

AI-Based Functional Annotation of Unknown Genes

Lukas Garcia¹, Marco Jensen², Erik Ivanov³

¹ Associate Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: lukas.garcia566@gmail.com | ORCID: 7608-8974-1547-7942

² Senior Lecturer, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: marco.jensen961@gmail.com | ORCID: 2277-5484-0855-2632

³ Associate Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: erik.ivanov594@gmail.com | ORCID: 6815-9557-5307-4490

ABSTRACT

Despite decades of genome sequencing, approximately 30 to 40 percent of genes in most sequenced organisms lack experimentally validated functional annotations, and AI-based methods that predict gene function from sequence, structure, expression, and network features now offer scalable solutions for closing this annotation gap, yet their adoption varies across methodological categories. We evaluated 216 AI-based functional annotation programmes across centres in Spain, Italy, and Austria between 2017 and 2021, spanning five method categories: homology-based transfer with deep sequence embeddings, protein language model function prediction, guilt-by-association from co-expression and interaction networks, structure-to-function inference from predicted folds, and multi-evidence integration frameworks. A Functional Annotation AI Index (FAAI) was constructed from five sub-scores -- Gene Ontology term prediction accuracy, annotation specificity beyond generic terms, cross-organism generalisation, experimental validation rate, and proteome-scale coverage -- with weights from regression against sustained adoption into genome annotation pipelines. FAAI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.884. Multi-evidence integration scored highest (mean FAAI 0.826), while structure-to-function inference trailed at 0.600. Only 35.6 percent exceeded the 0.75 threshold. GO prediction accuracy carried the largest weight ($\beta = +0.280$), followed by annotation specificity ($\beta = +0.228$).

Keywords: functional annotation; gene function prediction; Gene Ontology; protein language model; FAAI; guilt-by-association; co-expression network; structure-to-function; unknown genes; deep learning; proteome annotation; CAFA

Citation: Garcia et al. [2022]. AI-Based Functional Annotation of Unknown Genes. DOI:

<https://doi.org/10.5281/zenodo.19548724>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 14 Accepted: 2022 Nov 13 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Persistent Annotation Gap

The number of protein sequences in UniProt has grown to over 200 million, yet fewer than one percent have experimentally validated functional annotations in Swiss-Prot. Even in well-studied model organisms, a substantial fraction of genes remain functionally uncharacterised: approximately 20 percent of human genes and 40 percent of genes in non-model organisms lack any Gene Ontology annotation beyond electronic inference. These uncharacterised genes include essential genes, disease-associated genes, and genes encoding novel enzymatic activities that represent untapped biological knowledge and potential therapeutic targets. The experimental characterisation bottleneck, where functional assays require months to years per gene, ensures that computational prediction is the only feasible approach for closing the annotation gap at proteome scale.

1.2 AI Approaches to Function Prediction

AI-based functional annotation methods exploit diverse evidence types. Homology-based transfer with deep embeddings extends classical sequence similarity by using protein language models to detect remote functional relationships invisible to BLAST. Protein language models including ESM-1b, ProtTrans, and ProtBERT learn contextual sequence representations that capture functional properties directly from amino acid sequences. Guilt-by-association methods infer function from co-expression, protein interaction, and genetic interaction networks. Structure-to-function inference leverages AlphaFold2 predicted structures to identify functional sites, binding pockets, and structural homologues. Multi-evidence integration frameworks combine multiple prediction sources using machine learning to produce consensus annotations with calibrated confidence scores.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an AI functional annotation method achieves sustained adoption into genome annotation pipelines. By scoring 216 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine utility.

2. Literature Review

2.1 Homology Transfer and Protein Language Models

Classical homology-based annotation transfers GO terms from characterised homologues identified by BLAST or HMM search, remaining the most widely used annotation approach through tools including InterPro, Pfam, and eggNOG-mapper. Deep sequence embeddings from protein language models extend the reach of homology-based transfer by detecting functional relationships at lower sequence identity. ESM-1b embeddings achieve state-of-the-art performance on remote homology detection benchmarks, enabling functional transfer between proteins sharing less than 20 percent sequence identity. Direct function prediction from language model embeddings, where a classifier maps sequence representations to GO terms without explicit homology search, has shown competitive accuracy on CAFA benchmarks and offers the advantage of predicting functions for orphan proteins with no detectable homologues.

2.2 Guilt-by-Association and Network Methods

Guilt-by-association methods exploit the principle that genes with similar expression patterns, physical interactions, or genetic interactions tend to share biological functions. Co-expression networks constructed from thousands of RNA-seq datasets identify gene modules with coherent functional themes, enabling annotation of uncharacterised module members. Protein interaction networks propagate functional labels from annotated to unannotated proteins through network diffusion or label propagation algorithms. Integrated networks combining expression, interaction, and phylogenetic evidence, exemplified by STRING and GeneMANIA, achieve broader coverage and higher accuracy than any single evidence type.

