

# Unifying proteomic technologies with ProteinProjector

Leah V Schaffer<sup>1,\*</sup>, Mayank Jain<sup>1</sup>,  
Rami Nasser<sup>2</sup>, Roded Sharan<sup>2</sup>,  
Trey Ideker<sup>1,3,4,\*</sup>

<sup>1</sup> Department of Medicine, University of California San Diego, La Jolla, CA, USA

<sup>2</sup> Blavatnik School of Computer Science and AI, Tel Aviv University, Tel Aviv 69978, Israel

<sup>3</sup> Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

<sup>4</sup> Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

\*Correspondence:

[leahvschaffer@gmail.com](mailto:leahvschaffer@gmail.com) (L.V.S.),  
[tideker@health.ucsd.edu](mailto:tideker@health.ucsd.edu) (T.I.)

## Abstract

Proteomics has developed many approaches to inform the subcellular organization of proteins, each with differing coverage and sensitivity to distinct scales. Here we develop a self-

## Introduction

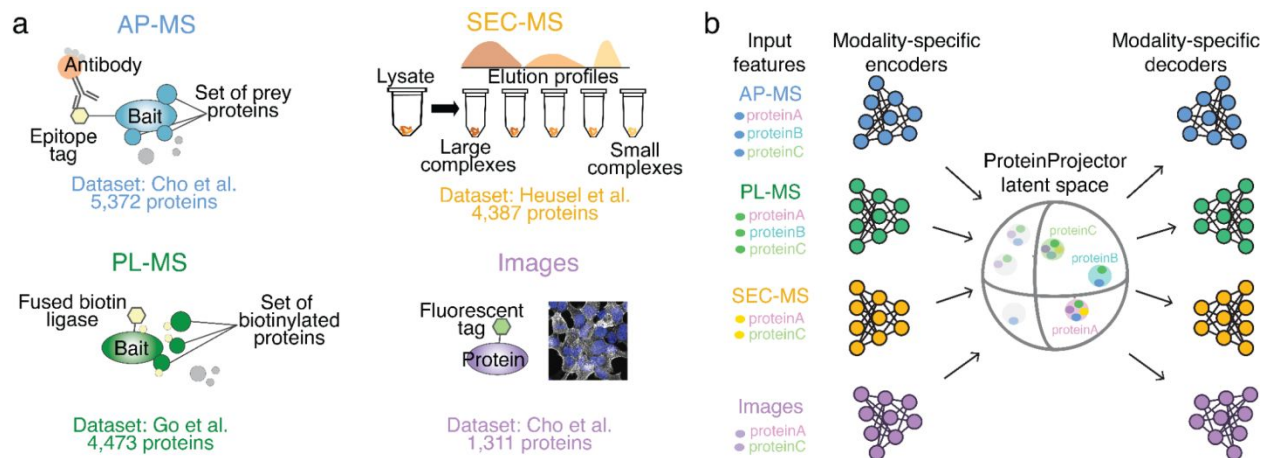
The last few decades have witnessed enormous advances in proteomic technologies for charting the protein assemblies of human cells (**Figure 1a**) (Richards *et al.*, 2021; Thul and Lindskog, 2018; Wilhelm *et al.*, 2014; Luck *et al.*, 2020; Skinnider *et al.*, 2021; Mulvey *et al.*, 2017). For example,

supervised deep learning framework, ProteinProjector, that flexibly integrates all available data for a protein from any number of modalities, resulting in a unified map of protein position. As initial proof-of-concept we integrate four proteome-wide characterizations of HEK293 human embryonic kidney cells, including protein affinity purification, proximity ligation, and size-exclusion-chromatography mass spectrometry (AP-MS, PL-MS, SEC-MS), as well as protein fluorescent imaging. Map coverage and accuracy grow substantially as new data modes are added, with maximal recovery of known complexes observed when using all four proteomic datasets. We find that ProteinProjector outperforms individual modalities and other integration methods in recovery of orthogonal functional and physical associations not used during training. ProteinProjector is available as part of the Cell Mapping Toolkit at [https://github.com/idekerlab/cellmaps\\_coembedding](https://github.com/idekerlab/cellmaps_coembedding), providing a foundation for integration of diverse modalities that characterize subcellular structure.

**Key words:** proteomics, deep learning, representation learning, subcellular components, protein interactions affinity purification mass spectrometry (AP-MS) isolates a tagged protein of interest from whole-protein extracts, allowing for the identification of neighboring proteins with biophysical interactions (Gordon *et al.*, 2020; Huttlin *et al.*, 2015, 2021); proximity labeling mass spectrometry (PL-MS) employs an enzyme fused to a protein of interest to label nearby proteins

covalently (Kim *et al.*, 2014; Go *et al.*, 2021); and size exclusion chromatography mass spectrometry (SEC-MS) identifies groups of proteins with similar elution profiles during chromatography (Fossati *et al.*, 2023; Havugimana *et al.*, 2012; Bludau *et al.*, 2020). Adding to these MS-based approaches, protein fluorescence coupled to confocal microscopy reveals the spatial distribution of a target protein within the cell as well as other proteins that share this distribution (Cho *et al.*, 2022; Thul *et al.*, 2017).

These and numerous other techniques (Johnson *et al.*, 2021; Luck *et al.*, 2020; Geladaki *et al.*, 2019) each reveal complementary aspects of how proteins are organized in cells. Integrating across these multiple approaches could substantially increase proteome coverage and fidelity over what is obtained with any single technique for mapping cell structure, advancing towards the goal of providing a complete view of protein assemblies (Schaffer *et al.*, 2025).



**Figure 1 | ProteinProjector Overview.** a) Overview of data modalities obtained from application of proteomic technologies to HEK293 cells: affinity purification (AP-MS), proximity labeling (PL-MS), size exclusion chromatography (SEC-MS), and fluorescence imaging. b) Schematic diagram of ProteinProjector encoder-decoder neural network architecture designed for multi-modal data integration and interpolation.

In recent years, the field of machine learning has developed a powerful arsenal of approaches for combining multiple types of data collected for a sample (i.e. data modalities) into a general unified representation (i.e. sample embedding) (Radford *et al.*, 18--24 Jul 2021; Girdhar *et al.*, 2023). In the emerging class of approaches known as Foundation models, this representation is learned in a general

task-agnostic manner, without being specifically trained for any particular downstream application. Recently, Foundational models have been applied in biology to create integrated embeddings of single-cell sequencing and imaging data (Bao *et al.*, 2022; Yang *et al.*, 2021) as well as molecular interactions (Forster *et al.*, 2022; Nasser and Sharan, 2023). To apply these concepts to datasets for mapping subcellular organization, we developed

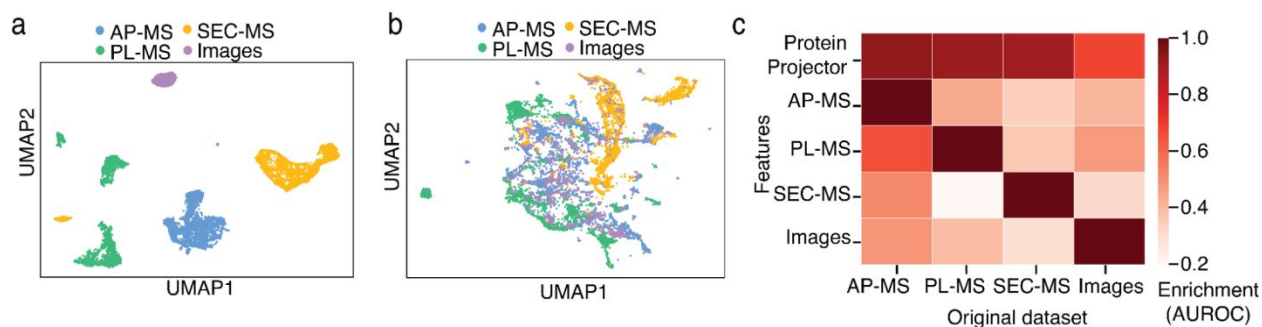
a framework called ProteinProjector, which integrates any number of proteomic datasets to learn a unified protein representation that captures information from each of the original modalities (**Figure 1b**).

