

# Predicting Protein Functions with Ensemble Deep Learning and Protein Language Models

Giacomo Frisoni<sup>a,\*</sup>, Marcello Fuschi<sup>a,1</sup> and Gianluca Moro<sup>a,1</sup>

<sup>a</sup>Department of Computer Science and Engineering (DISI), University of Bologna, Cesena Campus, Via dell'Università 50, I-47522 Cesena, Italy

**Abstract.** Understanding protein functions enables deciphering cellular mechanisms and improving healthcare outcomes, from disease diagnosis to targeted therapy. We present GOMIX, an ensemble learning method for predicting the functions of newly discovered proteins, packaged within an easy-to-use web application. By combining seven complementary base predictors—including sequence homology and protein language models, GOMIX achieves competitive or state-of-the-art performance in the CAFA-3 challenge. Unlike existing solutions, GOMIX is entirely open-source, modular, and computationally low-resource. The code is publicly available at <https://github.com/disi-unibo-nlp/gomix> (MIT License).

## 1 Introduction

Proteins serve as the building blocks and functional drivers of all living organisms. They take on diverse roles, from serving as catalysts in biochemical reactions to providing structural support within cells and tissues. However, exponential growth in protein discovery has outpaced our capacity for experimental verification of their functions.<sup>2</sup> Developing computational techniques for automated protein function prediction (PFP) is crucial to gain insight into biological processes and advance healthcare innovation.

To organize known functions and foster accessibility, bio-curators devised the Gene Ontology (GO) [3], referencing information found in academic papers. GO is structured as three distinct directed acyclic graphs, each representing a different aspect of protein biology: Molecular Function Ontology (MFO), Biological Process Ontology (BPO), and Cellular Component Ontology (CCO). Formally, *the PFP task involves mapping a given protein to valid GO terms.*

In this work, we propose a Stacked Ensembling Learning method for PFP, termed GOMIX. Mechanically, we fuse the predictions of a set of classifiers, from conventional methodologies to protein language embeddings. GOMIX is trained to correlate protein data with GO functions and is evaluated on a popular dataset constructed according to CAFA standard practices [32]. In a CAFA competition, participants aim to predict the functions of a specific set of target proteins, which are then compared against experimental annotations released after the deadline. Comparative analyzes and ablation studies reveal the effectiveness of GOMIX, while elucidating the contribution of each component. We openly release the code and pro-

vide a web application where users can input a protein sequence and obtain confidence-sorted scores for each GO term. This tool facilitates the rapid function determination for previously *unseen* proteins, potentially expediting the research process for scientists. Moreover, GOMIX is inherently flexible and, beyond the classifiers explored in this demo paper, can be readily extended by the community to accommodate new predictive strategies as they emerge.

## 2 Related Work

**Protein function predictors** Early methods relied on sequence similarity, juxtaposing input proteins and known sequences under the assumption that proteins with matching amino acids likely exhibit shared functions [1, 14]. Although efficient, they fail in the cases of convergent evolution or distant homologues. Alternative approaches include analysis of evolutionary relationships and genomic context [25], protein-protein interaction (PPI) nets [7], and structural characteristics [24]. The growing scale of multimodal, biomedical data requires methods for large-scale knowledge extraction, often complicated by decentralized storage [22, 23]. Deep neural networks autonomously extract pertinent features from raw sequence data and discern patterns and latent relationships from vast datasets. They have been applied to cross-organism transfer [10], annotation discovery through random perturbations [9], and literature-based GO term assignment [8]. DeepGO [20] integrates protein sequences and the STRING PPI graph, merging 1D convolutional representations of amino acids with node embeddings. DeepGOPlus [19] extends DeepGO with a sequence similarity search, bypassing the need for PPI data. DeepText2GO [29] computes the consensus between sequence-based and text-based techniques, augmenting dense protein representations with vectorized scientific publications. In recent years, ensemble methods that amalgamate multiple information sources have shown significant promise, exemplified by GO-Labeler [30], NetGO 2 [28], and PANDA2 [31]. Nevertheless, their implementation is closed, and the explored features (e.g., term frequency, amino acid trigrams, and biophysical properties) overlook representations from advanced protein language models (PLMs).

**Protein language embedding** The past two years have witnessed remarkable success in the application of transformer-based language models to protein sequences [11, 16, 27]. The embeddings produced by self-supervised PLMs, such as ESM2 [21] and ESM3 [15], encapsulate protein structure at the 2D and 3D levels, as well as residue-residue contacts, sequence homology (shared evolutionary ancestry),

\* Corresponding Author. Email: [giacomo.frisoni@unibo.it](mailto:giacomo.frisoni@unibo.it).

