Decoding the Functional Interactome of Non-Model Organisms with PHILHARMONIC

Samuel Sledzieski[*] [◦] Charlotte Versavel[†] Rohit Singh[‡] Faith Ocitti[†]
Kapil Devkota[‡] Lokender Kumar[§] Polina Shpilker[†] Liza Roger[¶] Jinkyu Yang[‖]
Nastassja Lewinski[**] Hollie Putnam[††] Bonnie Berger[‡‡◦] Judith Klein-Seetharaman[¶] [◦]
Lenore Cowen[†] [◦]

Abstract

Protein-protein interaction (PPI) networks are a fundamental resource for modeling cellular and molecular function, and a large and sophisticated toolbox has been developed to leverage their structure and topological organization to predict the functional roles of under-studied genes, proteins, and pathways. However, the overwhelming majority of experimentally-determined interactions from which such networks are constructed come from a small number of well-studied model organisms. Indeed, most species lack even a single experimentally-determined interaction in these databases, much less a network to enable the analysis of cellular function, and methods for computational PPI prediction are too noisy to apply directly. We introduce PHILHARMONIC, a novel computational approach that couples deep learning *de novo* network inference with robust unsupervised spectral clustering algorithms to uncover functional relationships and high-level organization in non-model organisms. Our clustering approach allows us to de-noise the predicted network, producing highly informative functional modules. We also develop a novel algorithm called ReCIPE, which aims to reconnect disconnected clusters, increasing functional enrichment and biological interpretability. We perform remote homology-based functional annotation by leveraging hmmscan and GODomainMiner to assign initial functions to proteins at large evolutionary distances. Our clusters enable us to newly assign functions to uncharacterized proteins through "function by association." We demonstrate the ability of PHILHARMONIC to recover clusters with significant functional coherence in the reef-building coral *P. damicornis*, its algal symbiont *C. goreaui*, and the well-annotated fruit fly *D. melanogaster*. We perform a deeper analysis of the *P. damicornis* network, where we show that PHILHARMONIC clusters correlate strongly with gene co-expression and investigate several clusters that participate in temperature regulation in the coral, including the first putative functional annotation of several previously uncharacterized proteins. Easy to run end-to-end and requiring only a sequenced proteome, PHILHARMONIC is an engine for biological hypothesis generation and discovery in non-model organisms.

PHILHARMONIC is available at `https://github.com/samsledje/philharmonic`

[*]Center for Computational Biology, Flatiron Institute, New York, NY, USA
[†]Department of Computer Science, Tufts University, Medford MA, USA
[‡]Departments of Biostatistics & Bioinformatics and Cell Biology, Duke University, Durham, NC, USA
[§]Shoolini University, Solan, Himachal Pradesh-173229- India
[¶]School of Molecular Sciences, Arizona State University, Phoenix, AZ, USA
[‖]Department of Mechanical Engineering, Seoul National University, Seoul, South Korea
[**]Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA, USA
[††]Department of Biological Sciences, University of Rhode Island, Kingston, RI, USA
[‡‡]Computer Science & Artificial Intelligence Laboratory and Department of Mathematics, MIT Cambridge, MA, USA
◦ co-corresponding authors: ssledzieski@flatironinstitute.org, bab@csail.mit,edu, Judith.Klein-Seetharaman@asu.edu, cowen@cs.tufts.edu

1 Introduction

There exists a rich ecosystem of methods that leverage known protein-protein interaction (PPI) networks to cluster genes and proteins into functionally-enriched modules [18, 61, 82], or that use network-propagation to directly infer protein labels based on the graph-theoretic structure [22]. PPI databases such as STRING [90] and BioGRID [70] play a crucial role by aggregating experimentally determined PPIs into networks, which then enable the deployment of these powerful tools for functional annotation. However, an overwhelming proportion of known PPIs in such databases are from humans and a small number of well-studied model organisms (Section A.1), and for the vast majority of non-model organisms experimentally determined PPI interactions are too rare to establish a useful network for these types of inference.

There is likewise no existing computational method that can infer a sufficiently useful network. It is well-known that even over relatively short evolutionary distances there is substantial network re-wiring [7], so that even when sequence and protein structure conservation can allow confident ortholog assignment from a well-studied species, the interaction patterns are often not preserved. Even at a moderate evolutionary distance, the presence of an interaction can differ between two species, and the probability of shared edges decreases as the evolutionary distance between two species increases—so the networks from well-studied organisms such as human cannot trivially be used to transfer interactions to non-model organisms. Existing deep learning methods that predict PPIs are either too slow to make genome-wide predictions [29, 4, 102, 48] or are not accurate enough to use the raw predicted network as ground truth interactions [17, 84, 91].

Here, we introduce **PHILHARMONIC** (Protein Human-Transferred Interactome Learns Homology And Recapitulates Model Organism Network Interaction Clusters), a novel bioinformatics framework for network, community, and functional inference in non-model organisms. We show that PHILHARMONIC makes meaningful predictions of PPI and protein function when applied to coral, algae, and fruit fly proteomes. The key insight behind PHILHARMONIC is that PPI networks predicted by high-throughput computational methods, while not at the level of experimental correctness, nonetheless contain substantial signal of functional relatedness. **With innovative downstream processing, these *de novo* networks can serve as an information-rich scaffold from which to perform pathway-level network analysis and extract functional genomics.**

PHILHARMONIC has four primary stages (Figure 1a). First, we use the deep learning method D-SCRIPT (a deep learning PPI prediction method shown to have state-of-the-art performance [86] in non-model organisms) to infer the initial, noisy PPI network. We then build a novel hierarchical clustering algorithm based on the Double Spectral method, which couples a Diffusion State Distance (DSD)-based similarity metric [14] with several rounds of spectral clustering [68] to convert the PPI network into putative functionally-enriched non-overlapping clusters of balanced sizes amenable to easy biological analysis. These clusters are by definition non-overlapping and can be highly disconnected, as opposed to in biology where proteins can play many functional roles. To rectify this, we develop a novel method, ReCIPE (Reconnecting Clusters In Protein Embedding), a greedy algorithm that optimizes intra-cluster connectivity by re-adding high degree nodes and creates coherent, better connected, and explainable clusters. We employ remote homology methods that give sequence-based gene ontology (GO) annotations to proteins based on Pfam domains [28], and our clusters inherit the functions that are enriched in its constituent proteins. On a gene-by-gene basis, we have previously demonstrated that some functional inference can still be accomplished by remote homology approaches using profile-profile HMMs [52]. We demonstrate here that when combined with functionally-enriched clusters, these annotations can be successfully transferred to non-annotated proteins to infer larger functional pathways. Finally, we use a generative AI tool to summarize the function of each cluster in paragraph form, providing easily readable summaries of cluster functions for the user.

Corals are important natural resources that are key to the ocean's vast biodiversity and provide economic, cultural, and scientific benefits [9, 12, 74]. While they cover only 0.1% of the ocean floor, coral reefs are home to the largest density of animals on earth, rivaling rain forest habitats in species diversity [37]. As a result of anthropogenic activities, coral holobionts are declining rapidly (Section A.2). This trajectory of loss highlights the extreme urgency to assist corals in overcoming the impacts of pollution and global warming before the damage to these vital reef ecosystems is irreparable. New efforts are underway to harness the growing amount of genomic information that is becoming available for the coral animal, its hosted symbiotic algae, and other elements of the coral holobiont, to discover ameliorative strategies that could slow coral decline [23]. However, these efforts are hampered by the great evolutionary

distance of corals to any species with functionally well-annotated genes and genomes.

The importance of corals led us to apply PHILHARMONIC to the proteome of the coral *Pocillopora damicornis*, [24] (Section 3.4) and to one symbiont species, *Cladocopium goreaui* [62] (Section A.7), newly assigning predicted functional labels to hundreds of un-annotated proteins. We spotlight two clusters enriched with functions important for coral resilience, temperature sensing (Section 3.4.1) and response to environmental stimuli (Section 3.4.2). These clusters include proteins that homology methods fail to annotate, shedding some light on the so-called "dark proteome" [88] of proteins that are too evolutionary distant from model organisms for remote homology methods to predict their functions. We support these results by showing significant functional coherence in our clusters with both GO annotations (Section 3.1) and gene co-expression (Section 3.2) from Connelly et al. [20]. We perform a similar analysis on *C. goreaui* in Section A.7. As orthogonal validation of our method, we evaluate ReCIPE on recent DREAM Challenge networks (Section 3.3) where we demonstrate that the reconnection step yields more functionally enriched clusters. We also run PHILHARMONIC on the fruit fly proteome (*Drosophila melanogaster*), where we show significant enrichment for canonical GO functions from FlyBase [94] (Section A.8). Our work demonstrates the ability of methods like PHILHARMONIC to realize the promise of systems biology, particularly with respect to understanding non-model organisms.

