DeepGO-SE: Protein function prediction as Approximate Semantic Entailment

Maxat Kulmanov[1,2*], Francisco J. Guzmán-Vega[2,3],
Paula Duek Roggli[4,5], Lydie Lane[4,5], Stefan T. Arold[2,3],
Robert Hoehndorf[1,3*]

[1*]Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology (KAUST), 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900, Makkah, Saudi Arabia.
[2]Bioscience Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900, Makkah, Saudi Arabia.
[3*]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), 4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900, Makkah, Saudi Arabia.
[4]CALIPHO group, SIB Swiss Institute of Bioinformatics, CMU, 1 rue Michel Servet, Geneva 4, 1211, Switzerland.
[5]Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CMU, 1 rue Michel Servet, Geneva 4, 1211, Switzerland.

*Corresponding author(s). E-mail(s): maxat.kulmanov@kaust.edu.sa; robert.hoehndorf@kaust.edu.sa;
Contributing authors: francisco.guzmanvega@kaust.edu.sa; paula.duek@sib.swiss; lydie.lane@sib.swiss; stefan.arold@kaust.edu.sa;

Abstract

The Gene Ontology (GO) is one of the most successful ontologies in the biological domain. GO is a formal theory with over 100,000 axioms that describe the molecular functions, biological processes, and cellular locations of proteins in three sub-ontologies. Many methods have been developed to automatically predict protein functions. However, only few

1

of them use the background knowledge provided in the axioms of GO for knowledge-enhanced machine learning, or adjust and evaluate the model for the differences between the sub-ontologies.

We have developed DeepGO-SE, a novel method which predicts GO functions from protein sequences using a pretrained large language model combined with a neuro-symbolic model that exploits GO axioms and performs protein function prediction as a form of approximate semantic entailment. We specifically evaluate DeepGO-SE on proteins that have no significant similarity with training proteins and demonstrate that DeepGO-SE can improve function prediction for those proteins.

Keywords: Protein Function, Gene Ontology, Ontology Embedding, Machine Learning

1 Introduction

Protein function prediction is one of the key challenges in modern biology and bioinformatics as it enables better understanding of the roles and interactions of proteins within living systems. Accurate functional descriptions of proteins are necessary for tasks such as identification of drug targets, understanding disease mechanisms, and improving biotechnological applications in industry. While predicting protein structures has become increasingly accurate in recent years [1], predicting protein function remains challenging due to the small number of known functions combined with their complexity and interactions.

Functions of proteins are described using the Gene Ontology (GO) [2] which is one of the most successful ontologies in biology. GO includes three sub-ontologies for describing molecular functions (MFO) of a single protein, biological processes (BPO) to which proteins can contribute, and cellular components (CCO) where proteins are active. Researchers identify protein functions based on experiments and generate scientific reports which are then taken by database curators and added to knowledge bases. These annotations are generally propagated to homolog proteins. As a result, the UniProtKB/Swiss-Prot database [3] contains manually curated GO annotations for thousands of organisms and more than 550,000 proteins.

Recent protein function prediction methods rely on different sources of information such as sequence, interactions, protein tertiary structure, literature, coexpression, phylogenetic analysis, or the information provided in GO [4–20]. The methods may use sequence domain annotations [5, 6, 8, 11, 21], directly apply deep convolutional neural networks (CNN) [13] or language models such as LSTMs [9] and transformers [14], or use pretrained protein language models [10, 15] to represent amino acid sequences. Models may also incorporate protein–protein interactions through knowledge graph embeddings [12, 16], approaches using k-nearest neighbors [21], and graph convolutional neural networks [6]. Also, natural language models applied to scientific literature have been successful in automated function prediction [8].

One of the major limitations of many function prediction methods is their reliance on sequence similarity to predict functions. While this approach has been effective when applied to proteins that have similar proteins with well-characterized functions, it can be less reliable for proteins with little or no sequence similarity to known functional domains. Molecular

functions arise largely from structure, and proteins with similar structures might have different sequence [22]. Importantly, proteins with similar sequences can have a different set of functions depending on their active sites and the organisms in which they are a part. Consequently, methods that use the same sources of information for all three sub-ontologies of GO are limited; while functions from the MFO sub-ontology can be predicted by a protein sequence or structure, functions from BPO and, to a lesser degree, CCO, inherently rely on multiple proteins being present and interacting in particular ways; therefore, predicting BPO and CCO annotations requires different sources of information than predicting MFO annotations. In general, predicting whether a protein participates in a biological process requires knowledge of an organisms proteome, or at least its annotated genome so that proteins can be predicted; as a result, two proteins may have 100% sequence identity but participate in different processes, depending on the presence or absence of other proteins within the organism's proteome. Protein–protein interaction networks can encode the proteome as well as limit the search space for potential interactions between proteins that give rise to biological processes.

Ontologies are another source of information rarely exploited for predicting protein functions. Ontologies are not simply collections of classes; rather, ontologies are formal theories that specify some aspects of the intended meaning of a class using a logic-based language [23]. The background knowledge that is contained in the axioms of GO can be used by some machine learning models to improve predictions through knowledge-enhanced machine learning [11, 12, 14, 15]. By incorporating the formal axioms into machine learning models, it becomes possible to leverage prior knowledge during the learning or prediction process, and to put constraints on the parameter search space that can improve the accuracy and efficiency of the learning process, and, ultimately, make better predictions [24, 25]. While there are different approaches of how formal background knowledge can be incorporated in machine learning methods, *approximate entailment* aims to explicitly and provably perform "semantic entailment" as optimization objective, and therefore reproduce many of the formal properties of deductive systems [26]. Only few function prediction methods utilize the formal axioms that are in GO. Hierarchical classification methods for predicting protein functions such as GoStruct2 [27], DeepGO [12] and DeePred [28] use subsumption axioms to extract hierarchical relations between classes but ignore the rest of the axioms stored in GO that can be used to reduce the search space and improve predictions.

We have developed DeepGO-SE, a novel protein function prediction method which predicts functions from protein sequences using a pretrained large protein language model combined with a neuro-symbolic model that performs function prediction as approximate semantic entailment. We use the ESM2 protein language model [29] to generate representations of single proteins. Similarly to DeepGOZero [11], we project the ESM2 embeddings into an embedding space (ELEmbeddings) that is generated from the axioms in the GO [30]. ELEmbeddings encode ontology axioms based on geometric shapes and geometric relations, and corresponds to a Σ algebra, or "world model", in which we can determine whether statements are true or false. In contrast to DeepGOZero, we use these world models to perform "semantic entailment": statement ϕ is entailed by theory T ($T \models \phi$) if and only if ϕ is true in every world model in which all statements in T are true [31]. While there are, in general, infinitely many such world models for a theory T or a statement ϕ, we learn multiple, but finitely many, such models and generate predictions of functions as "approximate" semantic entailment where we test for truth in each of the generated world models. Using this form

3

of approximate semantic entailment, we show that the axioms in the extended version of GO enhance the predictions of molecular functions.

Furthermore, we improve predictions for complex biological processes and cellular components by incorporating information about an organisms proteome and interactome in the form of protein–protein interaction networks. We show that, unlike molecular functions, predictions of annotations to biological processes and cellular components can significantly benefit from protein–protein interactions. For biological processes, we found that integrating predicted molecular functions and interactions considerably improves the performance of the predictions; this finding indicates that prediction of biological process annotations does not require knowledge of specific proteins but only their molecular functions, thereby substantially expanding the generality of our method.

We train and evaluate our model on a dataset with experimental annotations which is split based on sequence similarity to make sure that the evaluations are reported using a test set that does not share similar protein with the training set. We find that methods which rely on sequence similarity perform poorly in this setting, whereas DeepGO-SE significantly improves the prediction performance for all sub-ontologies of GO.

Overall, the contributions of our work are as follows:

- We developed a novel method for knowledge-enhanced machine learning as approximate semantic entailment over multiple generated world models.
- We developed a novel method for predicting protein functions which improves prediction performance of sub-ontologies of GO by using knowledge-enhanced learning and a combination of different sources of information.
- We improve function prediction performance for novel proteins by using sequence features generated by a pretrained protein language model ESM2.

2 Results

2.1 DeepGO Semantic Entailment

In the DeepGO-SE model, we use ESM2 [29] to represent protein sequence and project them into multiple geometric interpretations (i.e., models) of GO that have been generated with ELEmbeddings [30]; we then test the degree of truth of statements assigning a function to a protein in each interpretation of GO, and aggregate over all interpretations. The EMS2 embeddings of proteins are used as input to an MLP model that projects the embedding into ELEmbeddings space by matching the dimensionality of the EMS2 embedding with the dimension of the ELEmbedding space:

$$f_\eta(p) = MLPBlock(esm2(p)) \tag{1}$$

Within a model, given a protein p and GO class c, we test for the truth of the statement $p \sqsubseteq \exists hasFunction.c$ using the following formula:

$$y'_c = \sigma(f_\eta(p) \cdot (f_\eta(hF) + f_\eta(c))^T + r_\eta(c)) \tag{2}$$
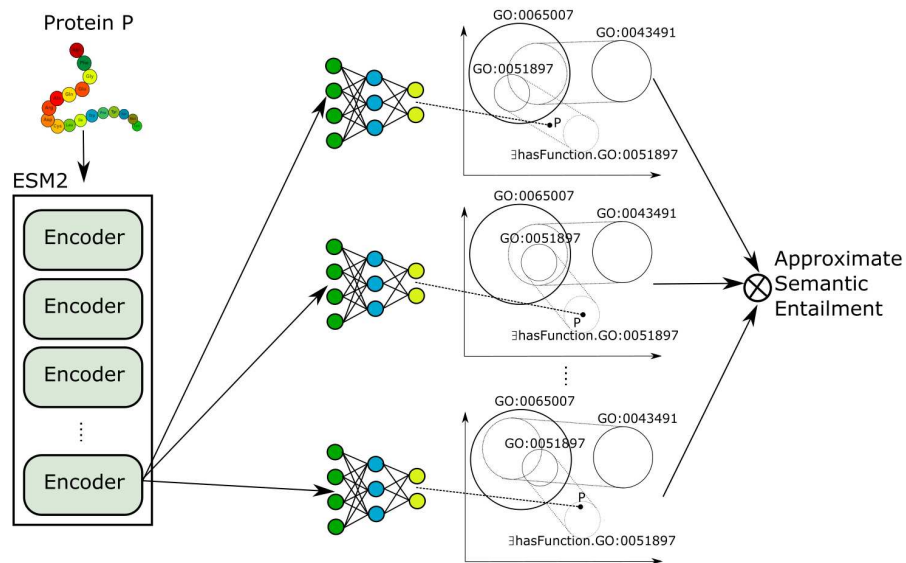
4

Fig. 1: The figure provides a high-level overview of the DeepGO-SE model. On the left, a protein P is embedded in a vector space using ESM2 model whereas the right side shows multiple models with an MLP that embeds the protein in the same space as the GO axioms. Furthermore, predictions from multiple models are combined to perform approximate semantic entailment.

where $f_\eta(p)$ is the projection function, $f_\eta(hF)$ is the embedding of the *hasFunction* relation, $f_\eta(c)$ is the center embedding of an n-ball representing class c, $r_\eta(c)$ is the radius of the n-ball representing class c, and σ is a sigmoid activation function.