<sup>1</sup> Equal contribution (co-first authorship).

<sup>2</sup> As of April 2023, the UniProtKB/Swiss-Prot database contains  $\approx 600K$  entries, considerably less than the  $>200M$  protein sequences discovered [5].

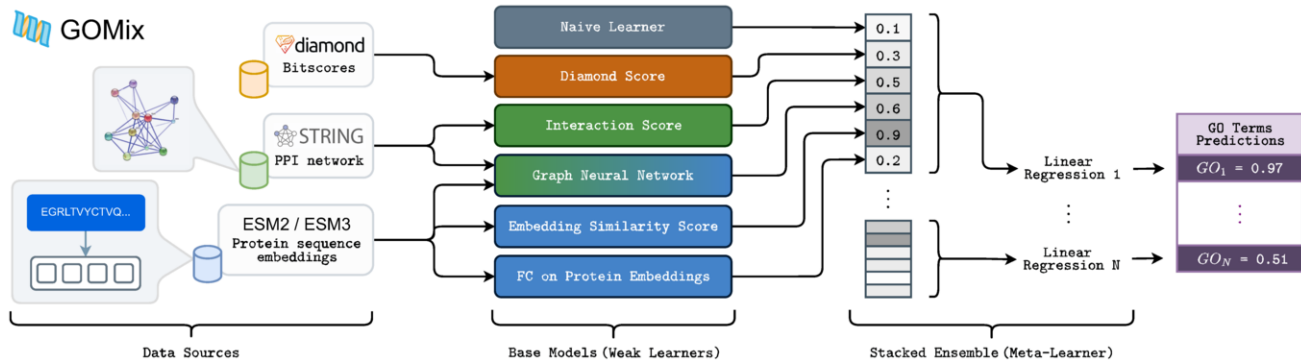


Figure 1: GOMIX architecture.

and physicochemical properties (e.g., polarity, hydrophobicity). Notably, proteins with similar structure or function, yet differing in sequence, tend to exhibit proximity in terms of linear vector arithmetic.

### 3 Method

GOMIX is composed of seven **base learners**, whose predictions are utilized by a **meta-learner** to determine a score for each target GO term within the training data (Figure 1).

#### 3.1 Base Learners

Our framework incorporates established classifiers from previous research [19], namely Naive and Diamond Score, which focus on sequence homology. In addition, we introduce four novel strategies inspired by case-based reasoning principles [4]. Each base learner outputs a  $[0, 1]$  confidence score for each protein-GO term pair.

**(1) Naive** GO terms are assigned based on the frequency of annotations observed in the training data. For each query protein  $p$  and GO term  $f$ , the predicted score is  $N_f/N_{\text{total}}$ , where  $N_f$  is the number of proteins labeled with  $f$ , and  $N_{\text{total}}$  is the number of dataset proteins.

**(2) Diamond Score** The Diamond tool [4] is a high-speed alternative to the BLAST algorithm [2], designed to identify proteins in a database similar to a given query protein  $q$ . For each match, Diamond reports a bitscore, which quantifies sequence similarity while remaining independent of query length and database size. To predict whether  $q$  is associated with a given GO term  $f$ , we normalize the sum of bitscores from all similar sequences identified by Diamond:

$$\frac{\sum_{p \in D} \text{bitscore}(q, p) \cdot I(f \in T_p)}{\sum_{p \in D} \text{bitscore}(q, p)}$$

Here,  $D$  is the set of proteins with a significant similarity to  $q$ , filtered using an e-value threshold of 0.001.  $T_p$  denotes the known GO annotations for protein  $p$ ; the indicator function  $I$  equals 1 if the term  $f$  is among them, 0 otherwise. This component allows PFP to be guided by the aggregate evidence of multiple homologous sequences.

**(3) Interaction Score** This classifier exploits PPI data from the STRING network [26], which unifies known and predicted PPIs from various sources such as experimental data, computational predictions, and public text collections. For a given query protein  $q$ , the prediction score for a GO term  $f$  is calculated based on the annotations of proteins that interact with  $q$ , weighted by their confidence:

$$\frac{\sum_{p \in S} w(q, p) \cdot I(f \in T_p)}{\sum_{p \in S} w(q, p)}$$

$S$  is the set of proteins that interact with  $q$  as identified by STRING.  $w(q, p)$  indicates the probability that the interaction between proteins  $q$  and  $p$  is biologically meaningful, based on supporting evidence.

