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 "unknownme". 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 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

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. 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, integration of multiple types of genomic and/or genome-wide experimental evidence can lead to predictions that can be tested experimentally[21].

Machine learning-based methods are currently revolutionizing many fields of biology[22,23] including prediction of protein properties such as 3D structure, ligand binding[24], molecular function[25], and function propagation accuracy[26–33]. However, these models are not, in essence, designed to discover novelty, as they primarily rely on training sets derived from existing knowledge. Hence, machine learning-based methods such as those reported by Kim et al[31] that claim to label true "unknowns" with substrate-level specificity enzymatic roles, should be carefully evaluated, as discussed below.

Methods

Bioinformatic analyses and data mining

The Uniprot (www.uniprot.org)[34], NCBI (www.ncbi.nlm.nih.gov)[35], BioCyc (www.biocyc.org)[36] and KEGG[37] knowledgebases were routinely used in this study. Crystal structures were

retrieved from the Protein Data Bank (www.rcsb.org)[38] 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)[39] and Espript (espript.ibcp.fr/ESPript/ESPript). The percent conservation scores were calculated using the program AL2CO (http://prodata.swmed.edu/al2co/al2co.php)[40] 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)[41]. 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)[42]. More SSNs were then 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/)[43]. To generate the phylogenetic tree, proteins were aligned using MUSCLE[44]. The alignment was trimmed using BMGE[45] and used to build the maximum likelihood tree using FastTree[46] with LG+CAT model with bootstrap (1000 replicates) and visualized using iTOL[47].

Results

Prevalent functional annotation errors are over-propagations to non-isofunctional paralog subgroups.

The process of functional annotations of protein entries in databases starts with capturing information in the literature by biocurators[48]. This process links experimental characterizations of specific proteins in specific organisms to controlled vocabularies that describe validated functions such as the Gene Ontology (GO)[49], the IUPAC Enzyme Commission (EC) numbers (https://iubmb.qmul.ac.uk/enzyme/), or biochemical reaction descriptors (e.g. Rhea[50]). Even though text-mining tools have accelerated the flow of information captured[51–53], very few of the proteins in UniprotKB[54], the most widely used protein function database[55], have been linked to

experimental data[56], 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[12,13,57–59]. From the very beginning, red flags were raised on many sources of errors (listed in Table 1) that can occur in this process[60–66]. One prevalent error type affecting most protein families[66] is the over-annotation of paralogs[64,66,67] caused by the inherent functional complexity of enzyme families[68,69]. Fifty years of classification of evolutionary and structurally related superfamilies, have shown that functional diversification often occurs through duplication and divergence[70–72]. Closely related proteins can have different functions, resulting in a major source of annotation errors where proteins are assigned incorrect functional roles [64,66]. 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[73–75] (Fig. 1). 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[76] or incorporated into rules such as the Unirules used by UniprotKB[77]. 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 overpropagation errors is increasing over time[66].

Computational models could help propagate the experimentally validated functional annotations to the correct portion of the protein space.

Recently, there has been an explosion of publications reporting the use of pre-trained Protein Language Models (PLM) to predict protein functions[29–33,78–80] (**Fig. 2**). Fueled by the availability of PLMs from UniProt, ESM Atlas, and others (see references in Fig. 2 legend), models have been developed to link protein sequences to vocabularies such as GO terms [28,81,82] or to EC numbers[83]. Some models predicting EC numbers have integrated structural data to identify active site signature residues, improving the separation of non-isofunctional paralogous subgroups when these harbor experimentally validated members[17,29,84]. While EC prediction models outperform BLAST[78,80,83], they are less accurate than the curated annotation databases such as SwissProt or KEGG[85] and not accurate enough to reliably annotate enzymes for many practical purposes, particularly when trying to reach the level of substrate specificity (or the fourth EC number[86]). Based on the precedent of the “AlphaFold moment” that occurred for structure predictions[87], it is highly probable that as these PLM-driven approaches improve, they

could become mainstream tools to help correctly propagate known functional knowledge among isofunctional proteins. This will require improving training data, reducing bias caused by oversampling of model organisms and overrepresented functions[88] and integrating other types of evidence such as gene neighborhoods[79,89] or chemical reactions to better separate non-isofunctional paralog subgroups[62,70].

Embeddings from pre-trained protein language models can incorporate universal contextualized representations of protein sequences in an unsupervised manner from large unlabeled datasets, effectively capturing information from protein sequences and structures. Integrative embedding models can leverage distinct data sources such as protein sequences, InterPro fingerprints, and protein-protein interaction data within a unified embedding framework for multi-label function prediction, allowing integration and standardization of diverse sources of evidence beyond just sequences[90]. Proteins that have similar functions and very similar protein-protein interactions tend to be evolutionarily conserved across species. By creating a shared embedding space that captures the functional and interaction similarities between proteins within and across species, we can transfer functional annotations from well-studied proteins in one species to their functionally similar counterparts in other species[91]. While embedding-based models have shown promise in leveraging multiple data sources for protein function prediction, they have primarily relied on supervised learning approaches that transfer labels from proteins with known functions to unlabeled ones. To truly uncover novel, unexplored functional groups, there is a need for more unsupervised learning techniques that can cluster proteins into isofunctional groups without relying on existing annotations[92].

As highlighted in the CAFA challenges, the quality of the benchmark sets, annotation databases, and ontologies can also 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. 2). Explainable AI (XAI) techniques can help to dissect ‘black-box’ models and understand what input features are driving correct and incorrect predictions[93].

To date machine learning models largely fail at discovering the exact enzymatic activities of 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[30]. More specifically, as shown using the ProtInfer platform, EC number predictions of unknowns break down at the level of the 4th digit[86]. These failures are illustrated here by dissecting the predictions from a recent study that used deep learning with transformer layers (DeepECtransformer) to predict EC numbers for a third (452) of the ~1500 unknowns present in the model organism *E. coli* K12[31]. We started by analyzing the three cases (YciO, YgfF, and YjdM) where the predictions were further validated by *in vitro* assays.

YgfF is a member of the large Short-Chain Dehydrogenase/Reductase (SDR) superfamily (IPR002347). The Oppermann group developed a nomenclature system and HMMs based classification website (http://www.sdr-enzymes.org/) that distinguish different functional subgroups of the SDR superfamily[94,95]. This resource predicts YgfF is part of the SDR63C/Glucose 1-dehydrogenase subgroup, the exact activity predicted and validated in the Kim et al.[31] 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 DeepECtransformer platform.

