npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00358-0

GO2Sum: generating human-readable functional summary of proteins from GO terms

Check for updates

Swagarika Jaharlal Giri[1], Nabil Ibtehaz[1] & Daisuke Kihara[1,2]

Understanding the biological functions of proteins is of fundamental importance in modern biology. To represent a function of proteins, Gene Ontology (GO), a controlled vocabulary, is frequently used, because it is easy to handle by computer programs avoiding open-ended text interpretation. Particularly, the majority of current protein function prediction methods rely on GO terms. However, the extensive list of GO terms that describe a protein function can pose challenges for biologists when it comes to interpretation. In response to this issue, we developed GO2Sum (Gene Ontology terms Summarizer), a model that takes a set of GO terms as input and generates a human-readable summary using the T5 large language model. GO2Sum was developed by fine-tuning T5 on GO term assignments and free-text function descriptions for UniProt entries, enabling it to recreate function descriptions by concatenating GO term descriptions. Our results demonstrated that GO2Sum significantly outperforms the original T5 model that was trained on the entire web corpus in generating Function, Subunit Structure, and Pathway paragraphs for UniProt entries.

Elucidating the function of proteins is one of the most essential and important tasks in molecular biology, biochemistry, genetics, as well as bioinformatics. Conventionally, the function of proteins is described in free-text, such as those which we see in public protein and gene databases[1,2] as it is natural and versatile for biologists to describe various aspects of functions and behaviors of proteins. However, a drawback of text representation is that it is challenging for computer programs to extract and use function information in various tasks, such as finding proteins with a particular function from a genome or comparing functions of different proteins and defining functional similarity. To achieve machine-readable unified function description, gene ontology (GO)[3,4] a structured vocabulary for describing protein functions, has been developed about two decades ago, which is now well-established. GO classifies protein functions into three primary categories: biological process (BP), which pertains to pathway information; cellular component (CC), describing subcellular locations; and molecular function (MF), focusing on biochemical reactions of proteins. Using GO, the function of a protein can be represented as a list of GO terms. As each functional term is tokenized, GO has significantly facilitated computational studies of protein functions, such as quantitative comparison of protein functions and GO enrichment analysis[5], which is indispensable for proteomics and genomics studies.

Protein function prediction is one of the research areas which has substantially benefited from GO[6]. Although the function of a protein needs to be ultimately determined by wet lab experiments, computational prediction serves as a valuable tool by providing hypotheses and guiding biologists in designing experiments. Conventionally, computational function annotation (prediction) uses sequence database search[7,8] as a source of function information. Methods were developed that use sequence information with more thorough function information mining techniques[9–12]. Other information used includes protein domain composition in protein sequences[13,14], protein tertiary structures[15], protein networks[16], literature[16–19], and combinations of multiple sources[19–21]. The progress of computational function prediction has been objectively monitored by the community-wide function prediction assessment, the Critical Assessment of Function Annotation (CAFA)[22].

Most of the contemporary protein function prediction methods rely on GO. An output of protein function prediction comprises a list of GO terms, which can often be long, sometimes exceeding several dozen terms, and difficult for researchers to comprehend. In fact, the presentation of a list of GO terms has frequently led to questions and confusion among users of our protein function prediction web servers[23].

In this study, our primary objective is to transform a list of GO terms into human-readable text, utilizing a recent advancement in natural

[1]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [2]Department of Biological Sciences, Purdue University, West Lafayette, IN, USA.
e-mail: dkihara@purdue.edu

language processing (NLP) techniques. NLP, a field within computer science, has experienced significant progress, thanks to deep learning technologies[24]. With the use of transformer models, particularly large language models (LLMs), we can now perform a multitude of NLP tasks, including translation, summarization, and question answering, at a practical and efficient level[25]. The work we present here, GO2Sum, takes as its input short text descriptions of a set of GO terms that annotate a protein and produce a text summary describing the protein's function. To achieve this, we fine-tuned an LLM, the base version of Text-To-Text Transfer Transformer (T5)[26], on UniProt[27] entries with their GO annotations and functional descriptions.

LLMs have been increasingly used in bioinformatics domains. A pre-trained BERT language model[28] that was trained on the general text was further fine-tuned on biomedical literature to perform tasks such as sentence classification, dependency parsing[29], biomedical relation extraction, biomedical question answering[30], sentence similarity, and relation extraction[31]. PubMedBERT[32] is a BERT model trained on PubMed abstracts and performs tasks including relation extraction, question answering, and document classification. Instead of BERT, a text-to-text transfer transformer model, T5[26], as used in SciFive[33], which was trained on biomedical corpora and performed language generation tasks such as relation extraction, natural language inference, and question answering.

LLMs were also used for summarization tasks in the bioinformatics domain. Xie et al.[34] used a knowledge infusion training framework to enhance multiple PLMs for summarizing biomedical literature. Bidirectional and Auto-Regressive Transformers (BART)[35] and the domain-aware pre-trained language model in BioBERTSum[36] were used for abstractive and extractive summarization of biomedical evidence, respectively.

In GO2Sum, we employ T5 to summarize the functional descriptions from a list of GO terms. To the best of our knowledge, his work represents the pioneering effort in performing summarization within the protein function prediction domain. When selecting a technique for this general summarization task, we opted for T5 due to its superior performance demonstrated by the encoder-decoder-based architecture, outperforming earlier techniques[26,37,38]. From UniProt entries, we extracted three functional descriptions, Function, Subunit Structure, and Pathway. For each of these descriptions, we trained a separate T5 model. We showed that the fine-tuned T5 model performed significantly better than the pre-trained vanilla T5 in reproducing these function descriptions when we evaluated with three embedding-based metrics, BERT[28], MiniLM[39], and BioSentVec[40] as well as three sentence distance-based metrics, WMS[41], SMS[41], and W + SMS[41]. Moreover, in GO2Sum, we provide a confidence level of output summary by considering the probability of multiple variations of summary texts in beam search. Finally, we applied GO2Sum to the predicted GO terms generated by Phylo-PFP[9] and demonstrated that, in most cases, the summaries produced by GO2Sum exhibit sufficient accuracy even when derived from predicted GO terms.

Results

Framework of GO2Sum

The GO2Sum workflow, illustrated in Fig. 1, begins with a set of input GO terms, each accompanied by a text description. For instance, GO:0000049 is associated with the description 'Binding to a transfer RNA.' These GO term descriptions are concatenated into a document, serving as the input for the summarizer model, T5. The summarizer then generates a paragraph that elucidates the function of the input protein.

The dataset of GO terms and protein text descriptions was sourced from SwissProt (Release: 11 February 2022). Out of the 542,953 proteins included in this SwissProt release, we selected 518,422 proteins that had at least one GO term annotation. For each GO term, we obtained the text description from the Gene Ontology Consortium (Release: 16 November 2021). The length of these GO term descriptions ranged from 3 to 7,836 words, with an average of 310 words. Meanwhile, we gathered three different paragraphs that described the function of each protein from their SwissProt entries. These paragraphs included the general function description, usually presented at the top of the UniProt entry (referred to as "Function in UniProt"). We also collected two additional paragraphs: one related to "Subunit Structure", which details molecular interactions with other proteins, and the other providing "Pathway Information", explaining the metabolic pathways associated with the protein. These three paragraphs served as the ground truth for the protein's function description. For reference, some examples of these three function description paragraphs can be found in Supplementary Information 1.

To reduce the redundancy of proteins of similar function, we filtered out proteins that had 90% or more identical GO term annotations. This process reduced the dataset to 109,658 proteins from 542,953. As an entry often does not have all three functional descriptions of Function, Subunit Structure, and Pathway, we constructed three separate datasets for each. The dataset for Function, Subunit Structure, and Pathway had 97,600, 62,340, 14,600 proteins, respectively.

