The Gene Ontology

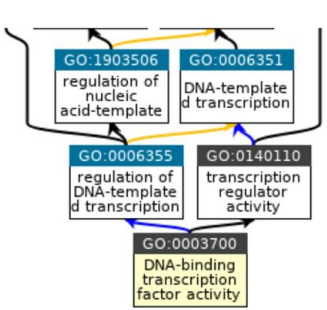
The Gene Ontology Consortium

Genetics 224.1: iyad031 (2023)
doi.org/10.1093/genetics/iyad031

Gökçe Uludoğan
PhD Candidate
LifeLU Reading Group | 24 October 2024

The Gene Ontology

(a)



black = ***is a***, **blue** = ***part of***, and **orange** = ***regulates***

(b)

] Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform	Reference	Date
] ZNF410	Zinc finger protein 410		RNA polymerase II cis-regulatory region sequence-specific DNA binding	has input UniProtKB:Q14839 occurs in erythroid lineage cell	UniProt	Homo sapiens	IMP		zinc finger protein pthr46179	protein		PMID:33301730	20210629
] ZNF410	Zinc finger protein 410		sequence-specific double-stranded DNA binding		ARUK-UCL	Homo sapiens	IDA		zinc finger protein pthr46179	protein		PMID:28473536	20200608
] ZNF410	Zinc finger protein 410		sequence-specific double-stranded DNA binding	has input UniProtKB:Q14839 occurs in erythroid lineage cell	UniProt	Homo sapiens	IMP		zinc finger protein pthr46179	protein		PMID:33301730	20210629

The Gene Ontology

- The terms used to describe **functional characteristics of gene products**, which are linked together by relations into a **labeled directed acyclic graph** (like a hierarchy but with multiple parents allowed).
- It also includes term definitions, synonyms, and relations to terms from external ontologies.
- The GO is available in different editions, including
  - (1) the **“basic” edition**, which includes only core relationship types;
  - (2) the **core ontology,** including additional relationship types; and
  - (3) the **“go-plus”** edition which also includes relationships to terms in other ontologies

GO versions

Name	Description	Permanent URL
go-basic.obo	The basic version of the GO, filtered such that the graph is guaranteed to be acyclic and annotations can be propagated up the graph. The relations included are *is a*, *part of*, *regulates*, *negatively regulates* and *positively regulates*. This version excludes relationships that cross the 3 GO hierarchies. This version should be used with most GO-based annotation tools.	http://purl.obolibrary.org/obo/go/go-basic.obo
go.obo & go.owl	Core ontology. This view includes relationships not in the filtered version of GO including *has_part* and *occurs_in*. Many of these relationships may not be safe for propagating annotations across, so this version should not be used with legacy GO tools. This version excludes relationships to external ontologies.	http://purl.obolibrary.org/obo/go.obo / http://purl.obolibrary.org/obo/go.owl
go-plus.owl	This is the fully axiomatised version of the GO. It includes cross-ontology relationships (axioms) and imports additional required ontologies including ChEBI, Cell Ontology and Uberon. It also includes a complete set of relationship types including some not in go.obo/go.owl. This version is only available in OWL format.	http://purl.obolibrary.org/obo/go/extensions/go-plus.owl

https://geneontology.org/docs/download-ontology/

Statistics

- The ontology contains about **40K terms**, linked together by **88,099 relationships** in the basic edition.
- When relationships to external terms are included, there are **121,698 relationships.**

Ontology

Property	Value
Valid terms	40939 (Δ = -1154)
Obsoleted terms	6965 (Δ = 1202)
Merged terms	2436 (Δ = 0)
Biological process terms	26552
Molecular function terms	10365
Cellular component terms	4022

