Protein Link Augmentation for Functional Prediction: Combining Spectral and Machine Learning Techniques

A thesis

submitted by

Andrew DelMastro

In partial fulfillment of the requirements
for the degree of

Master of Science

in

Computer Science

TUFTS UNIVERSITY

August 2023

ADVISOR: Prof. Lenore J. Cowen

Abstract

Recent advances in network-based methods for functional annotation of proteins have proved effective in well annotated species. However, network information is lacking in many species, so these methods cannot be effectively applied. Some methods, such as MUNK and MUNDO, attempt a co-embedding of well and lesser annotated species to boost performance. We explore the use of ML based approaches to augment these networks, with the intent of increasing the performance of co-embedding methods. A model that can predict interactions between two proteins was used to add potentially missing interactions or remove existing false interactions. We found that too much noise is added to the network to be useful. This does not rule out future endeavors in the area or specific applications of ML in instances where the model is known to perform well, only that generalized models are currently not capable of protein-protein interaction network augmentations.

Acknowledgments

I would like to thank Dr. Lenore Cowen of the Tufts University Department of Computer Science for her guidance on the project, and Donna Slonim for her participation in my thesis committee. Additionally, I appreciate all of the help provided by Kapil Devkota and insight of Mert Erden throughout the project.

List of Tables

List of Figures

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Protein Function Prediction

In an organism, a protein can serve one or more purposes, and each protein can be assigned functional labels, by some ontology, that categorizes the roles that it performs in the cell [ABB+00]. Knowledge of a protein's function can be leveraged in many ways. For example, knowing that a certain protein is important in a specific disease related pathway might suggest it as a candidate for potential treatment targets. Many studies have been conducted to assign each protein its appropriate labels; however, even amongst the species with the most complete annotations, there are still many unlabeled proteins, and beyond that most species have very few functional labels.

Another important area of protein research, is to discover the ways in which they interact with each other. A protein-protein interaction (PPI) network is a graph-like structure that describes the set of protein interactions within an organism [KPA+12]. Some studies have leveraged PPI networks to transfer functional labels to interacting proteins, while others have even investigated cross-species transfer of information [CZP+13].

One standard ontology used for protein function annotation is the Gene Ontology

(GO) hierarchies [ABB+00]. This is a set of functional labels that are divided into three groups: Biological Process, Molecular Function, and Cellular Component, that can be used to categorize all the functions of a protein in the cell. "Biological process refers to a biological objective to which the gene or gene product contributes." "Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product." "Cellular component refers to the place in the cell where a gene product is active. These terms reflect our understanding of eukaryotic cell structure [ABB+00]." During this research, functional prediction will be the task of assigning each protein its set of true GO labels.

1.1.2 Diffusion State Distance

A common trait of PPI networks is that they tend to have low diameter. This is, in part, due to the existence of "hub nodes", or proteins that interact with a large number of other proteins [CZP+13]. These nodes offer a very short path between two proteins in the network that would never interact with each other or have a similar purpose, yet a shortest-path distance based metric would see them as very close. This led to the motivation of a metric known as Diffusion State Distance (DSD). This metric is based on random walks from a node, in a network and attempts to capture how likely other nodes are to be reached on this random walk. DSD offers a distance metric that reduces the importance of these "hub nodes", allow for better information to be extracted from the network [CZP+13].

1.1.3 DSCRIPT

Recently, machine learning has been used to approach many biological problems. Google's AlphaFold has had major success in predicting the 3D structure of many proteins [JEP+21]. Other models have been trained to predict protein interactions for use in generating PPI networks [CJZ+19]. While significant effort has been put into experimental approaches for discovering PPI edges, they are hindered by their high monetary and labor intensive costs. Even if these experiments could be easily carried out, they are prone to many false positives in terms of protein

interactions [YCH15]. As such, these machine learning based approaches offer a reasonable alternative.

One of these models, DSCRIPT, reports to be transferable between species [SSCB21b]. The model is designed in a way to take in only the sequences to build a representative 2D contact map as a way to capture the mechanical mechanisms by which the two would interact. Building off of a robust sequence embedding, this allows for a more generalized approach to the prediction process and lends itself to work across species.

1.1.4 Topsy-Turvy

Since DSCRIPT focused solely on a sequences-based approach to learn interactions, it ignores what we know to be a very valuable source of knowledge, the interaction network [SSCB21b]. Both local and global properties of the network contain valuable information about protein interactions, so a model that completely ignores the network misses out on this information [DMC20]. In an attempt to learn from this network, an alternative to DSCRIPT called Topsy-Turvy was proposed. The model mimics the sequence embedding of DSCRIPT, however, during the training process, in addition to the DSCRIPT loss term, an additional term corresponding with a network based prediction method's score is factored in [DMC20] [SDS$^+$22]. GLIDE was the network-based model chosen for Topsy-Turvy. This allows the model to capture some of the information contained within the PPI network. Since one of the goals of Topsy-Turvy was to make a model transferable to less annotated species, the model does not require network data for prediction, and instead relies only on sequence data for that.

1.2 Overview

In this thesis, we ask if a Topsy-Turvy model can be applied to augment current PPI networks with predicted interactions, in such a way that improves the performance of existing functional annotation methods, or if the noise produced by the model is

still too high to provide any benefit. We investigate different application strategies for the ML model to test which method, if any, produces the best results.

1.3 Outline of This Work

The following is the outline of individual chapters in this thesis. In Chapter 2, we describe the methods used. In Chapter 3, we detail the experimental plan and evaluation metrics. In Chapter 4, we present the results, and in Chapter 5, we discuss the results and future directions for the work.

Chapter 2

Methods

2.1 DSD

The diffusion state distance (DSD), a "metric based on a graph diffusion property, designed to capture finer-grained distinctions in proximity for transfer of functional annotation in PPI networks [CZP$^+$13]", from initial node i to target node j is based on a random walk of length k from i and the expected number of times j will be reached on that walk [CZP$^+$13]. Figure 2.1 provides an example of how DSD differs from the traditional idea of graph distance by looking at two nodes that are connected by either a hub node neighbor or a neighbor with few other neighbors.

The DSD matrix can be computed on a graph $G(V, E)$, where V is a set of n nodes, and E is the set of edges, and a walk length t. This will produce an $n \times n$ matrix, M, where M_{ij} represents the distance from i to j. If $He^t(i, j)$ is the expected number of times a walk of a fixed length t, starting from v_i will reach v_j, then $He^t(i) = (He^t(i, v_1), ...He^t(i, v_n)$ is an array of the expected number of times each node will be reached in that walk of length t. As noted in Cao et al. as the length of the random walk goes to infinity, the DSD metric converges, meaning it can be defined without referring to t and $He(i)$ can be used in place of $He^t(i)$. Thus if the distance between two nodes is the L_1 norm of their respective expectation arrays, then $M_{ij} = ||He(i) - He(j)||_1$ [CZP$^+$13].

