STREAMLINE: Structural and Topological Performance Analysis of Algorithms for the Inference of Gene Regulatory Networks from Single-Cell Transcriptomic Data

Niclas Popp [1,2,3], Marco Stock [1,2,3],
Jonathan Fiorentino*† [1,2,3,4], Antonio Scialdone* [1,2,3,4]

[1]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München – German Research Center for Environmental Health, Munich, 81377, Germany
[2]Institute of Functional Epigenetics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, 85764, Germany
[3]Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, 85764, Germany

[4]These authors jointly supervised this work

†Current address: Center for Life Nano- & Neuro- Science, Istituto Italiano di Tecnologia, Rome, 00161, Italy

*Correspondence: jonathan.fiorentino@iit.it, antonio.scialdone@helmholtz-muenchen.de

Abstract

In recent years, many algorithms for inferring gene regulatory networks from single-cell transcriptomic data have been published. Several studies have evaluated their accuracy in estimating the presence of an interaction between pairs of genes. However, these benchmarking analyses do not quantify the algorithms' ability to capture structural properties of networks, which are fundamental, for example, for inferring the robustness of a gene network to external perturbations. Here, we devise a three-step benchmarking pipeline called STREAMLINE that quantifies the ability of algorithms to capture topological properties of networks and identify hubs. To this aim, we use data simulated from different types of networks as well as experimental data from three different organisms. We apply our benchmarking pipeline to four algorithms and provide guidance on which algorithm should be used depending on the global network property of interest.

1 Introduction

Single-cell transcriptomics techniques allow probing patterns of gene expression on an increasingly larger scale, with recent studies including million of cells and thousands of genes [1]. Such rapid progress in expanding the scale of data available makes single-cell datasets more appealing for tasks like the inference of gene regulatory networks (GRNs), with the goal of achieving a mechanistic

understanding of the systems at hand and going beyond purely descriptive characterizations [2]–[4]. However, GRN inference from single-cell data entails many computational challenges, such as high levels of technical noise in the data [5], the extreme sparsity of the ground truth network to be inferred [6] and the increasing scale of gene expression data [7]. For this reason, many algorithms for GRN inference from single-cell data have been published in the last few years. The increasingly large number of such algorithms demands benchmarking studies that can guide the user in the choice of the methods that perform best under various conditions [8]–[10].

While these studies offer some guidance for users, they are affected by important limitations. First, the quantification of the performance is obtained from a limited number and types of networks. Moreover, the available benchmarking studies mostly focus on the ability of the GRN algorithms to predict local features of networks, like the interactions between pairs of genes, using, for example, area under the curve metrics, or the presence of specific sub-graphs (network motifs). Nevertheless, these metrics do not assess the algorithms' ability to infer the structural properties of the GRN, which quantify important features like the robustness to perturbations [11] and the presence of network hubs representing master regulators. These structural properties can be quantified by topological measurements [12], including, for instance, the network efficiency and the assortativity. So far, the performance of GRN inference algorithms on the estimation of topological properties has been only assessed with bulk RNA-seq data, and employing a limited number of synthetic networks [13], which makes it hard to reach robust conclusions, especially when applying the methods to single-cell data.

In this work, we developed STREAMLINE, a three-step benchmarking framework to score the performance of GRN inference algorithms in estimating structural properties of networks from single-cell RNA-seq (scRNA-seq) datasets. Such properties quantify the network's robustness to perturbations and the presence of hubs. We use data simulated from hundreds of networks belonging to four classes with different structural properties that have been shown to be biologically relevant [14], [15], as well as from a set of curated networks extracted from real GRNs [8]. In addition to simulated data, we also use real datasets from yeast, mouse, and human [16].

We apply STREAMLINE to four GRN inference algorithms chosen amongst the top-performing ones [8]. Our benchmarking analysis provides guidance in the choice of the algorithm for the prediction of network robustness and the identification of hubs. Moreover, our results point to systematic biases in some algorithms, which could indicate ways of improving them.

