RESEARCH

Open Access

Kernelized multiview signed graph learning for single-cell RNA sequencing data

Abdullah Karaaslanli[1*†], Satabdi Saha[2†], Tapabrata Maiti[3] and Selin Aviyente[1]

†Abdullah Karaaslanli and Satabdi Saha have contributed equally to this work

*Correspondence: karaasl1@msu.edu

[1] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA
[2] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3] Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

Abstract

Background: Characterizing the topology of gene regulatory networks (GRNs) is a fundamental problem in systems biology. The advent of single cell technologies has made it possible to construct GRNs at finer resolutions than bulk and microarray datasets. However, cellular heterogeneity and sparsity of the single cell datasets render void the application of regular Gaussian assumptions for constructing GRNs. Additionally, most GRN reconstruction approaches estimate a single network for the entire data. This could cause potential loss of information when single cell datasets are generated from multiple treatment conditions/disease states.

Results: To better characterize single cell GRNs under different but related conditions, we propose the joint estimation of multiple networks using multiple signed graph learning (scMSGL). The proposed method is based on recently developed graph signal processing (GSP) based graph learning, where GRNs and gene expressions are modeled as signed graphs and graph signals, respectively. scMSGL learns multiple GRNs by optimizing the total variation of gene expressions with respect to GRNs while ensuring that the learned GRNs are similar to each other through regularization with respect to a learned signed consensus graph. We further kernelize scMSGL with the kernel selected to suit the structure of single cell data.

Conclusions: scMSGL is shown to have superior performance over existing state of the art methods in GRN recovery on simulated datasets. Furthermore, scMSGL successfully identifies well-established regulators in a mouse embryonic stem cell differentiation study and a cancer clinical study of medulloblastoma.

Keywords: Gene regulatory networks, Single cell, Graph signal processing, Graph learning

Background

Gene expression arises from a network of regulatory interactions between transcription factors, co-factors and signaling molecules [1, 2]. Elucidating the topology of this underlying transcriptomic network is essential for understanding the mechanisms that govern complex biological processes in human physiology and pathology. Identifying the differences in transcriptional regulation between normal and disease states helps in revealing the specific biological and biochemical pathways relevant to disease mechanisms and progression [3, 4].

A major focus area in clinical research lies in studying the changes in gene coexpression networks across different tissues, cell types/states, and conditions. For example, in the extensively studied breast cancer datasets from the cancer genome atlas, there are four main subtypes of breast cancer [5]. The variation between these subtypes holds the key to inferring how genes transcriptionally regulate each other and how their expressions and interactions change across subgroups. In addition one would expect the gene relationships corresponding to different subtypes to be similar to each other since they originate in the same tissue, but also posses crucial differences since they are in different stages of disease progression [6–8]. Thus, instead of estimating a single network for all the subtypes, constructing class-specific graphical models for different conditions will provide a more robust and deeper understanding of group-specific characteristics.

Recent advances in next generation sequencing technologies have made it possible to profile the transcriptomes of individual cells, hence capturing expressions of thousands of genes at a cellular resolution. Dozens of algorithms have been proposed for the reconstruction of gene regulatory networks from single cell RNA sequencing (scRNA-seq) datasets [9, 10]. This has further enabled novel insights into the transcriptional regulation underlying various biological processes, including cancer progression [11] and embryonic development [12]. Most of these algorithms, however estimate a single gene regulatory network, assuming the data samples to be identically and independently distributed; hence ignoring the presence of natural subgroups that may be present within the data. Given the assumption of a grouped dataset, one should be able to apply these algorithms to estimate networks from each subgroup separately; but this procedure of independent group-wise network estimation will fail to model the shared structures between the subgroups, eventually leading to information loss. Therefore, there is a pressing need to develop joint graph estimation models that would allow information borrowing across subgroups while retaining subgroup specific heterogenity.

Multiple algorithms have been proposed for joint estimation of networks from high dimensional data. Most of these methods assume that the data has a Gaussian distribution. Seminal papers by [6, 13] paved the way for penalized estimation of multiple Gaussian graphical models, and demonstrated the use of lasso based penalty functions for better estimation across multiple groups. Later, Danaher et al. [6] proposed the fused graphical lasso and the group graphical lasso penalties for better estimation. These methods however are not directly applicable to single cell datasets. Despite many advantages, scRNA-seq datasets are undermined by a series of technical limitations, such as drop-out events (expressed genes undetected by scRNA-seq) and a high level of noise, which renders void the assumption of gaussianity [14–16]. Few methods have been proposed for joint estimation of multiple networks from scRNA-seq datasets. Mukherjee et al. [17] developed PIPER, a penalized local Poisson graphical model [18] for joint estimation of multiple networks in scRNA-seq datasets. One of the main limitations of PIPER is that the Poisson distribution has one single parameter characterizing both the mean and the standard deviation. Single cell datasets would be better characterized by a negative binomial distribution which has a separate dispersion parameter or a zero inflated negative binomial distribution which could account for the excessive zeroes. To account for the non-Gaussian nature of the scRNA-seq datasets, Wu et al. [19] proposed a modification of the joint Gaussian copula graphical model based on the Gaussian copula transformation proposed in [20]. To facilitate estimation of Kendall's τ correlation matrix in the presence of dropouts they propose a modified Kendall's τ metric that only utilizes the completely observed values, and excludes the missing values. Dong et al. [21] proposed a three step hybrid joint estimation strategy that relies on (a) integrated application of a Bayesian zero inflated Poisson based model imputation strategy and single cell imputation technique McImpute [22, 23], (b) data Gaussianization [24] and eventually (c) joint estimation of a Gaussian graphical model [6]. Contrary to [17], the last two proposed approaches estimate graphical models for continuous data and rely on a data transformation step for making the data continuous.

Recent work in graph signal processing (GSP) extends classical signal processing concepts to data defined on nodes of a graph, i.e. *graph signals* [25]. GSP based graph learning (GL) approaches infer the graph structure from the observed graph signals based on assumptions made about the relation between the signals and the unknown graph [26]. Since graph signals are represented explicitly in the graph frequency domain, GSP based GL has more flexibility in modeling signals compared to previous network inference methods, such as statistical models reviewed above for GRN inference. Therefore, in this work, we focus on GSP based GL for the joint inference of multiple GRNs, where gene expressions from cells are considered as graph signals on the unknown GRNs. Existing GL algorithms [27–30] have two important shortcomings for multiple GRN learning. First, they cannot learn signed graphs, which is a more suitable model for GRNs as they include activating and inhibitory edges. Second, with the exception of [30], they can only learn a single graph. Thus, they are not applicable to the joint inference of multiple GRNs problem.

In this paper, we present a multiple signed graph learning algorithm (scMSGL) for joint inference of GRNs from multiple classes (conditions/disease states). Based on the method developed in [31], scMSGL learns multiple GRNs by deriving an optimization problem using three assumptions: (i) expressions of genes connected with activating edges are similar to each other, (ii) expressions of genes connected with inhibitory edges are dissimilar to each other, and (iii) GRNs corresponding to the different datasets are related to each other. Thus, scMSGL optimizes the total variation of graph signals to learn signed graphs while ensuring that the learned signed graphs are similar to each other through regularization with respect to a learned signed consensus graph. The proposed method has several advantages over existing approaches. First, it performs joint GRN inference taking advantage of the shared information across datasets while not making any specific parametric assumptions about the data. Second, during application to single cell data, scMSGL is kernelized as in [31] to take the structure of scRNA-seq data into account. For instance, it can employ proportionality measures to reflect relative rather than absolute abundance or zero-inflated Kendall's tau to handle drop-outs [32]. Finally, the proposed method learns an additional consensus graph, which captures the common structure across all graphs.

