MultiPredGO: Deep Multi-Modal Protein Function Prediction by Amalgamating Protein Structure, Sequence, and Interaction Information

Swagarika Jaharlal Giri, Pratik Dutta, *Student Member, IEEE*, Parth Halani, and Sriparna Saha, *Senior Member, IEEE*

***Abstract*—Protein is an essential macro-nutrient for perceiving a wide range of biochemical activities and biological regulations in living cells. In this work, we have presented a novel multi-modal approach, named *MultiPredGO*, for predicting protein functions by utilizing two different kinds of information, namely protein sequence and the protein secondary structure. Here, our contributions are threefold; firstly, along with the protein sequence, we learn the feature representation from the protein structure. Secondly, we develop two different deep learning models after considering the characteristics of the underlying data patterns of the protein sequence and protein 3D structures. Finally, along with these two modalities, we have also utilized protein interaction information for expediting the efficiency of the proposed model in predicting the protein functions. For extracting features from different modalities, we have utilized various variations of the convolutional neural network. As the protein function classes are dependent on each other, we have used a neuro-symbolic hierarchical classification model, which resembles the structure of Gene Ontology (GO), for effectively predicting the dependent protein functions. Finally, to validate the goodness of our proposed method (*MultiPredGO*), we have compared our results with various uni-modal along with two well-known multi-modal protein function prediction approaches, namely, INGA and DeepGO. Results show that the overall performance of the proposed approach in terms of accuracy, F-measure, precision, and recall metrics are better than those by the state-of-the-art methods. *MultiPredGO* attains an average 13.05% and 30.87% improvements over the best existing comparing approach (DeepGO) for cellular component and molecular functions, respectively.**

***Index Terms*—Protein function prediction, multi-modality, deep learning, gene ontology, protein sequence, protein structure.**

Manuscript received January 4, 2020; revised May 2, 2020, July 3, 2020, and August 14, 2020; accepted September 4, 2020. Date of publication September 8, 2020; date of current version May 11, 2021. *(Swagarika Jaharlal Giri and Pratik Dutta contributed equally to this work.) (Corresponding author : Pratik Dutta.)*

Swagarika Jaharlal Giri, Pratik Dutta, and Sriparna Saha are with the Department of Computer Science, and Engineering, Indian Institute of Technology, Patna 801103, India (e-mail: swagarika95@gmail.com; pratik24111991@gmail.com; sriparna.saha@gmail.com).

Parth Halani is with the Department of Computer Science and Engineering, Indian Institute of Information Technology Guwahati, Guwahati 781015, India (e-mail: parthhalani05@gmail.com).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2020.3022806

Availability and implementation: Source code: https://github.com/SwagarikaGiri/Multi-PredGO, Web-server: http://multipred.co.in.

I. Introduction

INTERPRETING the protein function is a critical and indispensable aspect to comprehend an organism at a molecular level, and it is also a help in pharmaceutical implications. Protein functions or GO terms are not independent classes, but they are naturally dependent on each other. They are arranged in many-to-many parent-child relationships, i.e., hierarchical structures of the GO terms and a protein can be annotated to any ontology level and more than one GO terms within an ontology.

There are various approaches present in the literature in predicting protein functions. Protein sequence plays a crucial role in understanding the protein functions. One of the oldest approaches to anticipate the function of a new protein is by finding a substantial sequence similarity to the known protein sequence [1], [2]. Also, Anfinsen *et al.* [3] experimentally proved that protein folding is reversible, i.e., the underlying sequence determines the tertiary protein structure under native condition. Hence the observation suggests, similar sequence determines a similar structure and this has a lead to surge of utilizing protein structure and sequence for predicting protein functions [4]–[6]. Along with the sequence and structure of proteins, the protein-protein interaction network [7] plays a significant role in understanding the functionalities of the proteins. The protein interaction network helps to understand the higher level Gene Ontology term or function. Hence, the protein interaction information has emerged as an imperative knowledge for solving different computational biology problems [8]–[12].

These above facts have motivated computational biologists to build AI-based model for efficiently predicting the protein functions. Recent advancements in deep learning [13][14] have unleashed new avenues in solving different well-known problems ranging from computational biology [15], machine translations [16], speech recognition [17], image captioning [18]. Subsequently, there is a notable trend of using deep learning for solving different problems related to bioinformatics [19]–[21] including protein function prediction [22], [23].

As these AI technologies mature and become more accessible to researchers, the next frontier is to collect "multi-modal" data

for the same set of subjects and conduct integrative analyses using multi-level views on the same phenomena [24]–[26]. Multi-modal approaches show promising results compared to the conventional single modal based models as the prior can capture cumulative insightful information from the underlying multifaceted datasets. Simultaneously, the integrative research of multi-modal data is embraced by the biomedical research community due to its promising success for the diagnosis of the disease along with personalized disease prediction [23], [27].