**(4-5) Embedding Similarity Score** We compute embeddings for query protein sequences using models from the ESM family—ESM2 (15B) and ESM3-open (1.4B). PFP relies on  $k$ -nearest neighbor search, with cosine similarity used to measure proximity between query and training proteins. This approach is motivated by the hypothesis that proteins with similar embeddings are likely to share functional characteristics. Scalable retrieval is enabled by FAISS [18], which partitions the embedding space to reduce time and memory costs. To improve robustness, particularly when  $k$  is large, we apply a reweighting scheme to the cosine values, preventing marginal neighbors from exerting disproportionate influence. For a given GO term  $f$ , the prediction score is defined as:

$$\frac{\sum_{p \in E} \text{cos\_sim}(q, p) \cdot I(f \in T_p)}{\sum_{p \in E} \text{cos\_sim}(q, p)}$$

$E$  is the set of the top  $k = 256$  proteins in the training set with the highest cosine similarity to  $q$ .

**(6) Fully Connected Model on Protein Embeddings** Multilayer perceptron (MLP) network with  $\sim 220\text{M}$  parameters, designed to take ESM-3 embeddings as input and output prediction scores across all GO terms. The architecture consists of seven linear layers, each followed by batch normalization [17] and ReLU as non-linearity.

**(7) Graph Neural Network on STRING PPIs** Graph neural networks (GNNs) provide a differentiable state transition mechanism to condense structured biological networks into low-dimensional node representations. We employ a two-layer GraphSage model [13], comprising  $\sim 795\text{M}$  params, to operate over a graph constructed from STRING. In this discrete representation, each node describes a protein and is initialized with its ESM-3 embedding, while edges encode functional or physical interactions derived from STRING. The output layer applies a sigmoid activation function to produce independent probabilities for each GO term, enabling multi-label classification. The model is trained using binary cross-entropy loss.

#### 3.2 Meta-Learner



The meta-learner consists of multiple independent linear regressors, each dedicated to predicting a specific GO term. Each regressor is implemented as an *ordinary least squares* linear model, tasked with learning an optimal combination of outputs from the base learners for its assigned term. In this way, the meta-learner captures term-specific

**Table 1:** GOMIX dataset details and results averaged across proteins within each ontology.

(a) Dataset statistics.

	Type	Count
Proteins	Training	65,028
	Testing	1,788
GO Terms	MFO	6,232
	BPO	19,760
	CCO	2,410
	Leaf Terms	14,225
	Non-Leaf Terms	14,177

(b) Test set performance comparison. Bold and underline denote the best and second-best scores.

Method	$F_{max}$ ( $\uparrow$ )			$S_{min}$ ( $\downarrow$ )			AUPR ( $\uparrow$ )		
	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
◇ Naive	0.304	0.318	0.605	12.112	38.890	9.646	0.146	0.214	0.542
◇ Diamond Score	0.550	0.440	0.625	8.617	34.027	7.945	0.375	0.257	0.395
◇ Interaction Score	0.406	0.421	0.644	11.214	35.911	8.175	0.332	0.352	0.626
◇ Embedding Similarity Score (ESM-2)	0.579	0.471	0.705	8.698	33.631	7.513	0.549	0.425	<u>0.747</u>
◇ Embedding Similarity Score (ESM-3)	0.599	<u>0.491</u>	0.722	8.056	30.413	6.933	0.598	0.442	<b>0.766</b>
◇ FC on protein embeddings	0.531	0.450	0.690	9.734	34.859	7.933	0.440	0.361	0.700
◇ GNN on PPI & embeddings	0.435	0.361	0.606	11.108	38.407	8.914	0.336	0.233	0.545
UDSMProt	0.582	0.475	0.697	8.787	33.615	7.618	0.548	0.422	0.728
DeepText2GO	<b>0.627</b>	0.441	0.694	<u>5.240</u>	<u>17.713</u>	<b>4.531</b>	<b>0.605</b>	0.336	0.729
GOLabeler	0.580	0.370	0.687	<b>5.077</b>	<b>15.177</b>	<u>5.518</u>	0.546	0.225	0.700
DeepGOPlus	0.585	0.474	0.699	8.824	33.576	7.693	0.536	0.407	0.726
PANDA2	0.598	0.478	0.709	9.670	40.229	9.558	0.564	<u>0.436</u>	0.744
 GOMIX: Naive + *Score	0.601	<b>0.498</b>	<u>0.728</u>	8.217	29.620	6.098	<u>0.604</u>	<b>0.444</b>	<u>0.747</u>
 GOMIX: full	<u>0.604</u>	<b>0.498</b>	<b>0.732</b>	8.251	29.850	6.048	0.601	<u>0.436</u>	0.738

$\uparrow$  The higher, the better     $\downarrow$  The lower, the better    ◇ GOMIX components

weighting patterns, allowing it to modulate the relative contribution of each base learner according to its informativeness for that function. The total number of trainable parameters in the meta-learner amounts to  $N_L \times N_C$ , where  $N_L$  is the count of base learners and  $N_C$  is the number of distinct GO terms in the training annotations.

