A Multi-Species Functional Embedding Integrating Sequence and Network Structure

Mark D.M. Leiserson[1], Jason Fan[1], Anthony Cannistra[2], Inbar Fried[3], Tim Lim[4], Thomas Schaffner[5], Mark Crovella[4], and Benjamin Hescott[6]

[1] Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park
[2] Department of Biology, University of Washington
[3] University of North Carolina Medical School
[4] Department of Computer Science, Boston University
[5] Department of Computer Science, Princeton University
[6] College of Computer and Information Science, Northeastern University

Abstract. A key challenge to transferring knowledge between species is that different species have fundamentally different genetic architectures. Initial computational approaches to transfer knowledge across species have relied on measures of heredity such as genetic homology, but these approaches suffer from limitations. First, only a small subset of genes have homologs, limiting the amount of knowledge that can be transferred, and second, genes change or repurpose functions, complicating the transfer of knowledge. Many approaches address this problem by expanding the notion of homology by leveraging high-throughput genomic and proteomic measurements, such as through network alignment.

In this work, we take a new approach to transferring knowledge across species by expanding the notion of homology through explicit measures of functional similarity between proteins in different species. Specifically, our kernel-based method, HANDL *(Homology Assessment across Networks using Diffusion and Landmarks)*, integrates sequence and network structure to create a *functional embedding* in which proteins from different species are embedded in the same vector space. We show that inner products in this space and the vectors themselves capture functional similarity across species, and are useful for a variety of functional tasks. We perform the first whole-genome method for predicting phenologs, generating many that were previously identified, but also predicting new phenologs supported from the biological literature. We also demonstrate the HANDL embedding captures pairwise gene function, in that gene pairs with synthetic lethal interactions are significantly separated in HANDL space, and the direction of separation is conserved across species. Software for the HANDL algorithm is available at `http://bit.ly/lrgr-handl`.

Keywords: biological networks, kernel learning, network embedding, phenologs, genetic interactions

1 Introduction

A primary challenge of research with model organisms is to transfer knowledge of genetics – i.e. a mapping of genotype to phenotype – between model organisms and humans. The main promise of researching model organisms stems from researchers' ability to measure the organisms in ways that are infeasible in humans. To realize the promise of this research, it is crucial to transfer knowledge between species – ideally, in two directions. First, discoveries in model organisms can be transferred to improve knowledge of human genetics (e.g via homology[1]). Second, knowledge of human genetics can be transferred to design better experiments in model organisms (e.g. for disease models).

More specifically, cross-species knowledge transfer can enable a wide variety of applications. First and foremost has been the large-scale annotation of protein function by transferring function annotations (e.g. from the Gene Ontology [3]). Even in the era of high-throughput genomics, addressing this problem remains valuable, as fewer than 1% of protein sequences in Uniprot have experimentally-derived functional annotations [56]. A second but less well-explored use is in predicting human disease models through 'orthologous phenotypes' (*phenologs*) [33]. McGary, et al. [33] reasoned that while conserved genes may retain their *molecular* functions across species, these may manifest as different "species-level" phenotypes, and they introduced a statistical test to identify such phenologs. Another application of cross-species knowledge transfer is for *pairwise* gene function (*genetic interactions*). Knowledge of synthetic lethal genetic interactions is crucial for the study of functional genomics and disease [39,40]. Since measuring synthetic lethal interactions in humans is currently infeasible, computationally transferring knowledge of these interactions from yeast to humans (and human cancers) has become a focus of recent research.

Nonetheless, cross-species knowledge transfer is quite challenging, since many model organisms diverged from humans millions of years ago and have fundamentally different genetic architectures. Initial computational approaches to transfer knowledge across species relied on measures of heredity such as *homology*. Using genetic heredity to characterize genes in different species is foundational to genetics and comparative genomics, as there is widespread and long-standing evidence that conserved genes tend to share the same function [27]. Consequently, methods for inferring homology are crucial for transferring knowledge, prompting researchers to develop sophisticated algorithms to infer homology from DNA or protein sequences (e.g. [41]), and protein-protein interaction (PPI) networks.

The most common method to infer homology is the Basic Local Alignment Search Tool (BLAST) which computes a sequence similarity score between pairs of proteins (or genes) in different species, and reports the statistically significant pairs as orthologs [50]. A second class of methods expands beyond sequence by using protein structure to infer remote homology with the goal of classification and function prediction [20].

Despite their widespread use, current approaches that infer homology from sequence data face several challenges, limiting the amount of knowledge that can be transferred. Structure-based methods are computationally expensive and do not currently scale to entire genomes. In many cases, only a relatively small subset of genes between species have sequence homologs. Further, as species diverge, protein functions change and are repurposed (e.g. [18]), and genetic interactions are often rewired [49,55]. Thus, more recent approaches aim to expand the notion of homology to capture both convergent (analogous) and divergent (homologous) evolutionary mechanisms.

As a result, recent methods expand beyond sequence homology by using widely-available high-throughput proteomic datasets. Many methods use genome-scale protein-protein interaction (PPI) networks, as genes with similar functions tend to have similar topology (e.g. as measured by network

[1] A pair of genes is *homologous* if the pair share a common ancestor – the pair is called *orthologous* if they are in different species [14].

propagation [12]). The comparison of PPI networks across species is well studied and commonly referred to as the network alignment problem [31,43,46,25,34,32,1,58,24,53,36,22]. The goal of network alignment is to establish a mapping between nodes in different networks. The dominant paradigm for this work is based on matching nodes; in some cases, this is done using cross-species node similarity scores in an intermediate step (e.g., [52,31,34,32,38,21]). Working in another direction, Jacunski, et al. [23] introduced the notion of *connectivity homology* to relate genes in different species by their position in corresponding PPI networks. While not the focus of their paper, they show that pairs of genes with the same or similar functions have lower connectivity homology dissimilarity scores.

1.1 Contributions

In this work, we go beyond matching-oriented network alignment, and generalize the notion of similarity scores by creating a biologically meaningful vector space into which we embed the nodes of multiple networks. Both the similarity scores derived from this embedding, as well as the node vectors themselves, are useful for a variety of tasks. To our knowledge, no existing approach has attempted to address combining both sequence similarity and topological similarity between PPI networks into a general-purpose representation such as ours.

Our method makes only two assumptions. First, it assumes that connectivity homology can be captured using a similarity function that is a *kernel,* which encompasses a broad class of useful metrics. Second, it assumes that sequence homology is known for some subset of proteins across the different species. We illustrate the method in this paper using pairs of species, but extension to simultaneous treatment of multiple species is straightforward.

We note that network alignment and other network-based homology methods do in some cases compute similarity scores internally. However, such methods construct these scores only for the purpose of node matching, limiting their applicability beyond alignment and protein function annotation. It is not obvious how to extend these methods for problems that go beyond node matching – such as transferring knowledge of *pairwise* gene function (e.g. synthetic lethality). More recently, Jacunski, et al. [23] introduced a general purpose cross-species node vector representation, but their work is focused on training classifiers using the vectors as features. Although they show good performance on these prediction tasks, their representations are derived from summary statistics, and we show the distances in their vector space are not strongly correlated with functional similarity.

