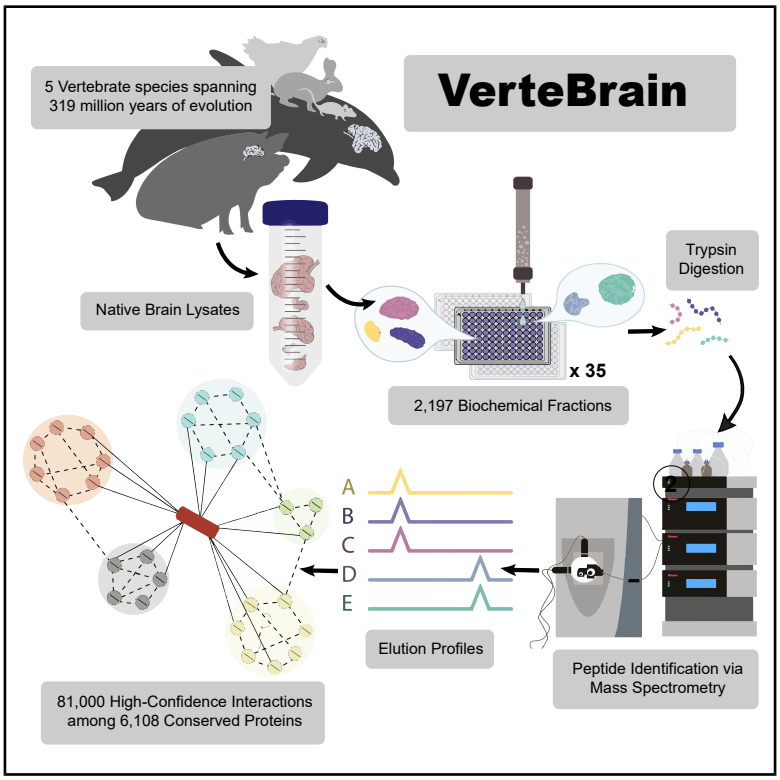
Cell Reports

Resource

A cross-vertebrate brain protein interaction map identifies conserved neural and non-neural complexes

Graphical abstract



Authors

Vy Dang, Brittney Voigt, David Yang, ..., Benjamin A. Neely, Ryan S. Gray, Edward M. Marcotte

Correspondence

marcotte@utexas.edu

In brief

Dang et al. use co-fractionation and immunoprecipitation mass spectrometry to map protein interactions conserved across vertebrate brains. This resource, dubbed VerteBrain, offers insights into brain protein function and new links between proteins and disease, identifying candidate genes and pathways involved in epilepsy, deafness, and developmental disorders.

Highlights

- Highly conserved interactome among brain proteins across 5 vertebrate species
- Over 81,000 protein interactions define conserved brain multi-protein assemblies
- Interactions between Rab effectors and across trafficking systems
- VerteBrain identifies candidate disease genes, including RELCH and ARHGEF1

Dang et al., 2026, Cell Reports *45*, 117422
June 23, 2026 © 2026 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2026.117422

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

A cross-vertebrate brain protein interaction map identifies conserved neural and non-neural complexes

Vy Dang,[1] Brittney Voigt,[1,5] David Yang,[2] Gabriel Hoogerbrugge,[1] Muyoung Lee,[1] Rachael M. Cox,[1] Ophelia Papoulas,[1] Claire D. McWhite,[3] Raksha Pradeep,[1] Janelle C. Leggere,[1] Benjamin A. Neely,[4] Ryan S. Gray,[5] and Edward M. Marcotte[1,6,*]

[1]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712, USA
[2]University of California, Berkeley, Center for Computational Biology, Berkeley, CA 94720, USA
[3]Department of Molecular and Cellular Biology, The University of Arizona, Tucson, AZ 85721, USA
[4]Chemical Sciences Division, National Institute of Standards and Technology, NIST, Charleston, SC 29412, USA
[5]Department of Nutrition and Pediatrics, Dell Pediatric Research Institute, The University of Texas at Austin, Dell Medical School, Austin, TX 78712, USA
[6]Lead contact
*Correspondence: marcotte@utexas.edu
https://doi.org/10.1016/j.celrep.2026.117422

SUMMARY

Protein-protein interactions underlie core brain functions, including neurotransmitter release, receptor activation, and intracellular signaling. Here, we present a compendium of protein-protein interactions conserved across vertebrate brains, spanning over 300 million years of evolution. From 2,197 biochemical fractions across five vertebrate species, we identify over 81,000 high-confidence interactions among 6,108 conserved proteins. This interaction map (VerteBrain) reveals both regulatory and structural complexes, including extensive synaptonemal protein associations likely involved in inter-neuronal coordination. Conservation across species underscores essential roles in neuronal and glial function, as well as in additional tissues for more widely expressed complexes. The VerteBrain dataset uncovers candidate disease mechanisms, including roles for ARHGEF1 in short stature syndromes, synaptic vesicle trafficking complexes in epilepsy, and RELCH (Rab11-binding and LisH domain, coiled-coil and HEAT repeat-containing) in congenital deafness. VerteBrain provides a resource for investigating brain protein interactions and their relevance to human neurological disorders.

INTRODUCTION

Protein-protein interactions (PPIs) organize cellular processes, particularly in the brain, where they regulate synaptic transmission, plasticity, and memory. Disruptions in these assemblies underlie diverse neurodevelopmental disorders, underscoring the need to systematically define brain protein interactions. Beyond informing about direct binding interfaces, PPIs provide more general insights into protein function and associate genes with genetic traits and diseases through guilt by association. Moreover, as estimates suggest that 80%–95% of human genes are expressed at some stage in the human brain,[1] brain PPIs should also broadly inform studies in other tissues. Mapping these assemblies should thus help guide our understanding of the molecular logic of brain and other tissue function, as well as provide a framework for interpreting the impact of disease-linked mutations and ultimately facilitate targeted treatments.

While transcriptome and connectome atlases have advanced rapidly,[2–5] proteomic atlases, particularly those capturing PPIs, remain limited due to the proteome's inherent complexity: proteins are dynamic, often extensively post-translationally modified, and exist within specialized spatial and temporal contexts, all of which introduce technical challenges. Global neuroproteomic studies face additional challenges due to the brain's heterogeneity and tissue complexity and the difficulties of capturing endogenous protein interactions on a large scale.

Despite such challenges, recent advances in proteomics have made it possible to systematically survey protein interactions in complex systems. Various *in vitro* methods from yeast two-hybrid (Y2H) assays[6] to recently developed orthogonal mass spectrometry (MS)-based techniques have been applied to define the interactomes of various biological systems. To date, large-scale (high-throughput) physical protein interaction maps have been reported for systems ranging from yeast[7] and plants[8] to human cell lines[9,10] and specific cell types such as red blood cells.[11] For the brain, the largest interaction datasets have been measured from postmortem mouse[12,13] and human brain[14] tissues, as well as from specific compartments such as mouse synaptosomes[15] (recently reviewed in Dang et al.[16]). While these studies offer key insights, comprehensive brain interactomes

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

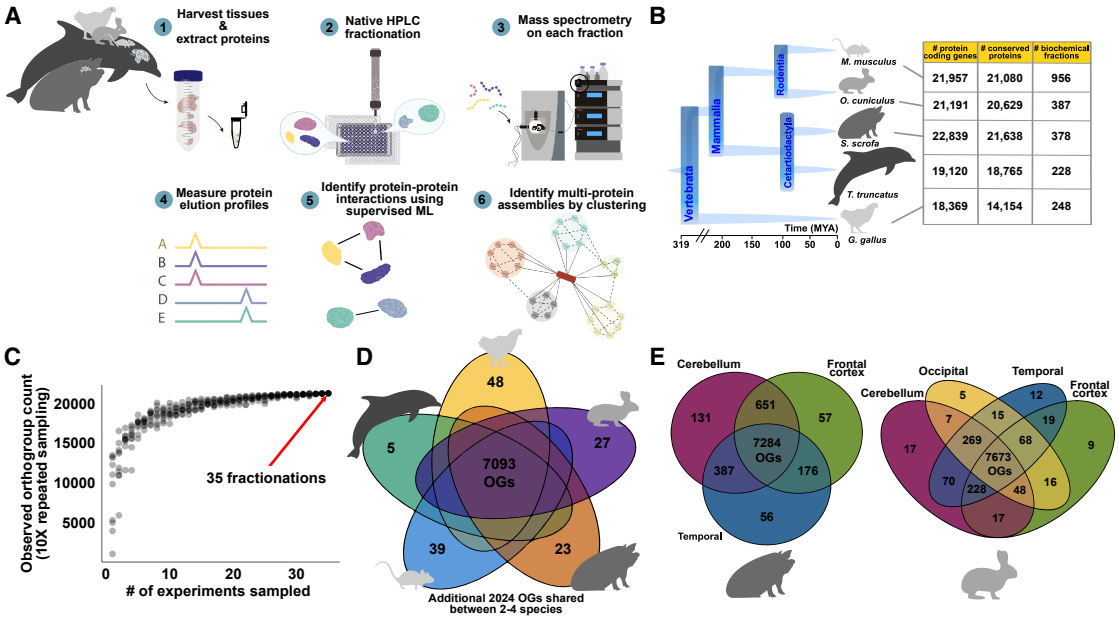


Figure 1. Overview of methods, samples, data collection, and proteome coverage

(A) Schematic of co-fractionation mass spectrometry experimental and computational workflows.

(B) Phylogenetic relationship of the species used in the study, their respective proteome sizes, and the number of mass spectrometry experiments performed for each species.

(C) The total number of orthogroups detectable by mass spectrometry is saturated by 35 CF-MS experiments used in the study. Each dot represents the number of OGs identified (*y* axis) in a subsample of *n* experiments (*x* axis), with sampling repeated ten times per *n* to evaluate saturation.

(D) A core set of 7,093 conserved OGs is detected across all 5 species.

(E) Core sets of 7,284 and 7,673 OGs are detected across all tested distinct brain regions of pig and rabbit, respectively.

remain limited, especially given the difficulties in accessing and preserving healthy human brain tissue, suggesting an important opportunity for leveraging the relative ease of obtaining healthy tissue from evolutionarily diverse animals to better define conserved brain interactions. Such proteomic data could then be integrated with transcriptomic and connectomic atlases to enhance our understanding of brain biology and uncover pathways relevant to neurological disease.

To address the need for more comprehensive brain protein interaction networks, we applied co-fractionation MS (CF-MS), an ultra-high-throughput technique that enables proteome-wide mapping of endogenous protein interactions without the need for antibodies or genetic tagging, unlike related techniques such as affinity purification MS (AP-MS), making it uniquely suited for comparative studies across species. We analyzed native soluble protein extracts from the postmortem brains of four vertebrate species—chicken, dolphin, pig, and rabbit—integrating these datasets with previously published mouse brain CF-MS experiments.[12,13] We focused exclusively on evolutionarily conserved proteins and interactions, as conservation significantly increases confidence that these interactions are critical to brain function. This study reports the resulting cross-species interactome of the vertebrate brain: VerteBrain.

To validate our observed interactions, we conducted extensive immunoprecipitation MS (IP-MS) analyses and experimental validation in vertebrate model organisms. Notably, we observed protein complexes active both in the brain and in non-neural tissues, underscoring the broader biological relevance of the detected assemblies. While one of the ultimate goals for neuroscience research is to develop therapeutics for debilitating neurological diseases, access even to healthy brain tissue remains inherently limited. VerteBrain demonstrates that comparative analysis across evolutionarily diverse species can yield robust insights into conserved mechanisms of brain function and their disruption in disease.

RESULTS AND DISCUSSION

A comprehensive dataset of brain protein abundance and purification profiles from five vertebrates

We generated a proteomics dataset from five vertebrate species spanning 319 million years of evolution: *Mus musculus* (mouse), *Sus scrofa* (pig), *Oryctolagus cuniculus* (rabbit), *Gallus gallus* (chicken), and *Tursiops erebennus* (Tamanend's bottlenose dolphin) (Figure 1A). We performed 20 biochemical fractionations of native protein lysates from chicken and dolphin whole-brain tissues and dissected pig and rabbit brain regions (frontal cortex, cerebellum, hippocampus, and temporal lobe). Combined with published mouse brain CF-MS data,[12,13] this dataset comprises 2,197 native biochemical fractions from five vertebrates (Figure 1B).

Native protein extracts were separated by size-exclusion chromatography (SEC), ion exchange (IEX) chromatography, or weak anion exchange (WAX) chromatography. Each

chromatographic fraction was analyzed by high-resolution, high-sensitivity liquid chromatography-MS (LC-MS), with replicate separations performed to ensure reproducibility. Each species contributed at least 200 MS experiments, ensuring balanced species representation (Figure 1B). Further description of species, tissues, and fractionation procedures can be found in the STAR Methods, the key resources table, or Table S1. Across all datasets, we identified 14,653,734 interpretable peptide mass spectra (sum of all peptide spectral matches [PSMs] across all protein groups and fractions), controlling peptide and protein identifications to <1% false discovery rate (FDR) using three different MS search algorithms.[17] Proteins were mapped to vertebrate orthogroups (OGs) using eggNOG[18] to enable cross-species comparisons. This dataset thus captures vertebrate brain proteins and enables neurobiological and evolutionary analyses of the vertebrate nervous system.

Overall, our approach resulted in high protein coverage, with 35 CF-MS experiments saturating OG detection (Figure 1C). To focus downstream analyses on well-observed proteins, we required ≥60 PSMs per OG across fractions and treated missing values as zero. The filtered dataset covers 11,025 human proteins across 9,259 Vertebrata OGs.

Multiple metrics indicate high data quality. First, OGs already known to be associated with neurodegenerative disorders and disease, or essential for neurogenesis, modulation of neurotransmitter release at the pre-synaptic terminal, or maintenance of dendritic architecture,[19] are among the most abundant. These include microtubule-associated protein 2 (MAP2),[20] dihydropyrimidinase-related protein 2 (DPYSL2),[21] spectrin alpha chain, non-erythrocytic (SPTAN1), and the synapsin protein family (SYN1–SYN3).

Second, over 75% (7,093 of 9,259) of high-abundance OGs were detected across all 5 species (Figure 1D), underscoring the evolutionary support within this dataset. Quantification of the full set of 9,259 high-abundance OGs can be found in Table S2.

Third, highly observed proteins correlated with higher mRNA expression. We found that the robustly observed (≥60 PSMs) vertebrate proteins were more likely to correspond to higher mRNA expression levels across eight distinct human brain regions from the Adult Genotype Tissue Expression Portex (GTEx) v.8.0 (Figure S1). This finding highlights the nuanced and imperfect correlation between protein-level detection and transcriptional activity,[22,23] even on a cell-by-cell basis,[24] while also reinforcing the reliability of our dataset. Finally, we examined the diversity of proteins recovered from different brain regions within the same species and observed a core set of proteins observed broadly across different brain regions (Figure 1E). Together, these results indicate a high-quality, deeply sampled vertebrate brain proteome.

Defining a well-conserved interactome among vertebrate brain proteins

Many stable complexes exhibited clustered elution patterns detectable by eye. Examples include the 19S regulatory subunits of the proteasome, the Commander holocomplex, and the actin-related protein 2/3 complex (Figure 2A). However, to systematically identify these patterns, we applied supervised machine learning.[8,11,25] This method leverages the fact that known protein complexes are observed within the separations and can serve as internal standards and assigns a probabilistic CF-MS score between 0 and 1 to each observed PPI, where a score of 1 represents the highest level of confidence that a pair of proteins interacts and 0 indicates no evidence of interaction.

In order to train the classifier and score the pairwise interactions, we first computed pairwise similarities between proteins' elution profiles, considering a variety of measures and normalizations (STAR Methods). The resulting numerical measures served as features for a supervised classification algorithm, selected using automated machine learning[26] to best recover a training set of known protein complexes as defined by the CORUM protein complex database,[27] a compendium of complexes manually curated from the literature. We observed the best performance by using an ExtraTreeClassifier with 5-fold cross-validation and a set of the 70 best-performing features. We evaluated the model's performance using a test set of 2,342 known pairwise PPIs that were fully withheld from model training (Figure 2B). For interactions with CF-MS scores exceeding 0.88, the model achieved 92% precision and 18% recall (see Zenodo repository for data), indicating high precision for confident interactions.

Out of the 12,961,622 pairs of proteins evaluated by the model (STAR Methods), 81,225 PPIs had scores above the 0.88 threshold, representing high-confidence protein interactions. In total, the resulting brain protein interactome included 6,108 OGs, each with at least one interaction above the 8% FDR threshold based on the ExtraTreeClassifier model. Pairwise interactions are provided in Table S3.

Defining and validating multi-protein assemblies in VerteBrain

To define coherent protein complexes from the high-confidence pairwise interactions, we used the unsupervised community detection algorithm walktrap,[28] previously shown to perform well for protein complex identification[11,25] (STAR Methods). This algorithm organizes interactions into a dendrogram that defines a nested hierarchy of complexes. Depending on the granularity of the cutoff, the algorithm identifies 145 (less granular) and 5,380 (more granular) protein "neighborhoods" (Figure 2C), which at finer granularity correspond strongly to discrete protein complexes. This hierarchical organization thus enables exploration of both broad and fine-grained protein interactions, allowing analysis of interactome organization. The set of protein assemblies defined in this way is reported in Table S4.