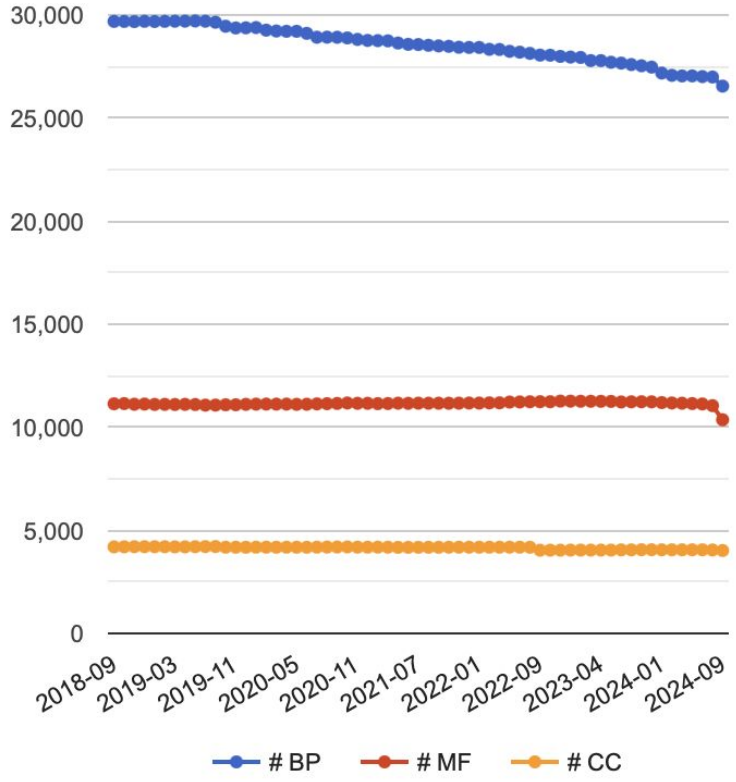
Annotations

Property	Value
Number of annotations	7,894,411
Annotations for biological process	2,862,559
Annotations for molecular function	2,531,611
Annotations for cellular component	2,500,241
Annotations for evidence PHYLO	3,908,659
Annotations for evidence IEA	1,746,505
Annotations for evidence EXP	1,037,435
Annotations for evidence OTHER	913,093
Annotations for evidence ND	229,103
Annotations for evidence HTP	59,616
Number of annotated scientific publications	180,792

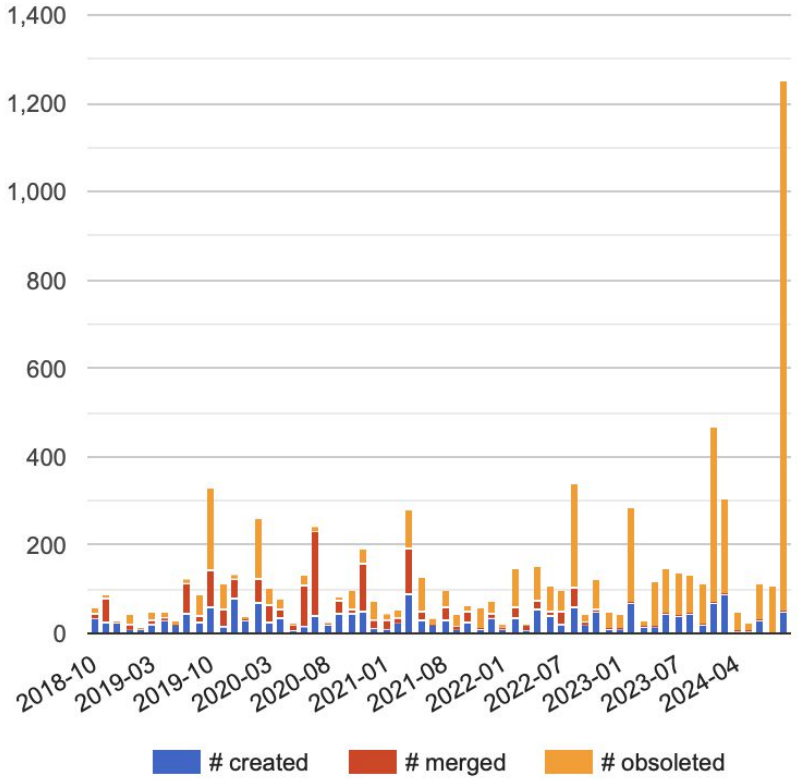
Gene products and species

Property	Value
Annotated gene products	1,573,444
Annotated species	5,426
Annotated species with over 1,000 annotations	184