Methods

Graphs

A weighted undirected graph can be denoted as $G = (V, E, \mathbf{W})$ where V is the node set with $|V| = n$, E is the edge set and $\mathbf{W}$ is the adjacency matrix with W_{ij} the weight of the edge between nodes i and j. G is an unsigned graph, if edge weights are constrained to only positive values. Combinatorial Laplacian matrix of the unsigned graph G is $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D}$ is the diagonal matrix with node degrees, i.e. $D_{ii} = \sum_{j=1}^{n} W_{ij}$. The Laplacian matrix is positive semi-definite, thus its eigendecomposition is $\mathbf{L} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. We assume the eigenvalues of $\mathbf{L}$ are ordered such that $0 = \Lambda_{11} \leq \Lambda_{22} \leq \cdots \leq \Lambda_{nn}$.

If the edge weights are allowed to take on negative values, G is a signed graph. A signed graph G can be decomposed into two unsigned graphs, $G^+ = (V, E^+, \mathbf{W}^+)$ and $G^- = (V, E^-, \mathbf{W}^-)$, where $W^+_{ij} = W_{ij}$ ($W^-_{ij} = |W_{ij}|$) if $W_{ij} > 0$ ($W_{ij} < 0$), and 0, otherwise.

Graph signals

A graph signal over an unsigned graph G is a function $x : V \rightarrow \mathbb{R}$ and can be represented by a vector $\mathbf{x} \in \mathbb{R}^n$ where each x_i is the signal value on node i. Graph Fourier transform (GFT) of $\mathbf{x}$ can be defined using the spectrum of $\mathbf{L}$ as its eigenvalues and eigenvectors provide a notion of frequency, i.e., small eigenvalues correspond to low frequencies and large ones to high frequencies [25, 33]. GFT of $\mathbf{x}$ is defined as $\widehat{\mathbf{x}} = \mathbf{V}^\top\mathbf{x}$ and inverse GFT is $\mathbf{x} = \mathbf{V}\widehat{\mathbf{x}}$. Thus, $\mathbf{x}$ is the linear combination of eigenvectors of $\mathbf{L}$ with the coefficients determined by $\widehat{\mathbf{x}}$. If most of the energy of $\widehat{\mathbf{x}}$ is concentrated in the entries corresponding to the small eigenvalues, $\mathbf{x}$ has a low-frequency representation in the graph Fourier domain. On the other hand, if its energy is concentrated in the entries corresponding to large eigenvalues, it has a high-frequency representation. The total variation of $\mathbf{x}$ with respect to G can then be quantified as:

$$\mathrm{tr}(\widehat{\mathbf{x}}^\top \boldsymbol{\Lambda}\widehat{\mathbf{x}}) = \mathrm{tr}(\mathbf{x}^\top \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathrm{T}\mathbf{x}) = \mathrm{tr}(\mathbf{x}^\top \mathbf{L}\mathbf{x}), \tag{1}$$

whose small(large) values indicate that $\mathbf{x}$ has low(high) frequency representation on G.

For a graph signal $\mathbf{x}$ defined on a signed graph G, total variation can be quantified based on the decomposition of G into G^+ and G^-. Namely, let $\mathbf{L}^+$ and $\mathbf{L}^-$ be the Laplacian matrices of G^+ and G^-, then we define two total variation values for $\mathbf{x}$: $\mathrm{tr}(\mathbf{x}^\top\mathbf{L}^+\mathbf{x})$ and $\mathrm{tr}(\mathbf{x}^\top\mathbf{L}^-\mathbf{x})$.

Single view signed graph learning

Given a set of graph signals $\{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^p$ that are defined on an unknown unsigned graph G, Dong et al. [27] proposed to learn the structure of G with the assumption that signals admit low-frequency representation in the graph Fourier domain of G. Thus, G can be learned by minimizing (1) with respect to $\mathbf{L}$ as follows:

$$\underset{\mathbf{L}\in\mathbb{L}}{\text{minimize}}\ \mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) + \alpha\|\mathbf{L}\|_\mathrm{F}^2 \text{ subject to } \mathrm{tr}(\mathbf{L}) = 2\mathrm{n}, \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the data matrix whose columns are $\mathbf{x}_i$'s, $\mathbb{L} = \{\mathbf{L} : L_{ij} = L_{ji} \leq 0\ \forall i \neq j,\ \mathbf{L}\mathbf{1} = \mathbf{0}\}$ is the set of valid Laplacian matrices. The first term in (2) measures the total variation of the graph signals. The second term is the Frobenius norm of $\mathbf{L}$ and controls the density of the learned graph. Finally, the last constraint is added to prevent the trivial solution $\mathbf{L} = \mathbf{0}$.

In [31], (2) is extended to learn an unknown signed graph G based on the assumption that the graph signals admit (i) low-frequency (smooth) representation over G^+, and (ii) high-frequency (nonsmooth) representation over G^-. Smoothness and non-smoothness of the graph signals with respect to signed graphs are defined as follows: (1) Signal values on nodes that are connected with positive edges are similar to each other; (2) Signal values on nodes that are connected with negative edges are dissimilar to each other. These assumptions imply that if genes i and j are connected by an activating edge, their gene expressions should be similar, i.e low-frequency. On the other hand, if i and j are connected by an inhibitory edge, their expressions should be dissimilar, i.e., high frequency. These assumptions are biologically reasonable and have been validated in [31]. Based on these assumptions, the signed graph G is learned by minimizing $\text{tr}(\mathbf{X}^\top\mathbf{L}^+\mathbf{X})$ with respect to $\mathbf{L}^+$ and maximizing $\text{tr}(\mathbf{X}^\top\mathbf{L}^-\mathbf{X})$ with respect to $\mathbf{L}^-$. This leads to the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{L}^+\in\mathbb{L},\mathbf{L}^-\in\mathbb{L}}{\text{minimize}} \sum_{s\in\{+,-\}} \text{tr}(\mathbf{K}^s\mathbf{L}^s) + \alpha_s\|\mathbf{L}^s\|_F^2 \\ &\text{subject to } \ \text{tr}(\mathbf{L}^s) = 2n\ \forall s, \text{ and } (\mathbf{L}^+,\mathbf{L}^-) \in \mathbb{C}, \end{aligned} \tag{3}$$

where $\mathbf{K}^+ = \mathbf{X}\mathbf{X}^\top$, $\mathbf{K}^- = -\mathbf{X}\mathbf{X}^\top$ and we used the cyclic property of trace operation, i.e. $\text{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) = \text{tr}(\mathbf{X}\mathbf{X}^\top\mathbf{L})$. $\mathbf{L}^+$ and $\mathbf{L}^-$ are constrained to be in the set $\mathbb{C} = \{(\mathbf{L}^+,\mathbf{L}^-) : L^+_{ij} = 0 \text{ if } L^-_{ij} \neq 0 \text{ and } L^-_{ij} = 0 \text{ if } L^+_{ij} \neq 0\}$ to ensure that they are not non-zero at the same indices.

The optimization problem in (3) can be kernelized to exploit various (nonlinear) relations between graph signals. Kernelization is important for GRN inference as it is unclear which association measure between gene expressions is best for various scRNA-seq data analysis [32]. Therefore, (3) is kernelized by changing $\mathbf{X}\mathbf{X}^\top$ with any positive semi-definite kernel matrix. In [31], three kernels are considered: correlation coefficient (r), proportionality measure (ρ) [34] and zero-inflated Kendall's tau (τ_{zi}) [35]. These kernels are considered because r is a commonly used metric for network inference, ρ is found as the best performing measure in [32] and τ_{zi} can handle the dropouts in scRNA-seq.