2 Methods

2.1 Data Collection

In this work, we focus on the coral *Pocillopora damicornis* and its dinoflagellate symbiont *Cladocopium goreaui* (formerly named *Symbiodinium goreaui*, Clade C type C1). Coral protein sequences were obtained from Cunning et al. [24], while symbiont sequences were obtained from ReefGenomics [60, 100], based on Liu et al. [62]. After filtering to proteins with between 50 and 800 residues due GPU VRAM limitations, we were left with 22,541 *P. damicornis* sequences and 28,271 *C. goreaui* sequences. We also perform analysis on the fly *Drosophila melanogaster* genome, using sequences from FlyBase [94]. We keep 11,487 fly sequences with between 50 and 800 residues.

2.2 Functional Annotation of Individual Genes

Remote homology methods can find related genes that might preserve functional roles over greater evolutionary distance than standard sequence-based or ordinary-HMM-based searches. We employ a two-step functional annotation to maximize conserved function. We construct an HMM database from Pfam [28] domains using hmmer [27], then use hmmscan to search our input sequences against this database. Each protein is then annotated with some number of domains. Finally, we use mappings between domains and GO biological process (GO:BP) functions from GODomainMiner [1] to transfer functional terms to candidate sequences. Because structured domains are more likely to be conserved across long evolutionary distance, this procedure ensures the highest quality functional annotations for single proteins—but still results in some proteins which are unable to be functionally characterized, or may yield false positive annotations.

2.3 Synthetic Network Prediction

D-SCRIPT [85] is a deep learning method that was trained on human PPIs and has been shown to generalize broadly across species. It requires only the amino acid sequences of two proteins and predicts a probability of interaction. D-SCRIPT can either be run on all pairs of proteins, or, if there are computational constraints, instead the set of proteins can be curated based on a list of GO terms. In our case, we use a hand-curated list of diverse high-level GO:BP terms potentially related to coral metabolism, resilience, and regulation (see Section A.4 for this full list). We keep any proteins which are annotated with these high-level GO terms or any child GO terms. We generate a list of candidate interactions by considering all pairwise interactions between the filtered set of proteins, resulting in 45.9 million coral candidates, 52.7 million symbiont candidates, and 15.2 million fly candidates. We use D-SCRIPT because it is both highly accurate for non-model organisms and computationally efficient enough to be run genome-wide on the proteins in a new organism—it takes approximately a week to predict ≈ 50 million interactions on a single GPU. We consider a positive interaction to be anything predicted with a probability greater than 0.5.

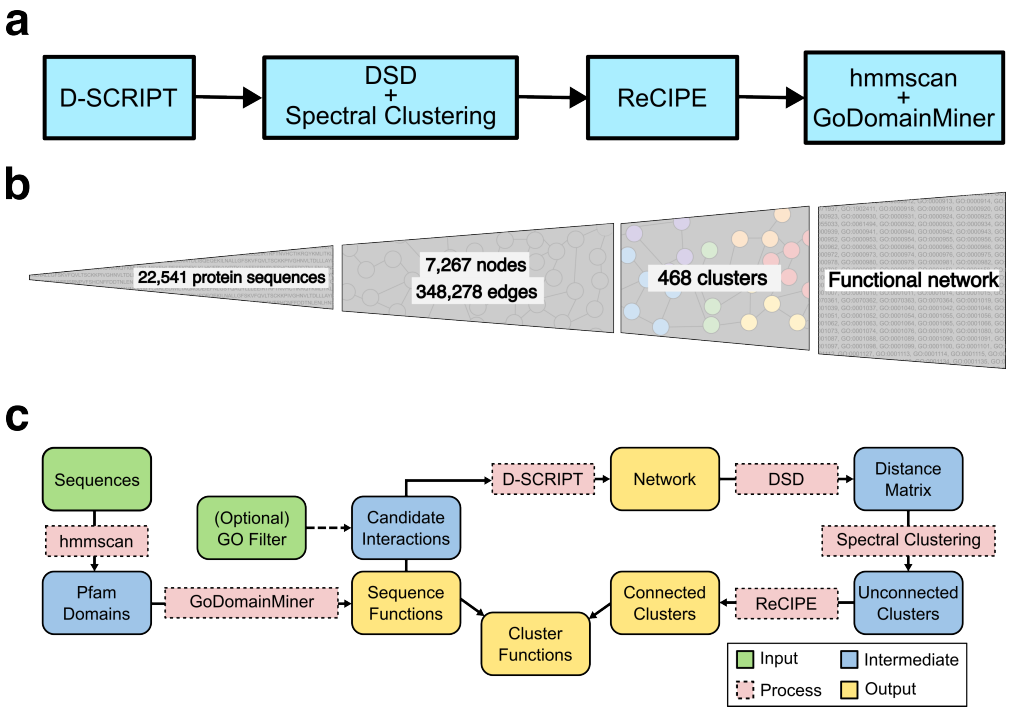


Figure 1: PHILHARMONIC is an engine for biological discovery in non-model organisms. (a) High-level overview of PHILHARMONIC. We use D-SCRIPT [85] to predict a *de novo* protein-protein interaction network (optionally on a smaller set of GO-filtered proteins, if computational cost is an issue), then cluster this network using DSD [13], recursive spectral clustering [68], and ReCIPE (see Figure 3). We use hmmscan [27] and GoDomainMiner [1] to map sequence functions, then use the individual sequence functions to assign cluster functions, which allows for network-level discovery and hypothesis generation. **(b)** PHILHARMONIC requires only a sequenced proteome, which provides on its own little functional information. We integrate several other sources of information and downstream processing to build up to a functional network. **(c)** More detailed diagram of the PHILHARMONIC method, where we show individual inputs, outputs, intermediate data, and the tools we use to transform them, including the newly-developed ReCIPE.

2.4 Diffusion State Distance

DSD is an effective way to measure distance between nodes in a PPI network. We first consider it in the case the edges are unweighted as in [14]. For two nodes (proteins) a and b, we measure the expected number of times a k-step random walk starting at a visits b (including the 0-step random walk), and call this $He^k(a, b)$. If $v_1, \ldots v_n$ represent all the nodes in the network, we then associate to each node v_i the vector consisting of these He^k values to all n nodes, namely

$$He^k(v_i) = (He^k(v_i, v_1), He^k(v_i, v_2), \ldots He^k(v_i, v_n)) \tag{1}$$

Then the DSD distance between two nodes v_i and v_j is defined as the L_1 norm of the vectors $He^k(v_i)$ and $He^k(v_j)$, i.e.

$$DSD^k(v_i, v_j) = ||He^k(v_i) - He^k(v_j)||_1 \tag{2}$$

Cao et al. (2013) [14] show that this converges (rapidly, in small world networks) as the length of the random walk $k \to \infty$, and denote the $DSD(v_i, v_j)$ without the k superscript to mean the converged DSD. This can be computed exactly with the closed form in [14]. Furthermore, when there are weights on the edges, a straightforward modification presented in Cao et al. (2014) [13] can take into account confidence weights on the edges by instead biasing the underlying random walk to more likely go along edges of more confidence (in direct proportion to the confidence ratios among the edges from the node). Everything else in the weighted DSD definition stays the same. We set the confidences according to the edge weights predicted by D-SCRIPT.

2.5 Community Detection

Given the DSD-smoothed node distance matrices, we convert this to a similarity measure between nodes using the RBF kernel [43]. We cluster this distance matrix using a modified version of the spectral clustering algorithm described in [68], recursively splitting any cluster that was bigger than assigned threshold. This DSD + spectral clustering method won the disease module identification challenge in DREAM 2016 [18].Our updated version of this algorithm, which gives us more control over the final distribution of cluster sizes, requires three parameters K, D, M. For the default PHILHARMONIC implementation, we initially call spectral clustering with $K = 500$ clusters. The initial set of clusters is recursively split, where clusters are recursively divided into $round(\frac{n}{D})$ clusters (where n is the size of the cluster) until $round(\frac{n}{D}) < 2$. Since the network will display cluster organization at multiple scales, we chose the default $D = 20$ to create clusters of a size reasonable for biologists to further investigate by hand (resulting in a maximum cluster size of $(1.5 \times D) - 1 = 29$). We finally discard clusters with fewer than $M = 3$ nodes.