YciO is a member of the same Pfam family (PF01300) as TsaC/Sua5. TsaC and Sua5 (or TsaC2) are two types of the L-threonylcarbamoyladenylate synthase (EC 2.7.7.87) that catalyzes the first step in the synthesis of the universal tRNA modification N-6-threonylcarbamoyladenosine or t^6A that differ by the presence of an additional domain[96,97]. The function of TsaC/Sua5 was first elucidated in 2009 [98]. It was already noted at the time that YciO was a paralog that did not perform the same function *in vivo*[98,99], even if a structure-based multi-sequence alignment comparing YciO and TsaC/Sua5 shows that the active site residues are very similar between the two proteins, consistent with similar catalytic function (Fig. 3A). Gene neighborhood and structural data suggest the actual function of YciO could be related to rRNA rather than tRNA metabolism; 1) the structure shows a broad positively charged surface predicted to interact with RNA ([100] and Fig. 3B and C); 2) *yciO* genes are colocalized with *rnm* genes in many species (Fig. 4) 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* [101]. Kim et al. predicted that YciO had the same function as TsaC, and showed it catalyzed the synthesis of L-threonylcarbamoyladenylate from ATP, L-threonine, and bicarbonate, *in vitro*[31]. 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, consistent with the possibility of a missing partner or a different biological substrate. Models of enzyme evolution go through promiscuous stages[75]. TsaC is an enzyme predicted to have been present in the Last Universal Common Ancestor (LUCA)[96,102]. Hence as YciO is a clear paralog of TsaC (Fig. 4) 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 if the existing data suggests a role in RNA metabolism it cannot be the same as TsaC and the EC number predicted by DeepECtransformer (2.7.7.87) is incorrect.

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. 5) but the initial publication that claimed this gene was involved in phosphonate catabolism was later refuted[103]. 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[104]. Gene neighborhoods of *phnA* genes show strong clustering with genes encoding phosphonoacetate transporters, phosphonoacetate sensing regulators, and in some cases 2-aminoethylphosphonate catabolism[105.] Still, except for *E. coli, yjdM* genes are generally not close to phosphonate catabolism or transport genes (Fig. 5). 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.

It is not possible to expand the extensive forensic analyses done on the YgjF, YciO, and YjdM proteins for all the 450 predictions as each analysis takes 5 to 10 hours. However, an initial review suggests that the conclusions drawn from these three cases may apply to the entire set. The predictions fall into four categories: correct predictions such as YgjF, obvious predictions because these proteins were not previously “unknown”, and those that are either outright erroneous at worst or uncertain requiring additional evidence at best.

Indeed, for 120 predictions, the same EC number was already present in the Uniprot annotation (Table S1A). These proteins were not of unknown function to start with and should not have been included in the “unknown” test set as they were most certainly part of the training set. Second, 75 other EC number predictions differ from the existing numbers in UniProt by the last digit(s) (Table S1B). Published experimental evidence shows that several of these are incorrect (Table 2) and, like YciO, fall in the over-propagation error category. We anticipate that

many other predictions in this group will fall into this category. Around half of the predictions (261) have no EC number in Uniprot (Table S1C). For a few of these (in red in Table S1C), the same function is listed in Uniprot but without an EC number. These predictions are again cases of propagation of knowns rather than solving unknowns. Adding biological context and/or literature analyses reveals that many of the other functional predictions in this set make no biological sense. 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[36] 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[106]. A few more examples of these nonsensical predictions are given in Table 2. Another type of obvious error observed was the repetition of identical EC numbers (up to 12 times, in blue Table S1C). This 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[107–110].

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[84]. 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 gene as some of the gap-filling studies have successfully done[111,112]. However, by essence, they cannot provide 4-digit EC numbers that do not yet exist to true unknowns as, as elegantly stated by Bromberg and Coll[30], “annotating new proteins based on available data seems to be an inherently flawed proposition” and “curiously, despite their seemingly unlimited by homology scope, novel deep learning methods remain far from capturing functional signal encoded in protein sequence”. In the few cases where AI tools have successfully predicted novel functions, other types of evidence were incorporated and reviewed by experts who brought their biochemical expertise and insights[33,113]. Supervised learning models always fit a solution, hence the absence of downstream checks that place the

predictions into a broader biological context can lead to erroneous calls even if those seem to be validated by *in vitro* tests, as *in vitro* function is not a reliable proof of the in-vivo functional role in the absence of biological validation[114,115]. These limitations are not well understood by authors, reviewers, readers, or even biocurators leading to the propagation of incorrect annotations in databases, making these harder to fix[116]. These mistakes can be easily avoided by incorporating biological context sanity checks to evaluate the results from predictive ML models. On the more constructive side, the field should avoid supervised models that were not developed to discover novelty but explore the use of unsupervised models that integrate different types of evidence that could help biochemical experts develop testable functional hypotheses for unknowns.

Table 1: Main types of errors that lead to erroneous functional annotations of proteins.

Error Types		
False unknowns	**Annotated as unknown or general but function is known and published**	
	Details	***Example***
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[85].
Naming issues	Inconsistent naming of the same entities.	See GroEL in [117].
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[118] 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***
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[119]. See also examples in[85].
Curation mistake	1) The data was incorrectly captured by biocurator; 2) Functional annotations may become outdated.	Ureidoglycolate lyase [65].
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[116].
Over-annotation of paralogs	Annotation wrongly propagated to non-isofunctional paralogous groups.	See examples in [64,66,67].

Table 2: Example where experimental evidence contradicts the DeepECtransformer predictions.

Id	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[80].
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[81].
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[82]

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 [120]
b1267	None	Uncharacterized protein YciO	2.7.7.87	L-threonylcarbamoyladenylate synthase	YciO is a paralog of TsaC but does catalyze the same activity [98,99].
Example of easily 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 mofification[83].
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[84].
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[106].
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).

Supplemental tables and data

Table S1: Analysis of EC predictions for 450 E. coli unknowns by DEEP EC.
Table S2: PF01300 family members used in SSN.
Table S3: Gene neighborhood information for selected PF01300 family genes.
Table S4: Gene neighborhood information for selected *yjdM* (IPR004624 family) genes.
Supplemental data 1: TsaC sequences used to generate the positional conservation scores for Figure 3B.
Supplemental data 2: YciO sequences used to generate the positional conservation scores for Figure 3C.

Competing interests

The authors declare no competing interests.

Data availability

Supplemental material is available at G3 online.

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 3 and wrote sections of the manuscript. YY helped generate the data for Table S1 and created Figures 1, 2, 4, and 5. 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.; Danko, C. G.; Gill, C. A.; Jarvis, E. D.; Koepfli, K.-P.; Koltes, J. E.; Lyons, E.; Ronald, P.; Ryder, O. A.; Schriml, L. M.; Soltis, P.; VandeWoude, S.; Zhou, H.; Ostrander, E. A.; Karlsson, E. K. 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.; Muschalik, N.; Shah, R. D.; Emran, S.; Robles, C.; Freeman, M.; Munro, S. 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.; Babor, J.; Bateman, A.; Blaby, I.; Blaby-Haas, C.; Bridge, A. J.; Burley, S. K.; Cleveland, S.; Colwell, L. J.; Conesa, A.; Dallago, C.; Danchin, A.; de Waard, A.; Deutschbauer, A.; Dias, R.; Ding, Y.; Fang, G.; Friedberg, I.; Gerlt, J.; Goldford, J.; Gorelik, M.; Gyori, B. M.; Henry, C.; Hutinet, G.; Jaroch, M.; Karp, P. D.; Kondratova, L.; Lu, Z.; Marchler-Bauer, A.; Martin, M.-J.; McWhite, C.; Moghe, G. D.; Monaghan, P.; Morgat, A.; Mungall, C. J.; Natale, D. A.; Nelson, W. C.; O'Donoghue, S.; Orengo, C.; O'Toole, K. H.; Radivojac, P.; Reed, C.; Roberts, R. J.; Rodionov, D.; Rodionova, I. A.; Rudolf, J. D.; Saleh, L.; Sheynkman, G.; Thibaud-Nissen, F.; Thomas, P. D.; Uetz, P.; Vallenet, D.; Carter, E. W.; Weigele, P. R.; Wood, V.; Wood-Charlson, E. M.; Xu, J. A Roadmap for the Functional Annotation of Protein Families: A Community Perspective. *Database* **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.

