AstraROLE & AstraSUIT: Multi-Task Annotation Models for Functional Profiling of Proteins

Çağlar Bozkurt[1,*], **Alexandra Vasilyeva**[1], **Aniruddh Goteti**[1]

[1]Orbion GmbH, Berlin, Germany

[*] Correspondence: caglar.bozkurt@orbion.life

Abstract

Most *in-silico* protein characterisation tools focus on only one aspect of protein function, forcing researchers to use multiple models or to bypass computational checks. Here we introduce AstraROLE and AstraSUIT, two transformer-based, multi-task annotators that deliver an integrated functional profile in a single pass.

A 1,351-dimensional input (ESM-2 CLS embeddings plus physicochemical Orbion enrichments) is mapped by a 512-unit encoder and task-specific linear heads: four in AstraROLE (EC class, GO term, molecular pathway, protein category) and six in AstraSUIT (cofactor, domain, host, membrane association type, transmembrane helix number, subcellular localization). Models were trained on 730k UniProt proteins with stratified 70/15/15 splits; class-weighted BCE and Optuna hyper-parameter search countered imbalance. On hold-out sets the heads reached macro F_1=0.82–0.98 and MCC=0.79–0.98. Highest scores were seen for membrane association type (F_1=0.98), top-level EC number (0.97) and cofactor binding (0.96); organism type association was hardest (0.82). Against recent comparators (incl. DeepGOPlus and TargetP 2.0), the Astra models matched or exceeded performance, especially on metal-ion binding and subcellular localisation. Additional tests on three novel proteins not included in initial dataset showed good predictions for most labels, underscoring the potential for hypothesis generation.

Overall, AstraROLE and AstraSUIT supplied fast, state-of-the-art multi-label protein annotation within one unified model network.

June 21, 2025

1 Introduction

Using AI-driven insights for biomedical experiments and protein design is rapidly gaining traction. Artificial intelligence is leveraged to accelerate drug discovery, identify novel biomarkers, process microscopy and scattering data, and facilitate the automation of laboratory workflows.(1) In light of the long time-scales and high failure rates that characterise the drug discovery pipeline, it is hoped that utilising AI can improve both the speed and the success rate of developing novel drug candidates and studying potential drug targets.(2; 3)

A number of tools are available to researchers that can assist with stabilising proteins and interpreting experimental outcomes.(1; 2; 3) However, many of the currently available models focus on only one aspect of protein biology, such as post-translational modifications (PTMs) or taxonomic-classification prediction, and integrating insights from a large number of models can be challenging. At the same time, multiple protein characteristics need to be considered: for instance, taxonomy and PTMs influence the choice of expression system, whereas the expected molecular function informs the necessary assays.

If modifications to the protein are introduced, it is desirable to assess all its parameters *in silico* before committing to experimental work; within a single structure-determination project, dozens of mutant variants may be screened for stability, making the use of multiple input models impractical or impossible. As a result, mutant proteins are often created via rational design and tested experimentally without a reliable intermediate assessment, introducing delays and extra costs into the process in case of failure.

Our goal is to create a family of Orbion AI models using state-of-the-art methods and to integrate them seamlessly within a single platform, thereby providing an accurate, multifaceted overview of protein features. Previously, we introduced the AstraPTM model, which employed a transformer-based architecture to predict PTMs with high reliability; the AstraPTM2 model is currently being prepared for publication. Here, we report on two additional AI models—AstraROLE and AstraSUIT—that aim to provide a more rounded picture of a protein by predicting its likely biological role along with key features such as metabolic-pathway membership and sub-cellular localisation.

2 Methods

Architecture

AstraROLE and AstraSUIT are both single-encoder, multi-task **transformers**(9) that turn the joint protein feature vector (protein-level embedding + ESM2 CLS embeddings) into functional annotations in one forward pass (Figure 1). Briefly, the input features are projected to a Linear Layer, sent to a Positional Encoding layer, and then sent through the Transformer encoder layers, which are connected to different heads for each task, with sigmoid functions. The step-by-step architecture of the models is described in more detail in the Appendix.

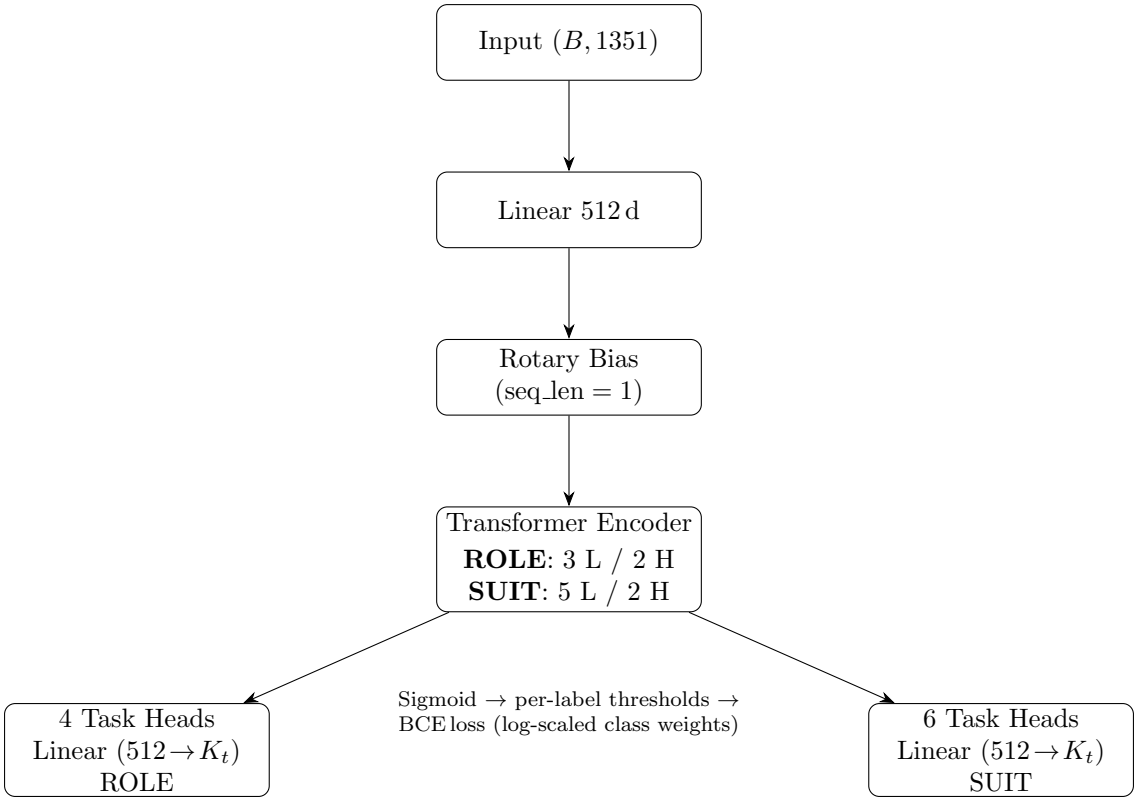

Figure 1: Vertical data-flow diagram for the Astra backbone and its two model-specific task-head branches. L = transformer layers, H = attention heads.

Training Data

Both models accept as input a 1,351 dimensional vector per protein, including protein sequence, ESM2 embeddings and Orbion enrichments. The dataset comprised 730,197 unique proteins curated from **UniProt**(8) (Swiss- and TrEMBL; database cut-off 31.12.2024; see Figure 2 for top-line dataset overview and Appendix for more detailed classification). The data from UniProt were retrieved and categorized via UniProt's API into individual JSON files. Then, through a proprietary classifier, comments, annotations, categorized information, references, and keywords were parsed to find the right values for each label.

ESM2 embeddings, which help to incorporate evolutionary information, were generated from `esm2_t33_650M_UR50D`(6). CLS versions were used here, as token_representations[0]. To add descriptive information on the protein, the enrichments from Orbion were added that include molecular weight, GRAVY score, isoelectric point, length and number of bonds.

Normalization was performed through a StandardScaler on the training data. Both mean and standard deviation values were saved into scaler files, which were then used on inference requests on the models. The data set was shuffled and split 70% for training, 15% for validation, and 15% for testing, independently for each model.