To facilitate the use of our benchmarking framework, we have made it compatible with an existing pipeline (BEELINE [8]), and we make all the code available in a GitHub repository (`github.com/ScialdoneLab/STREAMLINE`).

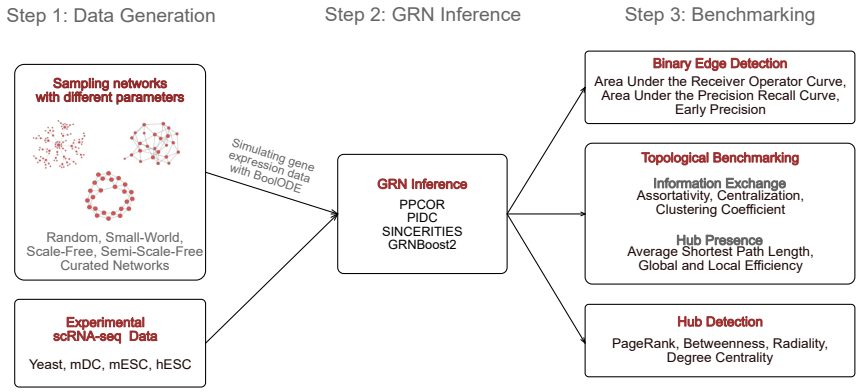

Figure 1: Schematic overview of STREAMLINE. STREAMLINE consists of three steps: first, synthetic scRNA-seq data are generated from different classes of networks. Then, GRN inference methods are applied to synthetic as well as real data. Finally, the methods' performance on the predictions of edges and of structural network properties (quantifying the network robustness and hub presence) are evaluated.

2 Results

2.1 Overview of STREAMLINE

The steps involved in STREAMLINE are schematically represented in Figure 1. We consider two types of datasets: simulated and real datasets. With the simulated datasets, we generate scRNA-seq data in silico from four classes of networks with well-defined and different structural properties: Random, Small-World, Scale-Free, and Semi-Scale-Free Networks. Random or Erdös-Renyi (ER) networks include a set of nodes in which each node pair has the same probability of being connected by an edge [17]. In Scale-Free (SF) networks, the edges are drawn such that the degree distribution follows a power law [14]. In Semi-Scale-Free (SSF) graphs only the out-degree distribution follows a power law, while the in-degree distribution is uniform. Such networks were introduced by Ouma et al. [14] to model real GRNs. Small-World (SW) networks have the property that the neighbors of any given node are likely to be neighbors of each other [15]. Additionally, we included four Curated (Cur.) networks that consist of sub-networks of known GRNs [8].

Networks from each class are defined by a set of parameters. To make our results independent of specific instances of networks, we sampled 20 networks from each class with different combinations of parameters and two sizes: a smaller (15 nodes and 50 edges) and a larger (25 nodes and 100 edges) size. All results shown below are averaged over all the instances of networks generated for a given

class. Details about the network classes and the parameters used for network sampling are provided in the Methods section, Supplementary Section S1 and Table S1. From each of these networks, we simulated scRNA-seq datasets using BoolODE, a recently developed software based on ordinary differential equations [8] (see Methods section).

In addition to simulated datasets, we also considered four real scRNA-seq datasets generated from different organisms and cell types: yeast [18], mouse dendritic cells (mDC) [19], mouse embryonic stem cells (mESC) [20] and human embryonic stem cells (hESC) [21]. These datasets were used in a previous benchmarking study [16], where the authors also provide estimations of ground truth networks.

The second step of our pipeline involves running the algorithms to infer GRNs from each of the datasets. We chose the four top-performing algorithms according to a recent study where the accuracy in predicting gene-gene interactions was evaluated [8]: PIDC [22], PPCOR [23], SINCERITIES [24] and GRNBoost2 [25]. Two of these methods (PIDC and PPCOR) give as output undirected networks, while SINCERITIES and GRNBoost2 provide directed networks. A brief description of each algorithm is provided in the Methods section.