(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). 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.

(17) Sajid, S.; Mashkoor, M.; Jørgensen, M. G.; Christensen, L. P.; Hansen, P. R.; Franzyk, H.; Mirza, O.; Prabhala, B. K. The Y-Ome Conundrum: Insights into Uncharacterized Genes and Approaches for Functional Annotation. *Mol Cell Biochem* **2023**. 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.

(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) 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* **2010**, *425* (1), 1–11.

(22) Sapoval, N.; Aghazadeh, A.; Nute, M. G.; Antunes, D. A.; Balaji, A.; Baraniuk, R.; Barberan, C. J.; Dannenfelser, R.; Dun, C.; Edrisi, M.; Elworth, R. A. L.; Kille, B.; Kyrillidis, A.; Nakhleh, L.; Wolfe, C. R.; Yan, Z.; Yao, V.; Treangen, T. J. 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.

(23) Bordin, N.; Dallago, C.; Heinzinger, M.; Kim, S.; Littmann, M.; Rauer, C.; Steinegger, M.; Rost, B.; Orengo, C. 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.

(24) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**. https://doi.org/10.1038/s41586-024-07487-w.

(25) 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.

(26) 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.

(27) 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.

(28) Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N.; Davis, L.; Dogan, T.; Atalay, V.; Rifaioglu, A. S.; Dalkıran, A.; Cetin Atalay, R.; Zhang, C.; Hurto, R. L.; Freddolino, P. L.; Zhang, Y.; Bhat, P.; Supek, F.; Fernández, J. M.; Gemovic, B.; Perovic, V. R.; Davidović, R. S.; Sumonja, N.; Veljkovic, N.; Asgari, E.; Mofrad, M. R. K.; Profiti, G.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Boecker, F.; Schoof, H.; Kahanda, I.; Thurlby, N.; McHardy, A. C.; Renaux, A.; Saidi, R.; Gough, J.; Freitas, A. A.; Antczak, M.; Fabris, F.; Wass, M. N.; Hou, J.; Cheng, J.; Wang, Z.; Romero, A. E.; Paccanaro, A.; Yang, H.; Goldberg, T.; Zhao, C.; Holm, L.; Törönen, P.; Medlar, A. J.; Zosa, E.; Borukhov, I.; Novikov, I.; Wilkins, A.; Lichtarge, O.; Chi, P.-H.; Tseng, W.-C.; Linial, M.; Rose, P. W.; Dessimoz, C.; Vidulin, V.; Dzeroski, S.; Sillitoe, I.; Das, S.; Lees, J. G.; Jones, D. T.; Wan, C.; Cozzetto, D.; Fa, R.; Torres, M.; Warwick Vesztrocy, A.; Rodriguez, J. M.; Tress, M. L.; Frasca, M.; Notaro, M.; Grossi, G.; Petrini, A.; Re, M.; Valentini, G.; Mesiti, M.; Roche, D. B.; Reeb, J.; Ritchie, D. W.; Aridhi, S.; Alborzi, S. Z.; Devignes, M.-D.; Koo, D. C. E.; Bonneau, R.; Gligorijević, V.; Barot, M.; Fang, H.; Toppo, S.; Lavezzo, E.; Falda, M.; Berselli, M.; Tosatto, S. C. E.; Carraro, M.; Piovesan, D.; Ur Rehman, H.; Mao, Q.; Zhang, S.; Vucetic, S.; Black, G. S.; Jo, D.; Suh, E.; Dayton, J. B.; Larsen, D. J.; Omdahl, A. R.; McGuffin, L. J.; Brackenridge, D. A.; Babbitt, P. C.; Yunes, J. M.; Fontana, P.; Zhang, F.; Zhu, S.; You, R.; Zhang, Z.; Dai, S.; Yao, S.; Tian, W.; Cao, R.; Chandler, C.; Amezola, M.; Johnson, D.; Chang, J.-M.; Liao, W.-H.; Liu, Y.-W.; Pascarelli, S.; Frank, Y.; Hoehndorf, R.; Kulmanov, M.; Boudellioua, I.; Politano, G.; Di Carlo, S.; Benso, A.; Hakala, K.; Ginter, F.; Mehryary, F.; Kaewphan, S.; Björne, J.; Moen, H.; Tolvanen, M. E. E.; Salakoski, T.; Kihara, D.; Jain, A.; Šmuc, T.; Altenhoff, A.; Ben-Hur, A.; Rost, B.; Brenner, S. E.; Orengo, C. A.; Jeffery, C. J.; Bosco, G.; Hogan, D. A.; Martin, M. J.; O'Donovan, C.; Mooney, S. D.; Greene, C. S.; Radivojac, P.; Friedberg, I. 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.

(29) 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.

(30) 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.

(31) 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.

(32) Durairaj, J.; Waterhouse, A. M.; Mets, T.; Brodiazhenko, T.; Abdullah, M.; Studer, G.; Tauriello, G.; Akdel, M.; Andreeva, A.; Bateman, A.; Tenson, T.; Hauryliuk, V.; Schwede, T.; Pereira, J. 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.

(33) 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* **2023**, in press. https://doi.org/10.1038/s41586-023-06955-z.

(34) Consortium, U. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res* **2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(35) 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.

(36) Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; Paley, S. M.; Subhraveti, P. The BioCyc Collection of Microbial Genomes and Metabolic Pathways. *Brief Bioinform* **2019**, *20* (4), 1085–1093. https://doi.org/10.1093/bib/bbx085.

(37) 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.

(38) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Fayazi, M.; Feng, Z.; Flatt, J. W.; Ganesan, S.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Henry, J.; Hudson, B. P.; Khokhriakov, I.; Lawson, C. L.; Liang, Y.; Lowe, R.; Peisach, E.; Persikova, I.; Piehl, D. W.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Vallat, B.; Voigt, M.; Webb, B.; Westbrook, J. D.; Whetstone, S.; Young, J. Y.; Zalevsky, A.; Zardecki, C. 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 Research D1* **2022**, *51* (D1), D488-D508. https://doi.org/10.1093/nar/gkac1077.

(39) 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.

(40) 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.

(41) 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.

(42) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **2003**, *13* (11), 2498–2504.

(43) 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.

(44) Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucl. Acids Res.* **2004**, *32* (5), 1792–1797. https://doi.org/10.1093/nar/gkh340.

(45) 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.

(46) 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.

(47) 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). https://doi.org/10.1093/nar/gkab301.

