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

Valérie de Crécy-Lagard[a,b*], Raquel Dias[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 gap prevents the biological community from fully leveraging the plethora of genomic data that is now available. Machine-learning approaches are showing some promise in propagating functional knowledge from experimentally characterized proteins to the correct set of isofunctional orthologs. However, they largely fail to predict enzymatic functions unseen in the training set, as shown by dissecting the predictions made for over 450 enzymes of unknown function from the model bacteria *Escherichia coli* using the DeepECTransformer platform. Lessons from these failures can help the community develop machine-learning methods that assist domain experts in making testable functional predictions for more members of the uncharacterized proteome.

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%, 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 from proteins of known functions to similar ones, but 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

Biology is now firmly in the genomic era. All fields, from ecology to agriculture to health care, rely on information derived from genome sequences to advance[1–5]. Proteins act as the executors of the genomic blueprint. Hence, the genomic revolution can only reach its full potential if accurate and precise descriptions of all known protein functions in forms that both machines and humans can exploit are captured and made freely available[5–7]. This is far from an easy task, and nearly thirty years after the first bacterial genome was sequenced, the functional annotation status of the proteome of most species is far from accurate and complete even for model organisms[8–11].

Predicting protein function involves two conceptually different challenges that are often confounded (Fig. 1). The first challenge is to correctly annotate the corpus "knowns" by capturing the existing protein functional knowledge and propagating it to the correct set of isofunctional orthologous proteins[12–14]. The second is to predict the function of the "unknowns," or proteins from families with no members that are linked to an experimentally validated function[15,16]. Estimates suggest that 30 to 70% of proteins in any given genome are "unknowns"[10,11,17,18]. Some of these "unknowns" must fill previously reported functional roles[7,15,19,20]. Others fill novel biological functions that have never been previously described. Addressing both challenges requires a combination of computational and experimental methods. For known proteins, the experiments come first, and the automatic propagation of functional knowledge follows [13]. For unknown proteins, the integration of multiple types of genomic and/or genome-wide experimental evidence can lead to predictions that can be tested experimentally[16].

The process of functional annotations of protein entries in databases starts with capturing information in the literature by biocurators[21]. This process links experimental characterizations of specific proteins in specific organisms to controlled vocabularies that describe validated functions such as the Gene Ontology (GO)[22], the IUPAC Enzyme Commission (EC) numbers (iubmb.qmul.ac.uk/enzyme), or biochemical reaction descriptors (e.g. Rhea[23]). Even though text-mining tools have accelerated the flow of information captured[24–26], very few of the proteins in UniprotKB[27], the most widely used protein function database[28], have been linked to experimental data[29]. Hence, most functions of proteins in Uniprot have been inferred computationally by sequence similarity[19].

The methods used over the past 30 years to automatically propagate functional annotations have been extensively reviewed[12,13,30–32]. From the very beginning, red flags were raised on

many sources of errors (listed in Table 1) that can occur in this process[33–39]. One prevalent error type affecting most protein families[39] is the over-annotation of paralogs[37,39,40] caused by the inherent functional complexity of enzyme families[41,42]. Fifty years of classification of evolutionary and structurally related superfamilies, have shown that functional diversification often occurs through duplication and divergence[43–45]. Closely related proteins can have different functions, resulting in a major source of annotation errors where proteins are assigned incorrect functional roles[37,39]. Methods based on phylogeny, active site analyses, metabolic reconstruction, and sequence similarity networks (SSNs) combined with gene neighborhood or co-expression information can be used to better separate functional subfamilies[46–48] (Fig. 2). When Hidden Markov Models (HMMs) or signature motifs are generated to separate the isofunctional subfamilies, they can be integrated into annotation pipelines such as RefSeq[49] or incorporated into rules such as the Unirules used by UniprotKB[50]. However, creating these HMMs or rules requires manual intervention akin to biocuration, which has not kept pace with the exponential increase of the sequenced proteome. Consequently, the number of over-propagation errors is increasing over time[39]. In summary, the two primary challenges that hinder a more comprehensive and accurate coverage of the functional space in sequenced proteomes are the incorrect annotation of paralogous subgroups within protein superfamilies and the difficulty in identifying the functions of truly unknown proteins.

Machine learning (ML)-based methods are currently revolutionizing many fields of biology[51,52] including prediction of protein properties such as 3D structure, ligand binding[53], molecular function[54], and function propagation accuracy[55–62]. In this work, we reviewed recent papers that utilize ML methods to predict Enzyme Commission (EC) numbers, evaluating how effectively these approaches address the challenges both in accurate annotation propagation and novel function discovery discussed above. We argue that while ML-based methods have shown steady improvement in correctly propagating EC numbers to uncharacterized proteins within the isofunctional protein space, they still fall short in predicting novel functions.

Methods

Bioinformatic analyses and data mining

The Uniprot (www.uniprot.org)[27], InterPro (www.ebi.ac.uk/interpro/)[63], NCBI (www.ncbi.nlm.nih.gov)[64], BioCyc (www.biocyc.org)[65] and KEGG[66] knowledgebases were routinely used in this study. PaperBlast was used to query literature (https://papers.genomics.lbl.gov/cgi-bin/litSearch.cgi)[67]. Crystal structures were retrieved from the Protein Data Bank (www.rcsb.org)[68] and visualized using Pymol (www.pymol.org).

Structure-based multi-sequence alignment derived from available crystal structures and AlphaFold structural models (https://alphafold.ebi.ac.uk/) were generated using PROMALS3D (prodata.swmed.edu/promals3d/promals3d.php)[69] and Espript (espript.ibcp.fr/ESPript/ESPript). The percent conservation scores were calculated using the program AL2CO (http://prodata.swmed.edu/al2co/al2co.php)[70] using 4264 TsaC/Sua5 sequences and 1955 YciO sequences (Supplemental data 1 and 2). Sequence Similarity Networks (SSNs) and the corresponding Gene Neighborhood Networks (GNT) were generated using EFI-EST (EFI Enzyme Similarity Tool, efi.igb.illinois.edu/efi-est)[71]. 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 figure legend and visualized using Cytoscape (3.10.1)[72]. 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/)[73]. To generate the phylogenetic tree, proteins were aligned using MUSCLE[74]. The alignment was trimmed using BMGE[75] and used to build the maximum likelihood tree using FastTree[76] with LG+CAT model with bootstrap (1000 replicates) and visualized using iTOL[77]. Open AI ChatGPT. 4.0 was used to extract data and generate figures from publications (https://chatgpt.com/).

Results

Recent progress and current limitations in using machine learning models to accurately propagate EC numbers to the correct portion of the protein sequence space.

There has been a recent explosion of publications reporting the use of pre-trained Protein Language Models (PLM) to predict protein functions[58–62,78–80] (Fig. 3). Fueled by the availability of PLMs from UniProt, ESM Atlas, and others (see references in Fig. 3 legend), models have been developed to link protein sequences to vocabularies such as GO terms [57,81,82] or to EC numbers[83].

It is difficult to keep up with the avalanche of these PLM-based EC-prediction studies but we extracted comparisons of tools and database metrics from recent manuscripts describing new EC prediction tools: 1) GloEC that has added the EC number hierarchical graph organization as

part of the training information[84]; 2) MAPred that integrates 3D tokens of proteins, to capture essential local functional sites[85]. This summary reveals the limitations of current ML-based EC prediction tools (Fig. 4). First, the results are dependent on the methods and datasets used ranging from 10 to 80% in F1 scores that represent a combination of the precision and recall metrics. When data sets are based on well-characterized proteins with structures (cofactor-237) or small datasets with well-separated functional groups (phosphorylase), the newest methods such as CLEAN or GloEC are quite effective in reaching macro-F1 scores of 60 to 75% (Fig. 4A). Surprisingly BlastP still performs better than most PLM-based predictors, even if there is a clear improvement in a few of the more recent ones (CLEAN, MAPred). The carbohydrate dataset comprises ~350 glycosyl hydrolase, highly similar proteins with different substrate specificities covering 7 EC numbers [84]. None of the EC prediction methods tested performed well on this dataset with F1 scores ranging from 6 to 30%, emphasizing the difficulty of automatic annotation of closely related enzyme subfamilies. Similarly, the EC numbers from the Price dataset that cover 100 enzymes that were annotated using high throughput phenotypic data and were not in any training sets[86] are very poorly covered with the best methods barely reaching 50% F1 scores (Fig. 4B).

As highlighted in this analysis as well as in the CAFA challenges results[57], the quality of the benchmark sets, annotation databases, and ontologies impact model outcomes. Efforts to improve the quality and consistency of training data and avoid data leakage between training and testing data can help to minimize annotation errors. Also, models should always provide the probabilities or confidence levels for predicted protein functions, instead of just binary labels, to provide transparency into the model's confidence. By providing probability estimates (or confidence scores) along the predicted functions, users can better interpret the model's outputs and reliability, guiding further experimental validation efforts. Many advanced machine learning models, particularly deep neural networks, function as ‘black boxes’, where the internal decision-making logic is difficult for humans to interpret (**Fig. 3**). Explainable AI (XAI) techniques can help to dissect ‘black-box’ models and understand what input features are driving correct and incorrect predictions[87]. Recent models like MAPred implement feature-dropping approaches that represent a step towards XAI, providing insights into feature importance and model behavior, albeit in a more targeted manner than comprehensive XAI techniques[85]. While not a full-fledged XAI method, MAPred's approach aligns with the goals of interpretability and transparency in machine learning models, offering a practical compromise between model performance and explainability.

