Bacterial protein function prediction via multimodal deep learning

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

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.
[2]Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland.
[3]Integrated Science Lab (IceLab), Umeå University, Umeå, Sweden.
[4]Department of Molecular Biology, Umeå University, Umeå, Sweden.
[5]Umeå Centre for Microbial Research (UCMR), Umeå University, Umeå, Sweden.
[6]Department of Machine Learning and Systems Biology, Max Planck Institute of Biochemistry, Martinsried, Germany.

*Corresponding author(s). E-mail(s): kemal.avican@umu.se;
†These authors contributed equally to this work.

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 developed Deep Expression STructure (DeepEST), a multimodal deep learning framework designed to accurately predict protein function in bacteria by assigning Gene Ontology (GO) terms. DeepEST comprises two modules: a multi-layer perceptron that takes gene expression and location as input features, and a protein structure-based predictor. Within DeepEST, we integrated these modules through a learnable weighted linear combination and introduced a novel masked loss function to fine-tune the structure-based predictor for bacterial species. We showed that DeepEST strongly outperforms existing protein function prediction methods relying solely on amino acid sequence or protein structure. Moreover, DeepEST predicts GO terms for unclassified hypothetical proteins across 25 human bacterial pathogens, facilitating the design of experimental setups for characterization studies.

Keywords: Protein function prediction, Transfer learning, Protein structures, Multimodal deep learning

Main

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 [1]. This issue is even more pronounced in prokaryotes, where up to 60% of protein sequences remain unknown in terms of function [2]. 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 [3, 4].

With roughly 10 million protein sequences of unknown function in bacteria [2], 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 [5, 6], or on gene location-aware techniques [7]. More advanced methods utilize deep learning approaches, for example, DeepGOplus [8], 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 [7, 9].

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 location data. The information contained within protein structures provides valuable insights into their functional groups and domains [10, 11]. The state-of-the-art in the field, DeepFRI [10], 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 [12]. 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 [13], TransFun [14], and HOPER [15]. 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.

In this study, 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. DeepEST fine-tunes DeepFRI [10], by leveraging high-quality structural models available today [16–18] and incorporates an additional neural network using gene expression levels and locus features, capable of providing functional context. We

evaluated the performance of DeepEST on a set of 25 bacterial species and predicted protein functions on a selection of nearly 7,000 functionally unannotated hypothetical proteins.

Results

Establishing a framework for bacterial protein function prediction

The DeepEST framework is dedicated to function prediction for bacterial proteins. This section starts with the definition of protein function utilized, followed by a multimodal strategy implemented to tackle the prediction problem.

GO terms as label for protein function

To define the function of a protein, we utilize the GO terms [19], 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 [20].

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

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

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 $\mathcal{P}$.

Integrating protein structure, gene-expression, and genomic localization

As introduced before, we use the combination of protein structures, gene expression and location data to enhance protein function prediction in bacteria. Hence, with DeepEST, we predict protein functions from a tuple of protein descriptors, $X = (X_{\mathrm{s}}, X_{\mathrm{e}})$. X_{s} represents the protein structures, while X_{e} comprises the gene expression and location data Fig 1.a,b. Given the nature of the labels, as highlighted before, we formulate

the GO term prediction as a multi-modal and multi-label binary classification problem,

$$f(X_{\mathrm{s}}, X_{\mathrm{e}}; \mathbf{w}) = \hat{Y}, \tag{2}$$

where the matrix of the predicted labels, $\hat{Y} \in \mathbb{R}^{n_{\mathrm{p}} \times |\mathcal{T} \cup \mathcal{S}|}$, is obtained by applying $f(\cdot)$ to the input data, $X = (X_{\mathrm{s}}, X_{\mathrm{e}})$. $f(\cdot)$, the non-linear function parameterized in $\mathbf{w}$ representing DeepEST, is detailed in Methods. 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 [21] Fig 1.c. Additional details regarding hyperparameter optimization and training – which is conducted individually for each species – are reported in Extended Data.

Performance evaluation and application of DeepEST on bacterial protein function prediction

We perform a comprehensive exploration of microbial protein function prediction with a specific focus on 25 diverse human bacterial pathogens from different phylogenetic orders and Gram staining groups (Extended Data Fig. S.1.a). In this section, we outline the experimental setup and present the results obtained from two analyses, namely:

(i) applying DeepEST to labeled data to assess performance, and
(ii) predicting the functions of unlabeled proteins.

Evaluating the performance of DeepEST

To assess the performance of DeepEST on predicting protein functions for the set of 25 bacterial species (Extended Data Fig. S.1), we obtained the label matrix Y and the tuple of protein descriptors, $X = (X_{\mathrm{s}}, X_{\mathrm{e}})$, which includes protein structures and the expression-location features (as described in Methods. 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*. We initially evaluated different data splitting strategies, accounting for protein structure and sequence separately. To perform these operations and evaluate the performance of our model, we employ 5-fold nested cross-validation. For each pathogen, we assess the performance of DeepEST with structure- and sequence-based methods: DeepFRI[10], BLAST [5], DIAMOND [6], DeepGOCNN and DeepGOplus [8], described in Extended Data. Moreover, 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 the application of our transfer learning technique.

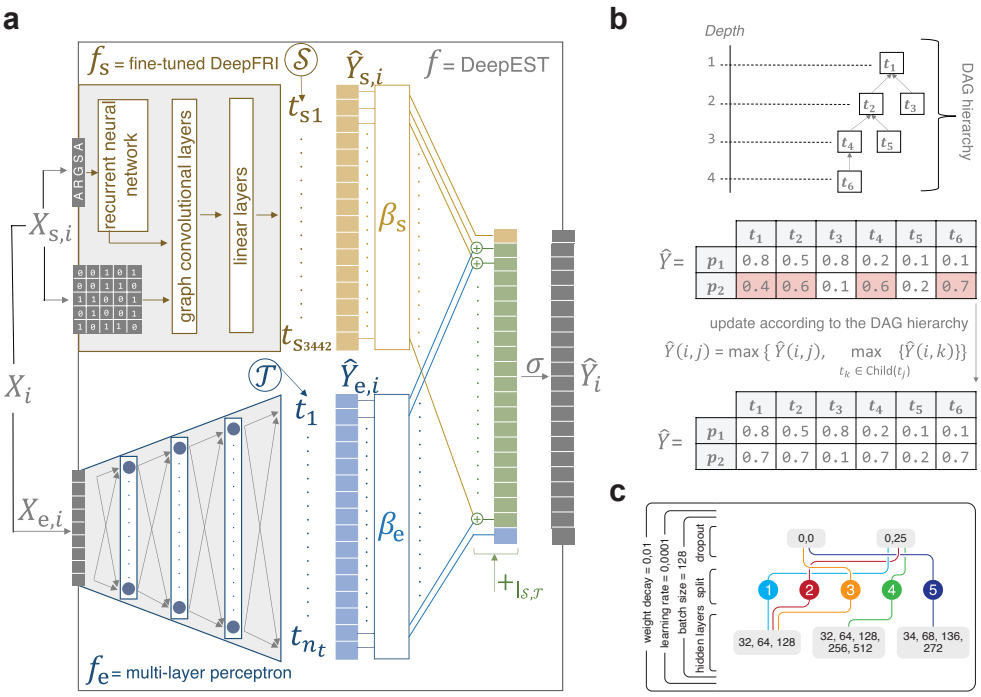

Fig. 1: Schematic illustration of DeepEST framework. **a** Our approach is composed of two modules: (i) the expression-location module f_e, a multi-layer perceptron, and (ii) the structure-based module f_s, DeepFRI [10] exploited in a transfer learning setting. A protein p_i is represented by the tuple X_i, composed of its expression-location features $X_{e,i}$, and its protein structure $X_{s,i}$, containing both the protein sequence and the adjacency matrix representing the amino acids contacts. f_e takes as input $X_{e,i}$ and outputs $\hat{Y}_{e,i} = \hat{Y}_e(i,:)$, while f_s takes as input $X_{s,i}$ and outputs $\hat{Y}_{s,i} = \hat{Y}_s(i,:)$. Predictions for protein p_i, $\hat{Y}_i$, are obtained by applying the sigmoid function (σ) on the masked linear combination ($+_{|\mathcal{S},\mathcal{T}}$) of $\hat{Y}_{s,i}$ and $\hat{Y}_{e,i}$. **b** At test step, predictions non-conforming the DAG hierarchy, visualized in red, are updated according to the rule from Eq. (9). **c** Hyperparameter configurations leading to the best validation performance for f_e in the five data splits.

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` [22], which enables us to cluster proteins with high structure similarity (local distance difference test score [23] $\geq 70\%$) 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 [24], 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

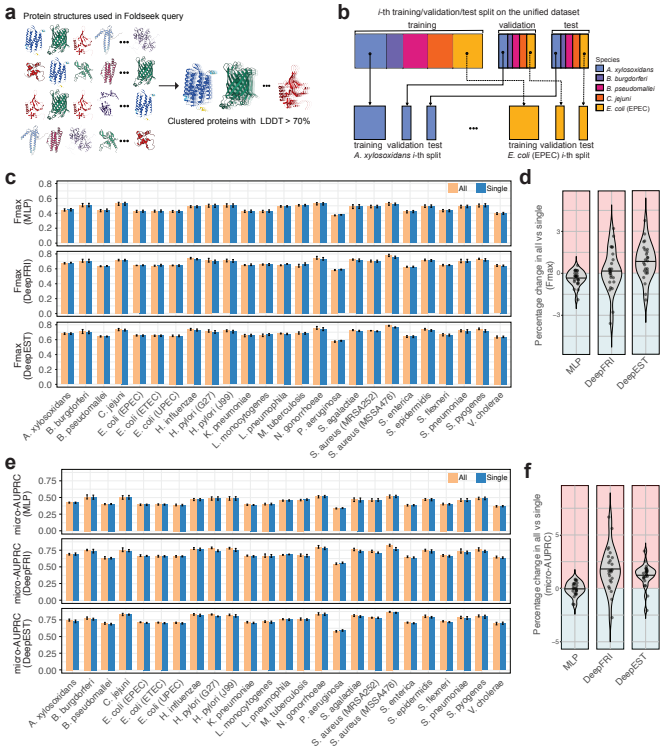

Fig. 2: Accounting structure similarity with a unified dataset from 25 species or a single species exhibits similar performance suggesting no data leakage in training, validation, and test sets. **a** Available protein structures from the 25 bacterial species retrieved from AlphaFold were used as queries 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 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 strait lines represent the mean values.

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

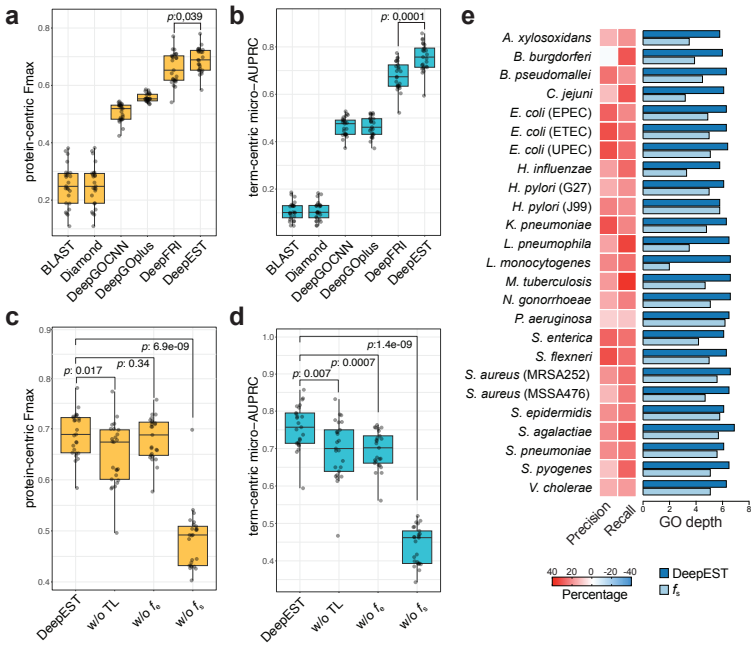

Fig. 3: DeepEST outperforms state-of-the-art methods in protein function predictions and achieves greater specificity in GO assignment by using expression location features. **a** The average protein-centric $F_{\max}$ and **b** term-centric *micro*-AUPRC values calculated for each species separately across the five sets for DeepEST and other methods: BLAST, Diamond, DeepGOCNN, DeepGOplus, 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 percentage change in precision and recall for each GO term across proteins (left panel), and average depth of the GO terms identified solely 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* in **a** and **b** were calculated with Wilcoxon rank-sum test.

sets (Extended Data Fig S.2, Fig S.3, Fig S.4, Fig S.5, Fig S.6). 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 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 Extended Data Fig S.6). 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 Extended Data.

DeepEST outperforms sequence and structure-based state-of-the-art methods

To compare DeepEST with state-of-the-art gene function annotation methods using different protein descriptors, we trained one model per species and compared to other methods. We used sequence-based traditional annotation methods (i.e., BLAST and Diamond), deep learning approaches (i.e., DeepGOCNN and DeepGOplus), and a structure-based deep learning method, DeepFRI. The $F_{\max}$ and *micro*-AUPRC scores, which are used to assess performance, were calculated for each species as described in Extended Data. DeepEST consistently outperforms all the comparison partners which employ sequences as protein descriptors (i.e., BLAST, Diamond, DeepGOCNN, and DeepGOplus), both in terms of term-centric *micro*-AUPRC and protein-centric $F_{\max}$ Fig. 3 (a and b and Extended Data Fig. S.7). These results highlight the importance of considering more than just protein sequences when predicting protein function in bacteria. Instead, the use of DeepFRI, which utilizes protein structure data, already enables superior performance in comparison to BLAST, Diamond, DeepGOCNN, and DeepGOplus. This suggests that considering protein structures is more insightful than solely relying on protein sequences for function prediction in bacteria. Moreover, DeepEST with usage of gene expression level 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. 3 a and b). These results are confirmed by the additional metrics we utilized, namely term-centric $s_{\min}$, term-centric *macro*-AUPRC, and protein-centric AUPRC, which are introduced in Extended Data and whose results are visualized in Fig S.8 and Fig S.9. These findings suggest that integrating the expression-location module along with harnessing the abundance of available protein structure data to refine the structure-based module allows for a more comprehensive analysis of protein functionality in bacteria.

All building blocks of DeepEST contributes to its performance

To understand the contributions of the various components of DeepEST, we performed an ablation study. We calculated protein-centric $F_{\max}$ and term-centric *micro*-AUPRC for each exclusion of the components: usage of transfer learning TL, fine-tuned structure-based module f_{s}, and the expression-location module f_{e}. The results in terms of protein-centric $F_{\max}$ and term-centric *micro*-AUPRC indicate that the use of transfer learning proves to be beneficial for DeepEST (Fig. 3 c and d), enhancing its performance across all the species and metrics (Extended Data, Table S.5 and Table S.4). Similarly, the use of transfer learning improves fine-tuned structure-based module f_{s} in comparison to DeepFRI (see Extended Data, Fig. S.10).

