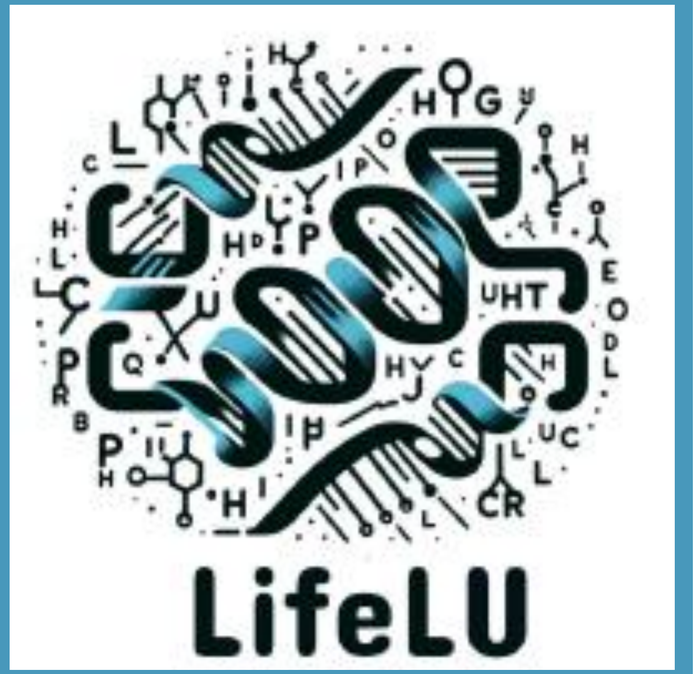
LifeLU - UNDERSTANDING THE LANGUAGE OF LIFE

READING GROUP



Presenter: Özlem Şimşek

Paper: PANDA-3D: protein function prediction based on AlphaFold models

Link: https://academic.oup.com/nargab/article/6/3/lqae094/7728018

Date: 10/10/2024

PANDA-3D:

protein function prediction based on AlphaFold models

Zhao, C., Liu, T., & Wang, Z. (2024). PANDA-3D: protein function prediction based on AlphaFold models. NAR Genomics and Bioinformatics, 6.

NAR Genomics and Bioinformatics, 2024, **6**, lqae094
https://doi.org/10.1093/nargab/lqae094
Advance access publication date: 6 August 2024
Pipelines and Workflows for Biological Data Analysis

OXFORD

PANDA-3D: protein function prediction based on AlphaFold models

Chenguang Zhao [1], Tong Liu [2] and Zheng Wang [2,*]

[1]Computer and Information Sciences Department, St. Ambrose University, 518 W Locust St, Davenport, IA 52803, USA
[2]Department of Computer Science, University of Miami, 1365 Memorial Drive, Coral Gables, FL 33124, USA
[*]To whom correspondence should be addressed. Tel: +1 305 284 3642; Email: zheng.wang@miami.edu

Abstract

Previous protein function predictors primarily make predictions from amino acid sequences instead of tertiary structures because of the limited number of experimentally determined structures and the unsatisfying qualities of predicted structures. AlphaFold recently achieved promising performances when predicting protein tertiary structures, and the AlphaFold protein structure database (AlphaFold DB) is fast-expanding. Therefore, we aimed to develop a deep-learning tool that is specifically trained with AlphaFold models and predict GO terms from AlphaFold models. We developed an advanced learning architecture by combining geometric vector perceptron graph neural networks and variant transformer decoder layers for multi-label classification. PANDA-3D predicts gene ontology (GO) terms from the predicted structures of AlphaFold and the embeddings of amino acid sequences based on a large language model. Our method significantly outperformed a state-of-the-art deep-learning method that was trained with experimentally determined tertiary structures, and either outperformed or was comparable with several other language-model-based state-of-the-art methods with amino acid sequences as input. PANDA-3D is tailored to AlphaFold models, and the AlphaFold DB currently contains over 200 million predicted protein structures (as of May 1st, 2023), making PANDA-3D a useful tool that can accurately annotate the functions of a large number of proteins. PANDA-3D can be freely accessed as a web server from http://dna.cs.miami.edu/PANDA-3D/ and as a repository from https://github.com/zwang-bioinformatics/PANDA-3D.

Introduction

Proteins, the essential functional units of life, play crucial roles in catalyzing biochemical reactions (1), providing structural support for anaphase spindle (2), and regulating gene expressions (3,4). Accurately annotating protein functions is important for understanding biological processes and discovering novel drug targets (5,6). However, experimentally determining the functions of proteins is both laborious and expensive (7), whereas machine learning approaches can decrease the time and cost required for this task making accurate and comprehensive annotations possible and offering a promising avenue for protein function prediction.

Amino acid sequences determine protein structures (8,9), and protein structures determine the function of proteins (10). Therefore, proteins that share similar sequences may have similar functions. For a considerable length of time, protein function predictors focus on using machine learning methods to leverage the sequence alignment, as revealed by the critical assessment of functional annotation (CAFA) challenge. The keyword analyses of the top ten methods in CAFA2 (11) and of all participating methods in CAFA3 (7) show that sequence alignment and machine learning are the two most frequently used approaches. Our previous tool PANDA (12) uses profile-profile alignments and PSI-BLAST (13) to find similar proteins, detects reserved protein domains, executes a Bayesian model to infer GO terms from domain architectures, and then combines the candidate GO terms from these approaches to make final predictions. DeepGOPlus (14) uses a one-dimensional (1D) convolutional neural network (CNN) to predict the protein functions from amino acid sequences. GODoc (15) applies a *k*-nearest-neighbor algorithm over sequence information, such as amino acid-coupling pattern representations, to predict protein functions. DEEPred (16) makes predictions by feeding the sequence features, such as subsequence profile map and pseudo amino acid composition, into stacked feed-forward deep neural networks (DNN) followed by a hierarchical post-processing method. ProLanGO (17) used a recurrent neural network (RNN)-based machine translation model to predict protein functions from protein sequences.

