Joint representation of molecular networks from multiple species improves gene classification

Christopher A Mancuso[1], Kayla A Johnson[2,3,4], Renming Liu[4], Arjun Krishnan[2,4,*]

[1]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO
[2]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO
[3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI
[4]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI
[*]To whom correspondence should be addressed (arjun.krishnan@cuanschutz.edu)

Abstract

Network-based machine learning (ML) has the potential for predicting novel genes associated with nearly any health and disease context. However, this approach often uses network information from only the single species under consideration even though networks for most species are noisy and incomplete. While some recent methods have begun addressing this shortcoming by using networks from more than one species, they lack one or more key desirable properties: handling networks from multiple species, incorporating many-to-many orthology information, or generating a network representation that is reusable across different types of and newly-defined prediction tasks. Here, we present GenePlexusZoo, a framework that casts molecular networks from multiple species into a single reusable feature space for network-based ML. We demonstrate that this multi-species network representation improves both gene classification within a single species and knowledge-transfer across species, even in cases where the inter-species correspondence is undetectable based on shared orthologous genes. Thus, GenePlexusZoo enables effectively leveraging the high evolutionary molecular, functional, and phenotypic conservation across species to discover novel genes associated with diverse biological contexts.

Introduction

Molecular interaction networks serve as genome-scale functional maps that help contextualize and expand our knowledge about the various roles genes play in health and disease. Specifically, networks help in 'gene classification', the problem of predicting novel genes associated with a biological context of interest — e.g., pathway, phenotype, or disease — based on a set of known genes previously identified in multiple low-throughput studies or in single high-throughput genetic screening or omics experiments. While network-based gene classification has been shown to be a powerful approach (Warde-Farley *et al.*, 2010; Ideker and Sharan, 2008; Vanunu *et al.*, 2010; Liu *et al.*, 2020), the gene/protein networks this approach relies on are quite noisy and largely incomplete in many individual species including humans. A potentially powerful way to address the shortcomings of individual networks is to leverage the

high evolutionary molecular, functional, and phenotypic conservation across species by combining network information from multiple species to improve gene classification within a single species.

Considering networks from multiple species also provides the opportunity to address another major biological and biomedical challenge, which is transferring knowledge from human to model species to human, or vice versa; a challenge that often causes clinical trials to fail (Mak *et al.*, 2014; Leenaars *et al.*, 2019; Bugelski and Martin, 2012; Hünig, 2012). Most current methods for transferring genetic and molecular knowledge across species are often restricted to using one-to-one orthologs, and not those with one-to-many and many-to-many relationships. Further, deletion of orthologous pairs does not always give rise to the same phenotypes in different species (Liao and Zhang, 2008), reinforcing the functional differences in the "same" genes between model species and humans (Beura *et al.*, 2016; Geifman and Rubin, 2013; Mestas and Hughes, 2004). While methods are being developed for cross-species knowledge transfer on the annotation level (Djordjevic *et al.*, 2016; Cai *et al.*, 2021) or using expression data (Brubaker *et al.*, 2019; Normand *et al.*, 2018), it is essential to develop methods that utilize genome-scale network information from multiple species to build predictive models that capture the global functional context of nearly every gene (even those uncharacterized) in all the species.

A few different approaches have been proposed to leverage network information across multiple species. Some of these methods directly find similar functional modules in the networks of each species; either through graph alignment (Guzzi and Milenković, 2018) or the use of gene families (Park *et al.*, 2013). Methods such as MUNK (Fan *et al.*, 2019), MUNDO (Arsenescu *et al.*, 2022), and ETNA (Li *et al.*, 2022) cast genes across pairs of species (*e.g.*, human and mouse) into a joint low dimensional embedding space, anchoring the embedding space together with one-to-one-orthologs. A recent deep learning approach, NetQuilt (Barot *et al.*, 2021), trains a neural network model using a multispecies network representation as input and predicts the full set of gene ontology (GO) biological process (Ashburner *et al.*, 2000) terms for each gene. There also exist many methods that combine information from multiple networks within the same species — using diffusion states (Warde-Farley *et al.*, 2010), autoencoders (Gligorijević *et al.*, 2018) or graph convolutional networks (Forster *et al.*, 2022) — that could be readily adapted for the use in cross species knowledge transfer. All of the above mentioned methods lack one or more desirable properties: 1) the ability to handle networks from multiple species, 2) incorporating many-to-many orthology information, or 3) generating a representation that is reusable across many tasks (*i.e.,* pathway, disease or phenotype prediction), including new tasks defined by users.

In this study, we explore the question of whether a joint representation of molecular networks from multiple species can be created so that this representation can be used on-the-fly for any gene classification task, both to improve predictions within a species and to prioritize novel genes across species, especially with the capability to seamlessly start with genes in any one species and prioritize genes in any other. To that end, we present GenePlexusZoo, a framework that casts molecular network information from multiple species (*i.e.*, human and five major

research organisms) into a single reusable feature space for network-based machine learning (ML). We report three key results enabled by this framework on modeling cross-species gene associations. First, we show that, even when restricting training and prediction on gene annotations from a single target species, deriving the features of the ML classifier from multiple species outperforms ML classifiers with features from only the network from the target species. Second, we examine the case of training with annotations from one species and then predicting annotations in another species, and show that creating joint representations using node embedding with degree-weighted random walks allows for better knowledge transfer across species than more naive methods. Lastly, we present several examples of training a model using human disease-associated genes and predicting genes and biologically-meaningful gene sets in research organisms, including multiple cases when the relevant gene sets (*e.g.*, pathways and phenotypes) in the other species share no common genes (orthologs) with the human disease. Overall, our work demonstrates that we can create effective joint representations of molecular networks from multiple species that can be easily reused for any within- and cross-species gene classification task. We propose GenePlexusZoo as a general framework where any previous (Fan *et al.*, 2019; Arsenescu *et al.*, 2022; Li *et al.*, 2022; Barot *et al.*, 2021), or future approaches for joint network embedding and learning can be adapted to take the place of our implementation to continue improving cross-species network-based gene classification.

Methods

Networks

Processing networks

We chose networks from six species — *Homo sapiens* (human), *Mus musculus* (mouse), *Danio rerio* (fish), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm), and *Saccharomyces cerevisiae* (yeast) — from two databases, BioGRID (Oughtred *et al.*, 2019) and IMP (Wong *et al.*, 2015), that offer molecular networks that have genome-wide coverage [**Table SM1**]. BioGRID is a sparse network consisting of binary edges (0 or 1) between genes based only on physical interactions between their protein products, whereas IMP is a dense network consisting of weighted edges (ranging continuously between 0 and 1) derived through computational integration of thousands of high-throughput genomic datasets. For each network, we used the Entrez IDs supplied by the databases to identify genes and, for each species, we removed any gene not present in that species “gene_info” file from the NCBI FTP server.