## Results

ProteinProjector employs an encoder-decoder neural network architecture, in which a bank of encoders distills features from each of the separate data modalities collected for a protein into a unified embedding, while a corresponding bank of decoders reconstructs the original features from this shared space. This system is trained in such a way as to minimize error between the reconstructed and original datasets (“reconstruction loss”, **Methods**), which encourages accurate data replication. Furthermore, the multiple modalities characterizing a protein are encouraged to occupy embedding coordinates that are similar to one another but distinct from the coordinates of other proteins (“triplet loss”, a type of contrastive learning; **Methods**). While ProteinProjector trains from all available data for each protein, a key feature is its tolerance to missing data (i.e., it does not require

that a protein is covered by every dataset).

As proof-of-concept, we applied this approach to integrate the growing wealth of protein physical association datasets generated in the human embryonic kidney (HEK)-293 cell line, a common model used for *in-vitro* studies of human cell biology. We collected datasets from multiple mass spectrometry techniques including AP-MS (Cho *et al.*, 2022), PL-MS (Go *et al.*, 2021), SEC-MS (Heusel *et al.*, 2019), and protein fluorescent images (Cho *et al.*, 2022) (**Figure 1a**, **Supplementary Table 1**). These four datasets were supplied to ProteinProjector, which used them to learn the unified embedding space (**Methods**). UMAP projection of the embeddings revealed that the modality embeddings are more unified in the latent space after integration with ProteinProjector (**Figure 2a,b**). As needed for later applications, we averaged the separate modality ProteinProjector embeddings to produce a single unified embedding per protein. This embedding is then used as a basis for computing protein-protein pairwise cosine similarities, which we call “protein proximities”, for comparison against other datasets.



**Figure 2 | ProteinProjector representations. a)** UMAP visualization of ProteinProjector embeddings for each protein prior to training, colored by input modality. **b)** UMAP visualization of

ProteinProjector embeddings for each protein after training, colored by input modality. **c)** Agreement (white-to-red color gradient) of each original data type embeddings (columns) to each other (rows) or to the ProteinProjector embedding (top row). Agreement measured by enrichment of most similar protein pairs in one embedding versus another, defined by AUROC (**Methods**).

We first investigated how the ProteinProjector embedding positions protein pairs in comparison to the original data modality embedding (**Methods**), prior to training. The ProteinProjector protein proximities showed high levels of enrichment for the most similar pairs in each of the original modalities (**Figure 2c**), demonstrating how the ProteinProjector embeddings retain protein proximity information from each original dataset.

We found that the ProteinProjector embedding increases coverage of the proteome, containing more proteins than any single modality alone (**Figure 3a**). In particular the ProteinProjector embedding included coordinates for 8,004 proteins, versus ~5,500 proteins for AP-MS (the most complete single modality) and ~1,200 proteins for imaging (the least complete modality). As a result of this expanded coverage, we found that ProteinProjector also markedly improves coverage of protein functions. In particular, ProteinProjector embeddings covered the highest fraction of Gene Ontology terms (GO, all three branches) compared to each individual modality (**Figure 3b, Methods**).

We next examined how well the ProteinProjector embedding similarities position protein pairs with high similarity in orthogonal functional and physical datasets not used in training. These datasets

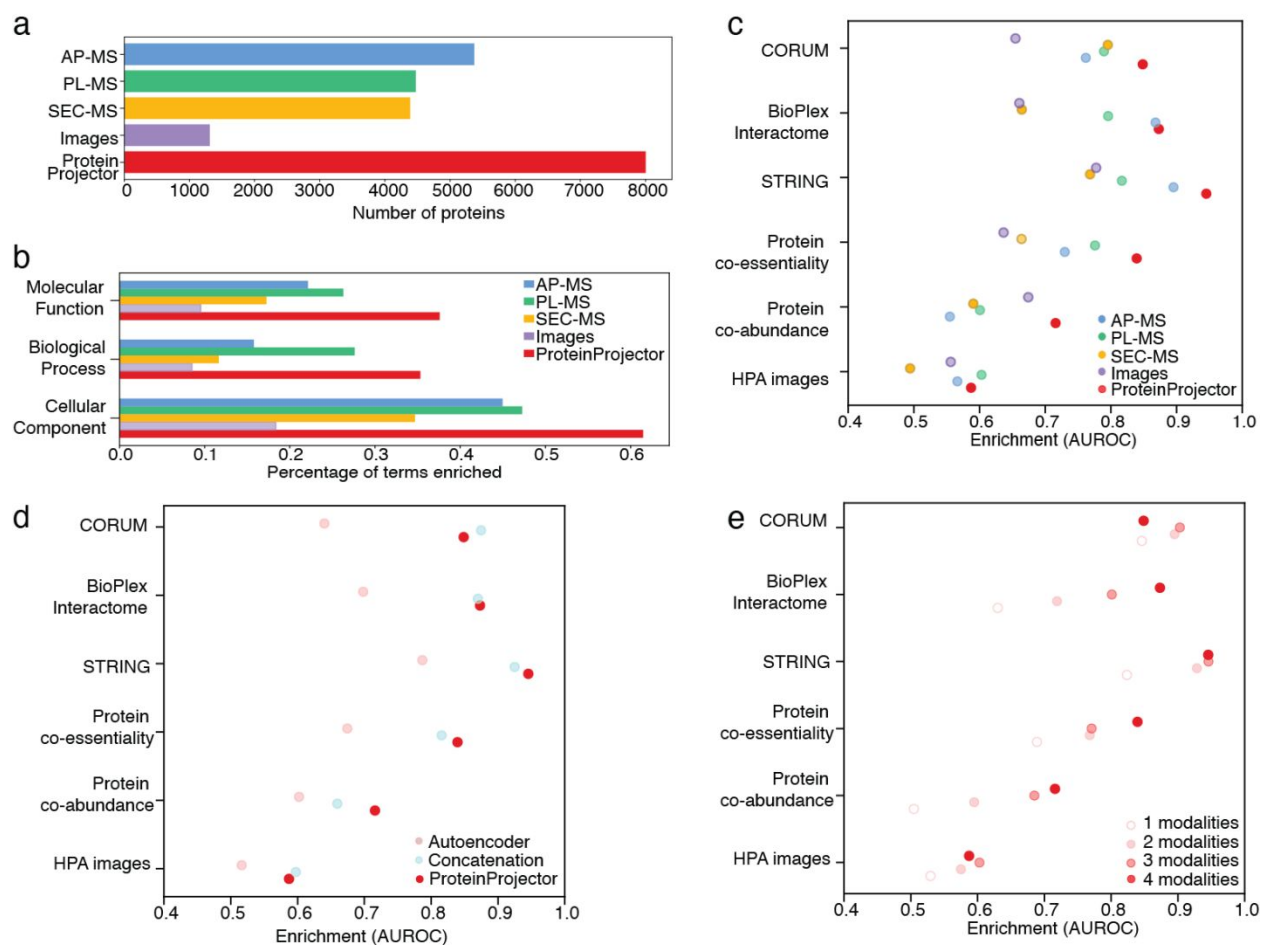
included a protein co-essentiality screen, defined as pairs of proteins with similar transcriptional profiles upon CRISPR perturbations (Tsherniak *et al.*, 2017); protein co-abundance, defined as pairs of proteins with similar abundances across cell types (Gonçalves *et al.*, 2022); interactions in independent AP-MS dataset, the BioPlex interactome in HEK-293 (Huttlin *et al.*, 2015, 2021); pairs of proteins with similar Human Protein Atlas immunofluorescence images (HPA) (Thul *et al.*, 2017); STRING interactions (Szklarczyk *et al.*, 2018); and co-membership in a CORUM complex (Tsitsiridis *et al.*, 2022). Compared to individual datasets, the ProteinProjector embedding markedly improved the enrichment (Area Under the Receiver Operating Characteristic, AUROC, **Methods**) between pairs in functional datasets like protein co-abundance and physical datasets such as pairs in the same CORUM complex (**Methods, Figure 3c**). We observed that for the subset of proteins present in all four modalities (580 proteins), concatenation performed close to, or on par with, ProteinProjector (**Figure 3d**). However, a core strength of ProteinProjector is that it does not require all modalities present to generate an embedding for each protein. Notably, when considering all proteins measured by at least one modality (8,004 proteins, **Methods**), we found that ProteinProjector markedly improves upon concatenation in recovery of all external standards

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

(**Supplementary Fig. 1**). This analysis suggested that the ProteinProjector embeddings better capture physical and functional relationships between proteins than any modality alone. Furthermore, ProteinProjector also tended to outperform other strategies for data integration, including simple concatenation of modality features and a standard autoencoder (**Figure 3d**). We noticed the observed trends held true at varying thresholds of top 5%, 10%, and 20% of pairs for both the original datasets and when using thresholds for orthogonal datasets (Protein co-essentiality and Protein co-abundance; **Supplementary Fig. 2 and Supplementary Fig. 3**).