To assess the effect of incorporating the additional information captured by the expression and location data within DeepEST, we excluded f_{e} module. Considering protein-centric $F_{\max}$ score, results show that the expression-location module is not beneficial (Fig. 3c, Extended Data Table S.4). Nevertheless, term-centric *micro*-AUPRC score showed that incorporation of f_{e} module is significantly beneficial for DeepEST performance (Fig. 3d, Extended Data Table S.5). This suggests that the inclusion of expression-location features primarily provides insights into specific GO terms across

different proteins. Indeed, the incorporation of gene expression and location descriptors yields valuable information regarding specific biological processes and pathways, thereby enhancing predictions related to certain GO terms. Accordingly, when analyzing the GO term annotations predicted by DeepEST through thresholding the scores in $\hat{Y}$ with the optimal τ identified on the validation set, we noted an increase in the average precision and recall calculated across the five splits for each GO term in comparison to f_{s}. This improvement was noted for all species except for the average precision value in *Borrelia burgdorferi* (Fig. 3e and Extended Data Table S.6). 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 and Extended Data Table S.6). These values suggest that exploiting the expression-location features leads to the prediction of more specific GO terms by DeepEST.

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 and Extended Data Table S.5 and Table S.4). 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 (Extended Data Fig. S.12). This finding suggests that the node features derived by the recurrent neural network of DeepFRI effectively capture the information enclosed within the amino acid sequence [25], [10].

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 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 (Fig S.13). Most

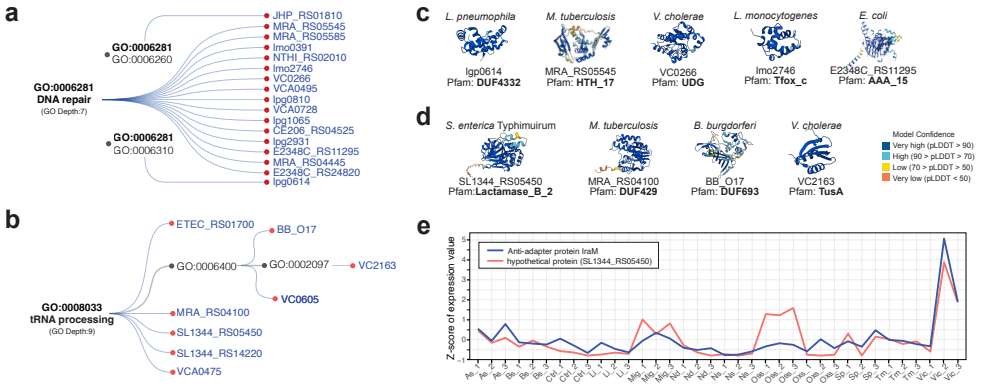

Fig. 4: DeepEST predicts functions via assigning specific biological process GO terms to bacterial hypothetical proteins with domains of known and unknown functions. Ontology tree of DeepEST-assigned **a** DNA repair (GO:00066281) with GO hierarchy depth of 7 and **b** tRNA processing (GO:0008033) with 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 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 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.

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 4a), while for RNA metabolism, we focused on "tRNA processing" (GO:0008033) and associated child terms (Fig 4b). DeepEST identified genes potentially involved in DNA repair across various species with diverse taxonomic and physiological traits (Extended Data Table S.7). 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 4c and Extended Data Table S.7) [26], [27], [28]. Notably, DeepEST predicted GO terms for proteins with domains of unknown function also, indicating its prediction capability for proteins with unknown functional signatures (Fig 4c Extended Data Table S.7). In the case of tRNA processing, we identified domains linked to beta-lactamase superfamily proteins and sulfurtransferase TusA (Fig 4d and Extended Data Table S.8), which are critical for tRNA modification and maturation [29], [30], [31]. 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 4 b and d, and Extended Data Table S.8), consistent with previous literature [31]. In *Salmonella*

enterica 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 4d). Additionally, using the Co-PATHOgenex web app [32] 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 4e. IraM is an 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* Typhimurium [33]. 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

This work introduces DeepEST, a deep learning framework that combines protein structures, gene expression levels and location data, in a transfer learning setting. DeepEST was applied to data from 25 bacterial species to predict protein functions. We evaluated two training strategies: developing species-specific models versus training a unified model across all species. When not accounting for structural similarity, both DeepEST and fine-tuned DeepFRI perform better due to potential data leakage, as similar protein structures appear in the training, validation, and test sets. Despite this, training on each species separately is preferred, as there is no clear advantage in training a unified model across species when structural similarity is considered. In addition, our analysis demonstrates that DeepEST significantly outperforms all comparison methods that predict protein functions solely from protein sequences. We used protein-centric $F_{\max}$ and term-centric *micro*-AUPRC scores to evaluate performance across all assays, as these metrics are commonly employed in protein prediction [34] Furthermore, findings from the strongest comparison partner, DeepFRI, indicate that incorporating protein structure data yields more insightful information than relying solely on amino acid sequences for predicting protein function in bacterial species. However, DeepEST enables superior performance by exploiting the wealth of high-quality predicted protein structure data available today to fine-tune its structure-based module, and by leveraging gene expression and location data. Interestingly, exploiting the information enclosed within the expression and location features permits retrieving annotations placed deeper in the GO hierarchy. This more specific indication of protein functionality could prove useful for planning targeted wet lab experiments aimed at validating the predicted functions, which could be highly challenging and time-consuming otherwise. Thus, when available, integrating contextual information such as gene expression levels and locations may lead to more accurate and specific functional annotations. Notably, gene expression datasets used in this study were obtained from explicitly designed conditions mimicking stress environments in the human host. These conditions are known to activate particular stress responses and biological processes through a change in gene expression, where this change could help predict functions of proteins involved in those biological processes such as proteins involved in DNA

repair, known to be activated under oxidative stress and low iron. Additionally, DeepEST's comprehensive exploration of gene ontology has revealed significant diversity and specificity in predicted GO terms using a set of functionally unannotated hypothetical proteins in clinically relevant human bacterial pathogens. Hence, DeepEST represents a valid method to annotate proteins and, in turn, a valuable support to plan experiments to investigate protein function. Future work could investigate the integration of alternative or additional data sources. Improvements could also be made via a more sophisticated integration procedure for the different data modalities. Finally, an intrinsic limitation of this study is the nature of the evidence of the annotations, which could cause circularity in our analysis. However, we hope that this will be mitigated in the future when more experimental annotations are available for prokaryotes.

Methods

Implementation of DeepEST

In this section, we present the implementation of DeepEST, describing its two distinct modules based on different data modalities and explaining their integration. We also outline how transfer learning is applied to enhance prediction accuracy, along with the equation used to estimate performance

Structure-based module

The current state-of-the-art in structure-based protein function prediction is DeepFRI [10]. This method 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:

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

where X_{s} is constituted by the protein sequences and structures, $f_{\mathrm{s}}(\cdot)$ corresponds to DeepFRI and is parameterized in $\mathbf{w}_{\mathrm{s}}$, 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 [35]. 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 retain the extracted features from the GCN layers and only retrain the final linear layers. Henceforth, $\mathbf{w}_{\mathrm{s}}$ denotes the subset of learnable weights of these final layers we retrain. The structure of f_{s} is visualized at the top of Fig. 1.a.

Expression-location module

As introduced before, more accurate protein function prediction in bacteria requires a multimodal approach. Hence, 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, designed to capture the gene proximity information. Given that the bacterial genomes, with rare exceptions [36], consist of one main chromosome with a circular structure and smaller circular plasmids, we encode each position into polar coordinates. We then generate features by calculating the genetic distance, which involves calculating the sine and cosine of the angles formed by the positioning of the start and end positions of each gene. Lastly, we encode whether a gene is found on the main chromosome or on any of the plasmids, and the coding strand the gene is located on. These n_{e} gene expression and location features, i.e. $X_{\mathrm{e}} \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{e}}}$, form a tabular dataset. We therefore use a multilayer perceptron (MLP) architecture to perform predictions according to this module,

$$f_{\mathrm{e}}(X_{\mathrm{e}}; \mathbf{w}_{\mathrm{e}}) = \hat{Y}_{\mathrm{e}} \ , \tag{4}$$

with $f_{\mathrm{e}}(\cdot)$ being the non-linear function representing the MLP and parameterized in $\mathbf{w}_{\mathrm{e}}$, 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. 1.a. Unlike the structure module, this module is trained *ab initio*.

Combining different modalities

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. (4), that allows to rewrite $f(\cdot)$, introduced in Eq. (2), as:

$$f(X_{\mathrm{s}}, X_{\mathrm{e}}; \mathbf{w}) = \mathrm{Sigmoid}\left(\beta_{\mathrm{s}} f_{\mathrm{s}}(X_{\mathrm{s}}; \mathbf{w}_{\mathrm{s}}) \ +_{|\mathcal{S},\mathcal{T}} \beta_{\mathrm{e}} f_{\mathrm{e}}(X_{\mathrm{e}}; \mathbf{w}_{\mathrm{e}})\right), \tag{5}$$

where the learnable parameters of DeepEST, denoted with $\mathbf{w}$, comprehend $\mathbf{w}_{\mathrm{s}}$, $\mathbf{w}_{\mathrm{e}}$, and the two integration parameters, β_{s} and β_{e}. $+_{|\mathcal{S},\mathcal{T}}$ denotes the masked sum performed accounting for the different sets of predicted GO terms. Indeed, exploiting $f_{\mathrm{s}}(\cdot)$ 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 $\hat{Y} \in \mathbb{R}^{n_{\mathrm{p}} \times |\mathcal{T} \cup \mathcal{S}|}$. Explicitly, the masked linear integration of the two modalities writes:

$$\hat{Y} = \mathrm{Sigmoid}\Bigg(\beta_{\mathrm{s}} \left[\hat{Y}_{\mathrm{s}}|_{\mathcal{T}\cap\mathcal{S}} \ \ \mathbf{0}_{n_{\mathrm{p}} \times |\mathcal{T}\setminus\mathcal{S}|} \ \ \hat{Y}_{\mathrm{s}}|_{\mathcal{S}\setminus\mathcal{T}}\right] + \tag{6}$$

$$\beta_{\mathrm{e}} \left[\hat{Y}_{\mathrm{e}}|_{\mathcal{T}\cap\mathcal{S}} \ \ \hat{Y}_{\mathrm{e}}|_{\mathcal{T}\setminus\mathcal{S}} \ \ \mathbf{0}_{n_{\mathrm{p}} \times |\mathcal{S}\setminus\mathcal{T}|}\right]\Bigg), \tag{7}$$

where $\hat{Y}_{\mathrm{e}}|_{(\cdot)}$ and $\hat{Y}_{\mathrm{s}}|_{(\cdot)}$ are the outputs of each module restricted to specific 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 $\mathcal{T}$ ($\mathcal{S} \setminus \mathcal{T}$); $\mathbf{0}_{n_{\mathrm{p}} \times |\mathcal{S} \setminus \mathcal{T}|}$ and $\mathbf{0}_{n_{\mathrm{p}} \times |\mathcal{T} \setminus \mathcal{S}|}$ 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. Finally, the output probabilities are calculated using a sigmoid layer Sigmoid$(\cdot)$. This operation is visualized in Fig 1.a.

Transfer learning on the structure-based module

Transfer learning is a machine learning technique that harnesses the learned features from a pre-trained model to enhance the learning process for a closely related task, where usually only a dataset of small size is available. This is commonly accomplished by fine-tuning only the output layers of the model. We apply this technique on DeepFRI [10], to tailor the model to the bacterial organism we are investigating. To this end, we use a masked loss function $\ell(\cdot)$,

$$\ell(\hat{Y}, Y) = \underbrace{\mathrm{BCE}\left(\hat{Y}|_{\mathcal{T}}, Y\right)}_{\text{experimental risk within } \mathcal{T}} + \underbrace{\lambda \sum_{\mathbf{w}} ||\mathbf{w}||_2}_{\text{regularization term}} , \tag{8}$$

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; $\lambda \sum_{\mathbf{w}} ||\mathbf{w}||_2$ is a weight decay regularization term with parameter λ [37]. 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 [19], which follows the scheme

$$\hat{Y}(i,j) = \max\Big\{\hat{Y}(i,j), \max_{t_k \in \mathrm{Children}(t_j)} \{\hat{Y}(i,k)\}\Big\}, \tag{9}$$

where Children(t_j) represents the children of the term t_j according to the DAG. A visualization is reported in Fig. 1.b. Note that at training and validation time, we do not perform such updating steps to retain a differentiable loss function.

Metrics

To assess the quality of the predicted protein labels $\hat{Y}$ in comparison to the ground truth labels Y, we use the term-centric and protein-centric metrics introduced in Critical Assessment of Functional Annotation (CAFA) challenge [38]. These two categories of measures capture different aspects of the protein function prediction performance, namely the capability of the model to predict for each GO term what proteins are annotated with it, and its ability in predicting the function given a protein.

As for the term-centric metrics, we employ the *micro*- averaging area under the precision-recall curve (*micro*-AUPRC), following [10]. This metric considers all the GO terms simultaneously and is obtained by:

(1) Flattening both Y and $\hat{Y}$ to obtain two vectors of size $n_{\mathrm{p}} n_{\mathrm{t}}$,
(2) Calculating the precision and recall of these vectors per each threshold τ, and
(3) Obtaining the AUPRC given the precision and recall vectors.

In regard to the protein-centric metric, we use the $F_{\max}-$score, defined as:

$$F_{\max} = \max_{\tau} \left\{ \frac{2 \cdot \mathrm{pr}(\tau) \cdot \mathrm{rc}(\tau)}{\mathrm{pr}(\tau) + \mathrm{rc}(\tau)} \right\} \tag{10}$$

where $\tau \in [0, 1]$ represents some decision threshold; $\mathrm{pr}(\tau)$ is obtained by averaging the precisions $\mathrm{pr}_i(\tau)$ across the proteins $\{p_i \in \mathcal{P} : \hat{Y}(i, j) \geq \tau,\ j = 1, \cdots, n_{\mathrm{t}}\}$; $\mathrm{rc}(\tau)$ is the average of the recall $\mathrm{rc}_i(\tau)$ calculated per each protein $p_i \in \mathcal{P}$. $\mathrm{pr}_i(\tau)$ and $\mathrm{rc}_i(\tau)$ are derived as follows:

$$\mathrm{pr}_i(\tau) = \frac{\sum_t I(t \in \hat{\mathcal{T}}_i(\tau) \wedge t \in \mathcal{T}_i)}{\sum_t I(t \in \hat{\mathcal{T}}_i(\tau))} \tag{11}$$

$$\mathrm{rc}_i(\tau) = \frac{\sum_t I(t \in \hat{\mathcal{T}}_i(\tau) \wedge t \in \mathcal{T}_i)}{I(t \in \mathcal{T}_i)} \tag{12}$$

with $\hat{\mathcal{T}}_i(\tau) = \{t_j \in \mathcal{T} : \hat{Y}(i, j) \geq \tau,\ j = 1, \cdots, n_{\mathrm{t}}\}$ being the set of predicted GO terms for protein p_i at threshold τ, $\mathcal{T}_i = \{t_j \in \mathcal{T} : Y(i, j) = 1,\ j = 1, \cdots, n_{\mathrm{t}}\}$ the set of GO terms protein p_i is annotated with, $I(\cdot)$ the standard indicator function, and t a GO term.