In addition to scoring each method's ability to predict edges as a binary classification problem, similarly to what has been done in previous benchmarking analyses (see [8], [9], Supplementary Section S2 and Figure S1), we analyzed the ability of each method to predict global properties of networks. Specifically, we computed topological properties that quantify how efficiently information is exchanged in the network and the tendency of networks to include hubs.

The efficiency of information exchange in a network can be a measure of how robust networks are to perturbations [26]. To quantify this, we calculated three topological measures: the Global Efficiency, the Local Efficiency and the Average Shortest Path Length (see Methods).

Another biologically important property of networks is the presence of hubs, i.e., nodes that have a degree much larger than the average. In GRNs, hubs are genes that regulate the expression levels of many other genes and can represent master regulators of a given biological process. We computed the Centralization, the Assortativity and the Clustering Coefficient, which are topological measures linked to the presence of hubs in a network.

In addition to quantifying the tendency of networks to possess hubs, it is important to correctly identify them. Hence, we tested the GRN inference algorithms for their ability in predicting which nodes constitute hubs. To do so, we computed four metrics used to detect hubs - Page Rank Centrality, Betweenness Centrality, Out-Centrality and Radiality (see Methods) - and we compared the values obtained from the ground truth networks versus those calculated from the inferred networks.

Below, we describe the detailed results of each of these benchmarking analyses.

2.2 Estimation of information exchange efficiency

As mentioned above, the topological properties we evaluated are the Average Shortest Path Length, the Global Efficiency, and the Local Efficiency (Figure 2A)

of the inferred and ground truth networks in order to quantify topological aspects of the efficiency in information exchange. In particular, we estimated the Mean Signed Error (MSE) between these quantities computed on the ground truth networks and the networks inferred from each of the algorithms (see Methods section).

First, we considered the simulated datasets generated from different classes of networks. Figure S2A displays the distributions of the ground truth values of all the topological measures we considered in this stage. These plots show how the different structural properties of each class of networks are reflected by different values of these topological measures. For example, the SW networks are characterized by larger Average Shortest Path Length and lower Global and Local Efficiency, as expected based on their properties [15].

In Figure 2B, we report the MSE for all the topological measures computed on the simulated datasets. Overall, we found that the accuracy of the predictions depends on the type of network, in addition to the algorithm. For instance, all the algorithms we tested tend to overestimate the Local Efficiency (MSE > 0), except for SINCERITIES, which underestimates it in SSF and Cur networks. The best predictions (corresponding to MSE~ 0) are obtained on SF and SSF networks.

Similarly, the Global Efficiency tends to be overestimated by all algorithms except for GRNBoost2 (Figure 2B), which underestimates it. These biases imply that the best-performing algorithm depends on the type of network: for example, GRNBoost2 provides an accurate estimation of the Global Efficiency in SW networks, which is lower than in other network classes (Figure 2B and Figure S2).

The best estimations of the Average Shortest Path Length (Figure 2B) are provided by the PIDC and PPCOR algorithms, especially in the ER, SF, and SSF networks. In the SW and the Cur networks, for which the Average Shortest Path Length is greater than for SSF graphs (Figure S2A), all algorithms underestimate this property, particularly GRNBoost2 and SINCERITIES.

We performed the same analysis on four real scRNA-seq datasets from three species (Figure 2C). The corresponding ground truth networks have lower Global and Local Efficiency and larger Average Shortest Path Length compared to the synthetic networks we considered (Figure S2B).

These differences likely affect the values of the MSE we computed, which are reported in Figure 2C. For example, unlike with the synthetic networks, we found an overall tendency of all algorithms to underestimate the Local Efficiency and overestimate the Global Efficiency. GRNBoost2 and PPCOR provide the most accurate predictions of the Global and Local Efficiency, respectively.

As for the Average Shortest Path Length, the MSE is mostly positive, indicating an overestimation, and it is smallest for SINCERITIES, which is the best-performing algorithm in this case.

