FAVA: High-quality functional association networks inferred from scRNA-seq and proteomics data

Mikaela Koutrouli[1], Pau Piera Líndez[1], Robbin Bouwmeester[2,3], Simon Rasmussen[1], Lennart Martens[2,3,*] and Lars Juhl Jensen[1,*]

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[2] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[3] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

* To whom correspondence should be addressed.

Email: lars.juhl.jensen@cpr.ku.dk

Email: lennart.martens@ugent.be

Abstract

Protein networks are commonly used for gaining insight into how proteins interact. Unfortunately, they typically have few interactions for understudied proteins. To address this bias, unbiased network inference methods are needed. We introduce a novel method, FAVA, which infers networks from high-dimensional omics data with improved accuracy compared to state-of-the-art methods, showing a minimum of ~4 times higher accuracy on real datasets and ~1.6 times on simulated datasets. Furthermore, FAVA is capable of processing large atlases with over 0.5 million conditions, leading to the prediction of 4,210 interactions between 1,039 understudied proteins.

Keywords

Co-expression, GCN, biological networks, understudied proteins, deep learning, single-cells RNA sequencing, proteomics, variational autoencoders

Background

Networks of physical and functional interactions among genes/proteins are widely used to understand the inner workings of cells and to visualize results from omics data, e.g., as obtained from transcriptomics and proteomics experiments. Unfortunately, most research is focused on the same 10% of human protein-coding genes (1–3). Thus, networks derived from the biomedical literature — whether through manual annotation or through automatic text mining — are heavily biased by this skewed availability of data. Networks obtained from databases such as STRING (4) consequently have many interactions for well-studied proteins. However, there are only very few interactions for understudied proteins, which are arguably the most interesting targets for network-based function prediction (3,5).

Recent publications highlight the need for new information about understudied proteins (5,6). These proteins have been largely ignored due to the Matthew effect in which proteins with proven disease importance are the subject of most studies (5). To create networks that also provide interactions for understudied proteins, one must focus on systematic high-throughput data, as these are inherently unaffected by literature bias.

One approach to address this is to use experimental approaches to systematically capture new information on, for example, physical protein interactions of understudied proteins. The hu.MAP 2.0 resource (7) is an example of this, providing protein complexes derived from over 15,000 proteomic experiments. By capturing physical interactions between proteins in a systematic manner, without literature bias, it provides protein interaction information for many completely uncharacterized proteins.

Complementary to this, technological advances such as RNA and mass spectrometry-based proteomics have enabled the discovery of protein functions and protein–protein relationships with greater accuracy on a large scale. Co-expression network analysis is one popular approach to this; it works by linking genes that show similar expression patterns across many samples (8,9). Many computational methods have been developed for Gene Co-expression Network (GCN) construction from bulk RNA-seq data (10), such as WGCNA (11). However, other types of data with higher resolution, such as single-cell RNA-seq (scRNA-seq) data, may allow better GCNs to be constructed. ScRNA-seq provides unbiased data on gene expression at the level of individual cells, thus

capturing differences between both cell types and cell states. However, methods optimized for bulk RNA-seq may not be successful in this technology due to its sparse and redundant nature.

As high-dimensional, sparse omics data are becoming increasingly common, in part due to single-cell technologies, new methods designed to handle such data are needed. Pearson correlation coefficient (PCC) is an often-utilized technique to establish GCNs. Recently, more sophisticated computational methods have been developed for GCN construction specifically from scRNA-seq data. Some of the state-of-the-art methods are the hdWGCNA (12) and the scLink (13). hdWGCNA makes use of the WGCNA (11) protocol to construct gene networks and to identify modules with specific provisions for high-dimensionality datasets. To enable this process in scRNA-seq data, hdWGCNA identifies similar cells, which are pooled, thereby dealing with sparsity. scLink (13) manages sparsity by not considering all samples for every pair of genes, but instead focusing on the cells in which both genes are accurately measured. Additionally, the authors recommend creating GCNs with scLink using only the top 500 most highly expressed genes.

While single-cell RNA expression data provide great detail, the correlation between RNA and protein levels is far from perfect (14). It thus makes sense to consider also using proteomics data for creating GCNs, although proteomics data at single-cell resolution currently remains a rarity (15). ScRNA-seq and proteomics can thus be viewed as two complementary types of data that can be used as starting points for predicting functional interactions, also for understudied proteins. Unfortunately, there are no computational methods designed to handle both types of data, taking advantage of their complementary nature. The biggest challenge remains that both scRNA-seq and proteomics datasets are very sparse (i.e. many transcripts/proteins are not observed in each cell/sample) and have high redundancy (i.e., many similar samples/cells are analyzed), both of which present problems for most analysis methods.

Dimensionality reduction can help address both sparsity and redundancy in data. By compressing the information into a lower dimensional space, sparsity is eliminated by combining data from across multiple cells/samples. Redundancy is also inherently reduced, because data compression is specifically achieved by not storing the same information multiple times (16). One can thus expect that applying dimensionality reduction to scRNA-seq and proteomics data provides a latent representation that is better suited for co-expression-based prediction of interactions than the original data matrices.

Dozens of dimensionality-reduction algorithms exist, which can be used to produce this latent space. These span linear methods such as truncated singular value decomposition (17,18), nonlinear transformations such as Uniform Manifold Approximation and Projection (UMAP) (19), and deep generative models such as variational autoencoders (VAEs) (20). VAEs use two deep neural networks — an encoder and a decoder (as shown in **Figure 1**) — to learn a representation from complex data without supervision (21). The encoder learns the input data and projects it into a normally distributed latent representation, parameterized by the mu (μ) and sigma (σ) layers, while the decoder attempts to reconstruct the input data from the instances sampled from the latent representation distributions. The encoder–decoder networks are simultaneously optimized to reconstruct the input data and regularize the latent representations. This ensures that the constructed latent representations follow Gaussian distributions, making them easier to interpret than those of other autoencoders, such as the tones described in (22). Unlike, e.g., singular value decomposition, VAEs can capture both linear and nonlinear relationships between the

cells/samples in scRNA-seq and proteomics data. As a result, VAEs have become popular in the field of expression data (e.g., for normalization and visualization of scRNA-seq).

Here, we present FAVA, a method to infer Functional Associations using Variational Autoencoders from omics data. We show that i) the method can efficiently handle large-scale scRNA-seq and proteomics data; ii) it outperforms traditional co-expression analysis and the state-of-the-art methods by predicting approximately 1.6 times more accurate associations in simulated data and 3.9 times in real data; iii) it can predict high-confidence functional associations, and iv) the resulting networks provide good coverage also for understudied proteins with 4,210 interactions between 1,039 understudied proteins. FAVA is available as a Python package on PyPI, via the command `pip install favapy`, and it can be used through direct command line or as a module in Python scripts. In addition, the Python module can take as input either a counts/abundance matrix or an Anndata object (23), making it compatible with the most popular single-cell analysis pipelines in Python (e.g. SCANPY (24)).

Results

Overview of the FAVA method

FAVA utilizes scRNA-seq data from the Human Protein Atlas (25) and proteomics data from the PRIDE database (26) as inputs to train two separate VAEs (**Figure 1**, top and main box). VAEs are then used to learn the distribution of the input data and create a meaningful low-dimensional representation (i.e. the latent space). Lastly, all-against-all PCCs are calculated in the latent space to quantify similarity in expression between any two genes/proteins. The resulting adjacency matrix is directly translated to a weighted GCN, one for each data type. To combine the two networks, we first convert the Pearson scores into posterior probabilities of belonging to the same biological pathway (KEGG map) by fitting a calibration function (**Figure 1**, equation plot). If an interaction is supported by both scRNA-seq and proteomics data, the two probabilities are combined. The resulting combined network (**Figure 1**, bottom) thus, takes advantage of the complementary nature of the two datasets and is enriched for understudied proteins.

