Limitations of Current Machine-Learning Models in Predicting Enzymatic Functions for Uncharacterized Proteins

Valérie de Crécy-Lagard[a,b*], Raquel Dias[a], Nick Sexson[a], Iddo Friedberg[c], Yifeng Yuan[a] and Manal A. Swairjo[d,e]

[a]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611
[b]Genetic Institute, University of Florida, Gainesville, FL 32611
[c]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA 50011
[d]Department of Chemistry and Biochemistry, San Diego State University, San Diego, CA 92182
[e]The Viral Information Institute, San Diego State University, San Diego, CA 92182

*Corresponding author: vcrecy@ufl.edu

Abstract

Thirty to seventy percent of proteins in any given genome have no assigned function and have been labeled as the protein "unknome". This large knowledge shortfall is one of the final frontiers of biology. Machine-Learning (ML) approaches are enticing, with early successes demonstrating the ability to propagate functional knowledge from experimentally characterized proteins. An open question is the ability of machine-learning approaches to predict enzymatic functions unseen in the training sets. Using a set of *Escherichia coli* unknowns, we evaluated the current state-of-the-art machine-learning approaches and found that these methods currently lack the ability to integrate scientific reasoning into their prediction algorithms. While human annotators can leverage the plethora of genomic data in making plausible predictions into the unknown, current ML methods not only fail to make novel predictions but also make basic logic errors in their predictions. This underscores the need to include assessments of prediction uncertainty in model output and to test for 'hallucinations' (logic failures) as a part of model evaluation. Explainable AI (XAI) analysis can be used to identify indicators of prediction errors, potentially identifying the most relevant data to include in the next generation of computational models.

Keywords

Paralogs, Orthologs, YciO, Protein Language Models, Enzyme Commission, Protein Family, Functional Annotation

Article Summary

Many proteins in any genome, ranging from 30% to 70% of the genome, lack an assigned function. This knowledge gap limits the full use of the vast available genomic data. Machine learning has shown promise in transferring functional knowledge within isofunctional families, but it largely fails to predict novel functions not seen in its training data. Understanding these failures can guide the development of better machine-learning methods to help experts make accurate functional predictions for uncharacterized proteins.

Introduction

Determining protein function is not an easy task, and thirty years after the first bacterial genome was sequenced, the functional annotation status of the proteome of most species is far from being accurate or complete, even for model organisms (Ghatak et al., 2019; Wood et al., 2019; de Crécy-lagard et al., 2022; Rocha et al., 2023). Experimental validation of protein function is a painstaking process, and with the explosion of whole genome sequences (Kyrpides, 1999; Peterson et al., 2025; Torres et al., 2025), the gap between experimentally validated functions and those predicted through computational methods continues to widen. In UniprotKB (Bateman et al., 2023), the most widely used protein function database (Ramola et al., 2022), the estimates are that less than 0.5% to 15% of proteins have been linked to experimental data(Škunca et al., 2017).

The process of functional annotation of protein entries in databases starts with capturing information in the literature by biocurators(International Society for Biocuration, 2018). This process links experimental characterizations of specific proteins in specific organisms to controlled vocabularies that describe validated functions, such as the Gene Ontology (GO)(Gene Ontology Consortium et al., 2023), the IUPAC Enzyme Commission (EC) numbers (https://iubmb.qmul.ac.uk/enzyme/), or biochemical reaction descriptors [e.g. Rhea(Bansal et al., 2022)]. Text-mining tools have accelerated the flow of information captured(Soldatos et al., 2015; Poux et al., 2017; Wei et al., 2019). However, this step remains a major bottleneck in the annotation workflow, which can result in mislabeling proteins as “unknown” when a function has been reported in the literature (see type 1 error in Table 1 and Fig. 1).

Using the principle of sequence similarity, also known as homology transfer, putative functions are assigned to proteins in newly sequenced genomes as a part of the genome annotation process (Seemann, 2014; Thibaud-Nissen et al., 2016; Olson et al., 2023). Indeed, most functions of proteins in UniProt have been inferred computationally based on sequence similarity (Bateman et al., 2023). The methods used over the past 30 years to automatically propagate functional annotations have been extensively reviewed (Storm and Sonnhammer, 2002; Friedberg, 2006; Lee et al., 2007; Satish Kumar et al., 2007; Devoid et al., 2013). This seemingly simple task can become quite difficult (Challenge 1 in Fig. 1), since relying on sequence similarity alone can result in significant annotation errors (Schnoes et al., 2009; Rembeza and Engqvist, 2021). One prevalent error type affecting protein families (Rembeza and Engqvist, 2021) is in the misannotation of paralogs (Schnoes et al., 2009; Zallot et al., 2016; Rembeza and Engqvist, 2021) caused by the inherent functional complexity of enzyme families (Glasner et al., 2006; Gerlt et al., 2012) (see type 6 error in Table 1 and Fig. 1). Functional diversification through duplication and

divergence (Todd et al., 2001; Das et al., 2015; Bordin et al., 2021) results in proteins with high degrees of similarity having different functional roles, creating non-isofunctional paralogous groups. Indeed, even within a single protein family, a difference in a few amino acids can affect substrate binding and/or catalysis, effectively changing the function (Roy et al., 2009; Ribeiro et al., 2020; Precord et al., 2023). Most current genome annotation pipelines annotate paralogs without considering the potential for functional divergence, leading to incorrect annotations for up to 80% of family members, with these types of errors increasing over time (Schnoes et al., 2009; Rembeza and Engqvist, 2021).

Methods for inferring function that incorporate additional information such as phylogeny, active site analyses, metabolic reconstruction, and sequence similarity networks (SSNs) combined with gene neighborhood or co-expression information can better separate non-isofunctional subfamilies (Zallot et al., 2016, 2021; Ribeiro et al., 2023) (Fig. 2). When Hidden Markov Models (HMMs) or signature motifs are generated to annotate the subfamilies, they can be integrated into annotation pipelines such as RefSeq(Li et al., 2021) or incorporated into rules such as the as the Unified Rule (UniRule) used by UniprotKB (MacDougall et al., 2020) additional curation, which has not kept pace with the exponential increase in sequenced genomes. In addition to mis-assignments among paralogs increasing with evolutionary distance, distinguishing between sub- and neo-functionalization can be problematic (Birchler, 2025).

A significant challenge is to predict the functions of the proteins that have not previously been functionally characterized (Fig. 1A). Estimates suggest that 30% to 70% of proteins in any given genome are in this “unknown” category (Ghatak et al., 2019; Wood et al., 2019; Lobb et al., 2020; Sajid et al., 2023). Functionally characterizing these unknown families is challenging because of the openness of the prediction space (Hanson et al., 2009; Niehaus et al., 2015). Some of these “unknowns” are likely to fill functional roles not yet linked to a specific gene (Karp, 2004; Lespinet and Labedan, 2006; Chen and Vitkup, 2007; Niehaus et al., 2015). Others are likely to be non-orthologous replacements (or convergent evolution), where different (structurally diverse) protein families overlap in functional roles (Ribeiro et al., 2023). Finally, we expect some novel biological functions and predicting these require extrapolation beyond the current knowledge base. As we sequence more organisms from more previously unexplored ecological niches, we expect to encounter proteins with novel functions.

In summary, the annotation of a new genome faces two challenges: 1) to identify genes whose likely function can be inferred by propagating known annotations; 2) to identify genes with potentially novel functions and make predictions. This is a thorny path with many different types of errors that can be made (Fig. 1 B and Table 1). Computational approaches, however

innovative, cannot prove a particular protein function; experimental validation is still necessary. Proof of function is not necessarily straightforward, as demonstrating a protein capable of participating in a chemical reaction is not proof that the protein has the responsibility of that function in the cellular context where it is identified (García-Contreras et al., 2012; Copley, 2015). For proteins with available experimental evidence of function in one context, the computational propagation of the assigned functional role is the logical next step (Lee et al., 2007), although as described above, similarity does not guarantee functional conservation. For unknown proteins, the increasing amounts of genome-wide/transcriptome-wide/metabolome-wide experimental data can be used in combination with structural knowledge to develop predictions that can be tested experimentally (Hanson et al., 2009). Scientists are increasingly successful in these types of endeavors in a range of systems (Dembech et al., 2023; Rodríguez del Río et al., 2023), and there is tremendous hope that AI methods will accelerate the pace at which predictions can be made.

For challenge 1, there has been a recent explosion of publications reporting the use of pre-trained Protein Language Models (PLM) to predict protein functions (Ardern et al., 2023; Ayres et al., 2023; Derry and Altman, 2023; Durairaj et al., 2023; Kim et al., 2023; Prabakaran and Bromberg, 2023; Rodríguez del Río et al., 2023; Hwang et al., 2024) (Fig. 3). Models have been developed to link protein sequences directly to vocabularies such as GO terms (Radivojac et al., 2013; Jiang et al., 2016; Zhou et al., 2019) or EC (Enzyme Commission Classification) numbers (Hamamsy et al., 2023). EC numbers are a set of 4 numbers that are hierarchical with the first number the most general classification and the last number the most specific, and whose purpose is to standardize the description of enzymatic activities. PLMs have been reported to accurately predict the first two digits of the EC number but not the last two (Sanderson et al., 2023). GloEC (Huang et al., 2024) and MAPred (Rong et al., 2024) are two recent PLM-based EC prediction tools (Fig. 3) that achieve precision and recall metrics combined into macro-F_1 or regular F_1 scores ranging from 10% to 80%, depending on the test set [see Tables 2-4 of (Huang et al., 2024) and Table 1 of (Rong et al., 2024) respectively, summarized in Fig 4]. The F1 score is the harmonic mean of precision and recall. Recall is the proportion of true positives correctly identified out of all actual positives, while precision is the proportion of true positives among all predicted positives. A "positive" refers to an instance that actually belongs to the class of interest (here, the correct EC number). The macro-F1 score is calculated by computing the F1 score for each EC number category independently and then averaging these scores across EC numbers. When training data contains well-characterized proteins with well-described structures (e.g.,

cofactor-237 set) or well-separated functional groups (e.g., phosphorylase set), F_1 scores range from 60% to 75% for the best methods (Fig. 4A). Once the structure of the training data is more complex, and the algorithm must distinguish among similar enzymes with different functions, such as the carbohydrate dataset comprised of ~350 glycosyl hydrolases, F_1 scores fall sharply, ranging from 6% to 30% (Fig. 4A and B). When using the Price test dataset, which comprises 100 enzymes that were annotated using high-throughput phenotypic data and are not included in any training sets(Price and Arkin, 2024), the best performance was a 50% F_1 score with the MAPred pipeline (Fig. 4B). As recently discussed by the Bromberg group (Prabakaran and Bromberg, 2023)PLM-based methods are not designed to discover novelty (Challenge 2), as they rely on training data derived from existing knowledge.

Several independent studies agree that the number of proteins in the model gram-negative *Escherichia coli* K12 not linked to precise molecular and/or biological functions is between 1,200 and 1,400, or a quarter of the encoded proteins (Ghatak et al., 2019). Some of these proteins are likely responsible for functions that have been described in *E. coli* but never linked to a gene [called orphan enzymes (Karp, 2004; Danchin et al., 2018; Kobras et al., 2021)] or functions that might have been described in another organism but could also be present in *E. coli.* What percentage of these currently unannotated proteins represent novel functions is difficult to determine. A recent study labeled over 450 of the unknowns in the model bacteria *Escherichia coli* with substrate-level specificity ECs (all 4 numbers) using a supervised ML-based DeepECTransformer (DeepECTF) platform and validated three of these predictions *in vitro* (Kim et al., 2023). If corroborated, this report could be a real breakthrough in predictive modeling by linking true unknowns with their function, a process that usually takes years (Hanson et al., 2009; Niehaus et al., 2015). We designed a study to compare human curation to the reported DeepECTF predictions and examined in detail the potential of this new approach.