In this regard, researchers of computational biology field have investigated extensively in a multi-modal approach. Kulmanov *et al.* [23] proposed a deep learning-based hybrid approach for predicting protein functions by utilizing protein sequence and protein-protein interaction networks. Also in [28], a novel probabilistic chain-graph-based approach is proposed for predicting protein functions by utilizing the Gene ontology and the knowledge of interspecies relationships. Recently, three orthogonal approaches, i.e., sequence similarity, domain architecture, and protein-protein interaction network data are integrated to predict the protein functions [29]. But none of the above methods has considered the 3D structural perspective of the proteins.

Drawing inspirations from these findings, in this article, we have developed a deep multi-modal protein function model named **MultiPredGO**. Here, along with the protein structure and underlying protein sequence, we have utilized protein interaction information for predicting protein functions. For extracting features from protein structure and sequence, we have used two separate deep learning models, while for protein interaction network we have used a pre-trained knowledge graph embedding. This higher-level integrated feature is fed to a neuro-symbolic classification model for the final protein function prediction. We have compared our proposed multi-modal architecture, **MultiPredGO**, with various single modal along with two well-known multi-modal protein function prediction approaches, INGA [29] and DeepGO [23]. For the comparative analysis, we have used the performance metrics which were used in the CAFA challenge [30] which are described in the section II of the **supplementary material**. The obtained results illustrate that the proposed multi-modal architecture performs better than other comparative models in terms of predicting the protein functions.

II. Methods and Materials

Our contributions to the current work are summarized in three stages.

1) In this study, we prepare a dataset where each gene spans over two modalities, i.e., underlying amino acid sequence and 3D PDB structure. Also, we have collected proteins and their respected Gene Ontology (GO) annotation information from SwissProt's[1] manually annotated data for better understanding the protein function and the relationships with different GO terms.
2) In the second stage, for each modality, we have developed two different deep learning based models to extract the features and accurately analyzed the extracted features.
   i) *Protein sequence:* For the protein sequence (FASTA), we have first converted the sequence in a trigram which is followed by an embedding layer. The embedding output is finally fed to the convolutional layer [31] to extract the features.
   ii) *3D structure:* For the 3D structure, we first extract the 3D structure from Protein Data Bank (PDB).[2] The 3D structure is then converted into four types of 3D voxelized representations which are further fed to ResNet-50, a popular convolutional neural network (CNN) model to extract the features from the protein structures.
   iii) Finally, the extracted features from two modalities are concatenated with protein interaction information to make the resultant features more informative. Finally, the resultant feature is pass through the neuro-symbolic hierarchical classification for predicting the final result.

The detailed descriptions of the important steps of the proposed method are provided in the subsequent subsections. The proposed deep multi-modal architecture is shown in Fig. 3

A. Problem Formalization

In computational bioinformatics domain, the function of a particular protein (P_i) is characterized as a set of biochemical functions $\{f_1, f_2, \ldots, f_n\}$, where n is the number of molecular functions related to a particular protein. Suppose, we are given a multi-modal protein ($\{P_i\}_{i=1}^{N}$) dataset, where $\forall i \mid P_i$ consists of two modalities; protein sequence, P^i_{Seq}, and protein structure, P^i_{Struc}. Formally, our deep multi-modal model ($\mathbb{M}$) predicts the functionalities of proteins ($P_{i|i\in\{1,2,..N\}}$) by utilizing protein interaction information (I_{PPI}) along with protein sequence (P^i_{Seq}) and structure (P^i_{Struc}) modality which can be mathematically formulated as

$$f_{HC}\left(f_{MLP}\left(\mathbb{M}_1(P^i_{Seq}) \oplus \mathbb{M}_2(P^i_{Struc}) \oplus I_{PPI}\right)\right) \quad (1)$$

Here, $\mathbb{M}_1$ and $\mathbb{M}_2$ are two different deep neural network based models for handling sequence and structure modality, respectively. $\oplus$ represents the concatenation operator. The integrated feature representation of protein sequence and structure along with PPI information is fed to a *multi-layer perceptron* (f_{MLP}) which is finally given as an input of *hierarchical classifier* (f_{HC}) for predicting protein functions.

B. Feature Extraction Using Multi-Modal Architecture

In this subsection, we have described three components of our proposed multi-modal architecture in detail: (1) detailed architecture for extracting features from the protein sequence; (2) detailed architecture for extracting features from protein 3D structures, and (3) detailed architecture of the proposed multi-modal approach, where protein-protein interaction information is used to predict the protein function.