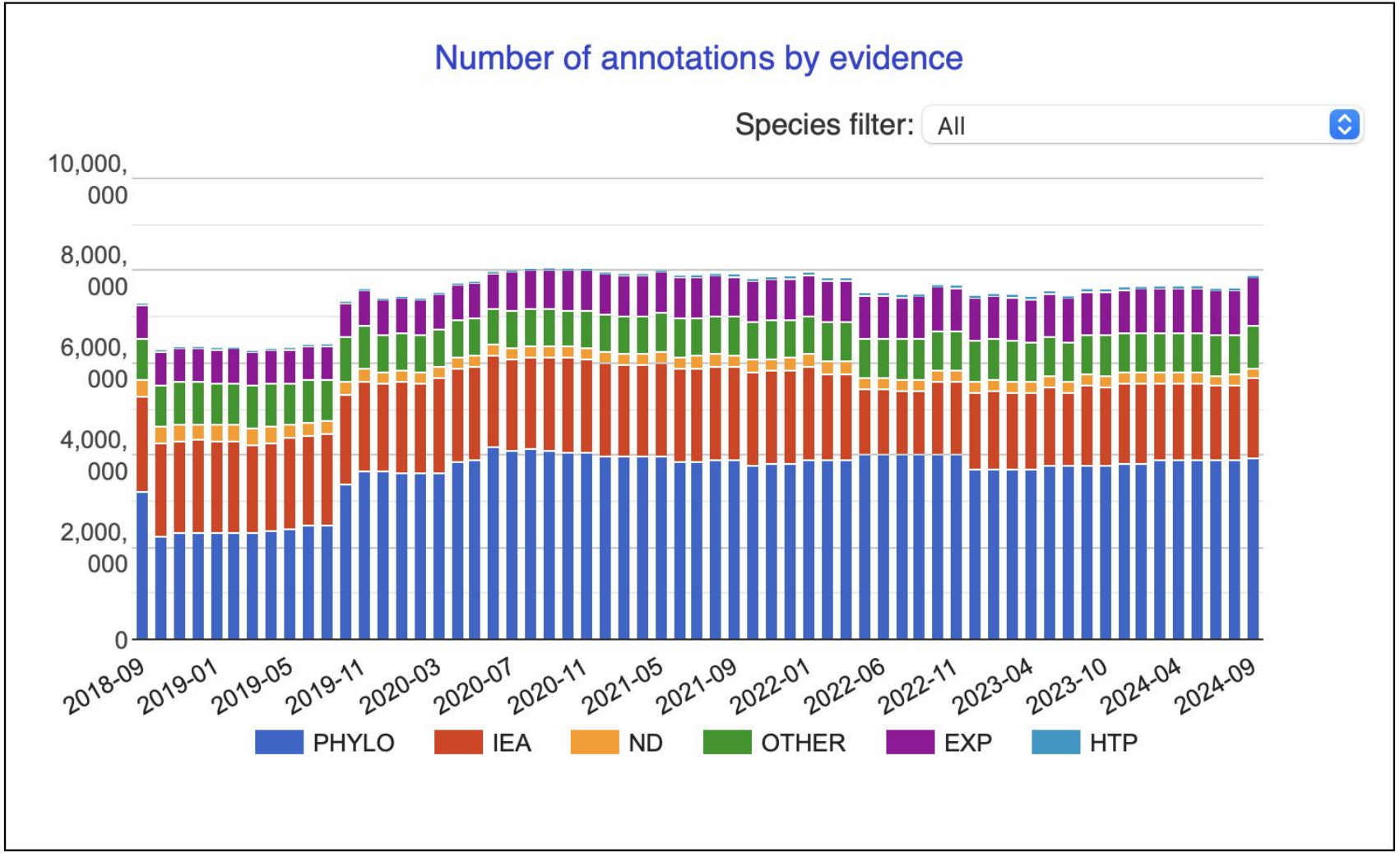
Number of GO terms by aspect



Changes in GO terms between releases



https://wiki.geneontology.org/Principles_for_term_obsoletion


Number of annotations by evidence
Species filter: All
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
2018-09
2019-01
2019-05
2019-11
2020-03
2020-07
2020-11
2021-05
2021-09
2022-01
2022-06
2022-11
2023-04
2023-10
2024-04
2024-09
PHYLO
IEA
ND
OTHER
EXP
HTP

Protein function prediction as approximate semantic entailment

Maxat Kulmanov, Francisco J. Guzmán-Vega, Paula Duek Roggli, Lydie Lane, Stefan T. Arold & Robert Hoehndorf

Nat Mach Intell 6, 220–228 (2024)
doi.org/10.1038/s42256-024-00795-w

Gökçe Uludoğan
PhD Candidate
LifeLU Reading Group | 24 October 2024

Motivation

Many function prediction methods rely on sequence similarity to predict functions.

Molecular functions arise largely from structure, and proteins with similar structures might have different sequence.

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.

Motivation

Methods that use the same sources of information for all three subontologies of GO are **limited**.