2.6 Reconnecting Clusters

We observed that because DSD groups together nodes that are most similar, these clusters often placed nodes in the same cluster because they were all connected to the same intermediate-degree minor hub nodes, but these hub nodes were instead grouped with other minor hub nodes, producing highly disconnected clusters. We develop a new method, ReCIPE (Reconnecting Clusters in Protein Embeddings) to return the minor hubs back into these clusters if they were otherwise disconnected, allowing some overlap between clusters (Figure 3). This is a substantial methodological innovation, because by relaxing the strictly non-overlapping requirements of spectral methods we maintain the strengths of those methods while gaining the flexibility and interpretability required to model complex biological systems. Proteins were considered for re-inclusion in the cluster ("candidate proteins") if they connected at least 10% of the disconnected cluster components. Then the candidate proteins are added back in increasing order of degree. If n is the number of proteins in the cluster, and c is the number of disjoint cluster components, then proteins are added until the connectivity score $s_c = 1 - \frac{(c-1)}{(n-1)} > 0.75$, up to a maximum of 20 re-added proteins (where we note that a fully connected cluster has a s_c of 1, and that since clusters contain at least 3 nodes, the denominator is always defined (and at least 2)). We selected these parameters from independent experiments on human PPI networks derived from the DREAM challenge [18], where they led to clusters that were better connected and more functionally enriched than the original (Section A.11). We perform a comprehensive evaluation of hyper-parameters both for community detection and reconnection in Section A.9, and find that PHILHARMONIC is robust to a wide variety of choices for these parameters. We also show that the ReCIPE method does not result in clusters that are massively overlapping and that even after the ReCIPE reconnection step, most nodes are not placed in too many clusters (Figure A18).

2.7 Creating the Cluster Graph

To aid in the analysis of functional clustering in the network, we compute a higher-order graph where each node is not a protein, but rather a PHILHARMONIC-determined cluster. Given cluster C with nodes $\{c_1, ..., c_n\}$, cluster D with nodes $\{d_1, ..., d_m\}$, and original adjacency matrix $A \in \{0, 1\}^{N \times N}, N > n, m$, we construct a new graph G with edges $(C, D) = \sum_{c_i, d_i} A_{c_i, d_i}$. Finally, we drop all edges with weight less than $t = 10$, although we later filter to higher levels of t for visualization. When constructing this graph, we do not consider ReCIPE-added nodes in the edge count, as ReCIPE results in potentially overlapping clusters, leading to potentially double-counting edges if the same node appears in both C and D. We assign each cluster node a single function, based on GO Slim [3] terms appearing most frequently within that cluster. A given GO term must appear in at least 3 proteins in the cluster to be assigned, otherwise the cluster is left with an unassigned function.

2.8 LLM Summarization

Inspired by work from Hu et al. [40], we sought to improve the interpretability of our clusters by using large language models (LLMs) to automatically annotate them with names. Not only does this more easily summarize cluster function, but we found that it eases communication within teams of users when referring to clusters. We utilize the `llm` command line utility [101], prompting the GPT 4o-mini model with a short text summary of each cluster containing the size, number of edges/triangles, and the top-represented GO terms within each cluster. The model is instructed

to act as an expert biologist, and to come up with a short, descriptive, human-readable name for each cluster. The full system prompt is include in Section A.5.1. This name is added to the text summary from above and output for biological analysis (Figures 5a, 6a). We note that, similarly to Hu et al., the correctness of cluster labels will be highly dependent on the prompt and the choice of LLM. These cluster names should be used as a helpful supplement rather than an absolute statement of truth about the function of the cluster.

2.9 Implementation

We implement PHILHARMONIC using Snakemake [65]. This allows for consistent and reproducible analysis, flexible configuration with a `.yaml` file, and flexible allocation of computational resources. We use HMMER v3.3.0, D-SCRIPT v0.2.8, and fastDSD v1.0.1. ReCIPE is a newly implemented and `pip`-installable Python package; we use v0.0.4 in this work. Each clustering and analysis step is implemented as a Python script which is run by Snakemake, and can be individually called through the PHILHARMONIC package. We use the spectral clustering implementation from scikit-learn version 1.5.0. Statistical tests were performed with scipy version 1.11.3. We use Python version 3.11 and Snakemake version 8.10. D-SCRIPT was run on a single A6000 NVIDIA GPU, and hmmscan was run using 32 cores. Cytoscape 3.10.2 was used for visualization of all networks, both protein-protein interaction and higher level cluster graphs. We provide a set of Cytoscape style files that can be used to visualize assigned cluster functions. PHILHARMONIC is available at `https://github.com/samsledje/philharmonic`. ReCIPE is available at `https://github.com/focitti/ReCIPE`.

3 Results

3.1 PHILHARMONIC clusters are functionally coherent

We set out to determine whether PHILHARMONIC generated functionally meaningful clusters by investigating the level of functional coherence within a cluster. We define functional similarity between two proteins as the Jaccard similarity between the sets of GO [3] terms assigned to each protein by our Pfam-based function annotation. Then, the cluster coherence is the mean functional similarity of all pairs of proteins in the cluster. We compare PHILHARMONIC clusters with a degree-preserving random clustering, constructed by randomly shuffling node function assignments. We find that *P. damicornis* clusters computed by PHILHARMONIC conserve significantly more functional relationships ($p = 1.15 \times 10^{-53}$, one-tailed independent samples t-test) (Figure 2a). We compare with several other state-of-the-art clustering methods in Section A.6.2. Using the most-represented GO Slim term assigned to each cluster (Section 2.7), we show the cluster coherence of clusters with each high-level function in Figure 2c. We find several highly-coherent clusters related to mitotic cell cycle (GO:0000278), DNA-templated transcription (GO:0006351), and inflammatory response (GO:0006954). We perform the same analysis using only GO Slim [3] terms in Section A.6.1. We replicate this analysis on the *C. goreaui* network (Section A.7.1), and the *D. melanogaster* network using gold-standard GO terms from FlyBase [94] (Section A.8). To evaluate the impact of network noise on coherence, we replicate this analysis using the gold-standard *D. melanogaster* network from the STRING database [89] (Section A.8.2). We perform a similar analysis using shared gene groups from FlyBase [94] (Section A.8.3).

3.2 PHILHARMONIC clusters correlate with gene expression

We likewise investigated whether the PHILHARMONIC *P. damicornis* clusters are more likely to group proteins which are co-expressed. We use gene expression data from Connelly et al. [20], consisting of 47 samples of RNAseq data from four *Pocillopora* coral specimens, collected in triplicate across four conditions (control, heat exposure, antibiotics exposure, heat+antibiotic exposure). Following [20], we compute gene co-expression scores by first computing a variance-stabilizing transform on expression values, then computing the bi-weighted mid-correlation statistic between genes. Genes with significant missing data (0 counts in $> \frac{1}{2}$ of samples) were removed. As a baseline, we re-shuffle expression, preserving the distribution of expression values. We find that proteins in PHILHARMONIC clusters are significantly more likely to be co-expressed across these tested conditions ($p = 1.27 \times 10^{-21}$, one-tailed related samples t-test) (Figure 2d,e). We compute an alternate method of co-expression within a cluster using the singular values of the cluster sub-matrix in Section A.6.3, with similar findings.