In this paper, we use a *diffusion* kernel to capture functional similarity and call the resulting method Handl *(Homology Assessment across Networks using Diffusion and Landmarks)*. Diffusion kernels are a natural tool for capturing aspects of local and global network structure that correlate with functional similarity of nodes. Our method may be used with other kernels to relate proteins for other biological applications, or to assess other network properties.

We demonstrate that Handl homology scores capture functional similarity between proteins in different species, and show that Handl outperforms connectivity homology at capturing functional similarity. We show that comparisons in the Handl vector space can replace standard sequence homology for two important applications:

1. **Phenologs.** We show the first whole-genome method for predicting phenologs. The method considers sets of phenotype-associated genes from two species, and evaluates their orthology using the statistical test from [33]. In contrast to [33], we do not only consider known homologs in the statistical test; instead we consider *any* protein pairs that Handl predicts are functionally close across both genomes. A subset of our predicted phenologs match those identified in [33]; however, we also predict many new phenologs and support these new predictions from the biological literature.

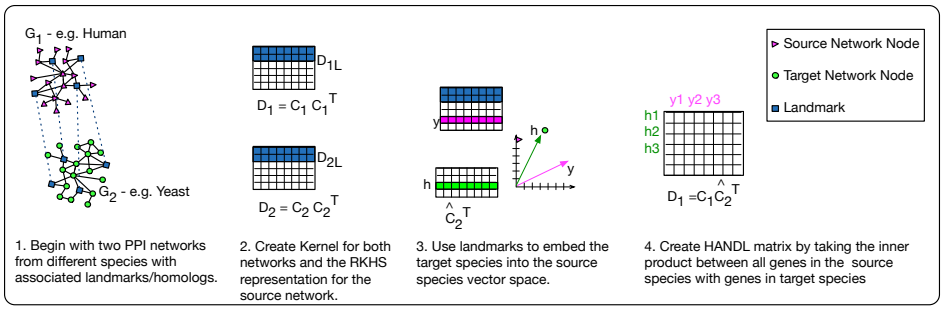


Fig. 1: Starting with a source and a target network, HANDL computes diffusion kernels for each network. The diffusion kernels are factored, resulting in a representation in which each node becomes a vector, and inner products correlate with functional similarities. HANDL then solves a linear system to create a single embedding of both vector sets, allowing for functional comparisons between proteins across the two networks.

2. **Cross-species synthetic lethality.** We demonstrate that HANDL-embeddings for *pairs* of genes in different species capture pairwise gene function in *S. cerevisiae* and *S. pombe*. We use a support vector machine to separate pairs with and without synthetic lethal interactions in one (source) species in the HANDL vector space, and show that the separating hyperplane retains significant predictive power for the SL interactions from a second (target) species projected into the common Hilbert space.

Together, these tasks encompass transferring knowledge both between model organisms, and between model organisms and humans. We conclude with a proof-of-concept experiment using HANDL to predict protein function in humans, simultaneously leveraging networks from humans and multiple model organisms.

2 Methods

The central contribution we make in this paper is to introduce a *network-based metric of functional similarity between proteins of different species.* In this section we introduce and motivate our metric.

HANDL, as shown in Figure 1, leverages properties of *kernel* functions as tools for measuring the similarity of nodes in a network. There are a large variety of kernels derived from networks [26], [16, Ch. 2], and they can model processes such as random walks, heat diffusion, PageRank, electrical resistance, and other ways of capturing node similarity in a network. Many kernels derived from networks have been applied successfully for a wide range of problems associated with biological network analysis (e.g., see review in [12]).

While kernels for individual networks are well studied, it remains an open problem to construct network-based kernels that capture the similarity of nodes in *different* networks. This is the challenge that HANDL addresses.

To do so, HANDL relies on a basic property of a kernel: any kernel is also an *inner product* in a particular space. That is, for any kernel $\kappa(\cdot,\cdot)$, there is a function $\phi(\cdot)$ that assigns vectors to nodes such that that $\kappa(i,j) = \phi(i)^T\phi(j)$. The corresponding vector space (termed the *reproducing kernel Hilbert space* (RKHS)) introduces a geometric interpretation for the kernel function. In the context of a kernel for network nodes, the RKHS representation can be thought of as an *embedding* of the network into a vector space, in a manner that captures node similarity via inner product.

Conceptually, HANDL approaches the multi-network challenge by constructing a *joint embedding* of the nodes of two (or more) networks in a *single* RKHS. The key to the HANDL method is that, within this RKHS, the similarity of nodes from *different* networks is still captured by inner product, resulting in a *multi-network kernel.* We now describe the method in detail.

2.1 Multi-Species Network Kernel Embedding

Given a *source* network G_1, a *target* network G_2 and a kernel κ, the strategy taken by HANDL is to start by embedding the nodes of the source network G_1 using the associated function, ϕ. As described above, this means that inner product between embedded G_1 nodes will capture similarity as described by κ. Next, HANDL makes use of *landmarks* – pairs of nodes in the source and target networks with identical function. The idea is to embed the nodes of the target network G_2 into the same space as the nodes of the source network G_1, such that their position in that space reproduces their similarity to the landmarks of G_2. Essentially, we posit that locating a node from G_2 based on its similarity to the set of landmarks in G_1 will also establish its similarity with the *non-landmark* nodes in G_1. As a result, HANDL creates a multi-network kernel – a single kernel function that both captures the similarity of nodes to each other in the source network G_1, and also captures similarity of nodes between the source and target networks G_1 and G_2.[2]

We now define the HANDL approach formally. Let the matrix $K \in \mathbb{R}^{n\times n}$ hold the values of the similarity function $\kappa(i,j)$ for all pairs of n proteins from a particular species. For any such kernel matrix, we can write $K = CC^T$ where C is an $n \times k$ matrix, uniquely defined up to an orthogonal transformation, with $k \leq n$. This follows from the fact that K is positive semidefinite, and means that $\kappa(i,j) = c_i^T c_j$, where c_i is the ith row of C, represented as a column vector. As explained above, the similarity between nodes v_i and v_j is exactly given by the inner product of their corresponding vectors, c_i and c_j.

Now consider a source network $G_1 = (V_1, E_1)$ and a target network $G_2 = (V_2, E_2)$, with $|V_1| = m$ and $|V_2| = n$. We assume the existence of some (small set of) nodes that correspond between G_1 and G_2. In the case where G_1 and G_2 are PPI networks, these can be orthologous proteins. For example, for orthologous proteins in different networks, it is well known that evolutionary rates differ over a wide range of magnitudes [6]. Some proteins are highly conserved, and their orthologs will have substantial sequence similarity between G_1 and G_2. Thus, there are generally a small subset of proteins that can be confidently mapped between G_1 and G_2 based on the magnitude and uniqueness of the similarity of their sequence information, which we refer to as landmarks.

We then proceed as follows. First, we construct kernel (similarity) matrices $D_1 \in \mathbb{R}^{m\times m}$ and $D_2 \in \mathbb{R}^{n\times n}$ corresponding to G_1 and G_2. Next, we construct RKHS vector representations C_1 for nodes in the source network G_1 from the factorization $D_1 = C_1 C_1^T$. Let C_{1L} be the subset of the rows of C_1 corresponding to landmarks, and let D_{2L} be the subset of the rows of D_2 corresponding to landmarks (in corresponding order).

