Multiomics Data Integration for Gene Regulatory Network Inference with Exponential Family Embeddings

Surabhi Jagtap, Abdulkadir Çelikkanat, Aurélie Pirayre, Frederique Bidard, Laurent Duval, Fragkiskos D. Malliaros

Multiomics Data Integration for Gene Regulatory Network Inference with Exponential Family Embeddings

Surabhi Jagtap*†, Abdulkadir Çelikkanat†, Aurélie Pirayre*, Frederique Bidard*,
Laurent Duval*, Fragkiskos D. Malliaros†
*IFP Energies nouvelles, Rueil-Malmaison, France
Email: {surabhi-vasantrao.jagtap, aurelie.chataignon, frederique.bidard-michelot, laurent.duval}@ifpen.fr
†Université Paris-Saclay, CentraleSupélec, Inria, Gif-Sur-Yvette, France
Email: {abdulkadir.celikkanat, fragkiskos.malliaros}@centralesupelec.fr

***Abstract*—The advent of omics technologies have enabled the generation of huge, complex, heterogeneous, and high-dimensional omics data. Imposing numerous challenges in data integration, these data could lead to a better understanding of the organism's cellular system. Omics data are typically represented as networks to study relations between biological entities, such as protein-protein interaction, gene regulation, and signal transduction. To this end, network embedding approaches allow us to learn latent feature representations for nodes of a graph structure. In this study, we propose a new methodology to learn embeddings by modeling the underlying interactions among biological entities (nodes) with exponential family distributions from a well chosen set of omics modalities. We evaluate our proposed method based on the gene regulatory network (GRN) inference problem. As the ground truth for evaluation, we use GRN available in public databases and demonstrate its effectiveness by comparing to other network integration approaches.**

***Index Terms*—omics data integration, multilayer network, random walks, network embedding, representation learning**

I. Introduction

The organism's cellular system is composed of multiple interacting entities like genes, proteins, and metabolites that are associated with different biological mechanisms. The advent of high throughput technologies has enabled us to study the relationships of these entities by analyzing huge omics data. Yet, independent analysis of such data is limited to correlations, mostly reflecting reactive processes rather than causative ones. Integration of different omics data types is often expected to elucidate potential causative changes such as gene regulation or signal transduction that lead to specific phenotypes or treatment targets. To this direction, networks are widely used to represent biological relations (edges) between individual entities (nodes) [1], [2]. The primary challenge here is to properly represent these networks in a way that they can effectively be used as input to machine learning models to perform downstream prediction tasks. In this paper, we are inspired by graph representation learning (GRL) algorithms that allow us to encode the graph structure into compact embedding vectors [3]. As a target application, we focus on the task of GRN inference, which will be described in detail in Sec. II. In particular, we aim to embed graph-based multiomic information in a lower-dimensional space towards inferring edges of GRN.

Although there is a plethora of graph representation learning approaches based on random walks [4], [5], matrix factorization, and neural networks [3], they have mostly been introduced for single-layer networks [6]. Nevertheless, for biological networks there are only a few existing network integration strategies that leverage GRL. As one of the first proposed models, MASHUP [7] is a network integration framework based on matrix factorization that builds compact low-dimensional vector representation of proteins. More recently, DEEPNF [8] is a network fusion method based on multimodal deep autoencoders. Both approaches consider a set of input networks and for each network they construct vector representations of proteins. Yet, they lack to extract features that represent the relationships of entities in a network and across all networks which could be essential while dealing with multiomics data, where information naturally flows from one layer to another. In such cases, it is necessary to obtain an informative representation of the nodes and their proximity that is not fully captured by features that are extracted directly from single input network.