Methods

Comparing human curation with DeepECTF predictions

453 protein predictions from DeepECTF were grouped into different categories with confidence scores based on evidence assembled from human curation. Concordant predictions for all 4 EC numbers were given a confidence score of 2. These were of two types. Proteins with the identical functional annotation already present in UniProt (with or without an EC number), and proteins with functions not captured by UniProt. Discordant predictions were cases where: 1) another protein

171was known to perform the same function and experimental data showed that there were no
172redundancies; 2) the predicted function was known to be absent in *E. coli;* 3) the published
173literature or comments in UniProt provided evidence for a different function; were given a
174confidence score of 0. When proteins were members of families with many paralogous subgroups
175but no literature disproving the prediction was found and in other cases of uncertain calls, a
176confidence score of 1 was given.
177
178**Bioinformatic analyses and data mining**
179The UniProt (www.uniprot.org)(Bateman et al., 2023), InterPro (www.ebi.ac.uk/interpro/)(Paysan-
180Lafosse et al., 2023), NCBI (www.ncbi.nlm.nih.gov)(NCBI Resource Coordinators, 2016), BioCyc
181(www.biocyc.org)(Karp et al., 2019) and KEGG (Kanehisa et al., 2021) knowledgebases were
182used to gather information on the 453 proteins. PaperBlast was used to query literature
183(https://papers.genomics.lbl.gov/cgi-bin/litSearch.cgi)(Price and Arkin, 2017). Open AI ChatGPT
1844.o was used to extract data from Tables in publications and generate figures
185(https://chatgpt.com/) on August 20 using the prompt: "Extract data from Table X in attached pdf
186file".
187
188**Structural and phylogenetic analyses**
189Crystal structures were retrieved from the Protein Data Bank (www.rcsb.org)(Burley et al., 2022)
190and visualized using Pymol (www.pymol.org)(The PyMOL Molecular Graphics System, Version
1911.8, Schrödinger, LLC). Structure-based multi-sequence alignment derived from available crystal
192structures and AlphaFold structural models (https://alphafold.ebi.ac.uk/) were generated using
193PROMALS3D (prodata.swmed.edu/promals3d/promals3d.php)(Pei et al., 2008) and Espript
194(espript.ibcp.fr/ESPript/ESPript). The percent conservation scores were calculated using the
195program AL2CO (http://prodata.swmed.edu/al2co/al2co.php)(Pei and Grishin, 2001) using 4264
196TsaC/Sua5 sequences and 1955 YciO sequences (Supplemental data 1 and 2). To generate the
197phylogenetic tree, proteins were aligned using MUSCLE (Edgar, 2004). The alignment was
198trimmed using BMGE (Criscuolo and Gribaldo, 2010) and used to build the maximum likelihood
199tree using FastTree (Price et al., 2010) with LG+CAT model with bootstrap (1000 replicates) and
200visualized using iTOL (Letunic and Bork, 2021).
201
202**Sequence similarity networks and gene neighborhood analyses**
203Sequence Similarity Networks (SSNs) and the corresponding Gene Neighborhood Networks
204(GNT) were generated using EFI-EST (EFI Enzyme Similarity Tool, efi.igb.illinois.edu/efi-

est)(Zallot et al., 2019). Briefly, 54,820 sequences of the PF01300 family between 150 and 350 aa in length were retrieved from Uniprot and subjected to EFI-EST. Each node in the network represents one or multiple sequences that share no less than 70% identity. The initial SSN was generated with an Alignment Score (AS) cutoff set such that each connection (edge) represented a sequence identity above 40%. The nodes of paralogs were colored as given in the legend and visualized using Cytoscape (3.10.1)(Shannon et al., 2003). More SSNs were created by gradually increasing the alignment score cutoff in small increments (usually by 5 AS units). This process was repeated until most clusters were homogeneous in color (AS = 70). The genome neighborhood graphs were generated using Gene Graphics (https://genegraphics.net/)(Harrison et al., 2018).

Explainable AI methods

To enhance the interpretability of DeepECTF's multilabel enzyme function predictions, we incorporated an explainable artificial intelligence (XAI) module based on the Local Interpretable Model-agnostic Explanations (LIME) framework (Ribeiro et al., 2016). The approach was specifically adapted to address the multilabel nature of EC number predictions and to provide both local and global insights into the protein sequence segments or residues driving model decisions. Traditional LIME is designed for binary or multiclass outputs, but enzyme function prediction often involves multilabel assignments. To address this, we implemented a multilabel adaptation of LIME, which constructs independent explanation pipelines for each possible label. For each protein sequence input, the XAI module isolates the probability output for a single EC label using the model's sigmoid activation, enabling LIME to generate label-specific explanations. To optimize computational efficiency, explanations are generated primarily for the label with the highest predicted probability for each input sequence, focusing interpretability efforts on the most relevant predictions.

To balance explanation quality with computational feasibility, local LIME explanations are computed for a representative subsample of up to 500 protein sequences per analysis batch. For each sequence, the module generates a local feature importance map, highlighting which residues in the protein sequence most influenced the model's prediction for the selected EC label. For proteins outside this subsample, residue-level importance is estimated based on the aggregate statistics from the explained subset. This approach enables the extraction of both local (residue-level) and global (dataset-level) feature importance profiles, which are subsequently exported for downstream analysis. To further dissect model behavior, feature importance scores

derived from LIME explanations were stratified by prediction type (correct predictions, paralog errors, non-paralog errors, and repetitions, see Table S1). Each protein sequence is mapped to a prediction category based on expert curation of the model outputs. Residue-level importance values are normalized within each error type, and summary plots are generated to visualize the distribution and magnitude of important sequence segments across different error categories.

The XAI module is fully integrated into the DeepECTF workflow, allowing users to generate explainability outputs alongside standard predictions without additional user intervention (https://github.com/Dias-Lab/XAI_DeepProZyme). All explanation results, including local and global feature importance scores, are made available in standard tabular formats for further interpretation or visualization. This explainable AI approach provides actionable insights into the decision-making process of DeepECTF, supporting both the validation of correct predictions and the systematic investigation of model failures, facilitating the development of more robust and trustworthy protein function prediction systems.

Results

Most correctly AI predicted EC numbers are generic or already in the training set.

We analyzed 453 *E. coli* proteins functionally annotated in the Kim et al study (Kim et al., 2023) (Table S1)121 had the same EC number in the August 2024 corresponding UniProt annotation(Bateman et al., 2023) and 87% of these were in the version used to generate the training dataset (Fig. 5 and Table S1A) making these an example of training data contamination. 16 proteins had the exact same function labeled in UniProt but without an EC number or with a partial EC number (Fig. 5 and Table S1BC) and are examples of successful annotation propagation (Fig,1). These 136 cases were all considered correct but not novel predictions (CNN) with confidence scores of 2 (Fig. 5).

The remaining 317 predictions can be split into proteins with partial or different EC numbers present in the UniProt annotation (63 cases, Table S1B) or with no EC number in the corresponding UniProt entry (254 cases, Table S1C). For 27 predictions, more generic annotations than the existing UniProt annotations were given (less precise cases or LSP in Table S1BC). These were also given a confidence score of 2. The 290 remaining cases were subject to human curation. We combined PaperBlast searches and UniProt and EcoCyc data analyses to group the 290 DeepECTF predictions into different categories with confidence scores from 0 to 2 (Table S1 and Fig. 5). The 42 predictions that could neither be validated nor refuted were labeled as uncertain with a confidence score of 1. Three cases, including YgfF discussed below, were validated by publications not captured in UniProt and can be considered successful predictions

(confidence scores of 2). Combining the novel (COR, 3), the obvious (CNN, 136) and the less precise or generic (LSP, 27) predictions brings the number of correct predictions to 166 (36.6%)**.** 245 predictions (54%) were inconsistent with the existing evidence and given confidence scores of 0 (Table S1BC).

Manual analyses reveal logical inconsistencies in AI predictions of unknowns.

The 245 predictions with confidence scores of 0 can be separated into two categories. Those with published evidence refuting the annotation (77 NPI and 42 PLI in Table S1BC) and those that were replications of identical EC numbers (126 REP in Table S1BC). Examples of the first type (n=119) are given in Table 2. For example, YjhQ/b4307 is predicted to be a mycothiol synthase (EC:2.3.1.189), but mycothiol is not a molecule synthesized by *E. coli* and the remaining pathway genes are absent from the genome {BioCyc ID: PWY1G-0). YrhB/b3446 is predicted to be a 6-carboxytetrahydropterin synthase (EC: 4.1.2.50) but *E. coli* already encodes this enzyme (QueD/b2765) and a *queD* mutant lacks this activity (Zallot et al., 2017).

Replication of identical EC numbers occurred for 126 proteins (Fig. 5 and Table S1BC). Repetitions of EC numbers do occur in bacterial genomes, particularly with families that, despite having 4 EC numbers, have generic functions. For example, histidine kinases with different substrate specificities are frequent, with 29 annotated in *E. coli* (Table S1D, top). However, analysis of the protein family domain membership showed that most of the repetitions in these data were errors, except for the less specific annotation cases classified as correct predictions above (Table S1C). For example, out of the 12 proteins annotated as histidine kinase (EC:2.7.13.3), none of them have sequence similarity to histidine kinase families, and eight have been annotated with different and experimentally validated functions (such as ferric enterobactin transport protein FepE for b0587)(Table S1C). For the 15 proteins annotated as “protein-Npi-phosphohistidine-sugar phosphotransferase (EC:2.7.1.69) (or PTS family proteins), 4 were indeed PTS transporters but were given less specific annotations than the ones in UniProt (Table S1C). The 11 remaining were part of transporter families not related to PTS (Table S1B). This type of error may be due to inherent limitations in how AI methods operate. Indeed, if the input features in the training data lack the biological structure and cannot leverage information to distinguish between different functions, the model is expected to make frequency-dependent predictions that reflect the training data, as demonstrated with histidine kinases.

Correct separation of paralogous group can confirm or refute AI based functional predictions

Many of the incorrect or uncertain (confidence scores of 0 or 1) were part of protein families with paralogs (Fig. 5 and Table S1BC). For example, b2100 was annotated as a dehydro-2-deoxygluconokinase (EC 2.7.1.92 using DeepECTF and as an uncharacterized sugar kinase YegV (EC 2.7.1.-) in UniProt (Table S1B). These two predictions differ by the 4th or last position of the EC number that specifies substrate specificity. This protein is a member of a superfamily of sugar kinases with multiple non-isofunctional paralogous subgroups that phosphorylate different substrates (Table S1D, bottom). The dehydro-2-deoxygluconokinase (EC 2.7.1.92) activity is encoded by another member of this superfamily KdgK/b3526 (Table S1D, bottom). Here, DeepECTF predicted correctly the first 3 digits of the EC number but not the last, making an over-propagation mistake (error 6 in Table 1 and Fig. 1). Correctly separating non-isofunctional paralogous subgroups in a superfamily is difficult, requiring extensive examination of all of the evidence and oftentimes experiments.

We performed an additional, in-depth analysis of the three proteins described using *in vitro* assays in the Kim *et al.* study (YgfF, YdjM, and YciO)(Kim et al., 2023) and show that *in vitro* functionality does not always correspond with *in vivo* functionality.

YggF analysis

YgfF is a member of the large Short-Chain Dehydrogenase/Reductase (SDR) superfamily (IPR002347). The Oppermann and Persson groups developed a nomenclature system and HMM-based classification (http://www.sdr-enzymes.org/) that distinguish different functional subgroups of the SDR superfamily(Persson et al., 2009; Kallberg et al., 2010). This resource predicts YgfF is part of the SDR63C/Glucose 1-dehydrogenase subgroup, the activity predicted and validated in the.(Kim et al., 2023) study. This prediction demonstrates the accurate propagation of functional annotation and is a successful prediction.

YjdM analysis

YjdM was predicted and shown to catalyze phosphonoacetate hydrolase (EC:3.11.1.2) activity *in vitro*. In *E. coli*, the *yjdM* gene is located upstream of the methylphosphonate catabolism operon (*phnCDEFGHIJKLMNOP)* (Fig. 6). Underscoring the challenges in annotations, the initial report that the protein was involved in phosphonate catabolism was later refuted with additional genetic analyses (Metcalf and Wanner, 1993). The experimentally validated phosphonoacetate hydrolase (PhnA) is part of a non-homologous family and expression of members of this family in

E. coli suggested that phosphonoacetate hydrolase activity was not present in this organism (Kulakova et al., 1997). Genome neighborhoods of *phnA* show strong clustering with genes encoding phosphonoacetate transporters, phosphonoacetate sensing regulators, and in some cases, enzymes involved in 2-aminoethylphosphonate catabolism (Kulakova et al., 2001). However, except for *E. coli, yjdM* genes are generally not close to phosphonate catabolism or transport genes (Fig. 6). In conclusion, the phosphonoacetate hydrolase activity observed *in vitro* is not supported, and additional *in vivo* experiments are required to confirm the biological role of this enzyme. This prediction was given a confidence score of 1

YciO analysis

YciO is a member of the same Pfam family (PF01300) as TsaC/Sua5 and DeepEC predicts YCiO has the same function as TsaC/Sua5. TsaC and Sua5 (the latter known in bacteria as TsaC2) are two types of the well characterized L-threonylcarbamoyladenylate synthase (EC 2.7.7.87) that catalyzes the first step in the synthesis of the universal tRNA modified nucleoside N-6-threonylcarbamoyladenosine or t^6A (TsaC and Sua5 have a common catalytic domain and differ by the presence of an additional domain in Sua5)(Su et al., 2022; Pichard-Kostuch et al., 2023). The function of TsaC/Sua5 was first elucidated in 2009 (El Yacoubi et al., 2009). A structure-based multi-sequence alignment comparing YciO and TsaC/Sua5 shows that the active site residues of TsaC are largely conserved in YciO, suggesting similar catalytic activities (Fig. 7A). YciO catalyzed the synthesis of L-threonylcarbamoyladenylate from ATP, L-threonine, and bicarbonate, *in vitro* (Kim et al., 2023). However, the activity reported (0.14 nM/min TC-AMP production rate) for *E. coli* YciO is more than four orders of magnitude weaker than that of *E. coli* TsaC (2.8 μM/min) at the same enzyme concentration and similar reaction conditions(Swinehart, W. Deutsch, C. Sarachan, K.L. Luthra, A.; de Crécy-Lagard, V. Swairjo, M. Agris, P.F. Iwata-Reuyl, 2019), consistent with the possibility of a missing partner or a different biological substrate for YciO.