The statement is approximately entailed if it is true in all interpretations generated by DeepGO-SE. We independently generate several ELEmbedding and projection functions $f_\eta(p)$, and combine the truth values for the tested axiom (y'_c) in each of the interpretations to obtain the final prediction scores (the degree of entailment). Given N interpretations, we combine truth values by computing the minimum, maximum, or arithmetic mean of the truth values in all N interpretations. Figure 1 provides an overview of the prediction model of DeepGO-SE.

2.2 Evaluation of DeepGO-SE against the UniProtKB/Swiss-Prot dataset

We evaluate and compare our method with the baseline methods using the UniProtKB/Swiss-Prot dataset split by sequence similarity. We use protein-centric evaluation measures such as Fmax, Smin, and AUPR, and class-centric AUC standardized by the CAFA challenge [32]. We provide detailed information about evaluation measures in the Supplementary materials.

We train and evaluate three sub-ontologies of GO seperately because they have different characteristics in terms of number of classes and their relations, number of proteins and sources of information they can benefit from. We compare with five baseline methods: Naïve, MLP, DeepGOCNN, DeepGOZero and DeepGraphGO. None of these methods relies on sequence similarity and, except for the naïve predictor, all assign functions based on sequence

5

features that are learned directly or using features derived from tools such as InterProScan [33].

In all evaluations, the DeepGO-SE model significantly outperformed all the baseline methods in terms of $F\mathrm{max}$, AUPR and AUC. In MFO, DeepGO-SE achieved an $F\mathrm{max}$ of 0.554 which is 7% larger than the result achieved by the MLP and DeepGOZero methods (Table 1). In predicting BPO annotations, the model achieves an $F\mathrm{max}$ of 0.432 which is around 8% higher than the best baseline method DeepGraphGO (Table 2), and in the CCO evaluation, DeepGO-SE model achieves $F\mathrm{max}$ of 0.721 (Table 3).

In our basic DeepGO-SE model, protein embeddings are generated from protein sequence by ESM2; however, we can modify $f_\eta(p)$ to encode more information about a protein. We argue that biological process and cellular component annotations cannot be predicted from a protein sequence alone because even sequence-identical proteins can legitimately be involved in different processes dependent on the presence or absence of other proteins. Therefore, we use $f_\eta(p)$ to also encode information about a proteome and its interactions (protein–protein interactions, PPIs). To combine PPIs with individual features of proteins we use Graph Attention Networks (GAT) [34] and embed the protein p in the ELEmbeddings space using the formula

$$f_\eta(p) = GATConv(MLPBlock(x), g) \tag{3}$$

where x is an input feature vector for p, g is the PPI graph, $MLPBLock$ is described in equation 6, $GATConv$ is a GAT layer. We use this embedding function and alter the input vector x to perform three experiments. First, in DeepGOGAT-SE, we use the ESM2 embeddings as input for each protein. Second, in DeepGOGATMESE, x consists of the experimental annotations of a protein p to its molecular functions using a binary vector of size 6851. Third, in DeepGOGATMESE-Pred, we use the prediction scores from the DeepGO-SE model for molecular functions as input. We train and evaluate these three models to determine the effect of incorporating interactions.

Combining PPIs and ESM2 embeddings in the DeepGOGAT-SE model reduced the MFO prediction performance to an $F\mathrm{max}$ of 0.525, but slightly improved $S\mathrm{min}$. Incorporating PPIs improves the performance in BPO predictions to $F\mathrm{max}$ of 0.435. The overall best performance in BPO is achieved when using experimental MFO annotations as features (DeepGOGATMF-SE), followed by MFO annotations predicted by DeepGO-SE (DeepGOGATMF-SE-Pred) (Table 2). For CCO, incorporating PPIs in the DeepGO-SE model increases $F\mathrm{max}$ from 0.721 to 0.736 (DeepGOGAT-SE) (Table 3).

Interestingly, including PPIs in our model did not improve MFO predictions (except for a slight improvement in $S\mathrm{min}$), demonstrating that molecular functions can be predicted from single proteins whereas information about multiple proteins needs to be used to predict BPO and CCO annotations.

2.3 Evaluation of DeepGO-SE against the neXtProt manual prediction dataset

In order to further evaluate the performance of our method and baseline methods, we used a dataset of manually predicted protein functions from neXtProt. neXtProt assigns functions to uncharacterized proteins based on expert curation of available evidence. We found that, for molecular functions, the best $F\mathrm{max}$ of 0.386 is achieved by our the DeepGO-SE method and

Table 1: Prediction results for Molecular Function on the UniProtKB/Swiss-Prot dataset

Method	Fmax	Smin	AUPR	AUC
Naive	0.321	14.568	0.180	0.500
MLP	0.484	12.650	0.441	0.673
DeepGOCNN	0.404	13.741	0.365	0.749
DeepGOZero	0.483	12.722	0.444	0.749
DeepGraphGO	0.416	14.077	0.357	0.673
DeepGO-SE	**0.554**	11.681	**0.552**	**0.874**
DeepGOGAT-SE	0.525	**11.137**	0.523	0.861

This table shows protein-centric F_{max}, S_{min}, and AUPR, and the class-centric average AUC.

Table 2: Prediction results for Biological Process on the UniProtKB/Swiss-Prot dataset

Method	Fmax	Smin	AUPR	AUC
Naive	0.294	43.934	0.195	0.500
MLP	0.350	42.059	0.303	0.668
DeepGOCNN	0.334	42.912	0.275	0.686
DeepGOZero	0.343	42.857	0.284	0.643
DeepGraphGO	0.354	42.100	0.303	0.736
DeepGO-SE	0.432	39.419	0.401	0.864
DeepGOGAT-SE	0.435	39.123	0.404	**0.876**
DeepGOGATMF-SE	**0.448**	**37.299**	**0.428**	0.831
DeepGOGATMF-SE-Pred	0.444	39.098	0.409	0.855

This table shows protein-centric F_{max}, S_{min}, and AUPR, and the class-centric average AUC.

the second best Fmax 0.375 is achieved by DeepGOGAT-SE. Surprisingly, similar performance is achieved by the Naïve method which only uses the term frequency. However, when we evaluate based on AUPR and term centric AUC, we find that DeepGO-SE performs significantly better. The discrepancy can be explained by the small number of annotations. In this dataset, the median number of annotations is one, meaning that most proteins have only one specific GO function prediction (Table 4).

For biological processes, our method DeepGOGAT-SE which combines PPIs into the model performs with the best Fmax of 0.350. DeepGO-SE achieves slightly lower Fmax of 0.349 and slightly better Smin, however DeepGOGAT-SE is substantially better in terms of AUPR and AUC. The third best Fmax and the best AUC is achieved by our method which uses predicted molecular functions to predict biological processes. We were not able to evaluate the DeepGOGATMF-SE method because many of the proteins are missing manual molecular function (Table 5).

Table 3: Prediction results for Cellular Component on the UniProtKB/Swiss-Prot dataset

Method	$F\max$	$S\min$	AUPR	AUC
Naive	0.620	11.879	0.490	0.500
MLP	0.623	11.584	0.622	0.636
DeepGOCNN	0.661	11.079	0.670	0.758
DeepGOZero	0.625	11.700	0.587	0.599
DeepGraphGO	0.667	10.020	0.666	0.814
DeepGO-SE	0.721	9.499	0.730	0.914
DeepGOGAT-SE	**0.736**	**8.634**	0.743	**0.930**
DeepGOGATMF-SE	0.668	9.809	0.679	0.884
DeepGOGATMF-SE-Pred	0.694	9.907	**0.753**	0.884

This table shows protein-centric $F_{\max}$, $S_{\min}$, and AUPR, and the class-centric average AUC.

Table 4: Prediction results for Molecular Function on the neXtProt dataset

Method	$F\max$	$S\min$	AUPR	AUC
Naive	0.360	10.340	0.165	0.500
MLP	0.370	10.555	0.298	0.516
DeepGOCNN	0.348	10.641	0.270	0.599
DeepGOZero	0.337	10.662	0.261	0.573
DeepGraphGO	0.330	10.573	0.270	0.558
DeepGO-SE	**0.386**	**10.093**	**0.324**	**0.744**
DeepGOGAT-SE	0.375	10.254	0.291	0.700

This table shows protein-centric $F_{\max}$, $S_{\min}$, and AUPR, and the class-centric average AUC.

2.4 Case study: Validation based on structural homologs

We further investigated some predictions of molecular functions for which DeepGO-SE and neXtProt were in agreement to test if we could find additional evidence for the predictions. Specifically, we investigated the Mab-21-like protein 4 (MAB21L4) protein which has a single MFO annotation *nucleotidyltransferase activity* (GO:0016779), which is both assigned by neXtProt and DeepGO-SE. DeepGO-SE predicts this annotation with a high score of 0.638. MAB21L4 was predicted by neXtProt to be a nucleotidyltransferase based on available information about the protein's activity in epidermal keratinocytes [35]. As part of investigating the role of MAB21L4 in keratinocytes, distant homology detection was used to assign MAB21L4 to the nucleotidyltransferase (NTase) fold superfamily [36]. The active site is described by the motifs hG[GS], [DE]h[DE]h and h[DE]h (h indicates a hydrophobic amino acid), where the three conserved aspartate/glutamate are involved in coordination of divalent ions and activation of acceptor hydroxyl group of the substrate, and the hG[GS] pattern is involved in holding the substrates within the active site [36]. Sequence alignment combined with structural data provided by AlphaFold suggests that the [DE]h[DE]h motif is conserved in MAB21L4 (Asp80-Met81-Glu-82-Val83) while the h[DE]h motif that

Table 5: Prediction results for Biological Process on the neXtProt dataset

Method	Fmax	Smin	AUPR	AUC
Naïve	0.308	32.987	0.183	0.500
MLP	0.311	32.140	0.251	0.580
DeepGOCNN	0.286	32.152	0.235	0.571
DeepGOZero	0.329	31.999	0.263	0.553
DeepGraphGO	0.322	31.861	0.240	0.558
DeepGO-SE	0.349	**30.170**	**0.312**	0.683
DeepGOGAT-SE	**0.350**	30.218	**0.312**	0.666
DeepGOGATMF-SE-Pred	0.339	30.653	0.293	**0.694**

This table shows protein-centric $F_{\max}$, $S_{\min}$, and AUPR, and the class-centric average AUC.

aligns with other family members may not be conserved as it is replaced by an histidine (Phe199-His200-Val201). An alternative h[DE]h motif is present in Val-236, Asp-237, Leu-238 with the two first residues in a loop and the third one at the beginning of a short β-strand. The hG[GS] motif is less conserved among the nucleotidyltransferase superfamily members and seems not conserved among the members of the Mab-21 group but is present in Mab-21-like protein 1 (MAB21L1); sequence-based methods like InterProScan identify a *Mab-21-like, nucleotidyltransferase domain* (`IPR046903`) in MAB21L1. We used foldseek [37] to compare MAB21L1 and MAB21L4 structurally, and find that both are structurally very similar despite a low sequence similarity. Furthermore, MAB21L4 is structurally very similar to Cyclic GMP-AMP synthase (CGAS), which is well-characterized as having the *nucleotidyltransferase activity*.