We benchmarked CF-MS scores against independent interactome datasets (STAR Methods). High-scoring PPIs were enriched across Y2H assays (in HuRI[7]), AP-MS experiments (in BioPlex 3.0[10]), chemical cross-linking (XL-MS) in mouse synaptosomes,[15] integration of AP-MS and CF-MS (in hu.MAP 2.0[29]), and by live-cell fluorescent microscopy coupled with IP-MS (as recorded in the OpenCell[30] database), when compared to randomly paired proteins (Figure 2D). Notably, PPIs surpassing the 8% FDR threshold displayed a very strong correlation with other methods, strongly supporting the accuracy of our reported

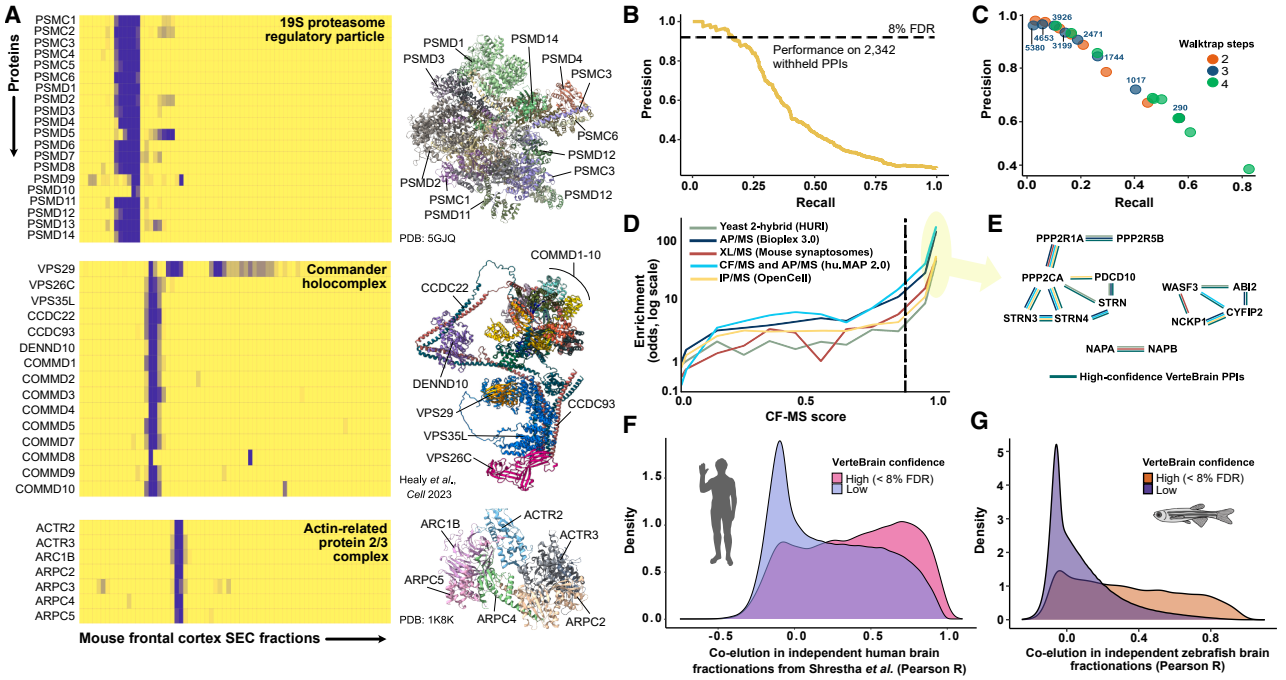


Figure 2. Identifying protein interactions from the CF-MS datasets and validating with independent experiments

(A) Heatmap showing normalized abundance profiles for three representative molecular assemblies from SEC fractionations of mouse frontal cortex: the 19S regulatory particle of the proteasome, the Commander holocomplex, and the actin-related protein 2/3 (Arp2/3) complex. These examples were selected from the full dataset of 9,259 well-sampled orthogroups across 2,197 biochemical fractions from five vertebrate species. Each column represents an individual biochemical fraction, and each row corresponds to a single subunit, which is depicted in the corresponding structural model to the right of the heatmap.

(B) Precision-recall performance of 2,342 withheld PPIs.

(C) Precision-recall curves for the reconstruction of known protein complexes defined by a walktrap algorithm, where pairwise PPI scores from each number of walks (colored as in legend) are used as input. Points are labeled with the total number of clusters (complexes) created at each point in the hierarchy.

(D) VerteBrain shows strong agreement with orthogonal techniques that also measure PPIs. Dotted line indicates the 8% FDR score cutoff (CF-MS score ≥0.88).

(E) Agreement of CF-MS PPI pairwise interactions (dark green) with Y2H (light green), AP/MS (dark blue), XL/MS (red), huMAP2.0 (purple), and IP/MS (yellow) for three example protein complexes.

(F) PPIs with high-confidence CF-MS scores (≥0.88) in VerteBrain are highly correlated with the CF-MS dataset of human brain published by Shrestha et al.,[14] plotted as Pearson correlation coefficients between the chromatographic elution vectors for 2,502,432 protein pairs.

(G) PPIs with high-confidence CF-MS scores (≥0.88) in VerteBrain are highly correlated with the independent CF-MS dataset measured from zebrafish brains, plotted as Pearson correlation coefficients between the chromatographic elution vectors for 3,714,022 protein pairs.

PPIs. Additionally, of the ~81K PPIs in VerteBrain, ~40K are entirely novel, indicating substantial novelty (Figure S2; Table S5).

As an independent test of whether high CF-MS scores indicate more frequent direct contacts between the interacting proteins, we performed an *in silico* PPI screen using AlphaFold-Multimer[31] on protein pairs identified in the VerteBrain dataset, grouping them into bins based on their CF-MS scores (in 0.1 increments). We then examined the relationship between CF-MS score bins and AlphaFold-Multimer's interface predicted template modeling (ipTM) score. Overall, we observed that protein pairs in score bins exceeding the FDR threshold exhibited higher AlphaFold confidence scores (Figure S3). However, while high ipTM scores have been shown to be reliable indicators of accurate protein complex predictions across multiple studies,[32–34] it should be noted that low ipTM scores (i.e., <0.25) do not exclude interactions, as subunits of the same protein complex may not directly contact each other but will still co-elute in CF-MS experiments. This is because they are part of the same larger structure and are therefore physically associated, even if they do not share a binding interface.

It is also important to note that no single method captures all PPIs. For instance, although all techniques are capable of capturing the majority of well-established pairwise interactions among members of the evolutionarily conserved STRIPAK complex (striatin-interacting phosphatase and kinase),[35] only Y2H, AP-MS, and CF-MS detected the interaction between PDCD10 (programmed cell death protein 10) and striatin (STRN) (Figure 2E), likely reflecting differential expression: STRN is highly enriched in the CNS, STRN4 is predominantly expressed in the brain and lung, while STRN3 is ubiquitously expressed across all tissues.[36,37] Additionally, VerteBrain demonstrates the capacity to detect neural-specific protein interactions across a broad spectrum of structural and functional categories. Examples range from large, multisubunit assemblies like the WAVE (WASP family verprolin homolog—also known as SCAR, suppressor of cAMP receptor) regulatory complex, to simpler heterodimeric interactions, such as the NSF

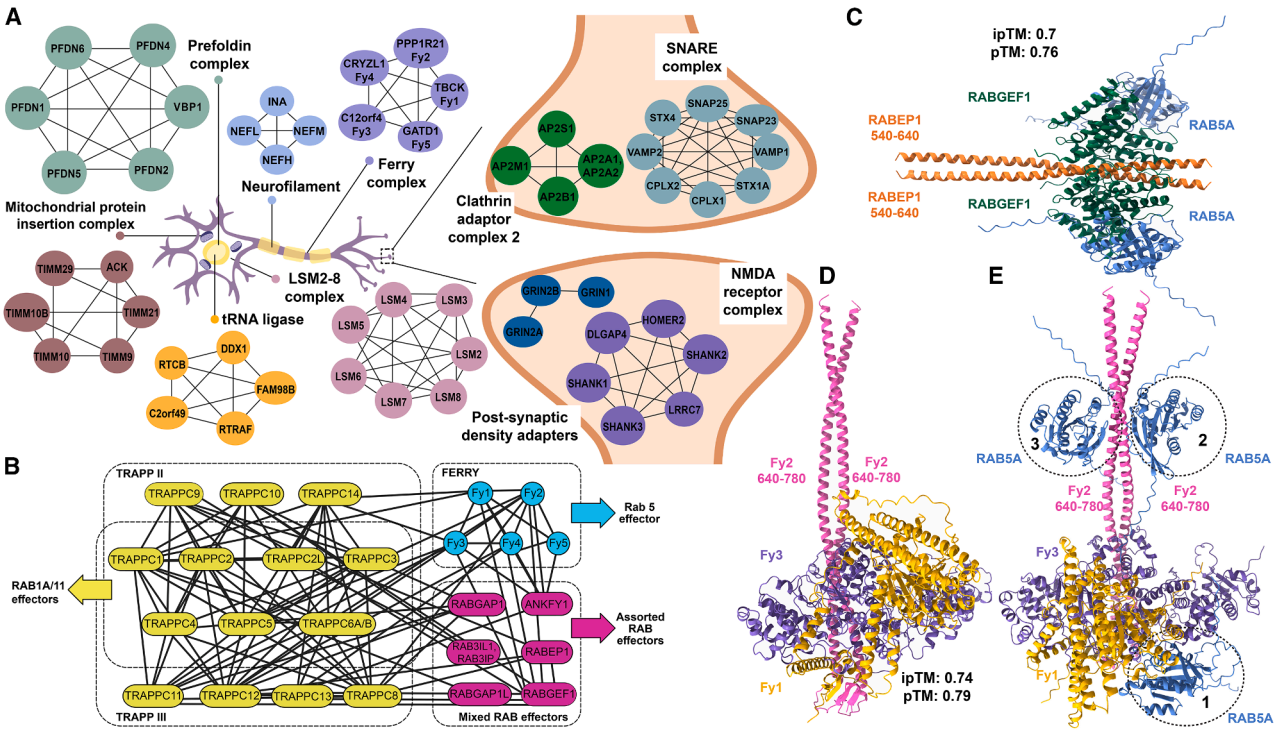


Figure 3. VerteBrain recapitulates many known neuronal protein complexes and interactions

(A) VerteBrain analysis identified numerous well-characterized neuronal protein complexes spanning a range of sizes, including the NMDA receptor complex, prefoldin complex, clathrin adaptor complexes, and additional assemblies essential for neuronal function.

(B) Detected interactions encompassed subunits of the TRAPPII and TRAPPIII complexes, as well as crosstalk among established Rab effectors.

(C) AlphaFold 3 Multimer modeling predicted the most favorable stoichiometry for the RABEP1 (residues 540–640; orange)-RABGEF1 (dark green)-RAB5A (blue) complex to be 2:2:2.

(D) The AlphaFold 3 Multimer model is in good agreement with the schematic representation of the FERRY complex by Quentin et al.,[45] capturing Fy1 (yellow) and Fy3 (purple) binding to a homodimer of Fy2 (residues 640–780; pink), consistent with hydrogen/deuterium exchange mass spectrometry data.

(E) Predicted partial structure of the FERRY complex highlights three potential Rab5A-binding (blue) sites, with a suggested order of preferential binding from site 1 to 3.

attachment proteins alpha (NAPA) and beta (NAPB) (Figure 2E). Some interactions, including those identified through our CF-MS observations, can only be captured by specialized techniques such as Y2H or targeted studies of specific cellular compartments, such as in mouse synaptosomes. This highlights the complementary nature of various PPI mapping approaches and highlights VerteBrain's strengths in detecting neural-specific interactions.

Cross-species validation confirms reliability and biological relevance of VerteBrain PPIs

We further validated VerteBrain using two independent species-specific PPI maps. First, we observed the same overall trend in CF-MS data from a human brain study published by Shrestha et al.[14] (Figure 2F). As a further test of conservation across vertebrates, we conducted two additional CF-MS experiments on adult zebrafish brain lysates using SEC and IEX chromatography, for which internal controls confirmed proper complex co-elution profiles (Figure S4). We observed that high-confidence VerteBrain PPIs tended to exhibit highly correlated elution profiles in the zebrafish brain data, supporting the generality of our interaction across vertebrates (Figure 2G). These analyses support the robustness and biological relevance of the detected interactions.

VerteBrain recapitulates the known organization of many neuronal proteins

To assess biological relevance, we evaluated VerteBrain's recovery of well-characterized multiprotein assemblies critical for neuronal function. As illustrated for representative examples in Figure 3A, we examined a diverse array of essential neuronal processes, spanning multiple regions of the nervous system, including the heterotetramic clathrin adaptor complex-2 (AP-2),[38] which plays a pivotal role in clathrin-dependent endocytosis, a process essential for retrieving synaptic vesicles at presynaptic terminals following neurotransmitter release. In neurons, AP-2 is also critical for neurodevelopment by regulating cell polarity during differentiation[39] and promoting neuronal survival,[40] axon growth cone outgrowth and guidance,[41] and dendrite formation and extension.[40,42] VerteBrain confidently identified high-scoring pairwise interactions between AP-2 subunits (e.g., AP2A1/AP2A2, AP2B1, and AP2M1) and recapitulated other broadly functioning complexes, such as the prefoldin complex and

the tRNA ligase complex. VerteBrain also identified specialized complexes including neurofilaments (NEFL, NEFM, and NEFH) involved in maintaining neuronal architecture[43] and the post-synaptic density adapters, such as the SHANK family (SH3 and multiple ankyrin repeat domains), which play critical roles in synaptic signaling and organization. Additionally, VerteBrain confidently recovers interactions within the SNARE (soluble NSF attachment protein receptors) complex, a key regulator of vesicle transport and neurotransmitter release.[44] These examples illustrate the breadth of the dataset in uncovering known ubiquitous and neuron-specific assemblies.

We also systematically measured VerteBrain's reconstruction of known systems by quantifying the extent to which VerteBrain clusters proteins known to function together. We calculated the area under the receiver operating characteristic curve (AUROC) for proteins known to belong to the same Gene Ontology (GO) in the same GO protein-containing complex (PCC) or cellular anatomical entity (CAE) (Figure S5). As expected, random protein sets yielded AUROC scores centered around 0.5, indicating no predictive power. In contrast, VerteBrain demonstrated significantly higher AUROC scores for proteins with shared GO term annotations, confirming biological coherence. Similar to previously reported large-scale interactome studies on brain tissues,[12–14] our dataset showed significant enrichment for protein networks associated with the cytoskeleton, extracellular matrix (ECM), and membranes (e.g., *trans*-Golgi network membrane, endosome membrane, and lysosomal membrane) (Figure S5). In general, we observed clear correspondence between specific VerteBrain protein clusters and well-established protein complexes or biological processes, as labeled in Table S4.

As an example, the COG (conserved oligomeric Golgi) complex is an evolutionarily conserved peripheral membrane protein complex that is primarily responsible for controlling membrane trafficking and ensuring Golgi homeostasis by orchestrating retrograde vesicle transport within the Golgi.[46] As expected, VerteBrain identified high-confidence interactions between the eight subunits (COG1–COG8). Additionally, our data revealed interactions between COG subunits and GOLGA1 (golgin subfamily A member 1), a member of the golgin protein family, which functions as a vesicle tethering factor to facilitate Golgi fusion events.[47,48] Specifically, VerteBrain showed that GOLGA1 interacts with COG3, COG5, and COG7, while GOLGA6D interacts with COG5 and COG7. To our knowledge, GOLGA1 pairwise interactions with COG1 and COG3 were previously reported in a proximity labeling study by Go et al.[49] in HEK293 cells. These data suggest coordinated roles for the COG complex and the golgin protein family in Golgi trafficking.

We also showed that VerteBrain is enriched for proteins annotated in other proteomics resources, including CNS essential genes[50] and genes implicated in neurological disorders[51] (Figures S6A and S6B). Additionally, the VerteBrain interactions themselves are significantly enriched among multiple specific Brain.GMT disease gene sets, including genes linked to malignant glioma, cortex genes downregulated in schizophrenia, or genes upregulated in Alzheimer's disease, etc. (Figure S6C).

VerteBrain reveals interactions between Rab effectors and across trafficking systems

CF-MS confirms known protein assemblies and reveals potential crosstalk between them. When combined with AlphaFold-Multimer modeling, this approach provides a robust framework for generating structural hypotheses and functional insights based on complex membership and interaction dynamics.[31,52] Leveraging VerteBrain's overall accuracy, we decided to investigate such newly observed interactions among established systems.

We examined endocytic trafficking systems, which regulate key physiological activities in the brain such as synaptic signaling, receptor dynamics, and nutrient uptake.[53] Rab (Ras-associated binding) GTPases and their effectors coordinate key steps such as vesicle budding, transport, and membrane fusion.[54] One such system is the five-subunit endosomal Rab5 and RNA/ribosome intermediary (FERRY) complex, involved in mRNA transport on early endosomes.[55] In two *in vitro* studies,[45,55] FERRY was shown to be a pentameric complex composed of TBCK (Fy1), PPP1R21 (Fy2), C12ORF4 (Fy3), CRYZL (Fy4), and GATD1 (Fy5). Our data recapitulate this assembly, suggesting that the FERRY complex forms a stable complex *in vivo*, across all species in this study (Figures 3A and 3B). Notably, mutations in three different members of the FERRY complex (Fy1, Fy2, and Fy3) are associated with rare neurological disorders in human patients.[56]

Similarly, VerteBrain captures high-confidence PPIs observed among subunits of the TRAPPII and TRAPPIII complexes. TRAPP (transport protein particle) complexes are Rab GTPase exchange factors with shared cores.[57] In metazoans, there are two TRAPP complexes: TRAPPII and TRAPPIII. TRAPPII and TRAPPIII have distinct specificity for GEF activity toward Rabs, with TRAPPIII acting on Rab1 and TRAPPII acting on Rab1 and Rab11.[57] We also observed previously undescribed interactions between these systems, with high CF-MS scores between TRAPPII/TRAPPIII subunits and the FERRY complex (Figure 3B), suggesting functional connectivity among Rab effectors.

