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[††] Judith Klein-Seetharaman [¶ ◦]
Bonnie Berger[‡‡◦] Lenore Cowen[† ◦]

Abstract

Despite the widespread availability of genome sequencing pipelines, many genes remain part of the genome's "dark matter," where existing inference tools cannot even begin to guess the biological function of their proteins from sequence alone. This challenge is especially pronounced in organisms that are highly evolutionarily distant from well-studied models, where homology-based methods break down. Here, we describe PHILHARMONIC, a computational method that combines deep learning–based *de novo* protein interaction network inference with robust unsupervised spectral clustering and remote homology to illuminate functional organization in any non-model organism. From only a sequenced proteome, we show PHILHARMONIC predicts protein functions, functional communities, and higher-order network structure with high accuracy. We validate its performance using experimental gene expression and pathway data in *D. melanogaster*, and we demonstrate its broad utility by analyzing temperature sensing and stress response pathways in the reef-building coral *P. damicornis* and its algal symbiont *C. goreaui*. PHILHARMONIC provides a general-purpose engine for functional discovery and biological hypothesis generation in non-model organisms, enabling systems-level insights across the full diversity of life.

[*]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: judith.klein-seetharaman@asu.edu, bab@csail.mit.edu, cowen@cs.tufts.edu

Protein-protein interaction networks are a fundamental resource for modeling molecular and cellular 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. Common functional inference workflows leverage known protein-protein interaction (PPI) networks to cluster genes and proteins into functionally-enriched modules [1–3], or use network-propagation to directly infer protein labels based on the graph-theoretic structure [4, 5]. However, the overwhelming majority of experimentally-determined physical protein 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 databases such as BioGRID [6] and STRING [7], much less a network complete enough to enable the analysis of cellular function.

There was likewise no existing computational method that could construct a sufficiently useful network. It is well-known that even over relatively short evolutionary distances there is substantial network re-wiring [8]. Even when sequence and protein structure conservation can allow confident ortholog assignment from a well-studied species, interaction patterns are often not preserved. 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 physical PPIs directly from protein amino acid sequence are either too slow to make genome-wide predictions [9–12] or are not accurate enough to use the raw predicted network as ground truth interactions [13–15].

Here, we introduce **PHILHARMONIC**, the first method that successfully enables systems-level functional genomics in non-model organisms and accurately infers protein functions and communities genome-wide. With PHILHARMONIC, we annotate proteins that are too evolutionary distant from model organisms for existing methods to handle. Our key insight is that although protein-protein interaction networks predicted by high-throughput computational methods may not reach the level of experimental correctness, they nonetheless contain enough signal so that with innovative downstream processing, these *de novo* networks can serve as an information-rich scaffold, enabling pathway-level network analysis and functional genomics. We validate our method using experimental data and gold-standard functions in the fruit fly, and then show that PHILHARMONIC makes meaningful predictions of protein function and credibly sketches interacting modules in coral and algae. PHILHARMONIC provides annotations across hundreds of millions of years of evolution, shedding light on the "dark proteome," [16] and bringing powerful systems biology approaches to non-model organisms.

Results

Overview of PHILHARMONIC

To overcome the lack of experimental functional data, we deploy lightweight, protein language model-based PPI prediction methods at the genome scale and build upon robust methods for remote homology functional inference. Then, our novel clustering and reconnection approach allows us to de-noise the predicted network, producing highly informative functional modules. This modular decomposition of the network then enables functional assignment at a community level, illuminating the function of proteins outside the reach of traditional homology-based methods.

PHILHARMONIC (Protein Human-transferred Interactome Learns Homology And Recapitulates Model Organism Network Interaction Clusters) requires only a set of protein sequences to perform *de novo* inference of functional communities (Figure 1). First, we use the deep learning method D-SCRIPT [17] to infer the initial, noisy PPI network. We develop a novel hierarchical clustering algorithm based on the Double Spectral method, which couples a Diffusion State Distance (DSD)-based similarity metric [18] with several rounds of recursive spectral clustering [19] to convert the PPI network into putative functionally-enriched non-overlapping clusters of balanced sizes that are amenable to easy biological analysis. These clusters are by definition non-overlapping and can be highly disconnected, as opposed to in real biological systems where proteins often play many functional roles. To rectify this, we develop a novel method, ReCIPE (Reconnecting Clusters In Protein Embedding) that creates coherent, better connected, and more explainable clusters by re-adding high degree nodes with a greedy algorithm, thus optimizing intra-cluster connectivity. We leverage robust remote homology methods that assign sequence-based gene ontology (GO) annotations to proteins based on Pfam domains [20], 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 [21]. We demonstrate here that these annotations can be successfully transferred to non-annotated proteins to infer larger functional pathways when combined with functionally-enriched clusters. Finally, we summarize the function of each cluster in paragraph form using generative AI to create easily readable summaries of cluster functions and to add biological interpretability. PHILHARMONIC provides a user-friendly Python package and Colab notebook that seamlessly runs the entire pipeline end-to-end, requiring only a sequenced proteome. We have designed PHILHARMONIC to be accessible to users even with limited programming expertise. Additional technical details are provided in Methods, and the full workflow is shown in Figure A1. PHILHARMONIC is available as a free and open-source tool that can be easily applied to a wide-variety of non-model organisms at `https://github.com/samsledje/philharmonic`.