The key step then is to construct the vector representations of the nodes in the target network G_2. To do this, we treat the similarity scores D_{2L} in the target network as if they applied to the landmarks in the source network G_1. Then it is straightforward to construct RKHS vectors, $\hat{C}_2$, for nodes in the target network G_2 by solving for $\hat{C}_2$ in linear system $D_{2L} = C_{1L}\hat{C}_2^{\,T}$. This has solution set

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

where $C_{1L}^{\dagger}$ is the Moore-Penrose pseudoinverse of C_{1L}, and W is an arbitrary matrix. We choose the solution corresponding to $W = 0$.

The resulting $\hat{C}_2$ represents the embedding of the nodes of G_2 (the target) into the same space as the nodes of G_1 (the source). We can then compute similarity scores for all pairs of nodes across

[2] We note that this is a fundamentally different strategy than has been used in past manifold-alignment methods [19,59], in which alignment is based on Euclidean distance. Aligning on Euclidean distance does not respect inner product, and so the similarity captured by the kernel is not preserved in the alignment.

the two networks as $D_{12} = C_1 \hat{C_2}^T$. This yields D_{12}, an $m \times n$ matrix of similarity scores between nodes in the source and target networks.

2.2 HANDL and the Regularized Laplacian

While our method can be used with any kernel, in this paper we focus on using a kernel intended to capture functional similarity of proteins. To motivate our choice of kernel function $k(\cdot,\cdot)$, we consider the function prediction problem on a single network, $G = (V, E)$ with $|V| = n$. G has adjacency matrix A with entries a_{ij}. For simplicity we consider G to be unweighted, so $a_{ij} \in \{0, 1\}$. Extensions of our arguments to weighted graphs are straightforward.

A central idea used throughout network-based functional prediction methods is that of *guilt by association* – that is, two nodes that are near each other are more likely to share the same label than two nodes that are far apart. In the context of protein function prediction, this principle has been well established. For example, the authors in [30] show that two neighbors in the protein interaction network are more likely to have the same function than a randomly chosen pair.

Consider the case of determining whether nodes should receive a particular function label, where we label a node with a 1 if it should receive the label, and 0 otherwise. We are interested in the case in which the label is rare, and we believe that nodes may be mislabeled (e.g., some nodes labeled 0 should actually be labeled 1). We assume that there is some current labeling which is incomplete; that is, most nodes are currently labeled 0, and some nodes labeled 0 should actually be labeled 1. Define the vector y such that $y_i = 1$ if node v_i has label 1, and zero otherwise. The goal of the function prediction problem is to estimate a new $\hat{y}$, which is a better labeling of nodes in V.

To address this problem, we proceed as follows [4,15]. First, we posit that y should not differ too much from $\hat{y}$ – nodes should tend to be given the labels they already have. Second, we also posit that neighbors in G should tend to be given the same label – this is the guilt by association principle.

Note that these two goals are in conflict: fully following the first principle leaves all labels unchanged, while fully following the second principle makes all labels the same (either 0 or 1). To balance these, we define the following optimization:

$$\hat{y} = \arg\min_{y'} \underbrace{\sum_{i=1}^{n} (y'_i - y_i)^2}_{\text{quality of fit}} + \lambda \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} (y'_i - y'_j)^2}_{\text{smoothness}}, \tag{2}$$

in which we use λ to control the tradeoff between the two principles.

This expression can be compactly expressed using the Laplacian of G: $L = D - A$ in which D is a diagonal matrix with node degree on the diagonal: $D_{ii} = \sum_j a_{ij}$. Then,

$$\begin{aligned} f(\hat{y}) &= \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} (\hat{y}_i - \hat{y}_j)^2 \\ &= ||\hat{y} - y||^2 + \lambda \hat{y}^T L \hat{y}, \\ \frac{df}{d\hat{y}} &= 2\hat{y} - 2y + 2\lambda L \hat{y} = 0, \\ \hat{y} &= (I + \lambda L)^{-1} y. \end{aligned}$$

The matrix $(I + \lambda L)^{-1}$ is the *regularized Laplacian* of G [60]. It is a positive semidefinite matrix and hence a kernel.[3]

The combination of the multi-network kernel embedding (Section 2.1) with the Regularized Laplacian constitutes HANDL, and we refer to the cross-species similarity scores computed as HANDL *homology scores.* We denote the HANDL homology score of two proteins p_i and p_j as d_{ij}, we refer to the RKHS in which G_1 and G_2 are embedded as HANDL-*space*, and we refer to the vectors for nodes given by C_1 and $\hat{C}_2$ (Section 2.1) as HANDL-*embeddings*.

2.3 Scores for gene pairs

We also find it useful to develop a measure to capture functional similarity between two *pairs* of nodes across species. Pair-similarity can then be used to predict outcomes for pairs of genes (e.g. synthetic lethality. Given two pairs of nodes $(v_i, v_j), (v_k, v_\ell)$, we define a pairwise similarity metric such that the score is large only if d_{ik} and $d_{j\ell}$ (or $d_{i\ell}$ and d_{jk}) are both large. This reflects the hypothesis that synthetic lethal interactions occur within pathways, and between pathways that perform the same/similar essential biological function [5,35].

Hence, to compare pairs, we simply sum the HANDL-embeddings for the nodes in each pair, and then compute HANDL scores in the usual way. Given a matrix C of HANDL-embeddings for nodes, we define the HANDL-embeddings for a *pair* of nodes (v_i, v_j) as $P_C(v_i, v_j) = c_i + c_j$. Computing similarity for a two pairs (v_i, v_j) and (v_k, v_ℓ) then yields:

$$(c_i + c_j)^T (c_k + c_\ell) = c_i^T c_k + c_i^T c_\ell + c_j^T c_k + c_j^T c_\ell.$$

In general we expect each of the terms on the right hand side to be close to zero *unless* there is functional similarity between the corresponding nodes, because in high dimension, independent random vectors tend to be nearly orthogonal.

The utility of this approach is informed by recent work on predicting synthetic lethal interactions from network diffusion [10].

3 Results

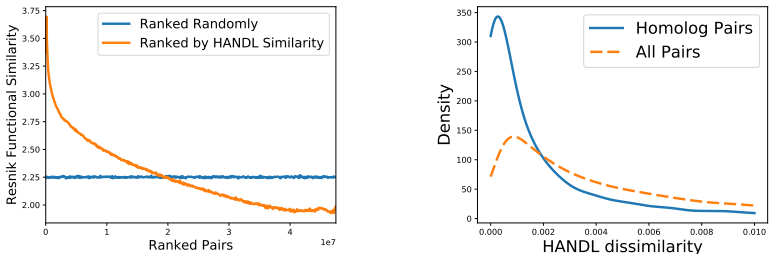
Our results show that the HANDL-embeddings encode functional relationships between proteins in different species. Leveraging HANDL homology scores as well as HANDL-embeddings, we successfully perform three cross-species tasks that traditionally use homologs: protein function annotation, phenolog discovery, and synthetic lethal classification.

3.1 Data

Sequence homologs. For each combination of two organisms, we identify homologous pairs of genes using NCBI's Homologene database [51] and chose 400 homolog-pairs at random, which constitute our landmarks.

Protein-protein interaction networks. We constructed protein-protein interaction (PPI) networks in *Sc*, *Sp*, mouse, and human (refer to Appendix B.1 for details). We restricted each network to the two-core of the largest connected component, and report summary statistics of the networks

[3] In addition to the "guilt by association" argument, we note an additional reason from [57,7] that the regularized Laplacian is an appropriate tool for functional inference on protein interaction networks: it also naturally discounts paths that pass through high-degree nodes.



