Degree project MSc Molecular Techniques in Life Science

Georgios Panos

Mass spectrometry-based functional annotation of proteoforms using network embeddings

PI Responsible: Prof. Janne Lehtiö

Supervisors: Ioannis Siavelis and Yanbo Pan

Spring 2025

Table of Contents

1. Abstract

Bottom-up mass spectrometry (MS)- based proteomics have finally reached near genomics scale, quantifying thousands of proteins and becoming a tour de force of clinical research. However, the cistronic view of gene's function is not the only perspective worth being investigated in systems biology and proteoform level analysis has yet to take the center stage. Proteoforms could resolve several setbacks of the gene centric analysis such as the confounding multilocalizing proteome, the extreme phenotypic variation and the missed hits from hypothesis testing. Nevertheless, the digestion of proteins into peptide precursors precludes direct identification of proteoforms. In this light, we suggest a peptide centric approach which incorporates the MS signal covariation assumptions for protein deconvolution into proteoform groups. We combine proteoforms with deep learning algorithms to compress MS signals and integrate networks inferred from bottom-up subcellular and total cell proteomic datasets. In the integrated network, proteoforms in proximity are assigned a putative function using enrichment analysis of the neighboring proteoforms under the guilt-by-association principle. As a case study, we set out to deconvolve protein members of the mTROC1 complex, mTOR and RPTOR. Enrichment analysis and subcellular profiling suggest that the spatial separation of the proteoform components of mTORC1 could reflect distinct temporal states that are differentially allocated across tissues rather than unique functional entities with their own repertoires.

2. Popular Science

'Show me who your friends are, and I will tell you who you are.'

The fundamental unit of our whole body is the cell, and it can't live without proteins. Through proteins our cells have their own skeletal structure, their own subcellular compartments – their inner decorations - they communicate with each other, feed, digest and faithfully perform their responsibilities to keep our body in check. For years we thought that only 20,000 proteins are behind this functional repertoire, but this estimation is quite erroneous. It seems that in our 3 trillion cells and among millions and millions of people, these proteins come in different lengths, flavors or so-called post translational modifications and reside in different locations in the same cell at the same time.

So, every one of these proteins is a family of hidden proteoforms, forms of the same protein with sometimes different functions and most importantly different associates. If we think about it, one protein could be a family of two proteoforms inhabiting different places. Then what is the chance these two proteoforms would be neighboring with different proteins and their proteoforms? Who are their friends and what function do they have inside one cell or even many different cells? These questions are important if we consider that almost all the drugs out there target proteins in a very precise manner. A minor difference in the same protein because of two proteoforms could be the reason why most cancers show drug resistance. One of the forms could be located somewhere else and still be active. To get a glimpse of how this happens, first we need to identify the proteoforms and monitor their friends to understand them.

This is exactly what we set out to answer about these proteoforms and their friends. We wrote an algorithm which scans how these proteoforms are related to each other and where they are located in different cells. We generated a massive subcellular map to find the different locations of these proteoforms, and then we zoomed in to check out the neighborhood they were residing in. Subsequently, we asked the simplest question. What exactly do their neighbors do? And we got some answers: We found two proteins that work together for cellular digestion, left possible tracks in totally different places, still cooperating and having different neighbors. Most importantly some cancer cells favor one pair of proteoforms while others favor the other, which begets a new question: Can we use this against them?

3. Introduction

The field of mass-spectrometry (MS)-based proteomics has evolved significantly in the last decade, becoming one of the most essential methods of assessing protein composition in biological samples. Bottom-up approaches, i.e., enzymatically digesting the protein samples into smaller peptides followed by LC-MS analysis to capture their unique spectra, have been the primary focus of clinical proteomics. The experimental MS/MS spectra of these peptides are searched against in silico spectra generated from simulated digestion of theoretical peptides from protein databases. A unique peptide can be fragmented multiple times and thereafter be associated with more than one PSMs – Peptide spectra Matches[1–3]. In gene-centric analysis the quantitative signals of these peptide fragments are collected, and based on coding sequence match, the peptide quantities are aggregated into unique measurements. In this way, the presence and quantity of specific proteins in a sample are inferred[4,5]. One of the key merits of this methodology is the genomic scale depth of analysis, approximating almost 10.000 proteins per MS run. In addition, the summarization of peptide mass spectra signals to the protein level abundance confers several benefits: (1) it mitigates the effect of the outliers and missing values, (2) downsizes the number of hypothesis tests and (3) simplifies biological interpretation[6]. From molecular portraits of a broad spectrum of cancers to an exhaustive interrogation of protein subcellular localization, bottom-up proteomics has been applied to a wide range of applications, pushing towards the field of precision oncology[7–10].

However, biological exuberance and phenotypic variation cannot be defined solely by the "cistronic" perspective of protein diversity, that of the gene allele heterogeneity. It has been almost three-quarters of a century since Benzer introduced the concept of the cistron – the gene as an immobile, central entity of information flow -, yet its essence remains in how we perform systems biology to this very day[11]. The representative size of the human proteome, from splice variants to post-translational modifications and proteolytic cleavage, exponentiates dramatically our 20.000 genes to the million scale. A modern perspective on phenotypic plasticity is the proteoform - any distinctive molecular form that corresponds to an expressed protein – and can be the representative summarization of a gene's function[12]. These unique molecular forms, or more precisely, the lack of their study, can be a major factor behind several shortcomings in proteomic analysis. For instance, it has been observed that proteins with high peptide coverage display very low to no fold-change in various MS-proteomic experiments. The absence of the effect size could be partly attributed to different proteoform species under the same protein not being differentially expressed in the same direction or extent. Hence, averaging the signal intensities of different forms may conceal statistically significant differences or even miss the underlying biology behind the adaptive evolution of disease phenotypes such as cancer. In precision oncology, drug resistance mechanisms can be orchestrated by a proteoform of the target protein that is not affected by the specific inhibitor due to a minor structural dissimilarity between these proteoform groups[13,14]. These discrepancies also involve subcellular biology. A mixture of signals from different proteoforms inhabiting diverse subcellular organelles can be the underlying cause of the multi-localizing proteome, which is estimated to be at least 30% of total protein count [8,15]. This complexity increases even further when we consider the protein-complex layer under the stable stoichiometry model, where proteins with common membership in a functional unit are co-abundant and co-regulated by the proteostatic surveillance mechanisms of the cell[16]. Protein relocalization and alterations of protein complex composition are a largely neglected source of proteoform entities and heterogeneity. With bottom-up proteomics, as advantageous as they are, we cannot directly interpret proteoforms from peptide aggregation because these precursors are fragmented and predominantly map to overlapping regions of splice isoforms. Aside from the sequence variation issue, digestion with proteolytic enzymes also decouples the fragmented peptide from its proteoform of origin.

In this context, the topic of my thesis and, by extension, the work of our group, concerns the deconvolution of proteins into multiple proteoform species at the bulk and subcellular level, followed by their functional annotation. Approaches of proteoform discovery in bulk proteomics have already been introduced[17] but this method relies on two novelties, the inclusion of MS-based subcellular proteomic profiles, and the use of deep learning algorithms for dimension reduction and network integration. By reconstructing a pan-cancer integrated proteoform network, we aimed to find the biological role of the deconvoluted proteoforms based on their local neighborhood using the guilt-by-association principle –proteins with similar functions tend to be in proximity in the network.

As this includes several bioinformatic pipelines, it is important to highlight the biological assumptions that underscore how this network was reconstructed. 1) In bottom-up proteomics, peptides that share a distinct covariation pattern of their mass spectra signals across multiple samples could potentially be grouped into distinct proteoform species. Especially, in the case of subcellular MS-based proteomics, aberration of signal patterns of peptides related to the same protein could be an indication of proteoforms residing in different subcellular compartments. The deconvolution is based on a graph-based clustering method that assigns different communities of peptides using the covariation of their signals as a similarity metric. A singular community of peptides is a proteoform by this definition. 2) The signal covariation premise can extend to functional association by the protein complex stoichiometry principle. This principle is predicated on the fact that protein complexes are assemblies of multiple subunits with defined stoichiometry, and cells tend to regulate the abundance of complex subunits in a coordinating manner[16,18,19]. At this point, an important link between the aggregated protein measurement and its deconvoluted proteoform siblings is the premise that for every protein there is one "similar" proteoform, referred in this thesis as Psim. After deconvolution, if we try to summarize all the peptides (rather than PSMs) into a single protein measurement, this unit measurement usually correlates with the Psim proteoform but not with the rest Palt-alternative proteoforms. We match this Psim proteoform per protein to the respective HUGO gene / ENSEMBL ID to link protein complex affiliation with proteoform measurements across datasets. As a result, co-abundance in bulk as well as in subcellular MS-based proteomics has the potential to be a strong indicator of functional association between a pair of proteins or proteoforms. This can be further facilitated using denoising dimension reduction algorithms (VAE-Variational autoencoders) and classifiers which use the MS-signal based features for PPI - proteoform-proteoform interaction prediction. These association networks are the basis of the final step, late integration. 3) The integration of co-localization and co-abundance proteoform networks can generate a final localization-aware association network, where proteoform species under a protein might be found at different locations and can be assigned putative functions based on their nearest neighbors. These proteoform associations can be corroborated by functional enrichment analysis of the extracted neighboring nodes. In the scope of this thesis, I am testing the potential of such an approach for biological exploration of the understudied proteoform.

4. Methods

4.1. Data Collection and Preprocessing

All preprocessing was conducted using the R programming language[20] (version 4.3.4) in the RStudio development environment. I extracted data from two different manually curated protein complex databases, CORUM[21] and COMPLEAT[22], where both were used as a ground truth for labeling true and false interactions in the training and test data. Protein complexes with a size ranging from 3 to 22 members were chosen to reduce large complex bias, and subsequently, a correction for PPI redundancy was done by extracting all possible and unique protein pairs from the annotated complexes. The ground truth pairs were labeled with a HUGO gene symbol convention, which was later matched to the respective proteoforms by their ENSEMBL ID. Furthermore, STRING interactions of high confidence (score > 0.7) were also curated following the procedure mentioned above[23].

ABMS (Antibody validation using Mass Spectrometry) is an LC-MS TMT-based 10-plex bulk proteomics dataset consisting of eighteen cancer cell lines in triplicate. The peptide input of the dataset accounts for 404539 unique mass spectra across the 54 samples[24]. The Lehtiö group has also recently expanded SCBC-SubCellBarCode –a TMT-based 10-plex quantitative subcellular proteomics resource– from 7 to 13 different cancer cell lines[25]. I analyzed this dataset both in a peptide- and protein-centric manner because a crucial objective of the project is to compare the predicted localizations of the deconvoluted proteoforms with their related proteins. The generation of MS-based subcellular proteomics data is illustrated and explained at **Fig.1a**. The SCBC data consists of 13 cell lines in duplicates at 5 subcellular fractions each, with a total of 130 samples and an estimate of 358923 peptides summarized to 15168 unique protein symbols. Protein levels of the SCBC dataset were calculated by median summarization of all the PSMs that have similar ENSEMBL ID after running them to the spectral library. Both proteomic datasets were based on DDA-Data Dependent Acquisition instrumentation, so during preprocessing each TMT-channel was log2-transformed, normalized on the mean value across all channels and rows were median-centered.