In addition to extracting knowledge from protein sequence alignment, some newer methods leverage that by using protein language models in the last few years. Our PANDA2 (18) utilizes a graph neural network (GNN) to model the GO-directed acyclic graph (GO-DAG) topology and incorporates features generated by the protein large language model (LLM) (19). UDSMProt (20) uses a self-supervised RNN to learn task-agnostic representations of sequences, which is then fine-tuned on the downstream task of protein function prediction. Littmann *et al.* (21) found that predicting GO terms based on proximity of embeddings from language models SeqVec (22) or ProtBert (23) outperformed naïve sequence-based transfer. Rives *et al.* (19) trained a deep transformer (24) on about 250 million sequences, and the embedding generated by this evolutionary-scale language modeling (ESM) contains information on protein structures, functions, and binding information, which outperformed others in a variety of downstream tasks (19). This pre-trained language model was used by many state-of-the-art methods, such as ATGO (25), DeepGO-SE(26), SPROF-GO (27) and NETGO3 (28).

Received: November 1, 2023. **Revised:** July 9, 2024. **Editorial Decision:** July 22, 2024. **Accepted:** July 23, 2024

SUMMARY

Aim: predict GO terms from AlphaFold models (uses tertiary structures)

Problem type: multi-label classification

Method: combination of graph neural networks and transformer decoder layers

GLOSSARY

Protein Structures

1. Primary str: The sequence of amino acids in a polypeptide* chain.

2. Secondary str: The local folded structures that form within a polypeptide due to interactions between atoms of the backbone***. The most common types of secondary structures are the α helix and the β pleated sheet.

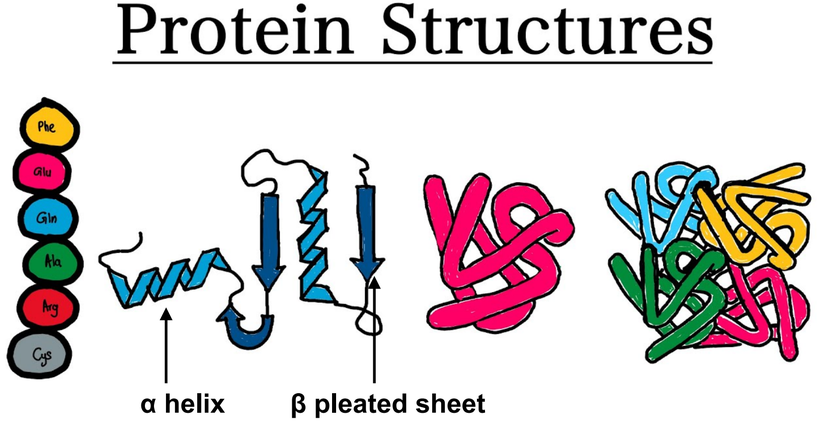
3. Tertiary str: The overall 3D structure of a polypeptide. Due to interactions between atoms of the side chains (R group).

4. Quaternary str: The 3D structure of multiple polypeptides.

***** Many proteins are made up of a single polypeptide chain and have only three levels of structure.

* Some proteins are made up of multiple polypeptide chains, also known as subunits.

Denaturation: When a protein loses its higher-order structure (due to the temperature or pH of a protein's environment is changed, or if it is exposed to chemicals), but not its primary sequence, it is said to be denatured. Denatured proteins are usually non-functional.



* A linear organic polymer consisting of a large number of amino-acid residues** bonded (with peptide bonds) together in a chain, forming part of (or the whole of) a protein molecule.

** When two or more amino acids combine to form a peptide, the elements of water are removed, and what remains of each amino acid is called an amino-acid residue.

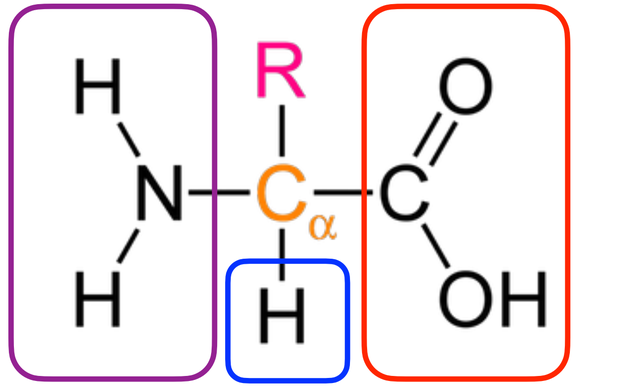
*** The backbone refers to the polypeptide chain apart from the R groups (explained in the next slide)

GLOSSARY

The structure of an amino acid

An amino acid is made up around an αCarbon. The central αCarbon is bonded with covalent bonds to 4 groups of molecules:

1. amino group (the amino)
2. carbonyl group (the acid)
3. hydrogen (H)
4. R group (causes the variation of amino acids, defines the shape of the protein)



A simple presentation: https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/v/introduction-to-amino-acids

GLOSSARY

Gene Ontology (GO)

Definition: The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms. It is designed for supporting the computational representation of biological systems. *