Another noticeable example is the Family With Sequence Similarity 151 Member B (FAM151B) protein which was predicted to be a phosphoric diester hydrolase (`GO:0008081`) based on structural similarity to a protein from *Sicarius terrosus* by the neXtProt database. DeepGO-SE predicted the same function with a high score of 0.846. Foldseek search resulted in many sequence and structure homologs. Structure homologs with high sequence identity were not annotated, however, we found several well annotated structural homologs with low sequence identity. For example, the human protein Lysophospholipase D (GDPD3) has a high structural similarity to FAM151B and has been annotated with *phosphoric diester hydrolase activity* (`GO:0008081`) based on experimental evidence (Supplementary Figure B1). In addition, DeepGO-SE predicts other functions such as *metal ion binding* (`GO:0046872`) which GDPD3 has been annotated with as well. These findings suggest that DeepGO-SE learned to predict functions, among others, based on structural information.

2.5 Ablation study

In order to evaluate the contribution of the individual components of our models, we performed an ablation study. First, for each of the models, we removed the ElEmbeddings axiom loss functions and only optimized function prediction loss to determine the effect of using background knowledge contained in the GO. In the DeepGO-SE model, removing axioms losses resulted in a performance drop in the MFO evaluation while the performance in the

9

Table 6: Summary of the UniProtKB/Swiss-Prot dataset

Ontology	GO Terms	Proteins	Groups	Training	Validation	Testing
MFO	6,851	43,279	6,963	52,072	2,964	4,221
BPO	21,356	58,729	9,463	52,584	2,870	3,275
CCO	2,829	59,257	10,019	48,318	4,970	5,969

The table shows the number of GO terms, total number of proteins, number of groups of similar proteins, number of proteins in training, validation and testing sets for the UniProtKB/Swiss-Prot dataset.

BPO and CCO evaluations was not affected. Second, we trained the models with only GO or only GO-PLUS axioms to further evaluate the effect of using more background knowledge for performing approximate semantic entailment. We found that the performance of the MFO model improves with GO-PLUS axioms compared to GO axioms whereas the performance of the BPO and CCO models slightly drop when using the additional axioms contained in GO-PLUS.

Using PPI information, in the DeepGOGAT-SE model, removing axioms and removing the semantic entailment module resulted in a slight performance increase in MFO evaluation but the performance dropped in the BPO and CCO evaluations. In models that use PPIs and molecular functions as protein features, performance is better for BPO and CCO when removing axioms and semantic entailment.

Overall, the ablation study shows that the ontology axioms and semantic entailment mostly contribute to MFO and CCO model performance whereas the performance of BPO model is not significantly affected. The PPIs with GAT noticeably contribute to CCO and BPO model performance and BPO model achieves the best performance without axioms and semantic entailment. Supplementary Table B5 provides the results of the ablation study for all four evaluation measures.

3 Methods

3.1 UniProtKB/Swiss-Prot Dataset

We use a dataset that was generated from manually curated and reviewed dataset of proteins from the UniProtKB/Swiss-Prot Knowledgebase [3] version 2021_04 released on 29-Sep-2021. We filtered all proteins with experimental functional annotations with evidence codes EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, HTP, HDA, HMP, HGI, HEP. The dataset contains $77,647$ reviewed and manually annotated proteins. For this dataset we use Gene Ontology (GO) released on 2021-11-16. We train and evaluate models for each of the sub-ontologies of GO separately.

We mainly aim to predict functions of novel proteins that have a low sequence similarity to existing proteins in the dataset. Therefore, we decided to split our dataset based on any similarity hit with maximum e-value score of 0.001. We computed pairwise similarity using Diamond (v2.0.9) [38] and grouped the sequences that have some similarity and split these groups into training, validation and testing sets. Table 6 summarizes the datasets for each sub-ontology. We train and evaluate a separate model for each sub-ontology.

3.2 neXtProt Dataset

In order to further evaluate the performance of our models we use a dataset of manually annotated predictions of uncharacterized human proteins from the neXtProt [39] database. neXtProt standardizes and integrates information on human proteins and provides users with an advanced search capability built around semantic technologies [39]. neXtProt contains free text summaries of the literature and standardized enzyme annotations from UniProtKB/Swiss-Prot, pathway annotations from KEGG [40] and Reactome [41], and GO MFO and BPO terms from a variety of resources, obtained either manually or by automatic procedures, and based on either experiments or computational analysis. Proteins lacking the above-mentioned annotations and those that are solely annotated with broad GO terms are considered as uncharacterized. They can be retrieved using the SPARQL [42] query `NXQ_00022` [43]. In the 2023-04-18 release of neXtProt, there are 1,521 such proteins. To stimulate the characterization of these poorly-studied proteins, neXtProt collects and reviews functional predictions from the literature and proposes their own function annotations based on a manual interpretation of different types of public data (phenotypes, expression, subcellular localization, protein and genetic interactions, phylogeny, structure, sequence and functional assays) [44]. These predictions are displayed in the function prediction pages as GO MFO or BPO terms, and the underlying evidence using the Evidence Code Ontology (ECO) [45].

Here we use the data retrieved from 113 publications together with different resources that were used to predict the functions of 239 uncharacterized human proteins. In total, the proteins collected 659 specific GO function annotations where 69 molecular functions were assigned to 53 proteins and 590 biological processes were assigned to 225 proteins. Roughly one third of the proteins (38%) are assigned to only one function that in most of the cases (85%) is a GO BPO term. Most of the functional predictions (78%) are based on one piece of evidence.

3.3 Protein Language Model ESM2

Protein language models are large transformer architectures trained on protein sequences. The Evolutionary Scale Model (ESM) [29, 46] has been trained on 250 million sequences and learned protein sequence representations that are predictive for biochemical and biological properties of proteins including their functions. The second version of ESM has been improved to learn better representations that are also predictive of tertiary structures of proteins. We use the pretrained model of ESM2 with 3 billion parameters (esm2_t36_3B_UR50D) to generate representations of proteins in our dataset. For a protein, we compute the output of the last layer and take the mean of embeddings for each amino acid, resulting in an embedding of size of 2560 for each protein.

3.4 GO-PLUS

The standard version of GO does not include relations between GO classes and external ontologies such as ChEBI [47], Uberon [48], the Cell Ontology [49], or to structured vocabularies such as the NCBI Taxonomy [50]. These relations and cross-ontology axioms exist in an extended version called GO-PLUS [51]. For example, in GO-PLUS a class *Atrioventricular bundle cell differentiation* (`GO:0003167`) is defined as equivalent to *Cell differentiation*

(GO:0030154) and *results in acquisition of features of* (RO:0002315) some *Atrioventricular bundle cell* (CL:0010005). We use the GO-PLUS ontology version released on 2021-11-16 which has over 260K axioms. Like GO, GO-PLUS uses the Web Ontology Language (OWL) 2 [52] to represent its axioms. The Description Logic fragment of OWL 2, OWL 2 DL, defines several profiles, i.e., restricted languages with specific computational properties. GO is formalized using the OWL EL profile [53]. However, GO-PLUS contains axioms that are not part of the OWL EL profile; therefore, it cannot directly be used with reasoning or machine learning methods that are based on OWL EL. We identify around $1,500$ axioms that do not fit in the OWL EL profile and filtered them out using the EL Vira tool [54].

3.5 Approximate Semantic Entailment

Suppose $\mathcal{O}$ is an ontology composed of a set of class symbols $\mathbf{C}$, relation symbols $\mathbf{R}$, and individual symbols $\mathbf{I}$, and that it is expressed in the Description Logic $\mathcal{ALC}$ [55]. In this logic, each class symbol is considered a class description. If C and D are class descriptions and R is a relation symbol, then the expressions $C \sqcap D$, $C \sqcup D$, $\neg C$, $\forall R.C$, and $\exists R.C$ are also considered as class descriptions.

In the $\mathcal{ALC}$ Description Logic, axioms can be classified as TBox or ABox axioms. If C and D are class descriptions, a and b are individual symbols, and r is a relation symbol, a TBox axiom has the form $C \sqsubseteq D$, while an ABox axiom has the form $C(a)$ or $r(a, b)$. A TBox is a set of TBox axioms, and an ABox is a set of ABox axioms. An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ in $\mathcal{ALC}$ comprises a non-empty domain $\Delta^{\mathcal{I}}$ and an interpretation function $\cdot^{\mathcal{I}}$ that satisfies $C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ for all $C \in \mathbf{C}$, $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ for all $R \in \mathbf{R}$, and $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ for all $a \in \mathbf{I}$. The interpretation function is extended to concept descriptions as follows:

$$
\begin{aligned}
&(C \sqcap D)^{\mathcal{I}} := C^{\mathcal{I}} \cap D^{\mathcal{I}}, \ (C \sqcup D)^{\mathcal{I}} := C^{\mathcal{I}} \cup D^{\mathcal{I}}, \\
&(\forall R.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} | \forall e \in \Delta^{\mathcal{I}} : (d, e) \in R^{\mathcal{I}} \text{ implies } e \in C^{\mathcal{I}}\}, \\
&(\exists R.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} | \exists e \in \Delta^{\mathcal{I}} : (d, e) \in R^{\mathcal{I}} \text{ and } e \in C^{\mathcal{I}}\}, \\
&(\neg C)^{\mathcal{I}} := \Delta^{\mathcal{I}} - C^{\mathcal{I}}.
\end{aligned}
\tag{4}
$$

An interpretation $\mathcal{I}$ is called a model of a TBox if, for all $C \sqsubseteq D$ in the TBox, $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$; and a model of an ABox if, for all $R(a, b)$, $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$ and for all $C(a)$, $a^{\mathcal{I}} \in C^{\mathcal{I}}$.

A statement ϕ is semantically entailed by ontology $\mathcal{O}$ (consisting of TBox and ABox), denoted $O \models \phi$, if and only if every model of $\mathcal{O}$ (i.e., an interpretation $\mathcal{I}$ that is a model of both ABox and TBox of $\mathcal{O}$) is also a model of ϕ ($Mod(\mathcal{O}) \subseteq Mod(\phi)$). Semantic entailment requires access to all models of $\mathcal{O}$ and ϕ which are usually infinite; approximate semantic entailment considers only a strict (usually finite) subset of $Mod(\mathcal{O})$ and tests whether ϕ is true in each of them [26, 56].

Here, we perform approximate semantic entailment by learning several models and determining whether a prediction (i.e., a statement that assigns a function to a protein) is true in all of them. For each sub-ontology of GO we train up to 10 models and combine the prediction scores using three different strategies. First, we take the maximum of the selected scores which means that if the predictions is made, it is true in all models. Second, we take an average of the scores. Here, the prediction is made if the prediction threshold is lower than average of all models. Lastly, we take the minimum of the scores where we make sure that the

prediction is true in at least one of the selected models. We select the best parameters of the approximate semantic entailment based on our validation set and use the same on our test set. Supplementary Table B1 summarizes the results of semantic entailment on our validation set.

3.6 Protein–protein interaction (PPI) networks

Molecular functions of proteins mainly depend on their sequences and structures. However, biological processes result from interactions between multiple proteins. Therefore, to accurately predict biological processes, it is necessary to include multiple proteins and their interactions.

For our experiments, we use functional interactions between proteins provided by the STRING database (v11.0). We filter out all the interactions with confidence score less than 700. Our dataset uses UniProtKB identifiers and we map them to STRING database identifiers with mappings provided by UniProtKB. We generate the protein interaction graph using all the proteins in our dataset and use the DGL [57] library to process it and train graph neural networks.