The FAVA software

The method is implemented in Python and makes use of the Keras deep-learning framework for training VAEs. The software is available via PyPI as ‘favapy’ and is distributed under the MIT open source license. It can be used either directly from the command line, taking an expression matrix file as argument, or as a module in Python scripts. In the Python module, the input can be either a matrix with counts or abundance values from any omics dataset or an Anndata (23) object. The latter is a generic class for handling annotated data matrices making it simple to perform co-expression analysis from scRNA-seq data using popular pipelines.

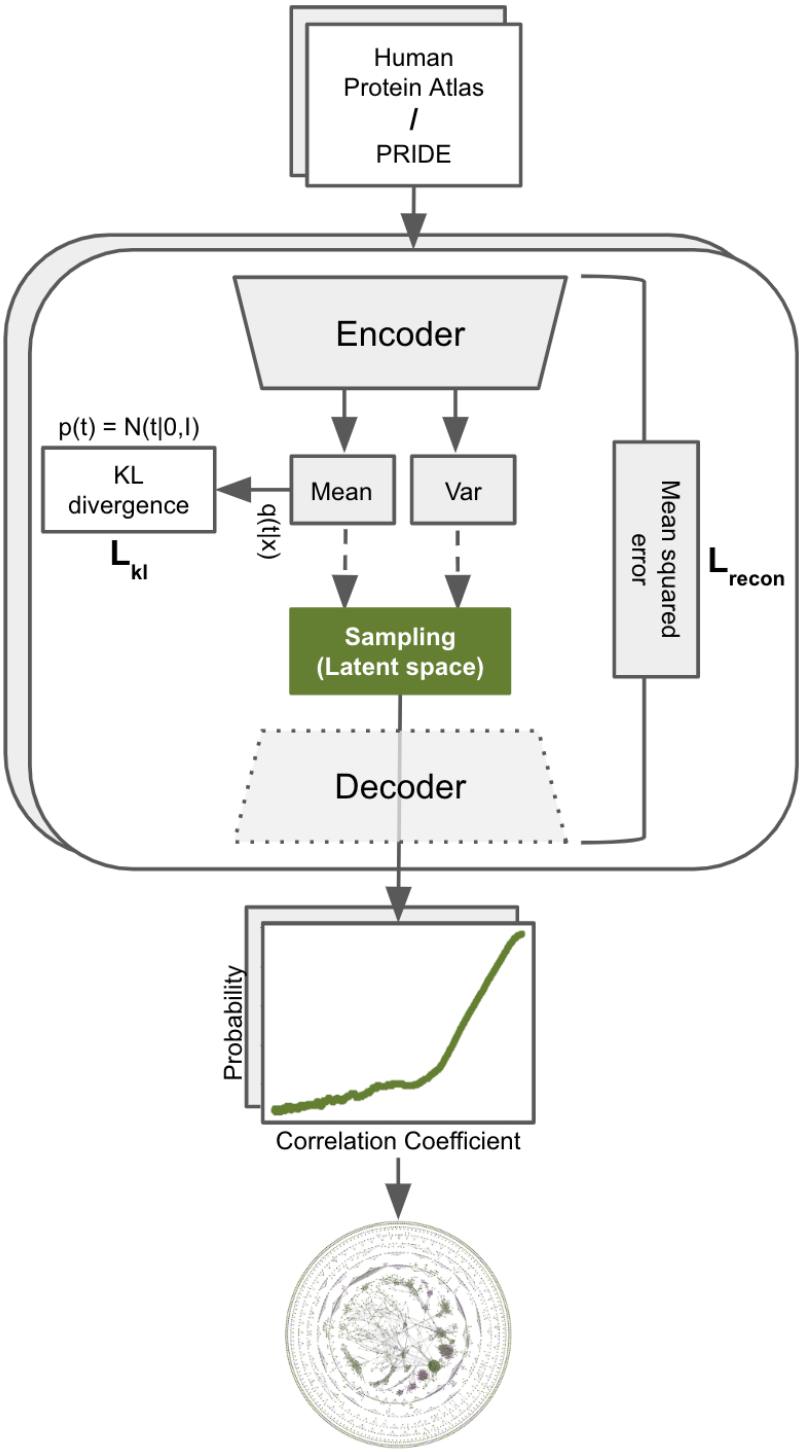


Figure 1. Overview of the FAVA method: from expression matrices to a combined network. The first step is to pre-process each input matrix before it is fed to a variational autoencoder (VAE). After pre-processing, the input matrix is used to train the VAE model, from which we obtain the latent space from the bottleneck layer. In this latent space, we calculate the Pearson correlation coefficient (PCC) for each pair of proteins, resulting in a GCN. Next, we fit a calibration function to convert the PCC scores to posterior probabilities based on KEGG pathways. This pipeline is applied separately to each of the two matrices obtained from the Human Protein Atlas and PRIDE. Finally, we combine the posterior probabilities from the two networks to produce a single, combined network.

FAVA outperforms state-of-the-art methods for GCN construction from individual scRNA-seq datasets

We first evaluated the performance of FAVA on scRNA-seq data, due to their sparse and redundant characteristics even at low dimensions. Specifically, we tested the method on two scRNA-seq datasets, a small one with 2,700 cells (PBMCs from 10x–v3.1 chemistry) and a large one with 30,478 cells (PBMCs from Parse Biosciences–v2 chemistry). The two scRNA-seq datasets were chosen to be complementary with respect to their size and origin as well as to have been used as benchmarks for earlier methods (24,27). The resulting GCNs were benchmarked on two complementary gold standards: the manually annotated pathways from KEGG (28) and the experimental physical protein interactions from BioGRID (29). For comparison, we also generated networks using three other GCN methods that are either commonly used or specifically designed for scRNA-seq data, namely Pearson correlation, hdWGCNA (12), and scLink (13).

Figure 2 shows that FAVA outperforms the other methods by a wide margin on both datasets and with respect to both gold standards. For a simplified interpretation of the curves, we here focus on the true positive/false positive (TP/FP) ratios in the top 10,000 interactions from each method. On the small dataset, FAVA gave a TP/FP ratio of 3.9 (4,588/1,168) according to the KEGG benchmark (**Figure 2a**), which is 3.9 times better than the second best method, scLink (1,878/1,860=1.0). For the BioGRID benchmark, the corresponding ratios are 0.29 (2,266/7,734) and 0.1 (1,014/8,986) for FAVA and scLink, respectively (**Figure 2b**). On the large dataset, which was too large for scLink to process, FAVA also produces the best results with TP/FP ratios of 0.55 (988/1,800) and 0.04 (430/9,570) for the KEGG and BioGRID benchmarks, respectively (**Figure 2c,d**). By comparison, two other methods barely manage to perform better than random.

We thus conclude that the GCNs produced by FAVA based on scRNA-seq data are better than the corresponding networks from other methods at predicting which proteins function in the same pathway as well as which proteins bind to each other. While neither gold standard is a perfect ground truth for evaluating GCNs, the consistency of the results on two very complementary gold standards and datasets is reassuring.