Here we explain more about the rationale behind this GO term-centric dataset construction process. Usually, when we construct a non-redundant dataset for bioinformatics studies, such as a dataset for protein structure prediction or protein function (GO term) prediction, redundant proteins are identified with a sequence identity of 25–40%[15,42]. However, through a close examination, we noticed that there are noticeable number of protein pairs where protein sequence similarity is low but have a very similar set of GO term set annotation and function description. For instance, Protein Q88YN8 (phosphonoacetaldehyde hydrolase) and Q183T0 (bifunctional phosphonoacetaldehyde hydrolase) have a sequence identity of 15.2%, but their GO annotations and function description are quite similar. The overlap between their GO annotations was 100% when considering the annotation of the former protein and 75% using the latter protein. To remove such redundancy, we decided to define redundancy based on GO term annotations, which would be more reasonable because GO

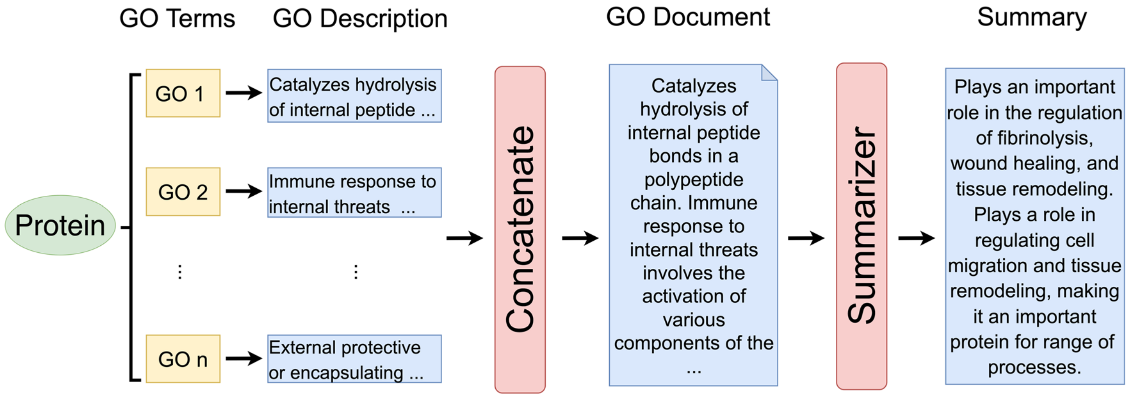
GO Terms
GO Description
GO Document
Summary
Protein
GO 1
GO 2
GO n
Catalyzes hydrolysis of internal peptide ...
Immune response to internal threats ...
External protective or encapsulating ...
Concatenate
Catalyzes hydrolysis of internal peptide bonds in a polypeptide chain. Immune response to internal threats involves the activation of various components of the ...
Summarizer
Plays an important role in the regulation of fibrinolysis, wound healing, and tissue remodeling. Plays a role in regulating cell migration and tissue remodeling, making it an important protein for range of processes.

Fig. 1 | GO2Sum workflow.

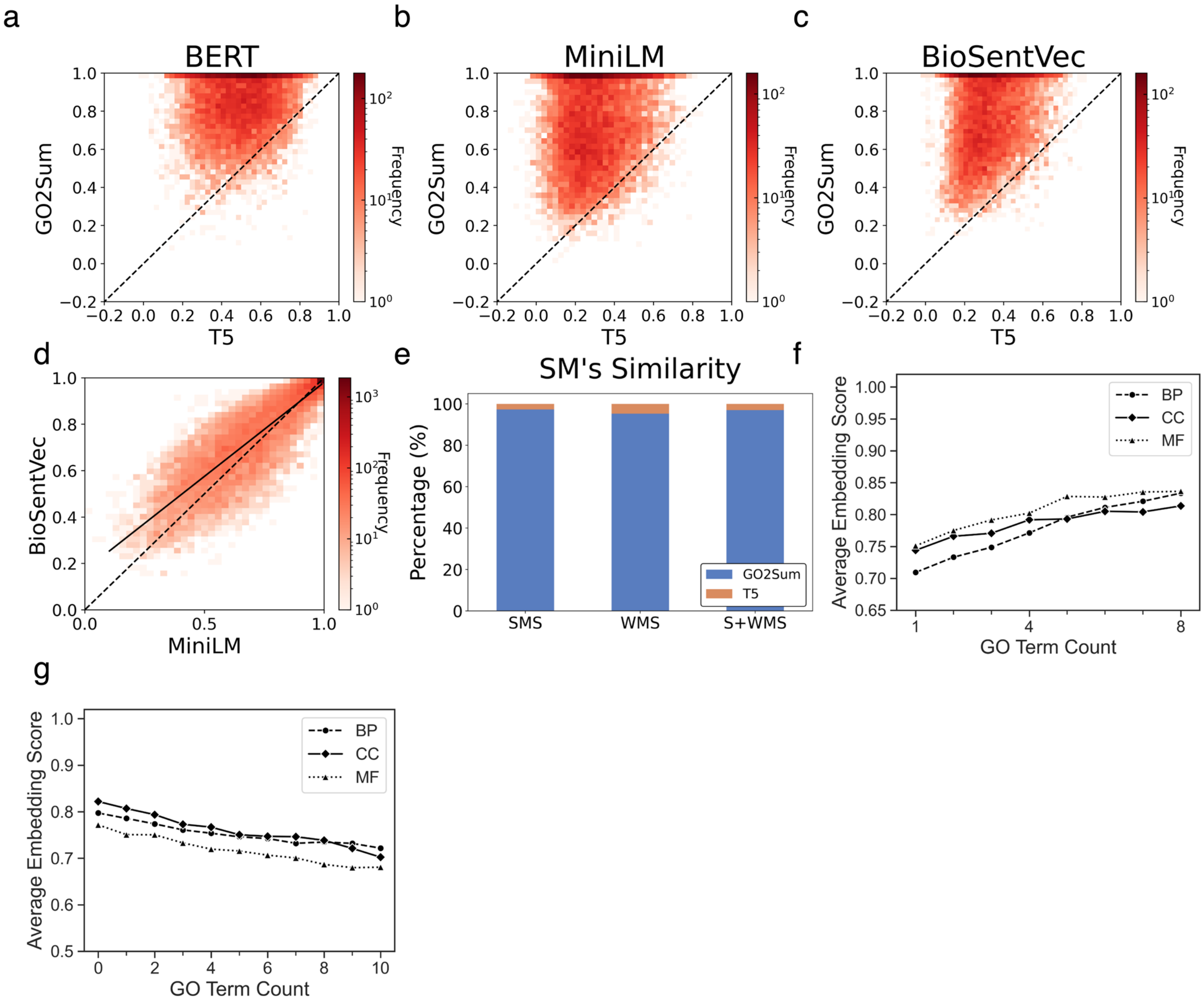

Fig. 2 | Evaluation scores for UniProt function paragraphs. **a**–**c** Comparison between GO2Sum and the vanilla T5 using the three embedding-based scores, **a** BERT, **b** MiniLM, and **c** BioSentVec. The number of UniProt entries used was 9760. **d** Comparison between MiniLM and BioSentVec. The correlation coefficient was 0.886. The dashed line is $y = x$ line, and the solid line is the regression line. **e** Comparison between GO2Sum and the vanilla T5 using the three Sentence Mover (SM)'s similarity-based metrics, WMS, SMS, and S + WMS. The number of wins (i.e., entries where GO2Sum produced closer paragraph to ground truth than the vanilla T5) are shown in blue. **f** impact of reducing GO terms. From a set of proteins, we reduced GO terms from their annotations randomly and observed how the average embedding score changed. The number of proteins in the dataset for BP, CC, and was 215, 130, and 55, respectively. The x axis shows the number of remaining GO terms in GO annotations after removal of a certain number of GO terms. For each target, 1, 2, 3, …, and 7 GO terms were removed randomly three times, and the embedding score was averaged over the three trials. Then, the value from each target was averaged across all the targets and plotted along the y axis. **g** impact of adding GO terms. We added randomly selected GO terms to their annotations. The number of proteins in the dataset for BP, CC, and was 296, 165, and 227, respectively. The x axis shows the number of GO terms added randomly. The y axis is the average embedding score from three trials.

terms are the actual input information for GO2Sum. To determine the cutoff for GO term annotation similarity, we collected protein pairs with varied degrees of GO term overlap, ranging from 50% to 100% with intervals of 5%. (We computed the GO term overlap of two proteins with GO terms of either protein as the denominator, choosing the larger value.) Upon examination, we found that when the GO term overlap exceeded 90%, the textual descriptions of the functions became substantially similar. Therefore, we clustered the proteins using the 90% GO term overlap and randomly picked one protein from each cluster to construct the dataset. This reduces the dataset to 109,658.

Following earlier studies of text summarization[43,44] we opted for an 80/10/10 split for training/validation/testing. Therefore, the training/validation/test sets of the Function dataset included 78,080/9760/9760 entries, Pathway had 11,680/1460/1460 entries, and Subunit had 49,872/6234/6234 entries, respectively. Due to the limitation of the size of GPU memory, we used, up to 1024 tokens or words were used from a concatenated GO document if it exceeded the size.