(a) (b)

Fig. 2: (a) Relationship between cross-species Resnik score and HANDL homology score, for human (source) - mouse (target) comparison. (b) Distribution of HANDL dissimilarity scores for homologs, compared to distribution for all protein pairs: human (source) - mouse (target) comparison.

in Table A1. We use the two-core of the graph because topologically indistinguishable nodes (nodes that participate in an automorphism of the graph) will necessarily have identical HANDL homology scores. A large class of topologically indistinguishable nodes includes many of the leaf nodes in the graph (degree-1 nodes). That is, in any case where there are two or more leaf nodes attached to the same parent, the nodes are topologically indistinguishable. Protein names were standardized by mapping to Uniprot Accession IDs using mappings provided by the Uniprot consortium [56].

Protein function annotations. Protein functions were determined using the Gene Ontology database (GO) [3]. Currently, GO contains more than 40,000 biological concepts over three domains: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). We use GO annotation corpora downloaded from SGD [9] for yeast, and Uniprot [56] for all other species. We exclude annotations based on IEA or IPI evidence due to their lower associated confidence levels.

Synthetic lethal interactions. We constructed datasets of synthetic lethal (SL) interactions and non-interactions (non-SLs) from two studies of analogous proteins in baker's (*Sc*) and fission (*Sp*) yeast [11,49]. We report the size of the datasets in Table A2, and additional details in the Appendix B.2.

3.2 HANDL homology scores capture functional similarity across species

Our results show that high HANDL homology scores are strongly correlated with functional similarity between mouse and human proteins. As a quantitative measure of functional similarity, we use Resnik score [47]. The Resnik score between two GO terms is the information content of their most informative common ancestor in the GO hierarchy; to compare two proteins we take the maximum Resnik score over all pairs of GO terms. The Resnik score has been shown to be one of the best performing metrics for capturing functional similarity within the GO hierarchy [44].

To demonstrate the relation between HANDL homology score and functional similarity, we look at all pairs of proteins (p_i, p_j), where p_i and p_j are from human (source) and mouse (target), respectively. We exclude from this plot pairs involving the homologs used as landmarks. We order all pairs according to their HANDL homology scores, and plot against the Resnik score of the pair. The results (smoothed over non-overlapping windows of 100,000 observations) are shown in Figure 2a. Other cases are shown in Appendix C.1.

The figure shows that HANDL homology scores are strongly correlated with functional similarity across the entire range of scores. Furthermore, the very largest HANDL homology scores are indicative of protein pairs with particularly high functional similarity.

Homologs have distinct HANDL **homology scores.** Next we show that individual pairs that are known to be functionally related are distinguishable by their HANDL homology scores. For this purpose we look at protein pairs (p_i, p_j), where p_i and p_j are homologous proteins in different organisms. Note that these homolog pairs do not include any proteins used as landmarks.

In Figure 2b, we show the distribution of HANDL dissimilarity among known homologs, as compared to the distribution of scores across all protein-pairs. In this figure, we use reciprocal scores (dissimilarities), meaning that small scores are associated with high functional similarity. Only the left side of the distributions are shown, as the distribution of all pairs extends far to the right and obscures the homolog distribution on the left. As suggested by the plots, the mean HANDL homology scores for human-mouse homologs are 36% lower than the mean HANDL homology scores across all pairs. Other cases are shown in Appendix C.1. These results provide additional evidence that HANDL homology scores are correlated with functional similarity across species.

HANDL **homology scores outperform** SINATRA **homology scores.** As a final illustration of HANDL's close connection to functional similarity, we compare HANDL to the SINATRA method for computing "connectivity homology" [23] between pairs of nodes. Among other results, [23] showed that the Euclidean distance between rank-normalized connectivity profiles are smaller for pairs of nodes with the same or similar function, compared to those with different functions. For a given pair, we take reciprocal Euclidean distances of connectivity profiles to obtain a similarity, which we call the SINATRA homology score.

We compare the ability of HANDL and SINATRA homology scores to predict functional similarity. We define a pair of proteins (p_i, p_j) from two species to be *k-functionally similar* if both p_i and p_j are annotated by the same GO term and, in each species, that GO term is associated with at most k proteins. This definition was also used (without naming it) in [23]; we set $k = 100$, also one of the values used in that paper. We then evaluated the HANDL and SINATRA homology scores by computing the enrichment of k-functionally similar pairs among pairs with high homology scores. To do so, we treat the homology scores as the output of a classifier for k-functional similarity, and evaluate classifications by computing the area under the ROC (AUROC).

Table A3 shows that for human and mouse proteins, HANDL homology scores result in a significant increase in enrichment for k-functionally similar protein pairs, with a 27% higher AUROC (0.625) than using SINATRA homology scores (0.514). This advantage is consistent both for homolog pairs (excluding those used as landmarks) and non-homolog pairs, and regardless of the source-target choice used in HANDL. Table A3 also shows the advantage persists when considering a more closely related species pair (*Sc* and *Sp*). The results demonstrate that HANDL is better able to capture functional similarity between species compared to earlier methods that rely solely on network features.

3.3 Using HANDL-homologs to find known and novel phenologs

In this section we show that HANDL can yield new methods for *phenolog* discovery. The notion of a phenolog was defined in [33] as two phenotypes in different species that are related by the orthology of their associated genes. The method used in [33] to identify phenologs is based on detecting an over-representation of orthologous proteins associated with each phenotype in the pair. Reliance on orthologs is a good start for determining phenotypic preservation, but the requirement of sequence preservation may be too restrictive if the primary goal is to study function across species [2,48,45].

Our goal is to investigate whether additional phenologs may be discovered through the expanded notion of homology that is provided by HANDL (rather than reproduce the results in [33] with a different methodology). As a proof of concept, we show results using human-mouse HANDL

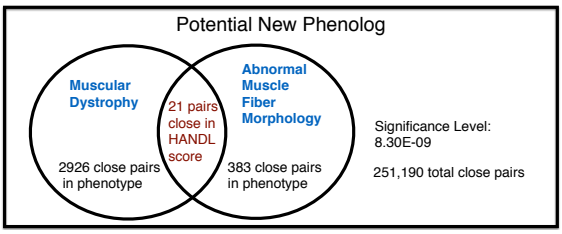


Fig. 3: A potential new phenolog, not found by [33], relating the phenotypes Muscular Dystrophy (human) and Abnormal Muscle Fiber Morphology (mouse).

Human Disease	Obvious Match	Similar Match	Novel Match
Systemic lupus erythematosus	1	0	0
Dilated cardiomyopathy	4	0	1
Zellweger syndrome	2	1	2
Dysfibrinogenemia	0	0	1
Myopathy	1	0	0
Bardet-Biedel syndrome	6	2	1
Adrenoleukodystrophy	0	3	1
Muscular Dystrophy	2	0	1
Total	**16**	**6**	**7**

Table 1: Matched mouse phenotypes per human disease among newly-identified phenologs.

homology scores. To make consistent comparisons with [33] we use the same phenotype to gene association datasets as in that study.

