

EvoWeaver: large-scale prediction of gene functional associations from coevolutionary signals

Received: 5 February 2025

Accepted: 9 April 2025

Published online: 24 April 2025

 Check for updatesAidan H. Lakshman¹ & Erik S. Wright^{1,2} 

The known universe of uncharacterized proteins is expanding far faster than our ability to annotate their functions through laboratory study. Computational annotation approaches rely on similarity to previously studied proteins, thereby ignoring unstudied proteins. Coevolutionary approaches hold promise for injecting new information into our knowledge of the protein universe by linking proteins through ‘guilt-by-association’. However, existing coevolutionary algorithms have insufficient accuracy and scalability to connect the entire universe of proteins. We present EvoWeaver, a method that weaves together 12 signals of coevolution to quantify the degree of shared evolution between genes. EvoWeaver accurately identifies proteins involved in protein complexes or separate steps of a biochemical pathway. We show the merits of EvoWeaver by partly reconstructing known biochemical pathways without any prior knowledge other than that available from genomic sequences. Applying EvoWeaver to 1545 gene groups from 8564 genomes reveals missing connections in popular databases and potentially undiscovered links between proteins.

Our ability to capture the protein universe with genome sequencing far outpaces our ability to investigate individual proteins. A select few proteins have historically received a disproportionate amount of study^{1–3}. This annotation inequality hinders biomedical progress by neglecting many proteins that could be important determinants of health⁴. Only a small fraction of uncharacterized proteins can be automatically annotated via similarity to experimentally investigated proteins of known function^{5–7}. The sparsity of high-quality annotations exacerbates the problem of non-specific and low-confidence annotations that proliferate across genomes^{8,9}. Thus, computational approaches to infer function without dependence on prior knowledge are acutely needed.

Computationally annotating the remainder of the protein universe requires establishing connections with characterized proteins to generate hypotheses about function through ‘guilt-by-association’¹⁰. Shared function necessitates that protein-encoding genes coevolve in the same cell, thereby leaving behind a molecular signal of

coevolution¹¹. Four primary approaches are used to identify coevolution: phylogenetic profiling^{12–14}, phylogenetic structure^{15–17}, gene organization^{18–20}, and sequence level methods^{21–23}. Each of these coevolutionary signals is an outcome of a shared selection pressure acting on groups of genes. To date, these four coevolutionary approaches have primarily been applied independently. Even large databases of functional associations, such as STRING, only consider evidence from a small subset of coevolutionary approaches²⁴.

Although coevolutionary analyses have shown great potential for predicting functional associations^{25–32}, scalability is a major impediment to comprehensive application on large datasets. The era of big data holds the promise of distinguishing coevolution from other drivers of molecular evolution¹³. Additionally, holistic evaluation of many coevolutionary signals offers a means of amplifying weaker signals to make higher-accuracy predictions. For example, conserved genes may not display a phylogenetic profiling signal but can still show patterns of gene organization. Combining disparate coevolutionary signals and

¹Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. ²Center for Evolutionary Biology and Medicine, Pittsburgh, PA, USA.

✉ e-mail: eswright@pitt.edu

scaling to larger datasets requires inventing new approaches for discerning signal from noise.

Coevolutionary analyses have the potential to infer functional associations directly from sequencing data in a way that is agnostic to prior annotations, thereby overcoming the current reliance on extrapolating from existing knowledge that compounds annotation inequality. Here, we set out to develop a scalable approach to extract and combine coevolutionary signals for predicting functional associations between protein-coding genes. This required improving upon existing approaches to scale to larger input data and incorporating statistical testing. We unite these signals of coevolution using machine learning models to quantify the degree of functional association between genes. Our approach, named EvoWeaver, is available within the *SynExtend* package (v1.19.0) for R and serves as a high-quality hypothesis generator to help extend our knowledge of the protein universe.

Results

Existing coevolutionary algorithms have widespread issues with scalability, interoperability, and interpretability¹³. We chose to implement

all our coevolutionary analyses from scratch within a single software package to standardize user interaction and allow for easy application of ensemble methods. Our approach, named EvoWeaver, takes as input a set of phylogenetic gene trees and optional metadata (Fig. 1a). EvoWeaver then performs four types of coevolutionary analyses, comprised of 12 algorithms optimized for scalable performance. The output of EvoWeaver is 12 scores ranging from -1 to 1 that quantify the strength of coevolution between a pair of gene groups. These scores can be combined using a machine learning classifier to generate inferences or hypotheses about gene function.

The first type of coevolutionary analysis, Phylogenetic Profiling, investigates patterns of presence/absence (P/A) or gain/loss (G/L) of genes, which manifest when multiple genes work in concert (Fig. 1b). While P/A analyses have been successfully used to predict gene function^{12–14,33–35}, existing approaches can be susceptible to biases from small sample sizes or low evolutionary divergence³⁶. We addressed these biases by introducing an algorithm (G/L Distance) that examines the distance between G/L events to measure compensatory changes rather than extant patterns. We also implement clade-wise phylogenetic profiling (P/A Jaccard), which corrects for bias from oversampled

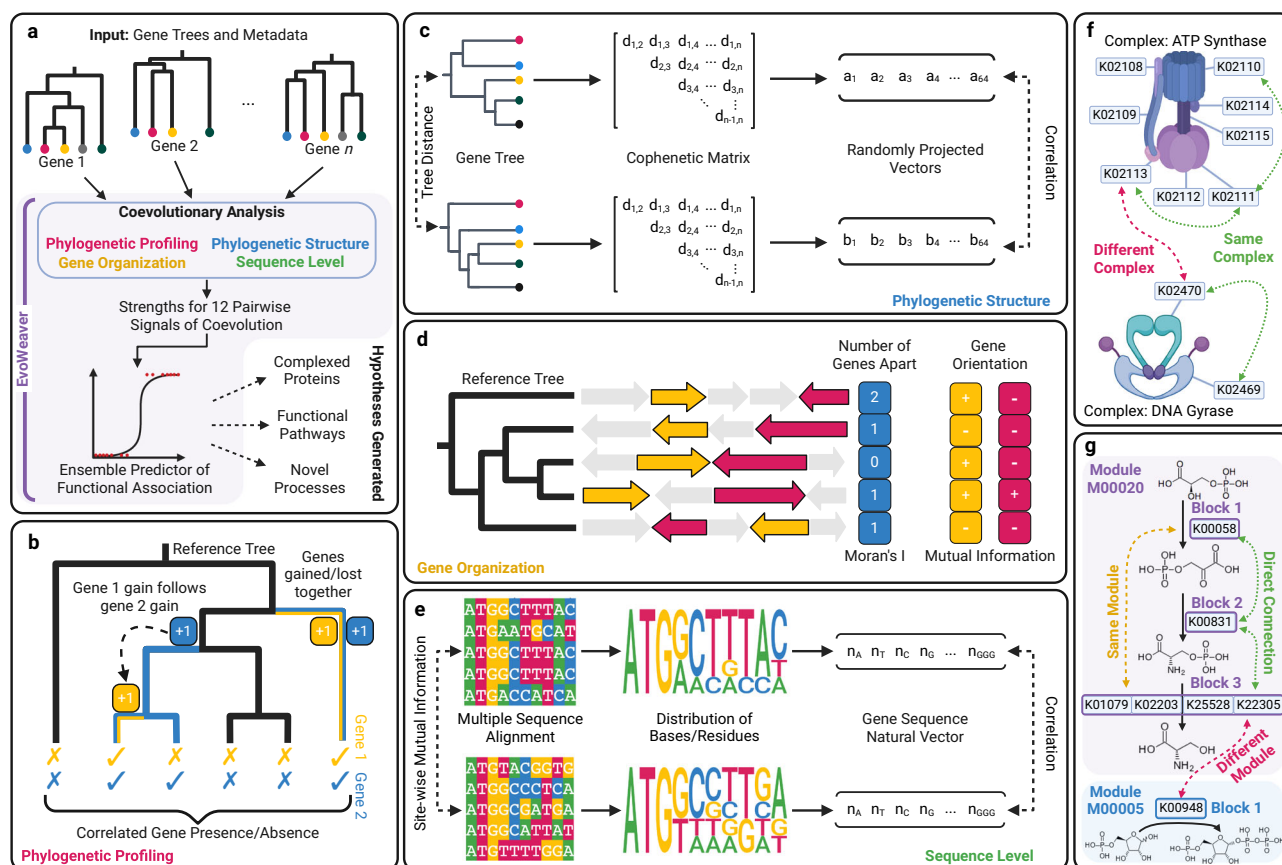


Fig. 1 | Overview of the EvoWeaver algorithm and benchmarking.

a Phylogenetic trees from groups of orthologous genes serve as the primary input to EvoWeaver. Four categories of coevolutionary signal are quantified for each pair of genes. These signals are combined in an ensemble classifier to predict functional relationships between gene pairs. EvoWeaver provides as output its 12 predictions for signals of coevolution, and can optionally provide an ensemble prediction using built-in pretrained models. **b** Functional associations often result in correlated gain/loss patterns on a reference phylogenetic tree (e.g., a species tree). EvoWeaver assesses the presence/absence patterns, correlation between gain/loss events, and distance between gain/loss events as signals of coevolution. **c** Similarity in phylogenetic structure is another indicator of coevolution between genes. EvoWeaver computes topological distance as well as correlation in patristic distances following dimensionality reduction using random projection. **d** Functionally associated genes sometimes cluster on the genome due to co-

regulation or horizontal gene transfer. EvoWeaver derives signals from the conservation in gene orientation and the distance between gene pairs. **e** Functional associations sometimes cause concerted changes in sequences that are interrogated by EvoWeaver. EvoWeaver can analyze nucleotide sequences or amino acid sequences, though nucleotide sequences are pictured here. **f** Proteins involved in the same complex are functionally associated and can be identified through signals of coevolution. The goal of the Complexes benchmark is to distinguish orthology groups in the same complex (i.e., positives) from those in different complexes (i.e., negatives). **g** Functional associations between proteins that are adjacent in the same module are stronger than those between different modules. The goal of the Modules benchmark is to distinguish adjacent proteins in the same module from independent modules. Created in BioRender. Lakshman, A. (2025) <https://BioRender.com/m73q207>.

taxa³⁷. Finally, we analyze the mutual information of ancestral state transitions (G/L MI), as well as the conservation of mutual presence in ancestral states (P/A Overlap). The end result is a category of algorithms for identifying coevolution between gene groups that are not highly conserved.

The second type of coevolutionary analysis, Phylogenetic Structure, uses the fact that functionally associated genes tend to evolve in tandem³⁸, giving rise to similar genealogies (Fig. 1c). Commonly used phylogenetic structure approaches include MirrorTree and ContextTree^{39–41}, although these approaches scale poorly due to high computational complexity. We addressed this issue by using random projection (RP MirrorTree, RP ContextTree) to decrease computational overhead and improve accuracy by reducing redundant information. Random projection provides the added advantage that computation can be distributed across computers, unlike in SVD-phy⁴², allowing EvoWeaver to process very large datasets on compute clusters. Additionally, we introduce the use of tree distance metrics (Tree Distance) to analyze coevolution via topological differences in genealogies⁴³. Taken together, these algorithms facilitate inference of coevolution among more conserved gene groups.