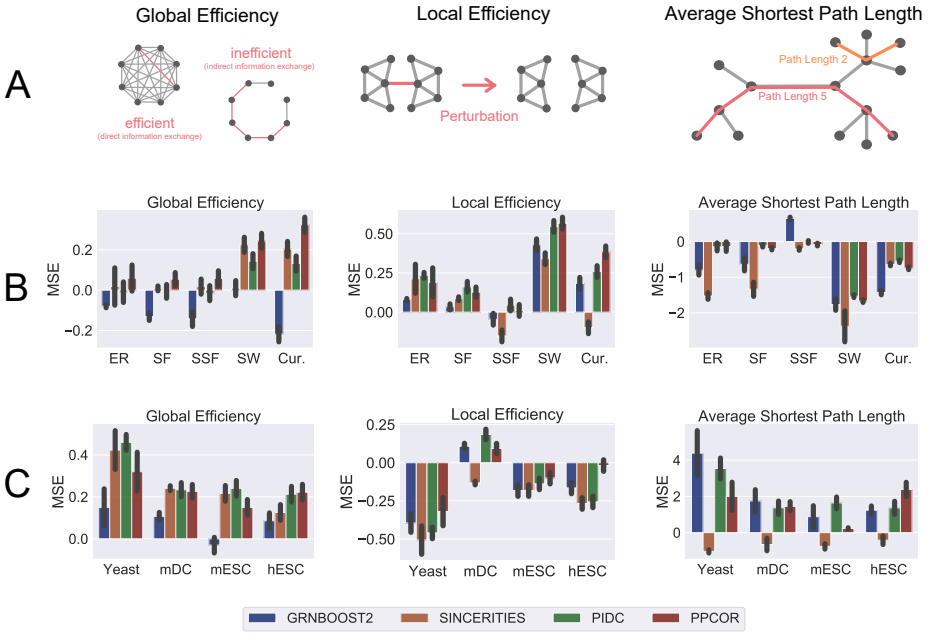

Figure 2: Results of the topological benchmarking of GRN inference algorithms with respect to information exchange both on synthetic and experimental scRNA-seq datasets. (A) Schematic representations of the three topological measures we computed (see Methods). Global Efficiency quantifies how well information can be distributed in the entire network. Local Efficiency measures how robust the network is to perturbation on a small scale. The Average Shortest Path Length specifies how many links are necessary to go from one node to another on average. (B) Barplots showing the Mean Signed Error (MSE) for the estimations of the topological properties written at the top in different types of synthetic networks (indicated on the x-axis) and for different algorithms (marked by colors). (C) Same as B, for networks estimated from real scRNA-seq datasets (indicated on the x-axis). The error bars display the standard deviations.

2.3 Hub analysis

One important downstream analysis on GRN is the identification of genes with a number of links much larger than the average. These are known as network hubs, and they can play key roles in differentiation and reprogramming [27]. Moreover, GRN hubs have been identified as potential disease regulators or drug targets [28]. The presence of such nodes depends on several topological properties that are influenced by the type of network. In particular, we expect the hubs in SF and SSF networks to be more easily identifiable due to their node degree distribution. This is reflected, for example, by the higher Centralization values of SF and SSF networks compared to other classes of networks (Figure S2).
Here, we first analysed how well each algorithm predicts the value of topological measures that quantify the tendency of networks to include hubs. Then, we test how well the inferred GRN preserves hub identities.