To match the methods used in [33] we threshold HANDL homology scores, leading to a binary classification of each cross-species gene pair as either a HANDL-homolog or not. We set the threshold so as to output a small set consisting of the most confident phenolog predictions. To compare a phenotype in one species to a phenotype in a different species we count the number of HANDL-homolog pairs across species in the two associated phenotypes. We then use the same procedure as in [33], with HANDL-homologs playing the role that orthologs did in that work; details are provided in Appendix A.3.

Our results show 47 pairs of phenotypes to be significantly associated, and therefore candidates for an expanded set of phenologs. Within this set were 18 phenologs reported by [33], which is not surprising given that many ortholog pairs are ranked highly by HANDL (e.g. see Figure 2b). However our primary interest are the matches that are not part of the ortholog-based phenologs reported in [33].

We find that HANDL-based similarity can uncover many new phenologs that are not statistically significant when using orthologs. As an example, we show in Figure 3 details of a phenolog found using HANDL, but not found in [33]. This phenolog matches *Abnormal Muscle Fiber Morphology* in the mouse with *Muscular Dystrophy* in human. The example is illustrative as it shows that false negatives can occur when using only orthologs, even for straightforward matches such as this one. In fact the ortholog-based method used in [33] finds only 1 ortholog in common, while HANDL detects 21 functionally-close pairs of proteins. The larger set of functionally-close protein pairs uncovered by HANDL gives greater statistical power for cases such as this one, where there is only a small number of orthologs shared between the two phenotypes.

We find that many of the remaining 29 phenotype pairs identified using HANDL are potentially valid phenologs. To demonstrate, we compare the human phenotype (disease) with mouse phenotype (symptom). The 29 pairs of phenotypes ranged over 8 unique human diseases and 24 unique mouse phenotypes. We obtained the description of each disease from the Genetics Home Reference[4]

Table 1 groups the results into three categories: obvious symptom match, possible symptom match, and novel match. For each disease term and mouse phenotype pair, if the name of the mouse phenotype was indicated as a symptom in the disease description, we considered an obvious match; if the disease description contained a symptom that was similar to the mouse phenotype, we considered it a possible match; and if there was no similarity in the phenotype and the disease description we considered it as a novel match. Table 1 shows that, for over 75% of the new phenologs predicted using HANDL ,there is evidence in the literature supporting the association.

[4] `https://ghr.nlm.nih.gov/`and compared the disease description to the matched mouse phenotypes.

In summary, we find that HANDL-homologous pairs can indeed provide a basis for expanding the set of phenologs discovered by previous methods. We anticipate future research developing methods to uncover phenologs may use previous methods in tandem with HANDL-based methods to more thoroughly explore the space of possible phenologs.

3.4 Synthetic lethal interactions are significantly separated in HANDL-space

In this section we show that HANDL homology scores for gene pairs in different species are associated with pairwise gene function. We hypothesize that gene pairs with genetic interactions are separated in HANDL-space from those without genetic interactions, and that the direction of separation is preserved when projecting other species into the same HANDL-space. That is, gene pairs with genetic interactions in P_{C_1} are co-located with gene pairs with genetic interactions in $P_{\hat{C}_2}$. To test this hypothesis, we consider a particular form of genetic interaction – the synthetic lethal (SL) interaction – among analogous genes in two different species of yeast. We use datasets consisting of a collection of almost all SL and non-SL pairs among 743 *Sc* genes [11] and 550 *Sp* genes [49] involved in chromosome biology (see Section 3.1).

To separate gene pairs with SLs from those without SLs in HANDL-space in a (source) species, we use a support vector machine (SVM). A linear SVM is well suited to this task, because when applied to the HANDL-embeddings (in C_1, or $\hat{C}_2$) it finds the maximum-margin hyperplane with respect to the inner products of each gene pair, thus exactly using the HANDL homology scores for classification. We trained SVMs using 4-fold cross-validation on the representation of gene pairs in the source species in HANDL-space (given by P_{C_1}), and tested on the held-out representations for the target species (given by $P_{\hat{C}_2}$). We assessed the degree of separation between SLs and non-SLs in HANDL-space in the target species by evaluating the SVM classifications with F_1 scores (the harmonic mean of precision and recall) and the area under the ROC curve, and report the average across the 4-folds.

We find that the same hyperplane that separates SL from non-SL gene paris in the source species, also separates SL gene pairs from those without SLs in a second (target) species, after projecting the target species' gene pairs into HANDL-space. With no knowledge of SLs in the target species, the SVM achieves an average maximum F_1 score of 0.185 (average AUROC: 0.70) with *Sc* as the target and 0.165 (average AUROC: 0.66) with *Sp* as the target in 4-fold cross-validation.

The significant predictive power the SVMs achieve using the HANDL stands in stark contrast to the results achieved using sequence homologs to transfer knowledge. Following earlier approaches [13], we predicted SLs across species using orthologs, obtaining a single point on the ROC (or precision-recall) curve. For a pair (u, v) in the target species, we predicted (u, v) to have a SL if u and v each have orthologs (u', v') in the source species, and (u', v') is a SL in the source species. This approach achieves modest precision of 0.10, but recall of only 0.0007, resulting in an F_1-score of 0.001. These results support our hypothesis that the HANDL-space encodes functional relationships between gene pairs across species.

3.5 Leveraging multiple model organisms for function prediction

As a final exploration of the utility of HANDL for cross-species functional inference, we study the potential for leveraging annotations from multiple model organisms simultaneously. This problem has not previously been explored extensively, with the exception of [37], which took a Bayesian approach. Our intent is not to propose a new method for function prediction, but to demonstrate the value of cross-species information as obtained via HANDL.

We focus on function prediction in three species: human, mouse, and yeast (Sc). Following the approach taken in [30], we focus on predicting rare GO terms – in particular we form predictions

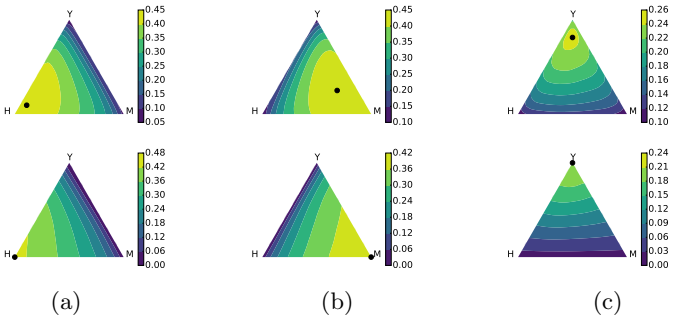


Fig. 4: F-score of functional Inference using multiple additional species. Inferring protein functions for (a) human, (b) mouse, (c) yeast. Upper row: using HANDL scores; Lower row: null hypothesis, using randomly permuted HANDL scores.

Target Species	AUC	F score
Human	0.3%	2.6%
Mouse	1.0%	8.6%
Yeast	0.3%	16.0%

Table 2: Improvement in Functional Prediction Using Two Other Species.

for all GO terms that occur between 2 and 300 times in the annotation corpus of one species. We study the information content of multiple HANDL scores by constructing a convex combination $\alpha_H, \alpha_Y, \alpha_M$ of a given species' diffusion score (D_1) with the HANDL homology scores (D_{12}) of the two other species.