The third type of coevolutionary analysis, Gene Organization, leverages the fact that functionally linked genes tend to colocate on the genome to facilitate gene regulation and horizontal gene transfer^{44–46} (Fig. 1d). These approaches most commonly employ profile hidden Markov models, such as antiSMASH^{47–49}. While these methods perform well at functional prediction, they rely on a priori knowledge about genes that colocate. We circumvented this limitation by introducing an algorithm that compares the number of coding regions separating genes (Gene Distance). Our approach is similar to STRING's colocalization metric, which measures the number of nucleotides separating genes²⁴, but STRING's approach fails to consider that low rates of evolutionary divergence can inflate evidence of colocalization. We address this issue by using Moran's I to calculate the extent to which evolutionary divergence affects the observed colocalization of genes. Additionally, EvoWeaver analyzes the conservation of relative gene orientation (Orientation MI), since this also indicates functional association⁵⁰. Collectively, these algorithms provide evidence of coevolution among conserved gene groups on the same chromosome.

The last type of coevolutionary analysis, Sequence Level methods, looks at sequence patterns across gene groups, which are sometimes indicative of physical interactions between gene products⁵¹ (Fig. 1e). Direct coupling analysis is a well-known approach in this category^{52–54}, but it suffers from high computational complexity. Instead, we extended a prior approach based on mutual information to predict interacting sites between sequences⁵⁵. EvoWeaver analyzes the extent of these site-wise interactions to construct an overall score (Sequence Info). Additionally, EvoWeaver compares gene sequence natural vectors (Gene Vector), which carry evidence of functional association and can be quickly computed⁵⁶. These algorithms provide additional evidence of coevolution for physically interacting gene products.

The four categories of analysis span levels of coevolution from the organism (Phylogenetic Profiling) to the genome (Gene Organization) to the gene (Phylogenetic Structure) to the sequence. Since our component analyses individually capture different facets of coevolution, we sought to combine their strengths into a single comprehensive estimate of evidence for functional association between gene pairs. To this end, we trained three machine learning classifiers (logistic regression, random forest, and neural network) on sets of protein-coding gene pairs with known functional associations (Fig. 1a). While these ensemble models require a priori knowledge to calibrate their predictions, after training they permit the extension of this knowledge to gene pairs without previously known associations. More details about each algorithm are provided in section SII of the Supplemental Information.

Ensemble methods accurately identify functionally associated genes

Selection of high-quality ground truth datasets for coevolutionary analysis is a challenging task¹³. As with previous studies^{42,57}, we relied upon the Kyoto Encyclopedia of Genes and Genomes database (KEGG) because it is well-curated and experimentally validated^{58,59}. KEGG provides a hierarchical ontology of biochemical pathways consisting of orthologous gene groups (KO groups) participating in protein complexes (Fig. 1f) and/or enzymatic reactions within modules (Fig. 1g). Modules are the building blocks of larger biochemical pathways. We first sought to validate the performance of EvoWeaver at identifying KO groups that participate in the same complex, since physical interactions are a form of functional association. We anticipated a strong coevolutionary signal for these pairs because of their mutual dependence. Each algorithm's performance was graded on its ability to distinguish 867 pairs of KO groups that complex (positives) versus 867 randomly selected pairs of unrelated KO groups (negatives). The negative set was constructed from a weighted random sample of 57,321 unrelated KO groups. Weighted sampling reduces the risk of overfitting by matching the distribution of data features in the negative set to the positive set.

Almost all coevolutionary algorithms performed well at identifying KO groups involved in the same complex (Fig. S1). Sequence Level methods performed slightly worse than other categories of coevolutionary signal. This outcome was expected because many non-interacting proteins appear to physically interface similarly to interacting proteins⁶⁰. The predictions of most algorithms were weakly correlated with each other, which suggests combining signals could further improve performance (Fig. S1). To this end, we evaluated three ensemble methods (Logistic Regression, Random Forest, and Neural Network) using five-fold cross-validation. All ensemble methods displayed predictive power exceeding component coevolutionary signals, with Logistic Regression performing the best (Fig. S1).

Given EvoWeaver's strong performance on the Complexes benchmark, we next sought to establish its ability to identify functionally associated protein-coding genes that were not involved in the same protein complex. To this end, we developed the Modules benchmark as a set of 899 pairs of gene groups acting in adjacent steps of a biochemical pathway in KEGG (positives) and 899 randomly selected pairs from disconnected pathways (negatives). This task is more challenging because proteins involved in the same module need not physically interact (Fig. 1g). As shown in Fig. 2, the performance of component algorithms on the Modules benchmark was slightly worse than on the Complexes benchmark. However, ensemble methods retained high performance (AUROC of 0.955 for Random Forest) and outperformed individual coevolutionary signals. The gap between ensemble and component predictors highlights the importance of using multiple coevolutionary signals to infer functional associations.

Next, we sought to determine whether EvoWeaver's ensemble predictions were transferrable to a different prediction task. To this end, we used the CORUM database to test EvoWeaver's ability to identify human proteins that participate in common complexes (Fig. S2). The 12 component algorithms were less accurate on CORUM than KEGG but outperformed other commonly used approaches (Fig. S3). We observed a similar change in relative performance when restricting EvoWeaver to only eukaryotic sequences in KEGG (Fig. S4), indicating that the diminished performance on CORUM is likely due to its stronger focus on eukaryotes.

EvoWeaver's ensemble method trained on the KEGG Modules benchmark did not show an advantage over the best component algorithm on CORUM. We attributed this discrepancy to differences between the two databases and CORUM's sole emphasis on physical protein-protein interactions. EvoWeaver is designed to predict functional associations, which encompass direct (i.e., physical) and indirect interactions. Retraining the ensemble classifier on the CORUM

database resulted in a substantial increase in accuracy above that of any component algorithm (Fig. S2). This result underscores the fact that transferability of accurate ensemble predictions relies upon a shared prediction objective. More detail on the discrepancies between the CORUM and KEGG benchmarks is discussed in sections SI2 and SI3 of the Supplementary Information.

Even in the absence of training an ensemble classifier, a naïve combination of EvoWeaver’s component algorithms resulted in performance approximately equivalent to the best predictor on all benchmarks, especially at low false positive rates (Fig. S5). This outcome highlights a unique benefit of EvoWeaver over previous

approaches that require users to know which evolutionary patterns will be most informative on their dataset. The combination of EvoWeaver’s algorithms produces a prediction approximately as good as the best choice without any information about which algorithm is best suited to a particular use case.

EvoWeaver infers hierarchical relationships among genes
Functional relationships among genes exist at multiple levels, ranging from physically interacting to merely being part of the same cellular environment. For this reason, it would be ideal to predict the strength of coevolution across a hierarchy of multi-level relationships among

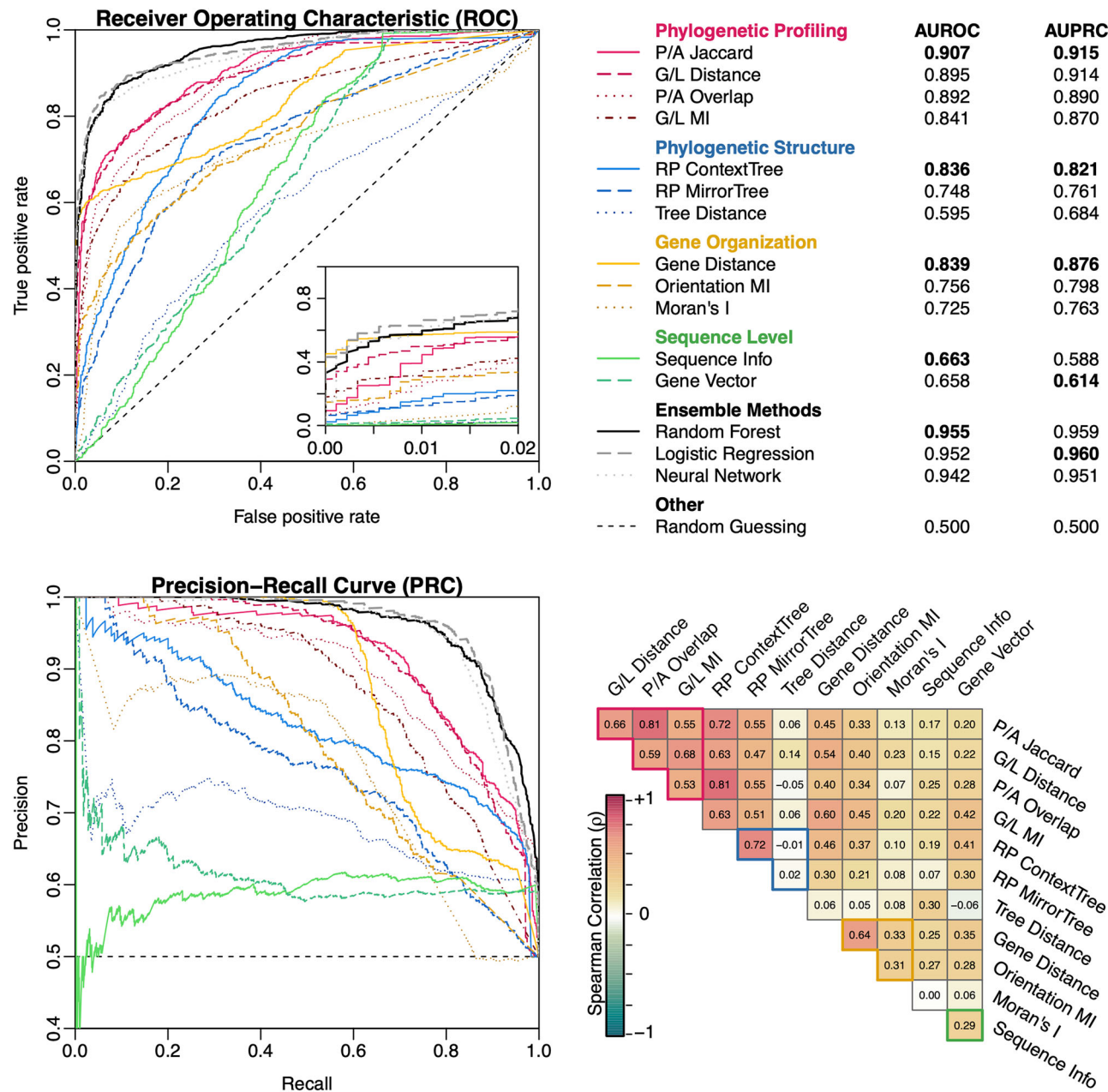


Fig. 2 | EvoWeaver’s ensemble predictions outperform individual algorithms on the Modules benchmark. Coevolutionary approaches were compared for their ability to discern adjacent proteins in KEGG modules (i.e., 899 positives) from proteins in distinct modules (i.e., 899 negatives). No single source of coevolutionary signal greatly outcompeted all other sources. However, EvoWeaver’s ensemble predictions that combine all component sources of coevolutionary signal substantially improved predictive accuracy, as seen by larger areas under the

curves. Inset of the receiver operating characteristic highlights the region with low false positive rates. Scores from individual algorithms tended to have low correlation except within similar categories of coevolutionary signal (i.e., boxed groups in the heatmap), suggesting that the ensemble approach is superior because it combines semi-orthogonal coevolutionary signals. Spearman’s correlation from positive and negative sets is averaged to correct for artificial correlation among high-performing algorithms. Source data are provided as a Source Data file.