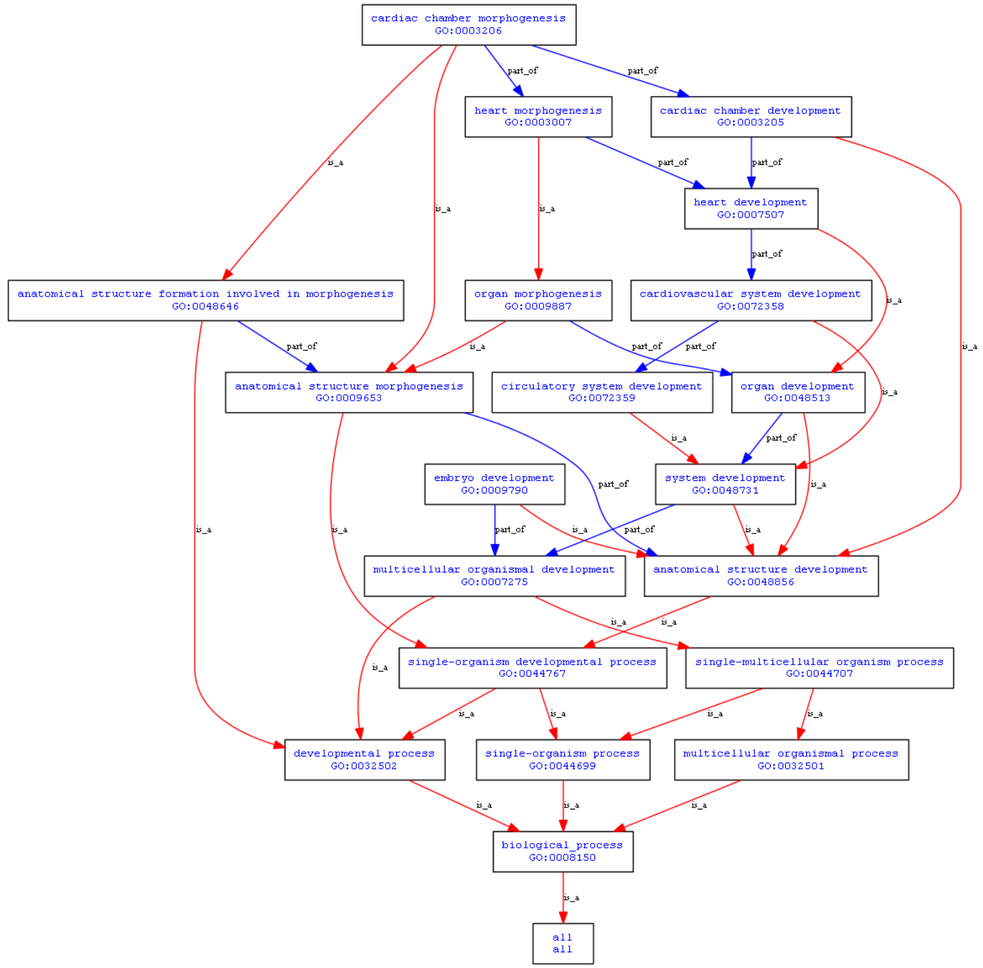
Each concept in the GO relates to the activity of a gene product or complex, as they carry out cellular processes.

Gene product/complex: A macromolecule produced according to the instructions from a gene. Can be of two types, a protein (the most common type) or a non-coding RNA. Gene products from different genes can combine into a larger molecular machine, called a macromolecular complex.

The GO defines the “universe” of possible functions a gene might have, but it makes no claims about the function of any particular gene.

GO annotation: A statement about the function of a particular gene.

A gene encodes a gene product, and that gene product carries out a molecular-level process or activity (molecular function) in a specific location relative to the cell (cellular component), and this molecular process contributes to a larger biological objective (biological process) comprised of multiple molecular-level processes.



* Thomas, P.D. (2017). The Gene Ontology and the Meaning of Biological Function. *Methods in molecular biology, 1446*, 15-24 .

** The GO Resource: https://geneontology.org/

SUMMARY

Aim: predict GO terms from AlphaFold models

Problem type: multi-label classification

Method: combination of graph neural networks and transformer decoder layers

DATASET

- Swiss-Prot
  - manually-reviewed protein sequences
  - experimentally determined protein functions in the format of GO terms
- AlphaFold DB
  - the predicted protein tertiary structures
- All three ontologies of GO terms
  - molecular function ontology (MFO)
  - biological process ontology (BPO)
  - cellular component ontology (CCO)

Total: **68,523 proteins** with both GO terms and AlphaFold models

- %80 train, %10 validation, %10 test

Target classes: the GO terms that had been annotated with at least 50 proteins in the training dataset