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

(49) Gene Ontology Consortium; Aleksander, S. A.; Balhoff, J.; Carbon, S.; Cherry, J. M.; Drabkin, H. J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N. L.; Hill, D. P.; Lee, R.; Mi, H.; Moxon, S.; Mungall, C. J.; Muruganugan, A.; Mushayahama, T.; Sternberg, P. W.; Thomas, P. D.; Van Auken, K.; Ramsey, J.; Siegele, D. A.; Chisholm, R. L.; Fey, P.; Aspromonte, M. C.; Nugnes, M. V.; Quaglia, F.; Tosatto, S.; Giglio, M.; Nadendla, S.; Antonazzo, G.; Attrill, H.; Dos Santos, G.; Marygold, S.; Strelets, V.; Tabone, C. J.; Thurmond, J.; Zhou, P.; Ahmed, S. H.; Asanitthong, P.; Luna Buitrago, D.; Erdol, M. N.; Gage, M. C.; Ali Kadhum, M.; Li, K. Y. C.; Long, M.; Michalak, A.; Pesala, A.; Pritazahra, A.; Saverimuttu, S. C. C.; Su, R.; Thurlow, K. E.; Lovering, R. C.; Logie, C.; Oliferenko, S.; Blake, J.; Christie, K.; Corbani, L.; Dolan, M. E.; Drabkin, H. J.; Hill, D. P.; Ni, L.; Sitnikov, D.; Smith, C.; Cuzick, A.; Seager, J.; Cooper, L.; Elser, J.; Jaiswal, P.; Gupta, P.; Jaiswal, P.; Naithani, S.; Lera-Ramirez, M.; Rutherford, K.; Wood, V.; De Pons, J. L.; Dwinell, M. R.; Hayman, G. T.; Kaldunski, M. L.; Kwitek, A. E.; Lauledеrkind, S. J. F.; Tutaj, M. A.; Vedi, M.; Wang, S.-J.; D'Eustachio, P.; Aimo, L.; Axelsen, K.; Bridge, A.; Hyka-Nouspikel, N.; Morgat, A.; Aleksander, S. A.; Cherry, J. M.; Engel, S. R.; Karra, K.; Miyasato, S. R.; Nash, R. S.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Bakker, E.; Berardini, T. Z.; Reiser, L.; Auchincloss, A.; Axelsen, K.; Argoud-Puy, G.; Blatter, M.-C.; Boutet, E.; Breuza, L.; Bridge, A.; Casals-Casas, C.; Coudert, E.; Estreicher, A.; Livia Famiglietti, M.; Feuermann, M.; Gos, A.; Gruaz-Gumowski, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Pedruzzi, I.; Pourcel, L.; Poux, S.; Rivoire, C.; Sundaram, S.; Bateman, A.; Bowler-Barnett, E.; Bye-A-Jee, H.; Denny, P.; Ignatchenko, A.; Ishtiaq, R.; Lock, A.; Lussi, Y.; Magrane, M.; Martin, M. J.; Orchard, S.; Raposo, P.; Speretta, E.; Tyagi, N.; Warner, K.; Zaru, R.; Diehl, A. D.; Lee, R.; Chan, J.; Diamantakis, S.; Raciti, D.; Zarowiecki, M.; Fisher, M.; James-Zorn, C.; Ponferrada, V.; Zorn, A.; Ramachandran, S.; Ruzicka, L.; Westerfield, M. The Gene Ontology Knowledgebase in 2023. *Genetics* **2023**, *224* (1), iyad031. https://doi.org/10.1093/genetics/iyad031.

(50) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea, the Reaction Knowledgebase in 2022. *Nucleic Acids Res* **2022**, *50* (D1), D693–D700. https://doi.org/10.1093/nar/gkab1016.

(51) Poux, S.; Arighi, C. N.; Magrane, M.; Bateman, A.; Wei, C.-H.; Lu, Z.; Boutet, E.; Bye-A-Jee, H.; Famiglietti, M. L.; Roechert, B.; UniProt Consortium, T. 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.

(52) 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.

(53) 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.

(54) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Garmiri, P.; da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.;

Coudert, E.; Cuche, B.; de Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51* (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

(55) 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.

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

(57) 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.

(58) Storm, C. E.; Sonnhammer, E. L. L. Automated Ortholog Inference from Phylogenetic Trees and Calculation of Orthology Reliability. *Bioinformatics* **2002**, *18*, 92–99.

(59) 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.

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

(61) Devos, D.; Valencia, A. Intrinsic Errors in Genome Annotation. *Trends in Genetics* **2001**, *17* (8), 429–431.

(62) 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/doi: 10.1093/bioinformatics/18.12.1641.

(63) 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.

(64) 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/DOI: 10.1371/journal.pcbi.1000605.

(65) 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–bat071. https://doi.org/10.1093/database/bat071.

(66) 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.

(67) 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.

(68) 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.

(69) 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.

(70) 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.

(71) 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.

(72) 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.

(73) Zallot, R.; Harrison, K.; Kolaczkowski, B.; de Crécy-Lagard, V. Functional Annotations of Paralogs: A Blessing and a Curse. *Life* **2016**, *6* (3), 39.

(74) 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/https://doi.org/10.1016/j.copbio.2020.12.004.

(75) 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.

(76) Li, W.; O'Neill, K. R.; Haft, D. H.; DiCuccio, M.; Chetvernin, V.; Badretdin, A.; Coulouris, G.; Chitsaz, F.; Derbyshire, M. K.; Durkin, A. S.; Gonzales, N. R.; Gwadz, M.; Lanczycki, C. J.; Song, J. S.; Thanki, N.; Wang, J.; Yamashita, R. A.; Yang, M.; Zheng, C.; Marchler-Bauer, A.; Thibaud-Nissen, F. 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.

(77) MacDougall, A.; Volynkin, V.; Saidi, R.; Poggioli, D.; Zellner, H.; Hatton-Ellis, E.; Joshi, V.; O'Donovan, C.; Orchard, S.; Auchincloss, A. H.; Baratin, D.; Bolleman, J.; Coudert, E.; de Castro, E.; Hulo, C.; Masson, P.; Pedruzzi, I.; Rivoire, C.; Arighi, C.; Wang, Q.; Chen, C.; Huang, H.; Garavelli, J.; Vinayaka, C. R.; Yeh, L.-S.; Natale, D. A.; Laiho, K.; Martin, M.-J.; Renaux, A.; Pichler, K.; Consortium, U. 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.