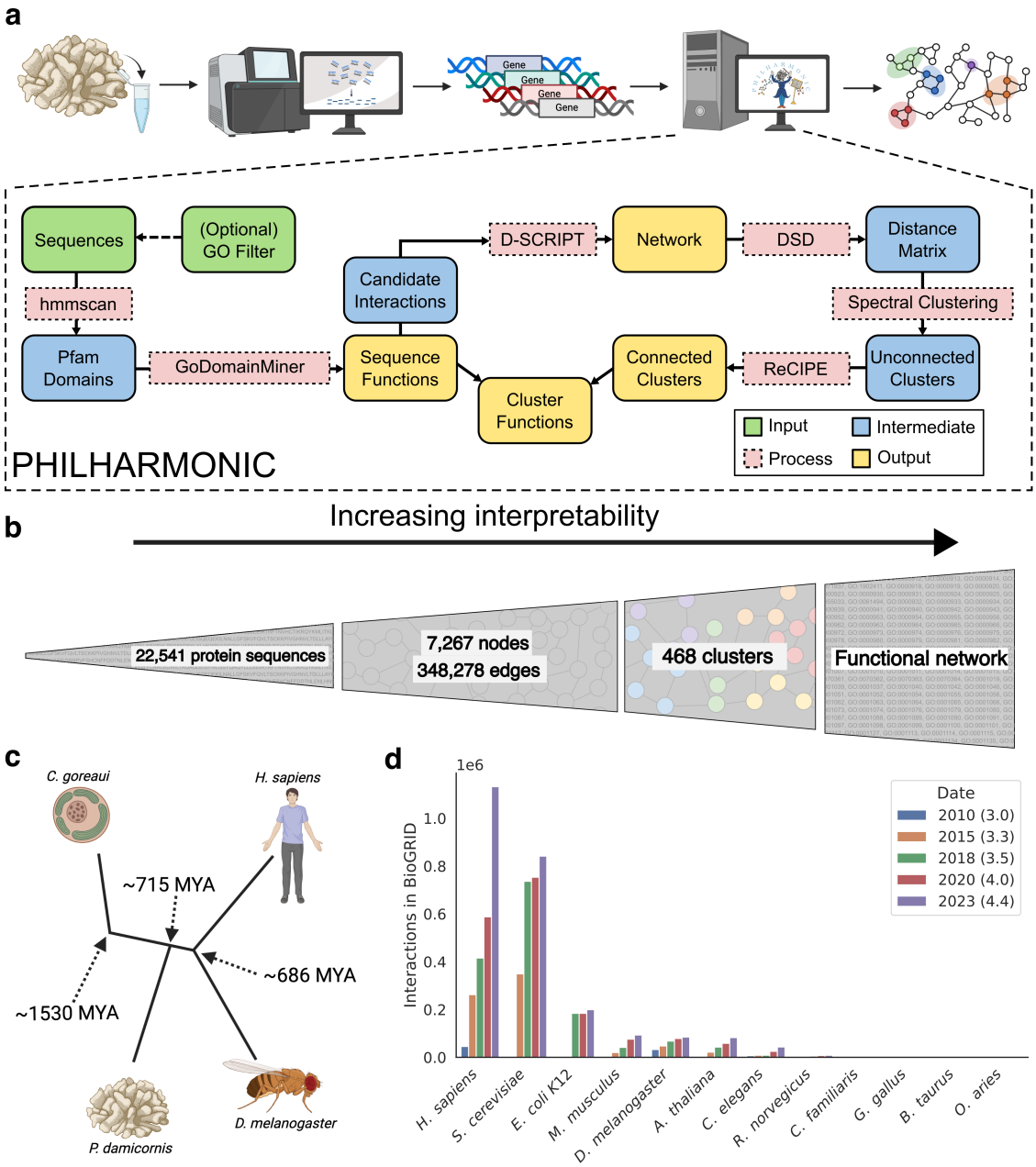

Figure 1: PHILHARMONIC is an engine for biological discovery in non-model organisms. (a) We address the problem of functional network inference in non-model organisms with PHILHARMONIC. Given a newly-sequenced proteome from a non-model organism, we use D-SCRIPT [22] to predict a *de novo* protein-protein interaction network (optionally on a smaller set of GO-filtered proteins, if computational cost is an issue). We cluster this network using DSD [23], hierarchical spectral clustering [19], and ReCIPE (Figure 4a). We use hmmscan [24] and GoDomainMiner [25] to map sequence functions, then aggregate the *individual* sequence functions to assign *cluster* functions, which allows for network-level discovery and hypothesis generation. **(b)** PHILHARMONIC requires only protein sequences to run, which on their own provide little functional information. Starting with 22,541 *P. damicornis* sequences, we integrate several other sources of information, denoising, and downstream processing to build up to a much more interpretable functional network with meaningful communities. **(c)** For species like the cauliflower coral *P. damicornis* and its algal symbiont *C. goreaui*, hundreds of millions of years of evolutionary distance from well-studied model organisms make naive homology-based transfer of interactions or protein functions difficult. **(d)** We show the number of protein interactions in the BioGRID [6] database for several different organisms over several years (database versions). Nearly all known PPIs are concentrated in six well-studied model organisms, while most species have no known interaction networks.

Dissecting the functional network of the coral *P. damicornis*

Corals are a prime example of the type of non-model organism for which functional systems genomics has traditionally been difficult. Corals are key to the ocean's vast biodiversity and provide significant economic, cultural, and scientific benefits [26–28]. 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 [29]. As a result of anthropogenic activities, coral holobionts are declining rapidly (Appendix 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 [30]. However, these efforts are hampered by the great evolutionary distance of corals to any species with functionally well-annotated genomes. We thus sought to assess whether PHILHARMONIC could generate functionally meaningful clusters in the coral *P. damicornis*. We define functional similarity between two proteins as the Jaccard similarity between the sets of GO [31] 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 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 2c,d), compared to a degree-preserving random clustering constructed by shuffling node function assignments (Figure 2a). We also compare with several other state-of-the-art clustering methods (Appendix A.7.2). Using the most-represented GO Slim term assigned to each cluster, we show the cluster coherence of clusters with each high-level function in Figure 2d. We identify several highly-coherent clusters related to mitotic cell cycle (GO:0000278, green), DNA-templated transcription (GO:0006351, light blue) and its regulation (GO:0006355, dark blue), and inflammatory response (GO:0006954, orange). We perform the same analysis using only GO Slim [31] terms (Appendix A.7.1) and we replicate this analysis on the *C. goreaui* network (Appendix A.8).

We next 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. [32], 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). We find that proteins in PHILHARMONIC clusters are significantly more likely to have correlated expression across these tested conditions ($p = 1.27 \times 10^{-21}$, one-tailed related samples t-test) (Figure 2f,g, Appendix A.7.1). We compute an alternate method of co-expression within a cluster using the singular values of the cluster sub-matrix with similar findings (Appendix A.7.3).

