TITLE

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

AUTHORS

Aidan Lakshman[1] and Erik S. Wright[1,2,*]

[1]Department of Biomedical Informatics, University of Pittsburgh
[2]Center for Evolutionary Biology and Medicine, Pittsburgh, PA
*address correspondence to eswright@pitt.edu

## ABSTRACT

The 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, an algorithm 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 1,545 gene groups from 8,564 genomes reveals missing connections in popular databases and potentially undiscovered links between proteins.

# INTRODUCTION

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 health[4]. 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 coevolution[11]. Four primary approaches are used to identify coevolution: phylogenetic profiling[12], phylogenetic structure[13], gene organization[14], and sequence-level methods[15]. 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 approaches[16].

Although coevolutionary analyses have shown great potential for predicting functional associations[17-24], 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 evolution[25]. Additionally, holistic evaluation of many coevolutionary signals offers a means of amplifying weaker signals to make higher accuracy predictions. For example, conserved genes will not display a phylogenetic profiling (i.e., presence/absence) signal but may show patterns of gene organization. Combining disparate coevolutionary signals and 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 agonistic 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 incorporate 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, 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 interpretability[25]. 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 analysis, comprised of 12 algorithms optimized for scalable performance. These component predictors are combined using a machine learning classifier to compute a strength of coevolution between every pair of gene groups. From this, users can generate novel inferences or hypotheses about gene function.

The first type of coevolutionary analysis, phylogenetic profiling, investigates patterns of presence/absence or gain/loss of genes, which manifests when multiple genes work in concert (Fig. 1b). While presence/absence analyses have been successfully used to predict gene function[12,25-27], existing approaches are susceptible to biases from small sample sizes or low evolutionary divergence[28]. We addressed these biases with a novel algorithm (G/L Distance) that examines the distance between gain/loss events to measure compensatory changes rather than extant patterns. We also incorporated statistical testing into existing measures of presence/absence patterns[12,29] (P/A Info, P/A Jaccard) and correlation of ancestral states[30] (G/L Correlation). 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 tandem, giving rise to similar genealogies (Fig. 1c). Commonly used phylogenetic structure approaches include MirrorTree and ContextTree[31-33], although these approaches scale poorly due to high computational complexity. We addressed this issue by introducing novel algorithms (RP MirrorTree, RP ContextTree) that use random projection 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-phy[34], 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 genealogies[35]. Taken together, these algorithms facilitate inference of coevolution among 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[36-38] (Fig. 1d). These approaches most commonly employ profile hidden Markov models, such as antiSMASH[39-41]. While these approaches perform well on functional prediction, they rely on *a priori* knowledge about genes that colocalize. 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 genes[16], 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 genes remain colocalized in spite of evolutionary divergence. Additionally, EvoWeaver analyzes the conservation of relative transcriptional direction (Transcription Info), since this also indicates functional association[42]. 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 products[43] (Fig. 1e). Direct coupling analysis is a well-known approach in this category[44-46], but it suffers from high computational complexity. Instead, we extended a prior approach based on mutual information to predict interacting sites between sequences[47]. 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 computed[48]. These algorithms provide additional evidence of coevolution for physically interacting gene products.

These four categories 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 with previously unknown associations and no relationship to the training set.

## Ensemble methods accurately identify functionally associated genes

Selection of high-quality ground truth datasets for coevolutionary analysis is a challenging task[25]. As with previous studies[34,49], we relied upon the Kyoto Encyclopedia of Genes and Genomes database (KEGG) because it is well-curated and experimentally validated[50,51]. 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 within the same complex. 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 risk of overfitting by matching the distribution of data features in the negative set to the positive set.

Almost all coevolution 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 proteins[52]. 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 Random Forest performing the best (Fig. S1).

Given EvoWeaver's excellent 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 1,948 pairs of gene groups acting in adjacent steps of a biochemical pathway (positives) and 1,948 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 Figure 2, performance of component algorithms on the Modules benchmark was slightly worse than on the Complexes benchmark. However, ensemble methods retained high performance (AUROC of 0.981 for Random Forest) and greatly outperformed individual coevolutionary signals. The large gap between ensemble and component predictors highlights the importance of using multiple coevolutionary signals to infer functional associations.

## EvoWeaver infers hierarchical relationships among genes

Coevolutionary relationships are stratified across a gradient of associations within the cell. For this reason, it would be ideal to predict a strength of coevolution across a hierarchy of multi-level relationships among gene groups. We evaluated the Random Forest model on pairs of KEGG module blocks belonging to each of five classes: Direct Connection, Same Module, Same Pathway, Same Global Pathway, and Unrelated module blocks. These classes are arranged in a hierarchy of decreasing functional association. 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 1,018,353 pairs of module blocks. Most Random Forest predictions were assigned to the correct class or the adjacent class (Fig. S2), 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.

EvoWeaver is based on the premise that a comprehensive view of coevolution is preferable to any single source of coevolutionary signal. Along these lines, all 12 predictors contributed substantially to the ensemble classifier's accuracy (Fig. 3b). The three top predictors (G/L Correlation, RP ContextTree, and Gene Distance) were also the top predictors in each of the three highest performing categories in the Modules benchmark (Fig. 2). We attributed this observation to the fact that distinct categories of coevolution were generally more weakly correlated with each other (Fig. 2), suggesting they provide complementary information.

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 parameter-free label propagation to detect communities within this network[53]. A randomly selected community is shown in Figure 3c-d, which included all module blocks involved in the prodigiosin biosynthesis pathway. EvoWeaver correctly identified all but two Direct Connections within the pathway and properly distinguished the two modules within the pathway. However, EvoWeaver incorrectly classified some Same Module pairs as Direct Connection, and predicted an element of the actinorhodin biosynthetic pathway (*actIV2,4*) to be involved in this pathway. This was likely a spurious connection due to many *Streptomyces* species producing both actinorhodin and undecylprodigiosin. This result suggests EvoWeaver's predictions can be used to hypothesize biochemical pathways, although EvoWeaver's predictions do not provide directionality to biochemical steps.

## EvoWeaver outperforms 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[49] is to predict genes belonging to the same non-global pathway in KEGG, which corresponds to EvoWeaver's Direct Connection, Same Module, and Same Pathway classifications. STRING's Total Score is a composite of seven evidence streams[16]. We applied STRING's formula for Total Score to quantify the marginal benefit of each evidence stream. External data, including mining the literature for cooccurrence of terms (Text Mining) and knowledge bases such as KEGG (Databases), provided the majority of STRING's predictive performance (Fig. 4a). EvoWeaver outperformed STRING at its stated goal of predicting pairs of gene groups sharing a functional pathway in KEGG using purely coevolutionary signal without relying on KEGG itself (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]. As expected, STRING's coevolutionary evidence streams (Cooccurrence, Gene Neighborhood) were correlated with comparable signals derived by EvoWeaver (Fig. 4b).

## EvoWeaver can inform novel hypotheses

EvoWeaver's primary purpose is to serve as a generator for novel hypotheses about functional associations. As a proof of concept, we investigated the top 15 misclassifications wherein a gene pair was assigned to Direct Connection or Same Module with high confidence when it ostensibly belonged to Same Global Pathway or Unrelated in KEGG (Supplemental Data). While some putative mispredictions had no clear evidence for or against a functional relationship in the literature, several were actually correct predictions between clearly related gene groups that have yet to be connected in the same KEGG module. Several purported mispredictions were for genes encoding proteins involved in closely linked plant biochemical pathways, such as gibberellin and abscisic acid biosynthesis, which are both known to regulate plant dormancy and germination[54]. Other alleged mispredictions were for gene pairs implicated in the same diseases, although there was insufficient experimental evidence to validate their functional association. The existence of quasi-mispredictions implies EvoWeaver can be used to identify errors or voids in our current understanding of molecular biology.

As a case study, we examined EvoWeaver's top misprediction, which was between human genes *B3GNT5* and *ST6GAL1*. These genes belong to the "Glycosphingolipid biosynthesis – lacto and neolacto series" and "N-glycan biosynthesis" pathways, respectively. Despite their connection being absent from the KEGG or STRING databases (Fig. 5a), *B3GNT5* was experimentally shown to directly promote the expression of *ST6GAL1* in ovarian cancer cell lines[55]. EvoWeaver predicted this pair to be Direct Connection with probability 0.72 or Same Module with probability 0.27 (Fig. 5b). This prediction was supported by weak phylogenetic profiling evidence because of the high conservation of both genes (Fig. 5c), but there was strong evidence for gene organization due to conservation in gene proximity 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). This proof of concept demonstrates that EvoWeaver can be used to generate reasonable hypotheses about functional relationships.

