Rethinking bacterial relationships in light of their molecular abilities

Yannick Mahlich[1]*, Chengsheng Zhu[1,2], Henri Chung[3], Pavan K. Velaga[1], M. Clara De Paolis Kaluza[4], Predrag Radivojac[4], Iddo Friedberg[3], Yana Bromberg[1,5,6]*

Affiliations

[1] Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA.
[2] Xbiome Inc., 1 Broadway, 14th fl, Cambridge, MA 02142, USA.
[3] Interdepartmental program in Bioinformatics and Computational Biology and Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA 50011, USA.
[4] Khoury College of Computer Sciences, Northeastern University, 177 Huntington Avenue, Boston, MA 02115, USA.
[5] Department of Biology, Emory University, 1510 Clifton Road NE, Atlanta, GA 30322, USA.
[6] Department of Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA.
* Corresponding author: yana@bromberglab.org; ymahlich@bromberglab.org

Abstract

Determining the repertoire of a microbe's molecular functions is a central question in microbial genomics. Modern techniques achieve this goal by comparing microbial genetic material against reference databases of functionally annotated genes/proteins or known taxonomic markers such as 16S rRNA. Here we describe a novel approach to exploring bacterial functional repertoires without reference databases. Our *Fusion* scheme establishes functional relationships between bacteria and thus assigns organisms to Fusion taxa that differ from otherwise defined taxonomic clades. Three key findings of our work stand out. First, Fusion profile comparisons outperform existing functional annotation schemes in recovering taxonomic labels. Second, Fusion-derived functional co-occurrence profiles reflect known metabolic pathways, suggesting a route for discovery of new ones. Finally, our alignment-free nucleic acid-based Siamese Neural Network model, trained using Fusion functions, enables finding shared functionality of very distant, possibly structurally different, microbial homologs. Our work can thus help annotate functional repertoires of bacterial organisms and further guide our understanding of microbial communities.

Introduction

Exploring the molecular functional capabilities of microbes is key to understanding their lifestyles and contributions to the biogeosphere cycles that run our world(*1-6*). Microbial communities are often analyzed by taxonomically categorizing their members, defining their functional capabilities, and using this knowledge as a proxy for the community's overall functional abilities(*7-10*). The gold standard for taxonomic classification of newly sequenced organisms, and reclassification of existing ones, is DNA-DNA hybridization (DDH)(*11, 12*). DDH can be approximated using 16S rRNA similarity and bacterial morphology and physiology (*13-15*). More recent approaches analyze genome sequence properties such as average nucleotide identity and multilocus sequence similarity(*16-20*). These sequence-based methods promise to match DDH's taxonomic precision while being simpler and cheaper.

Notably, the above methods adopt a primarily phylogenetic view of bacterial relationships, assessing microorganisms' likely evolutionary lineage based on genetic similarity. Horizontal gene transfer (HGT), i.e. the exchange of genetic material across taxonomic lineages, complicates this approach to bacterial classification(*21-23*). HGT is the primary way for evolutionarily distant organisms to acquire similar functional capabilities encoded by similar sequences(*24-26*). Conversely, evolutionarily close sequence-similar organisms can functionally diverge under environmental pressure. Given a shift towards analyzing the functional capabilities of microbes(*8, 27-29*), i.e. "What are they doing?" instead of "Who are they?", one might ask the question "Are these bacteria functionally related?" as opposed to "Are they evolutionary cousins?" The former question can be answered well, if incompletely, by phenetic approaches based on, for example, differentiation of cell wall composition, guanine-cytosine content, and the presence of lipids amongst others (*30, 31*). We propose that genome-inferred bacterial functional annotations may further improve the resolution of these methods.

We previously developed Fusion, a method for evaluating microbial similarities based on shared functionality encoded in their genomes(*27, 32*). This approach revealed relationships between organism groups that are overlooked when using taxonomic or DNA similarity alone. Here, in addition to updating our classification scheme for a faster and more precise way of dealing with a flood of microbial genomes, we made five key discoveries: (1) We established that Fusion outperformed other function definitions in reconstituting the current state-of-the-art bacterial taxonomies (*33*) (*34*). Furthermore, using only a few common Fusion functions was sufficiently descriptive of these taxonomic assignments. (2) We also found that functional similarity could complement 16S rRNA sequence identity in assigning taxonomic classification. (3) In light of these findings, we proposed that a functional similarity-based classification scheme for Prokaryotes may be more robust than evolution-based taxonomic classifications. (4) We further found that collections of Fusion functions co-occurring within organisms highlight known metabolic pathways. We note that, unlike existing techniques (*35-38*) our approach allows for discovery of novel pathways. (5) Finally, we trained a Siamese Neural Network (SNN)(*39*) model to label two gene sequences as encoding proteins of the same Fusion function. In contrast to function transfer via sequence-derived homology, we expect that this model will be useful for further generalization of function concepts. We also note that this approach could potentially be optimized to label functional profiles of microbial metagenomes directly from sequencing reads, i.e. without the need of assembly or metagenomic binning(*40-43*).

Results & Discussion

Sequenced bacterial proteomes are significantly redundant. We retrieved from GenBank (*44, 45*) (Methods) a set of 8,906 genomes/proteomes of bacterial organisms representing 3,005 species. This set comprised all fully sequenced bacterial genomes available at the time of extraction (2018). It is notably redundant with 65% (5,754) of the proteomes belonging to only 25% (753) of the species. An extreme case of this observation is the 360 proteomes of *Bordetella pertussis,* contributing 360 copies of almost every *B.pertussis* protein (~3,620 proteins per proteome) to our collection. Overall, nearly 60% (18.8 of 31.6 million) of proteins in our set were identical to others. Of the ~15.6M sequence-

unique proteins in our set (*sequence-unique protein set*, Methods), ~2.8M (~18%) were found in multiple proteomes, while ~12.8M (82%) were proteome-specific.

As expected, much of the sequence redundancy occurred between strains of the same species, emphasizing the difficulty of distinguishing organism classes. When the set of organisms was phylogenetically balanced (*balanced organism set*, Methods), much of the protein redundancy was removed. This collection contained ~4.75 million proteins, of which 99% (~4.69M) were sequence- and organism- unique. We also note that these proteins still recapitulated nearly two thirds of the functions identified in the complete set of proteins (Methods). Most of the analyses presented here are based on the balanced organism set.

Fusion reflects and augments known functionality. We computed functional pairwise similarities (edges; using HFSP (*46*), Homology-derived Functional Similarity of Proteins) between sequence-unique proteins (vertices) and clustered the resulting network to determine the molecular functions likely carried out by proteins in our set (Methods; Fig. 1). We obtained 433,891 clusters of functionally similar proteins, dubbed *Fusion functions*, ranging in size from 2 to 118,984 proteins (Fig. S1).

This collection of Fusion functions, particularly the large number of small functions, i.e. containing few proteins, is contrary to expectations of functional diversity as compared to, e.g. 19,179 Pfam-A families/clans (Pfam v34, Methods) (*47*) and 11,185 molecular function GO terms (GeneOntology version 2021-09-01; Methods) (*48, 49*). This discrepancy between the annotations is likely accounted for by functional definitions. Pfam-A, for example, needs many sequences per family to build multiple sequence alignments (MSAs) for Hidden Markov Model (HMM) construction; thus, some of our functions may simply have not contained enough sequences to recapitulate a Pfam family. Furthermore, Pfam domains are not functionally precise as the same domain is often reused in different functions (*50-53*) and one protein can have more than one domain. Of the Fusion functions, only 15% (65,663) have at least 20 sequence-unique proteins, i.e. the lower limit for even the less-precise MSAs (*54*). Of these functions, 80% (52,678) contain proteins with one or more non-overlapping Pfam domains, i.e. ~1.6 domains per protein, 10,114 unique domains overall, and ~11 Fusion functions per domain. Of the smaller functions (size < 20 proteins; ~370K in total), 128,128 have at least one Pfam-A domain. We hypothesize that the remaining ~240K functions, not identifiable by Pfam, may be responsible for highly specific bacterial activity.

We calculated homogeneity (Eqn. 1) and completeness (Eqn. 2) for how well the Fusion functions (180,806 functions of >1 sequence) of proteins with at least one Pfam domain (12,611,237) compared to Pfam-A domain assignments (Methods). An optimal homogeneity (=1) would indicate that each function only contains proteins with one domain. An optimal completeness (=1) indicates that all proteins with a specific Pfam domain only fall into a single function. Neither optimal completeness nor heterogeneity are, as described above, possible for our data. However, both homogeneity (=0.9) and completeness (=0.79) were still fairly high for our data. That is, Fusion captured much of the Pfam-like functional diversity.

We further compared the Fusion functions with their respective Pfam domain sets, i.e. collections of Pfam domains without accounting for domain order in sequence (57,165 sets). This comparison marginally increased completeness (=0.8, homogeneity=0.94) as compared to single domain-based evaluations (completeness=0.78, homogeneity=0.9). Additionally considering domain order (91,113 arrangements), we observed that each Fusion function most

often only contained proteins of one arrangement (homogeneity =0.93) and further increased completeness over set comparisons (=0.81). Thus, while each Fusion function is highly specific to a given Pfam domain arrangements (high functional specificity), each domain set or arrangement might encode multiple functions.

While Pfam domain arrangements are more precise than individual domains, they do not always report experimentally defined functionality(*55*). Fusion functions are somewhat more precise. For example, the *Geobacter sulfurreducens* acyltransferases (R)-citramalate synthase (AAR35175, EC 2.3.1.13) and *Salmonella heidelberg* 2-isopropylmalate synthase (ACF66296, EC 2.3.3.182/2.3.3.21) have the same domain arrangement (HMGL-like pyruvate carboxylase domain, PF00682, followed by a LeuA allosteric dimerization domain, PF08502) but have a different 4th digit Enzyme Commission classification (EC) number (*56*), indicating their different substrate specificities. Notably, these proteins fall into two different Fusion functions. To evaluate Fusion functional mappings more broadly we collected, where available, the experimentally derived EC annotations for proteins in our set (4,206 proteins, 1,872 unique EC numbers) and measured the similarity of these with the corresponding 1,893 Fusion functions. Fusion functions more closely resembled annotations of enzymatic activity (homogeneity = 0.95, completeness = 0.94) than those of Pfam domains. This finding suggests that our Fusion functions capture aspects of molecular function better than domain-based annotations.