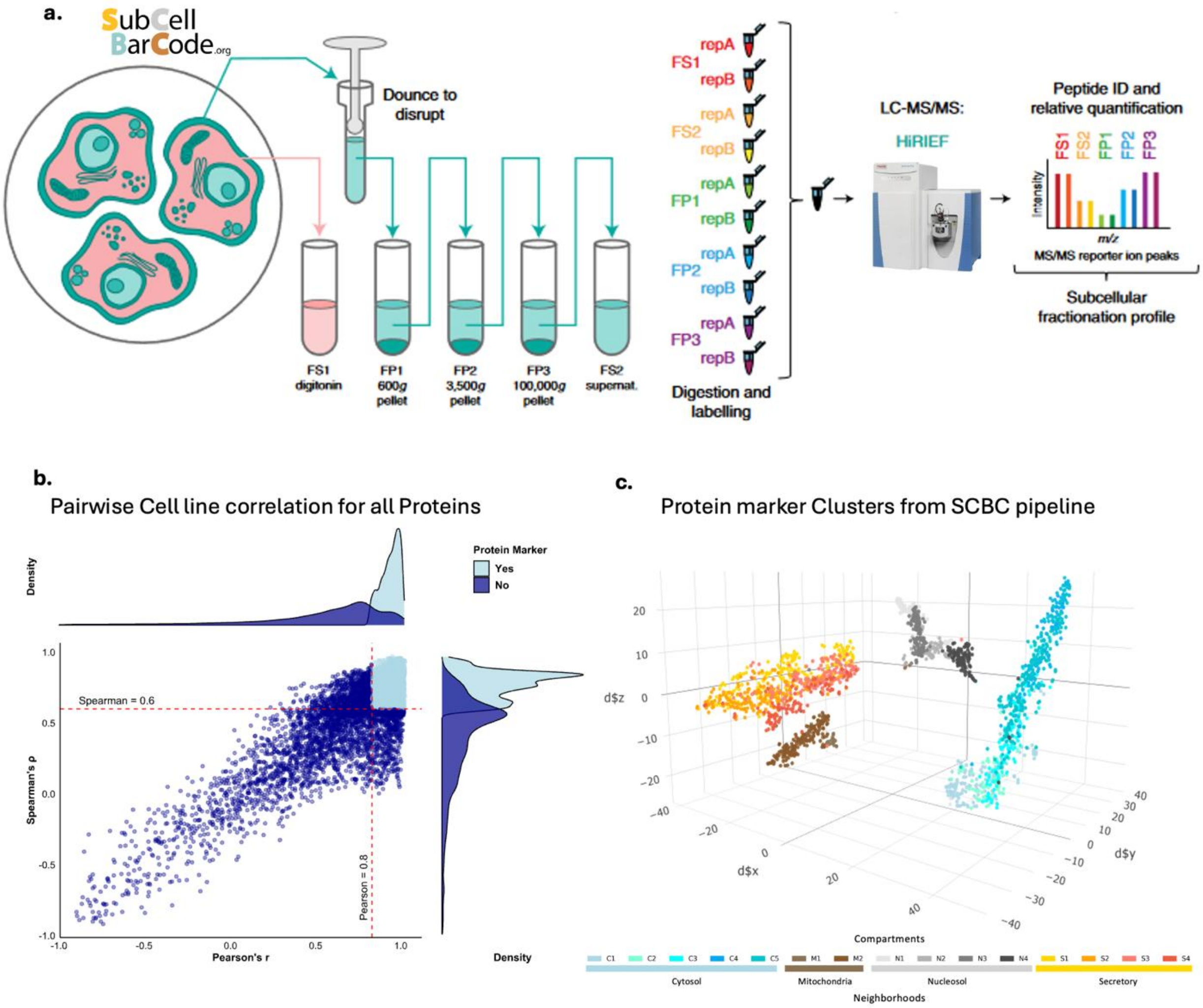


Figure 1: a. The Subcellular proteomics wet-lab workflow. In the original paper, the authors performed subcellular fractionation for in-vitro cell lines in duplicates, generating in total 10 samples per cell type. Currently, this framework has expanded to 13 cancer cell lines. Briefly, the samples were cleaned, digested with trypsin and labeled with tandem mass tags (TMT-10plex). HiRIEF, a peptide prefractionation method based on isoelectric focusing and elution of the peptides in 72-subsmamples, was applied to increase the depth of the spectra acquisition. A data-dependent acquisition approach was followed where each peptide is matched to a spectral library database by a searching algorithm. Finally for each peptide with a spectra match, a unique mass-spec profile of 10 abundance values is generated. **b. Protein Subcellular Marker discovery in SCBC pipeline.** The proteins were correlated cell-line-wise across five cell lines and the minimum correlation coefficient (Pearson and Spearman) across the 10 pairwise coefficient was plotted for each protein. Protein markers are the proteins whose coefficients were robust across all the pairwise correlations above the indicated thresholds. **c. Clustering of Protein Markers and Prediction of Subcellular Localization:** The protein markers were clustered based on the first three t-SNE coordinates into 15 compartments and 4 neighborhoods, which were identified by heuristics and GO enrichment analysis of each compartment. The next step of the pipeline dictated that protein markers from any SCBC cell line which overlap with the clustered markers are used to train a Support Vector Machine (SVM). The SVM model assigns for any peptide or protein membership probabilities for each neighborhood using directly the subcellular fractionation profile as a predictor.

4.2. Proteoform deconvolution and subcellular localization prediction.

An in-house pipeline was used for the deconvolution of the proteoforms, called DepMS. It takes as an input the normalized peptide tables of each cell line, either SCBC or ABMS. Before the analysis, peptides were filtered to those corresponding to one unique protein symbol and assigned to proteins with more than one unique peptide. According to DepMS, peptides that have deviating quantitative profiles and covary across multiple samples can be grouped into the same "proteoform". The similarity metric used was Pearson correlation, and Louvain clustering algorithm[26,27] was used to group the peptides into the proteoform communities with a modularity threshold of 0. In this scheme, each unique peptide is a node, and their Pearson similarity is represented by an edge. Furthermore, the communities per protein are first organized into two basic categories, Psim which share a similar profile with the median summarization of all the peptides of the same protein they originate, and Palt – the other proteoform groups of the same protein and usually do not correlate with the summarized peptide measurement. Subsequently Palt are deconvoluted further into proteoform species. This pipeline was developed by Dr. Ioannis Siavelis and is available here[28].

The second in-house pipeline takes only the SCBC table of each cell line as an input and is thoroughly explained in **Fig.1b,c**. The SCBC pipeline involves the localization prediction of any proteoform or protein across the 10 samples of different cell lines into 4 unique subcellular neighborhoods, and 15 compartments. The principle is also built upon the premise of the covariation of mass-spec signals. In this pipeline, a Support Vector Machine (SVM) is trained to classify each protein in the sample into these distinct neighborhoods. The SVM was trained by subcellular markers, which have unique properties and are explained in **Fig.1b.** It is important to note that the SCBC matrices were analyzed in a long format. All cell lines were first collapsed into a single table (10 columns – 5 fractions x 2 replicates) for proteoform deconvolution, then the proteoform table was converted into wide format for each cell line to perform the SCBC pipeline.

4.3. Dimension reduction with Variational Autoencoder - VAE.

I developed the following framework, which utilizes two basic components: 1) a compression algorithm, the VAE, and 2) random forest classifiers. In this section, I will be describing the first component, the compression by the VAE (**Fig.2**). Generally, VAEs are unsupervised neural network algorithms with encoder-decoder architecture and their primary task is to reduce the dimension of an input vector X (here, the proteoform abundances) to a shorter space Z, known as a latent space which before training is modeled by a standard normal distribution prior *p(Z)-N (0, I)*. Then, the decoder, by sampling from these latent variables/Z-vectors, reconstructs the original input using variational inference[29,30]. It generates a reconstruction vector $\hat{X}$ with distribution $p(\hat{X}|Z)$. Optimizing this task, the encoder maps each proteoform vector into the latent space, whose variables have a mean Zμ(x), variance log(Zσ^2(x)) and all samples follow a Gaussian Posterior q(Z|X) - N(μ(x), diagσ^2(x)). The formation of the latent space during the training is controlled by two objective functions i) a *Kullback–Leibler (KL) divergence* that penalizes the sample-wise posterior *q(Z|X)* whenever it drifts away from the standard gaussian prior, ii) *Masked Gaussian Reconstruction Loss error* that accounts how well the reconstructed vector $\hat{X}$ matches the original proteoform *X* while ignoring entries with missing values with a binary mask. The KL-term acts as an information bottleneck, and a regularization parameter was introduced to avoid the posterior collapse phenomenon[31]. The algorithm's learning rate was optimized using the Leslie Smith range test[32], whereas the value of the KL-regularization term was estimated using heuristics (**Supp.Fig.1)**. To determine how many latent and hidden dimensions were meaningful, I coupled the latent variables to a logistic classifier and evaluated their predictive value (**Supp.Data.File**). Smaller architectures with the best reconstruction error loss and AUC metrics were prioritized. The model runs unscaled TMT-labeled proteomics data with dropout rate of 0.2, and Leaky Rectified Linear Unit as the activation function in the hidden layer. I wrote the pipeline in Python, and the source code, along with the tech stack, is available here[33]

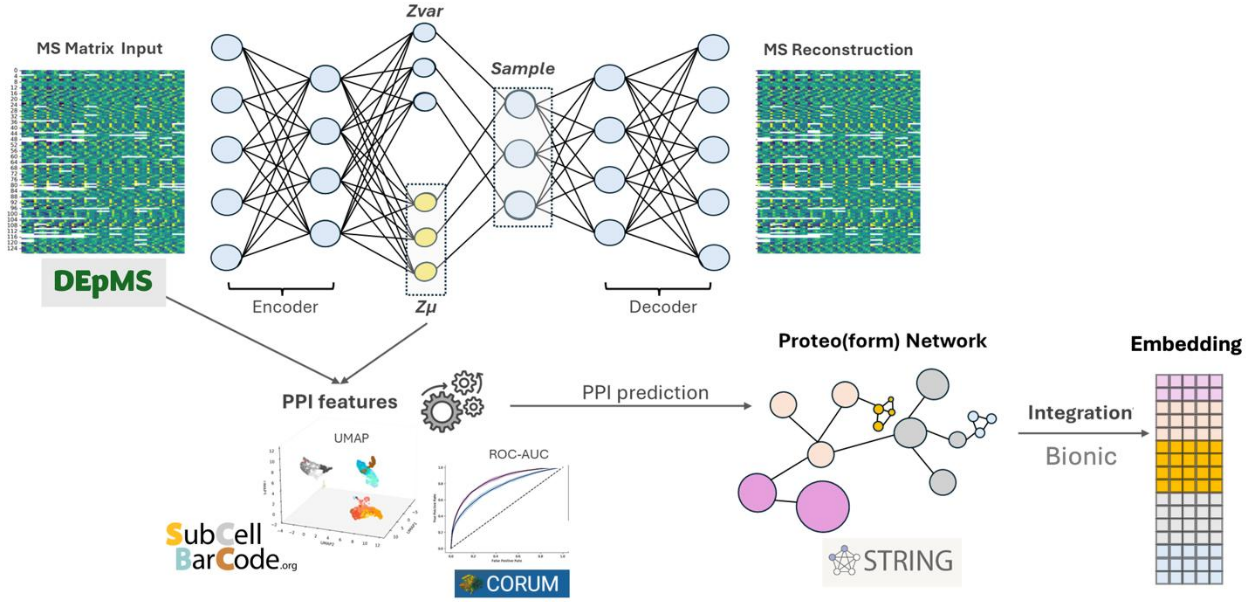


Figure 2: The initial input in the pipeline is a proteoform table in a wide format that could be from the ABMS or the SCBC dataset. In this example, we suppose it is the SCBC table with proteoforms as rows and 130 samples as columns. The VAE, through the encoder, sequentially reduces the dimensions of the table from 130 -> 90 -> 45, where 45 is the size of the latent space. Each proteoform has been mapped to 45 random variables, which are estimated by a mean Zμ (yellow-colored neurons) and variance Zvar. The Decoder samples from these 45 random variables and sequentially reconstructs the input by similar stepping, from 45 -> 90 -> 130. The dimensions are an optimizable hyperparameter and data sensitive. After training, the latent variables (a vector of 45 values/proteoform) along with the original raw values (130 samples) were used for feature engineering (pairwise proteoform correlation, raw or/and VAE based distances, UMAP embeddings of the latent variables, etc.). A random forest classifier is tuned

with 10-fold Cross-Validation, followed by importance analysis, then, the final classifier and predictors are used to assign proteoform-proteoform association probabilities. The association probabilities define the edges of the SCBC proteoform network, and 5% FDR threshold is applied. Protein-Protein interactions reported in CORUM/COMPLEAT are used as ground truth for positives. The Psim-similar proteoforms whose abundance profile correlates with the median summarization of all the deconvoluted proteoforms at the protein level are matched to HUGO gene symbols of the database. The Palt-alternative proteoforms are not included in the training set as they are different entities of the same summarization/protein that Psim is matched to. Negative examples are considered proteoforms inside CORUM/COMPLEAT that do not share membership to any protein complex. All probability scores above the threshold that correspond to 5% FDR were assigned to a value of 1 and the rest were discarded. The final output is an adjacency matrix. Once all the datasets pass through the pipeline, the association networks are integrated with a graph convolutional neural network algorithm called BIONIC. The final output of the integration is an embedding, with a total of 512 dimensions for each proteoform. The embedding was further reduced to 2 dimensions with UMAP for downstream analysis and visualization of the integrated proteoform network.

4.4. Proteoform – Proteoform interaction prediction and association network reconstruction

In the second part of my pipeline (**Fig.2**), I attempt to link the proteoform-proteoform signal covariation to an assignment of putative proteoform interactions. The final output is an association network for each dataset where each node is a proteoform and each edge is an interaction probability. A 5% FDR probability threshold is later applied to filter out the noisy edges from each network. After compression, raw data and VAE embedding-based feature engineering and selection were done with the aim to increase the performance of RF -random forest classifiers. RF classifiers were tuned with 10-fold cross validation and optimized to predict proteoform-proteoform interactions. Proteoform pairs that existed within CORUM/COMPLEAT and corresponded to a protein complex were considered as ground truth positives, whereas negative examples were proteoform pairs within the database but lacking protein complex membership. As the negatives are far more than the true positives, the majority class was undersampled. The protein complex members with HUGO conventions were matched to Psim proteoform counterparts by their ENSEMBL ID. Feature importance analysis was performed in multiple iterations with the tuned RF-predictor for consistency and indicated the most relevant features for the classification. After hyperparameter tuning and feature selection, the models were trained with each dataset following 80-20 data split. Areas under the curve (AUC) for the receiver operating characteristic (ROC) and precision recall curve (PRC) were computed during the 10-fold CV of the models with the final selection of features. I wrote the pipeline in Python, and the source code, along with the tech stack, is available here[33]

4.5. Network integration and functional annotation of proteoform using lower-dimensional embeddings

The two output networks of the previous pipeline were integrated using the BIONIC software with adapted hyperparameters. BIONIC is a graph convolutional network (GCN) algorithm, whose architecture and benchmark are thoroughly described here[34]. The output is a table with proteoforms as rows and the embedding size as columns. Each proteoform post-integration is designated to a graph embedding vector, which will be used for downstream analysis. The embedding dimensions were reduced with UMAP[35] to construct a 2D profile of the pan-cancer proteoform association network. UMAP was run with number of neighbors equal to 40, and a minimum distance to 0.2. Information such as SCBC predicted localization and expression levels of different proteoforms across cell lines was overlaid on the profile. This information was based on the original raw proteoform tables. Finally, using Euclidean distances in UMAP space, each target proteoform's nearest neighbors were extracted, whose functional enrichment was evaluated with gene ontology terms[36] and STRING. The analysis was performed in R with clusterProfiler (v.4.8.3)[37], adjusting p-values of one-sided Fischer tests against GO terms with Benjamini–Hochberg (BH) procedure.