## DISCUSSION

EvoWeaver represents a marked advancement in employing coevolutionary principles to the discovery of functional associations. In this work, we showed that EvoWeaver can capitalize on multiple sources of coevolutionary signal to outcompete individual algorithms at identifying relationships between gene groups. EvoWeaver's accuracy permitted us to construct a multi-level 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. EvoWeaver's predictive performance was higher than STRING's for the same objective without any dependence on external data. Moreover, we demonstrated how EvoWeaver's predictions can be leveraged to infer novel 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 1,545 gene groups from 8,564 genomes across the tree of life. To our knowledge, this is the largest coevolutionary analysis to date, exceeding the 2,167 genomes analyzed in previous work[12,25]. Unlike popular prior approaches, such as

ContextTree or SVD-phy[34,56], EvoWeaver's pairwise comparisons are independent and can be easily distributed across a cluster of computers. Second, EvoWeaver's predictions are higher accuracy because they incorporate multiple sources of coevolutionary signal, and each component algorithm incorporates statistical testing that mitigates spurious signals. 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 analyses[25].

Coevolution differs from protein-protein interactions in that it does not require any physical interaction. There exist many prior approaches to predicting protein-protein interactions, along with databases of known interactors[45,46,57,58]. Benchmarking functional association algorithms presents its own challenges, as proteins that do not physically interact may nevertheless be functionally associated. This renders common benchmarks for protein-protein interactions insufficient for benchmarking coevolutionary algorithms[58-60]. 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.

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.

ONLINE METHODS

## Construction of Benchmark Datasets

The goal of the Complexes benchmark is to judge algorithms' ability to discern genes encoding proteins involved a complex versus genes encoding unrelated proteins. To this end, we identified all orthology groups belonging to a complex in KEGG[61], 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 that are not in the same pathway (unrelated pairs) and 867 pairs participating as required or optional components of the same complex. 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 algorithms' 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 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. 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 1,545 blocks with 1,948 positive pairs. An equivalent number of negative pairs were sampled in the same manner as the Complexes benchmark.

Having constructed two binary benchmarks, we sought 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 category than their sum of probabilities across all other categories. The confusion matrix at 0% confidence is shown in Figure S2. To look for novel connections (Fig. 5), we examined pairs belonging to Unrelated and Same Global Pathway groups that EvoWeaver predicted as being Direct Connection or Same Module.

## 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 species 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[62] to trim paralogs, align sequences, and construct neighbor joining gene trees. In total, there were 8,564 genomes with at least one gene present in the benchmarks. Species trees were estimated using the ASTRID algorithm[63]. To find each gene's index within its genome, we downloaded complete genomes and coding sequences from NCBI following the reference links provided in KEGG. Of the 8,564 genomes present in the benchmarks, 7,535 had genome sequences available. Coding sequences were matched to locations on the genome with the *Biostrings* (v2.68.1) package in R[64,65] (v4.3.0).

## 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 values imply greater coevolutionary signal. Some algorithms can provide scores between -1 and 1, in which case 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 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 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 score for all gene organization algorithms is zero. The 12 algorithms implemented fall into four categories: phylogenetic profiling, phylogenetic structure, gene organization, and sequence-level methods (Fig. 1a). Of these, four algorithms are completely novel (G/L Distance, RP ContextTree, RP MirrorTree, and Gene Distance), three are novel applications of existing algorithms (TreeDistance, Moran's I, Gene Vector), and the remaining five are refinements on existing algorithms.

### *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 absence[66]. The first P/A approach used Hamming distances on binary profiles as a score[67]. Later, Jaccard index and mutual information were applied to score P/A profiles[12,68]. Subsequent work transformed P/A profiles into ancestral gain/loss (G/L) events and scored the correlation between events[30]. This transformation reduces redundancy for sets of organisms with low rates of gene gain and loss[28,30].

EvoWeaver includes four phylogenetic profiling algorithms (Fig. 1b). The first algorithm, P/A MI, calculates bidirectional mutual information of binary P/A profiles using a recently introduced weighting scheme[69]. The second algorithm, P/A Jaccard, uses the Jaccard index of P/A profiles. The third algorithm, G/L Correlation, applies Fitch Parsimony[70] to infer ancestral states on the species tree from P/A profiles. These G/L profiles include three states: -1 for gene loss, 0 for no change, and +1 for gene gain. The G/L Correlation score is defined as Pearson's correlation coefficient of the ternary G/L profiles.

G/L Correlation fails to account for compensatory changes that do not occur on the same branch of the species tree, which are common in sequence evolution[71]. The fourth algorithm, G/L Distance, quantifies 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 $we^{-d(v_1,v_2)}$, 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 species 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 the expected value of distance between two points randomly placed on a line segment (i.e., 1/3rd the branch length). 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 MI, P/A Jaccard, and G/L Correlation are calculated using Fisher's Exact Test (two-way for P/A and three-way for G/L), and a p-value for G/L Distance is calculated using empirical values from permutation testing with 100 replicates.

### *Phylogenetic Structure*

Gene tree structural comparisons were pioneered by MirrorTree[32], 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 analysis[72]. These extensions, often referred to as ContextTree or ContextMirror, use different approaches to remove shared signal represented by the species tree[31,56,73]. More recently, SVD-phy was introduced as an alternative approach using BLAST to measure distance between sequences[34,74]. SVD-phy uses singular value decomposition 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 is a surjective mapping that approximately preserves distances between vectors[75]. While traditional random projection uses a large matrix of random values, this requirement can be circumvented by generating values of the matrix on the fly 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 random projection to patristic distances and scores pairs of vectors using Spearman's correlation coefficient. The RP ContextTree algorithm also subtracts the randomly projected species tree from each vector prior to scoring. RP ContextTree's final scores are multiplied by the Jaccard index of overlap in organism membership to correct for spurious correlations caused by minimally overlapping sets. Statistical significance for both RP ContextTree and RP MirrorTree are 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 metrics[35] were implemented as measures of functional similarity, all of which were highly correlated in their tree distances. By default, EvoWeaver's TreeDistance predictor uses normalized Robinson-Foulds Distance due to its low memory requirement and closed form solution for significance[76]. The score for each pair of genes was defined as $1 - TD(T_1, T_2)$, where *TD* is the tree distance and $T_1$ and $T_2$ are gene trees.

*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 antiSMASH[39], 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 association[16]. 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 are possibly transcribed together, 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, Transcription MI, examines the relative transcriptional direction of gene pairs. Conservation of transcriptional direction has been validated in prior work to be indicative of shared function[77]. The score for Transcription MI is defined as the bidirectional mutual information[69] between transcriptional directions of gene pairs, 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, 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, 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 segment[78].

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 distances[79] and values in the form of gene distances ($d$). Moran's *I* distinguishes between genes that are colocated purely due to low evolutionary divergence versus genes that have maintained a consistent relative distance in spite of evolutionary divergence. Statistical significance is calculated using the closed form solution to the expected value and variance of Moran's *I* (ref. [80]).

*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 analysis[46], 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[81,82]. While some previously developed approaches improved scaling[83,84], 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. The first of these, Gene Vector, uses the gene sequence natural vector approach, developed to predict protein-protein interactions[48]. 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 high 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 group[47]. For every 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[62]. 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 correction[85]. 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 each of the aforementioned coevolutionary algorithms into a final prediction using an ensemble machine learning method. All 12 algorithms were used as features for ensemble prediction (Fig. 2). For ensemble methods, we tested logistic regression, random forest, and neural network models in R[65]. Logistic regression was performed with the *glm* function, random forests using default parameters in the *randomForest* package[86] (v4.7-1.1), and neural networks using the *neuralnet* package (v1.44.2). 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 5-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 5-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 1,948 members. This meant that each class in the train set for each fold had 1,558 members (i.e., 80%). Testing was done on the complete (unbalanced) test set, which comprised 203,669 - 203,674 members (i.e., ~20%) per fold. Each pair was in exactly one test set. 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 estimates[87].

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 nodes in KEGG are directly connected to two neighbors, and other nodes in the same module are less important than direct connections. We then used label propagation implemented in the *igraph* package[88] (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. We further validated our results by reevaluating our ensemble classifier using 10-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 5-fold cross validation with pair holdouts. The results of this classification were virtually identical to prior results (Fig. S3), implying that EvoWeaver is not heavily relying on features of the individual gene groups themselves when making predictions. This is consistent with the notion that most gene groups have few direct connections and thus learning to distinguish highly connected gene groups gives little predictive power.

## 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, 6,849 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 (≥ 50%) KEGG KO group. A total of 6,311 COGs had a majority KO group, and 4,481 (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"[49]. 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 3,446 pairs of COGs in the matched dataset belonged to the Same Pathway, Same Module, or Direct Connection categories in KEGG. An equivalent number of negatives were randomly sampled from the remaining pairs in a similar manner to the Modules benchmark. 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 Figure 4a was obtained by sequentially adding the evidence stream with the lowest impact on AUROC to the Total Score calculation.