2.3.1 Hub-related topological quantities

The topological measures we chose to quantify the presence of hubs are Assortativity, Clustering Coefficient and Centralization (Figure 3A). In networks with larger values of Assortativity, nodes with lower degrees tend to be linked to nodes that feature a higher degree; hence, in these networks, hubs tend to be present and clearly identifiable. Networks with a large Clustering Coefficient feature groups of nodes with high interconnectivity that, thus, have similar node degrees. In this situation, hubs are less dispersed. The Centralization quantifies how centralized a graph is around a small number of nodes, which will have a large number of links, and will therefore tend to be strong and clearly identifiable hubs.
Among the networks we simulated data from, the Assortativity of SW and ER networks is relatively closer to 0 on average compared to other networks, while their Clustering Coefficient is lowest (Figure S2A). Conversely, SF and SSF have lower (negative) Assortativity and larger Clustering Coefficients together with the Cur networks. This is in agreement with the known properties of each class of networks [14]. As previously mentioned, the Cur networks are a group of four small subnetworks of real GRNs, therefore their properties are influenced by their small size and individual peculiarities rather than general characteristics.
The three network properties assessed during this step uncover multiple algorithm-specific behaviours (Figure 3B). Firstly, the networks inferred by SINCERITIES have lower Assortativity and larger Centralization than the corresponding ground truth graphs. However, the estimations of the Clustering Coefficient obtained with SINCERITIES tend to be more accurate than those coming from other algorithms.
As in the previous section, we repeated the analysis on real datasets (Figure 3C). The corresponding networks have Assortativity values that are lower than those of most synthetic networks and thus best comparable to SF and SSF networks. On average, the Clustering Coefficients are similar to those of ER or SW networks (Figure S2B). The distribution of the values of the Assortativity is

in line with the Scale-Free hypothesis for real GRN [14]. In summary, the gold standards display a higher tendency to contain hubs that are more clustered together.
Similarly to what happens with the synthetic datasets, here the Centralization is overestimated by SINCERITIES (Figure 3C). In contrast, the Assortativity is now preserved better by SINCERITIES rather than being underestimated.
GRNBoost2, PIDC and PPCOR show similar performances. These three algorithms overestimate the Assortativity as well as the Centralization, albeit less than SINCERITIES. Furthermore, the Clustering Coefficient is underestimated by all algorithms.
The difference in performance between the real and the synthetic data might be due to a number of factors. First, the gold standards provide only estimates for GRNs of the three organisms while the synthetic data are simulated from fully specified networks. Furthermore, we found that the algorithms tend to output networks that feature similar topological properties, regardless of whether the data is experimentally or artificially generated. This might explain, for example, the opposite trends in the estimations of the Assortativity and the Clustering Coefficient with the synthetic versus the real datasets.

2.3.2 Hub identification

While hubs are loosely defined as nodes having degrees higher than average, there is no consensus on the best metric to identify them. For this reason, here we compute four centrality measures that have been previously adopted to find hubs in GRNs [29]: the Betweenness [30], the Out-Centrality [31], the Radiality [32], and the Page Rank [33] (see Methods). We verify how accurately these measures are estimated by the four inference algorithms introduced above. More specifically, we select the set of top 10% nodes according to the centrality measure computed in the ground truth network, Ω_{true}, and in the inferred network, $\Omega_{inferred}$. Then, we quantify the similarity between the two sets of nodes with the Jaccard similarity index [34] (see Methods). Finally, we compute the ratio between J and J_{rand}, i.e., the Jaccard index between Ω_{true} and a set of randomly selected nodes Ω_{rand} (see Methods). Hence, the ratio J/J_{rand} shown in Figure 4 represents the improvement in the prediction of hubs in the inferred network compared to a random guess.
In general, we find that better scores are achieved on networks featuring stronger hubs (i.e., with larger values of Centralization, like the GRN associated with hESC; see Figure S2). However, the scores depend on the algorithm as well as the network type. In real networks, there is a wide range of Centralization values, with the hESC networks featuring higher values, which indicate the presence of more clearly identifiable hubs. Conversely, the yeast GRN is characterized by lower values of Centralization, and, thus, by fewer hubs or hubs that are less isolated (Figure S2). As a possible consequence, the best performance is achieved on hESC networks. Similarly, the performance on synthetic data is best on SF networks which are likely to possess hubs that are easier to identify due to their degree distribution.