5. Results

5.1 Implementation of Proteoform deconvolution.

Starting from the peptide tables of each dataset, the ABMS – total cell proteomics dataset of 13 cancer cell lines in triplicates, and SCBC – subcellular proteomics dataset of 13 different cell lines in duplicates, I implemented the DepMS pipeline to assign peptides into separate proteoforms. The deconvolution algorithm is built upon the predicate that, across multiple samples, peptides originating from the same gene with co-varying quantitative patterns can be clustered into unique communities. Conceptually, this included pairwise peptide correlations based on sample intensity and Louvain clustering using the Pearson coefficients as a similarity metric. The first decisive point was how the 13 cell lines of the SCBC dataset were to be analyzed. We decided to run the pipeline by merging all the abundance values of the cell lines into a long format instead of analyzing each cell type independently – i.e. 10 samples each time. The rationale behind this choice relies on the following facts: i) Peptides with similar subcellular profiles across different tissues will be grouped together in a cell independent manner. ii) this grouping allows systematic identification of highly pronounced proteoform siblings across multiple tissues that might change subcellular patterns from cell type to cell type, and iii) it is also easier to integrate different proteoforms from different TMT batches. However, the drawback of this method is that we lose cell type specificity for these proteoforms, because not all peptides follow consistent pattern across different tissues.

So, in this model we aimed at generalizability rather than specificity, which also suits the ABMS dataset because it contains samples with diverse tissue origins[24]. In total, 9187 Psim and 509 Palt proteoforms were identified in ABMS while from almost 27073 unique entries in SCBC, 15058 belonged to the Psim group and the rest 12015 into the Palt. At this point, we establish the term Psim, the proteoform with the "similar" quantitative profile to the protein (unit measurement) profile that comes from aggregating the peptides of each gene together (**Fig.3a**). The identification of multiple proteoforms in the SCBC dataset was expected, since subcellular fractionation prior LC-MS/MS increases the depth of peptide coverage. An example of the deconvolution of the TMPO from the SCBC dataset into a Psim and Palt proteoforms is shown at **Fig.3b**. Next, I investigated the cell specificity issue in the SCBC proteoforms tables. To proceed with the downstream analysis, I needed to use a threshold of data completeness because the missingness in the dataset across the 150-sample ranged from 0 to 125 NAs horizontally. Obviously, tissue specific proteoforms or noise, while they were captured, could not be studied further as the covariation of signals requires a decent number of samples. A 70% data-completeness threshold was used for both datasets. Hence, the final proteoform tables included 12622 entries for SCBC and all from ABMS with an overlap estimated around 7000 Psims between the datasets.

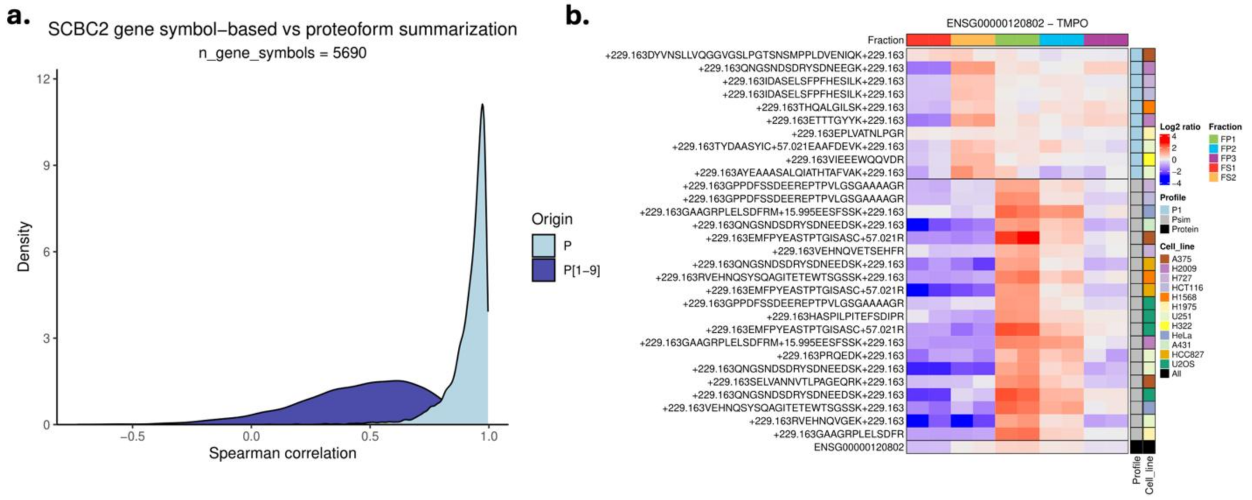


Figure 3: a. Proteoform – Gene level summarization correlations distribution. In the SCBC dataset, 5690 proteins have at least 2 proteoforms; one of them is considered as the Psim, here P (light blue), the rest are designated as P# (dark blue) in the figure. Psim proteoform correlates well with the gene-level summarization of the proteins. **b. Heatmap plot of TMPO peptides.** The SCBC abundance profile of a subset of the identified TMPO peptides from all the cell lines. The profiles are split by proteoform origin and sorted by subcellular fraction. The profile of the Psim is closer to the gene-level summarization of the TMPO protein (last row with ENSEMBLE ID) The relative abundance is log-transformed MS signals, normalized by the average of each TMT-channel. Signal abundance (Log2 ratio) is expressed as log-transformed mean-centered values of the TMT labels across the 10 subcellular fractions for each peptide.

5.2 Unsupervised Compression of mass-spectrometry based signals with VAE

Based on prior knowledge, there is a certain limit to how protein signal covariation can lead to functional association networks. Proteomic co-abundance levels are better predictors for PPI tha mRNA level, but not necessarily under all circumstances[16,18,19,38]. I developed and tested a small pipeline that compresses MS signals under the assumption that this 'compression' will generate complete tables available for discovery. Also, I assumed they would serve as a data structure that might have additive predictive value for proteoform/protein association prediction. Currently VAEs have been used for reduction of sparse single cell transcriptomic matrices and recent literature highlighted their potential for high-dimensional proteomics data analysis[39,40]. Even though the dataset is not high-dimensional sample-wise, the extreme subcellular patterns of the SCBC dataset might generate a unique compressed vector for each proteoform entry. Basically, I am making a case for a deep neural network-based representation learning method to abstract meaning from proteomics signals. The use of embeddings as proxies of the original data for downstream analysis have also been introduced in image-based proteomics[41].

The models parameters including hidden and latent dimensions, learning rate, batch size for training with the dataset as well as the regularization terms were optimized based on the Reconstruction error loss and a secondary classification task. Although Reconstruction error loss indicates how informative these latent variables are for data generation, the generative properties of the VAE were of less importance in this approach. I ensured the VAE works more as dimension reduction algorithm by limiting its generative properties and enhancing its compression stages. More specifically, I invested in the optimization of the encoder and harnessing the probabilistic perspective of the model that allows to reduce incomplete data to representative vectors. After I matched the Psim with HUGO genes, I used a logistic regression classifier which predicted Psims pairs that existed within the ground truth (**Sup.Fig.2**) using their Pearson coefficients as features. In **Supp.Data.File**, ten different initializations illustrate the performance of the VAE-based correlation compared to raw data-based correlations. To be more precise, the VAE model with (90,45) dimensions for SCBC and (54-30) for ABMS showed the most promising results. The VAE-based Pearson predictors scored an AUC of 0.78 in comparison with the raw-based, which scored 0.74 for SCBC. In ABMS the difference was within the error margin, both scoring approximately 0.72. Moreover, I also calculated the distances of UMAP 3D reduced latent variables and observed for the SCBC an AUC score of 0.79 (**Supp.Data.File**). The prediction metrics demonstrated that the current model architecture and hyperparameters were optimal to continue. The observed improvement of the AUC metric for SCBC is mostly due to the unique patterns while the lack of improvement for the ABMS dataset could be possibly attributed to low sample-wise dimensionality and high heterogeneity of tissues.

SCBC model's loss metrics for validation and tests sets are displayed for each training epoch in **Fig.4a**. The parameters of the model are optimized as the hold-out test set errors are slightly better than those of the validation set (**Fig.4b**). This is also explained by the regularization of the deep model, which is based on layer dropout. Aside from that, to evaluate if the dimensionality reduction preserved biology, I extracted the latent variables from all proteoforms, matched the Psims with 'ProteinMarkers' of the original SCBC pipeline, applied UMAP and observed whether the markers are clustered in the same way as in the original article[8] (**Fig.4c**). I use the term clustered because I labeled the Psim proteoforms with the cluster memberships of the SCBC pipeline as they are presented in **Fig.2c** and derived from the original publication**.** This confirms that the model abstracted subcellular information because the original clusters were reconstructed, albeit in different shape. For the ABMS, I observed the same tendency in its metrics (data not shown – analysis here[33]), however there wasn't any prior knowledge or any biological validation to be followed for visual inspection of protein clusters. Finally, I proceeded with my raw proteoform tables as well as with new tables of 45 and 30 dimensions per proteoform for SCBC and ABMS datasets, respectively.

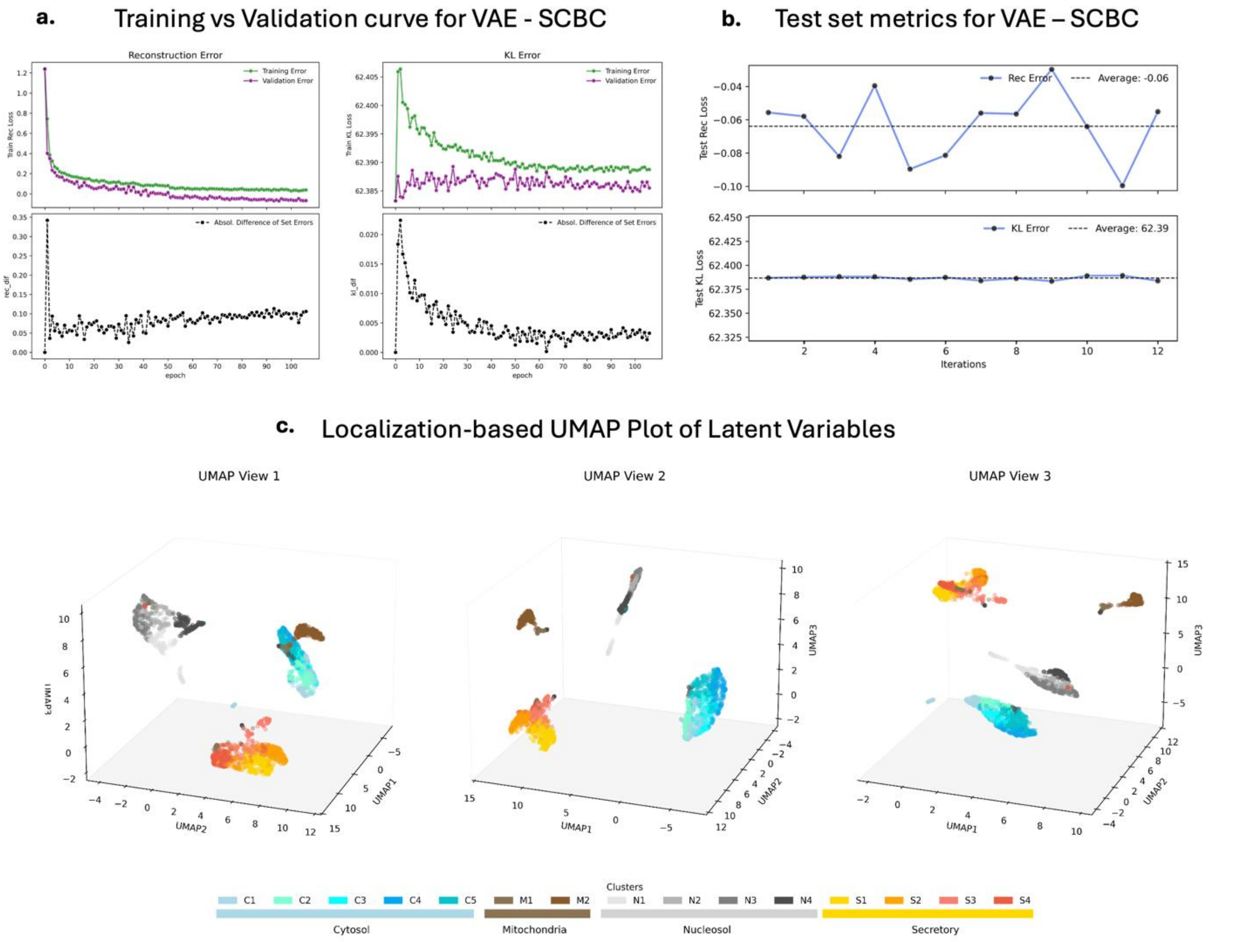


Figure 4 a. VAE-SCBC training vs validation curve. The per-epoch training and validation curves of the reconstruction error loss (left) and Kullback–Leibler (KL) divergence (right) across 100 epochs (1 epoch equals one pass of all the training batches of the dataset). The lower subplot displays the absolute difference between training and validation error losses for each epoch as an estimate of the model's generalization. **b. Test set metrics.** Reconstruction loss (upper subplot) and KL loss (lower subplot) are shown for each mini-batch of the hold-out test set. The average of all the batches is indicated with a horizontal line. **c. UMAP plot of the latent variables of the SCBC data.** The 45 latent variables inferred for every proteoform were further embedded with UMAP into 3 dimensions. The proteoforms Psim were matched to their protein counterparts and were retained if they belonged to the curated Subcellular Protein Marker set of the SCBC pipeline. These proteins have an established compartment and neighborhood membership, and their labels are used as a reference in the 3D scatter plot. There is strong concordance between the a-priori cluster labels and the spatial arrangement of the proteoform using the VAE based embeddings.