The additional metrics we utilize in some analyses are described in Extended Data.

Data

This section is devoted to introducing various data modalities utilized in our study, along with detailing the steps involved in data collection and preprocessing. Our analysis is specifically centered on the 25 bacterial species highlighted in Fig. S.1.a. For each of those, we conduct investigations using two datasets:

(1) A labeled dataset, used to assess performance, for which we are able to collect the matrix of the ground truth labels, and
(2) The dataset of hypothetical proteins, for which we are able to collect protein structures, gene expression and location features, while GO terms annotations are not available since these proteins have not been functionally characterized yet.

Concerning the labeled dataset, 32,573 proteins present both data modalities and GO term annotations. Instead, there are 6,997 hypothetical proteins for which both data views have been retrieved.

GO terms for protein function

As described in Results, we define protein function, i.e., the matrix of the ground truth labels Y, by utilizing GO terms annotations, with a specific focus on the Biological

Process hierarchy. To collect available GO annotations, we query the UniProt database [39] (accessed on July 12, 2023) using the RefSeq protein identifier [40] of every known protein and the taxonomic reference code [41] of a given pathogen's strain at hand. Note that the RefSeq protein identifier is a multi-species identifier and to obtain accurate annotations a taxonomy identifier has to be supplied. The number of unique GO terms obtained for each species can be found in Figure S.14. We used the GO ontology released on October 7, 2022 to retrieve a particular GO term's children or ancestors. We do not filter GO terms by experimental evidence codes due to the lack of experimentally confirmed GO annotations, as shown in Figure S.15. The distribution of the number of annotated genes and the number of GO terms in each of the 25 species is visualized in Figure S.14. Note that we retain a GO term in case it is annotated for at least 10 proteins per species.

Protein structure data

The development of protein structure prediction methods, such as AlphaFold [16, 17], and ESMFold [18], resulted in a wealth of high-quality protein structures across various species, including extensively investigated organisms as well as less studied species [17]. While experimentally confirmed protein structure may be available for some of the considered proteins, we only use structures in the AlphaFold database [17] in our approach, since obtaining all structures from one source mitigates the effect of covariate shifts in the input data. We query the AlphaFold database using the UniProt accession numbers of the proteins in question, thereby deriving the protein structure data $X_{\mathrm{s},i}$ for all $p_i \in \mathcal{P}$, i.e., for all the proteins we analyze, namely both the 32,573 proteins in the labeled dataset and the 6,997 hypothetical proteins.

Expression-location data

The expression-location features are derived from PATHOgenex dataset [9], which provides an in-depth global gene expression analysis across 11 host-mimicking stress conditions and control condition, listed in Fig. S.1.b. Specifically, as input features for our model, we consider the log-fold change values derived from the differential expression analysis of these 11 stress conditions compared to the control. As described before, we employed the engineered features derived from the genomic position and additional details on the preprocessing steps are reported in Extended Data. As done for the protein structure data, expression-location features $X_{\mathrm{e},i}$ are derived for all $p_i \in \mathcal{P}$, namely for the 32,573 proteins in the labeled dataset and the 6,997 hypothetical proteins.

Visualization

The visualization of model performance was conducted using ggplot2 [42] in R version 4.2.2 [43]. Separate datasets were prepared for each evaluation metric, including micro-AUPRC and F-max. Each dataset was structured to categorize data by model type, enabling comparison across models. For each plot, jittered points were overlaid on box plots to visualize data dispersion. Statistical significance was assessed using the Wilcoxon rank-sum test to calculate p-values.

Code availability

Accompanying code and data can be found under https://github.com/BorgwardtLab/DeepEST and https://github.com/avicanlab/DeepEST.

Supplementary information

Supplementary data, notes and results are included in the Extended data file.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement (No 813533). Swedish Research Council (No. 2021-02466), Swedish Research Council Excellence center grant (No. 2022-06543) (Center for modeling adaptive mechanisms in living systems under stress), Kempestiftelserna (JCK22-0017), Medical Faculty at Umeå University (FS 2.1.6-281-22) to K Avican and Icelab Multidisciplinary Postdoctoral Fellowship to L Fernandez.

References

[1] Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2017).

[2] Escudeiro, P., Henry, C. S. & Dias, R. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Current Research in Microbial Sciences* **3**, 100159 (2022).

[3] Ardern, Z. *et al.* Elucidating the functional roles of prokaryotic proteins using big data and artificial intelligence. *FEMS Microbiology Reviews* **47**, fuad003 (2023).

[4] Djemiel, C. *et al.* Inferring microbiota functions from taxonomic genes: a review. *GigaScience* **11**, giab090 (2022).

[5] Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

[6] Buchfink, B. *et al.* Fast and sensitive protein alignment using diamond. *Nature Methods* **12**, 59–60 (2015).

[7] P. Obregón, F. *et al.* Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning. *Scientific Reports* **12**, 11655 (2022).

[8] Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2019).

[9] Avican, K. *et al.* RNA atlas of human bacterial pathogens uncovers stress dynamics linked to infection. *Nat. Commun.* **12**, 3282 (2021).

[10] Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).

[11] Goodsell, D. S. *The machinery of life* (Springer, 2009).

[12] Ling, X. *et al.* Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics* **25** (2009).

[13] Kulmanov, M. *et al.* DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2017).

[14] Boadu, F. *et al.* Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics* **39**, i318–i325 (2023).

[15] Ünsal, S. *A deep learning based protein representation model for low-data protein function prediction.* Ph.D. thesis, Middle East Technical University (2023).

[16] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[17] Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2021).

[18] Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* (2022).

[19] Zhao, Y. J. *et al.* A Literature Review of Gene Function Prediction by Modeling Gene Ontology. *Front. Genet.* **11**, 583–589 (2020).

[20] Thomas, P. D. The gene ontology and the meaning of biological function. *The gene ontology handbook* 15–24 (2017).

[21] Biewald, L. Experiment tracking with weights and biases (2020). URL https://www.wandb.com/. Software available from wandb.com.

[22] van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 1546–1696 (2023).

[23] Mariani, V. *et al.* lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

[24] Kucera, T., Oliver, C., Chen, D. & Borgwardt, K. Proteinshake: building datasets and benchmarks for deep learning on protein structures. *Advances in Neural Information Processing Systems* **36** (2024).

[25] Fernández, M., Abreu, J. I., Caballero, J., Garriga, M. & Fernandez, L. 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. *Molecular Simulation* **33**, 1045–1056 (2007).

[26] Das, R. & Bajpai, U. Functional characterization of a dna-dependent aaa atpase in a f-cluster mycobacteriophage. *Virus Research* **323**, 198957 (2023).

[27] Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS microbiology reviews* **29**, 231–262 (2005).

[28] Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic acids research* **49**, D412–D419 (2021).

[29] Bhatta, A. *et al.* Molecular basis of human nuclear and mitochondrial trna 3'-processing. *bioRxiv* 2024–04 (2024).

[30] Nakanishi, K. & Nureki, O. Recent progress of structural biology of trna processing and modification. *Molecules and Cells* **19**, 157–166 (2005). URL https://www.sciencedirect.com/science/article/pii/S1016847823131517.

[31] Sato, Y. *et al.* Tusdcb, a sulfur transferase complex involved in trna modification, contributes to upec pathogenicity. *Scientific Reports* **14**, 8978 (2024).

[32] Fernandez, L., Rosvall, M., Normark, J., Fällman, M. & Avican, K. Copathogenex web application for assessing complex stress responses in pathogenic bacteria. *Microbiology Spectrum* **12**, e02781–23 (2024).

[33] Bougdour, A., Cunning, C., Baptiste, P. J., Elliott, T. & Gottesman, S. Multiple pathways for regulation of σs (rpos) stability in escherichia coli via the action of multiple anti-adaptors. *Molecular microbiology* **68**, 298–313 (2008).

[34] Zhang, C. & Freddolino, L. A large-scale assessment of sequence database search tools for homology-based protein function prediction. *Briefings in Bioinformatics* **25** (2024).

[35] Kipf, T. N. & Welling, M. *Semi-supervised classification with graph convolutional networks.* (2017).

[36] Hepner, S. *et al.* A high fidelity approach to assembling the complex Borrelia genome. *BMC Genomics* **24**, 401 (2023).

[37] Zhang, G. *et al.* Three mechanisms of weight decay regularization. *arXiv:1810.12281* (2018).

[38] Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).

[39] Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2022).

[40] O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).

[41] Federhen, S. *et al.* The NCBI Taxonomy database. *Nucleic Acids Research* **40**, D136–D143 (2012).

[42] Wickham, H. ggplot2. *Wiley interdisciplinary reviews: computational statistics* **3**, 180–185 (2011).

[43] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2022). URL https://www.R-project.org/.

[44] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv:1711.05101* (2017).

[45] Fu, L. *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

[46] Wilgenbusch, J. & Swofford, D. Inferring evolutionary trees with PAUP*. *Current Protocols in Bioinformatics* **6**, 4 (2003).

[47] Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

Extended Data

Preprocessing of expression-location module

As described in the main, we utilize gene expression levels across various stress conditions (acidic stress, bile stress, low iron, nitrosative stress, oxidative stress, osmotic stress, nutritional downshift, hypoxia, stationary phase, temperature and virulence inducing condition, as shown in Fig. S.1.b) and location features as input to the expression-location module of DeepEST. As a preprocessing step for these features X_e, we use `MinMaxScaler()` from `sklearn` to scale these features within the range $[0, 1]$.

Note that *B. burgdorferi*, *B. pseudomallei*, *H. influenzae*, *H. pylori (G27)*, *H. pylori (J99)*, and *M. tuberculosis* miss the measurements in one stress condition, while *L. pneumophila* lacks measurements in two stress conditions. When studying each species separately, these missing features are not considered. Instead, when analyzing multiple species simultaneously, these missing data are imputed with the average log fold expression under the same condition, across the other species.

Description of baselines

The additional baselines we use include methods taking protein sequences as input. We use traditional annotation methods (i.e., BLAST and Diamond) as well as deep learning-based approaches, such as DeepGOCNN and DeepGOplus. In the following, we provide a description of these tools.

- BLAST: the implementation of this baseline follows what done in [8]. Briefly, given a query protein from the validation or test set, its annotations are derived from the ones of the most similar protein in the training set, identified using sequence alignment algorithm BLAST [5] (we use its fastest implementation available under the Diamond tool [6]).
- Diamond: similarly to the baseline above, Diamond calculates annotation scores for a query protein by normalizing the sum of the alignment scores of similar proteins identified in the training set.
- DeepGOCNN: this is the convolutional part of DeepGOplus [8], and consists of several convolution filters, each with a different kernel size, to extract features directly from the input sequence. The resulting features are then pooled and concatenated, and given as input to a final linear layer with sigmoid activation that provides the output annotation scores.
- DeepGOplus: the annotations are obtained by weighting the predictions obtained using Diamond and DeepGOCNN.

Hyperparameter optimization and training

The hyperparameter optimization for f_e is performed by means of grid search executed using W&B [21]. The hyperparameter optimized are listed in Table S.1. This hyperparameter search leads to having 192 potential combinations for each data split, 960 in total. The best configuration minimizes the multi-target binary cross-entropy loss on the validation set. Note that since we are using 5-fold nested cross-validation, we

obtain a hyperparameter configuration per each of the five splits, as reported in Fig 1.c. Hence, this selection process is performed five times. However, to reduce computation and have a unified model per split across the 25 species, we perform the hyperparameter optimization on a unified dataset, which comprises all the proteins across the 25 species. Note that to avoid data leakage on this unified dataset as well, the structure-aware data splits generation using `foldseek` is performed on this dataset, to avoid very similar proteins from different species being in the training and test set at the same time. Hence, the species-specific splits are subsequently obtained by sub-setting the unified dataset, as visualized in Figure 1. After having defined the best hyperparameter configurations across species, the sizes of the input and output are adapted to the number of expression-location features and GO terms accordingly. For details on the different sizes of the input and output, refer to the main text and Figure S.14 respectively.

DeepEST, which features the combination of the MLP with the best hyperparameter configuration and DeepFRI subject to transfer learning, is trained on each species separately (ADAMW optimizer [44], cosine annealing with 10 epochs of linear warm up, base learning rate 0.0001, early-stopping based on validation loss with patience 4, maximum number of epochs 300). The reasoning behind the choice is that we did not observe any strong advantage in training across multiple species compared to training on each species separately.

Importance of a proper evaluation framework

A commonly used strategy to generate data splits in protein function prediction foresees accounting for sequence similarity while generating the training/validation/test sets. Such strategy is aimed at diminishing possible data leakage from the training to the validation and test sets by clustering proteins with *high* sequence identity in the same fold. This allows to circumvent situations where proteins with highly similar amino acid sequences are present simultaneously in the training, validation and test set, thereby avoiding positively biased performance assessment. Such a strategy is necessary when calculating the performance of methods relying on amino acid sequences as input data, like many traditional tools such as BLAST [5].

Given the recent availability of protein structure data and methods exploiting such data, it becomes necessary to modify this paradigm to account for structure similarities as well, to avoid overly positive performance assessment. To estimate the effect of not accounting for protein structure similarity when evaluating the performance of a method that exploits structures, we apply DeepEST on splits generated while accounting for sequence similarity only. Highly similar proteins, that is, proteins presenting sequence identity higher or equal than 70%, are clustered together in either the training, validation, or test set. This is realized through the use of `CD-HIT` [45] clustering algorithm. Analogously to the pipeline applied to structure-based splits, this split generation procedure and the hyperparameter optimization are performed on the unified dataset generated by merging the data from the 25 species. The optimal configurations for the expression-location module are summarized in Table S.2.

Results obtained according to the different splitting strategies are reported in Fig. S.2-S.3 and Fig.S.4-S.5, which present performance obtained with the structure-based and sequence-based data split procedure respectively. Note that since our study focuses on 25 different species, two different training strategies are possible, namely:

(i) considering each species separately, thereby training 25 species-specific models, and

(ii) considering all 25 species simultaneously, allowing to have a unique model trained on a unified dataset.