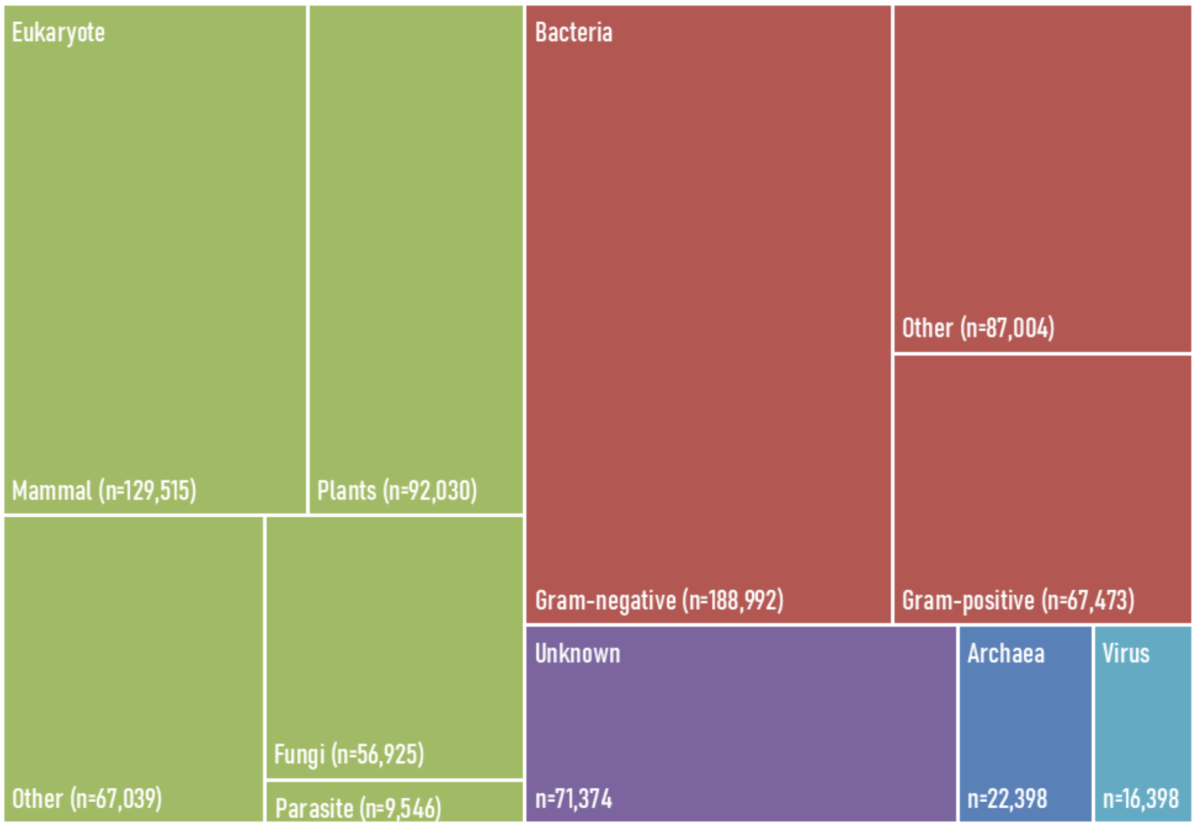

Figure 2: Overview of protein taxonomy labels in the UniProt dataset.

Training and Optimization

Multiple training iterations were run for each model to optimize the performance of the predictions. On the protein feature set approaches, CLS, LSE, and Mean ESM embedding versions were tested, where the F1 scores for all heads were checked. As these tests didn't show a significant difference between different features, CLS was used as the key ESM feature set approach. After this, hyper-parameter optimisation was made using **Optuna**(4).

For AstraROLE, this optimization process led to 39 optimization trials on hyperparameters, where 22 of them were pruned in the process, and 16 were completed (see Table 1 for the selected hyperparameters). Overall, 40, 45, 50, and 60 epochs were tested; 50 epochs was the optimal point, where a higher number of epochs lead to an overfit. After 54 training iterations (incl. hyperparameter optimization), the model has shown a net growth of 0.10% on EC Number F1, 3.80% on GO Terms F1, 1.10% on Protein Categories F1, and 2.10% on Pathway Memberships F1.

For AstraSUIT, the optimization led to 26 trials on hyperparameters, with 7 pruned in the process and 19 completed. 40 and 60 epochs were tested and 40 epochs was the optimal point. After 41 training iterations (incl. hyperparameter optimization), the model has shown a net growth of 0.40% on Cofactor Binding F1, 2.40% on Domain-level Classification F1, 0.40% on Host Association F1, 1.80% on Subcellular Location F1, and 1% on TM Helices F1.

Table 1: Key hyper-parameters chosen for AstraROLE and AstraSUIT, optimised with Optuna.

Hyper-parameter	**AstraROLE**	**AstraSUIT**
Hidden dimension	512	512
Number of attention heads	2	2
Number of layers	3	5
Dropout rate	0.10	0.21
Training epochs	50	40
Learning rate	5.96×10^{-5}	8.13×10^{-5}
Weight decay	9.16×10^{-6}	1.05×10^{-5}
Loss-scale factor	0.55	0.90
Batch size	64	128

3 Results

Overall Performance

Both AstraROLE and AstraSUIT models showed good overall performance, with macro F1 between 0.815 and 0.980 and macro Matthews Correlation Coefficient (MCC) between 0.788 and 0.976 (see Table 2). The most accurate category for prediction was membrane type (F1=0.980, MCC=0.976), followed closely by EC number (F1=0.966, MCC=0.961) and cofactor binding (F1=0.955, MCC=0.947). Lowest MCC and F1 were seen for host association (F1=0.815, MCC=0.788). GO terms showed satisfactory performance by these metrics but relatively higher false negative rate (FNR=0.160).

Table 2: Overall performance of AstraROLE and AstraSUIT models. Highest values (or lowest values for FPR/FNR) are highlighted in green.

Model	**Output**	**Acc**	**Prec**	**Rec**	**F1**	**MCC**	**FPR**	**FNR**
AstraROLE	EC Number	0.991	0.969	0.963	0.966	0.961	0.004	0.037
	GO Terms	0.996	0.860	0.840	0.850	0.848	0.002	0.160
	Protein Categories	0.964	0.873	0.858	0.866	0.845	0.020	0.142
	Pathway Memberships	0.997	0.934	0.900	0.916	0.915	0.001	0.100
AstraSUIT	Cofactor Binding	0.986	0.955	0.954	0.955	0.947	0.008	0.046
	Domain Classification	0.983	0.890	0.889	0.889	0.880	0.009	0.111
	Host Association	0.950	0.768	0.868	0.815	0.788	0.038	0.132
	Membrane Type	0.993	0.981	0.979	0.980	0.976	0.004	0.021
	Subcellular Location	0.982	0.809	0.873	0.881	0.871	0.009	0.127
	TM Helices Group	0.994	0.872	0.869	0.871	0.867	0.003	0.131

ACC: Accuracy, Prec: Precision, Rec: Recall; FPR: False Positive Rate; FNR: False Negative Rate; MCC: Matthews Correlation Coefficient. Green = Best, Yellow = Moderate, Red = Weakest

Per-Label Performance

Performance metrics (accuracy, precision, recall, F1, MCC, false positive and negative rates) for each label are provided in the Appendix. Here we describe the key outcomes, focusing on F1 as a representative metric due to the imbalance of the dataset, and on FNR due to false negative rates being much higher than false positives.

For EC number, F1 varied from 0.93 for hydrolases to 0.986 for ligases, with FNR 0.077 across the categories. In protein category prediction, enzymes (F1=0.949, FNR=0.055) and transporters (F1=0.927, FNR=0.092) performed best, and the worst performance was in motor proteins (F1=0.687, FNR=0.323).

Prediction for pathways was overall accurate, with somewhat lower performance in lipid-related pathways (F1=0.71 to 0.865, FNR=0.181 to 0.325) and the mevalonate pathway (F1=0.724, FNR=0.335). Interestingly, the best model performance was seen for the non-mevalonate (MEP) pathway (F1=0.983, FNR=0.012), which is an alternate route to isoprenoids found in bacteria and plants.

GO term prediction identified several terms with bad prediction performance, including antioxidant, hormone receptor, biotin binding, acetyltransferase and protein tyrosine kinase activity (see Figure 3). For other labels, model performance was satisfactory with F1 above 0.5, and some domains, including most binding and polymerase activity, performed well with F1 above 0.8.