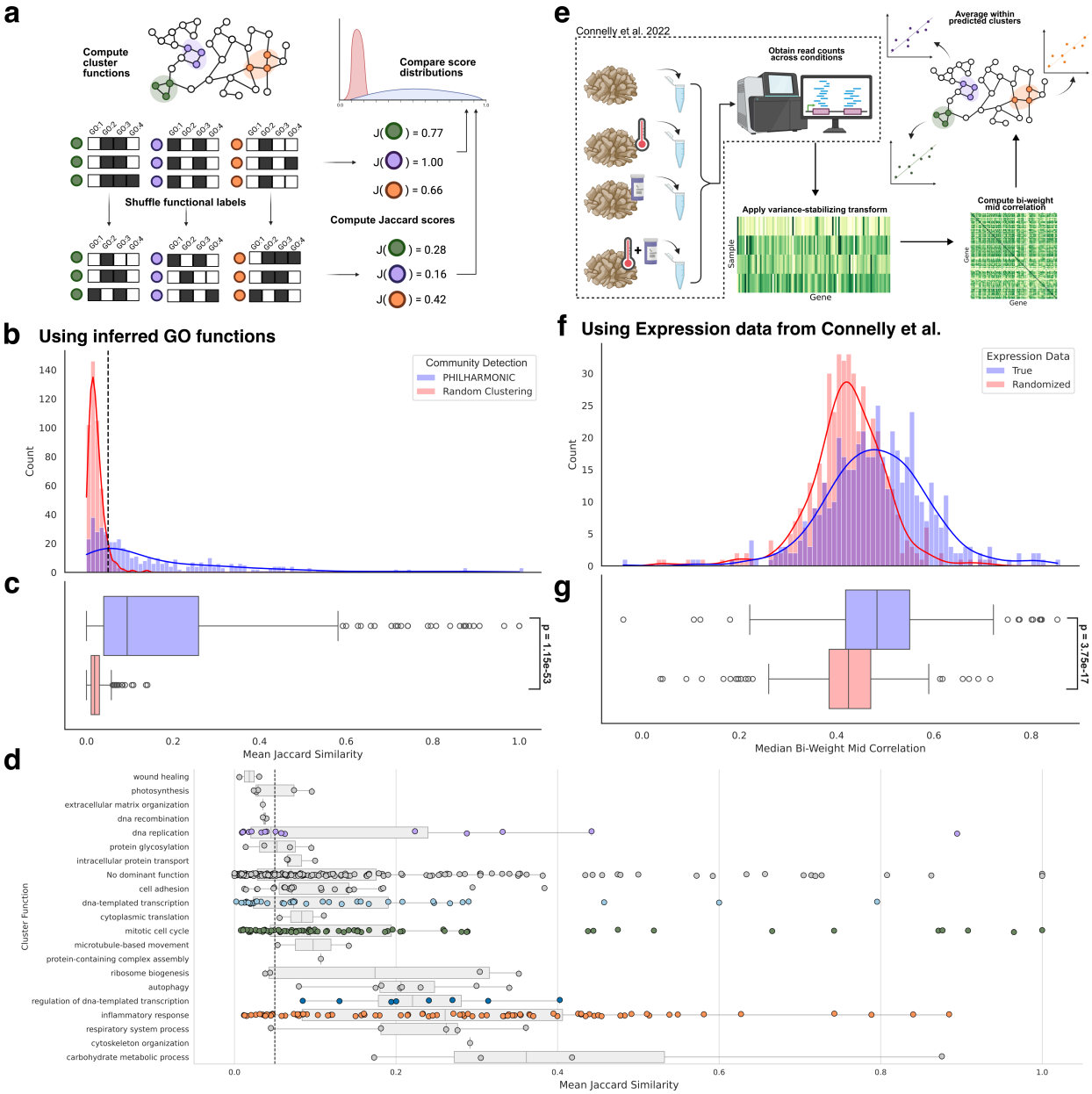

Figure 2: PHILHARMONIC clusters group functionally-related and co-expressed proteins in *P. damicornis*. (a) We compute the functional coherence of a cluster as the mean Jaccard similarity between the sets of GO functional terms assigned to each pair of proteins in the same cluster. We compare these coherence scores to a distribution-preserving random clustering of the network by shuffling functional labels. **(b, c)** 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. This can be used to select high-likelihood clusters even in absence of gold standard functions. **(d)** 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 DNA-templated transcription (GO:0006351, light blue) and its regulation (GO:0006355, dark blue), DNA replication (GO:0006260, purple), mitotic cell cycle (GO:0000278, green), and inflammatory response (GO:0006954, orange). **(e)** Using *Pocillopora* gene expression data from Connelly et al. [32], we compute the median bi-weight mid correlation between pairs of genes in the same cluster. We hypothesize that functionally related genes are more likely to be co-expressed across several conditions and samples. **(f, g)** We find that proteins that share a PHILHARMONIC cluster are significantly more likely to be co-expressed than a baseline with re-shuffled expression data ($p = 3.75 \times 10^{-17}$, one-tailed related t-test).

A global view reveals hierarchies of function

We next analyzed the global accuracy of cluster organization in the *P. damicornis* network. The *P. damicornis* network displays scale-free characteristics [33] typical of biological networks (Figure 3a), and the distribution of cluster sizes is balanced and intentionally constrained by our recursive splitting procedure (Figure 3b, Figure 4a). We construct a graph representation of the network where each node is itself a cluster, which we use to investigate high-level functional organization (Figure 3c, Methods). 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 transcription, gene regulation, and the cell cycle in cellular function [34], and we identify tightly-connected neighborhoods of the overall network dedicated to inflammatory response [35]. We perform a similar high-level analysis of the *C. goreaui* network, including two additional case studies (Appendix A.8).

Clustering identifies the regulation of thermal response

In addition to this global view, we investigated two clusters in depth, both of which involve ion channels communicating with G-coupled protein receptors (GPCRs). We analyze here one PHILHARMONIC cluster involved in cellular response to temperature stimulus (GO:0071502) (Figure 3d). Cnidarians, including coral animals, have a large and diverse family of voltage-gated potassium ion (K^+) channels, but little is known about how the function of these channels differs across different cnidarian species [36]. 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`, [37]), and a hyperpolarization-activated cyclic nucleotide-gated (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 [38]. Cnidarian HCN channels share all major functional features of vertebrate HCNs, including reversed voltage-dependence, activation by cAMP and PIP2, and blocking by extracellular cesium ions [39]. There is some evidence that HCN channels are also involved in rhythmic firing and the so-called pacemaker channels [40], and 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. Two of these, `pdam_00016481-RA` and `pdam_00016115-RA`, are similar to *ADRB1* and *ADRB2* respectively, which are involved in brown adipose tissue differentiation and linked to heat generation in humans [41]. Notably, many of these receptors are linked to neural peptides (Appendix A.7.4), including an orexin receptor (`pdam_00006261`) that has been implicated in systems involving sleep and light-sensing in mammals [42]. `pdam_00001720` resembles

the allatostatin-A receptor, to which the neuropeptide allatostatin-A binds, modulating feeding behavior in mosquitos [43].

We identify one protein that is completely uncharacterized (pdam_00017094, starred in Figure 3d). Not only were we unable to identify homologs through sequence similarity search, but we also could not find structural homologs using FoldSeek [44]. We predict this protein to bind to a subset of the GPCRs, including the putative orexin and allatostatin-A receptors. Using data from Connelly et al. [32], we find that 00017094 is over-expressed during heat stress. We use AlphaFold [9, 45] to predict the structure of 00017094 in complex with 0001720 (Figure 3e). 00017094 contains two likely flexible helices, predicted with low plDDT, that are predicted to sit alongside the barrel in the complex structure. 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. We hypothesize that its binding may regulate the allatostatin-A receptor and coral feeding behavior as a response to heat stress. Due to significant divergence of its sequence and structure, only our network-based approach could have arrived at this hypothesis.

Uncovering the neurophysiological response to environmental stimuli

We compared our community construction approach with several other popular clustering methods, and we find that not only does our approach result in more balanced cluster sizes and interpretable clusters, but that reconnection with ReCIPE actually further increases functional coherence of clusters (Figure 4b, Appendix A.7.2). Following this insight, we next highlight a cluster that has been strongly re-connected by ReCIPE, one enriched for putative neurological receptors and involved in the coral response to environmental stimuli. This cluster contains two proteins with high levels of homology to K^+ voltage-gated channels [46], namely pdam_00001381 in the original cluster and pdam_00006320 among the proteins added in by ReCIPE. Among the remaining proteins include several that are highly homologous to ligand-gated ion channels, previously recognized as key modifiers of behavior in vertebrate brains. We also identify four gamma-aminobutyric acid (GABA) receptors (pdam_00012376-RA, pdam_00001388, pdam_00001389 and pdam_00024051, the last added by ReCIPE), and four proteins that are homologous to nicotinic receptors in vertebrates (pdam_00006972, pdam_00015999, pdam_00006973, pdam_00012507). GABA has been identified as a signaling molecule in feeding behavior, orientation, and tentacle movement in other Cnidarian species [47]. Jing et al. showed that the roles of GABAergic inhibitory interneurons can also be extended to the Aplysia feeding motor network [48]. They are widely expressed in the central nervous system, while, in the periphery, they mediate synaptic transmission at the neuromuscular junction and ganglia. There