METHOD

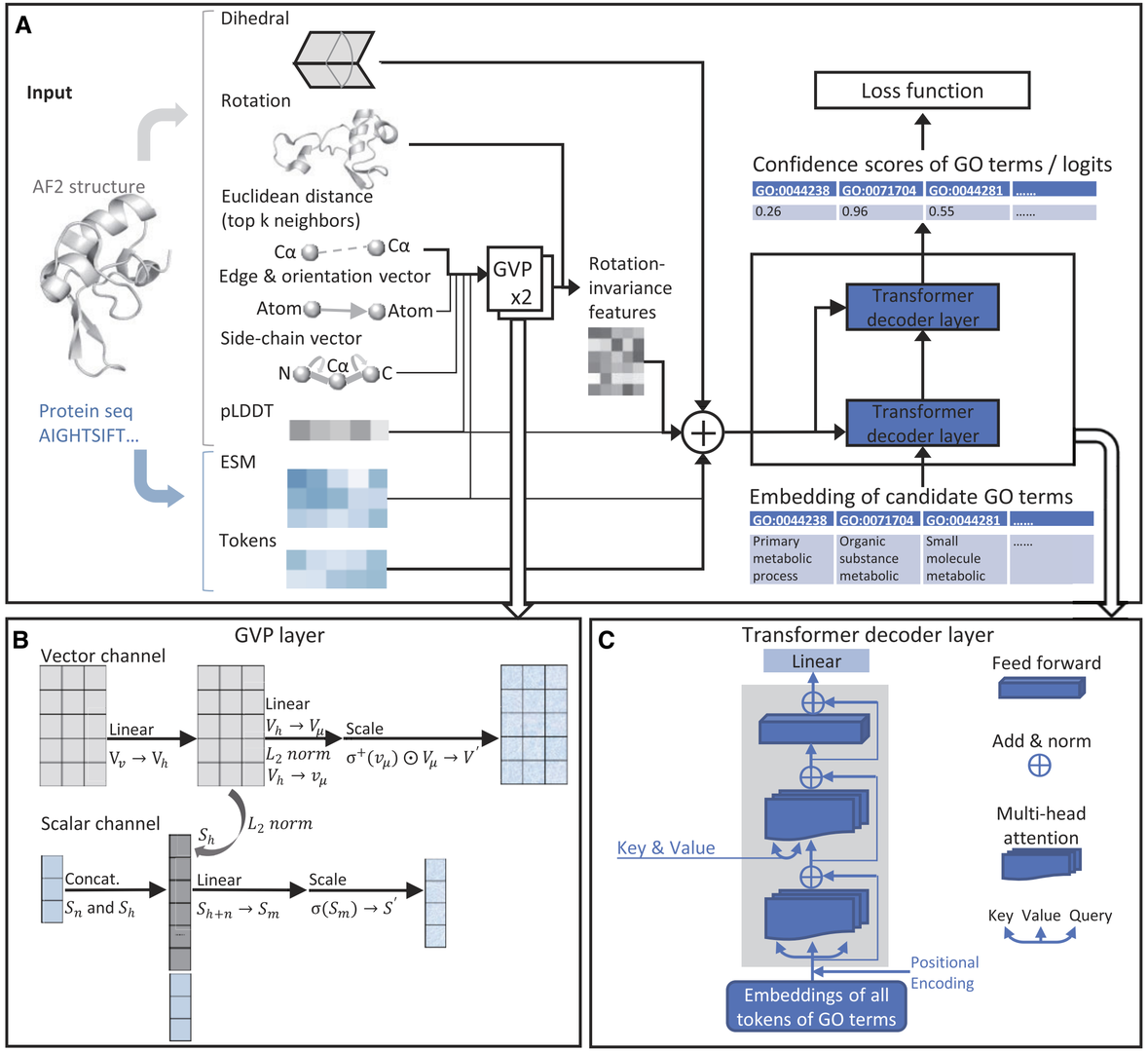


Figure 1. The overall architecture of PANDA-3D. Panel **A** shows that the GVP-GNN are used to extract information from predicted structures and protein sequence embeddings, succeeded by a variant transformer decoder producing confidence scores over query GO terms. Panel **B** shows the scalar and vector channels of a GVP layer. Panel **C** shows the architecture of a transformer decoder layers for multi-label classification.

METHOD

Features:

- Embedding dimension is 128, fixed for ease of calculations
- Structure-based features
  - Euclidean distances (between Cα) of the top k neighbors for each residue
  - Edge and orientation vectors
  - Side-chain vectors
  - pLDDT accuracy (measures confidence in the local structure)
- Structural-rotation and dihedral features
  - Rotation (due to the local reference frame)
  - Dihedral angles (the internal angle of polypeptide backbone)
- Sequential features
  - ESM of the amino acid sequence of the query protein
  - Tokens

* the predicted local-distance difference test

** the evolutionary-scale language modeling

METHOD

Model:

- GVP (Geometric Vector Perceptrons-proposed in ICLR 2021) layers:
  - a GVP layer has scalar and vector channels
  - Vector channel output V′
  - Scalar channel output S′ (also considers vector input)

- Transformer layers:
  - A self-attention layer: Feed all embedding of GO terms here to learn the relationships or co-occurrence patterns among them
  - A cross-attention layer: Combine the learnt GO relations with the encoder output
  - A feed-forward layer