5.3 Predicting proteoform interactions and reconstructing association networks

There are several methods available to build protein association networks: using Pearson or any other distance metric to coabundance data; network analysis from co-IP experiments; proceeding with more refined self-supervised vector representations from image data; association inference from genomics[42]. Here, I sought to leverage the information of the MS signals and the derived latent variables by feature engineering practices. In both datasets, I calculated for any proteoform pair correlation coefficients, Euclidean, Cosines and other types of distances of the raw or the

VAE vectors. I tested them with default random forest models and compared the raw data-based with the VAE-based features, and their combinations as well (**Sup.Fig.3**). I realized that the final RF classifiers should have a combination of raw data and VAE embedding-based features. Furthermore, I decided to add the Euclidean distances of the UMAP 3D reduced latent variables since they gave a promising predictive value in multiple seeds during the optimization of the VAE architecture (**Supp.Data.File**).

As illustrated in **Fig.5a** the combined SCBC features RF-model outperforms the model whose predictors are only based on the raw data. The probability threshold corresponding to 5% FDR as well as the results of the importance analysis are also providing justification for the combined classifier (**Fig.5b,c**). This intriguing enhancement of the classifier's performance was not observed in the ABMS dataset, (**Sup.Fig.4a**)**.** This result is consistent with previous observations, if we reflect the earlier results of the ABMS compression, where no differences were observed in the performance of the logistic regression between the raw and the VAE Pearson coefficients (**Supp.Data.File**). This is also corroborated by the feature comparisons, where the VAE based predictors in ABMS had almost the same performance as those from raw measurements (**Sup.Fig.3**). The improvement of the ABMS combined RF-model lies within the margin of the confidence intervals of AUROC of the 10-CV fold average metrics.

In both cases I used the combined RF-models and for validation I plotted the distribution of all the association probabilities within CORUM/COMPLEAT and STRING (**Fig.5d, Sup.Fig.4d**). There is a clear distribution drift for both classifiers validating that high association probabilities indicate more frequently a database curated proteoform association than a non-observed one. The 5% FDR threshold is used as a cutoff to account for anticipated false positives. The networks were represented as binary adjacent matrices in which probabilities above the threshold were designated as one and the rest as zero.

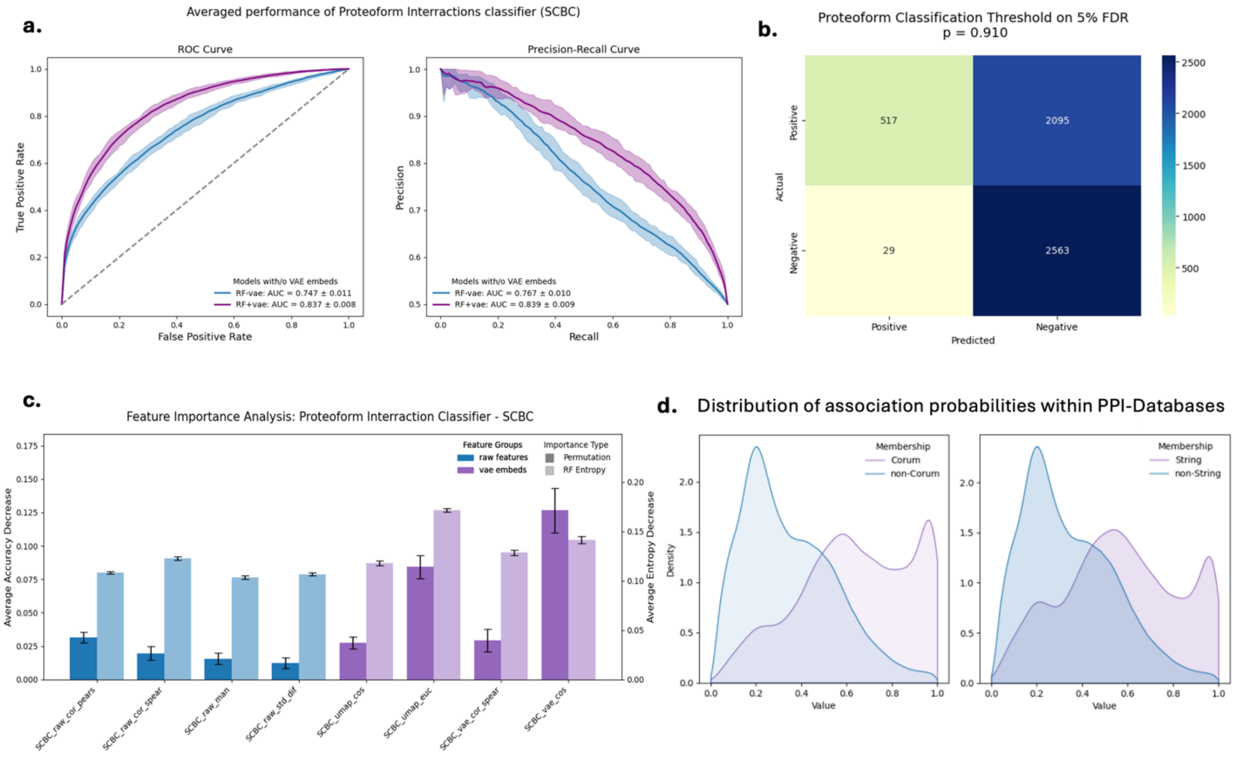


Figure 5: a. Averaged performance of proteoform interaction classifier – SCBC. Averaged ROC and PR curves (solid line) with their 95% CI (shaded regions) for the random forest classifier trained either only in raw-based features (blue) or in both raw- and VAE- based. AUC of each model was averaged over 10-fold stratified cross validation. The VAE model achieved greater performance. ROC curve shows true positive rate expressed as a function of false positive rate, interpolated in all probability thresholds. PR is computed in the same way. **b. Contingency table of the final SCBC classifier.** After 10-fold CV, the classifier has been retrained in 0.8:0.2 train-test split. Applying a 5% FDR threshold to the classifier during the prediction of the held-out test set yields a 0.91 probability threshold for predicting a positive class, a proteoform-proteoform interaction. **c. Feature Importance analysis of the cross-validated SCBC classifier**. After 10-fold CV, the tuned classifier was trained by randomly sampling the dataset 20 times and feature important analysis was performed. I run this analysis 20 times to account random processes introduced in the framework such as the undersampling of the negative class and the initialization of the classifier. The averages of accuracy decrease after permutation of each feature (left y-axis), and entropy loss before and after each tree node split (right y-axis) were calculated and plotted for each feature after the 20 iterations. **d. The distribution of assigned probabilities for all possible proteoform pairs in the SCBC dataset.** The kernel-density estimates compare the association probability scores of all possible pairs that exist within the databases (CORUM/COMPLEAT and STRING high confidence) and either share a membership in the same complex (defined as positives) or not (defined as non-Database pairs).

5.4 Edge pruning and network integration with BIONIC

Before integration, we evaluated the structure of the association networks and measured the number of interactions since we were uncertain if the integration could work with them being so densely connected. Furthermore, in the original implementation, BIONIC was used for networks with maximum 10,000 nodes with around 50,000 edges and an average node degree of 50. These are sparse networks[34]. Also, literature suggests that proteomic networks follow a scale free topology in consequence of the preferential attachment principle. As a scale free network constantly grows by the introduction of new interactors, the already densely connected nodes (e.g. signaling proteins) are favored to establish a connection with the newest additions. Thus, an inhomogeneous network with many low degree nodes and a few central hubs with extreme number of links is the current model of a protein interaction network[43,44]. The classifiers generated: 1) a coabundance proteoform association network with 9,000 nodes and approximately 225,000 edges from ABMS, and 2) a subcellular proteoform interaction network with 12,000 nodes 326,000 edges from SCBC. Both networks averaged node degrees above 200. The networks exceeded by far the requirements of BIONIC and did not meet the expectations of the assumed model of protein networks.

To tackle this challenge, we performed a refinement procedure by pruning the edges with a focus to retain node pairs sharing high number of common interactors calculated by a similarity metric. The similarity metric, i.e. the Jaccard index of every node pair, was plotted against three quantile-based node degree categories in each network (**Sup.Fig.5a**). We observed a positive relationship between highly connected nodes and high similarity metrics. So, during refinement we adjusted the thresholds by the quantiles of the distribution of Jaccard index in all three node categories. Consequently, for each node degree category (low, medium, high) we applied a similarity threshold at 53rd percentile and edges falling below were pruned. Through this process we maintained proteoform pairs of low degree and low count of common interactors after the refinement. Alternative proteoforms could have low degree centralities and still hold biological significance.

By sequentially increasing the percentile we calculated the linear fit of the log-log node degree distribution (**Sup.Fig.5b**). At the 53rd percentile both networks reached the 50 average node degree threshold and showed a suboptimal linear fit for a power law node distribution (**Fig.6a**). Aside from substantial edge reduction, there were also some nodes that were excluded from the network. The exponential fit was slightly better than the linear fit, and this can be attributed to the fact that these are primarily association networks and not physical ones. There is always noise in the predicted interactions, and other strategies of edge prioritization could have been added to the refinement process given more time.

The integration of the two networks was conducted with the BIONIC model using default parameters, and adaptations related to the network size. Post integration, each proteoform either from SCBC or ABMS was assigned a 512-dimensional embedding vector which was reduced to 2 dimensions with the UMAP algorithm. It is not the first time that UMAP coordinates will be used for visual inspection of protein network neighborhoods and distance calculations[45]. I plotted the Euclidean distance distribution of all the nodes of the integrated network using

CORUM/COMPLEAT to assess the frequency of physical associations across the measure (**Fig.6b,c**). According to probability density function (PDF) and cumulative distribution function (CDF) plots, there is a significant distribution drift between CORUM and non-CORUM proteoform pair memberships. At low distances (Euclidean distance < 1) the frequency of true positives is significantly higher than the that of noisy edges, false positives or undiscovered associations. On the other hand, the network has inherited noise given how the association probabilities were distributed in the constituent networks. The distribution differences before integration were far more prominent at the designated FDR thresholds (**Fig.5d, Sup.Fig.4d**).

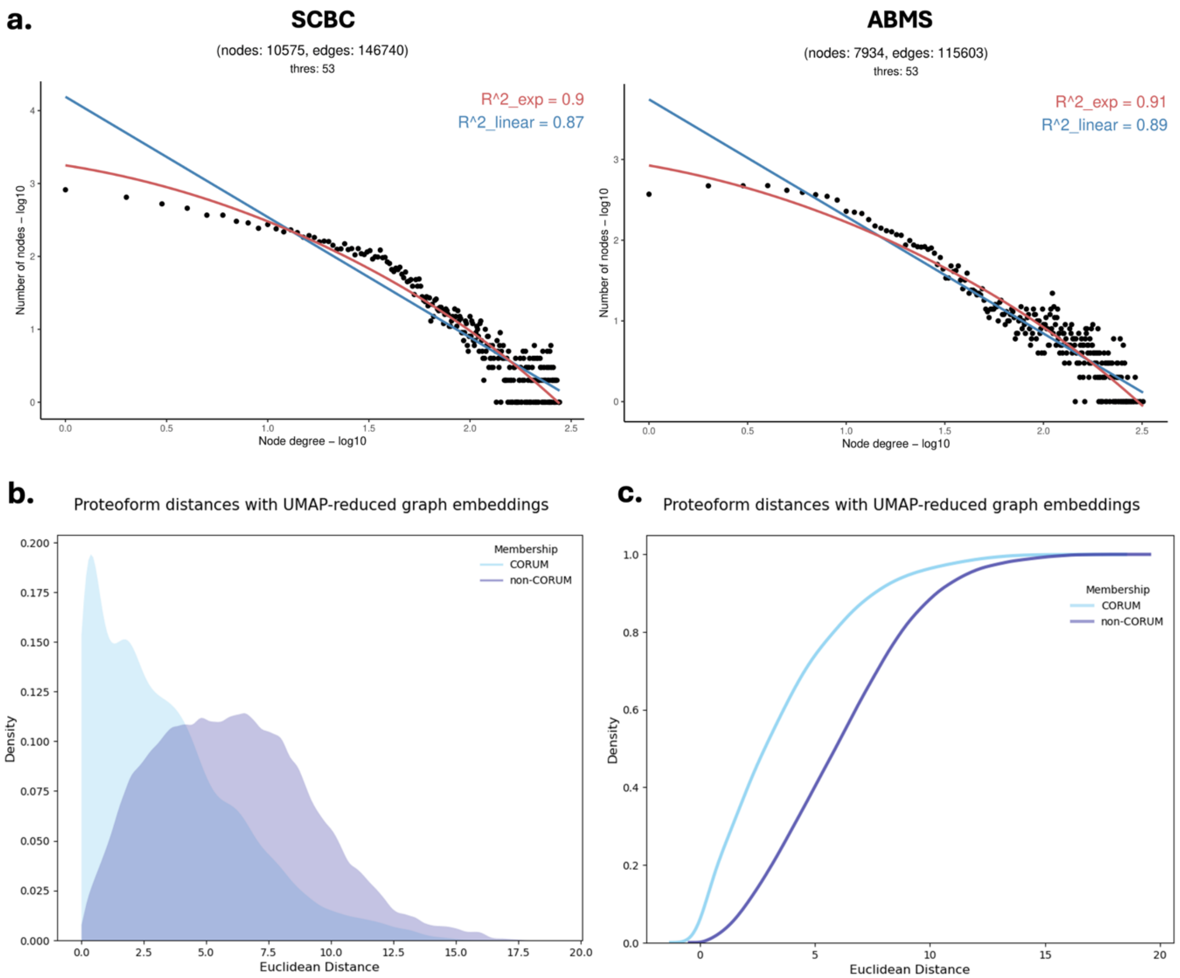


Figure 6: a. Node degree distribution of the pruned association networks before integration. Both networks show near power law fit (blue line indicates a linear fit on log-log scale whereas red is an exponential fit) of proteoform interactions before integration. It is assumed that protein networks follow a scale-free topology, so the global structures were optimized close to this topology during edge pruning. The exponential fit is slightly better due to the inherent noise of the association networks. **b. Distribution of proteoform-proteoform Euclidean distances.** These distances were based on UMAP 2D reduction of the graph embeddings post network integration. Unique CORUM/COMPLEAT pairs were used as ground truth for functional and physical associations between Psim proteoforms with corresponding ensemble id – hugo symbol. **c. CDF of proteoform-proteoform distances.** Cumulative density distribution of within vs non-within CORUM/COMPLEAT interaction partners.