Organism functional profiles capture taxonomy. For each organism of the balanced organism set, we extracted Fusion, Pfam-A domain arrangement, and GO term functional profiles. Briefly, a functional profile is the set of functions of a single organism, e.g. the set of Pfam-A domain arrangements encoded by the proteins of that organism (Methods). On average, per organism Fusion, Pfam-A and GO term profiles were of size 2,133, 1,479, 776 (Fig. S2). For each organism pair, we computed profile similarity, i.e. the count of functions found in both profiles divided by the larger functional profile (Methods; Eqn. 4). On average, the (larger) Fusion-based functional profiles were less similar than the (smaller) Pfam and GO -based profiles (Fig. S3). A pair of organisms were predicted to be of the same or different taxon based on whether their similarity exceeded a set threshold ([0,1] in steps of 0.01). Predictions were compared against NCBI(*33*) and GTDB(*57*) taxonomies at six levels (phylum through genus; Methods). Note that we could not assess the species level, since no two organisms of the same species were retained in the balanced organism set.

Both Fusion and Pfam outperformed GO annotations in assessing taxonomic similarity. Fusion profiles were better than Pfam (Fig. S4), e.g. at 50% recall (Eqn. 5) of identifying two organisms of the same GTDB phylum, Fusion and Pfam achieved 75% and 48% precision (Eqn. 5), respectively. This advantage was also present across deeper taxonomic ranks (Fig. S4). We note that Fusion's improvement over Pfam did not stem from the difference in the number of functions per organism (profile/function-ome size) as the predictive power of the function-ome size was only marginally better than random (Fig. S4).

These findings suggest that organism similarity established via comparison of functional profiles carries taxonomy-relevant information. Furthermore, comparing functional capabilities can reveal organism relationships that microbial taxonomy, muddled by horizontal gene transfer, is unable to resolve.

Functional profiles are more informative of taxon identity than 16S rRNA. The genetic marker most frequently used for organism taxonomic classification is the 16S rRNA

gene(*14*) – a non-coding gene that, by definition, can not be captured by Fusion. To evaluate its predictive power, we extracted 16S rRNA sequences for each genome in our complete set and calculated sequence identity for all 16S rRNA pairs (Methods).

Sequence similarity between 16S rRNA pairs below 97% is generally accepted as an indication that the organisms are of different species(*58*). Indeed, we found that 98.7% (663.7M) of the 16S rRNA pairs that originate from different species fall below the 97% sequence identity threshold, while only 2% of same species pairs do (Fig. 2, Fig. S5). That is, below this sequence identity threshold nearly all (99.96%) sequence pairs were of organisms of different species, confirming the 97% threshold as a good measure of organism taxonomic difference.

Using the 97% sequence identity threshold as an indicator of taxon identity, however, is impossible. Trivially, many genomes have multiple 16S rRNA genes (*59*). In our set, 625 pairs of 16S rRNAs extracted from the same genome were less than 97% identical (minimum similarity =75.8%); in these cases, the marker gene similarity could not even identify the same genome, let alone same species. Furthermore, while almost all of same-species 16S rRNA pairs were ≥97% identical, nearly half of all pairs above this threshold belonged to different species (recall=98%, precision=55%, Fig. S6). In contrast, at the optimal Fusion organism functional profile similarity threshold of 75.5% (Eqn. 4; threshold established via peak F1-measure, Eqn. 6; Fig. S7), organisms were correctly identified to be of the same species with 80% precision (recall=94%, Fig. S4). At a matched level of recall, function comparisons were also more precise than 16S rRNA (75% vs. 55% precision, at 98% recall). Furthermore, Fusion achieved 95% precision for more than a third (35%) of the organism pairs, whereas 16S rRNA measures were this precise for less than a fifth (17%). The ability of 16S rRNA to identify organisms of the same genus at the commonly used threshold of 95% also left much to be desired (43% precision, 78% recall). Fusion performance was significantly better (90% precision, 70% recall) when using optimal functional similarity threshold (72.3%) established for this task.

Functional profiles augmented 16S rRNA in determining organism species. For example, for all organism pairs sharing ≥97% 16S rRNA identity, additionally requiring a Fusion functional similarity of 75.5% lead to an increased precision of 86% vs. 55% for 16S rRNA or 80% for Fusion similarity alone; recall was slightly decreased to 92% vs. 98% for 16S rRNA and 94% for Fusion alone. These findings suggest that functional similarity is orthogonal to 16S rRNA similarity in defining taxonomic identity.

We note that the lack of precision in 16S rRNA has implications for metagenomic analysis, where 16S rRNA abundance is often used to assess sample taxonomic composition and functional diversity. Fusion, on the other hand, is specifically designed to enable sequence-based functional annotations and could directly inform a microbiome's functional composition.

Few functions are sufficient to accurately identify taxonomy. Earlier studies argue that a small number of carefully chosen marker genes/protein families are sufficient to determine taxonomic relationships of bacteria (*57, 60*). However, to be comparable across organisms, these genes should be ubiquitously present. We investigated whether a subset of Fusion functions could correctly identify two organisms of the same taxon. To this end, we progressively subset the number of Fusion functions used to generate organism functional similarities (100k, 50k, 25k, 10k, 5k, 1k, and 500 functions). We used two approaches for

function selection: (1) we chose the functions based on how frequently they appeared in the balanced organism set and (2) randomly sampled from the whole pool of functions. Importantly, our approach was based on the presence or absence of specific functional abilities encoded by these genes rather than their sequence similarity. We found that just 1,000 common Fusion functions were sufficient to classify organism pairs into the same taxon, outperforming a "complete Pfam"-based approach (Fig. S8). The same was true for taxonomic levels of order through genus with a set of 5,000 randomly selected functions (Fig. 3).

We further evaluated the overlap between the selected Fusion functions and the marker genes used for GTDB (bac120) classification (*57, 60*) (Methods). Each of the largest 1,000 functions of our balanced organism set contained at least one protein associated with one of the 120 GTDB marker protein families. However, only slightly more than half (70 of the bac120) of the marker families were present in the 1,000 sets of 5,000 randomly selected Fusion functions. The remaining functions were most likely unique to individual organisms.

Modularity-based taxonomic classification reflects phylogeny. Conventional taxonomic classification schemes rely on morphological and genetic markers (NCBI) or phylogenetic analysis of genetic data (GTDB). Genetic similarity, however, is not evenly spread across different sections of the taxonomy. Assuring that taxonomic groups at a given level are equally diverse is thus a well-known consideration when developing a taxonomy. GTDB, for example, tries to address this issue by breaking up the NCBI taxonomy's polyphyletic taxa and reassigning organisms to taxonomic ranks higher than species in order to better represent genetic diversity at the individual level(*60*).

We clustered our organism functional similarity network, where organisms are vertices and edges represent Fusion functional similarity, to extract groups of functionally related organisms – *Fusion-informed taxa* (Methods). We propose that this community detection-based taxonomy reflects functional similarity and metabolic/environmental preferences, and thus captures bacterial functional diversity better than phylogeny driven taxonomies. This is especially important when investigating environmentally specialized bacteria, e.g. symbionts or extremophiles, which are more likely to undergo convergent evolution and be functionally similar to other members of their environmental niche than to their phylogenetic relatives.

We identified resolution thresholds that influence the size and granularity of the Fusion-taxa such that the results best reflected existing taxonomic groupings at different taxonomic levels (Fig. 4). Note that for our balanced organism set, this excluded species and genus levels, as this set lacks pairs of organisms identical at these levels. To evaluate the similarity between Fusion-taxa and GTDB phylum/class/order/family levels we calculated the V-measure using GTDB-taxon designations for organisms as reference labels and Fusion-taxa as predicted labels. The V-Measure is the harmonic mean between homogeneity, a measure reflecting the number of organisms in a Fusion-taxon that belong to the same GTDB-taxon, and completeness, a measure reflecting the number of organisms of a GTDB-taxon are found within one Fusion-taxon. A high V-measure indicates that both homogeneity and completeness are high. For Fusion-taxa classifications, we selected the Louvain(*61*) clustering resolutions attaining the highest V-measures (Fig. 4, Methods). The distributions of GTDB phylum through order taxa and the corresponding-level Fusion-taxon sizes were similar, i.e. Kolmogorov-Smirnov *p*-values for GTDB phylum vs. Fusion resolution(0.68) = 0.89, class

vs. resolution(0.68) = 0.78, and order vs. resolution(0.5) = 0.68. This observation suggests some similarity between the larger organism groups captured by Fusion and GTDB despite differences in their approach to establishing organism relationships. However, the GTDB family-level taxa sizes were different from the corresponding Fusion-taxa, i.e. Kolmogorov-Smirnov *p*-value GTDB family vs. Fusion resolution(0.36) = 0.01, highlighting the (expected) divergence between the functional and phylogenetic approaches at finer taxonomic resolutions.

Modularity-based taxonomy is robust to the addition of novel organisms. As new organisms are added to taxonomies, organism assignments may need to be restructured. Here, updating the number of organisms per taxon or adding a new taxon containing only the novel organisms is far easier than reshuffling organisms from one taxon to others. Fusion-taxa appear robust to addition of organisms, favoring the first outcome. To demonstrate this quality, we created 50,000 new organism similarity networks by adding *n* organisms to the balanced organism set clusters, i.e. 100 networks for each *n*, where *n* ranges from 1 to 500 organisms randomly selected from the complete organism set, but not contained in the balanced organism set; each network was of size of 1,503 to 2,002 organisms (balanced organisms set + *n*). We re-clustered all networks at resolution=0.5 (Methods), the resolution we previously determined to correspond best to the GTDB order-level classifications. The resulting clusters (predicted labels) of the balanced set organisms were compared to the original clusters (reference labels).

We expected that addition of these new organisms, selected from the complete set, and thus similar to those already in the network, would reflect the "worst case" scenario for network stability. That is, while new organisms could be expected to form their own clusters, microbes similar to those already in the network could stimulate cluster re-definition. Our function-based clustering did not change significantly upon addition of new (existing taxon) microbes, demonstrating the stability of the identified taxa (predicted vs. reference labels; with one added organism, median V-measure=0.99; with 500 added organisms: V-measure=0.96; Methods, Fig. S9).

To further evaluate the (likely limited) effects of introducing organisms of novel taxa, we extracted ten genomes added to GenBank after the date of our set extraction (February 2018) and whose GTDB order was not represented in our collection. We annotated the Fusion functional profiles of these organisms by running alignments, as in Zhu et al(*32*), against our set of proteins, computed organism similarities to the 1,502 microbes of our balanced set, and re-clustered the resulting network. Eight of these ten organisms each formed their own cluster, as expected. The two remaining organisms clustered into an already existing Fusion-taxon. Interestingly, this taxon contained an organism of the same NCBI order as the two new bacteria, illustrating the subjectivity of GTDB vs. NCBI taxonomies and highlighting the importance of organism assignment standardization.