METHOD

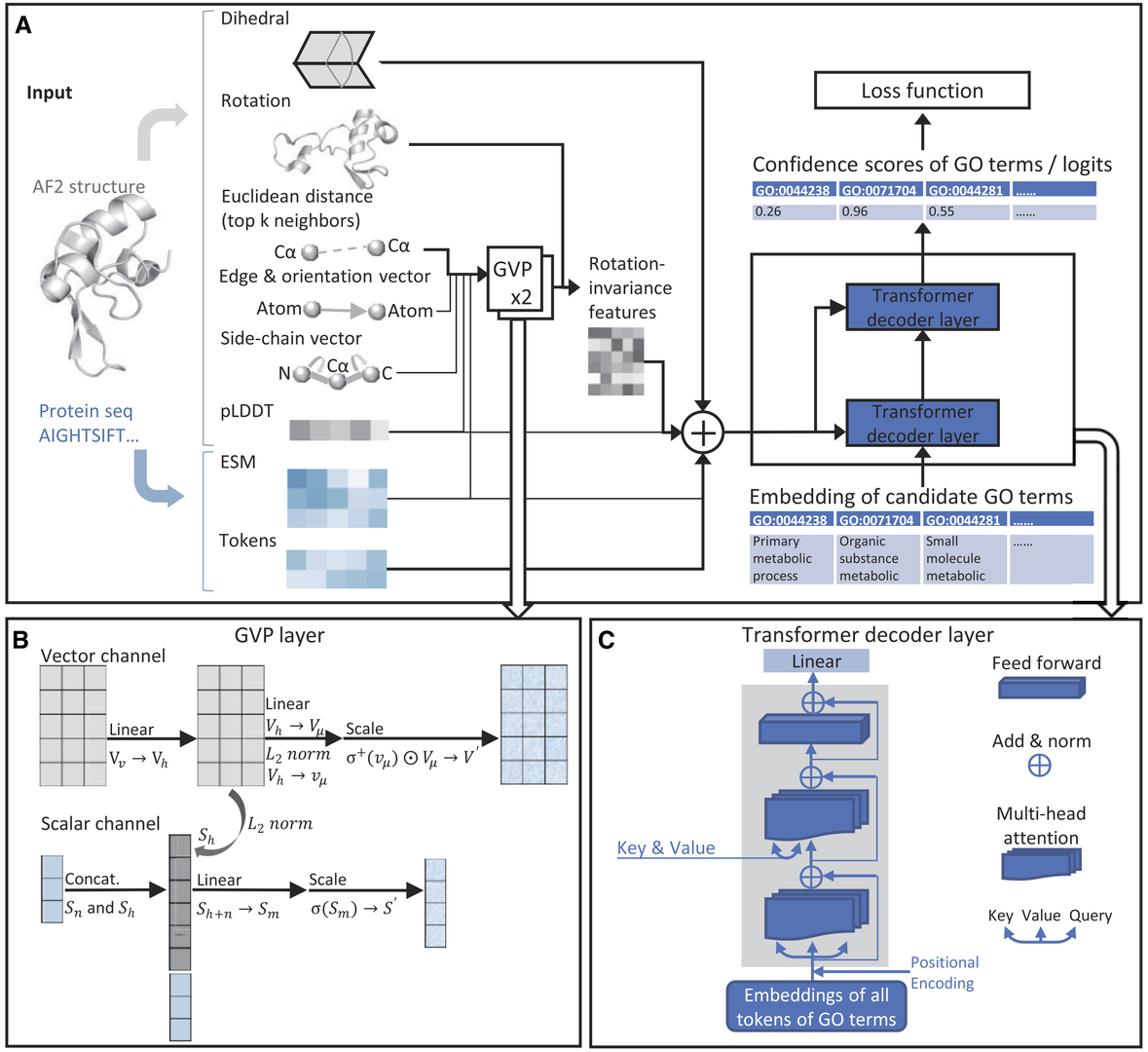


Figure 1. The overall architecture of PANDA-3D. Panel **A** shows that the GVP-GNN are used to extract information from predicted structures and protein sequence embeddings, succeeded by a variant transformer decoder producing confidence scores over query GO terms. Panel **B** shows the scalar and vector channels of a GVP layer. Panel **C** shows the architecture of a transformer decoder layers for multi-label classification.

COMPARED METHODS

- Baselines
  - **Naïve**: predicts GO terms based on the relative frequency of each GO term in the Swiss-Prot db
  - **BLAST**: predicts GO terms by transferring the experimental GO terms of similar sequences found in the training dataset

- **DeepFRI**: predict protein functions from tertiary structures and sequences
- **DeepGOCNN**: predict protein functions from amino acid sequences

- **DeepGO-SE**
- **SPROF-GO**

PERFORMANCE METRICS

- protein-centric evaluations
  - maximum F-measure (F_{max})
  - minimum semantic distance (S_{min})
  - area under the precision-recall curve (AUPR)

- GO-centric evaluations
  - the area under the ROC* curve (AUROC)

