Open access • Posted Content • DOI:10.1101/2020.06.21.163709

Mapping cell structure across scales by fusing protein images and interactions — Source link

Yue Qin, Casper F. Winsnes, Edward L. Huttlin, Fan Zheng ...+10 more authors

Institutions: University of California, San Diego, Royal Institute of Technology, Harvard University, Purdue University ...+1 more institutions

Published on: 22 Jun 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Mapping cell structure across scales by fusing protein images and interactions

Yue Qin[1,2], Casper F. Winsnes[3], Edward L. Huttlin[4], Fan Zheng[1], Wei Ouyang[3], Jisoo Park[1], Adriana Pitea[1], Jason F. Kreisberg[1], Steven P. Gygi[4], J. Wade Harper[4], Jianzhu Ma[5], Emma Lundberg[3,6,7,]*, Trey Ideker[1,2,]*

[1]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA.

[2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA 92093, USA.

[3]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH-Royal Institute of Technology, Stockholm, Sweden.

[4]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA.

[5]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA.

[6]Department of Genetics, Stanford University, Stanford, CA 94305, USA. †

[7]Chan Zuckerberg Biohub, San Francisco, San Francisco, CA 94158, USA. †

† visiting appointment

*Correspondence: emma.lundberg@scilifelab.se (E.L.), tideker@ucsd.edu (T.I.)

ABSTRACT

The eukaryotic cell is a multi-scale structure with modular organization across at least four orders of magnitude[1,2]. Two central approaches for mapping this structure – protein fluorescent imaging and protein biophysical association – each generate extensive datasets but of distinct qualities and resolutions that are typically treated separately[3,4]. Here, we integrate immunofluorescent images in the Human Protein Atlas[5] with ongoing affinity purification experiments from the BioPlex resource[6] to create a unified hierarchical map of eukaryotic cell architecture. Integration involves configuring each approach to produce a general measure of protein distance, then calibrating the two measures using machine learning. The evolving map, called the Multi-Scale Integrated Cell (MuSIC 1.0), currently resolves 69 subcellular systems of which approximately half are undocumented. Based on these findings we perform 134 additional affinity purifications, validating close subunit associations for the majority of systems. The map elucidates roles for poorly characterized proteins, such as the appearance of FAM120C in chromatin; identifies new protein assemblies in ribosomal biogenesis, RNA splicing, nuclear speckles, and ion transport; and reveals crosstalk between cytoplasmic and mitochondrial ribosomal proteins. By integration across scales, MuSIC substantially increases the mapping resolution obtained from imaging while giving protein interactions a spatial dimension, paving the way to incorporate many molecular data types in proteome-wide maps of cells.

Advances in confocal microscopy and immunofluorescence (IF) imaging have created systematic pipelines for mapping the spatial distribution of proteins and other molecules within single cells[3,7,8]. Based on these techniques, the Human Protein Atlas (HPA) has launched an extensive effort to map protein subcellular locations using a library of specific fluorescent antibodies targeting more than 13,000 human proteins[5,9]. The use of multiple dyes with separate emission spectra enables locations to be determined relative to known landmarks such as the nucleus, cytoskeleton and endoplasmic reticulum, with the result that most human proteins can be assigned relative positions at sub-micron resolution.

In parallel, advances in mass spectrometry (MS) have provided a complementary means of mapping protein coordinates through their biophysical associations with other proteins[4,10]. MS is now routinely combined with affinity purification (AP-MS)[11] or proximity-dependent labeling[12–16] to enumerate physical protein-protein interactions *in vitro* or *in vivo*. Combining AP-MS with epitope tagging, the BioPlex project has generated a systematic map of physical interactions covering approximately 7,500 epitope-tagged human proteins (BioPlex 2.0)[6], including approximately 56,000 candidate interactions organized into more than 400 multimeric protein complexes.

Given that protein imaging and biophysical association are leading approaches for mapping cell structure, with growing data resources and proteome-wide maps, a key question is whether and how they should be properly combined. One platform is sometimes used to validate results of the other, for example testing whether proteins associated by AP-MS co-localize within IF images[17]. The two platforms provide complementary measures of protein localizations, albeit at different physical scales, with IF positioning proteins globally relative to broad cellular landmarks, and AP-MS positioning proteins locally relative to other nearby proteins. In either case, such positioning has become increasingly quantitative in recent years, based on the ability of deep learning systems to recognize complex patterns in data[18–23]. Once an IF or AP-MS dataset has been analyzed to determine protein positions (or position distributions), a natural next step is to compute distances between these distributions, raising the possibility that the two types of protein distance might be calibrated and combined within integrated maps of cellular architecture (**Fig. 1a**).

Protein position and distance, two ways

We assembled a matched dataset consisting of IF images from HPA[5] and AP-MS data from BioPlex[6]. Both resources are partially based on HEK293 human embryonic kidney cells, leading us to 661 proteins with compatible imaging (1,451 images including replicates, **Extended Data Fig. 1**) and biophysical association data (291 proteins affinity-tagged as 'baits', the remaining 370 arising as interacting 'preys', **Supplementary Table 1**). While the number of studied proteins could be expanded by pooling additional IF and AP-MS datasets gathered in disparate cell types and conditions, we considered the importance of a controlled cellular context in prototyping any new approach. Using a deep convolutional neural network trained to recognize patterns in IF images[18], we embedded each protein as a 1024-dimension feature vector, capturing its spatial distribution relative to counter-stained cellular landmarks (**Methods**). Similarly, the node2vec deep neural network[20] was used to embed each protein as a second 1024-dimension feature vector based on its interaction neighborhood within the AP-MS data, including directly and indirectly associated proteins (**Methods**, **Extended Data Fig. 2**).

Next we computed protein-protein distances (cosine distance) for all pairs of proteins, separately in the IF and AP-MS high-dimensional embeddings. We found that the closest protein pairs measured by one technique were significantly enriched for those measured as close by the other, demonstrating that despite their differences, the two measurement types share significant information (**Fig. 1b, c**). As a means of calibrating the two distance measures, we sampled subcellular components of known physical size spread over several orders of magnitude in scale, from protein complexes of <20 nm to organelles >1 μm in diameter. The diameter of each component strongly correlated with its number of protein subunits documented in the Gene Ontology (GO)[24,25], suggesting a general means of converting current GO annotations for a protein pair to an approximate distance in nanometers (**Fig. 1d**). Using these approximated distances as training examples, we taught a supervised machine learning model (random forest regression, **Methods**) to translate patterns in the concatenated IF and AP-MS features to an integrated measure of pairwise protein distance (**Fig. 1e**).

A subcellular hierarchy at 20-20,000 nm

We analyzed this full set of protein distances to identify communities of proteins in close mutual proximity, suggesting distinct cellular components. Protein communities were identified at multiple resolutions, starting with those that form at the smallest protein-protein distances then progressively relaxing the distance threshold (multi-scale community detection[26,27], **Methods**). Communities at smaller distances were contained, in full or in part, inside larger communities as the threshold was relaxed, yielding a structural hierarchy (**Fig. 2a**). The sensitivity of community detection was tuned for best concordance with two independent datasets not used elsewhere in our study: a separate collection of protein interactions reported in the Human Cell Map[28] using proximity biotinylation, also in HEK293 cells, and patterns of gene co-essentiality observed in the Cancer Cell Dependency Map[29,30] (**Methods**). This exercise also provided an end-to-end validation of our analysis pipeline (**Fig. 1a**), as the agreement with the outside datasets was significant over a wide range of community detection parameters (**Extended Data Fig. 3**). The final hierarchy, which we call the Multi-Scale Integrated Cell (MuSIC 1.0), contained 69 protein communities representing putative subcellular systems organized by 87 hierarchical containment relationships (**Fig. 2a**, **Supplementary Table 2**). Sixteen systems were contained in multiple larger ones, suggesting these systems have pleiotropic roles or are intermediate subassemblies of a larger complex. To elucidate the biological roles of each system, we aligned the MuSIC hierarchy to the equivalent literature-curated hierarchy provided by GO (**Methods**). A total of 46% of systems had significant overlap with GO; the remaining 54% did not clearly correspond to a known component and were labeled as putative novel (**Fig. 2a**). Where possible, systems were labeled by synthesizing prior literature with our own biological knowledge and reasoning.

For example, MuSIC identified a community of nine proteins (labeled "Spliceosomal snRNP complex family", **Fig. 2b**), the majority of which were specific members of the U1 and U2 small nuclear ribonucleoproteins (snRNPs) or had snRNP-related activities in previous studies[31–33]. Co-location of these proteins was supported by IF, which showed a common nucleoplasm signal (**Fig. 2c**) and corresponding close distances in the IF embedding (**Fig. 2d**). The image distributions varied considerably, particularly across nucleoplasm and cytoplasm, such that most but not all protein pairs were closely associated. Close proximities of these proteins were reinforced by the AP-MS data, however (**Fig. 2e, f**), based on similar proteome-wide patterns of interaction. The combined support from both data types induced this spliceosomal complex to form a specific protein community at a resolution of around 150 nm (**Fig. 2g**, **h**).

Added sensitivity, resolution, context

We next examined the sensitivity with which these same systems could be identified using IF or AP-MS only, versus the full dataset. In general, we found that IF tends to robustly identify large systems such as organelles and sub-organellar parts that form the upper structure of MuSIC, whereas AP-MS robustly identifies smaller subcomponents within IF systems (**Fig. 3a**, **b**,

Methods). For the majority of systems, the highest robustness was observed when integrating both types of data (**Fig. 3c**, **d**). For instance, the ability to recover the spliceosomal snRNP complex family dropped considerably when using only AP-MS data compared to integrated analysis (**Fig. 2h**).

Further insight came from examining the complete distributions of distances inferred using each data type separately (**Fig. 3e**). Using IF only, the measured protein distances were generally >200 nm with a median of approximately 5000 nm – the scale of large intracellular assemblies and organelles. Using AP-MS only, the median protein distance was also around 5000 nm, as expected if most proteins do not physically interact; however, a fraction of pairs were assigned extremely short distances of 20-200 nm – the realm of protein interactions and complexes. By integrating both data types, MuSIC distance measurements spanned four orders of magnitude in physical scale, ranging from roughly 20 to 20,000 nm (**Fig. 3e**).