Connecting networks across species

Retaining the within-species edges as they were, we added edges between cross-species gene pairs using orthology information from the eggNOG database (Huerta-Cepas *et al.*, 2019), which contains pre-calculated orthologous groups for many species. For every pair of genes in a given orthologous group that come from different species, an edge was added between that pair of genes [**Fig. SM1**]. We examined two different ways to weight these cross-species edges [**Fig.**

SM2]: first, giving all cross-species edges the same edge weight and, second, creating a directed edge with an edge weight that is a function of the within-species degree of the source node and how many total cross-species edges the source node has. See the *Supplementary Material* for more information on the network properties and cross-species edge construction.

Creating feature representations

Let $G = (V, E, W)$ denote an undirected molecular network, where V is the set of vertices (genes from all species being considered), E is the set of edges (relationships between genes), and W is the set of edge weights (the strengths of the relationships). G can be represented as a weighted adjacency matrix $A_{i,j} = W_{i,j}$, where $A \in R^{|V|\times|V|}$ and every row in this matrix can be treated as a feature vector for each gene.

From G, we also derived a low-dimensional feature set through the process of node embedding. We used *PecanPy*, an optimized implementation of the *node2vec* algorithm (Grover and Leskovec, 2016; Liu and Krishnan, 2021). At a high-level, *node2vec* generates a set of random walks from the network and then uses these walks as input into a neural network that places nodes that often occur close to each other in the walk close together in a low dimensional embedding space. In *node2vec*, the random walks are controlled using two parameters p and q, in which a high value of q keeps the walk local (a breadth-first search), and a high value of p encourages outward exploration (a depth-first search). Additionally, the user can control the dimension of the embedding space (*d*), as well as the number and length of walks. Various combinations of these hyperparameters were examined in this study (see the *Supplemental Material*).

Gene set collections

Gene set collections for different prediction tasks

To assess the performance of the ML classifiers, we investigated three different tasks: i) gene function prediction (using annotations from MyGene.info (Wu *et al.*, 2013; Xin *et al.*, 2016) mapped to terms in the Gene Ontology (Ashburner *et al.*, 2000), ii) gene phenotype prediction (using annotations from and terms defined in Monarch (Mungall *et al.*, 2017)), and iii) disease gene prediction (using annotations from DisGeNet (Piñero *et al.*, 2015) mapped to terms in the Disease Ontology (Schriml *et al.*, 2019)). For function and phenotype prediction, we were able to generate geneset collections for all six model species. Disease prediction was evaluated only for humans.

For each task, we created a gene set collection by first converting all annotations to Entrez gene IDs. For GO and DisGeNet, the annotations were then propagated from the more specific terms to the more general terms of the Gene and Disease Ontologies, respectively. Next, we filtered the genes to retain only those that were found in a given network and filtered the terms to only those that had between a specified minimum and maximum number of annotations. Mouse and

human GO gene set collections were further filtered to only include terms in a previously curated set of specific terms (Greene *et al.*, 2015).

Deriving non-redundant gene sets

Next, we developed a procedure to take all gene sets within a collection and reduce them to a set of non-redundant terms. First, in a pairwise fashion, we calculated the set overlap and Jaccard scores of the genes annotated to each term. Then, we created a sparse term similarity graph by introducing edges between pairs of terms if the overlap and the Jaccard score were over predefined thresholds, and found connected components in this graph. If a connected component only had one term, that term was directly added to the non-redundant set. In the case where the component contained more than one term, we picked the term most similar to all others in the component as the representative term to be included in the non-redundant set.

Evaluation scheme

Instead of splitting genes randomly into the training and testing sets, we devised a study-bias holdout scheme to evaluate the scenario that is close to the real-world situation of learning from well-characterized genes to predict novel un(der)-characterized genes. In this scheme, genes annotated to each term were split into training and testing sets based on how well-studied they were, with well-studied genes being considered training genes and more novel genes assigned to the test set. Here, we defined study-bias for each gene as the number of articles in PubMed in which that gene was referenced in, as determined in the *gene2pubmed* file (downloaded on 2020-09-23) from the NCBI Gene database (Brown *et al.*, 2015). After creating these gene splits, only terms with at least ten genes in both the training and testing sets were retained. For each term, negatives were chosen as any gene that had at least one annotation to another term in the set of all possible training and testing genes, where we considered genes neutral if they were from gene sets that highly overlapped (Fisher exact test with FDR < 0.05).

Matched non-redundant gene sets across species

To evaluate GenePlexusZoo on its ability to transfer knowledge across species, we created gene set collections that contained only matched terms across species. Due to an insufficient number of matched terms in Monarch phenotypes, this was only possible for the GO biological processes, where matched terms were generated pairwise between humans and each of the other model species. These gene sets were constructed by first converting genes to Entrez IDs and propagating the annotations from the more specific terms to more general terms in GO. The GO term annotations were then subset to the genes in the given network and only terms with at least ten training and testing genes were retained, based on the study-bias validation split. Gene annotations from both species were combined for each term and, using the procedure described above (based on a term similarity graph), we derived a non-redundant set of terms.

Full gene set collections

For evaluation schemes where it is important to find the most relevant biological context, we use the full gene set collections instead of the non-redundant sets described above. For these analyses, we simply created gene set collections by converting gene annotations into Entrez

IDs, propagating the annotations through the ontology (not done for Monarch), subsetting to the given network, and retaining terms with a given minimum-to-maximum number of annotations.

Network-based gene classifiers

GenePlexus uses supervised-learning (SL) for network-based gene classification by using each gene's network neighborhoods as feature vectors, along with gene labels, in a classification algorithm. Here, we used logistic regression with L2 regularization with a strength of 1 as implemented by the *scikit-learn* (Pedregosa *et al.*, 2011) function *sklearn.linear_model.LogisticRegression*. For a more detailed explanation of the GenePlexus classifier see (Liu *et al.*, 2020; Mancuso, Bills, *et al.*, 2022; Mancuso, Liu, *et al.*, 2022).

Evaluation metrics

In this study, we present results in terms of area under the precision-recall curve (auPRC). Since, each geneset can have a different number of positive examples (and, hence, different positive-to-negative proportions), we normalized the standard auPRC by the prior. Specifically, the auPRC presented throughout this study is given by:

$$auPRC = \log_2\left(\frac{auPRC_S}{prior}\right) \qquad \text{(eqn. 1)}$$

where $auPRC_S$ is the standard area under the precision-recall curve, and the *prior* is *P*/(*P*+*N*) with *P* being the number of positive ground truth labels, and *N* being the number of negative ground truth labels. The log2 in eqn. 1 allows for the following interpretation: the number of 2-fold increases of the measured $auPRC_S$ over what is expected based on random chance (*e.g.*, a value of 1 indicates a 2-fold increase, a value of 2 indicates a 4-fold increase).

Precision-based metrics such as the auPRC are more suitable than the more popular area under the receiver-operating characteristic curve (auROC) for two reasons. First, gene classification is a highly imbalanced problem with many more negative examples than positive examples, and auROC is ill-suited for imbalanced problems (Saito and Rehmsmeier, 2015). Second, precision can control for Type-1 error (false positives) (Davis and Goadrich, 2006). Since the foremost reason for gene classification is to provide a list of candidate genes for further experimental follow-up, it is more important to make sure the top predictions are as correct as possible, as opposed to ensuring that, on average, positive examples are ranked higher than negative examples.