Motivated by the aforementioned limitation of current network integration models, in this paper we consider expressive conditional probability models to relate nodes within random walk sequences, towards extracting informative latent node representations. We capitalize on exponential family distributions to capture interactions between nodes in random walks that traverse nodes within and across input network layers. More precisely, we introduce network integration with the concept of exponential family graph embeddings [9], that generalizes multilayer random walk-based GRL methods to an instance of exponential family conditional distribution. Our method has the following conceptual advances: (i) it preserves both the intra-layer and inter-layer interactions, thereby learning rich features; (ii) it is an effective and scalable method as it uses the conditional probability distribution model to learn low-dimensional node features from all input networks. Besides, we define the objective function of the proposed

model in a way that is independent of downstream machine learning tasks, and the embeddings are learned in a purely unsupervised way. Although in this paper we aim to study the problem of gene regulation, the learned embeddings could be leveraged across a wide variety of omics integration tasks.

Source code: The source code and datasets are available at: https://github.com/Surabhivj/BRANE-EXP.

II. Problem Statement

Gene regulatory networks (GRN) impart how signals propagate through biomolecule pathways and result in transcriptomic modifications. These regulatory networks are computational modules of a biological system that carry out decision-making processes. They enable us to determine the ultimate response of an organism to a stimulus. Although there is an intense research effort on GRN inference using gene expression for more than a decade and much progress has been made, it remains a challenging problem [10], [11]. Even the most sophisticated inference techniques are far from the perfect. Mainly by leveraging gene co-expression networks with the relationships between regulators, such as transcription factors (TFs) and the target genes they control, one can achieve a better understanding of regulatory interactions, providing us the access points to modulate such responses [12], [13]. However, it is a challenging task to effectively combine this information in such a way that the rich and relevant features from the input datasets are preserved. Indeed, recent breakthroughs in graph representation learning have inspired us to solve this GRN inference task by modeling heterogeneous datasets as a multilayer graph and encode latent representations for them.

Here, we propose a novel GRN inference framework by integrating gene co-expression and TF-target networks. We formally define the task as a network embedding problem. Given a set of networks based on omics data, we aim to learn low-dimensional latent node representations (i.e., embeddings) so that the structure of the input networks is properly integrated and preserved in the new space. In other words, we aim to learn representations such that functionally related genes or co-regulated genes are placed close enough in the embedding space. Further, we define a similarity score for these embedding vectors in order to infer an integrated network. More formally, let $\mathcal{G} = (V_i, E_i)_{i=1}^{\mathcal{L}}$ be a multilayer graph, where V_i and E_i are the set of nodes and edges at layer i respectively, and $\mathcal{L}$ is the number of layers. Our model encodes $|\mathcal{V}|$ nodes, where $\mathcal{V}$ is defined as $V_1 \cap V_2 \cap \cdots \cap V_{\mathcal{L}}$. We first explore the multilayer neighborhood by simulating random walks per node. Based on the obtained node contexts, we learn a d-dimensional feature vector of each node, where $d \ll |\mathcal{V}|$.

III. Proposed Method

In this section, we first describe how relevant node pairs are sampled with random walks—a key step towards multiomics data integration. Then, we explain the methodology employed to learn node representations by modeling the underlying interactions among nodes with exponential family distributions.

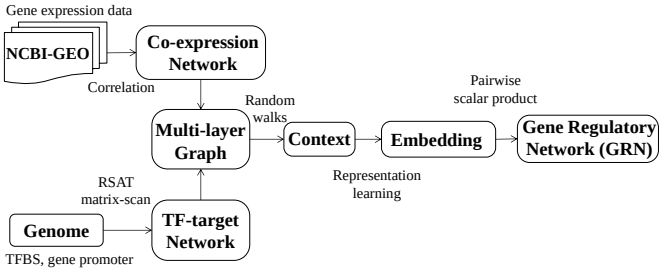


Fig. 1: The overall workflow of multinetwork integration for GRN inference using graph representation learning.

A general overview of the proposed methodology is depicted in Figure 1.

A. Context sampling using multilayer biased random walks

A multi-layer network can consist of various layers showing diverse characteristics so that each layer might possess a completely different topological structure. Our goal is to learn representations in a lower-dimensional space so that their embeddings reflect the underlying patterns commonly shared by these network layers. Therefore, we leverage random walks to sample nodes sharing similar characteristics across different omics layers. Although random walks have been used before in representation learning [14], here we introduce a flexible approach for multilayer network structures.