Notably, 30% of AP-MS protein interactions fell within a focused system of <100 proteins (**Fig. 3f**). In each of these cases, such knowledge validates and aids in biological interpretation of the protein-protein association. We found that such context also allows rescue of protein interactions with support in the raw AP-MS spectra but which were overlooked in previous proteome-wide analysis due to the stringent scoring thresholds necessary to control for false discoveries. Among protein pairs not reported to interact in the previous BioPlex study[6], pairs in smaller systems had significantly stronger AP-MS scores than pairs in larger systems ($p < 0.0001$, **Fig. 3g**), suggesting an untapped trove of bona-fide physical interactions.

Exploration of newly identified systems

Of the 661 proteins common to the IF and AP-MS datasets, 370 had not yet been affinity tagged as the central baits of an AP-MS experiment – rather, they had appeared in the list of preys isolated by another affinity-tagged protein. Therefore, as an immediate means of validating candidate systems in MuSIC, we created affinity tags for 134 former preys and performed AP-MS, resulting in the identification of 339 physical interactions (**Supplementary Table 1**). We found that 44/69 MuSIC systems were specifically enriched for the new interactions (64%, FDR < 0.1, **Fig. 4a**), including 23 "putative" novel systems. Conversely, 195 (58%) of new interactions fell into MuSIC systems of <100 proteins, placing these into specific subcellular contexts (**Extended Data Fig. 4**).

Among the novel systems validated by the additional interaction data was an assembly of seven proteins (**Fig. 4b-f**), which we named 'Pre-ribosomal RNA (pre-rRNA) processing complex 1' based on synthesizing information about known protein activities[34,35] (NVL, RPL13A) with results from human high-throughput genetic screens[36] (KRI1, NOC2L) and orthology to budding yeast[37] (REXO4, **Fig. 4e**). The close association of these proteins was based on IF similarity (predominantly nucleolar localizations, **Fig. 4b, c**) and common AP-MS interaction neighborhoods (**Fig. 4d**), although, notably, none of the protein pairs had been reported to directly interact in BioPlex (**Fig. 4f**). Our new affinity purifications targeted five of these proteins as baits, recovering direct interactions for the majority of protein pairs in the system (11/21, **Fig. 4f**). We noted that the recovery of interactions was specific to that system (52.4% interactions recovered), in comparison to interactions with proteins in closely related systems (23.6%) or elsewhere in MuSIC (0.5%, **Extended Data Fig. 5**).

A second example was a system of seven proteins featuring abundant crosstalk between canonical subunits of the cytoplasmic and mitochondrial ribosomes (**Fig. 4g-k**). All showed cytoplasmic staining, explaining their close IF distances (**Fig. 4g-h**, staining is dim for MRPS9, MRPS14 and MRPS31 compared to their predominant mitochondrial locations). The IF similarity was further supported by AP-MS neighborhood similarity, although majority of the protein pairs had not yet been reported to physically interact (**Fig. 4i-k**). In the new affinity pull-downs, we tagged four of these proteins as baits (two cytoribosomal, two mitoribosomal), identifying five direct within-system physical interactions, four interconnecting cytoplasmic and mitochondrial factors (**Fig. 4k**). Such crosstalk has not been previously reported but may play a role in mitoribosome biogenesis, a poorly understood process[38]. Based on these findings, we named this system ‘Mito-cyto ribosomal cluster’.

A number of systems captured novel protein associations apart from those we had targeted with the new AP-MS data. One example (**Fig. 4l-o)** was a system of six proteins, of which DMAP1, MORF4L1, and JAZF1 were histone acetyltransferases[39,40], SATB1 had been shown to interact with histone acetyltransferases[41], SRRM1 had been implicated in chromatin remodeling[42], and FAM120C had no previously assigned function (**Fig. 4o**). The specific association of these proteins was supported by their close measured distances in both IF (nucleoplasm and nuclear speckles, **Fig. 4l**, **m**) and AP-MS data (**Fig. 4n**). We thus named this system ‘Chromatin regulation complex 1’ based on the previously documented protein functions (**Fig. 4o**). The close associations of FAM120C, a candidate disease protein in autism spectrum disorder[43], with proteins involved in regulation of chromatin suggests a novel function for this protein, consistent with the previous identification of other chromatin factors in autism[44].

Discussion

In classical image analysis of cells, the locations of proteins are identified in reference to a small panel of known markers for organelles. Here, MuSIC moves away from this closed library of organellar locations to an open platform in which both existing and new structures are identified *de novo* from structure in data. Numerous synergies between the IF and AP-MS data support the identification of 69 hierarchically-organized protein systems, many of which are robustly identified only when integrating both data modalities (**Fig. 3d**).

What about when the data disagree? While nearly a third of AP-MS interactions fall within the same focused system of <100 proteins, more than two thirds do not (**Fig. 3f**). For example, PPP6R1, a phosphatase, and NPAS1, a helix-loop-helix transcription factor, interact directly by AP-MS but were placed in different organelles in MuSIC related to their distinct image locations (PPP6R1, Cytoplasm; NPAS1, Nucleus; **Extended Data Fig. 6a**). In general, such discrepancies may indicate rare physical association of the proteins in a common compartment despite their more abundant locations in distinct others, as might be expected from pleiotropic roles or stages of protein maturation[45]. Alternatively, the divergent measurements may be explained by temporal dynamics, by which the two proteins interact transiently or periodically (e.g. cell cycle, circadian rhythms). Discrepancies may also derive from the fact that IF detects endogenous proteins whereas AP-MS detects over-expressed tagged proteins. Barring the above explanations, one can always suspect errors in data, e.g., cases in which the physical interaction represents spurious or non-specific binding. Some disagreement between IF and AP-MS can clearly be tolerated by the system, such as MuSIC's correct assignment of GEMIN7 and SNRNP7 into the U1 small nuclear ribonucleoprotein particle[32,46] (U1 snRNP, **Fig. 2a**), despite only partial overlap in the image distributions of the two proteins (**Extended Data Fig. 6b**). In this case, correct assignment was facilitated by the direct association of these proteins by AP-MS.

While the imaging field is accustomed to thinking about physical sizes and distances between cellular objects, the idea that protein interactions can be calibrated to measure physical distance is, to our knowledge, new to this study. The estimates here suggest that protein networks can measure sizes of cellular components down to about 20 nm (**Fig. 2a**). In contrast, the minimum size resolution learned for IF imaging, at approximately 200 nm (**Fig. 3e**), is in good agreement with conventional estimates[47]. This type of integrated analysis, with distance calibration, allows discovery of biological structure across a wide range of physical scales, including those scales that fall between the windows of resolution for each individual assay. It will be interesting to explore the size resolutions of other omics technologies, many of which might be calibrated to measure molecular distances and, in turn, contribute to integrated maps of the multi-scale cell.

Acknowledgements

We gratefully acknowledge helpful discussion and comments from Cherie Ng, members of the Ideker laboratory, members of the Lundberg laboratory, the Human Protein Atlas, and the anonymous referees. This work was supported by the National Institutes of Health (NIH) under grants P41 GM103504 and R01 HG009979 to T.I., U24 HG006673 to E.L.H., S.P.G, and J.W.H., R50 CA243885 to J.F.K., and by the Knut and Alice Wallenberg Foundation (2016.0204) and the Swedish Research Council (2017-05327) to E.L..

Author Contributions

Y.Q., E.L. and T.I. designed the study and developed the conceptual ideas. C.F.W. and W.O. generated image features. E.L.H., J.W.H., and S.P.G. generated AP-MS data. Y.Q. and J.M. designed the data integration approach. Y.Q. and F.Z. designed the community detection approach. All authors contributed in developing ideas for data analysis. Y.Q. implemented all computational methods and analyses. Y.Q. and T.I. wrote the manuscript with input from all other authors.

Competing Interest Declaration

T.I. is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. T.I. is on the Scientific Advisory Board of Ideaya BioSciences, Inc., and has an equity interest. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies. J.W.H. is a co-founder of Caraway Therapeutics, is on the Scientific Advisory Board, and has equity interest. J.W.H. is a consultant for X-Chem Inc.

METHODS

Data sources

IF images interrogating protein locations in the HEK293 cell line were downloaded from the HPA Cell Atlas[5,18,19]. Physical protein interactions detected by AP-MS in the HEK293T cell line were downloaded from the BioPlex 2.0 protein interaction database[6]. We focused our study on the intersection of these IF and AP-MS datasets, i.e. selecting images of immunofluorescent proteins that had been affinity-tagged as baits or detected as preys in BioPlex. Each image had four channels, one for the protein of interest, and three for reference markers including nucleus, microtubules and endoplasmic reticulum. Images involving antibodies that targeted more than one protein, or which lacked annotated cellular localizations, were removed. The final imaging dataset contained 1,451 images covering 661 proteins and 726 antibodies, corresponding to a range of 2-6 images per protein (**Extended Data Fig. 1**). The final AP-MS dataset covered this same set of 661 proteins with 281 direct protein-protein interactions among the 661 proteins; we also retained the entire BioPlex 2.0 network of 10,961 proteins and 56,553 protein interactions, which provides significant information about the extended network neighborhoods of the 661 proteins (used in data embedding, see below).

Data embeddings

Three different image embedding methods (DenseNet, SLF, Paired Cell Inpainting) were compared for their ability to enrich for the 281 BioPlex protein-protein interactions among the 661 proteins (**Extended Data Fig. 7a**). DenseNet: For each image, a 1024-dimension feature embedding was generated from a DenseNet-121 model[48] optimized for annotating HPA images as previously described[18]. This method uses the protein channel and all three reference channels. SLF: Subcellular location features (SLF) were extracted at a per-cell level following the procedure described previously[19,49,50]. Here we used the same set of 719 features as in the previous SLF study[19], capturing the relative relationship between the immunostained protein and the two major HPA cellular landmarks, nucleus and microtubules. Paired Cell Inpainting: The Paired Cell Inpainting[51] unsupervised method uses a deep neural network to extract a 384-dimension feature vector for each protein, which was downloaded directly from http://hershey.csb.utoronto.ca/paired_cell_inpainting_features. For the SLF and Paired Cell Inpainting embeddings, per-image features were obtained by averaging the features of the cells within an image; the DenseNet embedding is already provided at the image level. We chose the DenseNet embedded features due to their ability to best enrich for physical protein interactions (**Extended Data Fig. 7a**).