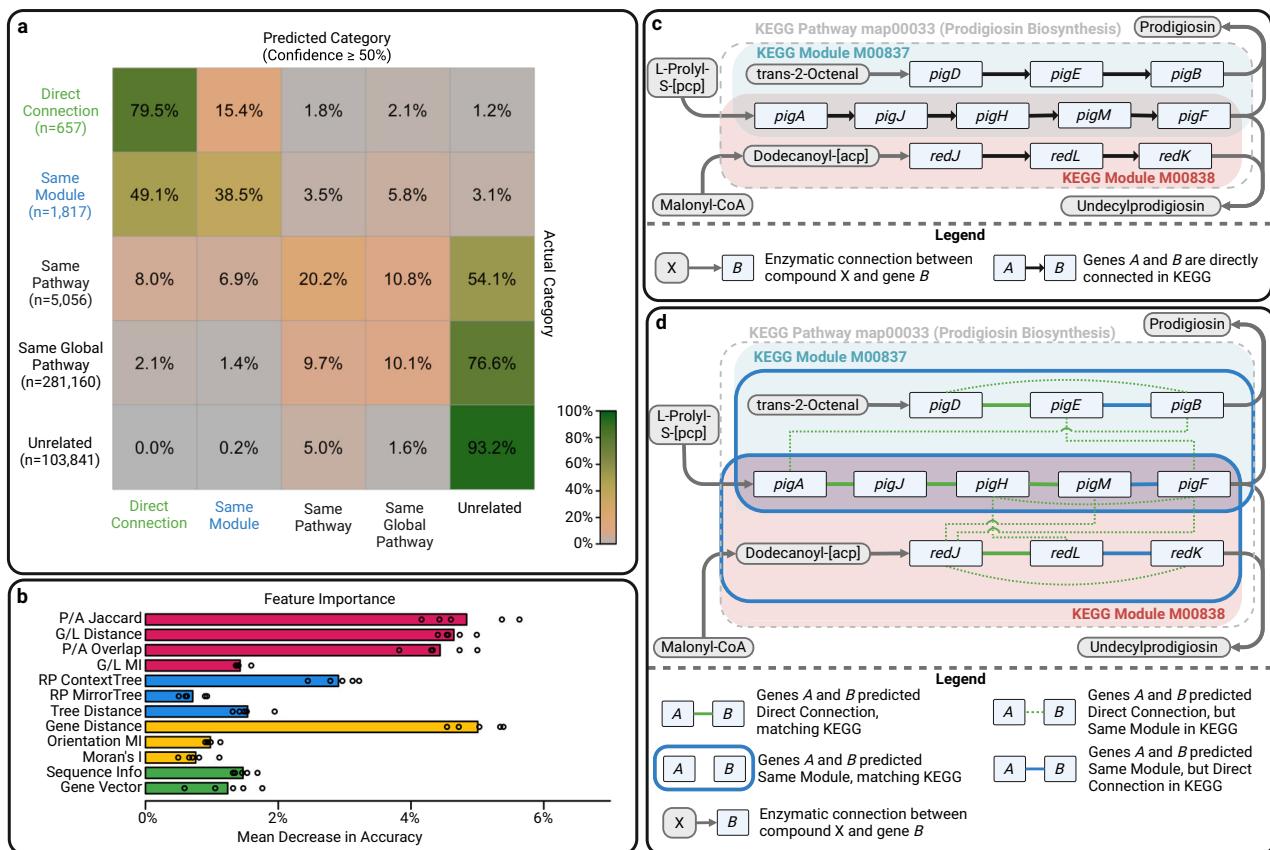


Fig. 3 | EvoWeaver is sufficiently accurate to hierarchically classify functional associations. **a** The confusion matrix of five-level classifications indicates that EvoWeaver's ensemble predictions (i.e., Random Forest) rarely confuse proteins within the same module with those from different modules. Values represent the percent of each actual class classified to each predicted class. **b** The best performing algorithm from each category on the Modules benchmark was also assigned greater feature importance by the random forest model in hierarchical classification. All features were important in the ensemble's predictions, further

underscoring the benefit of using multiple coevolutionary signals. Overlaid points denote importance from each of the five train/test folds. **c** A group of proteins randomly selected from hierarchical clustering exactly matches an existing tightly linked set of modules from KEGG. **d** EvoWeaver's ensemble predictions for genes involved in prodigiosin biosynthesis generally match experimentally verified connections in KEGG. Note that *pigA*, *pigJ*, *pigH*, *pigM*, and *pigF* belong to both modules. Source data are provided as a Source Data file. Created in BioRender. Lakshman, A. (2025) <https://BioRender.com/k26g262>.

gene groups. We created the Multiclass benchmark as pairs of KEGG module blocks from five classes arranged in a hierarchy of decreasing functional association: Direct Connection, Same Module, Same Pathway, Same Global Pathway, and Unrelated. Accurate classification would imply EvoWeaver can construct a hierarchical classification scheme of genes and recapitulate the relationships in KEGG. We then used five-fold cross-validation to predict class membership for 642,770 pairs of module blocks using a Random Forest model (Fig. 3). Notably, all 12 predictors contributed to the ensemble classifier's accuracy (Fig. 3b). Most Random Forest predictions were assigned to the correct class or the adjacent class (Fig. S6), even when requiring at least 50% confidence for prediction (Fig. 3a). Unsurprisingly, the model frequently confused the Same Global Pathway and Unrelated classes, which are both expected to contain weakly coevolving genes.

The Random Forest ensemble classifier was best at distinguishing the top two from bottom three hierarchical classes. Hence, we tested whether these predictions could be used to recapitulate KEGG pathways by building a network of module blocks with connections between pairs predicted as Direct Connection or Same Module. We applied Louvain clustering⁶¹ to detect communities within this network. A randomly selected community is shown in Fig. 3c, d, which included all module blocks involved in the prodigiosin biosynthesis pathway. EvoWeaver correctly identified most Direct Connection pairs and properly distinguished the two modules within the pathway. However, EvoWeaver incorrectly classified many Same Module pairs as

Direct Connection. This analysis suggests EvoWeaver's predictions can be used to hypothesize biochemical pathways, although they do not provide directionality to biochemical steps.

EvoWeaver rivals STRING without reliance upon external data

STRING is one of the most comprehensive databases of knowledge about functionally associated genes. One of STRING's stated goals is to predict genes belonging to the same pathway in KEGG⁵⁷, which corresponds to EvoWeaver's Direct Connection, Same Module, and Same Pathway classifications. STRING's Total Score is a composite of seven evidence streams²⁴. We applied STRING's formula for Total Score to quantify the marginal benefit of each evidence stream (Fig. S7). External data, including mining the literature for cooccurrence of terms (Text Mining) and knowledge bases such as KEGG itself (Databases), provided the majority of STRING's predictive performance (Fig. 4a). As expected, STRING's coevolutionary evidence streams (Cooccurrence, Gene Neighborhood) were correlated with comparable signals derived by EvoWeaver (Fig. 4b). Excluding Text Mining, EvoWeaver nearly matches STRING at its stated goal of predicting pairs of gene groups sharing a functional pathway in KEGG (Fig. 4a). This is especially notable given that STRING's Database evidence stream incorporates KEGG itself as a predictor, whereas EvoWeaver only relies on information extracted from genome sequences. EvoWeaver's ensemble methods greatly outperformed STRING when limiting both methods to only predictors that do not require prior knowledge, even

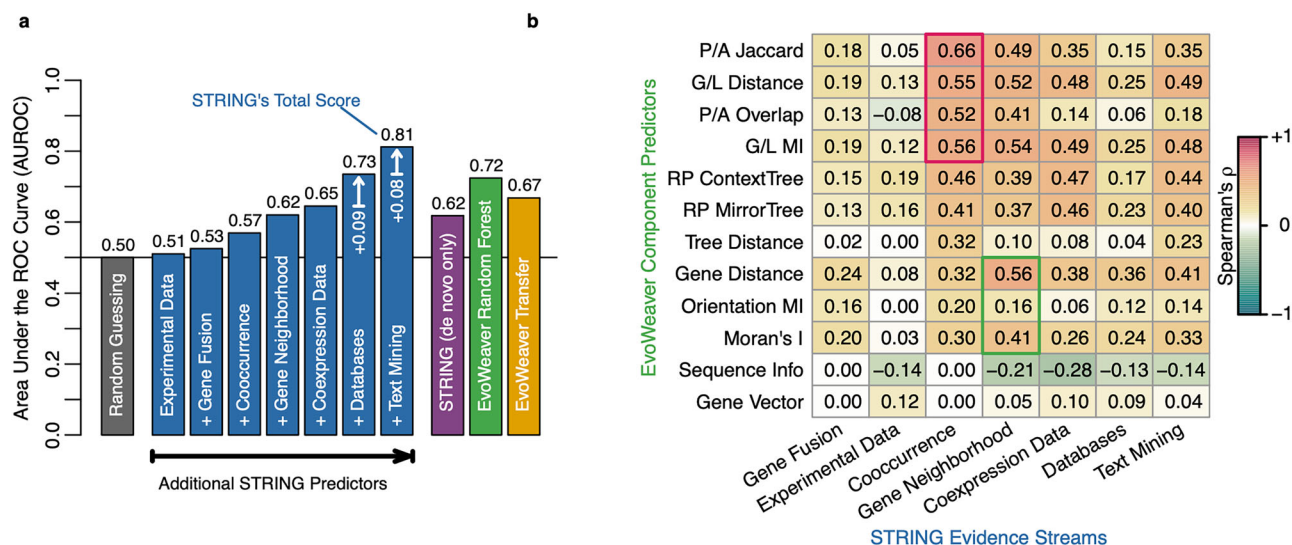


Fig. 4 | EvoWeaver rivals STRING without reliance on external data. a Predictive accuracy was compared on 1514 pairs of gene groups that overlapped between STRING and the Multiclass benchmark. Area under the ROC curve (AUROC) is shown for discerning between pairs sharing the same pathway in KEGG (i.e., positives) versus pairs in different pathways (i.e., negatives). STRING's predictions are a composite of seven evidence streams. Sequentially incorporating evidence streams from least to most beneficial demonstrates their marginal impact on STRING's reported Total Score. Text Mining and Databases were the most impactful STRING evidence streams. Despite STRING's predictions incorporating KEGG itself into its Databases evidence stream, EvoWeaver's Random Forest

predictions roughly match STRING's predictions without Text Mining while only using sequence information. EvoWeaver greatly outperforms STRING when both are limited to only de novo predictors (i.e., Gene Fusion, Cooccurrence, and Gene Neighborhood for STRING), even when trained on CORUM (EvoWeaver Transfer). **b** As expected, some of EvoWeaver's component predictors were modestly correlated with STRING's evidence streams. For example, STRING's Cooccurrence score is correlated with EvoWeaver's Phylogenetic Profiling algorithms (red box), and STRING's Gene Neighborhood score is correlated with EvoWeaver's Gene Organization algorithms (green box). Spearman's correlation is calculated in the same manner as in Fig. 2. Source data are provided as a Source Data file.

when using an ensemble model trained on a different dataset (Fig. 4a). This makes EvoWeaver particularly powerful for identifying unknown functional associations without reliance on prior knowledge, which may help to mitigate the problem of annotation inequality^{1,2}.

