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

Mikaela Koutrouli[1], Pau Piera Líndez[1], Robbin Bouwmeester[2,3], 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 understanding the interplay between proteins in the cell as well as for visualizing omics data. Unfortunately, most existing high-quality networks are heavily biased by data availability, in the sense that well-studied proteins have many more interactions than understudied proteins. To create networks that can help elucidate functions for the latter, we must start from data that are not affected by this literature bias, in other words, from omics data such as single cell RNA-seq (scRNA-seq) and proteomics. While networks can be inferred from such data through simple co-expression analysis, this approach does not work well due to high sparseness (many transcripts/proteins are not consistently observed in each cell/sample) and redundancy (many similar cells/samples are analyzed) of such data. We have therefore developed FAVA, Functional Associations using Variational Autoencoders, which deals with both issues by compressing these high-dimensional data into a dense, low-dimensional latent space. We demonstrate that calculating correlations in this latent space results in much improved networks compared to the original representation for large-scale scRNA-seq and proteomics data from the Human Protein Atlas, and from PRIDE, respectively. We show that these networks, which given the nature of the input data should be free of literature bias, indeed have much better coverage of understudied proteins than existing networks.

Introduction

Networks of physical and functional interactions among proteins are widely used to understand the inner workings of cells and to visualize results from omics results, e.g., as obtained from transcriptomics and proteomics experiments. Unfortunately, most research is focused on the same 10% of human protein-coding genes [1], and networks derived from the biomedical literature — whether through manual annotation or through automatic text mining — are thus heavily biased by this skewed availability of data. Networks obtained from databases such as STRING [2] consequently have many interactions for well-studied proteins but only very few interactions for understudied proteins, which are arguably the most interesting targets for network-based function prediction [3], [4].

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. Such types of data include single cell RNA-seq (scRNA-seq) and proteomics data, which each have different strengths and weaknesses. 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, the correlation between RNA and protein levels is far from perfect [5]. Mass spectrometry-based proteomics informs us about protein levels but, on the other hand, proteomics data at single-cell resolution currently remains a rarity [6]. These two types of data can thus be viewed as complementary starting points for predicting functional interactions, also for understudied proteins. However, both scRNA-seq and proteomics datasets are very sparse (i.e., many transcripts/proteins not consistently 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.

However, dimensionality reduction can help address both sparsity and redundancy in these data. By compressing the information into a lower dimensional space, sparsity is eliminated by combining data from across multiple cells/samples, while redundancy is

inherently reduced, because data compression is specifically achieved by not storing the same information multiple times [7]. 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 are available to produce this latent space. These span linear methods such as truncated singular value decomposition [8], [9], nonlinear transformations such as Uniform Manifold Approximation and Projection (UMAP) [10], and deep generative models such variational autoencoders (VAEs). The latter take advantage of recent advances in deep neural networks and provide a powerful framework for modeling data distributions in general and scRNA-seq data specifically [11]–[14]. VAEs use an encoder-decoder structure to learn meaningful compressed latent representations of the input data that follow Gaussian distributions, resulting in a latent space that is easier to interpret than those of other autoencoders [15], and do so in a completely unsupervised manner. 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 can predict high-confidence functional associations, iii) it outperforms simple co-expression analysis by a wide margin, and iv) the resulting networks provide good coverage also for understudied proteins. FAVA is available as a Python package on PyPI, via the command `pip install favapy`.

Materials and Methods

Single cell RNA-seq read-count data

We obtained the single cell dataset from the Human Protein Atlas [16], 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 [17].

Proteomics dataset

We obtained our proteomics dataset from The PRoteomics IDEntifications (PRIDE) database, the world's largest data repository of mass spectrometry-based proteomics data [18]. Specifically, we used 633 human proteomics project experiments with a total of 32,546 runs and reanalyzed them using ionbot [19] with an FDR threshold of 0.01 [20], 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).

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. In addition, the errors are usually proportional to the values, since the variance is not independent from the mean.

This kind of mean-variance relationship is usually absent on the log scale. The general recommendation is to ensure that the data lies in the range of the function we are using to approximate it. Therefore, afterwards, we divide each value by the maximum value of the row that belongs. With that, our input is within the range of 0 and 1, which is also the range of the output that we request.