* true-positive rate vs false-positive rate

RESULTS

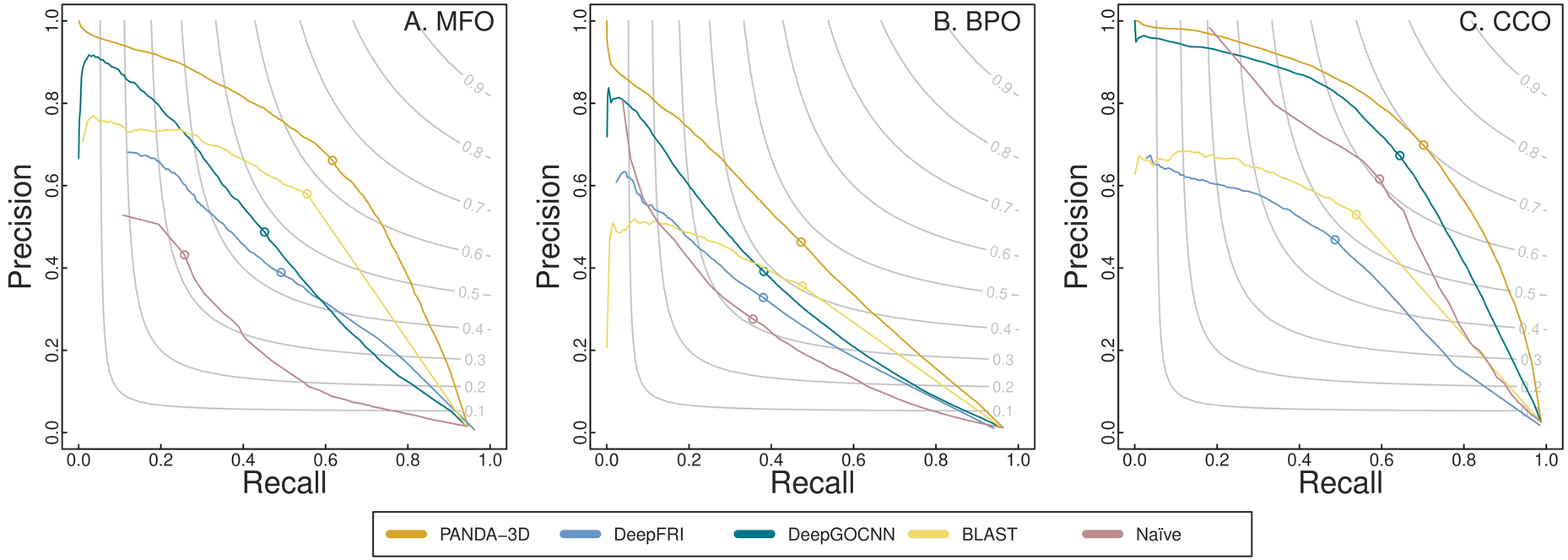


Figure 2. The F_{max} scores and precision-recall curves of PANDA-3D, DeepFRI, DeepGOCNN, Naïve and BLAST. The benchmark was performed on the testing dataset labeled as 'DeepFRI' in Table 1 with the maximum sequence identity cutoff of 0.95.

- PANDA-3D outperforms Deep-FRI, DeepGOCNN and all baseline methods.

RESULTS

Table 2. The performances of PANDA-3D, DeepFRI, DeepGOCNN, Naïve and BLAST for F_{max}, S_{min} and AUPR. The highest F_{max}, the smallest S_{min}, and the highest AUPR are in bold and italics. The benchmark was performed on the testing dataset labeled as 'DeepFRI' in Table 1

	F_{max}			S_{min}			AUPR		
Method	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
Naïve	0.322	0.311	0.605	11.381	49.404	12.728	0.197	0.227	0.525
BLAST	0.567	0.407	0.534	9.269	49.211	12.566	0.498	0.305	0.465
DeepGOCNN	0.469	0.387	0.658	9.834	46.861	11.635	0.444	0.336	0.69
DeepFRI	0.435	0.352	0.477	9.894	48.079	12.56	0.305	0.257	0.377
PANDA-3D	***0.642***	***0.471***	***0.705***	***7.29***	***43.601***	***10.027***	***0.654***	***0.445***	***0.766***

- PANDA-3D outperforms Deep-FRI, DeepGOCNN and all baseline methods.