In this experiment, we consider and compare both scenarios. The performance of both DeepEST and fine-tuned DeepFRI is generally higher when training on data splits generated *not* accounting for structure similarity (i.e., *sequence*-based data splits) in comparison to accounting for it. This applies to both the species-specific and the unified analysis. A possible explanation lays in the fact that there is data leakage, i.e., very similar protein structures are present in the training, validation and test set simultaneously, therefore leading to overly positive performance assessment since the validation and the test set contain protein structures already seen during training. In addition, when structure similarity is accounted for when generating the splits, there is no evident benefit in training DeepEST and fined-tuned DeepFRI on multiple species simultaneously, as visualized in Fig. S.2 and Fig. S.3. In fact, *micro*-AUPRC and $F_{\max}$ result comparable (i.e., within one standard deviation from the average) between the two training strategies in 60% and 76% of the species, respectively. For DeepFRI, these percentages are 56% and 80% respectively. When structure similarity is not taken into account during the process of generating the splits, instead, it appears that training on all species simultaneously is beneficial for both DeepEST and fine-tuned DeepFRI. As reported in Fig. S.4 and Fig. S.5, there is an increase in performance in 97% of the cases. Given the previously reported observations, this increase in performance seems to be due to data leakage: very similar structures, from different species, appear in both training, validation and test set, since structure similarity is not considered when generating the data splits. Finally, the different data splits generation and training strategies have negligible impact on the expression-location module. In light of these observations, we retain as default splitting strategy the one accounting for structure similarity. With this choice, we can reduce the phenomenon of data leakage from the training to the validation and test set. Such considerations have been recently reported in [24]. Furthermore, we decide to train on each species separately, as there is no evident benefit in training on all species simultaneously. Note that these decisions are made on the results obtained on the validation sets.

Additional evaluation metrics

In addition to $F_{\max}$ and *micro*-AUPRC, we complemented our analysis with additional metrics, namely term-centric minimum semantic distance ($s_{\min}$), term-centric macro-averaged area under the precision-recall curve (*macro*-AUPRC), and protein-centric area under the precision-recall curve (AUPRC).

The minimum semantic distance $s_{\min}$ is defined according to:

$$s_{\min} = \min_{\tau} \sqrt{ru(\tau)^2 + mi(\tau)^2} \tag{S.1}$$

where $\tau \in [0, 1]$ is the decision threshold, $ru(\cdot)$ and $mi(\cdot)$ represent the uncertainty and the misinformation respectively and are calculated as follows:

$$ru(\tau) = \frac{1}{n_{\mathrm{p}}} \sum_{i=1}^{n_{\mathrm{p}}} \sum_{t} \mathrm{IC}(t) \cdot I(t \notin \hat{\mathcal{T}}_i(\tau) \wedge t \in \mathcal{T}_i) \tag{S.2}$$

$$mi(\tau) = \frac{1}{n_{\mathrm{p}}} \sum_{i=1}^{n_{\mathrm{p}}} \sum_{t} \mathrm{IC}(t) \cdot I(t \in \hat{\mathcal{T}}_i(\tau) \wedge t \notin \mathcal{T}_i). \tag{S.3}$$

In the two equations above, $I(\cdot)$ is the standard indicator function, t a GO term, $\hat{\mathcal{T}}_i(\tau) = \{t_j \in \mathcal{T} : \hat{Y}(i, j) \geq \tau,\ j = 1, \cdots, n_{\mathrm{t}}\}$ the set of predicted GO terms for protein p_i at threshold τ, $\mathcal{T}_i = \{t_j \in \mathcal{T} : Y(i, j) = 1,\ j = 1, \cdots, n_{\mathrm{t}}\}$ the set of GO terms protein p_i is annotated with, and $\mathrm{IC}(t)$ the information content of term t. The latter represents the term's specificity by accounting for its frequency of annotations to the proteins in the dataset, and is calculated as $IC(t) = -\log_2 Pr(t)$. $Pr(t)$ is the probability of term t and is derived as the number of proteins annotated to t and its descendants, divided by the total number of annotated proteins.

The term-centric macro-averaged area under the precision-recall curve, *macro*-AUPRC, is derived by:

(i) calculating precision and recall for each term t at for each threshold $\tau \in [0, 1]$;
(ii) deriving the area under the precision-recall curve using the trapezoid rule for each term t;
(iii) averaging these areas under the precision-recall curve across the terms.

The protein-centric area under the precision-recall curve, AUPRC, is derived by calculating $\mathrm{pr}(\tau)$ and $\mathrm{rc}(\tau)$ introduced in Eq. (6) for each threshold τ, and subsequently calculating the area under the $\mathrm{pr}(\tau)$-$\mathrm{rc}(\tau)$ precision-recall curve by utilizing the trapezoid rule.

Supplementary tables

Table S.1: To choose the best hyperparameter configuration for the experimental module (i.e., the MLP), we perform a grid search using W&B. The hyperparameters and the values we tried are reported in this table, leading to 192 different combinations for each data split (960 total runs). We perform this hyperparameter optimization on both the structure-based and sequence-based splits.

Hyperparameter	**values**
hidden layers	- [32, 64, 128, 256, 512] - [34, 68, 136, 272] - [32, 64, 128] - [32, 64] - [512, 16] - [512, 16, 512] - [32, 256, 64, 512] - [32, 128, 32, 128]
dropout	- 0.0 - 0.25
activation	`ReLU`
batch size	- 32 - 64 - 128
base learning rate	- 0.01 - 0.0001
weight decay	- 0.0 - 0.01

Table S.2: Best hyperparameter configurations identified for the MLP in the five *sequence*-based data splits.

Split	**hidden layers**	**dropout**	**batch size**	**learning rate**	**weight decay**
1	[32, 128, 32, 128]	0.25	32	0.0001	0.01
2	[32, 128, 32, 128]	0.25	32	0.0001	0.01
3	[32, 128, 32, 128]	0.25	32	0.0001	0.01
4	[32, 64, 128, 256, 512]	0.25	32	0.0001	0.01
5	[32, 128, 32, 128]	0.25	32	0.0001	0.01

Table S.3: Sizes of the input and output layers for the MLP for each species.

Species	**Input size**	**Output size**
all species	17	1711
A. xylosoxidans	17	239
B. burgdorferi	16	159
B. pseudomallei	16	387
C. jejuni	17	210
E. coli (EPEC)	17	445
E. coli (ETEC)	17	435
E. coli (UPEC)	17	416
H. influenzae	16	242
H. pylori (G27)	16	179
H. pylori (J99)	16	203
K. pneumoniae	17	404
L. pneumophila	15	289
L. monocytogenes	17	376
M. tuberculosis	16	301
N. gonorrhoeae	17	233
P. aeruginosa	17	694
S. enterica	17	449
S. flexneri	17	420
S. aureus (MRSA252)	17	291
S. aureus (MSSA476)	17	198
S. epidermidis	17	271
S. agalactiae	17	229
S. pneumoniae	17	238
S. pyogenes	17	204
V. cholerae	17	474

Table S.4: Assessment of the contributions of the components of DeepEST, by removing one at a time: (i) the use of our transfer learning (TL) technique, (ii) the expression-location module, and (iii) the structure module. The results are in terms of the average $F_{\max}$. ± standard deviation across the five test sets. In bold, we highlight the highest average score achieved per each species.

Species	**Protein-centric F-max**			
	DeepEST	**w/o TL**	**w/o f_e**	**w/o f_s**
A. xylosoxidans	**0.689±0.026**	0.683±0.042	**0.688±0.028**	0.443 ± 0.025
B. burgdorferi	0.698±0.018	0.674±0.041	**0.704±0.032**	0.502±0.021
B. pseudomallei	**0.647±0.018**	0.585±0.027	**0.648±0.019**	0.438±0.009
C. jejuni	**0.725±0.016**	0.700±0.030	0.721±0.021	0.517±0.020
E. coli (EPEC)	**0.652±0.013**	0.599±0.027	0.644±0.017	0.430±0.011
E. coli (ETEC)	**0.652±0.009**	0.594±0.023	0.640±0.017	0.432±0.012
E. coli (UPEC)	**0.653±0.011**	0.609±0.021	0.646±0.016	0.432±0.008
H. influenzae	**0.728±0.210**	0.724±0.023	**0.728±0.022**	0.484±0.024
H. pylori (G27)	**0.716±0.016**	0.678±0.032	0.709±0.021	0.512±0.028
H. pylori (J99)	**0.717±0.020**	0.680±0.022	0.709±0.021	0.509±0.025
K. pneumoniae	**0.659±0.015**	0.624±0.021	0.651±0.013	0.430±0.014
L. pneumophila	**0.674±0.021**	0.621±0.037	0.659±0.029	0.492±0.019
L. monocytogenes	**0.673±0.021**	0.620±0.027	0.662±0.026	0.427±0.013
M. tuberculosis	**0.687±0.010**	0.619±0.015	0.664±0.004	0.504±0.017
N. gonorrhoeae	**0.742±0.017**	0.708±0.021	0.728±0.019	0.521±0.018
P. aeruginosa	**0.584±0.015**	0.496±0.030	0.577±0.019	0.541±0.022
S. enterica	**0.637±0.008**	0.583±0.021	0.624±0.011	0.426±0.004
S. flexneri	**0.667±0.003**	0.601±0.022	0.658±0.003	0.445±0.014
S. aureus (MRSA252)	**0.720±0.012**	0.675±0.022	0.704±0.012	0.492±0.023
S. aureus (MSSA476)	**0.780±0.021**	0.773±0.017	0.757±0.009	0.535±0.026
S. epidermidis	**0.724±0.021**	0.710±0.035	0.717±0.019	0.502±0.022
S. agalactiae	**0.722±0.026**	0.698±0.043	0.713±0.031	0.491±0.027
S. pneumoniae	**0.727±0.016**	0.700±0.024	0.715±0.018	0.698±0.014
S. pyogenes	**0.716±0.025**	0.689±0.029	0.711±0.020	0.504±0.021
V. cholerae	**0.641±0.014**	0.587±0.029	**0.640±0.011**	0.403±0.020

Table S.5: Assessment of the contributions of the components of DeepEST, by removing one at a time: (i) the use of our transfer learning (TL) technique, (ii) the expression-location module, and (iii) the structure module. The results are in terms of the average *micro*-AUPRC ± standard deviation across the five test sets. In bold, we highlight the highest average score achieved per each species.

Species	**Term-centric AUPRC**			
	DeepEST	**w/o TL**	**w/o f_e**	**w/o f_s**
A. xylosoxidans	**0.737±0.046**	0.695±0.048	0.702±0.041	0.417±0.019
B. burgdorferi	**0.768±0.035**	0.721±0.035	0.723±0.049	0.500±0.029
B. pseudomallei	**0.691±0.035**	0.626±0.024	0.635±0.030	0.397±0.012
C. jejuni	**0.834±0.030**	0.789±0.020	0.756±0.035	0.491±0.027
E. coli (EPEC)	**0.714±0.022**	0.621±0.023	0.666±0.023	0.391±0.015
E. coli (ETEC)	**0.710±0.025**	0.625±0.021	0.656±0.025	0.393±0.017
E. coli (UPEC)	**0.714±0.029**	0.639±0.019	0.661±0.026	0.392±0.015
H. influenzae	**0.816±0.029**	0.776±0.030	0.757±0.034	0.465±0.025
H. pylori (G27)	**0.809±0.014**	0.771±0.014	0.747±0.046	0.503±0.036
H. pylori (J99)	**0.813±0.017**	0.770±0.018	0.753±0.039	0.490±0.033
K. pneumoniae	**0.717±0.025**	0.650±0.020	0.655±0.021	0.393±0.017
L. pneumophila	**0.751±0.025**	0.700±0.025	0.676±0.034	0.452±0.019
L. monocytogenes	**0.723±0.027**	0.667±0.023	0.669±0.035	0.399±0.014
M. tuberculosis	**0.757±0.020**	0.694±0.011	0.673±0.009	0.470±0.017
N. gonorrhoeae	**0.831±0.019**	0.791±0.017	0.763±0.018	0.507±0.023
P. aeruginosa	**0.594±0.025**	0.467±0.027	0.561±0.027	0.343±0.010
S. enterica	**0.697±0.019**	0.629±0.021	0.635±0.021	0.385±0.008
S. flexneri	**0.726±0.019**	0.650±0.017	0.672±0.008	0.410±0.013
S. aureus (MRSA252)	**0.782±0.025**	0.717±0.027	0.714±0.031	0.463±0.026
S. aureus (MSSA476)	**0.857±0.021**	0.832±0.017	0.756±0.038	0.520±0.034
S. epidermidis	**0.785±0.029**	0.744±0.021	0.728±0.029	0.472±0.026
S. agalactiae	**0.795±0.033**	0.750±0.038	0.734±0.054	0.462±0.035
S. pneumoniae	**0.792±0.027**	0.744±0.027	0.726±0.028	0.464±0.019
S. pyogenes	**0.785±0.019**	0.739±0.027	0.720±0.031	0.480±0.035
V. cholerae	**0.708±0.041**	0.613±0.042	0.650±0.041	0.375±0.019

Table S.6: Comparison between f_s (i.e., fine-tuned DeepFRI) and DeepEST in terms of percentage change in precision, recall and F1-score calculated for each GO term across proteins. These annotations are generated by thresholding the annotation scores using the threshold that maximizes the F1-score on the validation set. We furthermore report the average and median of the GO terms that only DeepEST is capable to correctly retrieving, shown under "Depth DeepEST", in comparison to the average and median GO terms that both methods are capable to identify, under "Depth intersection", and to the average and median depth of the GO terms that solely f_s was capable of correctly identifying, i.e. "Depth f_s".