EvoWeaver can inform hypotheses about functional associations

EvoWeaver's primary purpose is to serve as a generator for hypotheses about functional associations. As a case study, we examined one of EvoWeaver's high-confidence mispredictions, which was between human genes *B3GNT5* and *ST6GAL1*. *B3GNT5* encodes an enzyme responsible for the synthesis of lactotriaosylceramide, the primary precursor for lacto- and neolacto-series glycosphingolipids, and this enzyme is known to play a role in a variety of human diseases^{62,63}. *ST6GAL1* is responsible for the α 2,6-sialylation of N-glycosylated proteins. Despite *B3GNT5* and *ST6GAL1* having no common modules or pathways in the KEGG database (Fig. 5a), EvoWeaver predicted this pair to be Direct Connection with probability 0.63 or Same Module with probability 0.36 (Fig. 5b). This finding is consistent with experimental evidence showing mutations in glycosphingolipid biosynthetic enzymes can cause changes in sialylation of N-glycosylated membrane-bound proteins⁶⁴, and specifically, mutations in *B3GNT5* modulate α 2,6-sialylation of membrane-bound glycoproteins in ovarian cancer cells by directly silencing the expression of *ST6GAL1* in several human cell lines⁶⁵. EvoWeaver's prediction was supported by Phylogenetic Profiling evidence because of the multiple inferred simultaneous gains of both genes (Fig. 5c) along with moderate evidence for Gene Organization due to conservation in gene orientation and relative distance across the phylogeny (Fig. 5d). *B3GNT5* and *ST6GAL1* also displayed strong similarity in their genealogies (Fig. 5e) and moderate evidence for coevolutionary signal at the sequence level (Fig. 5f). While both *B3GNT5* and *ST6GAL1* have functional associations with *B4GALT* family genes in KEGG (Fig. 5a), EvoWeaver's ensemble method did not

identify a connection between *ST6GAL1* and *B4GALT* family genes (Fig. S8), suggesting that the predicted linkage between *B3GNT5* and *ST6GAL1* is unlikely to have resulted from transitivity.

To further substantiate EvoWeaver's power as a hypothesis-generating tool, we investigated the top 100 mispredictions wherein a pair of genes was classified as Same Pathway in KEGG, but EvoWeaver predicted them to be Direct Connection. Many of these gene pairs were actually directly connected (19%) or separated by only a few genes in a KEGG pathway (Fig. S9), but were categorized as Same Pathway because they lacked connections in a common module. Therefore, the top mispredictions were partly artifacts of how the KEGG database defines modules within pathways. We also investigated the top mispredictions in which a pair was classified as Same Global Pathway in KEGG but EvoWeaver predicted Direct Connection. Of the top five misclassifications, three involved gene pairs between KEGG module M00892 (UDP-GlcNAc biosynthesis in eukaryotes) and KEGG module M00055 (N-glycan precursor biosynthesis). Coevolutionary (Fig. S8) and experimental evidence support the interconnectedness of these modules: N-glycan branching is hypersensitive to UDP-GlcNAc concentrations in mammals⁶⁶, UDP-GlcNAc transporters are involved in the delivery of N-glycan substrates in plants⁶⁷, and components of the UDP-GlcNAc biosynthetic pathway are required for complex N-glycan synthesis in *C. elegans*⁶⁸.

Next, we asked whether EvoWeaver can contribute in cases where a set of genes is implicated in a common function but their inter-relationships are unknown. We investigated EvoWeaver's predictions for six sets of genes comprising discrete biochemical pathways (Fig. 6). The four categories of coevolutionary algorithms often disagreed with each other and differed from the connections in KEGG. However, EvoWeaver's ensemble predictions generated more accurate connections, which reinforces the notion that merging evidence streams improves predictions. Taken together, these findings suggest that EvoWeaver can be used to augment existing biological knowledge by predicting credible gene functional associations.

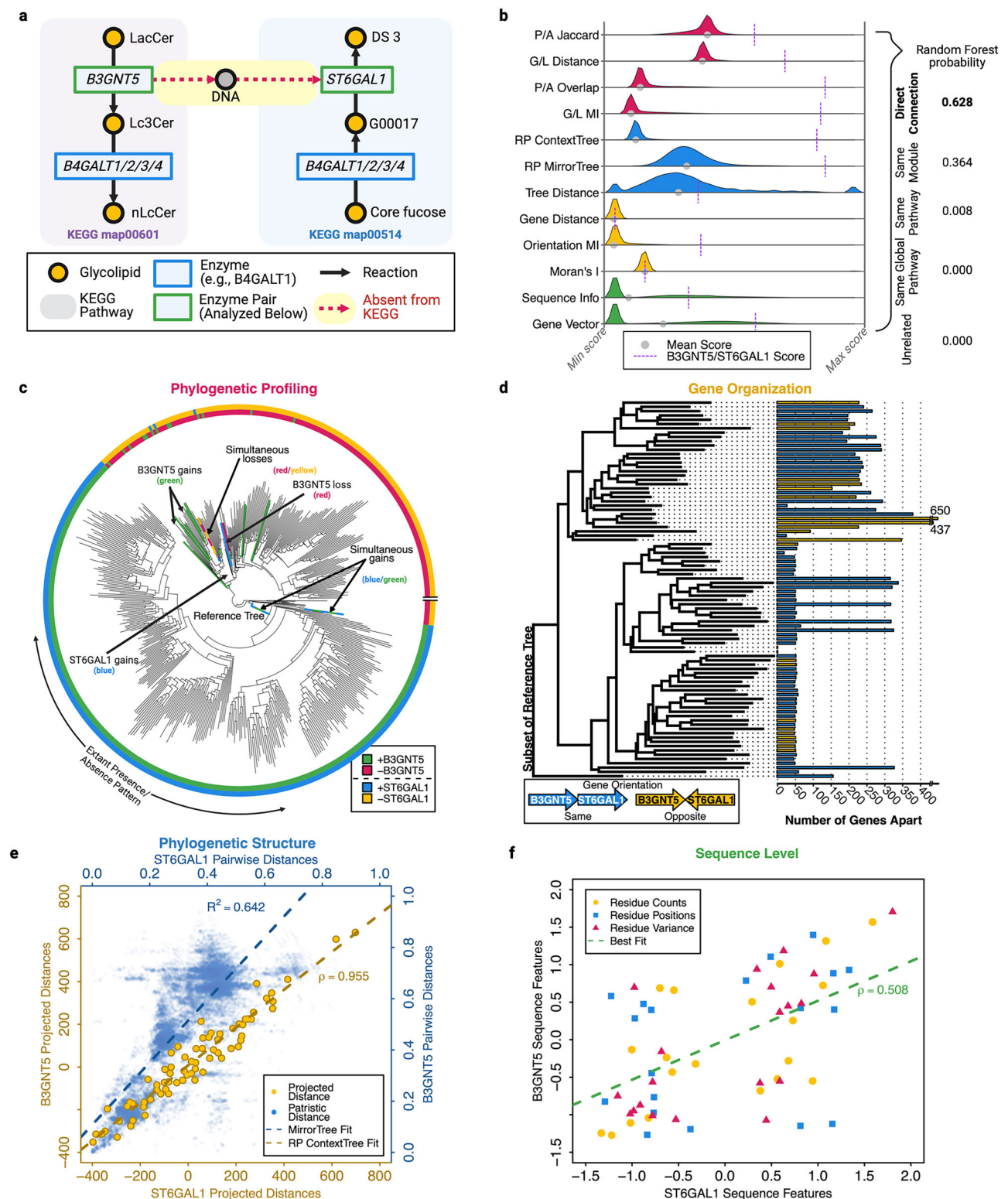


Fig. 5 | EvoWeaver's ensemble predictions can generate high-fidelity biological hypotheses. **a** The protein product of *B3GNT5* promotes the expression of *ST6GAL1*⁶⁵, although this connection is missing in KEGG and STRING. **b** EvoWeaver's component and ensemble predictions indicate that *B3GNT5* and *ST6GAL1* are functionally associated, which is supported by experiments in human cell culture⁶⁵. **c** Phylogenetic Profiling demonstrates a pattern of association between *B3GNT5* and *ST6GAL1*, although it is supported by relatively few gain/loss events on the reference tree. **d** Organisms with both *B3GNT5* and *ST6GAL1* on the same

chromosome display correlations in gene orientation and modest signal of colocalization. **e** Shared patristic distances from both gene trees are correlated, especially after compression with random projection, suggesting a high degree of coevolution between *B3GNT5* and *ST6GAL1*. **f** Gene sequence natural vectors for both *B3GNT5* and *ST6GAL1* are moderately correlated, implying similar residue compositions and providing further signal of coevolution. Source data are provided as a Source Data file. Created in BioRender. Lakshman, A. (2025) <https://BioRender.com/u57r428>.

