Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press



Bacterial protein function prediction via multimodal deep learning

Lukas Miklautz, Leyden Fernandez, Giulia Muzio, et al.

Genome Res. published online November 6, 2026
Access the most recent version at doi:10.1101/gr.282132.126

P<P Published online November 6, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Bacterial protein function prediction via multimodal deep learning

Lukas Miklautz[1,†], Leyden Fernandez[2,3,†], Giulia Muzio[4,5], Michael Adamer[4,5], Karsten Borgwardt[1,4,5,‡] and Kemal Avican[2,3,6,‡,*]

[1]Department of Machine Learning and Systems Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[2]Integrated Science Lab (IceLab), Umeå University, 901 87 Umeå, Sweden

[3]Department of Molecular Biology, Umeå University, 901 87 Umeå, Sweden

[4]Department of Biosystems Science and Engineering, ETH Zurich, 4056 Basel, Switzerland

[5]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland

[6]Umeå Centre for Microbial Research (UCMR), Umeå University, 901 87 Umeå, Sweden

[†]Equal contribution. [‡]Equal supervision.

[*]Corresponding author: Kemal Avican, Department of Molecular Biology, Umeå University, Försörjningsvägen 2A, 901 87 Umeå, Sweden; email: kemal.avican@umu.se; phone: +46 90 786 78 29

Running title: Protein function in bacteria

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Bacterial proteins are specialized with extensive functional diversity for survival in diverse and stressful environments. A significant portion of these proteins remains functionally uncharacterized, limiting our understanding of bacterial survival mechanisms. Hence, we develop Deep Expression STructure (DeepEST), a multimodal deep learning framework that predicts protein function in bacteria by assigning Gene Ontology (GO) terms. DeepEST comprises two modules: a multi-layer perceptron that uses gene expression and gene location, and a protein structure-based module that uses predicted structures and amino acid sequences. Within DeepEST, we integrate these modules through a learnable weighted linear combination and introduce a novel masked loss function to fine-tune the structure-based predictor for bacterial species. These choices are well suited for bacteria, in which functionally related genes frequently co-localize and are co-transcribed within operons, allowing transcription dynamics to serve as informative, condition-dependent signals. We show that DeepEST outperforms existing protein function prediction methods on a 25-species benchmark, relying solely on amino acid sequence or protein structure. Averaged over the 25 species, DeepEST reaches a protein-centric $F_{\max}$ of 0.689 and a term-centric *micro*-AUPRC of 0.756, a relative improvement of 4.4% and 11.6% over the second-best method, the structure-based predictor DeepFRI. Moreover, DeepEST predicts GO terms for unclassified hypothetical proteins across 25 human bacterial pathogens, supporting their experimental characterization. DeepEST bridges organism-specific data integration and structure-based transfer learning, providing a method tailored for bacterial protein function prediction in settings with structural and multi-condition expression data.

Introduction

A comprehensive characterization of protein function is a crucial aspect of understanding cellular biology. Despite this, a large majority of the known genes, even in extensively studied organisms, lack functional assessment for the encoded proteins (Zerbino et al. 2017). This issue is even more pronounced in prokaryotes, where up to 60% of protein sequences remain unknown in terms of function (Escudeiro et al.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

2022). Filling this knowledge gap is of crucial importance for advancements in various fields. For instance, it is essential for understanding bacterial physiology and their environmental adaptation. Moreover, it has the potential to drive advancements in biotechnological areas like bioremediation, bioenergy and novel medical treatments (Ardern et al. 2023; Djemiel et al. 2022). With roughly 10 million protein sequences of unknown function in bacteria (Escudeiro et al. 2022), it is evident that experimental methods alone are insufficient to determine the function of each protein. Standard computational tools often rely on sequence similarity to annotate proteins, such as sequence alignment-based tools like BLAST (Altschul et al. 1997; Buchfink et al. 2014), or on gene location-aware techniques (Pazos Obregón et al. 2022). More advanced methods utilize deep learning approaches, for example, DeepGOplus (Kulmanov and Hoehndorf 2019), which predicts protein functions by employing a convolutional neural network to automatically learn features from the amino acid sequences. However, despite their smaller and simpler genome, prokaryotes exhibit a high degree of functional redundancy and genetic diversity, making it challenging to predict protein functions based solely on sequences or gene location (Avican et al. 2021; Pazos Obregón et al. 2022).

To overcome this challenge, it is necessary to integrate multiple data modalities that can capture vital yet different aspects of protein functionality, including protein structures, gene expression and gene location data. The information contained within protein structures provides valuable insights into their functional groups and domains (Gligorijević et al. 2021; Goodsell 2009). DeepFRI (Gligorijević et al. 2021), annotates proteins based on their structures. Besides protein structure, gene expression and location can be important descriptors, as genes involved in the same processes often exhibit similar expression patterns and they tend to be located in proximity within bacterial genomes (Ling et al. 2009). While methods that combine different data sources already exist, they typically use protein sequences, protein structures, and protein-protein interaction data. Some examples include DeepGO (Kulmanov et al. 2017), TransFun (Boadu et al. 2023), HOPER (Ünsal 2023) and the protein language model ProstT5 (Heinzinger et al. 2024). Furthermore, the majority of the methods in the literature have primarily been applied to eukaryote benchmark datasets and have rarely been used to study prokaryotes.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

In this paper, we introduce DeepEST, a multimodal deep-learning approach specifically designed to predict protein functions in bacteria based on Gene Ontology (GO) annotations. DeepEST integrates structure, gene expression and location data as protein descriptors, offering the potential to enable accurate protein function prediction in bacteria. In bacteria, chromosomes are typically single and circular, so genes involved in related biological processes tend to be located close to each other on the chromosome and are often co-regulated. Moreover, many bacterial proteins are encoded in operons as polycistronic mRNAs, where multiple genes with related functions are transcribed together. DeepEST fine-tunes DeepFRI (Gligorijević et al. 2021), by leveraging high-quality structural models available today (Jumper et al. 2021; Varadi et al. 2021; Lin et al. 2023) and incorporates an additional neural network using gene expression levels and locus features, capable of providing functional context. Throughout, we focus on bacterial genomes with (i) a high-quality reference genome/organization, (ii) protein structures (predicted or experimental), and (iii) multi-condition gene expression data. We evaluated DeepEST on a set of 25 bacterial species and predicted protein functions on a selection of ≈ 7,000 functionally unannotated hypothetical proteins.

Results

Overview of DeepEST

DeepEST predicts protein functions from a tuple of protein descriptors, $X = (X_s, X_e)$, where X_s represents the protein structures and X_e comprises the gene expression and location data (Fig 1A and B). It integrates two modules: a structure module, for which we harness DeepFRI (Gligorijević et al. 2021) in a transfer learning setting, and an expression-location module, a multilayer perceptron that takes as input the experimentally measured gene expression values under different stress conditions and genetic locus descriptors. DeepEST predicts protein functions by integrating the expression-location and structure data through a masked linear integration with learnable parameters β_s and β_e, followed by a sigmoid, and predictions are updated at test time to conform to the hierarchical structure of the GO terms. As labels, we utilize the GO terms (Zhao et al. 2020), focusing on the Biological Process ontology as a more

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

comprehensive descriptor of protein function. Formal definitions, module architectures, and training are detailed in Methods.

To assess the performance of DeepEST on predicting protein functions for the set of 25 bacterial species, we obtained the label matrix Y and the tuple of protein descriptors, $X = (X_s, X_e)$, which includes protein structures and the expression-location features (preprocessing of expression features is described in Supplemental Section S4). We fine-tuned the pre-trained DeepFRI model, with the hyperparameter configuration identified in the original publication, as the structure module of DeepEST, f_s. The MLP constituting the expression-location module, f_e, is instead optimized and trained *ab initio*.

Protein structure-based splitting improves generalizability of DeepEST

Given that DeepEST relies on a structure-based module, we generate the five data splits by accounting for protein structure similarity. This is accomplished by utilizing foldseek (van Kempen et al. 2023), which enables us to cluster proteins with high structure similarity (local distance difference test score (Mariani et al. 2013) $\geq$ 70%) in the same fold (Fig 2.A). This strategy is to ensure that the structures in the validation and test sets sufficiently differ from the ones in the training set to prevent any data leakage, enabling the evaluation of DeepEST's generalizability to new structures (Kucera et al. 2023), and thus, to previously unseen proteins. Therefore, we employed structure- and sequence-based splitting strategies to evaluate performance over data leakage, which is the case for the sequence-based splitting strategy. When splitting, we unified a dataset from all species and also generated data splits for single species by simply taking the subset belonging to that species from the unified training, validation and test sets respectively (Fig 2.B). The performance of both DeepEST and DeepFRI in terms of F_{max} and *micro*-AUPRC is significantly higher when training on sequence-based splits, indicating a potential overperformance due to data leakage caused by similar protein structures in the split sets (Supplemental Figs S1, S2, S3, S4, S5). Noteworthy, this was more pronounced when using unified datasets, which are more likely to have similar protein structures. Hence, we compared the performance of the models with structure-based split datasets generated from unified data and subsets belonging to single species from the training, validation

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

and test sets respectively. The analysis showed no significant difference in the performance of the models when using splits from all and single species datasets (Fig 2.C-F and Supplemental Fig S5). This suggests improved generalizability of DeepEST by avoiding or minimizing similar structures in the training, validation, and test sets. We elaborate on the importance of employing this splitting strategy in the Supplement (Supplemental Section S5). Based on this, we apply structure-based splits to perform 5-fold nested cross-validation throughout the paper. Next, we compare DeepEST to several sequence-, structure- and pLM-based methods on the 25 pathogens.

DeepEST outperforms sequence, structure and pLM-based methods

To compare DeepEST with gene function annotation methods using different protein descriptors, we trained one model per species and compared to sequence-based traditional annotation methods (i.e., BLAST (Altschul et al. 1997) and Diamond (Buchfink et al. 2014)), a structure-based annotation baseline built on Foldseek (van Kempen et al. 2023) and a combined sequence and structure baseline (Foldseek+Diamond), deep learning approaches (i.e., DeepGOCNN and DeepGOplus (Kulmanov and Hoehndorf 2019)), a structure-based deep learning method, DeepFRI (Gligorijević et al. 2021), and ProstT5 (Heinzinger et al. 2024), a state of the art transformer-based protein language model (pLM) that combines both structure and sequence information. All methods and the rationale behind their selection are described in Supplemental Section S6. The F_{max} and *micro*-AUPRC scores described in Supplemental Section S2 were calculated for each species to assess performance.