Species	**% change vs. f_s**			**Depth DeepEST**		**Depth intersec.**		**Depth f_s**	
	Prec.	Recall	F1	Av.	Med.	Av.	Med.	Av.	Med.
A. xylosoxidans	+14.1	+20.5	+19.9	5.8	5.0	4.7	4.0	3.5	4.0
B. burgdorferi	-1.9	+30.8	+16.3	6.0	5.0	4.2	4.0	3.9	4.0
B. pseudomallei	+25.8	+20.2	+27.4	6.3	6.0	5.1	5.0	4.5	4.5
C. jejuni	+12.5	+30.8	+24.7	6.1	5.0	4.8	4.0	3.2	3.5
E. coli (EPEC)	+28.8	+22.1	+29.0	6.3	6.0	5.0	5.0	4.9	5.0
E. coli (ETEC)	+32.1	+26.7	+33.5	6.3	6.0	5.0	5.0	5.0	5.0
E. coli (UPEC)	+32.0	+26.5	+33.0	6.4	6.0	5.0	5.0	5.1	5.0
H. influenzae	+17.9	+27.5	+26.4	5.8	6.0	4.8	4.0	3.3	3.0
H. pylori (G27)	+15.2	+20.7	+22.1	6.1	6.0	4.5	4.0	5.0	5.0
H. pylori (J99)	+23.5	+21.2	+26.3	5.8	5.5	4.6	4.0	5.8	7.0
K. pneumoniae	+31.4	+22.7	+31.5	6.3	6.0	5.0	5.0	4.8	5.0
L. pneumophila	+17.4	+34.0	+27.4	6.5	6.0	5.0	5.0	3.5	3.5
L. monocytogenes	+22.2	+26.4	+27.9	6.6	7.0	5.2	5.0	2.0	2.0
M. tuberculosis	+18.9	+36.7	+32.4	6.6	7.0	4.9	4.0	4.7	4.0
N. gonorrhoeae	+15.8	+23.7	+24.5	6.6	7.0	4.8	4.0	5.1	4.0
P. aeruginosa	+8.9	+11.1	+12.0	6.5	6.0	5.4	5.0	6.2	6.0
S. enterica	+28.4	+26.0	+32.2	6.1	6.0	5.1	5.0	4.2	4.0
S. flexneri	+32.1	+26.5	+34.1	6.3	6.0	5.0	5.0	5.0	5.0
S. aureus (MRSA252)	+20.5	+26.3	+27.2	6.6	7.0	5.0	5.0	5.6	7.0
S. aureus (MSSA476)	+13.5	+24.6	+24.0	6.5	7.0	4.9	4.0	4.7	4.5
S. epidermidis	+21.0	+24.6	+27.1	6.1	6.0	4.9	4.0	5.8	5.5
S. agalactiae	+22.0	+29.8	+29.8	6.9	8.0	4.8	4.0	5.7	5.0
S. pneumoniae	+21.6	+25.6	+27.5	6.1	6.0	4.9	4.0	5.6	5.0
S. pyogenes	+11.9	+28.4	+22.2	6.5	7.0	4.7	4.0	5.1	5.0
V. cholerae	+15.2	+18.8	+20.8	6.3	6.0	5.2	5.0	5.1	5.0

Table S.7: Genes with GO:0006281 (DNA repair)

Species/Strain	Locus tag	Score	Pfam Domain	Description
A. xylosoxidans	CE206_RS04525	0.73	DUF4880 FecR	Unknown function FecR protein
E. coli (EPEC)	E2348C_RS11295 E2348C_RS24820	0.99 0.98	AAA_15 AAA_5	AAA ATPase domain AAA domain
E. coli ETEC	ETEC_RS15430	0.45		
H. pylori J99	JHP_RS01810	0.99	AAA_21	AAA domain
L. monocytogenes	lmo0391 lmo2746	0.71 0.53	NfeD TfoX_C	NfeD-like C-terminal TfoX C-terminal domain
L. pneumophila	lpg0614 lpg0810 lpg1065 lpg2931	0.88 0.73 0.45 0.69	DUF4332 HHH_3 YccV-like	Unknown function Helix-hairpin-helix motif YccV like
M. tuberculosis	MRA_RS04445 MRA_RS05545 MRA_RS05585	0.47 0.50 0.50	AbiEi_2 AbiEi_4 AbiEii HTH_17	AbiEi antitoxin Transcriptional regulator AbiEii toxin Helix-turn-helix domain
H. influenzae	NTHI_RS02010	0.73		
V. cholerae	VC0266 VCA0495 VCA0728	0.43 0.50 0.60	UDG DUF3800 AAA_15	Uracil DNA glycosylase Unknown function AAA ATPase domain

Table S.8: Genes with GO:0008033 (tRNA processing)

Species/Strain	Locus tag	Score	Pfam Domain	Description
B. burgdorferi	BB_O17	0.74	DUF693	Unknown function
M. tuberculosis	MRA_RS04100	0.50	DUF429	Unknown function
S. enterica	SL1344_RS05450 SL1344_RS14220	0.37 0.42	Lactamase_B_2 DUF4435	Beta-lactamase Unknown function
V. cholerae	VC0605 VC2163 VCA0475	0.69 0.84 0.34	UPF0231 TusA Fic	Uncharacterised protein Sulfurtransferase TusA Fic/DOC family

Supplementary figures

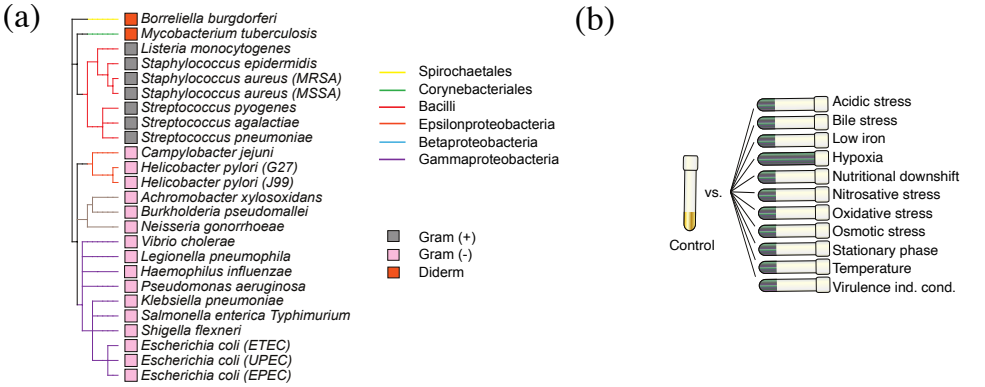

Fig. S.1: Overview of the bacteria and expression features utilized in this study: (a) Phylogenetic clustering of the set of 25 selected strains (obtained using PhyloT in Newick [46] format and represented using iTol [47]), and (b) the experimental setup of differential expression dataset from Avican et al. [9].

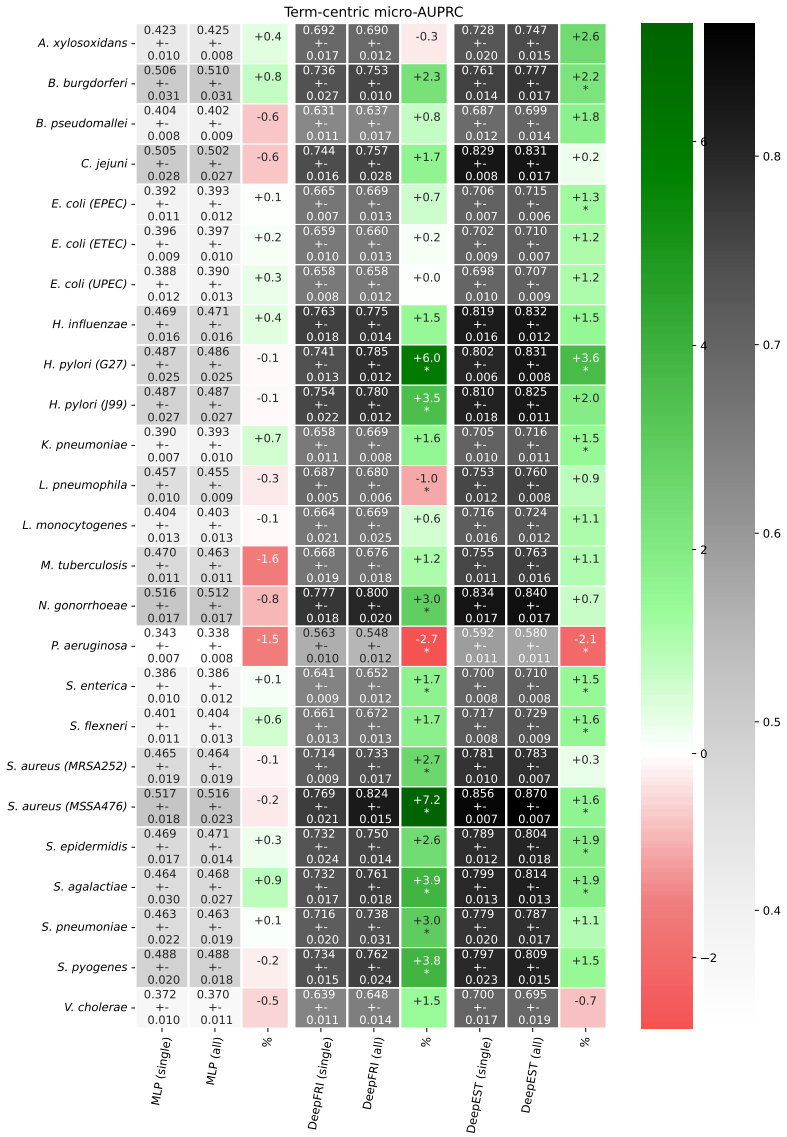

Fig. S.2: Comparison between training MLP (i.e., f_{e}), fine-tuned DeepFRI (i.e., f_{s}) and DeepEST on each species separately ("single") and on the unified dataset comprised of the 25 species ("all") in terms of average *micro*-AUPRC on the five *validation* sets, with *structure*-based splits. Each model presents three columns, namely "single", "all" and %. The latter is the percentage change obtained when training according to the "single" strategy in contrast to the "all", with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

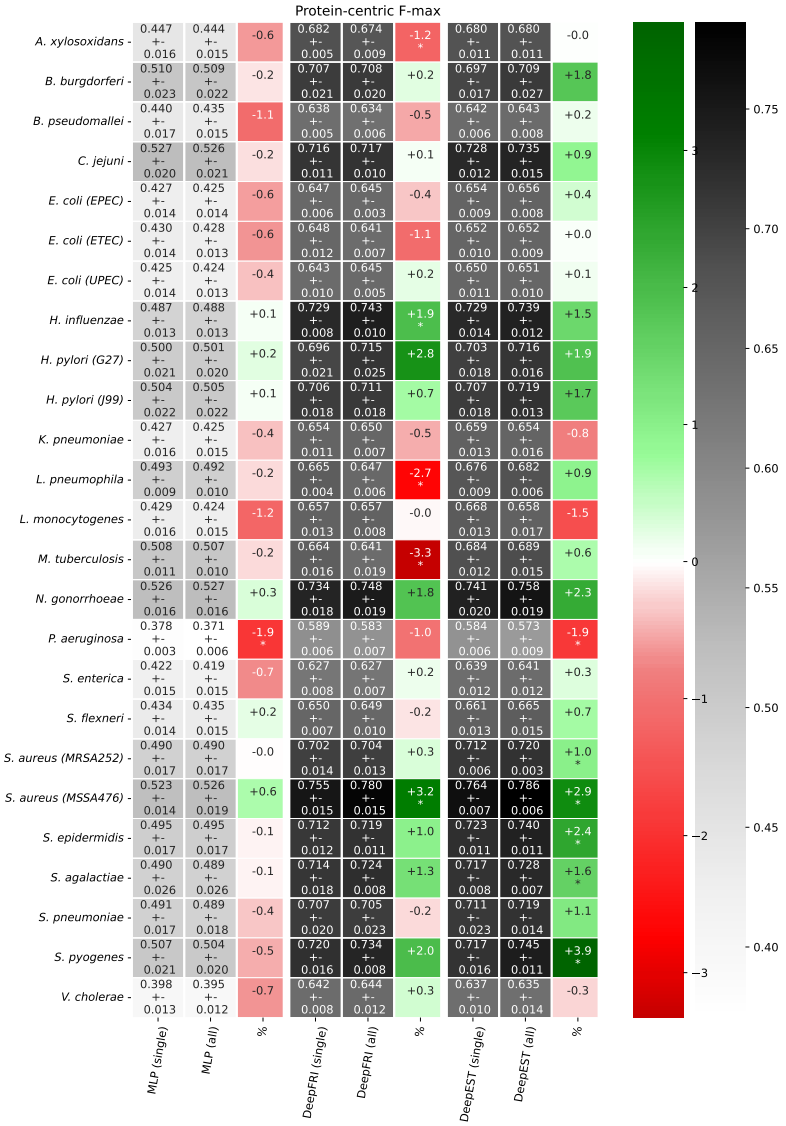

Fig. S.3: Comparison between training MLP (i.e., f_{e}), fine-tuned DeepFRI (i.e., f_{s}) and DeepEST on each species separately ("single") and on the unified dataset comprised of the 25 species ("all") in terms of average $F_{\max}$ on the five *validation* sets, with *structure*-based splits. Each model presents three columns, namely "single", "all" and %. The latter is the percentage change obtained when training according to the "single" strategy in contrast to the "all", with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

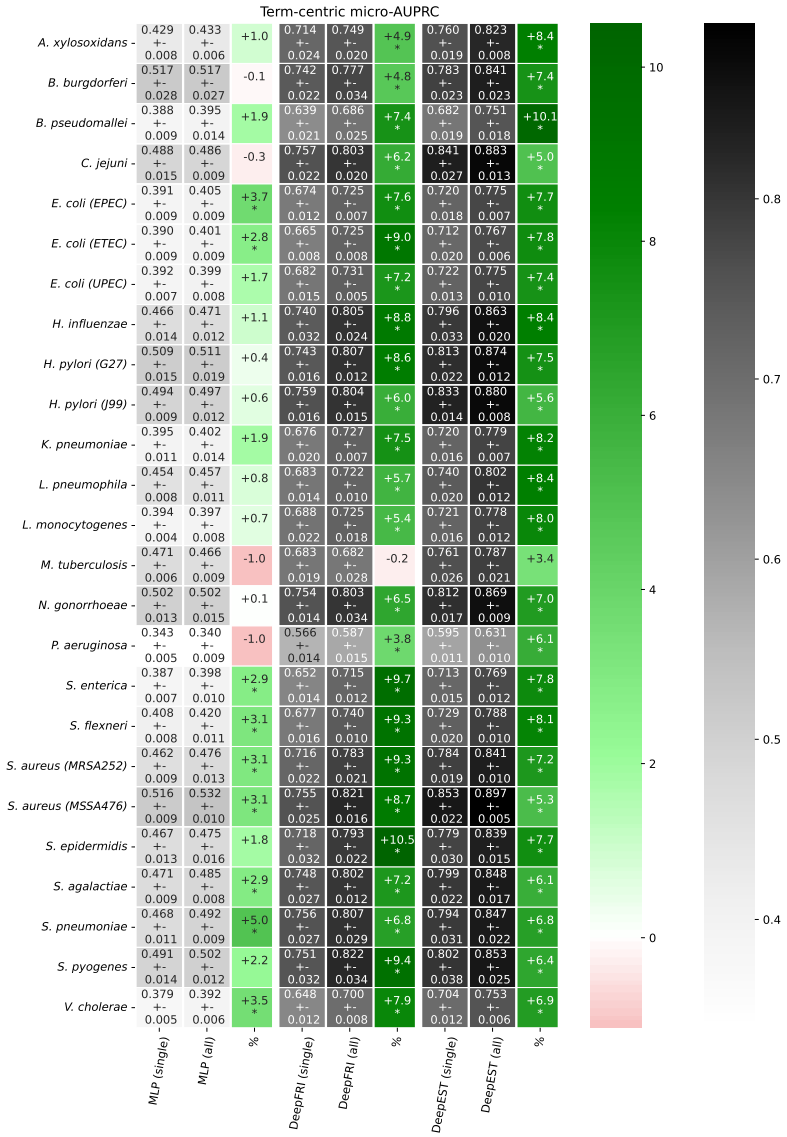

Fig. S.4: Comparison between training MLP (i.e., f_e), fine-tuned DeepFRI (i.e., f_s) and DeepEST on each species separately ("single") and on the unified dataset comprised of the 25 species ("all") in terms of average *micro*-AUPRC on the five *validation* sets, with *sequence*-based splits. Each model presents three columns, namely "single", "all" and %. The latter is the percentage change obtained when training according to the "single" strategy in contrast to the "all", with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