Based on the precedent of the "AlphaFold moment" that occurred for structure predictions[88], as training data improves, reducing bias caused by oversampling of model organisms and overrepresented functions[89] and other types of evidence such as structural data to identify active site signature residues [17,58,90], gene neighborhoods[79,91] or chemical reactions to better separate non-isofunctional paralog subgroups[62,70] are integrated, PLM-driven approaches could become mainstream tools to help correctly propagate known functional knowledge among isofunctional proteins.

By design, supervised ML-models cannot be used to predict the function of true unknowns.

Even if computational models can help automate the accurate propagation of known functions, they are still far from predicting the functions of true unknowns as discussed by the Bromberg group[59]. Indeed, these models are not, in essence, designed to discover novelty, as they primarily rely on training sets derived from existing knowledge. If a protein is truly of unknown function, then the EC number that described its function might not exist yet, and if it does as over 600 EC numbers are yet to be linked to a protein sequence[9,92], this EC number would, by definition, not be linked to a protein in the training set. If most groups that use ML focus only on the propagation of EC numbers and understand these limitations[93], the study by Kim et al.[60] claimed to label true "unknowns" with substrate-level specificity enzymatic roles using the supervised ML-based DeepECTransformer (DeepECTF) platform. Several independent studies agree that the number of proteins in the model gram-negative *E. coli* not linked to precise molecular and/or biological functions is around 1200 to 1400 hundred or a quarter of the encoded proteins[11]. The Kim et al. study has labeled 453 of those with a four-digit EC number using the DeepECTF platform and validated three of these predictions in vitro[60]. Because of the inherent logical flaw in the approach (how could proteins of novel biological functions be annotated with existing labels), we evaluated these predictions in detail (Fig. 5 and Table S1).

We first found that for around a quarter of these predictions (Fig. 5), the same EC number was already present in the Uniprot annotation (Fig. 5 and Table S1A), or the same function was present without an EC number (Fig. 5 and Table S1BC). These proteins were not of unknown function and should not have been included in the "unknown" test set as they were most certainly part of the training set. 72 DeepECTF predictions differ from the existing EC number in UniProt by the last digit(s) (Table S2B) and around half of the predictions (261) had no EC number in the corresponding Uniprot entry (Table S2C). We analyzed all these predictions

manually by combining PaperBlast searches and Uniprot and EcoCyc data analyses. We grouped them in different categories with confidence scores (Table S1 and Fig. 5). Three cases including YgfF discussed below, were validated by publications not captured in Uniprot and can be considered successful predictions. Twelve cases gave more generic annotations (less precise) than the existing Uniprot annotations and were discarded and 14 predictions could neither be validated nor refuted (Fig. 5 and Table S1). We could refute 142 DeepECTF predictions through literature or biological context analyses with around half of these cases being paralog over propagation mistakes (Fig. 5) including the case of YciO discussed more specifically below. Specific examples where available literature invalidates the predictions 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 (BioCyc[65] 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[94]. Other obvious errors were the repetition of identical EC numbers (up to 12 times, Table S2C). This was by far the most prevalent error with 147 suspected cases (Fig. 5 and Table S1BC). This type of error could be due to inherent issues with how AI methods work. Indeed, if the input features in the training data lack sufficient information to distinguish between different functions, the model may struggle to make diverse predictions and regress to assigning the same frequent labels seen during training. Biases, data imbalance (underrepresented domains, motifs, and functions), lack of relevant features, architectural limitations (e.g. incapacity of capturing complex patterns, inability to integrate diverse data sources, lack of regularization, etc.), and poor uncertainty calibration could potentially cause models to “force” common labels from the training data when they cannot make an informed prediction for a given protein[95–98].

Integrating different types of evidence can validate or invalidate ML-based predictions

Three predictions made using the DeepECTF platform were tested using *in vitro* assays. None were tested using genetic validation. It is well established that in vitro activity alone is not sufficient to validate the function of protein *in vivo*[99,100]. Indeed, enzymes evolve often by duplication and speciation. *In vitro* activities of many enzymes can show promiscuity and if this can be useful for biotechnological applications[101] it does not always reflect the *in vivo* function[102]. Best practices in functional annotations of enzymes strive to provide both *in vitro* (biochemical) *and in vivo* (genetic) evidence or to follow the Gene Ontology Consortium

definitions, to capture both the molecular and biological functions[103]. We therefore evaluated the biological context of the predictions made for three *E. coli* unknowns YgfF, YdjM and YciO that were validated with *in vitro* enzymatic assays in the Kim et al. study.

YgfF is a member of the large Short-Chain Dehydrogenase/Reductase (SDR) superfamily (IPR002347). The Oppermann and Persson groups developed a nomenclature system and HMMs based classification website (http://www.sdr-enzymes.org/) that distinguish different functional subgroups of the SDR superfamily[104,105]. This resource predicts YgfF is part of the SDR63C/Glucose 1-dehydrogenase subgroup, the exact activity predicted and validated in the Kim et al.[60] study. Hence this prediction falls more in the category of more accurately propagating knowns and not of predicting novel functions but still can be viewed as a success of the DeepECTF platform.

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) but the initial publication that claimed this gene was involved in phosphonate catabolism was later refuted[106]. 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[107]. Gene neighborhoods of *phnA* genes show strong clustering with genes encoding phosphonoacetate transporters, phosphonoacetate sensing regulators, and in some cases 2-aminoethylphosphonate catabolism[108]. Still, 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 by other types of evidence, and additional *in vivo* experiments are required to confirm the biological role of this enzyme.

Finally, YciO is a member of the same Pfam family (PF01300) 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)[109,110]. The function of TsaC/Sua5 was first elucidated in 2009 [111]. It was already noted at the time that YciO was a paralog that did not perform the same function *in vivo*[111,112], even though 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). Gene neighborhood and structural data suggest that the actual function of

YciO could be related to rRNA rather than tRNA metabolism: 1) the structure of YciO exhibits a broad positively charged surface predicted to interact with RNA ([113] and Fig. 7B and C). This 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* [114]. Kim et al. predicted that YciO had the same function as TsaC and showed that it catalyzed the synthesis of L-threonylcarbamoyladenylate from ATP, L-threonine, and bicarbonate, *in vitro*[60]. 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[115], consistent with the possibility of a missing partner or a different biological substrate for YciO. Models of enzyme evolution go through promiscuous stages[48]. TsaC is an enzyme predicted to have been present in the Last Universal Common Ancestor (LUCA)[109,116]. Hence, as YciO is a clear paralog of TsaC (Fig. 7) that must have evolved more recently, it is not surprising that residual ancestral catalytic activity can still 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 predicted by DeepECTF (2.7.7.87) is most certainly incorrect.

Conclusions

As discussed above, ML approaches can be used to improve the propagation of EC numbers within isofunctional groups starting with an experimentally validated input[90]. One can also envision that future ML approaches could allow to solve some of the cases where an existing EC number has never been linked to a gene as some of the gap-filling studies have successfully done[117,118]. However, by essence, they cannot provide 4-digit EC numbers that do not yet exist to true unknowns. The great majority (>90%) of the predictions claiming to solve true unknown case made in the Kim et al. study using the DeepECTF platform can be proven to be erroneous (>60 %) or not novel (30%) (Fig. 5). In the few cases where AI tools have successfully predicted novel functions unsupervised learning techniques that can cluster proteins into isofunctional groups without relying on existing annotations[119] were used. In addition, multiple types of evidence were incorporated and reviewed by experts who brought their biochemical expertise and insights[62,120]. More generally, biological common sense should be used to check AI-generated predictions until their reliability increases.

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 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[86].
2 Naming issues	Inconsistent naming of the same entities.	See GroEL in [121].
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[122] 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[123]. See also examples in[86].
4 Curation mistake	1) The data was incorrectly captured by biocurator; 2) Functional annotations may become outdated.	Ureidoglycolate lyase [38].
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[124].
6 Over-annotation of paralogs	Annotation wrongly propagated to non-isofunctional paralogous groups.	See examples in [37,39,40].

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[125].
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[126].
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[127]
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

					activity [128]
b1267	None	Uncharacterized protein YciO	2.7.7.87	L-threonylcarbamoyladenylate synthase	YciO is a paralog of TsaC but does catalyze the same activity [111,112].
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[129].
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[130].
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[94].
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 the 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. We thank Alan J. Bridge (SIB) for critical reading and input on early versions of 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 and IF wrote sections of the manuscript. All authors edited the final draft.

References

(1) Pourkheirandish, M., Golicz, A. A., Bhalla, P. L., Singh, M. B. Global role of crop genomics in the face of climate change. *Front Plant Sci* **2020**, *11*, 922. https://doi.org/10.3389/fpls.2020.00922

(2) Ungerer, M. C., Johnson, L. C., Herman, M. A. Ecological genomics: understanding gene and genome function in the natural environment. *Heredity (Edinb)* **2008**, *100*(2), 178–183. https://doi.org/10.1038/sj.hdy.6800992

(3) Forrester, S. J., Hall, N. The revolution of whole genome sequencing to study parasites. *Mol Biochem Parasitol* **2014**, *195*(2), 77–81. https://doi.org/10.1016/j.molbiopara.2014.07.008

(4) Stephan, T., Burgess, S. M., Cheng, H., et al. Darwinian genomics and diversity in the tree of life. *Proc Natl Acad Sci U S A.* **2022**, *119*(4), e2115644119. https://doi.org/10.1073/pnas.2115644119

(5) Kobras, C. M., Fenton, A. K., Sheppard, S. K. Next-generation microbiology: from comparative genomics to gene function. *Genome Biol* **2021**, *22*(1), 123. https://doi.org/10.1186/s13059-021-02344-9

(6) Danchin, A., Ouzounis, C., Tokuyasu, T., Zucker, J. D. No wisdom in the crowd: genome annotation in the era of big data – current status and future prospects. *Microbial Biotechnology.* **2018**, *11*(4), 588–605. https://doi.org/10.1111/1751-7915.13284.