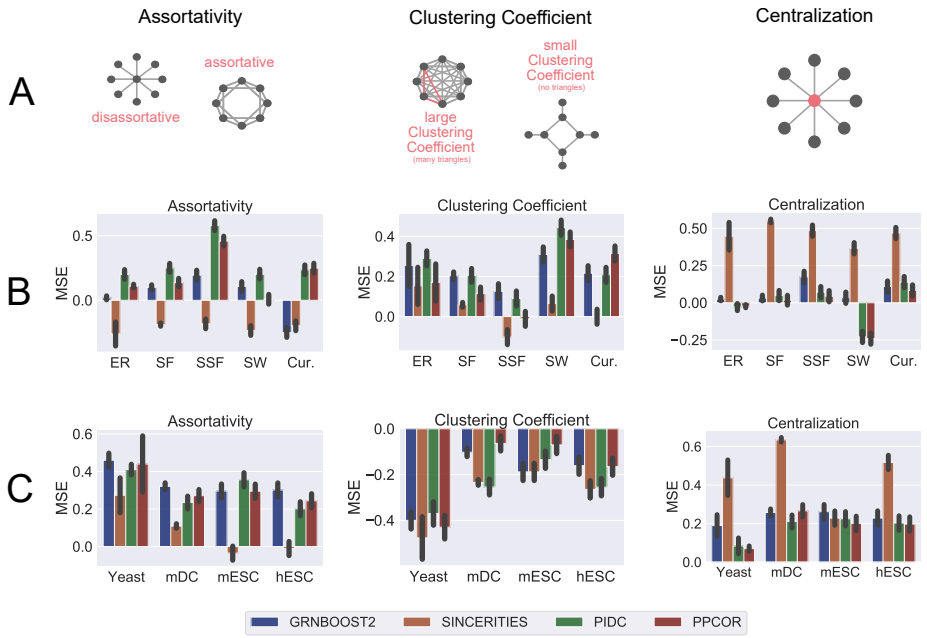

Figure 3: Results for the topological benchmarking of GRN inference assessing the presence of hubs. (A) Schematic representation of the three topological measures considered here (see Methods). The Assortativity quantifies the tendency of nodes in the networks to attach to others with similar degrees. The Clustering Coefficient reflects how much the nodes in a graph tend to cluster together. The Centralization indicates how strongly the network is arranged around a single centre. (B) Barplots showing the Mean Signed Error (MSE) for the estimations of the topological properties written at the top in different types of synthetic networks (indicated on the x-axis) and for different algorithms (marked by colors). (C) Same as B, for networks estimated from real scRNA-seq datasets (indicated on the x-axis). The error bars display the standard deviations.

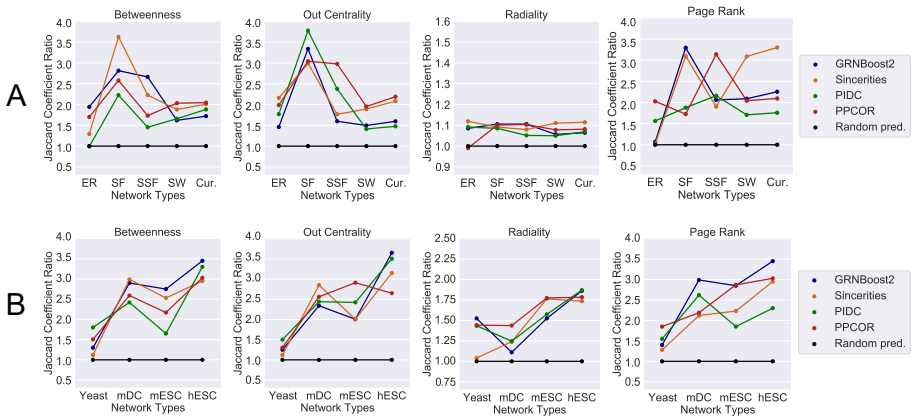

Figure 4: Accuracy of hub detection. The accuracy is measured by the Jaccard coefficient ratio with a random predictor as reference (see Methods and Eq. 15). (A) The Jaccard coefficient ratio is plotted with various hub metrics (written on top) as a function of the type of synthetic network indicated on the x-axes for different algorithms. (B) Same as (A), plotted as a function of the networks inferred from real scRNA-seq datasets (indicated on the x-axes). The Betweenness estimates the influence that a node has on the information exchange in a graph based on path lengths. The Out Centrality is the out-degree of a node in directed networks or its overall degree in undirected networks. The Radiality assigns high centrality values to nodes that have a short distance to all vertices in their reachable neighborhood compared to the graph diameter. PageRank is a generalization of the degree centrality that considers the eigenvalues of a modified adjacency matrix. We provide a detailed definition of the hub metrics in the text and in the Methods section.