## Experimental Details

All analysis and plotting was performed with R (v4.3.0). Area under receiver operator characteristic curves and precision-recall curves were calculated with the *auc* function in the *DescTools* package (v0.99.49) for R. Algorithms were implemented in EvoWeaver using R and C programming languages, with user-exposed methods available in R via the *SynExtend* package (v1.16.0). SynExtend is dependent on the *DECIPHER* package (v2.28.0) and is distributed via the Bioconductor software repository[89]. Users can run EvoWeaver by initializing an EvoWeaver object in R with the *EvoWeaver* function, and then using the *predict* function to run component algorithms. Local analyses were performed on a MacBook Pro with M1 Pro CPU and 32GB of RAM. Distributed computing was performed on the Open Science Grid[90]. Phylogenetic tree reconstruction used eight core nodes with 8 - 16GB RAM and 8GB disk space, and pairwise coevolutionary score calculations with EvoWeaver used single core nodes with 2 - 4GB

RAM and 2 - 4GB disk space. Computers matching these node specifications varied based on availability and Open Science Grid scheduling. Scripts for reproducing all analyses are available on GitHub (https://github.com/WrightLabScience/EvoWeaver-ExampleCode). Datasets are available from Zenodo (DOI: 10.5281/zenodo.10266140).

## 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, 3, and 5 were created using BioRender.

## REFERENCES

1 Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods* (2022). https://doi.org:10.1038/s41592-022-01454-x

2 Kustatscher, G. *et al.* An open invitation to the Understudied Proteins Initiative. *Nature Biotechnology* (2022). https://doi.org:10.1038/s41587-022-01316-z

3 Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, 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). https://doi.org:10.1002/pmic.201800093

4 Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Scientific Reports* **8** (2018). https://doi.org:10.1038/s41598-018-19333-x

5 Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biology* **20** (2019). https://doi.org:10.1186/s13059-019-1715-2

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

7 Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology* **16**, e2006643 (2018). https://doi.org:10.1371/journal.pbio.2006643

8 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 Computational Biology* **9**, e1003063 (2013). https://doi.org:10.1371/journal.pcbi.1003063

9 Gillis, J. & Pavlidis, P. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. *PLoS ONE* **6**, e17258 (2011). https://doi.org:10.1371/journal.pone.0017258

10 Aravind, L. Guilt by association: contextual information in genome analysis. *Genome Research* **10**, 1074-1077 (2000).

11 Codoñer, F. M. & Fares, M. A. Why should we care about molecular coevolution? *Evol Bioinform Online* **4**, 29-38 (2008).

12 Moi, D., Kilchoer, L., Aguilar, P. S. & Dessimoz, C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLOS Computational Biology* **16**, e1007553 (2020). https://doi.org:10.1371/journal.pcbi.1007553

13 Kann, M. G., Shoemaker, B. A., Panchenko, A. R. & Przytycka, T. M. Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree. *Journal of Molecular Biology* **385**, 91-98 (2009). https://doi.org:10.1016/j.jmb.2008.09.078

14 Umemura, 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). https://doi.org:10.3389/fmicb.2015.00371

15 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). https://doi.org:10.1371/journal.pone.0149166

16 Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605-D612 (2021). https://doi.org:10.1093/nar/gkaa1074

17 Stupp, D., Sharon, E., Bloch, I., Zitnik, M., Zuk, O. & Tabach, Y. Co-evolution based machine-learning for predicting functional interactions between human genes. *Nature Communications* **12** (2021). https://doi.org:10.1038/s41467-021-26792-w

18 Tabach, Y. *et al.* Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Molecular systems biology* **9**, 692 (2013).

19 Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493**, 694-698 (2013).

20 Sherill-Rofe, D. *et al.* Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Research* **29**, 439-448 (2019). https://doi.org:10.1101/gr.241414.118

21 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 (Basel)* **10** (2021). https://doi.org:10.3390/antibiotics10121533

22 Ding, D. *et al.* Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nature Ecology & Evolution* (2022). https://doi.org:10.1038/s41559-022-01688-0

23 Fongang, B., Zhu, Y., Wagner, E. J., Kudlicki, A. & Rowicka, M. *Co-evolutionary analysis accurately predicts details of interactions between the Integrator complex subunits* (Cold Spring Harbor Laboratory, 2019).

24 Ramani, A. K. & Marcotte, E. M. Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity. *Journal of Molecular Biology* **327**, 273-284 (2003). https://doi.org:https://doi.org/10.1016/S0022-2836(03)00114-1

25 Moi, D. & Dessimoz, C. Phylogenetic profiling in eukaryotes comes of age. *Proceedings of the National Academy of Sciences* **120** (2023). https://doi.org:10.1073/pnas.2305013120

26 Fukunaga, T. & Iwasaki, W. Inverse Potts model improves accuracy of phylogenetic profiling. *Bioinformatics* **38**, 1794-1800 (2022). https://doi.org:10.1093/bioinformatics/btac034

27 Cheng, Y. & Perocchi, F. ProtPhylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic Acids Research* **43**, W160-W168 (2015). https://doi.org:10.1093/nar/gkv455

28 Škunca, N. & Dessimoz, C. Phylogenetic profiling: how much input data is enough? *PloS one* **10**, e0114701 (2015).

29 Mering, C. V. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **31**, 258-261 (2003). https://doi.org:10.1093/nar/gkg034

30 Dembech, E. *et al.* Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proceedings of the National Academy of Sciences* **120** (2023). https://doi.org:10.1073/pnas.2218329120

31 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). https://doi.org:10.1016/j.jmb.2005.07.005

32 Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering, Design and Selection* **14**, 609-614 (2001). https://doi.org:10.1093/protein/14.9.609

33 Clark, G. W., Dar, V.-U.-N., Bezginov, A., Yang, J. M., Charlebois, R. L. & Tillier, E. R. M. 237-256 (Humana Press, 2011).

34 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).

35 Smith, M. R. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**, 5007-5013 (2020). https://doi.org:10.1093/bioinformatics/btaa614

36 Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology* **16**, 731-744 (2018). https://doi.org:10.1038/s41579-018-0075-3

37 Periwal, V. & Scaria, V. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics* **31**, 1-9 (2014). https://doi.org:10.1093/bioinformatics/btu600

38 Rocha, E. P. The organization of the bacterial genome. *Annual review of genetics* **42**, 211-233 (2008).

39 Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29-W35 (2021). https://doi.org:10.1093/nar/gkab335

40 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 research* **45**, W55-W63 (2017). https://doi.org:10.1093/nar/gkx305

41 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). https://doi.org:10.1093/nar/gkaa812

42 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).

43 Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins: Structure, Function, and Bioinformatics* **76**, 911-929 (2009). https://doi.org:10.1002/prot.22398

44 Morcos, F., Hwa, T., Onuchic, J. N. & Weigt, M. Direct coupling analysis for protein contact prediction. *Methods Mol Biol* **1137**, 55-70 (2014). https://doi.org:10.1007/978-1-4939-0366-5_5

45 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293-E1301 (2011).

46 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67-72 (2009). https://doi.org:10.1073/pnas.0805923106

47 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). https://doi.org:10.1093/bioinformatics/bti671

48 Zhao, N., Zhuo, M., Tian, K. & Gong, X. Protein–protein interaction and non-interaction predictions using gene sequence natural vector. *Communications Biology* **5** (2022). https://doi.org:10.1038/s42003-022-03617-0

49 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39**, D561-D568 (2011). https://doi.org:10.1093/nar/gkq973

50 Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000). https://doi.org:10.1093/nar/28.1.27

51 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). https://doi.org:10.1093/nar/gky962

52 Launay, G., Ceres, N. & Martin, J. Non-interacting proteins may resemble interacting proteins: prevalence and implications. *Scientific Reports* **7**, 40419 (2017). https://doi.org:10.1038/srep40419

53 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). https://doi.org:10.1103/PhysRevE.76.036106

54 Tuan, P. A., Kumar, R., Rehal, P. K., Toora, P. K. & Ayele, B. T. Molecular mechanisms underlying abscisic acid/gibberellin balance in the control of seed dormancy and germination in cereals. *Frontiers in Plant Science* **9**, 668 (2018).

55 Alam, S. *et al.* Altered (neo-) lacto series glycolipid biosynthesis impairs α2-6 sialylation on N-glycoproteins in ovarian cancer cells. *Scientific Reports* **7**, 45367 (2017). https://doi.org:10.1038/srep45367

56 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). https://doi.org:10.1093/bioinformatics/btl419

57 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020). https://doi.org:10.1038/s41586-020-2188-x

58 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 Research* **42**, D396-D400 (2014). https://doi.org:10.1093/nar/gkt1079

59 Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **46**, D649-D655 (2018).