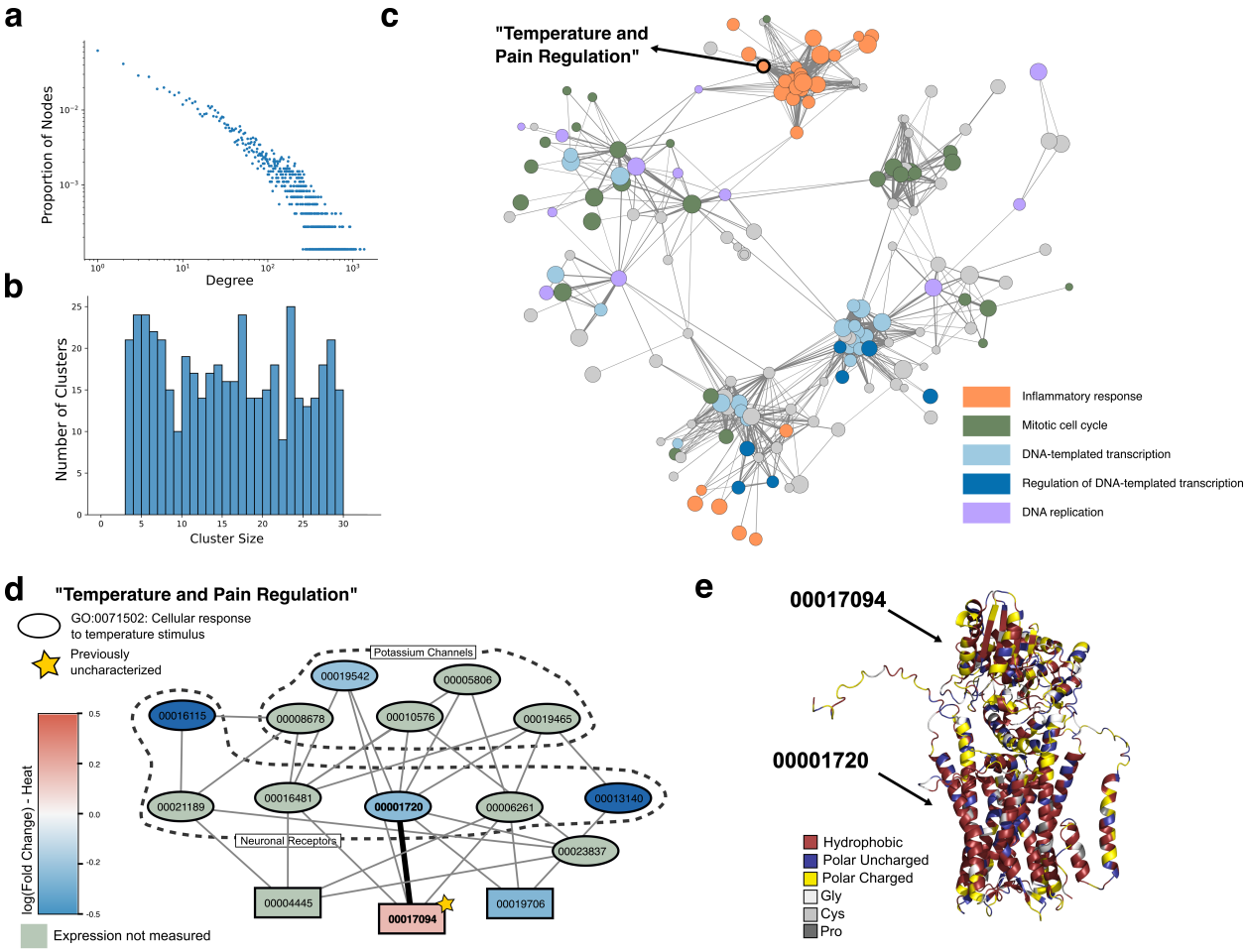

Figure 3: ***De novo*** **network inference and functional clustering in** ***P. damicornis*** **(a)** The synthetic PPI network predicted by PHILHARMONIC follows a scale-free degree distribution commonly found in biological networks [33]. **(b)** PHILHARMONIC yields balanced clusters with sizes between three and 29 nodes (see Figure 4a for more detail). **(c)** The *P. damicornis* cluster graph, where each node is a cluster and edges indicate high cross-connectivity between clusters. For ease of visualization, we have filtered out clusters that do not connect to others with at least 50 crossing edges. We highlight here five high-level cluster functions that were identified by PHILHARMONIC. We find neighborhoods of the network focused on inflammatory response (GO:0006954, orange), mitotic cell cycle (GO:0000278, green), and DNA-templated transcription (GO:0006351, light blue) and its regulation (GO:0006355, dark blue). Clusters focused on DNA replication (GO:0006260) are more dispersed throughout the network. **(d)** We highlight one case study, a cluster involved in temperature and pain regulation in *P. damicornis*. We find that 12 of the 15 proteins in this cluster are annotated with cellular response to temperature stimulus (GO:0071502) (indicated with ovals). Nodes are colored by log_2(Fold change) in expression when exposed to heat [32] (genes not in data are colored green). This cluster contains several putative potassium channels and neuronal receptors, and one previously uncharacterized protein (`pdam_00017094`, gold star). **(e)** AlphaFold-Multimer-predicted [9] structure of the predicted interaction between `pdam_00017094` and `pdam_00001720`. The complex is colored by amino acid chemical properties, highlighting the likely membrane association of the complex.

are three proteins in this cluster which have previously been uncharacterized—pdam_00011550, pdam_00018139, and pdam_00006109, which we highlight in Figure 4c with stars. Two of these, 00011550 and 00006109, are likely interacting with a GABA receptor candidate 00024051 that is strongly under-expressed in heat conditions. We hypothesize that proteins in this cluster are involved in the coral's response to its environment, acting as a sort of nervous system. Because this hypothesis rests on the organization of the cluster rather than any individual protein function, only our network-based approach could have arrived at this hypothesis.

Validated cluster functions in *D. melanogaster*

Precisely because *P. damicornis* is poorly annotated, it is difficult to compare PHILHARMONIC clusters with a "gold-standard," and thus our analysis in coral can only rely on coherence of assigned functions and correlation with gene expression experiments. However, the fruit fly *D. melanogaster* is significantly better characterized through efforts like the FlyBase project [50]. We therefore have access to high-confidence functional annotations, experimentally validated protein interactions, and extensive gene expression data. Thus, we are able to validate PHILHARMONIC in this model system. While we previously used GO function labels assigned by PHILHARMONIC (hmmscan + GODomainMiner), as an external validation we here use known GO function annotations from FlyBase [50]. We likewise find significant functional coherence in PHILHARMONIC clusters ($p = 5.2 \times 10^{-16}$, one-tailed independent t-test) in the fruit fly using these functional annotations (Figure 5e, f). We replicate this analysis using FlyBase pathway membership to evaluate a higher level of functional relatedness (Appendix A.9.2). To evaluate the impact of network noise on coherence, we replicate this analysis using the gold-standard *D. melanogaster* network from the STRING database [51] (Appendix A.9.1) and find that inference using the noisy predicted network is consistent with that using the STRING network. In addition to coherent function, we test whether genes in PHILHARMONIC clusters have correlated expression using dense time-series gene expression data in *D. melanogaster* from Schlamp et al. [52] (Appendix A.9.3). We show that PHILHARMONIC clusters are significantly more likely to have correlated expression than a random baseline ($p = 4.79 \times 10^{-21}$, one-tailed related t-test, Figure 5g,h).