To extract nodes' context in a multilayer graph, we propose a random walk-based sampling procedure that can explore local and global structures in the graph. Local exploration helps in discovering the clustering structure around the node of interest, whereas global exploration contributes in capturing global associations within nodes in the graph [15]. This point is particularly important within the biological context of our application domain. For instance, the relationship between a TF and its target is a local structure, while functional relationships among the co-expressed genes are more related to global structure in the graph (see also Figure 2).

To capture such local and global associations, we propose a *biased* version of random walks adapted to multilayer graphs [15]. It combines both types of exploration (i.e., local and global) with a decay parameter α to control the importance of nodes with respect to their distance from the node of interest. More formally, for each node $v_i \in \mathcal{V}$, a proximity score τ_{v_i} is computed to estimate how far the candidate node v_i is from the source node. When the i-th node in the walk is discovered, the proximity score of every node adjacent to that is increased by α^{i-1} and $\bar{\alpha}^{i-1}$, for nodes in the same and different layer respectively, where α, $\bar{\alpha} \in [0, 1]$. Then, the probability distribution of selecting the next node for the current walk is computed based on the proximity scores of the neighborhood nodes of the most recently visited node. For local explorations, the probability of a node being the next one in the random walk sequence should be proportional to its proximity score, i.e., $p_{v_i} = \frac{\tau_{v_i}}{\sum_{w \in \mathcal{V}_{(u)}} \tau_w}$. In the case of global exploration, the probability is set to be inversely proportional

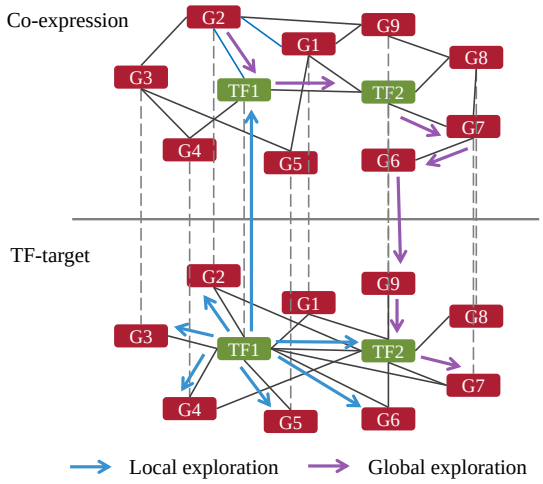


Fig. 2: Context sampling for a multilayer graph adapted to explore local and global structures in a graph.

to that score, i.e., $p_{v_i} = \frac{1/\tau_{v_i}}{\sum_{w \in \mathcal{V}_{(u)}} 1/\tau_w}$, where u is the most recently visited node, and $\mathcal{V}(u)$ defines the set of neighbors of u. Thus, given a desired exploration strategy, the context set for each node $\mathbf{w}_{(n,i)} \in \mathcal{V}$ is given by $C_\gamma(\mathbf{w}_{(n,i)}) := \{\mathbf{w}_{(n,j)} \in \mathcal{V} : -\max\{1, i-\gamma\} \leq j \neq i \leq \min\{i+\gamma, l\}\}$, where $\mathbf{w}_{(n,j)}$ indicates the node appearing at the j-th position of the n-th random walk, and γ is the window size. We call each element of $C_\gamma(\mathbf{w}_{(n,i)})$ as the *context* of a *center* node $\mathbf{w}_{(n,i)}$.

B. Learning embeddings with exponential family distributions

Random walk-based methods generate node sequences and learn node representations by maximizing the co-occurrence probability of nodes within a certain distance [4], [5], [15]. Similarly, we define our objective function as follows:

$$\mathcal{O}(\alpha, \beta) := \arg\max_{\Omega=(\alpha,\beta)} \sum_{n=1}^{N} \sum_{i=1}^{l} \sum_{u \in \mathcal{V}} \log p(\mathbf{x}^u_{n,i}; \Omega), \quad (1)$$