For AP-MS data, three different embedding methods (node2vec, RWR, node properties) were compared for their ability to enrich for the 281 protein pairs with highest cosine similarity calculated in the IF images using the DenseNet embedding (**Extended Data Fig. 7b**; the number 281 was selected to match the number of protein-protein interactions used above for the reciprocal analysis). All three embedding methods input a protein-protein interaction network, in which nodes represent proteins and edges represent measured protein-protein interactions drawn from all available BioPlex AP-MS data (n = 10,961 nodes, m = 56,553 edges, see above). Node2vec: This method uses a deep neural network to learn feature representations of the network neighborhood surrounding each node[20]. A 1024-dimension feature embedding for each protein was obtained by applying node2vec with parameters $p = 2$, $q = 1$. RWR: Random walk with restart (RWR) is carried out according to equation:

$$Q_{t+1} = (1-\alpha)Q_0 + \alpha Q_t A \quad (1)$$

α tunes the distance that a protein signal diffuses within the protein-protein interaction network, here set to α = 0.6, which was determined to be the optimal setting in a previous human protein network study[52]. Q_0 is an $n \times n$ identity matrix, and A is the column-normalized adjacency matrix representation of the network[53]. The embedding of each node was taken as the corresponding row of Q_t if eqn. (1) converged or after 30 iterations. Node properties: Each protein embedding was represented by a short feature vector formulated from the following graph-theoretic properties: (i) degree, (ii) clustering coefficient, (iii) closeness centrality, (iv) normalized betweenness centrality, (v) eigenvector centrality, (vi) stress centrality, (vii) bridging centrality, and (viii) information centrality, following definitions from a previous study[54]. Node2vec embeddings were chosen based on superior performance in enriching for protein pairs with similar image embeddings (**Extended Data Fig. 7b**).

Data integration

Protein annotations from the Cellular Component branch of the Gene Ontology (GO)[25] were used to formulate a training set of approximate protein pairwise distances. GO was downloaded on 25.9.2018 from http://geneontology.org. Annotations based on HPA images were removed from GO to avoid circularity. Among the 661 proteins under study, 602 had specific GO annotations (i.e. other than the cellular_component root term) which were used to supervise the calibration of distance from embedded features. The relative GO similarity (converse of distance) for a protein pair p_1 and p_2 was calculated according to a previously proposed metric[55] as:

$$Similarity(p_1, p_2) = -log_{10} \frac{|T(p_1,p_2)|}{N} \qquad (2)$$

where $|T(p_1, p_2)|$ is the number of proteins in the smallest term to which p_1 and p_2 are both annotated and N is the total number of proteins with specific annotations in GO.

Random forest regression[56] implemented in the Python scikit-learn package[57] was used to learn the approximate GO similarity, scaled to [0, 1], between a pair of proteins. The training features included the complete image and AP-MS embeddings of each protein in the pair as well as a panel of similarity measurements (cosine, Pearson, Spearman, Kendall, Manhattan, and Euclidean) of the protein pair in the image and AP-MS embeddings respectively. Because every protein had multiple images (ranging from two to six, see **Data sources** above), six different groups of image-to-protein mappings were randomly generated while requiring each image for a protein to be mapped at least once. We trained random forest regression with five-fold cross validation for each group. The final similarity of each protein pair (henceforth called the *integrated protein-protein similarity network*) was obtained by averaging the random forest predictions from the six groups. As a negative control, a 1024-dimension random vector sampled from a normal distribution was generated for each image embedding and each AP-MS embedding (**Fig. 1e**).

Pan-resolution community detection

The integrated protein-protein similarity network was analyzed to detect distinct communities of similar proteins using the Clique eXtracted Ontology algorithm[27] (CliXO v1.0, https://github.com/fanzheng10/CliXO-1.0). CliXO finds the maximal cliques in a weighted network while progressively decreasing the threshold on edge weights. Lower thresholds yield cliques that may contain, in full or in part, cliques identified at higher thresholds, resulting in a hierarchy of communities interrelated by community containment relations. CliXO has four parameters that control the depth (α), width (β), modularity (m) and modularity significance (z) of the community hierarchy (**Extended Data Fig. 3a**). We swept through 500 different combinations of these four parameters to obtain a pool of hierarchies. All communities, representing putative cellular systems, were required to have at least four proteins to further ensure quality and validity.

To select an optimal hierarchy, each hierarchy was evaluated based on its concordance with two independent datasets, the Human Cell Map[28] and the Cancer Cell Dependency Map (v18Q2)[29], which were not used elsewhere in this study (**Extended Data Fig. 3b, c**). From the Human Cell Map, 178 protein-protein interactions detected in HEK293 cells with high-confidence (FDR ≤ 0.01) were obtained, covering 293 proteins in MuSIC. From the Cancer Cell Dependency Map (DepMap)[29], we selected 9,030 protein pairs, covering 629 proteins in MuSIC, for which the CRISPR gene disruptions of the two proteins led to significantly correlated fitness profiles across the panel of 436 DepMap cell lines (so-called "co-essential" proteins). For each hierarchy, we recorded the number of Human Cell Map protein-protein interactions (x) or DepMap co-essential protein pairs (y) that were covered by systems that were significantly enriched (FDR ≤ 0.1) for those interactions (**Extended Data Fig. 3b, c**). The hierarchy having the highest number ($x \times y$) was selected for further study; among the several ties, we selected the hierarchy with the least number of systems (i.e. guided by the principle of parsimony).

The hierarchical structure was further matured by assigning additional hierarchical parent-child containment relations between pairs of systems having a containment index ≥0.75 and removing redundant systems having Jaccard index ≥0.9 with parent systems. The containment and Jaccard indexes between two systems S_1 and S_2 were calculated based on the following formulae:

$$Containment(S_1, S_2) = |S_1 \cap S_2| \, / \, min(|S_1|, |S_2|) \qquad (3)$$

$$Jaccard(S_1, S_2) = |S_1 \cap S_2| \, / \, |S_1 \cup S_2| \qquad (4)$$

with $|S|$ representing the number of proteins in a system. Redundant systems having only one protein difference from a parent system were also removed. When removing a system, the integrity of the hierarchical structure was maintained by adding containment relations from all children of that system to all of its parents.

We also analyzed the integrated protein-protein similarity network using Louvain clustering[26,58] (https://github.com/vtraag/louvain-igraph, v0.6.1) which partitions network nodes into a set of distinct clusters. We ran Louvain 1,000 times and selected the partition that maximized the global modularity, as output by the algorithm. The partition used for the main MuSIC model included just two clusters, which were strongly enriched for proteins with known subcellular locations in the cytoplasm versus nucleus, respectively (**Fig. 2a**). These clusters were added as parent systems of the top-layer systems found by CliXO (described above). In particular, a CliXO system was added as a child of a Louvain system if at least 50% of its proteins were in the Louvain system(s). A similar process was employed to construct the jackknifed hierarchies in the robustness analysis (**Fig. 3a-c**, see **Dependence of systems on data types** below). For systems highlighted in main figures (**Figs. 2a, 4a**), we introduced a further quality control step in which we manually inspected the corresponding IF images and raw AP-MS spectra. This process prompted us to remove the protein RPL6 from the system "Mito-cyto ribosomal cluster" out of concerns for antibody correctness.

To label MuSIC systems as "known" or "putative" (**Fig. 2a**), MuSIC was aligned to the Cellular Component branch of GO, filtered for the proteins under study. A system S was considered "known" if there existed a GO term T that was significantly enriched (FDR $\leq$ 0.001, hypergeometric statistic) for proteins in S with $Jaccard(S,T) \geq 0.4$, or if $Jaccard(S,T) = 1$, representing perfect agreement regardless of significance.

Calibration to physical distance

The following calibration function was obtained using linear regression on the list of subcellular compartments with known physical sizes (**Fig. 1d**):

$$log_{10} D_{nm}(p_1, p_2) = -3.946 \times Similarity(p_1, p_2) + 4.317 \qquad (5)$$

where $D_{nm}(p_1, p_2)$ is the physical distance in nm between proteins p_1 and p_2, and $Similarity(p_1, p_2)$ is calculated using eqn. (2) based on the number of proteins annotated to the compartment. The physical size of each MuSIC system (**Fig. 2a**) was calibrated based on the similarity threshold at which the system was found.

Dependence of systems on data types

To investigate the relative contributions of IF and AP-MS data in system formation, we evaluated the robustness of each system under perturbations to the integrated protein-protein similarity network (**Figs. 2h, 3a-c**). To assess the dependence of each system on imaging data, we created an alternative network using IF features only, with AP-MS features randomized (see **Data integration** above). Subsequently, 10% of the edges in this protein network were randomly removed, and community detection was performed to construct a hierarchy using the same parameters as MuSIC (see **Pan-resolution community detection** above). This randomization procedure was repeated 300 times to obtain a pool of perturbed hierarchies, similar to statistical jackknifing[59]. The “percentage recovery” of a MuSIC system S_{MuSIC} in a perturbed hierarchy H_P was calculated as:

$$\%\,recovery\,(S_{MuSIC}, H_P) = 100 \times Jaccard(S_{MuSIC}, S_P) \qquad (6)$$

where S_P is the perturbed system best enriching for S_{MuSIC} in H_P (**Fig. 2h**). S_{MuSIC} was considered to be “recovered” by H_P if the enrichment of S_{MuSIC} in S_P was significant ($FDR \le 0.001$, hypergeometric statistic) and $\%\,recovery(S_{MuSIC}, H_P) \ge 40\%$, or if $\%\,recovery(S_{MuSIC}, H_P) = 100\%$, representing perfect agreement regardless of significance. The image robustness score for S_{MuSIC} (**Fig. 3a**) was the fraction of all perturbed hierarchies that recovered S_{MuSIC}.

To assess the dependence of each system on AP-MS data, a reciprocal procedure was performed by generating an alternative integrated protein-protein similarity network using AP-MS features only, with imaging features randomized (**Fig. 3b**). As a benchmark, we also computed the robustness of the full MuSIC model using the original integrated protein-protein similarity network without IF or AP-MS feature randomization (**Fig. 3c**), using the same jackknifing procedure described above.