Results

Creating joint multi-species network representations using GenePlexusZoo

GenePlexusZoo was designed to build a robust reusable joint network-based feature representation for genes from multiple species for its machine learning (ML) classifier to improve gene classification both within and across species [**Fig. 1**]. It begins with genes and their within-species relationships from human, mouse, fish, fly, worm, and yeast followed by

combining them into a multi-species network by adding connections between pairs of genes across species based on orthology. Such a network built only using protein-protein interactions (PPI) from BioGRID contains 46,741 genes (across the six species) and nearly 1.59 million edges. The network built using predicted functional relationships from IMP contains 135,049 genes and nearly 566 million edges. Converting each of these large network objects into feature sets for ML entails making some choices that impact both gene classification performance and practical feasibility:

- Which species to include? For instance, which is better: the network from that species alone, a network combining that species and another well-studied species, or a network with all six species?
- How to connect genes across species? Should the new connections between cross-species gene pairs be weighted uniformly (with what weight?) or differently based on the degrees of the incident gene nodes?
- How to represent the network as features in the ML classifier? Can the adjacency matrix of the large multi-species network used as-is for ML or should the network be further smoothed (*e.g.*, using random walk with restart) or embedded into a lower dimensions (*e.g.*, using singular value decomposition or node embedding).

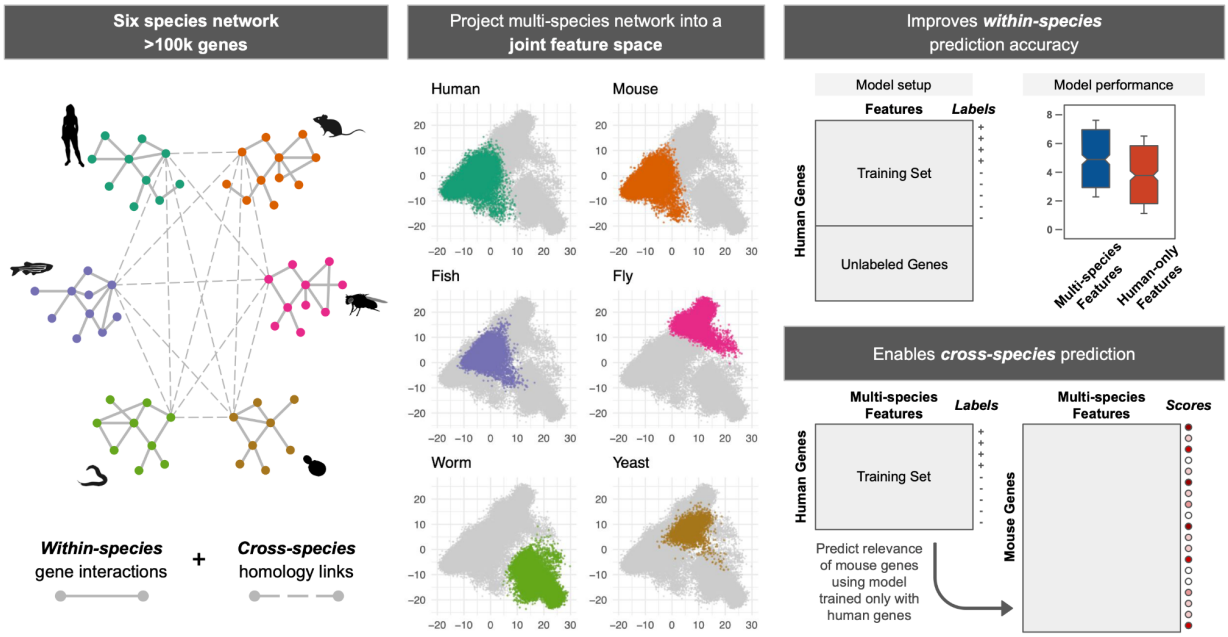

Figure 1: Overview of GenePlexusZoo. GenePlexusZoo is a framework for creating a joint representation of molecular networks from multiple species, which can in-turn be used for training machine learning (ML) classifiers for gene classification within and across species. The creation of the multi-species network representation entails: i) connecting gene networks from multiple species (in this study: human, mouse, fish, fly, worm, and yeast) so that pairs of genes belonging to the same orthologous group are connected across species (*cross-species* links; resulting in a multi-species network that can contain over 100k genes) and ii) projecting this network into a joint feature space by creating a low-dimensional embedding for each gene that captures its topological position and neighborhood. Here, we extensively test the ability of this joint network representation — and several other network-based

feature sets containing the genes of one or more species — in predicting genes associated with numerous functions and phenotypes.

To determine the best choices, we performed extensive evaluation for the human and mouse function and phenotype annotation tasks and arrived at an overall optimal choice of settings and hyperparameters from this analysis (described in detail in the *Supplementary Material*.). From these results, we found that the adjacency matrix (AdjMat) is the best representation among those that directly consider the edges of the network as features. *node2vec* (N2V) is the best-performing embedding method and has the additional properties of readily utilizing different weighting techniques and the ability to computationally handle combining networks from more than two species.

Multi-species network representations improve within-species gene classification

Once we had picked the best performing ways to create feature representations, we wanted to see if features derived from multiple species could offer improvement over features coming from a single species, even when the annotation task is restricted to only one species. For this analysis, we used non-redundant gene sets (see *Methods*) for function annotation (Gene Ontology) and phenotype annotation (Monarch) for the two networks considered in this study (BioGRID and IMP). First, we examined the effect of adding multi-species network information on human gene annotation tasks. For functional annotation, using either BioGRID or IMP, features derived from pairs of species [**Fig. 2**: blues] or features from all six species [**Fig. 2**: orange] perform significantly better than features derived from a single species [**Fig. 2**: greens]. Phenotype annotation is a much harder task in general and all methods perform roughly the same no matter how the features are derived.