When considered together, these observations suggest that functional similarity networks are stable when augmented with additional data points and present a viable alternative and/or addition to taxonomic classification of microorganisms.

Co-occurrence of functions informs joint participation in molecular pathways. Using the data from the balanced organism set, we assigned to each function a phylogenetic profile(*62*)(Fig. 1B). Each Fusion function was thus represented by a 1,502-length binary vector, where each entry reflected the presence or absence of the function in each organism

(Methods). We then calculated the Jaccard distance (Eqn. 7) between pairs of functions. Where available, we further annotated each function with the EC numbers of its member proteins; as above, most functions corresponded to only one EC. As a gold standard for our evaluations, we then retrieved 158 KEGG(*63, 64*) modules that encompassed at least three EC annotations resolving to Fusion functions (Methods). The median phylogenetic profile distance between pairs of Fusion functions (=0.63) co-occurring within any KEGG module was significantly lower (Wilcoxon Rank Sum, p-value $<2.2\text{x}10^{-16}$) than that of random (median distance=0.89) pairs (Fig. S10A). This observation supported our expectation that protein components of the same pathway have co-evolved in the same organism groups.

We note that the higher-than-expected distances between some functions co-occurring within a KEGG module were partially accounted for by functionally synonymous proteins (Fig S10B). That is, different proteins carrying out the same or similar molecular activity were likely part of different taxon-specific functional operons encoding the same generic molecular pathway. For example, the glycolysis module (M0001) enzymes phosphohexokinase (2.7.1.11) and pyrophosphate-fructose 6-phosphate 1-phosphotransferase (2.7.1.90) are functionally synonymous because they both of catalyze conversion of beta-D-Fructose 6-phosphate to beta-D-Fructose 1,6-bisphosphate. The phylogenetic profiles of these functions, however, were dissimilar (Jaccard distance = 0.83) as any given organism only uses one of these in its glycolytic pathway.

We also found that the median Jaccard distance between functions in a module reflected the combination of the number of organisms using the module, number of module enzymes, and the variance in function prevalence (Fig. 5). A lower Jaccard distance was expected of ubiquitous pathways, e.g. ribonucleotide synthesis (M00050, M00052; Figure 5A, bottom right corner) and small niche modules specific only to a few organisms, e.g. nitrification (M00528) and methanogenesis (M00567; (Figure 5A, bottom left corner). In contrast, pathways where some functions were more prevalent than others (coefficient of variation, CV, Eqn. 8) had a higher median distance. For example, the ectoine synthesis pathway (M00033) had a relatively high CV (1.2), partly due to the difference in prevalence of fructo-aldolase (EC 4.1.2.13, 407 organisms) and triose-phosphate isomerase (EC 5.3.1.1, 1,492 organisms; Figure 5A, red dot, upper left corner). These relationships were not observed for a set of randomized modules (Figure 5B).

Note that modules with a high median function distance and high median CV could differ from common modules by only a few enzymes. For example, the nitrification module M00804 (44 enzymes) differs from complete nitrification module M00528 (33 enzymes), solely by the absence of nitrate reductase (EC 1.7.5.1). However, this difference is enough to increase the median Jaccard distance from 0.39 in M00528 to 0.99 in M00804. Biologically, this is likely the result of divergence of the nitrification pathway in a small number of organisms, i.e. nitrate reductase is only found in nitrifying bacteria – a small subset of the original population. This observation suggests a means for tracking evolution of pathways via high median functional distances.

Machine learning-based sequence comparisons and sequence alignments capture different functional signals. We trained a Siamese Neural Network (SNN) to predict whether two nucleic acid (gene) sequences encode proteins of the same Fusion function. SNNs are specifically optimized to assess similarities of two objects (*65*) – in our case gene/protein functional similarity. This is critically different from traditional classifiers,

where the algorithm aims to predict which defined class an instance belongs to. In training (balanced set; ~300K gene pairs, 50% same vs. 50% different function), our model attained 73% overall accuracy at the default cutoff (score>0.5; area under the ROC curve, AUC_ROC =0.80). SNN prediction scores correlated with the precision of recognizing the pair's functional identity; thus, for example, at cutoff =0.98 the method attained 96% precision for the 19% of gene pairs that reached this threshold. Note that at this stringent cutoff, for an imbalanced test set with 10% same function pairs, the network still maintained high precision (82%) at a similar recall (24%). Importantly, increasing the size of the training data to one million gene pairs, improved the method performance (AUC_ROC = 0.81), suggesting that further improvements may be possible.

While somewhat correlated (Spearman rho=0.3, Fig. S11), the SNN similarity scores captured a different signal than the HFSP scores, i.e. values incorporating sequence identity and alignment length. Thus, a higher-dimensional representation of functional similarity of gene products beyond what can be detected through homology, may further improve functional annotations. To test this hypothesis, we compiled a set of Fusion functions where (1) the Fusion function was associated with only one EC number, (2) a number of different Fusion functions were associated with one EC number, and (3) different Fusion functions were associated with different EC numbers. As it was trained to do, SNN captured the similarity of genes from the first category (same Fusion function, same EC; Fig. S12 right green column, median SNN-score = 0.83) and the difference of the genes from the third category (different Fusion function, different EC; median SNN-score = 0.13; Fig. S12, left orange column). However, genes of the second category (different Fusion functions, same EC number) were scored significantly higher (median SNN-score = 0.7; Fig. 7, left green column) by the SNN than expected. We note that these different Fusion function gene pairs predicted to be of the same function would be considered false positives in SNN training. Thus, our SNN identified same enzymatic activity gene pairs that were NOT captured as same function by the homology-based Fusion.

Machine learning-based sequence comparisons and structure alignments capture orthogonal signals. What functional similarity does an SNN capture? We expected that functionally similar proteins that are not sequence similar should share structural similarity(*66, 67*). We compiled a set of Fusion proteins that have a structure in the PDB and then computed structural (TM-scores) and functional (SNN-scores) similarities for all pairs (Methods). Note that we did not use predicted protein structures(*68, 69*) to avoid compounding machine learning preferences.

First, we examined the relationship between the TM-score and SNN-score for sequence-similar protein pairs (HFSP score ≥0; Fig. S13). We found that 97% of these pairs (3,931 of 4,072) were structurally similar (TM-score ≥0.7; Table S1) and 94% (3,817) were predicted by the SNN to be of the same function (SNN-score ≥0.5; Table S2). These observations highlight HFSP's precision and confirm the expectation that high sequence similarity in most cases encodes for structural and functional identity.

It is worth noting that only a fifth (3,931 of 17,702) of all protein pairs with a TM-score ≥0.7 also had an HFSP≥0. SNN predictions, on the other hand, identified 77% (13,618 of 17,702) of the high TM-scoring pairs to be of the same function. Note that a quarter (3,544 of 13,618) of the SNN predictions had high reliability (SNN-score ≥ 0.98; Figure 6, Table S3) and many of these (2,028; 57%) were also sequence similar (HFSP≥0). These observations

suggest that function transfer by homology, while precise for the pairs it does identify, fails to find the more remote functional similarity of most protein pairs.

Most (73%, 21,412 of 29,213) of the reliably structurally dissimilar protein pairs (TM-scores <0.2 and excluding pairs that were filtered out by Foldseek(*70*), Methods) were predicted to be functionally different by SNN (score <0.5) and only 80 pairs (<1%) attained a high SNN score (≥0.98). Of pairs in the [0.2,0.5) and [0.5,0.7) TM-scores ranges, i.e. those that share minimal structural similarity, SNN labeled 45% and 53%, respectively, as having the same Fusion function; for both sets, only 4% reached SNN-score ≥0.98, which stands in contrast to the ~26% of the protein pairs with TM-score ≥0.7. These observations suggest that SNN, though not trained on protein structure, reliably identifies presence/absence of functional similarity at the extremes of structural similarity; it is significantly less certain for proteins that are only mildly structurally similar.

We further evaluated if protein pairs with known EC annotations (Methods) followed a similar structure-function relationship. As before (Fig. S12), we observed that the proteins of the same EC number were, on average, predicted with a higher SNN-score than different-EC pairs (Fig. S14). We then measured the ability of the SNN and the TM-score to predict the 3rd EC level of each protein pair. We found that while the SNN precision and recall were significantly above random, they were lower than simply using the TM-score (Figure 6). Importantly, we note that combining the TM and SNN predictions significantly improved recognition of genes of the same function. Adding an SNN-score evaluation of structurally similar protein pairs (TM-score ≥0.7) increased the precision to 90% at recall 30%. We thus suggest that the SNN reports a signal of functional similarity that is captured neither by sequence nor structure similarity alone.

To explore this signal further, we investigated outlier protein pairs in our set, i.e. structurally different (TM-score <0.2), sequence dissimilar (HFSP<0) pairs of proteins of the same 4th digit EC number attaining an SNNscore≥0.98, i.e. UniProt ids: P37870/P37871, O35011/O31718, and Q8RQE9 /P3787. For these, both TMAlign and the SNN were correct. That is, for each pair, the sequences were structurally different chains of the same heteromer structure (P37870/P37871 and O35011/O31718) or chains of different structures of the same protein complex (Q8RQE9/P37871) – all annotated with one EC number. While these three examples are anecdotal evidence they also clearly demonstrate the limitations of available chain-based functional annotations.

Going forward, our SNN can be further optimized and used for function prediction. We suspect that we will be able to create a functional ontology, combining Fusion functions that share a higher level of functional similarity not captured via sequence, or even structure, comparisons. We also see an exciting prospect for future use of our DNA-based predictor in metagenomics, where gene to fragment comparisons could potentially allow for forgoing assembly, to generate functional abundance profiles of microbial communities.

Summarizing the Findings. Understanding bacterial lifestyles requires describing their functional capabilities and critically contributes to research in medical, environmental, and industrial fields. The recent explosion in completely sequenced bacterial genomes has, simultaneously, created a deluge of functionally un-annotated and misannotated sequences and allowed for the development of new and informative sequence-based methods. Here, we optimized Fusion, a method for annotating the functional repertoires of bacteria, to recapitulate bacterial taxonomic assignments and create a novel functional taxonomy.

Importantly, we showed that bacterial functional profiles are significantly better at differentiating distinct species than 16S rRNA comparisons. We also found that using phylogenetic profiles of individual bacterial functions could provide insight into emergent functionality and potentially aid in the detection of novel metabolic pathways. Finally, we trained a Siamese Neural Network (SNN) to label pairs of genes whose product proteins are functionally similar. Notably, our SNN's ability to capture functional similarity signals that are orthogonal to sequence and structural signals may open the door to investigating remote homology. We propose that this method could elucidate a non-sequence or structure-driven functional ontology. Furthermore, it could potentially be optimized for extraction of functional annotation directly from metagenomic reads.