Among several T5 models with different parameter sizes, we used the T5-base model with 220 million parameters, 12 transformer layers, and 12 attention heads. Computation of training was performed on NVIDIA RTX5500 24GB memory, and inference was performed on NVIDIA RTX A6000 GPU with 48GB memory. We trained three models, each for producing paragraphs for Function, Subunit structure, and Pathway sections in UniProt, respectively. Supplementary Fig. 1 shows how loss changed during the training process of the models. For the loss function, token-level cross-entropy loss between the predicted summary and the ground truth summary was used.

Summarization performance assessment

In order to assess the summarization performance across the three categories, i.e., function, subunit structure, and pathway, we employed two types of scoring methods: embedding-based scores and sentence-mover distance-based similarities.

Embedding-based scores provide a quantitative comparison between the output generated by GO2Sum and the ground truth Function paragraphs obtained from UniProt entries. These scores are more suitable over traditional ones that rely on n-gram overlaps, such as ROUGE[45] and BLEU[46], as embedding takes into account word meaning and semantic context. We used three different embedding models, BERT[28], MiniLM[39], and BioSentVec[40] embeddings. BERT and MiniLM are pre-trained language models for general natural language processing (NLP) tasks, while BioSentVec is specially designed for biomedical texts. Using these embeddings, paragraph similarity is quantified through cosine similarity with a score range from -1.0 (complete irrelevance) to 1.0 (perfect match).

For the second type of score, mover distance-based similarity, we used three metrics, Word Mover's Similarity (WMS)[41], Sentence Mover's Similarity (SMS)[41], and Sentence and Word Mover's Similarity (S + WMS)[41]. WMS combines bag-of-word histogram representations with word embedding similarity. SMS treats documents as bags of sentence embeddings and solves a linear optimization problem. S + WMS represents each document as a collection of words and sentences, solving the same linear equation as WMS. These similarity scores, produced by SMS, WMS, and S + WMS, are relative measures of similarity between two documents. They fall within the range of 0 to 1 and indicate the minimum cost required to transform one document into the other.

GO summary for UniProt function

In Fig. 2a–c, we compared the performance of GO2Sum relative to the vanilla T5 in reproducing UniProt Function. Across all three embedding-based scores, GO2Sum's generated summaries achieved higher scores than vanilla T5 for the majority of UniProt entries. Specifically, GO2Sum's summaries outperformed vanilla T5 for 95.5%, 96.7%, and 98.0% of UniProt entries with BERT, MiniLM, and BioSentVec embeddings, respectively. These scores are highly correlated, as illustrated in Fig. 2d, where the correlation between MiniLM and BioSentVec had a Pearson's correlation coefficient of 0.886 (for additional score pairs, refer to Supplementary Fig. 2). Given this high correlation, we use the average score of the three for evaluation in the rest of the discussion.

In Fig. 2e, we used the three mover distance-based similarity scores to compare GO2Sum and vanilla T5. The results show that the summary by GO2Sum was more similar to the ground truth than vanilla T5 for almost all UniProt entries, 95.3%, 97.3%, and 97.0% of the cases by WMS, SMS, and S + WMS, respectively. When we took a close look at the remaining cases where GO2Sum is less similar to UniProt Function, typically those are cases in which both GO2Sum and T5 failed to produce meaningful paragraphs. For example, The UniProt entry Q4R7M2 describes the protein as a dipeptidase that is capable of hydrolyzing cystinyl-bis-glycine, but unable to hydrolyze leukotriene D4 into leukotriene E4. Neither T5 nor GO2Sum was successful for this target. T5 produced paragraph "a phospholipid bilayer and associated proteins. a phospholipid bi," while GO2Sum produced "Has a role in meiosis", both of which were irrelevant to the actual function. The average embedding-based score of GO2Sum and T5 were 0.21 and 0.28 for such cases where GO2Sum lost over T5, indicating that the summary of both GO2Sum and T5 was of a very poor quality.

We also compared GO2Sum with SciFive[33] in Supplementary Fig. 3. SciFive is a T5-based language model that was fine-tuned on biomedical and biological text corpus. Although SciFive was not trained specifically for summarization, the code is capable of performing summarization. GO2Sum had higher score for 91.52% of targets in terms of BERT, 98.01% for MiniLM, and 96.76% for BioSentVec, respectively. As shown in Supplementary Fig. 4, the vanilla T5 and SciFive performed quite similarly on the summarization task. To investigate the impact of the number of GO terms on GO2Sum's Function summary generation, we conducted an experiment depicted in Fig. 2f, g. We systematically removed (Fig. 2f) or added (Fig. 2g) a fixed number of GO terms from the target protein's GO annotation and examined the average of the three embedding scores. For this experiment, we curated datasets of protein targets for each of the three GO categories for the deletion and addition experiments. In the case of the BP category for the deletion experiment dataset, we selected proteins with exactly eight GO terms within their annotation, at least one GO term for each of the other categories (CC and MF, and a total count of GO terms ranging from 10 to 24. This procedure yielded datasets of 215, 130, and 59 proteins for BP, CC, and MF, respectively. From these datasets, we randomly removed a fixed number of GO terms (ranging from 1 to 7) from each protein's GO annotation. Subsequently, we ran GO2Sum with the remaining GO terms to generate Function summaries and evaluated the summaries using the average of the three embedding scores. To account for randomness, we repeated this experiment three times and averaged the results for each protein. For the addition experiment, we prepared a different dataset because adding more GO terms can exceed the length of tokens (up to 1024 tokens) our system can handle. For BP, we selected protein targets with exactly five BP GO terms, at least one GO term for each of CC and MF categories, and the total GO terms ranging from 7 to 15. A protein was removed if it still has >512 tokens in their original annotation because more GO terms were going to be added to it. This procedure yielded dataset of 296, 165, and 227 proteins for BP, CC, and MF, respectively. Similar to the deletion experiment, we added 1 to 10 randomly selected irrelevant GO terms three times and averaged the results.

In the GO deletion experiment (Fig. 2f), starting from the score of around 0.80 when there were eight GO terms (full annotation), the embedding score gradually declined as more GO terms were removed, ultimately reaching around 0.70 to 0.75 when only one GO term of that category remained. It's worth noting that the score did not drop further because GO terms of other categories were retained throughout the removal process. This effect varied among different ontologies. Removing BP terms had a more pronounced impact, resulting in a rapid decline in the embedding score. With an initial eight BP terms, the mean average embedding score was 0.83, decreasing to 0.70 when only one BP term remained. In contrast, CC terms had the least impact on Function annotation, with an average embedding score of 0.81 with all eight GO terms, decreasing to 0.74 when one GO term remained.

In the GO addition experiment (Fig. 2g), starting from the score of 0.79 (BP), 0.82 (CC), and 0.77 (MF), when there were exactly five original GO term annotations, the embedding score gradually declined as more GO terms were added, ultimately reaching 0.72 (BP), 0.70 (CC), and 0.68 (MF). The results are similar to what was observed in the deletion experiment. A score reduction of around 0.1 was observed after adding 10 irrelevant GO terms.

In Table 1, we present several examples of generated summaries. Each example includes a Function paragraph from UniProt, which is compared with the summaries produced by GO2Sum and the vanilla T5. The first three examples (UniProt ID: P0DMN7, Q86A79, and Q15438) showcase cases where GO2Sum achieved impressive average embedding-based scores ranging from 0.93 to 0.98. As demonstrated, with these high scores, GO2Sum generated summaries that closely matched the ground truth, with only negligible differences. These differences were typically minor, such as the absence of a specific protein name or slight variations in wording with the same meaning. In contrast, the vanilla T5 produced shorter, incomplete phrases, making the disparity between GO2Sum and T5 evident.

The next two examples, Q03233 and Q9ET67, are cases where GO2Sum had a moderate average embedding-based score of 0.69 to 0.74. These proteins have relatively brief function descriptions in UniProt. While GO2Sum's summaries correctly capture the protein functions, there are differences in wording and expression compared to the UniProt descriptions. In contrast, the vanilla T5 model provides fragmented phrases. Notably, T5's summary for Q03233 is incorrect as it fails to mention protein degradation.