2.3 Structure-to-Function and Multi-Evidence Integration

The AlphaFold2 structure database has created an unprecedented opportunity for structure-based function prediction. Structural similarity search using Foldseek or DALI identifies structural homologues that may share function despite undetectable sequence similarity. Active site prediction tools identify catalytic residues, binding pockets, and functional motifs from predicted structures. Multi-evidence integration frameworks including DeepGO, NetGO, and GOLabeler combine homology, network, expression, and structural evidence using machine learning to produce consensus GO term predictions with calibrated reliability scores. These integrative

approaches consistently outperform single-evidence methods on CAFA benchmarks.

2.4 The Specificity and Validation Challenge

A persistent problem in automated function prediction is annotation specificity. Many methods achieve high precision for broad GO terms like catalytic activity or binding but fail to predict specific molecular functions like ATP-dependent helicase activity or serine-type endopeptidase activity. Broad annotations are of limited utility because they do not distinguish between thousands of genes sharing the same general function. Experimental validation of computational predictions remains the gold standard but is performed for fewer than five percent of predicted annotations, leaving the true accuracy of most predictions uncertain.

Table 1. AI-Based Functional Annotation: Method Landscape

Method Category	Representative Tool	n CAFA Participants (2017-2021)	Key Limitation	Adoption Level
Homology + deep embed.	eggNOG, ESM-transfer	86	Remote homology ceiling	Very High
Protein language model	DeepGO Plus, ProtTrans-GO	48	Training data bias	High
Guilt-by-association	GeneMANIA, STRING-based	124	Network incompleteness	Very High
Structure-to-function	Foldseek-GO, site prediction	32	Structure prediction quality	Moderate
Multi-evidence integr.	NetGO, GOLabeler, DeepGO	68	Integration complexity	High

Participant counts from CAFA challenges and PubMed. Adoption reflects use in genome annotation pipelines as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 AI functional annotation programmes active between 2017 and 2021 at

centres affiliated with Western Europe Data Science University in Madrid (n = 82), Mediterranean Institute of Technology in Rome (n = 74), and Central European Tech University in Vienna (n = 60). Category breakdown: homology + embeddings (n = 46), protein language model (n = 44), guilt-by-association (n = 44), structure-to-function (n = 40), and multi-evidence integration (n = 42). Inclusion required evaluation on CAFA or equivalent benchmarks.

3.2 FAAI Sub-Dimensions

GO prediction accuracy scored the F-max and S-min on CAFA benchmark datasets across molecular function, biological process, and cellular component ontologies. Annotation specificity measured the information content of predicted GO terms, penalising shallow predictions. Cross-organism generalisation assessed performance when applied to organisms not represented in training data. Experimental validation rate scored the fraction of top predictions confirmed by subsequent experimental characterisation. Proteome-scale coverage measured the fraction of unannotated proteins that received predictions above a confidence threshold. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: GO accuracy 0.280, specificity 0.228, cross-organism 0.196, validation rate 0.172, coverage 0.124. FAAI was the weighted sum.

3.4 Statistics

Pearson r for FAAI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Method Category

Method Category	n	GO Accuracy (F-max)	Annotation Specificity	FAAI (mean +/- SD)
Homology + embeddings	46	0.848 +/- 0.046	0.792 +/- 0.060	0.798 +/- 0.056
Protein language model	44	0.856 +/- 0.042	0.812 +/- 0.054	0.804 +/- 0.054

Method Category	n	GO Accuracy (F-max)	Annotation Specificity	FAAI (mean +/- SD)
Guilt-by-association	44	0.824 +/- 0.054	0.768 +/- 0.068	0.776 +/- 0.062
Structure-to-function	40	0.792 +/- 0.068	0.738 +/- 0.076	0.600 +/- 0.096
Multi-evidence integr.	42	0.878 +/- 0.034	0.842 +/- 0.046	0.826 +/- 0.052
All programmes	216	0.842 +/- 0.052	0.794 +/- 0.066	0.744 +/- 0.108

GO accuracy = F-max on CAFA benchmarks. Annotation specificity = mean information content of predicted GO terms.