where $\mathbf{x}^u_{n,i}$ is the observed variable indicating the relationship between the pair of nodes $(\mathbf{w}_{(n,i)}, u) \in \mathcal{V}^2$, and $\Omega = (\alpha, \beta)$ are the parameters of the model, which correspond to the node embedding vectors. Note that, we obtain two different representations for each node. Here, $\alpha[v]$ indicates the embedding of node v if it is considered as *context*, and $\beta[v]$ denotes its representation if it is interpreted as *center* node. Although the conventional choice for modeling node relationships is the *softmax* function [4], it limits capturing possible intricate patterns in node interactions across the layers of the network structure. Therefore, here we extend it with a general framework based on *exponential families*, which is a set of parametric probability distributions satisfying the following form:

$$p(\mathbf{x}) = h(\mathbf{x}) \exp\Big(\eta T(\mathbf{x}) - A(\eta)\Big), \quad (2)$$

where $h(\mathbf{x})$ is the base measure, $T(\mathbf{x})$ is the sufficient statistic, and $A(\eta)$ is the log-normalizer function. Note that, many widely used distributions, such as the ones of *Bernoulli*, *Dirichlet*, and *Normal*, are actually exponential families. The main benefit of this generic formulation is that it provides an elegant and flexible way to model the complex interactions between center and context nodes in random walk sequences [9], [16]. By plugging the exponential form into the objective function provided in Eq. (1), we obtain the following:

$$\arg\max_{\Omega=(\alpha,\beta)} \sum_{n=1}^{N} \sum_{i=1}^{l} \sum_{u \in \mathcal{V}} \log h\big(\mathbf{x}^u_{n,i}\big) + \eta^u_{\mathbf{w}_{(n,i)}} T\big(\mathbf{x}^u_{n,i}\big) - A\big(\eta^u_{\mathbf{w}_{(n,i)}}\big).$$

Here, we define the natural parameter $\eta^u_{\mathbf{w}_{(n,i)}}$ as the product of embeddings, $\alpha[u]^\top \cdot \beta[\mathbf{w}_{(n,i)}]$.

In our approach, we employ the Bernoulli distribution to model node co-occurrences, by setting $h(\mathbf{x}) = 1$, $T(\mathbf{x}) = x$ and $A(\eta) = \log(1 + e^\eta)$. Let $X^u_{n,i}$ be a Bernoulli random variable indicating the occurrence of u in the context of node $\mathbf{w}_{(n,i)}$. Note that, this is equal to 1 if node u appears at any position index $i + j$, for $-\gamma \leq j \neq 0 \leq \gamma$. Then, we can rewrite our objective function as follows:

$$\arg\max_{\Omega=(\alpha,\beta)} \sum_{n=1}^{N} \sum_{i=1}^{l} \sum_{|j| \leq \gamma} \Bigg(\underbrace{\log p\Big(y^{\mathbf{w}_{(n,i+j)}}_{n,i+j}\Big)}_{\text{positive instances}} + \underbrace{\sum_{u \in \mathcal{V} \setminus \{\mathbf{w}_{(n,l+j)}\}} \log p\Big(y^u_{n,i+j}\Big)}_{\text{negative instances}} \Bigg),$$

where $y^u_{n,i+j}$ indicates the occurrence of node u at the $(i+j)$-th position of the n-th walk. However, the optimization step is very costly due to the size of the negative instances. Therefore, we approximate it by leveraging the *negative sampling* approach [17]:

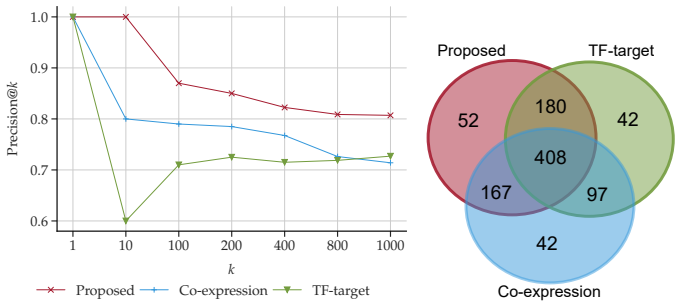
$$\arg\max_{\Omega=(\alpha,\beta)} \sum_{n=1}^{N} \sum_{i=1}^{l} \sum_{|j| \leq \gamma} \Bigg(\log p\Big(y^{\mathbf{w}_{(n,i+j)}}_{n,i+j}\Big) + k \mathop{\mathbb{E}}_{u \sim p^-} \Big[\log p\Big(y^u_{n,i+j}\Big) \Big] \Bigg)$$