3.7 Baseline methods

For our evaluations we selected methods that do not rely on predictions based on sequence similarity because our aim is to test the predictors on novel sequences. Therefore, we do not include methods as baselines that are primarily based on sequence similarity, such as predictions using BLAST or Diamond, or any other predictors that use their combinations.

3.7.1 Naive approach

Due to the imbalance in GO class annotations and propagation based on the true-path-rule, some classes have more annotations than others. Therefore, it is possible to obtain prediction results just by assigning the same GO classes to all proteins based on annotation frequencies. In order to test the performance obtained based on annotation frequencies, CAFA introduced a baseline approach called "naive" classifier [32]. Here, each query protein p is annotated with the GO classes with a prediction scores computed as:

$$S(p, f) = \frac{N_f}{N_{total}} \tag{5}$$

where f is a GO class, N_f is a number of training proteins annotated by GO class f and N_{total} is a total number of training proteins. We implement the same method.

3.7.2 MLP

The MLP baseline method predicts protein functions using a multi-layer perceptron (MLP) from a protein's InterPro domain annotations obtained with InterProScan [33]. We represent a protein with a binary vector for all the InterPro domains and pass it to two layers of MLP blocks where the output of the second MLP block has residual connection to the first block. This representation is passed to the final classification layer with sigmoid activation function.

13

One MLP block performs the following operations:

$$MLPBlock(\mathbf{x}) = DropOut(BatchNorm(ReLU(W\mathbf{x} + b))) \tag{6}$$

The input vector $\mathbf{x}$ of length $26,406$ represents InterPro domain annotations and is reduced to $1,024$ by the first MLPBLock:

$$\mathbf{h} = MLPBlock(\mathbf{x}) \tag{7}$$

This representation is passed to the second MLPBlock with the input and output size of $1,024$ and added to itself using residual connection:

$$\mathbf{h} = \mathbf{h} + MLPBlock(\mathbf{h}) \tag{8}$$

Finally, we pass this vector to a classification layer with a sigmoid activation function. The output size of this layer is the same as the number of classes in each sub-ontology:

$$\mathbf{y} = \sigma(W\mathbf{h} + b) \tag{9}$$

We train a different model for each sub-ontology in GO.

3.7.3 DeepGO-PLUS and DeepGOCNN

DeepGO-PLUS [13] predicts function annotations of proteins by combining DeepGOCNN, which predicts functions from the amino acid sequence of a protein using a 1-dimensional convolutional neural network (CNN), with the DiamondScore method. DeepGOCNN captures sequence motifs that are related to GO functions. Here, we only use CNN based predictions.

3.7.4 DeepGOZero

DeepGOZero [11] combines protein function prediction with a model-theoretic approach for embedding ontologies into a distributed space, ELEmbeddings [30]. ELEmbeddings represent classes as n-balls and relations as vectors to embed ontology semantics into a geometric model. It uses InterPro domain annotations represented as binary vector as input and applies two layers of MLPBlock as in our MLP baseline method to generate an embedding of size 1024 for a protein. It learns the embedding space for GO classes using ELEmbeddings loss functions and optimizes together with protein function prediction loss.

3.7.5 DeepGraphGO

The DeepGraphGO [6] method uses a neural network to combine sequence features (Inter-PRO domain annotations) with PPI networks by using graph convolutional neural networks. We have implemented DeepGraphGO based on the manuscript and provide the source code for our implementation. We trained and evaluated the model using our dataset.

14

4 Discussion

DeepGO-SE is a protein function prediction method that improves prediction performance for proteins by incorporating both protein sequence features generated by a pretrained protein language model, background knowledge from the GO, and interactions between proteins. Our results allow us to draw three main conclusions: knowledge-enhanced machine learning methods are now able to improve over methods that do not rely on background knowledge; GO function prediction is best formulated using a separate, hierarchical prediction approach; and function prediction models based on ESM2 can now generalize to largely unseen proteins.

Although DeepGO-SE can predict biological processes and cellular components using only a protein sequence, the best performance is achieved when the sequence is combined with PPIs. However, many novel proteins do not have known interactions which limits the application of the combined model on them. Therefore, there is a need for methods which can accurately predict PPIs for novel proteins based on only available sequence. In future, we plan to incorporate sequence ans structure based PPI predictors into our model.

In addition, DeepGO-SE is able to perform zero-shot predictions, similar to DeepGOZero, and is faster to obtain predictions than other methods that rely on similarity search or multiple sequence alignments. This is due to the fact that DeepGO-SE relies only on ESM2 embeddings, which are faster to compute. Overall, the DeepGO-SE model represents a significant improvement over existing protein function prediction methods, providing a more accurate, comprehensive, and efficient approach.

Declarations

Competing interests. Not applicable

Ethics approval. Not applicable

Consent to participate. Not applicable

Consent for publication. Not applicable

Availability of data and materials. All data underlying this work is freely available at https://github.com/bio-ontology-research-group/deepgo2.

Code availability. The source code of this work is freely available at https://github.com/bio-ontology-research-group/deepgo2.

Appendix A   Evaluation metrics

We use four different measures to evaluate the performance of our models. Three protein-centric measures $F_{\max}, S_{\min}$ and AUPR and one class-centric AUC.

$F_{\max}$ is a maximum protein-centric F-measure computed over all prediction thresholds. First, we compute average precision and recall using the following formulas:

$$pr_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in P_i(t))} \tag{A1}$$

$$rc_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in T_i)} \tag{A2}$$

$$AvgPr(t) = \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} pr_i(t) \tag{A3}$$

$$AvgRc(t) = \frac{1}{n} \cdot \sum_{i=1}^{n} rc_i(t) \tag{A4}$$

where f is a GO class, T_i is a set of true annotations, $P_i(t)$ is a set of predicted annotations for a protein i and threshold t, $m(t)$ is a number of proteins for which we predict at least one class, n is a total number of proteins and I is an indicator function which returns 1 if the condition is true and 0 otherwise. Then, we compute the $F_{\max}$ for prediction thresholds $t \in [0, 1]$ with a step size of 0.01. We count a class as a prediction if its prediction score is greater or equal than t:

$$F_{\max} = \max_t \left\{ \frac{2 \cdot AvgPr(t) \cdot AvgRc(t)}{AvgPr(t) + AvgRc(t)} \right\} \tag{A5}$$

$S_{\min}$ computes the semantic distance between real and predicted annotations based on information content of the classes. The information content $IC(c)$ is computed based on the annotation probability of the class c:

$$IC(c) = -log(Pr(c|P(c))) \tag{A6}$$

where $P(c)$ is a set of parent classes of the class c. The $S_{\min}$ is computed using the following formulas:

$$S_{\min} = \min_t \sqrt{ru(t)^2 + mi(t)^2} \tag{A7}$$

16

Fig. B1: The figure provides structural comparison between human proteins *FAM151B* and *Lysophospholipase D GDPD3*. Foldseek search score is 1.0, RMSD is 9.77 and TM-Score is 0.42952.
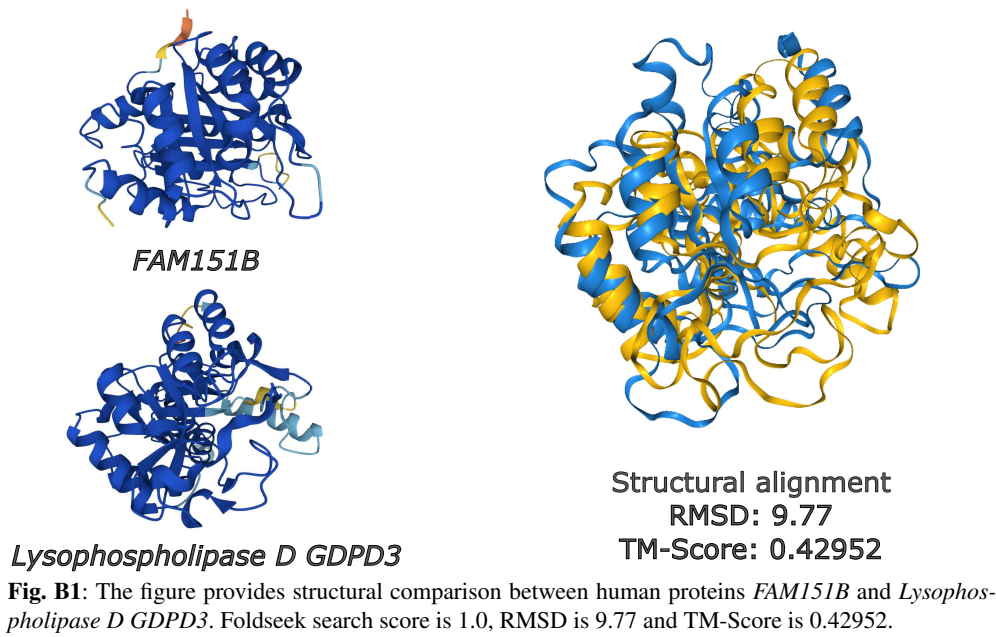
where $ru(t)$ is the average remaining uncertainty and $mi(t)$ is average misinformation:

$$ru(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{c \in T_i - P_i(t)} IC(c) \tag{A8}$$

$$mi(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{c \in P_i(t) - T_i} IC(c) \tag{A9}$$

AUPR is the area under the average precision ($AvgPr$) and recall ($AvgRc$) curve.

AUC is a class-centric measure where compute AUC ROC per each class and take the average.