DeepEST consistently outperforms all the sequence-only methods, i.e., BLAST, Diamond, DeepGOCNN, and DeepGOplus, both in terms of term-centric *micro*-AUPRC and protein-centric F_{max} (Fig. 3A and B; Supplemental Fig S6). This highlights the importance of considering more than just protein sequences when predicting protein function in bacteria (Fernández et al. 2007). Instead, the use of DeepFRI, which utilizes protein structure data, already enables superior performance compared to the sequence-only baselines, showing the higher relative importance of protein structure over protein sequences for function prediction in bacteria. Moreover, DeepEST with usage of gene expression level

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

and location data in addition to fine-tuning the structure-based module improved performance compared to DeepFRI, especially in terms of term-centric *micro*-AUPRC (Fig. 3A and B; Supplemental Fig S7). These results are confirmed by the additional metrics we utilized, namely term-centric s_{min}, term-centric *macro*-AUPRC, and protein-centric AUPRC (Supplemental Section S2;Supplemental Figs S8 and S9). These findings suggest that integrating the expression-location module along with the abundance of available protein structure data to refine the structure-based module allows for a more comprehensive analysis of protein functionality in bacteria.

DeepEST consistently outperforms ProstT5 fine-tuned on the 25 bacterial species, both in terms of protein-centric F_{max} and term-centric *micro*-AUPRC (Fig. 3A and B; per-species values in Supplemental Fig S10). This highlights the importance of developing specialized models like DeepEST compared to using and fine-tuning more generalist models like ProstT5. We further demonstrate that DeepEST empirically generalizes to a related CAFA5 strain, where PATHOgenex expression data is substituted with an independent stress-response expression compendium (Supplemental Section S7; Supplemental Fig S11).

Next, we assess the impact of DeepEST's building blocks, the MLP constituting the expression-location module f_e and fine-tuned DeepFRI f_s. We additionally evaluate DeepEST and DeepFRI with and without transfer learning.

All building blocks of DeepEST contribute to its performance and enable higher specificity in prediction

To understand the contribution of the various components of DeepEST, we performed ablation studies by excluding transfer learning (TL), the structure-based module f_s, or the expression-location module f_e. TL consistently improved both protein-centric F_{max} and term-centric *micro*-AUPRC across species (Fig. 3C,D; Supplemental Tables S1, S2), and fine-tuning f_s further enhanced performance compared to DeepFRI (Supplemental Fig S12). Removing f_e slightly affected F_{max} but reduced *micro*-AUPRC (Fig. 3C and D; Supplemental Table S2), indicating complementary information from expression-location

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

features. DeepEST achieved higher average precision and recall for most species except *B. burgdorferi* (Fig. 3E; Supplemental Table S3). For *B. burgdorferi*, inclusion of the expression-location module f_e reduced F_{max} from 0.704 to 0.698, the largest decrease among the 25 species but within the variation across test sets, whereas *micro*-AUPRC increased from 0.723 to 0.768 (Supplemental Tables S1 and S2). This was not explained by proteome size, the number of labeled proteins or the labeled fraction of the proteome, none of which was significantly correlated with the contribution of f_e (Supplemental Fig. S13A and B). *B. burgdorferi* is unusually plasmid-rich: its 21 plasmids carry 40.1% of its DNA but only 7.1% of its labeled proteins. Across 11 strains with plasmids, the plasmid representation gap was not significantly associated with the F_{max} contribution of f_e (Pearson $r = -0.67$, *p-value* = 0.07), with *B. burgdorferi* as an outlier (Supplemental Fig. S13C). The unusually large and poorly represented plasmid-associated repertoire of *B. burgdorferi* may therefore contribute to its atypical F_{max} response, but this interpretation remains exploratory. Exclusion of f_s from DeepEST yielded a significant reduction in performance considering both protein-centric F_{max} and term-centric *micro*-AUPRC scores (Fig 3C and D; Supplemental Tables S1 and S2). This implies that relying solely on location and expression patterns falls short in capturing the intricate nature of the protein function problem. However, when incorporated to protein structure data, it can provide meaningful insights to enhance our comprehension of protein functionality. Finally, a comparison between DeepEST with and without incorporation of DeepGOCNN showed a minimal, if any, advantage of this additional protein descriptor (Supplemental Fig S14). Notably, the newly identified GO terms by DeepEST for all species exhibited an average hierarchy depth of 6.3, whereas the same value was 4.7 for GO terms identified solely by f_s (Fig. 3E; Supplemental Table S3). These values suggest that exploiting the expression-location features leads to the prediction of more specific GO terms by DeepEST. To test whether the expression-location module works well only for well-studied organisms, we correlated DeepEST's performance with the number of genomes available for each species in NCBI. We found no significant correlation for complete genomes (Fig. 4A), assemblies at all levels (Fig. 4B) or assemblies at scaffold level or better (Fig. 4C), indicating that

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

DeepEST's performance does not depend on how well studied an organism is. Further, fully removing the six position features used in the model and retraining changes F_{max} by at most 0.0022 across the 25 species (Supplemental Fig S15), and simulated assemblies of *S. enterica* SL1344 stay within 0.005 F_{max} of the full model down to 60% of genes retained and 80 contiguous regions (Supplemental Fig S16, per-scenario numbers in Supplemental Table S4). DeepEST is therefore robust to fragmented and missing genome data, though not to metagenomic assemblies.

Performance by GO group and protein family

To show which annotations DeepEST predicts well, we broke the benchmark down by GO depth, protein family and broad functional category, scoring each group with the term-centric *micro*-AUPRC (Supplemental Section S8). The performance of DeepEST grows with annotation depth, grouping the terms in four bins: shallow (depth 0 to 3), medium (depth 4 to 6), deep (7 to 9) and very deep (depth of 10 or more). (Fig. 3F). Across the 1,131 protein families, DeepEST is ahead of DeepFRI in 58%. Figure 5A shows DeepEST gains most on the translation factor LepA, the cysteinyl- and glutaminyl-tRNA synthetases, the protease subunit HslU and the DNA-uptake protein Smf, and loses most on envelope transport and its regulation, ModA, LptD, ZnuA, PhoU and CopD. By functional category DeepEST leads on nucleotide metabolism, DNA recombination, cell-wall organisation, sulfur compound metabolism, cell division, tRNA metabolism, translation and DNA replication, each at $q < 0.001$ against both baselines (Fig. 5B). There are four cases where DeepEST performs clearly worse than DeepFRI: chemotaxis (−0.241), protein secretion (−0.216), cell motility (−0.146) and protein folding (−0.051), all at $q \leq 0.03$.

The model attends most to nitrosative stress, but no single condition is decisive

Applying DeepEST to a new organism requires transcriptomic data, so we investigated which perturbations are the most informative (Supplemental Section S10). We attributed each prediction to its inputs using integrated gradients (Sundararajan et al. 2017), a commonly used interpretability technique that identifies which features a deep neural network treats as important. Attribution points consistently at nitrosative stress, which takes the largest share in 16 of the 25 species, followed by nutritional downshift,

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

and this ordering holds per GO term (Fig. 4D and E; per species in Supplemental Fig S17). We then tested whether the conditions are necessary: removing one condition at a time and retraining DeepEST, over all 267 species-condition pairs, changes $F_{\max}$ by at most 0.0016, an order of magnitude below the cross-fold standard deviation, and attribution does not predict what removal costs ($\rho = -0.03$). The prominence of nitrosative stress in the attribution is nevertheless plausible on biological grounds. Nitric oxide and its derivatives in the host damage iron-sulfur clusters, oxidise cysteine thiols, inhibit respiration and generate DNA lesions, so a single perturbation engages iron and sulfur assimilation, thiol homeostasis, DNA repair and recombination, and nucleotide salvage at once, partitioning the transcriptome into many co-regulated modules and covering the categories on which DeepEST may gain most over the structure predictor alone. Given the leave-one-condition-out result, however, what the model exploits is the covariance structure across the panel, which several conditions can supply, so a sufficiently diverse panel should give most of the performance benefit of DeepEST.

Applying DeepEST on functionally unannotated proteins

Considering the diversity of protein function in bacteria and challenges in functional characterization of newly annotated proteins, a high percentage of bacterial proteins remains uncharacterized. These proteins are commonly known as hypothetical proteins, whose functions are unknown, and hence represent potential subjects for DeepEST. Within the 25 bacterial species, we identified 6,997 of such proteins that present both input features, X_s and X_e, and remain functionally unannotated at the time of collecting GO terms annotations. When employing DeepEST to those proteins, we retained the optimal hyperparameter configurations listed in Fig 1.C for the expression-location module of DeepEST. The model, however, is retrained on the entire labeled dataset, setting aside a validation set for early stopping and to determine the optimal prediction threshold τ. Specifically, we choose the τ that maximizes the average F1-score calculated for each GO term across proteins, thereby optimizing the harmonic mean of precision and recall. Annotations are reported as the average of the five models identified in the five splits. DeepEST predictions yielded diverse tiers of the GO hierarchy for the 25 bacterial strains (available on our GitHub

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

repository). At a GO depth greater than 5, DeepEST predicts 90 distinct GO terms, ranging from signal transduction and protein transport to aerobic respiration and RNA metabolism, for 727 gene products (Supplemental Fig S18). Most of these GO terms are associated with fewer than 25 gene products each, suggesting that our predictions are diverse and not biased towards fewer broader GO categories.

To further explore the biological significance of DeepEST-predicted GO terms with existing knowledge and literature, we conducted two case studies focusing on GO terms linked to DNA and RNA metabolism. For DNA metabolism, we analyzed “DNA repair” (GO:0006281) (Fig 6A), while for RNA metabolism, we focused on “tRNA processing” (GO:0008033) and associated child terms (Fig 6B). DeepEST identified genes potentially involved in DNA repair across various species with diverse taxonomic and physiological traits (Supplemental Table S5). DeepEST made high-confidence predictions for many genes containing Pfam domains such as AAA-ATPase, transcriptional regulator, and helix-turn-helix, all of which are associated with DNA binding and are typical of DNA repair proteins (Fig 6C and Supplemental Table S5) (Das and Bajpai 2023), (Aravind et al. 2005), (Mistry et al. 2020). Notably, DeepEST predicted GO terms for proteins with domains of unknown function also, indicating its prediction capability for proteins with unknown functional signatures (Fig 6C Supplemental Table S5). In the case of tRNA processing, we identified domains linked to beta-lactamase superfamily proteins and sulfurtransferase TusA (Fig 6D and Supplemental Table S6), which are critical for tRNA modification and maturation (Bhatta et al. 2025), (Nakanishi and Nureki 2005), (Sato et al. 2024). The term ’tRNA processing’ was found across multiple species with notable GO term specificity in *Vibrio cholerae* (locus tag: *VC2163*). DeepEST predicted “tRNA wobble base modification” (GO:0002097) for *V. cholerae* protein containing TusA domain (Fig 6B and D, and Supplemental Table S6), consistent with previous literature (Sato et al. 2024). In *Salmonella enterica* serovar Typhimurium, we identified a gene with locus tag *SL1344_RS05450*, classified under “tRNA processing” (GO:0008033) and possesses domains linked to the beta-lactamase superfamily (Fig 6D). Additionally, using the Co-PATHOgenex web app (Fernandez et al. 2024) and the same transcriptomic data as DeepEST, we identified a contiguous and co-expressed gene associated with the locus tag *SL1344_RS05450*, annotated as *iraM* (Fig 6E). IraM is an

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

anti-adaptor protein that regulates the degradation of RpoS, a sigma factor essential for stress adaptation, and its expression is facilitated by rare tRNA production in *S. enterica* serovar Typhimurium (Bougdour et al. 2008). Through this reverse engineering exercise, we assessed the biological relevance of the predictions and the input data used by DeepEST. This empirical approach demonstrates that by integrating multiple layers of biological and biochemical data, DeepEST can improve both the accuracy of predictions and the precision of the information it provides.