Table 1 | Examples of summary generated for UniProt function

Protein name	ID	UniProt	T5	GO2Sum	Average embedding score
S-adenosylmethionine decarboxylase proenzyme 1	P0DMN7	Essential for the biosynthesis of the polyamines spermidine and spermine. Promotes maintenance and self-renewal of embryonic stem cells, by maintaining spermine levels.	Spermine. spermidine. Binding to putrescine	Essential for biosynthesis of the polyamines spermidine and spermine. Promotes maintenance and self-renewal of embryonic stem cells.	0.98
Cleavage and polyadenylation specificity factor subunit	Q86A79	Component of the cleavage and polyadenylation specificity factor complex that play a key role in pre-mRNA 3'-end formation, recognizing the AAUAAA signal sequence and interacting with poly polymerase and other factors to bring about cleavage and poly addition. Has endonuclease activity, and functions as mRNA 3'-end-processing endonuclease.	The 3' end of an mRNA molecule.	Component of the cleavage and polyadenylation specificity factor complex that plays a key role in pre-mRNA 3'-end formation, recognizing the AAUAAA signal sequence and interacting with poly polymerase and other factors to bring about cleavage and poly addition. This subunit has endonuclease activity.	0.98
Cytohesin-1	Q15438	Promotes guanine-nucleotide exchange on ARF1, ARF5 and ARF6. Promotes the activation of ARF factors through replacement of GDP with GTP. Plays an important role in membrane trafficking, during junctional remodeling and epithelial polarization, through regulation of ARF6 activity.	Guanine triphosphate (GTP), which is more abundant in the cell	Promotes guanine-nucleotide exchange on ARF1 and ARF6. Promotes the activation of ARF factors through replacement of GDP with GTP. Plays a role in vesicular transport.	0.93
Peroxisome assembly protein 12	Q9ET67	Required for protein import into peroxisomes.	A zinc ion (Zn). a peroxisome is	Somewhat implicated in the biogenesis of peroxisomes.	0.74
Alpha1-proteinase inhibitor-degradation deficient protein 37	Q03233	Involved in ER-associated protein degradation.	not including the plasma membrane and nucleus, but including other subcellular structures.	May be involved in the endoplasmic reticulum quality control system called ER-associated degradation.	0.69
Large ribosomal subunit protein P1A	P05318	Component of the ribosome, a large ribonucleoprotein complex responsible for the synthesis of proteins in the cell. The small ribosomal subunit binds messenger RNAs and translates the encoded message by selecting cognate aminoacyl-transfer RNA molecules. The large subunit contains the ribosomal catalytic site termed the peptidyl transferase center, which catalyzes the formation of peptide bonds, thereby polymerizing the amino acids delivered by tRNAs into a polypeptide chain. The nascent polypeptides leave the ribosome through a tunnel in the LSU and interact with protein factors that function in enzymatic processing, targeting, and the membrane insertion of nascent chains at the exit of the ribosomal tunnel.	A ribosome. a ribosome. A	Plays an important role in the elongation step of protein synthesis.	0.59
Skn-1 dependent zygotic transcript 15 protein	P46500	May have a role in embryogenesis.	Embryonic life stage. zygote formation to the end of embryonic life	Required for zygotic survival.	0.47
Chromodomain-helicase-DNA-binding protein 6	A3KFM7	DNA-dependent ATPase that plays a role in chromatin remodeling. Regulates transcription by disrupting nucleosomes in a largely non-sliding manner which strongly increases the accessibility of chromatin. Activates transcription of specific genes in response to oxidative stress through interaction with NFE2L2.	The basal transcription machinery. a transcription factor. ATP hydrolysis.	Probable transcription regulator.	0.47
Homeobox protein HOX1A	P46605	Interacts with the shrunken 26 bp feedback control element.	The organellar chromosomes. the nucleus is the site.	May be involved in transcriptional regulation.	0.37
Inactive ribonuclease-like protein 9	P60154	Does not exhibit any ribonuclease activity	A nucleic acid. space outside of the plasma membrane. For cells without	Involved in sperm motility. Seems to act as a signaling molecule which is required for the flagellated sperm to enter fertilization competently.	0.19
T-cell antigen CD7	P09564	Not yet known.	A lipid bilayer.. a lipid bilayer.	Receptor for TNFSF13B/BLyS/BAFF and TNFSF13/APRIL. Promotes B-cell activation and differentiation.	0.16

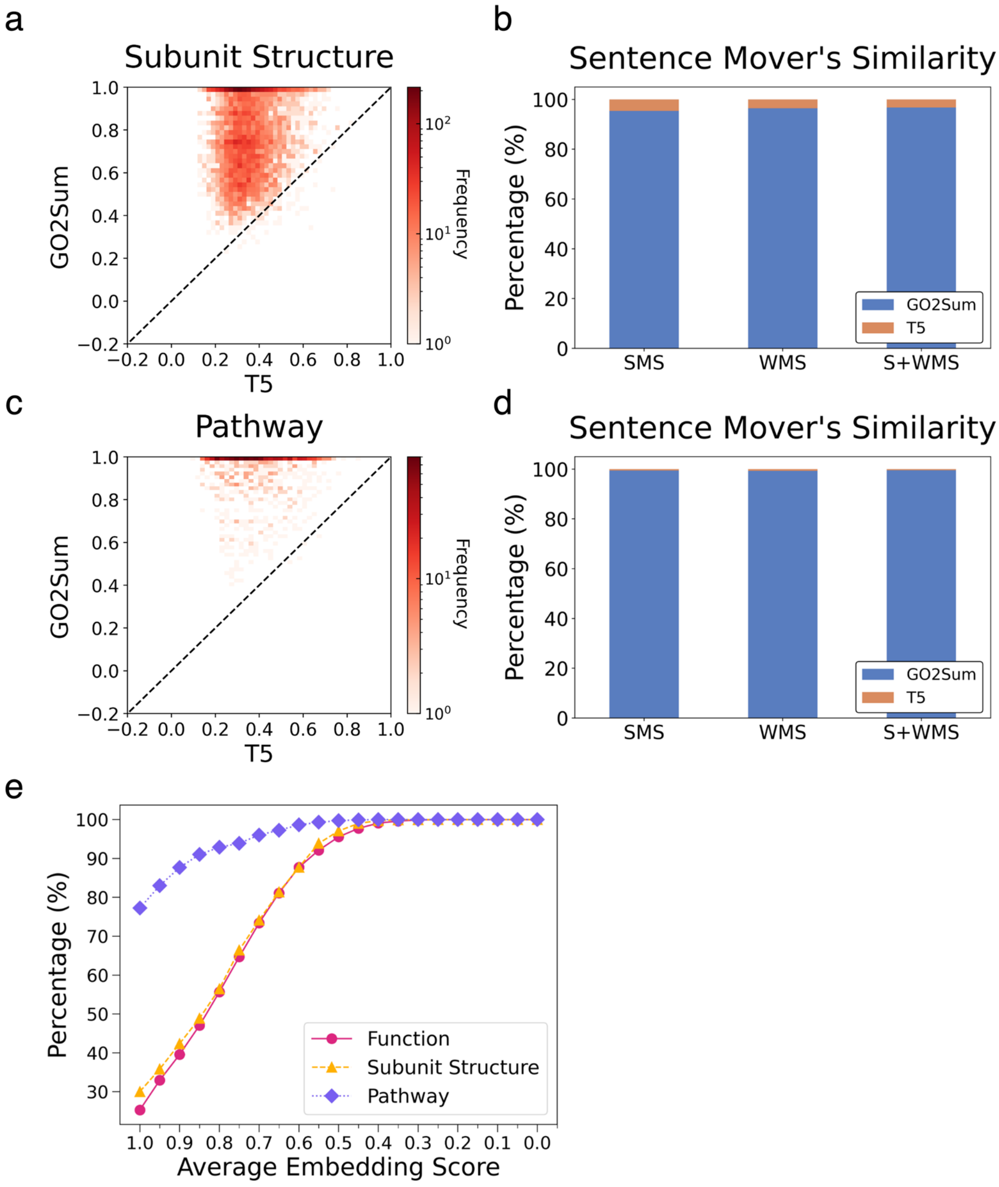

Fig. 3 | Comparison of GO2Sum and the vanilla T5 on the UniProt Subunit structure paragraphs and Pathway paragraphs. **a** The average embedding score for Subunit structure paragraphs. GO2Sum outperforms vanilla T5 for 99.2% of cases. **b** The Sentence mover's similarity for Subunit structure paragraphs. The number of wins for GO2Sum for SMS, WMS, and S + WMS is 95.3%, 96.4%, and 96.8%, respectively. **c** the average embedding score for Pathway paragraphs. GO2Sum outperforms vanilla T5 for 99.8% of cases. **d** The Sentence mover's similarity for Pathway paragraphs. The number of wins for GO2Sum for SMS, WMS, and S + WMS was 99.5%, 99.3%, and 99.5%, respectively. **e** The cumulative distribution of the average embedding score for function, subunit, and pathway paragraphs.