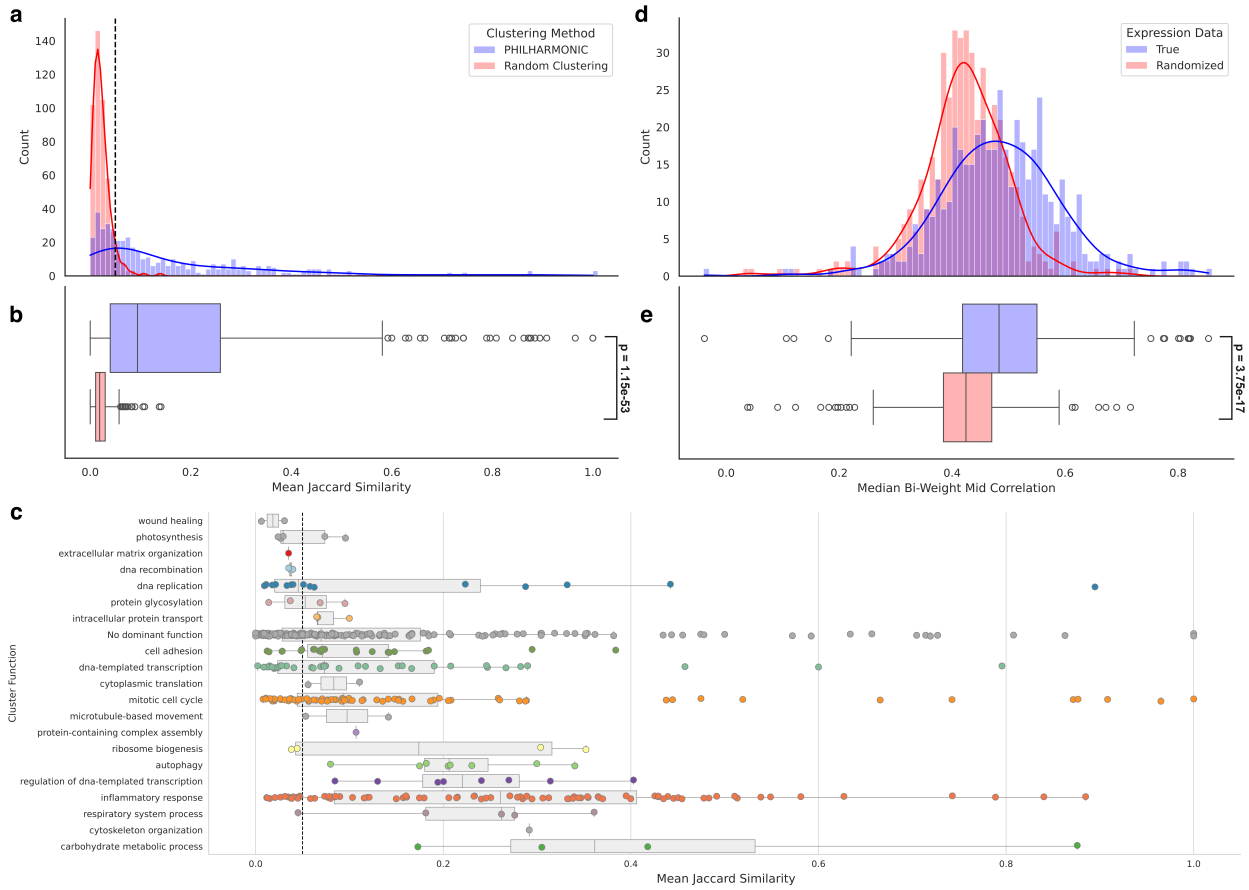


Figure 2: Validation of PHILHARMONIC clusters in *P. damicornis*. **(a,b)** We compute a cluster coherence as the mean Jaccard similarity between the sets of GO functional terms assigned to each pair of proteins in the same cluster. Compared to a distribution-preserving random clustering of the network, PHILHARMONIC clusters are significantly more functionally coherent ($p = 1.15 \times 10^{-53}$ one-tailed independent t-test). We use this test to set a cluster coherence threshold of 0.05 (dashed grey line) where the two curves cross, above which nearly every cluster is unlikely to be drawn from the random distribution. **(c)** We show each of the 486 clusters, separated by GO Slim functional assignment and plotted by cluster coherence. We find a substantial number of clusters above the 0.05 threshold across several diverse functions, especially mitotic cell cycle (GO:0000278), DNA-templated transcription (GO:0006351), and inflammatory response (GO:0006954). We show the same threshold at 0.05. Points are colored by cluster functions from Figure 4c. **(d,e)** Using *Pocillopora* gene expression data from Connelly et al. [20], we compute the median bi-weight mid correlation between pairs of genes in the same cluster, and compare this to shuffled expression data. We find that proteins that share a PHILHARMONIC cluster are significantly more likely to be co-expressed ($p = 3.75 \times 10^{-17}$, one-tailed relative t-test).

3.3 Reconnecting clusters improves functional enrichment

As true information about functional clustering is sparse outside of human data, we validate ReCIPE in the setting where the initial DSD+Spectral Clustering method was validated, namely three human PPI networks with known functions sourced from the Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenge [18] (Section A.11 for further network details). We compute the set of enriched GO terms for a cluster using v2.0.1 of the `FUNC-E` Python package [30]. We then assign GO functions based on the enriched terms, and compute a Jaccard similarity between the predicted and true sets of functional labels. We show in Figure 3 and Figure A16 that ReCIPE clusters perform at least as well as non-overlapping clusters for all three DREAM networks, and are often much better for functional annotation in addition to their improved interpretability. ReCIPE with a linear ratio of 10% overall performs the best, so we select this linear ratio for our analysis of the *P. damicornis*, *C. goreaui*, and *D. melanogaster* networks. We perform a similar analysis where we score function predictions not by Jaccard score, but by an F1 score computed over the top 10 most strongly enriched functions (by p-value) in Section A.11. In Figure A17, we show that the 10% linear ratio setting is significantly better than random additions of the same number of proteins to each

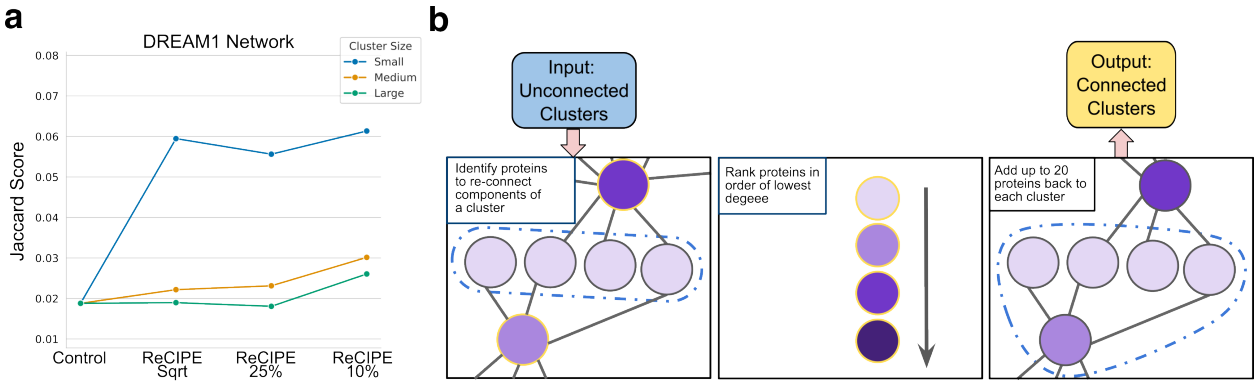


Figure 3: Reconnection with ReCIPE results in more functionally enriched clusters. **(a)** Jaccard score of function predictions for DREAM1 network when clusters are computed using ReCIPE at different linear ratio levels, compared to a control of disconnected clusters. We separate clusters into small (8-16 nodes), medium (17-32), and large (33-64) to better visualize performance. Generally, ReCIPE has the large impact on small clusters, likely because functionally informative hub nodes are preferentially placed in large clusters by the spectral algorithm. We find that a linear ratio of 10% yields the overall best performance, and thus use this setting for our coral experiments. We show a similar analysis using the percentage of enriched clusters, and on DREAM2 and DREAM3, in Section A.11. **(b)** Schematic overview of the ReCIPE method. ReCIPE identifies candidate nodes for re-addition, ranks them prioritizing low-degree nodes, and adds them back greedily until a connectivity threshold or maximum number of proteins is hit.

cluster.