Materials and Methods

Microbial proteomes. We retrieved a set of microbial proteomes from GenBank (*44, 45*) (NCBI public ftp - ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria; February 28, 2018) and extracted the corresponding coding sequences from the complete bacterial genome assemblies. As per NCBI, complete assemblies are complete gapless genomic assemblies for all chromosomes, i.e. in bacteria, the circular genome and any plasmids that are present. Our resulting dataset thus contained the proteomes of 8,906 distinct bacterial genome assemblies with a total of 31,566,498 proteins (*full protein set*). We further redundancy reduced this set at 100% sequence identity over the complete length of the two proteins using CD-Hit (*71, 72*). Our *sequence-unique protein set* contained 15,629,432 sequences. Sequences shorter than 23 amino acids (1,345 sequences) were removed from the set as this length is insufficient to determine functional similarity between proteins (*46*). All further processing was done on the resulting set of 15,628,087 sequences. Of these, 12.78M were truly unique, i.e. proteins for which no 100%-identical sequence exists in the original full protein set; the remaining 2.85M sequences represented the nearly 16M proteins that were redundant across organisms in our set.

Computing protein functional similarities. Functional similarities between our sequence-unique proteins were assessed using HFSP(*46*). Specifically, we generated a set of all-to-all alignments with MMSeqs2(*73*) (evalue ≤ 1e-3, inclusion evalue ≤ 1e-10, iterations = 3). Note that due to the specifics of MMSeqs2, the two alignments for a every pair of proteins P_i and P_j, i.e. P_i-to-P_j and P_j-to-P_i, are not guaranteed to be identical and thus may have different HFSP scores. We chose to conservatively represent each protein pair by only one, minimum, HFSP value. For every protein pair, we retained in our set only the alignments where this HFSP value was ≥0; at this threshold HFSP correctly predicts functional identity of proteins with 45% precision and 76% recall (*46*). Any protein without predicted functional similarity to any other protein in the sequence-unique protein set was designated as having a unique function, i.e. true singletons (766,050 proteins). Of these, 57,646 sequences represented 127,543 proteins in the full protein set, while 708,404 were truly unique. The remaining 14,862,037 proteins were connected by ~22.2 billion functional similarities.

Generating *Fusion functions*. We built a functional similarity network using the 22.2B similarities (edges) of the 14.86M proteins (vertices) as follows: For any protein pair P_iP_j, an edge was included if (1) HFSP(P_iP_j) was ≥ 30 or if (2) HFSP(P_iP_j) ≥ 0.7*max(HFSP(P_iP_k), HFSP(P_jP_l)), where proteins P_k and P_l are any other proteins in our set; note that P_k and P_l can but don't have to be the same protein. The first cutoff at HFSP≥30,

ensured that our protein pairs were often correctly assigned same function (precision = 95%). Our second criterion aimed to assuage the much lower recall (10%) and capture more distant relationships while introducing as little noise as possible, i.e. only reporting functionally similar pairs at specifically-targeted, stricter HFSP cutoffs. The resulting network contained 14,130,628 vertices connected by 780,255,934 edges; 731,409 proteins were disconnected from the network, i.e. *putative* functionally unique singletons. The network was composed of multiple connected components, where the largest contained 481,801 proteins (distribution of component sizes in Fig. S1).

We used HipMCL(*74*) (High-performance Markov Clustering), an optimized version of Markov Clustering(*75, 76*), to further individually cluster the components of this network into functional groups. Note that as HipMCL requires a directed graph as input, we converted each edge in our data into a pair of directed edges of the same weight. The key parameters chosen for each HipMCL run were S=4000, R=5000, and inflation (I) =1.1. This clustering resulted in 1,432,643 protein clusters as well as 1,235 clusters containing only one protein, i.e. additional *putative singletons* for a total of 732,644.

Each of the 1,432,643 MCL clusters was further clustered using CD-Hit at 40% sequence identity (with default parameters). Note that only 7% of the MCL clusters contained more than one CD-HIT cluster. A total of 1,632,986 CD-Hit cluster representatives, i.e. longest protein in each CD-HIT cluster, were thus extracted. To this set of representatives, we added the putative singletons for a total of 2,365,630 proteins. These were used to generate a new functional similarity network by including all edges with $HFSP(P_iP_j) \geq 0$. Note that 226,346 (~10%) of these were not similar to any other representative proteins; of these, ~40k were originally designated putative singletons. The resulting functional similarity network comprised 2,139,284 vertices and ~303M edges. The network was re-clustered with HipMCL (S=1500, R=2000, I=1.4; smaller inflation values did not generate results due to MPI segmentation faults that could not be resolved) generating 438,130 Fusion functions.

Enzymatic function annotation. Information about protein enzymatic activity (Enzyme Commission, EC number(*56*)) was extracted from Swiss-Prot(*77, 78*) (June 2021) as follows: for each protein there had to be (1) experimental evidence for protein existence at protein level, (2) experiment-based functional annotation, and (3) only one EC number, fully resolved to all 4 levels. The resulting dataset was redundancy reduced at 100% sequence identity across the entire protein length. Swiss-Prot entries sharing the same sequence, but assigned different EC annotations, were excluded from consideration. The final data set contained 18,656 unique proteins and 4,269 unique EC annotations. The overlap between the EC data and the Fusion protein set (*Fusion enzyme set*) comprised 4,206 unique proteins in 1,872 unique EC annotations.

Pfam data. Protein mappings to Pfam(*47*) domains (Pfam-A version 34) were generated using pfamscan v1.4(*79*) with default values; in hmmscan(*80*) (hmmer v3.3), HMM evalue (-E = 10) and domain evalue (--domE = 10) were used. If the sequence hit multiple Pfam domains belonging to the same clan/family, only the clan was reported. For 12,720,756 sequence-unique proteins (85% of our 14.86M) the set of non-overlapping Pfam domains and their order in sequence were extracted, e.g. given domains X and Y, the domain arrangements 'XYY', 'XY' and 'YX' are regarded as three individual occurrences; the remaining 15% of the proteins did not match any Pfam-A domain. We thus identified 92,321 unique Pfam

domain arrangements. These corresponded to 58,021 domain sets, where the domain arrangements 'XYY', 'XY' and 'YX' resolve to only one domain set representation (X,Y).

Overlap between Fusion clusters and GTDB. In order to compare Fusion functions to the set of 120 marker proteins/protein families that GTDB uses (TIGRFAM & Pfam families) to establish taxonomic relationships between organisms (bac120), Fusion proteins were associated with TIGRFAM (release 15.0 – September 2014) & Pfam (PFAM-A version 34) domains using hmmscan (hmmer v3.3). Only one best TIGRFAM/Pfam hit (i.e. smallest e-value) was extracted per protein. Results were limited to hits with HMM evalue (-E = 1) and domain evalue (--domE = 10). Fusion functions were assigned the set of TIGRFAMs/Pfams according to their proteins matches. Finally, the overlap between domain associations of Fusion functions and the TIGRFAMs/Pfams used by GTDB as marker genes was evaluated.

GeneOntology annotations. GO(*48, 49*) "molecular function" annotations were extracted from the GO 2021-09-01 release. For each protein, its set of GO annotations included all protein self-annotations, as well as annotations of its parent nodes, i.e. other nodes connected via an "is a" edge up to the root of the molecular function subgraph. This resulted in 25,825 sets of GO terms for 7,313,428 (49% of 14.9M) sequences-unique proteins.

Comparing Fusion functions to existing functional annotations. We compared Fusion functions to EC and Pfam annotations by calculating the homogeneity (h, Eqn. 1), completeness (c, Eqn. 2) and V-Measure (v, Eqn. 3) (*81*) values using scikit/python (*82*). When comparing Fusion functions to, for example, EC numbers, homogeneity describes how often a Fusion function is associated with multiple EC numbers. That is, a high homogeneity (close to 1) signifies a clustering where most Fusion functions have an association to only one EC number. Completeness describes how often a specific EC number can be found in different Fusion functions. A high completeness (close to 1) indicates that for most ECs, a specific EC number is associated with only one or a small number of functions. V-Measure represents the harmonic mean between homogeneity and completeness. A V-measure of 1 is indicative of an optimal clustering, where each function is only associated with one EC number, and an EC number is only associated with this one function.

(Eqn. 1)
$$h = \begin{cases} 1 & if\ H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & else \end{cases}$$
where

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\,log\,\frac{a_{ck}}{\sum_{c=1}^{|C|}a_{ck}}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|}a_{ck}}{n}\,log\,\frac{\sum_{k=1}^{|K|}a_{ck}}{n}$$

(Eqn. 2) $$c = \begin{cases} 1 & if\ H(K,C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & else \end{cases}$$ where

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}}$$

$$H(K) = -\sum_{k=1}^{|K|}\frac{\sum_{c=1}^{|C|} a_{ck}}{n} log \frac{\sum_{c=1}^{|C|} a_{ck}}{n}$$

(Eqn. 3) $$v = \frac{2*h*c}{h+c}$$

Taxonomy information. Our taxonomic analyses were conducted on the basis of two taxonomy schemes: the NCBI taxonomy(*33*) and the GTDB(*57*) (genome taxonomy database). NCBI taxonomy rank information for each assembly was retrieved during protein dataset extraction (Feb 2018) and is available for all 8,906 organisms in our set. GTDB taxonomy information was extracted from GTDB release rs202 (April 2021). Genbank assembly ids were mapped to bacterial assemblies available in GTDB. GTDB taxonomy information is available for 99% (8,817) of the organisms.

Balancing the assembly set. According to GTDB, the 8,906 assemblies/organisms in our set belong to 3,005 species. Of these species, 2,252 (75%) have only one associated organism, whereas others have hundreds; e.g. *E. coli* and *B. pertussis* have 472 and 360 assemblies, respectively. We generated a balanced organism set to reduce this unevenness. First, we reduced our full set of 8,906 assemblies to retain the 3,012 genomes that were representative of strains included in the GTDB bac120 phylogenetic tree. Note that of these, 2,206 genomes were in both GTDB and our data, while 806 genomes were not present in our set and were represented by other assemblies of these same strains. Using dendropy(*83*), we then extracted from the full GTDB bac120 tree (47,895 organisms) a subtree containing only these 3,012 representatives while retaining the original branch lengths. We used Treemmer(*84*) to determine which leaves to retain in our set such that the RTL (relative tree length) of the pruned tree was ≥0.90. RTL is used as an indicator of retained genetic diversity after pruning, reflected as the sum of all branch lengths in the pruned tree in relation to the full tree. We thus selected 1,502 assemblies (further referenced to as the *balanced organism set*) – a minimum set of organisms that retains at least 90% genetic diversity present in our complete set of 8,906 assemblies.