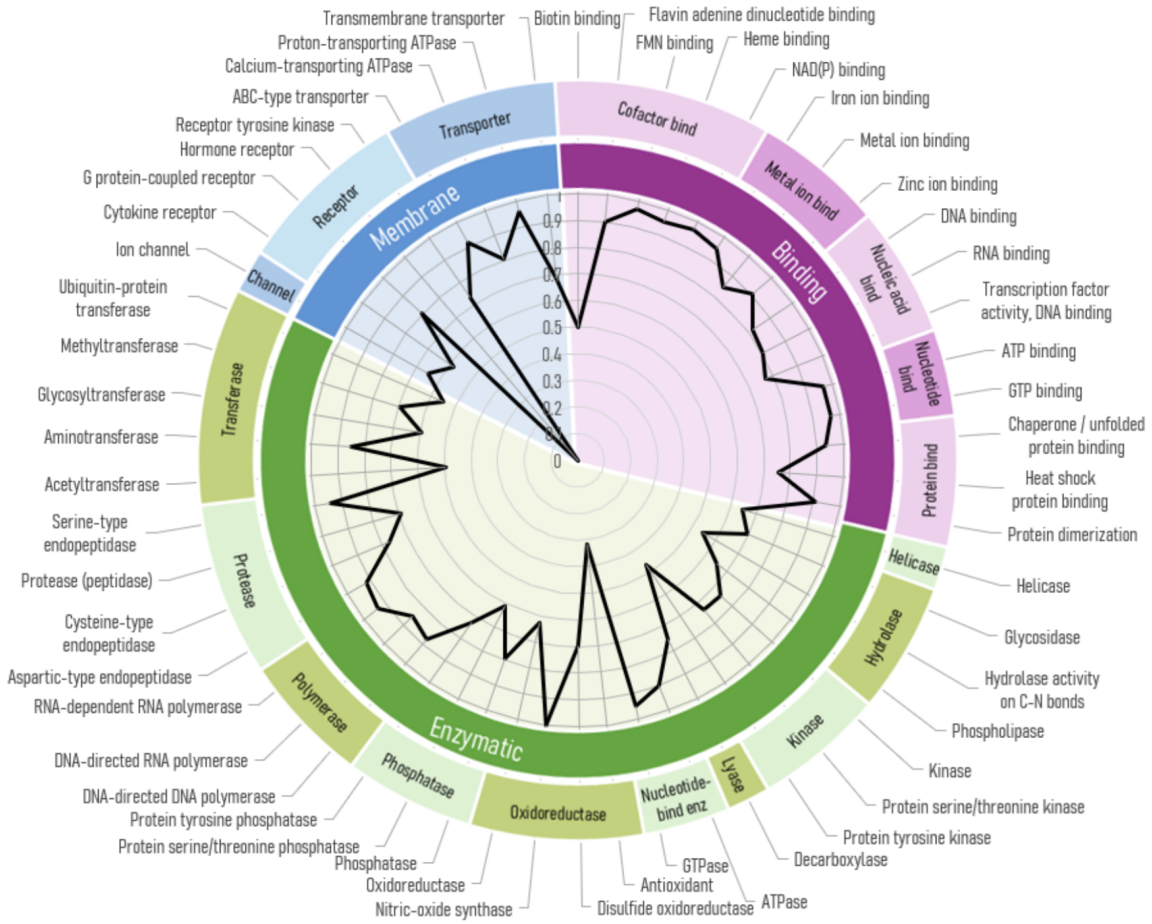

Figure 3: F1 factor for AstraROLE performance per GO term.

In AstraSUIT, cofactor binding showed positive results with F1 ranging from 0.94 (metal ions) to 0.98 (iron sulphur clusters). Domain classification yielded overall positive performance, with the highest F1=0.947 for gram-positive bacteria; within bacteria, eukaryote and virus sub-classification the lowest F1 (0.721–0.8) was seen for the catch-all "Other" subcategory. Organism type association yielded the worst prediction result within AstraSUIT, which was for the commensal/ symbiont label (F1=0.527, FNR=0.507); animal association prediction worked best (F1=0.887, FNR=0.095).

Membrane association type was predicted reliably (F1=0.792–0.99), aside from the peripheral membrane label (F1=0.182). Solubility label showed very good performance (F1=0.99, FNR=0.013). Number of transmembrane helixes was predicted with F1=0.826–0.92. In addition to membrane location classification, we separately evaluated prediction of subcellular localization, which showed F1=0.77 or above, and FNR=0.208 or less (see Figure 4).

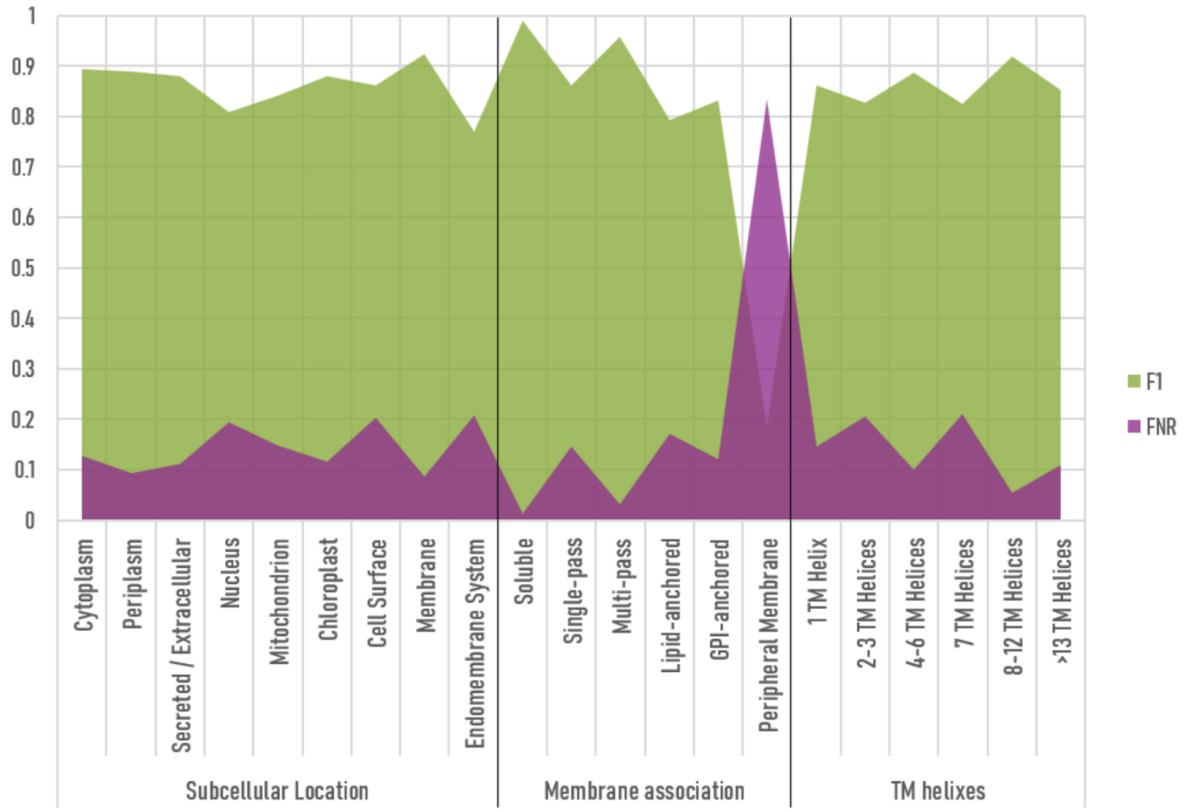

Figure 4: F1 and FNR prediction metrics for subcellular location labels.

Case Studies

In order to additionally assess the performance of the models, we identified three proteins which were not included in the UniProt database before the download date (31.12.2024) and that were reported to have a high degree of structural and/or functional novelty.

IgtG is a glycosyltransferase from gram-positive bacteria that catalyzes iterative addition of up to four glucose units to a single serine on mature lasso peptides.(10) IgtG and its homologs are evolutionally separate from other known ribosomally synthesized and post-translationally modified peptide (RiPP) glycosyltransferases. The Astra models correctly identified the bacterial taxonomy but suggested sub-classification as "other" (85.4%) or gram-negative (55.1%). Membrane type (as subcellular location and multi-pass type) was identified with almost 100% probability, which aligns with structural predictions. Transferase (EC2) function prediction (21.2%) did not reach the threshold for positive identification, and GO terms were a false negative. Function (enzyme 83.7%, structural 80.3%, peptidoglycan biosynthesis 89.2%) aligned with glycosyltransferase pathways.

esmGFP is a synthetic fluorescent protein that shares 58% sequence identity with the closest known natural fluorescent protein.(11) The models did not return a positive taxonomy but suggested animal interaction (72.2%), which aligns with a novel synthetic GFP variant. Other predictions were soluble (99.9%), located in cytoplasm (67.4%), non-enzyme (99.0%) with a specialized function (85.3%), with none of the evaluated GO

terms or molecular pathways returning a positive hit. All of these align with expectation and show no false positives.

TPM3P9 is a human microprotein previously believed to be associated with noncoding RNA; it regulates RNA splicing with a potentially oncogenic effect and lacks significant sequence similarity to known proteins.(12) It was identified by the Astra models as a mammalian protein (98.9%) interacting with humans (94.3%) and animals (34.4%), soluble (100%) but without a positive hit for subcellular location. Correctly identified as a non-enzyme (99.7%) without a false positive on pathways, but did not identify the regulatory function (predicting instead structural and motor function).

4 Comparisons

AstraROLE and AstraSUIT were benchmarked against the newest publicly-available protein-annotation models released between 2022–2025, where a comparison was possible. The following models were included as potentially relevant comparators:

- **GO-Slim**: TALE+(13), DeepGO-SE(14), GNN-GO(15), ProteInfer(21)
- **EC superclass**: GraphEC(16), DeepEC-Transformer(17), ECPred-GAT(18)
- **Cofactor / metal binding**: M-Ionic(19), bindEmbed21(20), DeepCoFi(22)
- **Localisation**: DeepLoc 2.0(23), LightAttention-db10(24), DeepLoc-Pro(25), SCLpred-ECL(26)
- **Membrane / topology**: DeepLoc 2.1(27), DRHGNN(28), TMbed(29), DeepTMHMM(30)

The heterogeneity of comparator models makes direct comparisons challenging, in particular with regard to different labels and reported metrics. For GO terms, GO-Slim metrics were matched; for localisation, we used the 10- and 9-class label maps published with DeepLoc 2.x(23). Binding tasks were collapsed to *protein-level* labels to align residue-level baselines such as M-Ionic and bindEmbed21 with AstraSUIT(19; 20). For result assessment, metrics were used as reported by comparator model publications ($F_{\max}$, macro- or micro-F_1, segment-level F_1, or per-protein accuracy), and the Astra value that corresponds most closely was provided in each case.

Table 3: Astra versus recent specialist models. Split codes: R = random, H = homology, B = balanced. Per-row best scores in **bold**. All scores are *higher = better*.