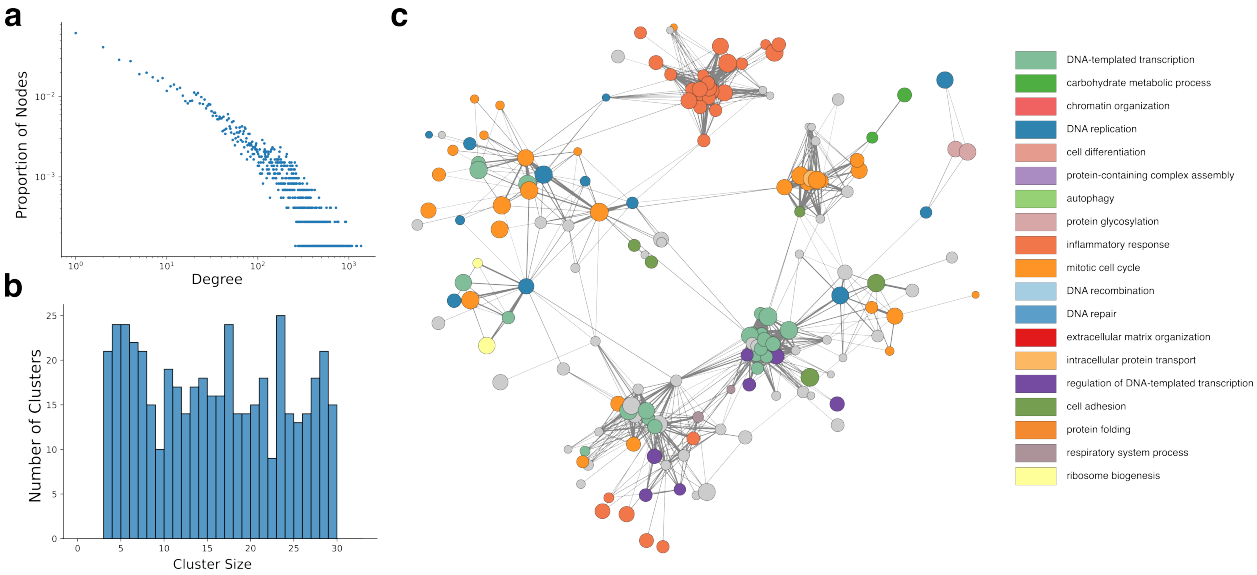
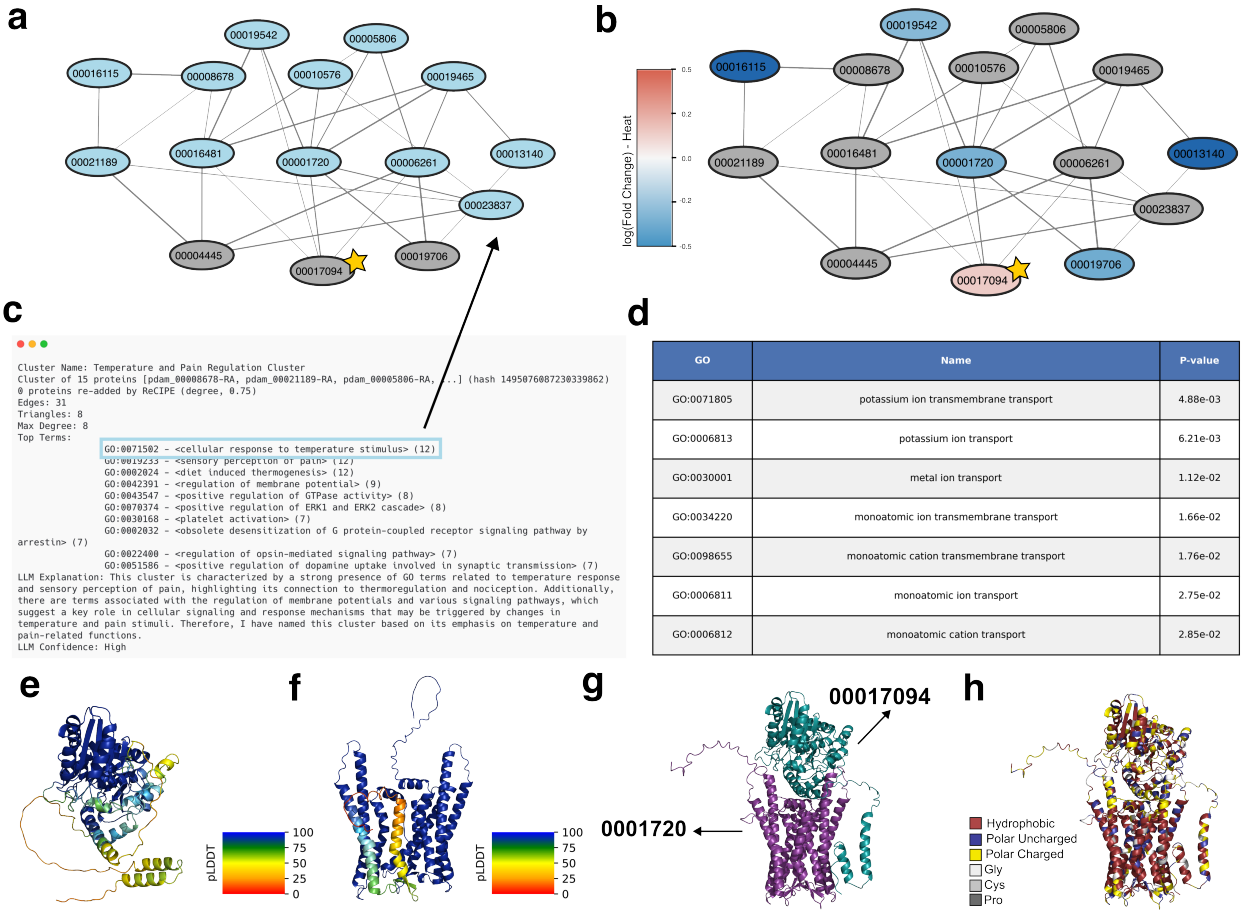


Figure 4: Dissecting the functional network of the coral *P. damicornis*. **(a)** The synthetic PPI network predicted by PHILHARMONIC follows a scale-free degree distribution commonly found in biological networks [2]. **(b)** We perform DSD followed by a hierarchical spectral clustering on the network, yielding clusters with sizes between three and 29 nodes. **(c)** The *P. damicornis* cluster graph, where each node is a cluster and edges indicate high cross-connectivity between clusters. Each cluster is colored by the predominant high-level function of constituent proteins. We find neighborhoods of the network focused on inflammatory response (GO:0006954, dark orange), mitotic cell cycle (GO:0000278, light orange), and DNA-templated transcription (GO:0006351, green) and its regulation (GO:0006355, purple).

3.4 Dissecting the *P. damicornis* network

Reef building coral play a crucial role in the reef ecosystem. We therefore chose to apply PHILHARMONIC on a case study in the coral *P. damicornis* (We also apply PHILHARMONIC to *C. goreaui* in Section A.7) and *D. exprenogaster* in Section A.8). The *P. damicornis* network displays scale-free characteristics [2] typical of biological networks (Figure 4a), and the distribution of cluster sizes is balanced and intentionally constrained by our recursive splitting



GO	Name	P-value
GO:0071805	potassium ion transmembrane transport	4.88e-03
GO:0006813	potassium ion transport	6.21e-03
GO:0030001	metal ion transport	1.12e-02
GO:0034220	monoatomic ion transmembrane transport	1.66e-02
GO:0098655	monoatomic cation transmembrane transport	1.76e-02
GO:0006811	monoatomic ion transport	2.75e-02
GO:0006812	monoatomic cation transport	2.85e-02

Figure 5: Example cluster involved in temperature and pain regulation in *P. damicornis*. (a) Graph of this cluster, where proteins annotated with cellular response to temperature stimulus (GO:0071502) are highlighted in blue. **(b)** Colored by log_2(Fold change) in expression when exposed to heat [20] (genes not in data are colored gray). **(c)** Human-readable PHILHARMONIC output for selected cluster. **(d)** Enriched GO terms identified by g:Profiler [50] **(e)** AlphaFold-predicted [45] structure of `pdam_00017094`, an uncharacterized protein (starred above), colored by pLDDT. **(f)** AlphaFold-predicted structure of `pdam_0001720`, a putative allatostatin-A receptor, colored by pLDDT. **(g)** AlphaFold-Multimer-predicted [29] structure of their interaction. **(h)** Colored by amino acid type, highlighting the likely membrane association of the complex.

procedure (Figure 4b). Further network statistics can be found in Table A2. Additionally, we show the high-level functional organization of the network where each node is itself a cluster (Figure 4c, see Section 2.7 for details). We find substantial organization within this network, with different functions partitioned to different "neighborhoods" of the network. This view of the network highlights the centrality of gene regulation and the cell cycle in cellular function [75], and we identify a tightly-connected neighborhood of the overall network dedicated to inflammatory response [95].

In addition to this general view, we investigate two clusters in depth, both of which involve ion channels communicating with G-coupled protein receptors (GPCRs). A large repertoire of GPCRs with homology to neuropeptide receptors that modify behavior in the vertebrate brain seem to have ancient origin, and predate the invertebrate-vertebrate split, and thus can be found in cnidaria such as hydra and coral animals [72], despite their lack of a traditional brain organ. The pathways involving these genes form a central nervous system in coral polyps and could effect coral behavior and plasticity to adapt to, e.g. heat stress. Indeed, we demonstrate heat-stress differential expression in multiple genes in these clusters using data from an independent study [20]. Using these clusters, we perform novel functional analysis of previously uncharacterized proteins.

3.4.1 Cluster involved in regulation of thermal response in *P. damicornis*

We analyze here one PHILHARMONIC cluster involved in cellular response to temperature stimulus (GO:0071502) (Figure 5). Cnidarians, including coral animals, have a large and diverse family of voltage-gated K+ channels, but little is known about how the function of these channels differs across different cnidarian species [55]. This cluster includes three proteins that are likely voltage-gated K+ channels (pdam_00019542, pdam_00019465, pdam_00010576), a cyclic nucleotide-gated K+ channel (pdam_0008678, [63]), and an HCN channel (pdam_00005806). HCN channels are involved in controlling neuronal excitability, dendritic integration of synaptic potentials, synaptic transmission, and rhythmic oscillatory activity in individual neurons and neuronal networks in humans, where they have been studied for their role in epilepsy and pain [8]. Cnidarian HCN channels share all major functional features of vertebrate HCNs, including reversed voltage-dependence, activation by cAMP and PIP2, and blocking by extracellular Cs+ [5]. There is some evidence that HCN channels are also involved in rhythmic firing and the so-called pacemaker channels [81]. This cluster also contains several GPCRs that are predicted to bind to and likely modulate the functions of the above diverse ion channels in response to different stimuli. Notably, many of these receptors are linked to neural peptides (Section A.6.4), including an orexin receptor (pdam_00006261) that has been implicated in systems involving sleep and light-sensing in mammals [47]. pdam_00001720 resembles the allatostatin-A receptor, to which the neuropeptide allatostatin-A binds, modulating feeding behavior in mosquitos [19].