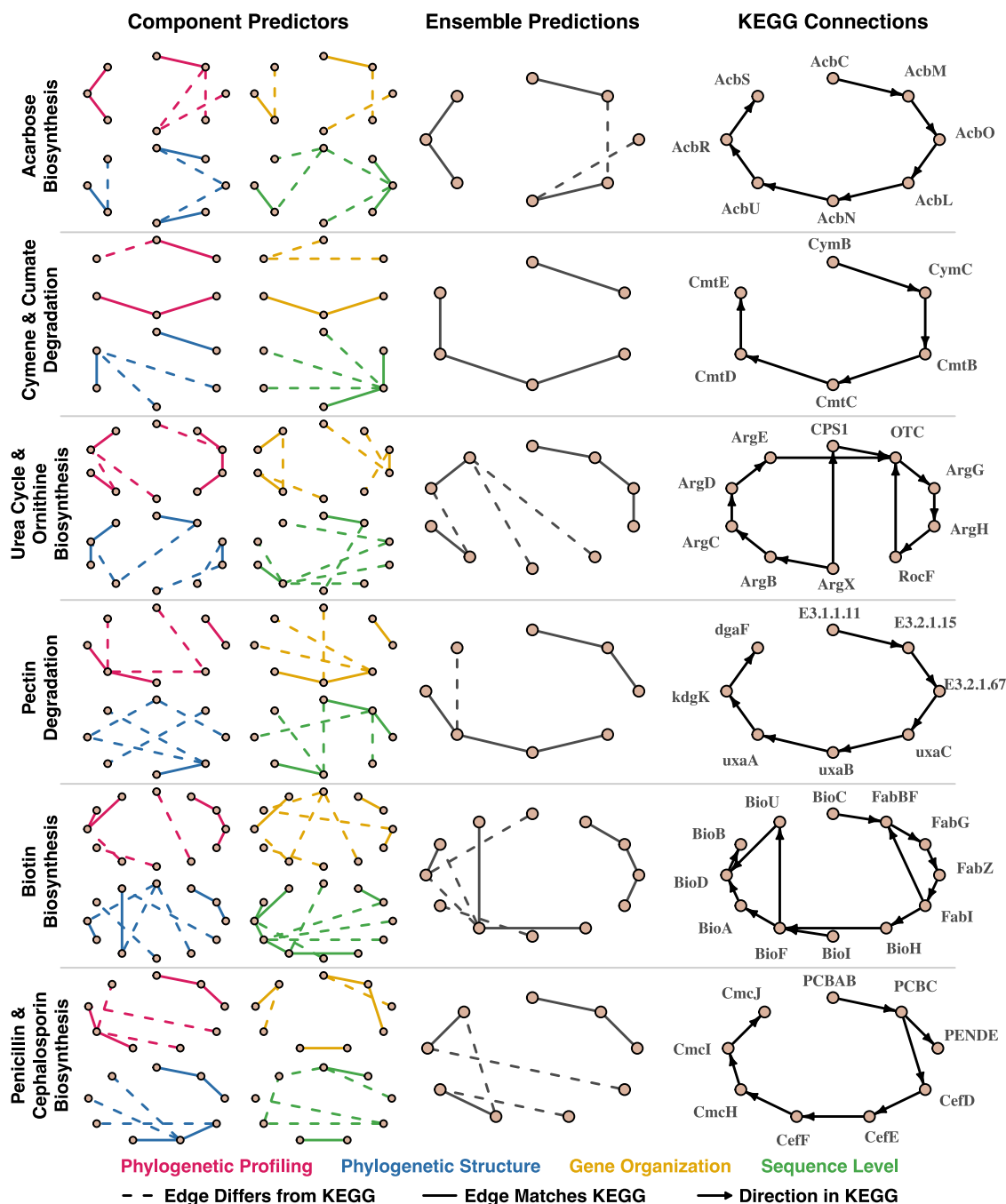


Fig. 6 | EvoWeaver partly identifies biochemical pathway connectivity. EvoWeaver's pairwise scores from component algorithms provide a ranking of functional association drawn from alternative categories of coevolutionary signal (colors). EvoWeaver combines 12 component scores into a single ensemble prediction for each pair of gene groups. The strongest predicted connection for every gene group shows high consistency (solid lines) with the actual connectivity of KEGG pathways (arrows). Discrepancies (dashed lines) between predicted and

actual connections are often caused by EvoWeaver incorrectly linking consecutive gene groups, such as AcbO-AcbL-AcbN in Acarbose Biosynthesis or ArgB-ArgC-ArgD in Ornithine Biosynthesis. Component predictors are connected according to the gene group with the highest mean rank among all algorithms in a category. Ensemble predictions are determined by connecting each gene group to the gene group with the highest probability of "Direct Connection" in the Multiclass Random Forest model.

Discussion

EvoWeaver showcases the power of employing coevolutionary principles to the discovery of functional associations. In this work, we introduced several algorithms to quantify coevolution and showed that EvoWeaver can capitalize on these signals to generate a more complete understanding of the functional relationships between gene groups. Importantly, EvoWeaver's ensemble predictions have the advantage that they do not require users to choose which

coevolutionary signals are appropriate for a particular context. EvoWeaver's accuracy permitted us to construct a hierarchical model of functional associations that was able to partly recapitulate experimentally validated KEGG pathways without any prior knowledge of the proteins other than their coding sequences and genomic locations. Moreover, we demonstrated how EvoWeaver's predictions can be leveraged to infer functional associations that are absent from large databases of biological knowledge.

EvoWeaver excels at three characteristics that are necessary for the practical application of coevolutionary analyses on large-scale datasets. First, EvoWeaver is highly scalable owing to its optimized algorithms. We demonstrated this by applying EvoWeaver to 1545 gene groups from 8564 genomes across the tree of life, comprising a total of 2,838,832 genes. To our knowledge, this is the largest coevolutionary analysis to date in terms of number of genomes analyzed, exceeding the 2167 genomes analyzed in previous work^{12,13}. Unlike popular prior approaches, such as ContextTree or SVD-phy^{42,69}, EvoWeaver's pairwise comparisons are independent and can be readily distributed across a cluster of computers. Second, EvoWeaver's predictions are more accurate because they incorporate multiple sources of coevolutionary signal, and each component algorithm incorporates statistical testing that mitigates spurious signals. This improves on prior work by allowing proteins to be classified by the extent of their functional relatedness rather than simple binary classifications. Third, EvoWeaver standardizes the application of multiple algorithms within a single software package with consistent inputs and outputs. This addresses usability issues previously identified in reviews of coevolutionary analyses¹³.

Coevolution differs from protein-protein interactions in that it does not require any physical interaction. Many prior approaches exist for predicting protein-protein interactions, along with databases of known interactors^{53,54,70,71}. Benchmarking functional association algorithms presents its own challenges, as proteins that do not physically interact may nevertheless be functionally associated¹⁷. This renders common benchmarks for protein-protein interactions insufficient for benchmarking coevolutionary algorithms^{71–73}. We chose to rely on the KEGG database as a source of experimentally validated functional associations within a multi-level hierarchy. Although KEGG is limited in size (i.e., 26,418 orthology groups), it is one of the few comprehensive sources of genomes and genes linked across pathways.

EvoWeaver is distinct from prior approaches for modeling functional associations. Alternative methods often depend on phylogenetic profiles calculated with similarity to a reference organism, limiting their generalizability. In contrast, EvoWeaver operates at the level of orthology groups, removing the dependence on a reference organism and increasing robustness to misspecifications in the underlying data. The application of a wide variety of coevolutionary algorithms further increases robustness by reducing dependence on any particular source of coevolutionary signal. These features circumvent the need to compare multiple approaches, streamlining the generation of hypotheses about the function of understudied proteins.

We anticipate EvoWeaver to be particularly useful for generating hypotheses that catalyze investigations into understudied proteins. EvoWeaver allows users to search through millions of gene pairs to find a comparatively small number of potential functional associations. EvoWeaver's predictions are particularly valuable when combined with network analyses or expert insights. In the future, EvoWeaver will assist in curating and supplementing large databases of biological knowledge to address errors and annotation inequality. We also expect EvoWeaver's predictions to be useful for other sequence features, such as non-coding RNAs, although protein-coding genes were the focus of this study. Most importantly, EvoWeaver empowers users to combat annotation inequality by predicting functional associations for the rapidly expanding collection of sequences with unknown function.

Methods

Experimental details

EvoWeaver is available as part of the *SynExtend* (v1.19.0) package⁷⁴ for R⁷⁵, which is distributed via the Bioconductor⁷⁶ platform. A comprehensive description of input/output and examples of running each algorithm are contained in the supplementary R Markdown file

available on GitHub (<https://github.com/WrightLabScience/EvoWeaver-ExampleCode>). Briefly, users first construct an EvoWeaver object with the *EvoWeaver* function using input gene groups, and then run the *predict* function to generate predictions using any of the 12 component algorithms. Depending on the algorithm, the input consists of a reference tree, gene trees, positional data, or sequences. The output is a matrix of scores, representing the pairwise strength of coevolution measured by each algorithm between each pair of gene groups. Scores range from -1 (strong negative association) to $+1$ (strong positive association). Detailed information about individual algorithms is described below and in SI1 of the Supplementary Information.

All analyses were performed with R (v4.4.1). Algorithms were implemented in EvoWeaver using the R and C programming languages, with user-exposed methods available in R via the *SynExtend* package (v1.19.0). *SynExtend* is dependent on the *DECIPHER* package (v3.2.0). Area under the receiver operator characteristic curves (AUROC) and Area under the precision-recall curves were calculated with the *AUC* function in the *DescTools* package (v0.99.49) for R.

Local analyses were performed on a MacBook Pro with M1 Pro CPU and 32GB of RAM. Runtimes were measured on a Dell PowerEdge T650 with an Intel Xeon processor (E5-2690 v4 2.6 GHz) and 792 GB of memory running Ubuntu 22.04.4 LTS. Distributed computing was performed on the Open Science Grid⁷⁷. Phylogenetic tree reconstruction used eight core nodes with 8–16 GB RAM and 8 GB disk space, and pairwise coevolutionary score calculations with EvoWeaver used single-core nodes with 2–4 GB RAM and 2–4 GB disk space. Computers matching these node specifications varied based on availability and Open Science Grid scheduling.

Coevolutionary algorithms in EvoWeaver

The goal of EvoWeaver is to capture a holistic view of coevolution for predicting functional associations between groups of genes. To achieve this, we implemented 12 algorithms from scratch that quantify different sources of coevolutionary signal. Each algorithm analyzes a pair of gene groups and returns a score between zero and one, where zero represents an absence of signal and more positive scores imply greater coevolutionary signal. Some algorithms can provide scores between -1 and 1 , where rare negative scores represent an inverse coevolutionary association. To correct for spurious signal resulting from insufficient information, we multiply all scores by their significance ($1 - p$ -value). The resulting final scores are combined into an overall prediction using an ensemble machine learning method. When an algorithm cannot make a prediction for a particular pair, the final score passed to the ensemble method for that algorithm is zero. For example, if a pair of genes do not cooccur in any organisms, then their final score for all Gene Organization algorithms is zero. The 12 algorithms we implemented fall into four categories: Phylogenetic Profiling, Phylogenetic Structure, Gene Organization, and Sequence Level methods (Fig. 1a). Of these, four algorithms are original to EvoWeaver (G/L Distance, P/A Overlap, RP ContextTree, and RP MirrorTree), four are new applications of existing algorithms (Tree Distance, Moran's I, Orientation MI, Gene Vector), and the remaining four are refinements on existing algorithms. Computational scaling for all algorithms in terms of number of gene groups and size of each gene group is available in SI3 of the Supplemental Information and Fig. S10. The problem of inferring functional associations among a set of gene groups is fundamentally quadratic in the number of gene groups. As such, we designed our algorithms to return the same results whether analyzing one pair or millions of pairs of gene groups to facilitate scalability through distributed computing.

Phylogenetic Profiling. Phylogenetic profiling is a common technique that uses presence/absence (P/A) profiles of genes to investigate shared function. The approaches previously introduced in the

literature use binary P/A profiles, where one represents the presence of a gene and zero represents its absence⁷⁸. The first P/A approach used Hamming distances on binary profiles as a score⁷⁹. Later, Jaccard index and mutual information (MI) were applied to score P/A profiles^{12,80}. Subsequent work accounted for clade-wise conservation²⁵ or transformed P/A profiles into ancestral gain/loss (G/L) events and scored the correlation between events⁸¹. These transformations reduce redundancy for sets of organisms with low rates of gene gain and loss^{36,81}.