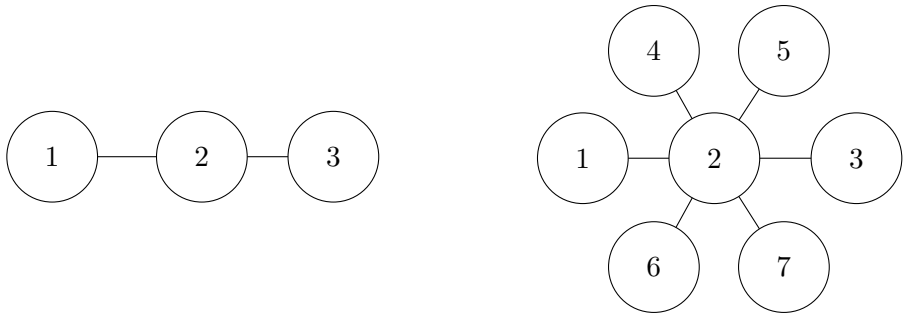

Figure 2.1: In the first example, nodes 1 and 3 are connected by node 2 which has no other neighbors. This results in a distance 2, and for DSD, with $k = 2$ one would expect to reach 3 on half of the random walks from 1. In the second case, we see that hub node 2 connects many other nodes together, which does not suggest much connection between all the other nodes despite the fact they all have a distance of 2 from each other. On random walks of length 2 starting at 1, one would expect to reach any of the non hub nodes 1 out of 6 times, so the DSD between 1 and 3 is now much lower.

2.1.1 DSD for Function Prediction

Once the DSD matrix for a PPI network has been computed, it can be used for protein function prediction. This can be done in several ways, but the most intuitive approach is a majority vote based on the k nearest neighbors. Each of a target protein's neighbors votes once for each of their own functional labels, and the labels are ordered by the number of votes received and assigned to the target protein.

2.2 MUNK

In Fan et al. [FCF+19], Multi-Species Network Kernel (MUNK) is proposed as a method of cross species transfer of protein function information. This transfer from source to target species is enabled by an embedding of both species PPI networks into the same vector space through a set of landmark genes. These landmark are chosen as homologous genes between species, identified by BLAST.

MUNK requires a graph $G_1 = (V_1, E_1)$, where V_1 represents the n proteins in the source species, and E_1 is the list of edges in the PPI network, and a similar graph $G_2 = (V_2, E_2)$, where $|V_1| = m$, for the target species. Additionally, a set of land-

mark genes for each species is needed, $L_1 \subset V_1$ and $L_2 \subset V_2$, such that $|L_1| = |L_2|$, such that L_{1n} is the landmark in species one that corresponds with the landmark L_{2n} in the other species. Their selection will be discussed in 3.1.3.1.

The first stage of MUNK is the construction of a kernel similarity matrix for each graph, $D_1 \in \mathbb{R}^{n \times n}$ and $D_2 \in \mathbb{R}^{m \times m}$ (the upper and lower matrices of Fig 2.2). For each node in the source species, a vector representation, C_{1i}, in the Reproducing Kernel Hilbert Space (RKHB), can be constructed such that $D_1 = C_1 C_1^\intercal$. Let C_{1L} be the subset of the rows for the landmark genes in the source species, and the subset of rows in the target species be D_{2L}. From this, we would like to generate a vector representation of nodes from V_2 in the same embedding RKHB space. To do this we have the similarity scores in D_{2L} apply to their corresponding landmarks in the source species. Under this assumption, the vector embeddings of the target proteins, $\hat{C}_2$, satisfies the equation $D_{2L} = C_{1L}\hat{C}_2$. This leads to the linear system:

$$\hat{C}_2 = \hat{C}_{1L}^\dagger D_{2L} + (I - \hat{C}_{1L}^\dagger C_{1L})W,$$

with $\hat{C}_{1L}^\dagger$ being the Moore-Penrose pseudoinverse of C_{1L}, and W being an arbitrary matrix. We can choose the solution of $W = 0$, leading to a $\hat{C}_2$ with minimum norm. This $\hat{C}_2$ allows us to compute a similarity score between the nodes in each network, $D_{12} = C_1 \hat{C}_2^\intercal$ (the red matrix in Fig 2.2), since $\hat{C}_2$ can be seen as an embedding of the nodes in G_2 into the same space as the nodes from G_1.

During trials, we found that instead of using a similarity matrix, a distance matrix could be used for increased performance, as long as further down the pathway, changes were made to account for this inversion. In this case, it meant that when choosing the k nearest neighbors, we chose the smallest distances instead of the greatest similarities. Therefore, in our experiments, the similarity matrices D_1 and D_2, were instead replaced with the distance based DSD matrices for these networks.

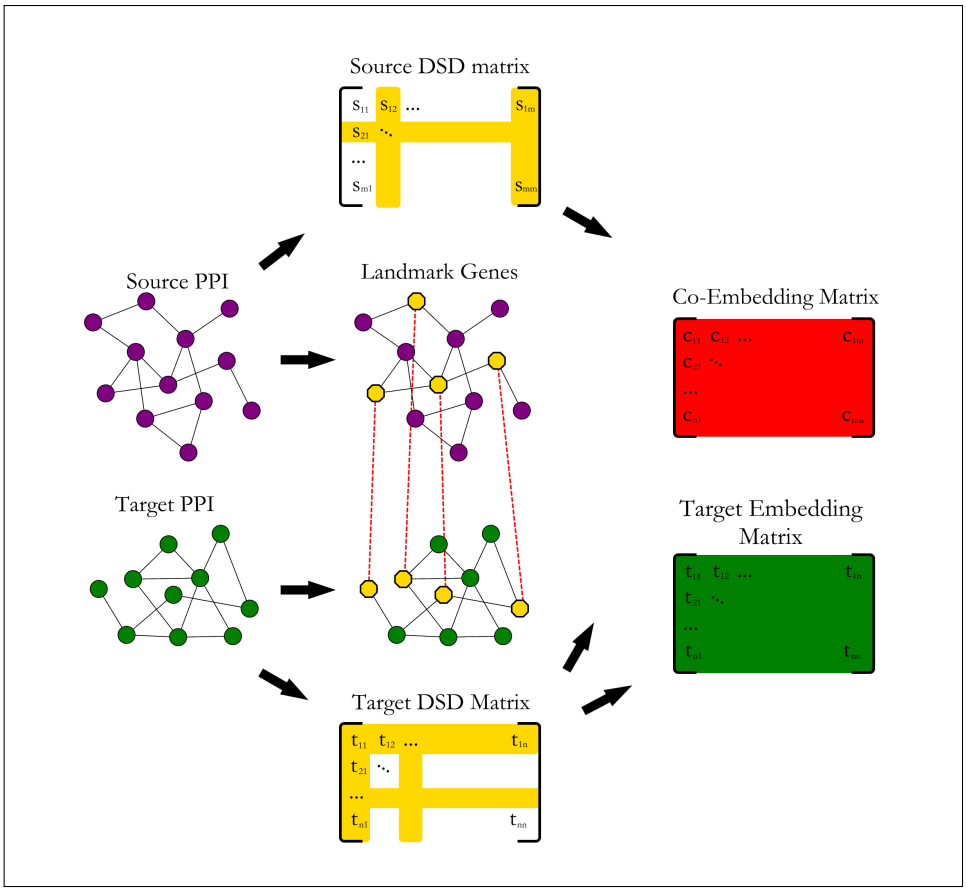

Figure 2.2: An overview of the MUNK and MUNDO methods. A PPI network and landmark genes are provided for both source and target species. First a similarity matrix, DSD, is generated for each network. The source DSD is used to produce the RKHB space, and the two networks are co-embedded yielding the red matrix. Additionally, in MUNDO the green target embedding is generated to use for majority vote amongst the neighbors in the original network.

2.2.1 Function Prediction With MUNK

Once the co-embedding of two species is generated, a majority vote algorithm can be used to assign a functional label to a target protein in the new space. Given a number of neighbors kB, and a target protein t, a majority vote will be conducted from the kB closest proteins to t in the co-embedding space.