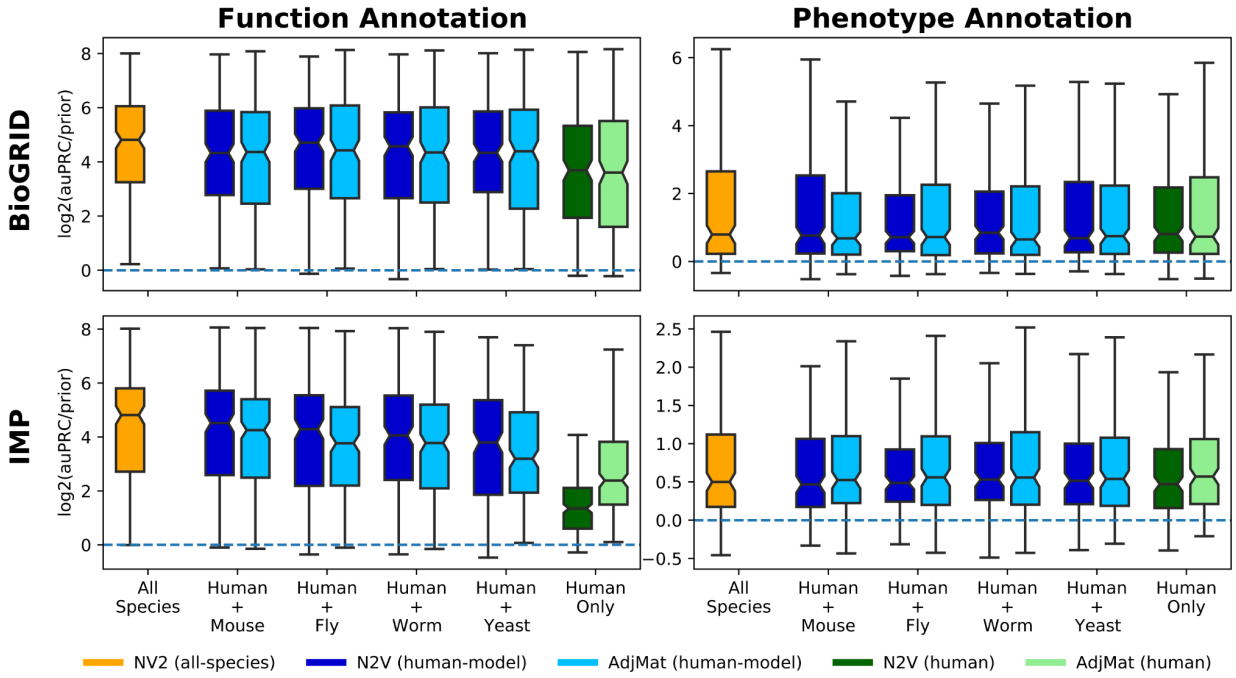

Figure 2: Predicting function- and phenotype-associated human genes using human, human+model, and all-species network representations. The boxplots show the $\log_2$(auPRC/prior) prediction performance of logistic regression classifiers trained to predict human gene annotation for two tasks: function (left column) and phenotype (right column). Features for the classifiers were created using all six species (orange), human plus one other species (blues), or human only (greens), and are represented using the adjacency matrix (AdjMat) or node embeddings (N2V). This analysis was performed using both the BioGRID (top row) and IMP (bottom row) networks.

Then, we examined the effect of adding multi-species network information on model-species gene annotation tasks, where we considered features only from the network for the species the annotation task pertains to [**Fig. 3**: greens], adding human network information to the features [**Fig. 3**: blues] and using features from all six species [**Fig. 3**: orange]. For function annotation, the trends we observed were generally similar to that for human gene annotation; features derived from multiple species typically increase performance. However, for model species annotation, the variation in performance using AdjMat or N2V increases substantially, especially for phenotype annotation.

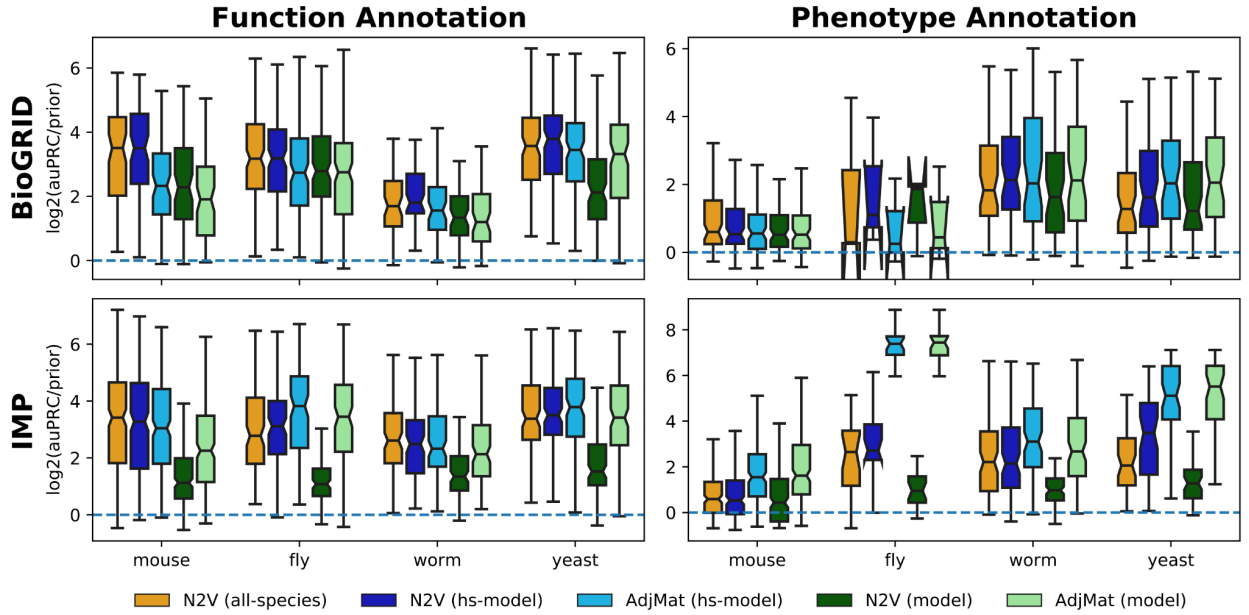

Figure 3: Predicting function- and phenotype-associated model species genes using model, human+model, and all-species network representations. The boxplots show the $\log_2$(auPRC/prior) prediction performance of logistic regression classifiers trained to predict model species gene annotations for two tasks: function (left column) and phenotype (right column). Features for the classifiers were created using networks from all six species (orange), human plus the other species (blue), or the model species only (green), and are represented using the adjacency matrix (AdjMat) or node embeddings (N2V). This analysis was performed using both the BioGRID (top row) and IMP (bottom row) networks.

As a final analysis of within-species gene classification, we checked the impact of explicitly marking orthologs of positive examples as additional positives when training the ML classifiers (on top of multi-species network implicitly containing this information through cross-species gene connections) [**Fig. SM5–6**]. We found that the addition of these potentially redundant

positive examples results in occasional significant increases and decreases in performance for classifiers using the adjacency matrix representation, but consistently results in decreased performance for classifiers using the node embeddings.

Multi-species network embeddings enable cross-species gene classification

Next, we investigated the ability of these feature representations to transfer knowledge between human and model species. To that end, we first established a matched non-redundant gene set collection (see *Methods*) for each model species paired with human. Then, we analyzed the ability of feature representations that combine human and model species networks to enable an ML classifier trained on a given human gene set (*e.g.*, a particular pathway) to accurately predict the matched model-species gene set (e.g., the same pathway in the other species). We summarized the results of this analysis in terms of the ability of the human ML classifier to prioritize the genes in the matched model-species gene set [**Fig. 4**; top row] as well as how the gene predictions in the matched gene set compared to all the other (dissimilar, unmatched) gene sets in the collection [**Fig. 4**; bottom row]. These summaries clearly indicate that feature representations using node embedding are more accurate transferring knowledge from human to model species. This trend holds for joint networks built using IMP and BioGRID [**Fig. 4**; **Fig. SM7**] and in the inverse case of transferring knowledge from model species to human (e.g., training on model species gene sets and predicting the matched human gene sets) [**Figs. SM8–9**].