EvoWeaver includes four Phylogenetic Profiling algorithms (Fig. 1b and S11). The first algorithm, P/A Jaccard, uses the centered Jaccard index⁸² of P/A profiles with conserved clades collapsed to mitigate bias from closely related organisms. The second algorithm, P/A Overlap, applies Fitch Parsimony⁸³ to infer ancestral states on the reference tree from P/A profiles and calculates the proportion of the tree where both genes appear together relative to their overall prevalence. The third algorithm, G/L MI, calculates weighted MI of G/L events (G/L profiles). G/L profiles include three states: -1 for gene loss, 0 for no change, and +1 for gene gain. G/L MI uses the weighted mutual information of four cases: simultaneous concordant transitions (i.e., gain/gain or loss/loss), simultaneous gain in gene one and loss in gene two, simultaneous gain in gene two and loss in gene one, and non-simultaneous transitions. MI is calculated by weighting the first case with +1, the second and third cases with -1, and the fourth case with 0.

G/L MI fails to adequately measure compensatory changes that do not occur on the same branch of the reference tree, which are common in sequence evolution⁸⁴. The fourth algorithm, G/L Distance, complements the previous algorithms by quantifying the evolutionary distance between G/L events assuming the time between gain or loss events is exponentially distributed. Thus, the score between a pair of events for two gene groups is calculated as $w e^{-d(v_1, v_2)}$, where w is +1 if the events are the same (i.e., both gain or both loss) and -1 if the events are different, and $d(v_1, v_2)$ is the distance between events v_1 and v_2 on the reference tree. The distance between events on separate branches is defined as the total distance between their branch midpoints. The distance between events on the same branch is defined as zero. For each pair of genes, events are paired to their closest event from the other group. The total score for the gene pair is the average score for all event pairs, and ranges from -1 to +1.

Statistical significance for P/A Jaccard is calculated using an empirical distribution of scores obtained from bootstrapping P/A vectors. Significance for G/L MI is calculated using Fisher's Exact Test on the contingency table of the four cases, and p -values for P/A Overlap and G/L Distance are calculated using empirical values from permutation testing.

Phylogenetic Structure. Gene tree structural comparisons were pioneered by MirrorTree⁴⁰, which scores each pair of gene groups by the correlation of their pairwise sequence distances. Subsequent improvements to MirrorTree attempted to correct for background evolutionary signal prior to analysis⁸⁵. These extensions, often referred to as ContextTree or ContextMirror, use different approaches to remove the shared signal represented by the reference tree^{39,69,86}. More recently, SVD-phy was introduced as an alternative approach using SIMAP⁸⁷ or BLAST to measure distance between sequences^{42,88}. SVD-phy uses singular value decomposition (SVD) to reduce redundant information contained in pairwise distances, which removes signal shared across all genes and improves overall predictions. However, this approach requires that all pairwise distances be simultaneously kept in memory.

EvoWeaver uses random projection in lieu of SVD for dimensionality reduction. Random projection (RP) is a surjective mapping that approximately preserves distances between vectors⁸⁹. While traditional RP uses a large matrix of random values, this requirement can be circumvented by generating values of the matrix on demand with a

preset random seed. Hence, this dimensionality reduction can be done with negligible memory overhead, allowing for efficient and replicable distribution across a compute cluster. The RP MirrorTree algorithm applies RP to patristic distances and scores pairs of vectors using Spearman's correlation coefficient. The RP ContextTree algorithm also subtracts the reference tree from each distance matrix prior to random projection and scoring. RP ContextTree's final scores are multiplied by the Hamming distance of overlap in organism membership to correct for spurious correlations caused by minimally overlapping sets. Statistical significance for both RP ContextTree and RP MirrorTree is calculated using the closed-form solution for significance of Spearman's correlation coefficient.

EvoWeaver also incorporates tree distance metrics to measure topological similarity. A variety of previously benchmarked metrics⁴³ were implemented as measures of functional similarity, all of which were highly correlated in their tree distances. By default, EvoWeaver's Tree Distance predictor uses normalized Robinson-Foulds Distance due to its low memory requirement and closed form solution for significance⁹⁰, though other tree distance metrics are also supported. The score for each pair of genes was defined as one minus the tree distance of the gene trees pruned to their common leaves. If two gene groups do not appear in any common genomes, their Tree Distance score is set to zero.

Gene Organization. Gene organization is commonly used as a signature of functional association. For example, a priori knowledge of genes that colocalize can be used to find biosynthetic gene clusters. Existing programs, such as antiSMASH⁴⁷, use profile hidden Markov models to search for clusters of genes with known functional associations. However, these approaches cannot be used to find gene clusters de novo. STRING makes use of the distance in nucleotides between genes as a de novo predictor of functional association²⁴. To our knowledge, analysis of gene organization is one of the most understudied approaches for de novo prediction of functional associations.

EvoWeaver incorporates three Gene Organization algorithms. Together, they provide a well-rounded view of gene organization: the first algorithm looks at whether genes possibly share regulation, the second measures how closely genes are located to each other, and the third quantifies the extent to which gene distances are preserved across phylogenies. The first algorithm, Orientation MI, examines the relative orientation of paired genes. Conservation of relative gene direction has been validated in prior work to be indicative of shared function¹⁹. The score for Orientation MI is defined as the bidirectional mutual information⁹¹ between the orientation of paired genes, with Fisher's Exact Test used to determine statistical significance.

The second algorithm, Gene Distance, examines the separation between genes. For each pair of genes on the same chromosome or contig, the distance d is calculated as the absolute value of the difference in gene index. The index of a gene is its gene order in the chromosome or contig, starting from one for the first gene. We used indices rather than nucleotide locations to mitigate the effect of variability in gene lengths. The score for each pair of sequences is defined as e^{1-d} , and the overall score for a pair of gene groups is the mean of their sequence pair scores. In this way, Gene Distance is maximized (1) when two genes are always adjacent ($d=1$). Statistical significance is derived from the distribution of distances between two random points on a line segment⁹². If a pair of gene groups never appears in the same organism on the same chromosome/contig, the score for the pairing is defined as zero.

The third algorithm, Moran's I, measures spatial autocorrelation among gene distances. Moran's I requires pairwise weights represented by the inverse exponential of the patristic distances⁹³ and values in the form of gene distances (d). Moran's I measures the extent to which the relative distances between genes are correlated with the

evolutionary trajectories of their respective organisms on the reference tree. Statistical significance is calculated using the closed form solution to the expected value and variance of Moran's I ref. 94.

Sequence Level methods. Covariation of residues is a common signal of protein-protein interactions, and numerous methods have been devised for this purpose. A popular approach is direct coupling analysis⁵⁴, which fits a Potts model to a multiple sequence alignment in order to parse direct effects from indirect effects. Other algorithms using deep learning have been successfully applied to sequencing data for finding interaction sites between proteins^{95,96}. While some previously developed approaches improved scaling^{97,98}, many of these algorithms have prohibitively high computational complexity for high-throughput analysis. Additionally, the focus of these algorithms is on finding interaction sites between small numbers of proteins or proteins known a priori to have a high likelihood of interacting.

EvoWeaver implements two Sequence Level methods that support either amino acid or nucleotide sequences, although amino acid sequences were utilized for all analyses in this work. The first of these, Gene Vector, uses the gene sequence natural vector approach, developed to predict protein-protein interactions⁵⁶. We extended this algorithm to amino acids following the same theoretical model as the initial nucleotide-based method. We chose to use the natural vector without 2-mers or 3-mers, since the full vector incurred higher computational overhead with a negligible difference in scores. For each pair of gene groups, we subset the sequences to the intersection of the organisms present in both groups. The natural vector for each group in the pair is the average of the natural vectors for each of its constituent sequences. We centered each natural vector assuming a null model of equally distributed nucleotides or amino acids. The final score and statistical significance for the pairing are calculated from Spearman's correlation coefficient of the natural vectors.

The second approach, Sequence Info, extends a prior approach to measure the mutual information between sites within sequence alignments of each gene group⁵⁵. For each pair of gene groups, we subset the sequences to the genomes that appear in both groups, and subset the sites to those with high information content (entropy ≥ 0.3 bits) using the MaskAlignment function in *DECIPHER*⁹⁹. Mutual information is calculated for each pair of sites (i.e., columns) across both alignments after applying a background entropy correction along with an average product correction¹⁰⁰. The final score is calculated as the average of the highest-scoring pairing for each site. Statistical significance is calculated by applying Fisher's combined probability test to the distribution of *p*-values across sites.

Ensemble methods. EvoWeaver combines the output of all 12 coevolutionary algorithms into a final prediction using an ensemble machine learning method (Fig. 1a). For ensemble methods, we tested logistic regression, random forest, and neural network models in R⁷⁵. Logistic regression was performed with the *glm* function with family = "binomial", random forests using the *randomForest* package¹⁰¹ (v4.7-1.1), and neural networks using the *neuralnet* package (v1.44.2)¹⁰². The random forest model used maxdepth = 25 for binary classification and maxdepth = 100 for multiclass classification to avoid overfitting trees of unlimited depth. The neural network architecture was a feed forward network with 12 inputs, one hidden layer of matched size (i.e., 12), two output nodes (i.e., class = 0 or class = 1), and sigmoid activation functions on each node. We intentionally chose relatively simple architectures with default parameters for our ensemble models to maintain interpretability of the predictions and mitigate overfitting to the dataset. All models were evaluated using five-fold cross-validation without hyperparameter tuning.

Only random forest was used for hierarchical classification due to its better performance in the binary classification benchmarks. Hierarchical classification was also evaluated using five-fold cross-

validation. Members of each class were distributed equally among each train/test fold. To prevent overfitting from high class imbalance in the complete dataset, we downsampled classes in each training set to match the size of the smallest class, Direct Connection, with 899 members. This meant that each class in the train set for each fold had 719 members (i.e., 80%). Testing was done on the complete set of data partitioned for testing, which comprised 128,552–128,557 members (i.e., ~20%) per fold. Each pair was in exactly one test set, and no pairs belonged to both the train and test set for any fold. Feature importance for the random forest model was calculated using permutation importance, which was chosen over mean decrease in Gini impurity since the latter has been shown to produce biased estimates¹⁰³.

To construct an example network, we first created a weighted adjacency matrix from the random forest predictions. Each node represented a single gene group and was connected to its top two Direct Connection predictions with edges of weight 1.0. All predicted Same Module pairs were connected with edges of weight 0.5. Our basis for this approach is that most module blocks in KEGG are directly connected to two neighbors, and other nodes in the same module are less important than direct connections. We then used Louvain clustering implemented in the *igraph* package¹⁰⁴ (v1.5.0.1) to perform community detection. The network in Fig. 3c was randomly chosen from the resulting communities.

A possible concern with holding out pairs in cross-validation is that ensemble methods could use spurious signals to simply distinguish highly connected gene groups from less connected groups. On binary benchmarks, we further validated our results by reevaluating our ensemble classifier using ten-fold cross-validation with gene group holdouts rather than pair holdouts. Within each fold, 10% of gene groups were randomly selected, and all pairs involving at least one of these groups was taken as the test set. The resulting train/test sets each comprised roughly 80/20% of the data (respectively), which forms a comparable scenario to five-fold cross-validation with pair holdouts. We also evaluated the impact of module/complex holdouts, which were performed similarly to gene group holdouts. The results of these analyses were virtually identical to prior results (Figs. S6, 12, 13), implying that EvoWeaver is not heavily relying on spurious signals when making predictions. More details on factors impacting EvoWeaver's performance are available in section SI3 of the Supplementary Information.