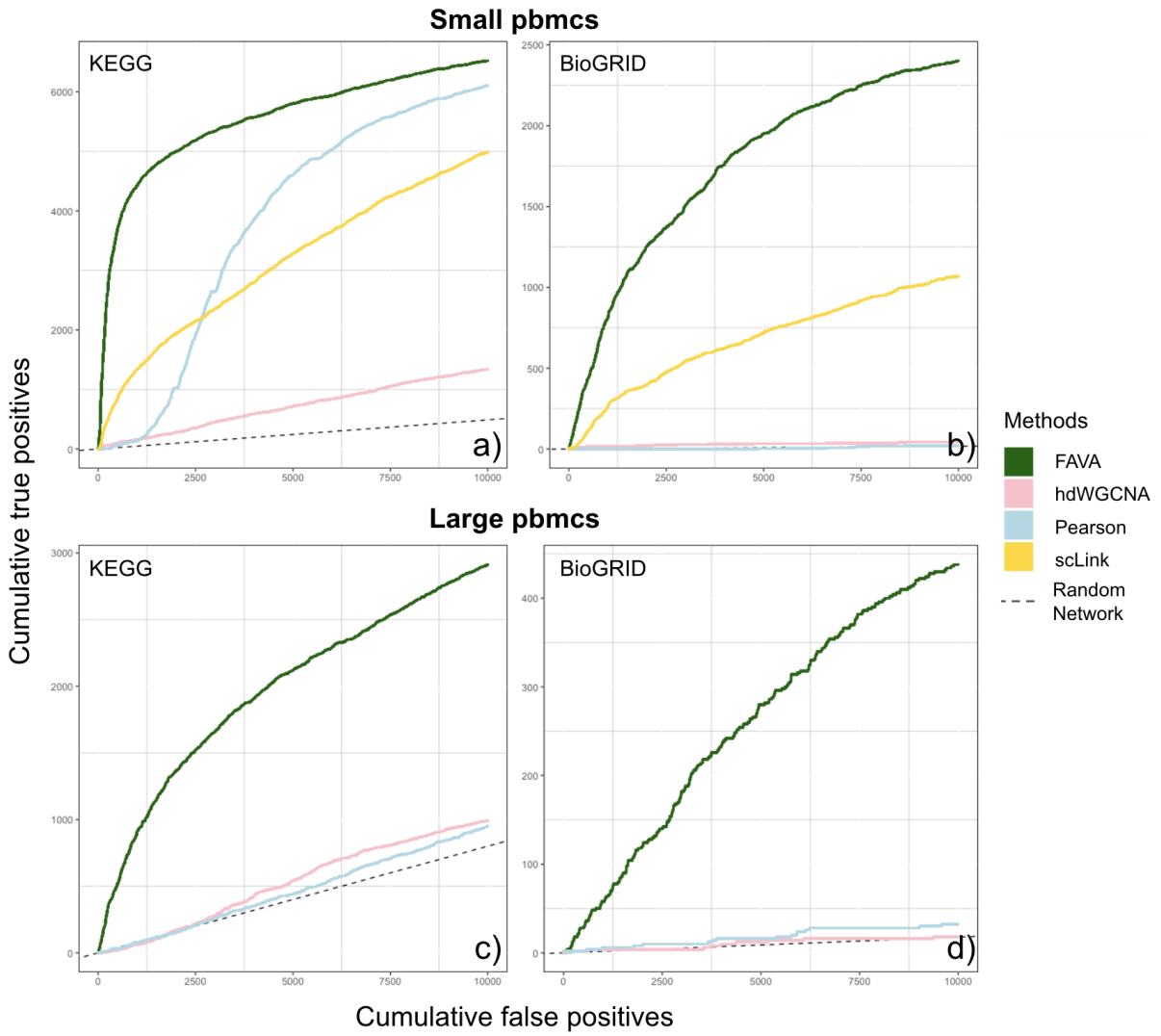


Figure 2. Comparison of the performance of FAVA and the other gene network inference methods on individual scRNA-seq expression data. The plots show method accuracy on GCNs based on different benchmark sets. a) and b) show performance on the small dataset in KEGG and BioGRID, respectively. c) and d) show performance on the large dataset in KEGG and BioGRID, respectively. scLink was unable to create a network from the large dataset and was therefore excluded from the Figures 2c and 2d. For KEGG, a True Positive means that both proteins/genes in a pair are found in the same KEGG map, and a False Positive is when both genes/proteins are found in different KEGG maps. For BioGRID, True Positive is a pair of genes/proteins that is also present in the BioGRID interactome, while False Positive is a pair not present in BioGRID. e) and f) show the total common interactions and the true common interactions between FAVA and scLink in a window of 10,000 False Positives in KEGG, respectively.

FAVA performs better on simulated datasets for Gene Regulatory Network predictions

Another common way to assess the performance of computational methods is using simulated data. A few simulators exist that can create simulated expression matrices from Gene Regulatory Networks (GRNs). Even though GRNs remain outside of the scope of methods constructing GCNs, we decided to evaluate their performance on simulated data created with two algorithms, namely scMultiSim (30), SERGIO (31). The results show that FAVA outperforms hdWGCNA, scLink, and Pearson correlation for prediction of GRNs from both simulation algorithms (**Figure 3)**. In both cases, FAVA has the best performance with a TP/FP ratio of 0.011 (116/9,884) and 0.02 (208/9,792) for the datasets simulated by

scMultisim and SERGIO, respectively. For the scMultisim dataset, scLink gave the second best performance (74/9,926=0.007) (**Figure 3a)**, whereas the ranking of the other methods is less clear on the SERGIO dataset (**Figure 3b)**.

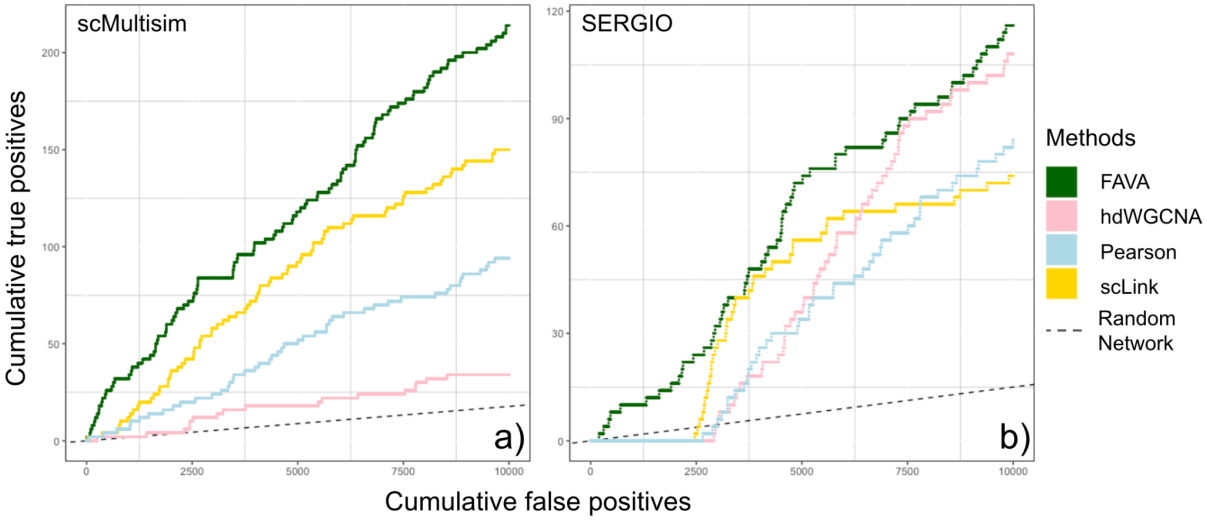


Figure 3. Performance of the methods on simulated data. a) Method performance on simulated data from the scMultisim package. FAVA outperforms the other algorithms with ~1.6 times better performance than the second best (scLink). b) Methods performance on simulated data from the SERGIO package. FAVA outperforms the other algorithms, whereas the ranking of the other methods is complicated.