We analyzed how agreement with the orthogonal datasets varied when comparing ProteinProjector embeddings for proteins covered by all

four data modalities versus proteins covered by only one, two or three (**Figure 3e**). This analysis revealed that in general, proteins present in greater numbers of modalities show more agreement with the functional and physical datasets. Proteins only present in one or two data modalities still tended to demonstrate a positive enrichment for physical associations such as CORUM, but integrating across modalities tended to improve this effect. To further investigate, we analyzed a set of experiments where ProteinProjector was trained using combinations of 2 or 3 modalities. While the ProteinProjector embedding trained with four modalities consistently performs within the top two, other subsets of the four modalities vary in performance depending on the dropped modalities and the external dataset (**Supplementary Fig. 4**).



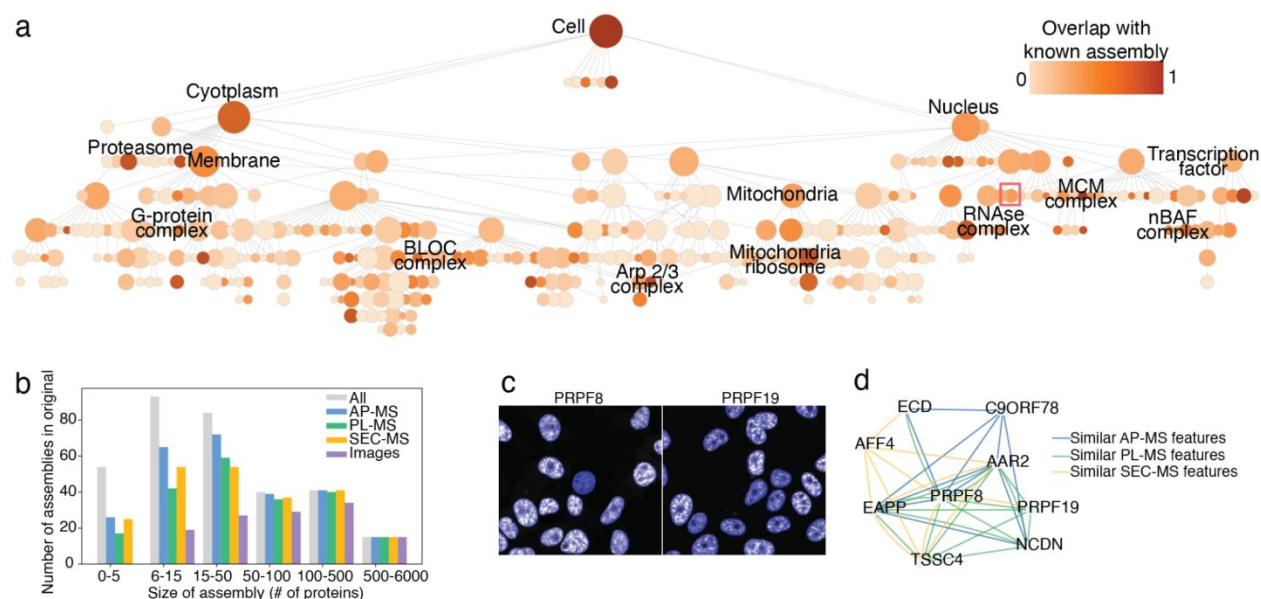
**Figure 3 | Evaluation of ProteinProjector integration** **a**) Number of proteins covered in each original modality vs. ProteinProjector. **b**) Fraction of Gene Ontology terms recovered in similar protein proximities for original modality embeddings and ProteinProjector (**Methods**, <1% FDR). **c**) Degree of enrichment (AUROC, **Methods**) of similar protein proximities (colored points, each data modality individually and combined with ProteinProjector) for orthogonal functional and physical association datasets not used in model training (rows), focused on proteins present in all four original datasets. **d**) Similar to panel (c), comparison of ProteinProjector embeddings to a standard autoencoder integration and to simple concatenation of input features (**Methods**). **e**) Similar to panel (c), but for ProteinProjector embeddings that incorporate increasing numbers of data modalities (colored points).

One downstream application of ProteinProjector is integration of modalities for mapping cell structure in order to robustly identify protein assemblies (Qin *et al.*, 2021). We performed multiscale community detection (Zheng *et al.*, 2021) on the ProteinProjector protein proximities to construct a global, integrated map of the cell with the union of proteins

across all data modalities. This global set of 8,004 proteins was organized into 359 protein assemblies (**Figure 4a**), including 147 with high overlap with a known GO cellular component or CORUM term (**Methods**). The remaining 212 assemblies were designated as putatively novel assemblies. We assessed each protein assembly for presence of similar original data modality features

(Methods), determining which assemblies were driven by different data modalities (Figure 4b, Supplementary Fig. 5). This analysis revealed 279 assemblies supported by AP-MS evidence, 228 by PL-MS evidence, 244 by SEC-MS evidence, and 125 by image evidence. Of these, 244 assemblies (of which 121 were putatively novel) were informed by more than one modality, providing

robust evidence for novel associations. For example, the map revealed a putative nuclear protein assembly involved in RNA splicing with support from multiple original data modalities, including similar images with nuclear localization (Figure 4c) and interactions across the MS datasets (Figure 4d), suggesting these protein associations are robustly recovered across experiments.



**Figure 4 | Cell map of protein assemblies in HEK293 based on ProteinProjector.** **a)** Hierarchy of protein assemblies constructed by performing multi-scale community detection on ProteinProjector embeddings. Each node is colored based on overlap with known subcellular components from GO, CORUM, and HPA (Methods). Red box denotes assembly highlighted in c,d. **b)** Number of assemblies with support from original data modalities (Methods) versus size of assembly in number of proteins. Gray bars denote the total number of assemblies in each size category. **c)** Live fluorescence cell images for proteins in spliceosome-associated complex present in the imaging dataset (PRPF8 and PRPF19). **d)** Spliceosome-associated complex supported by similar features (top 1% pairs by cosine similarity) from original data modalities.

## Discussion

ProteinProjector presents a flexible framework for integrating multiple proteomics data sources into a low dimensional representation of each protein which can be used for downstream analyses such as mapping

subcellular structure. Alternative approaches (e.g. feature concatenation, Supplementary Figure 1) require all modalities to be present for each protein. Additionally, ProteinProjector implements a self-supervised approach which avoids some of the pitfalls of

1 supervised approaches, such as biases  
2 towards well-studied proteins.

3 While an individual proteomics  
4 data modality may perform strongly in  
5 recovery of a particular external  
6 dataset, we find that overall  
7 ProteinProjector provides the best  
8 performance across a range of external  
9 standards (Fig. 3) , including the  
10 recovery of the largest number of  
11 documented subcellular components  
12 (Fig. 3b). We note that dropping out  
13 modalities has variable effects that  
14 depend on the particular modality and  
15 external standard used for evaluation.  
16 The ability to integrate new data  
17 modalities as they become available  
18 will enable further investigations of  
19 how these modalities excel in  
20 characterizing different types of  
21 interactions or classes of proteins.