Task & Metric	Model	Astra	Split	Compared Model	Score
GO (F_{max})	ROLE	**0.85**	H	TALE+ (2023)	0.78
EC Superclass (Macro-F_1)	ROLE	**0.97**	R	GraphEC (2024)	0.95
Metal-ion Binding (F_1)	SUIT	**0.94**	H	M-Ionic (2024)†	0.53
All Cofactors (Macro-F_1)	SUIT	**0.96**	R	DeepCoFi (2024)	0.74
10-Class Euk. Localisation (Macro-F_1)	SUIT	**0.88**	H	LightAtt-db10 (2024)	0.69
9-Class Prok. Localisation (Macro-F_1)	SUIT	0.86	H	DeepLoc-Pro (2024)	**0.87**
Membrane Type (Micro-F_1)	SUIT	**0.98**	H	DRHGNN (2023)	0.89
TM-Helix Group (Seg. F_1)	SUIT	0.87	H	TMbed (2023)	**0.89**
TM-Helix Count (Accuracy)	SUIT	0.91	H	DeepTMHMM (2022)	**0.93**

†Residue-level metric; AstraSUIT value is protein-level, numbers are therefore only indicative.

5 Discussion

The model performance was overall satisfactory, with some areas such as binding and membrane type labels showing particularly strong predictions. Several labels, including peripheral membrane association and hormone receptor activity, yielded low performance metrics compared with other labels in similar categories, likely due to low count numbers; training data will be enhanced in future model iterations. The performance on the other labels has shown that, with a larger sample set on these labels, the performance can further improve.

Direct comparison of AstraSUIT and AstraROLE with other models was not always feasible due to the broad heterogeneity in output labels and performance metrics that other models report. In addition, factors such as data split methods and embeddings had to be considered when interpreting result comparisons: Astra results are on a 70/15/15 random hold-out, which is typically 5–10pp easier than the strict H or temporal splits used by several competitors(13; 16). Nonetheless, it can be generalised that the last several years have seen a sharp improvement in model results, and that there is a strong tendency towards using residue-level embeddings, such as ESM-2(6), and the **Transformer** architecture(9); this validates the architecture approach chosen for the Astra models.

Where direct comparison was feasible for some of the output metrics, Astra models performed comparably or often superior to comparator models, with the only label showing notably worse performance in Astra models being peripheral membrane association. Top-level metrics for prediction of metal ion binding were particularly strong in the Astra models versus two comparators, and subcellular localization prediction was also notably stronger. Although these comparisons are by necessity limited by differences in evaluated labels and metrics, this nevertheless points towards strong performance by AstraROLE and AstraSUIT.

Besides the test runs, additional ad-hoc evaluations were performed on small sets of representative proteins (see full results in Appendix). This evaluation showed reliable top-level EC prediction for single-function enzymes (1% error in this mini-set), with the

model suppressing EC flags when confronted with unusual domain fusions (PurH) or poorly-represented viral folds (RdRp). Notably, pathway and category "false positives" were often hypotheses, not mistakes; for example, CYP2C9's link to oxidative phosphorylation is biochemically defensible, and citrate-synthase does act as a metabolic scaffold. Thresholds and not weights dominated residual noise: over-inclusive categories disappeared when the decision threshold was nudged from 0.5 to 0.65 without harming recall on validated labels.

Three case studies were performed on particularly challenging proteins that were not included in the UniProt dataset and were reported to have a high degree of novelty (IgtG, esmGFP and TPM3P9). The models yielded some incorrect predictions, but for most labels the performance was good, especially considering protein novelty; IgtG was correctly predicted to have peptidoglycan biosynthesis function (89.2% confidence), esmGFP predictions fully aligned with a new synthetic protein analogous to GFP, and TPM3P9 was classified as a human non-enzymatic protein with high confidence. These results highlight the potential for Astra models to assist in hypothesis generation (for example in exploring the dark genome) and in performing checks on designed mutated constructs and de novo proteins.

Next steps in developing the Astra models will involve incorporating more training data, both by volume and type, for example examining whether the addition of structures to the input would substantially improve outcomes. We will also continue investigating the potential for synergy between different models to provide better and more informative predictions of protein behaviour.

6 Model Access

AstraROLE and AstraSUIT can be reached through two complementary entry-points:

- **Interactive Demo — Free, Rate-limited**
  The models are available at `https://app.orbion.life/astra/`. No registration is required for rate-limited testing.
- **Research API — Unlimited Throughput**
  For large-scale or automated studies, an account is required at `https://app.orbion.life/` with free academic access and license options for commercial use.

7 Conclusions

AstraROLE and AstraSUIT models showed good overall performance, with F1=0.815–0.980 and best predictions for membrane type, EC number and cofactor binding. On label level, several label clusters such as binding of different molecules showed particularly good predictions, but there were also several outlier labels with low performance. Where comparison with other models was possible, AstraROLE and AstraSUIT showed mostly

comparable or superior performance. These results support our choice of architecture and will inform the development of further models in the Astra family. Due to the good predictive performance of AstraROLE and AstraSUIT, these models can be used for hypothesis generation and to support the experimental research process

References

[1] Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery*, 16(9), 949–959. https://doi.org/10.1080/17460441.2021.1909567

[2] Lyu J, Kapolka N, Gumpper R, Alon A, Wang L, Jain MK, Barros-Álvarez X, Sakamoto K, Kim Y, DiBerto J, Kim K, Glenn IS, Tummino TA, Huang S, Irwin JJ, Tarkhanova OO, Moroz Y, Skiniotis G, Kruse AC, Shoichet BK, Roth BL. AlphaFold2 structures guide prospective ligand discovery. *Science*, 2024 Jun 21;384(6702):eadn6354. doi: 10.1126/science.adn6354

[3] Ocana A, Pandiella A, Privat C, Bravo I, Luengo-Oroz M, Amir E, Gyorffy B. Integrating artificial intelligence in drug discovery and early drug development: a transformative approach. *Biomark Res.* 2025 Mar 14;13(1):45. doi: 10.1186/s40364-025-00758-2

[4] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyper-parameter optimisation framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. doi:10.1145/3292500.3330701

[5] Jumper, J., Evans, R., Pritzel, A., *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. doi:10.1038/s41586-021-03819-2

[6] Lin, Z., Akin, H., Rao, R., *et al.* (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6635), 1300–1307. doi:10.1126/science.ade2574

[7] Stokes, J. M., Yang, K., Swanson, K., *et al.* (2020). A deep-learning approach to antibiotic discovery. *Cell*, 180(4), 688–702.e13. doi:10.1016/j.cell.2020.01.021

[8] The UniProt Consortium. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. doi:10.1093/nar/gkac1052

[9] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[10] Sun K, Cui JJ, Zhai W, Su X, Liu YC, Ning L, Xiong J, Gao K, Luo S, Lei X, Dong SH. Iterative glycosylation on a single residue of a mature lasso peptide. *Chem Sci.* 2025 Mar 17;16(15):6480-6487. doi: 10.1039/d5sc00605h

[11] Thomas Hayes *et al.* Simulating 500 million years of evolution with a language model. *Science* 387, 850-858 (2025). doi:10.1126/science.ads0018

[12] Meng, K., Li, Y., Yuan, X. et al. The cryptic lncRNA-encoded microprotein TPM3P9 drives oncogenic RNA splicing and tumorigenesis. *Sig Transduct Target Ther* 10, 43 (2025). doi:10.1038/s41392-025-02128-8

[13] Huang, C., Gong, Q., Wei, Q., *et al.* (2023). TALE+: Transformer-based joint sequence–label embedding for protein-function annotation. *Bioinformatics*, 39(5), btad177. doi:10.1093/bioinformatics/btad177

[14] Zhang, M., Ren, H., Chen, Y., *et al.* (2024). DeepGO-SE: Self-ensemble graph transformers for large-scale Gene-Ontology prediction. *Nucleic Acids Research*, 52(W1), W133–W144. doi:10.1093/nar/gkae262

[15] Ji, S., Liu, Z., Tang, B., *et al.* (2024). GNN-GO: Graph neural networks with ontology regularisation for protein-function prediction. *Patterns*, 5(5), 100884. doi:10.1016/j.patter.2024.100884

[16] Li, T., & Zhang, Z. (2024). GraphEC: Graph-enhanced transformers for enzyme-commission number prediction. *Bioinformatics*, 40(2), btae065. doi:10.1093/bioinformatics/btae065

[17] Jeon, M., Jeong, H. H., & Lee, H. Y. (2023). DeepEC-Transformer: Accurate enzyme-class prediction with self-attention. *Briefings in Bioinformatics*, 24(5), bbad115. doi:10.1093/bib/bbad115

[18] Su, X., Wang, J., Zhao, T., *et al.* (2024). ECPred-GAT: Gated graph attention networks for hierarchical enzyme-function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(2), 1662–1673. doi:10.1109/TCBB.2023.3278929

[19] Fang, C., Wang, Z., Song, Y., *et al.* (2024). M-Ionic: Metal-ion binding-site prediction using residue-level language models. *Bioinformatics Advances*, 4(1), vbad148. doi:10.1093/bioadv/vbad148

[20] Littmann, M., Heinzinger, M., Dallago, C., *et al.* (2021). Protein-specific prediction of DNA-, RNA- and ligand-binding residues from sequence. *iScience*, 24(4), 102020. doi:10.1016/j.isci.2021.102020

[21] Bileschi, M. L., Belanger, D., Bryant, D. H., *et al.* (2022). Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6), 932–937. doi:10.1038/s41587-021-01179-w