We next examined RAB5A interactions. RAB5A, a key regulator of early endosomal fusion, interacts with multiple effectors including RABEP1[58,59] (Rabaptin, Rab GTPase-binding effector protein 1) and RABGEF1 (Rab guanine nucleotide exchange factor 1), which form a ternary complex that activates Rab5A more effectively than either protein alone.[60]

Using AlphaFold 3, we modeled this interaction and found the most favorable stoichiometry to be 2:2:2, supporting the role of RABEP1 dimerization in stabilizing the complex (Figure 3C). VerteBrain also recovered high-confidence interactions between RAB5A and RABGEF1, as well as between RAB5A and RABEP1. While the precise determinants of RAB5A binding remain unclear, these findings suggest that Rab5A binding is context dependent, influenced by cellular and molecular conditions.

Given these results, we hypothesized that RAB5A might also bind to multiple sites on the FERRY complex. Although direct RAB5A-FERRY interactions were not strongly supported by CF-MS scores, we found interactions between Rab effectors (RABEP1 and RABGEF1) and FERRY subunits (Fy1–Fy4). To investigate these findings, we performed structural modeling

using AlphaFold 3 to test the alignment of Rab5A binding with the site proposed by Quentin et al.[45] (Figure 3D). Testing multiple configurations with increasing copies of Rab5A consistently revealed three distinct favorable binding sites, as illustrated by dotted circles labeled 1 through 3 (Figure 3E). These findings highlight the dynamic and potentially promiscuous nature of Rab5A binding within the FERRY complex and illustrate VerteBrain's utility for hypothesis generation.

An extensive interaction network centered on Rabconnectin-3 and WDR37-PACS1/2 complexes

We examined a particular case where VerteBrain identified previously unknown interactions between two reasonably well-conserved and defined systems: the Rabconnectin-3 complex and the WDR37-PACS1/2 complex.

Rabconnectin-3, originally identified from purified synaptic vesicle fractions in rat brain,[61,62] comprised Rabconnectin-3α (DMXL1 or brain-enriched DMXL2) and Rabconnectin-3β (WDR7).[63] Despite limited characterization of individual subunit functions, DMXL1, DMXL2, and WDR7 were identified as accessory proteins specific to V-ATPase, implicating Rabconnectin-3 in synaptic vesicle acidification—a critical process in protein sorting and degradation.[63] While the original discovery of the Rabconnectin-3 complex only identified two subunits, recent studies leveraging IP-MS confirmed the inclusion of ROGDI,[30,64] a highly conserved metazoan protein with structural homology to yeast Rav2.[65] Since DMXL1/2 share partial homology with Rav1, another component of the yeast RAVE (regulator of ATPase of vacuoles and endosomes) complex, this supports the idea that Rabconnectin-3 may function analogously to the RAVE complex as a V-ATPase chaperone.[30,64] Consistent with these observations, we constructed a 3D structural model of Rabconnectin-3 with AlphaFold 3 that supports similarity to the yeast RAVE complex, with the exception of WDR7 and Skp1, which as of yet have no known equivalents (Figure 4A).

Similarly, the WDR37-PACS1/2 axis comprised WDR37, a highly conserved WD40-repeat-containing protein with unknown function in vertebrates, and two paralogs of PACS (phosphofurin acidic cluster sorting) proteins, PACS1 and PACS2, which are involved in sorting cargo proteins to various organelles.[66,67] Published Y2H assay, co-immunoprecipitation,[68] and AP-MS[10] experiments have confirmed the physical interaction between WDR37 and PACS1/2, although the stoichiometry remains unclear. Using AlphaFold 3, we modeled the interaction interface, which suggests that PACS1 and PACS2 compete for the same WDR37-binding site, hinting at dynamic, mutually exclusive interactions (Figures 4B and 4C). A patient-derived non-synonymous D220G mutation in WDR37, linked to a neuro-oculocardiogenitourinary syndrome and shown to disrupt both PACS1 and PACS2 binding,[68] falls directly at this predicted interface (Figure S7), supporting their likely competitive binding as well as the clinical relevance of this interaction. However, the stoichiometric and functional implications of this competition remain to be elucidated.

While VerteBrain accurately reconstructs subunit-subunit interactions for each complex, to our surprise, it also revealed previously unknown interactions between the Rabconnectin-3 (DMXL1/2, WDR7, and ROGDI) and the WDR37-PACS1/2 (WDR37, PACS1, and PACS2) axis, as well as numerous shared interaction partners, suggesting potential functional connections between these systems. Motivated by these observations, we independently tested the interactions using IP-MS from pig and mouse brain lysates, targeting the Rabconnectin-3 subunits (WDR7, DMXL1, DMXL2, and ROGDI), the WDR37-PACS1/2 complex subunits (WDR37 and PACS2), and the VerteBrain-identified interactor RELCH (Rab11-binding and LisH domain, coiled-coil and HEAT repeat-containing) as bait proteins (Figure 5A). While we confirmed interactions within and between these two complexes (Figure 5B), we further observed a rich interaction network centered on these proteins, which we visualized with a hierarchical clustering heatmap depicting the top 100 most enriched "prey" proteins from our immunoprecipitation experiments (Figure 5C; Table S6). Further supporting evidence for the enrichment of these interactors is provided in Figures S8 and S9 and the Zenodo repository. Notably, Rabconnectin-3 and WDR37-PACS1/2 share many common interactors, spanning elements of intracellular trafficking and signaling, and including the GABA (gamma-aminobutyric acid) and ryanodine receptors, Na^+/K^+-transporting ATPases, the E3 ubiquitin-protein ligases TRIM21 and HUWE1, as well as core cellular machinery, such as the eIF3 complex (Figures 5C and 5D).

VerteBrain reveals connections between protein interactions and human disorders

Clinical associations support the significance of these findings. Mutations in Rabconnectin-3 components have been implicated in various neurodevelopmental disorders. For example, DMXL2 mutations and structural variants are associated with developmental delay, intellectual disability (ID), autism spectrum disorder (ASD), and epilepsy,[71,72] while ROGDI mutations cause Kohlschütter-Tönz syndrome (KTS), characterized by epilepsy, spasticity, psychomotor regression, and amelogenesis imperfecta.[65,73] Similarly, *de novo* variants in WDR37 have been linked to multisystemic syndromes characterized by various developmental anomalies, including epilepsy, developmental delay, ID, cerebellar hypoplasia, colobomas, and dysmorphic facial features.[74,75] PACS1 mutations cause ID with facial dysmorphism,[76] while PACS2 mutations are linked to developmental and epileptic encephalopathy.[77] Consistent with the protein interactions identified in VerteBrain, these disorders share overlapping clinical features, including epilepsy, ID, and developmental delay (Figure 5E), reinforcing the hypothesis that Rabconnectin-3 and WDR37-PACS1/2 participate in shared biological pathways crucial for neurodevelopment.

Interestingly, previous genetic studies in humans have identified mutations in DMXL2 as pathogenic causes of dominant non-syndromic hearing loss (deafness, autosomal dominant 71 [DFNA71]), without accompanying neurological deficits in affected patients[78,79] (Figure 5E). Consistent with these findings, the zebrafish ortholog of DMXL2 localizes to the basal region of hair cells, and the corresponding mutant alleles of *dmxl2* cause auditory and vestibular defects.[80] The role of DMXL2 in auditory processes was further confirmed in a recent study by Peng et al.,[81] demonstrating its role in endocytosis and recycling of auditory synaptic vesicles in a hair cell-specific knockout mouse model.

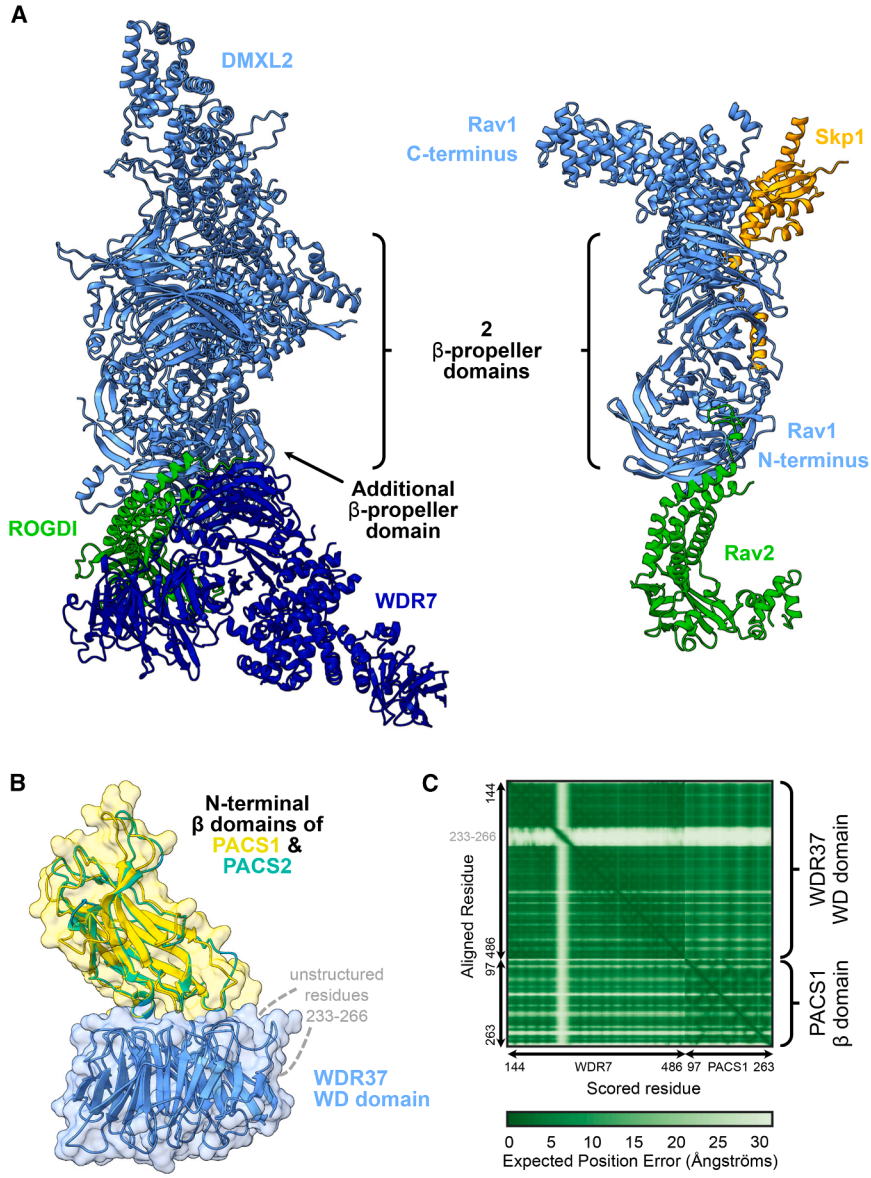


Figure 4. Structural predictions from AF3 reveal structural homology between the Rabconnectin-3 complex in animals and the RAVE complex in yeast

(A) AlphaFold 3 models reveal strong structural homology between the Rabconnectin-3 complex in higher eukaryotes, comprising DMXL2, WDR7, and ROGDI, and the RAVE complex in yeast, highlighting conserved architecture and potential functional parallels in V-ATPase regulation.

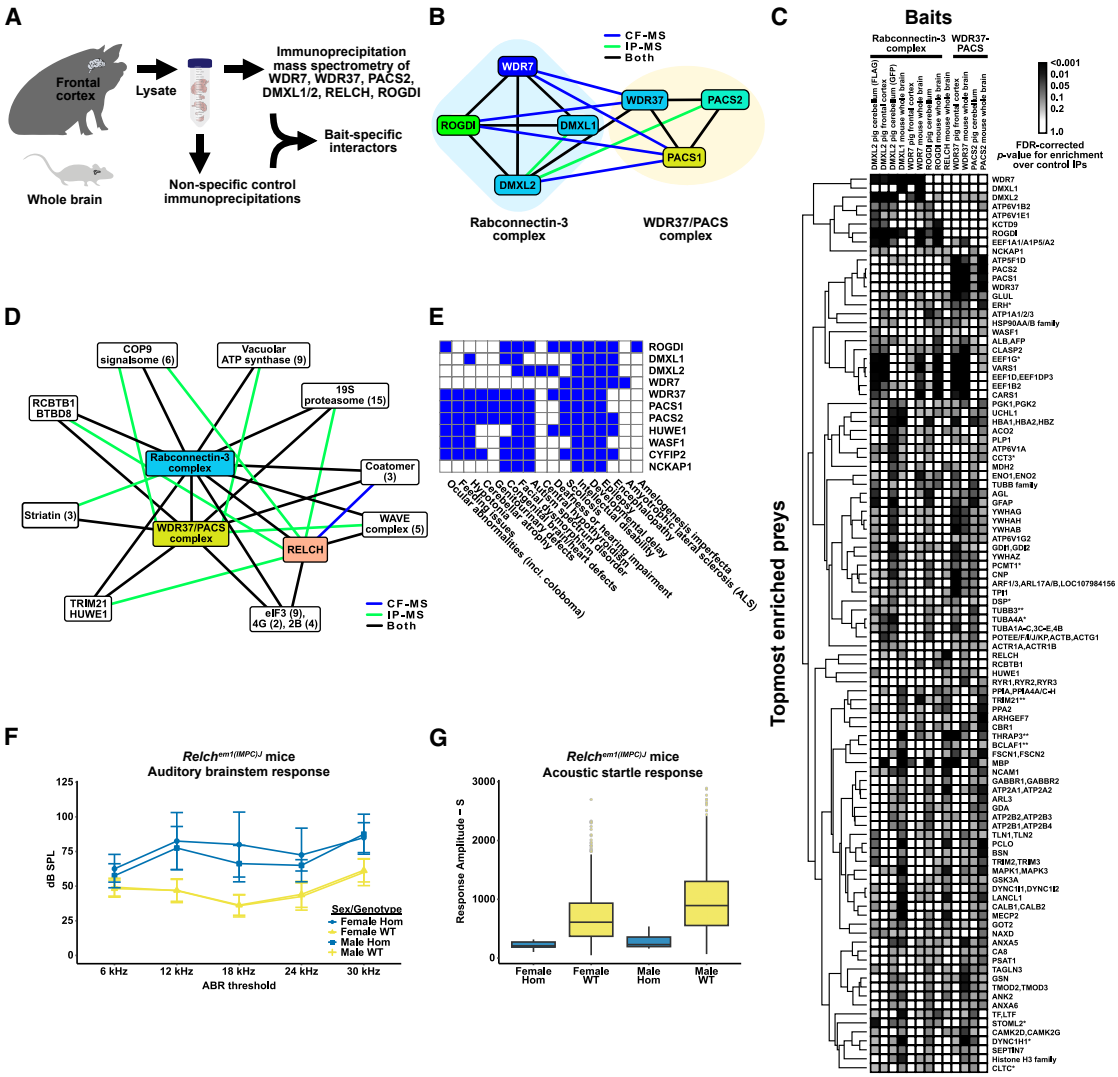
(B) Structural modeling shows PACS1 (yellow) and PACS2 (cyan) competing for the same WDR37-binding site, suggesting dynamic, mutually exclusive interactions.

(C) Predicted aligned error (PAE) plot showing the well-predicted WD40 domain of WDR37 interacting with the β-domain of PACS1, demonstrating high-confidence structural alignment and potential interaction stability.

Building on these observations, our VerteBrain analysis reports a perfect pairwise score of 1 between RELCH and DMXL2, suggesting a strong functional or physical association. In IP-MS experiments using Rabconnectin-3 subunits as bait proteins, RELCH emerged as a high-confidence interactor (Figure 5D). Likewise, homozygous RELCH mutant mice displayed abnormal auditory responses (Figure 5F) and a reduced startle reflex (Figure 5G), as reported by the International Mouse Genetic Consortium[70] (STAR Methods). Together, these findings drawn from VeterBrain observations as well as IP-MS validation bridge evidence from human genetics, zebrafish models, and mouse studies, to collectively point to RELCH as a potential causative gene for deafness in humans.