RESULTS

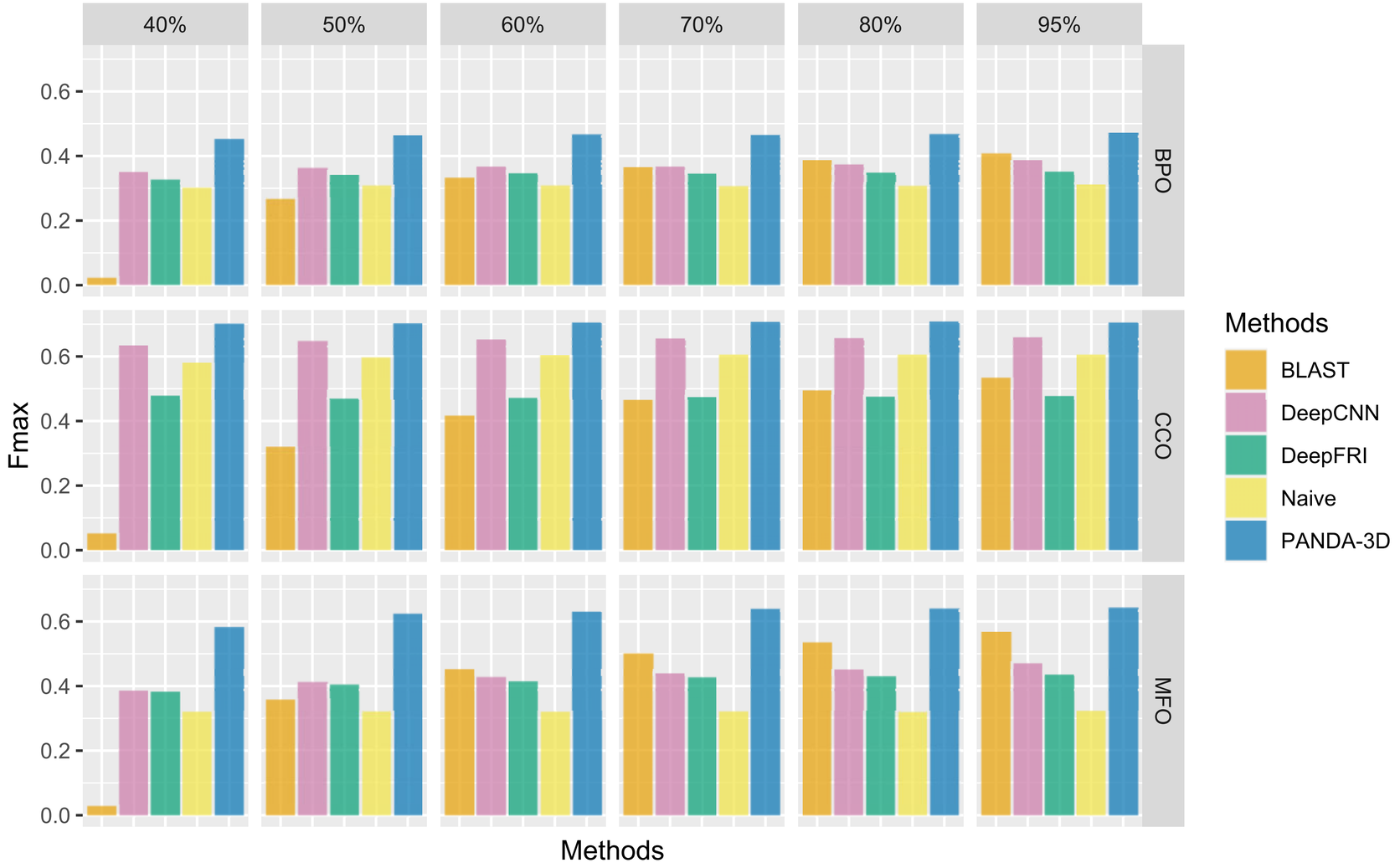


Figure 5. The Fmax scores of PANDA-3D, DeepFRI, DeepGOCNN, Naïve and BLAST at different maximum sequence identity cutoffs. 40%, 50%, 60%, 70%, 80% and 95% are maximum sequence identity cutoffs. Sequences from the testing set having a maximum identity score greater than the maximum sequence identity were removed. The benchmark was performed on the testing dataset labeled as 'DeepFRI' in Table 1.

RESULTS

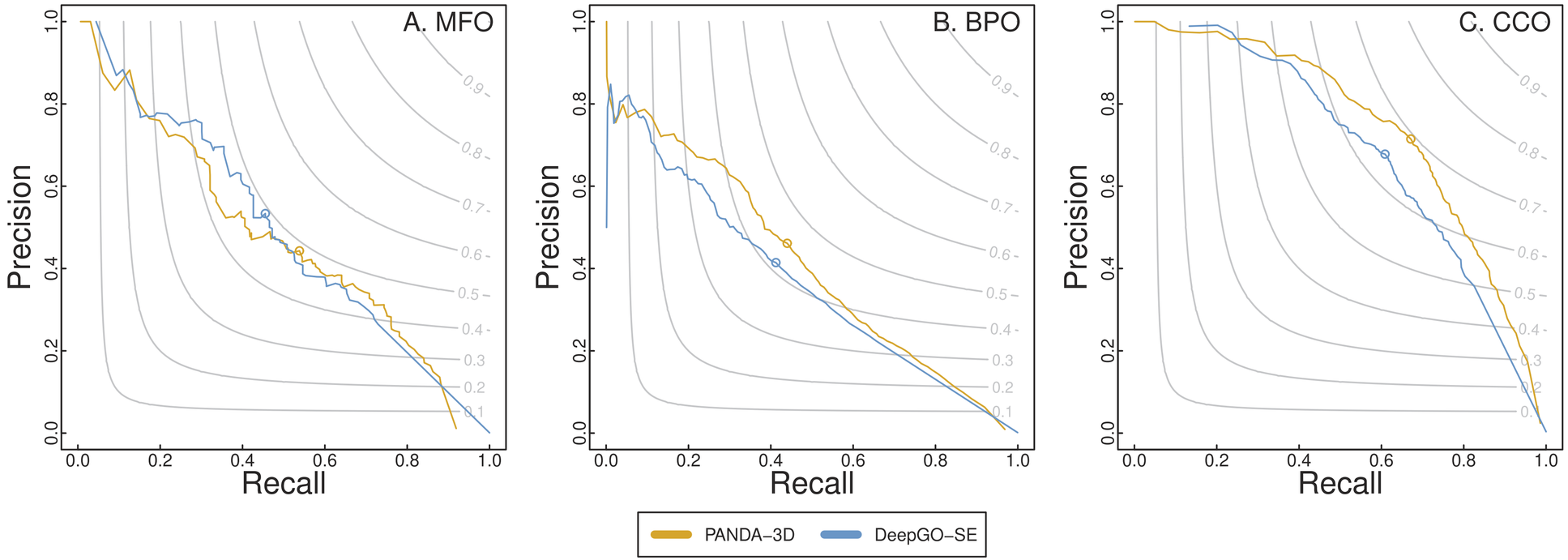


Figure 3. The F_{max} scores and precision-recall curves of PANDA-3D and DeepGO-SE. The benchmark was performed on the testing dataset labeled as 'DeepGO-SE' in Table 1 with the maximum sequence identity cutoff of 0.95.

- PANDA-3D outperforms DeepGO-SE for BPO and CCO almost all the time and provides comparable performance in terms of MFO.