Discussion

We introduced **DeepEST**, a multimodal deep learning framework for bacterial protein function prediction. DeepEST combines protein structure with gene expression profiles and genomic location to assign Gene Ontology (GO) terms, extending function prediction beyond sequence similarity approaches. DeepEST outperforms sequence- and structure-based baselines (Kulmanov and Hoehndorf 2019; Gligorijević et al. 2021; Altschul et al. 1997; Buchfink et al. 2014; Heinzinger et al. 2024). We find that structure is the dominant modality. DeepFRI alone outperforms every sequence-only method, and transferring annotations between structural neighbors with Foldseek outperforms the same transfer by sequence search (BLAST/Diamond) in all 25 species and, on *micro*-AUPRC, the sequence-based deep learning methods as well. DeepEST outperforms the strongest transfer baseline (Foldseek) in all 25 species, by 0.116 F_{max} and 0.194 *micro*-AUPRC on average. ProstT5, a protein language model fine-tuned on the same data does not reach DeepEST, so the gain comes from specializing to the bacterial setting rather than from a larger model. DeepEST further annotated nearly 7,000 previously uncharacterized proteins, revealing processes such as DNA repair and RNA metabolism and demonstrating its potential to guide targeted experimental validation.

The improvements of DeepEST come from fine-tuning the structure module on the target species and adding organism-specific expression and location context, and it is largest where annotation is least redundant, on the deepest and most specific GO terms (Fig. 3E and F). The categories in which DeepEST

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

does not improve are chemotaxis, protein secretion, cell motility and protein folding, where it is outperformed by DeepFRI (Fig. 5).

As in most protein function prediction benchmarks, our GO and UniProt labels are dominated by computationally inferred annotations rather than experimentally validated ones (Supplemental Fig S19). A label can therefore derive from the same kind of evidence a method takes as input, which would inflate that method's apparent performance. The overlap differs by method. It is closest between similarity-propagated annotations and the alignment- and structure-based methods, and between expression-inferred annotations and our expression-location module. Absolute performance is likely inflated for every method, and unevenly so, so the ranking under a shared label set is the quantity these results support, and the absolute values should be read as benchmark-relative. Resolving this would require experimentally validated annotations at a scale not yet available in these species.

In practice, neither an incomplete assembly nor a poorly studied species limits DeepEST (Fig 4 and Supplemental Section S9). What DeepEST requires is sufficiently diverse multi-condition expression data, and the individual conditions are dispensable (Supplemental Section S10), so a new organism likely needs fewer of them, with how many depending on which perturbations it responds to. DeepEST is not intended for metagenomic assemblies, which lack the single-organism genome organization and matched transcriptome that the expression-location module relies on. Overall, DeepEST shows that multimodal, organism-specific data improves bacterial protein annotation and advances interpretable functional genomics in prokaryotes.

DeepEST is lightweight to train and to apply. On 12 CPU cores (Intel Xeon Platinum 8568Y+), training a species-specific model takes about 2.3 min per species for 5-fold cross-validation and 57 min for all 25 species. Using a recent GPU (NVIDIA H200) shortens the full benchmark only to 43 min, because the network is small (about 5.1 million parameters) and most of the time is spent on CPU-side data preparation. Inference alone is fast. Annotating proteins with a trained DeepEST model takes about 18 ms per protein (about the same on CPU and GPU) and almost all of it for propagating the predictions

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

through the GO hierarchy. For comparison, Foldseek, the strongest transfer baseline, needs 13 min for the full benchmark on the same CPU cores and about 25 ms per protein at inference.

Methods

Datasets

This study analyses previously published, publicly available data. Gene Ontology annotations were retrieved from UniProt (https://www.uniprot.org; accessed 12 July 2023). Protein structures were obtained from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk). Source genomes and gene coordinates were taken from NCBI RefSeq, assembly and replicon accessions for all 25 strains are listed in Supplemental Table S7 (complete version with every replicon: Supplemental_Table_S7.xlsx). Multi-condition gene expression data are from the PATHOgenex resource (Avican et al. 2021). Processed data and trained-model inputs are archived on Zenodo (https://zenodo.org/records/18695434). Details on data collection and preprocessing are given in Supplemental Sections S3 and S4. The numbers of annotated proteins and GO terms per species are shown in Supplemental Fig S20.

Establishing DeepEST framework for bacterial protein function prediction

To define the function of a protein, we utilize the GO terms (Zhao et al. 2020), which are organized in three different ontologies, each represented as a directed acyclic graph (DAG). These ontologies include: Molecular Function, Cellular Components, and Biological Processes. Given the multifunctional nature of proteins and DAG structure of GO terms, a single protein can be associated with multiple GO terms simultaneously. Molecular Function terms are limited by precise semantics. We utilized Biological Process terms as a more comprehensive alternative to address this limitation. Biological Process encompasses a broader spectrum of protein activities, capturing diverse roles of proteins in cellular and physiological contexts. This holistic perspective accommodates proteins with multifaceted functions and integrates those whose specific molecular function remains elusive (Thomas 2017). Given a set of n_p

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

proteins $\mathcal{P} = \{p_1, p_2, \ldots, p_{n_p}\}$ and a set of n_t terms $\mathcal{T} = \{t_1, t_2, \ldots, t_{n_t}\}$, we define the matrix of the labels, $Y \in \mathbb{R}^{n_p \times n_t}$, as

$$Y(i,j) = \begin{cases} 1 & \text{if } p_i \text{ is annotated with } t_j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Hence, Y denotes the ground truth for protein function in the dataset we are studying, representing the GO term annotations within the set $\mathcal{T}$, which consists of n_t unique GO terms annotated for at least one of the n_p proteins in $\mathcal{P}$.

Prediction Problem

We use the combination of protein structures, gene expression and location data to enhance protein function prediction in bacteria. DeepEST predicts protein functions from a tuple of protein descriptors, $X = (X_s, X_e)$. X_s represents the protein structures, while X_e comprises the gene expression and location data (Fig 1.A and B). Given the nature of the labels, we formulate the GO term prediction as a multi-modal and multi-label binary classification problem,

$$f(X_s, X_e; w) = \hat{Y}, \quad (2)$$

where the matrix of the predicted labels, $\hat{Y} \in \mathbb{R}^{n_p \times |\mathcal{T} \cup \mathcal{S}|}$, is obtained by applying $f(\cdot)$ to the input data, $X = (X_s, X_e)$. $f(\cdot)$, the non-linear function parameterized in w representing DeepEST, is detailed below. Noteworthy, the matrix of the ground truth labels Y reports GO terms annotations for the terms in the set $\mathcal{T}$, as introduced in the previous section. Instead, the matrix of the predicted labels $\hat{Y}$ comprises predictions for the terms in $|\mathcal{T} \cup \mathcal{S}|$, with $\mathcal{S}$ being the set of terms predicted by the fine-tuned DeepFRI, i.e., the component of DeepEST that leverages protein structure data. The hyperparameter optimization for f_e is performed through grid search executed using W&B (Biewald 2020), with the best configurations (shown in Fig 1C) obtained after hyperparameter optimization on the unified dataset of all 25 species, followed by training on each species separately (Supplemental Section S1; searched grid in Supplemental Table S8, input and output sizes per species in Supplemental Table S9, and best configurations on the sequence-based splits in Supplemental Table S10).

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

To assess the quality of the predicted protein labels $\hat{Y}$ relative to the ground truth annotations Y, we follow the evaluation protocol established by the Critical Assessment of Functional Annotation (CAFA) challenge (Radivojac et al. 2013). In the term-centric setting, we compute the micro-averaged area under the precision–recall curve (*micro*-AUPRC), following (Gligorijević et al. 2021). For the protein-centric evaluation, we use the $F_{\max}$ −score (Supplemental Section S2).

DeepFRI (Gligorijević et al. 2021) predicts a set of 3,442 GO terms in the biological process hierarchy, $\mathcal{S} = \{t_{s_1}, t_{s_2}, \cdots, t_{s_{3,442}}\}$, from the protein structures:

$$f_{\mathrm{s}}(X_{\mathrm{s}}; \mathrm{w}_{\mathrm{s}}) = \hat{Y}_{\mathrm{s}}\,, \quad (3)$$

where X_{s} is constituted by the protein sequences and structures, $f_{\mathrm{s}}(\cdot)$ corresponds to DeepFRI and is parameterized in w_{s}, and $\hat{Y}_{\mathrm{s}} \in \mathbb{R}^{n_{\mathrm{p}} \times |\mathcal{S}|}$ are the predicted labels. The architecture of $f_{\mathrm{s}}(\cdot)$ consists of two stages. Firstly, a recurrent neural network extracts representations from the protein sequences, generating a feature vector for each amino acid residue. Subsequently, a graph is constructed on an amino acid level, putting two residues into contact if they are less than 10Å apart. A protein is therefore represented through a graph with sequence-generated node features and is then fed through three graph convolutional network (GCN) layers (Kipf and Welling 2017). Finally, the updated node features are concatenated, and the final predictions are computed with two linear layers. For the structure module of DeepEST, we harness DeepFRI in a transfer learning setting. Specifically, we keep the extracted features from the GCN layers and only retrain the weights of the final linear layers, henceforth denoted as w_{s}. The structure of f_{s} is visualized at the top of Fig. 1A. Contacts are derived from AlphaFold2 (Varadi et al. 2021) (model version_4), and the resulting graph is passed to the GCN without pLDDT-based node/edge masking or edge reweighting.

Expression-location module

Our method adds another neural network, referred to as expression-location module, to the protein structure module. This module takes as input the experimentally measured gene expression values under different stress conditions and genetic locus descriptors (i.e., the gene's *genomic locus* on the

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

chromosome/plasmids, not protein subcellular localization), designed to capture the gene proximity information. Given that the bacterial genomes, with rare exceptions (Hepner et al. 2023), consist of one main chromosome with a circular structure and smaller circular plasmids, we encode each position into polar coordinates. Concretely, for the gene encoding protein p_i let start and end be its coordinates in base pairs on a replicon of length L. Each coordinate is turned into an angle,

$$\theta_{\mathrm{start}} = 2\pi \frac{\mathrm{start}}{L}, \quad \theta_{\mathrm{end}} = 2\pi \frac{\mathrm{end}}{L}, \quad (4)$$

so that coordinate 0 and coordinate L coincide, and is encoded by its sine and cosine. Together with the coding strand and an indicator of the replicon carrying the gene, this gives the $n_{\mathrm{loc}} = 6$ location features of protein p_i,

$$X_{\mathrm{loc}}(i,:) = [\cos\theta_{\mathrm{start}},\ \sin\theta_{\mathrm{start}},\ \cos\theta_{\mathrm{end}},\ \sin\theta_{\mathrm{end}},\ \mathrm{strand},\ \mathrm{replicon}], \quad (5)$$

where strand is 1 on the forward strand and 0 on the complement, and $\mathrm{replicon}$ is 0 on the main chromosome and 1 on a plasmid. Fig. 1D and E illustrates the six features. Functionally related genes lying close together on the chromosome is what motivates these features, while their co-transcription within operons is what the expression features capture. We assess the contribution of these features to the trained models in Supplemental Section S9.