We tried to also use a third simulation algorithm, BEELINE (32), which is a well-known framework for benchmarking the performance of computational methods in constructing GRNs. BEELINE contains pre-implemented algorithms dedicated to GRN prediction and suitable expression datasets and benchmarks. Unfortunately, we were not able to make BEELINE run correctly.

Application of FAVA to ~0.5M single cells and ~32k proteomics studies

To assess the ability of FAVA to process huge, diverse omics datasets and infer novel functional associations, we applied the method to the single-cell dataset from the Human Protein Atlas, and to proteomics data on human samples from the PRIDE database. We evaluated the quality of the resulting GCNs by benchmarking them against pathways from the KEGG database (Methods).

Here we show that the network derived from scRNA-seq data (**Figure 4**, green continuous line) captures a very large number of high-confidence interactions in KEGG; it is possible to obtain more than 5,000 TPs with less than 500 FPs, corresponding a precision of more than 90%. The proteomics-based network (**Figure 4**, blue continuous line), by contrast, provides fewer high-confidence interactions but more interactions at lower confidence, with the two methods performing on par at ~3,000 FPs. Comparing the FAVA results to those obtained from Pearson correlation on the single-cell and proteomics data (**Figure 4**, dashed lines), shows that FAVA performs better by a wide margin on both types of data, and especially on scRNA-seq. We did not include scLink and hdWGCNA in this comparison, because the former is unable to handle so large datasets, and both are designed specifically for scRNA-seq data and are thus not applicable to the proteomics dataset.

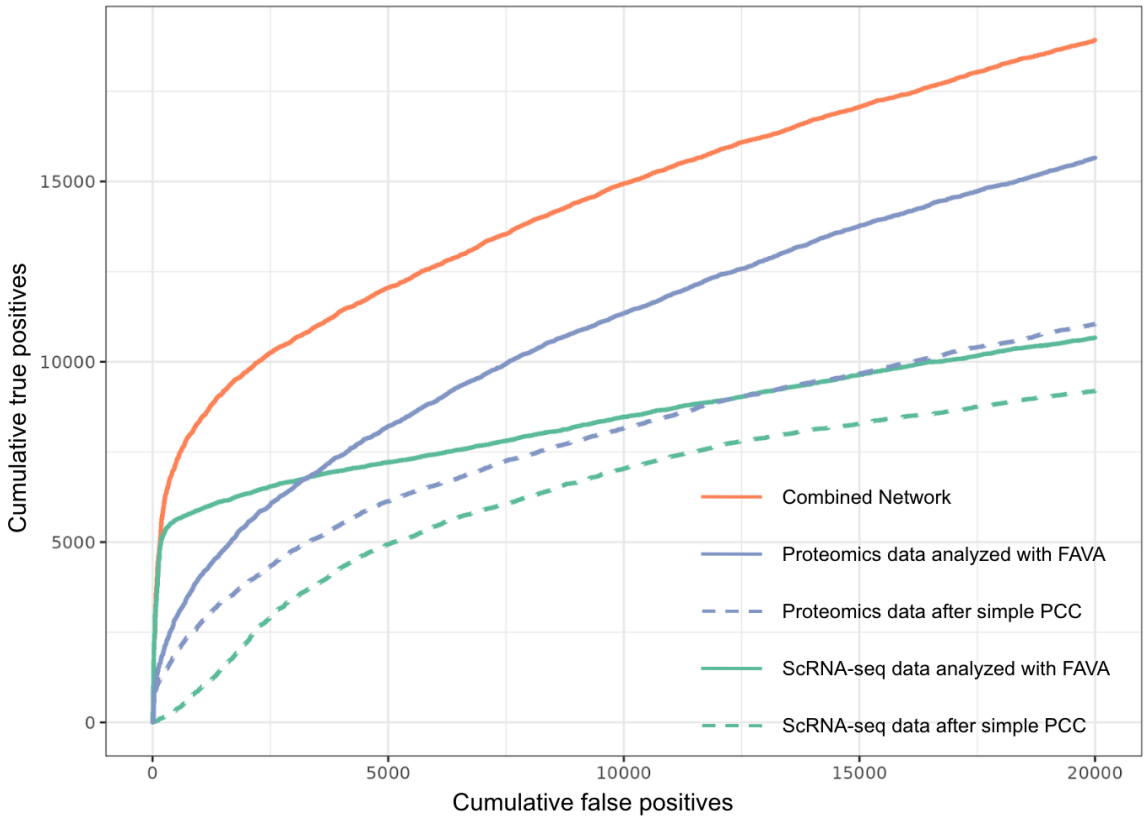


Figure 4. Comparison of FAVA's results against networks obtained by calculating PCC. The plots show how many known interactions can be predicted according to the KEGG database with the two different methods. True Positive is a pair of proteins when both proteins are found in the same KEGG map. False Positive is a pair of proteins when the two proteins are found in different maps of KEGG. Blue continuous line: Benchmark network from proteomics data after applying FAVA; Blue dashed line: Benchmark network from proteomics data after applying traditional PCC; Green continuous line: Benchmark network from scRNA-seq data after applying FAVA; Green dashed line: Benchmark network from scRNA-seq data after applying traditional PCC; Orange: Resulting network after combining scRNA-seq and proteomics networks coming from FAVA, using KEGG for score calibration (Methods).

Combined network from scRNA-seq and proteomics data

Given the complementary nature of the networks based on scRNA-seq and proteomics data individually, we decided to combine them into a single network. As the PCC scores from FAVA cannot be assumed to be directly comparable across the two networks, we converted them to probabilistic scores by calibrating both on the KEGG benchmark (Methods). These calibrated scores were then combined to produce a single network based on scRNA-seq as well as proteomics data. **Figure 4** shows the performance of the combined network (orange continuous line) derived from both the scRNA-seq and the proteomics networks (green and blue continuous lines, respectively). As should be expected, the combined network shows better agreement with KEGG than either of the individual networks that it was created from.

As was the case for the individual networks, the benchmark against KEGG can also be used to assign probabilistic scores to the interactions in the combined network. At a 15% confidence cutoff (corresponding to the low-confidence cutoff of STRING), the combined network consists of a total of 511,048 associations for 16,790 proteins. The network can be further filtered to obtain higher confidence networks, depending on the concrete use case; in case of proteins with many associations, one will generally want to focus on the highest scoring ones. The combined network has 52,953 associations between 8,191 proteins at the

medium confidence cutoff (40%), and even at the high confidence cutoff (70%), it provides 24,182 interactions among 4,166 proteins. The network of all with a confidence score of 15% or better is available for download (doi: 10.5281/zenodo.6803472).

No overfitting on scRNA and proteomics data

In FAVA, separate VAEs are trained to learn the distribution of individual datasets in an unsupervised manner, and the trained models are not applied to unseen data. As no fitting takes place on benchmark datasets, any overfitting of the VAEs to the input data should thus not result in the good performance that we observe. Rather, overfitting should result in a less meaningful latent space and consequently worse performance. To nonetheless evaluate if overfitting happens, we chose a scRNA-seq dataset (33) and randomly split it into a training and a test set, consisting of 67% and 33% of the original dataset, respectively. We trained the VAE only on the train set and tried to predict interactions from both the training and the test set. Similar performance was obtained for both training and test sets (Supplementary figure **S1**), suggesting that our models do not overfit the data and that they are even able to generalize to unseen data, although that is not required use cases presented.

Associations for understudied proteins

Given the nature of the data that the combined network is based on, it should be free of the inherent literature bias, which many other protein networks suffer from. This, combined with it containing many high-confidence interactions, should make the network well suited to help elucidate functions of understudied proteins.