Appendix B   Supplementary figures

Appendix C   Supplementary tables

References

[1] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu,

Axioms	Strategy	Models	$F_{\max}$			$S_{\min}$			AUPR			AUC		
			MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
GOPlus	min	1	0.526	0.448	0.700	11.071	38.842	9.716	0.516	0.417	0.707	0.864	0.870	0.912
		2	0.530	0.447	0.701	11.029	38.699	9.685	0.522	0.419	0.712	0.867	0.872	0.914
		3	0.530	0.446	0.700	11.010	38.666	9.709	0.523	0.420	0.715	0.868	0.872	0.915
		4	0.529	0.446	0.701	10.997	38.724	9.695	**0.524**	0.419	0.717	0.869	0.872	0.916
		5	0.529	0.446	0.701	10.999	38.679	9.676	**0.524**	0.419	0.719	0.874	0.872	0.916
		6	0.531	0.446	0.701	11.007	38.655	9.698	**0.524**	0.419	0.719	0.874	0.872	0.915
		7	0.529	0.446	0.700	11.047	38.744	9.704	**0.524**	0.419	0.720	0.874	0.872	0.915
		8	0.527	0.446	0.701	11.046	38.710	9.692	**0.524**	0.419	0.720	0.875	**0.873**	0.915
		9	0.528	0.444	0.701	11.044	38.773	9.679	**0.524**	0.418	0.721	0.874	0.872	0.915
		10	0.527	0.444	0.702	11.053	38.806	9.676	0.523	0.419	**0.722**	0.874	0.872	0.915
	max	1	0.526	0.448	0.700	11.071	38.842	9.716	0.516	0.417	0.707	0.864	0.870	0.912
		2	0.533	0.448	0.701	10.948	38.678	9.673	0.511	0.419	0.695	0.867	0.871	0.915
		3	0.530	0.448	0.701	10.980	38.731	9.712	0.508	0.421	0.692	0.872	0.870	0.915
		4	0.531	0.449	0.702	11.009	38.743	9.694	0.505	0.421	0.689	0.870	0.869	0.915
		5	0.532	0.448	0.701	10.986	38.755	9.687	0.505	0.419	0.685	0.870	0.869	0.915
		6	0.532	0.448	0.702	10.997	38.818	9.667	0.504	0.419	0.684	0.870	0.870	0.915
		7	0.530	0.448	0.702	11.027	38.810	9.664	0.502	0.419	0.682	0.870	0.870	0.915
		8	0.530	0.448	0.702	11.023	38.782	9.672	0.501	0.419	0.682	0.870	0.870	0.915
		9	0.529	0.448	0.702	11.020	38.782	9.686	0.500	0.419	0.679	0.871	0.870	0.915
		10	0.528	0.448	0.702	11.021	38.774	9.692	0.500	0.419	0.679	0.871	0.870	0.915
	avg	1	0.526	0.448	0.700	11.071	38.842	9.716	0.516	0.417	0.707	0.864	0.870	0.912
		2	0.535	0.448	0.702	10.938	38.605	9.660	0.520	0.421	0.708	0.868	0.872	0.916
		3	0.533	0.450	0.702	10.941	38.592	9.696	0.520	0.423	0.708	0.872	0.872	0.916
		4	**0.534**	0.449	0.702	**10.932**	38.613	9.673	0.520	0.423	0.708	0.872	0.872	0.917
		5	0.533	0.450	0.702	10.926	38.613	9.677	0.521	0.422	0.708	0.872	0.872	**0.917**
		6	0.533	0.449	0.702	10.948	38.579	9.659	0.521	0.421	0.708	0.872	**0.873**	**0.917**
		7	0.531	0.450	0.702	10.969	38.609	9.655	0.521	0.422	0.708	0.873	**0.873**	**0.917**
		8	0.531	0.449	0.702	10.963	38.595	9.657	0.521	0.422	0.708	0.874	**0.873**	**0.917**
		9	0.530	0.449	0.702	10.986	38.602	9.666	0.520	0.422	0.708	0.874	**0.873**	**0.917**
		10	0.529	0.449	**0.703**	10.993	38.606	9.668	0.520	0.422	0.707	0.874	**0.873**	**0.917**
GO	min	1	0.523	0.446	0.702	11.137	38.838	9.656	0.508	0.417	0.698	0.871	0.868	0.913
		2	0.519	0.449	0.701	11.190	38.746	9.683	0.514	0.420	0.702	0.876	0.869	0.911
		3	0.518	0.448	0.700	11.186	38.746	9.699	0.511	0.421	0.709	0.875	0.869	0.910
		4	0.520	0.449	0.700	11.176	38.650	9.684	0.515	0.423	0.715	0.875	0.871	0.910
		5	0.519	0.449	0.700	11.153	38.688	9.668	0.515	0.423	0.716	0.876	0.871	0.910
		6	0.520	0.448	0.700	11.150	38.708	9.668	0.516	0.423	0.717	0.875	0.871	0.910
		7	0.520	0.447	0.701	11.135	38.799	9.652	0.517	0.422	0.717	0.876	0.871	0.910
		8	0.520	0.448	0.701	11.141	38.774	9.664	0.518	0.422	0.717	**0.877**	0.872	0.910
		9	0.519	0.447	0.701	11.170	38.822	9.660	0.517	0.421	0.721	**0.877**	0.872	0.910
		10	0.519	0.448	0.702	11.170	38.819	9.664	0.518	0.422	0.721	0.876	0.872	0.909
	max	1	0.523	0.446	0.702	11.137	38.838	9.656	0.508	0.417	0.698	0.871	0.868	0.913
		2	0.521	0.447	0.702	11.175	38.803	9.646	0.506	0.418	0.687	0.874	0.869	0.912
		3	0.520	0.448	0.700	11.208	38.773	9.681	0.503	0.420	0.684	0.873	0.869	0.912
		4	0.519	0.451	0.700	11.142	38.728	9.706	0.501	0.422	0.683	0.874	0.870	0.912
		5	0.522	0.451	0.700	11.137	38.754	9.706	0.500	0.422	0.679	0.874	0.870	0.913
		6	0.521	0.450	0.701	11.143	38.707	9.682	0.498	0.423	0.678	0.874	0.870	0.913
		7	0.521	0.450	0.701	11.124	38.726	9.666	0.498	0.423	0.676	0.874	0.869	0.913
		8	0.522	0.450	0.702	11.120	38.729	9.673	0.498	0.423	0.674	0.874	0.870	0.914
		9	0.522	0.450	0.702	11.155	38.744	9.673	0.497	0.423	0.675	0.873	0.870	0.914
		10	0.521	0.450	0.702	11.169	38.717	9.661	0.516	0.422	0.674	0.873	0.870	0.913
	avg	1	0.523	0.446	0.702	11.137	38.838	9.656	0.508	0.417	0.698	0.871	0.868	0.913
		2	0.520	0.450	0.702	11.141	38.741	9.664	0.514	0.420	0.697	0.876	0.870	0.913
		3	0.521	0.450	0.701	11.151	38.653	9.676	0.512	0.422	0.700	0.876	0.870	0.913
		4	0.521	0.453	0.702	11.088	**38.574**	9.670	0.514	**0.425**	0.704	0.877	0.872	0.913
		5	0.522	0.453	0.702	11.086	38.625	9.664	0.515	**0.425**	0.704	0.877	0.872	0.914
		6	0.523	**0.453**	0.702	11.081	38.577	9.646	0.515	**0.425**	0.703	**0.877**	0.872	0.914
		7	0.522	0.452	0.702	11.066	38.591	**9.626**	0.516	**0.425**	0.703	**0.877**	0.872	0.915
		8	0.523	0.452	0.702	11.066	38.585	9.637	0.516	**0.425**	0.703	**0.877**	0.872	0.915
		9	0.523	0.452	0.702	11.066	38.590	9.638	0.517	**0.425**	0.704	**0.877**	0.872	0.915
		10	0.524	0.451	0.702	11.072	38.586	9.634	0.517	**0.425**	0.705	**0.877**	0.872	0.914

Table C1: Performance of semantic entailment depending on axioms, combination strategy and number of models for Deep-GOSE model

18

Axioms	Strategy	Models	F_{max}			S_{min}			AUPR			AUC		
			MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
GO	min	1	0.493	0.448	0.712	11.601	38.115	8.755	0.480	0.423	0.713	0.869	0.876	0.924
		2	0.494	0.448	0.715	11.585	38.076	8.713	0.484	0.424	0.723	0.870	0.877	0.925
		3	0.496	0.449	0.714	11.584	38.079	8.738	0.484	0.425	0.725	0.870	0.877	0.925
		4	0.497	0.449	0.714	11.559	38.096	8.710	0.484	0.425	0.727	0.869	0.878	0.923
		5	0.499	0.449	0.713	11.531	38.077	8.720	0.485	0.425	0.728	0.869	0.878	0.923
		6	0.498	0.448	0.713	11.530	38.109	8.735	0.485	0.424	0.730	0.868	0.878	0.923
		7	0.498	0.448	0.713	11.518	38.117	8.747	0.486	0.423	0.731	0.868	0.878	0.922
		8	0.498	0.448	0.713	11.515	38.123	8.779	0.486	0.424	0.732	0.868	0.878	0.922
		9	0.498	0.448	0.656	11.509	38.158	11.144	0.486	0.424	0.696	0.868	0.878	0.893
		10	0.499	0.448	0.652	11.499	38.163	11.456	0.487	0.424	0.682	0.867	0.878	0.879
	max	1	0.493	0.448	0.712	11.601	38.115	8.755	0.480	0.423	0.713	0.869	0.876	0.924
		2	0.500	0.450	0.715	11.506	38.072	8.706	0.483	0.424	0.706	0.870	0.877	0.926
		3	0.499	0.449	0.714	11.528	38.074	8.726	0.482	0.423	0.703	0.871	0.878	0.928
		4	0.500	0.451	0.715	11.504	38.070	8.718	0.482	0.424	0.699	0.871	0.878	0.928
		5	0.499	0.450	0.714	11.520	38.104	8.727	0.480	0.423	0.697	0.869	0.878	0.927
		6	0.500	0.450	0.714	11.528	38.143	8.731	0.480	0.423	0.696	0.869	0.878	0.927
		7	0.501	0.450	0.713	11.523	38.146	8.785	0.480	0.423	0.693	0.870	0.878	0.928
		8	0.501	0.450	0.713	11.516	38.195	8.792	0.480	0.423	0.691	0.870	0.879	0.928
		9	0.501	0.450	0.694	11.526	38.215	9.095	0.478	0.422	0.667	0.869	0.879	0.898
		10	0.500	0.450	0.694	11.540	38.228	9.115	0.478	0.422	0.664	0.870	0.879	0.894
	avg	1	0.493	0.448	0.712	11.601	38.115	8.755	0.480	0.423	0.713	0.869	0.876	0.924
		2	0.498	0.449	0.715	11.532	38.019	8.684	0.486	0.425	0.718	0.871	0.878	0.927
		3	0.498	0.450	0.715	11.483	37.984	8.681	0.487	0.426	0.719	0.872	0.879	0.928
		4	0.500	0.451	0.715	11.469	37.930	8.664	0.488	0.426	0.720	0.872	0.879	0.928
		5	0.500	0.450	0.715	11.448	37.934	8.672	0.488	0.426	0.720	0.871	0.879	0.928
		6	0.500	0.450	0.715	11.446	37.998	8.655	0.488	0.425	0.721	0.871	0.879	0.928
		7	0.502	0.450	0.715	11.431	37.987	8.678	0.489	0.425	0.720	0.871	0.879	0.929
		8	0.502	0.450	0.715	11.432	37.981	8.691	0.489	0.425	0.720	0.871	0.880	0.929
		9	0.502	0.450	0.714	11.421	37.975	8.710	0.489	0.425	0.774	0.871	0.880	0.918
		10	0.502	0.451	0.712	11.419	37.970	8.731	0.490	0.425	0.769	0.872	0.880	0.915
GOPlus	min	1	0.496	0.454	0.714	11.645	37.990	8.668	0.488	0.430	0.712	0.861	0.875	0.927
		2	0.498	0.453	0.715	11.529	37.973	8.672	0.488	0.429	0.725	0.867	0.877	0.927
		3	0.500	0.451	0.714	11.506	37.976	8.680	0.490	0.427	0.730	0.868	0.877	0.927
		4	0.500	0.452	0.714	11.509	37.948	8.673	0.491	0.429	0.731	0.868	0.878	0.928
		5	0.500	0.452	0.715	11.478	37.959	8.678	0.491	0.429	0.733	0.868	0.877	0.928
		6	0.501	0.452	0.715	11.450	37.996	8.705	0.491	0.429	0.733	0.867	0.877	0.928
		7	0.501	0.452	0.715	11.460	38.005	8.706	0.491	0.429	0.733	0.867	0.877	0.928
		8	0.501	0.452	0.714	11.462	38.024	8.707	0.491	0.428	0.734	0.867	0.877	0.928
		9	0.500	0.451	0.715	11.468	38.026	8.712	0.490	0.428	0.734	0.867	0.877	0.928
		10	0.500	0.451	0.715	11.468	38.041	8.710	0.490	0.428	0.735	0.867	0.878	0.928
	max	1	0.496	0.454	0.714	11.645	37.990	8.668	0.488	0.430	0.712	0.861	0.875	0.927
		2	0.501	0.454	0.714	11.527	37.988	8.672	0.489	0.429	0.708	0.865	0.878	0.928
		3	0.502	0.454	0.715	11.498	37.982	8.665	0.489	0.428	0.703	0.866	0.879	0.928
		4	0.503	**0.456**	0.715	11.477	37.930	8.662	0.489	0.429	0.701	0.866	0.880	0.929
		5	0.504	**0.456**	0.715	11.494	37.959	8.683	0.487	0.429	0.699	0.867	0.880	0.929
		6	0.502	**0.456**	0.715	11.504	37.954	8.692	0.485	0.429	0.697	0.867	0.880	0.929
		7	0.502	**0.456**	0.715	11.523	37.967	8.726	0.484	0.429	0.696	0.868	0.880	0.929
		8	**0.503**	**0.456**	0.715	11.514	37.988	8.732	0.485	0.429	0.695	0.868	0.879	0.928
		9	0.504	**0.456**	0.715	11.512	37.975	8.741	0.485	0.429	0.694	0.867	0.880	0.928
		10	0.504	**0.456**	0.715	11.519	37.958	8.742	0.484	0.429	0.693	0.867	0.880	0.928
	avg	1	0.496	0.454	0.714	11.645	37.990	8.668	0.488	0.430	0.712	0.861	0.875	0.927
		2	0.501	0.454	0.715	11.491	37.950	8.609	0.492	0.430	0.719	0.867	0.879	0.928
		3	0.501	0.454	0.716	11.452	37.911	8.629	0.492	0.429	0.720	0.868	0.880	0.929
		4	0.502	0.455	0.716	11.411	37.833	8.616	0.494	0.432	0.722	0.869	0.881	0.929
		5	0.502	**0.456**	**0.717**	11.409	**37.811**	**8.607**	0.494	0.432	0.722	0.869	0.881	0.929
		6	**0.503**	**0.456**	**0.717**	**11.404**	37.817	8.618	0.494	0.432	0.722	0.870	0.881	0.929
		7	0.502	0.455	**0.717**	11.443	37.839	8.624	0.494	0.432	0.722	0.870	0.881	0.929
		8	0.502	0.455	**0.717**	11.450	37.829	8.643	0.494	0.431	0.722	0.870	0.880	0.929
		9	0.502	**0.456**	**0.717**	11.461	37.820	8.639	0.493	0.431	0.722	0.869	0.880	0.929
		10	0.502	0.455	0.716	11.447	37.827	8.654	0.493	0.431	0.722	0.869	0.881	0.929