YciO does not perform the same function as TsaC/Susa5 *in vivo* experiments(El Yacoubi et al., 2009; Gerdes et al., 2011). Genome neighborhood and structural data suggest that the function of YciO may be related to rRNA rather than tRNA metabolism. The evidence related to rRNA is as follows: 1) the structure of YciO exhibits a large positively charged surface predicted to interact with RNA ((Jia et al., 2002) and Fig. 7B and C). This large positively charged surface is conserved in YciO proteins and is absent in TsaC proteins. 2) in many species, *yciO* genes are colocalized with *rnm* genes (Fig. 8), which encode the recently characterized RNase AM, a 5′ to 3′ exonuclease that matures the 5′ end of all three ribosomal RNAs in *E. coli*(Jain, 2020).

Models of enzyme evolution go through promiscuous stages(Ribeiro et al., 2023). TsaC is an enzyme predicted to have been present in the Last Universal Common Ancestor (LUCA)(Gagler et al., 2022; Pichard-Kostuch et al., 2023). YciO is a likely paralog of TsaC (Fig. 7), and as such, it is likely to have residual ancestral catalytic activity that can be detected *in vitro*. In summary, the functional puzzle is far from being solved for proteins of the YciO subgroup, and even if the existing data suggest a role in RNA metabolism, it cannot be the same as TsaC, and the EC number 2.7.7.87 prediction was given a confidence score of 0.

Feature Importance Patterns identify limitations in the DeepEC model

Application of the XAI module to DeepECTF predictions enabled residue-level analysis of feature importance across protein sequences, stratified by prediction type: correct predictions, paralog errors, non-paralog errors, and repetition errors (Fig. 9). For correct predictions, the feature importance profiles identified a small number of residues-often corresponding to known catalytic or conserved domain positions, showing high contribution scores, indicating that features whose function relies on localized residues for function are more likely to be accurately predicted. In contrast, for all categories of erroneous predictions, feature importance profiles had small deviations from the null values, indicating that the model is not able to leverage information about the protein sequence in making its predictions.

Discussion

Our expert curation of ML-based EC number predictions for *E. coli* “unknowns” proteins reveals that these methods still have a lot of room to improve. Biases, data imbalance (under- or overrepresented domains, motifs, and functions), generic EC numbers, lack of data structures that enable the inclusion of feature information, architectural limitations (e.g. incapacity of capturing complex patterns, inability to integrate diverse data sources, lack of regularization, etc.), and poor uncertainty calibration all contribute to challenges in training computational models and are likely contributors in frequency dependent predictions (Pucci et al., 2018; Urban et al., 2020; Mardikoraem and Woldring, 2023; Shahbazi et al., 2023). Efforts to improve the quality and consistency of training data and avoid data leakage between training and testing data can help minimize errors but metrics evaluating the uncertainty in predicted protein functions, or an assessment of the likelihood of the label assignment, need to become standard model outputs. In the absence of these metrics, XAI can be used to provide insights into feature importance and model behavior. Many advanced machine learning models, particularly deep neural networks such as DeepECTF, function as ‘black boxes’, where the prediction algorithm is difficult to

interpret. Explainable AI (XAI) techniques can help to dissect ‘black-box’ models and understand what input features are driving correct and incorrect predictions (Chennam et al., 2023) and provide insights about which data to include in future models to improve predictions.

Our study also shows that the model’s lack of inclusion of the broader accumulated knowledge in the fields of protein structure and function, biochemical pathways and the absence of a mechanism to identify ‘illogical’ predictions led to erroneous predictions for most proteins of unknown functions, although the model was able to successfully propagate functions among proteins that were included in the training data. We also want to emphasize that *in vitro* activity alone is not sufficient to validate the function of a protein *in vivo* (García-Contreras et al., 2012; Punekar, 2018). Indeed, enzymes evolve by duplication, divergence, and subsequent sub, neo or hypo functionalization (Ribeiro et al., 2023; Birchler, 2025). *In vitro* activities of many enzymes can show promiscuity, which is useful for biotechnological applications (Robinson et al., 2020) but does not guarantee that the protein plays that specific role *in vivo* (Copley, 2015) as shown here with the YciO example. Best practices in functional annotations of enzymes couple biochemical and contextual evidence and reflect the Gene Ontology Consortium definitions, capturing the cellular component as well as the molecular and biological functions (Ashburner et al., 2000).

PLM-driven approaches are becoming mainstream tools for propagating known functional annotations among isofunctional proteins (https://www.uniprot.org/help/ProtNLM) . Computational models will be further enabled by the integration of complementary evidence such as structural data to identify active site signature residues (Derry and Altman, 2023; Sajid et al., 2023; Yu et al., 2023), gene neighborhood context (Urhan et al., 2023; Hwang et al., 2024), and chemical reaction specificity (Qian et al., 2024), all of which help to distinguish non-isofunctional paralogous subgroups(Ribeiro et al., 2023). Recent models like MAPred (Rong et al., 2024) align with modern goals of interpretability and transparency in machine learning models (Chennam et al., 2023; Derry and Altman, 2023; Rong et al., 2024) and implements a feature-dropping approach that represents a step toward a more quantitative and nuanced ‘self’ evaluation of model predictions. The future in this space is exciting, even though the limits of these models are just now beginning to be understood.

Table 1: Main types of errors that lead to erroneous functional annotations of proteins. Error types 1-6 error types correspond to the errors numbered in Figure 1.

Error Types		
False unknowns	**Annotated as unknown or general but function is known and published**	
	Details	***Example***
1 Failure to capture the literature	Not captured in any database or captured in some databases but not others. Also annotated as vague when precise annotation is known.	CT_611 is captured as folylpolyglutamate synthase in KEGG (ctr:CT_611) but this annotation is still not in Uniprot as of April 2024 (O84617). See also examples in (Price and Arkin, 2024).
2 Naming issues	Inconsistent naming of the same entities.	See GroEL in (Lockwood et al., 2019) .
Propagation failures	Captured for one member of a family but not propagated to others.	The MptE protein that encodes 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase in most Archaea (De Crécy-Lagard et al., 2012) is captured in BioCyc for *Methanocaldococcus jannaschii DSM 2661* (https://biocyc.org/gene?orgid=MJ&id=MJ_RS08700-MONOMER) but not propagated to any archaeal homolog.
False knowns	**Annotated as precise but wrong (should be general or another annotation) or incomplete**	
	Details	***Example***
3 Multiple functions	Protein has multiple functions because of fusions moonlighting or promiscuity and only one of the functions is captured.	Very common few fusions have been linked to the multiple encoded functions. For example, A5I019 has two functions QueD and PTPS-III only one is called in Uniprot (de Crécy-Lagard, 2014). See also examples in (Price and Arkin, 2024).
4 Curation mistake	1) The data was incorrectly captured by biocurator; 2) Functional annotations may become outdated.	Ureidoglycolate lyase (Percudani et al., 2013).
5 Experimental mistake	Finding has been refuted by other studies: 1) the published data is inconclusive; 2) Inconsistency occurs when different databases or resources provide conflicting functional annotations for the same protein.	DUF34 family was annotated as GTP cyclohydrolase 1B (Reed et al., 2021).
6 Over-annotation of paralogs	Annotation wrongly propagated to non-isofunctional paralogous groups.	See examples in (Schnoes et al., 2009; Zallot et al., 2016; Rembeza and Engqvist, 2021).

Table 2. Example where experimental evidence contradicts the DeepECTF predictions.

Locus tag	Uniprot EC	Uniprot Function	DeepEC EC	DeepEC Function	Notes
Example of overpropagation mistakes					
b2736	1.1.1.411	L-threonate dehydrogenase	1.1.1.60	2-hydroxy-3-oxopropionate reductase	In vivo and in vitro data validate these proteins as involved D-threonate degradation (Zhang et al., 2016).
b2738	4.1.1.104	3-oxo-tetronate 4-phosphate decarboxylase	4.1.2.17	L-fuculose-phosphate aldolase	
b2739	5.3.1.35	2-oxo-tetronate isomerase	5.3.1.22	Hydroxypyruvate isomerase	
b4403	2.1.1.-	Uncharacterized tRNA/rRNA methyltransferase LasT	2.1.1.200	tRNA (cytidine32/uridine32-2'-O)-methyltransferase	This EC:2.1.1.200 activity is catalyzed by TrmJ/b2532 and the mutant is devoid of the Um/Cm_{32} modification (Purta et al., 2006).
b2160	2.7.1.-	Uncharacterized sugar kinase YeiI	2.7.1.83	Pseudouridine kinase	This activity is catalyzed by PsuK/b2166 and the mutant cannot use Psi as a C source (Preumont et al., 2008)
b3038	6.3.1.-	Putative acid--amine ligase YgiC	6.3.1.8	Glutathionylspermidine synthase	This activity is catalyzed by Gsp and YgiC does not have the same activit (Sui et al., 2012).
b1267	None	Uncharacterized protein YciO	2.7.7.87	L-threonylcarbamoyladenylate synthase	YciO is a paralog of TsaC but does catalyze the same activity (El Yacoubi et al., 2009; Gerdes et al., 2011).
Example of refuted predictions					
b2438	None	Bacterial microcompartment shell protein EutK (Ethanolamine utilization protein EutK)	2.1.1.223	tRNA1Val (adenine37-N6)-methyltransferase	This activity is catalyzed by TrmN6/b2575 and the mutant does not make the modification (Golovina et al., 2009).
b0254	None	HTH-type transcriptional regulator PerR (Peroxide resistance protein PerR)	2.4.2.29	tRNA-guanosine34 $preQ_1$ transglycosylase	This activity is catalyzed by Tgt/b0406 and the mutant does not insert $preQ_1$ in tRNA (Noguchi et al., 1982).
b3446	None	Uncharacterized protein YrhB	4.1.2.50	6-carboxytetrahydropterin synthase	This activity is catalyzed by QueD/b2765) and the mutant lacks this activity (Zallot et al., 2017).
b4307	None	Uncharacterized N-acetyltransferase YjhQ	2.3.1.189	mycothiol synthase	Mycothiol is not synthesized by *E. coli* and the remaining pathway enzymes are absent (BioCyc ID: PWY1G-0).

Competing interests

The authors declare no competing interests.

Data availability

Supplemental files are available at FigShare (10.1101/2024.07.01.601547). Table S1 contains the manual review of EC number predictions for 450 *E. coli* unknowns made by DEEP EC. Table S2 contains gene neighborhood information for selected *yjdM* (IPR004624 family) genes shown in Fig. 6. Table S3 includes the identifiers for the PF01300 family members used in Fig. 8 SSN and Table S4 the gene neighborhood information for selected PF01300 family genes. Supplemental data 1 contains the TsaC sequences used to generate positional conservation scores for Fig. 7. Supplemental data 2 contains YciO sequences used to generate the positional conservation scores for Figure 7.

Acknowledgments

This work was funded by NIGMS (GM110588 to MAS and VDC and GM145937 to IF) and DARPA (HR0012530305 to VDC and RD). We thank Alan J. Bridge (SIB) for critical reading and input on early versions of the manuscript and the peer editors and reviewers who helped us greatly improve the manuscript.

Contributions

VDC conceived the study, wrote the first draft of the manuscript, and did the analyses for Tables 1-2 and S1. MAS created figure 7 and wrote sections of the manuscript. YY helped generate the data for Table S1-4 and created Figure 1-6 and 8. RD ran the XAI analysis and wrote sections of the manuscript and created Fig. 9. NS ran the mapping and preprocessing of the input data for the XAI analysis and contributed to the implementation of the XAI module in DeepECTF. IF wrote sections of the manuscript. All authors edited the final draft.

References

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29.