Each of the kB nearest neighbor of t, votes once for all its proteins, then all the votes are tallied and the labels are ordered by their number of votes.

2.3 MUNDO

The MUNDO algorithm is a continuation of the ideas presented in the MUNK paper. While MUNK produces a multi species embedding of the two sets of proteins, and performs a majority vote in this new space, MUNDO uses both a co-embedding and a single species embedding for a weighted majority vote algorithm in its prediction of protein function.

The original distance matrix for each species can be generated in many manners, but to be consistent we chose to use the DSD matrix for both MUNK and MUNDO. One downside of MUNK is that the co-embedding process is a linear transformation. This results in an inability to capture nonlinear relationships between the two species embeddings. In order to create a more robust model that can deal with this, an attention based transformer can be used to generate an embedding based on the landmark gene similarity scores similar to the optimization goals of the MUNK co-embedding.

2.3.1 Function Prediction With MUNDO

Function prediction with the MUNDO algorithm is a combination of the other two methods. The DSD approach of a majority vote by the nearest neighbors in the single species embedding and the MUNK approach of a majority vote in the co-embedding space are both weighted and combined for a more accurate prediction of the true GO labels for each protein.

This voting process is similar to the MUNK voting, however in addition to the votes in the co-embedding space with a weight of 1, all the kA neighbors in the single species embedding get to vote with weight α. The total votes are again tallied, and an ordered list of predictions is produced. The selection of parameters kA and α are described more in Section 3.3

2.4 DSCRIPT/Topsy-Turvy

A trained DSCRIPT model can be used for link prediction between two protein sequences. When the model is given two protein sequences, it will produce a probability, that can be interpreted as the how likely the model believes those two proteins will interact [SSCB21b]. During this prediction step, the model will first create the language model embedding, so either the sequences or the pre-generated embeddings can be provided for prediction.

A DSCRIPT model can be trained by providing a set of positive or negative protein interactions and the protein sequences. These interactions are typically in the form of a triple of two proteins and a Boolean marking if they interact or not. The model will then attempt to learn these interaction mechanics based on the sequences of the two proteins alone.

2.4.1 Topsy-Turvy

The DSCRIPT training relies solely on pairwise interaction between two sequences, but global network information cannot be used in this process. However, there is important information in the network that can give context to protein interactions. To leverage this information, a new version of DSCRIPT called Topsy-Turvy, was introduced to allow for this information to be taught to the model during the training phase. In training instead of just the pairs of proteins and a boolean of whether or not they interact, the whole PPI network is provided as well. This allows for a network-based prediction score to be incorporated into the loss function for the model training [SDS+22].

2.4.2 Experimental Design

Currently, one of the major limiting factors in the efficacy of function prediction algorithms is the lack of reliable PPI data. However, with the emergence of new machine learning models mentioned above, it would be possible to augment the current networks with links from predicted from the model to produce a more complete

network that when passed to the already existing function prediction algorithms will allow them to perform better. We will attempt to use Topsy-Turvy to both add and remove edges from the existing PPI networks, and compare the results of DSD, MUNK, and MUNDO on these new networks.

Chapter 3

Experiments

3.1 Data Overview

The *R. norvegicus* PPI network consists of the 10,792 unique nodes and 23,315 unique interaction edges downloaded from BioGRID version 4.2.193, excluding self-loops. Taking the largest connected component yields 9,640 nodes 22,486 edges.

The *S. cerevisiae* PPI network consists of the 6,478 unique nodes and 70,638 unique interaction edges downloaded from BioGRID version 4.2.192, excluding self-loops. Taking the largest connected component yields 6,451 nodes and 70,616 edges.

The *D. Melanogaster* PPI network consists of the 11,247 unique nodes and 42,555 unique interaction edges downloaded from BioGRID version 4.2.192, excluding self-loops. Taking the largest connected component yields 11,046 nodes and 42,411 edges.

3.1.1 GO Annotations

For consistency with the MUNDO methods, GO annotations were obtained from EMBL-EBI's UniProt GOA database, version 201, and labels with a shortest path to the root with a length of 5 or less were filtered out [ADE+21].

3.1.2 Sequence Data

S. cerevisiae (baker's yeast), *D. melanogaster* (fly), and *M. Musculus* (mouse) protein sequences were obtain from the DSCRIPT data repository at https://doi.org/10.5281/zenodo.5140612 with version number v1.0 [SSCB21a]. The baker's yeast data contains 5,664 protein sequences, the fly contained 19,310, and the mouse data contains 40,606 protein sequences.

As *R. norvegicus* (rat) sequences were not included with this data set, rat sequences were obtained from the STRING database on March 1st, 2023 [SFW$^+$15]. It contains 22,763 protein sequences.

3.1.2.1 Topsy-Turvy model

A Topsy-Turvy model, pretrained on human sequences, was additionally acquired from the same location with version V1.0 [SSCB21a]. This was done to avoid the long training time and computational resources required to generate a reliable model.

3.1.3 Protein ID conversion

To convert between Swiss-Prot and STRING IDs for all the proteins, the UniProt ID Mapping was used [Con22]. A query from Swiss-Prot to STRING ID was made. This returned a 1-to-1 mapping between the two set, for the IDs that existed in both databases. Some proteins did not have a label in both sets, in which case, they were not used for the PPI augmentation step.

3.1.3.1 Landmark Selection

In order for the information in the source network to be helpful for labeling in the target network, the landmarks used to connect them have to be chosen properly. For each pair of species, a BLASTP query was performed between the protein sequences of each species resulting in a set of bit-scores. These were then filtered to contain only reciprocal matches and normalized from 0 to 1. From these matches, the top 500 were chosen as landmarks between the two species.

3.1.4 Data Split For Prediction

Trials were performed for each of the following hierarchies biological process, cellular component, and molecular function. As the number of terms and meaning of each hierarchy was vastly different, they were scored individually, so each experiment resulted in 3 hierarchy scores.

Within each trial, the data was split into 5 folds for the prediction step. The labels of one fold at a time were removed, and the remaining 4 folds were used to make predictions on the removed labels. The predictions were then scored and averaged together for the overall score of the trial.

3.2 Evaluation Metrics

Each trial was evaluated using two separate metrics: Top 1 Accuracy, and F1 Max score. Other studies have implemented metrics to focus on different aspects of the problem [JOC+16]. For example, the original MUNDO paper brings up a metric called Resnik similarity, which accounts for the hierarchical nature of GO labels [ADE+21]. However, since the trials we performed were already divided into the three main GO hierarchies and the terms were filtered beyond a certain depth, there was less of a need to focus on this aspect here. Since the main goal was to investigate if the changes in the data improved the performance of already existing methods, looking at the precision and recall, which are summarized by the F1 score, can serve as effective indicator for model performance.

3.2.1 Top k Accuracy

The simplest way to evaluate the labels predicted for a protein is to take the k predictions that the method was the most confident in, and check to if they are in the true set of labels for that protein. Choosing $k = 1$ allows you to determine how accurate the model's most confident functional label for a given protein is.

3.2.2 F1 Max Score

While accuracy focuses on the model's best guess performance, F measures looks at the overall performance. The ideal model would be able to correctly predict all of the labels for a protein while predicting no additional labels. This idea is captured by two metrics: precision, and recall. If P_i is the set labels the model predicts for the i^{th} protein, and T_i is the set of true labels then,

$$\text{precision}_i = \frac{|P_i \cap T_i|}{|P_i|},$$

$$\text{recall}_i = \frac{|P_i \cap T_i|}{|T_i|}.$$

These metrics can be averaged over all the proteins to produce an average precision and recall score.