Table C2: Performance of semantic entailment depending on axioms, combination strategy and number of models for Deep-GOGATSE model

Axioms	Strategy	Models	$F_{\max}$		$S_{\min}$		AUPR		AUC	
			BPO	CCO	BPO	CCO	BPO	CCO	BPO	CCO
GO	min	1	0.429	0.658	37.441	9.778	0.422	0.661	0.820	0.842
		2	0.433	0.659	37.336	9.736	0.421	0.667	0.819	0.843
		3	0.434	0.659	37.351	9.766	0.423	0.668	0.813	0.843
		4	0.434	0.660	37.360	9.784	0.424	0.669	0.813	0.843
		5	0.433	0.659	37.463	9.808	0.423	0.669	0.811	0.842
		6	0.434	0.659	37.470	9.808	0.423	0.670	0.811	0.840
		7	0.434	0.659	37.570	9.821	0.423	0.670	0.811	0.840
		8	0.434	0.659	37.596	9.801	0.423	0.671	0.810	0.841
		9	0.434	0.659	37.602	9.802	0.423	0.671	0.809	0.839
		10	0.434	0.659	37.614	9.807	0.423	0.672	0.808	0.839
	max	1	0.429	0.658	37.441	9.778	0.422	0.661	0.820	0.842
		2	0.436	0.658	37.445	9.749	0.425	0.658	0.824	0.845
		3	0.443	0.657	37.384	9.839	0.427	0.654	0.825	0.847
		4	0.445	0.657	37.399	9.833	0.426	0.652	0.825	0.849
		5	0.446	0.657	37.389	9.818	0.425	0.651	0.824	0.848
		6	0.447	0.657	37.521	9.824	0.425	0.651	0.825	0.848
		7	0.447	0.657	37.574	9.843	0.425	0.698	0.824	0.848
		8	**0.448**	0.657	37.615	9.851	0.424	0.697	0.824	0.848
		9	**0.448**	0.658	37.682	9.848	0.424	0.697	0.823	0.848
		10	**0.448**	0.658	37.733	9.857	0.424	0.697	0.823	0.849
	avg	1	0.429	0.658	37.441	9.778	0.422	0.661	0.820	0.842
		2	0.431	0.660	37.268	9.726	0.419	0.665	0.824	0.845
		3	0.433	0.659	**37.146**	9.743	0.423	0.664	0.825	0.847
		4	0.432	0.660	37.175	9.752	0.422	0.665	0.825	0.848
		5	0.432	0.660	37.201	9.753	0.421	0.665	0.825	0.848
		6	0.432	0.660	37.185	9.765	0.421	0.664	0.826	0.847
		7	0.432	0.660	37.170	9.768	0.421	0.664	0.826	0.847
		8	0.432	0.660	37.182	9.765	0.421	0.664	0.826	0.849
		9	0.432	0.659	37.192	9.771	0.421	0.664	0.825	0.847
		10	0.432	0.660	37.196	9.763	0.421	0.664	0.824	0.848
GOPlus	min	1	0.428	0.658	37.313	9.809	0.414	0.659	0.823	0.847
		2	0.429	0.659	37.344	9.779	0.417	0.665	0.822	0.847
		3	0.431	0.659	37.340	9.789	0.420	0.668	0.821	0.842
		4	0.432	0.659	37.368	9.776	0.420	0.670	0.820	0.841
		5	0.432	0.659	37.440	9.763	0.420	0.671	0.819	0.840
		6	0.432	0.659	37.511	9.755	0.421	0.672	0.818	0.839
		7	0.433	0.660	37.495	9.752	0.421	0.672	0.816	0.839
		8	0.434	0.660	37.511	9.765	0.421	0.673	0.816	0.839
		9	0.434	0.660	37.545	9.773	0.422	0.673	0.815	0.839
		10	0.434	0.660	37.553	9.779	0.422	0.673	0.815	0.839
	max	1	0.428	0.658	37.313	9.809	0.414	0.659	0.823	0.847
		2	0.430	0.657	37.313	9.819	0.413	0.655	0.827	0.851
		3	0.436	0.657	37.336	9.844	0.413	0.653	0.827	0.852
		4	0.437	0.657	37.382	9.851	0.411	0.652	0.828	0.852
		5	0.439	0.657	37.355	9.854	0.410	0.650	0.827	0.851
		6	0.442	0.657	37.386	9.856	0.412	0.649	0.828	0.852
		7	0.442	0.658	37.424	9.860	0.412	0.699	0.827	0.852
		8	0.444	0.657	37.429	9.867	0.411	0.699	0.827	0.851
		9	0.445	0.657	37.456	9.873	0.411	0.699	0.827	0.850
		10	0.445	0.657	37.515	9.877	0.411	0.699	0.827	0.850
	avg	1	0.428	0.658	37.313	9.809	0.414	0.659	0.823	0.847
		2	0.429	0.658	37.263	9.784	0.416	0.662	0.826	0.851
		3	0.430	0.659	37.205	9.768	0.418	0.662	0.827	0.851
		4	0.430	0.660	37.279	9.752	0.418	0.664	0.828	0.850
		5	0.429	0.660	37.254	9.737	0.418	0.663	0.827	0.851
		6	0.430	0.660	37.227	9.727	0.418	0.664	0.827	0.850
		7	0.431	0.660	37.182	**9.724**	0.419	0.663	0.827	0.850
		8	0.431	0.659	37.153	9.729	0.419	0.664	0.827	0.849
		9	0.432	0.660	37.164	9.734	0.419	0.663	0.827	0.848
		10	0.432	0.660	37.174	9.731	0.419	0.663	0.826	0.848

Table C3: Performance of semantic entailment depending on axioms, combination strategy and number of models for DeepGOGATMFSE model

Axioms	Strategy	Models	$F_{\max}$		$S_{\min}$		AUPR		AUC	
			BPO	CCO	BPO	CCO	BPO	CCO	BPO	CCO
GO	min	1	0.452	0.673	38.317	10.062	0.427	0.685	0.868	0.892
		2	0.451	0.671	38.199	10.087	0.427	0.687	0.869	0.893
		3	0.454	0.671	38.136	10.052	0.429	0.688	0.868	0.891
		4	0.454	0.671	38.086	10.051	0.430	0.689	0.869	0.891
		5	0.453	0.671	38.163	10.062	0.428	0.690	0.869	0.889
		6	0.452	0.671	38.109	10.046	0.428	0.690	0.868	0.888
		7	0.452	0.671	38.123	10.041	0.428	0.691	0.868	0.888
		8	0.453	0.671	38.138	10.046	0.428	0.691	0.868	0.888
		9	0.452	0.671	38.180	10.059	0.428	0.691	0.868	0.888
		10	0.452	0.670	38.185	10.077	0.428	0.690	0.868	0.888
	max	1	0.452	0.673	38.317	10.062	0.427	0.685	0.868	0.892
		2	0.451	0.674	38.405	10.011	0.426	0.681	0.869	0.897
		3	0.454	0.674	38.353	10.036	0.429	0.679	0.870	0.898
		4	0.453	0.673	38.439	10.076	0.429	0.676	0.870	0.900
		5	0.453	0.675	38.587	10.093	0.428	0.676	0.870	0.900
		6	0.453	0.675	38.551	10.127	0.427	0.676	0.870	0.900
		7	0.452	0.674	38.653	10.136	0.426	0.675	0.869	0.900
		8	0.452	0.673	38.742	10.172	0.426	0.674	0.869	0.898
		9	0.452	0.674	38.813	10.164	0.425	0.674	0.869	0.898
		10	0.452	0.673	38.890	10.163	0.424	0.673	0.868	0.898
	avg	1	0.452	0.673	38.317	10.062	0.427	0.685	0.868	0.892
		2	0.452	0.673	38.271	10.007	0.427	0.686	0.870	0.897
		3	0.454	0.674	38.188	9.990	0.430	0.686	0.871	0.898
		4	0.454	0.673	38.177	9.987	0.430	0.686	0.871	0.899
		5	0.454	0.673	38.201	9.992	0.430	0.687	0.872	0.900
		6	0.454	0.674	38.124	10.007	0.429	0.686	0.872	0.900
		7	0.454	0.674	38.092	9.998	0.429	0.686	0.872	0.899
		8	0.454	0.674	38.121	9.996	0.429	0.686	0.872	0.899
		9	0.454	0.673	38.171	9.991	0.429	0.686	0.872	0.899
		10	0.454	0.673	38.187	9.996	0.429	0.686	0.871	0.898
GOPlus	min	1	0.452	0.669	38.383	10.050	0.428	0.729	0.871	0.897
		2	0.452	0.673	38.380	9.960	0.430	0.690	0.872	0.896
		3	0.453	0.673	38.305	9.970	0.431	0.691	0.872	0.896
		4	0.453	0.674	38.262	9.963	0.431	0.692	0.873	0.896
		5	0.454	0.673	38.224	9.957	0.431	0.692	0.873	0.896
		6	0.454	0.673	38.210	9.955	0.431	0.692	0.872	0.897
		7	0.454	0.673	38.203	9.951	0.431	0.693	0.872	0.897
		8	0.454	0.673	38.222	9.960	0.431	0.693	0.871	0.897
		9	0.454	0.673	38.182	9.978	0.431	0.693	0.872	0.896
		10	0.453	0.673	38.188	9.977	0.431	0.693	0.871	0.896
	max	1	0.452	0.669	38.383	10.050	0.428	0.729	0.871	0.897
		2	0.453	0.672	38.705	10.028	0.428	0.730	0.872	0.897
		3	0.453	0.673	38.895	10.046	0.428	0.731	0.871	0.897
		4	0.453	0.672	38.958	10.087	0.427	0.730	0.871	0.898
		5	0.453	0.671	38.972	10.112	0.427	0.730	0.871	0.897
		6	0.453	0.672	39.066	10.097	0.426	0.730	0.871	0.898
		7	0.453	0.672	39.080	10.077	0.426	0.730	0.871	0.898
		8	0.452	0.672	39.108	10.085	0.426	0.730	0.871	0.898
		9	0.452	0.671	39.273	10.088	0.426	0.730	0.871	0.898
		10	0.451	0.671	39.288	10.098	0.426	0.730	0.870	0.897
	avg	1	0.452	0.669	38.383	10.050	0.428	0.729	0.871	0.897
		2	0.453	0.673	38.471	9.968	0.431	0.687	0.873	0.898
		3	0.454	0.673	38.359	9.980	0.431	0.687	0.873	0.899
		4	0.454	0.672	38.307	10.018	0.432	0.687	0.873	0.899
		5	0.454	0.672	38.309	10.016	0.432	0.686	0.873	0.899
		6	0.454	0.672	38.307	10.020	0.432	0.686	0.873	0.899
		7	0.454	0.672	38.295	9.988	0.432	0.687	0.873	0.900
		8	0.454	0.672	38.292	9.993	0.431	0.687	0.873	0.900
		9	0.454	0.672	38.285	10.011	0.432	0.687	0.873	0.899
		10	0.454	0.672	38.296	9.995	0.431	0.687	0.873	0.899