The next three examples, A3KFM7, P46500, and P05318 are cases where GO2Sum had average embedding scores between 0.4 to 0.6 and were lower than T5. Typical situations for this are that both T5 and GO2Sum captured the general idea of the function correctly, but T5's score was higher because it output a particular keyword. For example, A3KFM7 has a function summary that reads, "DNA-dependent ATPase that plays a role in chromatin remodeling. Regulates transcription by disrupting nucleosomes in a largely non-sliding manner which strongly increases the accessibility of chromatin. Activates transcription of specific genes in response to oxidative stress through interaction with NFE2L2." GO2Sum's summary for this entry was "Probable transcription regulator", which had a score of 0.479, while T5 output "The basal transcription machinery. A transcription factor. ATP hydrolysis." with a score of 0.583. In this case, T5's mention of ATP hydrolysis probably made the higher score than GO2Sum. Protein P46500, also known as Skn-1 dependent zygotic transcript 15 protein, has function "May be involved in embryogenesis". GO2Sum's summary stating it's "Required for zygotic survival" is accurate, supported by its annotation GO:0009792, which indicates its necessity for embryo development leading to birth or hatching. Despite this accuracy, GO2Sum scored lower (0.47) compared to T5, possibly due to the presence of the keyword "embryonic". Similarly, for protein P05318, T5 scored higher (0.61) than GO2Sum (0.58), possibly because T5 mentioned the keyword "ribosome." However, GO2Sum's summary, highlighting the protein's significance in the elongation step of protein synthesis, would be more informative.

For P46605 and P60154, GO2Sum's summaries obtained low scores of 0.37 and 0.19, respectively. In the case of P46605, GO2Sum's summary captures essentially correct information, although the ground truth provides more specific details, indicating its interaction with a particular DNA element. The next entry, Inactive ribonuclease-like protein 9 (P60154), is an interesting example, where GO2Sum has a low average embedding score of 0.19, indicating that the generated summary is dissimilar to the ground truth. For this entry, UniProt only added a short Function description that

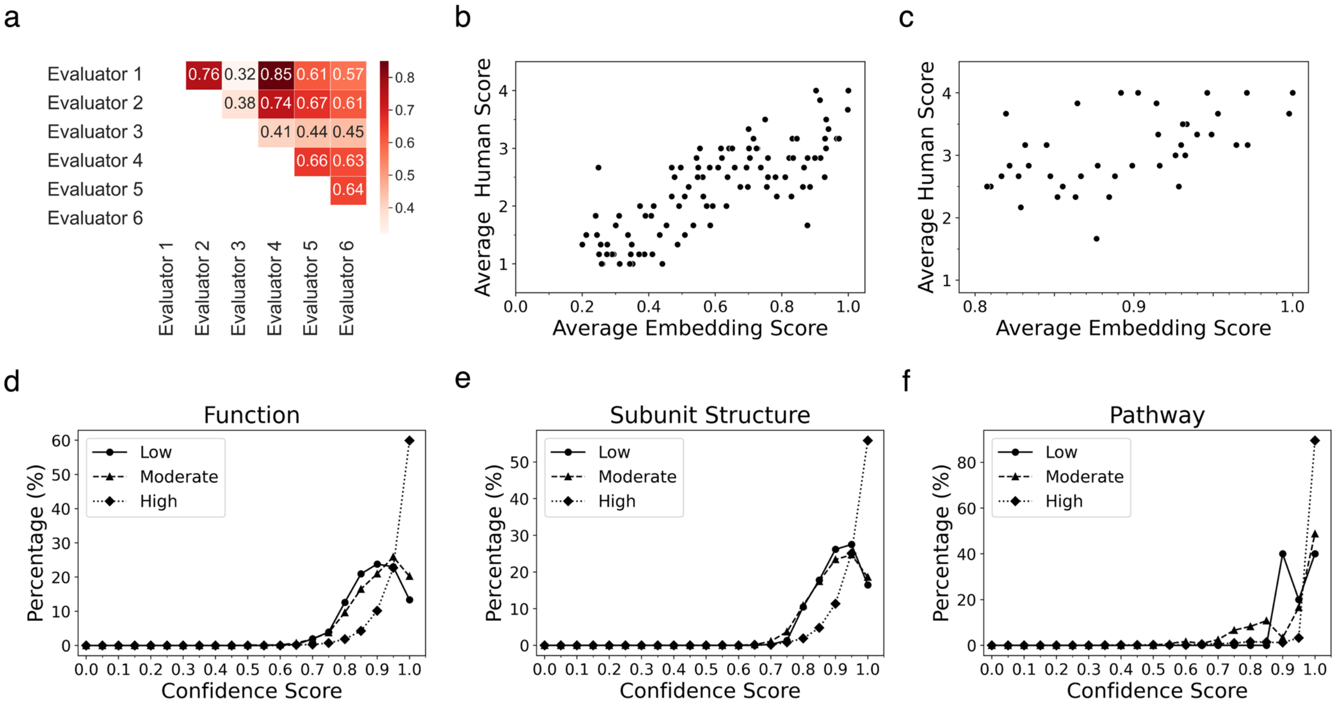

Fig. 4 | Expert human evaluation and the confidence score of paragraphs generated by GO2Sum. **a** Pearson correlation coefficient between pairs of human evaluators for the dataset of 100 evenly distributed entries. **b** Comparison between the average embedding score and the average of human evaluators score. The score correlation for the 100 proteins entry dataset has an average embedding score between 0.1 to 1.0. **c** the second dataset of 45 protein entries with a high average embedding score between 0.8 to 1.0. **d** Distribution of the confidence score for entries classified into three classes based on the average embedding score. High, [0.8, 1.0]; moderate, [0.5, 0.8); low, [0, 0.5). Function paragraphs. 4591, 4400, and 769 proteins were included in the high, moderate, and low score class, respectively. **e** Subunit Structure paragraphs. 3046, 2806, and 382 in high, moderate, and low. **f** Pathway paragraphs. 1329, 121, and 10 proteins in high, moderate, and low.

negates the ribonuclease activity despite the sequence similarity, which may not be comprehensive summary of the known functional activity of this protein. On the other hand, the GO document of this entry, i.e., the concatenated explanations of GO terms read "Binding to a nucleic acid. The space is external to the outermost structure of a cell. For cells without external protective or external encapsulating structures, this refers to space outside of the plasma membrane. This term covers the host cell environment outside an intracellular parasite. The process in which the controlled movement of a flagellated sperm cell is initiated as part of the process required for flagellated sperm to reach fertilization competence.". Thus, there are publications[47] that indicate the involvement of sperm, specifically sperm mobility, from which these GO terms are assigned. Therefore, the summary generated by GO2Sum would actually be a better summary for the mentioned document as it accurately captures the main concepts related to the function of this protein, which is involved in sperm motility and acts as a signaling molecule necessary for the proper movement of spermatozoa.

The last example, T-cell antigen CD7 (P09564) has a similar story as the previous example. UniProt describes its function as "Not yet known", which is obviously uninformative, despite the fact that this protein has seven GO term annotations (GO:0016020, GO:0005886, GO:0038023, GO:0002250, GO:0006955, GO:0042110, GO:0007169). The summary by GO2Sum has a low average embedding-based score of 0.16. However, considering the annotated GO terms and references[48] of this entry, we see that GO2Sum's summary, "Receptor for TNFSF13B/BLyS/BAFF and TNFSF13/APRIL. Promotes B-cell activation and differentiation." is accurate and more informative than UniProt Function description.

GO summary for UniProt subunit structure

Next, we discuss the summary for the Subunit structure section in UniProt. As shown in Supplementary Information 1, paragraphs of Subunit structure tend to be shorter compared to the Function paragraphs. The paragraphs contain information related to subunit structure, such as protein interactions (e.g., interaction with some protein), stoichiometry e.g., homotetramers and homotrimers. They often use specific protein names and IDs that can make it challenging for the models to produce it accurately.

Figure 3a illustrates the performance of GO2Sum in comparison to the vanilla T5, using the average of the three embedding-based scores. Similar to the findings in the Function section reported in Fig. 2, GO2Sum achieved a higher score than T5 in 99.2% of the entries. The results for the three individual scores (BERT, MiniLM, and BioSentVec) are detailed in Supplementary Figure 6. In comparison with SciFive, shown in Supplementary Fig. 5b, GO2Sum showed a higher average embedding score than SciFive for 98.89% of the entries. In Fig. 3b, we present an evaluation using mover distance-based similarity scores. GO2Sum's summaries were found to be closer to the ground truth in UniProt entries compared to the vanilla T5 for 96.4% using Word Mover's Similarity (WMS), 95.3% using SMS, and 96.8% using Sentence and WMS (S + WMS).