Other noteworthy interactors of the Rabconnectin-3 and WDR37/PACS1-2 complexes identified in VerteBrain include the GABA receptors, the major inhibitory neurotransmitter

CellPress
OPEN ACCESS


A
Frontal cortex
Lysate
Whole brain
Immunoprecipitation mass spectrometry of WDR7, WDR37, PACS2, DMXL1/2, RELCH, ROGDI
Bait-specific interactors
Non-specific control immunoprecipitations
B
CF-MS
IP-MS
Both
WDR7
WDR37
PACS2
ROGDI
DMXL1
PACS1
DMXL2
Rabconnectin-3 complex
WDR37/PACS complex
C
Baits
Rabconnectin-3 complex
WDR37-PACS
<0.001
0.01
0.05
0.1
0.2
1.0
FDR-corrected p-value for enrichment over control IPs
Topmost enriched preys
WDR7
DMXL1
DMXL2
ATP6V1B2
ATP6V1E1
KCTD9
ROGDI
EEF1A1/A1P5/A2
NCKAP1
ATP5F1D
PACS2
PACS1
WDR37
GLUL
ERH*
ATP1A1/2/3
HSP90AA/B family
WASF1
ALB,AFP
CLASP2
EEF1G*
VARS1
EEF1D,EEF1DP3
EEF1B2
CARS1
PGK1,PGK2
UCHL1
HBA1,HBA2,HBZ
ACO2
PLP1
ATP6V1A
CCT3*
MDH2
ENO1,ENO2
TUBB family
AGL
GFAP
YWHAG
YWHAH
YWHAB
ATP6V1G2
GDI1,GDI2
YWHAZ
PCMT1*
CNP
ARF1/3,ARL17A/B,LOC107984156
TPI1
DSP*
TUBB3**
TUBA4A*
TUBA1A-C,3C-E,4B
POTEE/F/I/J/KP,ACTB,ACTG1
ACTR1A,ACTR1B
RELCH
RCBTB1
HUWE1
RYR1,RYR2,RYR3
PPIA,PPIA4A/C-H
TRIM21**
PPA2
ARHGEF7
CBR1
THRAP3**
BCLAF1**
FSCN1,FSCN2
MBP
NCAM1
GABBR1,GABBR2
ATP2A1,ATP2A2
ARL3
GDA
ATP2B2,ATP2B3
ATP2B1,ATP2B4
TLN1,TLN2
PCLO
BSN
TRIM2,TRIM3
MAPK1,MAPK3
GSK3A
DYNC1I1,DYNC1I2
LANCL1
CALB1,CALB2
MECP2
GOT2
NAXD
ANXA5
CA8
PSAT1
TAGLN3
GSN
TMOD2,TMOD3
ANK2
ANXA6
TF,LTF
STOML2*
CAMK2D,CAMK2G
DYNC1H1*
SEPTIN7
Histone H3 family
CLTC*
D
COP9 signalsome (6)
Vacuolar ATP synthase (9)
19S proteasome (15)
RCBTB1 BTBD8
Rabconnectin-3 complex
Coatomer (3)
Striatin (3)
WDR37/PACS complex
WAVE complex (5)
RELCH
TRIM21 HUWE1
eIF3 (9), 4G (2), 2B (4)
CF-MS
IP-MS
Both
E
ROGDI
DMXL1
DMXL2
WDR7
WDR37
PACS1
PACS2
HUWE1
WASF1
CYFIP2
NCKAP1
Ocular anomalies (incl. coloboma)
Hypotonia
Cerebellar atrophy
Cerebellar hypoplasia
Cerebellar ataxia
Cerebellar vermis hypoplasia
Facial dysmorphism
Abnormal facial shape
Abnormality of the nervous system
Abnormality of the eye
Abnormality of the face
Abnormality of the musculature
Abnormality of the cerebellum
Abnormality of the cerebellar vermis
Abnormality of the cerebellar hemisphere
Abnormality of the cerebellar cortex
Abnormality of the cerebellar morphology
Abnormality of the cerebellar atrophy
Abnormality of the cerebellar hypoplasia
Abnormality of the cerebellar ataxia
Abnormality of the cerebellar vermis hypoplasia
Abnormality of the cerebellar hemisphere hypoplasia
Abnormality of the cerebellar cortex hypoplasia
Abnormality of the cerebellar morphology hypoplasia
Abnormality of the cerebellar atrophy hypoplasia
Abnormality of the cerebellar ataxia hypoplas

receptors in vertebrate brains. Specifically, we observed both complexes to interact with GABA receptor subunits, an observation replicated with 4 independent IP-MS baits (ROGDI, DMXL1, RELCH, and PACS2). Mutations in different subunits have been associated with genetic epilepsy syndromes with different clinical phenotypes.[82,83]

Our identification of RELCH and GABA receptors as shared interactors within these conserved complexes provides a framework for studying their roles in human disorders. These findings expand our understanding of intracellular trafficking networks and lay the groundwork for future research into their implications for epilepsy, ID, and hearing impairment. While many of the clinical features described here are common across diverse, unrelated neurodevelopmental disorders, the involvement of a shared cellular mechanism between these proteins underscores the importance of exploring a potential convergent pathological pathway.

Conserved interactions extending beyond the brain: INPPL1, ARHGEF1, and short-stature disorders

Neurological disorders, such as ASD, ID, and developmental delay, often co-occur with skeletal defects like scoliosis or hypotonia[84,85] (Figure 5E). Although these clinical manifestations are seemingly distinct, growing evidence highlights the shared molecular pathways and cellular mechanisms underpinning both brain and skeletal development. PPI networks in the brain can reveal critical regulatory nodes that extend their influence beyond neural phenotypes, shedding light on systemic processes like cytoskeletal organization, cellular signaling, and tissue homeostasis. Notably, skeletal dysplasia is increasingly recognized to co-occur with neurological symptoms,[86,87] suggesting that disruptions in shared developmental pathways can simultaneously impact both the brain and the skeleton. Motivated by this broader perspective, we turned our focus to *INPPL1* (inositol polyphosphate phosphatase-like 1), a gene with reported mutations linked to the rare human endochondral bone disorder opsismodysplasia (OPS), which affects growth plate expansion and skeletal development.[88–90] Clinical features of OPS include short stature with short limbs, small hands and feet; macrocephaly with a prominent anterior fontanel; and facial dysmorphism, although severity varies at birth.

The *INPPL1* gene encodes the 5′-inositol phosphatase SHIP2 (src homology 2 domain-containing inositol phosphatase 2), which regulates phosphoinositide signaling by removing the 5′-phosphate from substrates such as $PI_{(3,4,5)}P_3$ and $PI_{(4,5)}P_2$ in cellular membranes.[91,92] Previous work by Voigt et al.[93] demonstrated that homozygous mutations in *inppl1a*, the zebrafish ortholog of *INPPL1*, cause defects in notochord vacuolated cell expansion and hypertrophic chondrocyte differentiation, leading to vertebral malformations and scoliosis and endochondral bone defects in zebrafish. VerteBrain revealed several high CF-MS scoring protein interactions with INPPL1. We filtered these candidate interactors using AlphaFold 3 to test for putative direct binding and found a medium level of confidence for direct binding with ARHGEF1 (Rho guanine nucleotide exchange factor 1) (Figures 6A–6C). ARHGEF1 is an intracellular signaling molecule that has been shown to modulate signaling from G protein-coupled receptors to RhoA, and as such, it is an important regulator of the actin cytoskeleton with pleiotropic functions.[94] Interestingly, ARHGEF1 is primarily implicated in immunological functions, as its loss causes a primary antibody deficiency in humans.[95] This observation highlights the heterogeneity of disease etiology, as disruption of a single gene can affect multiple pathways, resulting in diverse clinical manifestations.

To further explore the functional relationship between INPPL1 and ARHGEF1, we generated a null allele of *ARHGEF1* in zebrafish. Although zebrafish have two paralogs of *ARHGEF1* (*arhgef1a* and *arhgef1b*), single-cell RNA sequencing data from Daniocell indicate that *arhgef1a* is primarily expressed in hematopoietic cells,[96] similar to the reported expression patterns in mouse.[97] In contrast, *arhgef1b* is more broadly expressed in developing zebrafish, suggesting that this paralog functions in extra-hematopoietic systems. To probe these potential roles of *arhgef1b*, we generated a null allele using a CRISPR method with small guide RNAs (sgRNAs) targeting the 5′ and 3′ ends of the gene (Figures 6D–6F; STAR Methods). To better visualize potential notochord defects, we performed the CRISPR injections in a transgenic line expressing membrane-localized GFP in notochord vacuolated cells.[97] This approach successfully generated what will be referred to as the *arhgef1b* deficiency (*df*) allele, in which much of the coding region is deleted, as confirmed by PCR (Figures 6D, 6E, and 6G).

At 6 days post-fertilization (dpf), 7% of the genetically mosaic F0 *arhgef1b* crispants displayed body curvatures and notochord defects (Figures 6H and 6I; $n = 292$), and confocal microscopy revealed the collapse of notochord vacuolated cells at the site of body curvature (Figure 6I′), supporting its role in vacuolated cell development. F1 $arhgef1b^{df/+}$ heterozygotes were in-crossed to assess vacuolated cell and notochord morphology in $arhgef1b^{df/df}$ homozygous mutants (Figures 6J and 6K). Analysis of vacuolated cell volumes using semi-automated cell segmentation (Imaris) (Figures 6J′ and 6K′) demonstrated a significant decrease in vacuolated cell volume in $arhgef1b^{df/df}$ mutants compared to wild-type siblings at 5 dpf (Figures 6M and 6N), which were genotyped using a duplex PCR strategy to identify zygosity (Figure 6L; STAR Methods).

Homozygous $arhgef1b^{df/df}$ mutants did not survive to adulthood, and we were therefore unable to assess skeletal phenotypes in $arhgef1b^{df/df}$ adults. We suspect that the broad expression of *arhgef1b* indicates a role in a wide variety of tissues, any of which could be essential for survival. Regardless, these embryonic results strongly support a role for *arhgef1b* in

(F) Auditory brainstem response (ABR) thresholds for RELCH null mice, generated by the International Mouse Consortium,[70] with a homozygous deletion of exons 3–5 and 454 bp of flanking intronic sequence ($n = 8$ per sex, $n = 16$ total), compared to healthy control mice (female, $n = 671$; male, $n = 673$). RELCH-deficient mice show significantly elevated ABR thresholds at all stimulus levels, indicating profound hearing impairment and reinforcing the role of RELCH in auditory function. (G) Acoustic startle response and pre-pulse inhibition phenotypic assay for RELCH mutant mice with homozygous exon deletion (female, $n = 7$; male, $n = 8$; $n = 15$, total) compared to control mice (1,949 females, 1,920 males, $n = 3,869$ total). Mutant mice exhibit significant differences in response amplitude—a decrease in startle reflex—suggesting deficits in auditory processing.

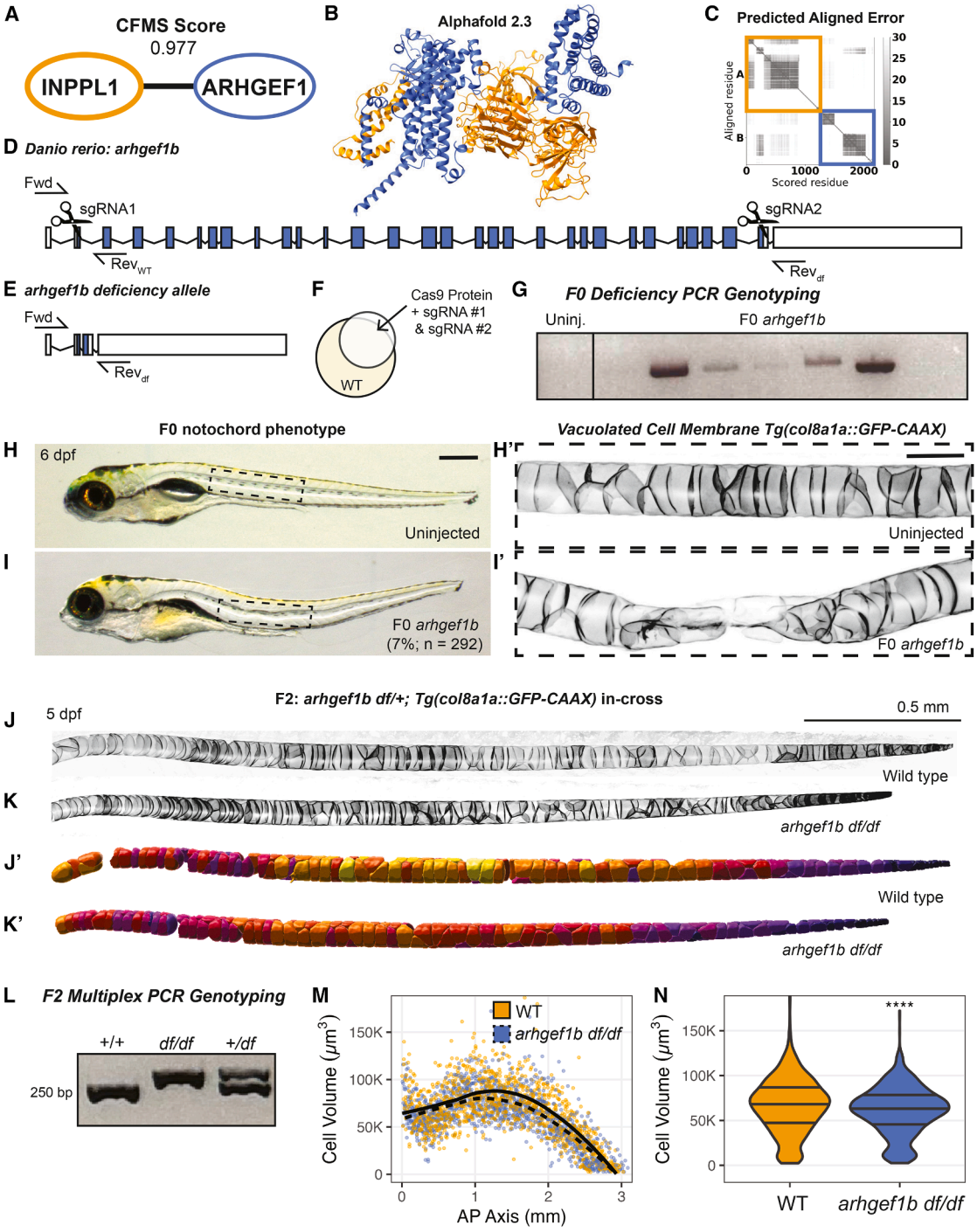


Figure 6. VerteBrain identifies interactors of INPPL1 important for skeletal development and disorders

(A) INPPL1 VerteBrain cluster identified an interaction between INPPL1 and ARHGEF1, with a CF-MS score of 0.977.

(B) Structural modeling of INPPL1-ARHGEF1 interaction using AlphaFold 2.3 Multimer. INPPL1 is depicted in gold, while ARHGEF1 is depicted in blue.

(C) Predicted aligned error (PAE) plot of INPPL1-ARHGEF1 interaction. Gold box indicates INPPL1 residues, while the blue box indicates ARHGEF1 residues.

(D) Genomic structure of *D. rerio arhgef1b* locus (white boxes represent untranslated exonic regions, blue boxes represent translated exonic regions, and lines represent intronic regions) with CRISPR mutagenesis strategy (scissors represent CRISPR sgRNA-targeted loci, and arrows represent approximate locations of the various genotyping primers: Fwd, universal forward primer; Rev-WT, reverse primer to detect the wild-type allele; Rev-df, reverse primer to detect the deficiency allele).

(E) Genomic structure of the *D. rerio arhgef1b* deficiency (*df*) allele construct with alignment of relevant genotyping primers (Fwd, universal forward primer; Rev-df, reverse primer to detect the deficiency allele).

(legend continued on next page)

vacuolated cell size regulation, similar to that of *inppl1a*, and suggest a role for INPPL1-ARHGEF1 interactions in notochord development in zebrafish. Notably, a protein interaction observed in the brain successfully predicted a phenotype in a second tissue, indicating that many brain interactions may be relevant to multiple tissue types. This example also illustrates the value of VerteBrain for candidate gene discovery for diseases and traits.