22 A future avenue for exploration  
23 will be to study cases that present  
24 conflicts among the different  
25 modalities. For example, two proteins  
26 may have very similar  
27 immunofluorescent images but  
28 completely disjoint patterns of protein  
29 interactions in AP-MS data, placing  
30 tension on the ProteinProjector  
31 embedding. Such conflicts raise the  
32 question of whether one of the  
33 modalities is correct and the other is in  
34 error, in which cases such preferences  
35 might be learned during model  
36 training. Alternatively, inter-modality  
37 conflicts could point to different aspects  
38 of protein biology, for example stable  
39 versus time-dependent properties of  
40 the protein or variations in protein  
41 localization across cell types.

42 Deep learning architectures like  
43 ProteinProjector could also be useful  
44 for translating across data modalities.

45 For example, one might wish to use a  
46 relatively rapid proteome profiling  
47 with SEC-MS to predict the protein  
48 interactions that would be expected to  
49 result from lower throughput  
50 techniques such as AP-MS. Future  
51 studies may further explore the  
52 ProteinProjector modeling framework  
53 to determine which aspects of the  
54 multimodal architecture are critical to  
55 its performance, to compare different  
56 architectures, and to assess which  
57 models apply best to specific biological  
58 datasets or problems. ProteinProjector  
59 can also be readily extended to add  
60 even greater numbers of data  
modalities or protein features,  
including direct incorporation of  
protein sequence or structure.

## Methods

### Compilation of HEK293 features

AP-MS interactions generated by the OpenCell project (Cho *et al.*, 2022) were downloaded from <https://opencell.czbiohub.org/>. SEC-MS profiles were generated in a previous study (Heusel *et al.*, 2019) and downloaded from the publication site Supplementary Material. To generate a network, we processed SEC-MS data using the PrInCE software (Stacey *et al.*, 2017) and selected protein pairs with precision >0.75. PL-MS interactions generated in the Human Cell Map project (Go *et al.*, 2021) were downloaded from [humancellmap.org](http://humancellmap.org) (saint-080922.txt) and filtered for high confidence interactions (BFDR < 0.01). For each of these network-based HEK293 datasets, we used node2vec to generate a 1024-dimensional embedding for each protein ([https://github.com/idekerlab/cellmaps\\_ppi\\_embedding](https://github.com/idekerlab/cellmaps_ppi_embedding), p=2, q=1, walk length=80, number of walks=10). The images of protein subcellular distribution were also generated by the OpenCell project (Cho *et al.*, 2022). An embedding for each image was generated using cytocelf (Kobayashi *et al.*, 2022); these 9217-dimensional embeddings were directly downloaded from GitHub (<https://github.com/royerlab/cytocelf>).

### ProteinProjector model architecture

ProteinProjector consists of an encoder and a decoder for each modality. The separate embedding vector inputs ( $x_{m,i}$  for protein  $i$  in modality  $m$ ) are

compressed by modality-specific encoders ( $f_m$ ), yielding 128-dimension vectors  $z_{m,i}$ :

$$z_{m,i} = f_m(x_{m,i})$$

$$f_m = \text{Dropout}(\text{Linear}(\text{ReLU}(\text{Dropout}(\text{Linear}(\text{ReLU}(\text{Linear}(\text{L2Norm}(\quad))))))))$$

‘Dropout’ indicates dropout layers (Srivastava *et al.*, 2014); ‘Linear’ indicates linear transformation layers; L2Norm indicates L2-normalization; ReLU indicates the Rectified Linear Unit function. The dimensionality of the linear layers is calculated based on the dimensions of the embedding vector inputs. The  $z_{m,i}$  embeddings are passed to modality-specific decoders ( $g_m$ ), yielding the reconstructed inputs ( $y_{m,i}$ ):

$$y_{m,i} = g_m(z_{m,i})$$

$$g_m = \text{Dropout}(\text{Linear}(\text{ReLU}(\text{Linear}(\text{ReLU}(\text{Linear}(\quad))))))$$

### Loss functions

To compute the reconstruction loss  $R$ , the outputs  $y$  are compared to the original inputs  $x$  for each combination of modality pairs. For example, for modalities  $a, b$ :

$$R_{a,b} = \frac{1}{n} \sum_{i=1}^n (x_{a,i} - y_{b,i})^2$$

where  $n$  is the total number of proteins present in both modalities  $a, b$ . The overall reconstruction loss is then sum of reconstruction losses across each

combination of pairs of modalities in the set of all modalities  $M$ :

$$R = \sum_{a \in M} \sum_{b \in M} R_{a,b}$$

To compute triplet loss  $T$  for modalities  $a, b$ , each  $z_{a,i}$  is compared to one positive example  $z_{b,i}$  (the same protein in the other modality) and one negative example  $z_{b,k}$  (a different protein in the other modality;  $k$  is a randomly sampled protein where  $k \neq i$ ):

$$T_{a,b} = \frac{1}{N} \sum_{i \in N} \max(\|z_{a,i} - z_{b,i}\|_2 - \|z_{a,i} - z_{b,k}\|_2 + \varepsilon, 0)$$

where  $N$  is the set of all proteins shared between  $a, b$  and  $\varepsilon$  is the triplet margin. For each batch, the loss  $T_a$  for modality  $a$  is computed by randomly selecting a single separate modality  $b$  (where  $a \neq b$ ); this selection is performed independently for each batch (see section “Model training”). The overall triplet loss is then:

$$T = \sum_{a \in M} T_a$$

The full loss function  $L$  is a weighted sum of the reconstruction and triplet losses:

$$L = \lambda R + (1 - \lambda) T$$

## Model training

Model parameters were trained with standard neural network learning procedures provided by PyTorch (Paszke *et al.*, 2019) v2.0.1, based on backpropagation using the Adam stochastic gradient descent method (Kingma and Ba, 2014). Values of hyperparameters were set based on

previous work (Bao *et al.*, 2022; Schroff *et al.*, 2015) without fine-tuning: batch size = 64,  $\lambda = 0.5$ , Adam optimization learning rate = 0.0001,  $\varepsilon = 0.2$ , dropout = 0.5.

## Comparison with original data modalities

To compare ProteinProjector embeddings with the original data modalities’ embeddings (**Figure 2c**) the primary metric used was the Area Under the Receiver Operating Characteristic (AUROC), comparing the distribution of ProteinProjector protein proximities for positive vs. negative pairs in each original modality. For each original modality, positive protein pairs were defined as the top 1% of most similar pairs (cosine similarity) in the embeddings (see section “Compilation of HEK293 features”), and negative pairs were all other pairs.

## Protein functional analysis

For each branch of the Gene Ontology (January 2024 release), the number of GO terms covered in ProteinProjector protein proximities was determined as follows. For the set of proteins in each GO term, we determined the distribution of protein proximities for all pairs of these proteins. This similarity distribution was then compared to a null distribution (all pairs of proteins not in any GO term, i.e. assigned to root node only) using a one-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction (**Figure 3a**, <1% FDR). For single modalities, the cosine similarities between original embeddings were

used (see section “Compilation of HEK293 features”).