- Functions from the **MFO** subontology 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.

- Predicting BPO and CCO annotations requires **different sources of information** than predicting MFO annotations.

Motivation

Ontologies are another source of information **rarely exploited** for predicting protein functions.

Ontologies are not just collections of classes but formal theories that define **the meaning of classes** using logic-based languages.

Integrating these formal axioms into ML models allows for **leveraging prior knowledge**, constraining the parameter search space, and improving both the accuracy and efficiency of the learning process, leading to better predictions.

Description Logic provides the framework for ontologies

Description Logic is a family of formal knowledge representation languages used to represent structured knowledge and reasoning in a domain.

Fragments are subsets or variants of DL tailored to specific reasoning tasks.

Examples of fragments:

- **AL** (Attributive Language): Basic fragment, includes conjunction, universal restrictions, and atomic concepts.
- **ALC**: Adds concept negation to AL.
- **SHOIN**: Adds more expressive power (roles, cardinality restrictions) and is the basis for OWL-DL.

DL **fragments** offer varying levels of expressiveness and reasoning efficiency.

Description Logic

Description Logic is a formalism for representing knowledge with clear distinctions between schema (TBox), facts (ABox), and relationships (RBox).

TBox (Terminological Box)

- Represents the **schema** or **ontology**.
- Contains **axioms** about how concepts and roles relate.
    - Example: $\text{Parent} \sqsubseteq \text{Person}$ (Every parent is a person).
- Used for defining the vocabulary of a domain and relationships among concepts.

ABox (Assertional Box)

- Represents the **data** or **assertions**.
- Contains **individuals** (instances) and their relationships.
    - Example: John: Parent (John is a parent).
    - Example: $\text{hasChild(John, Mary)}$ (John has a child named Mary).
- Describes facts about individual instances and their roles in the ontology.

RBox (Role Box)

- Describes the **roles** or **relationships** between concepts.
- Includes axioms about the properties of roles:
    - **Symmetry**: $\text{hasSibling}(x, y) \rightarrow \text{hasSibling}(y, x)$
    - **Transitivity**: $\text{hasAncestor}(x, y) \land \text{hasAncestor}(y, z) \rightarrow \text{hasAncestor}(x, z)$.
    - **Inverse roles**: $\text{hasChild}(x, y) \rightarrow \text{hasParent}(y, x)$.

Function prediction methods utilize the formal axioms

GoStruct

DeepGO

DeePred

SPROF-GO

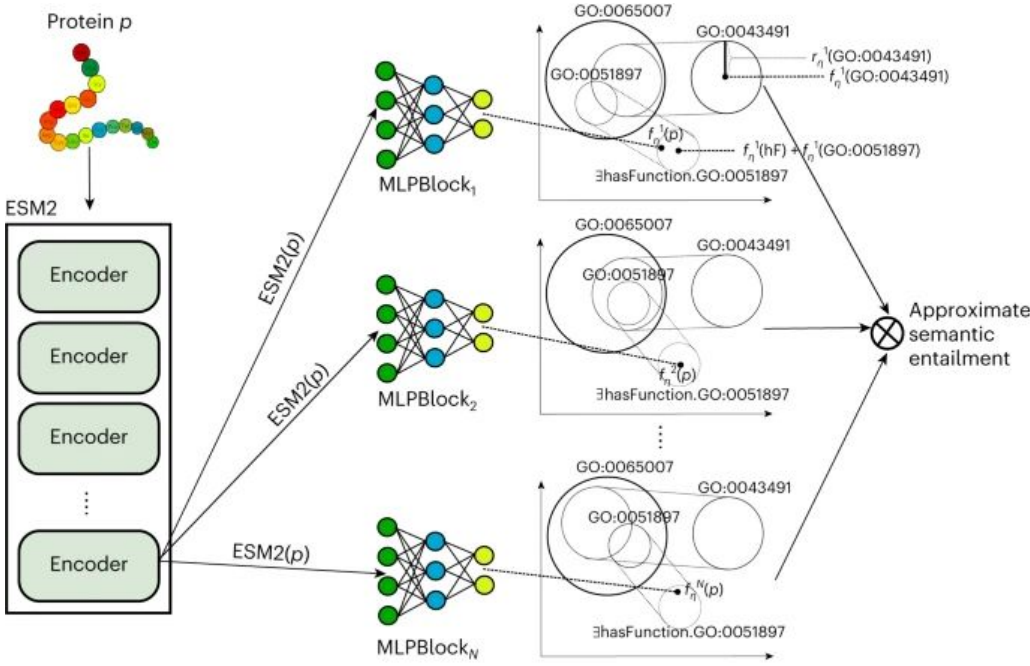
TALE

Uses **subsumption axioms (i.e. *is-a*)** to extract hierarchical relations between classes but **ignore other axioms** in GO.