(7) Karp, P. Call for an Enzyme Genomics Initiative. *Genome Biol* **2004**, *5*(8), 401. https://doi.org/doi: 10.1186/gb-2004-5-8-401

(8) Rocha, J. J., Jayaram, S. A., Stevens, T. J., et al. Functional unknomics: systematic screening of conserved genes of unknown function. *PLoS Biol* **2023**, *21*(8), e3002222. https://doi.org/10.1371/journal.pbio.3002222

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

(10) Wood, V., Lock, A., Harris, M. A., Rutherford, K., Bähler, J., Oliver, S. G. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol* **2019**, *9*(2), 180241. https://doi.org/10.1098/rsob.180241

(11) Ghatak, S., King, Z. A., Sastry, A., Palsson, B. O. The y-ome defines the 35% of *escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Res* **2019**, *47*(5), 2446–2454. https://doi.org/10.1093/nar/gkz030

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

(13) Lee, D., Redfern, O., Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **2007**, *8*(12), 995. https://doi.org/10.1038/nrm2281

(14) Koo, D. C. E., Bonneau, R. Towards region-specific propagation of protein functions. *Bioinformatics* **2019**, *35*(10), 1737–1744. https://doi.org/10.1093/bioinformatics/bty834

(15) Niehaus, T. D., Thamm, A. M. K., de Crécy-Lagard, V., Hanson, A. D. Proteins of unknown biochemical function: a persistent problem and a roadmap to help overcome it. *Plant Physiol* **2015**, *169*(3), 1436–1442. https://doi.org/10.1104/pp.15.00959

(16) Hanson, A. D., Pribat, A., Waller, J. C., de Crécy-Lagard, V. “Unknown” proteins and “orphan” enzymes: the missing half of the engineering parts list--and how to find it. *Biochem J* **2009**, *425*(1), 1–11. https://doi.org/10.1042/BJ20091328

(17) Sajid, S., Mashkoor, M., Jørgensen, M. G., et al. The y-ome conundrum: insights into uncharacterized genes and approaches for functional annotation. *Mol Cell Biochem* **2023**. *479*(8),1957-1968. https://doi.org/10.1007/s11010-023-04827-8

(18) Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G., Doxey, A. C. An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* **2020**, *6*(3), e000341. https://doi.org/10.1099/mgen.0.000341

(19) Lespinet, O., Labedan, B. Puzzling over orphan enzymes. *Cell Mol Life Sci* **2006**, *63*(5), 517–523. https://doi.org/ 10.1007/s00018-005-5520-6

(20) Chen, L., Vitkup, D. Distribution of orphan metabolic activities. *Trends Biotechnol* **2007**, *25*(8), 343–348. https://doi.org/10.1016/j.tibtech.2007.06.001

(21) International Society for Biocuration. Biocuration: Distilling data into knowledge. *PLoS Biol* **2018**, *16*(4), e2002846. https://doi.org/10.1371/journal.pbio.2002846

(22) Gene Ontology Consortium., Aleksander, S. A., Balhoff, J., et al. The Gene Ontology knowledgebase in 2023. *Genetics* **2023**, *224*(1), iyad031. https://doi.org/10.1093/genetics/iyad031

(23) Bansal, P., Morgat, A., Axelsen, K. B., et al. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* **2022**, *50*(D1), D693–D700. https://doi.org/10.1093/nar/gkab1016

(24) Poux, S., Arighi, C. N., Magrane, M., et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **2017**, *33*(21), 3454–3460. https://doi.org/10.1093/bioinformatics/btx439

(25) Wei, C.-H., Allot, A., Leaman, R., Lu, Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* **2019**, *47*(W1), W587–W593. https://doi.org/10.1093/nar/gkz389

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

(27) Bateman, A., Martin, M.-J., Orchard, S., et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

(28) Ramola, R., Friedberg, I., Radivojac, P. The field of protein function prediction as viewed by different domain scientists. *Bioinformatics Advances* **2022**, *2*(1), vbac057. https://doi.org/10.1093/bioadv/vbac057

(29) Škunca, N., Roberts, R. J., Steffen, M. Evaluating Computational Gene Ontology Annotations. *Methods Mol Biol .* **2017**, 1446, 97–109. https://doi.org/10.1007/978-1-4939-3743-1_8

(30) Satish Kumar, V., Dasika, M. S., Maranas, C. D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **2007**, *8*(1), 212. https://doi.org/10.1186/1471-2105-8-212

(31) Storm, C. E., Sonnhammer, E. L. L. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **2002**, *18*, 92–99. https://doi.org/10.1093/bioinformatics/18.1.92

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

(33) Brenner, S. E. Errors in Genome Annotation. *Trends in Genetics* **1999**, *15*(4), 132–133. https://doi.org/10.1016/s0168-9525(99)01706-0

(34) Devos, D., Valencia, A. Intrinsic errors in genome annotation. *Trends in Genetics* **2001**, *17*(8), 429–431. https://doi.org/10.1016/s0168-9525(01)02348-4

(35) Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S., Ouzounis, C. A. modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **2002**, *18*, 1641–1649. https://doi.org/10.1093/bioinformatics/18.12.1641

(36) Galperin, M. Y., Koonin, E. V. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* **1998**, *1*(1), 55–67

(37) Schnoes, A. M., Brown, S. D., Dodevski, I., Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **2009**, *5*(12), e1000605. https://doi.org/10.1371/journal.pcbi.1000605

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

(39) Rembeza, E., Engqvist, M. K. M. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. *PLoS Comput Biol* **2021**, *17*(9), e1009446. https://doi.org/10.1371/journal.pcbi.1009446

(40) Zallot, R., Harrison, K. J., Kolaczkowski, B., de Crécy-Lagard, V. functional annotations of paralogs: a blessing and a curse. *Life* **2016**, *6*(3). https://doi.org/10.3390/life6030039

(41) Glasner, M. E., Gerlt, J. A., Babbitt, P. C. Evolution of enzyme superfamilies. *Curr Opin Chem Biol* **2006**, *10*(5), 492–497. https://doi.org/doi: 10.1016/j.cbpa.2006.08.012

(42) Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., Almo, S. C. Divergent evolution in enolase superfamily: strategies for assigning functions. *J Biol Chem* **2012**, *287*(1), 29–34. https://doi.org/10.1074/jbc.R111.240945

(43) Todd, A. E., Orengo, C. A., Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **2001**, *307*(4), 1113–1143. https://doi.org/10.1006/jmbi.2001.4513

(44) Bordin, N., Sillitoe, I., Lees, J. G., Orengo, C. Tracing evolution through protein structures: nature captured in a few thousand folds. *Front Mol Biosci* **2021**, *8*, 668184. https://doi.org/10.3389/fmolb.2021.668184

(45) Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., Orengo, C. A. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* **2015**, *31*(21), 3460-7. https://doi.org/10.1093/bioinformatics/btv398

(46) Escudeiro, P., Henry, C. S., Dias, R. P. M. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Curr Res Microb Sci* **2022**, 3, 100159. https://doi.org/10.1016/j.crmicr.2022.100159

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

(48) Ribeiro, A. J. M., Riziotis, I. G., Borkakoti, N., Thornton, J. M. Enzyme function and evolution through the lens of bioinformatics. *Biochemical Journal* **2023**, *480*(22), 1845–1863. https://doi.org/10.1042/BCJ20220405

(49) Li, W., O'Neill, K. R., Haft, D. H., et al. RefSeq: Expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res* **2021**, *49*(D1), D1020–D1028. https://doi.org/10.1093/nar/gkaa1105

(50) MacDougall, A., Volynkin, V., Saidi, R., et al. UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* **2020**, *36*(17), 4643–4648. https://doi.org/10.1093/bioinformatics/btaa485

(51) Sapoval, N., Aghazadeh, A., Nute, M. G., et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* **2022**, *13*(1), 1728. https://doi.org/10.1038/s41467-022-29268-7

(52) Bordin, N., Dallago, C., Heinzinger, M., et al. Novel Machine Learning Approaches Revolutionize Protein Knowledge. *Trends Biochem Sci* **2023**, *48*(4), 345–359. https://doi.org/10.1016/j.tibs.2022.11.001

(53) Abramson, J., Adler, J., Dunger, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold□3. *Nature* **2024**. https://doi.org/10.1038/s41586-024-07487-w.