The resulting gene expression X_{expr} and genome-location features X_{loc}, with n_{expr} and n_{loc} number of features respectively, form a single tabular input for all n_{p} proteins, i.e., $X_{\mathrm{e}} \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{e}}}$ with

$$X_{\mathrm{e}} = \left[X_{\mathrm{expr}} \parallel X_{\mathrm{loc}} \right] \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{e}}}, \quad n_{\mathrm{e}} = n_{\mathrm{expr}} + n_{\mathrm{loc}}, \quad (6)$$

where n_{e} is the total number of features after concatenation and we denote horizontal concatenation with $[\,\cdot \parallel \cdot\,]$. Therefore, we use a multilayer perceptron (MLP) architecture to perform predictions according to this module,

$$f_{\mathrm{e}}(X_{\mathrm{e}}; \mathrm{w}_{\mathrm{e}}) = \hat{Y}_{\mathrm{e}}, \quad (7)$$

with $f_{\mathrm{e}}(\cdot)$ being the non-linear function representing the MLP and parameterized in w_{e}, and $\hat{Y}_{\mathrm{e}} \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{t}}}$ the matrix of predicted labels from the expression-location data. A visual representation of f_{e} is reported in the bottom of Fig. 1A. Unlike the structure module, this module is trained *ab initio*.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

DeepEST predicts protein functions by integrating the expression-location and structure data. This is achieved by performing a masked linear integration of Eq. (3) and Eq. (7), that allows to rewrite $f(\cdot)$, introduced in Eq. (2), as:

$$f(X_s, X_e; w) = \sigma\left(\beta_s f_s(X_s; w_s) +_{|\mathcal{S},\mathcal{T}} \beta_e f_e(X_e; w_e)\right), \quad (8)$$

where σ is the sigmoid function. The learnable parameters of DeepEST denoted by **w**, include w_s, w_e, and the two integration parameters, β_s and β_e. $+_{|\mathcal{S},\mathcal{T}}$ denote the masked sum performed accounting for the different sets of predicted GO terms. Indeed, exploiting $f_s(\cdot)$ in a transfer learning setting means that the structure module predicts the set of terms originally predicted by DeepFRI, i.e. $\mathcal{S}$. Instead, developing the expression-location module from scratch allows to predict the set $\mathcal{T}$ of GO terms annotations retrieved for the set of proteins $\mathcal{P}$ in our dataset. Hence, when integrating the two modalities, DeepEST predicts terms in $\mathcal{T} \cup \mathcal{S}$, generating the output $\hat{Y} \in \mathbb{R}^{n_p \times |\mathcal{T} \cup \mathcal{S}|}$. Explicitly, the masked linear integration of the two modalities writes $\hat{Y} =$

$$\sigma\left(\beta_s \left[\hat{Y}_s|_{\mathcal{T}\cap\mathcal{S}} \;\; 0_{n_p \times |\mathcal{T}\setminus\mathcal{S}|} \;\; \hat{Y}_s|_{\mathcal{S}\setminus\mathcal{T}}\right] + \beta_e \left[\hat{Y}_e|_{\mathcal{T}\cap\mathcal{S}} \;\; \hat{Y}_e|_{\mathcal{T}\setminus\mathcal{S}} \;\; 0_{n_p \times |\mathcal{S}\setminus\mathcal{T}|}\right]\right), \quad (9)$$

where $\hat{Y}_e|_{(\cdot)}$ and $\hat{Y}_s|_{(\cdot)}$ are the outputs of each module restricted to subsets of GO terms, namely the intersection of $\mathcal{T}$ and $\mathcal{S}$ ($\mathcal{T} \cap \mathcal{S}$), the difference between $\mathcal{T}$ and $\mathcal{S}$ ($\mathcal{T} \setminus \mathcal{S}$), and the difference between $\mathcal{S}$ and $\mathcal{T}$ ($\mathcal{S} \setminus \mathcal{T}$), $0_{n_p \times |\mathcal{S}\setminus\mathcal{T}|}$ and $0_{n_p \times |\mathcal{T}\setminus\mathcal{S}|}$ are the zero matrices with n_p rows and respectively $|\mathcal{S} \setminus \mathcal{T}|$ and $|\mathcal{T} \setminus \mathcal{S}|$ columns, $[\cdots]$ indicates the concatenation of three matrices. The output probabilities are given by sigmoid layer $\sigma(\cdot)$. This operation is visualized in Fig 1A.

We apply transfer learning on DeepFRI (Gligorijević et al. 2021), to tailor the model to the bacterial organism we are investigating. To this end, we use a masked loss function $\ell(\cdot)$,

$$\ell(\hat{Y}, Y) = \underbrace{\mathrm{BCE}(\hat{Y}|_{\mathcal{T}}, Y)}_{\text{experimental risk within } \mathcal{T}} + \underbrace{\lambda \sum_{w} \|w\|_2}_{\text{regularization term}}, \quad (10)$$

where $BCE(\cdot)$ is the multi-target binary cross-entropy loss, which calculates the experimental risk from the matrix of the predicted probabilities $\hat{Y}$, restricted to $\mathcal{T}$, and the matrix of the ground truth labels Y,

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

$\lambda \sum_w \| w \|_2$ is a weight decay regularization term with parameter λ (Zhang et al. 2019). The utilization of this loss allows for the transfer of label information for the set $\mathcal{T}$ of GO terms in our dataset.

Update according to the DAG

Due to label dependence, any prediction needs to conform to the hierarchical structure of the GO terms. To this end, we use the common approach of updating the predictions at the test step according to the DAG (Zhao et al. 2020), which follows the scheme

$$\hat{Y}(i,j) = \max\left\{\hat{Y}(i,j), \max_{t_k \in \text{Children}(t_j)}\{\hat{Y}(i,k)\}\right\}, \quad (11)$$

where $\text{Children}(t_j)$ represents the children of the term t_j according to the DAG Fig. 1A. Notably, we do not perform such updating steps to retain a differentiable loss function.

Software availability

The source code for DeepEST is available at https://github.com/BorgwardtLab/DeepEST. Further, the code is available in the accompanying Supplemental Material.

Competing interest statement

K.B. is co-founder and scientific advisor of Computomics GmbH, Tübingen, Germany.

Acknowledgments

GPU computational resources on Kebnekaise at HPC2N were provided through project hpc2n2025-144, awarded to Kemal Avican.

Author contributions. K.A. and K.B. conceived and supervised the study. L.M. and L.F. performed the main analyses and experiments, with L.M. carrying out the extended baseline, robustness, performance-breakdown and expression-condition analyses. G.M. and M.A. designed and developed the initial version of DeepEST, including the first computational experiments. L.F. and K.A. curated the expression and annotation data and performed the biological interpretation. All authors wrote the manuscript and approved the final version.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. doi:10.1093/nar/25.17.3389

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 29: 231–262.

Ardern Z, Chakraborty S, Lenk F, Kaster AK. 2023. Elucidating the functional roles of prokaryotic proteins using big data and artificial intelligence. FEMS Microbiol Rev 47. doi:10.1093/femsre/fuad003

Avican K, Aldahdooh J, Togninalli M, Mahmud AKMF, Tang J, Borgwardt KM, Rhen M, Fällman M. 2021. RNA atlas of human bacterial pathogens uncovers stress dynamics linked to infection. Nat Commun 12. doi:10.1038/s41467-021-23588-w

Bhatta A, Kuhle B, Yu RD, Spanaus L, Ditter K, Bohnsack KE, Hillen HS. 2025. Molecular basis of human nuclear and mitochondrial tRNA 3′ processing. Nat Struct Mol Biol 32: 613–624. doi:10.1038/s41594-024-01445-w

Biewald L. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Boadu F, Cao H, Cheng J. 2023. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. Bioinformatics 39: i318–i325. doi:10.1093/bioinformatics/btad208

Bougdour A, Cunning C, Baptiste PJ, Elliott T, Gottesman S. 2008. Multiple pathways for regulation of σS (RpoS) stability in Escherichia coli via the action of multiple anti-adaptors. Mol Microbiol 68: 298–313.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12: 59–60. doi:10.1038/nmeth.3176

Das R, Bajpai U. 2023. Functional characterization of a DNA-dependent AAA ATPase in a F-cluster mycobacteriophage. Virus Res 323: 198957.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Djemiel C, Maron PA, Terrat S, Dequiedt S, Cottin A, Ranjard L. 2022. Inferring microbiota functions from taxonomic genes: a review. Gigascience 11. doi:10.1093/gigascience/giab090

Escudeiro P, Henry CS, Dias RPM. 2022. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. Curr Res Microb Sci 3: 100159.

Fernandez L, Rosvall M, Normark J, Fällman M, Avican K. 2024. Co-PATHOgenex web application for assessing complex stress responses in pathogenic bacteria. Microbiol Spectr 12: e02781–23.

Fernández M, Abreu JI, Caballero J, Garriga M, Fernández L. 2007. Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks. Mol Simul 33: 1045–1056. doi:10.1080/08927020701564479

Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, et al. 2021. Structure-based protein function prediction using graph convolutional networks. Nat Commun 12. doi:10.1038/s41467-021-23303-9

Goodsell DS. 2009. The machinery of life. Springer New York. doi:10.1007/978-0-387-84925-6

Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M, Rost B. 2024. Bilingual language model for protein sequence and structure. NAR Genom Bioinform 6. doi:10.1093/nargab/lqae150

Hepner S, Kuleshov K, Tooming-Kunderud A, Alig N, Gofton A, Casjens S, Rollins RE, Dangel A, Mourkas E, Sheppard SK, et al. 2023. A high fidelity approach to assembling the complex borrelia genome. BMC Genomics 24. doi:10.1186/s12864-023-09500-4

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596: 583–589. doi:10.1038/s41586-021-03819-2

Kipf TN, Welling M. 2017. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations (ICLR).

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Kucera T, Oliver C, Chen D, Borgwardt K. 2023. ProteinShake: building datasets and benchmarks for deep learning on protein structures. In Adv Neural Inf Process Syst, pp. 58277–58289. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi:10.52202/075280-2539

Kulmanov M, Hoehndorf R. 2019. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36: 422–429.