1) Extracting Features From the Protein Sequence: In this study, the protein sequence (P^i_{Seq}) acts as one modality among the two modalities that we have considered for protein function

[1][Online]. Available: https://www.uniprot.org/

[2][Online]. Available: http://www.rcsb.org/

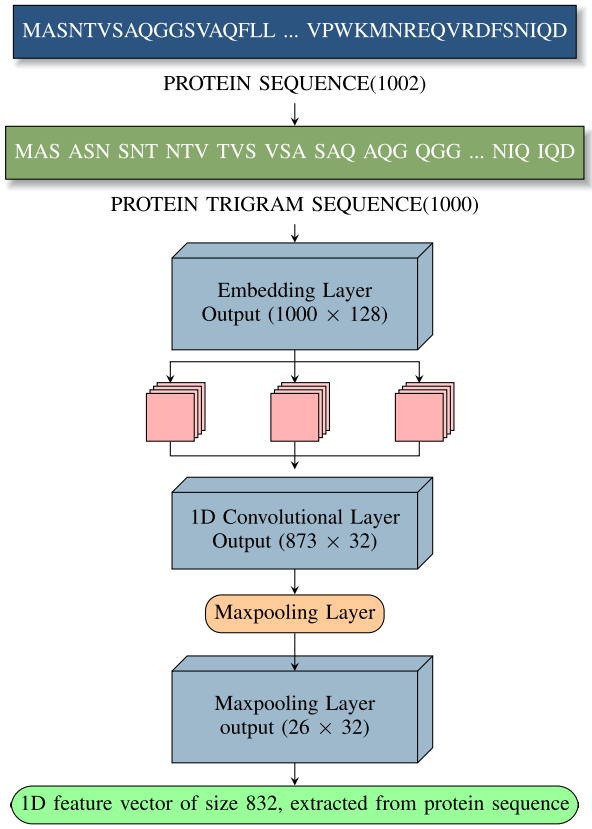


Fig. 1. The proposed deep architecture for extracting features from the protein sequence.

prediction. For the protein sequence, we have downloaded the FASTA format of the underlying amino acid(AA) sequence. In FASTA format, each protein is made up of a unique combination of 20 amino acids (a_k), i.e., $P^i_{Seq} = \{\{a_k\}_{k=1}^{20}\}^+$. The length of the FASTA sequence varies in the range of 2 to 35,213 but for computational limitations, we have only considered those proteins whose lengths are < 1002, i.e., in this study $2 \leq |P^i_{Seq}|1002$. We have then built an AA trigram vocabulary where each unique trigram is represented by 1-base index. If the length of the sequence is less than 1002, we have padded the vector with sequence of zeros. Using this trigram vocabulary with 8000 unique trigrams, we have converted the sequence of 1002 AAs into a vector of 1000 indices ($\{x_k\}_{k=1}^{1000}$). Each index was represented using a one-hot encoding, i.e., $\forall x_k = \{0,1\}^+$. However, due to the sparse nature of one-hot encoding, it leads to limited generalization performance [23]. To remove this computational bottleneck, we have generated a dense embedding of size 128 across each trigram, $x_k \in \mathbb{R}^{128}$ is the dense embedding across each trigram. Thus a protein sequence of 1002 is converted into a matrix of size 1000×128. On this matrix, we have used convolutional layer followed by the maxpool layer for extracting the features. The detailed hyper-parameter setting of the convolutional neural network is described in the section I of the **supplementary material**. The final feature representation of the sequence modality is described as follows

$$F^i_{Seq} : \mathbb{M}_1(P^i_{Seq}) = f_{maxpool}(f_{conv}(\{x_k\}_{k=1}^{1000})) \quad (2)$$

The detailed architecture for extracting features from the protein sequence using the embedding layer along with the convolutional layers is described in Fig. 1.

2) Extracting Features From the Protein 3D Structure: In this work, the 3D structure of the proteins is used as another modality of the proteins. Also, researchers of [29] proved that the 3D structure of proteins plays a more important role than the protein sequence for predicting the protein function. In this work, we have considered four 3D volumetric representations (binary(R_B), hydropathy(R_H), isoelectric(R_I) and charge(R_C)) of protein.

For extracting features from the 3D structure, firstly, we have mapped each raw 3D PDB structure into a grid structure. In this regard, we have utilized a binary volumetric shape with volume elements named **voxels**. Voxels are the 3D structure cube (V) with a fixed grid length l. One of the crucial modules in this process is determining the size of the cube, as all the proteins have different sizes and shapes [6]. The size of the cube has to be large enough to accommodate a sufficient number of proteins in consideration and also small enough to satisfactorily represent most of the proteins. Also, the length of the grid is another parameter that has to be tuned properly for the proper representation of the 3D structure.

Therefore, this scaling problem is analogous to evaluating a maximum radius ($r_{\max}$) that would be able to accommodate most protein structures of desirable sizes. Alternatively, it also has to be small enough so that most enzymes are represented at a satisfactory resolution. Hence, we consider the homothetic transformation ratio (λ) for maintaining above both conditions

$$\lambda = \lfloor \frac{l}{2} - 1 \rfloor \times \frac{1}{r_{\max}} \quad (3)$$

This transformation allows us to scale all proteins to their desired sizes. In this transformation, we have ignored the side chains of the protein structure and concentrated on their **backbone** atoms that are carbon, nitrogen, and calcium. In our experiment, we have used the grid length(l) as 32 and $R_{\max}$ as 40.