The F_β score is a way of weighing these two values together, where β is a weight applied to the precision. In this case, neither metric is more important than the other so the F_1 score is used, and is defined as

$$F_1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

Since the model outputs a confidence level for each label, a threshold is used to classify each prediction as either a positive or a negative. The F_1 score can be calculated at varying threshold levels, and the max taken to produce the F_1^* metric.

3.3 Experimental Setup

Several experiments were carried out to determine the most effective way to use the link prediction model to improve the performance of the other methods. For each test, three different methods would be compared against each other. These methods were DSD based knn, MUNK, and MUNDO.

3.3.0.1 Other Parameters

Once the landmarks are determined, 3 other parameters must be chosen for the methods: kA, kB, and α. kA determines the number of neighbors in the target species that will be used in the vote. This parameter is used by both DSD and MUNDO. kB, a parameter used by both MUNK and MUNDO, determines the number of neighbors in the combined embedding. Lastly, α is a weight used to determine how much neighbors in the combine embedding influence the vote. Based on previous results $kA = 10,\ kB = 20,\ \alpha = 1.5$ were the recommended settings for these parameters [ADE+21]. Other tested settings show a marginal decrease in performance, and since this investigation focuses more on the quality of data, these parameter spaces will not be explored. Lastly, for each experiment involving Topsy-Turvy, an additional threshold parameter, t, will be used to determine the likelihood cutoff that above which the model is confident that two proteins will interact or not. The threshold chosen for additional links was set at 0.5. A higher threshold was considered to lower the number of noisy links that were introduced, however raising the threshold significantly lowered the number of links added, such that it would be difficult to tell if the changes in performance were from random chance or the few added links. This threshold was chosen for the addition of links, however, various thresholds were testing during the removal of the links described more in section 3.3.3

Since we are testing DSD as a method for function prediction, we will also be using it as the kernel matrix for both MUNK and MUNDO as the local embeddings and similarity scores. No modifications of the process for generating the co-embeddings for MUNK were needed. The transformer used to generate the MUNDO co-embedding was a two layer multi-headed attention network.

For each experiment, a target species was chosen for GO prediction, and a source species was chosen as the basis for the co-embedding. *S. cerevisiae* was used as a target species to investigate how beneficial the methods would be in an already well

annotated species, while *R. norvegicus* was used as an example of a less annotated species. The *D. melanogaster* and *M. Musculus* were used as source species.

3.3.1 Baseline

To establish a control from which to compare, the first step was to determine how well the 3 methods performed on the initial datasets. We ran the 3 methods on each of the target species against the 3 other species.

For each species pairing experiment, a PPI network and GO labels for each species are the input. First, the DSD matrix was constructed for each species. This was then used directly for function prediction with DSD based knn and was later used to construct the co-embedding space for both MUNK and MUNDO.

Following the methods detailed in Chapter 2, we then performed functional label prediction for a set of proteins using each method. The results were then evaluated with the metrics described in Section 3.2.

3.3.2 Additional Links

With a baseline established, it is possible to investigate the effects of adding additional links suggested by the TT model. However, as a pairwise comparison of every protein with the model would be computationally expensive, some approach must be taken to limit the number of link predictions. One reasonable approach would be to take advantage of the DSD based similarity scores computed in the baseline experiment.

By only querying the model if some protein p will interact with its k nearest neighbors in DSD space, the total of number of queries will be greatly reduced while still having many relevant queries, as we expect that proteins that are closer to each other will be more likely to interact than proteins far from each other. We tested $k = 5, 10$ and 15.

This will produce a series of predictions that can be added to the previously known interactions, augmenting the existing PPI network. This network can then be fed into the discussed methods following the baseline comparison.

3.3.3 Removing Links

It is possible that some of the links in the original PPI network do not actually reflect real interactions between proteins and are thus adding noise to the network. Therefore, instead of being used to add additional links to the data, the TT model can also be used to prune links.

In a similar manner to before the number of queries will be limited based on DSD metrics. Since each node has relatively few neighbors, the model will run a query on all current neighbor pairs. For each of these pairings, the model will again produce a likelihood score, and based on a varying threshold t, these links will be removed if they are less likely than t. The pruned networks will then be passed into the 3 baseline methods and compared. For this experiment, threshold values of $0.1, 0.25$, and 0.5 were used.

Chapter 4

Results

4.1 Baseline Results

The *R. norvegicus* is an ideal target species, as its GO labels are not as well annotated as other species such as *S. cerevisiae*. In the baseline results (Figure 4.1), this can be seen as a slight improvement from DSD to MUNDO across both metrics and all hierarchies. In fact, we see the greatest variation between the three hierarchies. The cellular component scores the best, as there are fewer labels in this category, and as it relates to physical location in the cell, interaction data is extremely meaningful for this. As a result of the performance disparities between hierarchies, averaging across all of them might lead to inaccurately reporting the performance, so the scores will be reported for each hierarchy separately.

Comparing these results with the *S. cerevisiae* results, we can see that the performance increases dramatically. However, this is simply because baker's yeast GO labels are well annotated, so any algorithm will perform better. Using baker's yeast as a target species will let us examine whether removing noisy links might be beneficial.

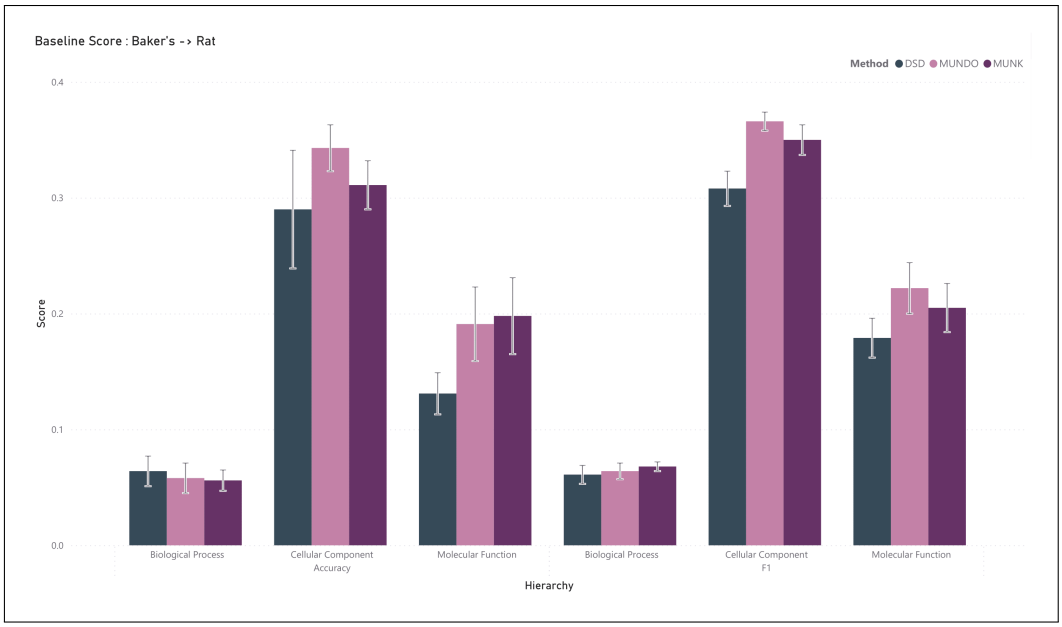

Figure 4.1: Baseline mean accuracy and F1-max reported with standard deviation for *S.Cerevisiae* → *R. norvegicus* on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