5.5 Deconvolution of mTORC1 components and functional annotation of neighboring nodes.

An interesting case study, where the deconvolution to proteoforms and the integrated network could potentially highlight missing information was the localization profile of two physically interacting partners mTOR-RPTOR, which are components of the mTORC1 complex[46,47]. First, we applied the SCBC pipeline and predicted the localization of all the proteins, including the mTORC1 members across the 13 cell lines of the SCBC dataset. The two proteins are multilocalizing because they belong to both secretory and cytosolic neighborhoods. Consistent with their shared protein complex membership, they do share the same localization profiles across most of the cell lines (**Fig.7a**). The localization prediction of the deconvoluted proteoforms of these proteins indicate an intriguing localization drift between some proteoform siblings. To elaborate, the same pipeline was applied at the proteoform level, and it seems that for both proteins there exists one cytosolic form with high peptide count in all the cell lines (**Fig.7b**). Although it is widely known that the mTOR is located at the cytoplasm and secretory pathway – to the surface of lysosomal vesicles for nutrient sensing and signal integration, cis/trans Golgi related endosomes and luminal compartments– the role of a putative soluble cytosolic form and its effects to mTORC1 has only just recently been stressed[48].

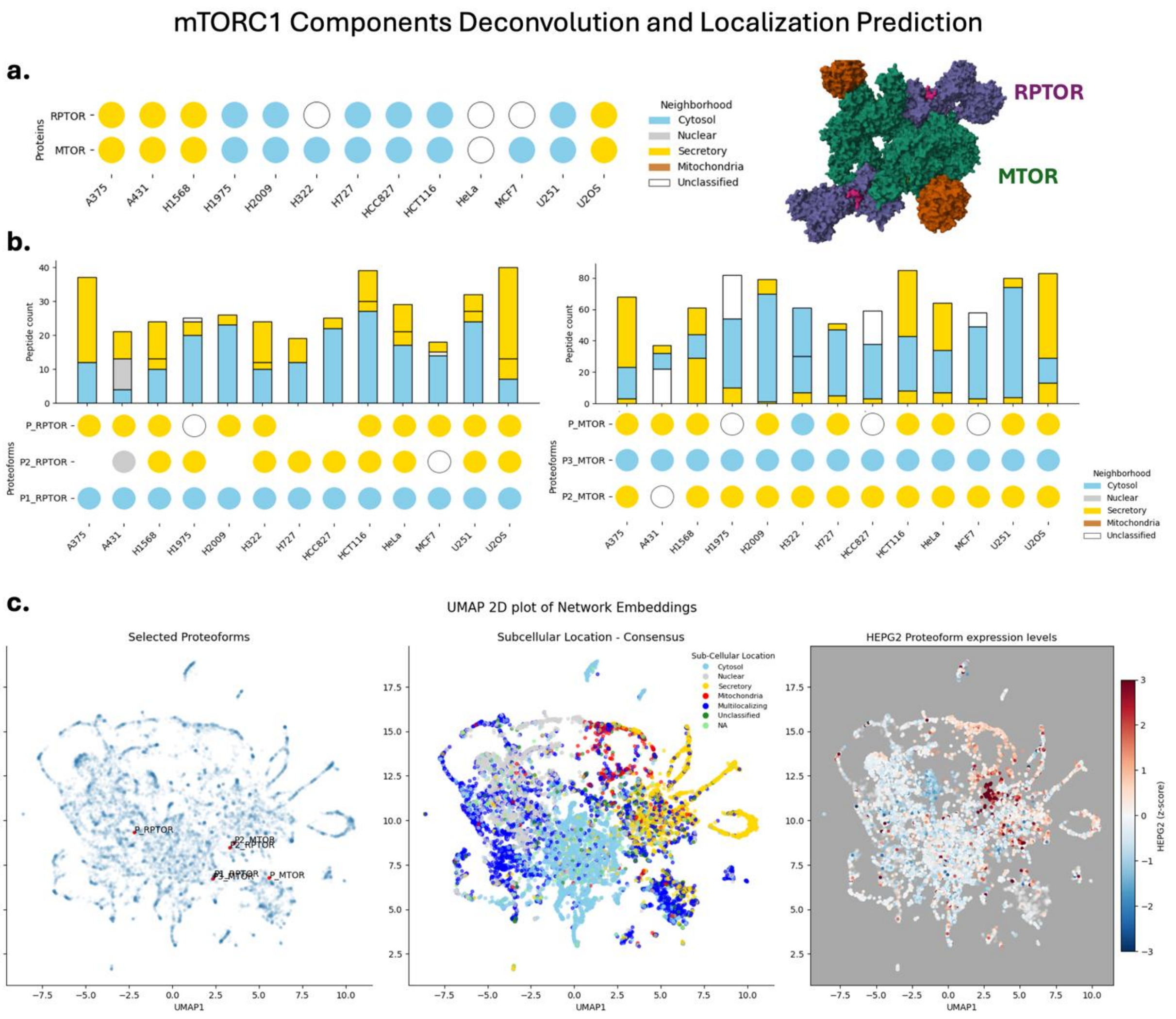


Figure 7: a. Localization Prediction of MTOR-RPTOR proteins with SubCellBarCode. The dot-plots display localization predictions of the SubCellBarCode pipeline for mTOR and RPTOR across all 13 current cell lines. On the right, a model structure for two members of the MTROC1 complex from PDB is illustrated. **b. Deconvolution and localization prediction of the MTOR-RPTOR proteoforms.** After the deconvolution of the peptides into their proteoform species, the SubCellBarCode pipeline was used to predict the localization of all the proteoform species. Furthermore, a stacked bar plot illustrates the peptide count of each proteoform group across all cell lines. **c. UMAP 2D visualization of the integrated network with overlaid localization predictions and expression levels.** UMAP was utilized to reduce the high-dimensional embeddings of each proteoform into two coordinates. Each proteoform is illustrated at the 2D plots where a consensus localization or expression levels, if applicable, is overlaid in the visualization. The spatial incoherence of the mTORC1 proteoforms is apparent, especially in two regions of different subcellular localization where both members of the complex share a notably similar neighborhood. The Psim proteoforms of both proteins are not proximal.

Next, I overlaid localization information to the integrated network and visualized the location of these proteoform species. To summarize the divergent cell-line predictions, I used a consensus rule-based system and overlaid this information to the UMAP plot of the reduced graph embeddings. According to the Consensus rule each proteoform is assigned: 1) a single Neighborhood localization if there is only one unique localization prediction across 13 cell lines, 2) a 'Multilocalizing' status if there are two unique localizations across the 13 cell lines. 3)

‘Unclassified’ status if there are unclassified and 4) NA if there is no available information at the subcellular covariation level. While the guilt-by-association principle based on single edge prediction does have its drawbacks[49], this proximal dual co-occurrence of the delocalized proteoform pairs in the UMAP strengthens the confidence of this finding and renders the alternative explanation of “artifact” origin less likely (**Fig.7c**). The proteoform scaled expression levels of HepG2 are illustrated to highlight the point that even with some inherited noise the integrated network retains the biological assumptions of colocalization and coexpression. Finally, functional enrichment analysis of the top 60 nearest neighbors of the two differentially localized mTOR proteoforms P2 and P3 identified the RPTOR member as well as other STRING curated associates of the protein (**Fig.8**). The secretory P2 proteoform is primarily associated with nutrient sensing, autophagy regulation and the lysosomal signaling hub, whereas the function of the cytosolic counterpart remains more elusive as it is associated with cytosolic COPII vesicles coat polymers, peroxisome enzymes and function that is reminiscent of cell structure organization. It is also worth noting that the Psim of both proteins are in different places in the network (**Fig.7c**). Specifically, mTOR showed relevant functional enrichment with the P2-secretory form (**Sup.Fig.6**).

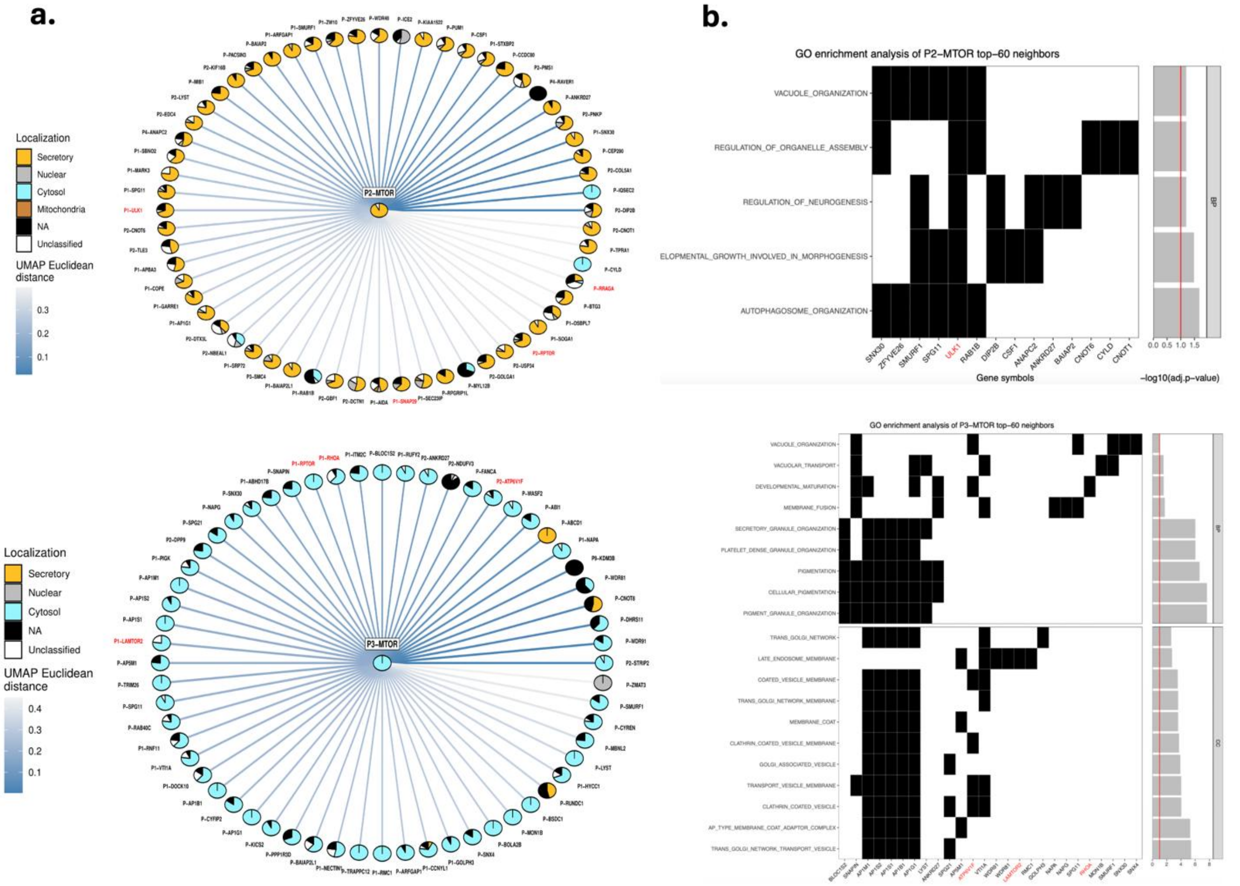


Figure 8: a. Neighborhood extraction of mTOR proteoforms P2 and P3. Star-shaped networks indicating localization for each proteoform and strength of the similarity measure between the proteoform and the neighbor. Proteoforms assigned with red color are STRING curated associates. **b. Functional Enrichment analysis of the neighborhoods.** GO keywords enrichment analysis for Biological Processes – BP and Cellular Components – wherever applicable. Adjusted P-value threshold at 0.01 is highlighted in red.