Now to capture the features of each 3D structure, we capture four different attributes (binary(R_B), hydropathy(R_H), isoelectric(R_I) and charge(R_C)) of the volumetric representation. To extract features from each volumetric representation, it is fed to a well-known pre-trained deep neural network model named *Resnet-50* [32] followed by a **dense layer**(d_1) with output $= 1024$ neurons. The output of the **dense layer** is batch normalized to $[-1,1]$ and applied to the next **dense layer**(d_2) of 256 neurons which is followed by a **sigmoid** *activation function* for the classification. Hence the mathematical formulation of extracting features from the structure modality is defined as follows

$$F^i_{Struc} : M_2(P^i_{Struc}) = f_{MLP(1024,256)}(f_{Resnet-50}(I_1) \oplus f_{Resnet-50}(I_2) \oplus f_{Resnet-50}(I_3) \oplus f_{Resnet-50}(I_4))$$

In this study, the size of the final feature vector of structure modality is 1024, i.e., $F^i_{Struc} \in \mathbb{R}^{1024}$. The detailed architecture

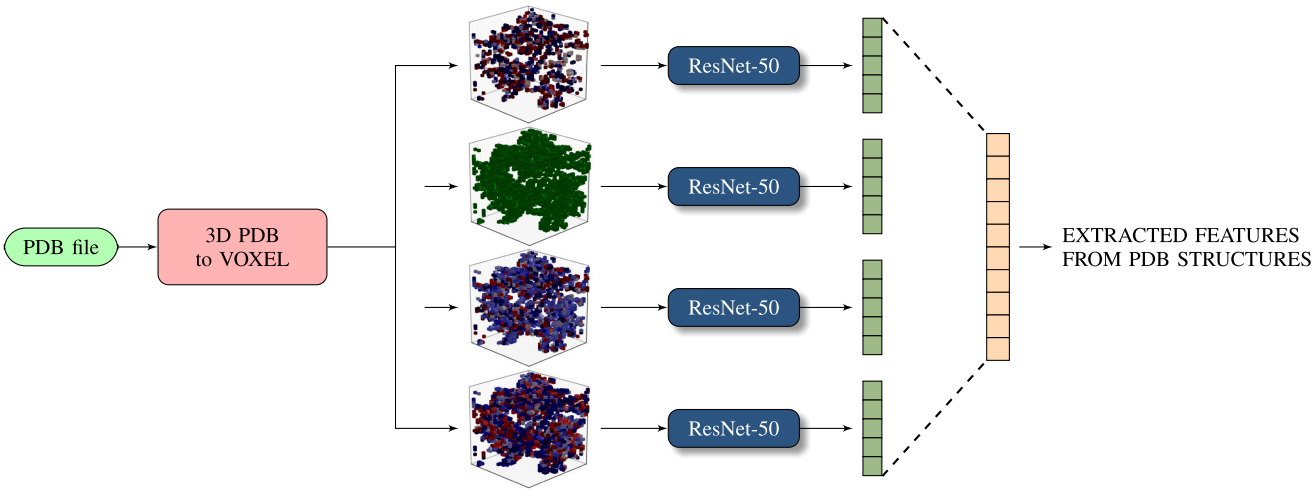


Fig. 2. Flowchart of extracting features from the 3D PDB structures using ResNet-50. The 3D PDB structure first convert into four voxel representation which fed to ResNet-50 to extract the features.

of extracting the features from the 3D structure is depicted in the Fig. 2.

3) Incorporating PPI Information With the Multi-Modal Architecture: In this section, we have briefly discussed two key points of the proposed architecture. One is incorporating the PPI information (I_{PPI}) with the extracted features and the other is a detailed description of the multi-modal architecture.

1) **Extracting features from protein-protein interaction network:** As protein-protein interaction information (I_{PPI}) is also essential to understand the protein functions, we have integrated the protein interaction information with the extracted features of the above two modalities. To do that, firstly, we have extracted the knowledge graph embedding based feature for the protein-protein interaction network. The protein-protein interaction network for multi-species is obtained from STRING [33] database. EggNOG [34] database was used to extract the confidence score and orthological relationships between the proteins. A confidence score of 300 was used to filter less frequent connections. The final network had 84,78,935 proteins with 1,90,649 edge types and a total of 11,58,66,95,610 edges. A knowledge graph embedding of 256 vectors across each protein is generated following the approach mentioned in [35]. This obtained protein interaction network based embedding is combined with the feature vectors extracted from the protein sequence and 3D structures.
2) **Multi-modal architecture:** The core step of this study is to integrate features from the three sources ($F^i_{Struc}, F^i_{Seq}, I_{PPI}$). The final concatenated feature vector ($\mathbb{F} = \{F^i_{Struc} \oplus F^i_{Seq} \oplus I_{PPI}\}$) is then passed to a fully connected dense layer which serves as an input to the neuro-symbolic model. Due to the large size of GO, we developed three separate models, one for each sub-ontology with selected GO terms to imitate the hierarchical structure of the GO, a series of fully connected layers is also considered, one for each class. Two types of layers are considered in this architecture.
   - The **Classification layer** that has input from the output of the first fully connected layer. The Classification layer has sigmoid as the activation function and it is associated with each class of the gene ontology. This layer is responsible for classifying proteins for the prediction classes.
   - The **Maximum merge layer** is used to select the maximum value of the classification layer of the classes and their children. It enables us to maintain consistency in the hierarchical classification and also preserves the hierarchical relationships between the classes. The details of the hierarchical classification model is described in section III of the **supplementary material**.