Multiview signed graph learning

Let $\{\mathbf{X}^i\}_{i=1}^N$ be a given set consisting of N matrices. $\mathbf{X}_i \in \mathbb{R}^{n\times p_i}$ is a data matrix constructed from p_i graph signals defined on an unknown signed graph $G^i = (V, E^i, \mathbf{W}^i)$ with $|V| = n$. It is assumed that E^i's and associated edge weights are different but similar to each other. Based on this assumption, when learning G^i's, one can have better performance by borrowing information across graphs. For example, when analyzing scRNA-seq expressions from different disease states/conditions, the datasets generated from the varying groups are generally assumed to share a common gene-coexpression structure. Thus, jointly learning cell-type specficic graphs can improve inference by allowing information sharing across cell-types. To this end, we propose an optimization problem (scMSGL) that learns G^i's simultaneously. In the proposed approach, the learned G^i's are regularized to be close to a consensus graph G, which is also learned by combining information from G^i's. Thus, the proposed formulation ensures that information is shared across graphs when learning G^i's. Furthermore, the structure of G reflects the common connections shared across G^i's, whose inference may be beneficial if one is interested in learning the common gene-coexpression structure over the different cell-types/disease-stage subgroups.

Let $\mathbf{L}^{i,+}$ and $\mathbf{L}^{i,-}$ be the Laplacian matrices of the positive and negative parts of G^i, respectively. Similarly, define $\mathbf{L}^+$ and $\mathbf{L}^-$ for the consensus graph $\mathbf{G}$. Let $\mathcal{L}^+ = \{\mathbf{L}^{1,+}, \ldots, \mathbf{L}^{N,+}, \mathbf{L}^+\}$ and $\mathcal{L}^- = \{\mathbf{L}^{1,-}, \ldots, \mathbf{L}^{N,-}, \mathbf{L}^-\}$. The optimization problem for jointly learning G^i's and G is then:

$$\begin{aligned}
&\underset{\mathcal{L}^+,\mathcal{L}^-}{\text{minimize}} \sum_{s\in\{+,-\}} \sum_{i=1}^{N} \left\{\mathrm{tr}(\mathbf{K}^{\mathrm{i,s}}\mathbf{L}^{\mathrm{i,s}}) + \alpha_{\mathrm{s}}\|\mathbf{L}^{\mathrm{i,s}}\|_{\mathrm{F}}^2 + \beta_{\mathrm{s}}\|\mathbf{L}^{\mathrm{i,s}} - \mathbf{L}^{\mathrm{s}}\|_{\mathrm{F,off}}^2\right\} \\
&\qquad + \gamma_+\|\mathbf{L}^+\|_{1,\mathit{off}} + \gamma_-\|\mathbf{L}^-\|_{1,\mathit{off}} \qquad\qquad (4) \\
&\text{subject to } \mathbf{L}^{i,s} \in \mathbb{L},\ \mathrm{tr}(\mathbf{L}^{\mathrm{i,s}}) = 2\mathrm{n},\ \forall \mathrm{i},\ \forall \mathrm{s} \in \{+,-\} \\
&\qquad (\mathbf{L}^{i,+}, \mathbf{L}^{i,-}) \in \mathbb{C}\ \forall i, \mathbf{L}^+,\ \mathbf{L}^- \in \mathbb{L},\ (\mathbf{L}^+, \mathbf{L}^-) \in \mathbb{C},
\end{aligned}$$

where $\mathbf{K}^{i,+} = \mathbf{K}^i$, $\mathbf{K}^{i,-} = -\mathbf{K}^i$, and $\mathbf{K}^i$ is a kernel matrix constructed from $\mathbf{X}^i$ as described in "Single view signed graph learning" section. $\|\cdot\|_{F,\mathit{off}}$ and $\|\cdot\|_{1,\mathit{off}}$ are the Frobenius norm and the ℓ_1-norm of the off-diagonal entries, respectively. The first term in the summation measures the smoothness and non-smoothness of $\mathbf{X}$ over $G^{i,+}$ and $G^{i,-}$, respectively. The second term controls the density of the learned $G^{i,+}$ ($G^{i,-}$) such that for larger values of α_+ (α_-), we learn denser graphs. The third term ensures that $G^{i,+}$ ($G^{i,-}$) are close to the positive (negative) part of consensus graph G with β_+ (β_-) controlling how close they should be. The last term is a regularizer that controls the sparsity of positive and negative parts of G with larger values of γ_+ and γ_- resulting in a sparser consensus graph. Finally, the constraints are the same as in (3). Algorithm 1 gives the workflow of scMSGL to learn multiple graphs jointly. See Additional file 1 for an ADMM based optimization for (4).

Algorithm 1 Joint inference of multiple related signed graphs with scMSGL

```
Input: $\{\mathbf{X}^i\}_{i=1}^N$, $\alpha_s$'s, $\beta_s$'s, $\gamma_s$'s, $\kappa$: Kernel function
Output: $\{W^{i,+}, W^{i,-}\}_{i=1}^N$: Learned view adjacency matrices, $W^+, W^-$: Learned consensus adjacency matrices
    Construct $\mathbf{K}^i$ where $K^i_{jk} \leftarrow \kappa(\mathbf{X}^i_{j\cdot}, \mathbf{X}^i_{k\cdot})\ \forall i$
    $\{\mathbf{L}^{i,+}\}_{i=1}^N, \mathbf{L}^+, \{\mathbf{L}^{i,-}\}_{i=1}^N, \mathbf{L}^- \leftarrow \textsc{OptimizeEq4}(\{\mathbf{K}^i\}_{i=1}^N, \alpha_s, \beta_s, \gamma_s)$
    $\mathbf{W}^{i,s} \leftarrow |\mathbf{L}^{i,s}|$, $W^{i,s}_{jj} \leftarrow 0$, $\forall i, s$          ▷ Construct view adjacency matrices
    $\mathbf{W}^s \leftarrow |\mathbf{L}^s|$, $W^s_{jj} \leftarrow 0$, $\forall s$          ▷ Construct consensus adjacency matrices
```

Hyperparameter selection procedure

scMSGL requires the selection of six hyperparameters, three of which control the properties of the positive parts of the learned graphs while the remaining control the negative parts. As mentioned above, α_+ (α_-) and γ_+ (γ_-) control the edge density of positive (negative) parts of the learned G^i's and G, respectively. β_+ (β_-) controls how similar the learned $G^{i,+}$'s ($G^{i,-}$'s) are to the consensus graph. We select these hyperparameters similar to that suggested in [6], where hyperparameter selection is guided to learn graphs with desired properties. Alternative to other model selection approaches, such as cross-validation or Bayesian information criterion, this approach can achieve a model that is interpretable and plausible in practice. Thus, we tune the hyperparameters such that the obtained graphs have a desired edge density and view similarity. In particular, assume that one wants the densities of positive and negative edges in the learned G^i's and G to be d_+ and d_-, respectively. Furthermore, assume that the pairwise similarity between $G^{i,+}$ and $G^{j,+}$, $\forall i \neq j$ is desired to be c_+, where the similarity is quantified by the correlation coefficient. Similarly, let c_- be the desired similarity amount for the negative edges of the graphs. Once d_+, d_-, c_+, c_- are fixed, we select the six hyperparameters accordingly. The values of d_+, d_-, c_+, and c_- are selected based on prior knowledge on the datasets under study as detailed in "Results" section.

Results

The performance of scMSGL is evaluated on both simulated and two real scRNA-seq datasets. For simulated data, learned graphs are compared to ground truth networks to quantify the performance of scMSGL. Signed version of area under precision recall curve (AUPRC) is used as the performance metric during this analysis. We report AUPRC ratio, which is the ratio of AUPRC value of scMSGL to that of a random predictor. More details on how AUPRC is calculated can be found in Additional file 1. Simulated data are used to benchmark the performance of scMSGL against scSGL and three GRN inference algorithms, GENIE3, GRNBOOST2 and PIDC, which are found to be the best performing algorithms for scRNA-seq data [10]. These methods and scSGL can only learn a single graph from each dataset at a time. Therefore, they are applied to each $\mathbf{X}^i$ separately and the learned graphs are compared to ground truth G^i's. In addition, we benchmark against Joint Graphical Lasso with fused lasso penalty (JGL-Fused) method [6], which learns multiple related Gaussian graphical models, and Joint Gene Networks with scRNA-seq data (JGNsc) [21] algorithm, which jointly learns the graphs for multiple classes of single cell data. Other single cell joint graph learning algorithms discussed in "Background" section [17, 19] have not been considered due to the absence of publicly/on request available code.