A potential limitation of ensemble methods is their generalizability to new datasets. Training an ensemble method necessitates having access to a gold-standard training set. Furthermore, distinct sets of organisms may exhibit fundamentally different coevolutionary patterns, leading to decreased performance on out-of-distribution data (e.g., Fig. S2). To analyze EvoWeaver performance under these constraints, we investigated the quality of naively combining EvoWeaver's component algorithms (Fig. S5). A simple sum of the 12 coevolutionary algorithms' scores produced results similar to that of the best component algorithm and substantially better than the median performing algorithm (i.e., the 6th best performing algorithm on each dataset). Notably, this result was most pronounced at low false positive rates and requires no a priori knowledge or training. This demonstrates that EvoWeaver generates informative predictions with minimal knowledge about the underlying data.

Construction of benchmark datasets

The goal of the Complexes benchmark is to judge each algorithm's ability to discern genes encoding proteins involved in a complex versus genes encoding unrelated proteins. To this end, we identified all orthology groups belonging to a complex in KEGG¹⁰⁵, for a total of 372 gene groups. We computed pairwise coevolutionary scores between orthology groups with at least three sequences that were involved in a complex, for a total of 358 orthology groups. This resulted in 57,321

pairs of orthology groups that are not in the same pathway (unrelated pairs) and 867 pairs participating as required or optional components of the same complex. Importantly, there was negligible similarity between distinct orthology groups (Fig. S14), which might have otherwise resulted in data leakage. Positive pairs were defined as the 867 pairs from the same complex, and an equivalent number of negative pairs were drawn to create a balanced dataset for benchmarking. Random sampling of negative pairs was weighted in order to match the distribution in number of sequences per gene group to that of the positive pairs. This weighted sampling was used to mitigate the ability of algorithms to use the number of sequences per group as a proxy for functional association.

Next, we constructed the Modules benchmark to test each algorithm's ability to discern proteins acting in subsequent steps of a biochemical pathway versus unrelated proteins. We first identified all module blocks within the KEGG MODULES database. Each module block is a set of one or more orthology groups that perform a discrete step within a biochemical pathway (Fig. 1g). Each module was parsed from its definition on KEGG (Table S1), for a total of 1547 module blocks from 369 modules. Positive test cases were defined as successive blocks in a module, and negative cases were defined as module blocks in separate modules not sharing a pathway in KEGG. KEGG's Global and Overview Pathways were not considered, since their broad definition encompasses most proteins in KEGG. Blocks containing complexes were also excluded to prevent overlap with the Complexes benchmark. Since some orthology groups belong to multiple blocks, only pairs of blocks without overlap in orthology groups were assessed. The final Modules benchmark was comprised of 1187 blocks with 899 positive pairs. An equivalent number of negative pairs were sampled in the same manner as in the Complexes benchmark.

Having constructed two binary benchmarks, we constructed the Multiclass benchmark to explore EvoWeaver's ability to distinguish interaction strengths among proteins. Accordingly, we used the relationships encoded in the KEGG PATHWAYS database to define multiple hierarchical levels of functional association. We assigned all pairs of module blocks into one of five categories: Direct Connection, Same Module, Same Pathway, Same Global Pathway, or Unrelated. The Same Pathway group comprises pairs of module blocks that share a pathway not in the Global and Overview Pathways category in KEGG, and the Unrelated group comprises pairs with no modules or pathways in common. We chose 50% confidence as the cutoff for classification (Fig. 3a) because these predictions have higher probability assigned to their predicted class than their sum of probabilities across all other classes. The confusion matrix at 0% confidence is shown in Fig. S6. To look for connections absent from KEGG (Fig. 5), we examined pairs belonging to Unrelated and Same Global Pathway groups that EvoWeaver predicted as being Direct Connection. More details on benchmark datasets and data preprocessing are available in section S12 of the Supplemental Information. A list of all misclassifications is available in Supplemental Datafile 1.

Preparing gene groups for analysis

EvoWeaver takes as input a set of two or more gene trees, which may include sequences, gene indexes, and/or a reference tree. It then applies the set of component algorithms for which it has the necessary input data types. We obtained amino acid sequences for each gene group from KEGG and used *DECIPHER*⁹⁹ to align sequences and construct neighbor-joining gene trees. In total, there were 8564 genomes with at least one gene present in the benchmarks. Reference trees were estimated using the ASTRID algorithm¹⁰⁶. Impact of error in the reference tree is discussed in sections S11 and S13 of the Supplementary Information and shown in Fig. S15. To find each gene's index within its genome, we downloaded the set of all genes available for each organism from KEGG, along with their chromosome/contig, orientation, and location. We mapped locations to indices by calculating the

index of each gene relative to all other genes on the same chromosome/contig available for that genome. Of the 8564 genomes present in the benchmarks, 8136 had location data available in KEGG. A taxonomic breakdown of the genomes used and their location data is available in Supplemental Datafile 2.

Comparison with STRING

Data for STRING's clusters of orthologous genes (COGs) and interactions were downloaded from STRING v12.0. Since STRING's COG membership sometimes did not perfectly correspond to KEGG's KO groups, we tabulated the KO group assignments for sequences belonging to each STRING COG. Overall, 6849 COGs had at least one sequence that could be mapped to a KO group in KEGG. Each STRING COG was mapped to KEGG Module blocks using its majority ($\geq 50\%$) KEGG KO group. A total of 6311 COGs had a majority KO group, and 4481 (71%) of these COGs had perfect consensus. Only 538 STRING COGs lacked a consensus KO group, and these COGs were excluded from analysis.

STRING's stated goal for its Total Score is to estimate how likely a reported functional linkage between two proteins "is at least as specific as that between an average pair of proteins annotated on the same 'map' or 'pathway' in KEGG"⁵⁷. Therefore, EvoWeaver's analogous predictions were made by summing the probabilities predicted for Direct Connection, Same Module, and Same Pathway in the hierarchical classification (Fig. 3). A total of 757 pairs of COGs in the matched dataset belonged to the Same Pathway, Same Module, or Direct Connection categories in KEGG. Note that this differs from the pairs used in the Modules Benchmark, which only included pairs in the Direct Connection category. An equivalent number of negatives were randomly sampled from the remaining pairs. STRING provides its Total Score calculation within a Python script available on their website. We used this formula to calculate the hypothetical Total Score using subsets of STRING's evidence streams. The sequence of AUROCs in Fig. 4a was obtained by sequentially adding evidence streams from lowest to highest marginal impact on AUROC to the Total Score calculation (Fig. S7). The "de novo only" prediction in Fig. 4a is STRING's Total Score using only the three evidence streams that do not depend on prior knowledge (i.e., Gene Fusion, Cooccurrence, and Gene Neighborhood), while "EvoWeaver Transfer" refers to using an EvoWeaver ensemble model trained on the CORUM prediction task (Fig. S2).

KEGG case studies

Case studies for Fig. 6 were manually constructed from KEGG data for biologically meaningful sets of KEGG modules belonging to the same KEGG pathway. Gene groups were connected according to the directed connections available in the corresponding KEGG pathway. Only gene groups used in the Multiclass benchmark were included. For the ensemble network, we connected each node to its top connection, where each connection's ranking is determined from its probability of Direct Connection according to the Multiclass classifier used in Fig. 3. The component predictor networks were constructed similarly to the ensemble network, but used the average rank of the component score ranks for calculating each node's top connections. For example, the Phylogenetic Profiling connections are determined by the mean rank of P/A Jaccard, P/A Overlap, G/L MI, and G/L Distance.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Source data are provided with this paper. Data for reproducing figures and pretrained ensemble models used in this work are available on GitHub (<https://github.com/WrightLabScience/EvoWeaver>).

ExampleCode, <https://doi.org/10.5281/zenodo.15027870>). All other datafiles are available from Zenodo (<https://doi.org/10.5281/zenodo.14205427>). The list of EvoWeaver's misclassifications and relevant algorithm scores are available in the Supplemental Datafiles. Unique identifiers for data downloaded from KEGG are available on Zenodo as well as in Supplemental Datafile 1. Accession codes for genomes used are available in Supplemental Datafile 2. Source data are provided with this paper.

Code availability

EvoWeaver is available under a GPL-3 license as part of the *SynExtend* (v1.19.0) package⁷⁴ for R⁷⁵, which is distributed via the Bioconductor⁷⁶ platform. The code used to develop the model, perform the analyses and generate results in this study is publicly available and has been deposited on GitHub at <https://github.com/WrightLabScience/EvoWeaver-ExampleCode>¹⁰⁷, under an MIT license (<https://doi.org/10.5281/zenodo.15027870>).