## Comparison with orthogonal datasets

CORUM complexes were obtained from NDEx (v4.1, NDEx uuid 764f7471-9b79-11ed-9a1f-005056ae23aa), and pairs of proteins with co-presence in a complex were extracted. BioPlex protein pairs were obtained from NDEx (uuid 6b995fc9-2379-11ea-bb65-0ac135e8bacf). High-confidence STRING v12 pairs were obtained from NDEx (uuid 0b04e9eb-8e60-11ee-8a13-005056ae23aa). For protein co-essentiality pairs, the K562 day-8 perturb-seq dataset was acquired at gwps.wi.mit.edu (BioProject ID PRJNA831566); we computed a pairwise Pearson correlation matrix and extracted the top 1% most similar pairs as interactions. Protein co-abundance data was downloaded from a previous study (Gonçalves *et al.*, 2022); we computed a pairwise Pearson correlation matrix and extracted the top 1% most similar pairs as interactions. Other thresholds (top 5%, 10%, and 20%) were evaluated in Supplementary Fig. 3. Human Protein Atlas (HPA) data images were downloaded

([https://github.com/idekerlab/cellmaps\\_image downloader](https://github.com/idekerlab/cellmaps_image downloader)) and classified using Densenet (Ouyang *et al.*, 2019)([https://github.com/idekerlab/cellmaps\\_image embedding](https://github.com/idekerlab/cellmaps_image embedding)); we determined pairs of proteins with a shared subcellular compartment. To compare ProteinProjector embeddings with these orthogonal datasets (**Figure 3c,d,e**), the primary metric used was the AUROC for protein pairs

positive in the evaluation set (e.g. pairs of proteins in CORUM complex) vs. negative pairs (e.g. pairs of proteins not in CORUM complex). For single modalities, the cosine similarities between original embeddings were used (see section “Compilation of HEK293 features”). For the concatenation comparison (Supplementary Fig. 1), if a protein was missing from a modality, a feature from that modality was randomly selected. For the standard autoencoder comparison, embeddings from each modality were first encoded separately with a linear layer and Rectified Linear Unit (ReLU) activation function to produce a 128-dimensional latent vector. These vectors were then concatenated and passed through an additional linear layer to produce a 128-dimensional latent vector. This latent vector was subsequently decoded back into the respective modalities, and the reconstruction accuracy was evaluated to ensure effective integration. The model was trained with data present in all four modalities for 50 epochs using a batch size of 64 and the Adam optimizer.

## Cell Map Construction

We used the Cell Mapping toolkit ([https://github.com/idekerlab/cellmaps\\_generate\\_hierarchy](https://github.com/idekerlab/cellmaps_generate_hierarchy)) to generate a hierarchical cell map of protein assemblies (**Figure 4a**). The ProteinProjector protein proximities were used to generate a series of protein-protein proximity networks in which edges were defined from the most similar 0.2, 0.4, 0.6, 0.8, 1.0, or 5.0% pairs, respectively, yielding 6

networks total. Pan-resolution community detection was performed in each of these networks using the Hierarchical community Decoding Framework (HiDeF, <https://github.com/fanzheng10/HiDeF>) (Zheng *et al.*, 2021), with a persistence threshold ( $k$ ) of 10 and a maximum resolution (maxres) of 80, with other parameters kept at default settings. The Cell Mapping toolkit ([https://github.com/idekerlab/cellmaps\\_hierarchy](https://github.com/idekerlab/cellmaps_hierarchy)) was also used to determine overlap with GO and CORUM terms via a hypergeometric test with Benjamini-Hochberg

correction (<1% FDR and Jaccard index > 0.2). To determine which assemblies were driven by the original data modalities, we performed the following. For the set of proteins in each assembly we determined the cosine similarity between original embeddings for all pairs of these proteins (see section “Compilation of HEK293 features”). This similarity distribution was then compared to a null distribution (all pairs of proteins not in any common assembly, i.e. assigned to root node only) using a one-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction (**Figure 4a,b**, <1% FDR).

## Competing interests

T.I. is a co-founder, advisor, and holder of equity for Data4Cure and Serinus Biosciences, and he is an advisor and shareholder for Ideaya BioSciences. The terms of these arrangements have been reviewed and approved by UC San Diego in accordance with its conflict of interest policies.

## Author contributions

L.V.S. and M.J. developed ProteinProjector. L.V.S., M.J., and R.N. performed analyses. L.V.S., R.N., R.S., and T.I. conceived analyses. L.V.S., M.J., and T.I. wrote the manuscript with input from all authors.

## Acknowledgements

We gratefully acknowledge funding from Schmidt Futures (T.I.), the Bridge2AI Program (NIH Common Fund; OT2 OD032742; T.I.) and the Israel Science Foundation (1692/24; R.S.). We also thank Mengzhou Hu, Gege Qian, Li Zhang, and Han Guo for helpful discussions.

1200–1209.

Bludau, I. *et al.* (2020) Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes. *Nat. Protoc.*, **15**, 2341–2386.

Cho, N.H. *et al.* (2022) OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, **375**, eabi6983.

Forster, D.T. *et al.* (2022) BIONIC: biological network integration using convolutions. *Nat. Methods*, **19**, 1250–1261.

Fossati, A. *et al.* (2023) Next-generation proteomics for

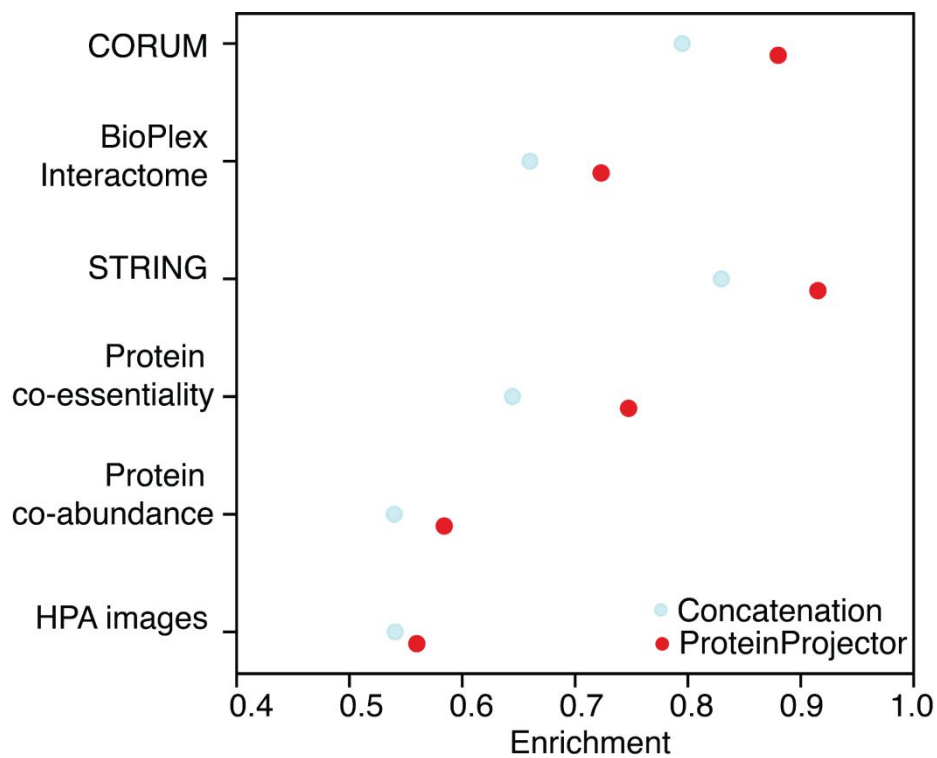
## References

Bao, F. *et al.* (2022) Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. *Nat. Biotechnol.*, **40**,