Selected hyperparameter values

Hyperparameters of scMSGL are set as described in "Methods" section with $d_+ = d_- = d$ and $c_+ = c_- = c$. We used the BEELINE [10] pipeline to run GENIE3, GRNBOOST2 and PIDC. GENIE3 and GRNBOOST2 employs random forest and gradient boosting regressors, respectively and hyperparameters of these regressors are set to the default values used in GENIE3 and GRNBOOST2 toolboxes. PIDC uses mutual information to learn gene regulations and it requires a discretizer and an estimator for probability distribution estimation. We used the discretizer and estimator recommended by PIDC toolbox. scSGL requires α_+ and α_-, which are determined the same way as α_s's of scMSGL, i.e., they are set to values such that learned graphs have desired edge densities of $d_+ = d_- = d$. JGL-Fused requires two parameters λ_1 and λ_2, which are analogous to the parameter of scMSGL, α_s and β_s, respectively. Therefore, they are set the same way, i.e. we choose λ_1 and λ_2 such that the learned graphs' desired edge densities satisfy

$d_+ + d_- = 2d$[1] and view similarity of $c_+ = c_- = c$. Finally, JGNsc consists of three steps: imputation, Gaussian transformation and GRN inference with JFL-Fused method. The hyperparameters of the first two steps are set to the default values provided in JGNsc toolbox and λ_1 and λ_2 of JGL-Fused step are set as described above.[2] Exact hyperparameter values of all methods are provided in Table 1 of Additional file 1.

For all datasets, we use $c = 0.5$. For simulated data, since benchmarking GRN inference methods (GENIE3, GRNBOOST2 and PIDC) learn fully connected graphs, we set $d = 0.4$ for fair comparison. For real data, we set $d = 0.1$ for ease of analysis. See Additional file 1 for a discussion on the sensitivity of scMSGL to the selection of d and c.

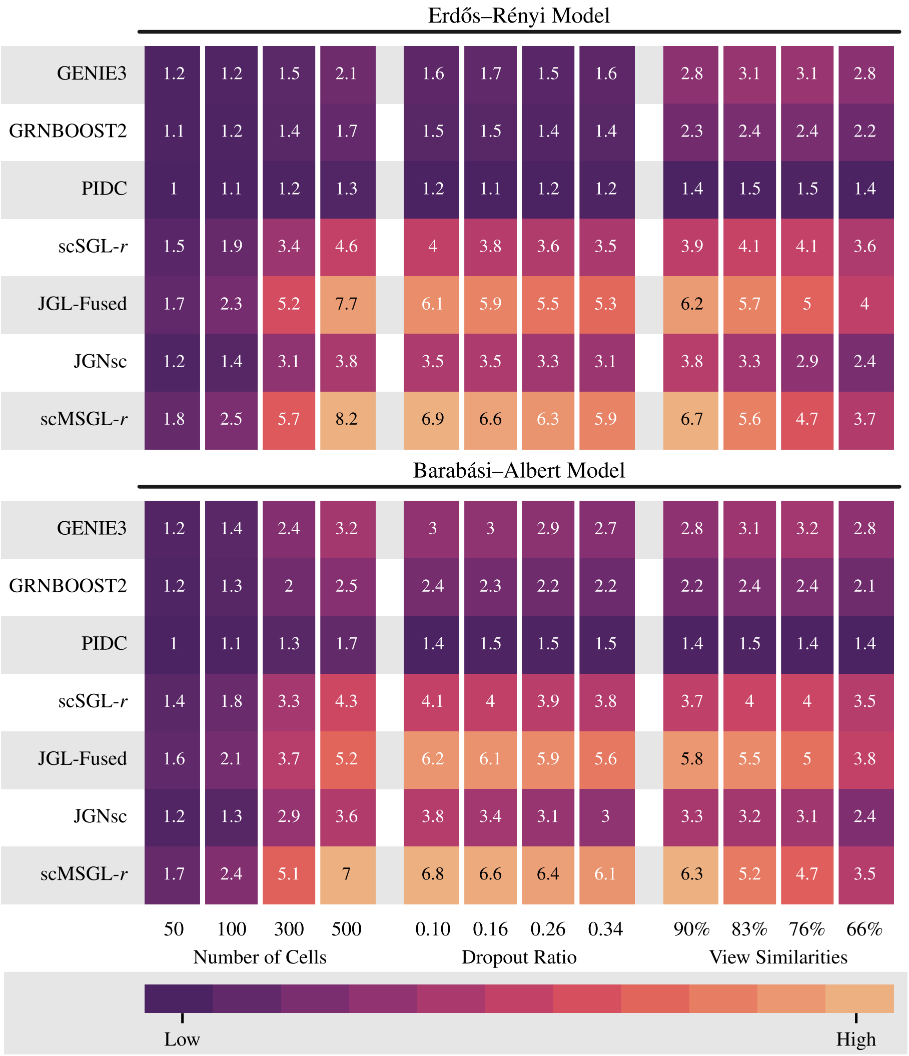
Simulated data

Data generation: To validate the performance of scMSGL, we simulate gene expression data from a multivariate zero-inflated negative binomial (ZINB) distribution. The ZINB distribution has been shown to accurately capture the characteristics of single cell datasets in several published studies [36, 37]. Given a known graph structure, we generate synthetic datasets using an algorithm developed by [38] and illustrated in [22, 31]. Two graph structures are considered for creating the baseline graph G with $n = 100$ genes: random graphs following an Erdős-Rényi (ER) model with an edge density of 0.1 and hub graphs following a Barabási-Albert (BA) model with a degree distribution that follows the power-law. We then convert G to a signed graph by randomly selecting half of the edges and assigning a negative sign to them while assigning a positive sign to the other half. Next, we generate $N = 5$ individual networks $\{G_i\}_{i=1}^N$ by adding $0.9 \times \binom{n}{2} \times \eta$ new edges to the baseline graph G. Half of the added edges are set as negative edges, while the other half are set as positive. The ZINB simulator is then used to generate datasets $\{X_i\}_{i=1}^N$ from the underlying graphs $\{G_i\}_{i=1}^N$. The three parameters of the ZINB distribution; λ, k and ω, which control its mean, dispersion and degree of zero-inflation, respectively were determined using a real scRNA-seq dataset [39]. Each simulation is repeated 10 times and the average performance over 10 realizations is reported. More details for data generation process can be found in Additional file 1.

Sensitivity to the number of cells: We first study the performance of the methods with varying number of cells when the dropout ratio is set to 0.26, $\eta = 0.1$, i.e. 90% of the edges are common across views and the correlation kernel is used for both scSGL and scMSGL. From left panel of Fig. 1, it can be seen that for the different cell numbers, scMSGL has higher AUPRC ratios than methods that learn from a single dataset. This indicates that the proposed method incorporates valuable information across views, which improves the performance. scMSGL also performs better than both joint graph learning methods JGL-Fused and JGNsc. Although JGNsc is based on JGL-Fused, its performance is worse than JGL-Fused. This could be due to a change in the data structure owing to multiple imputation and data transformation steps, which form a part

[1] JGL-Fused does not allow edge densities of the negative and positive parts of the learned graph to be controlled separately, therefore we learned a graph with edge density equal to $2d$, which is the same edge density for scMSGL if the edge signs are not considered.

[2] JGNsc [21] recommends to use Akaike information criterion for selection of λ_1 and λ_2. In our analysis, we found this selection technique does not perform well and its time complexity was high.