Global system validation using new AP-MS data

We constructed stable HEK293T cell lines for 134 bait proteins (**Supplementary Table 1**) with C-terminal FLAG-HA-tags based on the human ORFeome v8.1 (http://horfdb.dfci.harvard.edu/)[60] as previously described[6,61,62]. Cell pellets were lysed using 50 mM Tris-HCl pH 7.5, 300 mM NaCl, 0.5% (v/v) NP40 buffer, and cell debris were removed with centrifugation and filtration. Mouse monoclonal anti-HA agarose resins (Sigma-Aldrich, clone HA-7), immobilized and pre-washed, were incubated with cell lysates at 4 °C for 4 hours. After removing supernatant, precipitates were washed four times with lysis buffer and two times with PBS (pH 7.2). Elution was performed in two steps by adding 250 μg/mL HA peptide in PBS at 37°C followed by TCA precipitation. Eluted samples were analyzed by LC-MS using Q-Exactive mass spectrometers (Thermo Fisher). Each bait protein was analyzed with biological duplicates for the affinity purification step and technical duplicates for the LC-MS step, yielding four replicates in total. MS/MS spectra were analyzed using the Sequest algorithm[63] to match peptide sequences from the Uniprot database[40] supplemented by signatures for green fluorescent protein (negative control), the FLAG-HA-tag, and common contaminants. Identified peptides and proteins were further filtered using the target-decoy method[64] to control FDR. High confidence protein interactions were identified using the ComPASS algorithm[65,66] on merged technical duplicates, followed by ComPASS-Plus analysis[62].

To examine per-system enrichment for the new AP-MS data in MuSIC, the assignment of proteins to systems was shuffled while keeping the overall hierarchy structure and number of proteins per system the same, resulting in 1,000 random hierarchies. For each system, we calculated the empirical p-value[67] for the number of new AP-MS interactions among the "system-specific" protein pairs, defined as protein pairs in that system but not in children of that system. P-values were corrected using Benjamini-Hochberg multiple test correction to obtain an FDR for per-system enrichment (**Fig. 4a**).

Statistical analyses

All statistical tests were performed using SciPy[68] with Benjamini-Hochberg multiple test correction as appropriate. Statistics involving comparison between two groups of data to assess differences in data distributions were calculated using either a Mann-Whitney U test, when samples were not paired (**Figs. 2-4**), or a Wilcoxon signed-rank test for paired samples (**Fig. 1e**). Statistics for assessing the enrichment of proteins or protein pairs were calculated using hypergeometric tests (**Extended Data Fig. 3**) unless stated otherwise.

Data Availability

A web portal is available at http://nrnb.org/music with links to all major resources used for this study. These include the MuSIC systems map; the IF (HPA) and AP-MS data (BioPlex 2.0) on which the map is based; and data for the AP-MS pulldown experiments performed as follow-up. The follow-up AP-MS data have also been included as part of the larger compendium of protein interactions to be included in the next version of the BioPlex resource (BioPlex 3.0, *manuscript in review*).

Code Availability

Links to code for performing feature embedding and multi-scale community detection are available at http://nrnb.org/music.

REFERENCES

1. Harold, F. M. Molecules into cells: specifying spatial architecture. *Microbiol. Mol. Biol. Rev.* **69**, 544–564 (2005).
2. Kirschner, M. & Mitchison, T. Beyond self-assembly: from microtubules to morphogenesis. *Cell* **45**, 329–342 (1986).
3. Mori, H. & Cardiff, R. D. Methods of Immunohistochemistry and Immunofluorescence: Converting Invisible to Visible. *Methods Mol. Biol.* **1458**, 1–12 (2016).
4. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
5. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).
6. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
7. Stadler, C. *et al.* Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods* **10**, 315–323 (2013).
8. Chen, X. *et al.* Visualizing RNA dynamics in live cells with bright and stable fluorescent RNAs. *Nat. Biotechnol.* **37**, 1287–1293 (2019).
9. Colwill, K., Renewable Protein Binder Working Group & Gräslund, S. A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* **8**, 551–558 (2011).

10. Cristea, I. M., Williams, R., Chait, B. T. & Rout, M. P. Fluorescent proteins as proteomic probes. *Mol. Cell. Proteomics* **4**, 1933–1941 (2005).

11. Lee, C.-M., Adamchek, C., Feke, A., Nusinow, D. A. & Gendron, J. M. Mapping Protein-Protein Interactions Using Affinity Purification and Mass Spectrometry. *Methods Mol. Biol.* **1610**, 231–249 (2017).

12. Varnaitė, R. & MacNeill, S. A. Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics* **16**, 2503–2518 (2016).

13. Gingras, A.-C., Abe, K. T. & Raught, B. Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Curr. Opin. Chem. Biol.* **48**, 44–54 (2019).

14. Youn, J.-Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol. Cell* **69**, 517–532.e11 (2018).

15. Rhee, H.-W. *et al.* Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013).

16. Kalocsay, M. APEX Peroxidase-Catalyzed Proximity Labeling and Multiplexed Quantitative Proteomics. *Methods Mol. Biol.* **2008**, 41–55 (2019).

17. Trisciuoglio, D. *et al.* Affinity purification-mass spectrometry analysis of bcl-2 interactome identified SLIRP as a novel interacting protein. *Cell Death Dis.* **7**, e2090 (2016).

18. Ouyang, W. *et al.* Analysis of the Human Protein Atlas Image Classification competition. *Nat. Methods* **16**, 1254–1261 (2019).

19. Sullivan, D. P. *et al.* Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).

20. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**,

855–864 (2016).

21. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).

22. Weigert, M. *et al.* Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nat. Methods* **15**, 1090–1097 (2018).

23. Cho, H., Berger, B. & Peng, J. Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks. *Cell Syst* **7**, 185–191.e4 (2018).

24. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

25. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

26. Fortunato, S. Community detection in graphs. *arXiv [physics.soc-ph]* (2009).

27. Kramer, M., Dutkowski, J., Yu, M., Bafna, V. & Ideker, T. Inferring gene ontologies from pairwise similarity data. *Bioinformatics* **30**, i34–42 (2014).

28. Go, C. D. *et al.* A proximity biotinylation map of a human cell. *bioRxiv* 796391 (2019) doi:10.1101/796391.

29. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

30. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

31. Wahl, M. C. & Lührmann, R. SnapShot: Spliceosome Dynamics I. *Cell* **161**, 1474–e1 (2015).

32. Gubitz, A. K., Feng, W. & Dreyfuss, G. The SMN complex. *Exp. Cell Res.* **296**, 51–56 (2004).

33. Blencowe, B. J., Issner, R., Nickerson, J. A. & Sharp, P. A. A coactivator of pre-mRNA splicing. *Genes Dev.* **12**, 996–1009 (1998).

34. Yoshikatsu, Y. *et al.* NVL2, a nucleolar AAA-ATPase, is associated with the nuclear exosome and is involved in pre-rRNA processing. *Biochem. Biophys. Res. Commun.* **464**, 780–786 (2015).

35. Chaudhuri, S. *et al.* Human ribosomal protein L13a is dispensable for canonical ribosome function but indispensable for efficient rRNA methylation. *RNA* **13**, 2224–2237 (2007).

36. Tafforeau, L. *et al.* The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of Pre-rRNA processing factors. *Mol. Cell* **51**, 539–551 (2013).

37. Eppens, N. A. *et al.* Deletions in the S1 domain of Rrp5p cause processing at a novel site in ITS1 of yeast pre-rRNA that depends on Rex4p. *Nucleic Acids Res.* **30**, 4222–4231 (2002).

38. De Silva, D., Tu, Y.-T., Amunts, A., Fontanesi, F. & Barrientos, A. Mitochondrial ribosome assembly in health and disease. *Cell Cycle* **14**, 2226–2250 (2015).

39. Piunti, A. *et al.* CATACOMB: An endogenous inducible gene that antagonizes H3K27 methylation activity of Polycomb repressive complex 2 via an H3K27M-like mechanism. *Sci Adv* **5**, eaax2887 (2019).

40. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

41. Pavan Kumar, P. *et al.* Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. *Mol. Cell* **22**, 231–243 (2006).

42. McCracken, S. *et al.* Proteomic analysis of SRm160-containing complexes reveals a conserved association with cohesin. *J. Biol. Chem.* **280**, 42227–42236 (2005).

43. De Wolf, V. *et al.* A complex Xp11.22 deletion in a patient with syndromic autism: exploration of FAM120C as a positional candidate gene for autism. *Am. J. Med. Genet. A* **164A**, 3035–3041 (2014).

44. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).

45. Alberts, B. *et al. The Compartmentalization of Cells*. (Garland Science, 2002).

46. Yong, J., Pellizzoni, L. & Dreyfuss, G. Sequence-specific interaction of U1 snRNA with the SMN complex. *EMBO J.* **21**, 1188–1196 (2002).

47. Fouquet, C. *et al.* Improving axial resolution in confocal microscopy with new high refractive index mounting media. *PLoS One* **10**, e0121096 (2015).

48. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *arXiv [cs.CV]* (2016).

49. Boland, M. V. & Murphy, R. F. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **17**, 1213–1223 (2001).

50. Coelho, L. P., Peng, T. & Murphy, R. F. Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics* **26**, i7–12 (2010).

51. Lu, A. X., Kraus, O. Z., Cooper, S. & Moses, A. M. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLoS Comput. Biol.* **15**, e1007348 (2019).

52. Huang, J. K. *et al.* Systematic Evaluation of Molecular Networks for Discovery of Disease

Genes. *Cell Syst* **6**, 484–495.e5 (2018).

53. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
54. Paladugu, S. R., Zhao, S., Ray, A. & Raval, A. Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics* **9**, 426 (2008).
55. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *arXiv [cmp-lg]* (1995).
56. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
57. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
58. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
59. Efron, B. *The Jackknife, the Bootstrap and other resampling plans*. (1982).
60. Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
61. Huttlin, E. L. *et al.* Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *bioRxiv* 2020.01.19.905109 (2020) doi:10.1101/2020.01.19.905109.
62. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
63. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
64. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-

scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

65. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009).