60 Oughtred, R. *et al.* TheBioGRIDdatabase: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187-200 (2021). https://doi.org:10.1002/pro.3978

61 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457-D462 (2015). https://doi.org:10.1093/nar/gkv1070

62 Wright, E. S. Using DECIPHER v2. 0 to analyze big biological sequence data in R. *R Journal* **8** (2016).

63 Vachaspati, P. & Warnow, T. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics* **16**, S3 (2015). https://doi.org:10.1186/1471-2164-16-s10-s3

64 Biostrings: Efficient manipulation of biological strings. v. 2.68.0 (Bioconductor, 2023).

65 Team, R. C. R: A language and environment for statistical computing. (2013).

66 Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Liò, P. & Fani, R. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* **9**, 551 (2008). https://doi.org:10.1186/1471-2105-9-551

67 Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**, 4285-4288 (1999). https://doi.org:10.1073/pnas.96.8.4285

68 Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology* **21**, 1055-1062 (2003). https://doi.org:10.1038/nbt861

69 Beckley, A. M. & Wright, E. S. Identification of antibiotic pairs that evade concurrent resistance via a retrospective analysis of antimicrobial susceptibility test results. *The Lancet Microbe* **2**, e545-e554 (2021).

70 Fitch, W. M. On the problem of discovering the most parsimonious tree. *The American Naturalist* **111**, 223-257 (1977).

71 Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLoS Genetics* **7**, e1001301 (2011). https://doi.org:10.1371/journal.pgen.1001301

72 Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences* **105**, 934-939 (2008). https://doi.org:10.1073/pnas.0709671105

73 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). https://doi.org:10.1093/bioinformatics/bti564

74 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). https://doi.org:10.1016/s0022-2836(05)80360-2

75 Achlioptas, D. in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* 274-281.

76 Steel, M. A. & Penny, D. Distributions of Tree Comparison Metrics—Some New Results. *Systematic Biology* **42**, 126-141 (1993). https://doi.org:10.1093/sysbio/42.2.126

77 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. *Nature biotechnology* **22**, 911-917 (2004).

78 Philip, J. The probability distribution of the distance between two random points in a box. *KTH Mathematics, Royal Institute of Technology* (2007).

79 Gittleman, J. L. & Kot, M. Adaptation: Statistics and a Null Model for Estimating Phylogenetic Effects. *Systematic Biology* **39**, 227-241 (1990). https://doi.org:10.2307/2992183

80 Cliff, A. & Ord, J. Spatial Processes (London: Pion). *Google Scholar* (1981).

81 Pesaranghader, A., Matwin, S., Sokolova, M., Grenier, J.-C., Beiko, R. G. & Hussin, J. deepSimDEF: deep neural embeddings of gene products and Gene Ontology terms for functional analysis of genes. *Bioinformatics* (2022). https://doi.org:10.1093/bioinformatics/btac304

82 Soleymani, F., Paquet, E., Viktor, H. L., Michalowski, W. & Spinello, D. ProtInteract: a Deep Learning Framework for Predicting Protein—Protein Interactions. *Computational and Structural Biotechnology Journal* (2023). https://doi.org:https://doi.org/10.1016/j.csbj.2023.01.028

83 Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics* **276**, 341-356 (2014). https://doi.org:10.1016/j.jcp.2014.07.024

84 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). https://doi.org:10.1093/bioinformatics/btr638

85 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).

86 Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).

87 Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007). https://doi.org:10.1186/1471-2105-8-25

88 Csárdi, G. & Nepusz, T.

89 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, 1-16 (2004).

90 OSG. (ed OSG) (2015).

# FIGURES

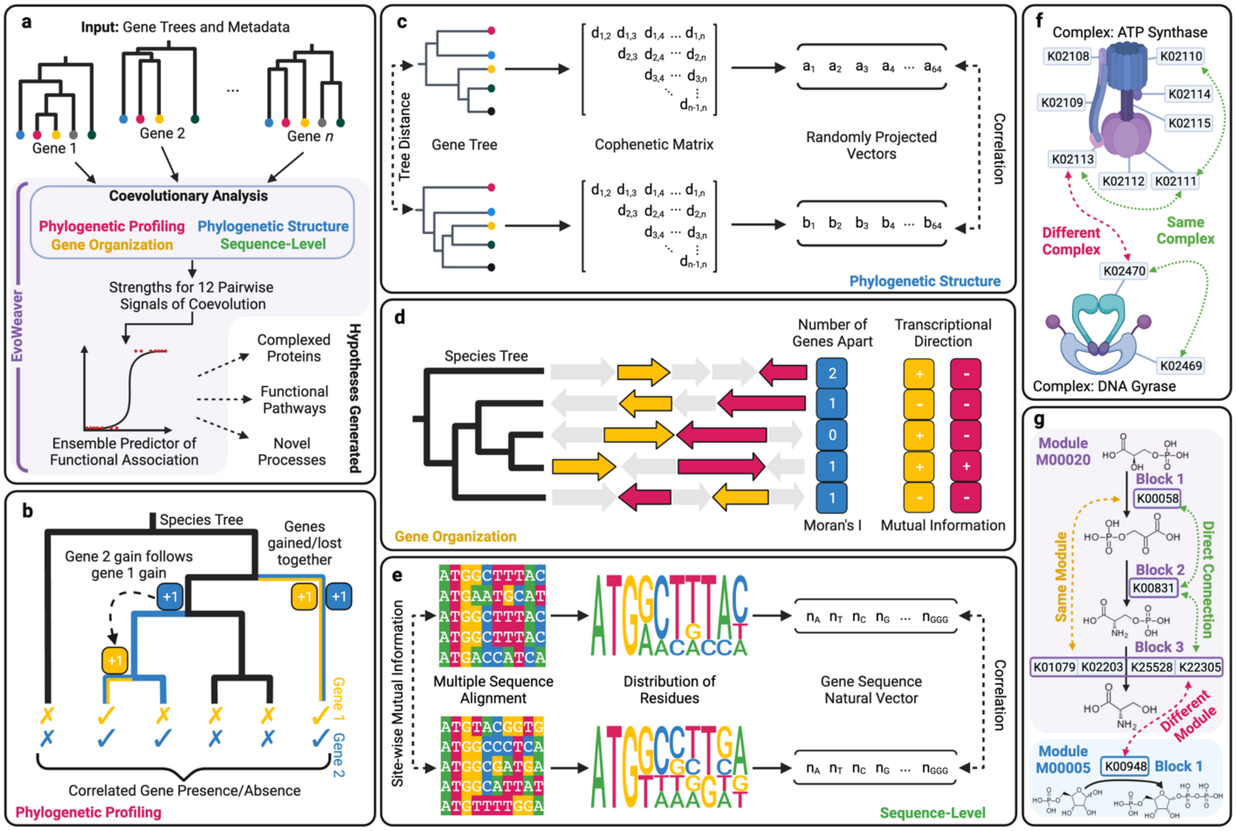
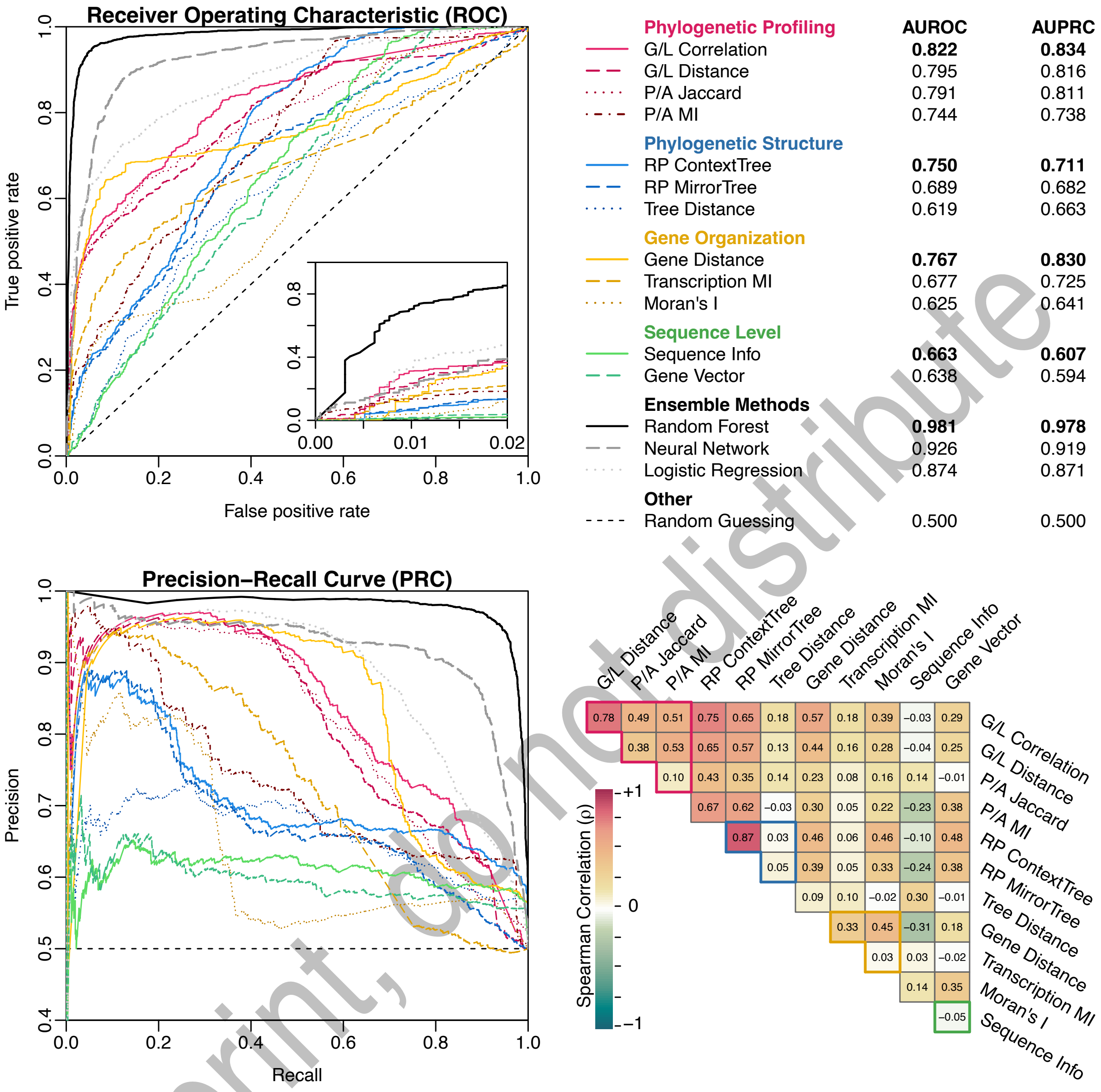