Erdős–Rényi Model

	Number of Cells				Dropout Ratio				View Similarities			
	50	100	300	500	0.10	0.16	0.26	0.34	90%	83%	76%	66%
GENIE3	1.2	1.2	1.5	2.1	1.6	1.7	1.5	1.6	2.8	3.1	3.1	2.8
GRNBOOST2	1.1	1.2	1.4	1.7	1.5	1.5	1.4	1.4	2.3	2.4	2.4	2.2
PIDC	1	1.1	1.2	1.3	1.2	1.1	1.2	1.2	1.4	1.5	1.5	1.4
scSGL-*r*	1.5	1.9	3.4	4.6	4	3.8	3.6	3.5	3.9	4.1	4.1	3.6
JGL-Fused	1.7	2.3	5.2	7.7	6.1	5.9	5.5	5.3	6.2	5.7	5	4
JGNsc	1.2	1.4	3.1	3.8	3.5	3.5	3.3	3.1	3.8	3.3	2.9	2.4
scMSGL-*r*	1.8	2.5	5.7	8.2	6.9	6.6	6.3	5.9	6.7	5.6	4.7	3.7

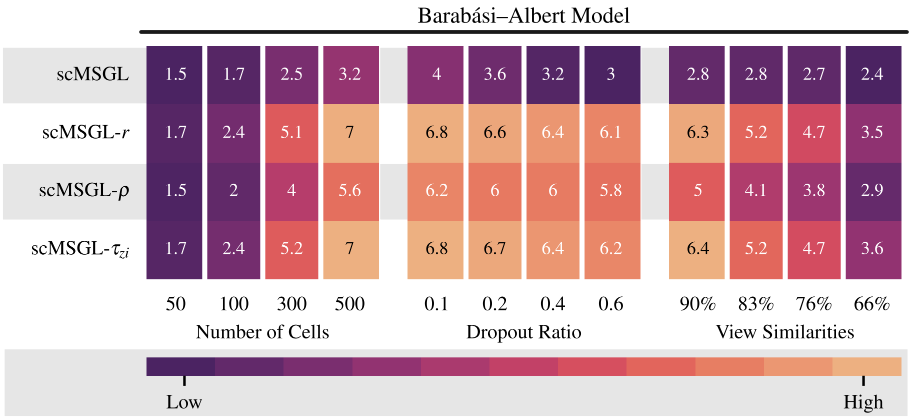
Barabási–Albert Model

	Number of Cells				Dropout Ratio				View Similarities			
	50	100	300	500	0.10	0.16	0.26	0.34	90%	83%	76%	66%
GENIE3	1.2	1.4	2.4	3.2	3	3	2.9	2.7	2.8	3.1	3.2	2.8
GRNBOOST2	1.2	1.3	2	2.5	2.4	2.3	2.2	2.2	2.2	2.4	2.4	2.1
PIDC	1	1.1	1.3	1.7	1.4	1.5	1.5	1.5	1.4	1.5	1.4	1.4
scSGL-*r*	1.4	1.8	3.3	4.3	4.1	4	3.9	3.8	3.7	4	4	3.5
JGL-Fused	1.6	2.1	3.7	5.2	6.2	6.1	5.9	5.6	5.8	5.5	5	3.8
JGNsc	1.2	1.3	2.9	3.6	3.8	3.4	3.1	3	3.3	3.2	3.1	2.4
scMSGL-*r*	1.7	2.4	5.1	7	6.8	6.6	6.4	6.1	6.3	5.2	4.7	3.5

Fig. 1 Performance of different methods on various datasets quantified by AUPRC ratio. All datasets have 100 genes. Left panel reports the results for varying number of cells. Middle one reports the results for varying dropout ratios. Right panel report results for varying degrees of view similarities, which is measured by the percentage of common edges across views in the ground truth graphs. Top plot shows the results for Erdős-Rényi model and the bottom plot shows the results for Barabási-Albert model

of the JGNsc algorithm. As expected, the performance of all methods improves with increasing number of cells. These observations hold for both random graph models.

Sensitivity to dropout ratio: In the second analysis, we evaluate the performance of the different methods with increasing dropout ratio while fixing the number of cells to 400 and $\eta = 0.1$. Results are shown in the middle panel of Fig. 1 for both random graph models, with correlation kernel used for scSGL and scMSGL. Similar to cell sensitivity analysis, scMSGL performs better compared to all other methods irrespective of which graph model is used to generate the datasets. Except for PIDC, AUPRC ratios of all methods drop with increasing dropout ratio as expected. Performance of PIDC mostly remains the same. Since PIDC performs poorly at all drop-out levels, this result does not imply robustness against dropouts.



	Number of Cells				Dropout Ratio				View Similarities			
	50	100	300	500	0.1	0.2	0.4	0.6	90%	83%	76%	66%
scMSGL	1.5	1.7	2.5	3.2	4	3.6	3.2	3	2.8	2.8	2.7	2.4
scMSGL-r	1.7	2.4	5.1	7	6.8	6.6	6.4	6.1	6.3	5.2	4.7	3.5
scMSGL-ρ	1.5	2	4	5.6	6.2	6	6	5.8	5	4.1	3.8	2.9
scMSGL-τ_{zi}	1.7	2.4	5.2	7	6.8	6.7	6.4	6.2	6.4	5.2	4.7	3.6

Fig. 2 Performance of scMSGL without any kernel (first row) and with different kernels on datasets generated from BA model and studied in Fig. 1

Sensitivity to view similarity: Next, we study the effect of view similarity on the performance of algorithms. Datasets are generated with varying η values while fixing the number of cells to 400 and the dropout ratio to 0.26. Results are reported in right panel of Fig. 1, where the correlation kernel is employed for scSGL and scMSGL. When view similarity is 90%, the best performing algorithm is scMSGL, while for lower view similarity values JGL-Fused performs slightly better than scMSGL. The reason that JGL-Fused performs better than scMSGL for smaller view similarity values could be due to the difference in the regularization terms used to impose similarity across views. JGL-Fused uses a ℓ_1 norm penalty, while we employ a squared Frobenius norm. Compared to fused lasso, squared Frobenius norm is susceptible to outliers, which can degrade the performance. The performance of single-view algorithms does not get affected by the changes in view similarity, as they learn each view independently. On the other hand, there is a drop in the performances of all joint graph learning methods with decreasing view similarity. This is an expected behaviour, since both methods assume the dependence of views.

Kernel comparison: Formulation of scMSGL allows us to use various kernels. Therefore, we study how the performance changes with respect to the kernel type. Datasets are created using the BA model and results are shown in Fig. 2 for varying number of cells, dropout ratios and view similarities. The best performing kernel is τ_{zi}, followed by the correlation kernel. When Figs. 1 and 2 are compared, scMSGL has higher AUPRC ratios than single-view approaches and JGNsc irrespective of the kernel choice. The change in the performance of τ_{zi} and ρ with varying cell numbers, dropout ratios and view similarity are very similar to that of the correlation. Finally, to better understand the effect of kernels, the performance of scMSGL without any kernels, i.e. $\mathbf{K}^i = \mathbf{X}^i{\mathbf{X}^i}^\top$, is also reported. Figure 2 shows all kernels have significantly higher performance compared to when no kernel is used, which indicates the importance of kernel usage in GRN inference.

Time complexity comparison: We compare the different methods based on their run time complexity. We generated datasets using BA model with varying number of cells and number of genes. Table 1 reports the run time of scSGL, scMSGL, JGL-Fused and

Table 1 Run time of scMSGL and benchmarking methods in seconds with respect to number of cells and genes

	Number of cells				Number of genes			
Method	**50**	**100**	**300**	**500**	**50**	**100**	**300**	**500**
scSGL-*r*	1.10	0.54	0.35	0.38	0.15	0.37	5.88	26.68
JGL-Fused	175.64	117.65	95.95	98.03	10.02	95.98	1703.66	–
JGNsc	196.76	160.37	233.06	373.15	130.13	373.06	2541.45	–
scMSGL-*r*	14.00	12.39	10.00	8.51	0.35	3.89	110.49	304.71

All methods run on the same computing cluster with compute nodes that have similar compute power. Run times of JGL-Fused and JGNsc for 500 genes are not reported, we were not able to run them in a reasonable time limit (4 h)

JGNsc in seconds. Run times of GENIE3, GRNBOOST2 and PIDC are not reported as they are shown to have higher time complexity than scSGL in [31]. Reported run times correspond to one run without hyperparameter search. Run time of scSGL is the total run time to infer all views.