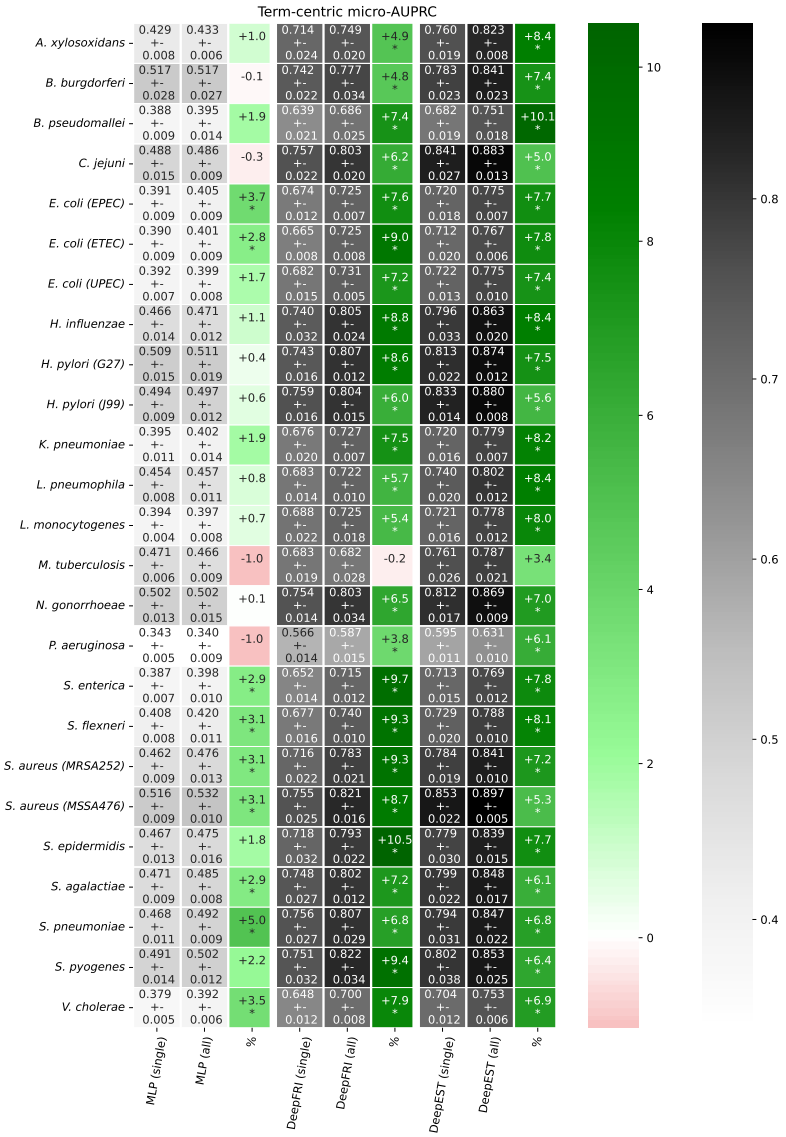

Fig. S.5: Comparison between training MLP (i.e., f_e), fine-tuned DeepFRI (i.e., f_s) and DeepEST on each species separately ("single") and on the unified dataset comprised of the 25 species ("all") in terms of average $F_{\max}$ on the five *validation* sets, with *sequence*-based splits. Each model presents three columns, namely "single", "all" and %. The latter is the percentage change obtained when training according to the "single" strategy in contrast to the "all", with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

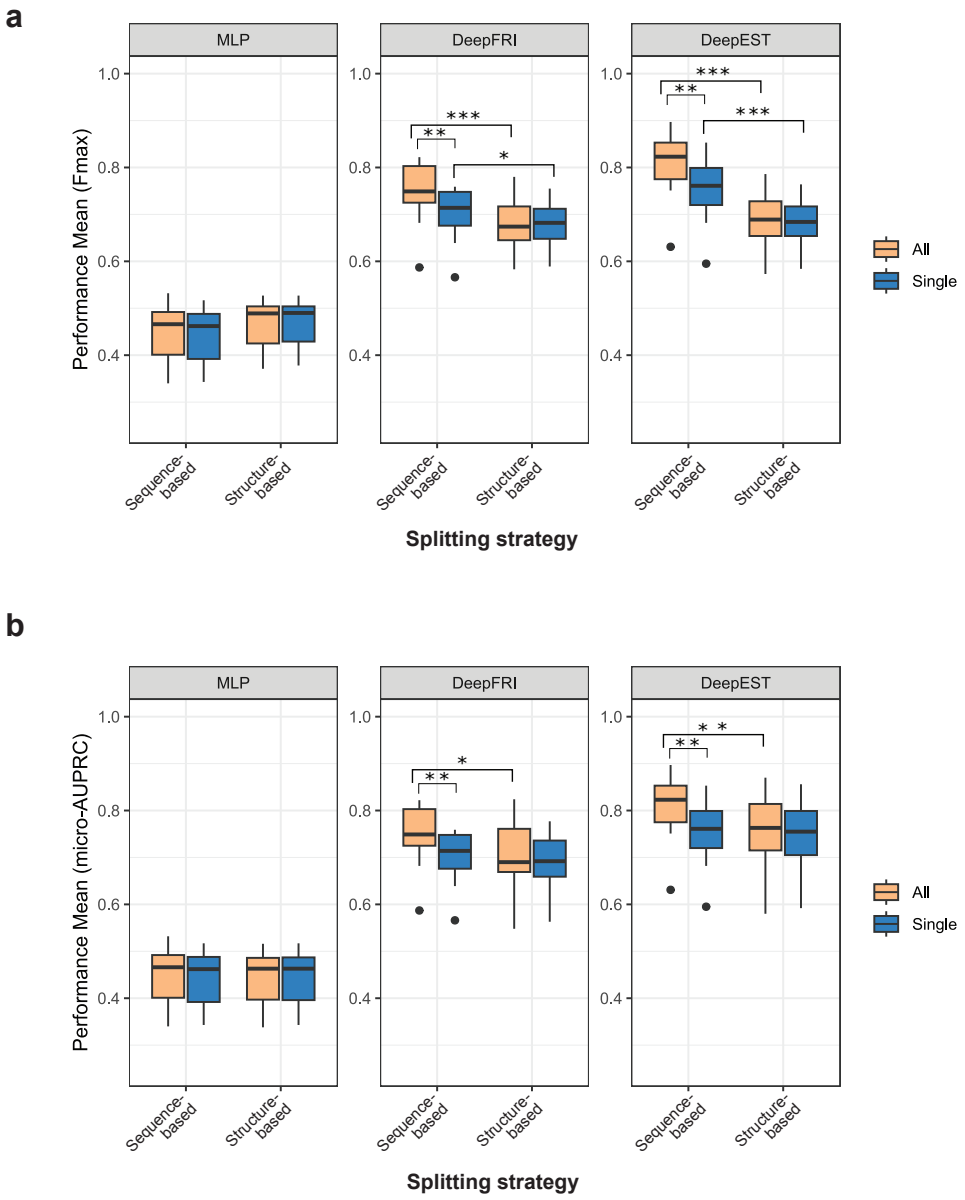

Fig. S.6: Boxplot comparing sequence (Seq) and structure (Str) accounting performance for three models: MLP, DeepFRI, and DeepEST. Performance is measured as **a** $F_{\max}$ and **b** *micro*-AUPRC. The models are displayed as separate facets in the specified order (MLP, DeepFRI, DeepEST). Each box represents the distribution of mean values under two conditions (colored bars): "All" (orange) and "Single" (blue). Error bars show the range of performance, and the y-axis is restricted to a range of 0.25 to 1 for clarity. Statistically significant results from the Wilcoxon rank-sum test (unpaired) are indicated by "*" (pvalue $<$ 0.05), "**" (pvalue $<$ 0.01) and "***" (pvalue $<$ 0.001).

(a) Term-centric micro-AUPRC

	DeepEST	DeepFRI	BLAST	Diamond	DeepGOCNN	DeepGOplus
A. xylosoxidans	0.737 +- 0.046	0.671 +- 0.038	0.084 ± 0.046	0.082 ± 0.045	0.454 ± 0.039	0.447 ± 0.019
B. burgdorferi	0.768 +- 0.035	0.698 +- 0.046	0.063 ± 0.027	0.063 ± 0.027	0.528 ± 0.040	0.521 ± 0.024
B. pseudomallei	0.691 +- 0.035	0.604 +- 0.020	0.177 ± 0.026	0.178 ± 0.025	0.431 ± 0.017	0.453 ± 0.023
C. jejuni	0.834 +- 0.030	0.740 +- 0.030	0.062 ± 0.007	0.062 ± 0.006	0.490 ± 0.029	0.480 ± 0.008
E. coli (EPEC)	0.714 +- 0.022	0.633 +- 0.018	0.132 ± 0.020	0.133 ± 0.020	0.430 ± 0.020	0.431 ± 0.017
E. coli (ETEC)	0.710 +- 0.025	0.631 +- 0.011	0.134 ± 0.018	0.134 ± 0.018	0.430 ± 0.021	0.418 ± 0.023
E. coli (UPEC)	0.714 +- 0.029	0.634 +- 0.018	0.129 ± 0.022	0.130 ± 0.021	0.431 ± 0.018	0.426 ± 0.012
H. influenzae	0.816 +- 0.029	0.741 +- 0.025	0.079 ± 0.011	0.079 ± 0.011	0.482 ± 0.026	0.461 ± 0.031
H. pylori (G27)	0.809 +- 0.014	0.755 +- 0.017	0.045 ± 0.008	0.046 ± 0.007	0.508 ± 0.034	0.497 ± 0.036
H. pylori (J99)	0.813 +- 0.017	0.741 +- 0.032	0.071 ± 0.025	0.071 ± 0.025	0.495 ± 0.027	0.499 ± 0.026
K. pneumoniae	0.717 +- 0.025	0.638 +- 0.014	0.127 ± 0.018	0.127 ± 0.018	0.431 ± 0.018	0.418 ± 0.013
L. pneumophila	0.751 +- 0.025	0.653 +- 0.030	0.100 ± 0.013	0.101 ± 0.011	0.455 ± 0.019	0.453 ± 0.026
L. monocytogenes	0.723 +- 0.027	0.645 +- 0.017	0.129 ± 0.010	0.129 ± 0.010	0.428 ± 0.028	0.431 ± 0.028
M. tuberculosis	0.757 +- 0.020	0.674 +- 0.015	0.168 ± 0.023	0.170 ± 0.022	0.476 ± 0.018	0.518 ± 0.008
N. gonorrhoeae	0.831 +- 0.019	0.752 +- 0.025	0.065 ± 0.014	0.065 ± 0.014	0.517 ± 0.023	0.514 ± 0.023
P. aeruginosa	0.594 +- 0.025	0.522 +- 0.029	0.186 ± 0.018	0.183 ± 0.015	0.373 ± 0.015	0.372 ± 0.013
S. enterica	0.697 +- 0.019	0.620 +- 0.015	0.128 ± 0.015	0.129 ± 0.015	0.421 ± 0.015	0.416 ± 0.005
S. flexneri	0.726 +- 0.019	0.636 +- 0.018	0.130 ± 0.023	0.131 ± 0.023	0.445 ± 0.018	0.436 ± 0.013
S. aureus (MRSA252)	0.782 +- 0.025	0.697 +- 0.006	0.101 ± 0.010	0.102 ± 0.010	0.479 ± 0.033	0.488 ± 0.026
S. aureus (MSSA476)	0.857 +- 0.021	0.774 +- 0.022	0.046 ± 0.010	0.046 ± 0.010	0.513 ± 0.031	0.515 ± 0.029
S. epidermidis	0.785 +- 0.029	0.724 +- 0.008	0.097 ± 0.024	0.097 ± 0.023	0.481 ± 0.030	0.486 ± 0.029
S. agalactiae	0.795 +- 0.033	0.713 +- 0.030	0.102 ± 0.013	0.103 ± 0.013	0.491 ± 0.048	0.483 ± 0.050
S. pneumoniae	0.792 +- 0.027	0.708 +- 0.017	0.100 ± 0.013	0.101 ± 0.012	0.487 ± 0.034	0.497 ± 0.041
S. pyogenes	0.785 +- 0.019	0.718 +- 0.022	0.092 ± 0.013	0.092 ± 0.015	0.503 ± 0.044	0.520 ± 0.042
V. cholerae	0.708 +- 0.041	0.626 +- 0.041	0.144 ± 0.031	0.142 ± 0.029	0.408 ± 0.022	0.400 ± 0.015

(b) Protein-centric F-max

	DeepEST	DeepFRI	BLAST	Diamond	DeepGOCNN	DeepGOplus
A. xylosoxidans	0.689 +- 0.026	0.653 +- 0.030	0.239 ± 0.053	0.239 ± 0.054	0.497 ± 0.043	0.555 ± 0.018
B. burgdorferi	0.698 +- 0.018	0.695 +- 0.026	0.157 ± 0.016	0.158 ± 0.016	0.541 ± 0.028	0.567 ± 0.019
B. pseudomallei	0.647 +- 0.018	0.595 +- 0.017	0.381 ± 0.023	0.382 ± 0.023	0.482 ± 0.021	0.583 ± 0.014
C. jejuni	0.725 +- 0.016	0.710 +- 0.019	0.167 ± 0.011	0.167 ± 0.011	0.529 ± 0.023	0.552 ± 0.019
E. coli (EPEC)	0.652 +- 0.013	0.618 +- 0.017	0.283 ± 0.023	0.284 ± 0.023	0.482 ± 0.019	0.548 ± 0.011
E. coli (ETEC)	0.652 +- 0.009	0.613 +- 0.011	0.295 ± 0.022	0.295 ± 0.021	0.480 ± 0.019	0.545 ± 0.006
E. coli (UPEC)	0.653 +- 0.011	0.616 +- 0.012	0.281 ± 0.035	0.282 ± 0.036	0.482 ± 0.018	0.546 ± 0.012
H. influenzae	0.728 +- 0.021	0.726 +- 0.018	0.189 ± 0.032	0.189 ± 0.032	0.519 ± 0.020	0.546 ± 0.014
H. pylori (G27)	0.716 +- 0.016	0.709 +- 0.025	0.153 ± 0.029	0.153 ± 0.029	0.531 ± 0.026	0.554 ± 0.020
H. pylori (J99)	0.717 +- 0.020	0.700 +- 0.030	0.188 ± 0.036	0.188 ± 0.036	0.526 ± 0.022	0.554 ± 0.021
K. pneumoniae	0.659 +- 0.015	0.623 +- 0.008	0.293 ± 0.030	0.293 ± 0.031	0.487 ± 0.025	0.553 ± 0.012
L. pneumophila	0.674 +- 0.021	0.620 +- 0.026	0.232 ± 0.025	0.233 ± 0.025	0.508 ± 0.016	0.554 ± 0.013
L. monocytogenes	0.673 +- 0.021	0.631 +- 0.017	0.296 ± 0.021	0.296 ± 0.022	0.478 ± 0.025	0.543 ± 0.012
M. tuberculosis	0.687 +- 0.010	0.642 +- 0.014	0.361 ± 0.030	0.363 ± 0.031	0.519 ± 0.019	0.586 ± 0.015
N. gonorrhoeae	0.742 +- 0.017	0.725 +- 0.016	0.149 ± 0.026	0.149 ± 0.026	0.536 ± 0.013	0.552 ± 0.014
P. aeruginosa	0.584 +- 0.015	0.541 +- 0.022	0.371 ± 0.018	0.371 ± 0.017	0.424 ± 0.012	0.536 ± 0.006
S. enterica	0.637 +- 0.008	0.599 +- 0.013	0.286 ± 0.020	0.287 ± 0.020	0.473 ± 0.015	0.534 ± 0.005
S. flexneri	0.667 +- 0.003	0.620 +- 0.017	0.287 ± 0.028	0.288 ± 0.029	0.491 ± 0.019	0.553 ± 0.012
S. aureus (MRSA252)	0.720 +- 0.012	0.687 +- 0.017	0.241 ± 0.021	0.241 ± 0.021	0.533 ± 0.027	0.569 ± 0.018
S. aureus (MSSA476)	0.780 +- 0.021	0.771 +- 0.014	0.109 ± 0.035	0.109 ± 0.035	0.544 ± 0.023	0.553 ± 0.023
S. epidermidis	0.724 +- 0.021	0.708 +- 0.015	0.241 ± 0.035	0.242 ± 0.035	0.531 ± 0.020	0.571 ± 0.014
S. agalactiae	0.722 +- 0.026	0.696 +- 0.026	0.261 ± 0.020	0.261 ± 0.020	0.529 ± 0.037	0.580 ± 0.020
S. pneumoniae	0.727 +- 0.016	0.698 +- 0.014	0.248 ± 0.039	0.248 ± 0.039	0.529 ± 0.028	0.575 ± 0.017
S. pyogenes	0.716 +- 0.025	0.703 +- 0.025	0.216 ± 0.032	0.216 ± 0.033	0.536 ± 0.029	0.571 ± 0.025
V. cholerae	0.641 +- 0.014	0.611 +- 0.024	0.334 ± 0.051	0.335 ± 0.051	0.448 ± 0.019	0.541 ± 0.017

Fig. S.7: We report the results, as average ± standard deviation across the five test sets, for DeepEST and the comparison partners (DeepFRI, BLAST, Diamond, DeepGOCNN, DeepGOplus): (a) *micro*-AUPRC (b) $F_{\max}$.