## 4 Experiments

### 4.1 Dataset and Evaluation Metrics

We evaluated GOMIX on the 2016 dataset provided by DeepGOPlus [19], which is compliant with the time-delayed CAFA-3 criteria [32]. Proteins with experimental annotations preceding Jan 2016 are included in the training set, while those obtained between Jan 2016 and Oct 2016 are allocated to the test set. We propagated annotations considering the True Path Rule [12]: if a GO term is annotated to a protein, its ancestor GO terms are also annotated to that protein. We train base learners in a subset (80%) of the train set, and the meta-learner in the left-out split (20%). Because models output confidence scores rather than binary labels, evaluation proceeds by thresholding scores to obtain discrete per-protein predictions. We sweep thresholds over  $[0, 1)$  with a step of 0.01 and calculate the CAFA metrics by varying them:  $F_{max}$  (maximum F1 attainable),  $S_{min}$  (minimum semantic distance between actual and predicted GO labels), AUPR (area under the precision-recall curve [6]).

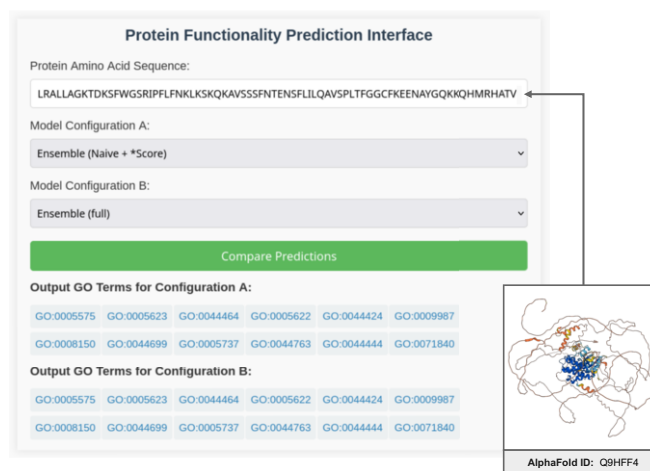
### 4.2 Results

The results obtained after three training epochs are delineated in Table 1, head-to-head compared to previous solutions and individual base components. We find that PLMs contribute significantly to overall performance. The ensemble of base learners (1-5) achieves generalization performance that is remarkably close to the full system. Given that the remaining components are long to train and VRAM demanding, restricting the ensemble to the first five modules offers a substantial gain in efficiency. Inference on the entire test set using this reduced ensemble takes only 389 seconds.

## 5 Demonstration

We provide an interactive web-based system to use GOMIX (Figure 2). Users are guided to input the amino acid sequence in FASTA format and choose the base learners for direct prediction or comparison. Function scores are estimated and presented in a panel, sorted by confidence level, along with a link to GO nodes. Broadly, our

tool empowers users to anticipate the functional characteristics of proteins, including those recently uncovered where no additional information is available apart from their sequence.

**Figure 2:** GOMIX web application.

## 6 Conclusion

In this demonstration paper, we present GOMIX, an open-source tool for protein function prediction. GOMIX follows a lightweight architecture that combines the output of multiple classifiers leveraging heterogeneous knowledge sources. We hope that our work will promote accessibility to biomedical research. In the future, we will integrate 3D structure-aware base learners grounded on AlphaFold-3.

## Acknowledgements

Research partially supported by: AI-PACT (CUP B47H22004450008, B47H22004460001); National Plan PNC-I.1 DARE (PNC0000002, CUP B53C22006450001); PNRR Extended Partnership FAIR (PE00000013, Spoke 8); 2024 Scientific Research and High Technology Program, project “AI analysis for risk assessment of empty lymph nodes in endometrial cancer surgery”, the Fondazione Cassa di Risparmio in Bologna; Chips JU TRISTAN (G.A. 101095947).