In Supplementary Table 1, we provided examples of summaries generated by GO2Sum and T5. In some cases, GO2Sum's summaries are generally in good agreement with the ground truth but lack specific protein names. This is attributed to the absence of such specific information in the GO descriptions.

GO summary for UniProt pathway

The last UniProt sections to examine is Pathway descriptions. As shown in Fig. 3c, d, paragraphs made by GO2Sum clearly have a higher average embedding score than the vanilla T5 (Fig. 3c) and also in terms of the SMS scores (Fig. 3d) for almost all the entries. In comparison with SciFive, GO2Sum surpassed SciFive in about 99.79% of the entries for the average embedding score (Supplementary Fig. 5c). Therefore, to summarize, consistently for all three UniProt sections, Function, Subunit, and Pathway, GO2Sum made more meaningful and correct paragraphs. Examples of GO2Sum's outputs are provided in Supplementary Table 2. Similar to the examples shown in Supplementary Table 1 for Subunit structure, GO2Sum

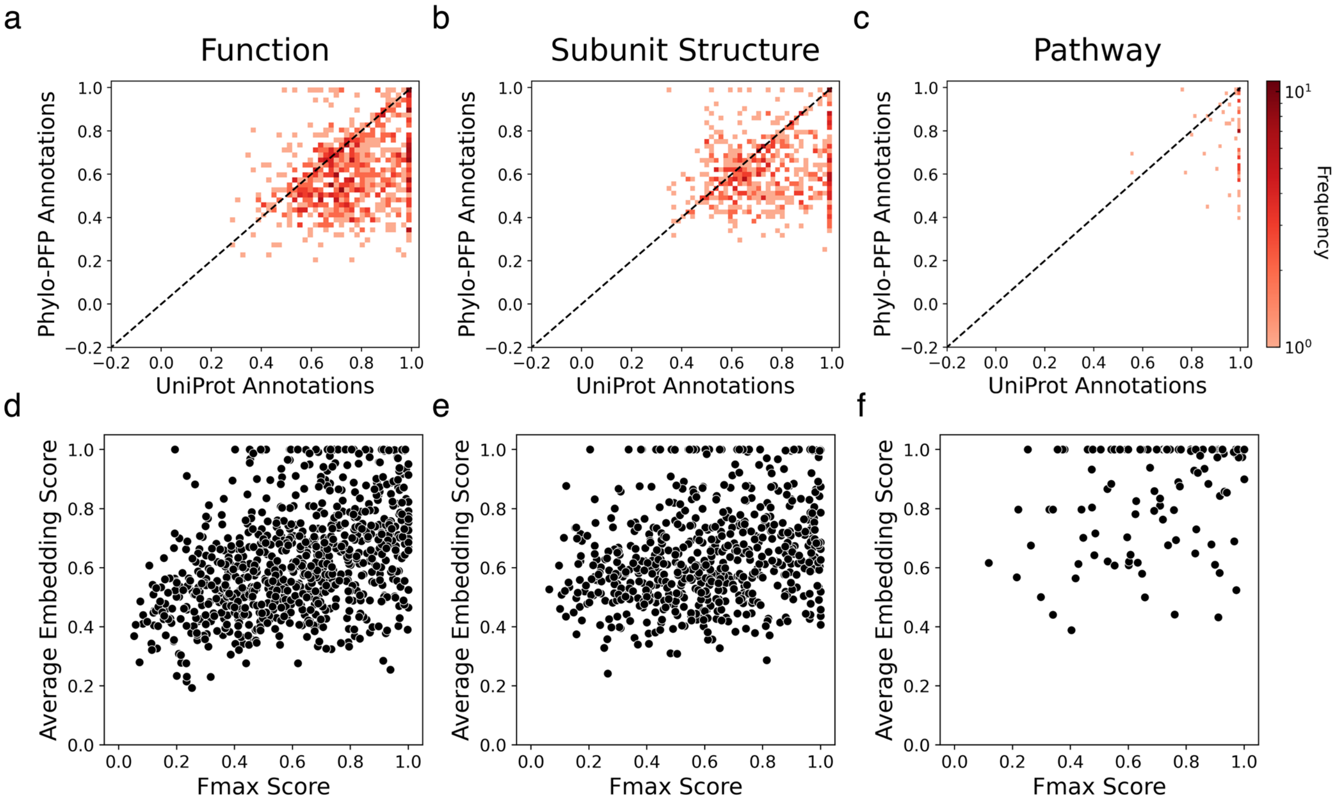

Fig. 5 | Average embedding scores of predicted GO terms. Predictions were produced by Phylo-PFP. 843, 116, and 632 proteins were used for function (**a**, **d**), subunit structure (**b**, **e**), and pathway (**c**, **f**) paragraphs, respectively. **a**–**c** The average embedding scores of summaries computed for predicted GO terms relative to the ground truth GO terms taken from UniProt. The color scale shows the number of cases at each point. **d**–**f** The average embedding score relative to *F*max score, which indicates the accuracy of predicted GO terms by Phylo-PFP.

performed fundamentally better than the vanilla T5. Often the average embedding score of GO2Sum's summary was low because it lacked specific information.

Average embedding score distribution for the three UniProt sections

In Fig. 3e, we present the distribution of average embedding scores for GO2Sum's outputs in the three UniProt sections to summarize its performance. Notably, the Pathway section exhibited a distinct score distribution, with 77.3% of entries achieving high scores between 1.0 and 0.95. In contrast, both Function and Subunit Structure summaries had only 25.3% and 30.0%, respectively, falling within the high score range. This divergence can be primarily attributed to paragraph length, with Pathway paragraphs being substantially shorter than those in the other two sections. On average, the word lengths for Pathway, Subunit, and Function summaries were 7.8, 28.4, and 56.1, respectively. Longer paragraphs, in general, tend to be more complex and challenging to reproduce accurately than shorter ones.

Human evaluation

To validate the embedding scores used in our work, we conducted a comparison with evaluations by human biology experts. We focused on assessing how the evaluations by human experts correlate with the embedding scores, particularly when the embedding score indicates high quality, i.e., favorable GO2Sum outputs. We engaged six human biology experts, one of whom holds a doctorate degree, three of whom hold master's degrees in biology, while the other two possess bachelor's degrees in biology or related fields such as biotechnology. The human evaluators assigned scores between 1 and 4, with 4 indicating the highest quality for each GO2Sum output.

We prepared two sets of protein targets for evaluation. The first set included 100 proteins, evenly distributed based on their average embedding scores. We randomly selected 10 proteins for each 0.1 interval, ranging from 0.1 to 0.2, 0.2 to 0.3, and so on. The second set comprised 20 randomly selected proteins with high average embedding scores between 0.8 and 1.0. We provided human experts with instructions, as detailed in Supplementary Information 2. In Fig. 4a, we present the Pearson correlation coefficients between human evaluators. Evaluator 3 exhibited a moderate correlation, ranging from 0.32 to 0.45, with other evaluators. On the other hand, the remaining five evaluators demonstrated strong agreement, with correlations ranging from 0.57 to 0.85.

In Fig. 4b, c, we assessed the agreement between the average embedding score we used and the evaluations by human experts. The *y* axis in Fig. 4b, c represent the average score assigned by the six human evaluators. As shown in Fig. 4a, the average embedding score demonstrates a strong correlation with the assessments made by human experts, with a Pearson correlation coefficient of 0.804. In Fig. 4b, we focused on 45 proteins with high average embedding scores ranging from 0.8 to 1.0. This selection included 25 high-scoring entries from dataset 1 and 20 proteins from dataset 2. Nearly all entries received an average human evaluator score of 2 or higher, with an overall average of 3.15. When considering entries with an even higher embedding score range of 0.9 to 1.0, the average human score rose to 3.52. Consequently, we can conclude that, on the whole, the average embedding score aligns sufficiently with evaluations by human biologists. Notably, high-scoring summaries generated by GO2Sum are highly reliable, as confirmed by human evaluators.

Although the embedding score agrees well with human evaluation, we observed two outliers where human evaluation differed from the embedding score. In the first case, Protein A3DNG9, a ribosomal RNA small subunit methyltransferase (Nep1), had a UniProt Function description: 'Methyltransferase involved in ribosomal biogenesis. Specifically catalyzes the N1-methylation of the pseudouridine corresponding to position 914 in *M. jannaschii* 16 S rRNA.' GO2Sum summarized the GO terms as: 'Methyltransferase involved in ribosomal biogenesis.

Specifically catalyzes the N1-methylation of pseudouridine at position 967 in 16 S rRNA. Is not able to methylate uridine at this position.' Despite the high average embedding score of 0.87, human evaluators assigned an average score of 1.66 because the last sentence contradicted the preceding sentences.