(a) Term-centric macro-AUPRC

	DeepEST	DeepFRI	BLAST	Diamond	DeepGOCNN	DeepGOplus
A. xylosoxidans	0.650 +- 0.052	0.614 +- 0.034	0.098 ± 0.028	0.097 ± 0.027	0.210 ± 0.014	0.295 ± 0.008
B. burgdorferi	0.673 +- 0.052	0.632 +- 0.041	0.063 ± 0.027	0.063 ± 0.027	0.231 ± 0.026	0.296 ± 0.012
B. pseudomallei	0.575 +- 0.019	0.550 +- 0.008	0.145 ± 0.024	0.145 ± 0.023	0.218 ± 0.005	0.317 ± 0.033
C. jejuni	0.744 +- 0.049	0.653 +- 0.039	0.070 ± 0.016	0.070 ± 0.015	0.244 ± 0.007	0.284 ± 0.014
E. coli (EPEC)	0.598 +- 0.015	0.563 +- 0.019	0.118 ± 0.014	0.119 ± 0.015	0.223 ± 0.007	0.275 ± 0.018
E. coli (ETEC)	0.591 +- 0.018	0.559 +- 0.021	0.102 ± 0.017	0.103 ± 0.018	0.225 ± 0.009	0.271 ± 0.025
E. coli (UPEC)	0.600 +- 0.016	0.562 +- 0.016	0.103 ± 0.016	0.104 ± 0.016	0.228 ± 0.008	0.268 ± 0.021
H. influenzae	0.738 +- 0.048	0.664 +- 0.018	0.068 ± 0.012	0.069 ± 0.013	0.214 ± 0.010	0.263 ± 0.020
H. pylori (G27)	0.726 +- 0.014	0.649 +- 0.024	0.050 ± 0.005	0.050 ± 0.005	0.224 ± 0.008	0.269 ± 0.009
H. pylori (J99)	0.714 +- 0.010	0.647 +- 0.021	0.089 ± 0.014	0.089 ± 0.014	0.228 ± 0.006	0.306 ± 0.007
K. pneumoniae	0.606 +- 0.022	0.565 +- 0.022	0.100 ± 0.012	0.099 ± 0.011	0.228 ± 0.009	0.283 ± 0.019
L. pneumophila	0.630 +- 0.022	0.587 +- 0.019	0.093 ± 0.015	0.094 ± 0.014	0.214 ± 0.010	0.279 ± 0.015
L. monocytogenes	0.612 +- 0.027	0.567 +- 0.024	0.100 ± 0.009	0.101 ± 0.009	0.236 ± 0.010	0.272 ± 0.012
M. tuberculosis	0.622 +- 0.028	0.602 +- 0.008	0.173 ± 0.007	0.173 ± 0.007	0.239 ± 0.010	0.347 ± 0.015
N. gonorrhoeae	0.727 +- 0.035	0.647 +- 0.022	0.077 ± 0.031	0.077 ± 0.031	0.235 ± 0.008	0.297 ± 0.033
P. aeruginosa	0.390 +- 0.021	0.492 +- 0.017	0.140 ± 0.025	0.136 ± 0.025	0.223 ± 0.004	0.238 ± 0.034
S. enterica	0.572 +- 0.017	0.552 +- 0.020	0.102 ± 0.011	0.102 ± 0.010	0.217 ± 0.004	0.261 ± 0.026
S. flexneri	0.614 +- 0.016	0.575 +- 0.016	0.099 ± 0.015	0.099 ± 0.015	0.221 ± 0.009	0.261 ± 0.023
S. aureus (MRSA252)	0.711 +- 0.037	0.615 +- 0.012	0.094 ± 0.008	0.094 ± 0.009	0.223 ± 0.010	0.278 ± 0.016
S. aureus (MSSA476)	0.834 +- 0.017	0.724 +- 0.017	0.044 ± 0.029	0.044 ± 0.029	0.253 ± 0.022	0.275 ± 0.036
S. epidermidis	0.699 +- 0.043	0.642 +- 0.027	0.088 ± 0.025	0.089 ± 0.025	0.226 ± 0.016	0.275 ± 0.027
S. agalactiae	0.737 +- 0.008	0.636 +- 0.014	0.105 ± 0.028	0.105 ± 0.028	0.230 ± 0.020	0.306 ± 0.039
S. pneumoniae	0.728 +- 0.027	0.629 +- 0.020	0.100 ± 0.023	0.101 ± 0.023	0.228 ± 0.018	0.311 ± 0.035
S. pyogenes	0.729 +- 0.019	0.644 +- 0.015	0.096 ± 0.022	0.096 ± 0.022	0.220 ± 0.017	0.304 ± 0.021
V. cholerae	0.572 +- 0.032	0.555 +- 0.036	0.123 ± 0.015	0.118 ± 0.017	0.214 ± 0.009	0.274 ± 0.012

(b) Term-centric s-min

	DeepEST	DeepFRI	BLAST	Diamond	DeepGOCNN	DeepGOplus
A. xylosoxidans	21.692 +- 1.365	26.293 +- 1.549	42.344 ± 3.154	42.331 ± 3.161	40.065 ± 3.057	37.861 ± 2.497
B. burgdorferi	23.426 +- 1.629	27.703 +- 1.702	50.653 ± 3.432	50.651 ± 3.433	42.075 ± 0.682	40.593 ± 1.795
B. pseudomallei	25.827 +- 1.773	30.973 +- 0.900	39.283 ± 1.264	39.209 ± 1.243	41.783 ± 1.294	37.655 ± 1.535
C. jejuni	19.605 +- 1.354	26.707 +- 2.863	50.903 ± 3.788	50.889 ± 3.770	45.214 ± 3.397	44.064 ± 2.999
E. coli (EPEC)	28.819 +- 1.404	34.543 +- 1.063	48.708 ± 0.193	48.562 ± 0.237	48.407 ± 1.418	45.188 ± 0.208
E. coli (ETEC)	28.804 +- 1.351	34.327 +- 0.824	47.950 ± 0.363	47.808 ± 0.506	47.725 ± 1.388	45.494 ± 0.366
E. coli (UPEC)	28.261 +- 1.416	33.874 +- 1.132	47.953 ± 1.796	47.756 ± 1.902	47.421 ± 1.359	44.866 ± 1.355
H. influenzae	20.204 +- 2.306	27.204 +- 1.635	52.661 ± 1.831	52.531 ± 1.793	47.599 ± 2.302	46.861 ± 1.688
H. pylori (G27)	20.794 +- 1.729	24.185 +- 2.487	50.326 ± 3.773	50.315 ± 3.780	42.466 ± 4.062	41.616 ± 3.783
H. pylori (J99)	21.975 +- 2.383	26.769 +- 3.382	52.010 ± 4.090	52.010 ± 4.090	46.147 ± 3.818	44.357 ± 3.590
K. pneumoniae	27.160 +- 0.999	32.491 +- 0.790	46.584 ± 1.141	46.428 ± 1.204	46.024 ± 1.793	43.861 ± 1.083
L. pneumophila	25.142 +- 1.432	31.345 +- 1.551	48.153 ± 1.155	48.106 ± 1.157	46.266 ± 1.289	44.109 ± 1.092
L. monocytogenes	30.442 +- 2.286	37.887 +- 2.337	54.645 ± 1.838	54.575 ± 1.853	53.490 ± 1.495	50.556 ± 1.577
M. tuberculosis	24.720 +- 1.841	30.688 +- 1.979	43.656 ± 2.950	43.574 ± 2.963	44.462 ± 2.962	39.879 ± 2.493
N. gonorrhoeae	20.640 +- 0.956	27.233 +- 2.400	53.192 ± 2.601	53.187 ± 2.604	46.655 ± 2.361	45.215 ± 2.165
P. aeruginosa	47.686 +- 1.903	52.113 +- 1.034	63.115 ± 2.672	62.121 ± 2.397	63.178 ± 1.530	61.948 ± 2.435
S. enterica	29.466 +- 1.019	34.629 +- 0.592	48.435 ± 1.030	48.229 ± 1.082	47.933 ± 1.226	45.848 ± 0.776
S. flexneri	26.923 +- 0.837	33.138 +- 0.466	47.636 ± 1.888	47.442 ± 1.876	46.786 ± 1.153	44.482 ± 1.537
S. aureus (MRSA252)	24.270 +- 2.269	30.789 +- 1.457	52.281 ± 2.629	52.276 ± 2.626	48.104 ± 2.769	46.506 ± 2.512
S. aureus (MSSA476)	16.663 +- 1.749	23.140 +- 0.622	55.857 ± 3.228	55.857 ± 3.228	47.830 ± 2.333	47.135 ± 2.599
S. epidermidis	24.191 +- 1.884	29.845 +- 0.997	54.312 ± 3.916	54.308 ± 3.915	49.579 ± 2.168	47.504 ± 3.016
S. agalactiae	22.020 +- 2.184	28.917 +- 2.383	50.727 ± 0.864	50.697 ± 0.888	46.692 ± 2.975	44.238 ± 1.523
S. pneumoniae	23.251 +- 1.846	30.708 +- 1.343	52.740 ± 2.807	52.731 ± 2.770	48.443 ± 3.456	45.578 ± 2.930
S. pyogenes	22.841 +- 1.811	28.116 +- 1.838	50.817 ± 2.680	50.795 ± 2.686	45.787 ± 2.798	43.438 ± 2.810
V. cholerae	34.490 +- 1.869	41.176 +- 1.169	57.388 ± 2.959	57.235 ± 3.057	57.288 ± 0.954	54.448 ± 2.544

Fig. S.8: We report the results, as average ± standard deviation across the five test sets, for DeepEST and the comparison partners (DeepFRI, BLAST, Diamond, DeepGOCNN, DeepGOplus): (a)*macro*-AUPRC (b) $s_{\min}$. Note that for $s_{\min}$, lower values represent a better performance in comparison to higher values.

(a) Protein-centric AUPRC

	DeepEST	DeepFRI	BLAST	Diamond	DeepGOCNN	DeepGOplus
A. xylosoxidans	0.759 +- 0.029	0.733 +- 0.025	0.143 ± 0.040	0.143 ± 0.040	0.520 ± 0.038	0.574 ± 0.025
B. burgdorferi	0.773 +- 0.026	0.765 +- 0.023	0.087 ± 0.010	0.087 ± 0.010	0.573 ± 0.031	0.597 ± 0.022
B. pseudomallei	0.713 +- 0.018	0.677 +- 0.016	0.252 ± 0.017	0.253 ± 0.017	0.493 ± 0.017	0.597 ± 0.012
C. jejuni	0.810 +- 0.020	0.778 +- 0.016	0.094 ± 0.008	0.094 ± 0.008	0.560 ± 0.024	0.584 ± 0.022
E. coli (EPEC)	0.719 +- 0.014	0.692 +- 0.014	0.174 ± 0.017	0.174 ± 0.017	0.488 ± 0.018	0.554 ± 0.011
E. coli (ETEC)	0.717 +- 0.013	0.689 +- 0.008	0.183 ± 0.017	0.184 ± 0.017	0.488 ± 0.019	0.555 ± 0.007
E. coli (UPEC)	0.720 +- 0.017	0.692 +- 0.013	0.173 ± 0.027	0.174 ± 0.027	0.492 ± 0.018	0.556 ± 0.011
H. influenzae	0.803 +- 0.022	0.785 +- 0.016	0.110 ± 0.023	0.111 ± 0.024	0.542 ± 0.024	0.570 ± 0.016
H. pylori (G27)	0.797 +- 0.012	0.783 +- 0.021	0.084 ± 0.018	0.084 ± 0.018	0.564 ± 0.024	0.589 ± 0.021
H. pylori (J99)	0.797 +- 0.014	0.770 +- 0.026	0.106 ± 0.023	0.106 ± 0.023	0.555 ± 0.022	0.588 ± 0.024
K. pneumoniae	0.729 +- 0.016	0.701 +- 0.010	0.181 ± 0.023	0.182 ± 0.023	0.496 ± 0.021	0.561 ± 0.005
L. pneumophila	0.744 +- 0.028	0.703 +- 0.019	0.136 ± 0.017	0.136 ± 0.017	0.528 ± 0.014	0.573 ± 0.014
L. monocytogenes	0.738 +- 0.021	0.702 +- 0.017	0.186 ± 0.018	0.188 ± 0.018	0.490 ± 0.027	0.551 ± 0.015
M. tuberculosis	0.754 +- 0.011	0.722 +- 0.015	0.231 ± 0.024	0.232 ± 0.024	0.541 ± 0.022	0.614 ± 0.022
N. gonorrhoeae	0.818 +- 0.014	0.786 +- 0.014	0.083 ± 0.016	0.083 ± 0.016	0.566 ± 0.013	0.584 ± 0.015
P. aeruginosa	0.631 +- 0.021	0.614 +- 0.016	0.268 ± 0.018	0.272 ± 0.017	0.425 ± 0.011	0.528 ± 0.005
S. enterica	0.702 +- 0.013	0.677 +- 0.011	0.177 ± 0.015	0.178 ± 0.015	0.480 ± 0.014	0.540 ± 0.003
S. flexneri	0.729 +- 0.007	0.694 +- 0.014	0.177 ± 0.022	0.177 ± 0.022	0.501 ± 0.017	0.563 ± 0.010
S. aureus (MRSA252)	0.788 +- 0.015	0.756 +- 0.012	0.141 ± 0.014	0.141 ± 0.013	0.554 ± 0.030	0.587 ± 0.022
S. aureus (MSSA476)	0.853 +- 0.015	0.822 +- 0.011	0.059 ± 0.021	0.059 ± 0.021	0.586 ± 0.024	0.597 ± 0.025
S. epidermidis	0.795 +- 0.020	0.774 +- 0.012	0.142 ± 0.024	0.142 ± 0.024	0.554 ± 0.024	0.595 ± 0.016
S. agalactiae	0.797 +- 0.027	0.765 +- 0.024	0.155 ± 0.015	0.156 ± 0.015	0.555 ± 0.044	0.609 ± 0.027
S. pneumoniae	0.802 +- 0.018	0.766 +- 0.011	0.146 ± 0.026	0.147 ± 0.025	0.552 ± 0.033	0.600 ± 0.025
S. pyogenes	0.789 +- 0.023	0.772 +- 0.020	0.124 ± 0.021	0.124 ± 0.021	0.565 ± 0.034	0.602 ± 0.033
V. cholerae	0.708 +- 0.016	0.683 +- 0.018	0.223 ± 0.043	0.225 ± 0.043	0.456 ± 0.020	0.547 ± 0.014

Fig. S.9: We report the results, as average ± standard deviation across the five test sets, for DeepEST and the comparison partners (DeepFRI, BLAST, Diamond, DeepGOCNN, DeepGOplus): (a) Protein-centric AUPRC.