**Figure 1: Overview of the EvoWeaver algorithm and benchmarking. (a)** Phylogenetic trees from gene orthologs serve as the primary input to EvoWeaver. Four categories of coevolutionary signal are quantified for each pair of genes. These signals are combined in an ensemble classifier to predict functional relationships between gene pairs. **(b)** Functional associations often result in correlated gain/loss patterns on a phylogenetic tree of the species. 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 transcriptional direction and the distance between gene pairs. **(e)** Functional associations sometimes cause concerted changes in sequences that are interrogated by EvoWeaver. **(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.



| | AUROC | AUPRC |
|---|---|---|
| **Phylogenetic Profiling** | | |
| G/L Correlation | **0.822** | **0.834** |
| G/L Distance | 0.795 | 0.816 |
| P/A Jaccard | 0.791 | 0.811 |
| P/A MI | 0.744 | 0.738 |
| **Phylogenetic Structure** | | |
| RP ContextTree | **0.750** | **0.711** |
| RP MirrorTree | 0.689 | 0.682 |
| Tree Distance | 0.619 | 0.663 |
| **Gene Organization** | | |
| Gene Distance | **0.767** | **0.830** |
| Transcription MI | 0.677 | 0.725 |
| Moran's I | 0.625 | 0.641 |
| **Sequence Level** | | |
| Sequence Info | **0.663** | **0.607** |
| Gene Vector | 0.638 | 0.594 |
| **Ensemble Methods** | | |
| Random Forest | **0.981** | **0.978** |
| Neural Network | 0.926 | 0.919 |
| Logistic Regression | 0.874 | 0.871 |
| **Other** | | |
| Random Guessing | 0.500 | 0.500 |

| | G/L Distance | P/A Jaccard | P/A MI | RP ContextTree | RP MirrorTree | Tree Distance | Gene Distance | Transcription MI | Moran's I | Sequence Info | Gene Vector |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G/L Correlation | 0.78 | 0.49 | 0.51 | 0.75 | 0.65 | 0.18 | 0.57 | 0.18 | 0.39 | −0.03 | 0.29 |
| G/L Distance | | 0.38 | 0.53 | 0.65 | 0.57 | 0.13 | 0.44 | 0.16 | 0.28 | −0.04 | 0.25 |
| P/A Jaccard | | | 0.10 | 0.43 | 0.35 | 0.14 | 0.23 | 0.08 | 0.16 | 0.14 | −0.01 |
| P/A MI | | | | 0.67 | 0.62 | −0.03 | 0.30 | 0.05 | 0.22 | −0.23 | 0.38 |
| RP ContextTree | | | | | 0.87 | 0.03 | 0.46 | 0.06 | 0.46 | −0.10 | 0.48 |
| RP MirrorTree | | | | | | 0.05 | 0.39 | 0.05 | 0.33 | −0.24 | 0.38 |
| Tree Distance | | | | | | | 0.09 | 0.10 | −0.02 | 0.30 | −0.01 |
| Gene Distance | | | | | | | | 0.33 | 0.45 | −0.31 | 0.18 |
| Transcription MI | | | | | | | | | 0.03 | 0.03 | −0.02 |
| Moran's I | | | | | | | | | | 0.14 | 0.35 |
| Sequence Info | | | | | | | | | | | −0.05 |

**Figure 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., 1,948 positives) from proteins in distinct modules (i.e., 1,948 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 quasi-orthogonal coevolutionary signals.

Spearman's correlation from positive and negative sets is averaged to correct for artificial correlation among high performing algorithms.

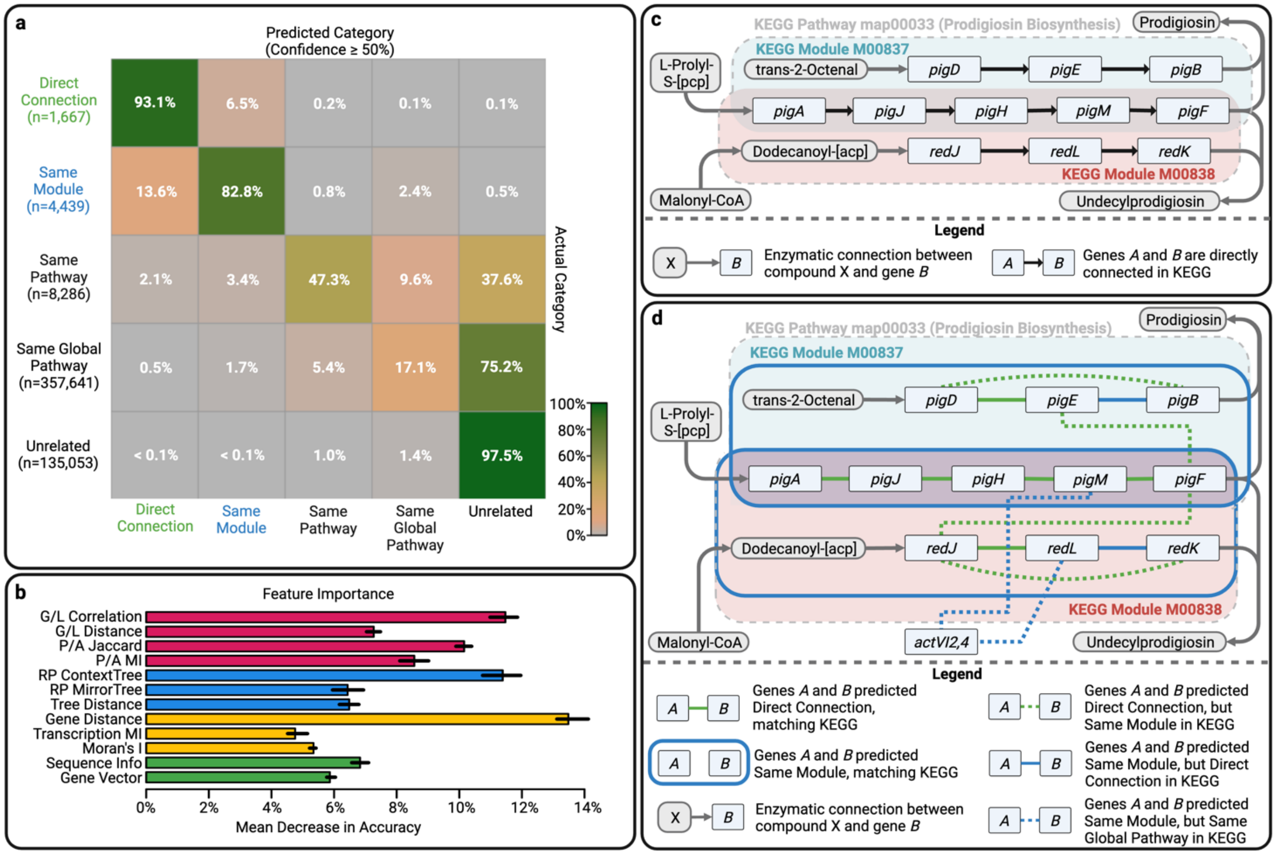
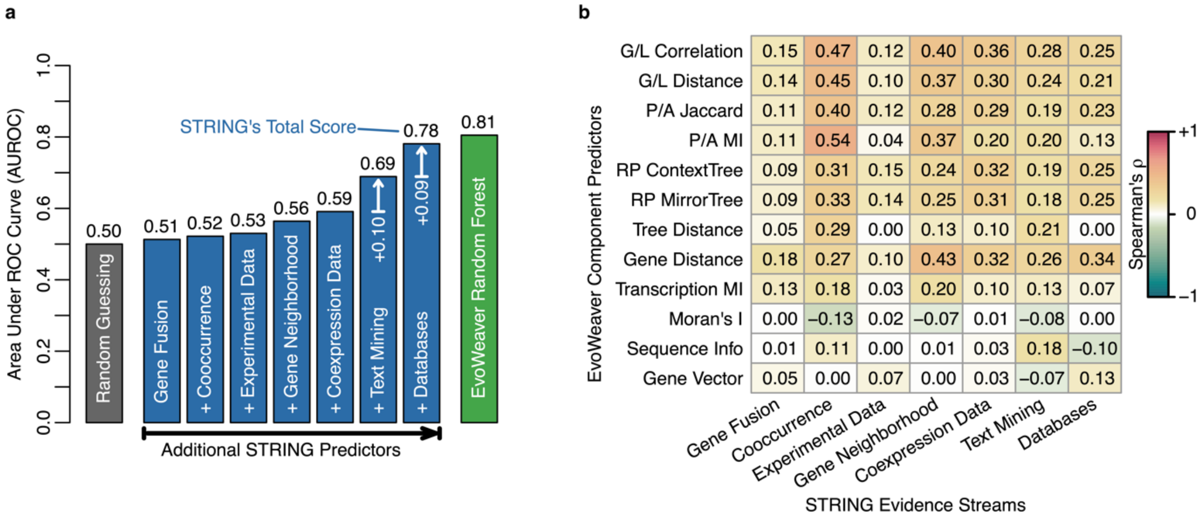