(a) Information Exchange

Network type	Global Efficiency	Local Efficiency	Shortest Avg. Path Length
ER	PIDC	GRNBoost2	PPCOR
SF	PIDC	GRNBoost2	PIDC
SSF	PIDC	PPCOR	PIDC
SW	GRNBoost2	SINCERITIES	PIDC
Cur.	PIDC	SINCERITIES	PIDC
yeast	GRNBoost2	PPCOR	SINCERITIES
mDC	GRNBoost2	PPCOR	SINCERITIES
mESC	GRNBoost2	PPCOR	PPCOR
hESC	GRNBoost2	PPCOR	SINCERITIES

(b) Hub Presence

Network type	Assortativity	Clustering Coefficient	Centralization
ER	GRNBoost2	SINCERITIES	PPCOR
SF	GRNBoost2	SINCERITIES	PPCOR
SSF	SINCERITIES	PPCOR	PPCOR
SW	PPCOR	SINCERITIES	GRNBoost2
Cur.	SINCERITIES	SINCERITIES	PPCOR
yeast	SINCERITIES	PIDC	PPCOR
mDC	SINCERITIES	PPCOR	PIDC
mESC	SINCERITIES	PPCOR	PPCOR
hESC	SINCERITIES	PPCOR	PPCOR

(c) Hub Identification

Network type	Betweenness	Out Centrality	Radiality	Page Rank
ER	GRNBoost2	SINCERITIES	SINCERITIES	PPCOR
SF	SINCERITIES	PIDC	GRNBoost2	GRNBoost2
SSF	GRNBoost2	PPCOR	GRNBoost2	PPCOR
SW	PPCOR	PPCOR	SINCERITIES	SINCERITIES
Cur.	PPCOR	SINCERITIES	SINCERITIES	SINCERITIES
yeast	PIDC	PIDC	GRNBoost2	PPCOR
mDC	SINCERITIES	SINCERITIES	PPCOR	GRNBoost2
mESC	GRNBoost2	PPCOR	PPCOR	PPCOR
hESC	GRNBoost2	GRNBoost2	PIDC	GRNBoost2

Table 1: The above table summarizes which algorithms performed best in each step of our benchmarking pipeline, subdivided by dataset and property type. The top performing methods were selected according to the lowest Mean Signed Error (MSE) in absolute value.

3 Discussion

Here, we significantly extended the benchmarking analyses to evaluate how well GRN inference algorithms can estimate the structural properties of the networks. More specifically, we quantify the ability of the algorithms to infer their robustness to perturbations as well as the presence and identification of network hubs. For this purpose, we computed six topological measures and tested four metrics for hub identification. Moreover, we considered scRNA-seq data simulated from different types of networks as well as real data collected from different organisms. The results of the benchmarking are summarized in Table 1.

Overall, we observed that there is no single best-performing algorithm, but the performance depends on the properties of the ground truth network and the topological metric being considered. For example, with real datasets, GRNBoost2 achieves the best performance in the estimation of Global Efficiency; whereas SINCERITIES produces the most accurate estimations of Local Efficiency and Average Shortest Path Length in almost all the real datasets (Figure 2). A similar situation emerged for the metrics quantifying Hub Presence (Figure 3), where the estimations of Assortativity obtained with SINCERITIES are the most accurate, while PPCOR performed best when estimating the Clustering Coefficient and the Centralization in almost all the real datasets.

The situation for the Hub Identification is more complex, as the best-performing algorithm seems more susceptible to the specific dataset analyzed (Figure 4). However, overall, all algorithms achieve good performance with all datasets and metrics. A notable exception to this is the yeast dataset (Figure 4B), where the accuracy of predictions for all algorithms (and especially SINCERITIES) was relatively low. This might be linked to the ground truth network of yeast having less clearly defined hubs, as, e.g., the value of Centralization suggests (Figure S2).