Ayres, G., Munsamy, G., Heinzinger, M., Ferruz, N., Yang, K., and Lorenz. Philipp (2023). HiFi-NN annotates the microbial dark matter with Enzyme Commission numbers., in *Machine Learning for Structural Biology Workshop, NeurIPS 2023.*

Bansal, P., Morgat, A., Axelsen, K. B., Muthukrishnan, V., Coudert, E., Aimo, L., et al. (2022). Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 50, D693–D700. doi: 10.1093/nar/gkab1016

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., et al. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 51, D523–D531. doi: 10.1093/nar/gkac1052

Birchler, J. A. (2025). When is it subfunctionalization and when is it not? *G3: Genes, Genomes, Genetics* 15, jkae269. doi: 10.1093/g3journal/jkae269

Bordin, N., Sillitoe, I., Lees, J. G., and Orengo, C. (2021). Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds. *Front Mol Biosci* 8, 668184. doi: 10.3389/fmolb.2021.668184

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., et al. (2022). RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res D1* 51, D488-D508. doi: 10.1093/nar/gkac1077

Chen, L., and Vitkup, D. (2007). Distribution of orphan metabolic activities. *Trends Biotechnol* 25, 343–348. doi: 10.1016/j.tibtech.2007.06.001

Chennam, K. K., Mudrakola, S., Maheswari, V. U., Aluvalu, R., and Rao, K. G. (2023). “Black Box Models for eXplainable Artificial Intelligence,” 1–24. doi: 10.1007/978-3-031-12807-3_1

Copley, S. D. (2015). An evolutionary biochemist’s perspective on promiscuity. *Trends Biochem Sci* 40, 72–78. doi: 10.1016/j.tibs.2014.12.004

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10, 210. doi: 10.1186/1471-2148-10-210

Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J. D. (2018). No wisdom in the crowd: genome annotation in the era of big data – current status and future prospects. *Microbial Biotechnology.* 11, 588–605. doi: 10.1111/1751-7915.13284

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., and Orengo, C. A. (2015). Functional classification of CATH superfamilies: A domain-based approach for protein function annotation. *Bioinformatics* 31, 3460–7. doi: 10.1093/bioinformatics/btv398

de Crécy-Lagard, V. (2014). Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. *Comput Struct Biotechnol J* 10, 41–50. doi: 10.1016/j.csbj.2014.05.008

de Crécy-lagard, V., Amorin de Hegedus, R., Arighi, C., Babor, J., Bateman, A., Blaby, I., et al. (2022). A roadmap for the functional annotation of protein families: a community perspective. *Database* 2022, baac062. doi: 10.1093/database/baac062

De Crécy-Lagard, V., Phillips, G., Grochowski, L. L., Yacoubi, B. E., Jenney, F., Adams, M. W. W., et al. (2012). Comparative genomics guided discovery of two missing archaeal enzyme families involved in the biosynthesis of the pterin moiety of tetrahydromethanopterin and tetrahydrofolate. *ACS Chem Biol* 7. doi: 10.1021/cb300342u

Dembech, E., Malatesta, M., De Rito, C., Mori, G., Cavazzini, D., Secchi, A., et al. (2023). Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc Natl Acad Sci U S A* 120, e2218329120. doi: 10.1073/pnas.2218329120

Derry, A., and Altman, R. B. (2023). Explainable protein function annotation using local structure embeddings. *bioRxiv*, 2023.10.13.562298. doi: 10.1101/2023.10.13.562298

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods Mol Biol* 985, 17–45. doi: 10.1007/978-1-62703-299-5_2

Durairaj, J., Waterhouse, A. M., Mets, T., Brodiazhenko, T., Abdullah, M., Studer, G., et al. (2023). Uncovering new families and folds in the natural protein universe. *Nature* 622, 646–653. doi: 10.1038/s41586-023-06622-3

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

El Yacoubi, B., Lyons, B., Cruz, Y., Reddy, R., Nordin, B., Agnelli, F., et al. (2009). The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. *Nucleic Acids Res* 37, 2894–2909. doi: 10.1093/nar/gkp152

Friedberg, I. (2006). Automated protein function prediction - the genomic challenge. *Brief Bioinformatics* 7, 225–242. doi: doi: 10.1093/bib/bbl004

Gagler, D. C., Karas, B., Kempes, C. P., Malloy, J., Mierzejewski, V., Goldman, A. D., et al. (2022). Scaling laws in enzyme function reveal a new kind of biochemical universality. *Proc Natl Acad Sci U S A* 119, e2106655119. doi: 10.1073/pnas.2106655119

García-Contreras, R., Vos, P., Westerhoff, H. V., and Boogerd, F. C. (2012). Why *in vivo* may not equal *in vitro* – new effectors revealed by measurement of enzymatic activities under the same *in vivo* -like assay conditions. *FEBS J* 279, 4145–4159. doi: 10.1111/febs.12007

Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., et al. (2023). The Gene Ontology knowledgebase in 2023. *Genetics* 224, iyad031. doi: 10.1093/genetics/iyad031

Gerdes, S., El Yacoubi, B., Bailly, M., Blaby, I. K., Blaby-Haas, C. E., Jeanguenin, L., et al. (2011). Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genomics* 12, S2. doi: 10.1186/1471-2164-12-S1-S2

Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. (2012). Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions. *J Biol Chem* 287, 29–34. doi: 10.1074/jbc.R111.240945

Ghatak, S., King, Z. A., Sastry, A., and Palsson, B. O. (2019). The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Res* 47, 2446–2454. doi: 10.1093/nar/gkz030

Glasner, M. E., Gerlt, J. A., and Babbitt, P. C. (2006). Evolution of enzyme superfamilies. *Curr Opin Chem Biol* 10, 492–497. doi: 10.1016/j.cbpa.2006.08.012

Golovina, A. Y., Sergiev, P. V., Golovin, A. V., Serebryakova, M. V., Demina, I., Govorun, V. M., et al. (2009). The *yfiC* gene of *E. coli* encodes an adenine-N6 methyltransferase that specifically modifies A37 of $tRNA_1^{Val}$ (cmo^5UAC). *RNA* 15, 1134–1141. doi: 10.1261/rna.1494409

Hamamsy, T., Barot, M., Morton, J. T., Steinegger, M., Bonneau, R., and Cho, K. (2023). Learning sequence, structure, and function representations of proteins with language models. *bioRxiv*, 2023.11.26.568742. doi: 10.1101/2023.11.26.568742

Hanson, A. D., Pribat, A., Waller, J. C., and de Crécy-Lagard, V. (2009). “Unknown” proteins and “orphan” enzymes: the missing half of the engineering parts list--and how to find it. *Biochem J* 425, 1–11.

Harrison, K. J., de Crécy-Lagard, V., and Zallot, R. (2018). Gene Graphics: A genomic neighborhood data visualization web application. *Bioinformatics* 34, 1406–1408. doi: 10.1093/bioinformatics/btx793

Huang, Y., Lin, Y., Lan, W., Huang, C., and Zhong, C. (2024). GloEC: a hierarchical-aware global model for predicting enzyme function. *Brief Bioinform* 25, bbae365. doi: 10.1093/bib/bbae365

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. *Nat Commun* 15, 2880. doi: 10.1038/s41467-024-46947-9

International Society for Biocuration (2018). Biocuration: Distilling data into knowledge. *PLoS Biol* 16, e2002846. doi: 10.1371/journal.pbio.2002846

Jain, C. (2020). RNase AM, a 5' to 3' exonuclease, matures the 5' end of all three ribosomal RNAs in *E. coli*. *Nucleic Acids Res* 48. doi: 10.1093/NAR/GKAA260

Jia, J., Lunin, V. V, Sauve, V., Huang, L. W., Matte, A., and Cygler, M. (2002). Crystal structure of the YciO protein from *Escherichia coli*. *Proteins* 49, 139–41.

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 17, 184. doi: 10.1186/s13059-016-1037-6

Kallberg, Y., Oppermann, U., and Persson, B. (2010). Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models. *FEBS J* 277, 2375–2386. doi: 10.1111/j.1742-4658.2010.07656.x

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res* 49, D545–D551. doi: 10.1093/nar/gkaa970

Karp, P. (2004). Call for an enzyme genomics initiative. *Genome Biol* 5, 401. doi: 10.1186/gb-2004-5-8-401

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., et al. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 20, 1085–1093. doi: 10.1093/bib/bbx085

Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., and Lee, S. Y. (2023). Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat Commun* 14, 7370. doi: 10.1038/s41467-023-43216-z

Kobras, C. M., Fenton, A. K., and Sheppard, S. K. (2021). Next-generation microbiology: from comparative genomics to gene function. *Genome Biol* 22, 123. doi: 10.1186/s13059-021-02344-9

Kulakova, A. N., Kulakov, L. A., Akulenko, N. V., Ksenzenko, V. N., Hamilton, J. T. G., and Quinn, J. P. (2001). Structural and Functional Analysis of the Phosphonoacetate Hydrolase (*phnA*) Gene Region in *Pseudomonas fluorescens* 23F. *J Bacteriol* 183, 3268–3275. doi: 10.1128/JB.183.11.3268-3275.2001

Kulakova, A. N., Kulakov, L. A., and P. Quinn, J. (1997). Cloning of the phosphonoacetate hydrolase gene from Pseudomonas fluorescens 23F encoding a new type of carbon–phosphorus bond cleaving enzyme and its expression in *Escherichia coli* and *Pseudomonas putida*. *Gene* 195, 49–53. doi: 10.1016/S0378-1119(97)00151-0

Kyrpides, N. C. (1999). Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide. *Bioinformatics* 15, 773–774.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 8, 995. Available at: http://dx.doi.org/10.1038/nrm2281

Lespinet, O., and Labedan, B. (2006). Puzzling over orphan enzymes. *Cell Mol Life Sci* 63, 517–523.

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 49, W293–W296. doi: 10.1093/nar/gkab301

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 49, D1020–D1028. doi: 10.1093/nar/gkaa1105

Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G., and Doxey, A. C. (2020). An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* 6, e000341. doi: 10.1099/mgen.0.000341

Lockwood, S., Brayton, K. A., Daily, J. A., and Broschat, S. L. (2019). Whole proteome clustering of 2,307 proteobacterial genomes reveals conserved proteins and significant annotation issues. *Front Microbiol* 10, 383.. doi: 10.3389/fmicb.2019.00383

MacDougall, A., Volynkin, V., Saidi, R., Poggioli, D., Zellner, H., Hatton-Ellis, E., et al. (2020). UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* 36, 4643–4648. doi: 10.1093/bioinformatics/btaa485

Mardikoraem, M., and Woldring, D. (2023). Protein Fitness Prediction Is Impacted by the Interplay of Language Models, Ensemble Learning, and Sampling Methods. *Pharmaceutics* 15, 1337. doi: 10.3390/pharmaceutics15051337

Metcalf, W. W., and Wanner, B. L. (1993). Evidence for a fourteen-gene, *phnC* to *phnP* locus for phosphonate metabolism in *Escherichia coli*. *Gene* 129, 27–32. doi: 10.1016/0378-1119(93)90692-V

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44, D7–D19. doi: 10.1093/nar/gkv1290

Niehaus, T. D., Thamm, A. M. K., De Crécy-Lagard, V., and Hanson, A. D. (2015). Proteins of unknown biochemical function: A persistent problem and a roadmap to help overcome it. *Plant Physiol* 169, 1436–1442. doi: 10.1104/pp.15.00959

Noguchi, S., Nishimura, Y., Hirota, Y., and Nishimura, S. (1982). Isolation and characterization of an *Escherichia coli* mutant lacking tRNA-guanine transglycosylase. Function and biosynthesis of queuosine in tRNA. *J Biol Chem* 257, 6544–6550.

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., et al. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* 51, D678–D689. doi: 10.1093/nar/gkac1003

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., et al. (2023). InterPro in 2022. *Nucleic Acids Res* 51, D418–D427. doi: 10.1093/nar/gkac993

Pei, J., and Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17, 700–712. doi: 10.1093/bioinformatics/17.8.700

Pei, J., Tang, M., and Grishin, N. V. (2008). PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* 36, W30–W34. doi: 10.1093/nar/gkn322

Percudani, R., Carnevali, D., and Puggioni, V. (2013). Ureidoglycolate hydrolase, amidohydrolase, lyase: how errors in biological databases are incorporated in scientific papers and vice versa. *Database (Oxford)* 2013, bat071–bat071. doi: 10.1093/database/bat071

Persson, B., Kallberg, Y., Bray, J. E., Bruford, E., Dellaporta, S. L., Favia, A. D., et al. (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem Biol Interact* 178, 94–98. doi: 10.1016/j.cbi.2008.10.040

Peterson, J. K., MacDonald, M. L., and Ellis, V. A. (2025). First whole-genome sequence of *Triatoma sanguisuga* (Le Conte, 1855), vector of Chagas disease. *G3: Genes, Genomes, Genetics* 15. doi: 10.1093/g3journal/jkae308

Pichard-Kostuch, A., Da Cunha, V., Oberto, J., Sauguet, L., and Basta, T. (2023). The universal Sua5/TsaC family evolved different mechanisms for the synthesis of a key tRNA modification. *Front Microbiol* 14, 1204045. doi: 10.3389/fmicb.2023.1204045

Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C.-H., Lu, Z., et al. (2017). On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* 33, 3454–3460. doi: 10.1093/bioinformatics/btx439

Prabakaran, R., and Bromberg, Y. (2023). Functional profiling of the sequence stockpile: a review and assessment of*in silico*; prediction tools. *bioRxiv*, 2023.07.12.548726. doi: 10.1101/2023.07.12.548726

Precord, T. W., Ramesh, S., Dommaraju, S. R., Harris, L. A., Kille, B. L., and Mitchell, D. A. (2023). Catalytic site proximity profiling for functional unification of sequence-diverse Radical *S* -

Adenosylmethionine enzymes. *ACS Bio & Med Chem Au* 3, 240–251. doi: 10.1021/acsbiomedchemau.2c00085

Preumont, A., Snoussi, K., Stroobant, V., Collet, J. F., and Van Schaftingen, E. (2008). Molecular identification of pseudouridine-metabolizing enzymes. *J Biol Chem* 283, 25238–25246.