(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. In *Machine Learning for Structural Biology Workshop, NeurIPS 2023.*; 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.; Bankapur, A. R.; D'Andrea, D.; Lepore, R.; Funk, C. S.; Kahanda, I.; Verspoor, K. M.; Ben-Hur, A.; Koo, D. C. E.; Penfold-Brown, D.; Shasha, D.; Youngs, N.; Bonneau, R.; Lin, A.; Sahraeian, S. M. E.; Martelli, P. L.; Profiti, G.; Casadio, R.; Cao, R.; Zhong, Z.; Cheng, J.; Altenhoff, A.; Skunca, N.; Dessimoz, C.; Dogan, T.; Hakala, K.; Kaewphan, S.; Mehryary, F.; Salakoski, T.; Ginter, F.; Fang, H.; Smithers, B.; Oates, M.; Gough, J.; Törönen, P.; Koskinen, P.; Holm, L.; Chen, C.-T.; Hsu, W.-L.; Bryson, K.; Cozzetto, D.; Minneci, F.; Jones, D. T.; Chapman, S.; BKC, D.; Khan, I. K.; Kihara, D.; Ofer, D.; Rappoport, N.; Stern, A.; Cibrian-Uhalte, E.; Denny, P.;

Foulger, R. E.; Hieta, R.; Legge, D.; Lovering, R. C.; Magrane, M.; Melidoni, A. N.; Mutowo-Meullenet, P.; Pichler, K.; Shypitsyna, A.; Li, B.; Zakeri, P.; ElShal, S.; Tranchevent, L.-C.; Das, S.; Dawson, N. L.; Lee, D.; Lees, J. G.; Sillitoe, I.; Bhat, P.; Nepusz, T.; Romero, A. E.; Sasidharan, R.; Yang, H.; Paccanaro, A.; Gillis, J.; Sedeño-Cortés, A. E.; Pavlidis, P.; Feng, S.; Cejuela, J. M.; Goldberg, T.; Hamp, T.; Richter, L.; Salamov, A.; Gabaldon, T.; Marcet-Houben, M.; Supek, F.; Gong, Q.; Ning, W.; Zhou, Y.; Tian, W.; Falda, M.; Fontana, P.; Lavezzo, E.; Toppo, S.; Ferrari, C.; Giollo, M.; Piovesan, D.; Tosatto, S. C. E.; del Pozo, A.; Fernández, J. M.; Maietta, P.; Valencia, A.; Tress, M. L.; Benso, A.; Di Carlo, S.; Politano, G.; Savino, A.; Rehman, H. U.; Re, M.; Mesiti, M.; Valentini, G.; Bargsten, J. W.; van Dijk, A. D. J.; Gemovic, B.; Glisic, S.; Perovic, V.; Veljkovic, V.; Veljkovic, N.; Almeida-e-Silva, D. C.; Vencio, R. Z. N.; Sharan, M.; Vogel, J.; Kansakar, L.; Zhang, S.; Vucetic, S.; Wang, Z.; Sternberg, M. J. E.; Wass, M. N.; Huntley, R. P.; Martin, M. J.; O'Donovan, C.; Robinson, P. N.; Moreau, Y.; Tramontano, A.; Babbitt, P. C.; Brenner, S. E.; Linial, M.; Orengo, C. A.; Rost, B.; Greene, C. S.; Mooney, S. D.; Friedberg, I.; Radivojac, P. 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) Yu, T.; Cui, H.; Li, J. C.; Luo, Y.; Jiang, G.; Zhao, H. Enzyme Function Prediction Using Contrastive Learning. *Science (1979)* **2023**, *379* (6639), 1358–1363. https://doi.org/10.1126/science.adf2465.

(85) 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.

(86) Sanderson, T.; Bileschi, M. L.; Belanger, D.; Colwell, L. J. ProteInfer, Deep Neural Networks for Protein Functional Inference. *Elife* **2023**, *12*. https://doi.org/10.7554/eLife.80942.

(87) 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.

(88) 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.

(89) Urhan, A.; Cosma, B.-M.; Manson, A. L.; Abeel, T. SAP: Synteny-Aware Gene Function Prediction for Bacteria Using Protein Embeddings. *bioRxiv* **2023**.

(90) Armah-Sekum, R. E.; Szedmak, S.; Rousu, J. Protein Function Prediction through Multi-View Multi-Label Latent Tensor Reconstruction. *BMC Bioinformatics* **2024**, *25* (1), 174. https://doi.org/10.1186/s12859-024-05789-4.

(91) Arsenescu, V.; Devkota, K.; Erden, M.; Shpilker, P.; Werenski, M.; Cowen, L. J. MUNDO: Protein Function Prediction Embedded in a Multispecies World. *Bioinformatics Advances* **2022**, *2* (1). https://doi.org/10.1093/bioadv/vbab025.

(92) Mahlich, Y.; Zhu, C.; Chung, H.; Velaga, P. K.; De Paolis Kaluza, M. C.; Radivojac, P.; Friedberg, I.; Bromberg, Y. 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.

(93) Chennam, K. K.; Mudrakola, S.; Maheswari, V. U.; Aluvalu, R.; Rao, K. G. Black Box Models for EXplainable Artificial Intelligence; 2023; pp 1–24. https://doi.org/10.1007/978-3-031-12807-3_1.

(94) Persson, B.; Kallberg, Y.; Bray, J. E.; Bruford, E.; Dellaporta, S. L.; Favia, A. D.; Duarte, R. G.; Jörnvall, H.; Kavanagh, K. L.; Kedishvili, N.; Kisiela, M.; Maser, E.; Mindnich, R.; Orchard, S.; Penning, T. M.; Thornton, J. M.; Adamski, J.; Oppermann, U. 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.

(95) 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.

(96) 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*. https://doi.org/10.3389/fmicb.2023.1204045.

(97) 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.

(98) El Yacoubi, B.; Lyons, B.; Cruz, Y.; Reddy, R.; Nordin, B.; Agnelli, F.; Williamson, J. R.; Schimmel, P.; Swairjo, M. A.; de Crécy-Lagard, V. 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.

(99) Gerdes, S.; El Yacoubi, B.; Bailly, M.; Blaby, I. K.; Blaby-Haas, C. E.; Jeanguenin, L.; Lara-Núñez, A.; Pribat, A.; Waller, J. C.; Wilke, A.; Overbeek, R.; Hanson, A. D.; de Crécy-Lagard, V. 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.

(100) 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.

(101) 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.

(102) Gagler, D. C.; Karas, B.; Kempes, C. P.; Malloy, J.; Mierzejewski, V.; Goldman, A. D.; Kim, H.; Walker, S. I. 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.

(103) 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.

(104) 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.

(105) 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.

(106) 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.

(107) 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.

(108) 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.
(109) 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.
(110) 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.
(111) 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.
(112) 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.
(113) Dembech, E.; Malatesta, M.; De Rito, C.; Mori, G.; Cavazzini, D.; Secchi, A.; Morandin, F.; Percudani, R. Identification of Hidden Associations among Eukaryotic Genes through Statistical Analysis of Coevolutionary Transitions. *Proc Natl Acad Sci U S A* **2023**, *120* (16). https://doi.org/10.1073/pnas.2218329120.
(114) Punekar, N. S. In Vitro Versus In Vivo: Concepts and Consequences. In *ENZYMES: Catalysis, Kinetics and Mechanisms*; Springer Singapore: Singapore, 2018; pp 493–519. https://doi.org/10.1007/978-981-13-0785-0_38.
(115) 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.
(116) 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.
(117) 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*. https://doi.org/10.3389/fmicb.2019.00383.
(118) De Crécy-Lagard, V.; Phillips, G.; Grochowski, L. L.; Yacoubi, B. E.; Jenney, F.; Adams, M. W. W.; Murzin, A. G.; White, R. H. 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.
(119) 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.
(120) 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).
(121) Li, Y. H.; Xu, J. Y.; Tao, L.; Li, X. F.; Li, S.; Zeng, X.; Chen, S. Y.; Zhang, P.; Qin, C.; Zhang, C.; Chen, Z.; Zhu, F.; Chen, Y. Z. 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.