Kulmanov M, Khan MA, Hoehndorf R. 2017. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 34: 660–668. doi:10.1093/bioinformatics/btx624

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379: 1123–1130. doi:10.1126/science.ade2574

Ling X, He X, Xin D. 2009. Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics 25: 571–577. doi:10.1093/bioinformatics/btp027

Mariani V, Biasini M, Barbato A, Schwede T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29: 2722–2728. doi:10.1093/bioinformatics/btt473

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2020. Pfam: the protein families database in 2021. Nucleic Acids Res 49: D412–D419. doi:10.1093/nar/gkaa913

Nakanishi K, Nureki O. 2005. Recent progress of structural biology of tRNA processing and modification. Mol Cells 19: 157–166. doi:10.1016/S1016-8478(23)13151-7

Pazos Obregón F, Silvera D, Soto P, Yankilevich P, Guerberoff G, Cantera R. 2022. Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning. Sci Rep 12. doi:10.1038/s41598-022-15329-w

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al. 2013. A large-scale evaluation of computational protein function prediction. Nat Methods 10: 221–227. doi:10.1038/nmeth.2340

Sato Y, Takita A, Suzue K, Hashimoto Y, Hiramoto S, Murakami M, Tomita H, Hirakawa H. 2024. TusDCB, a sulfur transferase complex involved in tRNA modification, contributes to UPEC pathogenicity. Sci Rep 14: 8978.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328.

Thomas PD. 2017. The gene ontology and the meaning of biological function. The gene ontology handbook: 15–24.

van Kempen M, et al. 2023. Fast and accurate protein structure search with Foldseek. Nat Biotechnol: 1546–1696.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. 2021. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50: D439–D444. doi:10.1093/nar/gkab1061

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2017. Ensembl 2018. Nucleic Acids Res 46: D754–D761. doi:10.1093/nar/gkx1098

Zhang G, Wang C, Xu B, Grosse RB. 2019. Three mechanisms of weight decay regularization. In ICLR (Poster). OpenReview.net. doi:10.48550/ARXIV.1810.12281

Zhao Y, Wang J, Chen J, Zhang X, Guo M, Yu G. 2020. A literature review of gene function prediction by modeling Gene Ontology. Front Genet 11. doi:10.3389/fgene.2020.00400

Ünsal S. 2023. A deep learning based protein representation model for low-data protein function prediction. PhD thesis, Middle East Technical University.

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure legends

Figure 1. Overview of the DeepEST framework. **A** DeepEST fuses two modules: the expression-location module f_e (a multi-layer perceptron) and the structure-based module f_s (DeepFRI used with transfer learning). A protein p_i is the tuple $X_i = (X_{e,i}, X_{s,i})$, where $X_{s,i}$ holds the sequence and the amino-acid contact map. The module outputs $\hat{Y}_{e,i}$ and $\hat{Y}_{s,i}$ are merged by a masked linear combination ($+_{|S,T}$) and a sigmoid (σ) into the prediction $\hat{Y}_i$. **B** At test time, predictions violating the GO DAG hierarchy (red) are corrected by Eq. (11). **C** Best-validation hyperparameter configurations for f_e across the five splits. **D** Genomic-location features: each gene coordinate becomes an angle on its replicon (Eq. (4); only the start angle shown), giving four of the six features. **E** Coding strand and chromosome-vs-plasmid binary features (Supplemental Section S9). **F** Phylogenetic clustering of the 25 strains. **G** Expression setup of the PATHOgenex dataset: control versus 11 host-mimicking stress conditions.

Figure 2. Accounting structure similarity with a unified dataset from 25 species or a single species exhibits similar performance suggesting no data leakage in training, validation, and test sets. **A** Available protein structures from the 25 bacterial species retrieved from AlphaFold were used as queries in Foldseek to cluster proteins with high structure similarity local distance difference test score (LDDT) $\geq$ 70%. **B** Visual representation of split generation for a representative subset of five species. The generation of the splits, which account for structure similarity between the proteins, is performed on the unified dataset. **C** Comparison of training MLP (i.e., f_e), fine-tuned DeepFRI (i.e., f_s) and DeepEST on each species separately ("single") and on the unified dataset comprised of the 25 species ("all") in terms of mean F_{max} on the five *validation* sets, with *structure*-based splits. Each panel corresponds to a specific model used in the analysis. The bar plot shows the mean F_{max} values for protein function prediction in each species. The colors represent datasets, namely "single" and "all", used in training. **D** Percentage change of mean F_{max} values when training the models according to the "all" strategy in contrast to the "single". **E** Comparison of training the three models in terms of average *micro*-AUPRC with same strategy explained in **C**. **F** Percentage change of mean *micro*-AUPRC values when training the models

Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

according to the “all” strategy in contrast to the “single”. The error bars in **C** and **E** represent the standard deviation for each species. Each dot in **D** and **F** represents a percentage change of F_{max} and *micro*-AUPRC, respectively, for each species and straight lines represent the mean values.

Figure 3. DeepEST outperforms established baselines in protein function predictions and achieves greater specificity in GO assignment by using expression location features. **A** The average protein-centric F_{max} and **B** term-centric *micro*-AUPRC values calculated for each species separately across the five sets for DeepEST and other methods: BLAST, Diamond, DeepGOCNN, DeepGOplus, Foldseek, Foldseek+Diamond, ProstT5, and DeepFRI. **C** The average protein-centric F_{max} and **D** term-centric *micro*-AUPRC values calculated for each species separately across the five sets for DeepEST and DeepEST excluding one component at a time: (i) excluding our transfer learning (TL) technique, (ii) excluding the expression-location module, and (iii) excluding the structure module. **E** Comparison between DeepEST and f_s in terms of average %-change in precision and recall for each GO term across proteins (left panel), and average depth of the GO terms identified by either of the methods (right panel). Each dot in **A,B,C,D** represents F_{max} and *micro*-AUPRC scores for each species. *P-values* were calculated with a paired Wilcoxon signed-rank test over the 25 species. **F** Term-centric *micro*-AUPRC by GO-term depth for DeepEST, DeepFRI and Foldseek, averaged over the 25 species, with the number of proteins carrying at least one term of each depth bin.

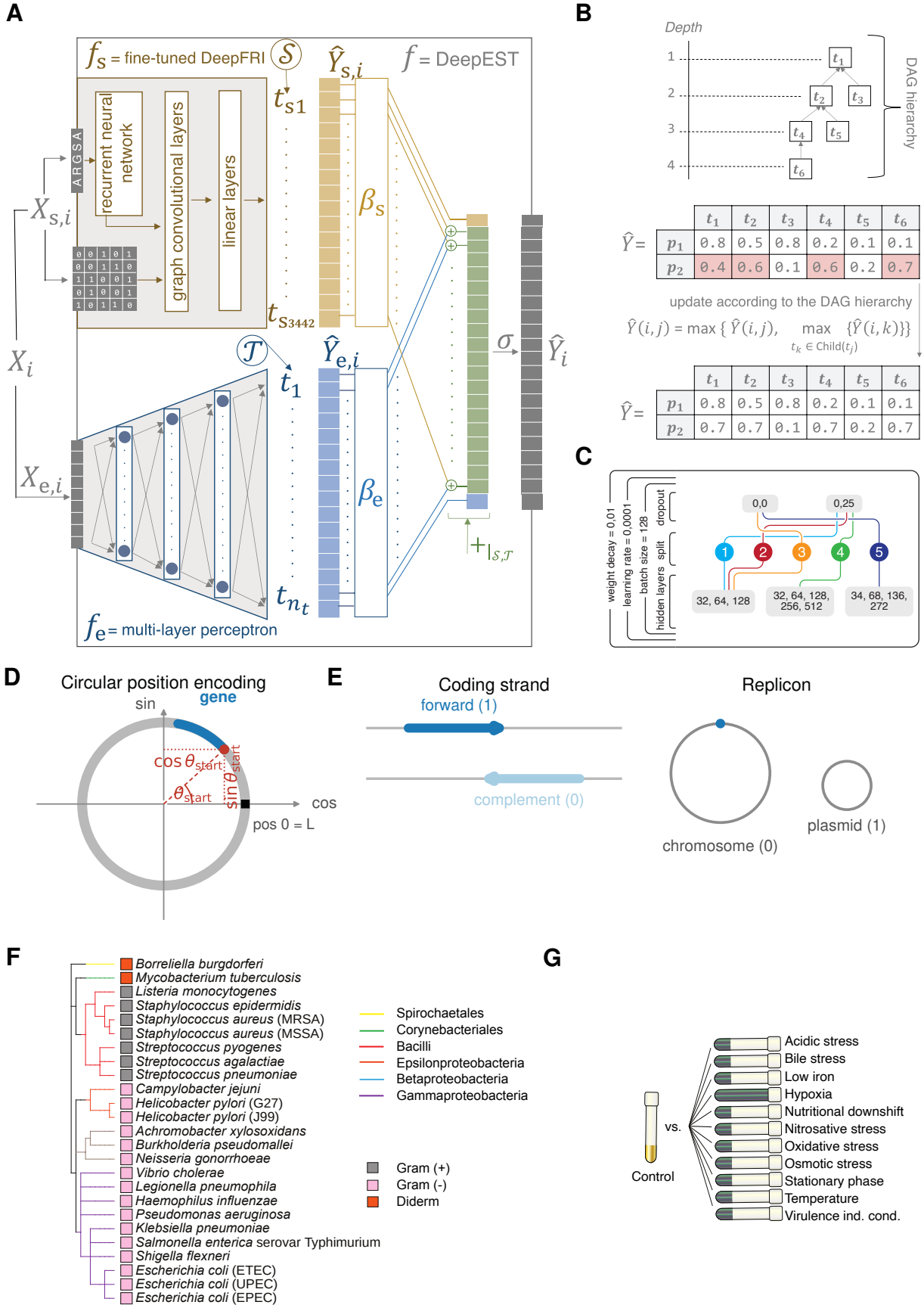
Figure 4. DeepEST performance is unrelated to how heavily a species has been sequenced, and integrated gradients point most to nitrosative stress. **A**–**C** Per-species protein-centric F_{max} against three measures of genome availability in NCBI: complete genomes (**A**), assemblies at all levels (**B**), and assemblies at scaffold level or better (**C**). Each dot is one of the 25 species, averaged over the five test sets, where fewer available genomes means a less studied species. Lines are least-squares fits in $\log_{10}$ of the genome count, and ρ is the Spearman rank correlation with its *p-value*, none of the correlations are significant. **D** Integrated-gradients attribution over the expression inputs: each condition’s share of the attribution within a species, one point is the average over the five folds per species. **E** The per-GO-term view for *S. enterica*