We perform a binary classification for each GO term. We rank proteins by the total scores contributed by other proteins – within the same and across species – annotated with that term, weighted by $\alpha_H, \alpha_Y, \alpha_M$. We then assess performance using the maximum F-score averaged over all GO terms. We test whether such a convex combination has greater predictive power for functional inference than just using information from any single species. Details of our method are given in Appendix A.2.

Figure 4 shows the prediction performance obtained over the simplex $(\alpha_H, \alpha_Y, \alpha_M)$ such that $\sum \alpha_i = 1$. Contours show F-score, and the point of maximum F-score is plotted. For comparison purposes we provide results for the case in which HANDL scores are randomly permuted. The figure shows that in each case, the greatest improvement in functional prediction comes when making use of information from both additional species. The improvements in functional prediction are shown in Table 2.

4 Discussion

We introduce a novel, kernel-based algorithm to embed proteins from different species into a shared vector space. Our approach combines standard sequence homology with the idea of network homology through the use of network *landmarks*. We use a particular kernel – the regularized Laplacian – to create a functional embedding, and use the resulting algorithm, HANDL, to embed proteins from humans, mice, and yeast into the same space. We interpret inner products in this vector space as homology scores, and show that the embedding itself is biologically meaningful.

We evaluate HANDL embeddings through comparisons with other types of genetic and network homology on a variety of cross-species tasks, validating predictions from HANDL with the biological literature and held-out data. First, we show that HANDL homology scores are strongly correlated with functional similarity. We then use HANDL embeddings for two cross-species tasks traditionally performed using homologs: identifying "orthologous phenotypes" (phenologs) [33] and classifying synthetic lethal interactions. The phenologs identified using HANDL include 18 previously discovered [33], but importantly include 29 new phenologs that were not statistically significant using orthologs. Most of these new phenologs missed by earlier methods are supported by the biological literature. We also find HANDL encodes pairwise gene function, as pairs with/without synthetic

lethal interactions are significantly separated in HANDL-space, and the direction of separation is conserved across species. Importantly, in these tasks, we transfer knowledge both from humans to model organisms – as is commonly done to derive disease models – and from model organisms to humans – as is commonly done for comparative genomics. Thus, HANDL represents a new direction towards realizing the crucial goal of algorithms for transferring knowledge of genetics across species.

We anticipate that combining network landmarks and diffusion kernels could also be useful for a broader class of data-driven tasks. The HANDL algorithm is based on the "guilt-by-association" hypothesis for functional annotation, but it may be possible to learn how to use this information for different prediction tasks. For example, many recent methods have successfully used vectors of diffusion-based similarity scores as features for off-the-shelf supervised learning [7,10], and achieved state-of-the-art performance on multiple tasks [10]. Seen in that light, the HANDL embeddings can be seen as a component of a *transfer learning* [42] approach for cross-species inference. Transfer learning – using knowledge gained in solving one task to aid in solving a different task – is often approached by finding appropriate transformations of data features. These transform a source dataset in a fashion such that a model trained on the source data can be usefully applied to a different, target dataset ('domain adaptation', e.g., see [54]). We showed that methods for transferring knowledge across species can be useful even when there is a wealth of data in the target species. Thus, to achieve optimal performance, supervised learners may need to train on multiple species *simultaneously*.

Beyond kernels derived from protein interaction, there are a wide range of other kernels that can inform biological function assessment, including kernels derived from co-expression, genetic interaction, metabolic pathways, domain structure, and sequence [28,7,29]. Because HANDL is a method for creating a new kernel encompassing the nodes of multiple networks, it holds potential as a new tool for kernel learning methods such as support vector machines in a wide variety of applications beyond cross-species function prediction.

References

1. Ahmet Aladag and Cesim Erten. Spinal: scalable protein interaction network alignment. *Bioinformatics*, 29(7):917–924, 2013.
2. Adrian M Altenhoff, Romain A Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol*, 8(5):e1002514, 2012.
3. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
4. Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. In Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 11. The MIT Press, 2006.
5. Charles Boone, Howard Bussey, and Brenda J Andrews. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–449, 7 2007.
6. Kevin R Brown and Igor Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome biology*, 8(5):R95, 2007.
7. Mengfei Cao, Hao Zhang, Jisoo Park, Noah Daniels, Mark Crovella, et al. Going the distance for protein function prediction: A new distance metric for protein interaction networks. *PLOS One*, 8(10):e76339, October 2013.
8. Andrew Chatraryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, et al. The biogrid interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, 2017.
9. JM Cherry, C Adler, C Ball, SA Chervitz, SS Dwight, ET Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. SGD: Saccharomyces genome database. *Nucleic Acids Res*, 26:73–79, 1998.
10. Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems*, 3(6):540–548.e5, 2016.
11. Sean R Collins, Kyle M Miller, Nancy L Maas, Assen Roguev, Jeffrey Fillingham, et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446(7137):806–810, 7 2007.
12. Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 2017.
13. Raamesh Deshpande, Michael K Asiedu, Mitchell Klebig, Shari Sutor, Elena Kuzmin, et al. A comparative genomic approach for identifying synthetic lethal interactions in human cancer. *Cancer Research*, 73(20):6128–6136, 2013.
14. W M Fitch. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113, 1970.
15. François Fouss, Kevin Francoisse, Luh Yen, Alain Pirotte, and Marco Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Netw.*, 31:53–72, July 2012.
16. François Fouss, Marco Saerens, and Masashi Shimbo. *Algorithms and Models for Network Data and Link Analysis*. Cambridge University Press, 2016.
17. Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2013.
18. Adam Frost, Marc G Elgort, Onn Brandman, Clinton Ives, Sean R Collins, et al. Functional repurposing revealed by comparing s.pombe and s.cerevisiae genetic interactions. *Cell*, 149(6):1339–1352, 11 2012.
19. Jihun Ham, Daniel D Lee, and Lawrence K Saul. Semisupervised alignment of manifolds. In *AISTATS*, pages 120–127, 2005.
20. Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, 9 2009.
21. Somaye Hashemifar and Jinbo Xu. Hubalign: an accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics*, 30(17):i438–i444, 2014.
22. Sohyun Hwang, Eiru Kim, Sunmo Yang, Edward M Marcotte, and Insuk Lee. Morphin: a web tool for human disease research by projecting model organism biology onto a human integrated gene network. *Nucleic Acids Research*, 42(Web Server issue):W147–W153, 07 2014.
23. Alexandra Jacunski, Scott J. Dixon, and Nicholas P. Tatonetti. Connectivity homology enables inter-species network models of synthetic lethality. *PLOS Computational Biology*, 11(10):e1004506, 2015.
24. Ehsan Kazemi, Hamed Hassani, Matthias Grossglauser, and Hassan Pezeshgi Modarres. Proper: global protein interaction network alignment through percolation matching. *BMC Bioinformatics*, 17(1):527, Dec 2016.
25. Gunnar W Klau. A new graph-based method for pairwise global network alignment. *BMC bioinformatics*, 10(Suppl 1):S59, 2009.
26. Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322, 2002.