Computing organism functional similarity. Each organism in our set can be represented by a functional profile, i.e. a set of corresponding Fusion functions, Pfam domains, or GO annotations. Functional similarity between the function-omes of two organisms, F_i and F_j, was calculated, as previously described (*27, 32*), by dividing the number of their shared functions by the size of the larger of the two profiles (Eqn. 4).

(Eqn. 4) $$FuncSim(F_i, F_j) = \frac{|F_i \cap F_j|}{\max(|F_i|,|F_j|)}$$

Fusion functional profiles for similarity calculations were generated at Fusion Level 1 with and, separately, without the inclusion of singletons. Pfam functional profiles were generated using Pfam domain arrangements and, separately, domain sets, as described above. GO functional profiles were generated using the GO terms extracted per proteins as described above. Note that Pfam and GO annotations are not available for all proteins, but each protein has an associated Fusion function. Thus, each method-based functional profiles (i.e. GO vs Pfam vs Fusion) of a single organism could be based on different sets of proteins.

We computed the precision/recall (Eqn. 5) values for correctly identifying two organisms as being of the same taxonomic rank based on their shared functional similarity. This was done at each taxonomic rank (phylum, class, order, family, genus, species) for both taxonomic definitions (NCBI and GTDB) and using a series of similarity thresholds ranging from 0 to 1 in increments of 0.01.

(Eqn. 5) $$Precision = \frac{TP}{TP+FP};\ Recall = \frac{TP}{TP+FN}$$

Here any pair of two organisms of the same taxonomic classification above the chosen threshold are true positives (TP), whereas pairs below the threshold are false negatives (FN). Any pair of two organisms of different taxonomic classifications above the similarity threshold are false positives (FP), while pairs below are true negatives (TN).

Grouping organisms by functional similarity. An organism similarity network was generated using Fusion functional profiles. Here assemblies (vertices) were connected by Fusion functional similarity edges; the resulting network is complete (all-to-all edges are present) as any two organisms share some similarity. We used Louvain clustering (*61*) to identify organism groups; implemented in 'python-louvain' (https://github.com/taynaud/python-louvain), an extension to 'networkx' (https://networkx.org). Organism groups at varying levels of granularity were generated by varying the resolution threshold parameter of Louvain clustering (resolution 0 to 1.5 in increments of 0.01), where larger resolution values lead to fewer but larger clusters. The V-measures (Eqn. 3) of the resulting partitions ("predicted labels") vs. GTDB taxa (reference labels) were calculated.

16S rRNA extraction and similarity calculations. 16S rRNA sequences were extracted from the NCBI GenBank database for 8,479 of the 8,906 organisms (427 organisms were missing annotated 16S rRNAs). From RDP (Ribosomal Database Project, v11.5)(*85*), we further extracted all 16S rRNA sequences and their corresponding multiple sequence alignment (MSA). The 16S rRNAs of the 8,479 organisms that were not contained in the RDP MSA were added using Infernal 1.1.4 (*86*) and the RDP bacterial covariance model. Using the resulting MSA we extracted gapless pairwise sequence identities for all 16S rRNA pairs (i.e. 683,261,061 pairs between 36,967 16S rRNA sequences).

We calculated the optimal F-measure (Eqn. 6) for both identifying organisms of the same species/genus using measures of 16S rRNA identity and Fusion organism similarity (Eqn. 4). Here, true positives (TP) are organisms of same taxon, attaining an identity or similarity measure at or above the chosen threshold, false negatives (FN) are organisms of same taxon but scoring below the threshold, and false positives (FP) are organisms of different taxa and scoring at or above the threshold.

(Eqn. 6) $$F1 - measure = \frac{2\times TP}{2\times TP+FP+FN}$$

Fusion function phylogenetic profiles. For all functions found in at least five organisms of the balanced organism set (1,502 organisms total), we created a profile indicating all assemblies containing a protein assigned to the function, akin to the Pelligrini et al study(*62*). Each functional profile was thus a 1,502-length binary vector; i.e. the presence or absence of the Fusion function in each organism was indicated with a 1 or 0. Furthermore, to be considered, each function had to have >5% of its proteins either belong to or have an HFSP score >20 with a protein in the Fusion enzyme set. Jaccard distance D_j was calculated for every pair of profiles F_1 and F_2 (Eqn. 7).

(Eqn. 7) $$D_j(F_1, F_2) = 1 - \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

Note that any pair of functions which had the same 4th level EC digit and a profile Jaccard distance >0.80, i.e. same enzyme found in very different organisms, was likely to represent only homologs of only slightly different functions; as such, profiles of these functions were merged. Jaccard distances were then recalculated for all resulting pairs of functions. This process was repeated until no two functions which match these conditions remained. The final profile matrix consisted of 1,420 functions, each represented by a 1,502-length profile vector.

For pairs of functions which co-occurred in a KEGG module, we calculated the coefficient of variation (CV) to assess the dispersion of the Jaccard Distances (Eqn. 7) in a module (Eqn. 8). A higher CV indicates that the functions present are found in different sets of organisms, while a low CV indicates that all functions in a module are found in nearly the same organisms.

(Eqn. 8) $$CV = \frac{\sigma}{\mu}$$

KEGG module annotations for Fusion functions. From KEGG (Kyoto Encyclopedia of Genes and Genomes)(*63, 64*), using R and the web scraping packing "rvest", we extracted the 280 unique KEGG modules and their corresponding enzymes (4th level EC numbers) found in our balanced organism set. We filtered these modules to retain those with at least three EC annotations mappable to Fusion functions. The resulting 158 modules were used for further analysis. Any pair of functions participating in the same module were labeled as co-occurring. To create a random set, we selected function pairs that were not present in the same module. We evaluated the median profile distances between co-occurring and random function pairs (Z-test at $\alpha = 0.05$, performed by bootstrapping subsamples of both sets of function pairs a thousand times). Note that a Fusion function may be mapped to more than one EC number and the same EC number may be assigned to more than one function. If a function was annotated with multiple EC numbers shared in a single module, the distance between the function and other shared functions was only considered once. A null set of profile distances was created by randomly permuting the EC numbers assigned to each module.

Machine learning-based predictor of shared protein functionality. We trained a Siamese Neural Network (SNN) (*39*) predictor to assess whether any two DNA sequences encoded proteins of the same Fusion function. SNNs are a class of neural network architectures that contain two identical subnetworks, i.e. the networks have the same configuration with the same parameters and weights. This type of network is often used to find the similarity of the inputs – in our case, two sequences encoding proteins of the same function. Because SNNs identify similarity levels, rather than predicting specific classes of

each input, they require significantly less data for training and are less sensitive to class imbalance. The latter was particularly a benefit here because the number of sequence pairs of different functions necessarily drastically exceeds the number of pairs of the same function. Additionally, as SNNs output a similarity metric rather than a probability score, they are likely specifically informative of the various levels of functional similarity, e.g. for a given pair of enzymes, whether two genes act upon the same bond vs. whether they use the same electron donor.

To train the model, we extracted 70 random Fusion functions, each containing at least ten different proteins from our sequence-unique set. The set of functions was split 50/10/10 for training, testing and validation. For training and validation, we balanced the dataset, i.e. we randomly selected gene sequence pairs such that 50% of the pairs included genes of same Fusion function and 50% were of different function. The final training set contained 20M gene sequence pairs generated from 29,907 sequences, the validation set contained 200,000 pairs and 9,982 sequences respectively. In testing we used balanced as well as imbalanced data sets. The imbalanced test set was generated to better resemble real-world data with a split of 90%/10% where 90% of the sequence pairs are between sequences of different function. The test set contained 100,000 sequence pairs generated from 1,000 gene sequences.

We tokenized protein-encoding genes to codons, i.e. split into non-overlapping 3-nucleotide chunks of sequence and projected each token into the LookingGlass(*87*) embedding space (length=104). The embeddings were then processed via an LSTM (*88*) and further used in SNN training. Note that at most the first 1,500 tokens were embedded per sequence. For sequences shorter 1,500 codons, the embedding vector was zero padded, i.e. any position in the vector after the last token was set to 0. The model was trained and validated in 50 iterations on our balanced training/validation data set. After 50 iterations performance of the model reached a precision of 0.72 and recall of 0.72 on the validation set at the default threshold of 0.5. The final model was tested on the imbalanced (90/10 split different/same function sequence pairs) attaining a precision of 0.22 and recall of 0.80 at the default prediction score cut-off of 0.5.

To further evaluate the model, we extracted a set of Fusion functions associated with only one level 4 EC annotation, but where the EC annotation was associated with multiple Fusion functions. We then predicted SNN scores for three sets of protein pairs: (1) proteins from the same Fusion function and same EC annotation, (2) proteins from different Fusion functions and same EC annotation, and (3) proteins of different Fusion functions and different EC annotation.

Structural alignments of Fusion proteins. We extracted from the PDB(*89, 90*) (May 2022) the available structure information for proteins in our set, i.e. 79,464 chains/entities mapping to 5,153 protein sequences in our sequence-unique protein set. Where multiple PDB structures mapped to one protein sequence we selected the PDB entry with the best resolution (lowest Å). For this set, we used foldseek(*70*) (–alignment-type 1, --tmscore-threshold 0.0) to identify structure pair TMscores(*91*) from TM-align(*92*). When a protein sequence pair resolved to multiple PDB entity (chain) pairs we selected the entity pair with the highest TMscore. Note that Foldseek was unable to generate TMscores for 498 PDB structures (mapping to 1,005 protein sequences) due to computational limitations and we excluded any structural/protein pair that included one of these from further consideration.

For the resulting 8,527,385 protein pairs we generated SNN prediction scores. For 8,080,324 of 8,527,385 (95%) pairs no TM-scores could be generated as they did not pass the pre-filtering step of Foldseek, i.e. they had no similar folds at all; for these we assumed a TMscore = 0. Notably, 143,347 (1.7%) of these were still predicted by the SNN to have high functional similarity (SNNscore≥0.98); we assume this percentage to be the approximate error rate of the SNN.

We also created subsets of PDB entity pairs where each protein was annotated with an E.C. number, i.e. proteins extracted for the Fusion enzyme set.