(122) 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.

(123) 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.

(124) 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.

(125) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; Xavier, R. J.; Knight, R.; Cho, K.; Bonneau, R. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat Commun* **2021**, *12* (1), 3168. https://doi.org/10.1038/s41467-021-23303-9.

(126) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep Learning Enables High-Quality and High-Throughput Prediction of Enzyme Commission Numbers. *Proceedings of the National Academy of Sciences* **2019**, *116* (28), 13996–14001. https://doi.org/10.1073/pnas.1821905116.

(127) 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.

(128) 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.

(129) 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.

(130) David, K. T.; Halanych, K. M. Unsupervised Deep Learning Can Identify Protein Functional Groups from Unaligned Sequences. *Genome Biol Evol* **2023**, *15* (5). https://doi.org/10.1093/gbe/evad084.

Figure legends

Figure 1. Computational workflow used to separate non-isofunctional paralogs in a protein superfamily. Arrows indicate the flow of information.

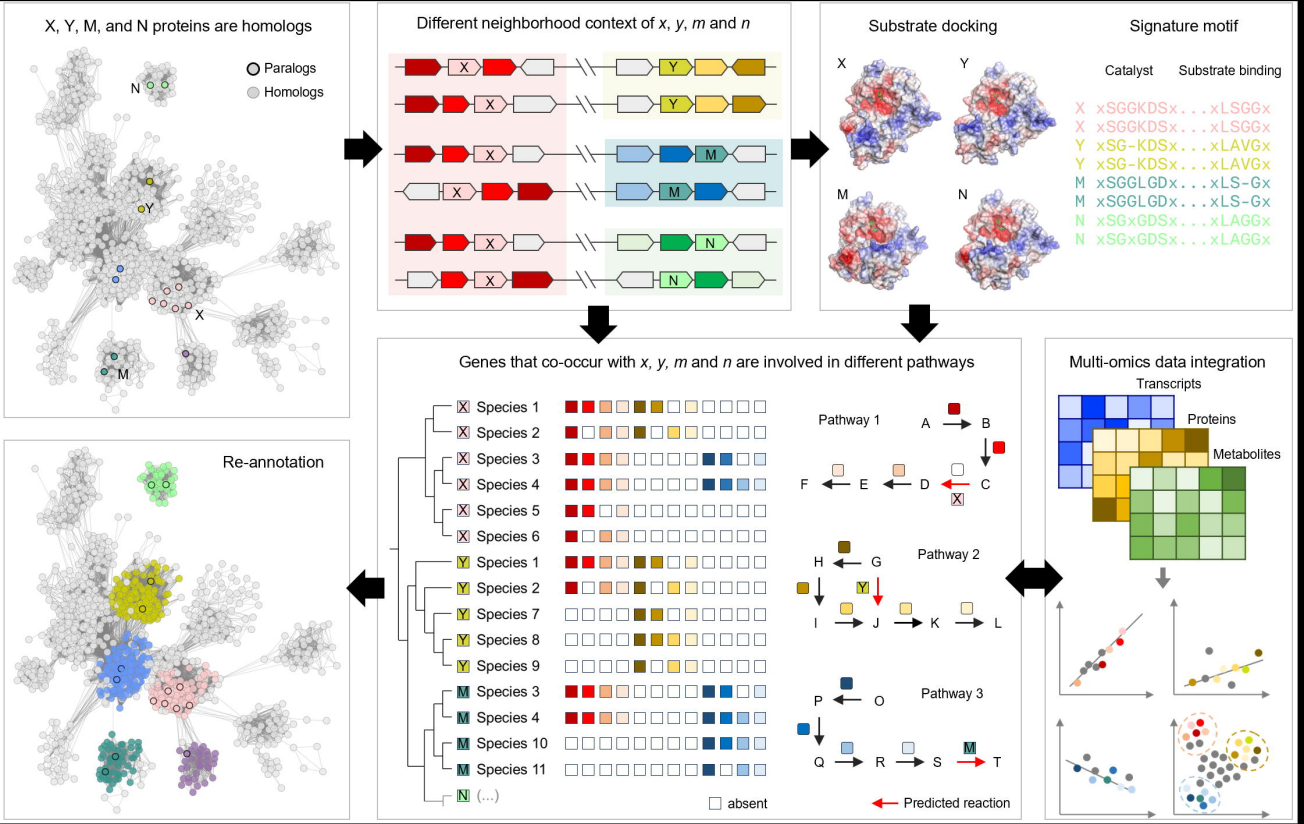
Figure 2. Workflow for functional annotation of proteins using machine learning. General steps are depicted: input data collection (training, validation, and query datasets), options for machine learning approach, model selection, training, and evaluation, and output data. Supervised, semi-supervised, and unsupervised learning approaches and 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 total publication numbers for each year was collected from PubMed (https://pubmed.ncbi.nlm.nih.gov/) at NCBI on June 26, 2024 using the search criteria: 'function prediction protein language model'. The corresponding reference number is given for each tool analyzed.