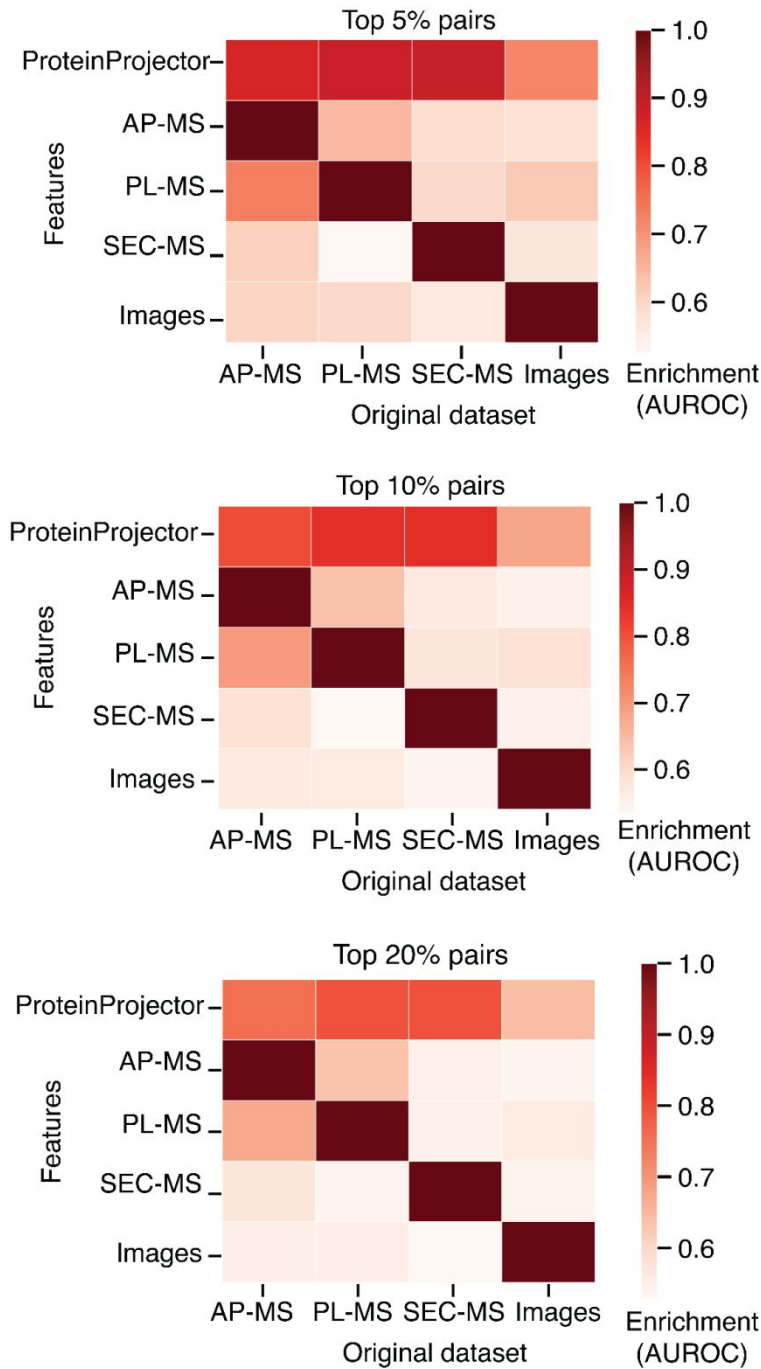
- quantitative Jumbophage-bacteria interaction mapping. *Nat. Commun.*, **14**, 5156.
- Geladaki,A. *et al.* (2019) Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.*, **10**, 331.
- Girdhar,R. *et al.* (2023) ImageBind one embedding space to bind them all. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 15180–15190.
- Go,C.D. *et al.* (2021) A proximity-dependent biotinylation map of a human cell. *Nature*, **595**, 120–124.
- Gonçalves,E. *et al.* (2022) Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*, **40**, 835–849.e8.
- Gordon,D.E. *et al.* (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, **583**, 459–468.
- Havugimana,P.C. *et al.* (2012) A census of human soluble protein complexes. *Cell*, **150**, 1068–1081.
- Heusel,M. *et al.* (2019) Complex-centric proteome profiling by SEC-SWATH-MS. *Mol. Syst. Biol.*, **15**, e8438.
- Huttlin,E.L. *et al.* (2021) Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, **184**, 3022–3040.e28.
- Huttlin,E.L. *et al.* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, **162**, 425–440.
- Johnson,K.L. *et al.* (2021) Revealing protein-protein interactions at the transcriptome scale by sequencing. *Mol. Cell*, **81**, 3877.
- Kim,D.I. *et al.* (2014) Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E2453–61.
- Kingma,D.P. and Ba,J. (2014) Adam: A method for stochastic optimization. In, Bengio,Y. and LeCun,Y. (eds), *3rd International Conference on Learning Representations (ICLR)*.
- Kobayashi,H. *et al.* (2022) Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods*, **19**, 995–1003.
- Luck,K. *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.
- Mulvey,C.M. *et al.* (2017) Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.*, **12**, 1110–1135.
- Nasser,R. and Sharan,R. (2023) BERTwalk for integrating gene networks to predict gene- to pathway-level properties. *Bioinform. Adv.*, **3**, vbad086.
- Ouyang,W. *et al.* (2019) Analysis of the Human Protein Atlas Image Classification competition. *Nat. Methods*, **16**, 1254–1261.
- Paszke,A. *et al.* (2019) PyTorch: an imperative style, high-performance deep learning library. In, Wallach,H.M. *et al.* (eds), *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., pp. 8026–8037.
- Qin,Y. *et al.* (2021) A multi-scale map

- of cell structure fusing protein images and interactions. *Nature*. Radford, A. *et al.* (18-24 Jul 2021) Learning Transferable Visual Models From Natural Language Supervision. In, Meila, M. and Zhang, T. (eds), *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, pp. 8748–8763.
- Richards, A.L. *et al.* (2021) Mass spectrometry-based protein–protein interaction networks for the study of human diseases. *Mol. Syst. Biol.*, **17**, e8792.
- Schaffer, L.V. *et al.* (2025) Multimodal cell maps as a foundation for structural and functional genomics. *Nature*, **In Press**.
- Schroff, F. *et al.* (2015) FaceNet: A unified embedding for face recognition and clustering. *arXiv [cs.CV]*.
- Skinnider, M.A. *et al.* (2021) An atlas of protein-protein interactions across mouse tissues. *Cell*, **184**, 4073–4089.e17.
- Srivastava, N. *et al.* (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.
- Stacey, R.G. *et al.* (2017) A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics*, **18**, 457.
- Szklarczyk, D. *et al.* (2018) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
- Thul, P.J. *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**.
- Thul, P.J. and Lindskog, C. (2018) The human protein atlas: A spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
- Tsherniak, A. *et al.* (2017) Defining a Cancer Dependency Map. *Cell*, **170**, 564–576.e16.
- Tsitsiridis, G. *et al.* (2022) CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Res.*, **51**, D539–D545.
- Wilhelm, M. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
- Yang, K.D. *et al.* (2021) Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.*, **12**, 31.
- Zheng, F. *et al.* (2021) HiDeF: identifying persistent structures in multiscale 'omics data. *Genome Biol.*, **22**, 21.