(54) Chandra, A., Tünnermann, L., Löfstedt, T., Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife* **2023**, *12*. https://doi.org/10.7554/eLife.82819

(55) Scheibenreif, L., Littmann, M., Orengo, C., Rost, B. FunFam protein families improve residue level molecular function prediction. *BMC Bioinformatics* **2019**, *20*(1), 400. https://doi.org/10.1186/s12859-019-2988-x

(56) Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., Rost, B. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom Bioinform* **2022**, *4*(2), lqac043. https://doi.org/10.1093/nargab/lqac043

(57) Zhou, N., Jiang, Y., Bergquist, T. R., et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through

experimental screens. *Genome Biol* **2019**, *20*(1), 244. https://doi.org/10.1186/s13059-019-1835-8

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

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

(60) Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., Lee, S. Y. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat Commun* **2023**, *14*(1), 7370. https://doi.org/10.1038/s41467-023-43216-z

(61) Durairaj, J., Waterhouse, A. M., Mets, T., et al. Uncovering new families and folds in the natural protein universe. *Nature* **2023**, *622*(7983), 646–653. https://doi.org/10.1038/s41586-023-06622-3

(62) Rodríguez del Río, Á., Giner-Lamia, J., Cantalapiedra, C. P., Botas, J., Deng, Z., Hernández-Plaza, A., Munar-Palmer, M., Santamaría-Hernando, S., Rodríguez-Herva, J. J., Ruscheweyh, H.-J., Paoli, L., Schmidt, T. S. B., Sunagawa, S., Bork, P., López-Solanilla, E., Coelho, L. P., Huerta-Cepas, J. Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **2024**, *626*, 377–384. https://doi.org/10.1038/s41586-023-06955-z

(63) Paysan-Lafosse T, Blum M, Chuguransky S, et al. InterPro in 2022. *Nucleic Acids Res* **2023**, *51*(D1), D418-D427. https://doi.org/10.1093/nar/gkac993

(64) NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **2016**, *44*(D1), D7–D19. https://doi.org/10.1093/nar/gkv1290

(65) Karp, P. D., Billington, R., Caspi, R., et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* **2019**, *20*(4), 1085–1093. https://doi.org/10.1093/bib/bbx085

(66) Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **2021**, *49*(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

(67) Price, M. N., Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2017**, *2*(4), e00039-17. https://doi.org/10.1128/mSystems.00039-17

(68) Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., et al. 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* **2022**, *51*(D1), D488-D508. https://doi.org/10.1093/nar/gkac1077

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

(70) Pei, J., Grishin, N. V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **2001**, *17*(8), 700–712. https://doi.org/10.1093/bioinformatics/17.8.700

(71) Zallot, R., Oberg, N., Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, *58*(41), 4169–4182. https://doi.org/10.1021/acs.biochem.9b00735

(72) Shannon, P., Markiel, A., Ozier, O., et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003**, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

(73) Harrison, K. J., de Crécy-Lagard, V., Zallot, R. Gene Graphics: a genomic neighborhood data visualization web application. *Bioinformatics* **2018**, *34*(8), 1406–1408. https://doi.org/10.1093/bioinformatics/btx793

(74) Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

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

(76) Price, M. N., Dehal, P. S., Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **2010**, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

(77) Letunic, I., Bork, P. Interactive Tree of Life (ITOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **2021**, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301.

(78) Ardern, Z., Chakraborty, S., Lenk, F., Kaster, A.-K. Elucidating the functional roles of prokaryotic proteins using big data and artificial intelligence. *FEMS Microbiol Rev* **2023**, *47*(1). https://doi.org/10.1093/femsre/fuad003

(79) Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., Girguis, P. R. Genomic language model predicts protein co-regulation and function. *Nat Commun* **2024**, *15*(1), 2880. https://doi.org/10.1038/s41467-024-46947-9

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

(81) Radivojac, P., Clark, W., Oron, T., Schnoes, A., Wittkop, T., Sokolov, A. A large-scale evaluation of computational protein function prediction. *Nat Methods* **2013**, *10*(2), e0147952. https://doi.org/10.1038/nmeth.2340

(82) Jiang, Y., Oron, T. R., Clark, W. T., et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* **2016**, *17*(1), 184. https://doi.org/10.1186/s13059-016-1037-6

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

(84) Huang, Y., Lin, Y., Lan, W., Huang, C., Zhong, C. GloEC: a hierarchical-aware global model for predicting enzyme function. *Brief Bioinform* **2024**, *25*(5), bbae365. https://doi.org/10.1093/bib/bbae365

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

(86) Price, M. N., Arkin, A. P. Interactive tools for functional annotation of bacterial genomes. *bioRxiv* **2024**, 2024.04.15.589591. https://doi.org/10.1101/2024.04.15.589591

(87) Chennam, K. K., Mudrakola, S., Maheswari, V. U., Aluvalu, R., Rao, K. G. Black Box Models for eXplainable Artificial Intelligence. In: Mehta, M., Palade , V., Chatterjee, I. (eds) Explainable AI: Foundations, Methodologies and Applications. Intelligent Systems Reference Library. Springer, Cham **2023**, *232*. https://doi.org/10.1007/978-3-031-12807-3_1

(88) Laurents, D. V. AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function. *Front Mol Biosci* **2022**, *9*. https://doi.org/10.3389/fmolb.2022.906437

(89) Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., Friedberg, I. Biases in the experimental annotations of protein function and their effect on our understanding of

protein function space. *PLoS Comput Biol* **2013**, *9*(5), e1003063. https://doi.org/10.1371/journal.pcbi.1003063

(90) Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., Zhao, H. Enzyme function prediction using contrastive learning. *Science* **2023**, *379*(6639), 1358–1363. https://doi.org/10.1126/science.adf2465

(91) Urhan, A., Cosma, B.-M., Manson, A. L., Abeel, T. SAFPred: synteny-aware gene function prediction for bacteria using protein embeddings. *Bioinformatics* **2024**, *40*(6), btae328. https://doi.org/10.1093/bioinformatics/btae328

(92) Siew, N., Azaria, Y., Fischer, D. The ORFanage: an ORFan database. *Nucleic Acids Res* **2004**, *32*, D281-3. https://doi.org/10.1093/nar/gkh116

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

(94) Zallot, R., Yuan, Y., de Crécy-Lagard, V. The *Escherichia coli* COG1738 Member YhhQ Is Involved in 7-Cyanodeazaguanine (preQ$_0$) Transport. *Biomolecules* **2017**, *7*(1), 12. https://doi.org/10.3390/biom7010012

(95) Mardikoraem, M., Woldring, D. Protein fitness prediction is impacted by the interplay of language models, ensemble learning, and sampling methods. *Pharmaceutics* **2023**, *15*(5), 1337. https://doi.org/10.3390/pharmaceutics15051337

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

(97) Pucci, F., Bernaerts, K. V., Kwasigroch, J. M., Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **2018**, *34*(21), 3659–3665. https://doi.org/10.1093/bioinformatics/bty348

(98) Shahbazi, N., Lin, Y., Asudeh, A., Jagadish, H. V. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput Surv* **2023**, *55*(13s), 1–39. https://doi.org/10.1145/3588433

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

(100) García‐Contreras, R., Vos, P., Westerhoff, H. V., Boogerd, F. C. 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* **2012**, *279*(22), 4145–4159. https://doi.org/10.1111/febs.12007

(101) Robinson, S. L., Smith, M. D., Richman, J. E., Aukema, K. G., Wackett, L. P. Machine learning-based prediction of activity and substrate specificity for olea enzymes in the thiolase superfamily. *Synth Biol* **2020**, *5*(1) ysaa004. https://doi.org/10.1093/synbio/ysaa004

(102) Copley, S. D. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem Sci* **2015**, *40*(2), 72–78. https://doi.org/10.1016/j.tibs.2014.12.004.

(103) Ashburner, M., Ball, C. A., Blake, J. A., et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**, *25*, 25–29. https://doi.org/10.1038/75556

(104) Persson, B., Kallberg, Y., Bray, J. E., et al. The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem Biol Interact* **2009**, *178*(1–3), 94–98. https://doi.org/10.1016/j.cbi.2008.10.040

(105) Kallberg, Y., Oppermann, U., Persson, B. Classification of the short‐chain dehydrogenase/reductase superfamily using hidden markov models. *FEBS J* **2010**, *277*(10), 2375–2386. https://doi.org/10.1111/j.1742-4658.2010.07656.x

(106) Metcalf, W. W., Wanner, B. L. Evidence for a fourteen-gene, phnC to phnP locus for phosphonate metabolism in *Escherichia coli*. *Gene* **1993**, *129*(1), 27–32. https://doi.org/10.1016/0378-1119(93)90692-V

(107) Kulakova, A. N., Kulakov, L. A., P. Quinn, J. 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* **1997**, *195*(1), 49–53. https://doi.org/10.1016/S0378-1119(97)00151-0

(108) Kulakova, A. N., Kulakov, L. A., Akulenko, N. V., Ksenzenko, V. N., Hamilton, J. T. G., Quinn, J. P. Structural and functional analysis of the phosphonoacetate hydrolase (*phnA*) gene region in *Pseudomonas fluorescens* 23F. *J Bacteriol* **2001**, *183*(11), 3268–3275. https://doi.org/10.1128/JB.183.11.3268-3275.2001

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

(110) Su, C., Jin, M., Zhang, W. Conservation and Diversification of tRNA t^6A-Modifying Enzymes across the Three Domains of Life. *Int J Mol Sci* **2022**, *23*(21), 13600. https://doi.org/10.3390/ijms232113600

(111) El Yacoubi, B., Lyons, B., Cruz, Y., et al. The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. *Nucleic Acids Res* **2009**, *37*(9), 2894–2909. https://doi.org/10.1093/nar/gkp152

(112) Gerdes, S., El Yacoubi, B., Bailly, M., et al. Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genomics* **2011**, *12*(S1), S2. https://doi.org/10.1186/1471-2164-12-S1-S2