**Figure 3: EvoWeaver is sufficiently accurate to hierarchically classify functional associations. (a)** The confusion matrix of five level classifications indicates that EvoWeaver's ensemble predictions (i.e., random forest) rarely confuse proteins within the same module with those from different modules. Values represent the percent of each actual category classified to each predicted category. **(b)** The best performing algorithm from each category on the Modules benchmark also was 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. Error bars denote the range of importances across each train/test fold. **(c-d)** Hierarchical classifications permit the partial inference of biochemical pathways directly from sequence information without any external biological knowledge. EvoWeaver's ensemble predictions for genes involved in prodigiosin biosynthesis generally match experimentally verified connections in KEGG. Panel (c) displays the original pathway from KEGG, and panel (d) overlays EvoWeaver's hierarchical classifications. Note that *pigA*, *pigJ*, *pigH*, *pigM*, and *pigF* belong to both modules.



| | Gene Fusion | Cooccurrence | Experimental Data | Gene Neighborhood | Coexpression Data | Text Mining | Databases |
|---|---|---|---|---|---|---|---|
| G/L Correlation | 0.15 | 0.47 | 0.12 | 0.40 | 0.36 | 0.28 | 0.25 |
| G/L Distance | 0.14 | 0.45 | 0.10 | 0.37 | 0.30 | 0.24 | 0.21 |
| P/A Jaccard | 0.11 | 0.40 | 0.12 | 0.28 | 0.29 | 0.19 | 0.23 |
| P/A MI | 0.11 | 0.54 | 0.04 | 0.37 | 0.20 | 0.20 | 0.13 |
| RP ContextTree | 0.09 | 0.31 | 0.15 | 0.24 | 0.32 | 0.19 | 0.25 |
| RP MirrorTree | 0.09 | 0.33 | 0.14 | 0.25 | 0.31 | 0.18 | 0.25 |
| Tree Distance | 0.05 | 0.29 | 0.00 | 0.13 | 0.10 | 0.21 | 0.00 |
| Gene Distance | 0.18 | 0.27 | 0.10 | 0.43 | 0.32 | 0.26 | 0.34 |
| Transcription MI | 0.13 | 0.18 | 0.03 | 0.20 | 0.10 | 0.13 | 0.07 |
| Moran's I | 0.00 | −0.13 | 0.02 | −0.07 | 0.01 | −0.08 | 0.00 |
| Sequence Info | 0.01 | 0.11 | 0.00 | 0.01 | 0.03 | 0.18 | −0.10 |
| Gene Vector | 0.05 | 0.00 | 0.07 | 0.00 | 0.03 | −0.07 | 0.13 |

**Figure 4: EvoWeaver outperforms STRING without reliance on external data. (a)** Predictive accuracy was compared on 6,892 pairs of gene groups that overlapped between STRING and the Modules benchmark. Area under the ROC curve (AUROC) is shown for discerning between pairs sharing the same non-global pathway in KEGG (i.e., positives) versus pairs in different non-global pathways (i.e., negatives). STRING's predictions are a composite of seven evidence streams, including three coevolutionary evidence streams (i.e., Gene Fusion, Cooccurrence, Gene Neighborhood). 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 evidence streams. Despite STRING's predictions incorporating KEGG into its Databases evidence stream, EvoWeaver's Random Forest predictions outperformed STRING's Total Score while only using sequence information. **(b)** Unsurprisingly, some of EvoWeaver's component predictors were modestly correlated with STRING's evidence streams. For example, STRING's Cooccurrence score, based on SVD-phy, is correlated with EvoWeaver's phylogenetic profiling methods, and STRING's Gene Neighborhood score is correlated with EvoWeaver's Gene Distance predictor. Spearman's correlation is calculated in the same manner as in Figure 2.