RESULTS

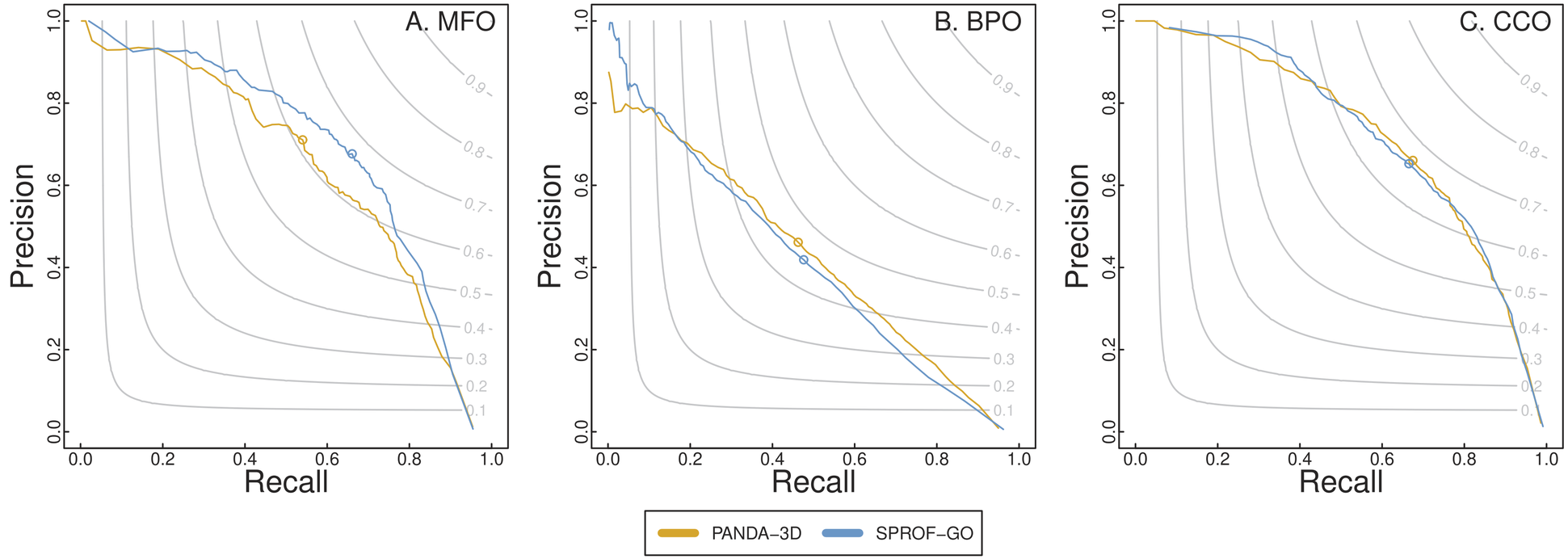


Figure 4. The F_{max} scores and precision-recall curves of PANDA-3D and SPROF-GO. The benchmark was performed on the testing dataset labeled as 'SPROF-GO' in Table 1 with the maximum sequence identity cutoff of 0.95.

- PANDA-3D is comparable to SPROF-GO for BPO and CCO and provides slightly worse performance in terms of MFO.

RESULTS

Table 5. The performance of PANDA-3D with permutated features. The top three lowest F_{max}, highest S_{min}, and the lowest AUPR are in bold and italics. The benchmark was performed on the testing dataset labeled as 'DeepFRI' in Table 1 with the maximum sequence identity cutoff of 0.95

	F_{max}			S_{min}			AUPR		
Permutated Features	**MFO**	**BPO**	**CCO**	**MFO**	**BPO**	**CCO**	**MFO**	**BPO**	**CCO**
None	0.642	0.471	0.705	7.29	43.601	10.027	0.654	0.445	0.766
ESM	***0.63***	***0.458***	***0.7***	***7.539***	***44.249***	***10.218***	***0.64***	***0.431***	***0.761***
Token	***0.634***	***0.465***	***0.7***	***7.421***	***43.934***	***10.151***	***0.647***	***0.437***	***0.763***
Dihedral	***0.631***	***0.463***	***0.701***	***7.457***	***44.024***	***10.177***	***0.643***	***0.435***	***0.762***
Euclidean distance	0.642	0.471	0.705	7.29	43.605	10.028	0.654	0.445	0.766
Edge vector	0.642	0.471	0.705	7.291	43.601	10.026	0.654	0.445	0.766
Orientation vector	0.641	0.47	0.704	7.295	43.627	10.023	0.651	0.443	0.764
pLDDT	0.643	0.471	0.704	7.286	43.58	10.034	0.654	0.445	0.765
Side-chain vector	0.642	0.47	0.705	7.291	43.611	10.034	0.654	0.444	0.765

- To interpret the significance of the features used in PANDA-3D, we shuffled the values of one feature randomly at the residue level to test its contributions towards accuracy.
- Top 3 features: ESM, dihedral angle and token

RESULTS

Table 6. Term-centric evaluations of PANDA-3D, DeepFRI, DeepGOCNN, Naïve and BLAST, in which the AUROCs on all candidate GO terms, biofilm formation (GO:0042710) and motility (GO:0001539) are reported. The benchmark was performed on the testing dataset labeled as 'DeepFRI' in Table 1 with the maximum sequence identity cutoff of 0.95

Methods	All GO-term classes	GO:0042710	GO:0001539
PANDA-3D	0.897924	0.942165	0.872102
DeepFRI	0.530414	0.418858	0.342771
BLAST	0.665574	0.525296	0.581225
Naïve	0.5	0.5	0.5
DeepGOCNN	0.809302	0.766282	0.721743

- term-centric evaluation

CONCLUSION

- PANDA-3D predicts protein functions from AlphaFold models and protein sequences.
- PANDA-3D combines GVP-GNN layers and decoder transformer layers.
- Feature importance analysis revealed that ESMs, tokens of sequence, and dihedrals are the top three most important features.

- A limitation of PANDA-3D is its reliance on AlphaFold models as input.