Limitations of the study

Several technical limitations of the methods used in this study are worth acknowledging. First, CF-MS favors abundant and stable protein complexes, potentially missing transient or low-abundance interactions. Because we employed a non-denaturing lysis buffer to preserve native interactions, certain assemblies, particularly those that are membrane associated, embedded in organelles, or confined within dense nuclear compartments, may have been underrepresented or incompletely solubilized under our extraction conditions. Because it detects co-elution rather than direct contact, additional analyses are required to distinguish direct from indirect associations. Also, as we analyze extracted proteins, detailed information about spatial and temporal dynamics of protein interactions is lost. Additionally, our reliance on evolutionary conservation of the interactions, while useful for prioritizing interactions likely to be functionally important for having been conserved across vertebrates, means that interactions specific to one or a subset of lineages will be missed. Similarly, the relatively high false-negative rate of these approaches suggests that some apparently tissue-specific PPIs may simply reflect undersampling. Together, these factors likely bias our dataset toward soluble, well-resolved, and well-conserved complexes and underrepresent more dynamic, compartmentalized, or condition-specific assemblies. Finally, as VerteBrain is derived entirely from animal data, its human relevance is inferred, not directly measured. However, our requirement that these interactions be conserved across multiple vertebrates strongly supports their likely occurrence in humans as well. Nonetheless, if only direct support in humans is strictly required, approximately 25% of VerteBrain PPIs have existing support in at least one human cell type (Table S5), and hence that subset could be considered a high-confidence human brain PPI network.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Edward M. Marcotte (marcotte@utexas.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- All raw and interpreted mass spectrometry data have been deposited at ProteomeXchange via the MassIVE partner repository as PXD061997, PXD062147, PXD062225, PXD062149, MSV000098441, MSV000098440, and PXD062174.
- Additional supporting data and files are available on Zenodo (https://doi.org/10.5281/zenodo.14807396).
- All original code has been deposited at GitHub and is publicly available at the links listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of the Knockout Mouse Program (KOMP2) at the Baylor College of Medicine and the International Mouse Phenotyping Consortium (IMPC) for the *Relch*$^{em1(IMPC)J}$ mice data; Wayne McFee at NOAA for providing dolphin brain samples; Juju Dessert for animal and brain cell figure illustrations; John Wallingford for manuscript feedback; Kevin Drew and Anna Battenhouse for extensive assistance with computational analyses; Maria Person, Michelle Gadoush, and Peter Faull at the UT Proteomics Facility for aiding in data collection; and the Texas Advanced Computing Center at The University of Texas at Austin for providing high-performance computing resources. Research was funded by grants from the National Institute of General Medical Sciences (R35GM122480 to E.M.M.), the National Institute of Arthritis & Musculoskeletal & Skin Diseases (R01AR071009 to R.S.G.), the National Institute of Child Health and Human Development (F31HD114419 to B.V.), the United States Army Research Office (W911NF-12-1-0390 to E.M.M.), and the Welch Foundation (F-1515 to E.M.M.). Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

(F) CRISPR mutagenesis was achieved via 1 nL injection of Cas9 protein, sgRNA1 targeting the *arhgef1b* start site, and sgRNA2 targeting the *arhgef1b* stop site of the locus into one-cell staged wild-type zebrafish embryos.
(G) PCR validation in 1 dpf F0 *arhgef1b* crispant embryos demonstrating that the *arhgef1b* deficiency allele was generated by the CRISPR mutagenesis strategy. Uninjected (Uninj.) control amplicons are absent when using the universal forward primer and the deficiency-specific reverse primer, while pooled F0 crispants display amplicons.
(H) Representative lateral view of uninjected control embryos at 6 dpf (scale bars, 0.5 mm), with (H′) inset showing maximum projection of confocal images of the notochord vacuolated cell membrane marker: *Tg(col8a1a::GFP-CAAX)* (scale bars, 50 μm).
(I) Representative lateral view of F0 *arhgef1b* crispant embryos at 6 dpf showing body curvatures (7%, $n = 292$), with (I′) inset showing maximum projection of confocal images of the notochord vacuolated cell membrane marker: *Tg(col8a1a::GFP-CAAX)*, showing vacuolated cell collapse at the region of body curvature.
(J) Representative maximum projections of confocal stacks of notochord vacuolated cell membranes labeled by the *Tg(col8a1a::GFP-Caax)* transgene in 5 dpf wild-type progeny of *arhgef1b*$^{df/+}$ in-cross, with (J′) Imaris cell segmentation colored by cell volume (scale bars, 0.5 mm).
(K) Representative maximum projections of confocal stacks of notochord vacuolated cell membranes labeled by the *Tg(col8a1a::GFP-Caax)* transgene in 5 dpf homozygous mutant progeny of *arhgef1b*$^{df/+}$ in-cross, with (K′) Imaris cell segmentation colored by cell volume (scale bars, 0.5).
(L) Representative agarose gel of the duplex PCR genotyping method used to distinguish wild-type (+/+), homozygous mutant (df/df), and heterozygous carrier (df/+) larvae used in imaging experiments.
(M) Scatterplot of cell volumes (μm^3) plotted against their anterior-posterior axis (mm) at 5 dpf in wild-type (orange color palette, $n = 922$ cells, $N = 7$ fish) and *arhgef1b*$^{df/df}$ homozygous mutants (blue color palette, $n = 1{,}318$ cells, $N = 10$ fish) from *arhgef1b*$^{df/+}$ in-crosses, with loess lines of best fit for each genotype.
(N) Violin plot of cell volumes (μm^3) of 5 dpf in wild-type (orange color palette, $n = 922$ cells, $N = 7$ fish) and *arhgef1b*$^{df/df}$ homozygous mutants (blue color palette, $n = 1{,}318$ cells, $N = 10$ fish) from *arhgef1b*$^{df/+}$ in-crosses (horizontal lines represent 25th, 50th, and 75th quartiles, $^{****}p < 0.0001$).

AUTHOR CONTRIBUTIONS

Design and co-supervision, V.D. and E.M.M.; proteomics samples and experiments, V.D., aided by G.H., O.P., R.P., J.C.L., and B.A.N.; data analysis, V.D. and E.M.M., aided by D.Y., M.L., R.M.C., and C.D.M.; zebrafish experiments, design, and supervision, B.V. and R.S.G.; manuscript initial draft, V.D., B.V., and E.M.M. All authors discussed the results and contributed to editing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Sources and handling of brain samples
  - Zebrafish husbandry
  - Summary of RELCH mouse data, sourced from KOMP/IMPC
- METHOD DETAILS
  - Sample preparation, protein extraction, and co-fractionation/mass spectrometry
  - Immunoprecipitation/mass spectrometry
  - Zebrafish genetics and imaging
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Computational analyses of protein mass spectrometry datasets
  - Identification and scoring of protein-protein interactions
  - Independent validation of CF-MS interactions by comparison to external protein interaction datasets
  - Evaluating pairwise interactions for evidence of direct binding using AlphaFold-Multimer
  - Evaluating the predictive power of the measured interactions for biological processes
  - Mapping of orthogroups to Disease Ontology terms
  - Mapping of Gene Ontology terms considering the gold standard training set

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2026.117422.

Received: July 15, 2025
Revised: April 1, 2026
Accepted: April 29, 2026
Published: June 4, 2026

REFERENCES

1. Bae, B.-I., Jayaraman, D., and Walsh, C.A. (2015). Genetic changes shaping the human brain. Dev. Cell *32*, 423–434.
2. Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.
3. Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. Cell *174*, 999–1014.e22.
4. Li, H., Janssens, J., De Waegeneer, M., Kolluru, S.S., Davie, K., Gardeux, V., Saelens, W., David, F.P.A., Brbić, M., Spanier, K., et al. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. Science *375*, eabk2432.
5. Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K.W., Ding, S.L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., Clark, M., et al. (2023). Transcriptomic diversity of cell types across the adult human brain. Science *382*, eadd7046.
6. Fields, S., and Song, O. (1989). A novel genetic system to detect protein–protein interactions. Nature *340*, 245–246.
7. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408.
8. McWhite, C.D., Papoulas, O., Drew, K., Cox, R.M., June, V., Dong, O.X., Kwon, T., Wan, C., Salmi, M.L., Roux, S.J., et al. (2020). A pan-plant protein complex map reveals deep conservation and novel assemblies. Cell *181*, 460–474.e14.
9. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex Network: a systematic exploration of the human interactome. Cell *162*, 425–440.
10. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022–3040.e28.
11. Sae-Lee, W., McCafferty, C.L., Verbeke, E.J., Havugimana, P.C., Papoulas, O., McWhite, C.D., Houser, J.R., Vanuytsel, K., Murphy, G.J., Drew, K., et al. (2022). The protein organization of a red blood cell. Cell Rep. *40*, 111103.
12. Pourhaghighi, R., Ash, P.E.A., Phanse, S., Goebels, F., Hu, L.Z.M., Chen, S., Zhang, Y., Wierbowski, S.D., Boudeau, S., Moutaoufik, M.T., et al. (2020). BraInMap elucidates the macromolecular connectivity landscape of mammalian brain. Cell Syst. *10*, 333–350.e14.
13. Liebeskind, B.J., Young, R.L., Halling, D.B., Aldrich, R.W., and Marcotte, E.M. (2020). Mapping functional protein neighborhoods in the mouse brain. Preprint at bioRxiv. https://doi.org/10.1101/2020.01.26.920447.
14. Shrestha, H.K., Lee, D., Wu, Z., Wang, Z., Fu, Y., Wang, X., Serrano, G.E., Beach, T.G., and Peng, J. (2024). Profiling protein-protein interactions in the human brain by refined cofractionation mass spectrometry. J. Proteome Res. *23*, 1221–1231.
15. Gonzalez-Lozano, M.A., Koopmans, F., Sullivan, P.F., Protze, J., Krause, G., Verhage, M., Li, K.W., Liu, F., and Smit, A.B. (2020). Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. Sci. Adv. *6*, eaax5783.
16. Dang, V., Voigt, B., and Marcotte, E.M. (2025). Progress toward a comprehensive brain protein interactome. Biochem. Soc. Trans. *53*, 303–314. https://doi.org/10.1042/BST20241135.
17. Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A.I., and Marcotte, E.M. (2011). MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. J. Proteome Res. *10*, 2949–2958.
18. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314.
19. Thiel, G. (1993). Synapsin I, synapsin II, and synaptophysin: marker proteins of synaptic vesicles. Brain Pathol. *3*, 87–95.
20. Harada, A., Teng, J., Takei, Y., Oguchi, K., and Hirokawa, N. (2002). MAP2 is required for dendrite elongation, PKA anchoring in dendrites, and proper PKA signal transduction. J. Cell Biol. *158*, 541–549.
21. Feuer, K.L., Peng, X., Yovo, C.K., and Avramopoulos, D. (2023). DPYSL2/CRMP2 isoform B knockout in human iPSC-derived glutamatergic

neurons confirms its role in mTOR signaling and neurodevelopmental disorders. Mol. Psychiatry *28*, 4353–4362.

22. de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Mol. Biosyst. *5*, 1512–1526.
23. Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.
24. Khan, S., Elcheikhali, M., Leduc, A., Huffman, R.G., Derks, J., Franks, A., and Slavov, N. (2024). Inferring post-transcriptional regulation within and across cell types in human testis. Preprint at bioRxiv. https://doi.org/10.1101/2024.10.08.617313.
25. Cox, R.M., Papoulas, O., Shril, S., Lee, C., Gardner, T., Battenhouse, A.M., Lee, M., Drew, K., McWhite, C.D., Yang, D., et al. (2024). Ancient eukaryotic protein interactions illuminate modern genetic traits and disorders. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.26.595818.
26. Le, T.T., Fu, W., and Moore, J.H. (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics *36*, 250–256.
27. Tsitsiridis, G., Steinkamp, R., Giurgiu, M., Brauner, B., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2023). CORUM: the comprehensive resource of mammalian protein complexes–2022. Nucleic Acids Res. *51*, D539–D545.
28. Pons, P., and Latapy, M. (2006). Computing communities in large networks using random walks. J. Graph Alg. Appl. *10*, 191–218.
29. Drew, K., Wallingford, J.B., and Marcotte, E.M. (2021). MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol. Syst. Biol. *17*, e10016.
30. Cho, N.H., Cheveralls, K.C., Brunner, A.D., Kim, K., Michaelis, A.C., Raghavan, P., Kobayashi, H., Savy, L., Li, J.Y., Canaj, H., et al. (2022). OpenCell: Endogenous tagging for the cartography of human cellular organization. Science *375*, eabi6983.
31. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. (2022). Protein complex prediction with AlphaFold-Multimer. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.04.463034.
32. O'Reilly, F.J., Graziadei, A., Forbrig, C., Bremenkamp, R., Charles, K., Lenz, S., Elfmann, C., Fischer, L., Stülke, J., and Rappsilber, J. (2023). Protein complexes in cells by AI-assisted structural proteomics. Mol. Syst. Biol. *19*, e11544.
33. Wang, B., Vartak, R., Zaltsman, Y., Naing, Z.Z.C., Hennick, K.M., Polacco, B.J., Bashir, A., Eckhardt, M., Bouhaddou, M., Xu, J., et al. (2024). A foundational atlas of autism protein interactions reveals molecular convergence. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.03.569805.
34. Weeratunga, S., Gormal, R.S., Liu, M., Eldershaw, D., Livingstone, E.K., Malapaka, A., Wallis, T.P., Bademosi, A.T., Jiang, A., Healy, M.D., et al. (2024). Interrogation and validation of the interactome of neuronal Munc18-interacting Mint proteins with AlphaFold2. J. Biol. Chem. *300*, 105541.
35. Moreno, C.S., Park, S., Nelson, K., Ashby, D., Hubalek, F., Lane, W.S., and Pallas, D.C. (2000). WD40 repeat proteins striatin and S/G2 nuclear autoantigen are members of a novel family of calmodulin-binding proteins that associate with protein phosphatase 2A. J. Biol. Chem. *275*, 5257–5263.
36. Castets, F., Bartoli, M., Barnier, J.V., Baillat, G., Salin, P., Moqrich, A., Bourgeois, J.P., Denizot, F., Rougon, G., Calothy, G., and Monneron, A. (1996). A novel calmodulin-binding protein, belonging to the WD-repeat family, is localized in dendrites of a subset of CNS neurons. J. Cell Biol. *134*, 1051–1062.
37. Castets, F., Rakitina, T., Gaillard, S., Moqrich, A., Mattei, M.G., and Monneron, A. (2000). Zinedin, SG2NA, and striatin are calmodulin-binding, WD repeat proteins principally expressed in the brain. J. Biol. Chem. *275*, 19970–19977.
38. Shih, W., Gallusser, A., and Kirchhausen, T. (1995). A clathrin-binding site in the hinge of the beta 2 chain of mammalian AP-2 complexes. J. Biol. Chem. *270*, 31083–31090.
39. Lin, Y.H., Currinn, H., Pocha, S.M., Rothnie, A., Wassmer, T., and Knust, E. (2015). AP-2-complex-mediated endocytosis of Drosophila Crumbs regulates polarity by antagonizing Stardust. J. Cell Sci. *128*, 4538–4549.
40. Zheng, J., Shen, W.H., Lu, T.J., Zhou, Y., Chen, Q., Wang, Z., Xiang, T., Zhu, Y.C., Zhang, C., Duan, S., and Xiong, Z.Q. (2008). Clathrin-dependent endocytosis is required for TrkB-dependent Akt-mediated neuronal protection and dendritic growth. J. Biol. Chem. *283*, 13280–13288.
41. Kamiguchi, H., and Yoshihara, F. (2001). The role of endocytic l1 trafficking in polarized adhesion and migration of nerve growth cones. J. Neurosci. *21*, 9194–9203.
42. Koscielny, A., Malik, A.R., Liszewska, E., Zmorzynska, J., Tempes, A., Tarkowski, B., and Jaworski, J. (2018). Adaptor complex 2 controls dendrite morphology via mTOR-dependent expression of GluA2. Mol. Neurobiol. *55*, 1590–1606.
43. Beaulieu, J.-M., Robertson, J., and Julien, J.-P. (1999). Interactions between peripherin and neurofilaments in cultured cells: disruption of peripherin assembly by the NF-M and NF-H subunits. Biochem. Cell. Biol. *77*, 41–45.
44. Südhof, T.C., and Rothman, J.E. (2009). Membrane fusion: grappling with SNARE and SM proteins. Science *323*, 474–477.
45. Quentin, D., Schuhmacher, J.S., Klink, B.U., Lauer, J., Shaikh, T.R., Huis In 't Veld, P.J., Welp, L.M., Urlaub, H., Zerial, M., and Raunser, S. (2023). Structural basis of mRNA binding by the human FERRY Rab5 effector complex. Mol. Cell *83*, 1856–1871.e9.
46. Smith, R.D., and Lupashin, V.V. (2008). Role of the conserved oligomeric Golgi (COG) complex in protein glycosylation. Carbohydr. Res. *343*, 2024–2031.
47. Witkos, T.M., and Lowe, M. (2015). The Golgin family of coiled-coil tethering proteins. Front. Cell Dev. Biol. *3*, 86.
48. Gillingham, A.K., and Munro, S. (2016). Finding the Golgi: Golgin coiled-coil proteins show the way. Trends Cell Biol. *26*, 399–408.
49. Go, C.D., Knight, J.D.R., Rajasekharan, A., Rathod, B., Hesketh, G.G., Abe, K.T., Youn, J.Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L.Y., et al. (2021). A proximity-dependent biotinylation map of a human cell. Nature *595*, 120–124.
50. Wertz, M.H., Mitchem, M.R., Pineda, S.S., Hachigian, L.J., Lee, H., Lau, V., Powers, A., Kulicke, R., Madan, G.K., Colic, M., et al. (2020). Genome-wide *in vivo* CNS screening identifies genes that modify CNS neuronal survival and mHTT toxicity. Neuron *106*, 76–89.e8.
51. Hagenauer, M.H., Sannah, Y., Hebda-Bauer, E.K., Rhoads, C., O'Connor, A.M., Flandreau, E., Watson, S.J., Jr., and Akil, H. (2024). Resource: A curated database of brain-related functional gene sets (Brain.GMT). MethodsX *13*, 102788.
52. Bryant, P., Pozzati, G., Zhu, W., Shenoy, A., Kundrotas, P., and Elofsson, A. (2022). Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nat. Commun. *13*, 6028.
53. Camblor-Perujo, S., and Kononenko, N.L. (2022). Brain-specific functions of the endocytic machinery. FEBS J. *289*, 2219–2246.
54. Zerial, M., and McBride, H. (2001). Rab proteins as membrane organizers. Nat. Rev. Mol. Cell Biol. *2*, 107–117.
55. Schuhmacher, J.S., Tom Dieck, S., Christoforidis, S., Landerer, C., Davila Gallesio, J., Hersemann, L., Seifert, S., Schäfer, R., Giner, A., Toth-Petroczy, A., et al. (2023). The Rab5 effector FERRY links early endosomes with mRNA localization. Mol. Cell *83*, 1839–1855.e13.
56. Riffe, R.M., and Downes, G.B. (2025). Neurogenetic disorders associated with mutations in the FERRY complex: a novel disease class? Biol. Open *14*, BIO061808.

57. Riedel, F., Galindo, A., Muschalik, N., and Munro, S. (2018). The two TRAPP complexes of metazoans have distinct roles and act on different Rab GTPases. J. Cell Biol. *217*, 601–617.