where k indicates the number of negative instances sampled from the noise distribution p^-. We employ the strategy described in [17] and negative instances are sampled from the whole vertex set with respect to the their number of occurrences in the generated walks raised to the power of 0.75. In the experimental evaluation, we generate $k = 5$ negative samples for each positive instance.

IV. Results

A. Experimental setup

Gene regulation could be examined by two major evidences. First, co-expression of genes and second, binding of TF to the TF binding site (TFBS) on the promoter of genes. The overall workflow of GRN inference is shown in Figure 1. We test our framework with the well-studied organism, *Saccharomyces cerevisiae S288c*. First, we build a gene co-expression network from (61) microarray experiments [18] using correlation information. Two genes are connected by an edge if the correlation between them is greater than 0.9 [19]. Then, we infer TF-target relationships by using TFBS deposited in JASPAR database [20] and promoter sequences of genes [21]. We scan the TFBS in the promoters using the *matrix-scan* tool in RSAT [22],



(a) Precision@k (b) Precision statistics

Fig. 3: Integrating networks outperform GRN inference when compared to individual ones. (a) Precision@k for the integrated network inferred by our approach (Proposed), TF-target, and co-expression network. (b) For the top $1,000$ edges from each network, we show the comparison of truly inferred edges.

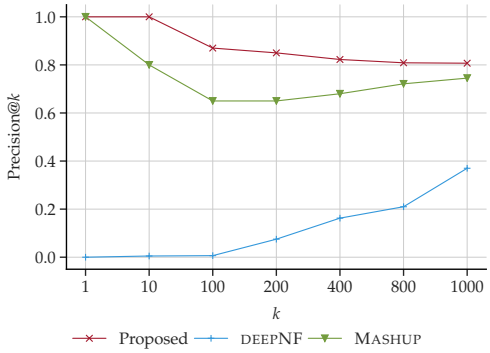


Fig. 4: Performance of our method in inferring the GRN, measured by the Precision@k for the top $1,000$ edges, compared to MASHUP and DEEPNF.

and we infer the edges based on the presence of binding sites in the promoter of the gene. For the same set of nodes, a multilayer network is constructed from the co-expression and TF-target networks. Considering it as input to our model, we learn node embeddings using the methodology described in Sec. III (Figure 1). The random walk parameters, such as the walk length, number of walks, and proximity score are set to 10, 20, and 1, respectively. The rest parameters for computing embeddings, i.e., the learning rate, window size, and embedding dimension are set to 0.025, 5, and 128, respectively. The same pipeline and embedding dimension size are used for the baselines MASHUP and DEEPNF. To infer regulatory interaction from the learned embeddings, we define the similarity between the embedding vectors by computing the scalar product for each TF-gene interaction.

B. Evaluation

We investigate the ability of our model to infer regulatory interactions by reconstructing the GRN for *Sacharomyces cerevsiae* [23]. In practice, biological networks show small-world properties, where nodes are linked by a short chain of acquaintances. These properties could be extracted by focusing on important edges in the graph. In our context of binary inference, the precision metric computes the accuracy to retrieve correctly inferred edges. Therefore, to evaluate the performance of graph inference and to retrieve such relevant information, we measure the precision at top k inferred edges (Precision@k), that corresponds to the number of correctly inferred edges among the top k ones [24]. We choose to study the top $1,000$ edges of the inferred GRN.