Dimensionality reduction

To compress the high-dimensional expression matrices into lower-dimensional latent spaces, we make use of VAEs. In general, the VAE framework uses two deep neural networks — an encoder and a decoder — 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 and sigma 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 as well as to regularize the latent representations. Input reconstruction is achieved by minimizing the mean squared error (MSE) between input and reconstructed instances, which ensures the encoder–decoder learning of the input data distribution. Regularization of the latent space is implemented by penalizing the divergence between the latent samples from the standard normal distribution (N(0,1)), quantified with the Kullback–Leibler (KL) divergence. Restricting the latent variables this way reduces the complexity of the latent space and encourages a meaningful latent representation of the input data, thereby distributing the input latent representations according to its underlying properties into a continuous compact space.

We tested VAE network configurations with zero, one, and two hidden layers and found that VAEs with one or two small hidden layers performed the best. Based on this we decided to use a single hidden layer in FAVA. In all cases, we used the Rectified Linear Unit (ReLU) function [22] for all layers, except for the sigma and mu encoding layers and the last layer in the decoder, where a sigmoid function was used to generate normalized values between 0 and 1. 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 Pearson Correlation Coefficients (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 [23] to quantify how well the predicted interactions agree with what is known. In Figure 1, we compare how well our method, FAVA, works in comparison with applying PCC directly on the single-cells and proteomics data, without prior dimensionality reduction with VAEs.

Co-expression scoring in the latent space

Having produced a regularized latent space that follows Gaussian distribution from the VAE, we calculate all pairwise Pearson Correlation Coefficients (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.

Benchmarking of functional associations

We benchmark the resulting ranked list against the KEGG database [23] to quantify how well the predicted interactions agree with what is known. To do this, we first map the protein pairs from FAVA 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.

Score calibration and combination of networks

To combine the two networks, we first convert the PCC scores from FAVA from each dataset into posterior probabilities of being on the same KEGG map given the PCC. 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 fittet, 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 the same way, it is done in the STRING database [2].

Results and Discussion

The FAVA software

We have developed a novel method for construction of co-expression networks 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. 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.

Performance on scRNA-seq and proteomics data

To assess the ability of FAVA to process huge, diverse omics datasets and infer functional associations, we applied the method to the single cell dataset from the Human Protein Atlas [16], and to proteomics data on human samples from the PRIDE database [16]. We evaluated the quality of the resulting co-expression networks by benchmarking them against pathways from the KEGG database [23], identical to how functional associations are benchmarked in the STRING database [2].

The results of the benchmark are shown in Figure 1. The network derived from scRNA-seq data contains a very large number of high-confidence interactions: it is possible to obtain more than 5,000 true positives with less than 500 false positives, corresponding to a precision of more than 90%. The proteomics-based network, by contrast, does not contain nearly as many high-confidence interactions, but provides more interactions at lower confidence. Comparing the FAVA results to those obtained by simple correlation analyses

without the same datasets (Figure 1, dashed lines) shows that FAVA performs better by a wide margin on both types of data, especially on scRNA-seq data.