While our primary goal in applying PHILHARMONIC to the fly was to validate its performance, we show that even in a well-studied organism our approach can add new insight that could not be obtained by looking only at individual proteins or clusters. We focus on a cluster of 9 proteins in the *D. melanogaster* network (Figure 5d). This cluster was specifically selected based on the high-level organization seen in the cluster graph (Figure 5c), as it is highly central in the graph and connects a diverse set of functions (a "bridge" cluster, betweenness centrality of 0.0291, 16th out of 285 clusters). We show all nodes that have at least 10 crossing edges to the bridge cluster (Figure 5d). Proteins in this bridge cluster interact with proteins annotated for inflammatory response (orange, GO:0006954), mitotic cell cycle

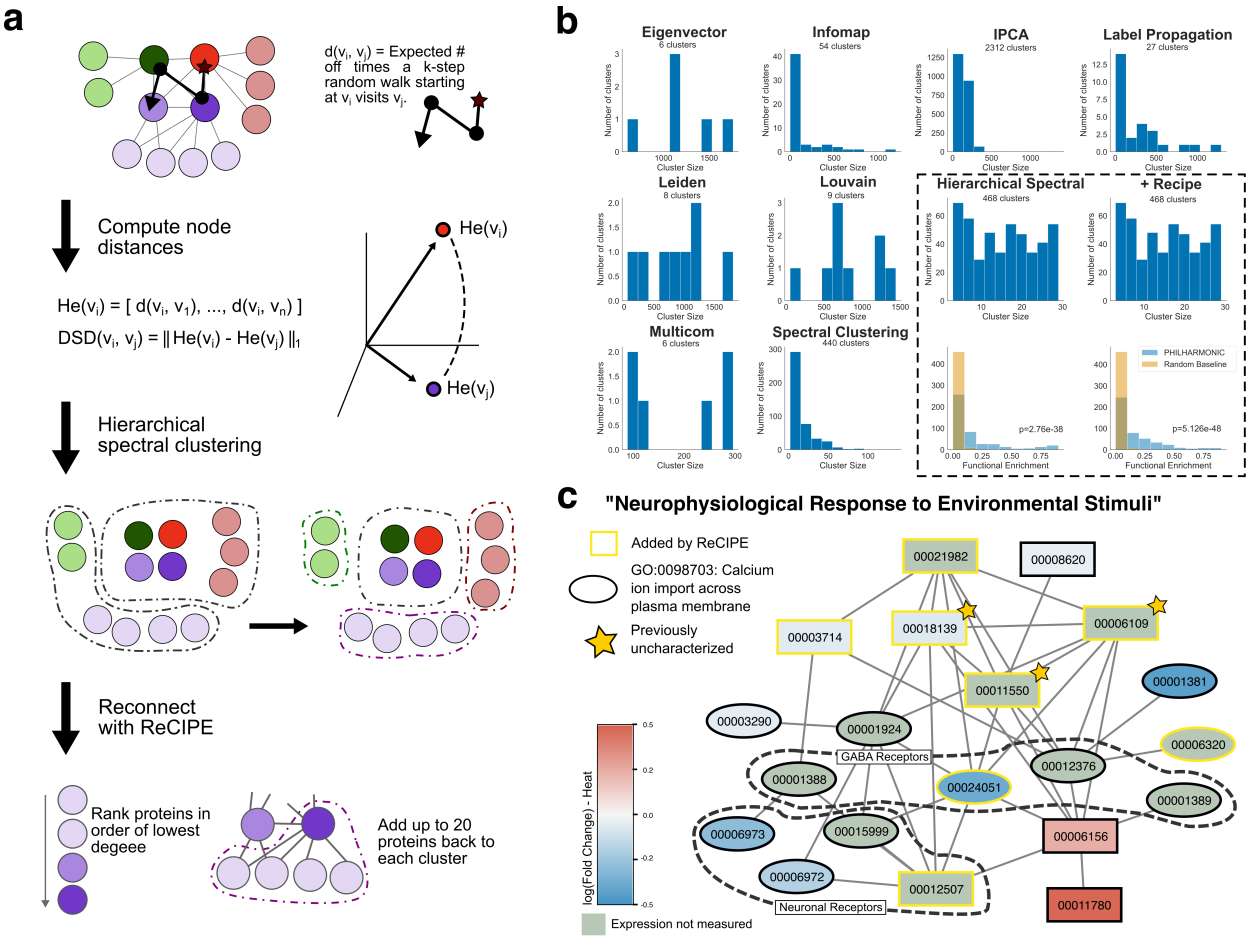

Figure 4: Reconnecting with ReCIPE increases interpretability and functional coherence of clusters. (a) Schematic overview of the DSD + Hierarchial Spectral + ReCIPE sections of PHILHARMONIC. DSD [18] defines pairwise node distances based on the expected distribution of random walks on the graph. We recursively perform spectral clustering on this distance matrix until a cluster size threshold is met. However, this often results in clusters that are difficult to interpret, because there are missing hubs that would connect the proteins. To address this, we use ReCIPE to identify candidate nodes for re-addition, rank them prioritizing low-degree nodes, and add them back greedily until a connectivity threshold or maximum number of proteins is hit. **(b)** We compare our clustering pipeline with several other popular clustering approaches (using CDlib [49]). Our approach achieves much more balanced and interpretable cluster sizes than these alternatives, as well as significantly better functional coherence (Appendix A.7.2). **(c)** We show another case study, a cluster involved in neurophysiological response to environmental stimuli in *P. damicornis*, where ReCIPE has re-added several important proteins to the cluster. Proteins annotated with calcium ion import across the plasma membrane (GO:0098703) are indicated with ovals. Nodes are colored by log_2(Fold change) in expression when exposed to heat [32] (genes not in data are colored green). This cluster contains several putative GABA and neuronal receptors, and three previously uncharacterized proteins (`pdam_00006109`, `pdam_00011550`, `pdam_00018139`).

(green, GO:0000278), extracellular matrix, DNA replication (purple, GO:0006260), and DNA-templated transcription (light blue, GO:0006352) among others. In addition, this cluster is connected to several clusters which themselves contain proteins with diverse functional annotations (no one dominant function).

This bridge cluster is primarily composed of Gα subunits and the Gα-like *concertina*/Gα11/12. In mice, Gαi and Gαo proteins are known to be involved in hearing [53] and give differential response in response to food deprivation [54]. In the fruit fly, the Gαs protein is involved in olfactory signal transduction [55] and Gαq is known to play a role in photoreceptor signaling [56, 57] and the gustatory response [58]. Interestingly, this cluster also contains DJ-1α and DJ-1β, both of which have been shown to convey a neuroprotective role against oxidative stress in multiple contexts [59, 60]. There is some evidence that Gαi and Gαo are target proteins of reactive oxygen species [61], making this cluster a good starting place to begin the search for the mechanism by which this protective role is played. We hypothesize that through diverse interaction partners, the proteins in this cluster are processing environmental signals and routing them to the appropriate systems to respond, acting as conductors of cellular function. In addition to the literature, there is experimental support for this role; several proteins in this cluster have correlated expression time series in response to immune stimulation (Figure 5i).

Discussion

Non-model organisms play a crucial role in our world, from the impact of coral on reef ecosystems [28], to the value of livestock and crops in agriculture, to the potential of microbes to combat climate change through carbon capture [62] or digestion of plastics [63]. Recent research suggests that certain biological mechanisms in these organisms may actually be better conserved with respect to human biology than those from the traditional model organisms [64]. While substantial previous work has advanced our understanding of functional networks in humans [65–69], limited experimental data and the large evolutionary distance from well-studied organisms has made similarly in-depth analysis of functional networks in non-model organisms difficult.