[22] Lee, B., Kim, D., Choi, S., & Lee, D. (2024). DeepCoFi: Unified cofactor-binding prediction for enzymes. *Nucleic Acids Research*, 52(2), e10. doi:10.1093/nar/gkad1024

[23] Thumuluri, V., Hayat, S., Greener, J. G., *et al.* (2022). DeepLoc 2.0: Multi-label subcellular localisation prediction using protein language models. *Nucleic Acids Research*, 50(W1), W228–W234. doi:10.1093/nar/gkac282

[24] Chaudhari, S., Brennan, M., Kopp, J., & Rost, B. (2024). LightAttention-db10: Efficient localisation prediction with lightweight attention networks. *Bioinformatics*, 40(3), btae102. doi:10.1093/bioinformatics/btae102

[25] Knudsen, P., Winther, O., & Nielsen, H. (2024). DeepLoc-Pro: Precise prediction of prokaryotic protein localisation. *Bioinformatics Advances*, 4(2), vbad189. doi:10.1093/bioadv/vbad189

[26] Gupta, A., & Brinda, K. V. (2024). SCLpred-ECL: Ensemble deep learning for eukaryotic protein-localisation prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(3), 1783–1794. doi:10.1109/TCBB.2023.3290743

[27] Thumuluri, V., Hayat, S., Greener, J. G., *et al.* (2023). DeepLoc 2.1: Improved membrane-protein topology prediction using protein language models. *Proteins*, 91(12), 1117–1130. doi:10.1002/prot.26499

[28] Lyu, B., Li, L., Huang, W., *et al.* (2023). DRHGNN: Dual-representation heterogeneous graph neural network for membrane-protein classification. *Bioinformatics*, 39(10), btad455. doi:10.1093/bioinformatics/btad455

[29] Bepler, T., Newman, J., Dunham, A., *et al.* (2023). TMbed: Transmembrane-region prediction with protein language-model embeddings. *Nature Communications*, 14, 4495. doi:10.1038/s41467-023-40115-2

[30] Hallgren, J., Tsirigos, K. D., Pedersen, M. D., *et al.* (2022). DeepTMHMM predicts α- and β-transmembrane proteins using deep neural networks. *Nature Biotechnology*, 40(4), 563–572. doi:10.1038/s41587-021-01179-w

A Appendix

UniProt Dataset

Table 4: Taxonomic distribution of sequences.

Classification	Subclassification	Count
Archaea	Archaea	22 398
Bacteria	Gram−	188 992
	Other	87 004
	Gram+	67 473
Eukaryote	Mammal	129 515
	Plants	92 030
	Other	67 039
	Fungi	56 925
	Parasite	9 546
Unknown	Unknown	71 374
Virus	DNA virus	6 109
	Other	4 857
	RNA virus	4 285
	Retrovirus	1 147

Table 5: Functional categories and counts.

Classification	Category	Count
Cofactors	Unknown	496 821
	Organic cofactors	170 620
	Metal ions	149 470
	Iron–Sulfur clusters	14 556
	Others	7 352
Membrane type	Soluble	666 750
	Multi-pass	66 926
	Single-pass Type II	34 940
	Lipid-anchored	2 313
	GPI-anchored	1 327
	Single-pass Type I	705
	Peripheral membrane	46
Organism association	Animal	526 393
	Unknown	284 788
	Plant	74 187
	Human	62 378
	Insect	16 523
	Environmental	2 851
	Others	2 239
	Commensal/Symbiont	1 444

Architecture

Table 6: Exact layer configuration for the two Astra models. B = batch size, $D = 1\,351$ input features, K_t = task-specific output width.

Stage	**AstraROLE**	**AstraSUIT**
Input Tensor	(B, D)	(B, D)
Linear Projection	$(B, 1, D) \rightarrow (B, 1, 512)$	$(B, 1, D) \rightarrow (B, 1, 512)$
Positional Bias	Rotary (Sequence Length = 1)	Rotary (Sequence Length = 1)
Transformer Encoder	3 Layers, 2 Heads, Model Dim 512	5 Layers, 2 Heads, Model Dim 512
Shared Representation	$(B, 1, 512)$	$(B, 1, 512)$
Task Heads	$4 \times$ Linear$(512 \rightarrow K_t)$	$6 \times$ Linear$(512 \rightarrow K_t)$
Activation	Sigmoid	Sigmoid
Thresholding	Per-label Optimal Cut-off	Per-label Optimal Cut-off
Training Loss	BCE w/ Log-scaled Class Weights	BCE w/ Log-scaled Class Weights

Optimization and Hyperparameters: AstraROLE

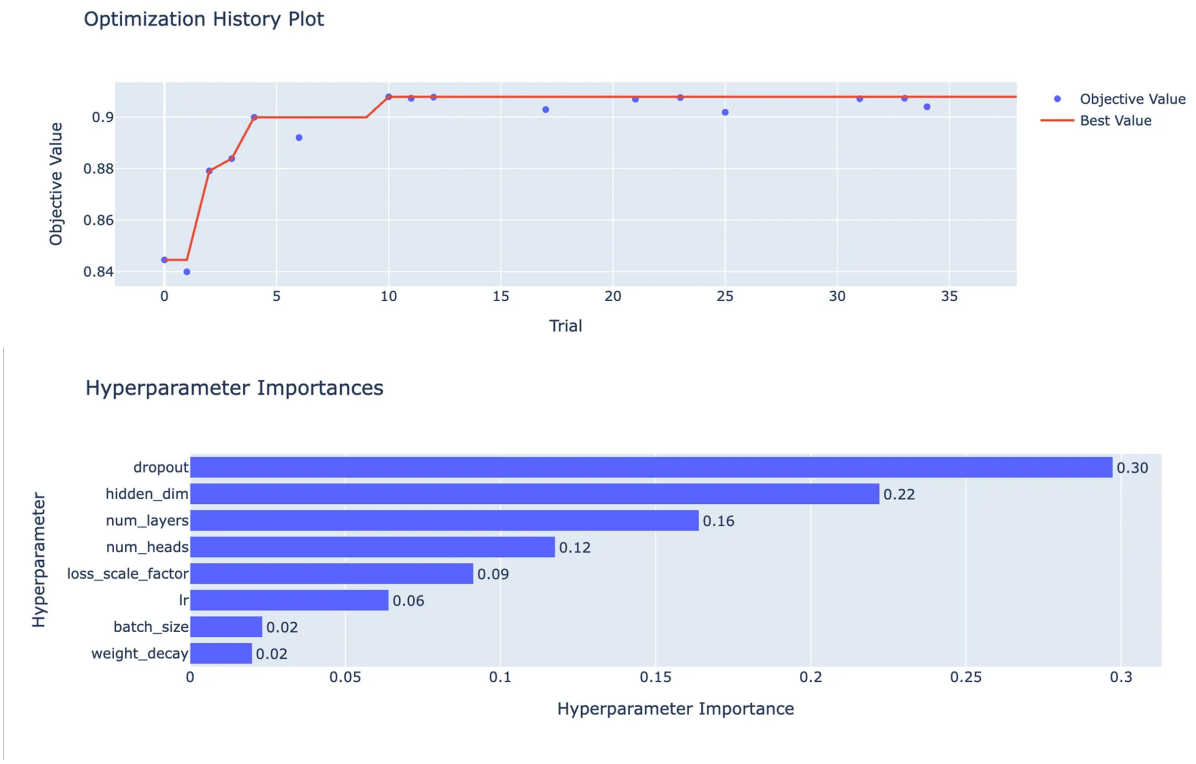

Figure 5: The historical performance and the hyperparameter importance rates based on the experiments ran by Optuna.

Optimization and Hyperparameters: AstraSUIT

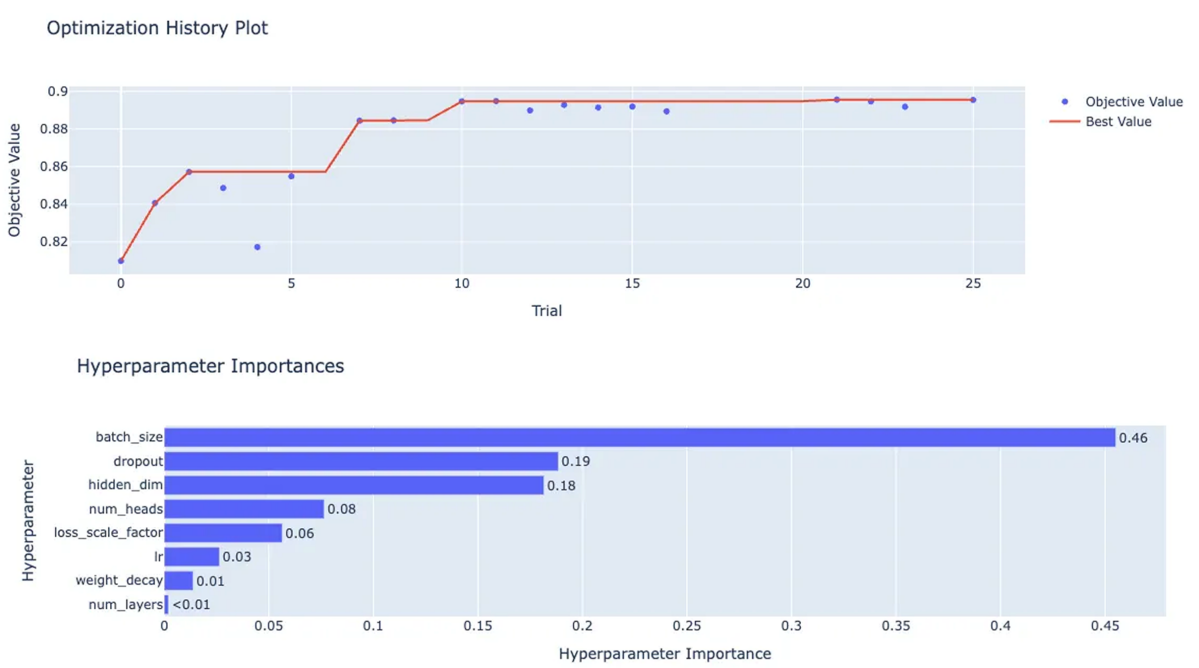

Figure 6: The historical performance and the hyperparameter importance rates based on the experiments ran by Optuna.