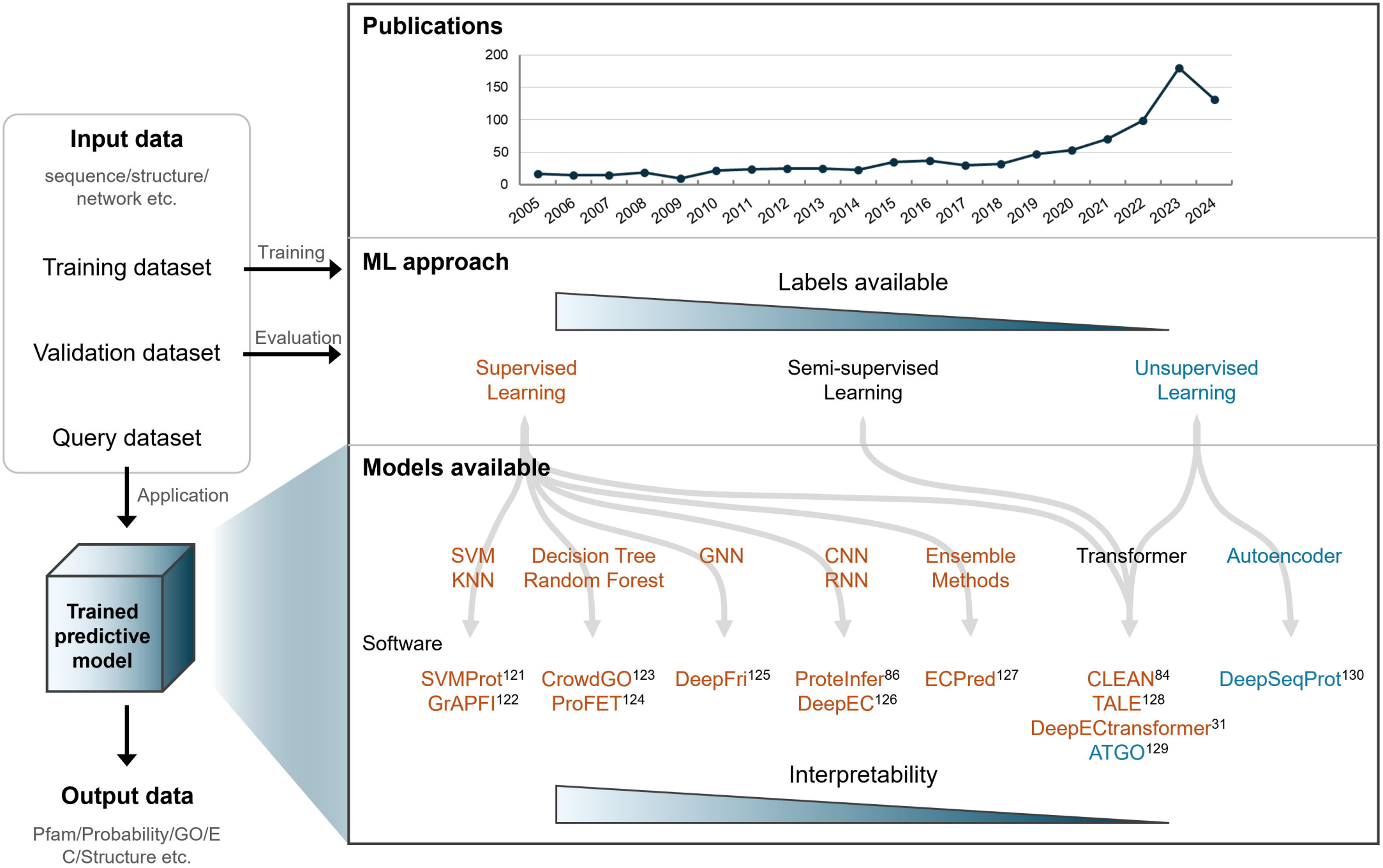
Figure 3. YciO harbors specific molecular surface features. (A) Structure-based multi-sequence alignment of TsaC proteins, the TsaC domains of Sua5 proteins and YciO proteins, derived from available crystal structures and AlphaFold structural models. For crystal structures, the PDB IDs are indicated 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 $_L$-Thr 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). Surface

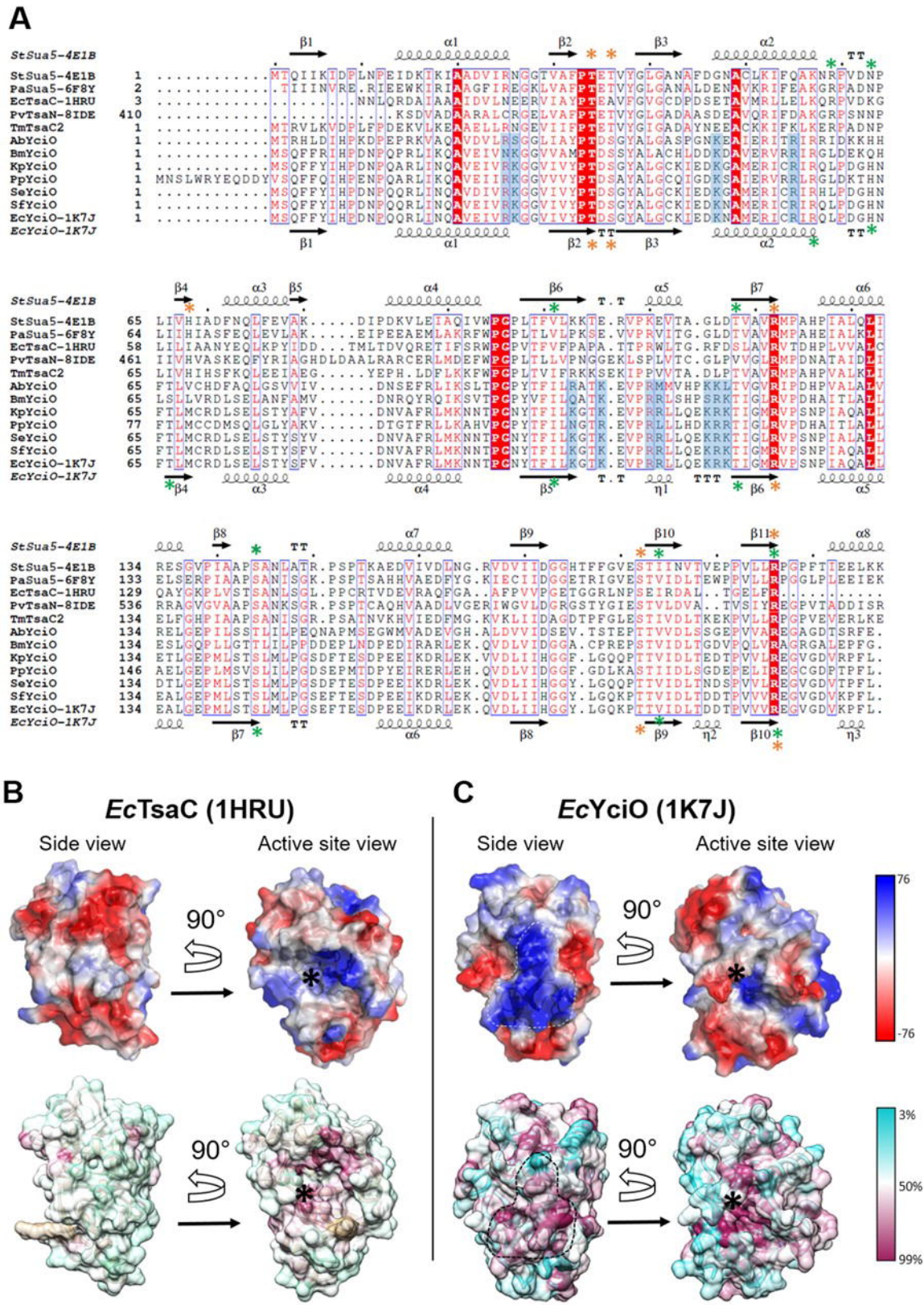
representations of the crystal structures of *Ec*TsaC (**B**) and *Ec*YciO (**C**), color-coded by electrostatic potential (top) and positional conservation score (bottom). The color keys for both panels are shown on the right. The conserved YciO-specific positively charged surface patch (6% of total molecular surface area) is encircled with a dashed line. The active center (or putative active center for YciO) is marked with an asterisk.