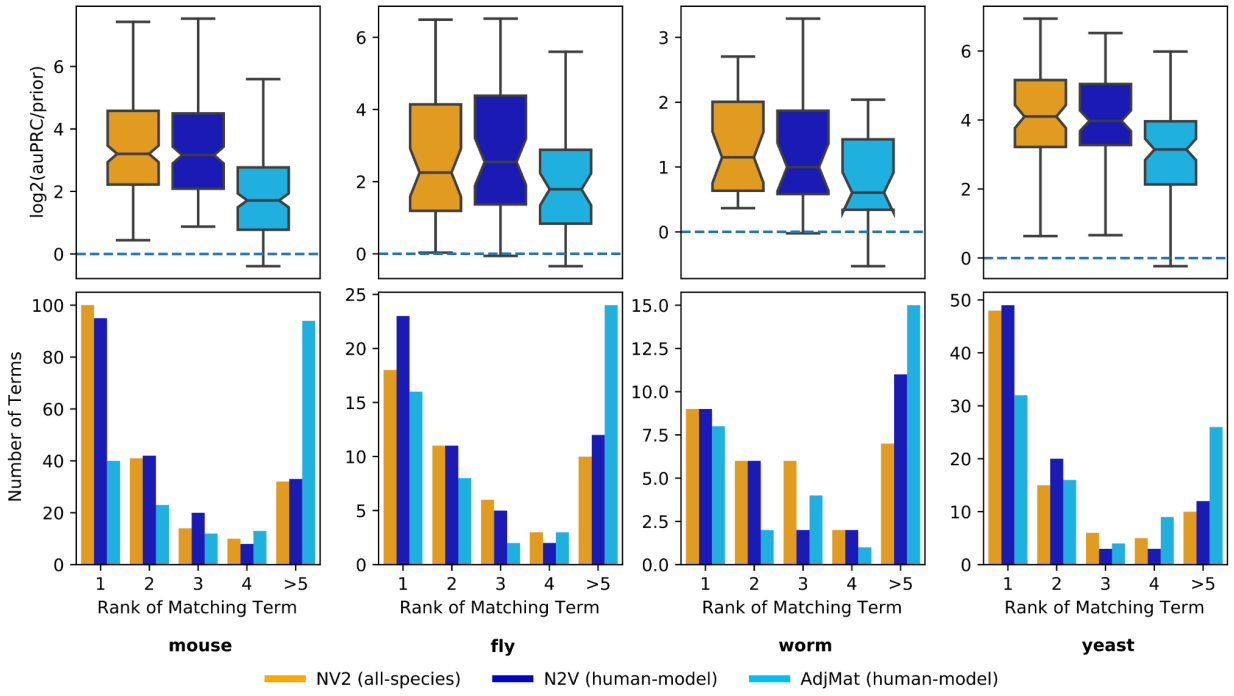

Figure 4: Transferring knowledge from human to model species using human+model and all-species network representations. The boxplots (top row) show the $\log_2$(auPRC/prior) prediction performance of logistic regression classifiers that were trained using only human genes annotated to a

given biological process and then used to predict model species genes annotated to the same biological process. The barplots (bottom row) show the counts of where the $\log_2$(auPRC/prior) for the matched biological process ranks in comparison to the other processes in the collection. Features for the classifier were created pairwise between human and a model species with IMP networks using the AdjMat (light blue) and N2V (dark blue) representations, as well as features from all species using the N2V (orange) representation.

Delving further into the benefits of multi-species network embeddings (involving pairs of species or all species), we asked how does an ML classifier for a particular gene set perform in a given species when it is trained on gene annotations in that species compared to when it is trained on annotations from another species [**Figs. SM10–13**]. We observed that, for human gene annotation tasks, there is invariably a substantial drop in performance when the classifier is trained with model-species annotations compared to human annotations. However, the performances for the model species gene annotation tasks are variable. Compared to training on model-species annotations, training on human annotations results in reduced performance in fly and worm, but on-par performance in mouse and yeast, indicating that the knowledge in human is often a good proxy for these species.

As a final evaluation of cross-species transfer, we checked if training the ML classifier on both human and model species annotations would improve its performance in either species [**Fig. SM14–17**]. An analysis on prediction tasks in both human and model showed that, when using the multi-species network embeddings as features, including annotations from multiple species in the same classifier does not improve gene classification performance in a target species (human or model). The similarity of performance is likely because the information from different species is already inherent to the multi-species feature representation. This hypothesis is supported by the clear significant jump in performance between using features from a single species and features from multiple species.

Mapping human diseases across species using all-species network embeddings

Having established the broad utility of all-species network embeddings in GenePlexusZoo, we explored the ability of this network representation to map human diseases across species and help discover meaningful and novel insights into diseases in all six model species. First, we identified ten human diseases that had top-performance in network-based disease-gene prediction (trained and tested only on human gene annotations and evaluated based on study-bias holdout for both BioGRID and IMP). Then, for each of these diseases, we trained a new ML classifier using all human genes annotated to that disease using the all-species network embeddings as features and made a prediction for every gene in all the five model species. For each model species, we could determine which processes and phenotypes were most enriched among the top-ranked “disease” genes in that species [**Supplemental File1**] (see *Supplemental Material* for more information on how the enrichment scores are calculated).

To illustrate how GenePlexusZoo can provide insights across species, we focus on the ciliopathic disorder Bardet-Biedl Syndrome 1 (BBS) (Forsythe and Beales, 2013). We choose

this syndrome because it has a strong genetic component and impacts many processes across the body. Upon training an ML classifier only on human BBS genes and predicting genes in all model species, we observed that the top-ranked genes [**Table SM5**], biological processes [**Table SM6**], and phenotypes in the model species are relevant to cilium function and manifestations of the BBS disorder such as rod/cone dystrophy, polydactyly, and hypogonadism [**Tables SM7**]. To better highlight the relationships between the top biological processes across species, we summarized them as a graph [**Fig. 5**] with nodes representing individual processes, node colors encoding the species the processes were identified in, and edges connecting semantically similar processes. Among these top processes were 13 where, in at least one species, the genes annotated to that biological process had no orthologs to any human gene annotated to BBS (see *Discussion* section for further discussion of this result).