66. Behrends, C., Sowa, M. E., Gygi, S. P. & Harper, J. W. Network organization of the human autophagy system. *Nature* **466**, 68–76 (2010).

67. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.* **71**, 439–441 (2002).

68. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

FIGURES

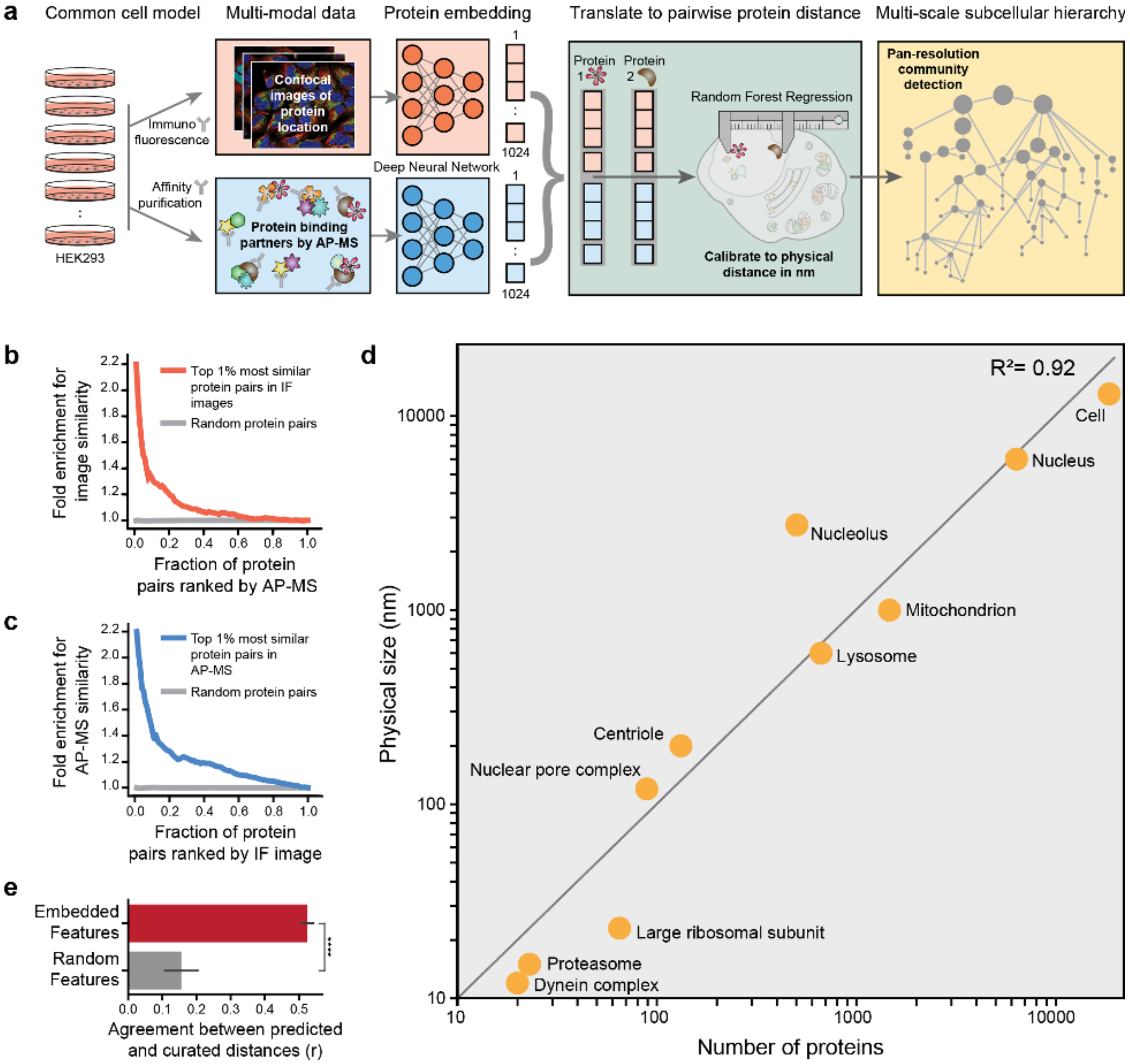


Fig. 1 | Fusing protein distances from immunofluorescence and affinity purification. a, Workflow. **b, c,** Protein pairs ranked by similarity in AP-MS data enrich for the most similar protein pairs in IF data (**b**), and vice versa (**c**). **d,** Calibrating size of subcellular components against the number of proteins assigned to the corresponding GO terms across four logs of physical scale. **e,** Performance of data-driven model in recovering protein-protein distances estimated using GO (red, Pearson *r*). Error bars show standard deviation in cross validation. Equivalent calculation for random feature sets (gray). ****, $p < 0.0001$.