To explore this, we use the definition of “understudied proteins” from the National Institutes of Health (NIH) funded consortium “Illuminating the Druggable Genome” (IDG) (34). IDG provides information on human proteins (targets), dividing them into four levels for drug development: Tclin are drug targets with known mechanisms of action approved by DrugCentral; Tchem are targets selected from ChEMBL/DrugCentral or manually curated from other sources; Tbio are biologically characterized targets without known drug or small molecule activities that meet specified activity thresholds; and Tdark are targets with little known information and not meeting the drug/small molecule activity criteria. In our analysis we used the list of 6,000 proteins that have been designated as Tdark (“dark” targets) by IDG (3). Looking these up in the combined FAVA network, based on both scRNA-seq and proteomics data, revealed 4,210 predicted interactions between 1,039 of the understudied proteins (**Figure 5**, purple nodes) and 611 other, better studied proteins (Tclin) with at least 70% confidence. Next, we compared this network to hu.MAP 2.0 (7) to evaluate both the agreement and the complementarity of the two networks. Since hu.MAP 2.0 is a physical interaction network, it should contain only a subset of the interactions from the combined FAVA network. In the Supplementary figure 2 (**S2**) we report the precision and recall calculations of our network on hu.MAP 2.0 edges, showing the agreement between the two networks. Furthermore, comparing the two networks, both at 70% confidence, we found that they agree on 1,284 edges (**Figure 5**, dark gray edges) for 520 nodes, which confirms the high quality of both networks and shows that the FAVA network includes physical interactions among the functional associations. Focusing on the understudied proteins, 49 of the 1,039 that are found in the FAVA network are also among the 648 that hu.MAP 2.0 provides physical interactions for (see Supplementary material S3). This shows that while there is good agreement between the two networks, they are also highly complementary, illustrating that FAVA provides much needed additional information about different understudied proteins.

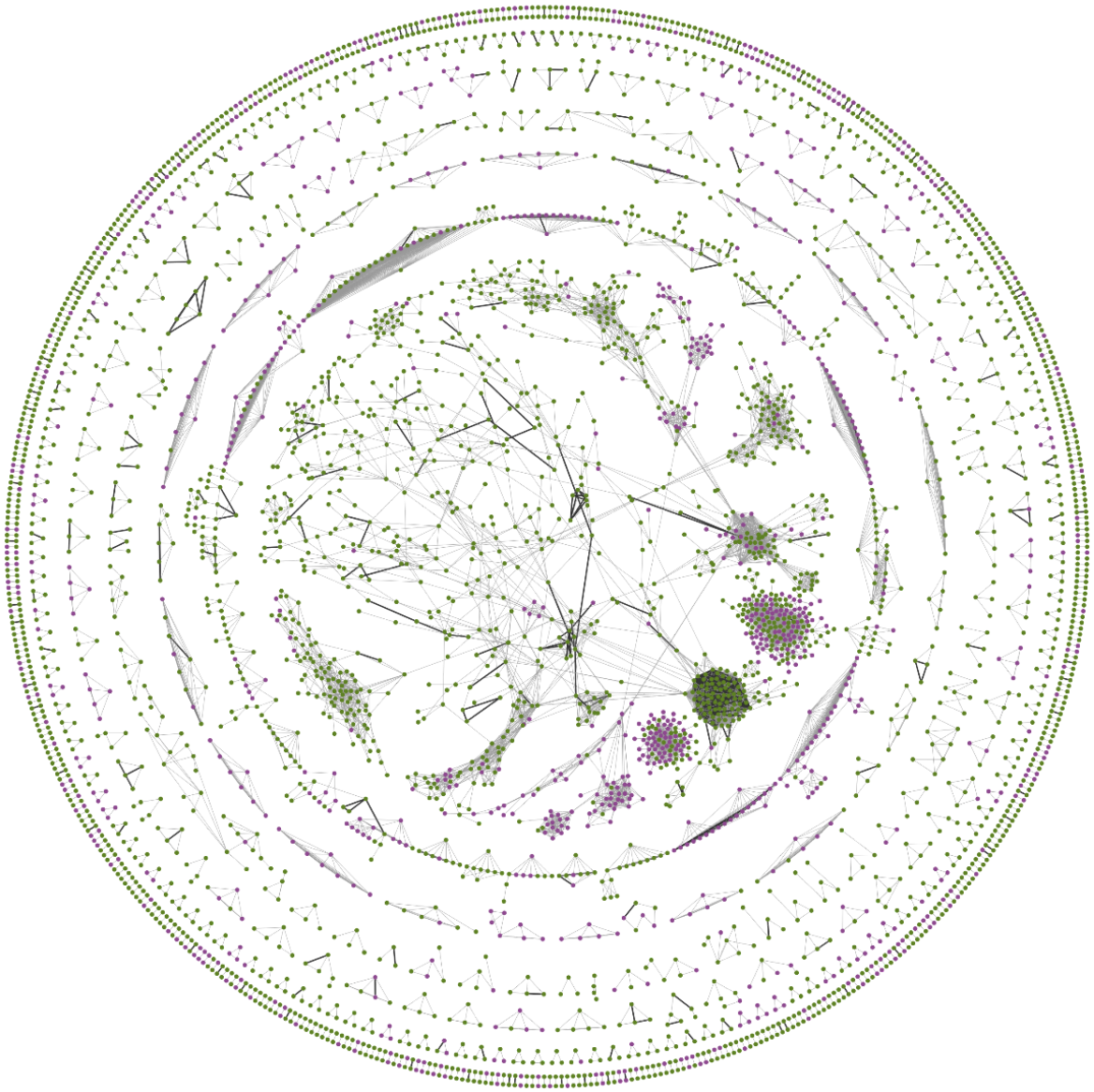


Figure 5. Understudied proteins in the combined FAVA Network. The figure shows all functional associations with a confidence score of at least 70%, based on FAVA analysis of both the scRNA-seq and proteomics data. Understudied proteins according to IDG are highlighted in purple. The edges in the network which are physical interactions according to hu.MAP 2.0 are highlighted with darker, thicker edges. The visualization of the network was made in Cytoscape (35) and the analysis for the IDG target classification was retrieved using Cytoscape stringApp (36).

Discussion

Here we introduced FAVA, a new method to build functional association networks from omics data using VAEs. FAVA takes an expression matrix from, for example, scRNA-seq or proteomics experiments. As input and uses VAEs to create a low-dimensional representation of it, known as the latent space. It then produces a weighted GCN by calculating PCCs in the latent space. To evaluate the ability of FAVA to handle sparse and redundant data, we have tested it on two scRNA-seq datasets of very different size and compared its performance to that of tools designed or commonly used for the same purpose, namely scLink, hdWGCNA and Pearson correlation. We evaluated the methods on two complementary benchmark sets,

namely pathways from KEGG and physical protein interactions from BioGRID. Overall, FAVA demonstrated the best performance.

FAVA is an efficient, unsupervised, stand-alone method, which does not rely on the results from other approaches, as the hdWGCNA does. Additionally, it considers the expression count of all genes, while scLink only takes into consideration the expression counts of certain highly expressed genes. With that, FAVA avoids the accumulation of "rich-get-richer" genes/proteins and produces networks that contain both studied and understudied genes/proteins, which is a major issue in the field. Compared to Pearson, FAVA is at least nine times faster (1 day ≈ 9 days) and is better at accurately identifying established patterns in the data.

The efficiency of FAVA allows the method to be applied not only to individual studies, but also to large expression atlases. This versatility of FAVA allows the construction of GCNs from vast amounts of data of different types. As illustrated in **Figure 5**, FAVA generates GCNs from scRNA-seq and proteomics data and outperforms conventional co-expression methods. The combination of these networks offers new perspectives on uncharacterized proteins. This highlights the ability of the method to handle data beyond scRNA-seq. Overall, the efficiency of FAVA makes it a powerful and flexible approach for gene expression studies.