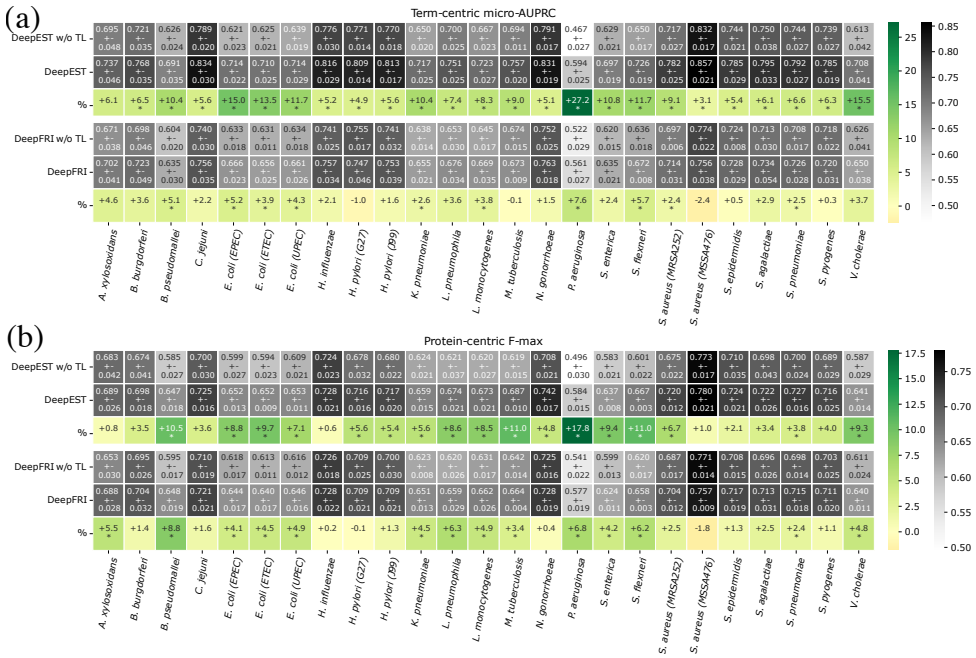

Fig. S.10: Comparison between not using (w/o TL) and using our transfer learning technique on both DeepEST and DeepFRI. We report the results as average ± standard deviation across the five test sets in terms of (a) *micro*-AUPRC and (b) $F_{\max}$. We additionally report the percentage change of performance when using transfer learning in comparison to not using it. These percentages are labeled with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

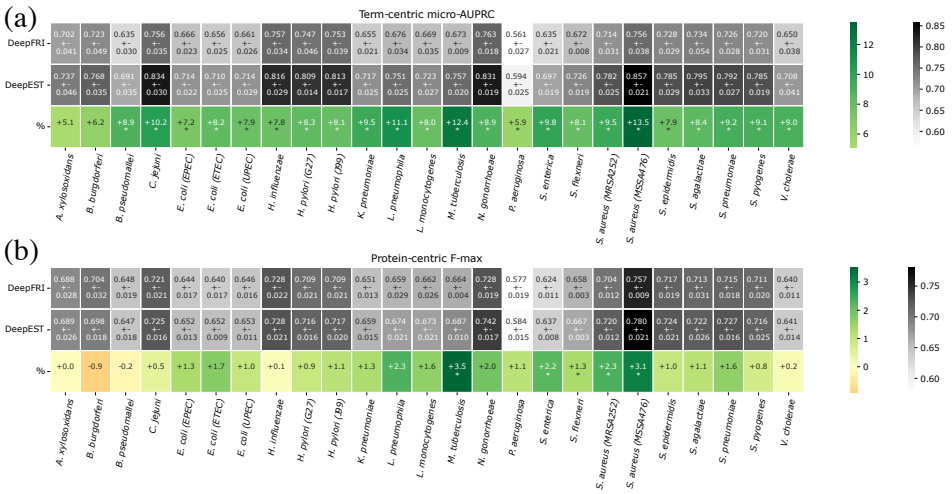

Fig. S.11: Comparison between f_s (i.e., DeepFRI with the use of transfer learning) and DeepEST in terms of (a) *micro*-AUPRC and (b) $F_{\max}$−score. The third line, labeled with "%", represents the percentage change in performance of DeepEST in comparison to DeepFRI. This percentage is marked with an asterisk when this change exceeds one standard deviation from the average performance of DeepFRI. These are the results on the test sets.

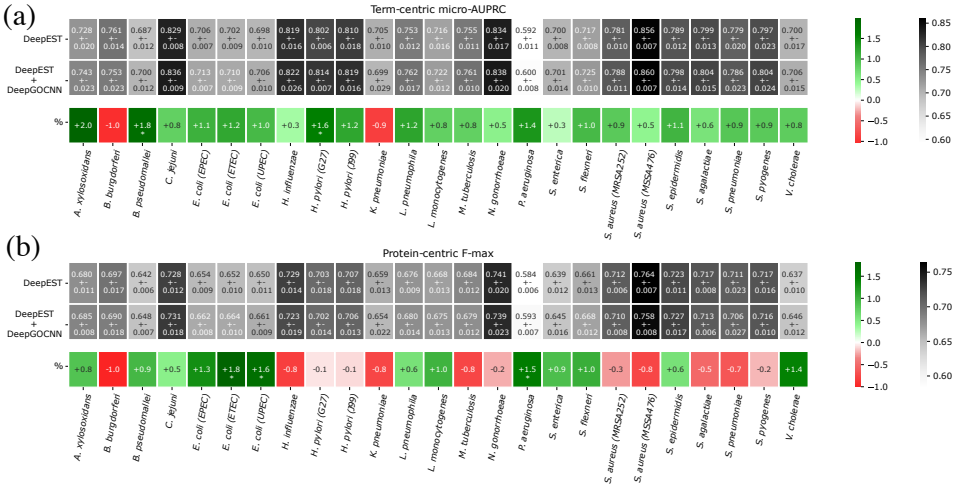

Fig. S.12: Comparison between DeepEST and DeepEST combined with DeepGOCNN (to include the linear sequence information) in terms of average (a) *micro*-AUPRC and (b) $F_{\max}$ on the five *validation* sets, with *structure*-based splits. % represents the percentage change in performance when using DeepEST + DeepGOCNN in comparison to DeepEST, with an "*" when the difference between the two averages exceeds one standard deviation. Each column represents a species.

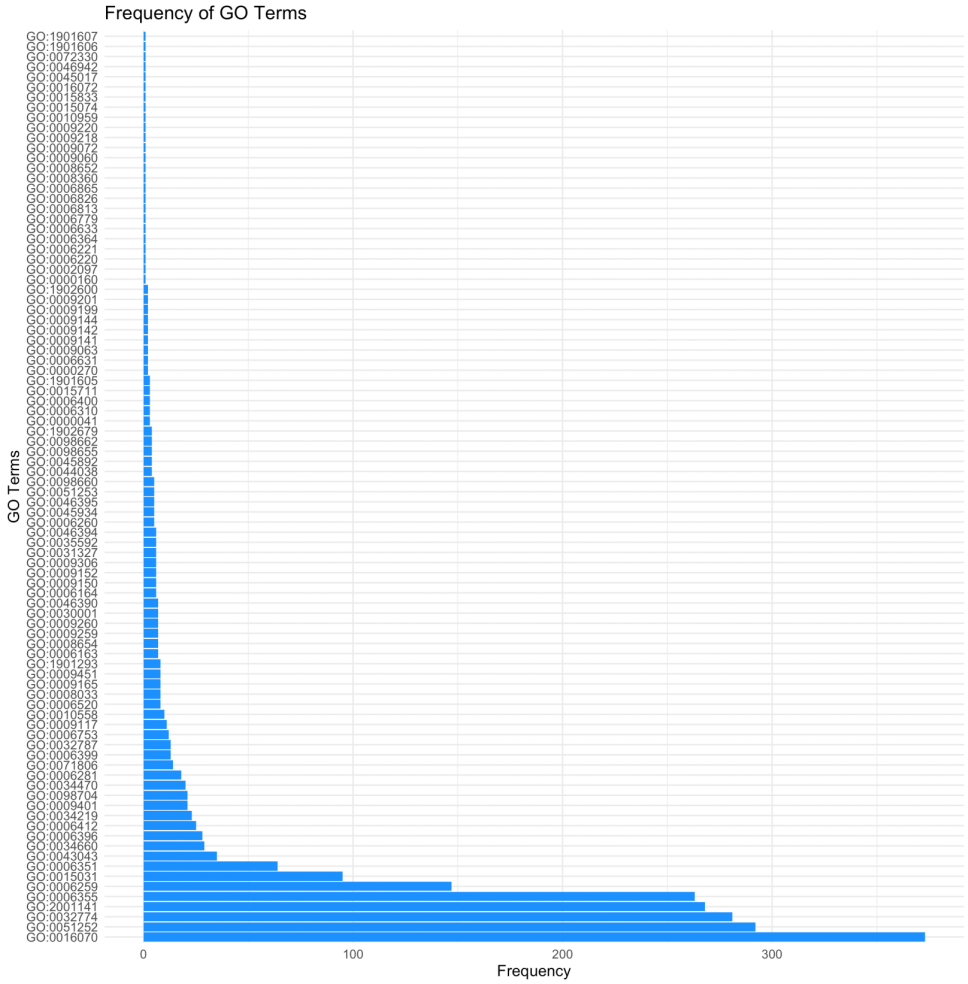

Fig. S.13: Histogram showcasing connections between GO terms and their corresponding target genes at depths greater than 5.

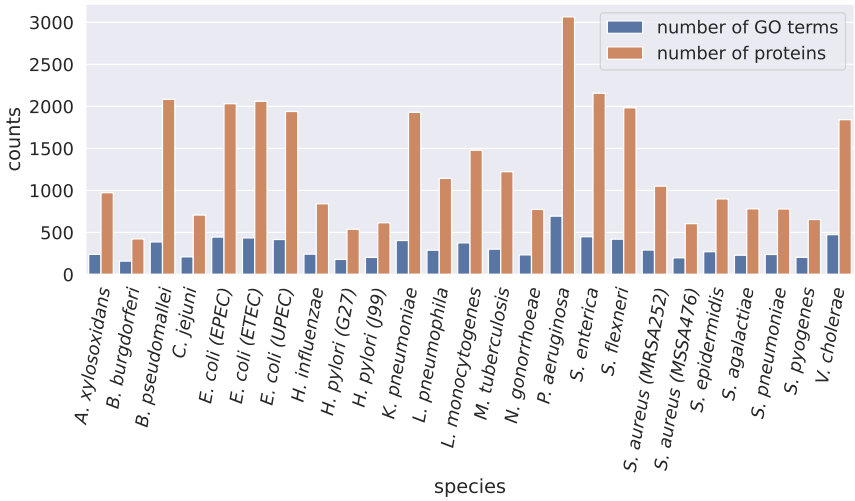

Fig. S.14: Number of annotated proteins and number of GO terms for each species obtained after having filtered out: (i) GO terms annotated for less than 10 proteins and (ii) proteins presenting no annotations.

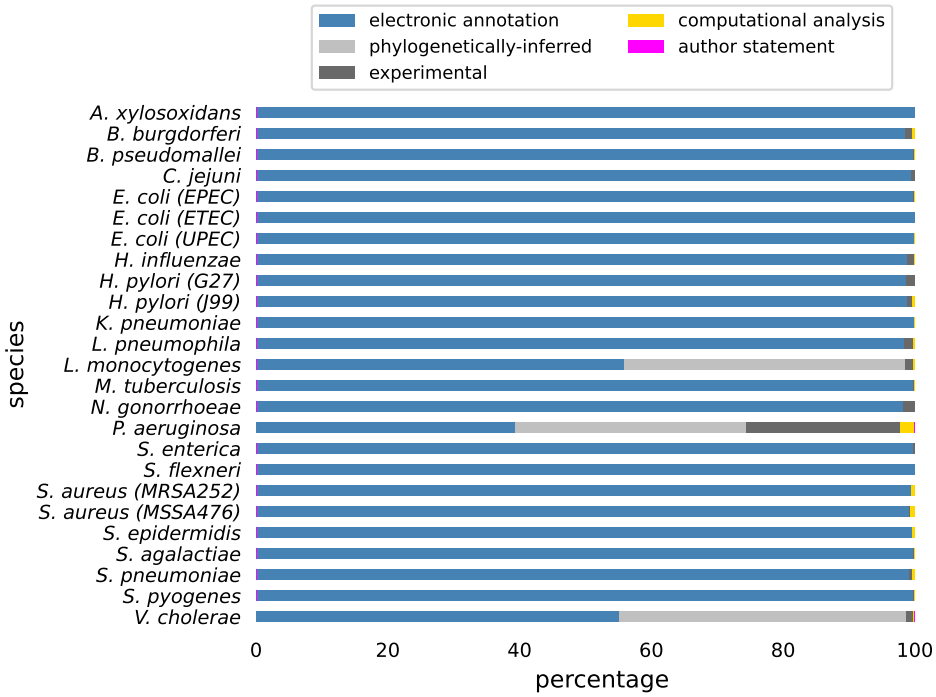

Fig. S.15: Distribution of the GO terms annotation categories for the selection of 25 species. Each annotation category maps to specific evidence codes, namely: (i) electronic annotation to IEA, (ii) phylogenetically inferred to IBA, (iii) experimental to EXP, IDA, IEP, IGI, IMP and IPI, (iv) computational analysis to IGC and ISS, and (v) author statement to NAS and TAS.