Price, M. N., and Arkin, A. P. (2017). PaperBLAST: text mining papers for information about homologs. *mSystems* 2, e00039-17. doi: 10.1128/mSystems.00039-17

Price, M. N., and Arkin, A. P. (2024). Interactive Tools for Functional Annotation of Bacterial Genomes. *bioRxiv*, 2024.04.15.589591. doi: 10.1101/2024.04.15.589591

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5, e9490. doi: 10.1371/journal.pone.0009490

Pucci, F., Bernaerts, K. V, Kwasigroch, J. M., and Rooman, M. (2018). Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* 34, 3659–3665. doi: 10.1093/bioinformatics/bty348

Punekar, N. S. (2018). “In Vitro Versus In Vivo: Concepts and Consequences,” in *ENZYMES: Catalysis, Kinetics and Mechanisms*, (Singapore: Springer Singapore), 493–519. doi: 10.1007/978-981-13-0785-0_38

Purta, E., van Vliet, F., Tkaczuk, K. L., Dunin-Horkawicz, S., Mori, H., Droogmans, L., et al. (2006). The *yfhQ* gene of *Escherichia coli* encodes a tRNA:Cm32/Um32 methyltransferase. *BMC Mol. Biol.* 7, 23.

Qian, W., Wang, X., Kang, Y., Pan, P., Hou, T., and Hsieh, C.-Y. (2024). A general model for predicting enzyme functions based on enzymatic reactions. *J Cheminform* 16, 38. doi: 10.1186/s13321-024-00827-y

Radivojac, P., Clark, W., Oron, T., Schnoes, A., Wittkop, T., and Sokolov, A. (2013). A large-scale evaluation of computational protein function prediction. *Nat Methods* 10, e0147952. doi: 10.1371/JOURNAL.PONE.0147952

Ramola, R., Friedberg, I., and Radivojac, P. (2022). The field of protein function prediction as viewed by different domain scientists. *Bioinformatics Advances* 2, vbac057. doi: 10.1093/bioadv/vbac057

Reed, C. J., Hutinet, G., and de Crécy-Lagard, V. (2021). Comparative genomic analysis of the duf34 protein family suggests role as a metal ion chaperone or insertase. *Biomolecules* 11, 1282. doi: 10.3390/biom11091282

Rembeza, E., and Engqvist, M. K. M. (2021). Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. *PLoS Comput Biol* 17, e1009446. doi: 10.1371/journal.pcbi.1009446

Ribeiro, A. J. M., Riziotis, I. G., Borkakoti, N., and Thornton, J. M. (2023). Enzyme function and evolution through the lens of bioinformatics. *Biochem J* 480, 1845–1863. doi: 10.1042/BCJ20220405

Ribeiro, A. J. M., Tyzack, J. D., Borkakoti, N., and Thornton, J. M. (2020). Identifying pseudoenzymes using functional annotation: pitfalls of common practice. *FEBS J* 287, 4128–4140. doi: 10.1111/febs.15142

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why Should I Trust You?”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY, USA: ACM), 1135–1144. doi: 10.1145/2939672.2939778

Robinson, S. L., Smith, M. D., Richman, J. E., Aukema, K. G., and Wackett, L. P. (2020). Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. *Synth Biol* 5, ysaa004. doi: 10.1093/synbio/ysaa004

Rocha, J. J., Jayaram, S. A., Stevens, T. J., Muschalik, N., Shah, R. D., Emran, S., et al. (2023). Functional unknomics: Systematic screening of conserved genes of unknown function. *PLoS Biol* 21, e3002222. doi: 10.1371/journal.pbio.3002222

Rodríguez del Río, Á., Giner-Lamia, J., Cantalapiedra, C. P., Botas, J., Deng, Z., Hernández-Plaza, A., et al. (2024). Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature*, 626, 377–384.doi: 10.1038/s41586-023-06955-z

Rong, D., Zheng, W., Zhong, B., Lin, Z., Hong, L., and Liu, N. (2024). *Autoregressive Enzyme Function Prediction with Multi-scale Multi-modality Fusion*. doi: 10.48550/arXiv.2408.06391

Roy, A., Srinivasan, N., and Gowri, V. S. (2009). Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: Implications for function annotation based on homology searches and structural genomics. *In Silico Biol* 9, S41-S55 . doi: 10.3233/ISB-2009-0379

Sajid, S., Mashkoor, M., Jørgensen, M. G., Christensen, L. P., Hansen, P. R., Franzyk, H., et al. (2023). The Y-ome Conundrum: Insights into Uncharacterized Genes and Approaches for Functional Annotation. *Mol Cell Biochem*. doi: 10.1007/s11010-023-04827-8

Sanderson, T., Bileschi, M. L., Belanger, D., and Colwell, L. J. (2023). ProteInfer, deep neural networks for protein functional inference. *Elife* 12 e80942., . doi: 10.7554/eLife.80942

Satish Kumar, V., Dasika, M. S., and Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 8, 212. doi: 10.1186/1471-2105-8-212

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 5, e1000605. doi: DOI: 10.1371/journal.pcbi.1000605

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shahbazi, N., Lin, Y., Asudeh, A., and Jagadish, H. V. (2023). Representation bias in data: a survey on identification and resolution techniques. *ACM Comput Surv* 55, 1–39. doi: 10.1145/3588433

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498–2504.

Škunca, N., Roberts, R. J., and Steffen, M. (2017). Evaluating Computational Gene Ontology Annotations. *Methods Mol Biol .*, 97–109. doi: 10.1007/978-1-4939-3743-1_8

Soldatos, T. G., Perdigão, N., Brown, N. P., Sabir, K. S., and O'Donoghue, S. I. (2015). How to learn about gene function: text-mining or ontologies? *Methods* 74, 3–15. doi: 10.1016/j.ymeth.2014.07.004

Storm, C. E., and Sonnhammer, E. L. L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18, 92–99.

Su, C., Jin, M., and Zhang, W. (2022). Conservation and diversification of tRNA t^6A-modifying enzymes across the three domains of life. *Int J Mol Sci* 23, 13600. doi: 10.3390/ijms232113600

Sui, L., Warren, J. C., Russell, J. P. N., and Stourman, N. V. (2012). Comparison of the functions of glutathionylspermidine synthetase/amidase from *E. coli* and its predicted homologues Ygic and Yjfc. *Int J Biochem Mol Biol* 3.

Swinehart, W. Deutsch, C. Sarachan, K.L. Luthra, A.; de Crécy-Lagard, V. Swairjo, M. Agris, P.F. Iwata-Reuyl, D. (2019). Specificity in the biosynthesis of the universal tRNA nucleoside threonylcarbamoyl adenosine (t6A) - TsaD is the gatekeeper. *Biochemistry* submitted.

Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P. A., Murphy, T. D., et al. (2016). P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *J Anim Sci* 94, 184–184. doi: 10.2527/jas2016.94supplement4184x

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective 1 1Edited by A. R. Fersht. *J Mol Biol* 307, 1113–1143. doi: 10.1006/jmbi.2001.4513

Torres, J., Rozo-Lopez, P., Brewer, W., Saleh Ziabari, O., and Parker, B. J. (2025). Draft genome sequence of the glasshouse-potato aphid *Aulacorthum solani*. *G3: Genes, Genomes, Genetics* 15, jkaf013. doi: 10.1093/g3journal/jkaf013

Urban, G., Torrisi, M., Magnan, C. N., Pollastri, G., and Baldi, P. (2020). Protein profiles: Biases and protocols. *Comput Struct Biotechnol J* 18, 2281–2289. doi: 10.1016/j.csbj.2020.08.015

Urhan, A., Cosma, B.-M., Manson, A. L., and Abeel, T. (2024). SAP: Synteny-aware gene function prediction for bacteria using protein embeddings. 40, btae328. https://doi.org/10.1093/bioinformatics/btae328

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 47, W587–W593. doi: 10.1093/nar/gkz389

Wood, V., Lock, A., Harris, M. A., Rutherford, K., Bähler, J., and Oliver, S. G. (2019). Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol* 9, 180241. doi: 10.1098/rsob.180241

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science (1979)* 379, 1358–1363. doi: 10.1126/science.adf2465

Zallot, R., Harrison, K. J., Kolaczkowski, B., and de Crécy-Lagard, V. (2016). Functional annotations of paralogs: A blessing and a curse. *Life* 6, :39 . doi: 10.3390/life6030039

Zallot, R., Oberg, N., and Gerlt, J. A. (2019). The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* 58, 4169–4182. doi: 10.1021/acs.biochem.9b00735

Zallot, R., Oberg, N., and Gerlt, J. A. (2021). Discovery of new enzymatic functions and metabolic pathways using genomic enzymology web tools. *Curr Opin Biotechnol* 69, 77–90. doi: https://doi.org/10.1016/j.copbio.2020.12.004

Zallot, R., Yuan, Y., and de Crécy-Lagard, V. (2017). The *Escherichia coli* COG1738 member YhhQ is involved in 7-cyanodeazaguanine (preQ$_0$) Transport. *Biomolecules* 7, 12. doi: 10.3390/biom7010012

Zhang, X., Carter, M. S., Vetting, M. W., San Francisco, B., Zhao, S., Al-Obaidi, N. F., et al. (2016). Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. *Proc Natl Acad Sci U S A* 113, E4161-9. doi: 10.1073/pnas.1605546113

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* 20, 244. doi: 10.1186/s13059-019-1835-8

Figures legends

Figure 1. The major two types of challenges in accurate protein functional annotation. (A) Challenge 1: to propagate the existing knowledge to the correct set of unannotated proteins. Challenge 2: to annotate proteins in a family with none of the members linked to any initial known information. Each circle represents a protein in a family with (black) or without (gray) initially known function. The edges that connect two circles represent protein similarity above a preset threshold. The squares represent known functions with E.C.a to E.C.n as examples. (B) The errors during propagation of annotation that could lead to erroneous results. The numbering of errors (red crosses) corresponds to the error types described in Table 1.

Figure 2. Computational workflow used to separate non-isofunctional paralogs in a protein superfamily. Sequence similarity networks of protein families can separate paralogs by integrating different types of information, Arrows indicate the flow of information. The final-colored network shows the paralog families that could be separated.

Figure 3. Workflow for functional annotation of proteins using machine learning. General steps are depicted: input data collection (training, validation, and query datasets), options for machine learning approach, model selection, training, and evaluation, and output data. Supervised, semi-supervised, and unsupervised learning approaches with examples of their respective modeling architectures and software, are colored differently. Blue scale bars represent the availability of labeled data (top) and the level of interpretability (bottom) of machine learning approaches and models available, respectively. The figure illustrates common associations between model architectures and machine learning approaches. However, it should be noted that models such as GNNs, CNNs, RNNs, and Transformers can be applied across supervised, semi-supervised, and unsupervised learning contexts. The total number of publications in each year was collected on June 10, 2024 from PubMed (pubmed.ncbi.nlm.nih.gov) using the search criteria: 'function prediction protein language model' in full data.

Figure 4. Comparison of different machine-learning methods to predict EC numbers. (A) Comparison of macro-F1 scores generated using four methods (GloEC, Protein Infer, DeepEC, and CLEAN) on four datasets (Carbohydrate esterase, New-432, Cofactor-237, and phosphorylase)(data extracted from (Huang et al., 2024)); (B) Comparison of F1-scores generated using eight methods including BlastP, DeepECTF, DeepEC, and CLEAN on three

datasets (New-815, New-392 and Price)(data extracted from (Rong et al., 2024)). Dotted lines mark the BlastP scores. See initial publications for details and references for methods and datasets.

Figure 5. **Classification of DeepECTF predictions**. The 453 EC number predictions for 453 *E. coli* unknowns were manually classified into categories by comparing the EC number in the UniProt database and labeled with confidence scores from 0 to 2. Data extracted from **Table S1.** Abbreviation: COR, Correct prediction; UNC: Uncertain; LSP, Less Precise; PLI, Paralogs incorrect; NPI, Not-paralog incorrect; REP, Repetitions; CNN, Correct but not novel; CS, Confidence score.