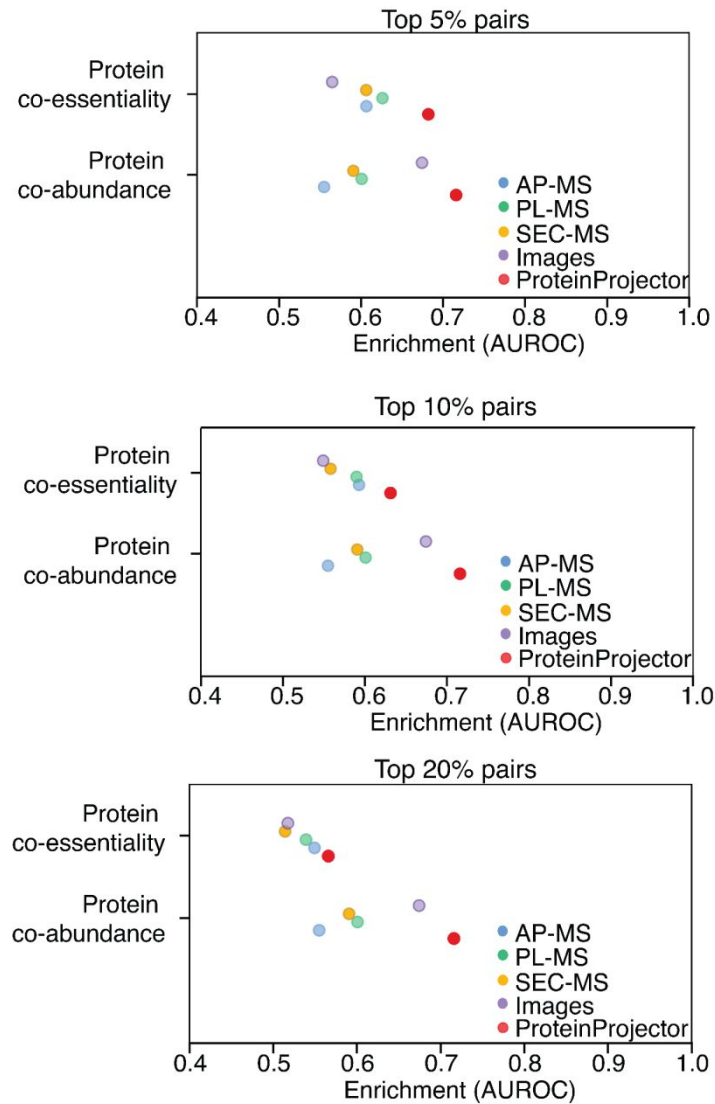
## Supplementary Figures



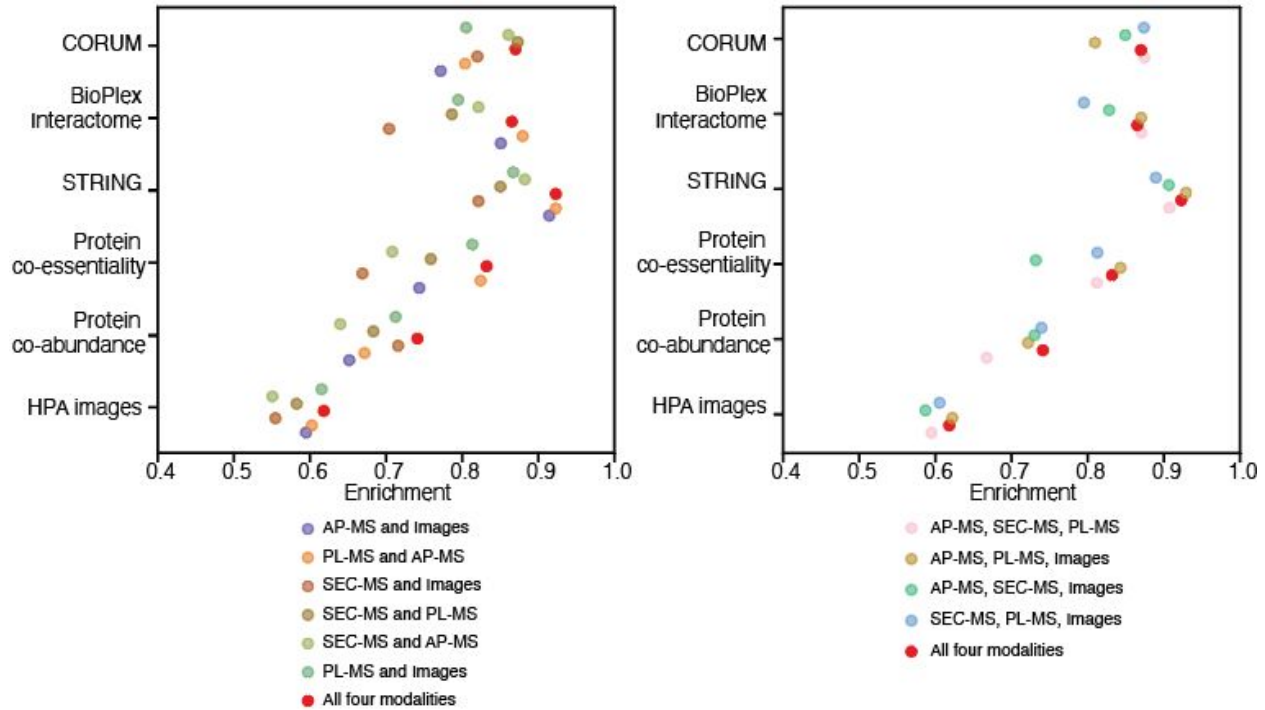
**Supplementary Fig. 1.** Comparison of ProteinProjector embeddings to simple concatenation of input features using the union of all proteins present in the four modalities (Methods).



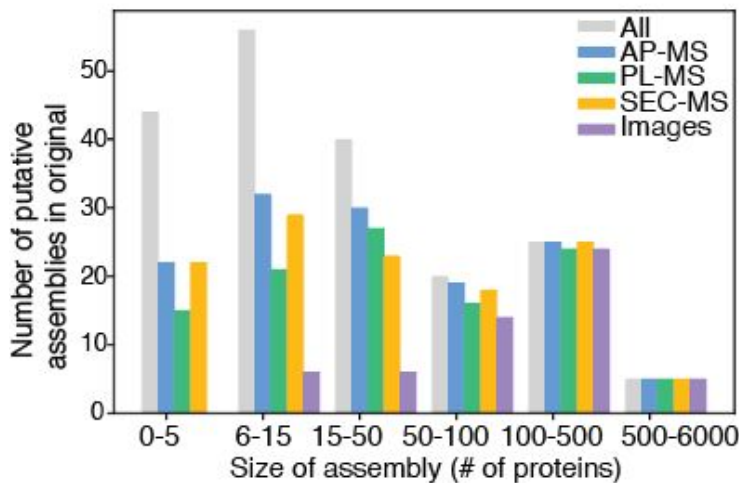
**Supplementary Fig. 2** Agreement (white-to-red color gradient) of each original data type embeddings (columns) to each other (rows) or to the ProteinProjector embedding (top row). Agreement measured by enrichment of most similar protein pairs at different thresholds (5%, 10%, and 20%) in one embedding versus another, defined by AUROC (Methods).



**Supplementary Fig. 3.** Degree of enrichment (AUROC, Methods) of similar protein proximities (colored points, each data modality individually and combined with ProteinProjector) for orthogonal functional and physical association datasets not used in model training (rows), focused on proteins present in all four original datasets. Different thresholds were used for orthogonal datasets Protein co-essentiality and Protein co-abundance to define protein pairs (top 5%, 10%, and 20% most similar pairs, Methods).



**Supplementary Fig. 4.** Degree of enrichment (AUROC, Methods) of similar protein proximities (colored points, each data modality individually and combined with ProteinProjector) for orthogonal functional and physical association datasets not used in model training (rows), using different subsets of the modalities in training ProteinProjector.



**Supplementary Fig 5.** Number of putative assemblies with support from original data modalities (Methods) versus size of assembly in number of proteins. Gray bars denote the total number of assemblies in each size category

## Supplementary Tables

Modality	Total	CORUM	BioPlex Interactome	STRING	Protein co-essentiality	Protein co-abundance	HPA images
<b>AP-MS</b>	5372	2351	4759	2862	783	1983	4266
<b>PL-MS</b>	4473	1838	4133	2163	607	1798	3775
<b>SEC-MS</b>	4387	1976	4199	2370	700	1961	3604
<b>Images</b>	1311	728	1252	850	226	587	1109

**Supplementary Table 1.** Number of proteins in each modality and overlap with each orthogonal dataset analyzed.

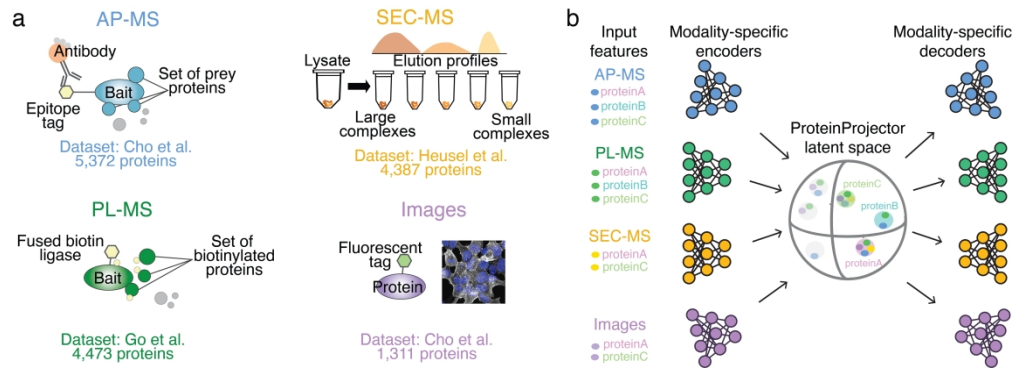


Figure 1 | ProteinProjector Overview. a) Overview of data modalities obtained from application of proteomic technologies to HEK293 cells: affinity purification (AP-MS), proximity labeling (PL-MS), size exclusion chromatography (SEC-MS), and fluorescence imaging. b) Schematic diagram of ProteinProjector encoder-decoder neural network architecture designed for multi-modal data integration and interpolation.