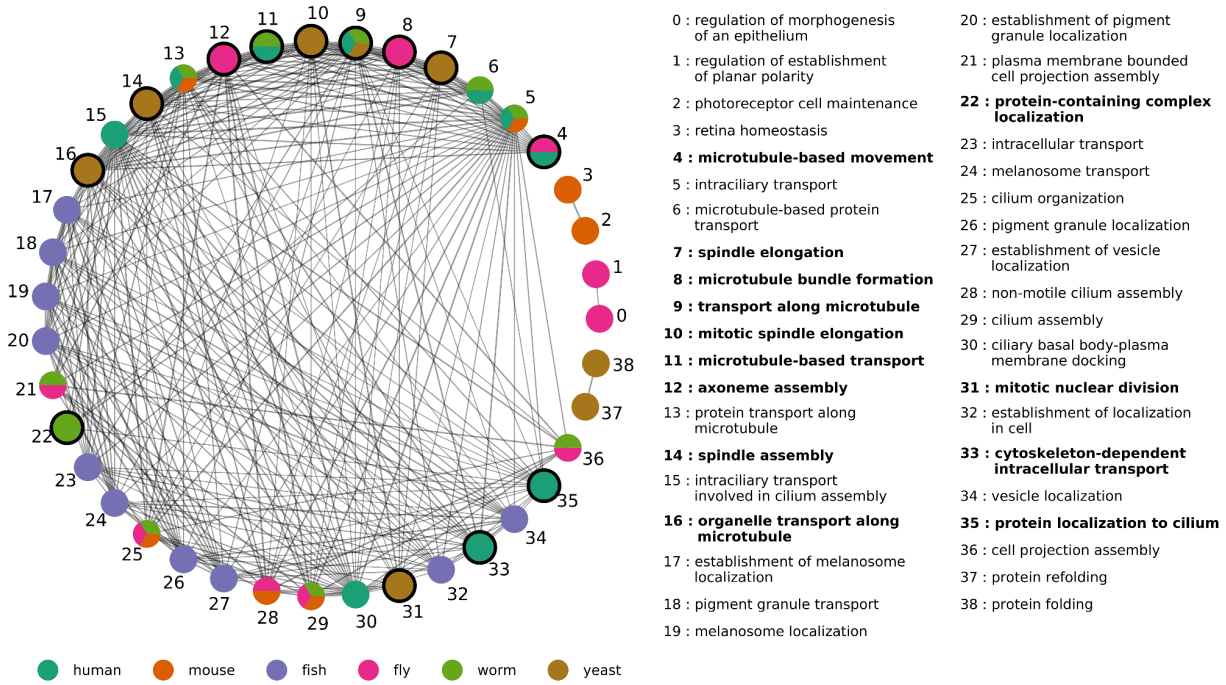

Figure 5: The ten most enriched biological processes associated with the top genes in each species predicted to be related to Bardet-Biedl Syndrome 1 (BBS). A classifier trained using human BBS genes and all-species N2V features was used to predict BBS-related genes in the five model species. The figure shows the ten most enriched biological processes associated with the top-ranked genes in each species. Nodes represent biological processes and are colored by which species they were identified in. Edges represent pairs of semantically similar processes (scaled Resnik similarity based on the Gene Ontology). Isolated nodes are not shown. Nodes with thick borders represent biological processes where, in at least one species, none of the annotated genes are orthologous to any human BBS gene.

Embeddings bring together genes from multi-species networks

All the results above demonstrate the power of multi-species network representations in improving gene classification performance both within and across species, and in enabling knowledge transfer about human diseases to model species. These results also show that

performance is influenced by how connections are made between genes across species in the multi-species network and how the multi-species network is turned into a feature representation (i.e., as is in the form of the adjacency matrix or as low-dimensional node embeddings; see *Methods* and *Supplementary Material*). Therefore, we performed visual exploratory analysis to inspect the feature spaces created through these different choices and to reason why the embeddings perform so well in cross-species tasks [**Fig. 6**; **Figs. SM19-21**].

The first observation from the plots comparing the features from different choices is that genes from different species are very disjoint in the adjacency matrix representation. This could perhaps explain why within species like fly and yeast, the adjacency matrix has such high performance, but for transferring knowledge across species, the adjacency matrix is the worst performing representation. On the other hand, multi-species network embeddings bring together genes from different species. Second, genes across species are brought much closer together when a degree-based weighting strategy is used to connect genes across species before embedding. These trends hold in pairwise settings (human and mouse [**Fig. 6**]; human and other model species [**Figs. SM19-21**]) and in the all-species setting.

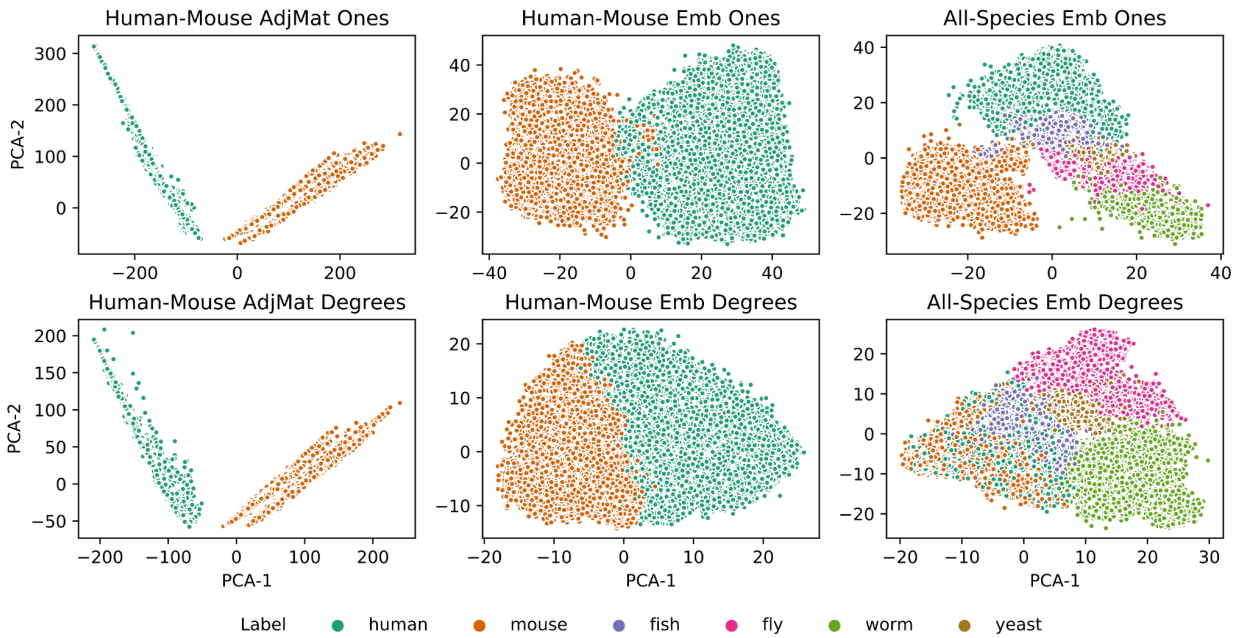

Figure 6: PCA of adjacency matrix and node embedding representations. Each PCA plot shows the top two PCA components from the human-mouse adjacency matrix (first column), the human-mouse node embedding (second column) and the all species node embedding (third column) representations. The cross-species edge weights were created by always using a uniform edge weight with a value of one (top row) or by using degree based edge weights (bottom row) for the IMP network.