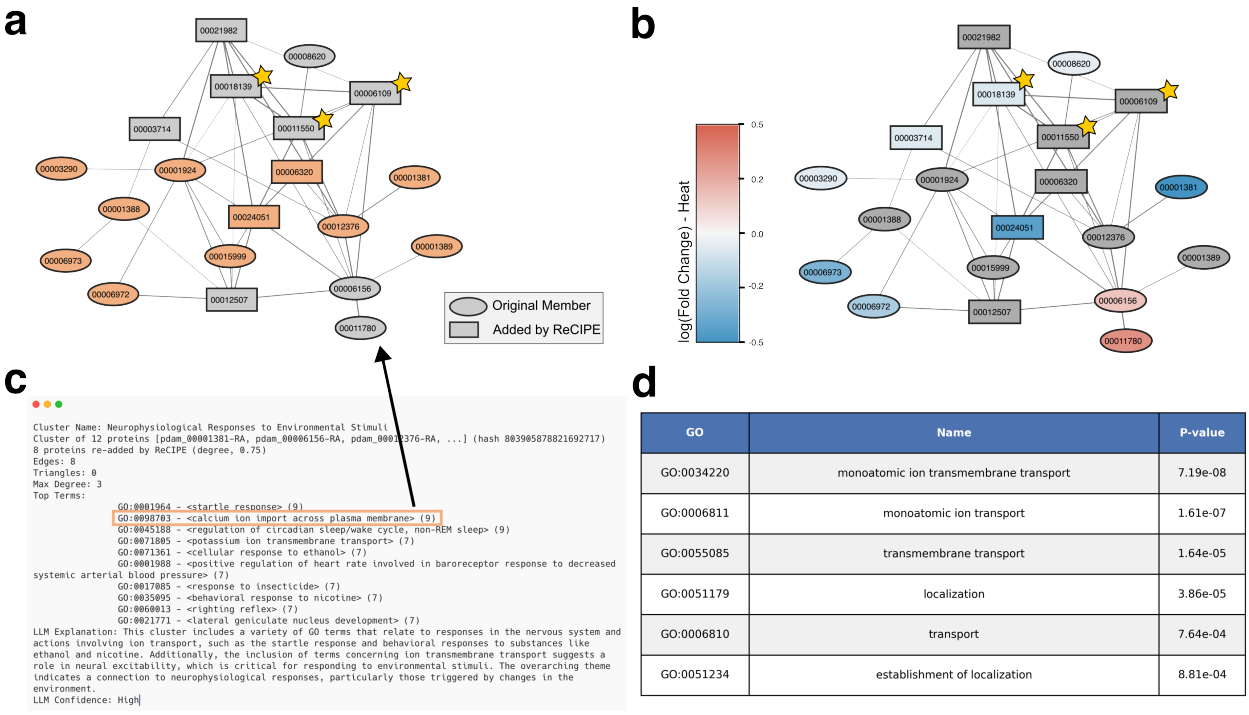
We identify one protein, predicted to bind to a subset of the GPCRs, including the putative orexin and allatostatin-A receptors, which is completely uncharacterized (pdam_00017094, starred in Figure 5b and c). Not only were we unable to identify homologs through sequence similarity search, but we also could not find structural homologs using FoldSeek [98]. Using data from Connelly et al. [20], we find that 00017094 is slightly up-regulated during heat stress. We use AlphaFold [45, 29] to predict the structure of 00017094 in complex with 0001720, as well as their individual structures. 00017094 contains two likely flexible helices, predicted with low plDDT (Figure 5e), that are predicted to sit alongside the barrel in the complex structure (Figure 5h). Due to the number of polar residues in these helices, it is more likely that these helices float on the surface of the membrane, or interact with another protein. These structures support the conclusion that 0001720 is likely a membrane-associated receptor, and suggest that 00017094 may bind on the outside of the receptor. It is tempting to speculate that its binding may regulate the allatostatin-A receptor and coral feeding behavior as a response to heat stress.

3.4.2 Cluster involved in neurophysiological response to environmental stimuli in *P. damicornis*

This cluster likewise contains two proteins with high levels of homology to K+ voltage-gated channels [73], namely pdam_00001381 in the original cluster and pdam_00006320 among the proteins added in by ReCIPE. Among the remaining proteins include several highly homologous to ligand-gated ion channels that have been recognized as key modifiers of behavior in vertebrate brains, four gamma-aminobutyric acid (GABA) receptors (pdam_00012376-RA, pdam_00001388, pdam_00001389 and pdam_00024051, the last added by ReCIPE), and four that are likely nicotinic receptors (pdam_00006972, pdam_00015999, pdam_00006973, pdam_00012507). GABA has been identified as a signalling molecule in feeding behaviour, orientation, and tentacle movement in other Cnidarian species [87]. Jing et al. [44] showed that the roles of GABAergic inhibitory interneurons can also be extended to the Aplysia feeding motor network. They are widely expressed in the central nervous system, while, in the periphery, they mediate synaptic transmission at the neuromuscular junction and ganglia. There are three proteins in this cluster which have previously been uncharacterized – pdam_00011550, pdam_00018139, and pdam_00006109, which we highlight in Figure 6 with stars. Two of these, 00011550 and 00006109, are likely interacting with a GABA receptor candidate 00024051 that is highly down-regulated in heat.

4 Discussion

While substantial previous work has advanced our understanding of functional networks in humans [31, 16, 32, 49, 42], limited experimental data and large evolutionary distance from well-studied organisms has made the analysis of functional networks in non-model organisms difficult. Methods for remote homology detection [52] or structure-based search [99] are important steps in this direction, but network re-wiring between species [83] limits their applicability to genome-scale pathway analyses. Recently developed machine learning methods to transfer PPIs from humans to



GO	Name	P-value
GO:0034220	monoatomic ion transmembrane transport	7.19e-08
GO:0006811	monoatomic ion transport	1.61e-07
GO:0055085	transmembrane transport	1.64e-05
GO:0051179	localization	3.86e-05
GO:0006810	transport	7.64e-04
GO:0051234	establishment of localization	8.81e-04

Figure 6: Example cluster involved in neurophysiological response to environmental stimuli in *P. damicornis*. (a) Graph of this cluster, where proteins annotated with calcium ion import across the plasma membrane (GO:0098703) are highlighted in orange. **(b)** Colored by log_2(Fold change) in expression when exposed to heat [20] (genes not in data are colored gray). Previously uncharacterized proteins are annotated with gold stars. **(c)** Human-readable PHILHARMONIC output for selected cluster. **(d)** Enriched GO terms identified by g:Profiler [50]

other species allows the generation of useful networks [85], but standalone these networks are too noisy to apply directly to function prediction. To address these limitations, we have developed PHILHARMONIC. Our approach combines protein language models, spectral algorithms, and traditional HMM-based sequence algorithms to generate functionally meaningful and biologically interpretable clusters. We validate the approach using the literature, gene expression data, and we couple our methods with protein 3D structure predictions to investigate several previously uncharacterized proteins related to thermal and environmental response in the coral *P. damicornis* that it would not be possible to annotate based on sequence or structural information alone. We show that our results extend beyond a single species by applying PHILHARMONIC to the algal symbiont *C. goreaui* and the well-annotated fruit fly *D. melanogaster*.

There still remain several challenges and opportunities to improve the understanding of functional networks in non-model organisms. We note that our predicted networks are 6-10x as dense as the signal-to-noise ratio (SNR) estimated in true PPI networks (1:1000, [105]); thus we still likely have false positive edges and spurious connections. Although our downstream clustering approach will help with denoising, or a higher threshold for interaction could be selected to more closely match this SNR, improvements in network inference will ultimately have the largest impact on downstream performance. Structure-based PPI methods such as AlphaFold-Multimer [29] remain too slow for most labs to apply at interactome scale, but any sufficiently fast PPI prediction method can be substituted into PHILHARMONIC, and should improve its performance as the accuracy of these fast PPI prediction methods advance. Moreover, it is well known that protein interaction networks differ across tissues [32], and that sub-cellular localization likewise plays a role in protein interaction [104]. Our current approach assumes a single PPI network, and orthogonal information such as localization prediction [92], gene co-expression, or tissue type could help further refine analysis of functional networks. Non-model organisms play a crucial role in our world, from the impact of coral on reef ecosystems [74], to the value of livestock and crops in agriculture, to the potential of microbes to combat climate change through carbon capture [36] or digestion of plastics [25]. The broad applicability and ease of use of PHILHARMONIC will enable domain experts to organize the proteome into functional clusters, to identify testable functional hypotheses, and to dissect and understand the complex cellular systems of these organisms.

Acknowledgements We thank NSF HDR: DIRSE-IL, Grants No. 1939249, No. 1940169, No. 1939699, No. 1939795, and No. 1939263 which supported this research when it was just beginning. S.S.'s early work on this project was supported by the NSF Graduate Research Fellowship under Grant No. 2141064. K.D. and L.C. were additionally supported by NSF No. 1934553. C.V. thanks NSF DIAMONDS REU No. 2149871. B.B. and S.S. were supported by the NIH grant No. R35GM141861. We thank Chris Edelmaier, Zhicheng Pan, Natalie Sauerwald, and Bargeen Turzo for helpful discussions and feedback on the manuscript.