   Such a hierarchical structure of fully connected layers ensures discrimination between each of the classes in a hierarchical manner. Each of the layers intends to learn features that would discriminate it from its sub-classes. The *maximum merge layer* of internal nodes and the *classification layer* of the leaf nodes form the final output model. The detailed diagram of the proposed multi-modal architecture is depicted in Fig. 3.

III. Results and Comparative Discussion

In this section, the details of the dataset formation and a brief comparative analysis of the proposed technique with different state-of-art methods are concisely illustrated.

A. Dataset Formation

In this study, we have collected the genes which span over two modalities, i.e., protein sequence and 3D structure. Along with this condition, we considered those genes for which we can obtain the relationships between the genes and the functional annotations in terms of Gene Ontology (GO).

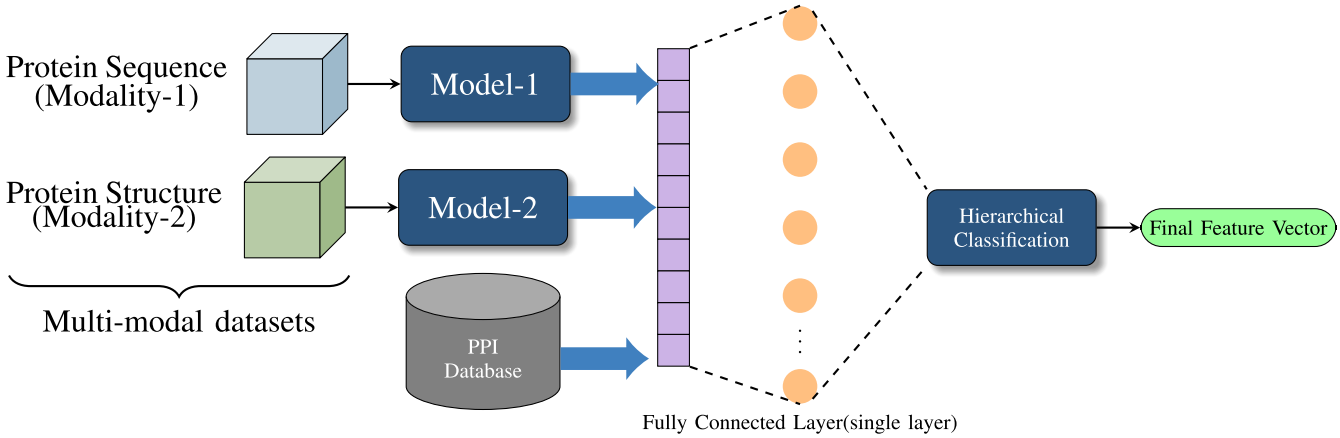


Fig. 3. The proposed deep multi-modal architecture for predicting protein functions. For different modalities, we use different deep uni-model architectures. Along with protein sequence and protein structure, we integrated protein interaction information with the extracted features of the deep learning models.

To understand the proteins and their relationships with their respective GO terms, we have exploited SwissProt's[3] manually annotated data. From there, the protein sequence, its respective accession number (entry), length and GO class annotations are extracted. Due to our computational limitations, we could not consider the entire protein dataset. We have only considered those proteins that have length equals to 1002, so that we can convert each protein sequence into a vector of 1000 indices. We have further selected only those entries that have at least one experimental evidence code and removed those protein sequences having any ambiguous amino acid code(B, J, X, Z, U, O). Here, we have selected only potential output class that has 250, 50, 50 protein annotations in *biological process* (BP), *molecular functions* (MF), *cellular component* (CC) classes, respectively. In this regard, the top 589 GO terms for MF, 932 GO terms for BP, 436 GO terms for CC are considered for final output class and three separate models are developed for them. In the next step, we have downloaded the protein 3D structure from Protein Data Bank (PDB) [4] for each gene sequence. But as not all the proteins have PDB ids corresponding to them, we have considered only those proteins which have 3D PDB structures.