27. Eugene V Koonin, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Dmitri M Krylov, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, 5(2):1–28, 3 2003.
28. G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Pacific Symposium on Biocomputing*, pages 300–311, 2004.
29. Christina S. Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467, 2004.
30. Stanley Letovsky and Simon Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl 1):i197–i204, 2003.
31. Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.
32. Noël Malod-Dognin and Nataša Pržulj. L-graal: Lagrangian graphlet-based network aligner. *Bioinformatics*, page btv130, 2015.
33. Kriston L McGary, Tae Park, John O Woods, Hye Cha, John B Wallingford, and Edward M Marcotte. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549, 10 2010.
34. Tijana Milenković, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.
35. Florian L. Muller, Elisa A. Aquilanti, and Ronald A. DePinho. Collateral lethality: A new therapeutic strategy in oncology. *Trends in Cancer*, 1(3):161–173, 2015.
36. E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(supplement 1):i302–i310, 2005.
37. Naoki Nariai. *Probabilistic Integration of Heterogeneous, Contextual, and Cross-Species Genome-Wide Data for Protein Function Prediction*. PhD thesis, Boston University, 2010.
38. Behnam Neyshabur, Ahmadreza Khadem, Somaye Hashemifar, and Seyed Shahriar Arab. Netal: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, 29(13):1654–1662, 2013.
39. Sebastian M.B. Nijman. Synthetic lethality: General principles, utility and detection using genetic screens in human cells. *FEBS Letters*, 585(1):1–6, 11 2011.
40. Nigel J O'Neil, Melanie L Bailey, and Philip Hieter. Synthetic lethality and cancer. *Nature Reviews Genetics*, 2017.
41. G Ostlund, T Schmitt, K Forslund, T Kostler, D N Messina, et al. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database):D196–D203, 9 2009.
42. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.
43. Rob Patro and Carl Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.
44. Catia Pesquita, Daniel Faria, Andr O. Falco, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7):1–12, 07 2009.
45. Mark E Peterson, Feng Chen, Jeffery G Saven, David S Roos, Patricia C Babbitt, and Andrej Sali. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, 18(6):1306–1315, 2009.
46. Hang TT Phan and Michael JE Sternberg. Pinalog: a novel approach to align protein interaction networks: implications for complex detection and function prediction. *Bioinformatics*, 28(9):1239–1245, 2012.
47. Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130, 1999.
48. Igor B Rogozin, David Managadze, Svetlana A Shabalina, and Eugene V Koonin. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome biology and evolution*, 6(4):754–762, 2014.
49. Assen Roguev, Sourav Bandyopadhyay, Martin Zofall, Ke Zhang, Tamas Fischer, et al. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322(5900):405–410, 8 2008.
50. Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011.
51. Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011.
52. Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.
53. Jimin Song and Mona Singh. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25(23):3143–3150, 2009.

54. Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2058–2065. AAAI Press, 2016.
55. Mark G. F. Sun, Martin Sikora, Michael Costanzo, Charles Boone, and Philip M. Kim. Network evolution: Rewiring and signatures of conservation in signaling. *PLoS Computational Biology*, 8(3):e1002411, 12 2012.
56. The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.
57. Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2016/11/07 2011.
58. Vipin Vijayan and Tijana Milenkovic. Multiple network alignment via multimagna++. *arXiv preprint arXiv:1604.01740*, 2016.
59. Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1120–1127, 2008.
60. D. Zhou and B. Schölkopf. A regularization framework for learning from graph data. In *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, pages 132–137, 2004.

Appendices

A Methods

For all regularized Laplacians, we used a value of $\lambda = 0.05$. We found that the resulting relationship between Resnik similarity and HANDL homology score did not vary significantly for λ values between 0.005 and 0.1.

A.1 Choice of SVM Using a linear SVM, we show that HANDL-embeddings capture functional similarity between gene pairs because the kernel used by the SVM explicitly computes HANDL homology scores between pairs.

Consider HANDL-embeddings C_1 and $\hat{C}_2$ of a source and target network. Let N HANDL-embeddings of node pairs in a source network given by P_{C_1} be $\mathbf{p}_1, \ldots, \mathbf{p}_N$ and the N targets corresponding to a classification task be $t_1, \ldots, t_N$ (target values are in the set $\{-1, 1\}$). Let the SVM be trained only on the HANDL-embeddings of the source network $\mathbf{p}_1, \ldots, \mathbf{p}_N$, the corresponding learned parameters be $a_1, \ldots, a_N$, and the linear kernel used be κ.

Notice that, to make a new prediction for a pair in the target network, with the corresponding HANDL-embedding $\hat{\mathbf{p}}$ (given by $P_{\hat{C}_2}$), $y(\hat{\mathbf{p}})$, the SVM computes the following:

$$y(\hat{\mathbf{p}}) = \sum_{n=1}^{N} a_n t_n \kappa(\hat{\mathbf{p}}, \mathbf{p_n}). \tag{3}$$

The values the kernel computes in this SVM are exactly the HANDL homology scores between the pair in target network and pairs in the source network (see Section 2.3).

A.2 Function Prediction Methods We assess prediction accuracy using leave-one-out cross validation. Let $K^s = \{k^s_{ij}\}$ denote the regularized Laplacian for species s. For GO term g, let $\mathcal{G}^s_g$ be the set of proteins in species s that are annotated with g. The same-species annotation score for a given protein p and GO term g is:

$$c^s(p, g) = \frac{1}{|\mathcal{G}^s_g|} \sum_{i \in \mathcal{G}^s_g} k^s_{pi}$$

in which p is excluded from the sum (i.e., if it is contained in $\mathcal{G}^s_g$). We also construct a cross-species annotation score for each protein, in which HANDL scores with respect to proteins in the other organism are used:

$$c^{s_0,s_1}(p, g) = \frac{1}{|\mathcal{G}^{s_1}_g|} \sum_{i \in \mathcal{G}^{s_1}_g} d_{pi}$$

where $d_{pi} = D_{12}(p, i)$ is the HANDL score for protein p in species s_0 and protein i in species s_1. The prediction score is then $h(p, g) = \alpha\, c^{s_0}(p, g) + (1 - \alpha) c^{s_0,s_1}(p, g)$. To use multiple cross-species annotations, say n, we generalize $h(p, g)$ to a convex combination of the same- and cross-species annotation scores: $h(p, g) = \alpha_0\, c^{s_0}(p, g) + \sum_{i=1}^{n} \alpha_i\, c^{s_0,s_i}(p, g)$ such that $\sum_{i=0}^{n} \alpha_i = 1$.

We evaluate predictions using area under the receiver operating curve (AUC) and maximal F-score (over all detection thresholds). Since we are concerned with predicting rare GO terms, we find that maximal F-score is generally a more discriminative metric. We set the convex coefficients $\{\alpha_i\}$ via cross-validation.

A.3 Phenolog Discovery Our method matches that used in [33], using HANDL-homologs rather than orthologs.[5] Specifically, let P_1 be the genes associated with the phenotype in species 1 and P_2 be the genes associated with the phenotype in species 2. Our contingency table consists of the counts of the number of HANDL-homologs involving $P_1 \cap P_2$, $P_1 \setminus P_2$, $P_2 \setminus P_1$, and $(\Omega \setminus P_1) \setminus P_2$, with Ω denoting the set of all close pairs. We used a Fisher exact test to measure significance, and considered the match significant if the uncorrected p value was less than 0.05. We corrected for multiple testing using a Bonferroni correction; there were $1,278,312$ possible phenotype matches so we set the significance level at 3.9×10^{-8}.