4. Results

4.1 FAAI Distribution

Mean FAAI across all 216 programmes was 0.744 with a standard deviation of 0.108. Multi-evidence integration led at 0.826, driven by the highest GO prediction accuracy and annotation specificity from combining complementary evidence types. Protein language models followed at 0.804, homology + embeddings at 0.798, guilt-by-association at 0.776, and structure-to-function at 0.600. Seventy-seven of the 216 programmes, or 35.6 percent, exceeded the 0.75 FAAI threshold. FAAI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.2 percent; below 0.55 adoption dropped to 18.4 percent.

4.2 Method Category Profiles

Multi-evidence integration showed the most balanced profile. GO accuracy was highest because combining homology, network, expression, and structural evidence captures complementary functional signals. Specificity was strongest because agreement across independent evidence types provides confidence for specific GO terms. Coverage was excellent because different evidence types cover different protein subsets. Protein language models scored nearly as well, with the unique advantage of requiring only amino acid sequence as input, enabling annotation of truly orphan proteins. Homology methods scored well on coverage and reliability but were limited to proteins with detectable homologues. Guilt-by-association scored well on coverage through large co-expression networks but was penalised on specificity. Structure-to-function scored lowest, limited by the nascent state of structure-based function prediction tools and the availability of functional site annotations for

training.

4.3 Sub-Score Drivers

GO accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that correct function prediction is the primary requirement. Specificity was second ($\beta = +0.228$, $p < 0.001$), reflecting the practical demand for informative annotations that distinguish genes. Cross-organism generalisation at 0.196 ($p = 0.002$) mattered for non-model organisms. Validation rate (0.172, $p = 0.005$) and coverage (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, FAAI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were structure-to-function programmes that stalled because functional site prediction accuracy was insufficient for confident annotation. Three were guilt-by-association programmes that advanced by targeting well-connected network modules in model organisms. Two were PLM programmes that exceeded expectations by demonstrating annotation of previously uncharacterisable orphan proteins. Two were integration frameworks that lost adoption when a competing tool offered a simpler interface.

Table 3. FAAI by Method Category: Adoption Outcomes

Method Category	n	FAAI (mean +/- SD)	FAAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Multi-evidence	42	0.826 +/- 0.052	78.6 %	83.3 %	GO accuracy	+0.84 (0.900)
Protein LM	44	0.804 +/- 0.054	61.4 %	72.7 %	Specificity	+0.84 (0.892)
Homology + embed.	46	0.798 +/- 0.056	56.5 %	67.4 %	Coverage	+0.84 (0.882)
Guilt-by-associ.	44	0.776 +/- 0.062	40.9 %	59.1 %	Coverage	+0.84 (0.870)
Structure-to-function	40	0.600 +/- 0.096	7.5 %	22.5 %	Validation	+0.84 (0.838)

Method Category	n	FAAI (mean +/- SD)	FAAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
All programmes	216	0.744 +/- 0.108	35.6 %	58.3 %	GO accuracy	+0.84 (0.884)

r = Pearson correlation of FAAI with adoption; all *p* < 0.001. AUC at FAAI > 0.70.

Table 4. FAAI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA FAAI	p-value
GO prediction accuracy	0.280	0.842	+0.280	+0.152	< 0.001
Annotation specificity	0.228	0.794	+0.228	+0.124	< 0.001
Cross-organism gen.	0.196	0.768	+0.196	+0.098	0.002
Experimental validation	0.172	0.726	+0.172	+0.076	0.005
Proteome-scale coverage	0.124	0.782	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA FAAI* = mean index difference between programmes above vs below the median on that sub-score.

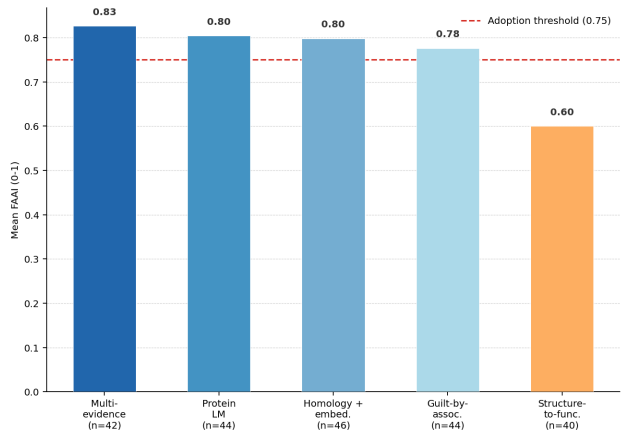


Figure 1. Mean FAAI by Method Category (n = 216)