58. Vitale, G., Rybin, V., Christoforidis, S., Thornqvist, P., McCaffrey, M., Stenmark, H., and Zerial, M. (1998). Distinct Rab-binding domains mediate the interaction of Rabaptin-5 with GTP-bound Rab4 and Rab5. EMBO J. *17*, 1941–1951.

59. Zhu, G., Zhai, P., Liu, J., Terzyan, S., Li, G., and Zhang, X.C. (2004). Structural basis of Rab5-Rabaptin5 interaction in endocytosis. Nat. Struct. Mol. Biol. *11*, 975–983.

60. Horiuchi, H., Lippé, R., McBride, H.M., Rubino, M., Woodman, P., Stenmark, H., Rybin, V., Wilm, M., Ashman, K., Mann, M., and Zerial, M. (1997). A novel Rab5 GDP/GTP exchange factor complexed to Rabaptin-5 links nucleotide exchange to effector recruitment and function. Cell *90*, 1149–1159.

61. Nagano, F., Kawabe, H., Nakanishi, H., Shinohara, M., Deguchi-Tawarada, M., Takeuchi, M., Sasaki, T., and Takai, Y. (2002). Rabconnectin-3, a novel protein that binds both GDP/GTP exchange protein and GTPase-activating protein for Rab3 small G protein family. J. Biol. Chem. *277*, 9629–9632.

62. Kawabe, H., Sakisaka, T., Yasumi, M., Shingai, T., Izumi, G., Nagano, F., Deguchi-Tawarada, M., Takeuchi, M., Nakanishi, H., and Takai, Y. (2003). A novel rabconnectin-3-binding protein that directly binds a GDP/GTP exchange protein for Rab3A small G protein implicated in Ca2+-dependent exocytosis of neurotransmitter. Genes Cells *8*, 537–546.

63. Merkulova, M., Păunescu, T.G., Azroyan, A., Marshansky, V., Breton, S., and Brown, D. (2015). Mapping the H+ (V)-ATPase interactome: identification of proteins involved in trafficking, folding, assembly and phosphorylation. Sci. Rep. *5*, 14827.

64. Jaskolka, M.C., Winkley, S.R., and Kane, P.M. (2021). RAVE and rabconnectin-3 complexes as signal dependent regulators of organelle acidification. Front. Cell Dev. Biol. *9*, 698190.

65. Lee, H., Jeong, H., Choe, J., Jun, Y., Lim, C., and Lee, C. (2017). The crystal structure of human Rogdi provides insight into the causes of Kohlschutter-Tönz syndrome. Sci. Rep. *7*, 3972.

66. Köttgen, M., Benzing, T., Simmen, T., Tauber, R., Buchholz, B., Feliciangeli, S., Huber, T.B., Schermer, B., Kramer-Zucker, A., Höpker, K., et al. (2005). Trafficking of TRPP2 by PACS proteins represents a novel mechanism of ion channel regulation. EMBO J. *24*, 705–716.

67. Nair-Gill, E., Bonora, M., Zhong, X., Liu, A., Miranda, A., Stewart, N., Ludwig, S., Russell, J., Gallagher, T., Pinton, P., and Beutler, B. (2021). Calcium flux control by Pacs1-Wdr37 promotes lymphocyte quiescence and lymphoproliferative diseases. EMBO J. *40*, e104888.

68. Sorokina, E.A., Reis, L.M., Thompson, S., Agre, K., Babovic-Vuksanovic, D., Ellingson, M.S., Hasadsri, L., van Bever, Y., and Semina, E.V. (2021). WDR37 syndrome: identification of a distinct new cluster of disease-associated variants and functional analyses of mutant proteins. Hum. Genet. *140*, 1775–1789.

69. Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.P., St-Denis, N.A., Li, T., Miteva, Y.V., Hauri, S., Sardiu, M.E., Low, T.Y., et al. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods *10*, 730–736.

70. Dickinson, M.E., Flenniken, A.M., Ji, X., Teboul, L., Wong, M.D., White, J.K., Meehan, T.F., Weninger, W.J., Westerberg, H., Adissu, H., et al. (2016). High-throughput discovery of novel developmental phenotypes. Nature *537*, 508–514.

71. Esposito, A., Falace, A., Wagner, M., Gal, M., Mei, D., Conti, V., Pisano, T., Aprile, D., Cerullo, M.S., De Fusco, A., et al. (2019). Biallelic DMXL2 mutations impair autophagy and cause Ohtahara syndrome with progressive course. Brain *142*, 3876–3891.

72. Costain, G., Walker, S., Argiropoulos, B., Baribeau, D.A., Bassett, A.S., Boot, E., Devriendt, K., Kellam, B., Marshall, C.R., Prasad, A., et al. (2019). Rare copy number variations affecting the synaptic gene DMXL2 in neurodevelopmental disorders. J. Neurodev. Disord. *11*, 3.

73. Aswath, N., Ramakrishnan, S.N., Teresa, N., and Ramanathan, A. (2018). A novel ROGDI gene mutation is associated with Kohlschutter-Tonz syndrome. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. *125*, e8–e11.

74. Reis, L.M., Sorokina, E.A., Thompson, S., Muheisen, S., Velinov, M., Zamora, C., Aylsworth, A.S., and Semina, E.V. (2019). De novo missense variants in WDR37 cause a severe multisystemic syndrome. Am. J. Hum. Genet. *105*, 425–433.

75. Kanca, O., Andrews, J.C., Lee, P.T., Patel, C., Braddock, S.R., Slavotinek, A.M., Cohen, J.S., Gubbels, C.S., Aldinger, K.A., Williams, J., et al. (2019). De novo variants in WDR37 are associated with epilepsy, colobomas, dysmorphism, developmental delay, intellectual disability, and cerebellar hypoplasia. Am. J. Hum. Genet. *105*, 672–674.

76. Schuurs-Hoeijmakers, J.H.M., Oh, E.C., Vissers, L.E.L.M., Swinkels, M.E.M., Gilissen, C., Willemsen, M.A., Holvoet, M., Steehouwer, M., Veltman, J.A., de Vries, B.B.A., et al. (2012). Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. Am. J. Hum. Genet. *91*, 1122–1127.

77. Olson, H.E., Jean-Marçais, N., Yang, E., Heron, D., Tatton-Brown, K., van der Zwaag, P.A., Bijlsma, E.K., Krock, B.L., Backer, E., Kamsteeg, E.J., et al. (2018). A recurrent de novo PACS2 heterozygous missense variant causes neonatal-onset developmental epileptic encephalopathy, facial dysmorphism, and cerebellar dysgenesis. Am. J. Hum. Genet. *102*, 995–1007.

78. Chen, D.-Y., Liu, X.F., Lin, X.J., Zhang, D., Chai, Y.C., Yu, D.H., Sun, C.L., Wang, X.L., Zhu, W.D., Chen, Y., et al. (2017). A dominant variant in DMXL2 is linked to nonsyndromic hearing loss. Genet. Med. *19*, 553–558.

79. Wonkam-Tingang, E., Schrauwen, I., Esoh, K.K., Bharadwaj, T., Nouel-Saied, L.M., Acharya, A., Nasir, A., Leal, S.M., and Wonkam, A. (2021). A novel variant in DMXL2 gene is associated with autosomal dominant non-syndromic hearing impairment (DFNA71) in a Cameroonian family. Exp. Biol. Med. *246*, 1524–1532.

80. Einhorn, Z., Trapani, J.G., Liu, Q., and Nicolson, T. (2012). Rabconnectin3α promotes stable activity of the H+ pump on synaptic vesicles in hair cells. J. Neurosci. *32*, 11144–11156.

81. Peng, H., Wang, L., Gao, Y., Liu, H., Lin, G., Kong, Y., Xu, P., Liu, H., Yuan, Q., Liu, H., et al. (2024). DMXL2 is required for endocytosis and recycling of synaptic vesicles in auditory hair cells. J. Neurosci. *44*, e1405232024.

82. Macdonald, R.L., Kang, J.-Q., and Gallagher, M.J. (2010). Mutations in GABAA receptor subunits associated with genetic epilepsies. J. Physiol. *588*, 1861–1869.

83. Feng, Y., Wei, Z.H., Liu, C., Li, G.Y., Qiao, X.Z., Gan, Y.J., Zhang, C.C., and Deng, Y.C. (2022). Genetic variations in GABA metabolism and epilepsy. Seizure *101*, 22–29.

84. Kaufman, L., Ayub, M., and Vincent, J.B. (2010). The genetic basis of non-syndromic intellectual disability: a review. J. Neurodev. Disord. *2*, 182–209.

85. Kurolap, A., Kreuder, F., Gonzaga-Jauregui, C., Duvdevani, M.P., Harel, T., Tammer, L., Xin, B., Bakhtiari, S., Rice, J., van Eyk, C.L., et al. (2022). Bi-allelic variants in neuronal cell adhesion molecule cause a neurodevelopmental disorder characterized by developmental delay, hypotonia, neuropathy/spasticity. Am. J. Hum. Genet. *109*, 518–532.

86. Cohn, D.H., Ehtesham, N., Krakow, D., Unger, S., Shanske, A., Reinker, K., Powell, B.R., and Rimoin, D.L. (2003). Mental retardation and abnormal skeletal development (Dyggve-Melchior-Clausen dysplasia) due to mutations in a novel, evolutionarily conserved gene. Am. J. Hum. Genet. *72*, 419–428.

87. Weisfelt, M., Poll-The, B.T., and Hennekam, R.C.M. (2006). Severe neurological complications in skeletal dysplasias: Two case reports. Eur. J. Paediatr. Neurol. *10*, 241–243.

88. Below, J.E., Earl, D.L., Shively, K.M., McMillin, M.J., Smith, J.D., Turner, E.H., Stephan, M.J., Al-Gazali, L.I., Hertecant, J.L., Chitayat, D., et al. (2013). Whole-genome analysis reveals that mutations in inositol polyphosphate phosphatase-like 1 cause opsismodysplasia. Am. J. Hum. Genet. *92*, 137–143.

89. Huber, C., Faqeih, E.A., Bartholdi, D., Bole-Feysot, C., Borochowitz, Z., Cavalcanti, D.P., Frigo, A., Nitschke, P., Roume, J., Santos, H.G., et al. (2013). Exome sequencing identifies INPPL1 mutations as a cause of opsismodysplasia. Am. J. Hum. Genet. *92*, 144–149.

90. Fradet, A., and Fitzgerald, J. (2017). INPPL1 gene mutations in opsismodysplasia. J. Hum. Genet. *62*, 135–140.

91. Erneux, C., Edimo, W.E., Deneubourg, L., and Pirson, I. (2011). SHIP2 multiple functions: a balance between a negative control of PtdIns(3,4,5)P_3 level, a positive control of PtdIns(3,4)P_2 production, and intrinsic docking properties. J. Cell. Biochem. *112*, 2203–2209.

92. Thomas, M.P., Erneux, C., and Potter, B.V.L. (2017). SHIP2: structure, function and inhibition. Chembiochem *18*, 233–247.

93. Voigt, B., Frazier, K., Yazdi, D., Klein, J., Gontarz, P., Zhang, B., Sepich, D.S., Mo, J., Smeeton, J., Solnica-Krezel, L., et al. (2025). Cell expansion for notochord mechanics and endochondral bone lengthening in zebrafish depends on the 5′-inositol phosphatase Inppl1a. Curr. Biol. *35*, 1949–1962.e6.

94. Mathew, D., Kremer, K.N., and Torres, R.M. (2019). ARHGEF1 deficiency reveals Gα13-associated GPCRs are critical regulators of human lymphocyte function. J. Clin. Investig. *129*, 965–968.

95. Bouafia, A., Lofek, S., Bruneau, J., Chentout, L., Lamrini, H., Trinquand, A., Deau, M.C., Heurtier, L., Meignin, V., Picard, C., et al. (2019). Loss of ARHGEF1 causes a human primary antibody deficiency. J. Clin. Investig. *129*, 1047–1060.

96. Sur, A., Wang, Y., Capar, P., Margolin, G., Prochaska, M.K., and Farrell, J.A. (2023). Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. Dev. Cell *58*, 3028–3047.e12.

97. Girkontaite, I., Missy, K., Sakk, V., Harenberg, A., Tedford, K., Pötzel, T., Pfeffer, K., and Fischer, K.D. (2001). Lsc is required for marginal zone B cells, regulation of lymphocyte motility and immune responses. Nat. Immunol. *2*, 855–862.

98. Garcia, J., Bagwell, J., Njaine, B., Norman, J., Levic, D.S., Wopat, S., Miller, S.E., Liu, X., Locasale, J.W., Stainier, D.Y.R., and Bagnat, M. (2017). Sheath cell invasion and trans-differentiation repair mechanical damage caused by loss of caveolae in the zebrafish notochord. Curr. Biol. *27*, 1982–1989.e3.

99. Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H., and Valen, E. (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. *47*, W171–W174.

100. Olson, R.S., Urbanowicz, R.J., Andrews, P.C., Lavender, N.A., Kidd, L.C., and Moore, J.H. (2016). Automating biomedical data science through tree-based pipeline optimization. Appl. Evol. Comp., 123–137.

101. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

102. Voigt, B., Minowa, R., and Gray, R.S. (2023). Screening sperm for the rapid isolation of germline edits in zebrafish. J. Vis. Exp. https://doi.org/10.3791/64686.

103. McWhite, C.D., Papoulas, O., Drew, K., Dang, V., Leggere, J.C., Sae-Lee, W., and Marcotte, E.M. (2021). Co-fractionation/mass spectrometry to identify protein complexes. STAR Protoc. *2*, 100370.

104. Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., et al. (2012). A census of human soluble protein complexes. Cell *150*, 1068–1081.

105. Powell, D. (2019). drpowell/degust 4.1.1. Zenodo. https://doi.org/10.5281/zenodo.3501067.

106. Chen, Y., Chen, L., Lun, A.T.L., Baldoni, P.L., and Smyth, G.K. (2025). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Nucleic Acids Res. *53*, gkaf018.

107. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

108. Meeker, N.D., Hutchinson, S.A., Ho, L., and Trede, N.S. (2007). Method for isolation of PCR-ready genomic DNA from zebrafish tissues. Biotechniques *43*, 610–614.

109. Voigt, B., Frazier, K., Yazdi, D., Gontarz, P., Zhang, B., Sepich, D.S., Solnica-Krezel, L., and Gray, R.S. (2024). A conserved regulation of cell expansion underlies notochord mechanics, spine morphogenesis, and endochondral bone lengthening. Preprint at bioRxiv. https://doi.org/10.1101/2024.08.12.607640.

110. Adusumilli, R., and Mallick, P. (2017). Data conversion with ProteoWizard msConvert. In Proteomics: Methods and Protocols, L. Comai, J.E. Katz, and P. Mallick, eds. (Springer), pp. 339–368. https://doi.org/10.1007/978-1-4939-6747-6_23.

111. Drew, K., Müller, C.L., Bonneau, R., and Marcotte, E.M. (2017). Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. PLoS Comput. Biol. *13*, e1005625.