References

[1] Seyed Ziaeddin Alborzi, David W Ritchie, and Marie-Dominique Devignes. Computational discovery of direct associations between GO terms and protein domains. *BMC Bioinformatics*, 19(14):53–66, 2018.

[2] Eivind Almaas, Alexei Vázquez, and Albert-László Barabási. Scale-free networks in biology. *Biological networks*, 3(1), 2007.

[3] 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.

[4] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[5] Emma C Baker, Michael J Layden, Damian B Van Rossum, Bishoy Kamel, Monica Medina, Eboni Simpson, and Timothy Jegla. Functional characterization of Cnidarian HCN channels points to an early evolution of i_h. *PloS one*, 10(11):e0142730, 2015.

[6] Sandra Beer-Hammer, Sze Chim Lee, Stephanie A Mauriac, Veronika Leiss, Isabel AM Groh, Ana Novakovic, Roland P Piekorz, Kirsten Bucher, Chengfang Chen, Kun Ni, et al. Gαi proteins are indispensable for hearing. *Cellular Physiology and Biochemistry*, 47(4):1509–1532, 2018.

[7] Pedro Beltrao and Luis Serrano. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS computational biology*, 3(2):e25, 2007.

[8] Eduardo E Benarroch. HCN channels: function and clinical implications. *Neurology*, 80(3):304–310, 2013.

[9] Charles Birkeland. *Coral reefs in the anthropocene*. Springer, 2015.

[10] Anneke C Blackburn, Klaus I Matthaei, Cindy Lim, Matthew C Taylor, Jean Y Cappello, John D Hayes, MW Anders, and Philip G Board. Deficiency of glutathione transferase zeta causes oxidative stress and activation of antioxidant response pathways. *Molecular pharmacology*, 69(2):650–657, 2006.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] Luke Brander and Pieter van Beukering. The total economic value of US coral reefs: a review of the literature. *United States, National Oceanic and Atmospheric Administration,;Coral Reef Conservation Program (U.S.);National Marine Protected Areas Center (U.S.)*, 2013.

[13] Mengfei Cao, Christopher M Pietras, Xian Feng, Kathryn J Doroschak, Thomas Schaffner, Jisoo Park, Hao Zhang, Lenore J Cowen, and Benjamin J Hescott. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, 30(12):i219–i227, 2014.

[14] 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.

[15] Sandra Casani, Rocío Gómez-Pastor, Emilia Matallana, and Nuria Paricio. Antioxidant compound supplementation prevents oxidative damage in a Drosophila model of Parkinson's disease. *Free Radical Biology and Medicine*, 61:151–160, 2013.

[16] Bolin Chen, Weiwei Fan, Juan Liu, and Fang-Xiang Wu. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Briefings in bioinformatics*, 15(2):177–194, 2014.

[17] 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.

[18] Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, et al. Assessment of network module identification across complex diseases. *Nature Methods*, 16(9):843–852, 2019.

[19] Peter Christ, Sharon R Hill, Joachim Schachtner, Frank Hauser, and Rickard Ignell. Functional characterization of the dual allatostatin-A receptors in mosquitoes. *Peptides*, 99:44–55, 2018.

[20] Michael T Connelly, Crystal J McRae, Pi-Jen Liu, Cecily E Martin, and Nikki Traylor-Knowles. Antibiotics alter Pocillopora coral-Symbiodiniaceae-bacteria interactions and cause microbial dysbiosis during heat stress. *Frontiers in Marine Science*, 8:814124, 2022.

[21] Gennaro Cordasco and Luisa Gargano. Community detection via semi-synchronous label propagation algorithms. In *2010 IEEE international workshop on: business applications of social network analysis (BASNA)*, pages 1–8. IEEE, 2010.

[22] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[23] Lenore J Cowen and Hollie M Putnam. Bioinformatics of corals: Investigating heterogeneous omics data from coral holobionts for insight into reef health and resilience. *Annual Review of Biomedical Data Science*, 5:205–231, 2022.

[24] R Cunning, RA Bay, P Gillette, AC Baker, and N Traylor-Knowles. Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral evolution. *Scientific reports*, 8(1):16134, 2018.

[25] Dominik Danso, Jennifer Chow, and Wolfgang R Streit. Plastics: environmental and biotechnological perspectives on microbial degradation. *Applied and environmental microbiology*, 85(19):e01095–19, 2019.

[26] Ying Deng, Weiyi Zhang, Katja Farhat, Sonja Oberland, Günter Gisselmann, and Eva M Neuhaus. The stimulatory Gαs protein is involved in olfactory signal transduction in Drosophila. *PLoS One*, 6(4):e18605, 2011.

[27] Sean R Eddy. Accelerated profile HMM searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[28] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The Pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2019.

[29] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv*, pages 2021–10, 2021.

[30] Stephen Ficklin. Func-e. `https://github.com/SystemsGenetics/FUNC-E`, 2023.

[31] Lude Franke, Harm Van Bakel, Like Fokkens, Edwin D De Jong, Michael Egmont-Petersen, and Cisca Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, 78(6):1011–1025, 2006.

[32] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, 2015.

[33] Steve Gregory. A fast algorithm to find overlapping communities in networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 408–423. Springer, 2008.

[34] Qiuxiang Gu, Jinglin Wu, Yao Tian, Shanshan Cheng, Zi Chao Zhang, and Junhai Han. Gαq splice variants mediate phototransduction, rhodopsin synthesis, and retinal integrity in Drosophila. *Journal of Biological Chemistry*, 295(17):5554–5563, 2020.

[35] Yushuang Guo, Changjun Huang, Yan Xie, Fengming Song, and Xueping Zhou. A tomato glutaredoxin gene SlGRX1 regulates plant responses to oxidative, drought and salt stresses. *Planta*, 232:1499–1509, 2010.

[36] Natalie Hicks, Unni Vik, Peter Taylor, Efthymios Ladoukakis, Joonsang Park, Frangiskos Kolisis, and Kjetill S Jakobsen. Using prokaryotes for carbon capture storage. *Trends in biotechnology*, 35(1):22–32, 2017.

[37] Ove Hoegh-Guldberg, Elvira S Poloczanska, William Skirving, and Sophie Dove. Coral reef ecosystems under climate change and ocean acidification. *Frontiers in Marine Science*, 4:158, 2017.

[38] Alexandre Hollocou, Thomas Bonald, and Marc Lelarge. Improving pagerank for local community detection. *arXiv preprint arXiv:1610.08722*, 2016.

[39] Alexandre Hollocou, Thomas Bonald, and Marc Lelarge. Multiple local community detection. *ACM SIGMETRICS Performance Evaluation Review*, 45(3):76–83, 2018.

[40] Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *ArXiv*, 2023.

[41] Terry P Hughes, Kristen D Anderson, Sean R Connolly, Scott F Heron, James T Kerry, Janice M Lough, Andrew H Baird, Julia K Baum, Michael L Berumen, Tom C Bridge, et al. Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science*, 359(6371):80–83, 2018.

[42] Sohyun Hwang, Chan Yeong Kim, Sunmo Yang, Eiru Kim, Traver Hart, Edward M Marcotte, and Insuk Lee. Humannet v2: human gene networks for disease research. *Nucleic acids research*, 47(D1):D573–D580, 2019.

[43] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2464–2477, 2015.

[44] Jian Jing, Ferdinand S Vilim, Jin-Sheng Wu, Ji-Ho Park, and Klaudiusz R Weiss. Concerted GABAergic actions of Aplysia feeding interneurons in motor program specification. *Journal of Neuroscience*, 23(12):5283–5294, 2003.

[45] 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.

[46] Takuya Kanai, Kenji Takahashi, and Hideshi Inoue. Three distinct-type glutathione S-transferases from Escherichia coli important for defense against oxidative stress. *Journal of biochemistry*, 140(5):703–711, 2006.

[47] Alex C Keene and Erik R Duboue. The origins and evolution of sleep. *Journal of Experimental Biology*, 221(11):jeb159533, 2018.

[48] Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.

[49] Vikram Khurana, Jian Peng, Chee Yeun Chung, Pavan K Auluck, Saranna Fanning, Daniel F Tardiff, Theresa Bartels, Martina Koeva, Stephen W Eichhorn, Hadar Benyamini, et al. Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. *Cell systems*, 4(2):157–170, 2017.

[50] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research*, 51(W1):W207–W212, 2023.

[51] Mickey Kosloff, Natalie Elia, Tamar Joel-Almagor, Rina Timberg, Troy D Zars, David R Hyde, Baruch Minke, and Zvi Selinger. Regulation of light-dependent gqα translocation and morphological changes in fly photoreceptors. *The EMBO Journal*, 2003.

[52] Lokender Kumar, Nathanael Brenner, Samuel Sledzieski, Monsurat Olaosebikan, Liza M Roger, Matthew Lynn-Goin, Roshan Klein-Seetharaman, Bonnie Berger, Hollie Putnam, Jinkyu Yang, et al. Transfer of knowledge from model organisms to evolutionarily distant non-model organisms: The coral Pocillopora damicornis membrane signaling receptome. *PloS one*, 18(2):e0270965, 2023.