Our evaluation strategy aims to demonstrate the added value of multiomics data integration in the problem of GRN inference, and the performance of multilayer network integration with respect to state-of-art methods. Firstly, to show the added value of integrating co-expression and TF-target networks in GRN inference, we compare the reconstruction performance of our method applied on individual networks. We have observed that integration outperforms the reconstruction for the top $1,000$ edges when compared to individual ones, as shown in Figure 3a. Moreover, from the top $1,000$ inferred edges after integration with the proposed model and from those inferred from individual networks, we observe that the integration process inferred 807 true edges whereas TF-target and co-expression networks alone inferred 727 and 714 true edges, respectively (Figure 3b). From these 807 correctly inferred edges, 52 are novel (i.e., not present in the input networks). This implies that during integration relevant and rich features are captured that are not directly seen in the individual input graphs. The visualization [25] of these truly inferred 807 edges is shown in Figure 5.

Second, we compare the performance of our model to two state-of-the-art network integration methods, namely MASHUP and DEEPNF. As shown in Figure 4, our method is able to outperform DEEPNF, inferring a larger number of actual edges than both state-of-the-art integration methods for the selected network size. Although we are mainly interested to study the most important edges, we withdraw the network size bias and further measure the performance of our model for all edges. To do this, we have computed the area under the Precision-Recall curve (AUPR), observing that the AUPR of the proposed model (0.734) is very similar to the one of MASHUP (0.735), and it outperforms DEEPNF (0.641).

To summarize the empirical analysis, the performance of our model (Figures 3 and 4) is especially appealing mainly because of three reasons. First, our approach shows that it can generate meaningful embeddings by preserving the inter- and across-layer interactions. Second, its objective function is independent of the downstream task (i.e., GRN inference), thereby our method is adaptable to various omics data analysis tasks. Third, it is a simple method and requires less parameter tuning as compared to other neural network-based embedding methods, which could reduce overfitting.

V. CONCLUSION

The recent wide application of high-throughput experimental techniques has provided complex, high-dimensional, and heterogeneous data. Their wide availability has, in turn, driven a need for integrative methods which could be utilized to

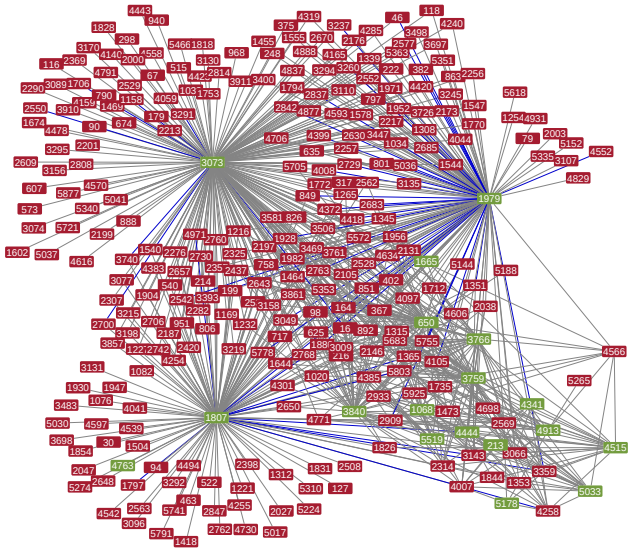


Fig. 5: Top $1,000$ true inferred edges (also in the reference network). Nodes in red correspond to genes, while TFs are indicated with green. The edges in blue are the newly inferred 52 true edges, which are not present in the input networks.

explore them effectively. In this paper, we have presented a graph embedding-based network integration method for GRN inference, by constructing a compact low-dimensional gene feature representation from the selected omics modalities. We have performed an empirical analysis of the proposed approach, comparing it to state-of-art baseline methods, and showing its ability to infer important edges for gene regulation.

Our method is not limited to only GRN inference. There are several directions for future work with focus on extending the integration to include other data types, such as protein sequences, protein structures, and epigentic marks, toward making more accurate predictions of gene regulations, TF targets, and protein functions.

Acknowledgements. Supported in part by ANR (French National Research Agency) under the JCJC project GraphIA (ANR-20-CE23-0009-01).

References

[1] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman, "Graphs in molecular biology," *BMC Bioinformatics*, vol. 8, no. S8, 2007.