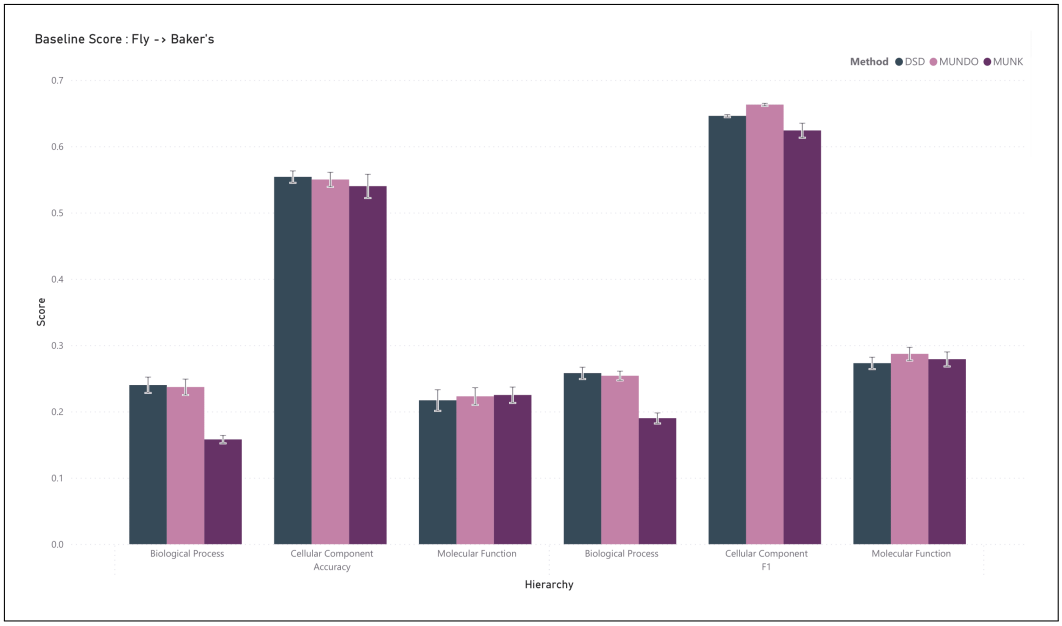

Figure 4.2: Baseline mean accuracy and F1-max reported with standard deviation for *D.Melanogaster* → *S.Cerevisiae* on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

4.2 Addition Results

With a baseline established, the next step was to investigate the addition of undiscovered protein interactions into the network. Since the addition of links can introduce more annotated neighbors, we believed this would help more in lesser annotated target species, so *R. norvegicus* was chosen as the target species.

While the changes in performance are marginal, a trend seems to hold over both DSD and MUNDO. The addition of more links universally reduces the performance across all categories. However, with $k = 10$, there appears to be an increase in performance above the other two threshold. If the links added were just random noise, then the performance should decrease as the number of links added increases, but the jump at $k = 10$, suggests that there is useful information added in these links, but there is also a lot of noise.

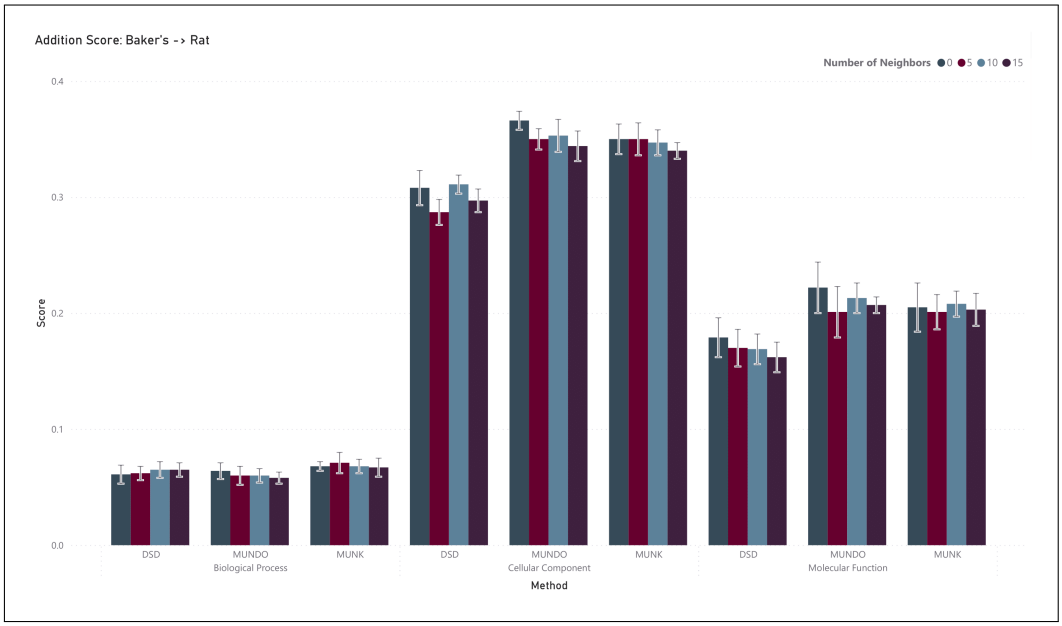

Figure 4.3: *S.Cerevisiae* $\rightarrow$ *R. norvegicus* with additional links predicted from the 5, 10, and 15 nearest neighbors compared with the baseline. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

4.3 Removal Results

The next experiment focused on pruning possibly noisy edges from a new annotated species (*S.Cerevisiae*), as these edges represented interactions that do not occur in reality. The first stategy used was to query edges with a similar strategy used in the addition set by looking at DSD neighbors. This removal strategy did see a slight overall increase in performance (Figure 4.4), however, the overlap between the k nearest DSD neighbors and the PPI neighbors was small, so the next was to investigate a larger selection of existing edges.

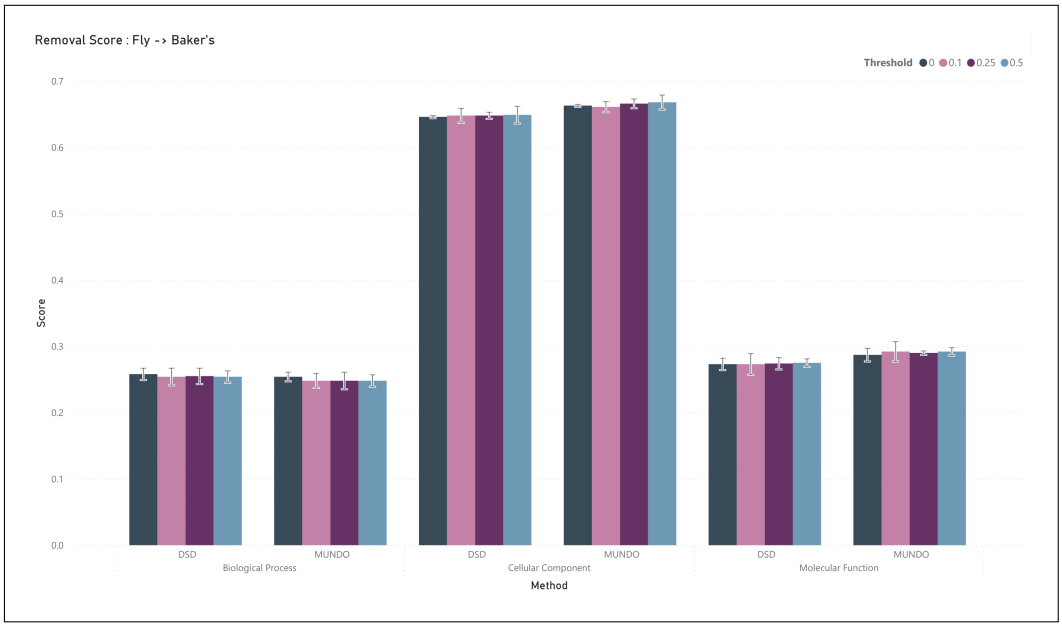

Figure 4.4: *D.Melanogaster* → *S.Cerevisiae* with links predicted from the 15 nearest neighbors were removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