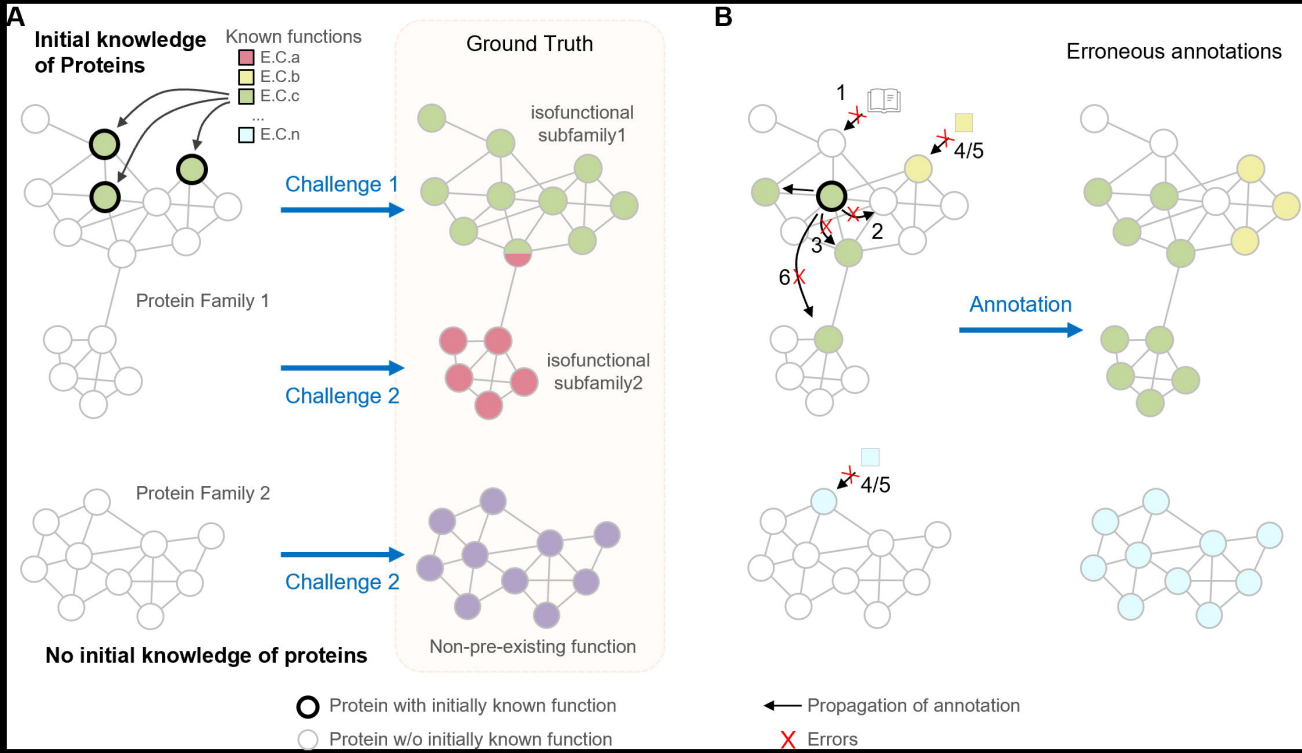
Figure 6. **Gene neighborhoods and metabolic reconstructions do not link YdjM to phosphonate degradation**. Most *ydjM* genes are not near phosphonate degradation operons. A Sequence Similarity Network of 11,986 IPR004624 family members was generated, and the corresponding genomic neighborhood information is available Table S4.

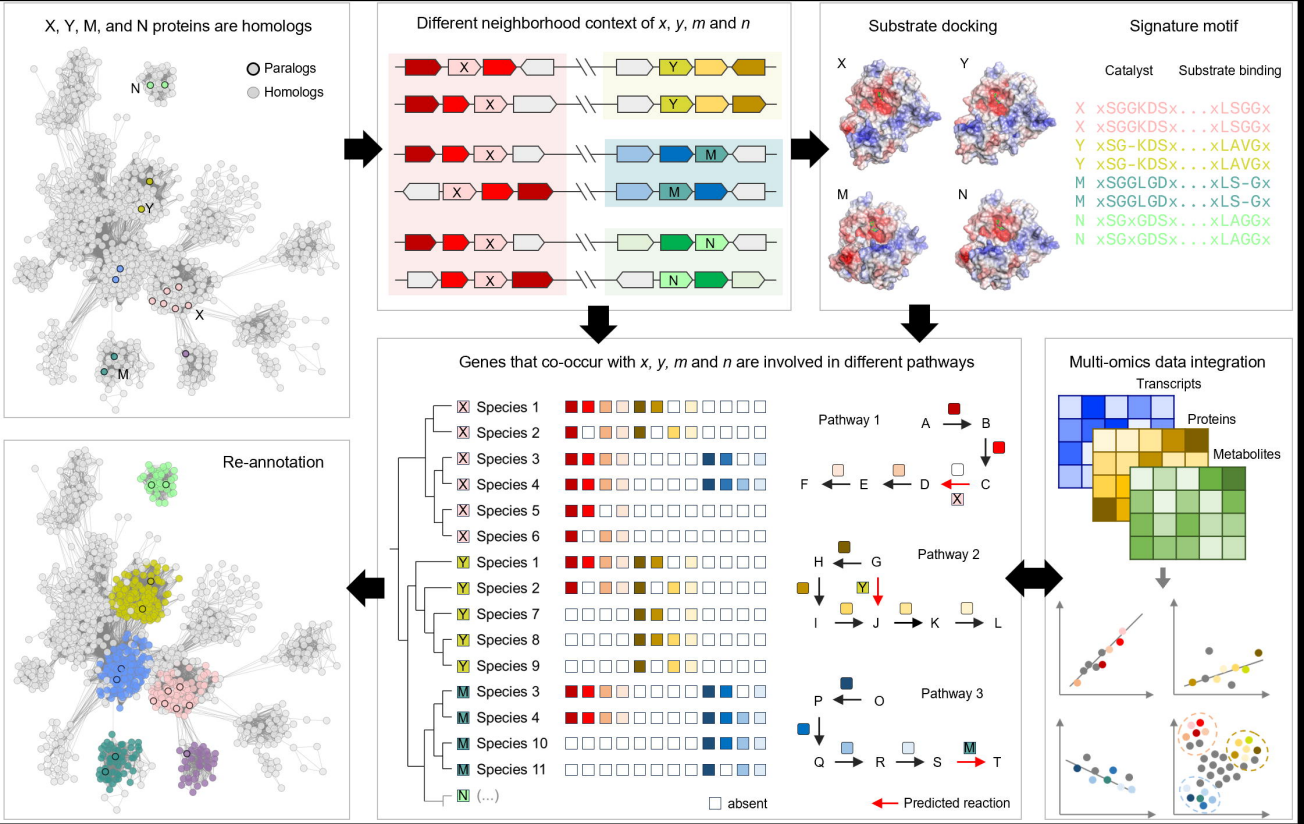
Figure 7. YciO harbors specific molecular surface features. (A) Structure-based multi-sequence alignment of TsaC proteins, the TsaC domains of Sua5 proteins and YciO proteins, derived from available crystal structures and AlphaFold structural models. For crystal structures, the PDB IDs are indicated in the sequence name after the hyphen. Secondary structure elements from the crystal structures of *E. coli* TsaC and *E. coli* YciO are displayed above and below the sequences, respectively. YciO-specific conserved basic residues forming the positively charged surface patch of YciO are shaded in blue. Stars above and below the alignment indicate the crystallographically observed substrate binding residues in *St*Sua5, and the corresponding putative substrate binding residues in *Ec*YciO, respectively. Green stars indicate the Mg^{2+}-ATP binding residues. Orange stars indicate the binding residues for the L-Threonine substrate. *St*Sua5: *Sulfurisphaera tokodaii* Sua5 (UniProt ID Q9UYB2), *Pa*Sua5: *Pyrococcus abyssi* Sua5 (UniProt Q9UYB2), *Ec*TsaC: *Escherichia coli* TsaC (UniProt P45748), *Pv*TsaN: TsaC domain of Pandoravirus TsaN (UniProt A0A291ATS8). *Tm*TsaC2: *Thermotoga maritima* TsaC2 (by the authors, unpublished, UniProt Q9WZV6), *Ab*YciO: *Actinomycetales bacterium* YciO (UniProt A0A1R4F4R9), *Bm*YciO: *Burkholderia multivorans* YciO (UniProt A0A1B4MSV9), *Kp*YciO: *Klebsiella pneumoniae* YciO (UniProt A6T7X1), *Pp*YciO: *Pseudomonas putida* YciO (UniProt A5W0B1), *Se*YciO: *Salmonella enterica* TciO (UniProt A0A601PQ14), *Sf*YciO: *Shigela flexneri* YciO (UniProt P0AFR6), *Ec*YciO: *Escherichia coli* YciO (UniProt P0AFR4). **(B, C)** Surface

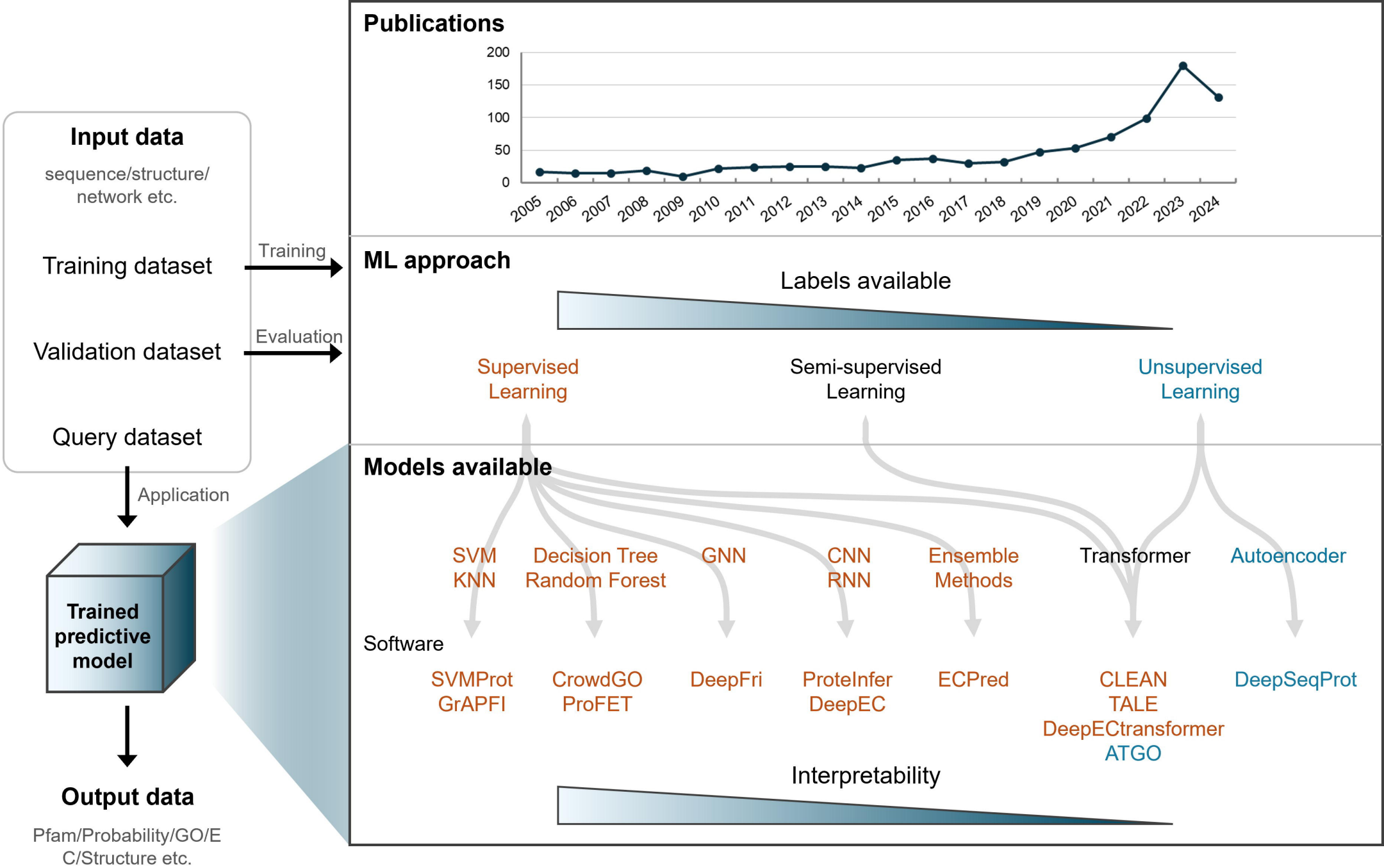
representations of the crystal structures of *Ec*TsaC (**B**) and *Ec*YciO (**C**), color-coded by surface electrostatic potential (top), and by positional sequence conservation score calculated from 4264 TsaC/Sua5 sequences and 1955 YciO sequences (bottom). The color keys for both panels are shown on the right. The conserved, YciO-specific, positively charged surface patch (6% of total molecular surface area) is encircled with a dashed line. The active center of TsaC and the putative active center of YciO are marked with asterisks. The figure illustrates that although the active site is conserved in both protein families, the positively charged surface patch is present and conserved only in the YciO family.