In the first dataset, number of genes, dropout ratio and η are fixed to 100, 0.26, and 0.1, respectively and number of cells varies. Results for this dataset indicate that scMSGL is faster than joint graph learning methods JGL-Fused and JGNsc. JGL-Fused also uses an ADMM based optimization, however it needs singular value decomposition at each ADMM iteration. scMSGL does not need this expensive operation; thus, it runs much faster than JGL-Fused. scSGL is faster than scMSGL, which is expected as scMSGL optimization takes longer time to converge due to added regularization terms and consensus graph learning. Finally, all methods except JGNsc are observed to run faster with increasing number of cells, since the inference problem becomes easier with higher number of cells, which makes iterative optimization procedure used by all methods converge faster. JGNsc runs slower with increasing number of cells, as its imputation step needs to handle a larger data matrix.

In the second dataset with increasing number of genes, the number of cells, dropout ration and η are fixed to 500, 0.26, and $\eta = 0.1$, respectively. As before, scMSGL is faster than joint graph learning methods and is slower than scSGL. Increasing the number of genes is observed to increase run time complexity of all methods, as it makes the problem harder.

Analysis of scRNA-seq data from mouse embryonic stem cell differentiation

Central to the differentiation process and many other cellular processes is the expression of right combination of genes or modules of genes. Accurate characterization of the co-expression networks for progenitor and multiple cell types can help in understanding the cascade of cellular state transitions [12]. In this section, we study the differentiation process of mouse embryonic stem cells (mESC) using single cell RNA sequencing datasets [40]. This data was generated using high-throughput droplet-microfluidic approach and was primarily used to study differentiation in mESC before and after leukemia inhibitory factor (LIF) withdrawal. Since LIF maintains pluripotency of mESC, LIF withdrawal is considered to initiate the differentiation process. The dataset contains cells sampled from 4 states (or natural subgroups): before LIF withdrawal, day 0 and after the withdrawal for days 2, 4 and 7. The subgroups contain 933, 303, 683 and 798 cells,

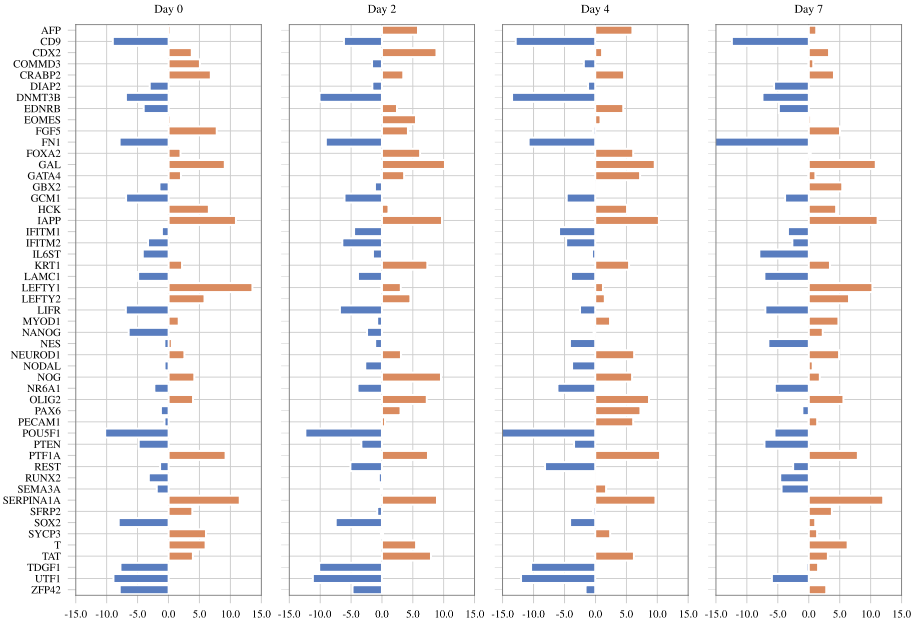


Fig. 3 Genes with the highest node degrees. Orange and blue bars indicate that the degree is calculated using activating and inhibitory edges, respectively. Only genes whose activating or inhibitory degrees is among the top 15 genes in any view are shown

respectively. This dataset has been previously analyzed using joint graphical estimation in [17, 19] and similar to them we only consider the 72 stem cell markers in our application [41].[3]

We first estimated the subgroup specific and the consensus graphs. Based on the results of simulated data, we employ the zero-inflated Kendall's tau kernel. Next, we calculate the signed node degrees of each gene, i.e., $D^{+}_{ii} = \sum_{j=1}^{n} W^{+}_{ij}$ and $D^{-}_{ii} = \sum_{j=1}^{n} W^{-}_{ij}$ from learned graphs G^{+} and G^{-}. We then consider the genes with top signed degrees as hub genes whose signed degrees are reported in Fig. 3. The result confirms the importance of regulator genes NANOG, SOX2, POU5F1, ZFP42, UTF1 in early stages of differentiation. NANOG has been reported to maintain pluripotency by inhibiting genes that activate differentiation to lineages associated with extraembryonic endoderm [43, 44]. Figure 3 clearly shows that the number of inhibitory relationships associated with NANOG decreases as the ES cells proceed to a matured state. POU5F1 and SOX2 also exhibit higher number of inhibitory relationships in the the first few days. SOX2, NANOG and POU5F1 are known to play a fundamental role in the self-renewal and pluripotency of mouse embryonic stem cells [45]. Reduction in expression of NANOG has been shown to be correlated with the induction of gene GATA4 which initiates differentiation of pluripotent cells [46] and therefore GATA4 has been correctly identified as a hub gene in Days 2 and 4. Collectively, these results confirm the fundamental roles of SOX2, NANOG and POU5F1 in the pluripotency stage and how an eventual reduction in their expression initiates differentiation.

[3] The dataset was downloaded from GEO database [42] (with ID GSE65525). For the preprocessing steps, please refer to "Data analysis" subsection of the "Experimental procedures" section in [40]. The only preprocessing we performed was log-transformation to make count data continuous.

Analysis of scRNA-seq data from medulloblastoma

Medulloblastoma (MB) is a highly malignant cerebellar tumor mostly affecting young children [47]. Several studies have been done to pinpoint the genetic drivers in each of the four distinct tumor subgroups: WNT-pathway-activated, SHH-pathway-activated, and the less-well-characterized Group 3 and Group 4 [47]. Among these subgroups, Group 3 and Group 4 tumors account for the majority of the MB diagnoses, with Group 3 MB having a metastatic diagnosis of approximately 50%. Transcription factors (TFs) MYC2 and OTX2 have commonly been identified as key oncogenic TFs in Group 3 and 4 tumorogenesis. Dong et al. [21] used the joint single cell network algorithm to study the roles of MYC and OTX2 utilizing the MB scRNA-seq data set (GSE119926) by [39].[4] Using the same selected samples from a subset of 17 individuals that were grouped into three subsets Group 3, Group 4 and an intermediate cell type, we estimate the joint gene regulatory network for the three groups for $\sim$ 750 genes among which most are enzyme-related genes from mammalian metabolic enzyme database [48]. Bulk profiling studies for MB cells have consistently observed overlapping transcriptional and epigenetic signatures in Group 3 and Group 4 tumors suggesting shared developmental origins [39, 49]. Based on this, we hypothesize that a joint analysis of the different MB cell-types would better capture the local functional interactions of MYC and OTX2 across different tumor subtypes and would eventually help in delineating their global role in regulating metabolic processes in MB cells.