6. Discussion

This study merges three fundamentally different algorithms with the aim to deconvolve proteins into proteoforms and make deductions about their putative function. Aside from technical aspects, the biological novelty is the convergence of MS-based subcellular and total cell proteomics data in the proteoform identification workflow at the network level. These late omics integration practices are used extensively in single cell genomics, and their effectiveness is still under investigation[50–52]. Here, I set out to harmonize the DepMS proteoform deconvolution algorithm, the SCBC pipeline for subcellular localization prediction and VAE-aided association network reconstruction by producing these graph convolution network embeddings post integration under the premise that they will serve as an 'integrated similarity metric' between proteoforms.

There are two important predicates involving the DepMS in this project. First, we decided to analyze all the 13 cell lines concomitantly to filter out proteoforms generated in the pipeline that were cell type specific. While this limits our findings, it reduces the noise due to the high incompleteness of the data and allows us to study proteoforms whose abundance is highly pronounced across multiple samples and change localization profiles in various cell lines. If we proceeded with a more cell-specific approach, it would be really challenging to integrate several TMT-sets after deconvolution. Secondly, we linked the proteoform to protein complex membership of genes using the Psim group of proteoforms. While high correlation between the Psim and the median summarization of all the peptides was proven here (**Fig.3a)**, this is not always the case and most importantly gene-level summarization in proteomics are based on PSMs and not peptides. These differences in summarization approaches led to minor conflicts when I compared the localizations of the proteins with their proteoform constituents and will be further discussed in mTOR analysis.

The robustness of the SCBC classifier with the VAE features was an interesting result (**Fig.5a**). These algorithms are mostly used for high dimensional data, with columns estimated from thousands to hundreds of thousands for single cells. It was surprising the 150 samples of the SCBC data were enough to generate latent variables with additive value for proteoform interaction prediction. I also believe that the high quality of the subcellular data was a major factor in this. On the other hand, the classification of the ABMS proteoform didn't reap the benefits of compression. Partly because of low dimension, heterogeneity of samples, and obviously the lack of pattern to abstract. From one angle, VAE opens new avenues for integration of multicohort data to generate co-abundance networks; from another angle, I need to account how many samples are required from global proteomics cohorts to leverage the compression algorithm for the discovery of new associations.

The true difference between physical and association networks became evident during the network refinement by edge pruning. Using association probabilities might be efficient for edge comparisons in different scenarios, or even to identify tissue related edges e.g. tissue specific protein interactions[18,19], but globally the association networks do not follow the expected scale free topology. One contributing factor could be the SCBC classifier. Association grounded in colocalization profiles can inflate the number of true interactors, even at the 5% FDR threshold. Concurrently, the multi-origin cell types of the ABMS dataset compounded the inherent noise — a constrain we alleviated through the edge pruning.

The final part of the thesis, the deconvolution of the mTORC1 components, begets the most captivating questions about the co-occurrence of mTOR-RPTOR proteoforms in two different subcellular neighborhoods. It has been recently shown that there is a cytoplasmic pool of mTOR and intact lysosomal activity, upstream regulators such as RAGA/B kinases and amino acid supplementation convert a fraction of that to lysosomes, where mTORC1 exerts its signaling function[48]. These are consistent with the enrichments of the secretory proteoforms of mTOR-RPTOR (**Fig.8, Sup.Fig.6**). It has been suggested that there are also RAG independent mechanisms behind the allocation of mTORC1 monomers in the cytoplasm and they are not strictly separated with the secretory form[48]. Hence our signals in both neighborhoods across all the cell lines are consistent with current research. The high number of peptides for both proteins and their proteoforms stands in favor of this finding. From the enrichment analysis of the cytosolic neighborhood, it is apparent that the proteoforms do not follow a specific function or cellular location. Soluble proteins that act as polymers, enzymes and membrane coating proteins in the cytosolic cluster indicate that the two proteoform co-localizations in the networks are not necessarily distinct functional entities but more likely temporal states of the overall mTORC1 pool. Obviously, the inherited fuzziness of the integrated network makes the interpretation here more elusive. Setting that aside, the dynamics between these proteoforms are not the same across samples and in some cell lines the secretory forms are more pronounced for both proteins, their proteoforms and in the same direction (**Fig.7b**). The peptide count is mostly in agreement with PSM-based gene level summarization to the extent that the localization prediction of the proteins can be explained by the kinetics of the proteoforms. This case is not always true though. Some cell line predictions at the protein level are not in agreement with the proteoforms (MCF7 and A431 for mTOR and the HeLa unclassified). This could be possibly the product of comparing two different summarization approaches, since DepMS is peptide based whereas protein-level summarization works at the PSM level. Providing the PSM counts for the proteoform constituent peptides could potentially indicate the origin of these discrepancies.

My conclusion is that across tissues the regulation of the mTORC1 pool is different, which builds a premise for perturbation analysis of the tissues that exhibit elevated levels of the different forms. For example, autophagy induction by starvation protocols to cell types with a highly expressed cytosolic form could demonstrate the conversion of that proteoform to the secretory sibling and validate the pool phenotype at the proteoform layer. Finally, the integrated network approach shows great potential for the identification of new proteoform species, and after more refinement, functional enrichment via guilt-by-association could highlight new knowledge in cancer biology.

7. Current Limitations and Future work

Although we were very vigilant with the network's inherent noise, it is unequivocal that the pan-cancer 'all in' approach has certain limitations. We included tissues of diverse origins, from lung cancer to glioblastoma and suspension cells. Protein covariation models for functional association prediction may not be robust to extremely different developmental lineages as many protein complexes follow tissue specific origins. Recent literature suggests that around 25% of protein complexes are compliant with the stable stoichiometry principle, given they are analyzed at certain groups of tissues[19]. In other words, protein association networks are not generalized for all cell types and biologically this makes sense. Of course, this builds the argument in favor of a second version of this pipeline following the path towards a cancer-specific proteoform network and poses a question about the source of variation we aim to model. Before delving into this, I would like to highlight how tissue specificity might limit some discoveries. For example, the mTOR protein exhibited differential localization across 13 cell types, and by examining these covariation differences we managed to identify the proteoform constituents in different locations because the profiles of a group of tissues guided this. In a specific cell type, given that the proteoform discovery is valid, the ratio of these proteoforms and the localization of the protein may be stable across a cohort. So, the question of the signal covariation from tissue biology-driven becomes clinical-driven and sometimes these two objectives are incongruent. That protein may have a distinct profile in a tissue such as lung, so its proteoform may not fluctuate at that specific cohort. However, there is a suspicion that even at the cohort level some proteoforms are concealing significant hits that classic hypothesis driven approaches are currently missing[6].

In view of the above, the cancer specific network should model the granularity of a clinical cohort using the SCBC dataset as a proxy to make it localization informed. Co-abundance profiles of the same tissue will reduce the noisy edges, and we can further attenuate false positive signals by including tissue specific markers from single cell atlases. More and more information is available about which markers and proteins are prevalent in each tissue. We could preserve putative interactions between proteoforms after the classification by grouping tissues whose origins are very close to the cancer we study. If we aim at lung cancer, instead of limiting and pruning the network, we can guide discovery by retaining putative edges of the lung network that exist in closely related mesodermic and endodermic tissues. At a more technical aspect, we can add some prior knowledge to the network integration step by using a third co-IP or database curated interaction network. The BIONIC model could prioritize more true positive edges in this way. Proteomics data becomes more available and clinical cohorts are increasing, and so does the intra-cohort variation between the MS signals due to batch effects and instrumentation. The compression algorithm could be used for high dimensional proteomics data to abstract representation vectors of various datasets of the same cancer. This can be achieved at any covariation level of analysis, be that proteoform deconvolution or proteoform association prediction. Finally, an essential part of the analysis that we pondered on is the UMAP algorithm. We used the Euclidean distances of the two UMAP coordinates derived from the graph embeddings. We need to evaluate this step by adding initializations of the algorithm at different hyperparameters and measure the shortest path between interactors of our interest to assess the effect that different settings might have on the proximal neighbors. Looking forward, there is potential of the integrated approach to elucidate novel findings if we balance the range of discovery and the specificity of the biological network.

8. Ethical Reflection

Raw data and analysis are available as the laboratory follows open-source policy. Samples were based on in-vitro experiments, so no animal or clinical samples were involved in this work. All code for the analysis is available at GitHub repositories.

9. Acknowledgements

First, I want to express my gratitude to Professor Janne Lehtiö for approving this project and opening the doors of his lab to me. Professor's Janne group is my first exposure in clinical proteomics, and I will never forget what I learned from there.

It has been a privilege and quite astounding to work with Ioannis Siavelis and Yanbo Pan simultaneously. There are not many memes enough to express the intensity of the last 3 months. Ioannis, I enjoyed our hardships, our wins and I wish for more.

Yanbo, personally I also want to thank you for introducing me to the Lehtiö Group back in the summer when we worked together, and I hope to see your SCBC2 up and running in no time.

I want to thank Professor Carsten Daub for the review and the constructive criticism of my report. It was a very interesting meeting. I would also like to mention the amazing discussions with Harris, Isabelle, Nora, and Luca. Isabelle's work on proteoforms was inspiring.

Vasilina Tsopela, Konstantinos Sideris, Jia, and Hilmar, thank you for being there to listen, even though sometimes you did not understand me – you were always patient.

I devote this work to my parents and my brother, who are always there, giving me immeasurable strength, never asking for anything in return.

"Untroubled, scornful, outrageous - that is how wisdom wants us to be: she is a woman and never loves anyone but a warrior."
— Friedrich Nietzsche, Thus Spoke Zarathustra

10. References

1. Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. *Nat Biotechnol*. 2010;28(7):659-664. doi:10.1038/nbt0710-659

2. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*. 2016;537(7620):347-355. doi:10.1038/nature19949

3. Wu G, Wan X, Xu B. A new estimation of protein-level false discovery rate. *BMC Genomics*. 2018;19(6):567. doi:10.1186/s12864-018-4923-3

4. Shuken SR. An Introduction to Mass Spectrometry-Based Proteomics. *J Proteome Res*. 2023;22(7):2151-2171. doi:10.1021/acs.jproteome.2c00838

5. Jiang Y, Rex DAB, Schuster D, et al. Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry. *ACS Meas Sci Au*. 2024;4(4):338-417. doi:10.1021/acsmeasuresciau.3c00068

6. Plubell DL, Käll L, Webb-Robertson BJ, et al. Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? *J Proteome Res*. 2022;21(4):891-898. doi:10.1021/acs.jproteome.1c00894

7. Stratmann S, Vesterlund M, Umer HM, et al. Proteogenomic analysis of acute myeloid leukemia associates relapsed disease with reprogrammed energy metabolism both in adults and children. *Leukemia*. 2023;37(3):550-559. doi:10.1038/s41375-022-01796-7

8. Orre LM, Vesterlund M, Pan Y, et al. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol Cell*. 2019;73(1):166-182.e7. doi:10.1016/j.molcel.2018.11.035

9. Branca RMM, Orre LM, Johansson HJ, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods*. 2014;11(1):59-62. doi:10.1038/nmeth.2732

10. Lehtiö J, Arslan T, Siavelis I, et al. Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms. *Nat Cancer*. 2021;2(11):1224-1242. doi:10.1038/s43018-021-00259-9

11. Benzer S. ON THE TOPOLOGY OF THE GENETIC FINE STRUCTURE. *Proc Natl Acad Sci*. 1959;45(11):1607-1620. doi:10.1073/pnas.45.11.1607

12. Aebersold R, Agar JN, Amster IJ, et al. How many human proteoforms are there? *Nat Chem Biol*. 2018;14(3):206-214. doi:10.1038/nchembio.2576

13. Kurzawa N, Leo IR, Stahl M, et al. Deep thermal profiling for detection of functional proteoform groups. *Nat Chem Biol*. 2023;19(8):962-971. doi:10.1038/s41589-023-01284-8

14. Leo IR, Kunold E, Audrey A, et al. Functional proteoform group deconvolution reveals a broader spectrum of ibrutinib off-targets. *Nat Commun*. 2025;16(1):1948. doi:10.1038/s41467-024-54654-8

15. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. *Science*. 2017;356(6340):eaal3321. doi:10.1126/science.aal3321

16. Buljan M, Banaei-Esfahani A, Blattmann P, et al. A computational framework for the inference of protein complex remodeling from whole-proteome measurements. *Nat Methods*. 2023;20(10):1523-1529. doi:10.1038/s41592-023-02011-w

17. Bludau I, Frank M, Dörig C, et al. Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat Commun*. 2021;12:3810. doi:10.1038/s41467-021-24030-x

18. Ryan CJ, Kennedy S, Bajrami I, Matallanas D, Lord CJ. A Compendium of Co-regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. *Cell Syst*. 2017;5(4):399-409.e5. doi:10.1016/j.cels.2017.09.011

19. Laman Trip DS, van Oostrum M, Memon D, et al. A tissue-specific atlas of protein–protein associations enables prioritization of candidate disease genes. *Nat Biotechnol*. Published online May 2, 2025:1-14. doi:10.1038/s41587-025-02659-z

20. R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R project.org/>.

21. Tsitsiridis G, Steinkamp R, Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Res*. 2023;51(D1):D539-D545. doi:10.1093/nar/gkac1015

22. Vinayagam A, Hu Y, Kulkarni M, et al. Protein Complex–Based Analysis Framework for High-Throughput Data Sets. *Sci Signal*. 2013;6(264):rs5-rs5. doi:10.1126/scisignal.2003629

23. Szklarczyk D, Nastou K, Koutrouli M, et al. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res*. 2025;53(D1):D730-D737. doi:10.1093/nar/gkae1113