(113) Jia, J., Lunin, V. V., Sauve, V., Huang, L. W., Matte, A., Cygler, M. Crystal structure of the YciO protein from *Escherichia coli*. *Proteins* **2002**, *49*(1), 139-41. https://doi.org/ 10.1002/prot.10178

(114) Jain, C. RNase AM, a 5' to 3' exonuclease, matures the 5' end of all three ribosomal rnas in *E. coli*. *Nucleic Acids Res* **2020**, *48*(10). https://doi.org/10.1093/NAR/GKAA260

(115) Swinehart, W., Deutsch, C., Sarachan, K.L., et al. Specificity in the biosynthesis of the universal tRNA nucleoside N^6-threonylcarbamoyl adenosine (t^6A) - TsaD is the gatekeeper. *RNA* **2020**, *26*(9), 1094–1103. https://doi.org/10.1261/rna.075747.120

(116) Gagler, D. C., Karas, B., Kempes, C. P., et al. Scaling laws in enzyme function reveal a new kind of biochemical universality. *Proc Natl Acad Sci U S A* **2022**, *119*(9). https://doi.org/10.1073/pnas.2106655119

(117) Smith, A. A. T., Belda, E., Viari, A., Medigue, C., Vallenet, D. The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput Biol* **2012**, *8*(5), e1002540. https://doi.org/10.1371/journal.pcbi.1002540

(118) Chitale, M., Khan, I. K., Kihara, D. Missing gene identification using functional coherence scores. *Sci Rep* **2016**, *6*(1), 31725. https://doi.org/10.1038/srep31725

(119) Mahlich, Y., Zhu, C., Chung, H., et al. Learning from the unknown: exploring the range of bacterial functionality. *Nucleic Acids Res* **2023**, *51*(19), 10162–10175. https://doi.org/10.1093/nar/gkad757

(120) Dembech, E., Malatesta, M., De Rito, C., et al. Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc Natl Acad Sci U S A* **2023**, *120*(16), e2218329120. https://doi.org/10.1073/pnas.2218329120

(121) Lockwood, S., Brayton, K. A., Daily, J. A., Broschat, S. L. Whole Proteome Clustering of 2,307 Proteobacterial Genomes Reveals Conserved Proteins and Significant Annotation Issues. *Front Microbiol* **2019**, *10,* 383. https://doi.org/10.3389/fmicb.2019.00383

(122) de Crécy-Lagard, V., Phillips, G., Grochowski, L. L., et al. 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* **2012**, *7*(11). https://doi.org/10.1021/cb300342u

(123) de Crécy-Lagard, V. Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. *Comput Struct Biotechnol J* **2014**, *10*(16), 41–50. https://doi.org/10.1016/j.csbj.2014.05.008

(124) Reed, C. J., Hutinet, G., de Crécy-Lagard, V. Comparative Genomic Analysis of the DUF34 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase. *Biomolecules* **2021**, *11*(9), 1282. https://doi.org/10.3390/biom11091282

(125) Zhang, X., Carter, M. S., Vetting, M. W., et al. 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* **2016**, *113*(29), E4161-9. https://doi.org/10.1073/pnas.1605546113

(126) Purta, E., van Vliet, F., Tkaczuk, K. L., et al. The *yfhQ* gene of *Escherichia coli* encodes a trna:Cm32/Um32 methyltransferase. *BMC Mol Biol.* **2006**, *7*(1), 23. https://doi.org/10.1186/1471-2199-7-23

(127) Preumont, A., Snoussi, K., Stroobant, V., Collet, J. F., Van Schaftingen, E. Molecular identification of pseudouridine-metabolizing enzymes. *J Biol Chem* **2008**, *283*(37), 25238–25246. https://doi.org/10.1074/jbc.M804122200

(128) Sui, L., Warren, J. C., Russell, J. P. N., Stourman, N. V. Comparison of the functions of glutathionylspermidine synthetase/amidase from *E. coli* and its predicted homologues YgiC and YjfC. *Int J Biochem Mol Biol* **2012**, *3*(3), 302–312.

(129) Golovina, A. Y., Sergiev, P. V., Golovin, A. V., et al. The *yfiC* gene of *E. coli* encodes an adenine-N^6 methyltransferase that specifically modifies A37 of $tRNA_1^{Val}$(Cmo^5UAC). *RNA* **2009**, *15*(6), 1134–1141. https://doi.org/10.1261/rna.1494409

(130) Noguchi, S., Nishimura, Y., Hirota, Y., Nishimura, S. Isolation and characterization of an *Escherichia coli* mutant lacking tRNA-guanine transglycosylase. Function and biosynthesis of queuosine in tRNA. *J Biol Chem* **1982**, *257*(11), 6544–6550.

(131) Li, Y. H., Xu, J. Y., Tao, L., et al. SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity. *PLoS One* **2016**, *11*(8), e0155290. https://doi.org/10.1371/journal.pone.0155290

(132) Sarker, B., Ritchie, D. W., Aridhi, S. GrAPFI: predicting enzymatic function of proteins from domain similarity graphs. *BMC Bioinformatics* **2020**, *21*(1), 168. https://doi.org/10.1186/s12859-020-3460-7

(133) Reijnders, M. J. M. F., Waterhouse, R. M. CrowdGO: Machine learning and semantic similarity guided consensus gene ontology annotation. *PLoS Comput Biol* **2022**, *18*(5), e1010075. https://doi.org/10.1371/journal.pcbi.1010075

(134) Ofer, D., Linial, M. ProFET: Feature engineering captures high-level protein functions. *Bioinformatics* **2015**, *31*(21), 3429–3436. https://doi.org/10.1093/bioinformatics/btv345

(135) Gligorijević, V., Renfrew, P. D., Kosciolek, T., et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **2021**, *12*(1), 3168. https://doi.org/10.1038/s41467-021-23303-9

(136) Ryu, J. Y., Kim, H. U., Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci U S A* **2019**,*116*(28), 13996–14001. https://doi.org/10.1073/pnas.1821905116

(137) Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V., Doğan, T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. *BMC Bioinformatics* **2018**, *19*(1), 334. https://doi.org/10.1186/s12859-018-2368-y

(138) Cao, Y., Shen, Y. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics* **2021**, *37*(18), 2825–2833. https://doi.org/10.1093/bioinformatics/btab198

(139) Zhu, Y.-H., Zhang, C., Yu, D.-J., Zhang, Y. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. *PLoS Comput Biol* **2022**,*18*(12), e1010793. https://doi.org/10.1371/journal.pcbi.1010793

(140) David, K. T., Halanych, K. M. Unsupervised deep learning can identify protein functional groups from unaligned sequences. *Genome Biol Evol* **2023**, *15*(5), evad084. https://doi.org/10.1093/gbe/evad084

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 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 number of total 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 or 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.[84]); (B) Comparison of F1-scores generated using eight methods including BlastP, DeepECTF, DeepEC, ProteinInfer, and CLEAN on three datasets (New-815, New-392 and Price)(data extracted from [85]). Dotted lines mark the BlastP scores. See initial publications for details and references for methods and datasets.

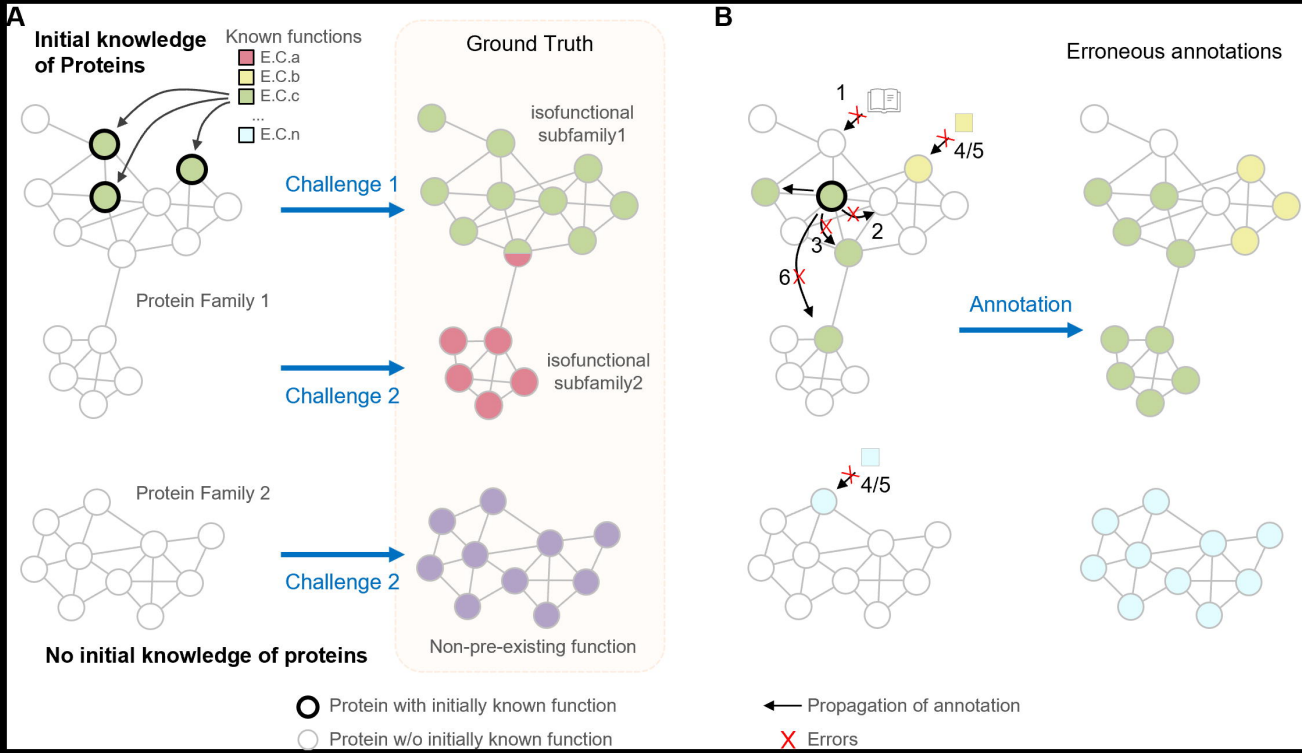
Figure 5. **Classification of DeepECTF predictions**. The 453 EC number predictions for 453 *E. coli* unknowns were manually classified in categories and labeled with confidence scores from 0 to 3. Data extracted from **Table S1.**