FAVA also achieves higher accuracy in gene network construction on simulated datasets, as presented on **Figure 3**. However, we remain skeptical about these results, because the simulation algorithms used were not designed for the same purpose as the methods discussed. Indeed, we utilized simulation algorithms that create expression datasets based on GRNs. Then we evaluated the performances of FAVA, scLink, hdWGCNA and Pearson, methods developed for building GCNs, regarding their accuracy on GRN predictions. We concluded that while none of the methods had accuracy suitable for this task, FAVA performed better than the other methods.

Conclusions

In this work, we have shown that variational autoencoders can be used to model single-cell RNA-seq and proteomics data for prediction of functional associations. We have applied this method, which we call FAVA, to large compendiums of scRNA-seq data as well as proteomics data. The results show that FAVA scales to such large datasets, and that the resulting networks are considerably more reliable than those obtained from traditional co-expression methods to the same data. We have moreover shown that combining FAVA results from both data types provides an even more comprehensive network, and that this network can be used to associate understudied proteins with better studied ones, thereby providing hints to their possible functions. We make these networks publicly available along with the Python implementation of FAVA, which can be installed as the PyPI package ‘favapy’.

Methods

Datasets

Small PBMCs dataset

This dataset consists of 2,700 single Peripheral Blood Mononuclear Cells (PBMC) that were sequenced using Chromium kit from 10x Genomics. We downloaded the counts matrix for this freely available dataset from 10X Genomics:
https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz

Large PBMCs dataset

This larger dataset of 30,478 Peripheral Blood Mononuclear Cells (PBMC v2) was sequenced after applying the Evercode™ WT kit. We downloaded the counts matrix for this freely available dataset from Parse Biosciences:
www.parsebiosciences.com/datasets

Human Protein Atlas – Single-cell RNA-seq read-count data

We obtained the single-cell dataset from the Human Protein Atlas (https://www.proteinatlas.org/humanproteome/single+cell+type), a public resource that provides transcriptomics and spatial antibody-based proteomics profiling of human tissues. Their single-cell transcriptomics atlas combines data from 26 datasets. The matrix gives read counts for 19,670 human protein-coding genes in 56,6109 individual cells grouped into 192 cell type clusters. Information about the external datasets and processing of the data is described in detail in (37).

PRIDE EMBL-EBI – Proteomics dataset

We obtained our proteomics dataset from The PRoteomics IDEntifications (PRIDE - https://www.ebi.ac.uk/pride/) database, the world's largest data repository of mass spectrometry-based proteomics data (26). Specifically, we used 633 human proteomics project experiments with a total of 32,546 runs and reanalyzed them using ionbot (38) with an FDR threshold of 0.01 (39), resulting in a total of 154,885,151 peptide spectrum matches for 18,846 proteins. For the full list of projects, runs, and general statistics of the search see supplementary table (doi: 10.5281/zenodo.6798182).

Simulated Datasets

To benchmark the GCN methods on simulated data, we used two recent simulators, SERGIO and scMultiSim, to create single-cell expression matrices based on GRNs. We then used FAVA, scLink, hdWGCNA, and Pearson correlation to create GCNs, and benchmarked these on the ground-truth GRNs.

scMultisim

To construct an expression matrix with scMultisim, we follow the pipeline described here: https://zhanglabgt.github.io/scMultiSim/vignettes/sim_new.nb.html. To allow for generation of a matrix with 2,000 cells and 1,000 genes we ran the following command: `options_ = list(rand.seed=0, GRN=GRN_params_100, num.cells=2000,`

`diff.cif.fraction=0.8, intrinsic.noise=1)`. The given GRN contains 1,432 interactions.

SERGIO

From SERGIO, we used the `SERGIO_noised_1200G_9T_300cPerT_dynamics_6_DS83` file, which is a simulated matrix populated with 300 cells and 1,200 genes and includes technical noise. The package also offers the equivalent GRN with 2,713 interactions, from which the simulated matrix is derived.

FAVA method

We have developed a novel method for construction of GCNs from huge omics datasets with high data sparseness and redundancy. The method makes use of VAEs to perform dimensionality reduction and subsequently scores co-expression in the latent space.

Pre-processing count data

The first step is to log_2 normalize the count matrices. By log-transforming the values, we model proportional changes rather than additive changes and we help the model focus on the biologically relevant differences rather than the extreme values. The general recommendation is to ensure that the data lies in the range of the function we are using to approximate it, in our case a neural network with sigmoid activation function on the output layer. Then, we normalized the values to the [0,1] range by dividing by the max value of the row that each value belongs.

Dimensionality reduction using Variational Autoencoders and Hyperparameters

To compress the high-dimensional expression matrices into lower-dimensional latent spaces, we make use of Variational Autoencoders (VAEs). Using VAEs, the latent representations are regularized by minimizing the regularization loss, implemented with the Kullback–Leibler (KL) divergence between the latent and a prior, in this case a normal distribution with mean 1 and 0.1 standard deviation. The VAE loss is composed by the reconstruction and regularization losses weighted sum, where the weights are set to 0.9 and 0.1 respectively, obtained through hyperparameter tuning. Regarding the VAEs architecture, VAE performance was evaluated with different layers. We use a single hidden layer in FAVA. In all cases, we used the Rectified Linear Unit (ReLU) function (40) for all layers, except for the sigma and mu encoding layers and the last layer in the decoder. In the output layer, considering the 0-1 scaling of the inputs, a sigmoid function was used to generate values between 0 and 1, whereas linear activation was set for the mu and sigma layers. To train the VAE, we used the Adam optimizer with a learning rate of 10^{-3}. The VAE model was implemented in Keras (https://keras.io/).

Pearson Correlation Coefficient pairwise on the latent space

Having produced a regularized latent space that follows Gaussian distribution from the VAE, we calculate all pairwise PCCs between proteins in the latent space. That outputs a list of protein pairs with an assigned score, showing the proximity of the two proteins in the latent space. Based on this score, we sort all protein pairs and create a ranked list, in which numbers closer to 1 represent higher proximity in the latent space and thus, expression similarity. Finally, we benchmark the resulting ranked list against the KEGG database (28) to quantify how well the predicted interactions agree with what is known.

Co-expression scoring in the latent space

Having produced a regularized latent space that follows Gaussian distribution from the VAE, we calculate all pairwise PCCs between proteins in the latent space. That outputs a list of protein pairs with an assigned score, showing the similarity of the two proteins in the latent space. Based on this score, we sort all protein pairs and create a ranked list, in which numbers closer to 1 represent higher similarity in the latent space and thus, expression similarity.

Generalization of FAVA algorithm on unseen data

To show that the FAVA method can even generalize to genes not included in training the VAEs we employed another dataset from GEO (GSE75748 (33)). We split it into a 67% train set and a 33% test set. We, then, trained our VAEs on the train set and we separately applied the trained model to both the train and test sets. Two different networks were generated by the two different latent spaces that were created. The accuracy of these networks was evaluated using KEGG (see Supplementary figure 1).