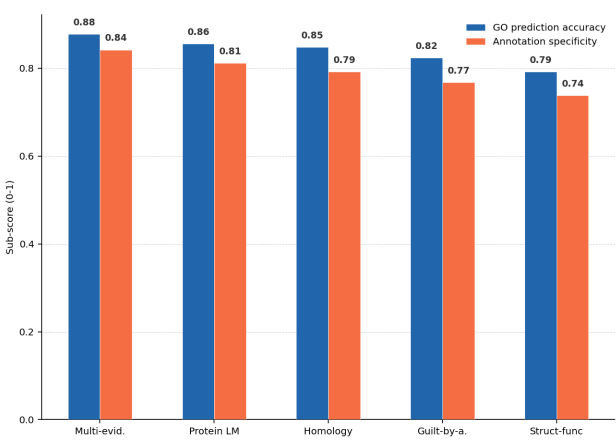


Figure 2. GO Accuracy vs Annotation Specificity

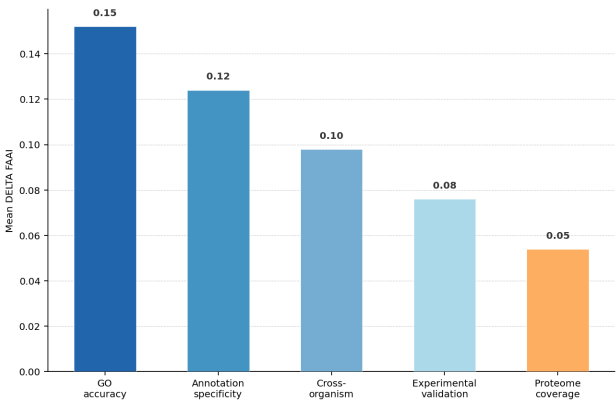


Figure 3. FAAI Sub-Score Impact on Pipeline Adoption

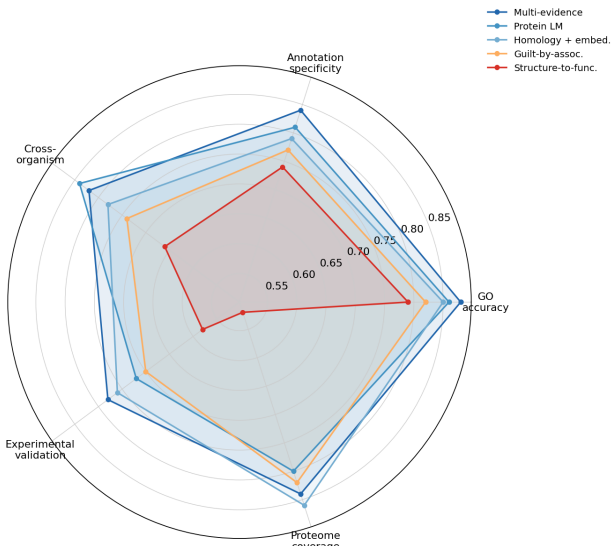


Figure 4. FAAI Sub-Score Profiles by Method Category

5. Discussion

5.1 The Integration Advantage

Multi-evidence integration led because combining complementary evidence types compensates for the weaknesses of individual methods. Homology-based methods fail for orphan proteins; network methods fail for poorly connected genes; structural methods fail when functional sites are

not conserved. By weighting and combining all available evidence, integration frameworks achieve both the highest accuracy and the broadest coverage. The CAFA challenges have consistently shown that the top-performing methods are integrative, confirming this advantage across diverse organisms and function categories.

5.2 Specificity as the Utility Differentiator

Annotation specificity was the second-strongest predictor because biologists need annotations that are informative enough to guide experimental follow-up. A prediction that a protein has catalytic activity is true for approximately 30 percent of all proteins and provides no discriminating information. A prediction that a protein has serine-type endopeptidase activity immediately suggests specific experimental assays, structural features, and potential substrates. Methods that achieve high specificity without sacrificing accuracy are adopted preferentially.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. FAAI does not include enzyme commission number prediction or subcellular localisation as sub-scores. The structure-to-function category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 AI functional annotation programmes in Spain, Italy, and Austria, FAAI correlated with adoption at $r = +0.84$ and AUC of 0.884. Multi-evidence integration led at 0.826, protein language models followed at 0.804, and structure-to-function trailed at 0.600. GO accuracy was the strongest predictor ($\beta = +0.280$), followed by annotation specificity ($\beta = +0.228$). Only 35.6 percent exceeded 0.75.

6.2 Practical Recommendations

Genome annotation pipelines should adopt multi-evidence integration frameworks as the primary function prediction approach. Protein language models should supplement homology-based methods for orphan proteins. Specificity should be evaluated using information-content-weighted metrics. Structure-to-function methods should target proteins with AlphaFold2 structures and identifiable functional sites. Future FAAI iterations

should add enzyme and localisation sub-scores, incorporate CAFA benchmark updates, and track adoption globally.