4.3.1 Extensive Removal Results

Since the total number of edges was below 100,000 for all the species, it was possible to use the TT model to query all the existing edges. From this, various thresholds were used to remove existing edges from the data. While retesting with *D.Melanogaster* → *S.Cerevisiae* (Figure 4.5) and *S.Cerevisiae* → *R. norvegicus* (Fig-

ure 4.6), as the threshold increased, there was clear decrease in performance.

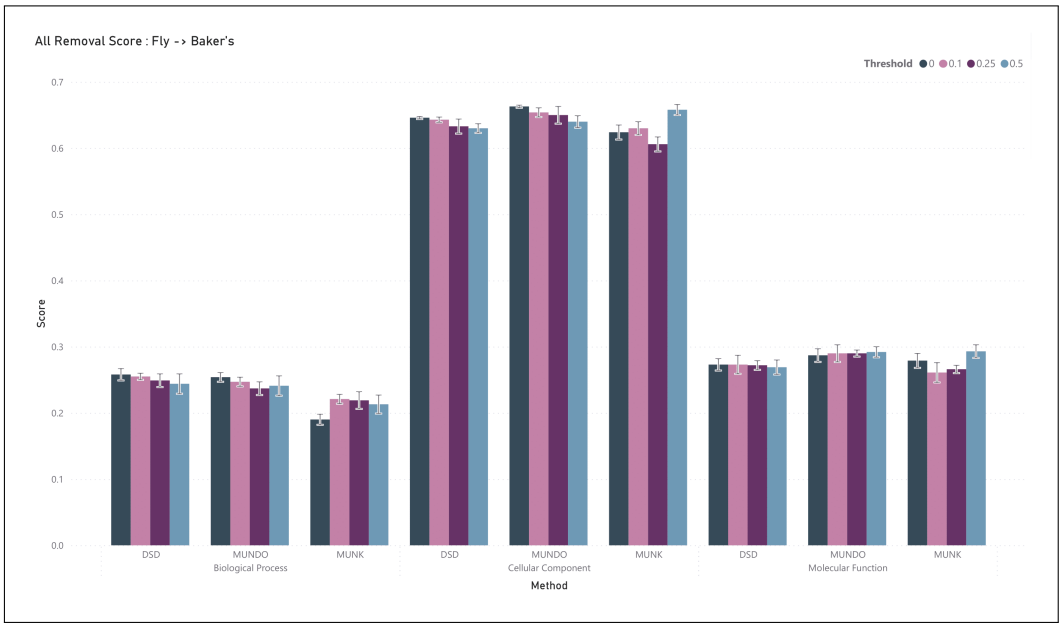

Figure 4.5: *D.Melanogaster* → *S.Cerevisiae* with links predicted from the all direct neighbors were removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

4.4 Summary

None of the methods were able to augment the data in a way that was beneficial to the performance of the function prediction methods. This is possibly because the strategy or where the machine learning model was applied was ineffective. Alternatively, the model itself was producing noise in its results that counteracted any of the gains from the true labels it predicted.

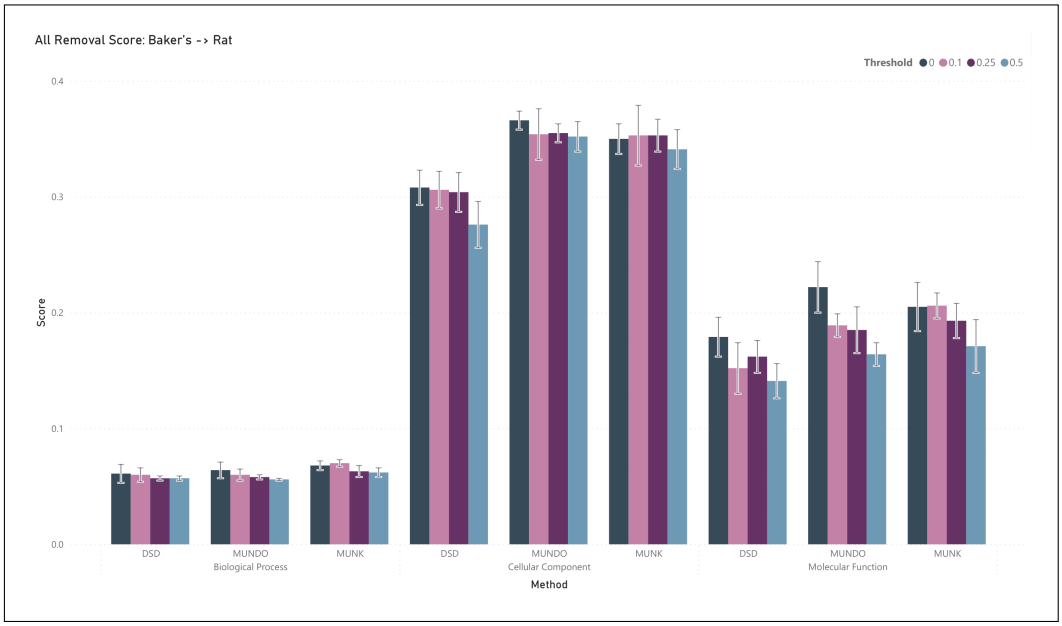

Figure 4.6: *S.Cerevisiae* $\rightarrow$ *R. norvegicus* with links predicted from the all direct neighbors removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Chapter 5

Conclusion

Based on the results observed across all the trials, the use of machine learning models to augment the protein-protein interaction networks does not currently increase the performance of protein function prediction. It appears that currently these models produce too much noise to be beneficial. However, since these models are always improving, with time, the amount of noise in their outputs will reduce, and this approach could become more effective. Furthermore, this research was not an extensive look at all currently possible options. Improvements could be made to a few aspects of the research that might have reduced the noise even with current iterations of the machine learning models.

There are some aspects of the implementation that could have contributed to the poor performance of both the addition and subtraction experiments. First, during the addition step, a threshold was chosen of 0.5. This may have been too low of a choice and could have contributed to the amount of noise added to the system. For each species about 16% of the predicted likelihoods were 0.5 or greater but choosing 0.6 would have lowered this to 7% and 2% at 0.9. A stronger threshold would have lowered the amount of noise introduced into the system but would also mean that the proportion of links in the final network that were added links would be less significant.

There is also a concern of the nature of link that Topsy-Turvy predicts compared to

the type of links that are contained within the PPI networks. For example, if there were a complex where each protein directly interacted with only the 2 proteins next to it, forming a ring of interactions, then in the original PPI network that would be displayed as a complete graph as they are all part of the same structure. However, as a Topsy-Turvy model only looks at the direct interactions, it would attempt to remove the non-direct interactions leaving only the ring. These edges contain valuable information and should not be removed but were mostly likely trimmed from the network resulting in worse performance.