Network construction with FAVA

FAVA is available as a PyPI package and can be used either in the terminal or as a module in a Python script. In both datasets, we applied FAVA from terminal with the following command: `favapy <path-to-data-file> <path-to-save-output>`, as described here: https://pypi.org/project/favapy/

Other methods and networks

scLink

To create the network of the small dataset using the scLink method, we followed the pipeline described here: https://cran.r-project.org/web/packages/scLink/scLink.pdf. Thus, we normalized the data with the package's function and filtered for the 500 most highly expressed genes, as recommended by the authors: `sclink_norm(count, scale.factor=1e+06, filter.genes = TRUE, n = 500)`. Then we calculated the correlation matrix with the function: `sclink_cor(expr = count.norm)`. We were not able to process the large dataset on a server with 180GB RAM.

hdWGCNA

To create the networks using hdWGCNA, we first had to analyze the data using the Seurat pipeline. Thus, we followed the tutorial provided by Seurat here: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html. That creates a Seurat object (seurat_obj) which can be utilized further by hdWGCNA. Afterwards, we followed the recommendations of the authors to construct the network as described in their tutorial here: https://smorabit.github.io/hdWGCNA/articles/basic_tutorial.html. Therefore, we ran in R the following commands:

```
seurat_obj = SetupForWGCNA(seurat_obj, gene_select = "fraction", fraction = 0.05)
seurat_obj = SetDatExpr(seurat_obj, assay = 'RNA', slot = 'data')
seurat_obj = TestSoftPowers(seurat_obj, networkType = 'unsigned')
```

Based on the results coming from the `GetPowerTable(seurat_obj)` function, we selected the soft_power parameter (4 for the small dataset and 7 for the large dataset). Final step was to construct the network:

```
seurat_obj = ConstructNetwork(seurat_obj, soft_power = soft_power,
setDatExpr = FALSE, overwrite_tom = TRUE, networkType = 'unsigned')
```

Pearson correlation coefficient

We calculated all-against-all Pearson correlation on the same log-transformed counts matrices used as input for FAVA. This was done in order to have a fair comparison with FAVA.

Benchmark of functional associations

KEGG

We benchmark the resulting ranked list against the KEGG database, identical to how functional associations are benchmarked in the STRING database (4), to quantify how well the predicted interactions agree with what is known. To do this, we first map the protein pairs from the methods to KEGG maps. If a KEGG map exists, which contains both proteins of a pair, the pair is counted as a true positive (TP). If both proteins can be mapped to KEGG, but there is no map containing both, the pair is counted as a false positive (FP). Pairs for which one or both proteins cannot be mapped to KEGG are disregarded for benchmarking purposes. Having defined which protein pairs are considered TPs and FPs, we plot the cumulative TP count as a function of the cumulative FP count for the sorted list of FAVA pairs.

BioGRID

To measure how accurately the networks from the different methods correlate with BioGRID's experimental results, we considered the interaction network from BioGRID as our ground truth. If a predicted pair existed in BioGRID, it was labeled as a True Positive (TP). Conversely, if it was not present in BioGRID, it was labeled as a False Positive (FP). The interaction network of BioGRID can be found here:
https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-4.4.217/

Score calibration and combination of FAVA networks from atlases

To combine the two networks, we decided to apply the pipeline used in the STRING database. Thus, we first convert the PCC scores from FAVA from each dataset into posterior probabilities of being on the same KEGG map given the PCC (**Figure 1**). We do this based on the benchmark results described above, by first plotting the local precision within a sliding window ($y = TP/(TP + FP)$) as function the average PCC within the window (x). We do this separately for scRNA-seq and proteomics data, and fit the following calibration function to each by minimizing the squared error using simplex optimization:

$y = a0 + a1 * x + a2/(1 + exp(a3 * (x - a4)))$,

where $a0$ through $a4$ are the parameters that are optimized to fit the points in the plot. Once fitted, we use the resulting calibration curves to convert all PCCs from each dataset to probabilities. When a pair is supported by both scRNA-seq and proteomics data, the two probabilities are combined.

Availability of data and materials

Data availability

Combined network

The Network: https://doi.org/10.5281/zenodo.6803472

The PRIDE database analysis

The full list of projects, runs, and general statistics about the analysis of the data in the PRIDE database: https://doi.org/10.5281/zenodo.6798182

Code availability

https://github.com/mikelkou/fava

PyPI: <pip install favapy>

References

1. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. Freeman T, editor. PLOS Biol. 2018 Sep 18;16(9):e2006643.
2. Oprea TI. Exploring the dark genome: implications for precision medicine. Mamm Genome. 2019 Aug;30(7–8):192–200.
3. Lin Y, Mehta S, Küçük-McGinty H, Turner JP, Vidovic D, Forlin M, et al. Drug target ontology to classify and integrate drug discovery data. J Biomed Semant. 2017 Dec;8(1):50.
4. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1):D605–12.
5. Kustatscher G, Collins T, Gingras AC, Guo T, Hermjakob H, Ideker T, et al. An open invitation to the Understudied Proteins Initiative. Nat Biotechnol. 2022 Jun;40(6):815–7.
6. Kustatscher G, Collins T, Gingras AC, Guo T, Hermjakob H, Ideker T, et al. Understudied proteins: opportunities and challenges for functional proteomics. Nat Methods. 2022 Jul;19(7):774–9.
7. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol [Internet]. 2021 May [cited 2022 Aug 1];17(5). Available from: https://onlinelibrary.wiley.com/doi/10.15252/msb.202010016
8. van Dam S, Craig T, de Magalhães JP. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Res. 2015 Jan 28;43(D1):D1124–32.
9. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2017 Jan 10;bbw139.
10. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. Learning from Co-expression Networks: Possibilities and Challenges. Front Plant Sci [Internet]. 2016 Apr 8 [cited 2023 Jan 21];7. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00444/abstract
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec;9(1):559.
12. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems [Internet]. Bioinformatics; 2022 Sep [cited 2023 Jan 3]. Available from:

http://biorxiv.org/lookup/doi/10.1101/2022.09.22.509094
13. Li WV, Li Y. scLink: Inferring Sparse Gene Co-expression Networks from Single-cell Expression Data. Genomics Proteomics Bioinformatics. 2021 Jun;19(3):475–92.
14. Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol. 2016 Oct 20;12(10):883.
15. Brunner A, Thielert M, Vasilopoulou C, Ammar C, Coscia F, Mund A, et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol Syst Biol [Internet]. 2022 Mar [cited 2022 Jun 27];18(3). Available from: https://onlinelibrary.wiley.com/doi/10.15252/msb.202110798
16. L. J. van der Maaten P EO,HJ van den Herik. Dimensionality reduction: a comparative review. Journal of Machine Learning Research. 2009;10.1(41):66–71.
17. Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, Apweiler R, et al. Analyzing large-scale proteomics projects with latent semantic indexing. J Proteome Res. 2008 Jan;7(1):182–91.
18. Franceschini A, Lin J, von Mering C, Jensen LJ. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. Bioinformatics. 2016 Apr 1;32(7):1085–7.
19. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 Sep 17 [cited 2021 Apr 8]; Available from: http://arxiv.org/abs/1802.03426
20. Kingma DP, Welling M. An Introduction to Variational Autoencoders. Found Trends® Mach Learn. 2019;12(4):307–92.
21. Kingma DP, Welling M. Auto-Encoding Variational Bayes [Internet]. arXiv; 2014 [cited 2022 Jun 10]. Available from: http://arxiv.org/abs/1312.6114
22. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 2006 Jul 28;313(5786):504–7.
23. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Annotated data [Internet]. Bioinformatics; 2021 Dec [cited 2023 Jan 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.12.16.473007
24. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018 Dec;19(1):15.
25. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.
26. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019 Jan 8;47(D1):D442–50.
27. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data [Internet]. Genomics; 2020 Oct [cited 2021 Mar 3]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.10.12.335331
28. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan 8;49(D1):D545–51.
29. Oughtred R, Rust J, Chang C, Breitkreutz B, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021 Jan;30(1):187–200.
30. Li H, Zhang Z, Squires M, Chen X, Zhang X. scMultiSim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks [Internet]. Bioinformatics; 2022 Oct [cited 2022 Dec 27]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.10.15.512320
31. Dibaeinia P, Sinha S. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. Cell Syst. 2020 Sep;11(3):252-271.e11.
32. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020 Feb;17(2):147–54.
33. Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq

reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. 2016 Dec;17(1):173.
34. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res. 2021 Jan 8;49(D1):D1334–46.
35. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003 Nov 1;13(11):2498–504.
36. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. 2019 Feb 1;18(2):623–32.
37. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single–cell type transcriptomics map of human tissues. Sci Adv. 2021 Jul 30;7(31):eabh2169.
38. Degroeve S, Gabriels R, Velghe K, Bouwmeester R, Tichshenko N, Martens L. ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification [Internet]. In Review; 2021 Aug [cited 2021 Dec 20]. Available from: https://www.researchsquare.com/article/rs-691927/v1
39. Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. J Proteome Res. 2008 Jan;7(1):40–4.
40. Nair, V. and Hinton, G.E. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). 2010;