The final dataset is prepared by considering the two modalities (protein sequence and protein 3D structure) of proteins. This dataset is not prepared from scratch, rather it is an exemplification of a benchmark dataset that is used in [23]. Thus the final dataset has 11536, 9982, 10741 proteins and their respective annotations are in BP, MF, CC, respectively. The final dataset can be accessed from the supporting online repository https://github.com/SwagarikaGiri/Multi-PredGO **supporting online repository**.

B. Results and Comparative Discussion

In this subsection, we analyzed the performance of the proposed deep multi-modal architecture with different state-of-the-art-methods along with the different uni-modal architectures.

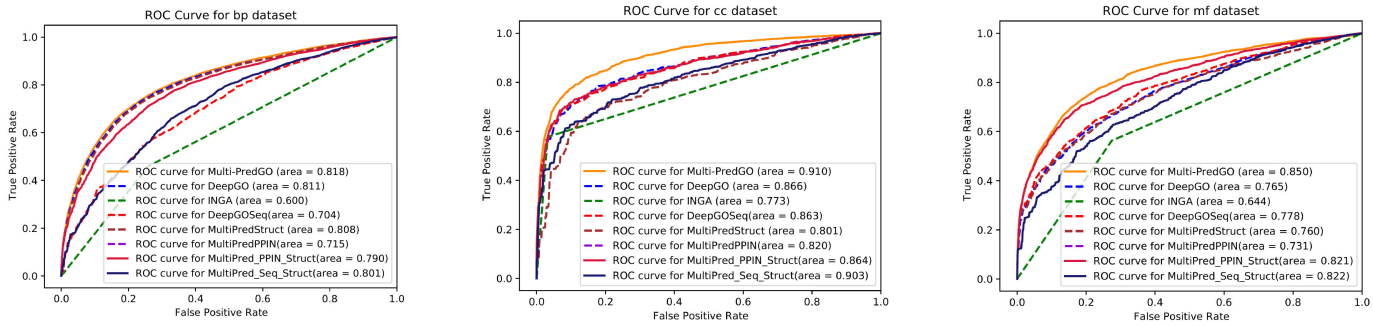
[3][Online]. Available: https://www.uniprot.org
[4][Online]. Available: https://www.rcsb.org/

TABLE I
COMPARATIVE ANALYSIS OF THE PROPOSED MULTI-MODAL ARCHITECTURE WITH DIFFERENT MODELS IN TERMS OF BIOLOGICAL PROCESS

	Biological Process				
	F_{max}	$AvgPre$	$AvgRc$	AUC	MCC
$DeepGO_{Seq}$	0.2756	0.2878	0.2745	0.7042	0.2465
$MultiPred_{Struct}$	0.3161	0.2917	0.3352	0.8032	0.2972
$MultiPred_{PPIN}$	0.2261	0.2072	0.2576	0.7174	0.2074
$DeepGO$	**0.3332**	0.3136	**0.3528**	**0.8161**	0.2805
$INGA$	0.1971	0.1576	0.2632	0.6034	0.1270
$MultiPred_{PPINStruct}$	0.2849	0.2928	0.2863	0.7945	0.2694
$MultiPred_{SeqStruct}$	0.3162	0.2987	0.3470	0.8037	0.2888
$MultiPredGO$	0.3278	**0.3257**	0.3439	**0.8169**	**0.2829**

The proposed method utilizes protein sequence and protein 3D voxel structures along with the protein interaction information. For understanding the benefits of each modality, we have also reported the results of each modality after analyzing/utilizing the extracted features of each modality, separately. Finally, we have performed 4-fold cross validation and the integrated multi-modal architecture shows better performance in terms of all five performance metrics. Also as comparing methods, we have used different deep learning-based uni-modal architectures for protein function prediction. Along with the uni-modal architecture, we have also compared the performance of the proposed method with two multi-modal methods e.g. INGA [29] and DeepGO [23]. The details of all these comparing methods are described in section IV of the supplementary material.

We have performed a comparative analysis of the performance of the above-mentioned models and the corresponding results are reported in Table I, II, and III. As we can sub-categorize the protein functionalities into three groups, i.e., biological process, cellular component, and molecular function, we have reported the comparative analysis based on these three categories in Table I, II, and III, respectively. From Table I, it is clear that for the biological process, the proposed method, **MultiPredGO**, performs well than other state-of-the-art methods in predicting the protein functions in terms of average precision (0.3257) and MCC (0.2829). For the biological process, the overall performance of the proposed multi-modal architecture is almost the



(a) ROC curve for the Biological Process. (b) ROC curve for Cellular Components. (c) ROC curve for Molecular Functions.

Fig. 4. Receiver Operating Characteristic (ROC) curve along with Area under Curve (AUC) value for the proposed deep multi-modal protein function prediction model, MultiPredGO. ROC curve of each sub-figure represents for the three specific protein class functions.