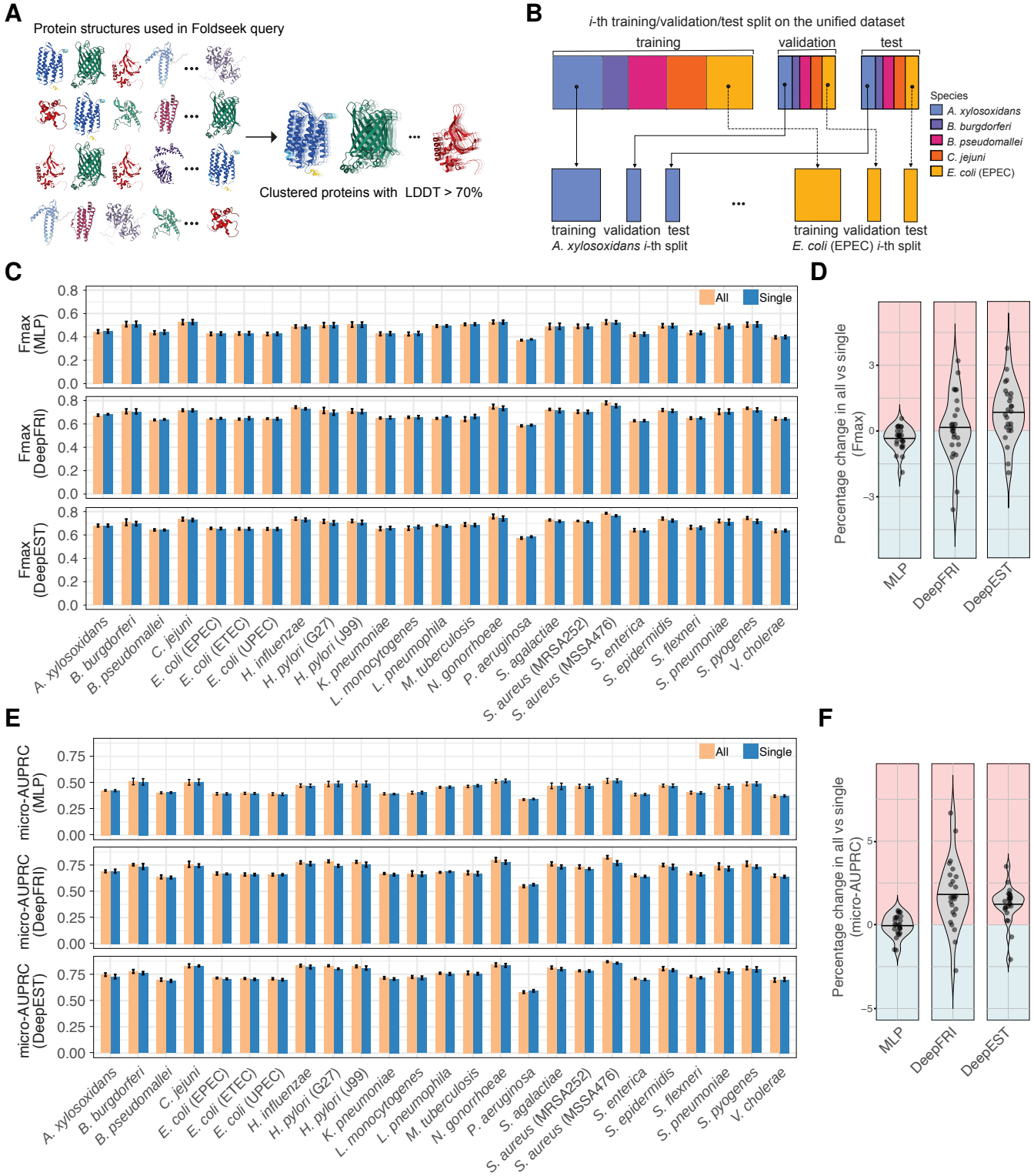
Downloaded from genome.cshlp.org on November 9, 2026 . Published by Cold Spring Harbor Laboratory Press

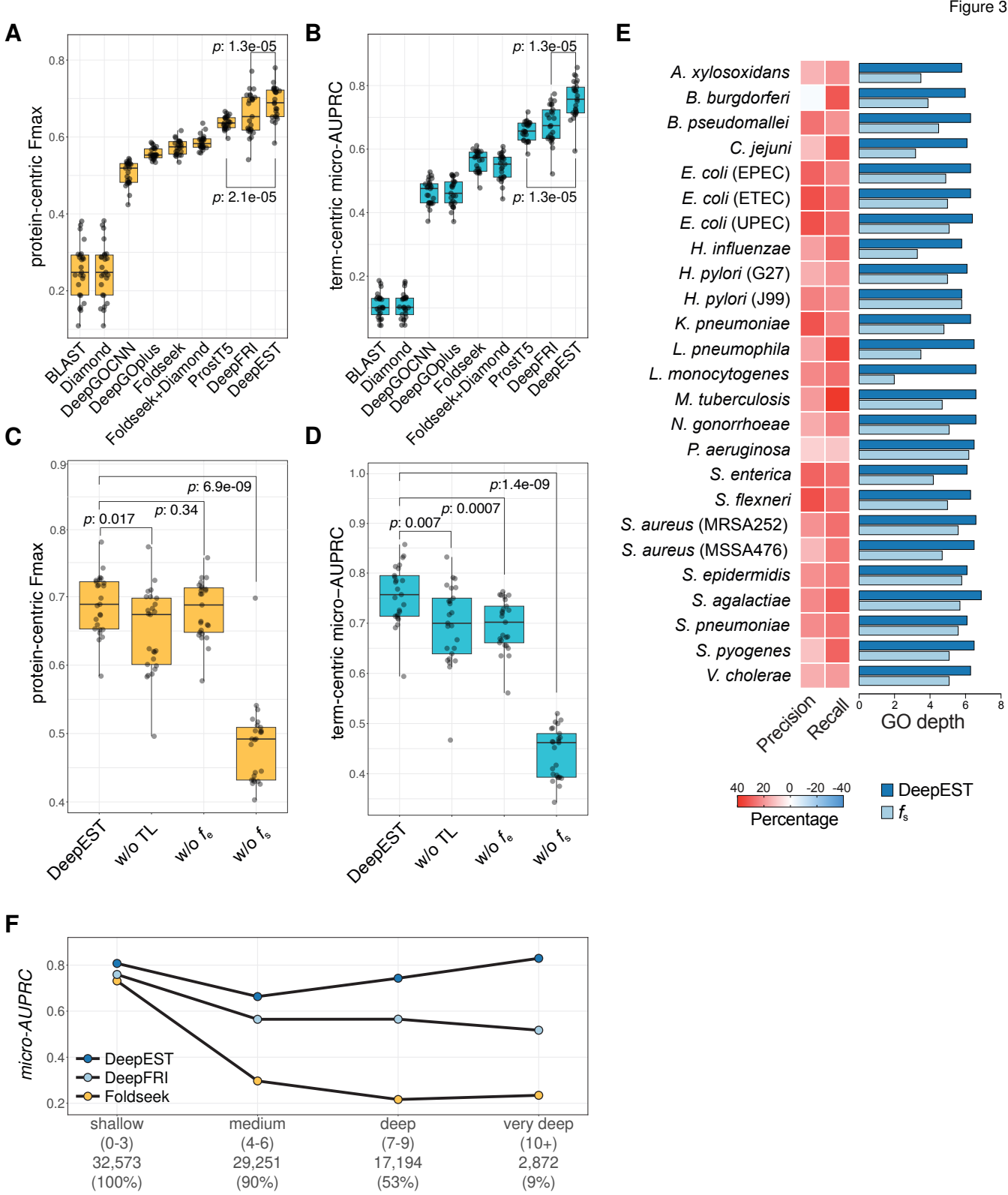
SL1344: the number of the 278 terms (with at least 20 positives) for which each condition receives the largest attribution. Rows in **D** and **E** follow the same order.

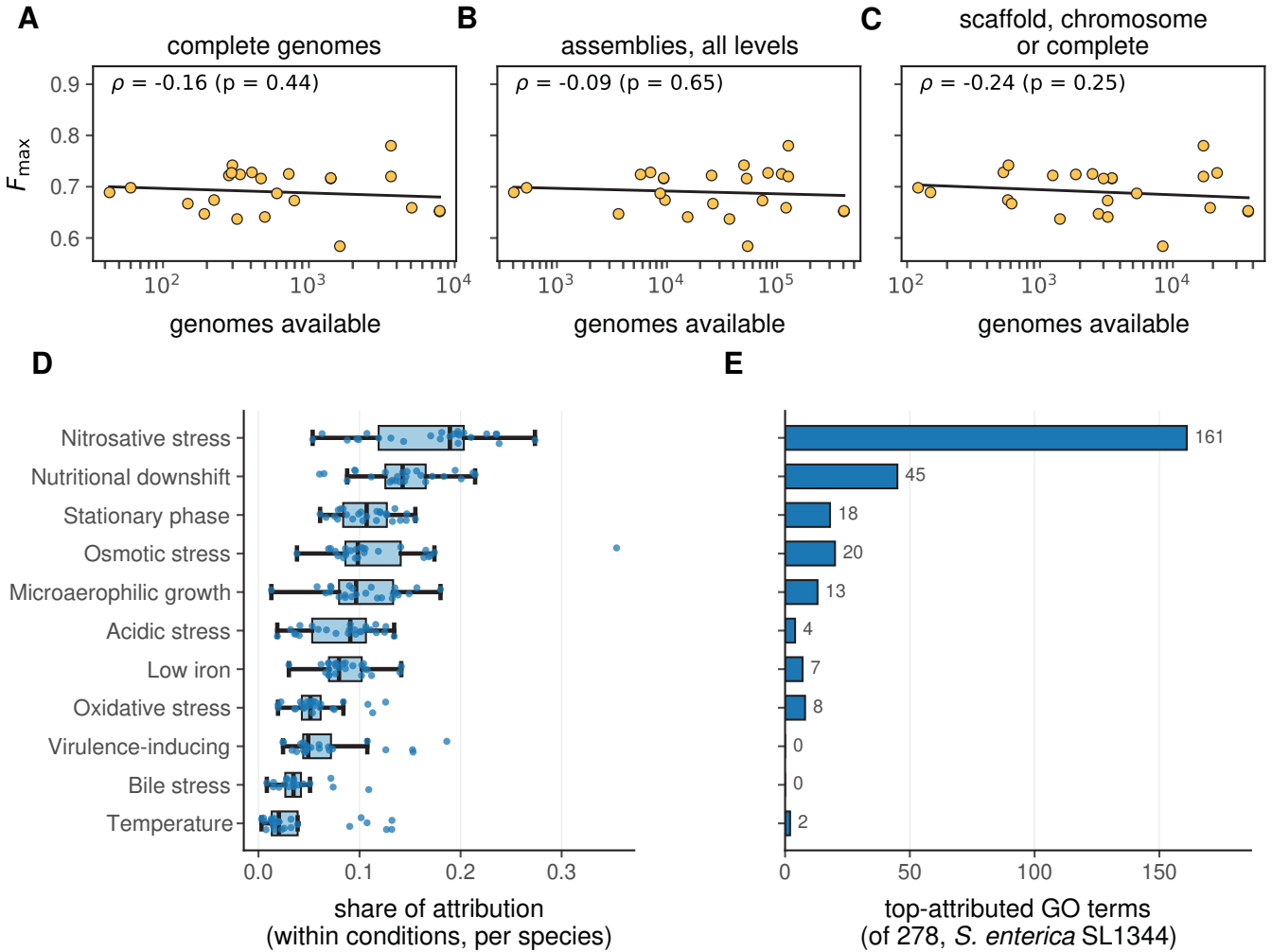
Figure 5. Performance by protein family and functional category, in term-centric *micro*-AUPRC (breakdown by GO depth in Fig. 3F). **A** The eight PATRIC protein families with the highest and the eight with the lowest difference against DeepFRI, of the 1,131 tested, labelled by protein name where possible: **GpmI** is 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, **TreC** is trehalose-6-phosphate hydrolase, and **Hmp** is flavohemoglobin / nitric oxide dioxygenase. **B** Likewise for functional categories, of the 34 present in at least ten species. Brackets give the number of proteins in a group, and bars run right and dark where DeepEST leads DeepFRI, left and light where it trails. Stars are a paired Wilcoxon signed-rank test over the species a group spans, Benjamini-Hochberg corrected within the panel ("***" $q < 0.001$, "**" $q < 0.01$, "*" $q < 0.05$, "ns" not significant). Values are computed per species and averaged (Supplemental Section S8).