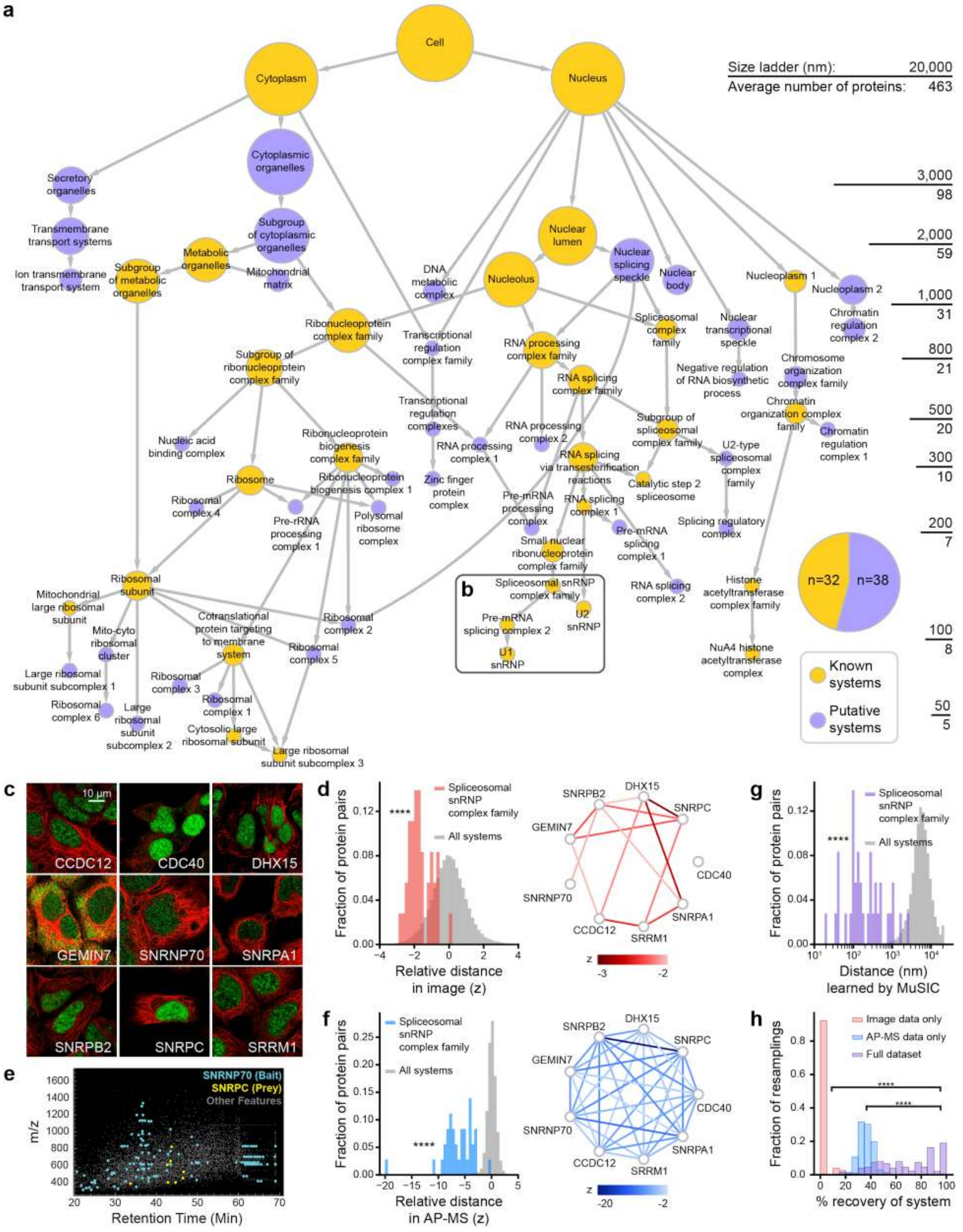

a
Cell
Cytoplasm
Nucleus
Size ladder (nm): 20,000
Average number of proteins: 463
3,000
98
2,000
59
1,000
31
800
21
500
20
300
10
200
7
100
8
50
5
Cytoplasmic organelles
Secretory organelles
Transmembrane transport systems
Ion transmembrane transport system
Subgroup of metabolic organelles
Metabolic organelles
Subgroup of cytoplasmic organelles
Mitochondrial matrix
DNA metabolic complex
Nuclear lumen
Nucleolus
Nuclear splicing speckle
Nuclear body
Nucleoplasm 1
Nucleoplasm 2
Chromatin regulation complex 2
Ribonucleoprotein complex family
Transcriptional regulation complex family
RNA processing complex family
Spliceosomal complex family
Nuclear transcriptional speckle
Negative regulation of RNA biosynthetic process
Chromosome organization complex family
Subgroup of ribonucleoprotein complex family
RNA splicing complex family
Chromatin organization complex family
Transcriptional regulation complexes
RNA processing complex 2
Subgroup of spliceosomal complex family
Chromatin regulation complex 1
Nucleic acid binding complex
Ribonucleoprotein biogenesis complex family
RNA processing complex 1
RNA splicing via transesterification reactions
U2-type spliceosomal complex family
Ribosome
Ribonucleoprotein biogenesis complex 1
Zinc finger protein complex
Catalytic step 2 spliceosome
Ribosomal complex 4
Pre-rRNA processing complex 1
Polysomal ribosome complex
Pre-mRNA processing complex
RNA splicing complex 1
Pre-mRNA splicing complex 1
Splicing regulatory complex
Small nuclear ribonucleoprotein complex family
b
Spliceosomal snRNP complex family
U2 snRNP
Pre-mRNA splicing complex 2
U1 snRNP
RNA splicing complex 2
Histone acetyltransferase complex family
NuA4 histone acetyltransferase complex
n=32
n=38
Known systems
Putative systems
Mitochondrial large ribosomal subunit
Ribosomal subunit
Mito-cyto ribosomal cluster
Cotranslational protein targeting to membrane system
Ribosomal complex 2
Ribosomal complex 5
Large ribosomal subunit subcomplex 1
Ribosomal complex 3
Ribosomal complex 6
Large ribosomal subunit subcomplex 2
Ribosomal complex 1
Cytosolic large ribosomal subunit
Large ribosomal subunit subcomplex 3
c
10 µm
CCDC12
CDC40
DHX15
GEMIN7
SNRNP70
SNRPA1
SNRPB2
SNRPC
SRRM1
d
Fraction of protein pairs
Spliceosomal snRNP complex family
All systems
****
0.12
0.08
0.04
0.00
−4
−2
0
2
4
Relative distance in image (z)
SNRPB2
DHX15
SNRPC
GEMIN7
CDC40
SNRNP70
SNRPA1
CCDC12
SRRM1
z
-3
-2
g
Fraction of protein pairs
Spliceosomal snRNP complex family
All systems
****
0.12
0.08
0.04
0.00
10^1
10^2
10^3
10^4
Distance (nm) learned by MuSIC
e
SNRNP70 (Bait)
SNRPC (Prey)
Other Features
m/z
1600
1400
1200
1000
800
600
400
20
30
40
50
60
70
Retention Time (Min)
f
Fraction of protein pairs
Spliceosomal snRNP complex family
All systems
****
0.25
0.20
0.15
0.10
0.05
0.00
−20
−15
−10
−5
0
5
Relative distance in AP-MS (z)
SNRPB2
DHX15
SNRPC
GEMIN7
CDC40
SNRNP70
SNRPA1
CCDC12
SRRM1
z
-20
-2
h
Fraction of resamplings
Image data only
AP-MS data only
Full dataset
****
****
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
% recovery of system

Fig. 2 | The Multi-scale Integrated Cell. a, MuSIC hierarchy, with nodes representing systems and arrows indicating containment of one system by another. Pie chart shows number of known (gold) or putative novel (purple) systems. **b,** Position of spliceosomal snRNP complex family. **c,** Immunostained spliceosomal snRNP complex proteins (green) or cytoskeleton (red). **d**, Corresponding IF protein-protein distances shown as z-scores (red), calibrated to all distances in IF data (gray). Results also shown as a network, with links for $z \leq -2$, intensity showing magnitude of z. **e,** MS1 features identified during LC-MS analysis of a selected pull-down. Features corresponding to SNRNP70 (bait, cyan) and SNRPC (prey, yellow) are highlighted. **f,** Histogram and network as in (**d**), showing AP-MS rather than IF data. **g,** Integrated distances learned by MuSIC for spliceosomal snRNP complex proteins. **h,** Robustness analysis showing percent recovery of system when perturbing the input data; results accumulated over 300 resamplings (**Methods**). ****, $p < 0.0001$.

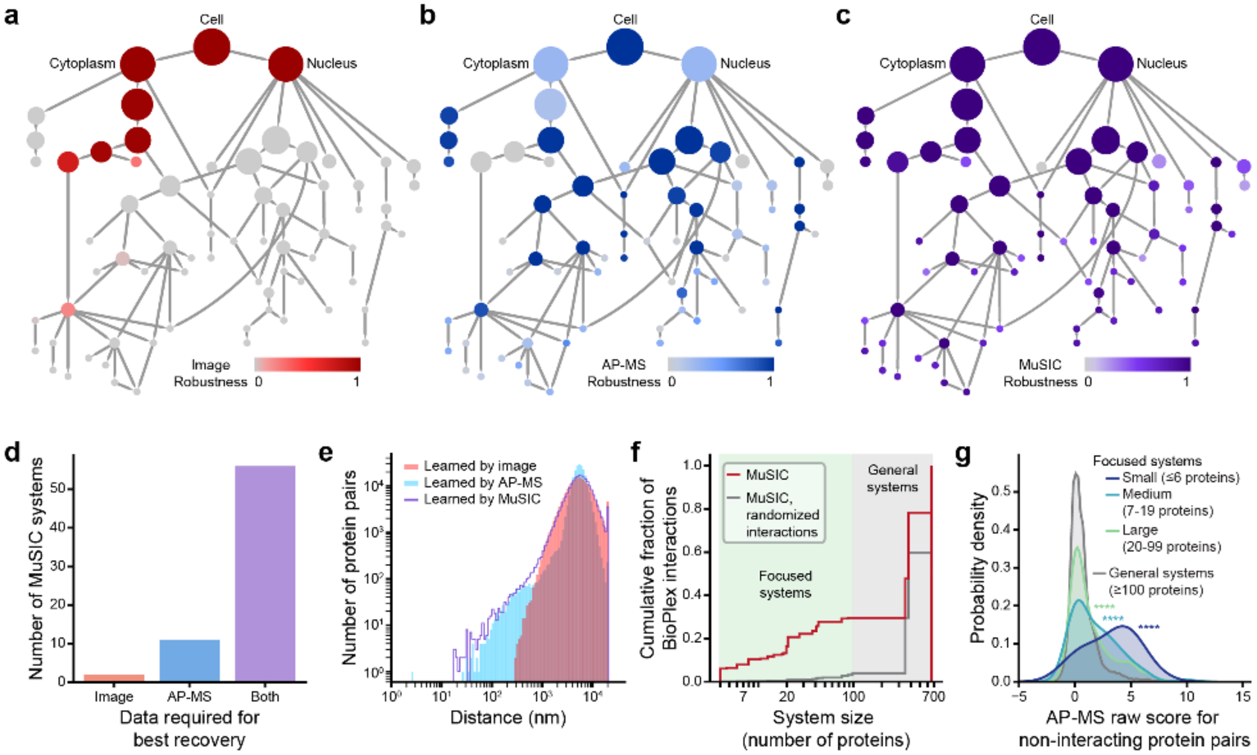


Fig. 3 | Different data, different scales of information. a-c, MuSIC hierarchy colored with system robustness using IF only (**a**), AP-MS only (**b**), or integrated data (full MuSIC, **c**). **d**, Number of systems for which the highest robustness was obtained with IF, AP-MS, or both data types (root system excluded). **e**, Distribution of protein-protein distances learned using only IF (red), only AP-MS (blue), or integrated analysis (purple). **f**, Cumulative fraction of BioPlex protein interactions within MuSIC systems (red) versus random protein pairs (gray, 1000 randomizations). **g**, Distribution of AP-MS raw z-scores for protein pairs not labeled as interacting by BioPlex. P-values calculated against general systems ≥100 proteins. ****: $p < 0.0001$.