Here, we developed PHILHARMONIC to enable inference of protein interactions and functions across the tree of life. 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 quality of our inferred networks and clusters with substantial support from functional coherence, gene expression, and the literature. These results demonstrate the power of applying a network lens to function, by identifying hierarchies of organization and propagating function throughout the genome. Using PHILHARMONIC, we identify clusters related to thermal and environmental response in the coral *P. damicornis*. For several proteins in these clusters, it was not previously possible to annotate their function using sequence or structural homology alone. PHILHARMONIC allows the first

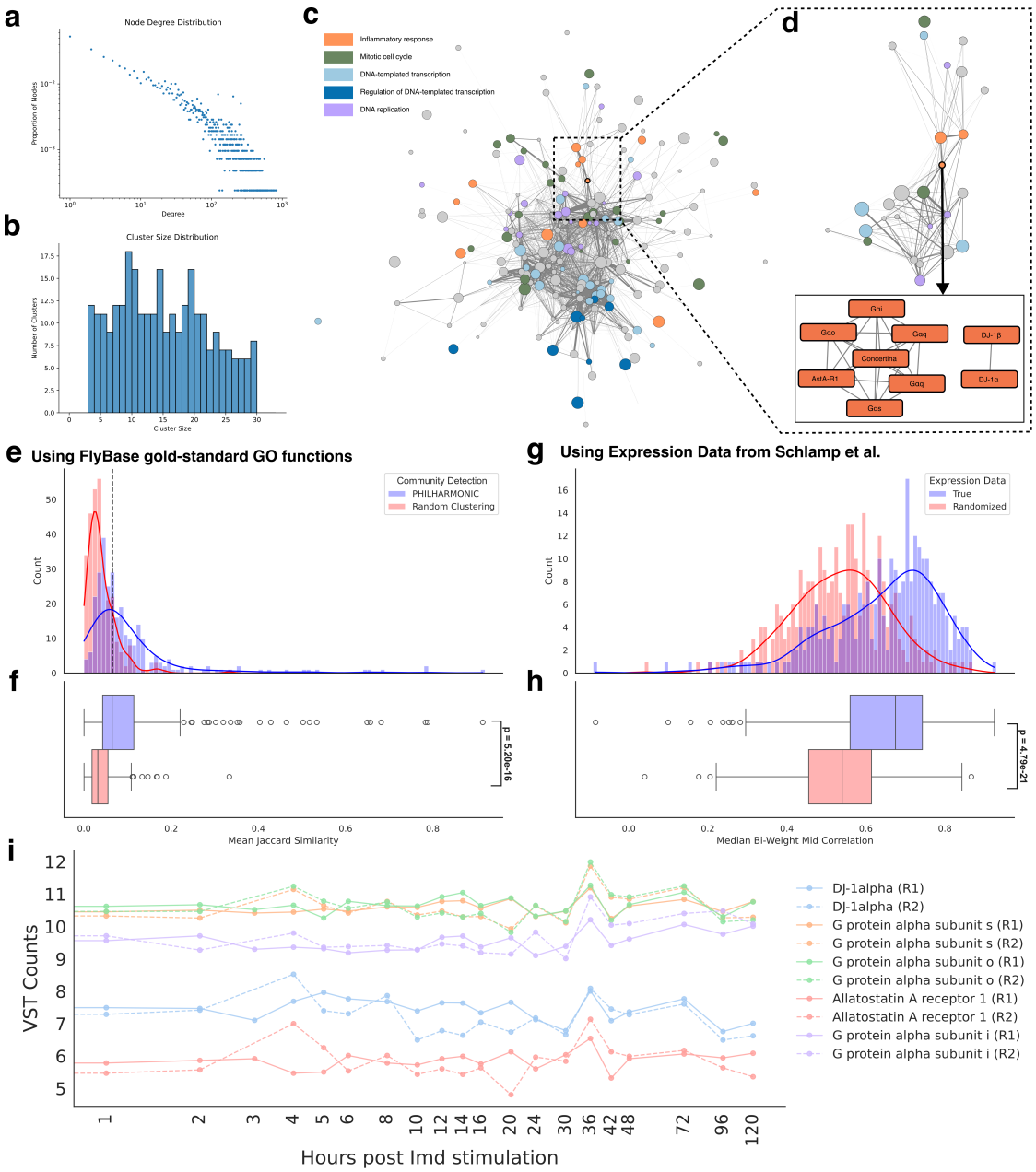

Figure 5: PHILHARMONIC recovers functional clusters in *D. melanogaster*. We show that PHILHARMONIC achieves similar results on the well-studied *D. melanogaster* proteome. This predicted network likewise is scale-free **(a)** with a balanced distribution of cluster sizes **(b)**. **(c)** We observe similar functional "neighborhoods" in the *D. melanogaster* network. We identify a central "bridge" cluster, shown in detail in **(d)**, that has one of the highest betweenness centralities in the cluster graph (Rank 16 out of 285 clusters; Betweeness centrality score: 0.0291). This cluster contains several G protein alpha subunits, as well as AstA-R1, DJ-1α, and DJ-1β, and is connected to a diverse set of functional clusters. We hypothesize that the proteins in this cluster play a crucial role in routing of environmental signals in *D. melanogaster*. **(e, f)** We compute functional coherence of clusters using gold-standard GO functions from FlyBase [50], as a higher-confidence annotation of protein function than Pfam domains as used in coral. PHILHARMONIC clusters contain proteins that are significantly more functionally related than the baseline ($p = 5.20 \times 10^{-16}$, one-tailed independent t-test). **(g, h)** Using time-series gene expression data from Schlamp et al [52], we find that genes in PHILHARMONIC clusters are significantly likely to have correlated patterns of expression ($p = 4.79 \times 10^{-21}$, one-tailed related t-test). **(i)** Members of the "bridge" cluster have correlated time-series expression after stimulation of the immune pathway [52].

functional characterization of these proteins, and we couple our methods with protein structure prediction to more deeply investigate these putative interactions. We show that our results extend beyond a single species by applying PHILHARMONIC not only to *P. damicornis*, but also to the algal symbiont *C. goreaui* and the well-annotated fruit fly *D. melanogaster*. In the fly, gold-standard functional data and independent gene expression data sets confirm the quality of our inference methods. As researchers work to decode the genomes of diverse organisms [70–72], PHILHARMONIC enables a network approach to understanding protein function. 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.

Acknowledgments We thank the US National Science Foundation which supported this research when it was just beginning under the ”Harnessing the Data Revolution” NSF Big Idea: Grants No. 1939249, No. 1940169, No. 1939699, No. 1939795, and No. 1939263. 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. is supported by the NIH grant No. R35GM141861. S.S. is supported by the Simons Foundation. We thank Chris Edelmaier, Zhicheng Pan, Natalie Sauerwald, and Bargeen Turzo for helpful discussions and feedback on the manuscript.

References

1. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nature Methods* **16,** 843–852 (2019).
2. Liu, C. *et al.* Computational network biology: data, models, and applications. *Physics Reports* **846,** 1–66 (2020).
3. Schaffer, L. V. & Ideker, T. Mapping the multiscale structure of biological systems. *Cell systems* **12,** 622–635 (2021).
4. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Molecular Systems biology* **3,** 88 (2007).
5. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18,** 551 (2017).
6. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic acids research* **47,** D529–D541 (2019).
7. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47,** D607–D613 (2019).
8. Beltrao, P. & Serrano, L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS computational biology* **3,** e25 (2007).
9. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv,* 2021–10 (2021).
10. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).
11. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv* (2022).
12. Ketata, M. A. *et al.* Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889* (2023).
13. Chen, M. *et al.* Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **35,** i305–i314 (2019).
14. Singh, R., Devkota, K., Sledzieski, S., Berger, B. & Cowen, L. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38,** i264–i272 (2022).
15. Szymborski, J. & Emad, A. RAPPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. *Bioinformatics* **38,** 3958–3967 (2022).

16. Stephens, T. G., Ragan, M. A., Bhattacharya, D. & Chan, C. X. Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. *Scientific reports* **8,** 17175 (2018).

17. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model. *bioRxiv* (2021).

18. Cao, M. *et al.* Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one* **8,** e76339 (2013).

19. Ng, A. Y., Jordan, M. I. & Weiss, Y. *On spectral clustering: analysis and an algorithm* in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (2001), 849–856.

20. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research* **47,** D427–D432 (2019).