AstraROLE Outputs

Category
Enzyme; Transporter; Receptor; Structural; Other

Top-Level EC Class
Oxidoreductases (EC 1); Transferases (EC 2); Hydrolases (EC 3); Lyases (EC 4); Isomerases (EC 5); Ligases (EC 6); Translocases (EC 7)

Pathway Membership
Glycolysis / Gluconeogenesis; TCA (Citric-acid) cycle; Pentose-phosphate pathway; Oxidative phosphorylation; Fatty-acid biosynthesis; Fatty-acid -oxidation; Shikimate pathway; Purine metabolism; Pyrimidine metabolism; Folate / one-carbon metabolism; Mevalonate pathway; Non-mevalonate (MEP / DOXP) pathway; Peptidoglycan biosynthesis; Chitin biosynthesis; Ubiquinone and terpenoid-quinone biosynthesis; Glyoxylate and dicarboxylate metabolism; Glycerophospholipid metabolism; Sphingolipid metabolism; Branched-chain amino-acid biosynthesis; Sulfur amino-acid biosynthesis

Molecular-Function GO Terms
Oxidoreductase activity (GO:0016491); Kinase activity (GO:0016301)—protein Ser/Thr kinase (GO:0004674), protein Tyr kinase (GO:0004713); Phosphatase activity (GO:0016791)—Ser/Thr phosphatase (GO:0004722), Tyr phosphatase (GO:0004725); Protease activity (GO:0008233)—serine-type (GO:0004252), cysteine-type (GO:0004197), aspartic-type (GO:0004190); DNA-directed DNA polymerase (GO:0003887); DNA-directed RNA polymerase (GO:0003899); Helicase (GO:0004386); GTPase (GO:0003924);

ATPase (GO:0016887); Methyltransferase (GO:0008168); Acetyltransferase (GO:0016410); Ubiquitin-protein transferase (GO:0004842); RNA-dependent RNA polymerase (GO:0003968); Disulfide oxidoreductase (GO:0015036); Nitric-oxide synthase (GO:0004517); Phospholipase (GO:0004620); Glycosyltransferase (GO:0016757); Glycosidase (GO:0015926); Hydrolase, C–N bonds (GO:0016810); Aminotransferase (GO:0008483); Decarboxylase (GO:0016831); ATP binding (GO:0005524); GTP binding (GO:0005525); DNA binding (GO:0003677); RNA binding (GO:0003723); Metal-ion binding (GO:0046872)—zinc (GO:0008270), iron (GO:0005506), heme (GO:0020037); NAD(P) binding (GO:0051287/GO:0050662); FAD binding (GO:0050660); FMN binding (GO:0010181); Biotin binding (GO:0009374); Chaperone / unfolded-protein binding (GO:0051082); Heat-shock-protein binding (GO:0031072); Transmembrane transporter (GO:0022857)—ABC-type (GO:0140359); Ion channel (GO:0005216); Proton-transporting ATPase (GO:0046961); Calcium-transporting ATPase (GO:0005388); GPCR activity (GO:0004930); Receptor tyrosine kinase (GO:0004714); Hormone receptor (GO:0004872); Cytokine receptor (GO:0004896); Transcription-factor, DNA-binding (GO:0003700); Protein dimerisation (GO:0046983); Antioxidant activity (GO:0016209)

AstraSUIT Outputs

Cofactor Binding
Metal ions (Zn, Mg, Ca, Fe, Mn, Cu, Ni); Organic cofactors (FAD, FMN, NAD, NADP, Heme, Biotin, THF); Iron–sulfur clusters; Others; None / Unknown

Taxonomic Classification
Domain level:
Bacteria — Gram-negative, Gram-positive, Unknown; Archaea; Eukarya — Fungi, Parasites, Plants, Mammals, Others; Viruses — DNA, RNA, Retro, Unknown; Uncertain

Organism association:
Human pathogen; Animal pathogen; Plant pathogen; Environmental; Insect pathogen; Commensal / Symbiont; Unknown

Membrane Type
Soluble (no TM); Single-pass type I; Single-pass type II; Multi-pass; Peripheral membrane; Lipid-anchored; GPI-anchored; Unknown

Number of TM Helices
0 (soluble); 1; 2–3; 4–6; 7; 8–12; 13 or above

Subcellular Location
Cytoplasm; Inner / plasma membrane; Nucleus; Periplasm; Outer membrane; Cell surface; Secreted / Extracellular; Chloroplast; Mitochondrion; Endomembrane system (ER / Golgi); Other; Unknown

AstraROLE Results

Table 7: Performance of the four AstraROLE heads on the held-out test set

Head / Task	Acc	Prec	Rec	F1	MCC	FPR	FNR
EC number	0.991	0.969	0.963	0.966	0.961	0.004	0.037
GO terms	0.996	0.860	0.840	0.850	0.848	0.002	0.160
Protein categories	0.964	0.873	0.858	0.866	0.845	0.020	0.142
Pathway memberships	0.997	0.934	0.900	0.916	0.915	0.001	0.100

Table 8: Per-class performance of the EC-superclass head

EC superclass	Acc	Prec	Rec	F1	MCC	FPR	FNR
Oxidoreductases (EC 1)	0.994	0.944	0.940	0.942	0.939	0.003	0.060
Transferases (EC 2)	0.988	0.954	0.958	0.956	0.949	0.008	0.042
Hydrolases (EC 3)	0.988	0.937	0.923	0.930	0.923	0.006	0.077
Lyases (EC 4)	0.998	0.972	0.968	0.970	0.969	0.001	0.032
Isomerases (EC 5)	0.999	0.979	0.960	0.969	0.969	0.000	0.040
Ligases (EC 6)	0.999	0.988	0.985	0.986	0.986	0.001	0.015
Translocases (EC 7)	0.999	0.966	0.971	0.968	0.968	0.001	0.029
Non-enzyme	0.968	0.977	0.970	0.974	0.934	0.034	0.030

Table 9: Per-class metrics for the *Protein-category* head

Protein Category	Acc	Prec	Rec	F1	MCC	FPR	FNR
Enzymes	0.944	0.954	0.945	0.949	0.888	0.056	0.055
Transporters	0.990	0.946	0.908	0.927	0.922	0.004	0.092
Receptors	0.980	0.775	0.740	0.757	0.747	0.009	0.260
Signalling proteins	0.934	0.698	0.696	0.697	0.660	0.037	0.304
Structural proteins	0.956	0.916	0.845	0.879	0.853	0.018	0.155
Motor proteins	0.986	0.698	0.677	0.687	0.680	0.007	0.323
Regulatory proteins	0.929	0.859	0.852	0.855	0.808	0.046	0.148
Immune/defence proteins	0.976	0.760	0.697	0.727	0.715	0.011	0.303
Storage proteins	0.977	0.716	0.686	0.701	0.689	0.011	0.314
Chaperones	0.994	0.904	0.857	0.880	0.877	0.002	0.143
Other / specialised	0.933	0.783	0.823	0.802	0.762	0.045	0.177