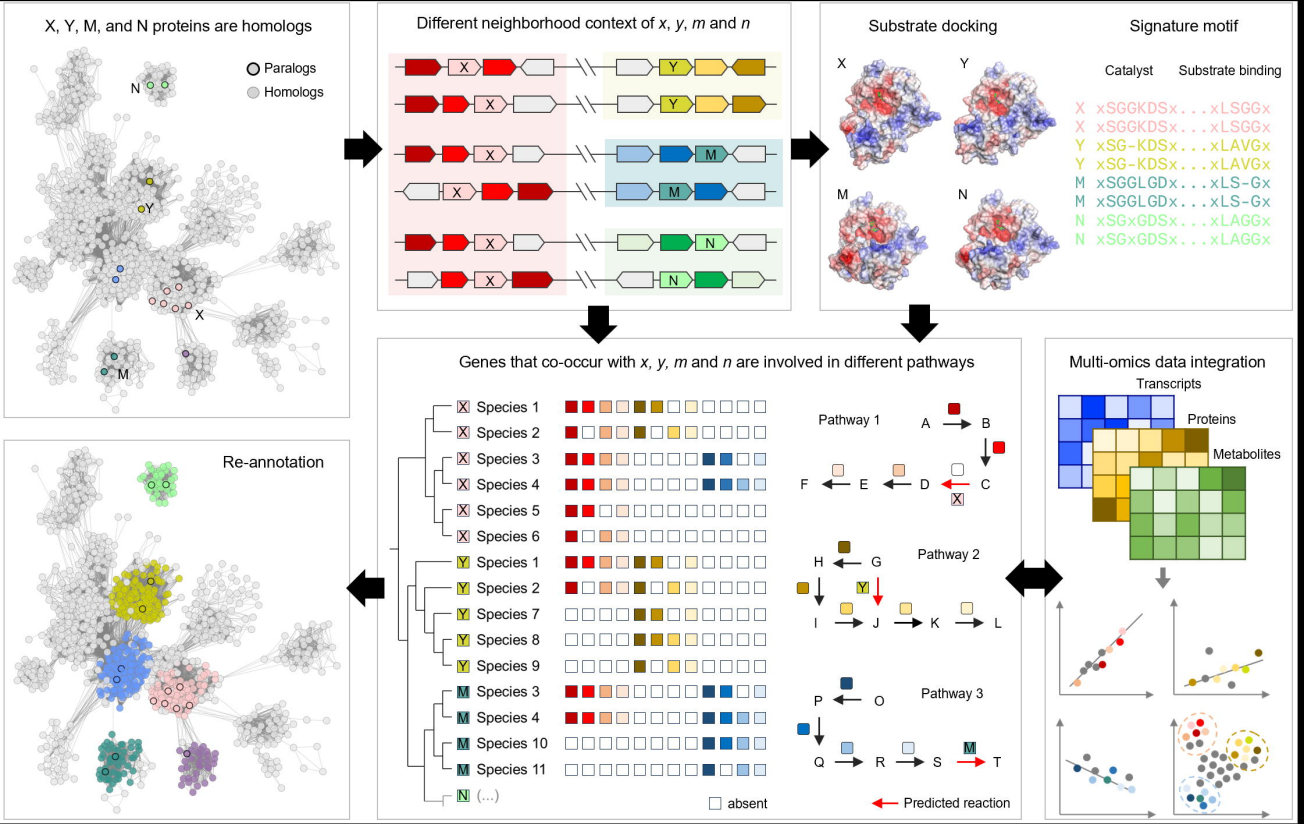
Figure 6. **Gene neighborhoods and metabolic reconstructions do not link YdjM to phosphonate degradation**. Most *yjdM* genes are not near phosphonate degradation operons. A Sequence Similarity Network of 11,986 IPR004624 family members was generated, and the corresponding genome neighborhood information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

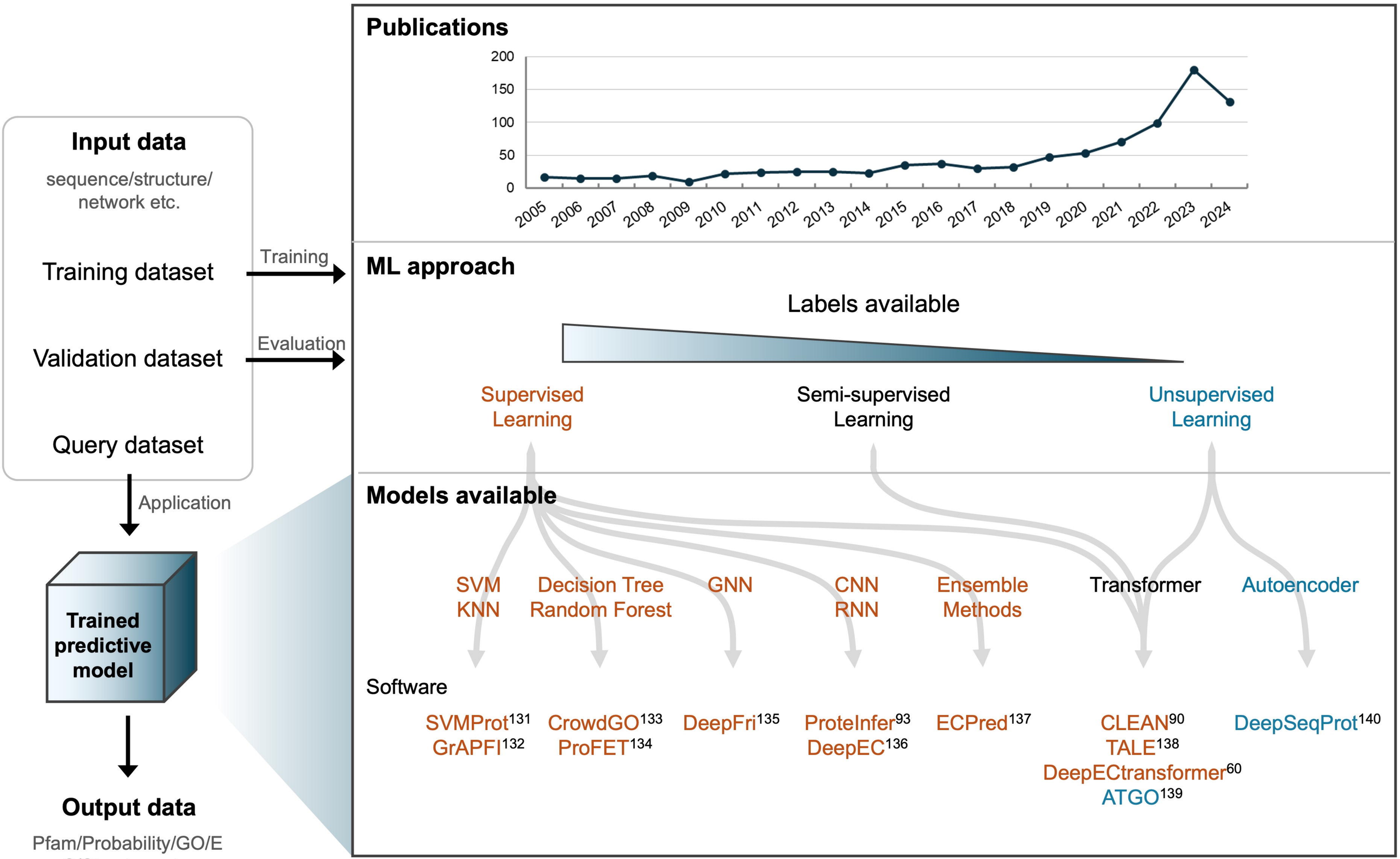
Figure 7. YciO harbors specific molecular surface features. (A) Structure-based multi-sequence alignment of TsaC proteins, the TsaC domains of Sua5 proteins and YciO proteins, derived from available crystal structures and AlphaFold structural models. For crystal structures, the PDB IDs are indicated 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