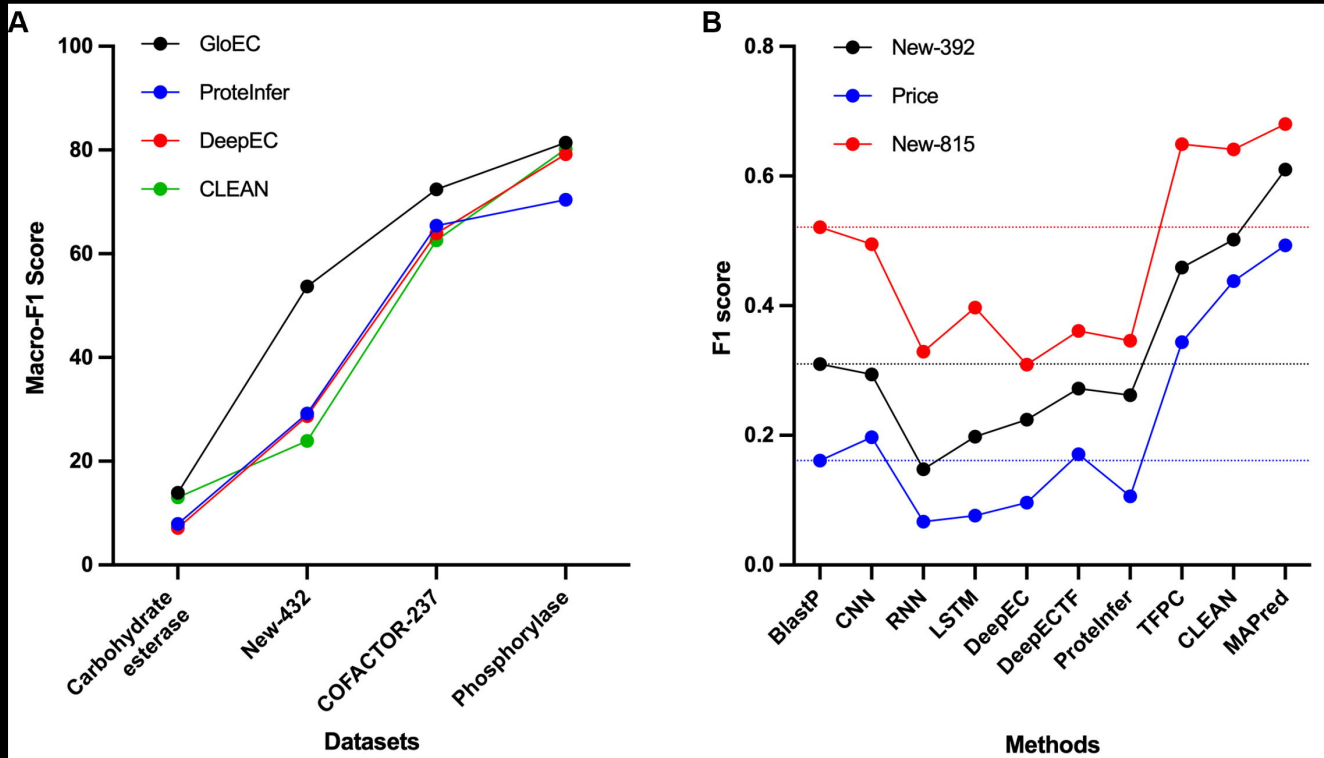
Figure 8. The genes encoding members of the *yciO* and *tsaC* have different genome neighborhood contexts. **(A)** Sequence Similarity Network of 54,820 PF01300 family members between 150 and 350 aa in length generated using EFI-EST. Each node in the network represents one or multiple sequences that share no less than 70% identity. An edge, represented as a line, is drawn between two nodes with alignment scores higher than 70 (similar in magnitude to the negative base-10 logarithm of a BLAST e-value of 1E-70). Paralogs were colored in pairs as indicated. Node borders were colored by gene neighborhood context, *yciO* and *rnm* cluster (red), *tsaC, dprA, smg, yrdD, aroE* and *yrdB* cluster (blue). For better visualization, clusters of less than 40 nodes and singular nodes were hidden. The proteins used in the SSN are available in Table S3. **(B)** Phylogenetic tree of TsaC and YciO proteins. 62 PF01300 proteins were selected from each major cluster in the family SSN and 3 RibB proteins (UniProt Q60364, Q46TZ9, and P0A7J0) were used as the outgroup. Bootstrap values less than 0.75 were indicated by blue dots. Branches and leaves were colored by phyla. TsaC2 proteins that contain YciO/TsaC (PF01300) and Sua5 (PF03481) domains are enclosed in a bracket. The genome neighborhood information is available in Table S4. Selected example organisms were indicated by arrows and a two-letter code in the SSN. Each corresponding genome neighborhood schematic was indicated by a circle in the same color as the node. As, *Arthrobacter saudimassiliensis*; Ec, *Escherichia coli*; Pr, *Pseudohaliea rubra*; Sd, *Shigella dysenteriae*; Sx, *Shewanella xiamenensis*.

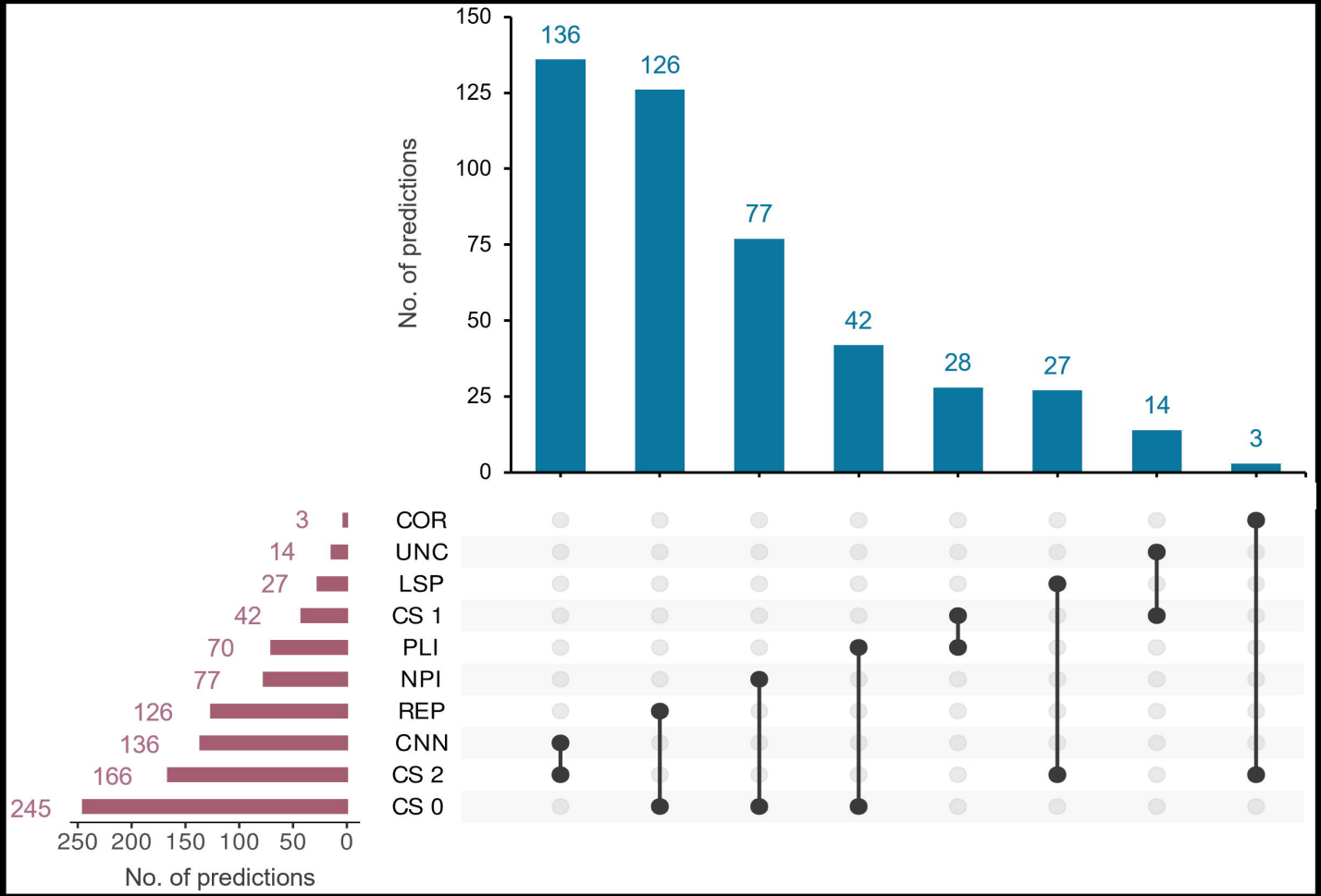
Figure 9. Residue-level feature importance profiles for DeepECTF highlight the distinct interpretability signatures associated with model predictions. Each line plot displays the normalized contribution (average feature importance computed via LIME) of individual residues across protein sequences for four categories: correct predictions (**A**), paralog errors (**B**), non-paralog errors (**C**), and hallucinations or repetitive predictions (**D**). Shaded regions represent standard deviation across instances within each category.

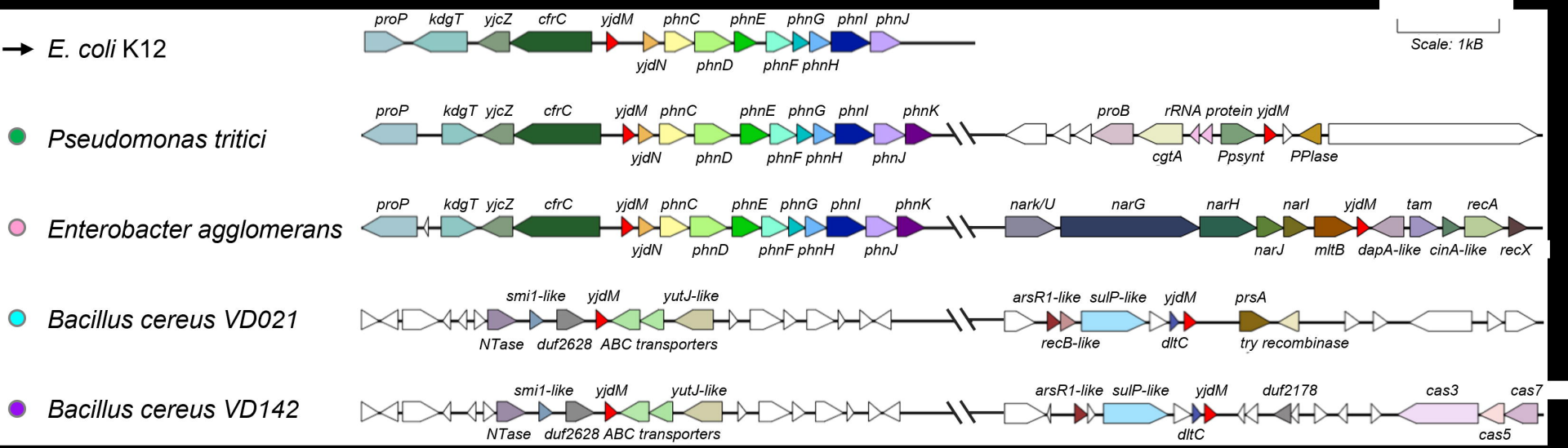

A
Initial knowledge of Proteins
Known functions
E.C.a
E.C.b
E.C.c
...
E.C.n
Ground Truth
isofunctional subfamily1
Challenge 1
Protein Family 1
isofunctional subfamily2
Challenge 2
Protein Family 2
Challenge 2
Non-pre-existing function
No initial knowledge of proteins
B
Erroneous annotations
1
4/5
2
3
6
Annotation
4/5
Protein with initially known function
Protein w/o initially known function
Propagation of annotation
Errors