Discussion

GenePlexusZoo is a network-based gene classification framework that derives features for a supervised machine learning classifier from molecular networks from multiple species, providing a powerful foundation for improving gene classification accuracy within species and for

transferring knowledge across species. GenePlexusZoo affords many ways to represent networks from multiple species including a novel method introduced here for generating a concise cross-species joint network representation that can be easily reused for any new gene classification task. This method involves connecting genes from different species together using orthology information and then jointly embedding them using a random-walk based algorithm, *node2vec* (Grover and Leskovec, 2016). We find that the strategy for setting the weights between genes across species can be very important and in particular we see that using a degree-based method that encourages walks to more readily traverse cross-species edges shows the best performance.

One of the goals of GenePlexusZoo is to offer a seamless way to handle one-to-many and many-to-many orthologs that goes beyond the typical workflows that use one-to-one orthologs to convert gene lists between human and model species. Previous computational approaches, such as MUNK (Fan *et al.*, 2019), anchor networks together by “connecting” genes across species as one-to-one pairs. However, genes often act as coherent modules and simple one-to-one matching can be limiting to an analysis. It is often daunting and non-intuitive for researchers to handle one-to-many or many-to-many mappings, as a gene of interest in one species may be in a “functional orthologous group” with many genes from another species, thus clouding the transfer of knowledge. GenePlexusZoo inherently incorporates many-to-many gene mappings by generating network-based feature representations where genes from different species are close to each other if they are similar based on both sequence (orthology) and function (network neighborhood). Upon a classifier is trained on such a joint representation using annotations from one species, the resulting predictions in another species are returned as a ranked list, removing much ambiguity as to which of the one-to-many or many-to-many mapped genes are important for the user's task.

We performed extensive evaluations comparing our new multi-species network embedding with many other ways to represent networks from single and multiple species. First, these evaluations show that the most suitable network representation is task dependent. For example, as yeast networks have been extensively mapped out experimentally, when training and predicting on only yeast gene sets, the yeast (single species) adjacency matrix representation usually has the same performance as the adjacency matrix representation derived from both human and yeast networks [**Fig. 3**]. However, these large feature representations that naively concatenate features from the different species appear to overfit and have trouble generalizing when transferring knowledge from one species to another [**Fig. 4**]. When creating low-dimensional representations using node embedding, building feature sets from more than one species almost always offers significant improvement over single species features [**Fig. SM14-17**]. With the cross-species gene relationships already encoded in these representations, explicitly adding genes from multiple species as extra positives in the classifier does not improve performance.

After establishing the merits through the various evaluations, we applied GenePlexusZoo’s multi-species network representation to a number of human disorders to explore the framework’s ability to identify cross-species counterparts of human diseases and gain insights

into the molecular and phenotypic basis of these conditions [**Supplementary File 1**]. The case of ciliopathic disorder Bardet-Biedl Syndrome 1 (BBS) (Forsythe and Beales, 2013) provides an illustration of this application. Here, we trained a classifier only using human BBS genes and predicted "BBS-associated" genes in the five model species. The genes predicted by GenePlexusZoo in the model species — and the biological processes and phenotypes associated with these top-ranked genes — collectively correspond to the various manifestations of this syndrome including rod/cone dystrophy, polydactyly, central obesity, hypogonadism and kidney dysfunction. This result indicates that the network-based classifier is general enough not to overfit to any one aspect of the syndrome and is able to recapitulate the diverse set of systems affected by BBS.

Further, this analysis has the potential to not only prioritize the best process/pathway to study in a given model organism, but also point to which model organism might be the best to study a given process/pathway. For example, ciliary behavior is related to the etiology of BBS and it has been studied extensively in fly (Avidor-Reiss *et al.*, 2004; zur Lage *et al.*, 2019). Previous studies like this often guide the choice of model organism and, in this case, suggest fly as a good model. However, the association scores calculated using GenePlexusZoo for processes and, in particular, phenotypes related to ciliary behavior are quite low in fly [**Supplementary File 1**]. On the other hand, the scores for these processes and phenotypes are much higher in worm, suggesting that it might be a better model than fly to study ciliary behavior. This hypothesis is supported by studies that show that Bardet-Biedl syndrome might be more driven by dysregulation of non-motile cilia, which is corroborated by GenePlexusZoo results [**Tables SM6-7**] (Tobin and Beales, 2009; Mitchison and Valente, 2017). Further, many more genes have been associated with non-motile cilia in worm (32) compared to fly (13), possibly due to the fact that worms only have non-motile cilia in one cell type (sensory neuron) (Inglis *et al.*, 2007), thus lending itself as a better system for studying non-motile cilia diseases.

The application of GenePlexusZoo to BBS also brought to light a fundamental challenge in identifying 'phenologs', phenotypes in different species that are 'orthologous' by virtue of having a similar genetic and molecular basis (*i.e.*, similar functional modules composed of orthologous genes) (McGary *et al.*, 2010). While phenologs are critical for studying cross-species models of human diseases, finding them solely based on shared (orthologous) genes is problematic because our knowledge about the genes involved in both human diseases and model species phenotypes is far from complete. This incomplete knowledge can lead to bona fide phenologs being missed (false negatives) because the genes experimentally characterized thus far for the human and model phenotype happen to be non-overlapping (or minimally overlapping) even if they are part of the same underlying functional module in the two species. This challenge can be deftly remedied by using a network-based approach like GenePlexusZoo because it can relate disjoint sets of genes across species if they are close to each other in the underlying networks. As evidence of this capability, GenePlexusZoo picks out highly similar processes and phenotypes in model species even when they share no orthologs with BBS in human [**Fig. 5**]. Another example is that in the case of organic acidemia (OA) [**Fig. SM18**], a class of inherited metabolic disorders that arise due to defects in intermediary metabolic pathways of carbohydrate, amino acids, and fatty acid oxidation (Vaidyanathan *et al.*, 2011). Here, we

observe that fatty acid biosynthetic process (FABP), a biological process highly relevant to OA, is top-ranked in mouse, but the two sets of genes (annotated to human OA and mouse FABP) have no shared orthologs.

Considering the similarities between all the top processes prioritized across species sheds additional light on the diversity of the molecular phenomena related to a particular human disease in terms of their biological relationships and the extent of our knowledge about them. For instance, the top processes for BBS in human and model species are highly similar to each other, with many of them being exactly the same across species [**Fig. 5,** multicolored nodes]. On the other hand, in OA, there are zero biological processes that are exactly overlapping between species. Moreover, the top human biological processes have strong connections within themselves and not with biological processes from other species. The extreme case of this scenario can be seen in zebrafish in which the processes are only connected to each other and have no ortholog overlap with the human gene annotations for OA. Depending on the research question, these observations from the two human diseases imply that it might be harder to transfer research results across species for OA than BBS, or that, conversely, there are many more novel research avenues for studying OA in model species. Though we only highlight BBS and OA here, GenePlexusZoo identified biologically-meaningful processes and phenotypes across all the diseases we analyzed, with every case including model counterparts that have no gene overlap with the human disease [**Supplementary File 2**].