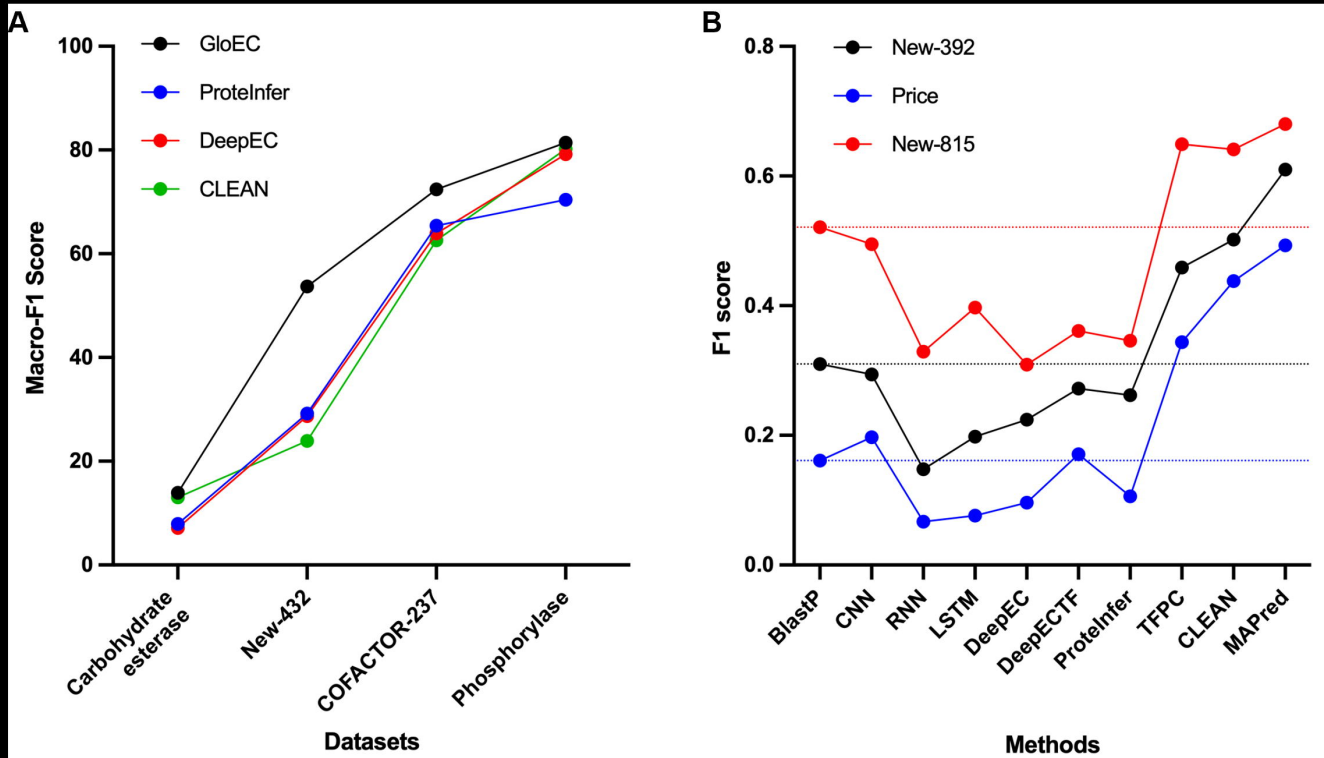
*Ec*YciO (**C**), color-coded by surface electrostatic potential (top), and by positional sequence conservation score calculated from 4264 TsaC/Sua5 sequences and 1955 YciO sequences (bottom). The color keys for both panels are shown on the right. The conserved, YciO-specific, positively charged surface 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.

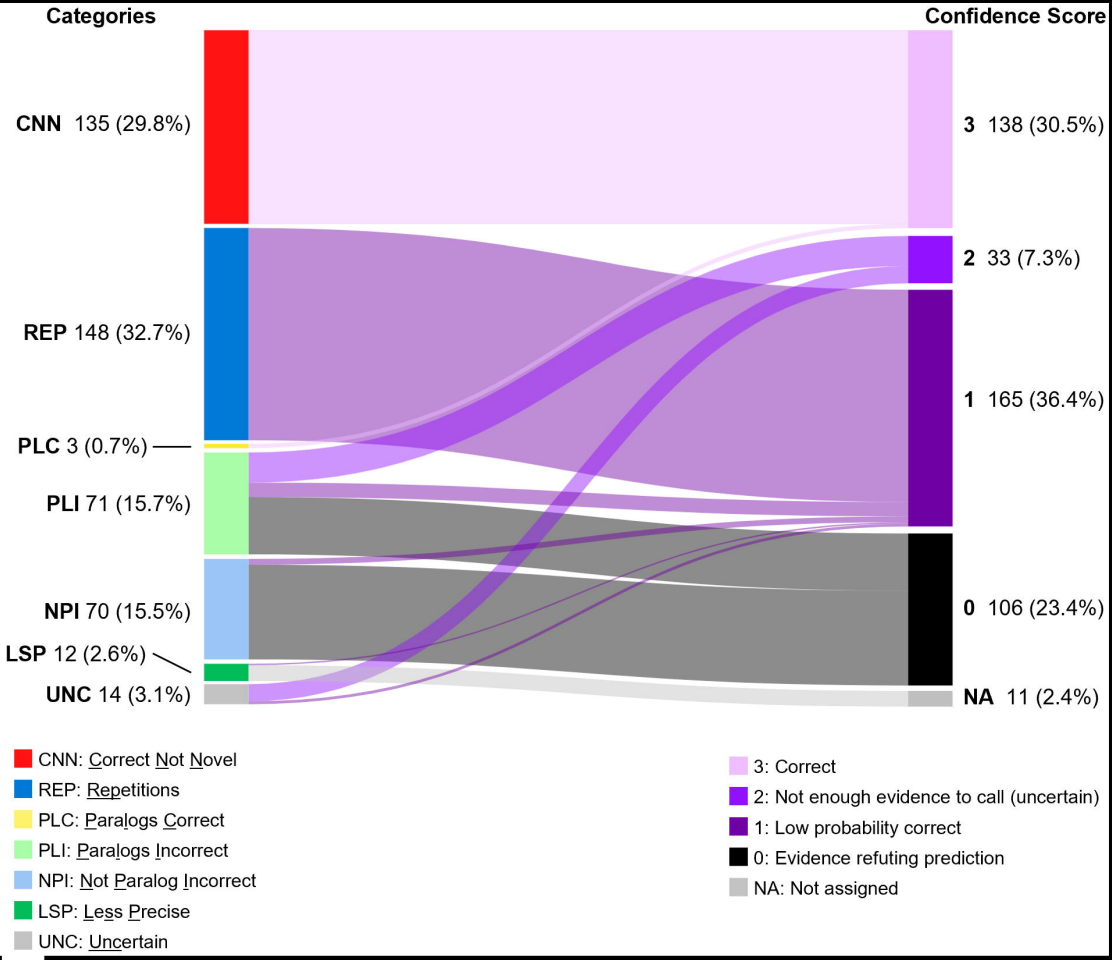
Figure 8. The genes encoding members of the *yciO* and *tsaC* have different genome neighborhood contexts. **(A)** Sequence Similarity Network of 54,820 PF01300 family members between 150 and 350 aa in length generated using EFI-EST. Each node in the network represents one or multiple sequences that 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 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*;

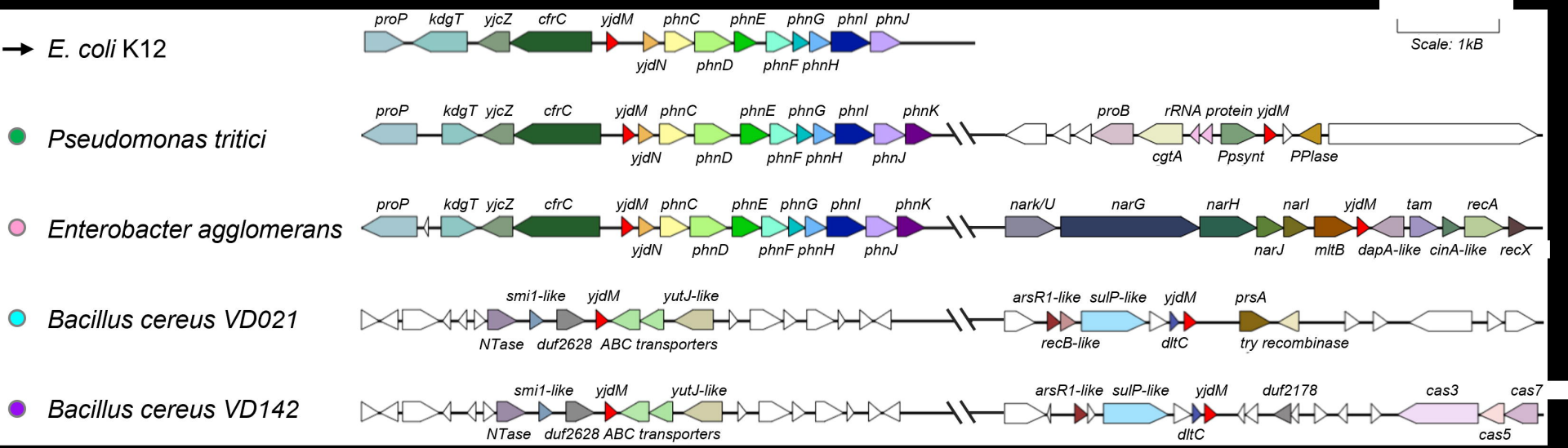

A
Initial knowledge of Proteins
Known functions
E.C.a
E.C.b
E.C.c
...
E.C.n
Ground Truth
isofunctional subfamily1
Challenge 1
Protein Family 1
isofunctional subfamily2
Challenge 2
Protein Family 2
Challenge 2
Non-pre-existing function
No initial knowledge of proteins
B
Erroneous annotations
1
4/5
2
3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