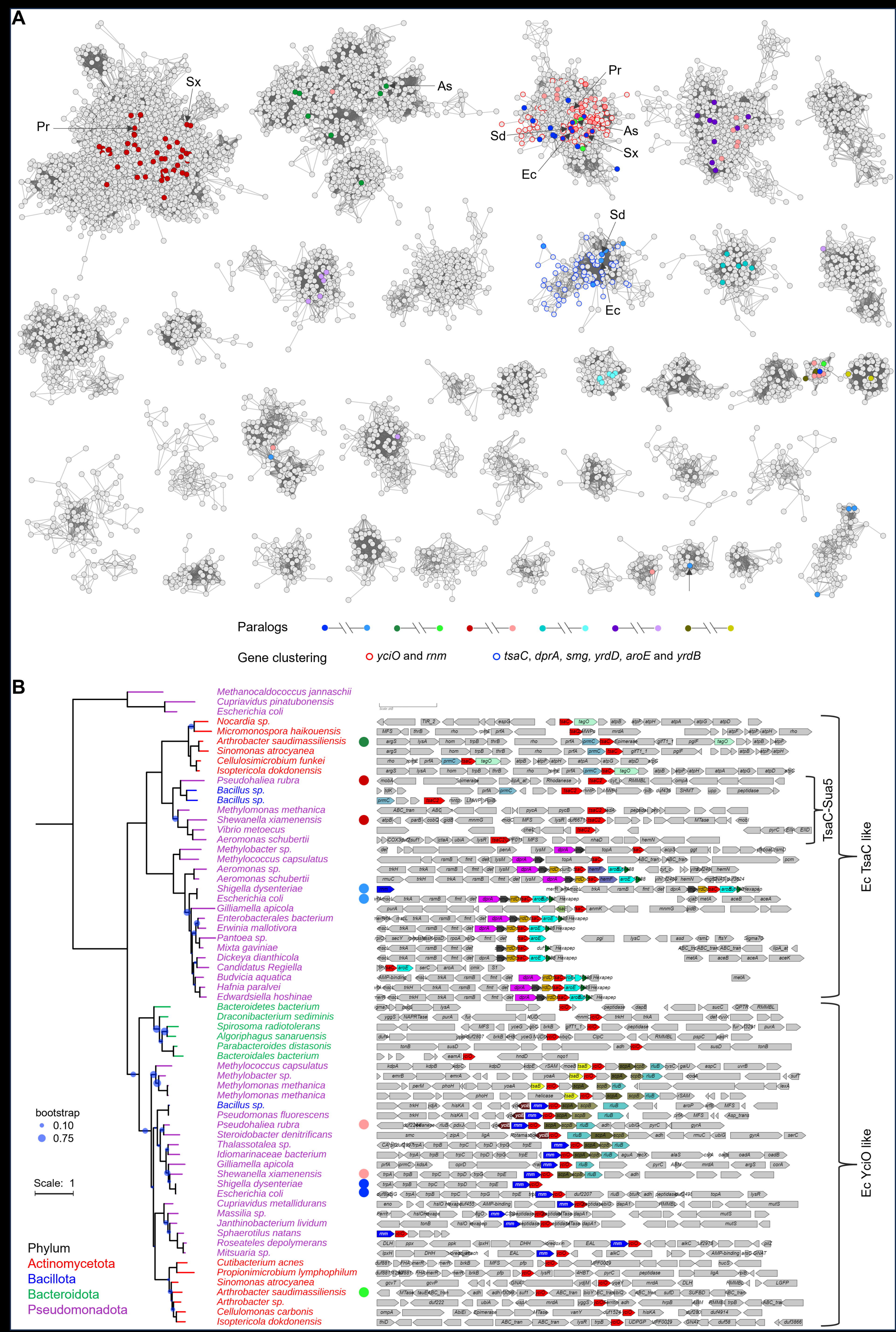
Figure 4. The genes encoding members of the *yciO* and *tsaC* have different genome neighborhood context. **(A)** Sequence Similarity Network of 54,820 PF01300 family members between 150 and 350 aa in length generated using EFI-EST. Each node in the network represents one or multiple sequences that share 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 S2. **(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 full genome neighborhood information is available Table S3. 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*;

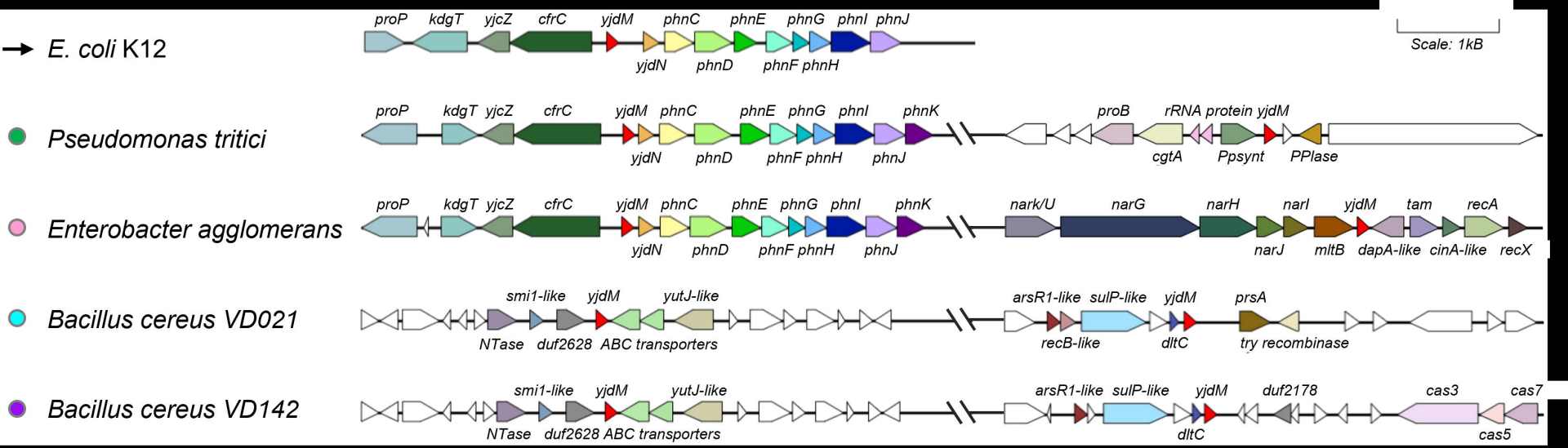
Figure 5. **Gene neighborhoods and metabolic reconstructions do not link YdjM to phosphonate degradation**. Most *yjdM* gene are not near phosphonate degradation operons. A Sequence Similarity Network of 11,986 IPR004624 family members was generated and the corresponding genome neighborhood information is available Table S4.


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


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
Input data
sequence/structure/ network etc.
Training dataset
Validation dataset
Query dataset
Training
Evaluation
Application
Trained predictive model
Output data
Pfam/Probability/GO/EC/Structure etc.
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
SVMProt121
GrAPFI122
CrowdGO123
ProFET124
DeepFri125
ProteInfer86
DeepEC126
ECPred127
CLEAN84
TALE128
DeepECtransformer31
ATGO129
DeepSeqProt130
Interpretability


A
StSua5-4E1B
StSua5-4E1B
PaSua5-6F8Y
EcTsaC-1HRU
PvTsaN-8IDE
TmTsaC2
AbYciO
BmYciO
KpYciO
PpYciO
SeYciO
SfYciO
EcYciO-1K7J
EcYciO-1K7J
B
EcTsaC (1HRU)
Side view
Active site view
90°
90°
C
EcYciO (1K7J)
Side view
Active site view
90°
90°
76
-76
3%
50%
99%


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


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