[2] B. Lee, S. Zhang, A. Poleksic, and L. Xie, "Heterogeneous multi-layered network model for omics data integration and analysis," *Frontiers in genetics*, vol. 10, p. 1381, 2020.

[3] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Data Eng. Bull.*, vol. 40, no. 3, pp. 52–74, 2017.

[4] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, 2014, pp. 701–710.

[5] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, 2016, pp. 855–864.

[6] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S. M. Lin, W. Zhang, and P. Zhang, "Graph embedding on biomedical networks: methods, applications and evaluations," *Bioinformatics*, vol. 36, pp. 1241–1251, Feb. 2020.

[7] H. Cho, B. Berger, and J. Peng, "Compact integration of multi-network topology for functional analysis of genes," *Cell systems*, vol. 3, no. 6, pp. 540–548, 2016.

[8] V. Gligorijević, M. Barot, and R. Bonneau, "deepNF: deep network fusion for protein function prediction," *Bioinformatics*, vol. 34, no. 22, pp. 3873–3881, 2018.

[9] A. Celikkanat and F. D. Malliaros, "Exponential family graph embeddings," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3357–3364.

[10] A. Pirayre, C. Couprie, F. Bidard, L. Duval, and J.-C. Pesquet, "Brane Cut: biologically-related a priori network enhancement with graph cuts for gene regulatory network inference," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–12, 2015.

[11] A. Pirayre, C. Couprie, L. Duval, and J.-C. Pesquet, "Brane Clust: Cluster-assisted gene regulatory network inference refinement," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 850–860, 2017.

[12] N. Ksouri, J. A. Castro-Mondragón, F. Montardit-Tardà, J. van Helden, B. Contreras-Moreira, and Y. Gogorcena, "Motif analysis in co-expression networks reveals regulatory elements in plants: The peach as a model," *Plant Physiology*, 2021.

[13] M. G. Van Der Wijst, D. H. de Vries, H. Brugge, H.-J. Westra, and L. Franke, "An integrative approach for building personalized gene regulatory networks for precision medicine," *Genome medicine*, vol. 10, no. 1, pp. 1–15, 2018.

[14] F. Chen, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "Graph representation learning: a survey," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[15] D. Nguyen and F. D. Malliaros, "BiasedWalk: Biased sampling for representation learning on graphs," in *IEEE Int. Conf. on Big Data (Big Data)*, 2018, pp. 4045–4053.

[16] M. Rudolph, F. Ruiz, S. Mandt, and D. Blei, "Exponential family embeddings," in *NIPS*. Curran Associates Inc., 2016, pp. 478–486.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[18] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets—update," *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2012.

[19] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, and D. Zhi, "Gene2vec: distributed representation of genes based on co-expression," *BMC Genomics*, vol. 20, no. 1, pp. 7–15, 2019.

[20] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. Van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić *et al.*, "JASPAR 2020: update of the open-access database of transcription factor binding profiles," *Nucleic acids research*, vol. 48, no. D1, pp. D87–D92, 2020.

[21] S. R. Engel, F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan, M. C. Costanzo, S. S. Dwight, B. C. Hitz, K. Karra, R. S. Nash *et al.*, "The reference genome sequence of saccharomyces cerevisiae: then and now," *G3: Genes, Genomes, Genetics*, vol. 4, no. 3, pp. 389–398, 2014.

[22] N. T. T. Nguyen, B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, R. Ossio, C. D. Robles-Espinoza, M. Bahin, S. Collombet, P. Vincens, D. Thieffry *et al.*, "RSAT 2018: regulatory sequence analysis tools 20th anniversary," *Nucleic acids research*, vol. 46, no. W1, pp. W209–W214, 2018.

[23] P. T. Monteiro, J. Oliveira, P. Pais, M. Antunes, M. Palma, M. Cavalheiro, M. Galocha, C. P. Godinho, L. C. Martins, N. Bourbon *et al.*, "Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts," *Nucleic acids research*, vol. 48, no. D1, pp. D642–D649, 2020.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, "scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[25] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.