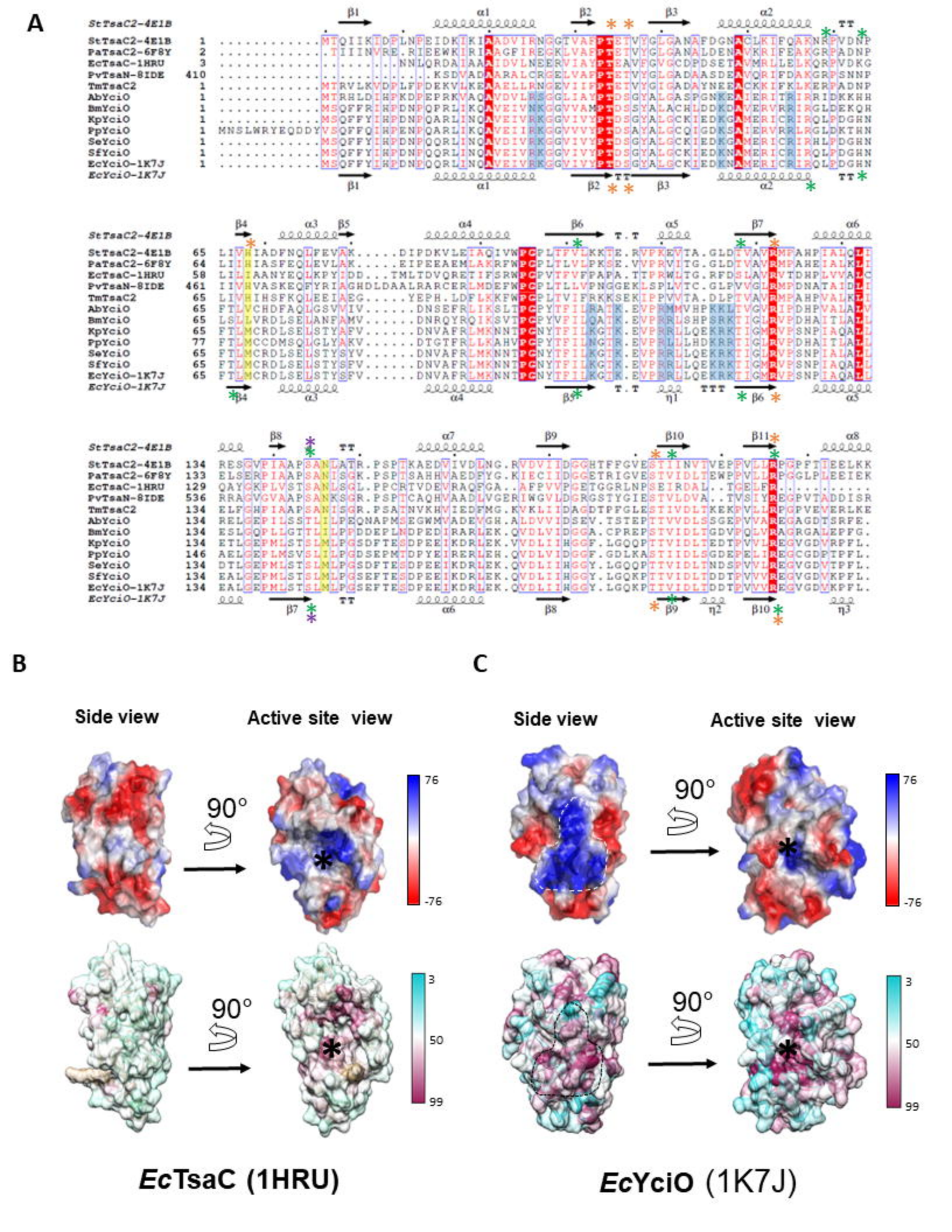

X, Y, M, and N proteins are homologs
Paralogs
Homologs
N
Y
X
M
Different neighborhood context of x, y, m and n
X
Y
M
N
Substrate docking
X
Y
M
N
Signature motif
Catalyst
Substrate binding
X xSGGKDSx...xLSGGx
X xSGGKDSx...xLSGGx
Y xSG-KDSx...xLAVGx
Y xSG-KDSx...xLAVGx
M xSGGLGDx...xLS-Gx
M xSGGLGDx...xLS-Gx
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

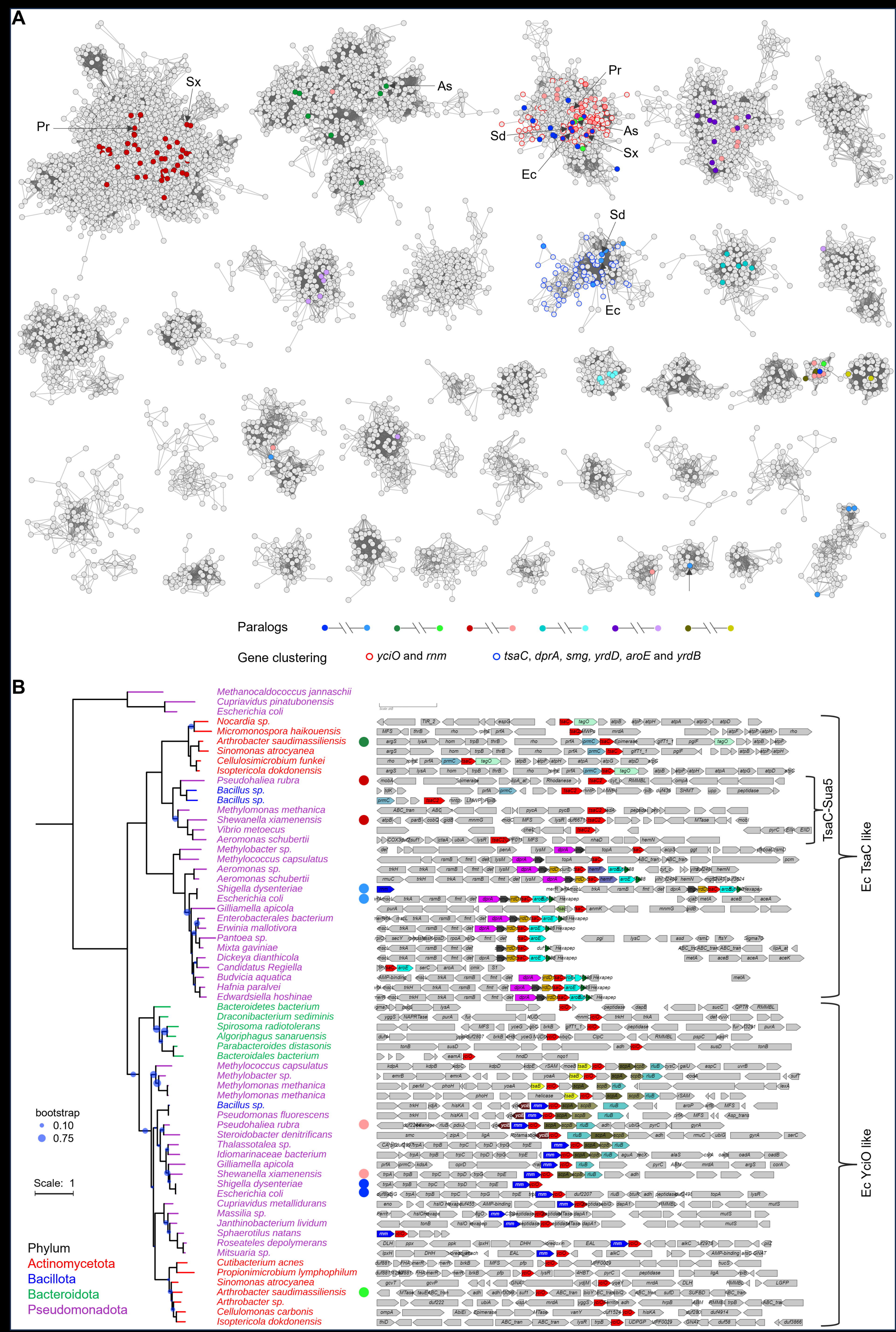

Input data
sequence/structure/
network etc.
Training dataset
Validation dataset
Query dataset
Training
Evaluation
Application
Trained predictive model
Output data
Pfam/Probability/GO/E
C/Structure etc.
Publications
200
150
100
50
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
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
SVMProt131
GrAPFI132
CrowdGO133
ProFET134
DeepFri135
ProteInfer93
DeepEC136
ECPred137
CLEAN90
TALE138
DeepECtransformer60
ATGO139
DeepSeqProt140


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


Categories
Confidence Score
CNN 135 (29.8%)
REP 148 (32.7%)
PLC 3 (0.7%)
PLI 71 (15.7%)
NPI 70 (15.5%)
LSP 12 (2.6%)
UNC 14 (3.1%)
3 138 (30.5%)
2 33 (7.3%)
1 165 (36.4%)
0 106 (23.4%)
NA 11 (2.4%)
CNN: Correct Not Novel
REP: Repetitions
PLC: Paralogs Correct
PLI: Paralogs Incorrect
NPI: Not Paralog Incorrect
LSP: Less Precise
UNC: Uncertain
3: Correct
2: Not enough evidence to call (uncertain)
1: Low probability correct
0: Evidence refuting prediction
NA: Not assigned


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
Sx
Pr
As
Pr
Sd
As
Sx
Ec
Sd
Ec
Paralogs
Gene clustering
yciO and rnm
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