B Data

B.1 Protein-protein interaction networks We constructed protein-protein interaction (PPI) networks in *Sc*, *Sp*, mouse, and human. The *Sc* and *Sp* networks were obtained from the Biological General Repository for Interaction Datasets (BioGRID) [8] release 3.4.149. Mouse and human PPIs were obtained from the STRING database version 9.1 [17]. PPI networks obtained were processed by mapping the protein names to the same namespace. Genes that could not be mapped via the UniProt database were removed from the PPI networks entirely. We provide further details of the network processing in Section 3.1. Table A1 shows summary statistics for the PPI networks before and after processing.

B.2 Synthetic lethal interactions We constructed datasets of synthetic lethal (SL) interactions in *Sc* and *Sp* from published epistatic miniarray profiles (E-MAPs). E-MAPs include genetic interactions scores for pairs of genes, where the magnitude of the score reflects the strength of the genetic interaction. We downloaded E-MAPs for *Sc* from the supplementary information of Collins, et al. [11], and for *Sp* from the supplementary information of Roguev, et al. [49]. We classified each pair of genes in each E-MAP as SL, non-SL, or uncertain. We used the thresholds from the Collins, et al. [11] supplementary information to classify pairs in *Sc*. Given a pair with E-MAP score ϵ, we classified it as SL if $\epsilon < -3$, uncertain if $-3 \leq \epsilon < -1$, and non-SL otherwise. Similarly, we used the threshold for synthetic lethality from the Roguev, et al [49] supplementary information and used the same threshold for uncertainty, classifying *Sp* pairs as SL if $\epsilon < -2.5$, uncertain if $-2.5 \leq \epsilon < -1$, and non-SL otherwise. We also remove pairs of genes in which either gene is not found in the corresponding PPI networks described in Sec 3.1. The resulting datasets included 7,165 SL and 123,507 non-SL interactions in *Sc* and 5,599 SL and 97,541 non-SL interactions in *Sp*. We report the size of the datasets pre- and post-processing in Table A2. Genes that could not be mapped via UniProt were excluded for this study. Table A2 shows summary statistics of the SL datasets before and after processing.

C Results

C.1 Associations of HANDL scores with functional similarity for other pairs of species We associated HANDL homology scores and functional similarity for pairs of proteins in additional pairs species, using the methodology described in Section 3.2. Figures 4.1 and 4.2 show the results for embedding human into yeast, and mouse into yeast.

Species	Processing	Nodes	Edges
Baker's yeast (*Sc*)	Before	6,669	497,915
	After	5,865	492,118
Fission yeast (*Sp*)	Before	4,206	57,551
	After	3,529	54,829
Human	Before	15,129	155,866
	After	12,872	153,609
Mouse	Before	6,596	18,697
	After	4,217	16,318

Table A1: Summary statistics of BioGRID [8] PPI networks. We processed the graphs to restrict to the two-core of the largest connected component.

Species	Reference	Processing	Total	SLs	Non-SLs	Uncertain
Baker's yeast (*Sc*)	[11]	Before	150,636	7,240	125,927	17,469
		After	150,309	7,165	123,507	17,466
Fission yeast (*Sp*)	[49]	Before	118,248	5,754	101,595	10,899
		After	115,315	5,599	97,541	10,620

Table A2: Summary statistics of synthetic lethal interaction datasets.

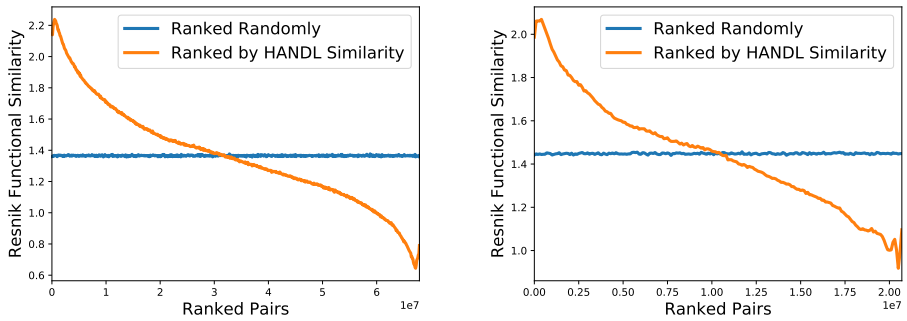
D Supplementary Tables

Data	All Proteins/Genes	Homologs	Non-homologs
SINaTRA (human↔mouse)	0.514	0.495	0.513
Handl (human→mouse)	0.624	0.556	0.624
Handl (mouse→human)	**0.625**	**0.561**	**0.625**
SINaTRA (*Sp*↔*Sc*)	0.500	0.465	0.500
Handl (*Sc*→*Sp*)	0.558	0.554	0.558
Handl (*Sp*→*Sc*)	**0.569**	**0.569**	**0.569**

Table A3: AUROC for k-functional similarity prediction using Handl and SINaTRA homology scores. Handl notation shows *source* → *target*, i.e., $C_1 \rightarrow \hat{C}_2$ in Equation (1).

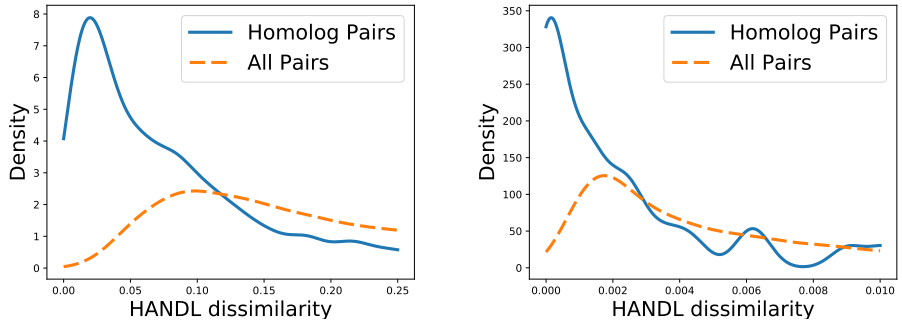
E Supplementary Figures

[5] Note that none of the homologs used to train Handl are retained in the dataset used for this study.



(a) (b)

Fig. 4.1: Relationship between cross-species Resnik similarity and HANDL homology score, for (a) human (source) - yeast (target) and (b) mouse (source) - yeast (target) comparison.



(a) (b)

Fig. 4.2: Distribution of HANDL dissimilarity scores for homologs, compared to distribution for all protein pairs, for (a) human (source) - yeast (target) and (b) mouse (source) - yeast (target) comparison.

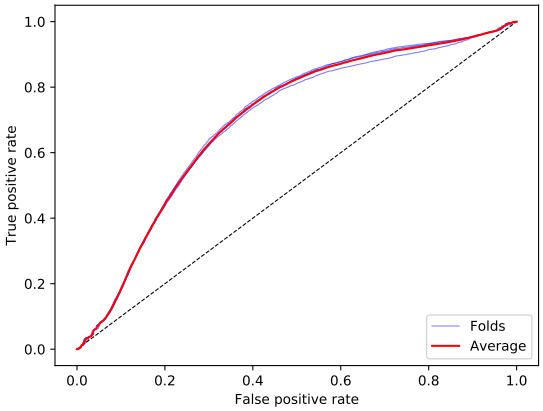


Fig. 4.3: ROC curves for prediction in *Sp* with SVM trained in *Sc* (AUROC: 0.70, F1: 0.185).