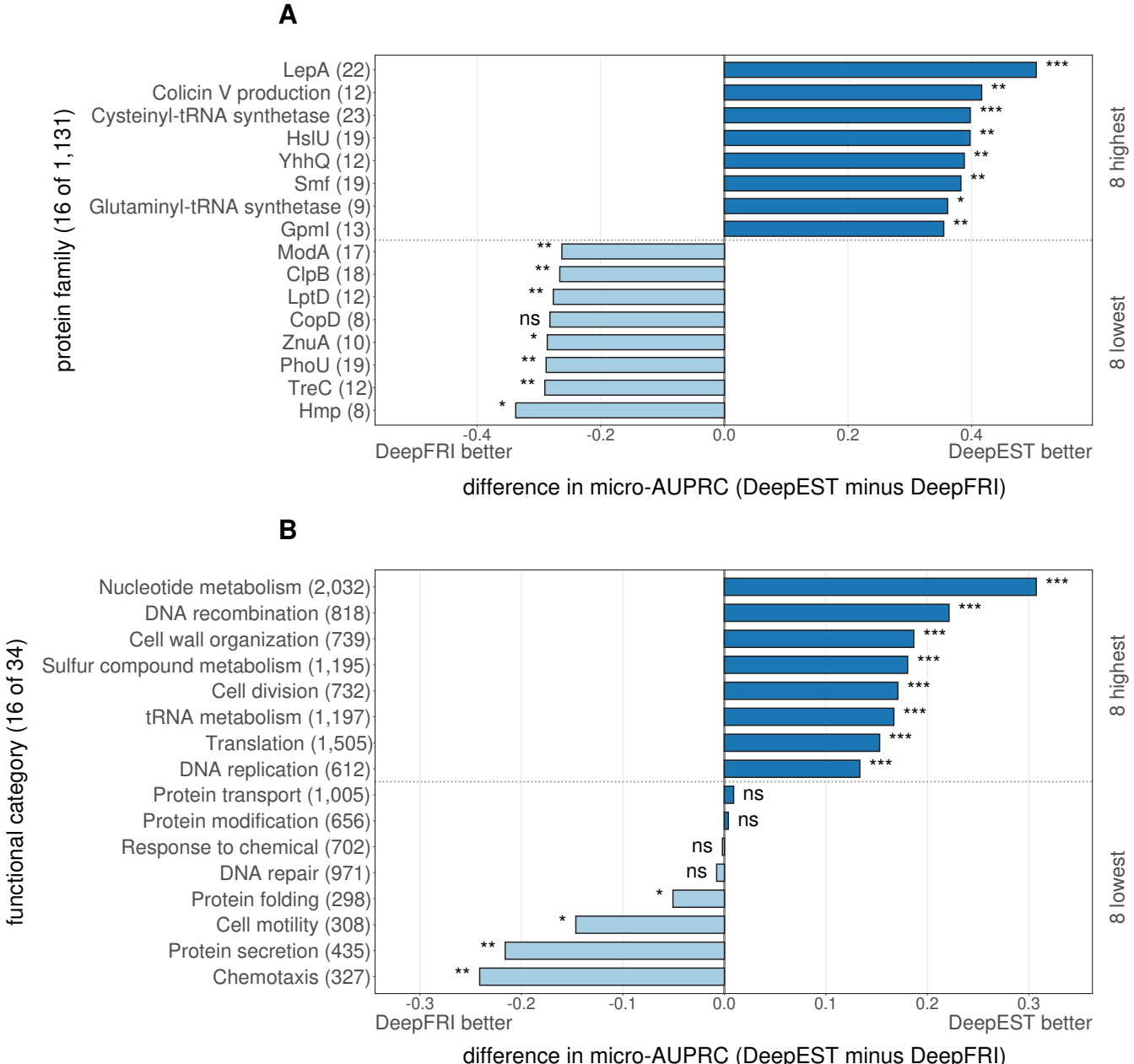
Figure 6. DeepEST predicts functions via assigning specific biological process GO terms to bacterial hypothetical proteins with domains of known and unknown functions. Ontology tree of DeepEST-assigned **A** DNA repair (GO:0006281) with GO hierarchy depth of 7 and **B** tRNA processing (GO:0008033) with its child nodes, along with its predicted proteins from different bacterial species. Representatives of proteins with domains of known and unknown functions together with AlphaFold-predicted structures (Varadi et al. 2021) for proteins **C** assigned with DNA repair and **D** tRNA processing. **E** Expression patterns of the gene encoding anti-adaptor protein IraM and *SL1344_RS05450* encoding a hypothetical protein in *S. enterica* serovar Typhimurium under each replicate of the 12 conditions. Abbreviations of the conditions in **E** are as follows: As; Acidic stress, Bs; Bile stress, Ctrl; Control, Li; Low iron, Mig; Microaerophilic growth, Nd; Nutritional downshift, Oss; Osmotic stress, Oxs; Oxidative stress, Sp; Stationary phase, Tm; Temperature; Vic. Virulence inducing condition.

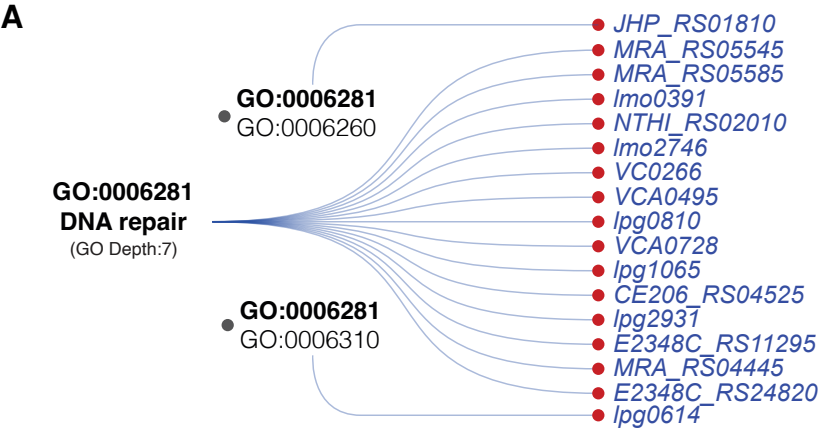

A
f_S = fine-tuned DeepFRI
f = DeepEST
ARGSA
recurrent neural network
graph convolutional layers
linear layers
$X_{s,i}$
X_i
$X_{e,i}$
$\mathcal{S}$
$\mathcal{T}$
t_{s1}
t_{s3442}
t_1
t_{nt}
$\hat{Y}_{s,i}$
$\hat{Y}_{e,i}$
β_s
β_e
σ
$\hat{Y}_i$
$+|_{S,\mathcal{T}}$
f_e = multi-layer perceptron
B
Depth
DAG hierarchy
update according to the DAG hierarchy
$\hat{Y}(i,j) = \max\{\hat{Y}(i,j), \max_{t_k \in \text{Child}(t_j)}\{\hat{Y}(i,k)\}\}$
C
weight decay = 0,01
learning rate = 0,0001
batch size = 128
dropout
split
hidden layers
0,0
0,25
1
2
3
4
5
32, 64, 128
32, 64, 128, 256, 512
34, 68, 136, 272
D
Circular position encoding
gene
sin
cos
pos 0 = L
$\cos\theta_{start}$
$\sin\theta_{start}$
θ_{start}
E
Coding strand
forward (1)
complement (0)
Replicon
chromosome (0)
plasmid (1)
F
Borreliella burgdorferi
Mycobacterium tuberculosis
Listeria monocytogenes
Staphylococcus epidermidis
Staphylococcus aureus (MRSA)
Staphylococcus aureus (MSSA)
Streptococcus pyogenes
Streptococcus agalactiae
Streptococcus pneumoniae
Campylobacter jejuni
Helicobacter pylori (G27)
Helicobacter pylori (J99)
Achromobacter xylosoxidans
Burkholderia pseudomallei
Neisseria gonorrhoeae
Vibrio cholerae
Legionella pneumophila
Haemophilus influenzae
Pseudomonas aeruginosa
Klebsiella pneumoniae
Salmonella enterica serovar Typhimurium
Shigella flexneri
Escherichia coli (ETEC)
Escherichia coli (UPEC)
Escherichia coli (EPEC)
Spirochaetales
Corynebacteriales
Bacilli
Epsilonproteobacteria
Betaproteobacteria
Gammaproteobacteria
Gram (+)
Gram (-)
Diderm
G
Control
vs.
Acidic stress
Bile stress
Low iron
Hypoxia
Nutritional downshift
Nitrosative stress
Oxidative stress
Osmotic stress
Stationary phase
Temperature
Virulence ind. cond.