24. Socciarelli F. *At the Intersection of Proteomics and Pathology : Application of Mass Spectrometry-Based Protein Quantification to Histopathology and Antibody Validation*. Inst för onkologi-patologi / Dept of Oncology-Pathology; 2021. Accessed October 15, 2023. http://openarchive.ki.se/xmlui/handle/10616/47821

25. Arslan T, Pan Y, Mermelekas G, Vesterlund M, Orre LM, Lehtiö J. SubCellBarCode: integrated workflow for robust spatial proteomics by mass spectrometry. *Nat Protoc*. 2022;17(8):1832-1867. doi:10.1038/s41596-022-00699-2

26. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008

27. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9(1):5233. doi:10.1038/s41598-019-41695-z

28. ioasia/DEpMS: DEpMS analysis. Accessed May 19, 2025. https://github.com/ioasia/DEpMS

29. Kingma DP, Welling M. An Introduction to Variational Autoencoders. *Found Trends® Mach Learn*. 2019;12(4):307-392. doi:10.1561/2200000056

30. Rezende DJ, Viola F. Taming VAEs. Published online October 1, 2018. doi:10.48550/arXiv.1810.00597

31. Yu R. A Tutorial on VAEs: From Bayes' Rule to Lossless Compression. Published online June 30, 2020. doi:10.48550/arXiv.2006.10273

32. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. Published online April 24, 2018. doi:10.48550/arXiv.1803.09820

33. Panos G. panos-gbio/Proteomics-in-Latent-Space. Published online April 28, 2025. Accessed May 14, 2025. https://github.com/panos-gbio/Proteomics-in-Latent-Space

34. Forster DT, Li SC, Yashiroda Y, et al. BIONIC: biological network integration using convolutions. *Nat Methods*. 2022;19(10):1250-1261. doi:10.1038/s41592-022-01616-x

35. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Published online September 18, 2020. doi:10.48550/arXiv.1802.03426

36. Gene Ontology Resource. Gene Ontology Resource. Accessed May 19, 2025. http://geneontology.org/

37. clusterProfiler. Bioconductor. Accessed May 19, 2025. http://bioconductor.org/packages/clusterProfiler/

38. Wu Z, Liao Q, Liu B. A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. *Brief Bioinform*. 2020;21(5):1531-1548. doi:10.1093/bib/bbz085

39. Webel H, Niu L, Nielsen AB, et al. Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. *Nat Commun*. 2024;15(1):5405. doi:10.1038/s41467-024-48711-5

40. Koutrouli M, Líndez PP, Nastou K, et al. FAVA: High-quality functional association networks inferred from scRNA-seq and proteomics data. Published online August 15, 2023:2022.07.06.499022. doi:10.1101/2022.07.06.499022

41. Cesnik A, Schaffer LV, Gaur I, Jain M, Ideker T, Lundberg E. Mapping the Multiscale Proteomic Organization of Cellular and Disease Phenotypes. *Annu Rev Biomed Data Sci*. 2024;7(1):369-389. doi:10.1146/annurev-biodatasci-102423-113534

42. Ding H, Douglass EF, Sonabend AM, et al. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat Commun*. 2018;9(1):1471. doi:10.1038/s41467-018-03843-3

43. Wuchty S, Ravasz E, Barabási AL. The Architecture of Biological Networks. In: Deisboeck TS, Kresh JY, eds. *Complex Systems Science in Biomedicine*. Topics in Biomedical Engineering International Book Series. Springer US; 2006:165-181. doi:10.1007/978-0-387-33532-2_5

44. Michaelis AC, Brunner AD, Zwiebel M, et al. The social and structural architecture of the yeast protein interactome. *Nature*. Published online November 15, 2023. doi:10.1038/s41586-023-06739-5

45. Dorrity MW, Saunders LM, Queitsch C, Fields S, Trapnell C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat Commun*. 2020;11(1):1537. doi:10.1038/s41467-020-15351-4

46. Kim J, Guan KL. mTOR as a central hub of nutrient signalling and cell growth. *Nat Cell Biol*. 2019;21(1):63-71. doi:10.1038/s41556-018-0205-1

47. Sabatini DM. Twenty-five years of mTOR: Uncovering the link from nutrients to growth. *Proc Natl Acad Sci*. 2017;114(45):11818-11825. doi:10.1073/pnas.1716173114

48. Fernandes SA, Angelidaki DD, Nüchel J, et al. Spatial and functional separation of mTORC1 signalling in response to different amino acid sources. *Nat Cell Biol*. 2024;26(11):1918-1933. doi:10.1038/s41556-024-01523-7

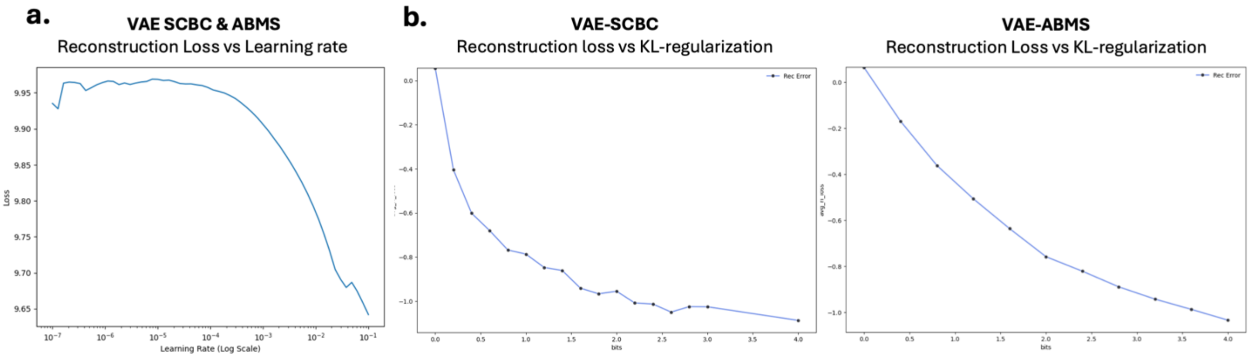
49. Gillis J, Pavlidis P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. *PLoS Comput Biol*. 2012;8(3):e1002444. doi:10.1371/journal.pcbi.1002444

50. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinforma Biol Insights*. 2020;14:1177932219899051. doi:10.1177/1177932219899051

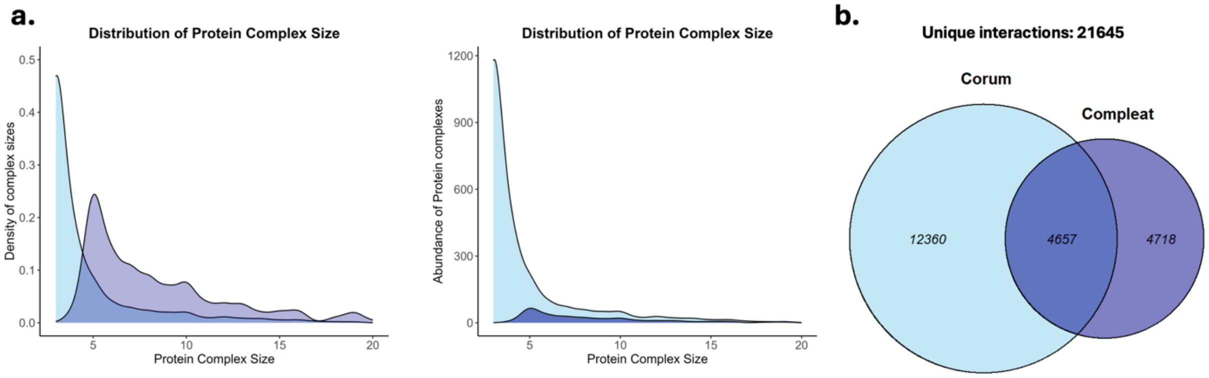
51. Integrating single-cell multi-omics and prior biological knowledge for a functional characterization of the immune system | Nature Immunology. Accessed May 30, 2025. https://www.nature.com/articles/s41590-024-01768-2

52. Dugourd A, Kuppe C, Sciacovelli M, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol*. 2021;17(1):e9730. doi:10.15252/msb.20209730

