPSS: A Deep Learning-BasedPredictor for Secondary Structureand Topology Structure Prediction ofAlpha-Helical TransmembraneProteins”, Frontiers in Bioengineering and Biotechnology, Jan 2021

[20] [20] YongChang Xu, TianJun ShangGuan, XueMing Ding&Ngaam J. Cheung, “Accurate prediction of proteintorsion angles using evolutionarysignatures and recurrent neuralnetwork”, Scientific Reports, Oct 2021

[21] [21] Robin Pearce and Yang Zhang, “Deep learning techniques have significantly impacted protein structure prediction and protein design”, Current Opinion in Structural Biology, Elsevier, Jul 2021

[22] [22] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J. Gray, “Deep Learning in Protein StructuralModeling and Design”, Pattern, Cell Press, Dec 2020

[23] [23] Yanbin Wang,Zhuhong You,Liping Li,Li Cheng,Xi Zhou,Libo Zhang,Xiao Li,and Tonghai Jiang, “Predicting Protein Interactions Using a Deep LearningMethod-Stacked Sparse Autoencoder Combined with aProbabilistic Classification Vector Machine”, Complexity, Wiley, Dec 2018

[24] [24] Lei Yang,Yukun Han,Huixue Zhang,Wenlong Li,and Yu Dai, “Prediction of Protein-Protein Interactions with LocalWeight-Sharing Mechanism in Deep Learning”, BioMed Research International, Hindawi, Jun 2020

[25] [25] PahalageDhanushkaSandaruwan, ChampiThusangiWannige, “An improved deep learning model forhierarchical classification of protein families”, PLOS ONE | https://doi.org/10.1371/journal.pone.0258625 October 20, 2021