21. Kumar, L. *et al.* Transfer of knowledge from model organisms to evolutionarily distant non-model organisms: The coral Pocillopora damicornis membrane signaling receptome. *PloS one* **18,** e0270965 (2023).

22. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems* **12,** 969–982 (2021).

23. Cao, M. *et al.* New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* **30,** i219–i227 (2014).

24. Eddy, S. R. Accelerated profile HMM searches. *PLoS computational biology* **7,** e1002195 (2011).

25. Alborzi, S. Z., Ritchie, D. W. & Devignes, M.-D. Computational discovery of direct associations between GO terms and protein domains. *BMC Bioinformatics* **19,** 53–66 (2018).

26. Birkeland, C. *Coral reefs in the anthropocene* (Springer, 2015).

27. Brander, L. & Beukering, P. v. 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.)* `https://repository.library.noaa.gov/view/noaa/8951` (2013).

28. Putnam, H. M., Barott, K. L., Ainsworth, T. D. & Gates, R. D. The vulnerability and resilience of reef-building corals. *Current Biology* **27,** R528–R540 (2017).

29. Hoegh-Guldberg, O., Poloczanska, E. S., Skirving, W. & Dove, S. Coral reef ecosystems under climate change and ocean acidification. *Frontiers in Marine Science* **4,** 158 (2017).

30. Cowen, L. J. & Putnam, H. M. 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).

31. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).

32. Connelly, M. T., McRae, C. J., Liu, P.-J., Martin, C. E. & Traylor-Knowles, N. Antibiotics alter Pocillopora coral-Symbiodiniaceae-bacteria interactions and cause microbial dysbiosis during heat stress. *Frontiers in Marine Science* **8,** 814124 (2022).

33. Almaas, E., Vázquez, A. & Barabási, A.-L. Scale-free networks in biology. *Biological networks* **3** (2007).

34. Rivera, H. E. & Davies, S. W. Symbiosis maintenance in the facultative coral, Oculina arbuscula, relies on nitrogen cycling, cell cycle modulation, and immunity. *Scientific Reports* **11,** 21226 (2021).

35. Toledo-Hernández, C. & Ruiz-Diaz, C. The immune responses of the coral. *Invertebrate Survival Journal* **11,** 319–328 (2014).

36. Lara, A., Simonson, B. T., Ryan, J. F. & Jegla, T. Genome-Scale Analysis Reveals Extensive Diversification of Voltage-Gated K+ Channels in Stem Cnidarians. *Genome biology and evolution* **15,** evad009 (2023).

37. Matulef, K. & Zagotta, W. N. Cyclic nucleotide-gated ion channels. *Annual review of cell and developmental biology* **19,** 23–44 (2003).

38. Benarroch, E. E. HCN channels: function and clinical implications. *Neurology* **80,** 304–310 (2013).

39. Baker, E. C. *et al.* Functional characterization of Cnidarian HCN channels points to an early evolution of I_h. *PloS one* **10,** e0142730 (2015).

40. Santoro, B. & Tibbs, G. R. The HCN gene family: molecular basis of the hyperpolarization-activated pacemaker channels. *Annals of the New York Academy of Sciences* **868,** 741–764 (1999).

41. Ishida, Y. *et al.* Genetic evidence for involvement of β2-adrenergic receptor in brown adipose tissue thermogenesis in humans. *International Journal of Obesity* **48,** 1110–1117 (2024).

42. Keene, A. C. & Duboue, E. R. The origins and evolution of sleep. *Journal of Experimental Biology* **221,** jeb159533 (2018).

43. Christ, P., Hill, S. R., Schachtner, J., Hauser, F. & Ignell, R. Functional characterization of the dual allatostatin-A receptors in mosquitoes. *Peptides* **99,** 44–55 (2018).

44. Van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* (2022).

45. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

46. Pongs, O. Voltage-gated potassium channels: from hyperexcitability to excitement. *FEBS letters* **452,** 31–35 (1999).

47. Snigirov, S. & Sylantyev, S. 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,** 1851–1858 (2021).

48. Jing, J., Vilim, F. S., Wu, J.-S., Park, J.-H. & Weiss, K. R. Concerted GABAergic actions of Aplysia feeding interneurons in motor program specification. *Journal of Neuroscience* **23,** 5283–5294 (2003).

49. Rossetti, G., Milli, L. & Cazabet, R. CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science* **4,** 1–26 (2019).

50. Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic acids research* **47,** D759–D765 (2019).

51. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49,** D605–D612 (Jan. 2021).

52. Schlamp, F. *et al.* Dense time-course gene expression profiling of the Drosophila melanogaster innate immune response. *BMC genomics* **22,** 1–22 (2021).

53. Beer-Hammer, S. *et al.* Gαi proteins are indispensable for hearing. *Cellular Physiology and Biochemistry* **47,** 1509–1532 (2018).

54. Sanna, M. D., Ghelardini, C. & Galeotti, N. Differential contribution of Gαi/o subunits in the response to food deprivation. *European Journal of Pharmacology* **750,** 27–31 (2015).

55. Deng, Y. *et al.* The stimulatory Gαs protein is involved in olfactory signal transduction in Drosophila. *PLoS One* **6,** e18605 (2011).

56. Gu, Q. *et al.* Gαq splice variants mediate phototransduction, rhodopsin synthesis, and retinal integrity in Drosophila. *Journal of Biological Chemistry* **295,** 5554–5563 (2020).

57. Kosloff, M. *et al.* Regulation of light-dependent Gqα translocation and morphological changes in fly photoreceptors. *The EMBO Journal* (2003).

58. Montanari, M. *et al.* Larval microbiota primes the Drosophila adult gustatory response. *Nature Communications* **15,** 1341 (2024).

59. Casani, S., Gómez-Pastor, R., Matallana, E. & Paricio, N. Antioxidant compound supplementation prevents oxidative damage in a Drosophila model of Parkinson's disease. *Free Radical Biology and Medicine* **61,** 151–160 (2013).

60. Yang, Y. *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,** 13670–13675 (2005).

61. Nishida, M. *et al.* Gαi and Gαo are target proteins of reactive oxygen species. *Nature* **408,** 492–495 (2000).

62. Hicks, N. *et al.* Using prokaryotes for carbon capture storage. *Trends in biotechnology* **35,** 22–32 (2017).

63. Danso, D., Chow, J. & Streit, W. R. Plastics: environmental and biotechnological perspectives on microbial degradation. *Applied and environmental microbiology* **85,** e01095–19 (2019).

64. Avasthi, P., McGeever, E., Patton, A. H. & York, R. Leveraging evolution to identify novel organismal models of human biology. *Arcadia Science* (2024).

65. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics* **78,** 1011–1025 (2006).

66. Chen, B., Fan, W., Liu, J. & Wu, F.-X. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Briefings in bioinformatics* **15,** 177–194 (2014).

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

68. Khurana, V. *et al.* Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. *Cell systems* **4,** 157–170 (2017).

69. Hwang, S. *et al.* HumanNet v2: human gene networks for disease research. *Nucleic acids research* **47,** D573–D580 (2019).

70. Benoit, M. *et al.* Solanum pan-genetics reveals paralogues as contingencies in crop engineering. *Nature,* 1–11 (2025).

71. Bhattacharya, D. *et al.* Comparative genomics explains the evolutionary success of reef-forming corals. *elife* **5,** e13288 (2016).

72. Li, M. *et al.* De novo assembly of 20 chicken genomes reveals the undetectable phenomenon for thousands of core genes on microchromosomes and subtelomeric regions. *Molecular Biology and Evolution* **39,** msac066 (2022).