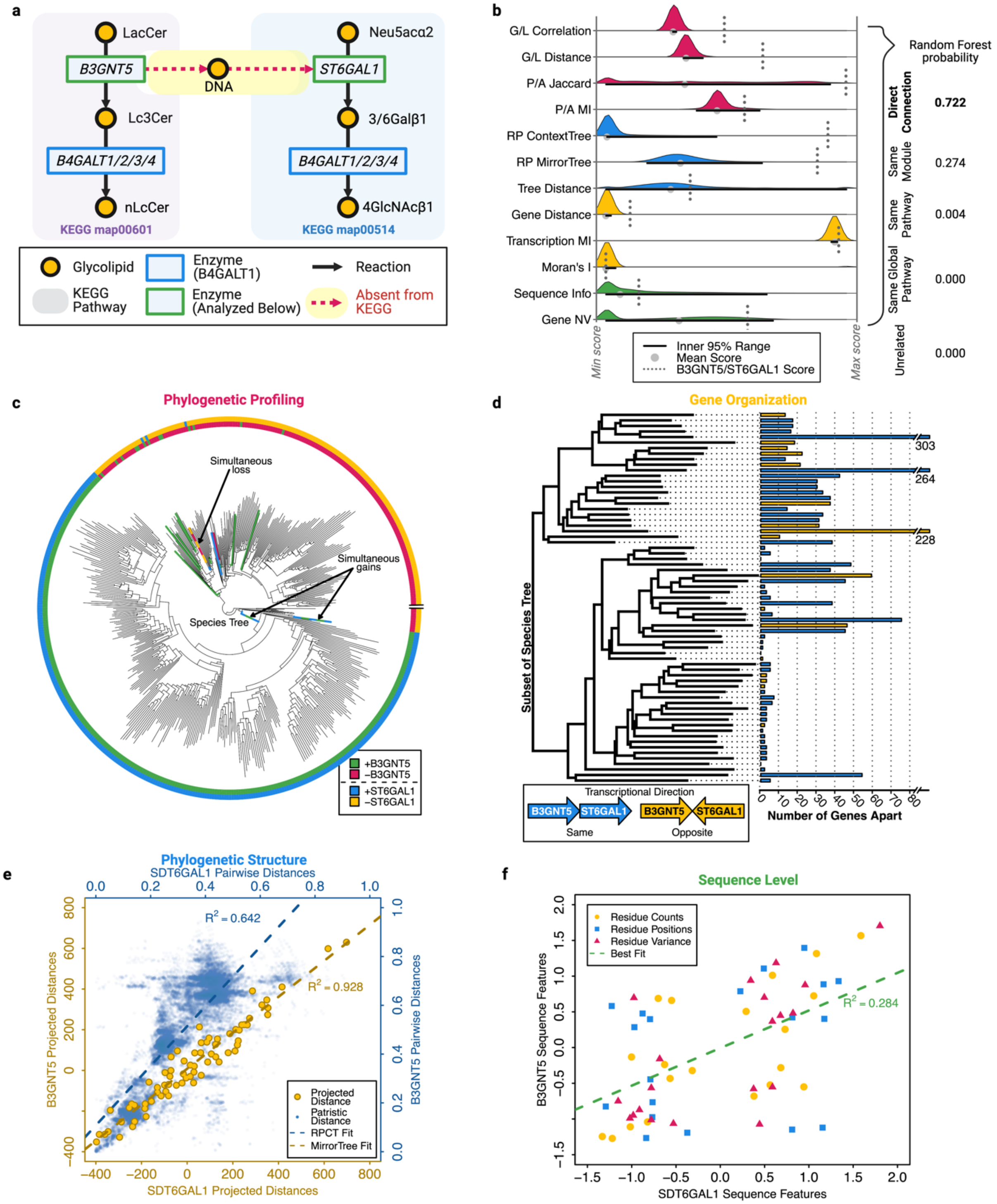
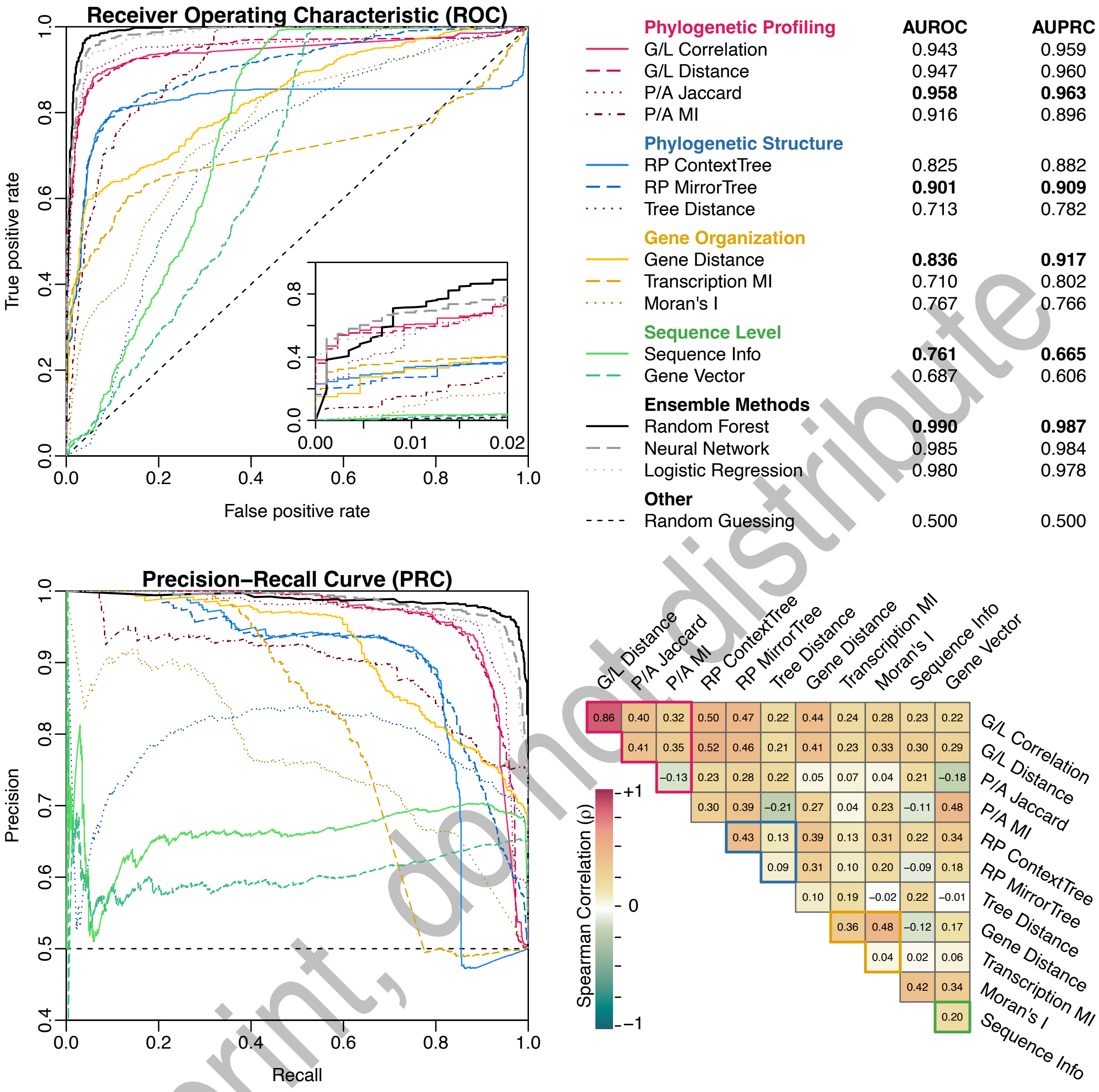


**Figure 5: EvoWeaver's ensemble predictions can generate high fidelity biological hypotheses. (a)** The protein product of *B3GNT5* promotes the expression of *ST6GAL1*[55], although this connection is missing in KEGG and STRING. **(b)** EvoWeaver's component and ensemble predictions indicate that *B3GNT5* and

*ST6GAL1* are functionally associated, which is supported by experiments in human cell culture[55]. **(c)** Phylogenetic profiling demonstrates a pattern of association between *B3GNT5* and *ST6GAL1*, although it is supported by relatively few gain/loss events on the species tree. **(d)** Organisms with both *B3GNT5* and *ST6GAL1* on the same chromosome display a clear linkage in gene distance and transcriptional direction. **(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.

| Symbol | Meaning | Example | Interpretation | Example Module |
|---|---|---|---|---|
| K12345 | Orthology group #12345 | K05308 | Each code is comprised of "K" followed by a unique 5-digit string. K05308 encodes gene *gnaD* | Any |
| Space | Direct connection | K05308 K18126 | K05308 performs/facilitates a chemical reaction immediately prior to K18126 | M00633 |
| Plus (+) | Complex | K02111+K02112 | K02111 and K02112 belong to the same complex | M00157 |
| Minus (-) | Optional Complex | -K03944 | K03944 is an optional component of the complex | M00143 |
| Parentheses | Optional Components | (K01681,K01682) | Either K01681 or K01682 performs/facilitates this chemical reaction | M00012 |
| Newline | Separate Components | K21428 K21778 K21779 K21787 | K21779 and K21787 are in the same module, but they participate in different stages of the module | M00837 |

**Table S1: Description of KEGG Modules.** Each module in KEGG is specified with a plain text definition composed of orthology groups and symbols specifying relationships.