Table C4: Performance of semantic entailment depending on axioms, combination strategy and number of models for DeepGOGATPredMFSE model

Method	$F_{\max}$			$S_{\min}$			AUPR			AUC		
	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
ESM	0.545	0.432	0.721	11.827	39.227	9.358	0.539	0.402	0.724	0.866	0.869	0.918
ESM + GO axioms	0.549	0.426	0.719	11.876	39.749	9.462	0.539	0.394	0.721	0.868	0.859	0.913
ESM + GO PLus axioms	0.552	0.426	0.717	11.750	39.686	9.645	0.550	0.393	0.728	0.867	0.861	0.907
DeepGOSE	**0.554**	0.432	0.721	11.681	39.419	9.499	**0.552**	0.401	0.730	0.874	0.864	0.914
ESM + GAT	0.535	0.432	0.730	11.978	39.201	8.809	0.536	0.404	0.802	0.837	**0.878**	0.927
ESM + GAT + GO axioms	0.521	0.430	0.727	12.229	39.460	8.743	0.516	0.398	0.733	0.860	0.873	0.927
ESM + GAT + GO PLus axioms	0.517	0.432	0.731	12.321	39.382	8.706	0.513	0.400	0.735	0.855	0.861	0.923
DeepGOGATSE	0.525	0.435	**0.736**	**11.137**	39.123	**8.634**	0.523	0.404	0.743	0.861	0.876	**0.930**
MF + GAT	-	0.453	0.671	-	**37.070**	9.693	-	**0.430**	0.721	-	0.833	0.844
MF + GAT + GO axioms	-	0.444	0.666	-	37.737	9.853	-	0.428	0.713	-	0.824	0.831
MF + GAT + GO PLus axioms	-	0.444	0.668	-	37.649	9.803	-	0.426	0.716	-	0.827	0.832
DeepGOGATMFSE	-	0.448	0.668	-	37.299	9.809	-	0.428	0.679	-	0.831	0.835
MF-Pred + GAT	-	**0.455**	0.699	-	38.943	9.868	-	0.422	**0.760**	-	0.864	0.895
MF-Pred + GAT + GO axioms	-	0.441	0.690	-	39.328	10.031	-	0.406	0.696	-	0.852	0.876
MF-Pred + GAT + GO PLus axioms	-	0.443	0.691	-	39.705	10.003	-	0.407	0.749	-	0.853	0.881
DeepGOGATMFSE-Pred	-	0.444	0.694	-	39.098	9.907	-	0.409	0.753	-	0.855	0.884

Table C5: Ablation study to analyze contributions of GO and GOPlus ontology axioms, PPIs, experimental and predicted MF annotations, and Semantic Entailment to the performance

K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021) https://doi.org/10.1038/s41586-021-03819-2

[2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, M.J., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Tarver, L.I., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000) https://doi.org/10.1038/75556

[3] Consortium, T.U.: UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research **51**(D1), 523–531 (2022) https://doi.org/10.1093/nar/gkac1052 https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf

[4] Zhou, N., Jiang, Y., Bergquist, T.R., Lee, A.J., Kacsoh, B.Z., Crocker, A.W., Lewis, K.A., Georghiou, G., Nguyen, H.N., Hamid, M.N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A.S., Dalkiran, A., Cetin-Atalay, R., Zhang, C., Hurto, R.L., Freddolino, P.L., Zhang, Y., Bhat, P., Supek, F., Fernández, J.M., Gemovic, B., Perovic, V.R., Davidović, R.S., Sumonja, N., Veljkovic, N., Asgari, E., Mofrad, M.R., Profiti, G., Savojardo, C., Martelli, P.L., Casadio, R., Boecker, F., Kahanda, I., Thurlby, N., McHardy, A.C., Renaux, A., Saidi, R., Gough, J., Freitas, A.A., Antczak, M., Fabris, F., Wass, M.N.,

Hou, J., Cheng, J., Hou, J., Wang, Z., Romero, A.E., Paccanaro, A., Yang, H., Goldberg, T., Zhao, C., Holm, L., Törönen, P., Medlar, A.J., Zosa, E., Borukhov, I., Novikov, I., Wilkins, A., Lichtarge, O., Chi, P.-H., Tseng, W.-C., Linial, M., Rose, P.W., Dessimoz, C., Vidulin, V., Dzeroski, S., Sillitoe, I., Das, S., Lees, J.G., Jones, D.T., Wan, C., Cozzetto, D., Fa, R., Torres, M., Vesztrocy, A.W., Rodriguez, J.M., Tress, M.L., Frasca, M., Notaro, M., Grossi, G., Petrini, A., Re, M., Valentini, G., Mesiti, M., Roche, D.B., Reeb, J., Ritchie, D.W., Aridhi, S., Alborzi, S.Z., Devignes, M.-D., Emily Koo, D.C., Bonneau, R., Gligorijević, V., Barot, M., Fang, H., Toppo, S., Lavezzo, E., Falda, M., Berselli, M., Tosatto, S.C., Carraro, M., Piovesan, D., Rehman, H.U., Mao, Q., Zhang, S., Vucetic, S., Black, G.S., Jo, D., Larsen, D.J., Omdahl, A.R., Sagers, L.W., Suh, E., Dayton, J.B., McGuffin, L.J., Brackenridge, D.A., Babbitt, P.C., Yunes, J.M., Fontana, P., Zhang, F., Zhu, S., You, R., Zhang, Z., Dai, S., Yao, S., Tian, W., Cao, R., Chandler, C., Amezola, M., Johnson, D., Chang, J.-M., Liao, W.-H., Liu, Y.-W., Pascarelli, S., Frank, Y., Hoehndorf, R., Kulmanov, M., Boudellioua, I., Politano, G., Di Carlo, S., Benso, A., Hakala, K., Ginter, F., Mehryary, F., Kaewphan, S., Björne, J., Moen, H., Tolvanen, M.E.E., Salakoski, T., Kihara, D., Jain, A., Šmuc, T., Altenhoff, A., Ben-Hur, A., Rost, B., Brenner, S.E., Orengo, C.A., Jeffery, C.J., Bosco, G., Hogan, D.A., Martin, M.J., O'Donovan, C., Mooney, S.D., Greene, C.S., Radivojac, P., Friedberg, I.: The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. bioRxiv (2019) https://doi.org/10.1101/653105 https://www.biorxiv.org/content/early/2019/05/29/653105.full.pdf

[5] You, R., Zhang, Z., Zhu, S., Sun, F., Xiong, Y., Mamitsuka, H.: GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics **34**(14), 2465–2473 (2018) https://doi.org/10.1093/bioinformatics/bty130 http://oup.prod.sis.lan/bioinformatics/article-pdf/34/14/2465/25138245/bty130.pdf. [PubMed:29522145] [doi:10.1093/bioinformatics/bty130]

[6] You, R., Yao, S., Mamitsuka, H., Zhu, S.: DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. Bioinformatics **37**(Supplement_1), 262–271 (2021) https://doi.org/10.1093/bioinformatics/btab270 https://academic.oup.com/bioinformatics/article-pdf/37/Supplement_1/i262/39620041/btab270.pdf

[7] You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., Zhu, S.: NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Research **47**(W1), 379–387 (2019) https://doi.org/10.1093/nar/gkz388 https://academic.oup.com/nar/article-pdf/47/W1/W379/28880124/gkz388.pdf. [PubMed:31106361] [PubMed Central:PMC6602452] [doi:10.1093/nar/gkz388]

[8] You, R., Huang, X., Zhu, S.: Deeptext2go: Improving large-scale protein function prediction with deep semantic text representation. Methods **145**, 82–90 (2018) https://doi.org/10.1016/j.ymeth.2018.05.026 . Data mining methods for analyzing biological data in terms of phenotypes

[9] Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen,

T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., Xavier, R.J., Knight, R., Cho, K., Bonneau, R.: Structure-based protein function prediction using graph convolutional networks. Nature Communications **12**(1), 3168 (2021) https://doi.org/10.1038/s41467-021-23303-9

[10] Lai, B., Xu, J.: Accurate protein function prediction via graph attention networks with predicted structure information. Briefings in Bioinformatics **23**(1) (2021) https://doi.org/10.1093/bib/bbab502 https://academic.oup.com/bib/article-pdf/23/1/bbab502/42747040/bbab502.pdf. bbab502

[11] Kulmanov, M., Hoehndorf, R.: DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. Bioinformatics **38**(Supplement_1), 238–245 (2022) https://doi.org/10.1093/bioinformatics/btac256 https://academic.oup.com/bioinformatics/article-pdf/38/Supplement_1/i238/44269467/btac256_supplementary_d

[12] Kulmanov, M., Khan, M.A., Hoehndorf, R.: DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics **34**(4), 660–668 (2017) https://doi.org/10.1093/bioinformatics/btx624 http://oup.prod.sis.lan/bioinformatics/article-pdf/34/4/660/25117339/btx624.pdf. [PubMed:29028931] [PubMed Central:PMC5860606] [doi:10.1093/bioinformatics/btx624]

[13] Kulmanov, M., Hoehndorf, R.: DeepGOPlus: improved protein function prediction from sequence. Bioinformatics (2019) https://doi.org/10.1093/bioinformatics/btz595 http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz595/29192371/btz595.p [PubMed:31350877] [doi:10.1093/bioinformatics/btz595]