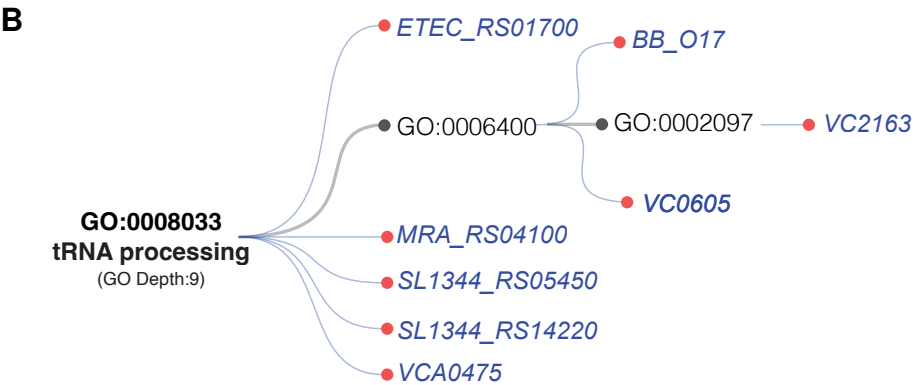

A
Protein structures used in Foldseek query
Clustered proteins with LDDT > 70%
B
i-th training/validation/test split on the unified dataset
training
validation
test
Species
A. xylosoxidans
B. burgdorferi
B. pseudomallei
C. jejuni
E. coli (EPEC)
training validation test
A. xylosoxidans i-th split
training validation test
E. coli (EPEC) i-th split
C
Fmax (MLP)
Fmax (DeepFRI)
Fmax (DeepEST)
All
Single
A. xylosoxidans
B. burgdorferi
B. pseudomallei
C. jejuni
E. coli (EPEC)
E. coli (ETEC)
E. coli (UPEC)
H. influenzae
H. pylori (G27)
H. pylori (J99)
K. pneumoniae
L. monocytogenes
L. pneumophila
M. tuberculosis
N. gonorrhoeae
P. aeruginosa
S. agalactiae
S. aureus (MRSA252)
S. aureus (MSSA476)
S. enterica
S. epidermidis
S. flexneri
S. pneumoniae
S. pyogenes
V. cholerae
D
Percentage change in all vs single (Fmax)
MLP
DeepFRI
DeepEST
E
micro-AUPRC (MLP)
micro-AUPRC (DeepFRI)
micro-AUPRC (DeepEST)
All
Single
A. xylosoxidans
B. burgdorferi
B. pseudomallei
C. jejuni
E. coli (EPEC)
E. coli (ETEC)
E. coli (UPEC)
H. influenzae
H. pylori (G27)
H. pylori (J99)
K. pneumoniae
L. monocytogenes
L. pneumophila
M. tuberculosis
N. gonorrhoeae
P. aeruginosa
S. agalactiae
S. aureus (MRSA252)
S. aureus (MSSA476)
S. enterica
S. epidermidis
S. flexneri
S. pneumoniae
S. pyogenes
V. cholerae
F
Percentage change in all vs single (micro-AUPRC)
MLP
DeepFRI
DeepEST

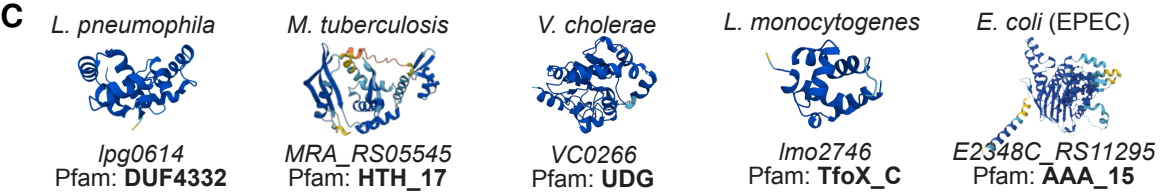

A
protein-centric Fmax
p: 1.3e-05
p: 2.1e-05
BLAST
Diamond
DeepGOCNN
DeepGOplus
Foldseek
Foldseek+Diamond
ProstT5
DeepFRI
DeepEST
B
term-centric micro-AUPRC
p: 1.3e-05
p: 1.3e-05
C
protein-centric Fmax
p: 6.9e-09
p: 0.017
p: 0.34
DeepEST
w/o TL
w/o f_e
w/o f_s
D
term-centric micro−AUPRC
p:1.4e-09
p: 0.0007
p: 0.007
E
A. xylosoxidans
B. burgdorferi
B. pseudomallei
C. jejuni
E. coli (EPEC)
E. coli (ETEC)
E. coli (UPEC)
H. influenzae
H. pylori (G27)
H. pylori (J99)
K. pneumoniae
L. pneumophila
L. monocytogenes
M. tuberculosis
N. gonorrhoeae
P. aeruginosa
S. enterica
S. flexneri
S. aureus (MRSA252)
S. aureus (MSSA476)
S. epidermidis
S. agalactiae
S. pneumoniae
S. pyogenes
V. cholerae
Precision
Recall
GO depth
40 20 0 -20 -40
Percentage
DeepEST
f_s
F
micro-AUPRC
DeepEST
DeepFRI
Foldseek
shallow
(0-3)
32,573
(100%)
medium
(4-6)
29,251
(90%)
deep
(7-9)
17,194
(53%)
very deep
(10+)
2,872
(9%)

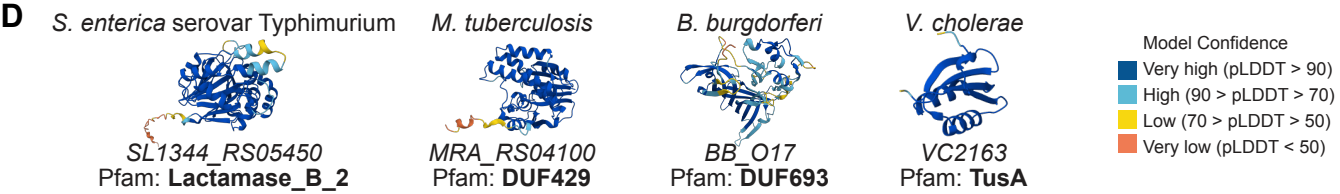

A
complete genomes
ρ = -0.16 (p = 0.44)
B
assemblies, all levels
ρ = -0.09 (p = 0.65)
C
scaffold, chromosome or complete
ρ = -0.24 (p = 0.25)
F max
genomes available
D
Nitrosative stress
Nutritional downshift
Stationary phase
Osmotic stress
Microaerophilic growth
Acidic stress
Low iron
Oxidative stress
Virulence-inducing
Bile stress
Temperature
share of attribution
(within conditions, per species)
E
161
45
18
20
13
4
7
8
0
0
2
top-attributed GO terms
(of 278, S. enterica SL1344)

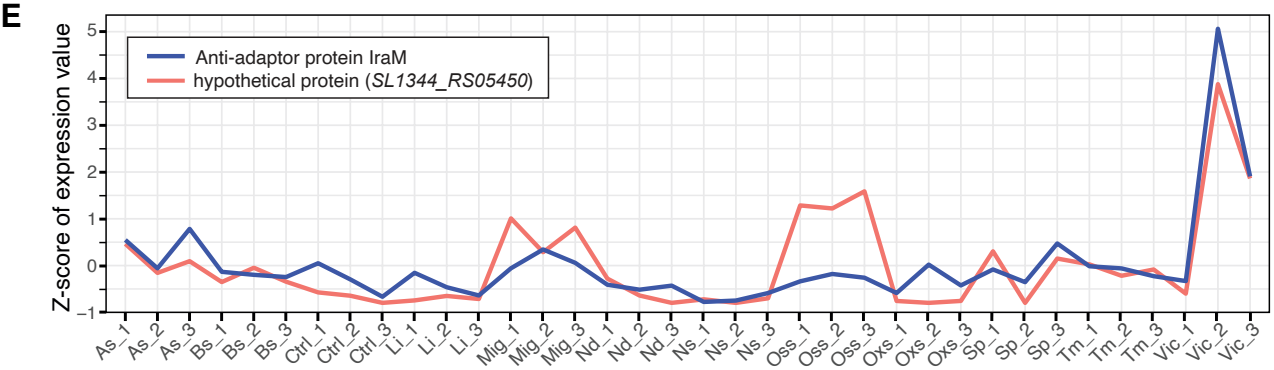

A
protein family (16 of 1,131)
LepA (22)
Colicin V production (12)
Cysteinyl-tRNA synthetase (23)
HslU (19)
YhhQ (12)
Smf (19)
Glutaminyl-tRNA synthetase (9)
GpmI (13)
ModA (17)
ClpB (18)
LptD (12)
CopD (8)
ZnuA (10)
PhoU (19)
TreC (12)
Hmp (8)
***
**
***
**
**
**
*
**
**
**
**
ns
*
**
**
*
8 highest
8 lowest
-0.4
-0.2
0.0
0.2
0.4
DeepFRI better
DeepEST better
difference in micro-AUPRC (DeepEST minus DeepFRI)
B
functional category (16 of 34)
Nucleotide metabolism (2,032)
DNA recombination (818)
Cell wall organization (739)
Sulfur compound metabolism (1,195)
Cell division (732)
tRNA metabolism (1,197)
Translation (1,505)
DNA replication (612)
Protein transport (1,005)
Protein modification (656)
Response to chemical (702)
DNA repair (971)
Protein folding (298)
Cell motility (308)
Protein secretion (435)
Chemotaxis (327)
***
***
***
***
***
***
***
***
ns
ns
ns
ns
*
*
**
**
8 highest
8 lowest
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
DeepFRI better
DeepEST better
difference in micro-AUPRC (DeepEST minus DeepFRI)


A
GO:0006281
GO:0006260
GO:0006281
DNA repair
(GO Depth:7)
GO:0006281
GO:0006310
JHP_RS01810
MRA_RS05545
MRA_RS05585
lmo0391
NTHI_RS02010
lmo2746
VC0266
VCA0495
lpg0810
VCA0728
lpg1065
CE206_RS04525
lpg2931
E2348C_RS11295
MRA_RS04445
E2348C_RS24820
lpg0614


B
GO:0008033
tRNA processing
(GO Depth:9)
ETEC_RS01700
BB_O17
GO:0006400
GO:0002097
VC2163
VC0605
MRA_RS04100
SL1344_RS05450
SL1344_RS14220
VCA0475


C
L. pneumophila
lpg0614
Pfam: DUF4332
M. tuberculosis
MRA_RS05545
Pfam: HTH_17
V. cholerae
VC0266
Pfam: UDG
L. monocytogenes
lmo2746
Pfam: TfoX_C
E. coli (EPEC)
E2348C_RS11295
Pfam: AAA_15


D
S. enterica serovar Typhimurium
SL1344_RS05450
Pfam: Lactamase_B_2
M. tuberculosis
MRA_RS04100
Pfam: DUF429
B. burgdorferi
BB_O17
Pfam: DUF693
V. cholerae
VC2163
Pfam: TusA
Model Confidence
Very high (pLDDT > 90)
High (90 > pLDDT > 70)
Low (70 > pLDDT > 50)
Very low (pLDDT < 50)


E
Anti-adaptor protein IraM
hypothetical protein (SL1344_RS05450)
Z-score of expression value
5
4
3
2
1
0
−1
As_1 As_2 As_3 Bs_1 Bs_2 Bs_3 Ctrl_1 Ctrl_2 Ctrl_3 Li_1 Li_2 Li_3 Mig_1 Mig_2 Mig_3 Nd_1 Nd_2 Nd_3 Ns_1 Ns_2 Ns_3 Oss_1 Oss_2 Oss_3 Oxs_1 Oxs_2 Oxs_3 Sp_1 Sp_2 Sp_3 Tm_1 Tm_2 Tm_3 Vic_1 Vic_2 Vic_3