112. UniProt Consortium (2025). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. *53*, D609–D617.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-DMXL1	Proteintech	Cat#: 24413-1-AP; RRID:AB_2879532
anti-DMXL2	Proteintech	Cat#: 24415-1-AP; RRID:AB_2879534
anti-WDR7	Proteintech	Cat#: 24431-1-AP; RRID:AB_2879546
anti-ROGDI	Proteintech	Cat#: 17047-1-AP; RRID:AB_2182328
anti-RELCH/KIAA1468	Proteintech	Cat#: 18993-1-AP; RRID:AB_10598164
anti-PACS2	Proteintech	Cat#: 19508-1-AP; RRID:AB_10643387
anti-WDR37	Proteintech	Cat#: 20916-1-AP; RRID:AB_10695501
anti-GFP	Proteintech	Cat#: 50430-2-AP; RRID:AB_11042881
anti-FLAG	Proteintech	Cat#: 20543-1-AP; RRID:AB_11232216
Biological samples		
Pig (Sus scrofa)	Sierra Medical, CA	N/A
Chicken (Gallus gallus)	Sierra Medical, CA	N/A
Dolphin (Tursiops erebennus)	NOAA/NCCOS laboratory - Charleston, SC	N/A
Rabbit (*Oryctolagus cuniculus*)	Sierra Medical, CA	N/A
Mouse (*Mus musculus*)	Sierra Medical, CA	N/A
Chemicals, peptides, and recombinant proteins		
GoTaq Green	Promega	Cat# M7123
AltR S.p. HiFi Cas9 Nuclease V3	IDT	Cat# 1081060
Critical commercial assays		
Pierce Quantitative Colorimetric Peptide Assay	Thermo Scientific	Cat# 23275
Pierce BCA Protein Assay Kit	Thermo Scientific	Cat# 23225
Pierce Co-Immunoprecipitation Kit	Thermo Scientific	Cat# 26149
Deposited data		
Pig (Sus scrofa) CF-MS data	This paper	MassIVE: PXD061997
Chicken (Gallus gallus) CF-MS data	This paper	MassIVE: PXD062147
Dolphin (Tursiops erebennus) CF-MS data	This paper	MassIVE: PXD062225
Co-Fractionation Mass Spectrometry: Mouse (Mus musculus)	Pourhaghighi et al.[12]; Liebeskind et al.[13]	PRIDE: PXD011304
Zebrafish (Danio rerio)	This paper	MassIVE: PXD062174
Immuno-precipication Mass Spectrometry: Mouse (Mus musculus)	This paper	To be deposited
Experimental models: Organisms/strains		
Zebrafish (*Danio rerio*: Tubingen strain)	NA	Wild-type Tubingen strain
Zebrafish (*Danio rerio*: AB strain)	NA	Wild-type AB strain
Zebrafish (*Danio rerio*: *Tg(col8a1a:GFP-CAAX)*)	Garcia et al.[98]	*pd1152* allele
Zebrafish (*Danio rerio*: *arhgef1b* deficiency)	This paper	*ut6* allele
Oligonucleotides		
arhgef1b 5′-targeting CRISPR sgRNA	Synthego	5′-TGGTAGGACCTGCGACAGCA-3 with modifications to contain 2'-O'Methyl at the first three and last three bases and 2' -phosphorothioate bonds at the first three and last two bases

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
arhgef1b 3′-targeting CRISPR sgRNA	Synthego	5′-GCTAAATACACACTTCAGGG-3′ with modifications to contain 2′-O'Methyl at the first three and last three bases and 2' -phosphorothioate bonds at the first three and last two bases
arhgef1b Forward genotyping primer	IDT	5′-CTTGGCTTCTGGCAGTATTTTT-3′
arhgef1b Reverse genotyping primer, wild-type allele	IDT	5′-ACCAAATCAGATACAGGCCACT-3′
arhgef1b Reverse genotyping primer, df allele	IDT	5′-CGACTCCAGCACATACAGTGAT-3′
Software and algorithms		
Orthogroup inferance	Huerta-Cepas et al.[18]	https://github.com/eggnogdb/eggnog-mapper/releases
Peptide identification	Kwon et al.[17]	Set of all positive external protein pairs in https://github.com/marcottelab/Msblender
CHOPCHOP	Labun et al.[99]	v3
IMARIS, Cells and Surfaces Modules	OXFORD Instruments	v10.2.0
Feature extraction	McWhite et al.[8]	https://github.com/marcottelab/protein_complex_maps/blob/master/protein_complex_maps/features/ExtractFeatures/canned_scripts/extract_features.py
Model optimization	TPOT[100]	https://epistasislab.github.io/tpot/latest/
Model optimization	Cox et al.[25]	https://github.com/marcottelab/leca-proteomics/scripts/run_tpot.py
Feature/model selection	Cox et al.[25]	https://github.com/marcottelab/leca-proteomics/scripts/select_features.py
Final model generatioN/Assessment	Cox et al.[25]	N/A
Pairwise protein interaction labels	Cox et al.[25]	https://github.com/marcottelab/leca-proteomics/scripts/label_featmat.py
Protein complex detection	Cox et al.[25]	https://github.com/marcottelab/leca-proteomics/scripts/detect_communities.py
NES Elements	Nikon	v6.02.03
Image analysis	ImageJ/FIJI[101]	v2.14.0, https://imagej.github.io
Other		
Confocal Microscope	Nikon	CSU-W1 Yokogawa

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Sources and handling of brain samples

A Dolphin (*Tursiops erebennus*) brain was obtained from a juvenile 180 cm female *Tursiops erebennus* (Individual # MMES2016042SC; SC1629; 6-25-16) according to established necropsy standard operation procedures at the NOAA/NCCOS laboratory in Charleston, South Carolina under federal employee authority as outlined in Section 109(h) of the Marine Mammal Protection Act. The animal was positive for *Brucella ceti* in the lung and brain. The brain sample was collected and stored at −80 C, prior to shipping on dry ice for proteomic analysis under NOAA/NMFS permit 24359 as an Authorized Recipient (June 8, 2023). Because this dolphin species was only recently recognized as a coastal ecotype of the common bottlenose dolphin (*Tursiops truncatus*), and a reference proteome for *T. erebennus* is not yet available, we used the genome of *T. truncatus* (common bottlenose dolphin) for all computational analyses.

Frozen brains were obtained from the following additional vertebrate species as follows: 5 whole pig brains (Sus scrofa), 20 whole chicken brains (*Gallus gallus*) and 50 whole mouse brains (Mus musculus) from animals of unknown age/sex were obtained from Sierra for Medical Sciences (Whittier, California). 20 frozen rabbit (*Oryctolagus cuniculus*) brains were obtained from Pel-Freez Biologicals (Rogers, Arkansas). These brains were all harvested and immediately flash frozen prior to shipping on dry ice. Dissection of specific brain regions was performed in house as described below. 35 whole zebrafish (*Danio rerio, Tubingen strain,* 8–10 months post fertilization) brains were harvested in house and immediately flash frozen (see STAR Methods below). Both male and female zebrafish were used in approximately equal numbers.

Zebrafish husbandry

Zebrafish Maintenance and Care

All zebrafish experiments were performed according to University of Texas at Austin IACUC standards. Wild-type AB strains were used unless otherwise stated. Embryos were raised at 28.5°C in fish water (0.15% Instant Ocean in reverse osmosis water) and then transferred to standard system water at 5 dpf. Both male and female zebrafish were used for all adult assessments. As zebrafish do not exhibit sexual dimorphism until the early adult stages of development, sex is not a factor in any of the embryonic assessments described in this manuscript.

Generation of the arhgef1b^{df} *allele*

We generated a deficiency in the *arhgef1b* locus by co-injecting one-cell stage embryos with CRISPR sgRNAs targeting the 5′- and 3′-ends of the gene. Specifically, CHOPCHOP v3[99] was used to design sgRNAs targeting exon 2 near the start site (5′-TGGTAG GACCTGCGACAGCA-3′) and exon 30 near the stop site (5′-GCTAAATACACACTTCAGGG-3′) of the *arhgef1b* locus. The sgRNAs were acquired from Synthego with modifications to contain 2'-O'Methyl at the first three and last three bases and 2′-phosphorothioate bonds at the first three and last two bases. One-cell stage embryos were injected with 1 nL of 5 μM equimolar sgRNA-Cas9 protein in 0.1 M KCl. For F0 notochord imaging, we injected into the transgenic vacuolated cell membrane marker line, *Tg(col8a1a::GFP-CAAX)*.[98] Once the F0 fish reached sexual maturity, sperm was collected from F0 males[102] to screen for deficiency founders (see below). F0 founders were outcrossed to the *Tg(col8a1a::GFP-CAAX)* line, and F1 carriers were identified using the same genotyping method on gDNA from finclip samples. F1 heterozygous deficiency carriers were in-crossed for all imaging experiments.

Summary of RELCH mouse data, sourced from KOMP/IMPC

With permission from the Knockout Mouse Program (KOMP), we sourced the RELCH mouse knockout data, which can be accessed through the website of the International Mouse Phenotyping Consortium (IMPC). Phenotype data collection procedures can be accessed in detail at https://www.mousephenotype.org/impress/index.

METHOD DETAILS

Sample preparation, protein extraction, and co-fractionation/mass spectrometry

Dissection of pig and rabbit brains

Unless stated otherwise, all steps were performed on ice, and prior to dissection, we removed the pia mater of the brains using surgical design Teflon thumb forceps (Fisher Scientific) and washed gently with PBS pH 7.4 (Gibco, Thermo Scientific #10010023).

Pig brains were dissected immediately after receiving from the supplier using a disposable scalpel with a carbon steel blade into the following regions: frontal cortex, temporal lobe, occipital lobe, and cerebellum. Rabbit brains were thawed on ice until tissues were soft enough to dissect (regions as above). Following dissection, brain regions were flash frozen in liquid nitrogen and stored in −80°C until the proteomics analysis.

Native protein extraction

Unless stated otherwise, all samples were ground to a fine powder in liquid nitrogen using a chilled Potter-Elvehjem PTFE pestle and glass tube (Sigma-Aldrich, P7859). Powder was resuspended in an approximately equal volume of ice-cold lysis buffer (HEPES lysis buffer with NP-40, 2X, Thermo Scientific), with the addition of an EDTA-free protease inhibitor cocktail tablet (cOmplete, Roche) and phosphatase inhibitor tablet (PhosSTOP, Roche). The mixture was homogenized on ice using a chilled 2mL glass dounce with approximately 20–30 strokes by hand and incubated on ice for 30 min. The following steps were then conducted at 4°C: The crude homogenate was clarified twice using a tabletop centrifuge (5430R, Eppendorf) at 15,000 for 20 min. Further clarification of the supernatant was performed by ultracentrifugation at 30,000 x g for 1.25 h in an NVT65.2 rotor (Beckman Coulter).

Zebrafish brain dissection and protein extraction

Sexually mature *Danio rerio* (8–10 mpf, Tubingen strain) were euthanized in chilled 0.4% MS-222 (Tricaine, $C_{10}H_{15}NO_5S$) before being transferred to chilled phosphate-buffered saline (PBS). The operculum and jaw bones were removed with sharp forceps. The parietal and frontal bones of the skull were chipped away to expose the brain, which was removed by severing the spinal cord posterior to the medulla. The surrounding tissue was pulled away, and the eyes were removed by severing the optic nerves. Once isolated, the dissected brain tissue was briefly rinsed in PBS, transferred to a Protein LoBind tube (Eppendorf #022431081), and snap frozen in dry ice. Samples were stored at −80°C prior to lysis. For each chromatographic separation, brain tissue was pooled from 15 to 20 individuals in an approximately equal volume of chilled Pierce IP Lysis Buffer (Thermo Scientific #87788) with freshly added protease (cOmplete Mini, EDTA-free; #11836170001) and phosphatase (Roche PhosSTOP; #04906837001) inhibitors. The suspension was homogenized with a motorized Teflon pestle (60 passes) and in a 2 mL glass dounce (15 passes of pestle A then 15 passes of pestle B). The homogenized lysate was incubated on ice for 30 min with extensive pipetting every 10 min before centrifugation at 15,000 x *g* for 15 min at 4°C. The supernatant was transferred to a pre-chilled LoBind Protein tube, snap-frozen in liquid nitrogen or dry ice, and stored at −80°C until further processing. Further clarification of the supernatant was performed by ultracentrifugation at 30,000 x g for 1.25 h in an NVT65.2 rotor (Beckman Coulter) directly prior to use.

HPLC chromatography

Brain lysates were fractionated on a Dionex UltiMate3000 HPLC system consisting of an RS pump, Diode Array Detector, and Conductivity Monitor, Automated Fraction Collector (Thermo Scientific, CA, USA) and a Rheodyne MXII Valve (IDEX Health & Science LLC, Rohnert Park, CA) using biocompatible PEEK tubing and either size exclusion chromatography or one of two ion exchange separations (mixed bed, or triple-phase WAXWAXCAT, see below). We loaded 1–4 mg protein per separation as measured by DC Protein Assay (BioRad), collecting chromatography fractions in 2mL 96-deep well plates (Eppendorf). Select support ribs in the base were notched with a single-edged razor blade prior to fraction collection to accommodate subsequent use of the Life Technologies magnetic plate for on-bead tryptic digest during mass spectrometry sample preparation.[103]

Size exclusion chromatography (SEC)

BioSep-SEC-s4000 600 3 7.8 mm ID, particle diameter 5 μm, pore diameter 500 Å (Phenomenex, Torrance, CA). Prior to chromatography of brain lysates, molecular weight standards (Sigma -Aldrich, MWGF1000, 2–5 μg each of Thyroglobulin (T9145), β-amylase (A8781), and bovine serum albumin (A8531)) were fractionated for molecular weight estimation. Each chromatography run was conducted with mobile phase PBS pH 7.4 (Gibco), 200μL load volume, flow rate 0.5mL/min, and fraction collection every 45 s (0.375 mL).

Mixed bed ion-exchange chromatography (IEX)

Poly CATWAX A (PolyLC Mixed-Bed WAX-WCX) 200 3 4.6 mm ID, Particle diameter 5 μm, pore diameter 1000 Å (PolyLC Inc., Columbia, MD) comprises cation-exchange (PolyCAT A) and anion-exchange (PolyWAX LP) materials in equal amounts. Each chromatography run was loaded with 250μL sample at ≤ 40 mM NaCl. Fractionation was performed with a flow rate of 0.5 mL/min, fraction collection every 45 s (0.375 mL) and a 60-min gradient elution of 0–70% Buffer B: Buffer A (10 mM Tris-HCl pH 7.5, 5% glycerol, 0.01% NaN3), Buffer B (1.5 M NaCl in Buffer A).

Triple phase ion-exchange chromatography (WWC)

Three columns, each 200 3 4.6 mm ID, particle diameter 5 μm, pore diameter 1000 Å, were connected in series in the following order: two PolyWAX LP columns followed by a single PolyCAT A (PolyLC, Inc, Columbia, MD). Loading, buffers, and fraction collection were as for mixed bed ion exchange above with slight modifications in flow rate and elution from the methods of Havugimana et al. (2012).[104] The flow rate was 0.25 mL/min with a 120-min gradient from 5% to 100% Buffer B.

Mass spectrometry sample preparation

Samples were prepared for mass spectrometry in 96-well plate format using either Method 1 (ultrafiltration and in-solution digest) or Method 2 (on-bead digest) MS preparation protocols from McWhite et al., 2021.[103]

Mass spectrometry data acquisition and processing

Acquisition. Mass spectra were acquired using one of two Thermo mass spectrometers: Orbitrap Fusion or Orbitrap Fusion Lumos. In all cases, peptides were separated using reverse phase chromatography on a Dionex Ultimate 3000 RSLCnano UHPLC system (Thermo Scientific) with a C18 trap to C18 column configuration. Peptides were eluted using a 3%–45% acetonitrile gradient in 0.1% formic acid over 60 min. For the Orbitrap Fusion chromatography was on an Acclaim C18 PepMap RSLC column (Dionex; Thermo Scientific) with a nano-electrospray source. For the Orbitrap Fusion Lumos both chromatography and electrospray were achieved with an EASY-Spray PepMap RSLC C18 column (Thermo Scientific, ES902).

Orbitrap Fusion

Top speed CID with full precursor ion scans (MS1) collected at 120,000 m/z resolution and a cycle time of 3 s. Monoisotopic precursor selection and charge-state screening were enabled, with ions of charge > + 1 selected for collision-induced dissociation (CID). Dynamic exclusion was active with 60 s exclusion for ions selected once within a 60 s window. For some experiments, a similar top speed method was used with dynamic exclusion of 30 s for ions selected once within a 30s window and high energy-induced dissociation (HCD) collision energy 31% stepped +/4%. All MS2 scans were centroid and done in rapid mode.

Orbitrap Lumos

Top speed HCD with full precursor ion scans (MS1) collected at 120,000 m/z resolution. Monoisotopic precursor selection and charge-state screening were enabled using Advanced Peak Determination (APD), with ions of charge > + 1 selected for high energy-induced dissociation (HCD) with collision energy 30% stepped +/3%. Dynamic exclusion was active with 20 s exclusion for ions selected twice within a 20s window. All MS2 scans were centroid and done in rapid mode.

Immunoprecipitation/mass spectrometry

All IP-MS experiments were conducted using the Pierce Co-Immunoprecipitation Kit (Thermo Scientific, #26149), according to the manufacturer's instructions. Additional information on antibody identification is provided in the key resources table.

Following elution of IP-MS with 25 μl of elution buffer, 25 μl of 2% SDS/50mM NH_4HCO_3 solution was added. Samples were heated at 95°C for 5 min on a heat block to dissolve any visible tissues, immediately followed by overnight precipitation with 300μL precooled acetone in 4°C. On the next day samples were centrifuged at 13,000 rpm (Eppendorf, 5430R) to pellet proteins, and the supernatant was discarded. The pellets were washed twice with 400μL precooled acetone, followed by the same centrifugation and supernatant aspiration steps. To resuspend protein pellets back to solution, to each sample 50μL 1% sodium deoxycholate (SDC)/50mM NH_4HCO_3 were added and sonicated twice in water bath (10 min each). Samples were reduced with 5mM TCEP (Thermo Scientific, REF #77720) at 56°C for 45 min, alkylated with 25mM iodoacetamide (Sigma, I1149-5G) in the dark at room temperature for 45 min, reactions were quenched with 12mM DTT subsequently. Proteins were digested overnight at 37°C using 2μg trypsin (Thermo Scientific, REF #90058) per sample. After protein digestion, samples were acidified with 1% formic acid, precipitated SDC was removed

by centrifugation of samples at 16,000 g for 10 min. Flowthrough was brought to the Amicon 10kDa filter (Millipore Sigma, UFC501024), and desalted using MilliporeSigma ZipTip Pipet Tips (Fischer Scientific, Catalog No. ZTC18S096), dried in speedvac and stored at −20°C until further use for mass spectrometry.

Specific protein interactors (“preys”) were identified by proteomics from biological replicate experiments with at least 3 replicate mass spectrometry analyses (technical replicates, conducted on the Orbitrap Lumos instrument as detailed above) each, identifying proteins using ProteomeDiscoverer/Percolator and maintaining FDR<1% at the PSM, peptide, and protein level. Experiments were accompanied with paired negative control IP experiments conducted using antibodies targeting either GFP or the FLAG epitope tag (neither of which was present in the samples) to control for non-specific binding of proteins during the immunoprecipitation protocol. Significant interaction partners were identified using the software tool Degust[105] and modeling differential enrichment between experimental and non-specific control IP-MS experiments using the edgeR quasi-likelihood method[106] in concert with the remove unwanted variation (RUVr) method[107] for the normalization of peptide spectral counts between samples, selecting differentially enriched proteins with a false discovery rate of <1%. IP-MS experiments in which the known target of the antibody was not significantly enriched were discarded prior to subsequent analyses.