References

1. M. J. Blaser *et al.*, Toward a Predictive Understanding of Earth’s Microbiomes to Address 21st Century Challenges. *mBio* **7**, e00714-00716 (2016).
2. P. G. Falkowski, T. Fenchel, E. F. Delong, The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* **320**, 1034-1039 (2008).
3. A. Jousset *et al.*, Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME Journal* **11**, 853-862 (2017).
4. J. A. Russell, N. Dubilier, J. A. Rudgers, Nature's microbiome: introduction. *Molecular Ecology* **23**, 1225-1237 (2014).
5. Y. Bromberg *et al.*, Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. *Science Advances* **8**, eabj3984 (2022).
6. A. Shade, Understanding Microbiome Stability in a Changing World. *mSystems* **3**, (2018).
7. F. Beghini *et al.*, Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
8. E. A. Franzosa *et al.*, Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* **15**, 962-968 (2018).
9. J. Kaminski *et al.*, High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLOS Computational Biology* **11**, e1004557 (2015).
10. E. Bolyen *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**, 852-857 (2019).
11. E. Stackebrandt, B. M. Goebel, Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International journal of systematic and evolutionary microbiology* **44**, 846-849 (1994).
12. D. J. Brenner, G. R. Fanning, K. E. Johnson, R. V. Citarella, S. Falkow, Polynucleotide sequence relationships among members of Enterobacteriaceae. *J Bacteriol* **98**, 637-650 (1969).
13. D. R. Boone, R. W. Castenholz, G. M. Garrity, J. Stanley, *Bergey's Manual® of Systematic Bacteriology: Volume One The Archaea and the Deeply Branching and Phototrophic Bacteria*. (Springer, 2001).
14. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**, 4576-4579 (1990).

15. J. S. Johnson *et al.*, Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 5029 (2019).

16. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, Toward a More Robust Assessment of Intraspecies Diversity, Using Fewer Genetic Markers. *Applied and Environmental Microbiology* **72**, 7286-7293 (2006).

17. K. T. Konstantinidis, J. M. Tiedje, Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* **10**, 504-509 (2007).

18. D. Gevers *et al.*, Re-evaluating prokaryotic species. *Nature Reviews Microbiology* **3**, 733-739 (2005).

19. R. Rosselló-Mora, Updating Prokaryotic Taxonomy. *Journal of Bacteriology* **187**, 6255-6257 (2005).

20. D. Gevers *et al.*, Stepping stones towards a new prokaryotic taxonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1911-1916 (2006).

21. E. Hilario, J. P. Gogarten, Horizontal transfer of ATPase genes — the tree of life becomes a net of life. *Biosystems* **31**, 111-119 (1993).

22. A. Babić, A. B. Lindner, M. Vulić, E. J. Stewart, M. Radman, Direct Visualization of Horizontal Gene Transfer. *Science* **319**, 1533-1536 (2008).

23. N. Goldenfeld, C. Woese, Biology's next revolution. *Nature* **445**, 369-369 (2007).

24. M. N. Price, P. S. Dehal, A. P. Arkin, Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biology* **9**, R4 (2008).

25. Z. He, S. Xu, S. Shi, Adaptive convergence at the genomic level—prevalent, uncommon or very rare? *National Science Review* **7**, 947-951 (2020).

26. M. R. Farhat *et al.*, Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature Genetics* **45**, 1183-1189 (2013).

27. C. Zhu, T. O. Delmont, T. M. Vogel, Y. Bromberg, Functional basis of microorganism classification. *PLoS Comput Biol* **11**, e1004472 (2015).

28. G. Rastogi, R. K. Sani, in *Microbes and Microbial Technology: Agricultural and Environmental Applications,* I. Ahmad, F. Ahmad, J. Pichtel, Eds. (Springer New York, New York, NY, 2011), pp. 29-57.

29. M. G. I. Langille *et al.*, Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814-821 (2013).

30. K. H. Schleifer, Classification of Bacteria and Archaea: Past, present and future. *Systematic and Applied Microbiology* **32**, 533-542 (2009).

31. J. M. Young, Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy. *International Journal of Systematic and Evolutionary Microbiology* **51**, 945-953 (2001).

32. C. Zhu, Y. Mahlich, M. Miller, Y. Bromberg, Fusion DB: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic acids research* **46**, D535-D541 (2018).

33. C. L. Schoch *et al.*, NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* **2020**, (2020).

34. D. H. Parks *et al.*, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785-D794 (2021).

35. J. Zimmermann, C. Kaleta, S. Waschina, gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology* **22**, 81 (2021).

36. H. Wang *et al.*, RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Computational Biology* **14**, e1006541 (2018).

37. P. D. Karp *et al.*, Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* **17**, 877-890 (2015).

38. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research* **46**, 7542-7553 (2018).

39. J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems* **6**, (1993).

40. J. N. Nissen *et al.*, Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* **39**, 555-560 (2021).

41. S. Pan, C. Zhu, X.-M. Zhao, L. P. Coelho, A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications* **13**, 2326 (2022).

42. D. D. Kang *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

43. Y.-W. Wu, B. A. Simmons, S. W. Singer, MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2015).

44. D. A. Benson *et al.*, GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).

45. E. W. Sayers *et al.*, GenBank. *Nucleic Acids Res* **47**, D94-d99 (2019).

46. Y. Mahlich, M. Steinegger, B. Rost, Y. Bromberg, HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics* **34**, i304-i312 (2018).

47. S. El-Gebali *et al.*, The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427-D432 (2018).

48. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**, D325-D334 (2021).

49. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

50. A. Barrera, A. Alastruey-Izquierdo, M. J. Martín, I. Cuesta, J. A. Vizcaíno, Analysis of the Protein Domain and Domain Architecture Content in Fungi and Its Application in the Search of New Antifungal Targets. *PLOS Computational Biology* **10**, e1003733 (2014).

51. E. V. Koonin, Y. I. Wolf, G. P. Karev, The structure of the protein universe and genome evolution. *Nature* **420**, 218-223 (2002).

52. M. Itoh, J. C. Nacher, K.-i. Kuma, S. Goto, M. Kanehisa, Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biology* **8**, R121 (2007).

53. S. G. Peisajovich, J. E. Garbarino, P. Wei, W. A. Lim, Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination. *Science* **328**, 368-372 (2010).

54. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412-D419 (2020).

55. P. Radivojac, Advancing remote homology detection: A step toward understanding and accurately predicting protein function. *Cell Syst* **13**, 435-437 (2022).

56. A. Bairoch, The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-305 (2000).

57. D. H. Parks *et al.*, A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 1-8 (2020).

58. R. Rosselló-Mora, R. Amann, The species concept for prokaryotes. *FEMS Microbiology Reviews* **25**, 39-67 (2001).

59. T. Větrovský, P. Baldrian, The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* **8**, e57923 (2013).

60. D. H. Parks *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996-1004 (2018).

61. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

62. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**, 4285-4288 (1999).

63. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

64. M. Kanehisa, Y. Sato, M. Kawashima, KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* **31**, 47-53 (2022).

65. D. Chicco, in *Artificial Neural Networks,* H. Cartwright, Ed. (Springer US, New York, NY, 2021), pp. 73-94.

66. E. Krissinel, On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **23**, 717-723 (2007).

67. B. Rost, Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection* **12**, 85-94 (1999).

68. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

69. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).

70. M. van Kempen *et al.*, Foldseek: fast and accurate protein structure search. *bioRxiv*, 2022.2002.2007.479398 (2022).

71. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

72. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

73. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).

74. A. Azad, G. A. Pavlopoulos, C. A. Ouzounis, N. C. Kyrpides, A. Buluç, HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Res* **46**, e33 (2018).

75. S. M. Van Dongen, (2000).

76. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).

77. A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med (Berl)* **75**, 312-316 (1997).

78. T. U. Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489 (2020).

79. F. Madeira *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research* **47**, W636-W641 (2019).

80. S. R. Eddy, Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).

81. A. Rosenberg, J. Hirschberg, in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. (2007), pp. 410-420.

82. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

83. J. Sukumaran, M. T. Holder, DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569-1571 (2010).

84. F. Menardo *et al.*, Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).

85. J. R. Cole *et al.*, Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**, D633-642 (2014).

86. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

87. A. Hoarfrost, A. Aptekmann, G. Farfanuk, Y. Bromberg, Shedding Light on Microbial Dark Matter with A Universal Language of Life. *bioRxiv*, (2020).

88. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory. *Neural Computation* **9**, 1735-1780 (1997).

89. S. K. Burley *et al.*, RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research* **49**, D437-D451 (2020).

90. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

91. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702-710 (2004).

92. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).

Acknowledgments

We want to thank Ariel Aptekmann, Maximilian Miller and Zishou Zheng (all Rutgers University) for the valuable feedback during the concept phase of the project. Many thanks to Martin Steinegger (Seoul National University) for assistance with MMSeqs2 and Foldseek. Lastly, we extend our gratitude to everyone in the scientific community for providing the tools, databases and datasources that were vital in producing this research.

Funding: This work was supported by:
National Science Foundation CAREER award 1553289 (YB, CZ, PKV, HC, MCP and YM)
NIH (National Institutes of Health) grant R01 GM115486 (YB and YM)
Iowa State University's Translational Artificial Intelligence Center (IF)

Author contributions:

Conceptualization: YM, YB, CZ
Methodology: YM, HC, MCP, PKV
Investigation: YM, YB, HC, MCP, IF, PR
Visualization: YM, HC, MCP
Writing—original draft: YM, YB
Writing—review & editing: YM, YB, IF, PR, HC, MCP, CZ

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text, the Supplementary Materials or referenced permanent online data repositories: Function dataset: 10.6084/m9.figshare.21599544, Organism similarities : 10.6084/m9.figshare.21637988, Structural similarities: 10.6084/m9.figshare.21637937, Jupyter & R notebooks containing the analysis of datasets: https://bitbucket.org/bromberglab/fusion-manuscript-analysis/, git-repository containing the code to generate the Fusion functions: https://bitbucket.org/bromberglab/fusion-updater/