One possible improvement could come from the use of specifically trained TT models for each species. Here only a model trained on humans was used for all species, however, it's possible that a model specifically trained for on each organism, or perhaps a more closely related species, would be able to produce better results at all stages. Sadly, the size of the networks used might be too small to effectively train a model on just that organism.

One drawback with the approaches used in these experiments is that the information captured by DSD is already contained within the network passed in during the training of Topsy-Turvy [SDS$^+$22]. As a result, by finding the closest DSD neighbors, we are not necessarily presenting the Topsy-Turvy model with new information. In order to combat this overlap of information, the search could be expanded to look at a broader range of neighbors. However, some limits would have to be placed to reduce the number of edges added through this method, as it would run the risk of introducing more noise.

Appendix A

Tabulated Results

Target: Rat Source: Baker's

Molecular Function	DSD	MUNK	MUNDO
Accuracy	0.131 +- 0.018	0.198 +- 0.033	0.191 +- 0.032
F1*	0.179 +- 0.017	0.205 +- 0.021	0.222 +- 0.022
Biological Process			
Accuracy	0.064 +- 0.013	0.056 +- 0.009	0.058 +- 0.013
F1*	0.061 +- 0.008	0.068 +- 0.004	0.064 +- 0.007
Cellular Component			
Accuracy	0.290 +- 0.051	0.311 +- 0.021	0.343 +- 0.020
F1*	0.308 +- 0.015	0.350 +- 0.013	0.366 +- 0.008

Table A.1: Baseline mean accuracy and F1-max reported with standard deviation for *S.Cerevisiae* $\rightarrow$ *R. norvegicus* on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Target: Baker's Source: Fly

Molecular Function	DSD	MUNK	MUNDO
Accuracy	0.217 +- 0.016	0.225 +- 0.012	0.223 +- 0.013
F1*	0.273 +- 0.009	0.279 +- 0.011	0.287 +- 0.010
Biological Process			
Accuracy	0.240 +- 0.012	0.158 +- 0.006	0.237 +- 0.012
F1*	0.258 +- 0.009	0.190 +- 0.008	0.254 +- 0.007
Cellular Component			
Accuracy	0.554 +- 0.009	0.540 +- 0.018	0.550 +- 0.011
F1*	0.646 +- 0.002	0.624 +- 0.011	0.663 +- 0.002

Table A.2: Baseline mean accuracy and F1-max reported with standard deviation for *D.Melanogaster* $\rightarrow$ *S.Cerevisiae* on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Target: Rat Source: Baker's Addition

Molecular Function	DSD	MUNK	MUNDO
(Accuracy) Baseline	0.131 +- 0.018	0.198 +- 0.033	0.191 +- 0.032
k=5	0.126 +- 0.027	0.190 +- 0.017	0.165 +- 0.018
k=10	0.131 +- 0.026	0.178 +- 0.019	0.186 +- 0.014
k=15	0.121 +- 0.013	0.196 +- 0.017	0.190 +- 0.013
(F1*) Baseline	0.179 +- 0.017	0.205 +- 0.021	0.222 +- 0.022
k=5	0.170 +- 0.016	0.201 +- 0.015	0.201 +- 0.022
k=10	0.169 +- 0.013	0.208 +- 0.011	0.213 +- 0.013
k=15	0.162 +- 0.013	0.203 +- 0.014	0.207 +- 0.007
Biological Process			
(Accuracy) Baseline	0.064 +- 0.013	0.056 +- 0.009	0.058 +- 0.013
k=5	0.059 +- 0.008	0.053 +- 0.010	0.047 +- 0.016
k=10	0.064 +- 0.011	0.040 +- 0.007	0.033 +- 0.006
k=15	0.064 +- 0.011	0.038 +- 0.013	0.015 +- 0.004
(F1*) Baseline	0.061 +- 0.008	0.068 +- 0.004	0.064 +- 0.007
k=5	0.062 +- 0.006	0.071 +- 0.009	0.060 +- 0.008
k=10	0.065 +- 0.007	0.068 +- 0.006	0.060 +- 0.006
k=15	0.065 +- 0.006	0.067 +- 0.008	0.058 +- 0.005
Cellular Component			
(Accuracy) Baseline	0.290 +- 0.051	0.311 +- 0.021	0.343 +- 0.020
k=5	0.256 +- 0.031	0.298 +- 0.022	0.316 +- 0.016
k=10	0.272 +- 0.024	0.286 +- 0.014	0.320 +- 0.028
k=15	0.246 +- 0.021	0.270 +- 0.026	0.304 +- 0.024
(F1*) Baseline	0.308 +- 0.015	0.350 +- 0.013	0.366 +- 0.008
k=5	0.287 +- 0.011	0.350 +- 0.014	0.350 +- 0.009
k=10	0.311 +- 0.008	0.347 +- 0.011	0.353 +- 0.014
k=15	0.297 +- 0.010	0.340 +- 0.007	0.344 +- 0.013

Table A.3: *S.Cerevisiae* → *R. norvegicus* with additional links predicted from the 5, 10, and 15 nearest neighbors compared with the baseline. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Target: Baker's Source: Fly

Molecular Function	DSD	MUNDO
(Accuracy) Baseline	0.217 +- 0.016	0.223 +- 0.013
T=0.1	0.216 +- 0.016	0.228 +- 0.023
T=0.25	0.212 +- 0.013	0.225 +- 0.005
T=0.5	0.217 +- 0.013	0.233 +- 0.021
(F1) Baseline	0.273 +- 0.009	0.204 +- 0.010
T=0.1	0.273 +- 0.016	0.292 +- 0.015
T=0.25	0.274 +- 0.009	0.290 +- 0.003
T=0.5	0.275 +- 0.006	0.292 +- 0.006
Biological Process		
(Accuracy) Baseline	0.240 +- 0.012	0.237 +- 0.012
T=0.1	0.239 +- 0.016	0.235 +- 0.014
T=0.25	0.239 +- 0.013	0.235 +- 0.015
T=0.5	0.241 +- 0.010	0.236 +- 0.011
(F1) Baseline	0.258 +- 0.009	0.254 +- 0.007
T=0.1	0.254 +- 0.013	0.248 +- 0.011
T=0.25	0.255 +- 0.012	0.248 +- 0.013
T=0.5	0.254 +- 0.009	0.248 +- 0.009
Cellular Component		
(Accuracy) Baseline	0.554 +- 0.009	0.550 +- 0.011
T=0.1	0.554 +- 0.008	0.543 +- 0.010
T=0.25	0.557 +- 0.005	0.550 +- 0.011
T=0.5	0.549 +- 0.011	0.550 +- 0.011
(F1) Baseline	0.646 +- 0.002	0.663 +- 0.002
T=0.1	0.648 +- 0.011	0.661 +- 0.008
T=0.25	0.648 +- 0.005	0.666 +- 0.007
T=0.5	0.649 +- 0.013	0.668 +- 0.011

Table A.4: *D.Melanogaster* $\rightarrow$ *S.Cerevisiae* with links predicted from the 15 nearest neighbors were removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Target: Baker's Source: Fly