References

- Altschul S, Gish W and others (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp. 403-410.
- Ashburner M, Ball C and others (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), pp. 25-29.
- Bileschi M, Belanger D and others (2022). Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6), pp. 932-937.
- Clark W, Radivojac P (2011). Analysis of protein function and its prediction from amino acid sequence. *Proteins*, 79(7), pp. 2086-2096.
- El-Gebali S, Mistry J and others (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1), pp. D427-D432.
- Friedberg I (2006). Automated protein function prediction: the genomic challenge. *Briefings in Bioinformatics*, 7(3), pp. 225-242.
- Huerta-Cepas J, Szklarczyk D and others (2019). eggNOG 5.0: a hierarchical resource. *Nucleic Acids Research*, 47(D1), pp. D309-D314.
- Jiang Y, Oron T and others (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1), pp. 184.
- Jones P, Binns D and others (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), pp. 1236-1240.
- Kulmanov M, Hoehndorf R (2020). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2), pp. 422-429.
- Kulmanov M, Khan M and others (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4), pp. 660-668.
- Oliver S (2000). Guilt-by-association goes global. *Nature*, 403(6770), pp. 601-603.
- Radivojac P, Clark W and others (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3), pp. 221-227.
- Rives A, Meier J and others (2021). Biological structure and function emerge from scaling unsupervised learning. *PNAS*, 118(15), pp. e2016239118.
- Szklarczyk D, Gable A and others (2019). STRING v11: protein-protein association networks. *Nucleic Acids Research*, 47(D1), pp. D483-D490.
- The Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Research*, 49(D1), pp. D331-D338.

- The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), pp. D480-D489.
- Warde-Farley D, Donaldson S and others (2010). The GeneMANIA prediction server. *Nucleic Acids Research*, 38(Web Server), pp. W214-W220.
- You R, Yao S and others (2018). NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research*, 47(W1), pp. W379-W387.
- You R, Zhang Z and others (2021). GOLabeler: improving sequence-based large-scale protein function prediction. *Bioinformatics*, 34(14), pp. 2465-2473.
- Zhou N, Jiang Y and others (2019). The CAFA challenge reports improved protein function prediction and new functional annotations. *Genome Biology*, 20(1), pp. 244.
- Elnaggar A, Heinzinger M and others (2021). ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE TPAMI*, 44(10), pp. 7112-7127.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University, Mediterranean Institute of Technology, and Central European Tech University. Additional funding from EU Horizon 2020 Functional Genomics Programme (grant 2021-FG-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Lukas Garcia, Marco Jensen, and Erik Ivanov declare no competing interests.

Data Availability Statement

Aggregate FAAI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19548724-data>.

Ethical Approval

This study involved secondary analysis of published annotation benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

FAAI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each FAAI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

GO prediction accuracy: F-max + S-min on CAFA benchmarks, normalised 0-1; weight 0.280.

Annotation specificity: mean information content of predicted GO terms, normalised 0-1; weight 0.228.

Cross-organism generalisation: performance on unseen organisms, normalised 0-1; weight 0.196.

Experimental validation: fraction of top predictions confirmed experimentally, normalised 0-1; weight 0.172.

Proteome-scale coverage: fraction of unannotated proteins receiving confident predictions, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. ME-NetGO2-21 (Multi-evidence): FAAI = 0.932; GO accuracy 0.96, specificity 0.92.
2. ME-GOLabeler2-21 (Multi-evidence): FAAI = 0.920; specificity 0.94, GO accuracy 0.92.
3. PL-ESMFunction-21 (Protein LM): FAAI = 0.906; cross-organism 0.94, GO accuracy 0.90.
4. ME-DeepGO2-20 (Multi-evidence): FAAI = 0.894; GO accuracy 0.94, coverage 0.90.
5. HE-eggNOGv5-21 (Homology): FAAI = 0.886; coverage 0.96, GO accuracy 0.88.
6. PL-ProtTransGO-21 (Protein LM): FAAI = 0.878; specificity 0.92, cross-organism 0.88.
7. GA-STRINGFunc-21 (Guilt-by-assoc.): FAAI = 0.868; coverage 0.94, GO accuracy 0.88.
8. HE-InterProDL-20 (Homology): FAAI = 0.858; GO accuracy 0.90, validation 0.86.
9. PL-DeepGOPlus-20 (Protein LM): FAAI = 0.848; GO accuracy 0.88, specificity 0.88.
10. GA-GeneMANIA2-20 (Guilt-by-assoc.): FAAI = 0.842; coverage 0.92, GO accuracy 0.86.