## References

- [1] S. Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sept. 1997. doi: 10.1093/nar/25.17.3389. URL <https://doi.org/10.1093/nar/25.17.3389>.
- [2] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
- [3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [4] B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.
- [5] U. Consortium. Uniprotkb/swiss-prot protein knowledgebase release 2023\_04 statistics, 2023. URL <https://web.expasy.org/docs/relnotes/relnote.html>. Accessed: May 10, 2024.
- [6] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.
- [7] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. In *Proceedings. IEEE Computer Society Bioinformatics Conference*, pages 197–206. IEEE, 2002.
- [8] P. di Lena, G. Domeniconi, L. Margara, and G. Moro. GOTA: GO term annotation of biomedical literature. *BMC Bioinform.*, 16:346:1–346:13, 2015. doi: 10.1186/S12859-015-0777-8. URL <https://doi.org/10.1186/s12859-015-0777-8>.
- [9] G. Domeniconi, M. Masseroli, G. Moro, and P. Pinoli. Discovering new gene functionalities from random perturbations of known gene ontological annotations. In A. L. N. Fred and J. Filipe, editors, *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 107–116. SciTePress, 2014. doi: 10.5220/0005087801070116. URL <https://doi.org/10.5220/0005087801070116>.
- [10] G. Domeniconi, M. Masseroli, G. Moro, and P. Pinoli. Cross-organism learning method to discover new gene functionalities. *Comput. Methods Programs Biomed.*, 126:20–34, 2016. doi: 10.1016/J.CMPB.2015.12.002. URL <https://doi.org/10.1016/j.cmpb.2015.12.002>.
- [11] A. Fallahpour, A. Magnuson, P. Gupta, S. Ma, J. Naimer, A. Shah, H. Duan, O. Ibrahim, H. Goodarzi, C. J. Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.
- [12] I. Friedberg and P. Radivojac. Community-wide evaluation of computational function prediction. *The Gene Ontology Handbook*, pages 133–146, 2017.
- [13] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [14] T. Hawkins, S. Luban, and D. Kihara. Enhanced automated function prediction using distantly related sequences and contextual association by pfp. *Protein Science*, 15(6):1550–1556, 2006.
- [15] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL <https://www.science.org/doi/abs/10.1126/science.ads0018>.
- [16] B. Hu, J. Xia, J. Zheng, C. Tan, Y. Huang, Y. Xu, and S. Z. Li. Protein language models and structure prediction: Connection and progression. *CoRR*, abs/2211.16742, 2022. doi: 10.48550/ARXIV.2211.16742. URL <https://doi.org/10.48550/arXiv.2211.16742>.
- [17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [18] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [19] M. Kulmanov and R. Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.
- [20] M. Kulmanov, M. A. Khan, and R. Hoehndorf. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2018.
- [21] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [22] S. Lodi, G. Moro, and C. Sartori. Distributed data clustering in multi-dimensional peer-to-peer networks. In H. T. Shen and A. Bouguet-taya, editors, *Database Technologies 2010, Twenty-First Australasian Database Conference (ADC 2010), Brisbane, Australia, 18-22 January, 2010, Proceedings*, volume 104 of *CRPIT*, pages 171–178. Australian Computer Society, 2010. URL <http://portal.acm.org/citation.cfm?id=1862264&CFID=17470975&CFTOKEN=71845406>.
- [23] G. Moro and G. Monti. W-grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. *J. Netw. Comput. Appl.*, 35(4):1218–1234, 2012. doi: 10.1016/J.JNCA.2011.05.002. URL <https://doi.org/10.1016/j.jnca.2011.05.002>.
- [24] D. Pal and D. Eisenberg. Inference of protein function from protein structure. *Structure*, 13(1):121–130, 2005.
- [25] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.
- [26] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.
- [27] C. Tran, S. Khadkikar, and A. A. Porollo. Survey of protein sequence embedding models. *International Journal of Molecular Sciences*, 24, 2023. URL <https://api.semanticscholar.org/CorpusID:256892679>.
- [28] S. Yao, R. You, S. Wang, Y. Xiong, X. Huang, and S. Zhu. Netgo 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic acids research*, 49(W1):W469–W475, 2021.
- [29] R. You, X. Huang, and S. Zhu. DeepText2Go: improving large-scale protein function prediction with deep semantic text representation. *Methods*, 145:82–90, 2018.
- [30] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka, and S. Zhu. Gola-beler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 2018.
- [31] C. Zhao, T. Liu, and Z. Wang. Panda2: protein function prediction using graph neural networks. *NAR genomics and bioinformatics*, 4(1):lqac004, 2022.
- [32] N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georgioui, H. N. Nguyen, M. N. Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20:1–23, 2019.