[14] Cao, Y., Shen, Y.: TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding. Bioinformatics **37**(18), 2825–2833 (2021) https://doi.org/10.1093/bioinformatics/btab198 https://academic.oup.com/bioinformatics/article-pdf/37/18/2825/40471543/btab198.pdf

[15] Pan, T., Li, C., Bi, Y., Wang, Z., Gasser, R.B., Purcell, A.W., Akutsu, T., Webb, G.I., Imoto, S., Song, J.: PFresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. Bioinformatics **39**(3) (2023) https://doi.org/10.1093/bioinformatics/btad094 https://academic.oup.com/bioinformatics/article-pdf/39/3/btad094/49394696/btad094.pdf. btad094

[16] Wu, Z., Guo, M., Jin, X., Chen, J., Liu, B.: CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. Bioinformatics (2023) https://doi.org/10.1093/bioinformatics/btad123 https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btad123/49454316/bta btad123

[17] Wekesa, J.S., Luan, Y., Meng, J.: Predicting Protein Functions Based on Differential Co-expression and Neighborhood Analysis. J Comput Biol **28**(1), 1–18 (2021)

[18] Makrodimitris, S., Reinders, M.J.T., Ham, R.C.H.J.: Metric learning on expression data for gene function prediction. Bioinformatics **36**(4), 1182–1190 (2020)

[19] Pellegrini, M.: Using phylogenetic profiles to predict functional relationships. Methods Mol Biol **804**, 167–177 (2012)

[20] Nevers, Y., Prasad, M.K., Poidevin, L., Chennen, K., Allot, A., Kress, A., Ripp, R., Thompson, J.D., Dollfus, H., Poch, O., Lecompte, O.: Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. Mol Biol Evol **34**(8), 2016–2034 (2017)

[21] Yao, S., You, R., Wang, S., Xiong, Y., Huang, X., Zhu, S.: NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. Nucleic Acids Research **49**(W1), 469–475 (2021) https://doi.org/10.1093/nar/gkab398 https://academic.oup.com/nar/article-pdf/49/W1/W469/38842382/gkab398.pdf

[22] Krissinel, E.: On the relationship between sequence and structure similarities in proteomics. Bioinformatics **23**(6), 717–723 (2007) https://doi.org/10.1093/bioinformatics/btm006 https://academic.oup.com/bioinformatics/article-pdf/23/6/717/49822004/bioinformatics_23_6_717.pdf

[23] Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics (2015) https://doi.org/10.1093/bib/bbv011 http://bib.oxfordjournals.org/content/early/2015/04/10/bib.bbv011.full.pdf+html

[24] Chowdhury, T., Ling, C., Zhang, X., Zhao, X., Bai, G., Pei, J., Chen, H., Zhao, L.: Knowledge-enhanced Neural Machine Reasoning: A Review (2023)

[25] Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R.: Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics **22**(4) (2020) https://doi.org/10.1093/bib/bbaa199

[26] Tang, Z., Hinnerichs, T., Peng, X., Zhang, X., Hoehndorf, R.: FALCON: Faithful Neural Semantic Entailment over ALC Ontologies (2023)

[27] Kahanda, I., Ben-Hur, A.: Gostruct 2.0: Automated protein function prediction for annotated proteins. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. ACM-BCB '17, pp. 60–66. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3107411.3107417 . https://doi.org/10.1145/3107411.3107417

[28] Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., Atalay, V.: Deepred: Automated protein function prediction with multi-task feed-forward deep neural networks. Scientific Reports **9**(1), 7344 (2019) https://doi.org/10.1038/s41598-019-43708-3

[29] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al.: Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv (2022)

[30] Kulmanov, M., Liu-Wei, W., Yan, Y., Hoehndorf, R.: El embeddings: Geometric construction of models for the description logic el++. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 6103–6109. International Joint Conferences on Artificial Intelligence Organization, ??? (2019). https://doi.org/10.24963/ijcai.2019/845 . https://doi.org/10.24963/ijcai.2019/845

[31] Henkin, L., Suppes, P., Tarski, A.: The Axiomatic Method with Special Reference to Geometry and Physics. Proceedings of an International Symposium Held at the University of California, Berkeley, December 26, 1957-January 4, 1958. North-Holland Pub. Co Amsterdam, ??? (1959)

[32] Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo, S., Souza, M.L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Toronen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D.W.A., Bryson, K., Jones, D.T., Limaye, B., Inamdar, H., Datta, A., Manjari, S.K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A.M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A.E., Bhat, P., Paccanaro, A., Hamp, T., Kaszner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Honigschmid, P., Hopf, T.A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjorne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M.N., Sternberg, M.J.E., Skunca, N., Supek, F., Bosnjak, M., Panov, P., Dzeroski, S., Smuc, T., Kourmpetis, Y.A.I., Dijk, A.D.J., Braak, C.J.F.t., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P.C., Brenner, S.E., Orengo, C., Rost, B., Mooney, S.D., Friedberg, I.: A large-scale evaluation of computational protein function prediction. Nat Meth **10**(3), 221–227 (2013) https://doi.org/10.1038/nmeth.2340 . [PubMed:23353650] [PubMed Central:PMC3584181] [doi:10.1038/nmeth.2340]

[33] Mitchell, A., Sangrador-Vegas, A., Quinn, A.F., McAnulla, C., Binns, D., Nuka, G., McWilliam, H., Chang, H.-Y., Maslen, J., Fraser, M., Scheremetjew, M., Jones, P., Lopez, R., Hunter, S., Pesseat, S., Yong, S.-Y., Li, W.: InterProScan 5: genome-scale protein function classification. Bioinformatics **30**(9), 1236–1240 (2014) https://doi.org/10.1093/bioinformatics/btu031 http://oup.prod.sis.lan/bioinformatics/article-pdf/30/9/1236/653449/btu031.pdf.

[PubMed:24451626] [PubMed Central:PMC3998142] [doi:10.1093/bioinformatic-s/btu031]

[34] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. International Conference on Learning Representations (2018)

[35] Ogami, T., Tamura, Y., Toss, K., Yuki, K., Morikawa, M., Tsutsumi, S., Aburatani, H., Miyazawa, K., Miyazono, K., Koinuma, D.: -induced expression of target genes in epidermal keratinocytes. J Biochem **171**(4), 399–410 (2022)

[36] Kuchta, K., Knizewski, L., Wyrwicz, L.S., Rychlewski, L., Ginalski, K.: Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. Nucleic Acids Res **37**(22), 7701–7714 (2009)

[37] Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., Steinegger, M.: Fast and accurate protein structure search with foldseek. Nature Biotechnology (2023) https://doi.org/10.1038/s41587-023-01773-0

[38] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using diamond. Nature Methods **12**, 59 (2014). [PubMed:25402007] [doi:10.1038/nmeth.3176]

[39] Zahn-Zabal, M., Michel, P.-A., Gateau, A., Nikitin, F., Schaeffer, M., Audot, E., Gaudet, P., Duek, P.D., Teixeira, D., Rech de Laval, V., Samarasinghe, K., Bairoch, A., Lane, L.: The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Research **48**(D1), 328–334 (2019) https://doi.org/10.1093/nar/gkz995 https://academic.oup.com/nar/article-pdf/48/D1/D328/31697996/gkz995.pdf

[40] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**(D1), 353–361 (2016) https://doi.org/10.1093/nar/gkw1092 https://academic.oup.com/nar/article-pdf/45/D1/D353/8846820/gkw1092.pdf

[41] Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., Beckman, L., Shao, X., Fabregat, A., Sidiropoulos, K., Murillo, J., Viteri, G., Cook, J., Shorser, S., Bader, G., Demir, E., Sander, C., Haw, R., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The reactome pathway knowledgebase 2022. Nucleic Acids Research **50**(D1), 687–692 (2021) https://doi.org/10.1093/nar/gkab1028 https://academic.oup.com/nar/article-pdf/50/D1/D687/42058295/gkab1028.pdf

[42] Seaborne, A., Prud'hommeaux, E.: SPARQL query language for RDF. W3C recommendation, W3C (January 2008). http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

[43] Duek, P., Gateau, A., Bairoch, A., Lane, L.: Exploring the uncharacterized human proteome using nextprot. Journal of Proteome Research **17**(12), 4211–4226 (2018) https://

doi.org/10.1021/acs.jproteome.8b00537

[44] Duek, P., Mary, C., Zahn-Zabal, M., Bairoch, A., Lane, L.: Functionathon: a manual data mining workflow to generate functional hypotheses for uncharacterized human proteins and its application by undergraduate students. Database **2021** (2021) https://doi.org/10.1093/database/baab046 https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab046/39353617/baab046.pdf. baab046

[45] Nadendla, S., Jackson, R., Munro, J., Quaglia, F., Mészáros, B., Olley, D., Hobbs, E., Goralski, S., Chibucos, M., Mungall, C., Tosatto, S.C.E., Erill, I., Giglio, M.: ECO: the Evidence and Conclusion Ontology, an update for 2022. Nucleic Acids Research **50**(D1), 1515–1521 (2021) https://doi.org/10.1093/nar/gkab1025 https://academic.oup.com/nar/article-pdf/50/D1/D1515/42057816/gkab1025.pdf

[46] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv (2020) https://doi.org/10.1101/622803 https://www.biorxiv.org/content/early/2020/12/15/622803.full.pdf

[47] Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research **36**(suppl_1), 344–350 (2007) https://doi.org/10.1093/nar/gkm791 https://academic.oup.com/nar/article-pdf/36/suppl_1/D344/7627418/gkm791.pdf

[48] Mungall, C., Torniai, C., Gkoutos, G., Lewis, S., Haendel, M.: Uberon, an integrative multi-species anatomy ontology. Genome Biology **13**(1), 5 (2012) https://doi.org/10.1186/gb-2012-13-1-r5

[49] Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C.E., Vasilevsky, N.A., Haendel, M.A., Blake, J.A., Mungall, C.J.: The cell ontology 2016: enhanced content, modularization, and ontology interoperability. Journal of Biomedical Semantics **7**(1), 44 (2016) https://doi.org/10.1186/s13326-016-0088-7

[50] Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I.: NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database **2020** (2020) https://doi.org/10.1093/database/baaa062 https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa062/33570620/baaa062.pdf. baaa062

[51] Consortium, T.G.O.: The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research **49**(D1), 325–334 (2020) https://doi.org/10.1093/nar/gkaa1113 https://academic.oup.com/nar/article-pdf/49/D1/D325/35364517/gkaa1113.pdf

[52] Grau, B., Horrocks, I., Motik, B., Parsia, B., Patelschneider, P., Sattler, U.: OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4), 309–322 (2008)

[53] Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C.: Owl 2 web ontology language: Profiles. Recommendation, World Wide Web Consortium (W3C) (2009)

[54] Hoehndorf, R., *et al.*: A common layer of interoperability for biomedical ontologies based on OWL EL. Bioinformatics **27**(7), 1001–1008 (2011)

[55] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, ??? (2003)

[56] Cadoli, M., Schaerf, M.: Approximate entailment. In: Ardizzone, E., Gaglio, S., Sorbello, F. (eds.) Trends in Artificial Intelligence, pp. 68–77. Springer, Berlin, Heidelberg (1991)

[57] Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., Zhang, Z.: Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315 (2019)