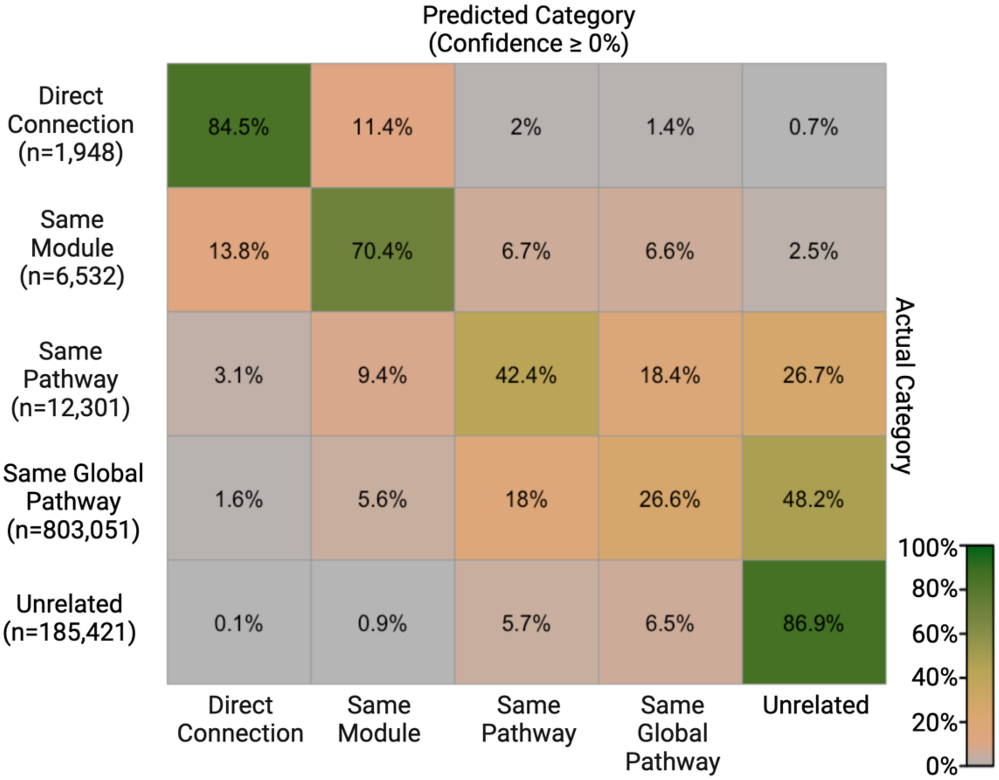


| | Phylogenetic Profiling | AUROC | AUPRC |
|---|---|---|---|
| ── | G/L Correlation | 0.943 | 0.959 |
| – – | G/L Distance | 0.947 | 0.960 |
| ····· | P/A Jaccard | **0.958** | **0.963** |
| –·– | P/A MI | 0.916 | 0.896 |
| | **Phylogenetic Structure** | | |
| ── | RP ContextTree | 0.825 | 0.882 |
| – – | RP MirrorTree | **0.901** | **0.909** |
| ····· | Tree Distance | 0.713 | 0.782 |
| | **Gene Organization** | | |
| ── | Gene Distance | **0.836** | **0.917** |
| – – | Transcription MI | 0.710 | 0.802 |
| ····· | Moran's I | 0.767 | 0.766 |
| | **Sequence Level** | | |
| ── | Sequence Info | **0.761** | **0.665** |
| – – | Gene Vector | 0.687 | 0.606 |
| | **Ensemble Methods** | | |
| ── | Random Forest | **0.990** | **0.987** |
| – – | Neural Network | 0.985 | 0.984 |
| ····· | Logistic Regression | 0.980 | 0.978 |
| | **Other** | | |
| - - - - | Random Guessing | 0.500 | 0.500 |

| | G/L Distance | P/A Jaccard | P/A MI | RP ContextTree | RP MirrorTree | Tree Distance | Gene Distance | Transcription MI | Moran's I | Sequence Info | Gene Vector |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G/L Correlation | 0.86 | 0.40 | 0.32 | 0.50 | 0.47 | 0.22 | 0.44 | 0.24 | 0.28 | 0.23 | 0.22 |
| G/L Distance | | 0.41 | 0.35 | 0.52 | 0.46 | 0.21 | 0.41 | 0.23 | 0.33 | 0.30 | 0.29 |
| P/A Jaccard | | | −0.13 | 0.23 | 0.28 | 0.22 | 0.05 | 0.07 | 0.04 | 0.21 | −0.18 |
| P/A MI | | | | 0.30 | 0.39 | −0.21 | 0.27 | 0.04 | 0.23 | −0.11 | 0.48 |
| RP ContextTree | | | | | 0.43 | 0.13 | 0.39 | 0.13 | 0.31 | 0.22 | 0.34 |
| RP MirrorTree | | | | | | 0.09 | 0.31 | 0.10 | 0.20 | −0.09 | 0.18 |
| Tree Distance | | | | | | | 0.10 | 0.19 | −0.02 | 0.22 | −0.01 |
| Gene Distance | | | | | | | | 0.36 | 0.48 | −0.12 | 0.17 |
| Transcription MI | | | | | | | | | 0.04 | 0.02 | 0.06 |
| Moran's I | | | | | | | | | | 0.42 | 0.34 |
| Sequence Info | | | | | | | | | | | 0.20 |

**Figure S1: EvoWeaver's ensemble predictions outperform individual algorithms on the Complexes benchmark.** Coevolutionary approaches were compared for their ability to discern pairs of KO groups that complex (i.e., 867 positives) from unrelated pairs of KO groups (i.e., 867 negatives). Phylogenetic profiling algorithms tended to outperform other methods, though all categories of analysis showed strong performance. EvoWeaver's ensemble predictions that combine all component sources of coevolutionary signal 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

quasi-orthogonal coevolutionary signals. Spearman's correlation from positive and negative sets is averaged to correct for artificial correlation among high performing algorithms.



| | Direct Connection | Same Module | Same Pathway | Same Global Pathway | Unrelated |
|---|---|---|---|---|---|
| Direct Connection (n=1,948) | 84.5% | 11.4% | 2% | 1.4% | 0.7% |
| Same Module (n=6,532) | 13.8% | 70.4% | 6.7% | 6.6% | 2.5% |
| Same Pathway (n=12,301) | 3.1% | 9.4% | 42.4% | 18.4% | 26.7% |
| Same Global Pathway (n=803,051) | 1.6% | 5.6% | 18% | 26.6% | 48.2% |
| Unrelated (n=185,421) | 0.1% | 0.9% | 5.7% | 6.5% | 86.9% |

**Figure S2: EvoWeaver's hierarchical classifications at any confidence level.** Each gene group pairing was assigned to its highest confidence predicted category without imposing a minimum confidence threshold. Results are analogous to Fig. 3a, except with more pairs predicted in the Same Global Pathway category and slightly higher misclassification rates.

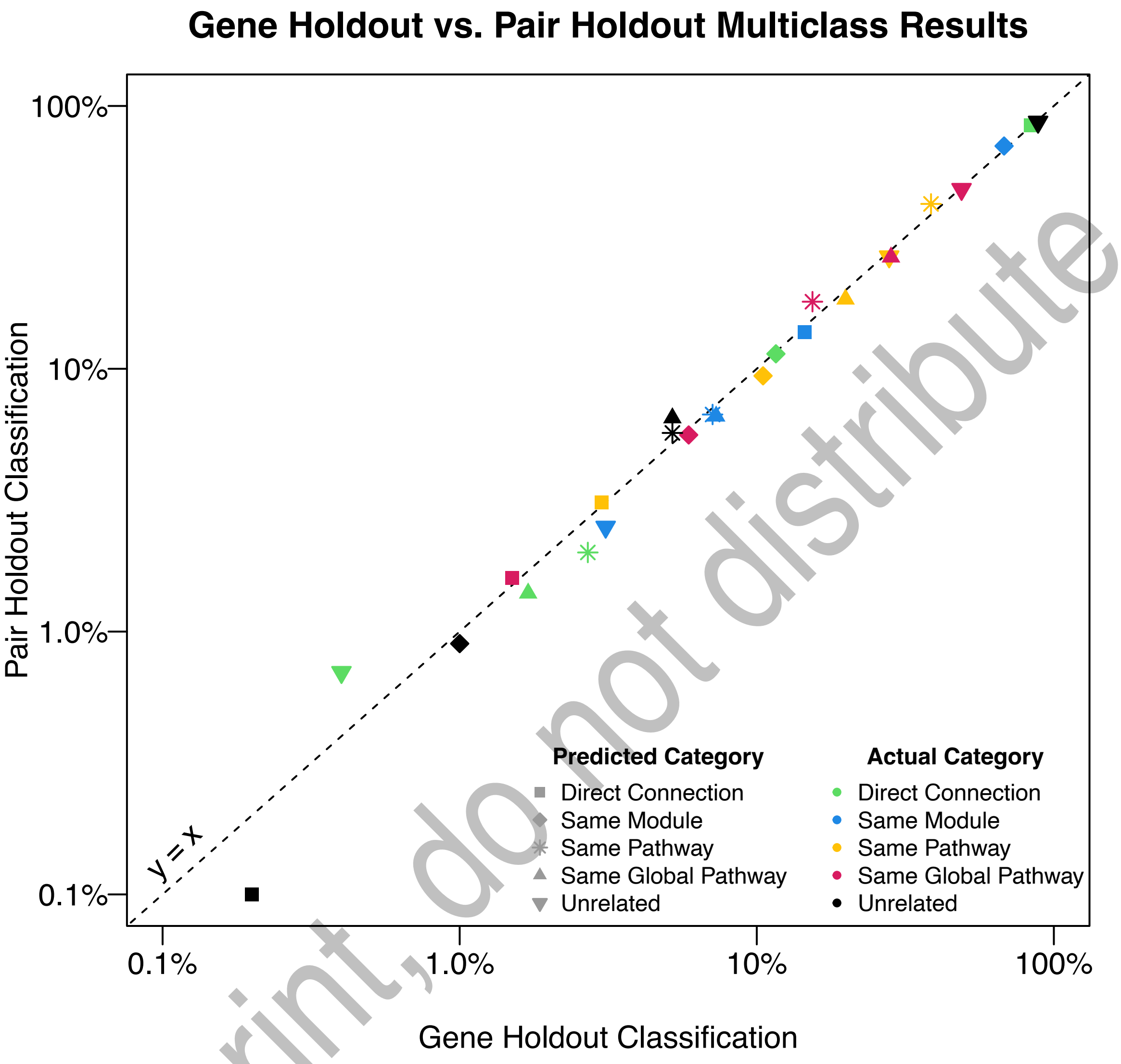


**Figure S3: EvoWeaver's hierarchical classifications are consistent using gene holdouts or gene pair holdouts.** Each point denotes the percentage of pairs in each actual category (point color) classified to each predicted category (point shape). The dashed identity line (i.e., y=x) represents a scenario of perfect consistency between the two evaluations. Note the log scaled axes used for visual clarity. Resulting classifications are almost identical between holdout approaches, implying that EvoWeaver is not simply learning to identify highly connected gene groups.