Subgroup specific networks along with the consensus graph were estimated with zero-inflated Kendall's tau kernel. Table 2 shows that the average edge weight for the MYC network is considerably higher for Group 3 compared to Group 4 and the intermediate subgroup. Figure 4 further shows that Group 3 MYC network has stronger edge connections and higher density in compared to the intermediate group. In Group 4, almost all the connections become activating except for Aldh3a2 and Eno2; which were found to be strongly downregulated in all the tumor subgroups confirming their role in cancer resistance [50, 51]. This varying network structure over the subgroups confirms the major role MYC plays in initiation, maintenance, and progression of Group 3 tumors [52]. In Fig. 4, it is shown that OTX2 has a denser network for Group 4 MB cells in comparison to the other groups. In Group 3 MB cells, OTX2's connections to the metabolic genes are very distinct from the MYC's. In addition, scMSGL detected relationships between OTX2 and metabolic genes PAICS and PPAT in Group 3 tumors. These genes related to the human purine biosynthesis pathways have been previously reported to be induced by MYC [53]. This confirms that OTX2 is functionally cooperating with MYC to regulate gene expression in medulloblastoma [52, 54]. Broadly these results suggest that MYC and OTX2 play significant roles in in the transcriptional regulation of the metabolic genes and the mechanisms underlying MYC and OTX2 mediated MB maintenance and progression likely vary in different subgroups of MB cells.

[4] A detailed overview of the MB scRNA-seq data generation and processing can be found in the "Methods" section of [39] under subsection "Human scRNA-seq data generation and processing". We log-transformed the obtained datasets to make count data continuous.

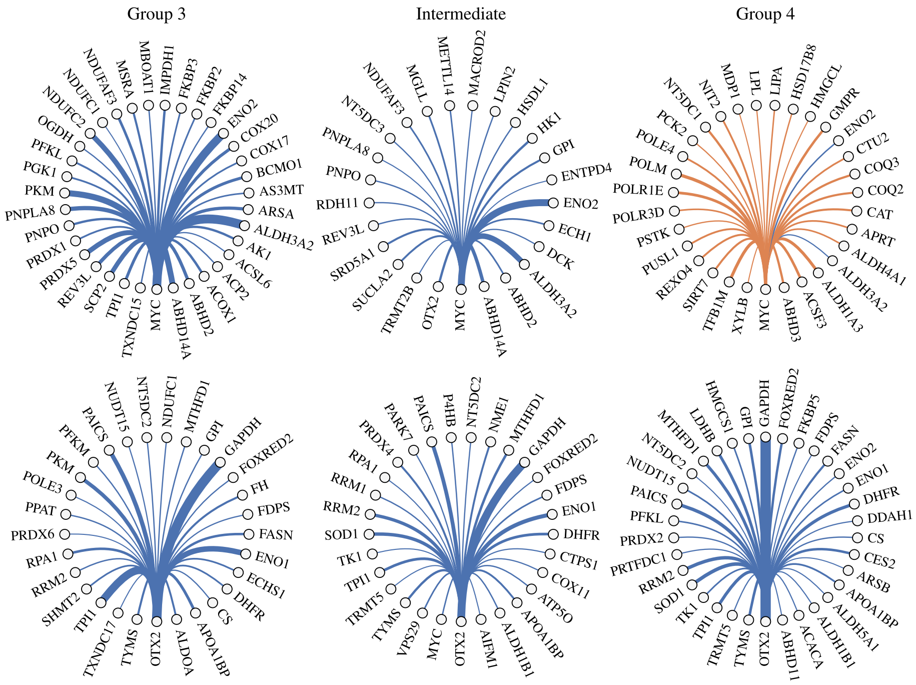


Fig. 4 Connections of MYC (top) and OTX2 (bottom) genes. Edge widths are proportional to connection weights. Orange and blue edge colors indicate that the connection is activating and inhibitory, respectively. Only the top one third of the connections in all views of the multiview graph are shown

Table 2 Node degree of MYC in learned graphs

	Group 3	Intermediate	Group 4
Total degree	5.436	3.334	4.180
Avg. edge weight	0.077	0.037	0.039

Conclusion

In this paper, we presented scMSGL for joint inference of multiple GRNs from scRNA-seq datasets having multiple classes. scMSGL learns functional relationships between genes across multiple related classes of single cell gene expression datasets under the assumption that there exists a shared structure across classes. The main novelty of our paper lies in the formulation of a highly efficient optimization framework that extends the signed graph learning [31] approach to high dimensional datasets with multiple classes. The kernelization trick embedded within the algorithm renders it capable of handling sparse and noisy features; expected to demonstrate highly non-linear relationships. Furthermore, the estimation of the consensus graph may help in understanding the joint structure existing within the multiple classes. Using simulation studies, we demonstrated the superior performance of scMSGL over single view learning and existing joint learning methods for ER and BA graph models. In addition, performance was ascertained by varying a number of simulation parameters such as dropout levels, cell numbers and view similarity and scMSGL demonstrated superior performance in all scenarios. Applying scMSGL to the mESC dataset, we robustly identified previously reported regulatory markers as the hub genes for the different days and captured the progression of the differentiation process by analyzing these changes in hubs over the days. For the medulloblastoma data, scMSGL efficiently captured the significant roles that key oncology markers MYC and OTX2 play in the transcriptional regulation of metabolic genes.

There are various aspects of the proposed method that can be considered for improvement as future work. One challenge in implementing scMSGL is how to select the kernel function. This challenge can be addressed by combining information from multiple kernels during learning. An open problem in graph learning literature is hyperparameter selection, which is also a limitation of the proposed method. Current work selects the hyperparameters by searching the values that would result in graphs with desired properties. Future work can improve the accuracy of the learned graphs through better hyperparameter selection and multi-kernel strategies. Computational complexity of scMSGL is quadratic with respect to the number of genes (similar to scSGL) and linear in number of views. Therefore, its application to datasets with very large number of genes is not feasible. However, recent developments in GSP to scale GL to large-scale problems [55] can be exploited to scale scMSGL. Finally, additional sources of data that help in identifying direct interactions between TFs and target genes, can provide a way to filter out false positives. The current availability of single-cell epigenomic datasets has made it easier to further explore the regulatory relationship between TF and genes. Single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq), for example, allows the identification of DNA regulatory elements within accessible genomic DNA regions in single cells, hence enabling the identification of direct regulations in GRNs. Integration of multiomics profiles within the framework of scMSGL could be an interesting avenue for future research.

Abbreviations

GRN	Gene regulatory networks
GSP	Graph signal processing
GL	Graph learning
GFT	Graph Fourier transform
scRNA-seq	Single cell RNA sequencing
scSGL	Single cell signed graph learning
scMSGL	Single cell multiple singed graph learning

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05250-y.

Additional file 1. Additional file includes optimization process for scMSGL, definition of Signed AUPRC, simulated data generation process, results on sensitivity of scMSGL to hyperparameter selection and details about selected hyperparameters values of all methods.

Acknowledgements

Not applicable.

Author contributions

AK devised the scMSGL and performed data processing. SS designed the simulation experiments and analyzed the results. TM and SA supervised the project. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by following grants: NSF CCF 2211645, NSF CCF 2006800, and NSF DMS 1924724.