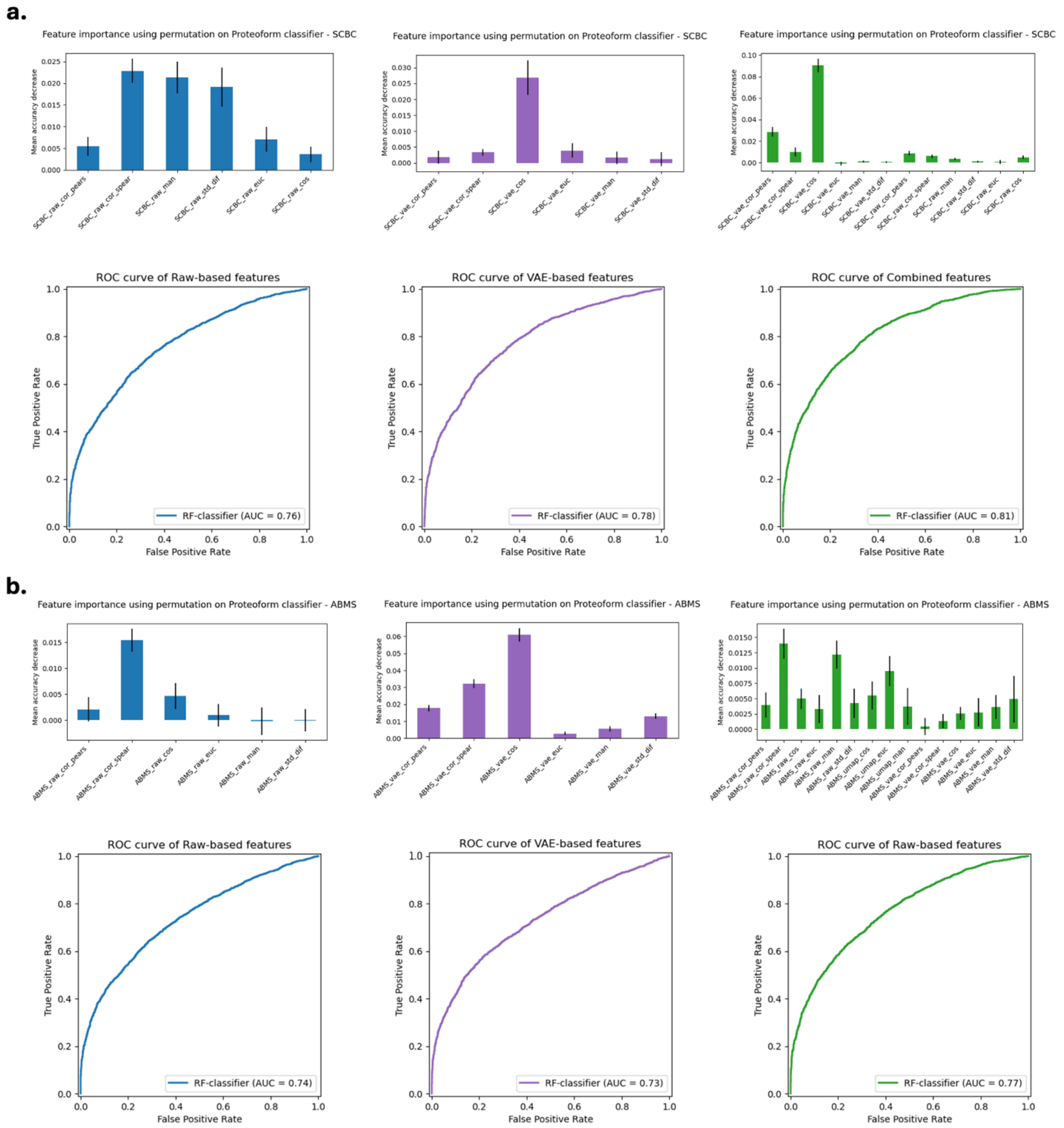
11. Supplementary Figures



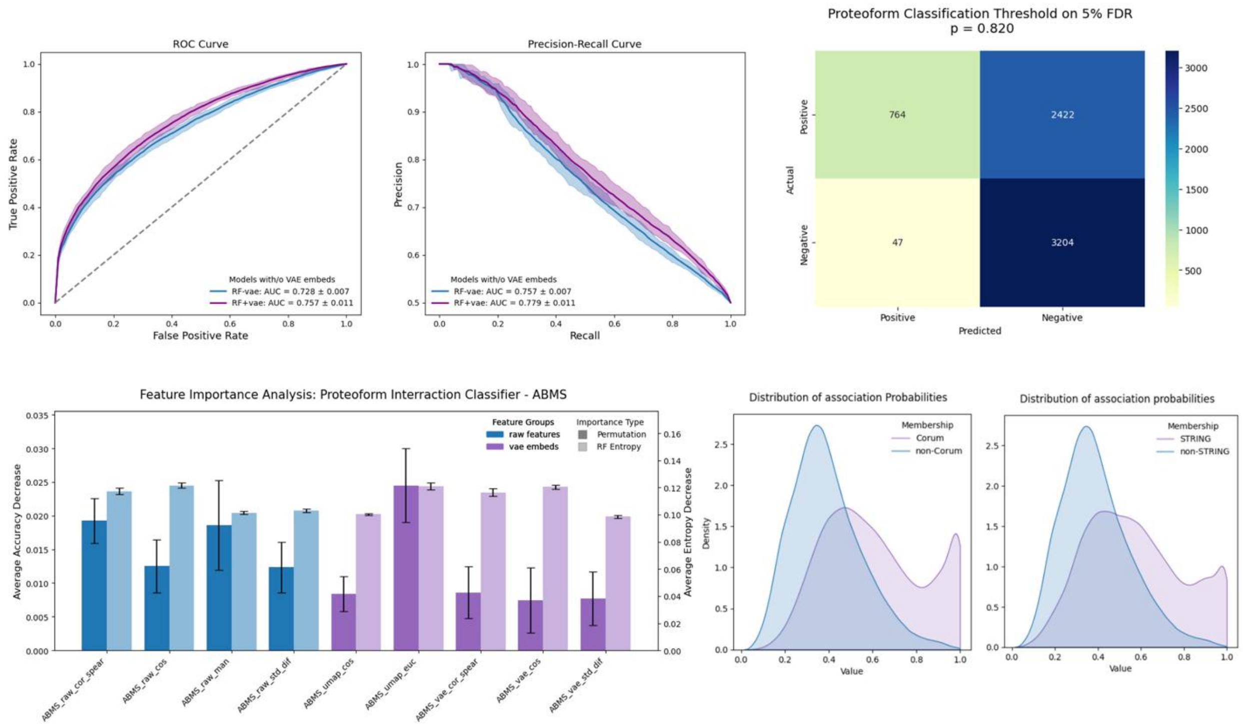
Sup.Figure 1 a. Optimization of Learning Rate for the VAE training: The VAE is initialized with a very low predefined learning rate and after each batch finishes the training loss is computed. In each new batch run, the learning rate increases, and the error is computed again. After multiple cycles of this approach the learning rate plateaus at 0.1. According to current literature, an optimal learning rate for training neural networks is the point of steepest descent, the point where loss shows the highest decreasing trend, approximately 0.0025 – 0.005 in this case. **b. Optimization of KL-Divergence regularization term:** The VAE model is initialized and the whole dataset is passed through the network with a different regularization value (bit) each time. Then the hold-out test set reconstruction loss is calculated. A plateau is visible for the SCBC dataset at bit equal to 2 and around 4 for the ABMS*.*



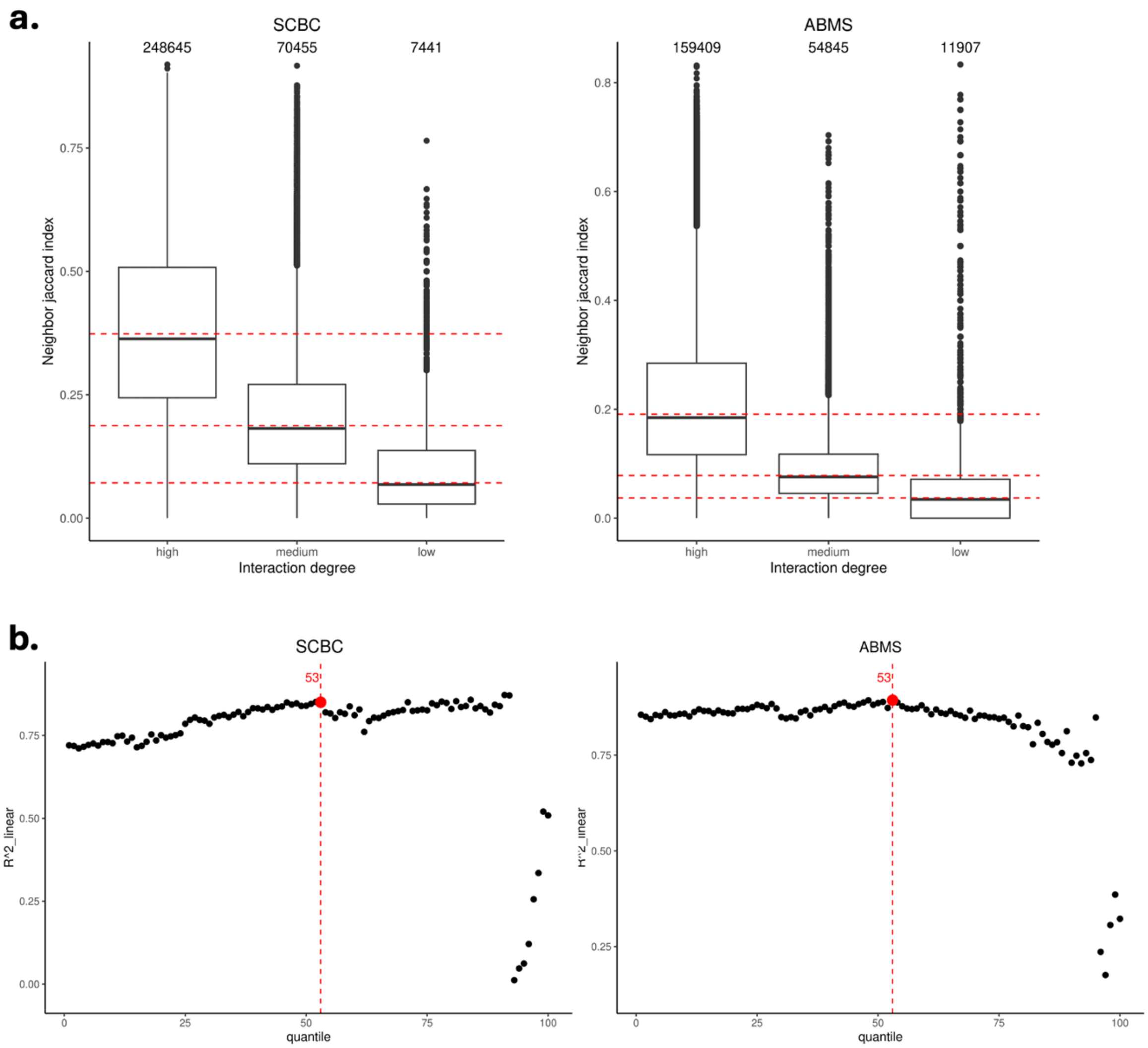
Sup.Figure 2 Database Comparison for Ground truth labeling: a. Distribution of protein complex size: Density plots illustrating the distribution of protein complex size in CORUM and COMPLEAT databases. The plot on the left demonstrates the total number of complexes in each size class while the plot on the right shows the relative frequency of each complex size in both databases. **b. Overlap of pairwise protein interactions on CORUM and COMPLEAT:** Venn Diagram highlighting the number of unique and common interacting protein pairs between the two protein complex databases, after correcting for redundancy and protein name degeneracy.



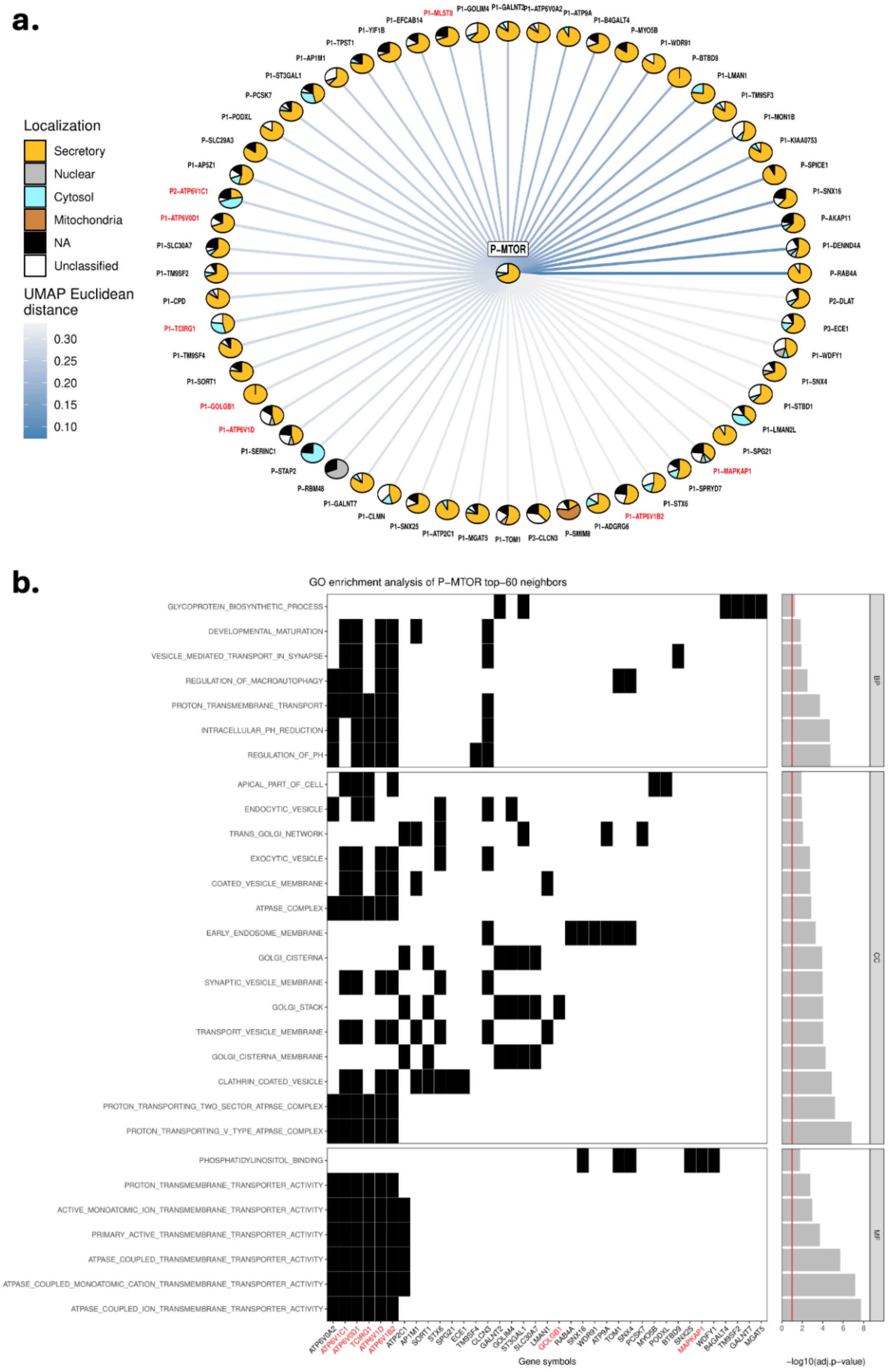
Sup.Figure 3 **a. Raw and VAE based Classifiers of the SCBC dataset.** Random forest classifiers were trained with i) raw-based ii) VAE-based features, and iii) all common features. Feature importance analysis was carried out for every classifier. Black lines indicate the range of values in 10 iterations. The purpose was to compare the predictive value of features of different origins and decide which to include in the final model. The UMAP-embeddings of the VAE-based features were excluded from the comparison since the raw data is incomplete and hence not amenable to dimension reduction. **b. Raw and VAE based Classifiers of the ABMS dataset.** I followed the same approach for this dataset too.



Sup.Figure 4 a. Averaged performance of proteoform interaction classifier – ABMS. Averaged ROC and PR curves (solid line) with their 95% CI (shaded regions) for the random forest classifier trained either only in raw-based features (blue) or in both raw- and VAE- based. AUC of each model was averaged over 10-fold stratified cross validation and the VAE model achieved greater performance. ROC curve shows the true positive rate expressed as a function of false positive rate, interpolated in all probability thresholds. PR curve metrics were computed in the same way. **b. Contingency table of the final ABMS classifier.** After 10-fold CV, the classifier has been retrained in 0.8:0.2 train-test split. Applying a 5% FDR threshold to the classifier during the prediction of the held-out test set yields a 0.81 probability threshold for predicting a positive class, a proteoform-proteoform interaction. **c. Feature Importance analysis of the cross-validated ABMS classifier**. After 10-fold CV, the tuned classifier was trained by randomly sampling the dataset 20 times and feature important analysis was performed. I run this analysis 20 times to account random processes introduced in the framework such as the undersampling of the negative class and the initialization of the classifier. The averages of accuracy decrease, and entropy loss were calculated and plotted for each feature after the 20 iterations. **d. The distribution of all the assigned probabilities for all possible proteoform pairs in the ABMS dataset.** The kernel-density estimates compare the association probability scores of all possible pairs that exist within the databases (CORUM/COMPLEAT and STRING) and either share a membership in the same complex (defined as positives) or not (defined as non-Database pairs).



Sup.Figure 5 a. Box plots of Jaccard index distribution of grouped edges by node degree class. The Jaccard index measures the fraction of shared neighbors between the nodes of an edge. In the different networks, we grouped the node pairs by their degree in three different categories (High, medium, low) and we plotted their Jaccard indices. A quantile-based threshold for network pruning was applied to different edge category. Red dashed horizontal lines indicate the pruning threshold for each edge class. Edges with Jaccard index below the threshold were discarded. **b. Plots of linear goodness of fit and retained edge quantile.** For each network, a quantile threshold of Jaccard indices between edge categories is applied, above of which the edges are retained. A local maximum is observed at the 53% quantile, which is adopted as a threshold for edge pruning.



Sup.Figure 6 a. Neighborhood extraction of mTOR Psim proteoform. Star-shaped networks indicating localization for each proteoform and strength of the similarity measure between the proteoform and the neighbor. Proteoforms assigned with red color are STRING curated associates. **b. Functional Enrichment analysis of the Psim neighborhoods.** GO keywords enrichment analysis for Biological processes (BP), Cellular Components (CC), and Molecular Functions (MF) – wherever applicable. Adjusted P-value threshold is highlighted in red.