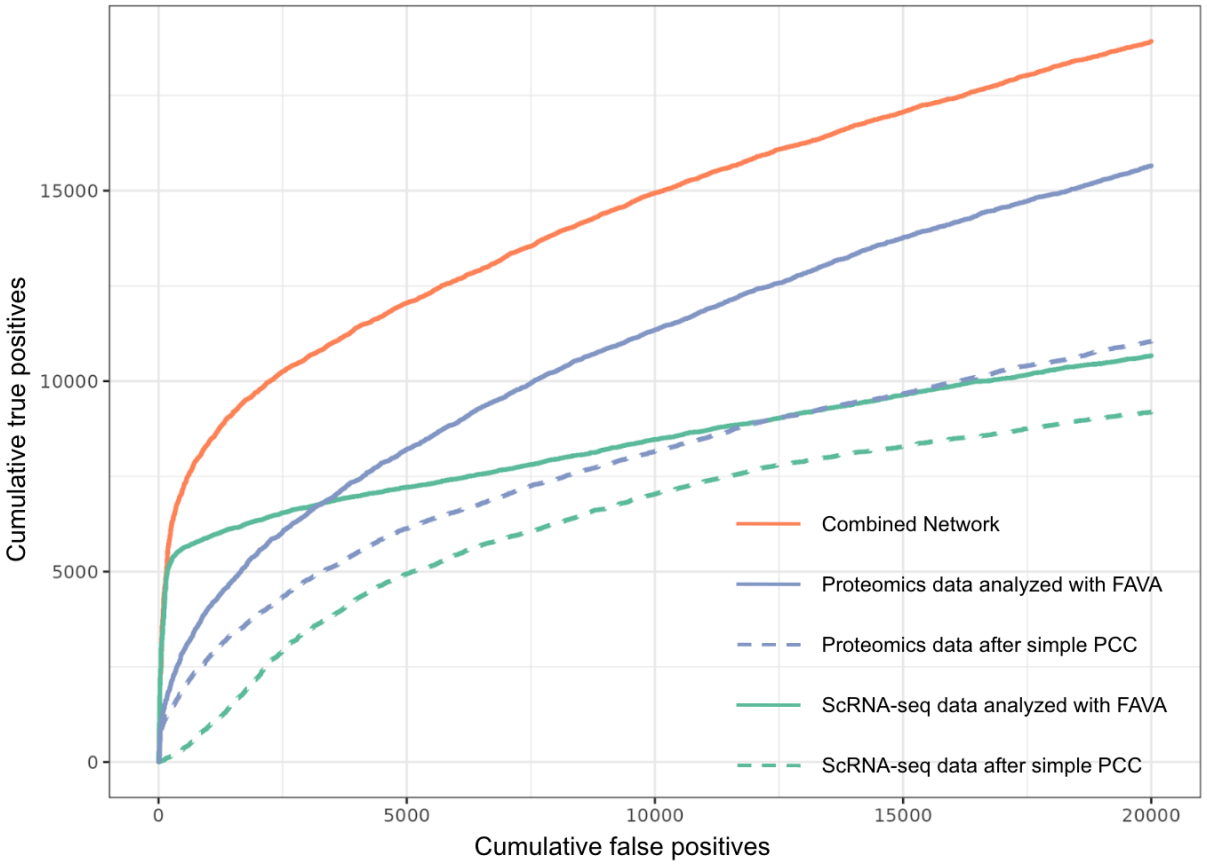

Figure 1. Comparison of FAVA's results against networks obtained by calculating Pearson Correlation Coefficient (PCC). The plots show how many known interactions can be predicted according to the KEGG database with the two different methods. True Positive is a pair of proteins when both proteins are found in the same KEGG map. False Positive is a pair of proteins when the two proteins are found in different maps of KEGG. Orange: Benchmark combined network from scRNA-seq and proteomics data after applying FAVA; Blue continuous line: Benchmark network from proteomics data after applying FAVA; Blue dashed line: Benchmark network from proteomics data after applying simple PCC; Green continuous line: Benchmark network from scRNA-seq data after applying FAVA; Green dashed line: Benchmark network from scRNA-seq data after applying simple PCC.

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 based on the KEGG benchmarks (see Methods). These calibrated scores were then combined to produce a single network based on scRNA-seq as well as proteomics data. As should be expected, this network outperforms the individual networks, combining the best aspects of both (Figure 1).

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 once. 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).

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 defined “understudied proteins” as the 10% least studied proteins in terms of number of publications, using the fractional publication count also featured in the Pharos resource [24]. This resulted in a list of 1,957 proteins, each with a fractional count of less than one publication. Looking these up in the combined FAVA network, based on both scRNA-seq and proteomics data, revealed 12,792 predicted interactions between 487 of the understudied proteins and 1,127 other, better studied proteins with at least 40% confidence. Further analyses are needed to show to what extent this allows functions to be assigned to understudied proteins and, if so, which functional categories they primarily fall into.

Conclusions

In this work, we have shown that variational autoencoders (VAEs) 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’.

Supplementary material

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

Data availability

Datasets

Human Protein Atlas
https://www.proteinatlas.org/humanproteome/single+cell+type

PRIDE - Proteomics Identification Database - EMBL-EBI
https://www.ebi.ac.uk/pride/

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

Code availability

https://github.com/mikelkou/fava
PyPI: <pip install favapy>

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].

References

[1] T. Stoeger, M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral, "Large-scale investigation of the reasons why potentially important genes are ignored," *PLOS Biol.*, vol. 16, no. 9, p. e2006643, Sep. 2018, doi: 10.1371/journal.pbio.2006643.

[2] D. Szklarczyk *et al.*, "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res.*, vol. 49, no. D1, Art. no. D1, Jan. 2021, doi: 10.1093/nar/gkaa1074.