Acknowledgments

The authors acknowledge Ralf Gabriels for input on the Python package ‘favapy’ and Enrico Massignani for the PRIDE data retrieval.

Funding

This work was supported by the Novo Nordisk Foundation [NNF14CC0001], [NNF20SA0035590] and EMBO Scientific Exchange Grant 9404 [STF_9404]. R.B. acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen under project number HBC.2020.2205. Research Foundation - Flanders (FWO) G028821N to L.M., by the European Union’s Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to L.M.], and Ghent University Concerted Research Action [grant number BOF21-GOA-033 to L.M.]. Funding for open access charge: Novo Nordisk Foundation [NNF14CC0001] and [NNF20SA0035590].

Author contributions

M.K. and L.J.J. conceptualized this study. The manuscript was written by M.K. and L.J.J. with assistance and approval from all authors. M.K. developed the favapy Python package. M.K. and P.P.L. designed the architecture of the VAEs. M.K. collected, processed, and performed network analysis on publicly available datasets. M.K. implemented, processed and performed network analysis on simulated datasets. R.B. and L.M. analyzed and provided the proteomics datasets from PRIDE. M.K. collected, analyzed and constructed networks from Human Protein Atlas (single-cells) and PRIDE (proteomics) data.

Corresponding authors

Correspondence to Lars Juhl Jensen or Lennart Martens.

Ethics declarations

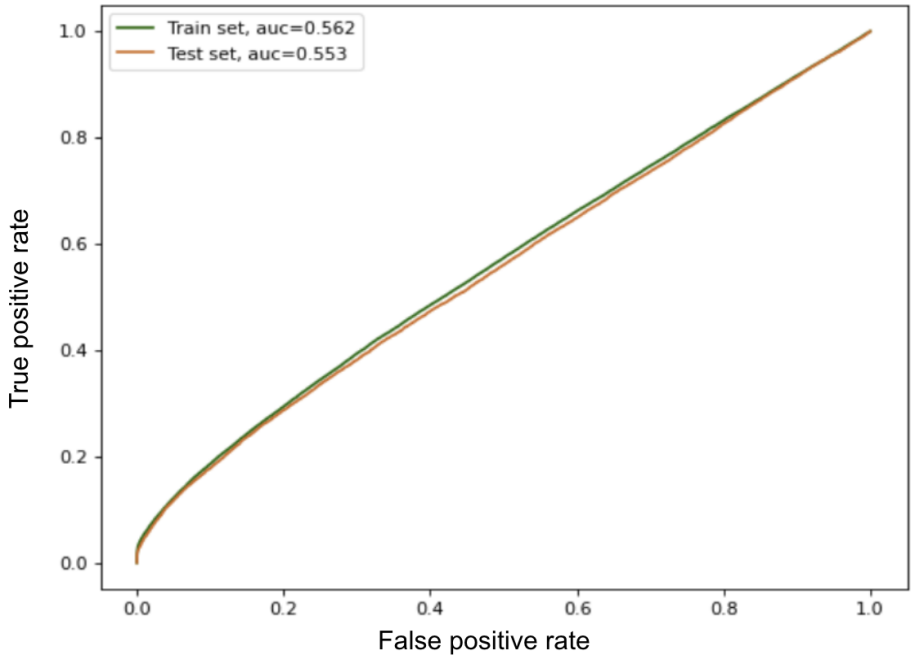
Conflict of interest

None declared.

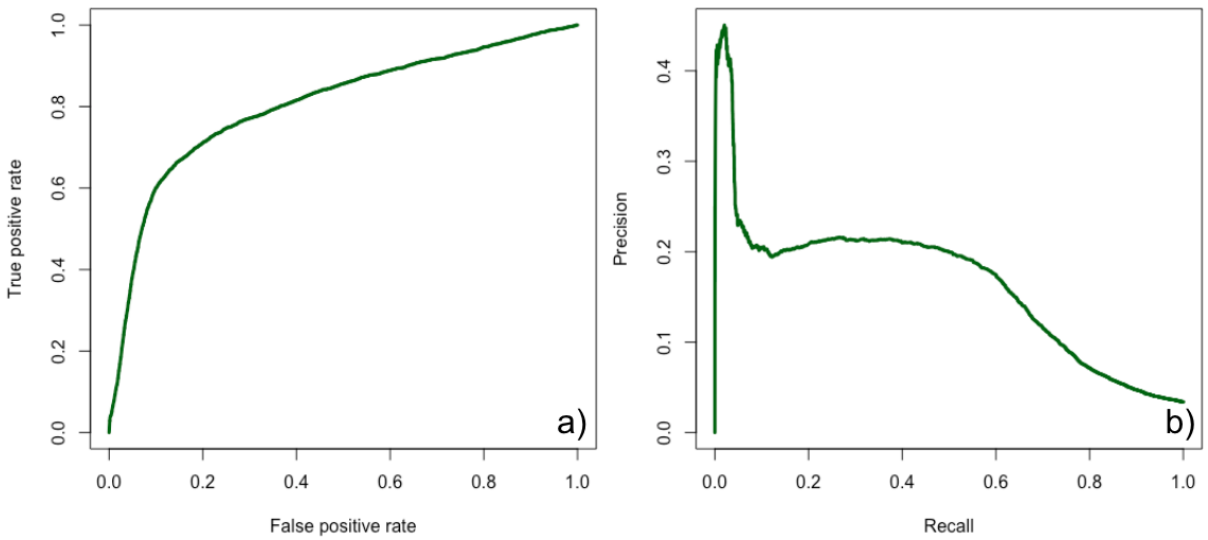
Ethics approval and consent to participate

Ethics approval was not applicable for this study.

Supplementary material



S1. Generalization on unseen data. We split the scRNA-seq dataset (GSE75748) into a training (67%) and a test (33%) set. After training FAVA, the algorithm was applied on both the training and the test set to predict associations. The performance of the predictions is evaluated on KEGG (Methods). The plot shows that our framework is able to generalize when applied on unseen data (i.e. the test set). This is evident from the similarity between the green line (train set) and brown line (test set) in terms of true positive predictions. It is clear that even in the unseen 33% of the dataset, the VAE is able to capture patterns and that the latent representation can be used to predict existing interactions.



S2. ROC curve, Precision and recall of FAVA's combined network based on hu.MAP 2.0. We conducted a comparison between our network from single-cells and proteomics data and the hu.MAP 2.0 physical interaction network. We report the a) roc curve and the b) precision and recall calculations of our network on hu.MAP 2.0 edges, showing the agreement between the two networks.

S3. The Network

A *.pdf* containing the network from Figure 5 with readable labels is provided.