DeepGO-SE

ESM2 embeddings are projected into an **embedding space** (ELEmbeddings) that is **generated from the axioms** in the GO

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.



Approximate semantic entailment

Suppose $\mathcal{O}$ is an ontology composed of a set of class symbols **C**, relation symbols **R** and individual symbols **I**, and that it is expressed in the Description Logic $\mathcal{ALC}$ (ref. 56). 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 nonempty 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:

$$
\begin{aligned}
&(C \sqcap D)^{\mathcal{I}} := C^{\mathcal{I}} \cap D^{\mathcal{I}}, (C \sqcup D)^{\mathcal{I}} := C^{\mathcal{I}} \cup D^{\mathcal{I}},\\
&(\forall R.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} | \forall e \in \Delta^{\mathcal{I}} : (d, e) \in R^{\mathcal{I}} \text{ implies } e \in C^{\mathcal{I}}\},\\
&(\exists R.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} | \exists e \in \Delta^{\mathcal{I}} : (d, e) \in R^{\mathcal{I}} \text{ and } e \in C^{\mathcal{I}}\},\\
&(\neg C)^{\mathcal{I}} := \Delta^{\mathcal{I}} - C^{\mathcal{I}}.
\end{aligned}
\tag{1}
$$

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 $\mathcal{O}$ (that is, an interpretation $\mathcal{I}$ that is a model of both ABox and TBox of $\mathcal{O}$) is also a model of ϕ ($\mathrm{Mod}(\mathcal{O}) \subseteq \mathrm{Mod}(\phi)$). Semantic entailment requires access to all models of $\mathcal{O}$ which are usually infinite; approximate semantic entailment considers only a strict (usually finite) subset of $\mathrm{Mod}(\mathcal{O})$ and tests whether ϕ is true in each of them[26,57].

EL-Embeddings

$$L = \frac{1}{N}\sum_{i=1}^{N} BCELoss(y_{c_i}, y'_{c_i}) + L_{NF1} + L_{NF2} + L_{NF3} + L_{NF4} \quad \text{(B10)}$$

ELEmbeddings normalizes TBox axioms in one of the following four normal forms:

NF1: $C \sqsubseteq D$, e.g., *binding* (GO:0005488) SubClassOf: *molecular function* (GO:0003674)
NF2: $C \sqcap D \sqsubseteq E$, e.g., *cutinase activity* (GO:0050525) and *biological regulation* (GO:0065007) SubClassOf: *positive regulation of protein kinase B signaling* (GO:0051897)
NF3: $C \sqsubseteq \exists R.D$, e.g., *positive regulation of arginine biosynthetic process* (GO:1900080) SubClassOf: *positively regulates* (RO:0002213) some *arginine biosynthetic process* (GO:0006526)
NF4: $\exists R.C \sqsubseteq D$, e.g., *part of* (BFO:0000050) some *conjugation* (GO:0000746) SubClassOf: *mammary stem cell proliferation* (GO:0002174)

$$L_{NF1} = \frac{1}{|NF1|} \sum_{c,d \in NF1} \max(0, \|f_\eta(c) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma) \tag{B11}$$

This loss goes to zero when the n-ball for class c is inside the n-ball for class d for all axioms of the first normal form.

$$\begin{aligned} L_{NF2} = \frac{1}{|NF2|} \sum_{c,d,e \in NF2} & \max(0, \|f_\eta(c) - f_\eta(d)\| - r_\eta(c) - r_\eta(d) - \gamma) + \\ & \max(0, \|f_\eta(c) - f_\eta(e)\| - r_\eta(c) - \gamma) + \\ & \max(0, \|f_\eta(d) - f_\eta(e)\| - r_\eta(c) - \gamma) + \\ & \max(0, \min(r_\eta(c), r_\eta(d)) - r_\eta(e) - \gamma) \end{aligned} \tag{B12}$$

$$L_{NF3} = \frac{1}{|NF3|} \sum_{r,c,d \in NF3} \max(0, \|f_\eta(c) - f_\eta(r) - f_\eta(d)\| - r_\eta(c) - r_\eta(d) - \gamma) \tag{B13}$$