Another example, O74523 (Putative ATPase inhibitor, mitochondrial), presents an opposite case, where the average embedding score was low (0.49), yet human evaluation scored it high at 2.66. The GO2Sum's output, 'Thought to be a regulatory component of the mitochondrial ATP-synthesizing complex in the mitochondria,' lacked a term indicating that the protein 'inhibits.' However, aside from this omission, the overall function was accurate.

Confidence score for summary outputs by GO2Sum

As T5 generates a paragraph using beam search, each generated paragraph is associated with a probability value. These probability values enable us to compute a score that reflects the confidence or dominance of a paragraph compared to alternative paragraphs explored during the generation process. The confidence score of a top-scoring paragraph i is defined as follows:

$$Confidence\ Score(i) = \frac{Prob(i)}{\sum_j^N Prob(j)} \tag{1}$$

where N is the number of paragraphs generated at the last step of a beam search, which was set to 4 in this work.

In Fig. 4d–f, we present the distribution of confidence scores for UniProt summaries, categorized into three classes based on accuracy determined by average embedding scores. The high-scoring class included proteins with an average embedding score between 0.8 and 1.0, the moderate class ranged from 0.5 up to 0.8, and the low class included scores below 0.5. The plots reveal a clear trend: highly accurate summaries are typically associated with high confidence scores. This trend holds true for all three types of summary paragraphs (Function, Subunit, and Pathway). For instance, in the Function paragraphs, 82.6% of proteins in the high-scoring class have a confidence score exceeding 0.9. The proportion of highly confident summaries with a score over 0.9 decreases to 45.0% for the moderate class and 36.4% for the low-scoring class.

Applying GO2Sum for GO term predictions by Phylo-PFP

Finally, we applied GO2Sum to predicted GO terms obtained from Phyo-PFP[9], one of the highly accurate protein function prediction methods. Predicted GO terms may not perfectly represent the actual functions of proteins, posing a challenge to GO2Sum's summarization. To assess GO2Sum's utility in describing predicted protein functions, we conducted tests across a wide range of accuracy levels using a realistic scenario of GO predictions. For the test set of proteins, we selected common proteins shared between the test set used in this study and the test set from the work of ContactPFP[15], encompassing a total of 9642 proteins. Among these, 843, 116, and 632 proteins were shared for function, pathway, and subunit structure paragraphs, respectively.

In Fig. 5a–c, the average embedding score of summaries generated from predicted GO terms was plotted relative to results from the ground truth GO terms taken from UniProt. Naturally, summaries generated from predicted GO terms have a lower score for most of the cases (82.7% for Function; 77.1% for Subunit, and 94.8% for Pathway). However, it is noteworthy that the majority of summaries generated from predicted GO terms still maintain a moderate average embedding score of 0.5 or higher, the level that is practically accurate and useful for users. Specifically, for Function paragraphs, 73.7% achieved a score of 0.5 or higher, while for Subunit paragraphs, the figure was 79.9%, and for Pathway paragraphs, it was 95.7%.

Figure 5d–f depict the correlation between the average embedding score and the function prediction accuracy, as measured by the Fmax score[49]. Fmax ranges from 0 to 1, with 1 for the perfect agreement with the ground truth GO terms. The average embedding score exhibits moderate to weak correlation with the Fmax score of GO terms, with the Pearson correlation coefficient of 0.44 for Function, 0.28 for Subunit, and 0.28 for Pathway. Of greater significance is the observation that even relatively low-scoring GO predictions, with an Fmax score ~0.2, can achieve a moderate average embedding score of 0.5. We manually examined such cases in details and found that those typically occur because predicted GO terms include one or more key informative GO terms.

Protein P04141 (Granulocyte-macrophage colony-stimulating factor), which has UniProt Function annotation of "Cytokine that stimulates the growth and differentiation of hematopoietic precursor cells from various lineages, including granulocytes, macrophages, eosinophils, and erythrocytes", is such an example. This protein is annotated with 36 GO terms in UniProt, among which only 6 GO terms, GO:0006955, GO:0002376, GO:0005615, GO:0008083, GO:0005125, and GO:0005129, were confidently predicted by Phylo-PFP. The Fmax score of the GO prediction was 0.2, due to the small coverage of correct GO terms. However, GO2Sum managed to generate a function summary of a high average embedding score, 0.91. This was possible because one of the predicted GO terms, GO:0005125 (cytokine activity) was informative, which describes the activity of a soluble extracellular gene product interacting with a receptor to control the survival, growth, differentiation, and effector function of tissues and cells. GO:0005129 (granulocyte-macrophage colony-stimulating factor receptor binding) was also helpful, which says binding to a granulocyte-macrophage colony-stimulating factor receptor. With these GO terms' information, GO2Sum produced the function description that is close to UniProt: "Cytokine that stimulates the growth and differentiation of granulocytes, macrophages, eosinophils, and erythrocytes.

A similar case occurred in Subunit Structure and Pathway summaries. For Protein P46988 (Prefoldin subunit 1), the Fmax score was low, 0.20, because only three out of seven GO terms were correctly predicted. But Subunit Structure summary, "Heterohexamer of two PFD-alpha type and four PFD-beta type subunits" was generated correctly by GO2Sum with an average embedding score of 1.0, mainly thanks to the information provided by the GO term GO:0016272 (prefoldin complex). This GO term contained the details needed to predict the specific information about PFD-alpha type and four PFD-beta type subunits. The other two correct GO terms, GO:0006457 and GO:0051082, probably also provided the functional context that goes well with GO:0016272. For Pathway summary generation, Protein Q9Y252 (E3 ubiquitin-protein ligase RNF6) is an illustrative example. A low coverage (five out of twenty-two GO terms) gave a low Fmax score of 0.35. Despite that, pathway summary was correctly generated as "Protein modification; protein ubiquitination", with an average embedding score of 1.0, probably by contributions from GO terms, GO:0044314 (protein K48-linked ubiquitination), GO:0085020 (protein ubiquitination), and GO:0045893 (protein K27-linked ubiquitination).

The results indicate that GO2Sum can provide readable and useful summaries even for low-scoring GO predictions. According to CAFA challenges[22,50] recent prediction methods achieve a Fmax score of about 0.5 to 0.6 on average, which is sufficient for GO2Sum to produce meaningful summary. For example, when the Fmax score is 0.5 or higher, 81.9% of proteins have a summary with an average embedding score of 0.5 or higher in Function paragraphs. Similarly, for Subunit Structure and Pathway, the figures are 82.8% and 96.7%, respectively.

Discussion

We have developed GO2Sum, a fine-tuned variant of the T5 model designed specifically for summarizing the descriptions of GO terms into Function, Subunit, and Pathway paragraphs in UniProt entries. The summaries generated by GO2Sum demonstrated a substantially higher level of agreement with UniProt data when compared to the results obtained using the standard T5 model. Our evaluation of the agreement with the ground truth UniProt paragraphs primarily relied on embedding-based scoring methods, which exhibited a strong correlation with assessments made by human evaluators.

While GO2Sum generally produces accurate summaries, a notable limitation is its inability to mention specific protein names. This limitation arises from the fact that GO descriptions typically lack such specific details. In contrast, UniProt's Subunit paragraphs often contain specific protein names. To address this limitation and enable the model to provide specific information, it would require access to additional information sources, such as protein-protein interaction data. Integrating GO terms with these external sources represents an intriguing avenue for future research. Another limitation of the current GO2Sum model is that it has a maximum input token size of 1024 due to the available computational resources. Applying pruning of GO terms or text compression can be a possible solution to this problem. A further potential future improvement is enhancing the model's ability to discern which part of a summary corresponds to each GO term description, possibly by implementing a Question and Answering framework[51,52].

For many years, protein function prediction has relied on GO terms to describe biological functions. However, recent advancements in language models have allowed us to translate these predictions into human-readable text, providing a more intuitive and user-friendly approach to describing predicted functions. We believe that this development contributes significantly to enhancing the interaction between humans and machine learning models, particularly in the context of assisting biologists and medical scientists in their daily research endeavors.

Methods

Network architecture

Text-to-Text Transfer Transformer, or T5[26] has been used as the neural network architecture in GO2Sum. The T5 model is based on standard transformer layers, which use self-attention to understand long-range dependencies in text. Furthermore, T5 follows an encoder-decoder architecture, where the encoder is assigned to extract a rich latent representation of the text, irrespective of the task, and the decoder aims at generating task-specific output based on the encoder embeddings. T5 is a competent model for several text-specific tasks including summarization, translation, question answering, etc. GO2Sum uses the baseline 220 million parameter version of the T5 model, i.e., T5-base. Both the encoder and decoder stages consist of 12 blocks each, where each block comprises a self-attention layer, a feed-forward layer, and an optional cross-attention layer. All the attention layers use 12 attention heads, and the dimensionality of keys, queries, and values is 64. The embedding dimensions of all the layers are fixed to 768. However, the feed-forward layers use an inner dimension of 3072 for increasing expressivity. Additionally, a dropout of 0.1 is used for regularization purposes.