Thus, GenePlexusZoo is a framework that will enable researchers to take any gene set they are interested in and build classifiers on-the-fly that leverage the molecular, functional, and phenotypic conservation across multiple species. The joint multi-species network representation of GenePlexusZoo simultaneously results in higher predictive power and offers a unique way to transfer knowledge directly across species compared to using a single species network. We note that the novelty of this work lies in the framework of this approach, and not so much on the exact manner in which the reusable network-based feature space is generated. Our implementation of this framework allows future work to explore how other similar methods for joint network representation could be adapted to continue improving cross-species knowledge transfer.

Data and code availability

To enable the research community to work on these developments, we have made the relevant code from this study available at https://github.com/krishnanlab/GenePlexusZoo-Mancusript and the data available at https://zenodo.org/record/7888044.

Acknowledgements

The authors wish to thank the members of the Krishnan Lab for valuable discussions.

Funding

This work was primarily supported by US National Institutes of Health (NIH) grant R35 GM128765 to AK and F32 GM134595 to CAM.

Conflict of interest

The authors state no conflict of interests.

References

Arsenescu,V. *et al.* (2022) MUNDO: protein function prediction embedded in a multispecies world. *Bioinforma. Adv.*, **2**, vbab025.

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

Avidor-Reiss,T. *et al.* (2004) Decoding Cilia Function: Defining Specialized Genes Required for Compartmentalized Cilia Biogenesis. *Cell*, **117**, 527–539.

Barot,M. *et al.* (2021) NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity. *Bioinformatics*, **37**, 2414–2422.

Beura,L.K. *et al.* (2016) Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature*, **532**, 512–516.

Brown,G.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.

Brubaker,D.K. *et al.* (2019) Computational translation of genomic responses from experimental model systems to humans. *PLOS Comput. Biol.*, **15**, e1006286.

Bugelski,P.J. and Martin,P.L. (2012) Concordance of preclinical and clinical pharmacology and toxicology of therapeutic monoclonal antibodies and fusion proteins: cell surface targets. *Br. J. Pharmacol.*, **166**, 823–846.

Cai,M. *et al.* (2021) XGSEA: CROSS-species gene set enrichment analysis via domain adaptation. *Brief. Bioinform.*, **22**, bbaa406.

Davis,J. and Goadrich,M. (2006) The Relationship Between Precision-Recall and ROC Curves. In, *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06. ACM, New York, NY, USA, pp. 233–240.

Djordjevic,D. *et al.* (2016) XGSA: A statistical method for cross-species gene set analysis. *Bioinformatics*, **32**, i620–i628.

Fan,J. *et al.* (2019) Functional protein representations from biological networks enable diverse cross-species inference. *Nucleic Acids Res.*, **47**, e51.

Forster,D.T. *et al.* (2022) BIONIC: biological network integration using convolutions. *Nat. Methods*, **19**, 1250–1261.

Forsythe,E. and Beales,P.L. (2013) Bardet–Biedl syndrome. *Eur. J. Hum. Genet.*, **21**, 8–13.

Geifman,N. and Rubin,E. (2013) The Mouse Age Phenome Knowledgebase and Disease-Specific Inter-Species Age Mapping. *PLOS ONE*, **8**, e81114.

Gligorijević,V. *et al.* (2018) deepNF: deep network fusion for protein function prediction. *Bioinformatics*, **34**, 3873–3881.

Greene,C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.

Grover,A. and Leskovec,J. (2016) Node2Vec: Scalable Feature Learning for Networks. In, *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, New York, NY, USA, pp. 855–864.

Guzzi,P.H. and Milenković,T. (2018) Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Brief. Bioinform.*, **19**, 472–481.

Huerta-Cepas,J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.

Hünig,T. (2012) The storm has cleared: lessons from the CD28 superagonist TGN1412 trial. *Nat. Rev. Immunol.*, **12**, 317–318.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Inglis,P.N. *et al.* (2007) The sensory cilia of Caenorhabditis elegans. *WormBook Online Rev. C Elegans Biol.*, 1–22.

zur Lage,P. *et al.* (2019) Survey of the Ciliary Motility Machinery of Drosophila Sperm and Ciliated Mechanosensory Neurons Reveals Unexpected Cell-Type Specific Variations: A Model for Motile Ciliopathies. *Front. Genet.*, **10**, 24.

Leenaars,C.H.C. *et al.* (2019) Animal to human translation: a systematic scoping review of reported concordance rates. *J. Transl. Med.*, **17**, 223.

Li,L. *et al.* (2022) Joint embedding of biological networks for cross-species functional alignment. 2022.01.17.476697.

Liao,B.-Y. and Zhang,J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci.*, **105**, 6987–6992.

Liu,R. *et al.* (2020) Supervised learning is an accurate method for network-based gene classification. *Bioinformatics*, **36**, 3457–3465.

Liu,R. and Krishnan,A. (2021) PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics*, **37**, 3377–3379.

Mak,I.W. *et al.* (2014) Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.*, **6**, 114–118.

Mancuso,C.A., Bills,P.S., *et al.* (2022) GenePlexus: a web-server for gene discovery using network-based machine learning. *Nucleic Acids Res.*, **50**, W358–W366.

Mancuso,C.A., Liu,R., *et al.* (2022) PyGenePlexus: A Python package for gene discovery using network-based machine learning. 2022.07.02.498552.

McGary,K.L. *et al.* (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 6544–6549.

Mestas,J. and Hughes,C.C.W. (2004) Of mice and not men: differences between mouse and human immunology. *J. Immunol. Baltim. Md 1950*, **172**, 2731–2738.

Mitchison,H.M. and Valente,E.M. (2017) Motile and non-motile cilia in human pathology: from function to phenotypes. *J. Pathol.*, **241**, 294–309.

Mungall,C.J. *et al.* (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.

Normand,R. *et al.* (2018) Found In Translation: a machine learning model for mouse-to-human inference. *Nat. Methods*, **15**, 1067–1073.

Oughtred,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.

Park,C.Y. *et al.* (2013) Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes. *PLOS Comput. Biol.*, **9**, e1002957.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Piñero,J. *et al.* (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**.

Saito,T. and Rehmsmeier,M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, **10**, e0118432.

Schriml,L.M. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

Tobin,J.L. and Beales,P.L. (2009) The nonmotile ciliopathies. *Genet. Med.*, **11**, 386–402.

Vaidyanathan,K. *et al.* (2011) Organic Acidurias: An Updated Review. *Indian J. Clin. Biochem.*, **26**, 319–325.

Vanunu,O. *et al.* (2010) Associating Genes and Protein Complexes with Disease via Network Propagation. *PLOS Comput. Biol.*, **6**, e1000641.

Warde-Farley,D. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.

Wong,A.K. *et al.* (2015) IMP 2.0: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.*, **43**, W128–W133.

Wu,C. *et al.* (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.*, **41**, D561–D565.

Xin,J. *et al.* (2016) High-performance web services for querying gene and variant annotation.

Genome Biol., **17**, 91.