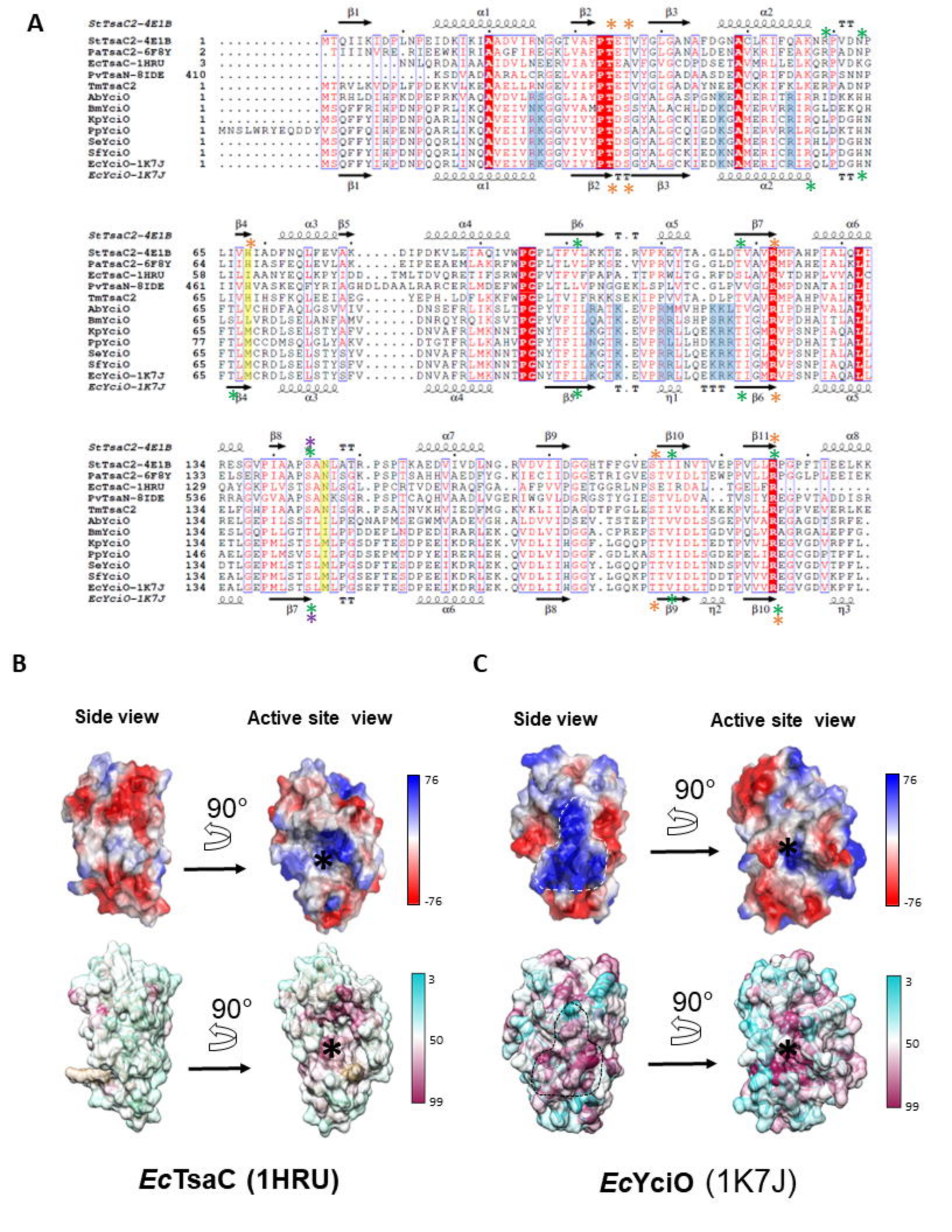

X, Y, M, and N proteins are homologs
Paralogs
Homologs
N
Y
X
M
Different neighborhood context of x, y, m and n
X
Y
M
N
Substrate docking
X
Y
M
N
Signature motif
Catalyst
Substrate binding
X xSGGKDSx...xLSGGx
X xSGGKDSx...xLSGGx
Y xSG-KDSx...xLAVGx
Y xSG-KDSx...xLAVGx
M xSGGLGDx...xLS-Gx
M xSGGLGDx...xLS-Gx
N xSGxGDSx...xLAGGx
N xSGxGDSx...xLAGGx
Re-annotation
Genes that co-occur with x, y, m and n are involved in different pathways
Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Species 1
Species 2
Species 7
Species 8
Species 9
Species 3
Species 4
Species 10
Species 11
(...)
absent
Pathway 1
A
B
C
D
E
F
Pathway 2
G
H
I
J
K
L
Pathway 3
O
P
Q
R
S
T
Predicted reaction
Multi-omics data integration
Transcripts
Proteins
Metabolites

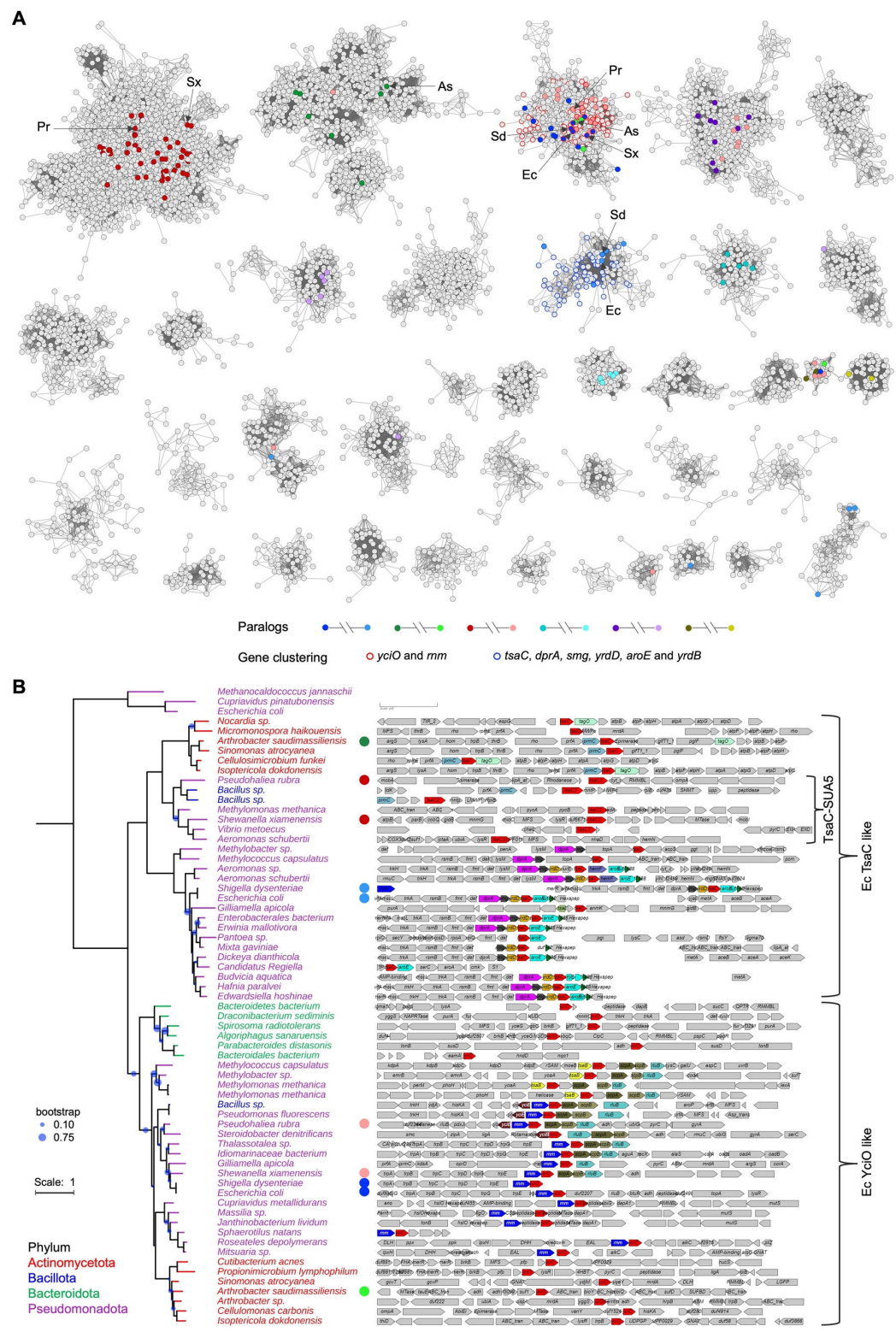

Input data
sequence/structure/
network etc.
Training dataset
Validation dataset
Query dataset
Training
Evaluation
Application
Trained predictive model
Output data
Pfam/Probability/GO/E
C/Structure etc.
Publications
200
150
100
50
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
ML approach
Labels available
Supervised Learning
Semi-supervised Learning
Unsupervised Learning
Models available
SVM
KNN
Decision Tree
Random Forest
GNN
CNN
RNN
Ensemble Methods
Transformer
Autoencoder
Software
SVMProt
GrAPFI
CrowdGO
ProFET
DeepFri
ProteInfer
DeepEC
ECPred
CLEAN
TALE
DeepECtransformer
ATGO
DeepSeqProt
Interpretability

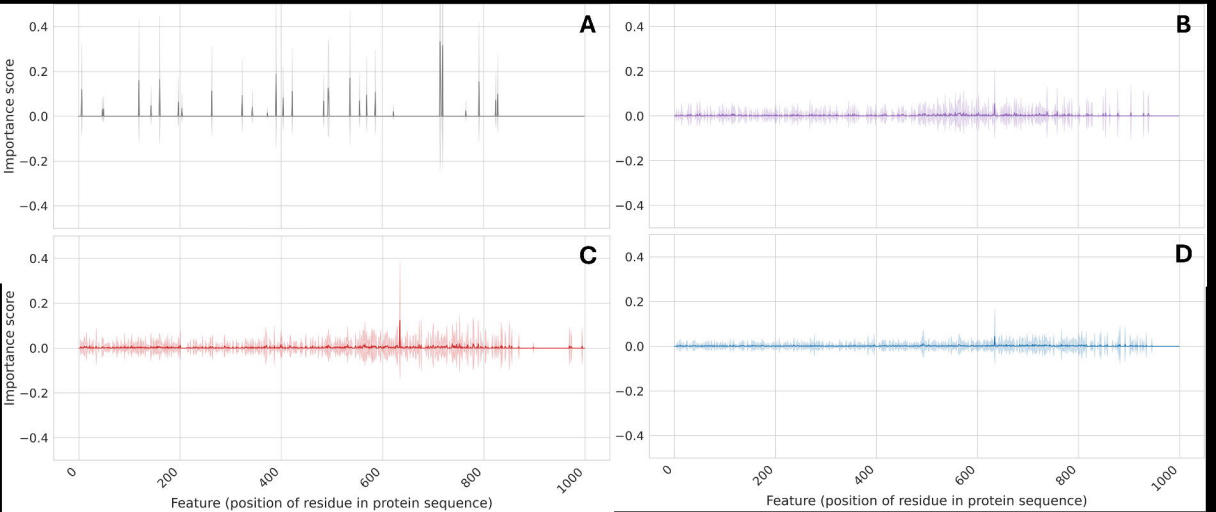

A
GloEC
ProteInfer
DeepEC
CLEAN
Macro-F1 Score
100
80
60
40
20
0
Carbohydrate esterase
New-432
COFACTOR-237
Phosphorylase
Datasets
B
New-392
Price
New-815
F1 score
0.8
0.6
0.4
0.2
0.0
BlastP
CNN
RNN
LSTM
DeepEC
DeepECTF
ProteInfer
TFPC
CLEAN
MAPred
Methods


No. of predictions
150
125
100
75
50
25
0
136
126
77
42
28
27
14
3
COR
UNC
LSP
CS 1
PLI
NPI
REP
CNN
CS 2
CS 0
3
14
27
42
70
77
126
136
166
245
250 200 150 100 50 0
No. of predictions


E. coli K12
proP
kdgT
yjcZ
cfrC
yjdM
phnC
phnE
phnG
phnI
phnJ
yjdN
phnD
phnF
phnH
Scale: 1kB
Pseudomonas tritici
phnK
proB
rRNA
protein
cgtA
Ppsynt
PPlase
Enterobacter agglomerans
nark/U
narG
narH
narI
tam
recA
narJ
mltB
dapA-like
cinA-like
recX
Bacillus cereus VD021
smi1-like
yutJ-like
arsR1-like
sulP-like
prsA
NTase
duf2628
ABC transporters
recB-like
dltC
try recombinase
Bacillus cereus VD142
duf2178
cas3
cas7
cas5


A
StTsaC2-4E1B
PaTsaC2-6F8Y
EcTsaC-1HRU
PvTsaN-8IDE
TmTsaC2
AbYciO
BmYciO
KpYciO
PpYciO
SeYciO
SfYciO
EcYciO-1K7J
EcYciO-1K7J
B
Side view
Active site view
90°
76
-76
3
50
99
EcTsaC (1HRU)
C
Side view
Active site view
90°
76
-76
3
50
99
EcYciO (1K7J)


A
Pr
Sx
As
Pr
Sd
As
Sx
Ec
Sd
Ec
Paralogs
Gene clustering
yciO and mm
tsaC, dprA, smg, yrdD, aroE and yrdB
B
Methanocaldococcus jannaschii
Cupriavidus pinatubonensis
Escherichia coli
Nocardia sp.
Micromonospora haikouensis
Arthrobacter saudimassiliensis
Sinomonas atrocyanea
Cellulosimicrobium funkei
Isoptericola dokdonensis
Pseudohaliea rubra
Bacillus sp.
Bacillus sp.
Methylomonas methanica
Shewanella xiamenensis
Vibrio metoecus
Aeromonas schubertii
Methylobacter sp.
Methylococcus capsulatus
Aeromonas sp.
Aeromonas schubertii
Shigella dysenteriae
Escherichia coli
Gilliamella apicola
Enterobacterales bacterium
Erwinia mallotivora
Pantoea sp.
Mixta gaviniae
Dickeya dianthicola
Candidatus Regiella
Budvicia aquatica
Hafnia paralvei
Edwardsiella hoshinae
Bacteroidetes bacterium
Draconibacterium sediminis
Spirosoma radiotolerans
Algoriphagus sanaruensis
Parabacteroides distasonis
Bacteroidales bacterium
Methylococcus capsulatus
Methylobacter sp.
Methylomonas methanica
Methylomonas methanica
Bacillus sp.
Pseudomonas fluorescens
Pseudohaliea rubra
Steroidobacter denitrificans
Thalassotalea sp.
Idiomarinaceae bacterium
Gilliamella apicola
Shewanella xiamenensis
Shigella dysenteriae
Escherichia coli
Cupriavidus metallidurans
Massilia sp.
Janthinobacterium lividum
Sphaerotilus natans
Roseateles depolymerans
Mitsuaria sp.
Cutibacterium acnes
Propionimicrobium lymphophilum
Sinomonas atrocyanea
Arthrobacter saudimassiliensis
Arthrobacter sp.
Cellulomonas carbonis
Isoptericola dokdonensis
bootstrap
0.10
0.75
Scale: 1
Phylum
Actinomycetota
Bacillota
Bacteroidota
Pseudomonadota
TsaC-SUA5
Ec TsaC like
Ec YciO like


A
B
C
D
Importance score
Importance score
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
Feature (position of residue in protein sequence)
Feature (position of residue in protein sequence)