Network training

The models were initialized with the pre-trained T5 weights and were fine-tuned for 100 epochs. The standard cross-entropy was used as the loss function, and the Adam optimizer was used with a learning rate of 0.0001. The models were trained on batch sizes of 4, and an early stopping criterion was used to prevent overfitting. Due to computational limitations, we limited the input token length to 1024, and the target maximum token length was kept at 256.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data used to train GO2Sum is publicly available at SwissProt https://www.uniprot.org/uniprotkb?query=reviewed:true. The trained GO2Sum models can be downloaded from https://kiharalab.org/GO2Sum/ and https://doi.org/10.5281/zenodo.10719085. All other data can be obtained from the corresponding author upon reasonable request.

Code availability

The source code of GO2Sum is available at https://github.com/kiharalab/GO2Sum.

Received: 14 December 2023; Accepted: 5 March 2024;
Published online: 15 March 2024

References

1. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
2. The UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
3. The Gene Ontology Consortium Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
4. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
5. Wei, Q., Khan, I. K., Ding, Z., Yerneni, S. & Kihara, D. NaviGO: Interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinformatics* **18**, 177 (2017).
6. Chitale, M.& Kihara, D. Computational protein function prediction: framework and challenges. *In:* Protein function prediction for omics era (ed. Kihara, D.) 1–17 (Springer, 2011).
7. McGinnis, S. L. & Madden, T. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).
8. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
9. Jain, A. & Kihara, D. Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences. *Bioinformatics* **35**, 753–759 (2019).
10. Hawkins, T., Chitale, M., Luban, S. & Kihara, D. PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins.* **74**, 566–582 (2009).
11. Chitale, M., Hawkins, T., Park, C. & Kihara, D. ESG: extended similarity group method for automated protein function prediction. *Bioinformatics* **25**, 1739–1745 (2009).
12. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2019).
13. You, R., Yao, S., Mamitsuka, H. & Zhu, S. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics* **37**, i262–i271 (2021).
14. Kulmanov, M. & Hoehndorf, R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics* **38**, i238–i245 (2022).
15. Kagaya, Y., Flannery, S. T., Jain, A. & Kihara, D. ContactPFP: protein function prediction using predicted contact information. *Front. Bioinformatics* **2**, 896295 (2022).
16. Jaeger, S., Gaudan, S., Leser, U. & Rebholz-Schuhmann, D. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics* **9**, S2 (2008).
17. Wong, A. & Shatkay, H. Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge. *BMC Bioinformatics* **14**, S14 (2013).
18. You, R., Huang, X. & Zhu, S. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. *Methods* **145**, 82–90 (2018).
19. Yao, S. et al. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res.* **49**, W469–W475 (2021).
20. Giri, S. J., Dutta, P., Halani, P. & Saha, S. MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information. *IEEE J. Biomed. Health Inform.* **25**, 1832–1838 (2021).
21. You, R. et al. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* **34**, 2465–2473 (2018).

22. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).
23. Khan, I. K., Wei, Q., Chitale, M. & Kihara, D. PFP/ESG: automated protein function prediction servers enhanced with gene ontology visualization tool. *Bioinformatics* **31**, 271–272 (2015).
24. Otter, D. W., Medina, J. R. & Kalita, J. K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 604–624 (2021).
25. Zhao, W. X. et al. A survey of large language models. *arXiv* https://arxiv.org/abs/2303.18223 (2023).
26. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
27. Bateman, A. et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
28. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *Proc. 2019 Conf. North Am. Chapter Assoc. Computat. Linguist. Hum. Lang. Technol.* **1**, 4171–4186 (2019).
29. Beltagy, I., Lo, K. & Cohan, A. SciBERT: a pretrained language model for scientific text. *In:* Proceedings of the 2019 Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (eds. Inui, K., Jiang, J., Ng, V. & Wan, X.) 3615–3620 (Association for Computational Linguistics, 2019).
30. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
31. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *In:* Proceedings of the 18th BioNLP Workshop and Shared Task (ed. Demner-Fushman, D.) 58–65 (Association for Computational Linguistics, 2019).
32. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2022).
33. Long P. N. & Gregoire A. B. SciFive: a text-to-text transformer model for biomedical literature. *arXiv* (2023).
34. Xie, Q., Bishop, J. A., Tiwari, P. & Ananiadou, S. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowl.-Based Syst.* **252**, 109460 (2022).
35. Lewis, M. et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *In:* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (eds. Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) 7871–7880 (Association for Computational Linguistics, 2020).
36. Du, Y., Li, Q., Wang, L. & He, Y. Biomedical-domain pre-trained language model for extractive summarization. *Knowl.-Based Syst.* **199**, 105964 (2020).
37. Nallapati, R. et al. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. 280–290 (Association for Computational Linguistics, 2016).
38. Dong, L. et al. Unified language model pre-training for natural language understanding and generation. *In:* Advances in neural information processing systems (ed. Wallach, H.) 32 (Curran Associates, Inc., 2019)
39. Wenhui W. et al. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers in proceedings of the 34th international conference on neural information processing systems (ed. Bertlett, P.) 13 (Curran Associates Inc., 2020).
40. Chen, Q., Peng, Y. & Lu, Z. BioSentVec: creating sentence embeddings for biomedical texts. *In:* 2019 IEEE International Conference on Healthcare Informatics (ICHI) 1–5 (IEEE, 2019).
41. Clark, E., Celikyilmaz, A., Smith, N. A. & Allen, P. G. Sentence Mover's similarity: automatic evaluation for multi-sentence texts. *In:* Proceedings of the 57th annual meeting of the association for computational linguistics (ed. Korhonen, A. & Traum, D.) 2748–2760 (Association for Computational Linguistics, 2019).
42. Jain, A. et al. Analyzing effect of quadruple multiple sequence alignments on deep learning based protein inter-residue distance prediction. *Sci. Rep.* **11**, 7574 (2021).
43. Zhang, J., Zhao, Y., Saleh, M. & Liu, P. J. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization *In*: Proceedings of the 37th international conference on machine learning (ICML) 12 (JMLR.org, 2020).
44. Shah, D. J., Yu, L., Lei, T. & Barzilay, R. Nutri-bullets hybrid: consensual multi-document summarization. *In:* proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL-HLT 2021) 5213–5222 (Association for Computational Linguistics, 2021).
45. Lin, C. Y. ROUGE: a package for automatic evaluation of summaries. *In:* text summarization branches out 74–81 (Association for Computational Linguistics, 2004).
46. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. BLEU: a method for automatic evaluation of machine translation in proceedings of the 40th annual meeting on association for computational linguistics 311–318 (Association for Computational Linguistics, 2002).
47. Westmuckett, A. D. et al. Impaired sperm maturation in Rnase9 knockout mice. *Biol. Reprod.* **90**, 1–10 (2014).
48. Stillwell, R. & Bierer, B. E. T cell signal transduction and the role of CD7 in costimulation. *Immunol. Res.* **24**, 31–52 (2001).
49. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
50. Jiang, Y. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* **17**, 184 (2016).
51. Zhang, R., Guo, J., Chen, L., Fan, Y. & Cheng, X. A review on question generation from natural language text. *ACM Trans. Inf. Syst.* **40**, 1–43 (2022).
52. Gupta, D., Kumari, S., Ekbal, A. & Bhattacharyya, P. MMQA: a multi-domain multi-lingual question-answering framework for english and hindi. *In:* Proceedings of the eleventh international conference on language resources and evaluation (European Language Resources Association, 2018).

Acknowledgements

This work was partly supported by the National Science Foundation (DBI2003635, DBI2146026, IIS2211598, DMS2151678, CMMI1825941, and MCB1925643) and by the National Institutes of Health (R01GM133840, 3R01GM133840-02S1).

Author contributions

D.K. conceived the study. S.G. designed and implemented the GO2Sum pipeline. S.G. performed the computation. S.G. and D.K. analyzed the data. N.I. participated in model interpretation and confidence score computation. S.G. drafted the manuscript, and D.K. edited it. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00358-0.

Correspondence and requests for materials should be addressed to Daisuke Kihara.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.