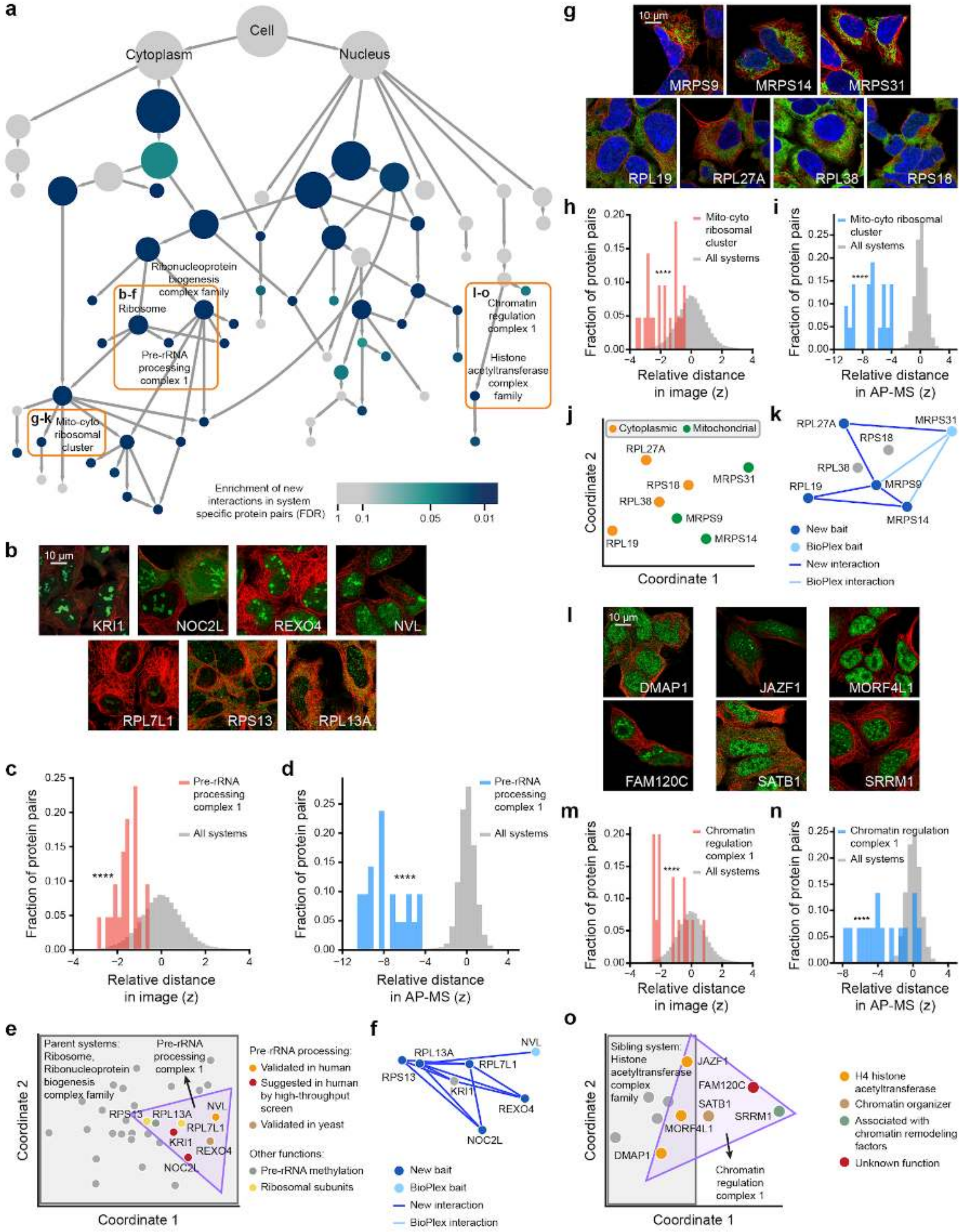
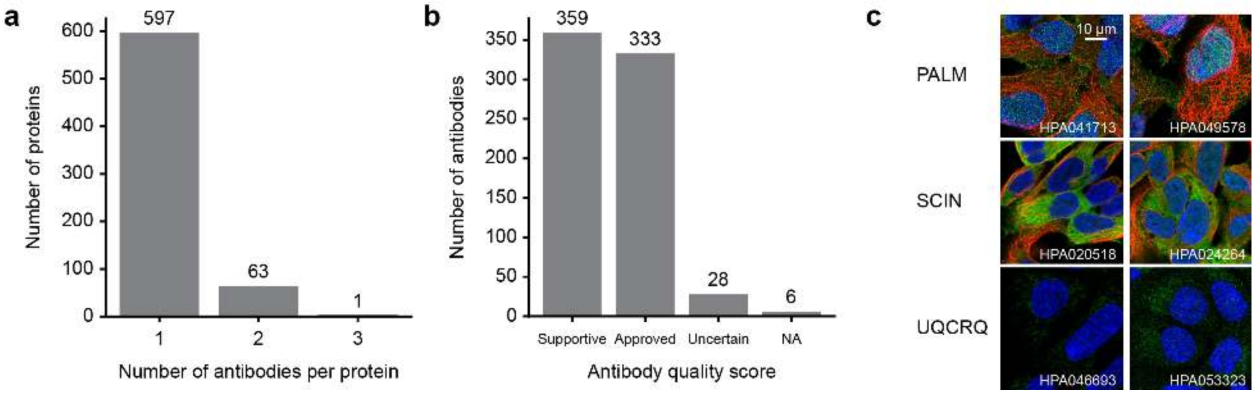
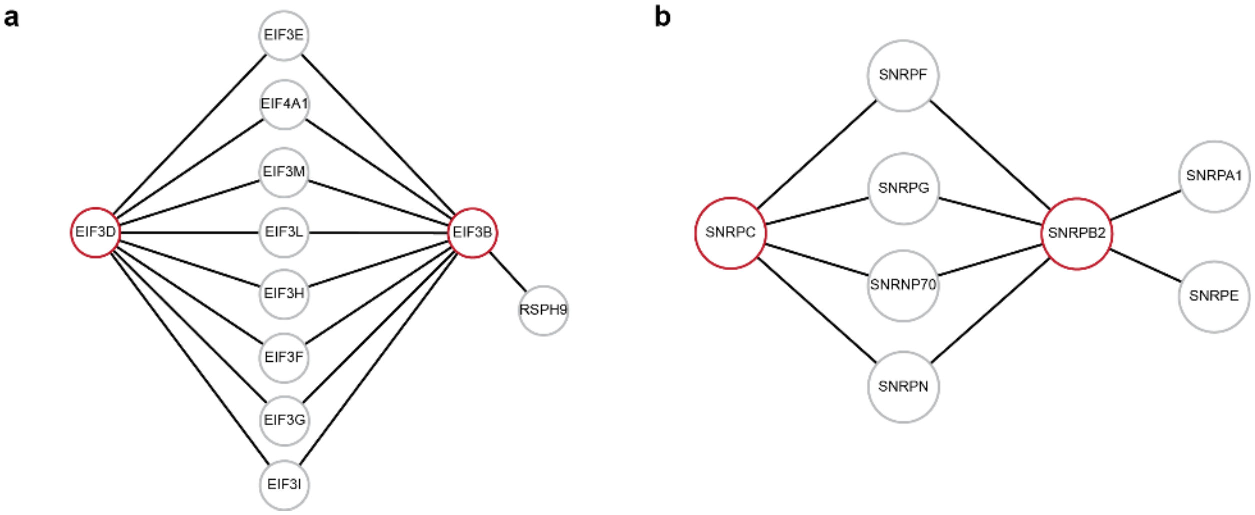

a
Cell
Cytoplasm
Nucleus
Ribonucleoprotein biogenesis complex family
b-f
Ribosome
Pre-rRNA processing complex 1
l-o
Chromatin regulation complex 1
Histone acetyltransferase complex family
g-k
Mito-cyto ribosomal cluster
Enrichment of new interactions in system specific protein pairs (FDR)
1
0.1
0.05
0.01
b
10 µm
KRI1
NOC2L
REXO4
NVL
RPL7L1
RPS13
RPL13A
c
Fraction of protein pairs
0.25
0.20
0.15
0.10
0.05
0.00
****
Pre-rRNA processing complex 1
All systems
Relative distance in image (z)
−4
−2
0
2
4
d
Fraction of protein pairs
Pre-rRNA processing complex 1
All systems
****
Relative distance in AP-MS (z)
−12
−8
−4
0
4
e
Coordinate 2
Coordinate 1
Parent systems: Ribosome, Ribonucleoprotein biogenesis complex family
Pre-rRNA processing complex 1
RPS13
RPL13A
NVL
RPL7L1
KRI1
REXO4
NOC2L
Pre-rRNA processing:
Validated in human
Suggested in human by high-throughput screen
Validated in yeast
Other functions:
Pre-rRNA methylation
Ribosomal subunits
f
RPL13A
NVL
RPS13
RPL7L1
KRI1
REXO4
NOC2L
New bait
BioPlex bait
New interaction
BioPlex interaction
g
10 µm
MRPS9
MRPS14
MRPS31
RPL19
RPL27A
RPL38
RPS18
h
Fraction of protein pairs
0.20
0.15
0.10
0.05
0.00
****
Mito-cyto ribosomal cluster
All systems
Relative distance in image (z)
−4
−2
0
2
4
i
Fraction of protein pairs
0.25
0.20
0.15
0.10
0.05
0.00
****
Mito-cyto ribosomal cluster
All systems
Relative distance in AP-MS (z)
−12
−8
−4
0
4
j
Cytoplasmic
Mitochondrial
Coordinate 2
Coordinate 1
RPL27A
RPS18
MRPS31
RPL38
MRPS9
RPL19
MRPS14
k
RPL27A
MRPS31
RPS18
RPL38
MRPS9
RPL19
MRPS14
New bait
BioPlex bait
New interaction
BioPlex interaction
l
10 µm
DMAP1
JAZF1
MORF4L1
FAM120C
SATB1
SRRM1
m
Fraction of protein pairs
0.20
0.15
0.10
0.05
0.00
****
Chromatin regulation complex 1
All systems
Relative distance in image (z)
−4
−2
0
2
4
n
Fraction of protein pairs
0.25
0.20
0.15
0.10
0.05
0.00
****
Chromatin regulation complex 1
All systems
Relative distance in AP-MS (z)
−8
−4
0
4
o
Coordinate 2
Coordinate 1
Sibling system: Histone acetyltransferase complex family
JAZF1
FAM120C
SATB1
SRRM1
MORF4L1
DMAP1
Chromatin regulation complex 1
H4 histone acetyltransferase
Chromatin organizer
Associated with chromatin remodeling factors
Unknown function

Fig. 4 | Validation of MuSIC systems using new AP-MS pull-downs. a, MuSIC hierarchy with system color showing enrichment for new AP-MS interactions at indicated false discovery rate (FDR, blue gradient, **Methods**). Orange boxes indicate systems highlighted in text. **b,** IF images showing similar nucleolar signals among proteins in the pre-rRNA processing complex 1. **c, d,** Corresponding distributions of protein-protein distance z-scores for IF (**c**, red) or AP-MS (**d**, blue), calibrated to all such distances, respectively (gray). **e,** Spring embedding of proteins in parent systems (grey box) of pre-rRNA processing complex 1 (purple triangle) based on pairwise distances learned by MuSIC. Proteins colored according to known functions. **f**, Validated physical interactions in pre-rRNA processing complex 1. **g-k**, Similar analyses for mito-cyto ribosomal cluster. **l-o,** Similar analyses for chromatin regulation complex 1. Image colors represent immunostained protein (green), cytoskeleton (red), or nucleus (blue). ****, p < 0.0001.