[53] Todd C LaJeunesse, John Everett Parkinson, Paul W Gabrielson, Hae Jin Jeong, James Davis Reimer, Christian R Voolstra, and Scott R Santos. Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Current Biology*, 28(16):2570–2580, 2018.

[54] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics*, 11(3):033015, 2009.

[55] Adolfo Lara, Benjamin T Simonson, Joseph F Ryan, and Timothy Jegla. Genome-scale analysis reveals extensive diversification of voltage-gated K+ channels in stem Cnidarians. *Genome biology and evolution*, 15(3):evad009, 2023.

[56] Chia-Wen Li, Shu-Hong Lee, Pei-Shan Chieh, Choun-Sea Lin, Yu-Chie Wang, and Ming-Tsair Chan. Arabidopsis root-abundant cytosolic methionine sulfoxide reductase B genes MsrB7 and MsrB8 are involved in tolerance to oxidative stress. *Plant and Cell Physiology*, 53(10):1707–1719, 2012.

[57] Min Li, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. Modifying the dpclus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics*, 9:1–16, 2008.

[58] Qing Li, Haishen Wen, Yun Li, Zhanxiong Zhang, Yangyang Zhou, and Xin Qi. Evidence for the direct effect of the NPFF peptide on the expression of feeding-related factors in spotted sea bass (Lateolabrax maculatus). *Frontiers in endocrinology*, 10:545, 2019.

[59] Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Stærfeldt, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature methods*, 14(1):61–64, 2017.

[60] Yi Jin Liew, Manuel Aranda, and Christian R Voolstra. Reefgenomics.org: A repository for marine genomics data. *Database*, 2016:baw152, 2016.

[61] Chuang Liu, Yifang Ma, Jing Zhao, Ruth Nussinov, Yi-Cheng Zhang, Feixiong Cheng, and Zi-Ke Zhang. Computational network biology: data, models, and applications. *Physics Reports*, 846:1–66, 2020.

[62] Huanle Liu, Timothy G Stephens, Raúl A González-Pech, Victor H Beltran, Bruno Lapeyre, Pim Bongaerts, Ira Cooke, Manuel Aranda, David G Bourne, Sylvain Forêt, et al. Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Communications biology*, 1(1):95, 2018.

[63] Kimberly Matulef and William N Zagotta. Cyclic nucleotide-gated ion channels. *Annual review of cell and developmental biology*, 19(1):23–44, 2003.

[64] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

[65] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable data analysis with Snakemake. *F1000Research*, 10, 2021.

[66] Martina Montanari, Gérard Manière, Martine Berthelot-Grosjean, Yves Dusabyinema, Benjamin Gillet, Yaël Grosjean, C Léopold Kurz, and Julien Royet. Larval microbiota primes the Drosophila adult gustatory response. *Nature Communications*, 15(1):1341, 2024.

[67] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(3):036104, 2006.

[68] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 849–856, 2001.

[69] Motohiro Nishida, Yoshiko Maruyama, Rie Tanaka, Kenji Kontani, Taku Nagao, and Hitoshi Kurose. Gαi and gαo are target proteins of reactive oxygen species. *Nature*, 408(6811):492–495, 2000.

[70] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The BioGRID interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2019.

[71] Stephan Pflugmacher and Heinrich Sandermann Jr. Cytochrome p450 monooxygenases for fatty acids and xenobiotics in marine macroalgae. *Plant physiology*, 117(1):123–128, 1998.

[72] Yves Pichon. *Comparative molecular neurobiology*, volume 63. Birkhäuser, 2013.

[73] Olaf Pongs. Voltage-gated potassium channels: from hyperexcitability to excitement. *FEBS letters*, 452(1-2):31–35, 1999.

[74] Hollie M Putnam, Katie L Barott, Tracy D Ainsworth, and Ruth D Gates. The vulnerability and resilience of reef-building corals. *Current Biology*, 27(11):R528–R540, 2017.

[75] Hanny E Rivera and Sarah W Davies. Symbiosis maintenance in the facultative coral, oculina arbuscula, relies on nitrogen cycling, cell cycle modulation, and immunity. *Scientific Reports*, 11(1):21226, 2021.

[76] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):1–26, 2019.

[77] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[78] Jori O Ruuskanen, Nina Peitsaro, Jan VM Kaslin, Pertti Panula, and Mika Scheinin. Expression and function of α_2-adrenoceptors in zebrafish: drug effects, mRNA and receptor distributions. *Journal of neurochemistry*, 94(6):1559–1569, 2005.

[79] Ana M Sánchez-Riego, Luis López-Maury, and Francisco J Florencio. Glutaredoxins are essential for stress adaptation in the cyanobacterium Synechocystis sp. PCC 6803. *Frontiers in plant science*, 4:63520, 2013.

[80] Maria Domenica Sanna, Carla Ghelardini, and Nicoletta Galeotti. Differential contribution of gαi/o subunits in the response to food deprivation. *European Journal of Pharmacology*, 750:27–31, 2015.

[81] Bina Santoro and Gareth R Tibbs. The HCN gene family: molecular basis of the hyperpolarization-activated pacemaker channels. *Annals of the New York Academy of Sciences*, 868(1):741–764, 1999.

[82] Leah V Schaffer and Trey Ideker. Mapping the multiscale structure of biological systems. *Cell systems*, 12(6):622–635, 2021.

[83] Chong Shou, Nitin Bhardwaj, Hugo YK Lam, Koon-Kiu Yan, Philip M Kim, Michael Snyder, and Mark B Gerstein. Measuring the evolutionary rewiring of biological networks. *PLoS computational biology*, 7(1):e1001050, 2011.

[84] 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, 2022.

[85] 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.

[86] Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model. *bioRxiv*, 2021.

[87] Sergii Snigirov and Sergiy Sylantyev. The regulatory role of GABA_a receptor in $a$ctinia equina nervous system and the possible effect of global ocean acidification. *Pflügers Archiv-European Journal of Physiology*, 473(12):1851–1858, 2021.

[88] Timothy G Stephens, Mark A Ragan, Debashish Bhattacharya, and Cheong Xin Chan. Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. *Scientific reports*, 8(1):17175, 2018.

[89] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, 49(D1):D605–D612, 01 2021.

[90] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[91] Joseph Szymborski and Amin Emad. RAPPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. *Bioinformatics*, 38(16):3958–3967, 2022.

[92] Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, 50(W1):W228–W234, 2022.

[93] Rebecca Vega Thurber, Dana Willner-Hall, Beltran Rodriguez-Mueller, Christelle Desnues, Robert A Edwards, Florent Angly, Elizabeth Dinsdale, Linda Kelly, and Forest Rohwer. Metagenomic analysis of stressed coral holobionts. *Environmental microbiology*, 11(8):2148–2163, 2009.

[94] Jim Thurmond, Joshua L Goodman, Victor B Strelets, Helen Attrill, L Sian Gramates, Steven J Marygold, Beverley B Matthews, Gillian Millburn, Giulia Antonazzo, Vitor Trovisco, et al. FlyBase 2.0: the next generation. *Nucleic acids research*, 47(D1):D759–D765, 2019.

[95] C Toledo-Hernández and CP Ruiz-Diaz. The immune responses of the coral. *Invertebrate Survival Journal*, 11(1):319–328, 2014.

[96] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[97] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966–967, 2016.

[98] Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, 2022.

[99] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, pages 1–4, 2023.

[100] Christian Robert Voolstra, David John Miller, Mark A Ragan, Ary Hoffmann, Ove Hoegh-Guldberg, David Bourne, Eldon Ball, Hua Ying, Sylvain Foret, Shunichi Takahashi, et al. The ReFuGe 2020 Consortium: Using "omics" approaches to explore the adaptability and resilience of coral holobionts to environmental change. *Frontiers in Marine Science*, 2:68, 2015.

[101] Simon Willison. llm. `https://github.com/simonw/llm`, 2024.

[102] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.

[103] Yufeng Yang, Stephan Gehrke, Md Emdadul Haque, Yuzuru Imai, Jon Kosek, Lichuan Yang, M Flint Beal, Isao Nishimura, Kazumasa Wakamatsu, Shosuke Ito, et al. Inactivation of Drosophila DJ-1 leads to impairments of oxidative stress response and phosphatidylinositol 3-kinase/Akt signaling. *Proceedings of the National Academy of Sciences*, 102(38):13670–13675, 2005.

[104] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[105] Jing Zhang, Ian R. Humphreys, Jimin Pei, Jinuk Kim, Chulwon Choi, Rongqing Yuan, Jesse Durham, Siqi Liu, Hee-Jung Choi, Minkyung Baek, David Baker, and Qian Cong. Computing the human interactome. *bioRxiv*, 2024.

[106] Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems*, 31, 2018.