Figures

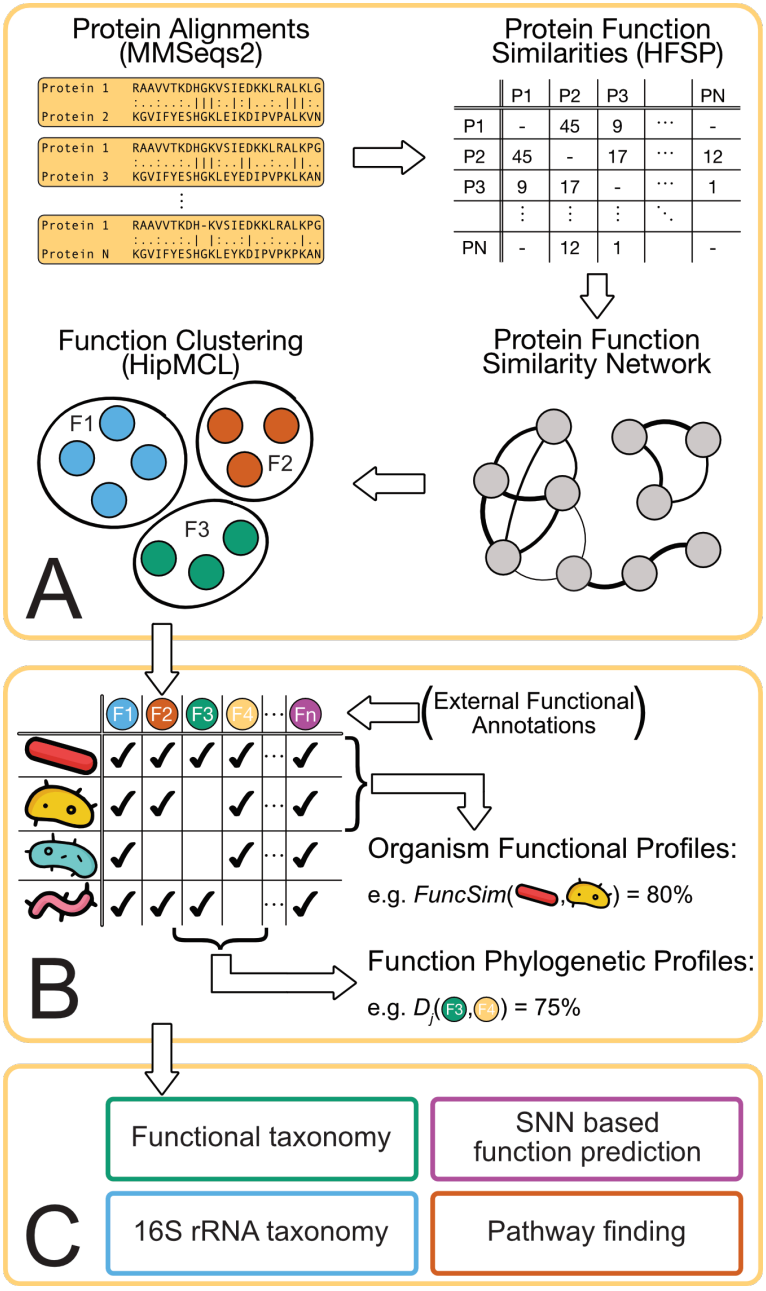

Fig. 1. Fusion workflow. (A) Fusion Functions are a result of an all-against-all protein alignment between all ~15.6M proteins in our set. **(B)** Organism (row-wise) comparisons net the organism functional profile similarities, while column-wise comparisons yield the functional phylogenetic profile similarities. **(C)** Analyzing organism similarities results in the functional taxonomy and contributes to the 16S rRNA analyses. Pathway finding uses functional profiles, while SNN function prediction relies on protein function annotations.

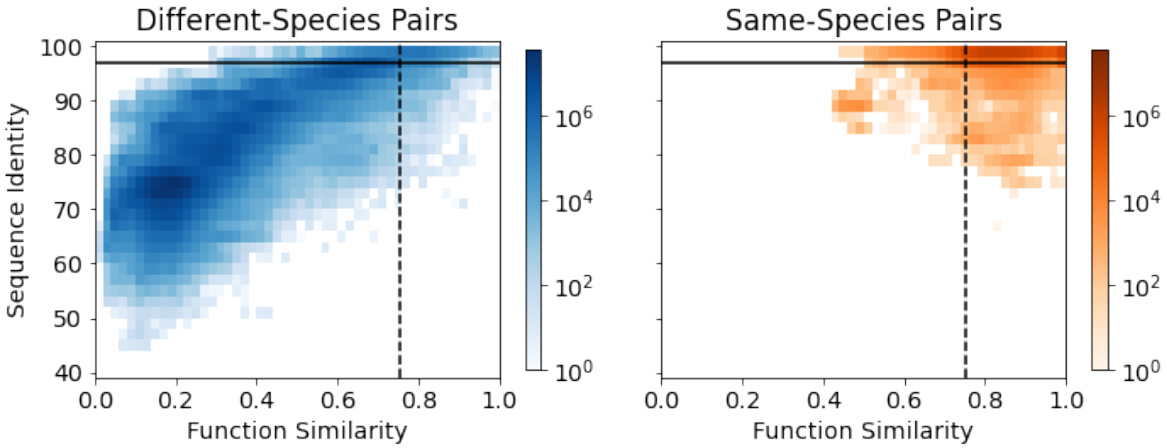

Fig. 2. 16S rRNA identity and functional similarity capture different taxonomic patterns. Density plots capture the location of pairs of different species (left, blue) and same species (right, orange) organisms in the space defined by the 16S rRNA identity (y-axis) and Fusion similarity (x-axis). Horizontal solid and vertical dashed lines represent the 16S rRNA and Function similarity thresholds of 97% and 75.5%, respectively.

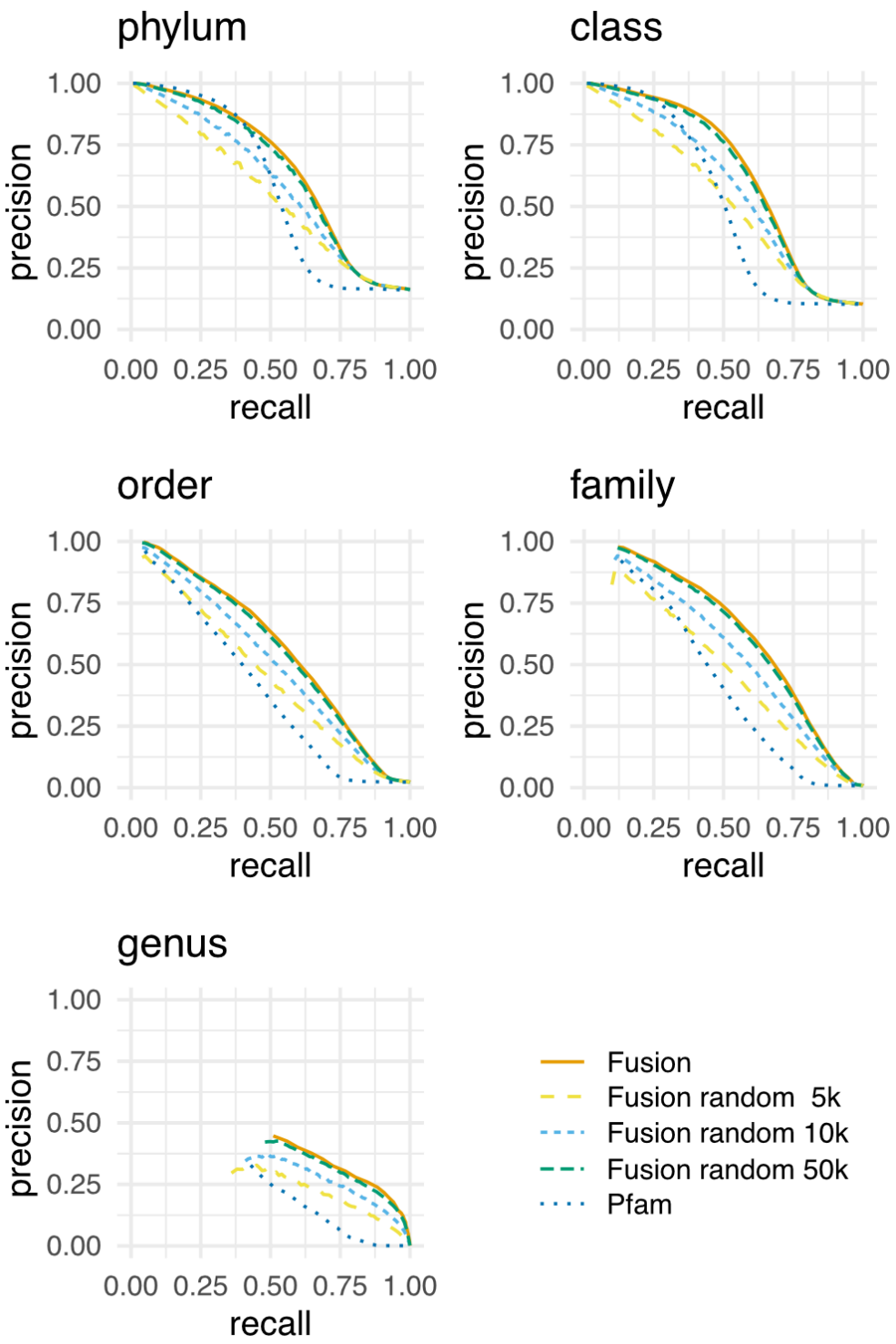

Fig. 3. Randomly selected Fusion functions identify organism taxonomic relationships. Each panel reflects the precision (y-axis) at a given recall (x-axis) for correctly identifying two organisms as sharing the same taxonomic rank (panel label). Line color indicates the functional samples. For example, using 5,000 Fusion functions (yellow) outperforms using all of Pfam

(darkblue) for most cutoffs across all panels. Displayed are only precision/recall pairs where predicted positives pairs (TP+FP) make up at least 0.1% of all possible pairs.