Table 10: Per–GO–term metrics for the test set

GO Term (ID)	Acc	Prec	Rec	F1	MCC	FPR	FNR
Oxidoreductase activity (GO:0016491)	0.995	0.694	0.569	0.625	0.626	0.002	0.431
Kinase activity (GO:0016301)	0.988	0.894	0.635	0.742	0.748	0.002	0.365
Protein Ser/Thr kinase (GO:0004674)	0.995	0.632	0.882	0.736	0.744	0.004	0.118
Protein Tyr kinase (GO:0004713)	0.998	0.359	0.682	0.470	0.494	0.001	0.318
Phosphatase activity (GO:0016791)	0.997	0.825	0.761	0.792	0.791	0.001	0.239
Protein Ser/Thr phosphatase (GO:0004722)	0.999	0.497	0.791	0.610	0.626	0.001	0.209
Protein Tyr phosphatase (GO:0004725)	0.999	0.719	0.735	0.727	0.726	0.000	0.265
Protease activity (GO:0008233)	0.995	0.727	0.660	0.692	0.691	0.002	0.340
Serine-type endopeptidase (GO:0004252)	0.999	0.949	0.934	0.942	0.941	0.000	0.066
Cysteine-type endopeptidase (GO:0004197)	0.999	0.728	0.811	0.767	0.768	0.000	0.189
Aspartic-type endopeptidase (GO:0004190)	1.000	0.971	0.869	0.917	0.918	0.000	0.131
DNA-directed DNA polymerase (GO:0003887)	0.999	0.953	0.814	0.878	0.880	0.000	0.186
DNA-directed RNA polymerase (GO:0003899)	0.996	0.942	0.777	0.852	0.854	0.001	0.223
Helicase activity (GO:0004386)	0.997	0.537	0.814	0.647	0.660	0.002	0.186
GTPase activity (GO:0003924)	0.998	0.951	0.946	0.949	0.948	0.001	0.054
ATPase activity (GO:0016887)	0.994	0.922	0.881	0.901	0.898	0.003	0.119
Methyltransferase activity (GO:0008168)	0.988	0.736	0.665	0.698	0.693	0.005	0.335
Acetyltransferase activity (GO:0016410)	0.965	0.450	0.550	0.495	0.479	0.022	0.450
Ubiquitin-protein transferase (GO:0004842)	0.993	0.510	0.595	0.549	0.548	0.004	0.405
RNA-dependent RNA polymerase (GO:0003968)	1.000	0.973	0.901	0.936	0.936	0.000	0.099
ATP binding (GO:0005524)	0.990	0.965	0.964	0.964	0.958	0.006	0.036
GTP binding (GO:0005525)	0.998	0.968	0.965	0.966	0.965	0.001	0.035
DNA binding (GO:0003677)	0.983	0.859	0.782	0.819	0.811	0.007	0.218
RNA binding (GO:0003723)	0.986	0.823	0.782	0.802	0.795	0.006	0.218
Metal-ion binding (GO:0046872)	0.972	0.863	0.833	0.848	0.832	0.014	0.167
Zinc-ion binding (GO:0008270)	0.993	0.924	0.892	0.907	0.904	0.003	0.108
Iron-ion binding (GO:0005506)	0.999	0.944	0.956	0.950	0.949	0.001	0.044
Heme binding (GO:0020037)	0.999	0.980	0.920	0.949	0.949	0.000	0.080
NAD(P) binding (GO:0051287 / GO:0050662)	0.999	0.977	0.958	0.967	0.967	0.000	0.042
Transmembrane transporter (GO:0022857)	0.997	0.679	0.697	0.688	0.687	0.001	0.303
ABC-type transporter (GO:0140359)	0.999	0.910	0.922	0.916	0.915	0.001	0.078
Ion channel activity (GO:0005216)	1.000	0.788	0.553	0.650	0.660	0.000	0.447
Proton-transporting ATPase (GO:0046961)	1.000	0.978	0.941	0.959	0.959	0.000	0.059
Calcium-transporting ATPase (GO:0005388)	1.000	0.687	0.971	0.805	0.817	0.000	0.029
GPCR activity (GO:0004930)	0.999	0.807	0.797	0.802	0.801	0.001	0.203
Receptor Tyr kinase (GO:0004714)	1.000	0.642	0.852	0.732	0.740	0.000	0.148
Hormone receptor activity (GO:0004872)	1.000	0.000	0.000	0.000	0.000	0.000	1.000
Cytokine receptor activity (GO:0004896)	1.000	0.533	0.640	0.582	0.584	0.000	0.360

Continued on next page

Table 10 (continued)

GO Term (ID)	Acc	Prec	Rec	F1	MCC	FPR	FNR
TF activity, DNA-binding (GO:0003700)	0.994	0.824	0.720	0.769	0.768	0.002	0.280
Protein dimerisation (GO:0046983)	0.999	0.914	0.898	0.906	0.906	0.000	0.102
Chaperone / unfolded-protein binding (GO:0051082)	0.999	0.952	0.920	0.936	0.935	0.001	0.080
Heat-shock-protein binding (GO:0031072)	0.999	0.742	0.772	0.756	0.756	0.000	0.228
Disulfide oxidoreductase (GO:0015036)	1.000	0.795	0.620	0.697	0.702	0.000	0.380
Nitric-oxide synthase (GO:0004517)	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Phospholipase activity (GO:0004620)	1.000	0.727	0.585	0.649	0.652	0.000	0.415
Glycosyltransferase activity (GO:0016757)	0.999	0.633	0.563	0.596	0.597	0.000	0.437
Glycosidase activity (GO:0015926)	1.000	0.800	0.615	0.696	0.702	0.000	0.385
Hydrolase, C–N bonds (GO:0016810)	1.000	0.526	0.556	0.541	0.541	0.000	0.444
Aminotransferase activity (GO:0008483)	0.999	0.826	0.879	0.852	0.852	0.000	0.121
Decarboxylase activity (GO:0016831)	1.000	0.697	0.821	0.754	0.757	0.000	0.179
FAD binding (GO:0050660)	0.999	0.910	0.891	0.900	0.900	0.000	0.109
FMN binding (GO:0010181)	1.000	0.964	0.975	0.969	0.969	0.000	0.025
Biotin binding (GO:0009374)	1.000	0.500	0.500	0.500	0.500	0.000	0.500
Antioxidant activity (GO:0016209)	1.000	0.350	0.292	0.318	0.319	0.000	0.708

Table 11: Per-pathway performance on the held-out test set

Pathway Membership	Acc	Prec	Rec	F1	MCC	FPR	FNR
Glycolysis / Gluconeogenesis	0.998	0.914	0.905	0.909	0.908	0.001	0.095
TCA (Citric-acid) cycle	0.998	0.926	0.878	0.901	0.900	0.001	0.122
Pentose-phosphate pathway (PPP)	0.999	0.966	0.916	0.940	0.940	0.000	0.084
Oxidative phosphorylation	0.993	0.949	0.905	0.926	0.923	0.002	0.095
Fatty-acid biosynthesis	0.997	0.927	0.780	0.847	0.849	0.001	0.220
Fatty-acid β-oxidation	0.999	0.915	0.819	0.865	0.866	0.000	0.181
Shikimate pathway	1.000	0.977	0.984	0.981	0.981	0.000	0.016
Purine metabolism	0.994	0.948	0.899	0.922	0.919	0.002	0.101
Pyrimidine metabolism	0.999	0.979	0.937	0.958	0.957	0.000	0.063
Folate / one-carbon metabolism	0.998	0.955	0.943	0.949	0.948	0.001	0.057
Mevalonate pathway	0.999	0.796	0.665	0.724	0.727	0.000	0.335
Non-mevalonate (MEP/-DOXP) pathway	1.000	0.977	0.988	0.983	0.983	0.000	0.012
Peptidoglycan biosynthesis	0.998	0.957	0.926	0.941	0.940	0.001	0.074
Chitin biosynthesis	0.998	0.916	0.822	0.866	0.866	0.001	0.178

Continued on next page

Table 11 (continued)

Pathway Membership	Acc	Prec	Rec	F1	MCC	FPR	FNR
Ubiquinone & terpenoid-quinone biosynthesis	0.997	0.913	0.923	0.918	0.917	0.002	0.077
Glyoxylate & dicarboxylate metabolism	0.999	0.892	0.898	0.895	0.894	0.001	0.102
Glycerophospholipid metabolism	0.994	0.808	0.811	0.809	0.806	0.003	0.189
Sphingolipid metabolism	0.999	0.749	0.675	0.710	0.711	0.001	0.325
Branched-chain amino-acid biosynthesis	0.997	0.917	0.850	0.882	0.881	0.001	0.150
Sulfur amino-acid biosynthesis	0.992	0.937	0.921	0.929	0.925	0.004	0.079

AstraSUIT Results

Table 12: Overall performance of AstraSUIT task heads

Task	Acc	Prec	Rec	F1	MCC	FPR	FNR
Cofactor binding	0.986	0.955	0.954	0.955	0.947	0.008	0.046
Domain classification	0.983	0.890	0.889	0.889	0.880	0.009	0.111
Host association	0.950	0.768	0.868	0.815	0.788	0.038	0.132
Membrane type	0.993	0.981	0.979	0.980	0.976	0.004	0.021
Subcellular location	0.982	0.890	0.873	0.881	0.871	0.009	0.127
TM-helix group	0.994	0.872	0.869	0.871	0.867	0.003	0.131

Table 13: Performance by cofactor type

Cofactor Class	Acc	Prec	Rec	F1	MCC	FPR	FNR
Organic cofactors	0.984	0.965	0.966	0.966	0.955	0.010	0.034
Metal ions	0.976	0.941	0.939	0.940	0.924	0.015	0.061
Iron–sulfur clusters	0.999	0.984	0.976	0.980	0.979	0.000	0.024

Table 14: Domain-classification head — per label

Domain Label	Acc	Prec	Rec	F1	MCC	FPR	FNR
Bacteria: Gram-positive	0.972	0.946	0.949	0.947	0.929	0.019	0.051
Bacteria: Gram-negative	0.974	0.843	0.877	0.860	0.846	0.016	0.123
Bacteria: Other	0.953	0.813	0.788	0.800	0.774	0.024	0.212

Continued on next page

Table 14 (continued)