446x162mm (300 x 300 DPI)

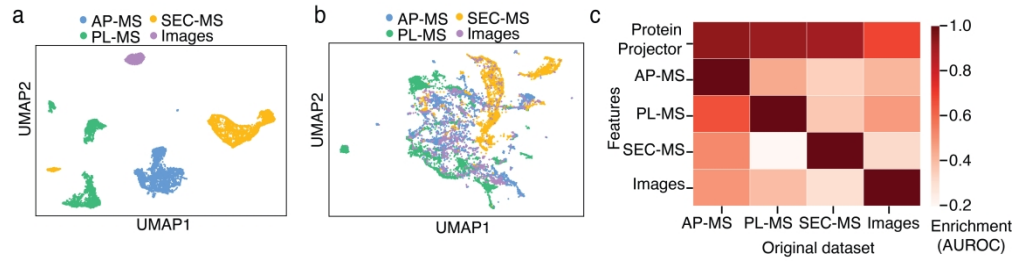


Figure 2 | ProteinProjector representations. a) UMAP visualization of ProteinProjector embeddings for each protein prior to training, colored by input modality. b) UMAP visualization of ProteinProjector embeddings for each protein after training, colored by input modality. c) Agreement (white-to-red color gradient) of each original data type embeddings (columns) to each other (rows) or to the ProteinProjector embedding (top row). Agreement measured by enrichment of most similar protein pairs in one embedding versus another, defined by AUROC (Methods).

445x115mm (300 x 300 DPI)

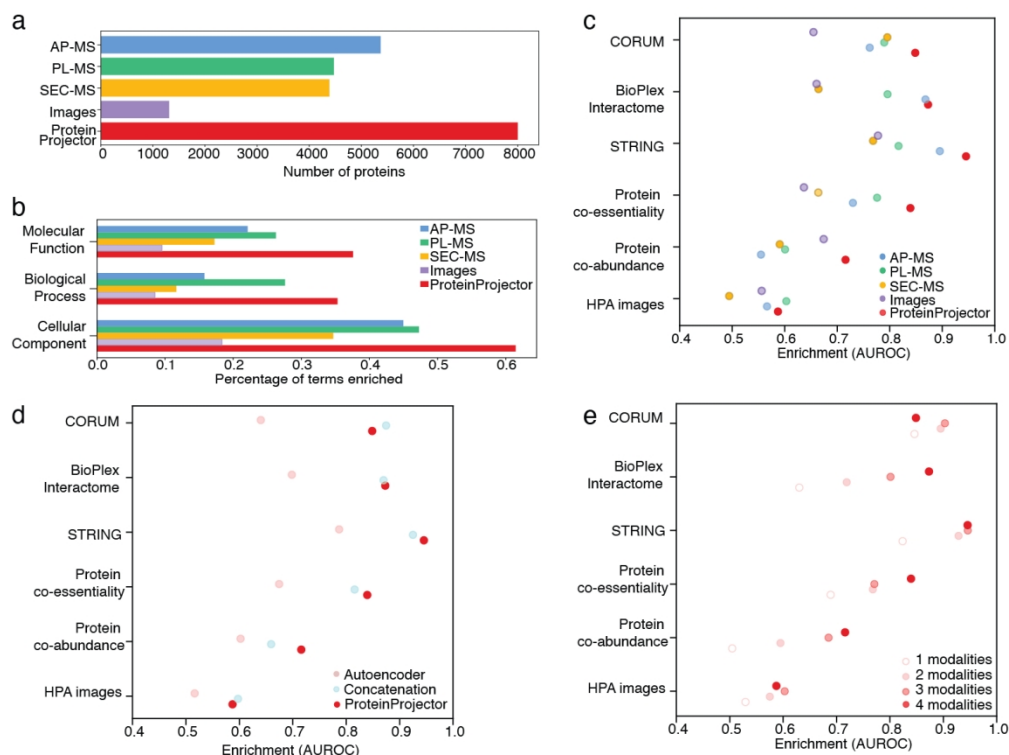


Figure 3 | Evaluation of ProteinProjector integration a) Number of proteins covered in each original modality vs. ProteinProjector. b) Fraction of Gene Ontology terms recovered in similar protein proximities for original modality embeddings and ProteinProjector (Methods, <1% FDR). c) Degree of enrichment (AUROC, Methods) of similar protein proximities (colored points, each data modality individually and combined with ProteinProjector) for orthogonal functional and physical association datasets not used in model training (rows), focused on proteins present in all four original datasets. d) Similar to panel (c), comparison of ProteinProjector embeddings to a standard autoencoder integration and to simple concatenation of input features (Methods). e) Similar to panel (c), but for ProteinProjector embeddings that incorporate increasing numbers of data modalities (colored points).

336x254mm (118 x 118 DPI)

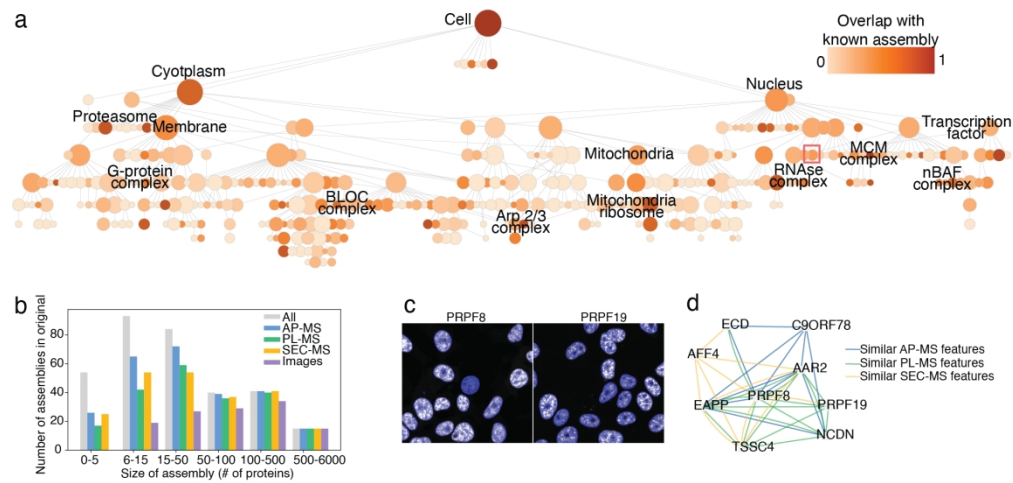


Figure 4 | Cell map of protein assemblies in HEK293 based on ProteinProjector. a) Hierarchy of protein assemblies constructed by performing multi-scale community detection on ProteinProjector embeddings. Each node is colored based on overlap with known subcellular components from GO, CORUM, and HPA (Methods). Red box denotes assembly highlighted in c,d. b) Number of assemblies with support from original data modalities (Methods) versus size of assembly in number of proteins. Gray bars denote the total number of assemblies in each size category. c) Live fluorescence cell images for proteins in spliceosome-associated complex present in the imaging dataset (PRPF8 and PRPF19). d) Spliceosome-associated complex supported by similar features (top 1% pairs by cosine similarity) from original data modalities.

456x217mm (118 x 118 DPI)