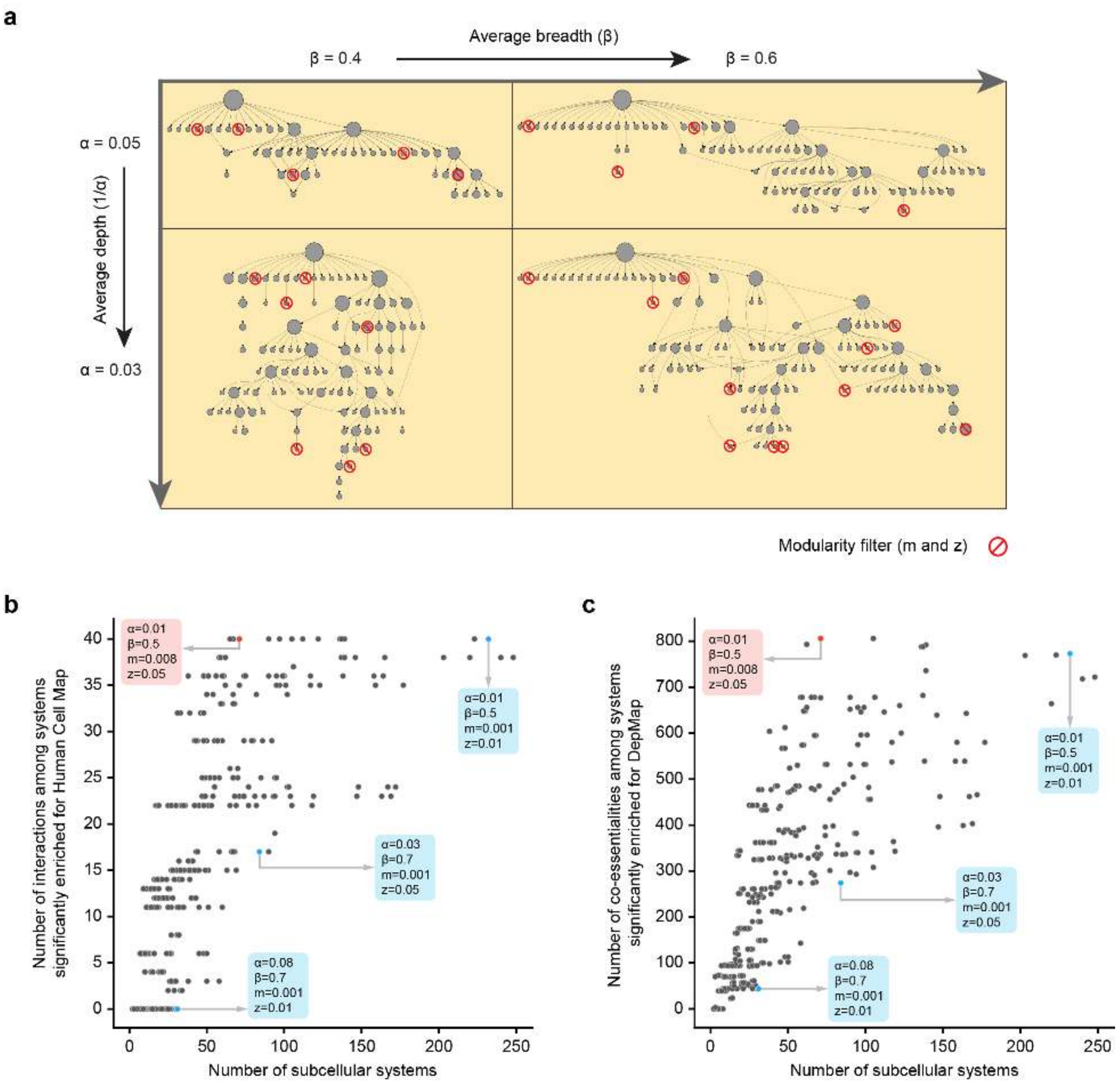
EXTENDED DATA



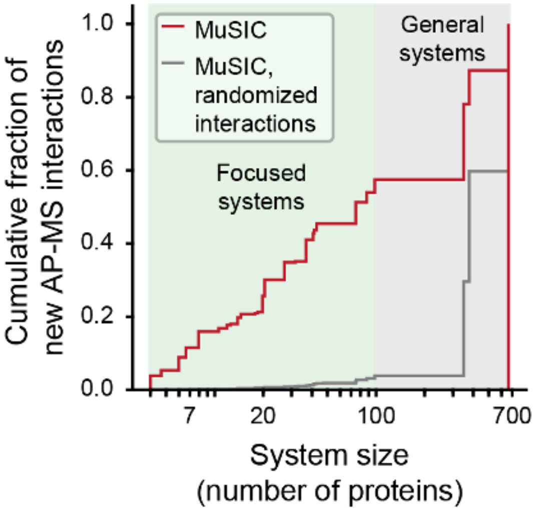
Extended Data Fig. 1 | Antibody count and quality. a, Histogram showing the distribution in number of antibodies per protein over the proteins included in this study. **b,** Histogram showing the distribution in antibody quality scores over the antibodies used in this study. **c,** IF images for three pairs of antibodies targeting the same protein, respectively. Colors represent immunostained protein (green), cytoskeleton (red), or nucleus (blue). The images show high reproducibility even when different antibodies for the same target protein are used.



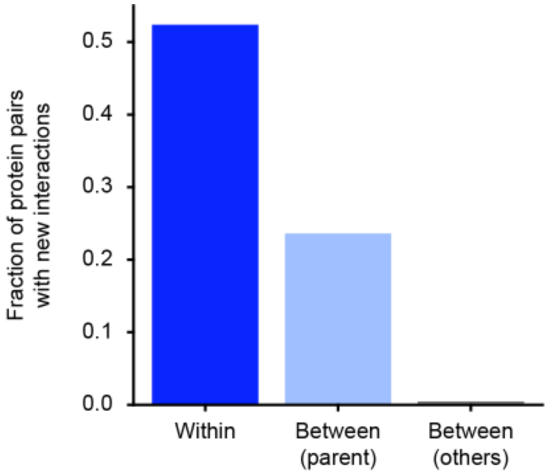
Extended Data Fig. 2 | Non-interacting protein pairs nonetheless assigned high similarity in the AP-MS embedding. The node2vec network embedding[20] can determine that two proteins have high similarity even if these proteins have not yet been recorded to have direct physical interaction. **a,** Network showing all proteins that physically interact with EIF3D and EIF3B (red) in BioPlex 2.0. EIF3D and EIF3B do not physically interact, but the cosine similarity of their embedded node2vec features, which consider the first interaction neighborhood (shown) and beyond (not shown), is 0.95. The maximum cosine similarity is 1. **b,** Network showing all proteins that physically interact with SNRPC and SNRPB2 (red) in BioPlex 2.0. SNRPC and SNRPB2 do not physically interact, but the cosine similarity of their embedded features is 0.93. In many cases where two proteins have high node2vec similarity, but do not directly interact, we found that neither protein had yet been tagged as bait for an affinity purification experiment. As the direct interaction has not yet been interrogated, the node2vec embedding serves to fill gaps in existing data. None of the proteins highlighted in red were tagged as bait proteins in BioPlex 2.0.



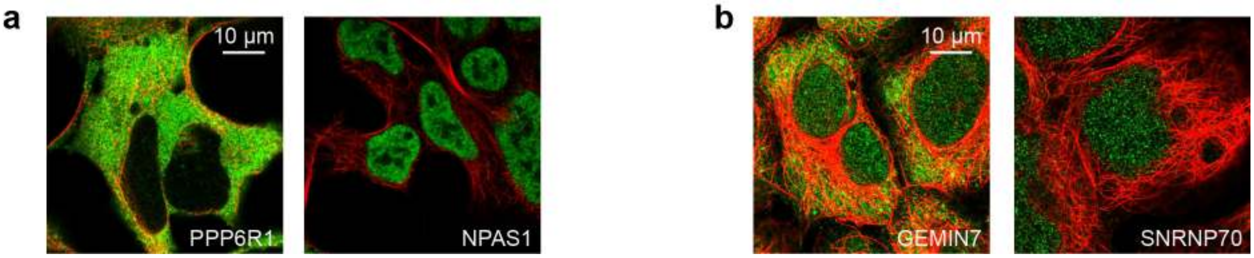
Extended Data Fig. 3 | Selection of parameters for community detection. a, CliXO community detection has several parameters (breadth, x-axis; depth, y-axis; minimum modularity, red circle backslash) that affect the sensitivity with which communities are identified and thus the size of the hierarchy. **b, c,** Dotplots in which each dot is a hierarchy generated with a particular set of parameters. The selection for MuSIC is highlighted in red. This selection was among several that were optimal, based on enrichment for protein-protein interactions in Human Cell Map (**b**) and co-essentialities from DepMap (**c**). Examples of other parameters are shown in blue.



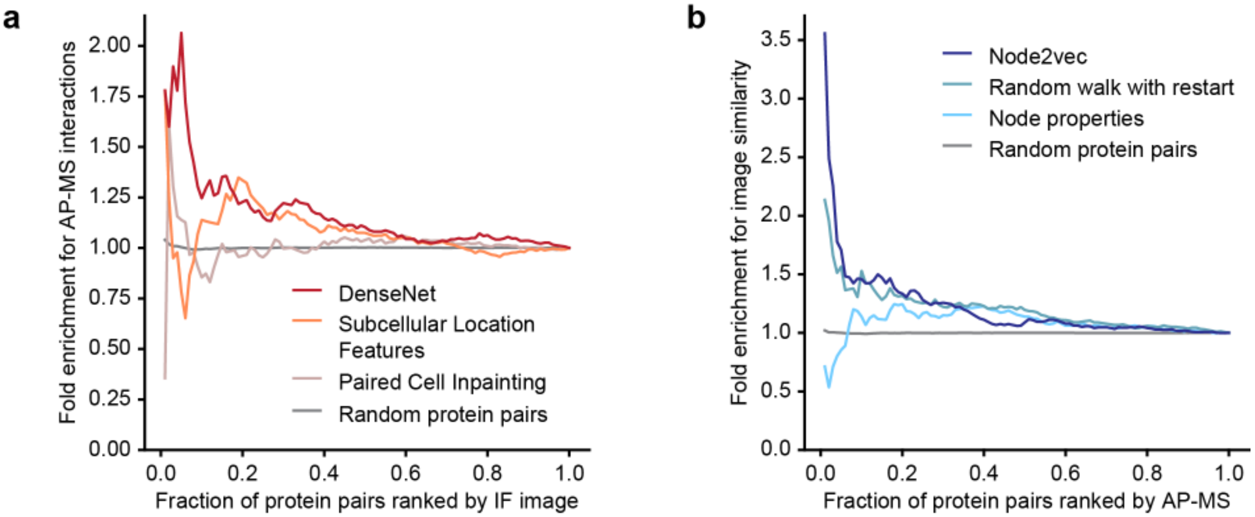
Extended Data Fig. 4 | Cumulative fraction of new AP-MS interactions in MuSIC. Related to Fig. 3f.



Extended Data Fig. 5 | Classification of newly identified physical interactions for proteins in the pre-rRNA processing complex 1. The specific recovery of interactions within that same system is shown (dark blue bar), in comparison to interactions with other proteins organized under the same parent systems in the MuSIC hierarchy (ribosome, ribonucleoprotein biogenesis complex family, light blue bar) or with proteins organized elsewhere in MuSIC (gray bar).



Extended Data Fig. 6 | Examples of proteins with strong AP-MS protein interactions that have very different IF localization patterns. a, PPP6R1 and NPAS1. **b,** GEMIN7 and SNRNP70. Colors represent immunostained protein (green) and cytoskeleton (red).



Extended Data Fig. 7 | Comparison of different data embedding methods. a, Ability of different image embedding methods (colored curves) to generate image-image similarities (cosine similarity) in agreement with the set of all 281 protein-protein interactions in BioPlex. **b,** Ability of different AP-MS embedding methods to generate protein-protein similarities (cosine similarity) in agreement with protein pairwise similarities computed from HPA images (considering the top 281 protein pairs by image cosine similarity). Embedding methods are described in **Methods**.