Here, we translate the n-ball for class d using relation vector r and minimize the non-overlap between the translated n-ball and the n-ball for class c.

$$L_{NF4} = \frac{1}{|NF4|} \sum_{c,r,d \in NF4} \max(0, \|f_\eta(c) + f_\eta(r) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma) \tag{B14}$$

DeepGO-SE variants

- DeepGATGO-SE
  - Integrating PPI information

- DeepGATGOMF-SE
  - Including MF annotations

- DeepGATGOMF-SE-Pred
  - Utilizing predicted MF terms

Experiments

- UniProtKB/Swiss-Prot split by sequence similarity (DIAMOND)
- neXtProt dataset
- Metrics
  - Protein-centric
    - F_max
    - S_min
    - AUPR
  - Class-centric
    - AUC

- Naïve
  - Sequence features that are learned directly or using features derived from tools such as InterProScan
- MLP
  - InterPro domains or ESM2 embeddings.
- DeepGOCNN
  - Sequence + CNN
- DeepGOZero
  - InterPro domains + EL-embeddings
- DeepGraphGO
  - InterPRO domains + PPI
- TALE
  - Transformers + hierarchical loss for GO
- SPROF-GO
  - ProtT5-XL-U50 + hierarchical loss for GO + label diffusion

Table 1 | Prediction results for molecular functions on the UniProtKB/Swiss-Prot dataset

Method	F max	S min	AUPR	AUC
Naive	0.321	14.568	0.180	0.500
MLP	0.321	14.606	0.195	0.500
MLP (ESM2)	0.517	12.197	0.508	0.830
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

Table 2 | Prediction results for biological processes on the UniProtKB/Swiss-Prot dataset

Method	F max	S min	AUPR	AUC
Naive	0.294	43.934	0.195	0.500
MLP	0.295	43.914	0.210	0.499
MLP (ESM2)	0.423	39.721	0.388	0.864
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

Table 3 | Prediction results for cellular components on the UniProtKB/Swiss-Prot dataset

Method	F max	S min	AUPR	AUC
Naive	0.620	11.879	0.490	0.500
MLP	0.620	11.879	0.552	0.500
MLP (ESM2)	0.717	9.489	0.708	0.909
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

Table 4 | Prediction results for molecular functions on the neXtProt dataset

Method	F max	S min	AUPR	AUC
Naive	0.360	10.340	0.165	0.500
MLP	0.347	10.371	0.194	0.493
MLP (ESM2)	0.382	**9.985**	0.292	0.730
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
TALE	0.344	10.673	0.238	0.640
SPROF-GO	0.352	10.331	0.270	0.652
DeepGO-SE	**0.386**	10.093	**0.324**	**0.744**
DeepGOGAT-SE	0.375	10.254	0.291	0.700

Table 5 | Prediction results for biological processes on the neXtProt dataset

Method	F max	S min	AUPR	AUC
Naïve	0.308	32.987	0.183	0.500
MLP	0.310	32.033	0.206	0.502
MLP (ESM2)	0.336	**30.044**	0.305	0.682
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
TALE	0.280	32.973	0.221	0.533
SPROF-GO	0.312	31.164	0.251	0.620
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**

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
DeepGO-SE	**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
DeepGOGAT-SE	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
DeepGOGATMF-SE	-	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
DeepGOGATMF-SE-Pred	-	0.444	0.694	-	39.098	9.907	-	0.409	0.753	-	0.855	0.884

Table D2: Ablation study to analyze contributions of GO and GOPlus ontology axioms, PPIs, experimental and predicted MF annotations, and Semantic Entailment to the performance

DeepGO-SE Overview

- Enhances protein function prediction by combining:
  - **Protein sequence features** from pretrained language model (ESM2)
  - **GO knowledge** and **protein-protein interactions (PPIs)**
- **Zero-shot prediction** similar to DeepGOZero.

Key Takeaways

- **Knowledge-enhanced models** outperform those without background knowledge.
- GO function prediction benefits from a **hierarchical, separate approach (?)**.
- Models based on **ESM2** generalize well to unseen proteins.

Challenges and Future Work

- Best results achieved by combining **sequence** and **PPIs**, but **novel proteins** often **lack known interactions**.

- Integrating **PPI prediction methods** based on sequence and structure for novel proteins.