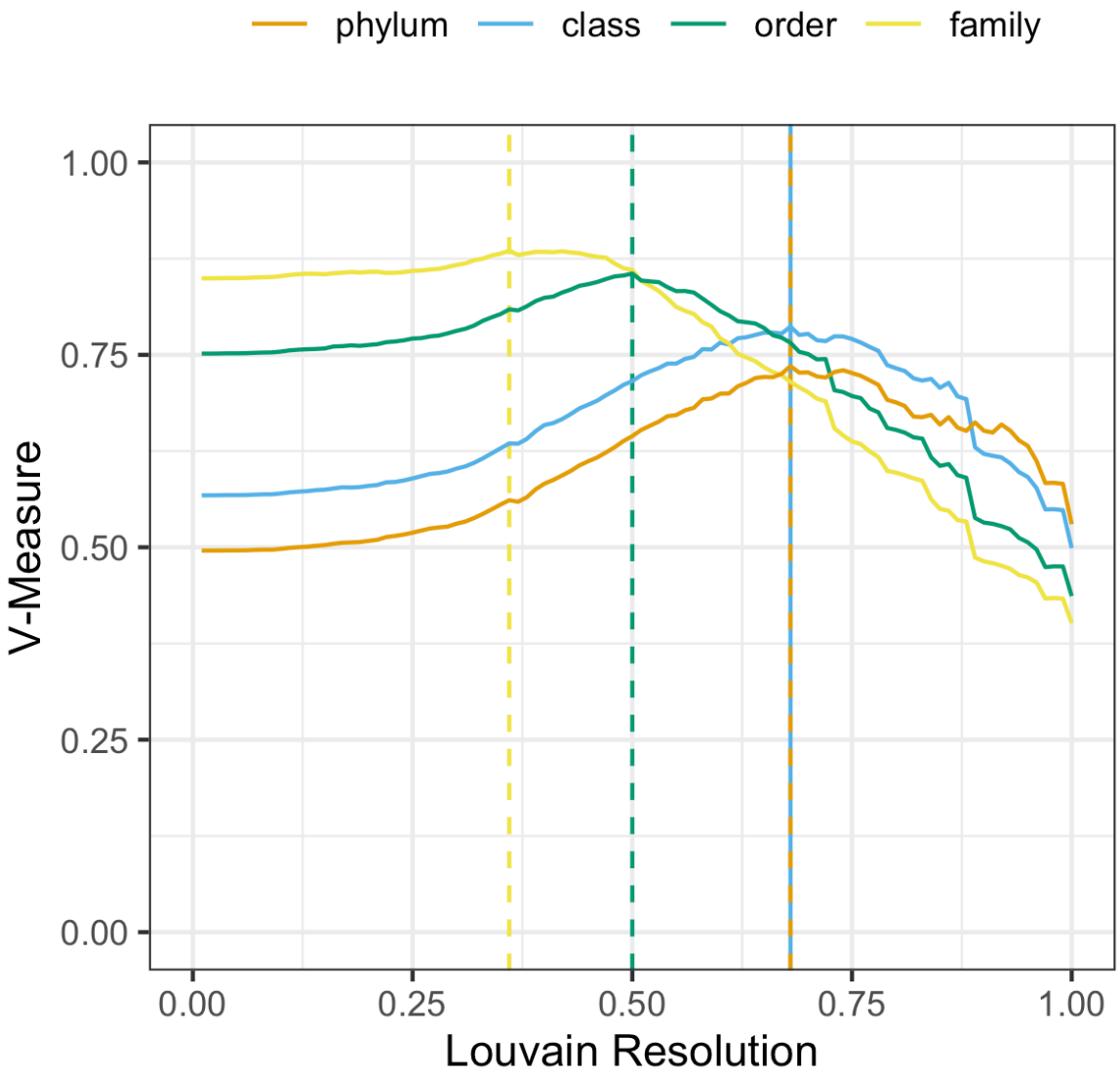

Fig 4. Community based organism classification using Fusion functional organism similarities recapitulates established taxonomy. Choosing different Louvain resolution parameters (x-axis) to establish communities of functionally similar organisms we can optimize the rate (y-axis) at which any two organisms are assigned to be in the same Fusion taxon vs. reference of GTDB-taxonomy assignment. For example, clustering the Fusion organism similarity network at a Louvain resolution parameter of 0.36 yields the best approximation of communities of organisms, corresponding to the family taxonomic level. Thresholds for order, class and phylum are 0.50, 0.68 and 0.68 respectively.

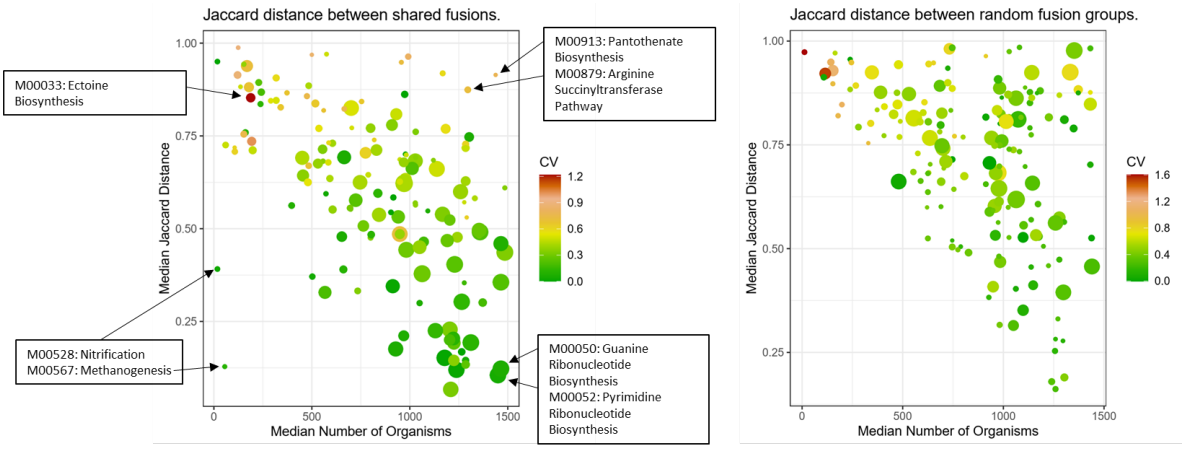

Fig 5. Median Jaccard distance and Fusion prevalence are inversely correlated. For each KEGG **(A)** and random **(B)** pathway module (dot in the plot), we calculated the median Jaccard distance between pairs of Fusion functions (y-axis) and the median number of proteomes each function is found in (x-axis). The dot color reflects the coefficient of variation (CV), or standard deviation over the mean for the assembly values, and the dot size captures the number of genomes encoding the given module (size). In **(A)**, modules with low median Jaccard values indicate either ubiquitous biological pathways (M00050, M00052), or pathways unique to specific niche communities (M00528, M00567). Modules with large distances tend to have high CVs, indicating a large difference in the prevalence of shared functions.

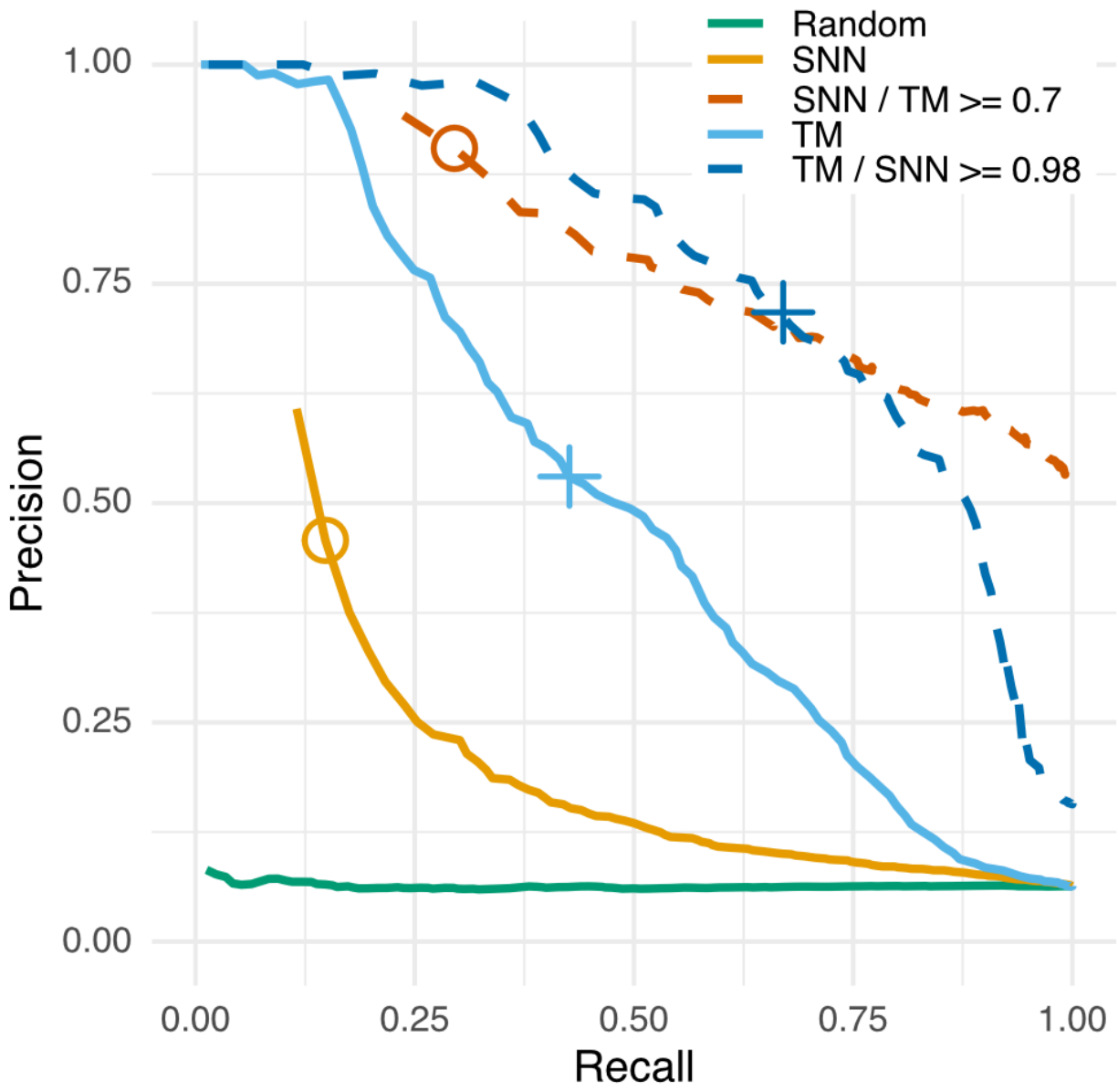

Fig 6. Combining TM and SNN scores improves annotation of functionally similar proteins. For proteins with available structures, the TM-score (blue solid line) was a better estimate of protein functional similarity (same EC number) than the SNN-score (orange solid line); even at the high reliability threshold of SNN-score ≥0.98 (circle), the method attained only 46% precision and 16% recall as compared 53% precision and 43% recall of the TM-score≥0.7 (cross). However, the combined SNN & TM-score metrics (dashed lines) were better than either of the methods alone. That is, for a subset of structurally similar proteins (TM≥0.7) the SNN score (orange dashed line) was a good indicator of functional similarity. Similarly for reliably functionally similar proteins (SNN≥0.98), the TM-score (blue dashed line) had a

significantly higher precision. Note that our dataset is representative of real life and thus, trivially, imbalanced as there are significantly fewer same EC (positive) pairs than different EC (negative) pairs; here, a ratio of ~1/15