Zebrafish genetics and imaging

DNA extraction from sperm, finclip, and whole embryo samples

DNA was extracted from sperm, finclip, and whole embryos using the HotShot method.[108] Briefly, samples were lysed in 50 mM NaOH and heated to 95°C for either 40 (sperm) or 20 min (finclip and whole embryos) before neutralization with a 1:4 volume of 1 M Tris-HCl (pH 8.0). Samples were either immediately used for PCR amplification or were stored at −20°C until further processing.

Single PCR genotyping method for identification of arhgef1b^{df} allele

Carriers of the *arhgef1b^{df}* allele were identified by PCR using a forward primer upstream of the 5′-sgRNA cut site (5′-CTTGGCTTCTGGCAGTATTTTT-3′) and a reverse primer downstream of the 3′-sgRNA cut site (5′-CGACTCCAGCACATACAGT GAT-3′). Importantly, this primer pair does not produce an amplicon in the wild-type condition; therefore, the presence of an amplicon indicates individuals carrying at least one copy of the *arhgef1b^{df}* allele. All PCR reactions were performed using GoTaq Green Master Mix (Promega #M7123) and the following thermocycler settings: an initial denaturation at 95°C for 3 min, followed by 34 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 30 s, followed by a final extension at 72°C for 5 min. Amplicons were run on 2% agarose gels for 20 min.

Single PCR genotyping method for identification of wild-type arhgef1b allele

To distinguish between PCR failure and the absence of the deficiency allele, we designed a genotyping method to identify the wild-type *arhgef1b* allele. For this we used primers surrounding the 5′-sgRNA cut site (the same upstream 5′-TGGTAGGACCTGCGA CAGCA-3′ primer and a new downstream 5′-ACCAAATCAGATACAGGCCACT-3′ primer). Importantly, this primer pair does not produce an amplicon in the homozygous *arhgef1b^{df}* condition; therefore, the presence of an amplicon indicates individuals carrying at least one copy of the wild-type *arhgef1b* allele.

Duplex PCR genotyping method for distinguishing homozygous and heterozygous arhgef1b^{df} carriers

To distinguish wild type, heterozygous *arhgef1b$^{df/+}$*, and homozygous *arhgef1b$^{df/df}$* mutants in a single PCR reaction, we developed a duplex genotyping method. Since both deficiency and wild-type PCR reactions used the same forward primer (5′-TGGTAG GACCTGCGACAGCA-3′), we combined this with half-molar concentrations of each of the reverse primers (5′-CGACTCCAGCACA TACAGTGAT-3′ and 5′-ACCAAATCAGATACAGGCCACT-3′). The PCR products were run on 4% agarose gels for at least one hour. Wild-type *arhgef1b* amplicons ran around 250 bp, while *arhgef1b^{df}* amplicons ran slightly higher on the gel, and the presence of both bands indicated a heterozygous *arhgef1b$^{df/+}$* individual.

Confocal microscopy and analysis

F0 *arhgef1b* or stable F2s (from F1 *arhgef1b$^{df/+}$; Tg(col8a1a::GFP-CAAX)* in-crosses) were raised until 5 dpf (F2s) or 6 dpf (F0s), at which point they were anesthetized in 0.0016% Tricaine and mounted laterally in 1% low-melt agarose for imaging. Whole notochord stacks were acquired on the Nikon CSU-W1 Yokogawa Confocal Microscope and cell segmentation was performed in IMARIS (v10.2.0) as described in Voigt et al., 2025.[109] Briefly, all images were taken at 20X magnification with a 1.2 μm step size, and cell segmentation was performed using the IMARIS Cells and Surfaces modules with the following settings: Source channel 488 nm, cell smallest diameter of 20 μm, membrane detail of 2.0 μm, local contrast, and manual thresholding. All volume and distance datasets were exported and loaded into R for statistical analysis. Representative images were taken in IMARIS and colored by volume measurements.

QUANTIFICATION AND STATISTICAL ANALYSIS

Computational analyses of protein mass spectrometry datasets

We adopted a strategy similar to McWhite et al.[8] and Cox et al.,[25] matching mass spectral observations to orthogroups rather than individual proteins. An orthogroup (OG) is defined as a set of orthologous genes—genes in different organisms that originate from a single gene in their last common ancestor, which often encode proteins performing the same function across species. We considered orthogroups at the vertebrate level as calculated by the eggNOG[18] (evolutionary genealogy of genes: non-supervised orthologous groups) database of orthology relationships, which provided the framework for comparing proteomics data across vertebrates

and ultimately to human gene annotations. This orthogroup-based proteomics approach allows peptides shared among multiple proteins within an OG, but not unique to a single protein, to contribute to quantification. The spectra were subjected to stringent database searching and filtering (false discovery rate <1% at both peptide- and protein-level) using three different MS-search algorithms.[17]

Reference proteome databases

Reference proteomes for *Sus scrofa* (Taxon ID: 9823, UP000008227), *Gallus gallus* (Taxon ID: 9031, UP000000539), *Oryctolagus cuniculus* (Taxon ID: 9986, UP000001811), *Mus musculus* (Taxon ID: 10090, UP000000589), *Tursiops truncatus* (Taxon ID: 9739, UP000245320) *Danio rerio* (Taxon ID: 7955, UP000000437) and *Homo sapiens* (Taxon ID: 9606, UP000005640) were downloaded from https://www.uniprot.org/(The UniProt Consortium, 2019) in August 2020.

Orthogroup assignment & curation of orthogroup-based protein reference databases

Each reference proteome protein was assigned to an eggNOG vertebrata (verNOG, taxonomic level = 40674) orthogroup using the eggNOG-mapper v5 software and the eggNOG v5.0 diamond mapper (Huerta-Cepas et al., 2019[18]). Each set of proteins across all species assigned to the same orthogroup were operationally considered to belong to the same orthogroup for subsequent analyses. We updated each mass spectrometry reference proteome database by concatenating protein sequences from the same orthogroup into single entries with triple-lysines separating each protein, thus computing proteomics support on an orthogroup-by-orthgroup basis, rather than a protein-by-protein basis, which allowed for direct comparisons of proteomics data across the full vertebrate dataset.

Initial assignment of peptide mass spectra

For downstream analysis Thermo Fisher raw files were converted into mzXML format using GUI MSConvert.[110] Peptide inference was performed with MSGF+, X!Tandem, and Comet-2013020, each run with 10ppm precursor tolerance, and allowing for fixed cysteine carbamidomethylation (+57.021464) and optional methionine oxidation (+15.9949). Peptide search results were integrated with MSBlender (Kwon et al., 2011[17]), available from https://github.com/marcottelab/run_msblender.

For a given sample and mass spectrometry experiment, peptide spectral matches (PSMs) were analyzed as follows: for each reference database orthogroup, we summed the PSMs that could be uniquely attributed to that orthogroup. In this way, we avoided double-counting PSMs across highly related paralogs and thus near duplicate proteins, and we did not otherwise have to consider variable numbers of proteins or their relative lengths within each orthogroup in a species.

Identification and scoring of protein-protein interactions

Assembly of features for scoring putative protein interactions

For each experiment, we assembled an elution matrix of identified proteins with verNOG ID (rows) by fractions (columns). Additionally, we concatenated the elution matrices across all experiments within each species as four additional matrices. Next, we calculated a series of all-by-all pairwise scores between orthogroups for individual matrices and the concatenated matrices.

We focused our analysis on well-observed proteins, so we additionally filtered for proteins with at least 60 total PSMs observed across the 30 combined fractionations. The scores/features were as follows: (1) Pearson's r, (2) Spearman's rho, (3) Euclidean distance, (4) Bray-Curtis similarity, (Drew et al., 2017).[111] All features/scores were calculated with added Poisson noise as in Drew et al.[111] Euclidean distance and Bray-Curtis similarity scores were inverted and normalized to a max score of 1. For each individual matrix from each fractionation, features 1–4 were calculated using the extract_features.py script from https://github.com/marcottelab/protein_complex_maps/

Construction of the gold standard protein complex training and test sets

We used known mammalian protein complexes (human, mouse and rat) from the CORUM 4.1[27] release database (available at https://mips.helmholtz-muenchen.de/corum/download) as a gold standard positive set of stable protein-protein interactions. Known complexes were divided into positive training and test complexes according to the scheme from Cox et al.[25] (2024) and complexes with over 30 members removed. To address class imbalance, the total number of negative labels was restricted to 3X the observed number of positive PPIs in our data, resulting in 3,123 total positive PPIs and 9,369 total negative PPIs labels in our feature matrix. Note that for computing training/test sets and for identification and scoring of protein-protein interactions, we used the software pipeline developed for the LECA protein interactome,[25] available in full at https://github.com/marcottelab/leca-proteomics.

Positive PPIs that participate in multiple complexes are a potential source of representation bias in the truth set, which can lead to under or overfitting during model training. To overcome this, we implemented a data stratification approach, as described in Cox et al.[25]

Identification of interacting proteins by supervised machine learning

We utilized TPOT to train our machine learning model to find pairwise interactions among members of protein complexes. We first used the scikit-learn ExtraTreesClassifier feature selection to reduce the dimensionality of the feature matrix to the top 100 features based on declining feature importance (see Zenodo repository for data). We used the TPOT[100] AutoML wrapper of scikit-learn machine learning functions for all subsequent training steps. A group-based split method (scikit-learn's ''GroupShuffleSplit'' class) was used to generate 5 sets of test and training data, where approximately 75% of the labeled data was used for training and 25% for testing. Each of the 5 training sets were used as input into TPOT (Olson and Moore, 2016) an automated machine learning pipeline built on top of scikit-learn, to find the best classification method, pre-processing steps, and parameters for our data.

Construction of the protein interaction network

Interaction scores above a 8% false discovery rate threshold (CF-MS score $\geq$ 0.88) were used to build a PPI-based orthogroup network. In the network, nodes are verNOG ID mapped for corresponding proteins. This resulted in a network with 6,108 nodes and 81,225 edges. Two nodes (orthogroups) connected by an edge in this network indicate at least one interacting protein pair between the two nodes. Cuts closer to the root of the dendrogram result in larger and less granular complexes and cut closer to the tips defined smaller and finer granular subcomplexes.

Independent validation of CF-MS interactions by comparison to external protein interaction datasets

We classified the external datasets into two categories: all-by-all and bait-prey. For all-by-all (HuRI, mouse synaptosome, hu.MAP 2.0) we treated all $\frac{n(n-1)}{2}$ possible protein pairs as tested interactions. For bait-prey (AP-MS, OpenCell) we treated all possible bait-prey pairs as tested interactions, resulting in a total of $|bait| \times |prey| - \frac{(n_{bp}-2)(n_{bp}-1)}{2}$ possible interactions where $n_{bp} = |bait \cap prey|$. Any possible protein pair that was not classified as a positive protein pair was then classified as a negative protein pair. We divided the score range into 10 bins, $X = 1, \ldots,10$, such that the coverage of each bin was $\{[0.1(x\text{-}1),0.1x|x \in X]\}$, except for the final bin, which covered the range [1.0, 1.0]. This final bin was included to capture the highest possible score from the classifier.

For each bin, we performed the following:

We obtained a set of protein pairs from CF-MS for the given bin. This was the set of positive CF-MS protein pairs for a bin $CFMS^+$. We extracted all individual proteins $CFMS_{single}$ from the protein pairs $CFMS^+$ and enumerated a set of all possible CF-MS pairs $CFMS_{all}$ using all-by-all ($\frac{n(n-1)}{2}$ pairs). For a given bin, we ended up with 2 protein pair sets - $CFMS^+$ and $CFMS_{all}$.

We then repeated the following for each external data source E and for each bin:

We obtained the set of individual proteins from the protein pairs from CF-MS for a given bin $CFMS_{single}$. We retrieved the set of individual proteins for E (this does not change for a given bin since the external data's PPI value is binary rather than score-based), which we denoted as E_{single}. Then, we calculated the intersection between the two sets of individual proteins $E_{single}^{CFMS} = E_{single} \cap CFMS_{single}$. From E_{single}^{CFMS}, depending on the methodology of all-by-all or bait-prey as described earlier, we obtained our set of all possible protein pairs E_{all}. We calculated another intersection between E_{all} and the set of protein pairs from E to determine the set of positive E protein pairs for a specific bin, which we denoted as E^+. The negative pairs were determined by $E_{all}\text{-}E^+ = E^-$.

In all, we determined the following sets for a given bin and external dataset of protein pairs E:

$CFMS^+$- Set of positive CFMS protein pairs.

$CFMS_{single}$ - Set of all observed individual CFMS proteins.

$CFMS_{all}$ - Set of all possible CFMS protein pairs.

E_{single} - Set of all observed individual external data proteins.

E_{single}^{CFMS}- Set of positive external data proteins E_{single} that are also found in $CFMS_{single}$

E_{all} - Set of all possible external data proteins from E_{single}^{CFMS}

E^+- Set of all positive external protein pairs in E_{all} i.e., protein pairs found in both E and E_{all}

E^-- Set of all negative external protein pairs in E_{all} *i.e.*, protein pairs in E_{all} that are not found in E

The enrichment score was calculated as $\frac{A}{B}$ where $A = \frac{|CFMS^+ \cap E^+|}{|CFMS_{all} \cap E^+|}$ and $B = \frac{|CFMS^+ \cap E^-|}{|CFMS_{all} \cap E^-|}$.

The enrichment score is thus an indicator of how well the set of CFMS positive PPI overlaps with the set of external positive PPI while minimizing the overlap between CFMS positive PPI and external negative PPI for a given bin.

Evaluating pairwise interactions for evidence of direct binding using AlphaFold-Multimer

We analyzed all candidate pairs using ColabFold 2.3, which uses the same architecture and weights as AlphaFold 2.3. We evaluated all ColabFold results for direct binding by observing 4 scores - ipTM (interface predicted TM-score), iPAE (interface predicted aligned error) 10th percentile, iPAE minimum, and pDockQ (predicted DockQ). The ipTM is produced directly from AlphaFold-Multimer. The PAE is produced directly from AlphaFold and represents an $n \times n$ matrix where n represents the combined length of the protein pair. We derived the iPAE by taking the inter-chain PAE values. We used a cutoff of pDockQ $\geq$0.23 for an acceptable direct interaction between two proteins, as was used in the original pDockQ and DockQ papers.

Evaluating the predictive power of the measured interactions for biological processes

To measure the prediction power of our network, we tested whether we could predict novel proteins related to specific biological annotations (GO biological processes) given already known human proteins (seed sets) for the annotation.

We made a list of human proteins for each annotation and mapped these proteins to vertebrate-level orthogroups. We filtered out orthogroups not belonging to the network. Annotations having at least five orthogroups (in the network) were used for the following steps. We ranked all orthogroups in the network for each annotation based on the sum of edge weights linked to the seed set. We annotated a new score s_i for each orthogroup i following this formula:

$$s_i = \sum w_{ij}$$

Where j is an orthogroup in the seed set, and w_{ij} is the weight of the edge connecting i and j. Then orthogroups were ranked by their s_i. (There is no self-link, i.e., $i \neq j$)

Then, we measure the predictability of the network for each annotation by calculating the area under the receiver operating characteristic curve (AUROC). This value ranges from 0 to 1. 0.5 means a random classifier, and 1 represents a perfect classifier. The true positive rate (TPR) and false positive rate (FPR) were calculated by this formula:

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

Where TP (true positive) is the number of orthogroups belong the seed set and ranked above a threshold, FP (false positive) is the number of orthogroups not in the seed set but ranked above a threshold, FN (false negative) is the number of orthogroup in seed set and ranked below a threshold, and TN (true negative) is the number of orthogroups not in seed set and ranked below a threshold. We also calculated AUROC values with random seed sets, which have the same sizes as the annotations. Then, we compared the distribution of AUROC values for annotations of interest and the random.

Mapping of orthogroups to Disease Ontology terms

Human disease-gene associations data were retrieved from the Alliance of Genome Resource (Ver 5.3.0, Oct 28, 2022). Among five different types of associations, only "Is Implicated In" genes were used, so genes whose variants are known to function in causing or modifying diseases were assigned. Then, vertebrate-level orthogroups were mapped to each disease term based on the assigned human genes.

We inherited orthogroups of ancestor terms to their child terms following the structure of Disease Ontology. When an ancestor term and its child terms have the same orthogroups, we remain only the ancestor term. If several terms with the same orthogroups share a common ancestor on the list, we also removed those terms to avoid using excessively detailed disease terms.

Mapping of Gene Ontology terms considering the gold standard training set

We retrieved GO terms assigned for each human gene from the UniProt database.[112] Based on these human genes, we mapped vertebrate-level orthogroups for each GO term. Similar to Disease Ontology terms, we inherited orthogroups of ancestor terms to their child terms. Among the three categories of GO terms, we used GO cellular component (GOCC) and biological process (GOBP) terms.

Especially for GOCC terms, we separated terms based on their top ancestors: protein-containing complex, virion component, and cellular anatomical entity. Our data had no term whose top ancestor was the virion component and had assigned orthogroups. Next, we divided the protein-containing complex and cellular anatomical entity terms into terms that highly overlapped with the Gold Standard training sets and less overlapped ones. To determine the overlap between GOCC terms and the training sets, we calculated the Jaccard Index between one GOCC term and every training set and used the maximum value for each GOCC term. 0.5 was used as the threshold between high and low overlaps.