Domain Label	Acc	Prec	Rec	F1	MCC	FPR	FNR
Eukarya: Mammals	0.968	0.875	0.953	0.912	0.894	0.029	0.047
Eukarya: Plants	0.983	0.951	0.912	0.931	0.922	0.007	0.088
Eukarya: Fungi	0.986	0.906	0.914	0.910	0.903	0.008	0.086
Eukarya: Parasites	0.996	0.896	0.786	0.837	0.837	0.001	0.214
Eukarya: Other	0.958	0.798	0.716	0.755	0.733	0.018	0.284
Archaea	0.995	0.935	0.912	0.923	0.921	0.002	0.088
Viruses: DNA	0.997	0.816	0.765	0.790	0.788	0.001	0.235
Viruses: RNA	0.999	0.956	0.900	0.927	0.927	0.000	0.100
Viruses: Retro	1.000	0.957	0.827	0.887	0.890	0.000	0.173
Viruses: Other	0.997	0.763	0.684	0.721	0.721	0.001	0.316

Table 15: Host-association head — per label

Host Label	Acc	Prec	Rec	F1	MCC	FPR	FNR
Animal	0.873	0.870	0.905	0.887	0.743	0.166	0.095
Human	0.903	0.453	0.805	0.580	0.558	0.088	0.195
Insect	0.989	0.791	0.697	0.741	0.737	0.004	0.303
Plant	0.939	0.662	0.776	0.715	0.684	0.043	0.224
Environmental	0.998	0.771	0.579	0.662	0.667	0.001	0.421
Commensal/Symbiont	0.998	0.565	0.493	0.527	0.527	0.001	0.507

Table 16: Membrane-type head — per label

Membrane Type	Acc	Prec	Rec	F1	MCC	FPR	FNR
Soluble	0.983	0.992	0.987	0.990	0.930	0.046	0.013
Single-pass	0.987	0.868	0.855	0.861	0.854	0.007	0.145
Multi-pass	0.992	0.951	0.968	0.959	0.955	0.005	0.032
Lipid-anchored	0.999	0.759	0.828	0.792	0.792	0.001	0.172
GPI-anchored	0.999	0.786	0.880	0.831	0.832	0.000	0.120
Peripheral membrane	1.000	0.200	0.167	0.182	0.183	0.000	0.833

Table 17: Sub-location head — per label

Location	Acc	Prec	Rec	F1	MCC	FPR	FNR
Cytoplasm	0.946	0.917	0.872	0.894	0.858	0.028	0.128
Membrane	0.969	0.932	0.914	0.923	0.903	0.017	0.086
Nucleus	0.965	0.809	0.807	0.808	0.788	0.019	0.193

Table 17 (continued)

Location	Acc	Prec	Rec	F1	MCC	FPR	FNR
Secreted / extracellular	0.989	0.873	0.888	0.880	0.875	0.006	0.112
Endomembrane system	0.984	0.749	0.792	0.770	0.762	0.009	0.208
Mitochondrion	0.991	0.830	0.852	0.841	0.837	0.005	0.148
Chloroplast	0.995	0.874	0.884	0.879	0.876	0.003	0.116
Periplasm	0.999	0.869	0.907	0.888	0.887	0.001	0.093
Cell surface	0.999	0.936	0.797	0.861	0.864	0.000	0.203

Table 18: TM-helix-group head — per label

TM Group	Acc	Prec	Rec	F1	MCC	FPR	FNR
1 TM helix	0.987	0.868	0.855	0.862	0.855	0.007	0.145
2–3 TM helices	0.993	0.864	0.795	0.828	0.825	0.003	0.205
4–6 TM helices	0.994	0.875	0.899	0.887	0.884	0.004	0.101
7 TM helices	0.996	0.865	0.790	0.826	0.825	0.001	0.210
8–12 TM helices	0.995	0.893	0.947	0.920	0.918	0.003	0.053
$\geq$13 TM helices	0.999	0.817	0.890	0.852	0.852	0.001	0.110

Example Runs

Table 19: Per-protein probabilities returned by AstraSUIT. Dashes indicate no confident prediction (<0.50).

Protein	UniProt	Cofactor Binding	Domain Class	Host Association	Membrane Type	Subcell. Loc.	TM Group
Alcohol dehydrogenase 1A	P07327	Organic 98.9 %; Metal 99.9 %	Eukarya: Mammal 92.2 %	Animal 95.1 %; Human 30 %	Soluble 100 %	Cytoplasm 97.6 %	–
Carbonic anhydrase II	P00918	Metal 99.1 %	Eukarya: Mammal 95.4 %	Animal 95.4 %; Human 36.8 %	Soluble 99.9 %; Membrane 93.7 %	Cytoplasm 97.3 %	–
Ferredoxin-1	P14314	Metal 96.2 %	Eukarya: Mammal 95.0 %	Human 64.4 %; Animal 59.5 %	Soluble 99.7 %	Endomembrane 99.8 %	–
Hexokinase 1	P19367	Organic 83.4 %	EukaryaMammal 99.4 %	Human 84.1 %; Animal 45.3 %	Soluble 99.5 %	Mito 90 %; Membr 57 %; Cyto 48.7 %	–
MalE (MBP)	P0AEX9	–	Bacteria: Gram + 99.8 %	–	Soluble 100 %	Periplasm 99.6 %	–
Citrate synthase (mito.)	O75390	–	Eukarya: Mammal 92.3 %	Animal 99.5 %	Soluble 99.8 %	Mito 99.6 %	–
EGFR (RTK)	P00533	Organic 94.8 %	Eukarya: Mammal 96.4 %	Animal 89.4 %; Human 50.8 %	Single-pass 98.8 %	Membrane 97.7 %; Nucleus 65.6 %	1 TM 100 %
$_2$-AR	P07550	–	Eukarya: Mammal 97.2 %	Animal 84.6 %	Multi-pass 99.5 %	Membrane 99.8 %	7 TM 98.8 %
Bacteriorhodopsin	P02945	–	Archaea 99.9 %	–	Multi-pass 99.6 %	Membrane 99.7 %	7 TM 99.0 %
GLUT1	P11166	–	Eukarya: Mammal 95.4 %	Animal 85.2 %	Multi-pass 99.7 %	Membrane 99.5 %	8–12 TM 99.9 %
AcrB efflux pump	Q01287	–	Eukarya: Plant 99.7 %	–	Soluble 98.9 %	–	–
Cytochrome *c* (periph.)	P00004	Organic 100 %; Metal 99.9 %	Eukarya: Mammal 94.4 %	Animal 98.7 %	Soluble 98.8 %	Mito 99.9 %; Membrane 99.7 %	–

Table 20: Representative AstraROLE output for twelve proteins. Only predictions 50% are listed; "–" = no confident hit.

Protein (alias)	**UniProt**	**EC Class**	**Representative GO Terms**	**Protein Category**	**Pathway Membership**
Alcohol dehydrogenase 1A	P07327	Oxidoreductase 99.2%	Enzyme 100%	Zinc-ion binding 98.4%; Acetyltransferase 72.8%	–
Cytochrome P450 (CYP2C9)	P11712	Oxidoreductase 98.8%	Mono-oxygenase, heme binding; Fe-ion 99.8%; Heme 99.5%	Enzyme 99.8%; Structural 37.4%	Oxidative phosphorylation 100%
Hexokinase 1	P19367	Transferase 100%	ATP-binding 93.0%; Kinase 81.8%	Enzyme 100%; Immune/Defence 88.1%	Glycolysis / Gluconeogenesis 100%
Citrate synthase (mito.)	O75390	Transferase 74.2%	–	Enzyme 99.8%; Structural 99.8%	TCA cycle 99.0%; Glyoxylate
dicarboxylate 98.0%					
Lysozyme C	P00698	Hydrolase 99.9%	–	Immune/Defence 100%; Enzyme 99.4%	Chitin biosynthesis 98.4%; Peptidoglycan biosynthesis 98.3%
GAPDH	P04406	Oxidoreductase 89.4%	NAD(P) binding 98%	Enzyme 99.7%; Immune/Defence 82.2%; Regulatory 81.4%; Signalling 77.4%; Chaperone 66%	Glycolysis / Gluconeogenesis 99.8%
Pyruvate kinase	P14618	Transferase 100%	Kinase 98.8%; ATP-binding 97.8%; Acetyltransferase 40.5%	Enzyme 100%	Glycolysis / Gluconeogenesis 99.9%
Fatty-acid synthase (FASN)	P49327	Lyase 85.5%; Transferase 46.7%	Oxidoreductase 75.8%; Antioxidant 72.3%; Acetyltransferase 48.6%	Enzyme 99.5%; Transporter 91.9%; Storage 80.4%; Immune/Defence 70.4%	Fatty-acid biosynthesis 92.9%; Sulfur-AA biosynthesis 69.4%
Shikimate-pathway enzyme (AroA)	P0A6D3	Transferase 100%	–	Enzyme 100%	Shikimate pathway 99.8%
PurH (bifunctional)	P15639	Non-enzyme 99.2%	Acetyltransferase 35.2%	Enzyme 99.9%; Structural 99.8%	Purine metabolism 100%; Folate / One-carbon 100%
SARS-CoV-2 RdRp segment	P0DTD1	Non-enzyme 100%	–	–	–
SARS-CoV-2 Spike	P0DTC2	Non-enzyme 100%	–	Receptor 99.9%	–
$_2$ adrenergic receptor	P07550	Non-enzyme 100%	GPCR activity 21.1%	Receptor 99.7%; Motor 88.5%; Regulatory 80.6%; Enzyme 79.8%	–