TABLE II
COMPARATIVE ANALYSIS OF THE PROPOSED MULTI-MODAL ARCHITECTURE WITH DIFFERENT MODELS IN TERMS OF CELLULAR COMPONENT

	Cellular Component				
	F_{max}	$AvgPre$	$AvgRc$	AUC	MCC
$DeepGO_{Seq}$	0.4367	0.4548	0.4363	0.8633	0.4052
$MultiPred_{Struct}$	0.4706	0.4865	0.4565	0.8019	0.4233
$MultiPred_{PPIN}$	0.3985	0.3562	0.4526	0.8262	0.2805
$DeepGO$	0.4767	0.4748	0.4866	0.8620	0.4240
$INGA$	0.2230	0.2300	0.2160	0.7700	0.1410
$MultiPred_{PPINStruct}$	0.4876	0.4758	0.4956	0.8593	0.4350
$MultiPred_{SeqStruct}$	0.4876	0.4407	0.6054	0.9062	0.4802
$MultiPredGO$	**0.5366**	**0.5684**	**0.5184**	**0.9102**	**0.5192**

TABLE III
COMPARATIVE ANALYSIS OF THE PROPOSED MULTI-MODAL ARCHITECTURE WITH DIFFERENT MODELS IN TERMS OF MOLECULAR FUNCTION

	Molecular Function				
	F_{max}	$AvgPre$	$AvgRc$	AUC	MCC
$DeepGO_{Seq}$	0.3039	0.3000	0.3027	0.7703	0.2434
$MultiPred_{Struct}$	0.2746	0.2632	0.2852	0.7652	0.2376
$MultiPred_{PPIN}$	0.2605	0.2656	0.2572	0.7332	0.2433
$DeepGO$	0.2605	0.2672	0.2572	0.7632	0.2433
$INGA$	0.2010	0.1583	0.275	0.6402	0.1547
$MultiPred_{PPINStruct}$	0.3234	0.3308	0.3263	0.8289	0.2883
$MultiPred_{SeqStruct}$	0.3274	**0.3664**	0.2845	0.8232	0.2943
$MultiPredGO$	**0.3671**	0.3630	**0.3685**	**0.8505**	**0.3123**

same as the best performing comparing method, i.e., *DeepGO*. Though for biological process, *DeepGO* performs same as the **MultiPredGO**, but the overall performance of *MultiPredGO* is better than other state-art-of-the-art methods for predicting protein functionalities in terms of cellular component and molecular function.

From Table-II, it is clearly evident that the performance of the **MultiPredGO** model is superior to other comparing methods with respect to all five performance metrics. For the cellular component process, *MultiPredGO* attains an improvement of 12.01%, 18.95%, 6.02%, 5.6% and 22.45% over the best existing comparative methods with respect to $F_{\max}$, $AvgPre$, $AvgRc$, AUC and MCC metrics, respectively. Hence, MultiPredGO shows an average of 13.05% improvements over the best existing protein function prediction method. Along with this, MultiPredGO also performs better than bi-modal architectures ($MultiPred_{PPINStruct}$ and $MultiPred_{SeqStruct}$) and uni-modal architectures, where we have considered protein structure ($MultiPred_{Struct}$) and protein-protein interactions ($MultiPred_{PPIN}$), separately.

Also, from the Table III, it is clearly evident that the proposed multi-modal architecture (*MultiPredGO*) performs better than other comparing methods for all performance metrics except average precision. For $F_{\max}$, $AvgRc$, AUC and MCC, *MultiPredGO* attains improvements of 40.09%, 43%, 8% and 28%, respectively, over the best comparing method for molecular function. Hence, after analyzing the results reported in Table I, II and III, we can infer that the overall performance of the proposed multi-modal architecture is better than other existing methods. The best values are highlighted in bold fonts in different tables. Also, in Fig. 4, we have also plotted the ROC curve of the proposed method along with other comparative methods.

The superiority of the proposed multi-modal architecture is happened due to the inherent power of the deep learning along with the voxel representation of the 3D protein structures. Deep learning models perform well if we represent the data in a very informative way. The more the model understands the data, it performs well accordingly. In this regard, we represent the 3D protein structures by four voxel representations which can be integrated to generate the final abstract representation of the 3D protein structure. Due to this informative representation, deep models accurately predict the protein functions than other existing methods. Also to prove that the better results attained by our proposed method are statistically significant, we have carried out the Welch's t-test [36]. The p-value of the test is less than 5% and the results for the cellular component are reported in Table-I of the **supplementary material**.

IV. CONCLUSION & FUTURE WORK

This paper presents a deep multi-source multi-modal architecture that can accurately predict the protein functions. Here, along with the two modalities we have also utilized the protein interaction information while predicting the protein functions. The two modalities are: the underlying amino acid sequences and the 3D protein data bank (PDB) structures. The main novelty of the work is to exploit 3D PDB structures as 2D voxels using ResNet-50. Finally, the results of the multi-modal architecture show that the multi-source multi-modal architecture performs

well than other uni-modal architectures and other state-of-the-art-methods in terms of different performance metrics.