Molecular Function	DSD	MUNK	MUNDO
(Accuracy) Baseline	0.217 +- 0.016	0.225 +- 0.012	0.223 +- 0.013
T=0.1	0.212 +- 0.015	0.221 +- 0.021	0.237 +- 0.018
T=0.25	0.214 +- 0.019	0.214 +- 0.008	0.239 +- 0.010
T=0.5	0.217 +- 0.013	0.237 +- 0.010	0.232 +- 0.009
(F1*) Baseline	0.273 +- 0.009	0.279 +- 0.011	0.204 +- 0.010
T=0.1	0.273 +- 0.014	0.261 +- 0.015	0.290 +- 0.013
T=0.25	0.272 +- 0.007	0.266 +- 0.006	0.290 +- 0.005
T=0.5	0.269 +- 0.011	0.293 +- 0.010	0.292 +- 0.008
Biological Process			
(Accuracy) Baseline	0.240 +- 0.012	0.158 +- 0.006	0.237 +- 0.012
T=0.1	0.240 +- 0.010	0.219 +- 0.010	0.232 +- 0.017
T=0.25	0.238 +- 0.017	0.220 +- 0.017	0.225 +- 0.015
T=0.5	0.233 +- 0.014	0.206 +- 0.012	0.226 +- 0.014
(F1*) Baseline	0.258 +- 0.009	0.190 +- 0.008	0.254 +- 0.007
T=0.1	0.255 +- 0.005	0.221 +- 0.007	0.247 +- 0.007
T=0.25	0.249 +- 0.010	0.219 +- 0.013	0.237 +- 0.010
T=0.5	0.244 +- 0.015	0.213 +- 0.014	0.241 +- 0.015
Cellular Component			
(Accuracy) Baseline	0.554 +- 0.009	0.540 +- 0.018	0.550 +- 0.011
T=0.1	0.546 +- 0.005	0.498 +- 0.020	0.549 +- 0.010
T=0.25	0.540 +- 0.016	0.526 +- 0.014	0.542 +- 0.018
T=0.5	0.541 +- 0.012	0.538 +- 0.014	0.544 +- 0.014
(F1*) Baseline	0.646 +- 0.002	0.624 +- 0.011	0.663 +- 0.002
T=0.1	0.643 +- 0.004	0.630 +- 0.010	0.654 +- 0.007
T=0.25	0.633 +- 0.011	0.606 +- 0.011	0.650 +- 0.013
T=0.5	0.630 +- 0.007	0.658 +- 0.008	0.640 +- 0.009

Table A.5: *D.Melanogaster* $\rightarrow$ *S.Cerevisiae* with links predicted from the all direct neighbors were removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Target: Rat Source: Baker's Subtraction

Molecular Function	DSD	MUNK	MUNDO
(Accuracy) Baseline	0.131 +- 0.018	0.198 +- 0.033	0.191 +- 0.032
T=0.1	0.104 +- 0.032	0.151 +- 0.016	0.167 +- 0.022
T=0.25	0.121 +- 0.020	0.160 +- 0.010	0.153 +- 0.034
T=0.5	0.115 +- 0.016	0.154 +- 0.031	0.137 +- 0.017
(F1*) Baseline	0.179 +- 0.017	0.205 +- 0.021	0.222 +- 0.022
T=0.1	0.152 +- 0.022	0.206 +- 0.011	0.189 +- 0.018
T=0.25	0.162 +- 0.014	0.193 +- 0.015	0.185 +- 0.021
T=0.5	0.141 +- 0.015	0.171 +- 0.023	0.164 +- 0.012
Biological Process			
(Accuracy) Baseline	0.064 +- 0.013	0.056 +- 0.009	0.058 +- 0.013
T=0.1	0.067 +- 0.009	0.061 +- 0.013	0.053 +- 0.007
T=0.25	0.059 +- 0.008	0.055 +- 0.015	0.058 +- 0.008
T=0.5	0.051 +- 0.008	0.048 +- 0.008	0.038 +- 0.007
(F1*) Baseline	0.061 +- 0.008	0.068 +- 0.004	0.064 +- 0.007
T=0.1	0.060 +- 0.006	0.070 +- 0.003	0.060 +- 0.005
T=0.25	0.057 +- 0.002	0.063 +- 0.005	0.058 +- 0.002
T=0.5	0.057 +- 0.002	0.062 +- 0.004	0.056 +- 0.001
Cellular Component			
(Accuracy) Baseline	0.290 +- 0.051	0.311 +- 0.021	0.343 +- 0.020
T=0.1	0.293 +- 0.027	0.307 +- 0.027	0.304 +- 0.026
T=0.25	0.259 +- 0.027	0.278 +- 0.027	0.265 +- 0.031
T=0.5	0.231 +- 0.012	0.283 +- 0.027	0.287 +- 0.028
(F1*) Baseline	0.308 +- 0.015	0.350 +- 0.013	0.366 +- 0.008
T=0.1	0.306 +- 0.016	0.353 +- 0.026	0.354 +- 0.022
T=0.25	0.304 +- 0.017	0.353 +- 0.014	0.355 +- 0.008
T=0.5	0.276 +- 0.020	0.341 +- 0.017	0.352 +- 0.013

Table A.6: *S.Cerevisiae* → *R. norvegicus* with links predicted from the all direct neighbors removed if their likelihood was below the given threshold. Mean accuracy and F1-max reported with standard deviation on 5-fold validation. Scores reported for DSD, MUNK, and MUNDO across the 3 GO hierarchies.

Bibliography

[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[ADE+21] Victor Arsenescu, Kapil Devkota, Mert Erden, Polina Shpilker, Matthew Werenski, and Lenore J Cowen. MUNDO: protein function prediction embedded in a multispecies world. *Bioinformatics Advances*, 2(1), 09 2021. vbab025.

[CJZ+19] Muhao Chen, Chelsea J-T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314, 2019.

[Con22] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022.

[CZP+13] Mengfei Cao, Hao Zhang, Jisoo Park, Noah M Daniels, Mark E Crovella, Lenore J Cowen, and Benjamin Hescott. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one*, 8(10):e76339, 2013.

[DMC20] Kapil Devkota, James M Murphy, and Lenore J Cowen. Glide: combining local methods and diffusion state embeddings to predict miss-

ing interactions in biological networks. *Bioinformatics*, 36(Supplement_1):i464–i473, 2020.

[FCF+19] Jason Fan, Anthony Cannistra, Inbar Fried, Tim Lim, Thomas Schaffner, Mark Crovella, Benjamin Hescott, and Mark DM Leiserson. Functional protein representations from biological networks enable diverse cross-species inference. *Nucleic acids research*, 47(9):e51–e51, 2019.

[JEP+21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[JOC+16] Yuxiang Jiang, Tal Ronnen Oron, Wyatt T Clark, Asma R Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S Funk, Indika Kahanda, Karin M Verspoor, Asa Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):1–19, 2016.

[KPA+12] Gavin CKW Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E Orchard. Analyzing protein–protein interaction networks. *Journal of proteome research*, 11(4):2014–2031, 2012.

[SDS+22] Rohit Singh, Kapil Devkota, Samuel Sledzieski, Bonnie Berger, and Lenore Cowen. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics*, 38(Supplement$_1$) : $i264 - -i272, 062022$.

[SFW+15] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2015.

[SSCB21a] Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. Associated data for d-script, cell systems, July 2021.

[SSCB21b] Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982, 2021.

[YCH15] Zhu-Hong You, Keith CC Chan, and Pengwei Hu. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one*, 10(5):e0125811, 2015.