[3] Y. Lin *et al.*, "Drug target ontology to classify and integrate drug discovery data," *J. Biomed. Semant.*, vol. 8, no. 1, p. 50, Dec. 2017, doi: 10.1186/s13326-017-0161-x.

[4] G. Kustatscher *et al.*, "An open invitation to the Understudied Proteins Initiative," *Nat. Biotechnol.*, vol. 40, no. 6, pp. 815–817, Jun. 2022, doi: 10.1038/s41587-022-01316-z.

[5] F. Edfors *et al.*, "Gene-specific correlation of RNA and protein levels in human cells and tissues," *Mol. Syst. Biol.*, vol. 12, no. 10, p. 883, Oct. 2016, doi: 10.15252/msb.20167144.

[6] A. Brunner *et al.*, "Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation," *Mol. Syst. Biol.*, vol. 18, no. 3, Mar. 2022, doi: 10.15252/msb.202110798.

[7] P. L. J. van der Maaten E. O. ,. H. J. van den Herik, "Dimensionality reduction: a comparative review," *Journal of Machine Learning Research*, vol. 10.1, no. 41, pp. 66–71, 2009.

[8] S. Klie *et al.*, "Analyzing large-scale proteomics projects with latent semantic indexing," *J. Proteome Res.*, vol. 7, no. 1, pp. 182–191, Jan. 2008, doi: 10.1021/pr070461k.

[9] A. Franceschini, J. Lin, C. von Mering, and L. J. Jensen, "SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles," *Bioinformatics*, vol. 32, no. 7, pp. 1085–1087, Apr. 2016, doi: 10.1093/bioinformatics/btv696.

[10] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv180203426 Cs Stat*, Sep. 2020, Accessed: Apr. 08, 2021. [Online]. Available: http://arxiv.org/abs/1802.03426

[11] Q. Ma and D. Xu, “Deep learning shapes single-cell data analysis,” *Nat. Rev. Mol. Cell Biol.*, vol. 23, no. 5, pp. 303–304, May 2022, doi: 10.1038/s41580-022-00466-x.
[12] C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, “scVAE: variational auto-encoders for single-cell gene expression data,” *Bioinformatics*, vol. 36, no. 16, pp. 4415–4422, Aug. 2020, doi: 10.1093/bioinformatics/btaa293.
[13] D. Wang and J. Gu, “VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder,” *Genomics Proteomics Bioinformatics*, vol. 16, no. 5, pp. 320–331, Oct. 2018, doi: 10.1016/j.gpb.2018.08.003.
[14] J. Ding, A. Condon, and S. P. Shah, “Interpretable dimensionality reduction of single cell transcriptome data with deep generative models,” *Nat. Commun.*, vol. 9, no. 1, p. 2002, Dec. 2018, doi: 10.1038/s41467-018-04368-5.
[15] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
[16] M. Uhlén *et al.*, “Tissue-based map of the human proteome,” *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015, doi: 10.1126/science.1260419.
[17] M. Karlsson *et al.*, “A single–cell type transcriptomics map of human tissues,” *Sci. Adv.*, vol. 7, no. 31, p. eabh2169, Jul. 2021, doi: 10.1126/sciadv.abh2169.
[18] Y. Perez-Riverol *et al.*, “The PRIDE database and related tools and resources in 2019: improving support for quantification data,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D442–D450, Jan. 2019, doi: 10.1093/nar/gky1106.
[19] S. Degroeve, R. Gabriels, K. Velghe, R. Bouwmeester, N. Tichshenko, and L. Martens, “ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification,” In Review, preprint, Aug. 2021. doi: 10.21203/rs.3.rs-691927/v1.
[20] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble, “Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin,” *J. Proteome Res.*, vol. 7, no. 1, pp. 40–44, Jan. 2008, doi: 10.1021/pr700739d.
[21] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes.” arXiv, May 01, 2014. Accessed: Jun. 10, 2022. [Online]. Available: http://arxiv.org/abs/1312.6114
[22] 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.
[23] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, “KEGG: integrating viruses and cellular organisms,” *Nucleic Acids Res.*, vol. 49, no. D1, Art. no. D1, Jan. 2021, doi: 10.1093/nar/gkaa970.
[24] T. K. Sheils *et al.*, “TCRD and Pharos 2021: mining the human proteome for disease biology,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1334–D1346, Jan. 2021, doi: 10.1093/nar/gkaa993.