In the future, we will explore different deep learning models (e.g., 3D convolutional neural network, capsule network, etc.) to extract features from 3D PDB structures. Also, along with these two modalities, we will add more modalities to perform a comprehensive study. We would also like to extract features from other popular pre-trained models and explore the possibility of better pre-trained models for feature extraction.

IV. Acknowledgment

Pratik Dutta acknowledges the Visvesvaraya Ph.D. Scheme for Electronics and IT, an initiative of the Ministry of Electronics and Information Technology (MeitY), Government of India, for fellowship support. Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (Formerly Media Lab Asia) for carrying out this research.

References

[1] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[2] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003.

[3] C. Anfinsen and H. Scheraga, "Experimental and theoretical aspects of protein folding," *Adv. Protein Chem.*. New York, NY, USA: Elsevier, 1975, vol. 29, pp. 205–300.

[4] J. C. Whisstock and A. M. Lesk, "Prediction of protein function from protein sequence and structure," *Quart. Rev. Biophys.*, vol. 36, no. 3, pp. 307–340, 2003.

[5] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: A unified platform for automated protein structure and function prediction," *Nat. Protoc.*, vol. 5, no. 4, pp. 725–738, 2010.

[6] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, and E. I. Zacharaki, "EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation," *PeerJ*, vol. 6, 2018, Art. no. e4750.

[7] P. Dutta and S. Saha, "Fusion of expression values and protein interaction information using multi-objective optimization for improving gene clustering," *Comput. Biol. Med.*, vol. 89, pp. 31–43, 2017.

[8] P. Dutta, S. Saha, and S. Gulati, "Graph-based hub gene selection technique using protein interaction information: Application to sample classification," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2670–2676, Nov. 2019.

[9] P. Dutta, S. Saha, S. Chopra, and V. Miglani, "Ensembling of gene clusters utilizing deep learning and protein-protein interaction information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2918523.

[10] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nat. Biotechnol.*, vol. 21, no. 6, p. 697, 2003.

[11] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein–protein interaction data," *J. Comput. Biol.*, vol. 10, no. 6, pp. 947–960, 2003.

[12] P. Dutta, S. Saha, S. Pai, and A. Kumar, "A protein interaction information-based generative model for enhancing gene clustering," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[13] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.

[14] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[15] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.

[16] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[17] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[18] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5659–5667.

[19] Q.-T. Ho *et al.*, "Classifying the molecular functions of rab gtpases in membrane trafficking using deep convolutional neural networks," *Anal. Biochem.*, vol. 555, pp. 33–41, 2018.

[20] N.-Q.-K. Le, Q.-T. Ho, and Y.-Y. Ou, "Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins," *J. Comput. Chem.*, vol. 38, no. 23, pp. 2000–2006, 2017.

[21] N. Q. K. Le, E. K. Y. Yapp, N. Nagasundaram, M. C. H. Chua, and H.-Y. Yeh, "Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture," *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 1245–1254, 2019.

[22] R. Cao, C. Freitas, L. Chan, M. Sun, H. Jiang, and Z. Chen, "Prolango: Protein function prediction using neural machine translation based on a recurrent neural network," *Molecules*, vol. 22, no. 10, pp. 1732–1745, 2017.

[23] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGo: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, 2017.

[24] P. Dutta and S. Saha, "Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguistics*, Jul. 2020, pp. 6396–6407.

[25] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[26] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[28] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 3, pp. 775–784, May/Jun. 2011.

[29] D. Piovesan, M. Giollo, E. Leonardi, C. Ferrari, and S. C. Tosatto, "INGA: Protein function prediction combining interaction networks, domain assignments and sequence similarity," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W134–W140, 2015.

[30] P. Radivojac *et al.*, "A large-scale evaluation of computational protein function prediction," *Nat. Methods*, vol. 10, no. 3, pp. 221–227, 2013.

[31] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[32] S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: Generalizing residual architectures," in *Int. Conf. Learn. Representations*, 2019.

[33] D. Szklarczyk *et al.*, "The string database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D561–D568, 2010.

[34] L. J. Jensen *et al.*, "eggNOG: Automated construction and annotation of orthologous groups of genes," *Nucleic Acids Res.*, vol. 36, no. suppl_1, pp. D250–D254, 2007.

[35] M. Alshahrani, M. A. Khan, O. Maddouri, A. R. Kinjo, N. Queralt-Rosinach, and R. Hoehndorf, "Neuro-symbolic representation learning on biological knowledge graphs," *Bioinformatics*, vol. 33, no. 17, pp. 2723–2730, 2017.

[36] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947. [Online]. Available: http://www.jstor.org/stable/2332510