Availability of data and materials

For real data, please see the cited references in "Results" section.The scMSGL code and simulated data are available at https://github.com/Single-Cell-Graph-Learning/scMSGL.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 24 December 2022 Accepted: 22 March 2023
Published online: 04 April 2023

References

1. Sanguinetti G, Huynh-Thu VA. Gene regulatory networks. Springer; 2019.
2. Yin W, Mendoza L, Monzon-Sandoval J, Urrutia AO, Gutierrez H. Emergence of co-expression in gene regulatory networks. PLoS ONE. 2021;16(4):0247671.
3. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun. 2014;5(1):1–9.
4. Van Der Wijst MG, de Vries DH, Brugge H, Westra H-J, Franke L. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Med. 2018;10(1):1–15.
5. 13, B..W.H..H.M.S.C.L....P.P.J..K.R., data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, G., for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, I., et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
6. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Ser B Stat Methodol. 2014;76(2):373.
7. Lee W, Liu Y. Joint estimation of multiple precision matrices with common structures. J Mach Learn Res. 2015;16(1):1035–62.
8. Ha MJ, Baladandayuthapani V, Do K-A. Dingo: differential network analysis in genomics. Bioinformatics. 2015;31(21):3413–20.
9. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinform. 2018;19(1):1–21.
10. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54.
11. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell. 2018;174(5):1293–308.
12. Matsumoto H, Kiryu H, Furusawa C, Ko MS, Ko SB, Gouda N, Hayashi T, Nikaido I. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. Bioinformatics. 2017;33(15):2314–21.
13. Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011;98(1):1–15.
14. Fiers MW, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S. Mapping gene regulatory networks from single-cell omics data. Brief Funct Genomics. 2018;17(4):246–54.
15. Chen G, Ning B, Shi T. Single-cell rna-seq technologies and related computational data analysis. Front Genet. 2019;10:317.
16. Akers K, Murali T. Gene regulatory network inference in single cell biology. Curr Opin Syst Biol. 2021;26:87–97.
17. Mukherjee S, Carignano A, Seelig G, Lee S-I. Identifying progressive gene network perturbation from single-cell rna-seq data. In: 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2018. p. 5034–40.
18. Allen GI, Liu Z. A local Poisson graphical model for inferring networks from sequencing data. IEEE Trans Nanobiosci. 2013;12(3):189–98.
19. Wu N, Yin F, Ou-Yang L, Zhu Z, Xie W. Joint learning of multiple gene networks from single-cell gene expression data. Comput Struct Biotechnol J. 2020;18:2583–95.
20. Liu H, Han F, Yuan M, Lafferty J, Wasserman L. High-dimensional semiparametric gaussian copula graphical models. Ann Stat. 2012;40(4):2293–326.
21. Dong M, He Y, Jiang Y, Zou F. Joint gene network construction by single-cell rna sequencing data. Biometrics. 2022.
22. Jia B, Xu S, Xiao G, Lamba V, Liang F. Learning gene regulatory networks from next generation sequencing data. Biometrics. 2017;73(4):1221–30.
23. Mongia A, Sengupta D, Majumdar A. Mcimpute: matrix completion based imputation for single cell rna-seq data. Front Genet. 2019;10:9.
24. Liu H, Lafferty J, Wasserman L. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. J Mach Learn Res. 2009;10(10):2295–328.

25. Shuman DI, Narang SK, Frossard P, Ortega A, Vandergheynst P. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process Mag. 2013;30(3):83–98.
26. Dong X, Thanou D, Rabbat M, Frossard P. Learning graphs from data: a signal representation perspective. IEEE Signal Process Mag. 2019;36(3):44–63.
27. Dong X, Thanou D, Frossard P, Vandergheynst P. Learning Laplacian matrix in smooth graph signal representations. IEEE Trans Signal Process. 2016;64(23):6160–73.
28. Kalofolias V. How to learn a graph from smooth signals. In: Artificial intelligence and statistics. PMLR; 2016. p. 920–9.
29. Segarra S, Marques AG, Mateos G, Ribeiro A. Network topology inference from spectral templates. IEEE Trans Signal Inf Process Netw. 2017;3(3):467–83.
30. Navarro M, Wang Y, Marques AG, Uhler C, Segarra S. Joint inference of multiple graphs from matrix polynomials. J Mach Learn Res. 2022;23(76):1–35.
31. Karaaslanli A, Saha S, Aviyente S, Maiti T. scsgl: kernelized signed graph learning for single-cell gene regulatory network inference. Bioinformatics. 2022;38(11):3011–9.
32. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. Nat Methods. 2019;16(5):381–6.
33. Sandryhaila A, Moura JM. Discrete signal processing on graphs: frequency analysis. IEEE Trans Signal Process. 2014;62(12):3042–54.
34. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: an r-package for identifying proportionally abundant features using compositional data analysis. Sci Rep. 2017;7(1):1–9.
35. Pimentel RS, Niewiadomska-Bugaj M, Wang J-C. Association of zero-inflated continuous variables. Stat Probab Lett. 2015;96:61–7.
36. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. Zinb-wave: a general and flexible method for signal extraction from single-cell rna-seq data. bioRxiv. 2017;125112.
37. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):1–15.
38. Yahav I, Shmueli G. On generating multivariate Poisson data in management science applications. Appl Stoch Model Bus Ind. 2012;28(1):91–102.
39. Hovestadt V, Smith KS, Bihannic L, Filbin MG, Shaw ML, Baumgartner A, DeWitt JC, Groves A, Mayr L, Weisman HR, et al. Resolving medulloblastoma cellular architecture by single-cell genomics. Nature. 2019;572(7767):74–9.
40. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
41. Przybyla LM, Voldman J. Attenuation of extrinsic signaling reveals the importance of matrix remodeling on maintenance of embryonic stem cell self-renewal. Proc Natl Acad Sci. 2012;109(3):835–40.
42. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
43. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A. Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. Cell. 2003;113(5):643–55.
44. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S. The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell. 2003;113(5):631–42.
45. Zhou Q, Chipperfield H, Melton DA, Wong WH. A gene regulatory network in mouse embryonic stem cells. Proc Natl Acad Sci. 2007;104(42):16438–43.
46. Hough SR, Clements I, Welch PJ, Wiederholt KA. Differentiation of mouse embryonic stem cells after rna interference-mediated silencing of oct4 and nanog. Stem Cells. 2006;24(6):1467–75.
47. Northcott PA, Robinson GW, Kratz CP, Mabbott DJ, Pomeroy SL, Clifford SC, Rutkowski S, Ellison DW, Malkin D, Taylor MD, et al. Medulloblastoma. Nat Rev Dis Primers. 2019;5(1):1–20.
48. Corcoran CC, Grady CR, Pisitkun T, Parulekar J, Knepper MA. From 20th century metabolic wall charts to 21st century systems biology: database of mammalian metabolic enzymes. Am J Physiol Renal Physiol. 2017;312(3):533–42.
49. Northcott PA, Korshunov A, Witt H, Hielscher T, Eberhart CG, Mack S, Bouffet E, Clifford SC, Hawkins CE, French P, et al. Medulloblastoma comprises four distinct molecular variants. J Clin Oncol. 2011;29(11):1408.
50. Moreb JS, Muhoczy D, Ostmark B, Zucali JR. Rnai-mediated knockdown of aldehyde dehydrogenase class-1a1 and class-3a1 is specific and reveals that each contributes equally to the resistance against 4-hydroperoxycyclophosphamide. Cancer Chemother Pharmacol. 2007;59(1):127–36.
51. Chang PM-H, Chen C-H, Yeh C-C, Lu H-J, Liu T-T, Chen M-H, Liu C-Y, Wu AT, Yang M-H, Tai S-K, et al. Transcriptome analysis and prognosis of ALDH isoforms in human cancer. Sci Rep. 2018;8(1):1–10.
52. Roussel MF, Robinson GW. Role of MYC in medulloblastoma. Cold Spring Harb Perspect Med. 2013;3(11): 014308.
53. Liu Y-C, Li F, Handler J, Huang CRL, Xiang Y, Neretti N, Sedivy JM, Zeller KI, Dang CV. Global regulation of nucleotide biosynthetic genes by c-Myc. PLoS ONE. 2008;3(7):2722.
54. Lu Y, Labak CM, Jain N, Purvis IJ, Guda MR, Bach SE, Tsung AJ, Asuthkar S, Velpula KK. OTX2 expression contributes to proliferation and progression in Myc-amplified medulloblastoma. Am J Cancer Res. 2017;7(3):647.
55. Kalofolias V, Perraudin N. Large scale graph learning from smooth signals. In: International conference on learning representations. 2018.

Publisher's Note