References

- Kustatscher, G. et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).
- Kustatscher, G. et al. An open invitation to the understudied proteins initiative. *Nat. Biotechnol.* **40**, 815–817 (2022).
- Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbajji, B. & Eisenhaber, F. Darkness in the human gene and protein function space: widely modest or absent illumination by the life science literature and the trend for fewer protein function discoveries since 2000. *Proteomics* **18**, 1800093 (2018).
- Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018).
- Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* **20**, 92 (2019).
- 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.* **6**, <https://doi.org/10.1099/mgen.0.000341> (2020).
- Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biol.* **16**, e2006643 (2018).
- 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.* **9**, e1003063 (2013).
- Gillis, J. & Pavlidis, P. The impact of multifunctional genes on “Guilt by Association” analysis. *PLoS ONE* **6**, e17258 (2011).
- Aravind, L. Guilt by association: contextual information in genome analysis. *Genome Res.* **10**, 1074–1077 (2000).
- Codoñer, F. M. & Fares, M. A. Why should we care about molecular coevolution? *Evol. Bioinform Online* **4**, 29–38 (2008).
- Moi, D., Kilchoer, L., Aguilar, P. S. & Dessimoz, C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLOS Comput. Biol.* **16**, e1007553 (2020).
- Moi, D. & Dessimoz, C. Phylogenetic profiling in eukaryotes comes of age. *Proc. Natl. Acad. Sci.* **120**, <https://doi.org/10.1073/pnas.2305013120> (2023).
- Canavati, C. et al. Using multi-scale genomics to associate poorly annotated genes with rare diseases. *Genome Med.* **16**, <https://doi.org/10.1186/s13073-023-01276-2> (2024).
- Kann, M. G., Shoemaker, B. A., Panchenko, A. R. & Przytycka, T. M. Correlated evolution of interacting proteins: looking behind the Mirrortree. *J. Mol. Biol.* **385**, 91–98 (2009).
- Chikina, M., Robinson, J. D. & Clark, N. L. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol. Biol. Evol.* **33**, 2182–2192 (2016).
- Little, J., Chikina, M. & Clark, N. L. Evolutionary rate covariation is a reliable predictor of co-functional interactions but not necessarily physical interactions. *eLife* **12**, RP93333 (2024).
- Umehura, M., Koike, H. & Machida, M. Motif-independent de novo detection of secondary metabolite gene clusters-toward identification from filamentous fungi. *Front. Microbiol.* **6**, 371–371 (2015).
- Korbel, J. O., Jensen, L. J., Von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22**, 911–917 (2004).
- Cotroneo, C. E., Gormley, I. C., Shields, D. C. & Salter-Townshend, M. Computational modelling of chromosomally clustering protein domains in bacteria. *BMC Bioinforma.* **22**, 593 (2021).
- Feinauer, C., Szurmant, H., Weigt, M. & Pagnani, A. Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the Trp operon. *PLOS ONE* **11**, e0149166 (2016).
- Clark, G. W., Ackerman, S. H., Tillier, E. R. & Gatti, D. L. Multi-dimensional mutual information methods for the analysis of covariation in multiple sequence alignments. *BMC Bioinforma.* **15**, 1–12 (2014).
- Bitbol, A.-F. Inferring interaction partners from protein sequences using mutual information. *PLOS Comput. Biol.* **14**, e1006401 (2018).
- Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
- Stupp, D. et al. Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.* **12**, <https://doi.org/10.1038/s41467-021-26792-w> (2021).
- Tabach, Y. et al. Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol.* **9**, 692 (2013).
- Tabach, Y. et al. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493**, 694–698 (2013).
- Sherill-Rofe, D. et al. Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res.* **29**, 439–448 (2019).
- Andreo-Vidal, A., Binda, E., Fedorenko, V., Marinelli, F. & Yushchuk, O. Genomic insights into the distribution and phylogeny of glycopeptide resistance determinants within the actinobacteria phylum. *Antibiotics* **10**, <https://doi.org/10.3390/antibiotics10121533> (2021).
- Ding, D. et al. Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat. Ecol. Evol.* **6**, 590–603 (2022).
- Fongang, B. et al. Coevolution combined with molecular dynamics simulations provides structural and mechanistic insights into the interactions between the integrator complex subunits. *Comput. Struct. Biotechnol. J.* **21**, 5686–5697 (2023).
- Ramani, A. K. & Marcotte, E. M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284 (2003).
- Fukunaga, T. & Iwasaki, W. Inverse Potts model improves accuracy of phylogenetic profiling. *Bioinformatics* **38**, 1794–1800 (2022).
- Cheng, Y. & Perocchi, F. ProtPhylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic Acids Res.* **43**, W160–W168 (2015).
- Ji, F. et al. DEPCOD: a tool to detect and visualize co-evolution of protein domains. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkac349> (2022).
- Škunca, N. & Dessimoz, C. Phylogenetic profiling: how much input data is enough? *PLOS ONE* **10**, e0114701 (2015).

37. Shin, J. & Lee, I. Co-inheritance analysis within the domains of life substantially improves network inference by phylogenetic profiling. *PLOS ONE* **10**, e0139006 (2015).
38. Clark, N. L., Alani, E. & Aquadro, C. F. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.* **22**, 714–720 (2012).
39. Pazos, F., Ranea, J. A., Juan, D. & Sternberg, M. J. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–1015 (2005).
40. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng., Des. Select.* **14**, 609–614 (2001).
41. Clark, G. W. et al. In *Network Biology: Methods and Applications* Vol. 781 (eds Gerard, C. & Andrew, E.) 237–256 (Humana Press, 2011).
42. Franceschini, A., Lin, J., von Mering, C. & Jensen, L. J. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics* **32**, 1085–1087 (2016).
43. Smith, M. R. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**, 5007–5013 (2020).
44. Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.* **16**, 731–744 (2018).
45. Periwal, V. & Scaria, V. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics* **31**, 1–9. <https://doi.org/10.1093/bioinformatics/btu600> (2014).
46. Rocha, E. P. The organization of the bacterial genome. *Annu. Rev. Genet.* **42**, 211–233 (2008).
47. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
48. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
49. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–d497 (2021).
50. Davila Lopez, M., Martinez Guerra, J. J. & Samuelsson, T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLOS ONE* **5**, e10654 (2010).
51. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of protein–protein interaction specificity from correlated mutations and interaction data. *Proteins Struct. Funct. Bioinforma.* **76**, 911–929 (2009).
52. Morcos, F., Hwa, T., Onuchic, J. N. & Weigt, M. Direct coupling analysis for protein contact prediction. *Methods Mol. Biol.* **1137**, 55–70 (2014).
53. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**, E1293–E1301 (2011).
54. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci.* **106**, 67–72 (2009).
55. Martin, L. C., Gloor, G. B., Dunn, S. D. & Wahl, L. M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116–4124 (2005).
56. Zhao, N., Zhuo, M., Tian, K. & Gong, X. Protein–protein interaction and non-interaction predictions using gene sequence natural vector. *Commun. Biol.* **5**, <https://doi.org/10.1038/s42003-022-03617-0> (2022).
57. Szklarczyk, D. et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
58. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
59. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–d595 (2019).
60. Launay, G., Ceres, N. & Martin, J. Non-interacting proteins may resemble interacting proteins: prevalence and implications. *Sci. Rep.* **7**, 40419 (2017).
61. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76**, 036106 (2007).
62. Wang, Z. et al. High expression of lactotriaosylceramide, a differentiation-associated glycosphingolipid, in the bone marrow of acute myeloid leukemia patients. *Glycobiology* **22**, 930–938 (2012).
63. Togayachi, A. et al. Molecular cloning and characterization of UDP-GlcNAc: lactosylceramide β 1, 3-N-acetylglucosaminyltransferase (β 3Gn-T5), an essential enzyme for the expression of HNK-1 and Lewis X epitopes on glycolipids. *J. Biol. Chem.* **276**, 22032–22040 (2001).
64. Boccuto, L. et al. A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. *Hum. Mol. Genet.* **23**, 418–433 (2013).
65. Alam, S. et al. Altered (neo-) lacto series glycolipid biosynthesis impairs α 2-6 sialylation on N-glycoproteins in ovarian cancer cells. *Sci. Rep.* **7**, 45367 (2017).
66. Lau, K. S. et al. Complex N-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. *Cell* **129**, 123–134 (2007).
67. Ebert, B. et al. A Golgi UDP-GlcNAc transporter delivers substrates for N-linked glycans and sphingolipids. *Nat. Plants* **4**, 792–801 (2018).
68. Zhang, W. et al. Synthesis of paucimannose N-glycans by *Caenorhabditis elegans* requires prior actions of UDP-N-acetyl-D-glucosamine: α -3-D-mannoside β 1, 2-N-acetylglucosaminyltransferase I, α 3, 6-mannosidase II and a specific membrane-bound β -N-acetylglucosaminidase. *Biochem. J.* **372**, 53–64 (2003).
69. Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M. & Toh, H. Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. *Bioinformatics* **22**, 2488–2492 (2006).
70. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
71. Blohm, P. et al. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.* **42**, D396–D400 (2014).
72. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
73. Oughtred, R. et al. TheBioGRIDdatabase: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
74. Cooley, N., Lakshman, A. & Wright, E. S. *SynExtend: Tools for Working with Synteny Objects*. **v1.17.1**. <https://doi.org/10.18129/B9.bioc.SynExtend> (2024).
75. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
76. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, 1–16. <https://doi.org/10.1186/gb-2004-5-10-r80> (2004).

77. OSG. Open Science Data Federation. OSG. <https://doi.org/10.21231/OKVZ-VE57> (2015).
78. Brilli, M. et al. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinforma.* **9**, 551 (2008).
79. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96**, 4285–4288 (1999).
80. Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062 (2003).
81. Dembech, E. et al. Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc. Natl. Acad. Sci.* **120**, e2218329120 (2023).
82. Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinforma.* **20**, 644 (2019).
83. Fitch, W. M. On the problem of discovering the most parsimonious tree. *Am. Nat.* **111**, 223–257 (1977).
84. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza surface proteins. *PLOS Genet.* **7**, e1001301 (2011).
85. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci.* **105**, 934–939 (2008).
86. Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**, 3482–3489 (2005).
87. Arnold, R., Goldenberg, F., Mewes, H.-W. & Rattei, T. SIMAP—the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. *Nucleic Acids Res.* **42**, D279–D284 (2013).
88. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
89. Achlioptas, D. Database-friendly random projections. In: *Proc. Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 274–281 (ACM, 2001).
90. Steel, M. A. & Penny, D. Distributions of tree comparison metrics—some new results. *Syst. Biol.* **42**, 126–141 (1993).
91. Beckley, A. M. & Wright, E. S. Identification of antibiotic pairs that evade concurrent resistance via a retrospective analysis of antimicrobial susceptibility test results. *Lancet Microbe* **2**, e545–e554 (2021).
92. Philip, J. *The Probability Distribution of the Distance Between Two Random Points in a Box*. KTH Mathematics (Royal Institute of Technology, 2007).
93. Gittleman, J. L. & Kot, M. Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Biol.* **39**, 227–241 (1990).
94. Cliff, A. D. & Ord, J. K. *Spatial Processes: Models and Applications* (Pion Limited, 1981).
95. Pesaranghader, A. et al. deepSimDEF: deep neural embeddings of gene products and Gene Ontology terms for functional analysis of genes. *Bioinformatics* **38**, 3051–3061 (2022).
96. Soleymani, F., Paquet, E., Viktor, H. L., Michalowski, W. & Spinello, D. ProtInteract: a deep learning framework for predicting protein–protein interactions. *Comput. Struct. Biotechnol. J.* **21**, 1324–1348 (2023).
97. Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* **276**, 341–356 (2014).
98. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2011).
99. Wright, E. S. Using DECIPHER v2.0 to analyze big biological sequence data in R. *J. R. J.* **8**, 352–359 (2016).
100. Buslje, C. M., Santos, J., Delfino, J. M. & Nielsen, M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* **25**, 1125–1131 (2009).
101. Liaw, A. & Wiener, M. Classification and regression by random Forest. *R. N.* **2**, 18–22 (2002).
102. Fritsch, S., Guenther, F. & Wright, M. neuralnet: Training of Neural Networks. R package version 1.44.2 <https://github.com/bips-hb/neuralnet> (2019).
103. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinforma.* **8**, 25 (2007).
104. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.*, 1695. <https://doi.org/10.5281/zenodo.7682609> (2006).
105. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2015).
106. Vachaspati, P. & Warnow, T. ASTRID: accurate species trees from internode distances. *BMC Genom.* **16**, S3 (2015).
107. Lakshman, A. & Wright, E. S. EvoWeaver: large-scale prediction of gene functional associations from coevolutionary signals. *EvoWeaver-ExampleCode*. <https://doi.org/10.5281/zenodo.15027870> (2025).

Acknowledgements

This study was funded by the NIAID at the NIH (grant number 1U01AI176418-01). AHL was supported by the NLM at the NIH (award number 5T15LM007059-35). This research was done using services provided by the OSG Consortium. Figures 1, 3c, d, and 5a, b were created using BioRender.

Author contributions

A.L. and E.S.W. designed algorithms, analyzed results, and contributed to the final version of the manuscript. A.L. wrote code and generated plots. E.S.W. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59175-6>.

Correspondence and requests for materials should be addressed to Erik S. Wright.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025