73. Cunning, R., Bay, R., Gillette, P., Baker, A. & Traylor-Knowles, N. Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral evolution. *Scientific reports* **8,** 16134 (2018).

74. Liew, Y. J., Aranda, M. & Voolstra, C. R. Reefgenomics.Org: A repository for marine genomics data. *Database* **2016,** baw152 (2016).

75. Voolstra, C. R. *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).

76. Liu, H. *et al.* Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Communications biology* **1,** 95 (2018).

77. Jayasumana, S., Hartley, R., Salzmann, M., Li, H. & Harandi, M. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE transactions on pattern analysis and machine intelligence* **37,** 2464–2477 (2015).

78. Ficklin, S. *FUNC-E* `https://github.com/SystemsGenetics/FUNC-E`. 2023.

79. Hu, M. *et al.* Evaluation of large language models for discovery of gene set function. *Nature Methods* **22,** 82–91 (2025).

80. Willison, S. *llm* `https://github.com/simonw/llm`. 2024.

81. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10** (2021).

82. LaJeunesse, T. C. *et al.* Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Current Biology* **28,** 2570–2580 (2018).

83. Thurber, R. V. *et al.* Metagenomic analysis of stressed coral holobionts. *Environmental microbiology* **11,** 2148–2163 (2009).

84. Hughes, T. P. *et al.* Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359,** 80–83 (2018).

85. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **74,** 036104 (2006).

86. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences* **105,** 1118–1123 (2008).

87. Cordasco, G. & Gargano, L. *Community detection via semi-synchronous label propagation algorithms* in *2010 IEEE international workshop on: business applications of social network analysis (BASNA)* (2010), 1–8.

88. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9,** 1–12 (2019).

89. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008,** P10008 (2008).

90. Zhang, Y. & Rohe, K. Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems* **31** (2018).

91. Gregory, S. *A fast algorithm to find overlapping communities in networks* in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2008), 408–423.

92. Li, M., Chen, J.-e., Wang, J.-x., Hu, B. & Chen, G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics* **9,** 1–16 (2008).

93. Lancichinetti, A., Fortunato, S. & Kertész, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11,** 033015 (2009).

94. Hollocou, A., Bonald, T. & Lelarge, M. Multiple local community detection. *ACM SIGMETRICS Performance Evaluation Review* **45,** 76–83 (2018).

95. Hollocou, A., Bonald, T. & Lelarge, M. Improving PageRank for local community detection. *arXiv preprint arXiv:1610.08722* (2016).

96. Ruuskanen, J. O., Peitsaro, N., Kaslin, J. V., Panula, P. & Scheinin, M. Expression and function of α_2-adrenoceptors in zebrafish: drug effects, mRNA and receptor distributions. *Journal of neurochemistry* **94,** 1559–1569 (2005).

97. Li, Q. *et al.* 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).

98. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature methods* **19,** 679–682 (2022).

99. Blackburn, A. C. *et al.* Deficiency of glutathione transferase zeta causes oxidative stress and activation of antioxidant response pathways. *Molecular pharmacology* **69,** 650–657 (2006).

100. Guo, Y., Huang, C., Xie, Y., Song, F. & Zhou, X. A tomato glutaredoxin gene SlGRX1 regulates plant responses to oxidative, drought and salt stresses. *Planta* **232,** 1499–1509 (2010).

101. Sánchez-Riego, A. M., López-Maury, L. & Florencio, F. J. Glutaredoxins are essential for stress adaptation in the cyanobacterium Synechocystis sp. PCC 6803. *Frontiers in plant science* **4,** 63520 (2013).

102. Pflugmacher, S. & Sandermann Jr, H. Cytochrome P450 monooxygenases for fatty acids and xenobiotics in marine macroalgae. *Plant physiology* **117,** 123–128 (1998).

103. Kanai, T., Takahashi, K. & Inoue, H. Three distinct-type glutathione S-transferases from Escherichia coli important for defense against oxidative stress. *Journal of biochemistry* **140,** 703–711 (2006).

104. Li, C.-W. *et al.* Arabidopsis root-abundant cytosolic methionine sulfoxide reductase B genes MsrB7 and MsrB8 are involved in tolerance to oxidative stress. *Plant and Cell Physiology* **53,** 1707–1719 (2012).

105. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature methods* **14,** 61–64 (2017).

106. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods* **13,** 966–967 (2016).

107. Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology,* 1–4 (2023).

108. Shou, C. *et al.* Measuring the evolutionary rewiring of biological networks. *PLoS computational biology* **7,** e1001050 (2011).

109. Zhang, J. *et al.* Computing the Human Interactome. *bioRxiv* (2024).

110. Yook, S.-H., Oltvai, Z. N. & Barabási, A.-L. Functional and topological characterization of protein interaction networks. *Proteomics* **4,** 928–942 (2004).

111. Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research* **50,** W228–W234 (2022).

Materials and Methods

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. [73], while symbiont sequences were obtained from ReefGenomics [74, 75], based on Liu et al. [76]. 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 [50]. We keep 11,487 fly sequences with between 50 and 800 residues.

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 [20] domains using hmmer [24], 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 [25] 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, necessitating a systems approach to function.

Synthetic Network Prediction

D-SCRIPT [22] 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 (Appendix A.4). We keep any proteins that 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.

Diffusion State Distance

DSD is an effective way to measure distance between nodes in a PPI network. We first consider the case where the edges are unweighted as in [18]. 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) [18] 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 [18]. Furthermore, when there are weights on the edges, a straightforward modification presented in Cao et al. (2014) [23] 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.

Community Detection by Hierarchical Spectral Clustering

Given the DSD-smoothed node distance matrices, we convert this to a similarity measure between nodes using the RBF kernel [77]. We cluster this distance matrix using a modified version of the spectral clustering algorithm described in [19], recursively splitting any cluster that is bigger than assigned threshold. The original DSD + spectral clustering method won the disease module identification challenge in DREAM 2016 [1]. Our updated version of this algorithm, which we call "Hierarchical Spectral," gives us more control over the final distribution of cluster sizes using 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 as uninformative.

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 4a). 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 [1], where they led to clusters that were better connected and more functionally enriched than the original (Appendix A.11). We perform a comprehensive evaluation of hyper-parameters both for community detection and reconnection and find that PHILHARMONIC is robust to a wide variety of choices for these parameters (Appendix A.6). 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 A14).

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 [1] (Appendix A.11). We compute the set of enriched GO terms for a cluster using v2.0.1 of the `FUNC-E` Python package [78]. We then assign GO functions based on the enriched terms, and compute a Jaccard similarity between the pre-

dicted and true sets of functional labels. We show in Figure A12 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) (Appendix A.11). In Figure A13, we show that the 10% linear ratio setting is significantly better than random additions of the same number of proteins to each cluster.

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 [31] 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.

LLM Summarization

Inspired by work from Hu et al. [79], 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 [80], 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 included in Appendix A.5.1. This name is added to the text summary from above and output for biological analysis. 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.

Implementation

We implement PHILHARMONIC using Snakemake [81]. 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. 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. We use version 0.16 of the llm command line utility. 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 itself is available as a `pip`-installable package. 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 provide job submission scripts for running on HPC environments, and we also provide Google Colab notebooks to easily run PHILHARMONIC in the browser.

Data Availability

Proteome sequences, predicted networks, clusters, and functional annotations for *P. damicornis*, *C. goreaui*, and *D. melanogaster* have been deposited on Zenodo at `https://zenodo.org/records/15066518`.

Code Availability

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