The benchmarking done with synthetic networks allowed us to check the performance of algorithms with networks having specific and tunable properties. In some cases, this has brought to light specific biases present in the networks estimated by each algorithm. For example, SINCERITIES produces more disassortative and centralized networks (i.e., networks with relatively low Assortativity and high Centralization), which causes an underestimation of Assortativity and overestimation of Centralization for all types of synthetic networks (Figure 3). Similar observations can be made, for example, with GRNBoost2, which tends to generate networks with lower Global Efficiency (Figure 2).

While these biases are likely to derive from the specific strategies that each algorithm follows to estimate GRNs, their knowledge could be used to improve the algorithms: for example, by guiding the design of objective functions that could lead to networks with global properties closer to real GRNs. This approach can be justified by the observation that GRNs happen to share certain topological features, such as a scale-free node-degree distribution [35], which could be assumed as prior knowledge for the inference process.

Finally, topological quantities can also be used to optimize community-based

inference schemes. Currently, consensus networks are derived from the outputs of different methods by taking into account only their performance in the estimation of single edges. Instead, new strategies could be devised that also consider the estimation of the network's global properties.

4 Methods

4.1 Ground Truth networks and Gold Standards

4.1.1 Synthetic networks

We use parameter-controlled networks from four different classes as well as the Curated GRNs that have been used in BEELINE [8]. The output of the network samplers is a graph G with n nodes and m edges.

Random networks Random networks were created by the Erdös-Renyi $G(n, p)$ model, which outputs a graph with n nodes where each pair is connected with probability p [17]. We set p so that the expected number of edges equals m.

Scale-Free networks Networks with a degree distribution that follows a power law are classified as Scale-Free [36]. Given the parameter α, the expected degree distribution follows

$$P(d) \sim d^{-\alpha} \ . \tag{1}$$

For directed networks the in-degree distribution and the out-degree distribution can feature different parameters α_{in} and α_{out}. We applied combinations of different in- and out degrees. The exact values can be found in Table S1.

Semi-Scale-Free networks Following the analysis of the degree distributions in known GRNs [14], we sample Semi-Scale-Free networks which feature an out-degree distribution that follows a power-law but a uniform in-degree distribution. Additionally, only 50% of the nodes have outgoing edges.

Small-World networks We use the Watts-Strogatz model to sample networks that feature Small-World topology [15]. The algorithm starts with n nodes with degree k in a regular lattice and then rewires edges with probability p.

Curated networks Curated networks are four known GRNs that were used in BEELINE to evaluate the statistical performance of the GRN inference algorithms [8]. These networks are simple models for mammalian cortical area development, ventral spinal cord development, hematopoietic stem cell differentiation and gonadal sex determination. We include them for comparison.

Network sampling We use the Julia package LightGraphs.jl [37] to sample the networks explained above. The parameters were chosen such that a large variety of structurally different networks is covered. We simulate single-cell RNA-sequencing data using BoolODE [8]. For every parameter set we sampled data from 100 cells for 10 smaller and 10 larger networks. The exact specifications and number of networks can be found in Table S1.

4.1.2 Experimental networks

For the benchmarking of GRN inference on experimental single-cell RNA-sequencing data we select four datasets from human [21], mouse [19], [20], and yeast [18] and compare the output networks to different types of gold standard networks that were collected by Stone et al. [16]. The properties of the networks and the number of corresponding gold standards can be found in Table S2. For the detailed description of the preprocessing we refer to [16].

4.2 Inference algorithms

We select the four top-performing algorithms from BEELINE [8] and examine the results using our three-step benchmarking pipeline.

GRNBoost2: GRNBoost2 [25] infers a GRN independently for each gene, by identifying the most important regulators using a regression model. It is an alternative to GENIE3, which uses a similar inference scheme but does not scale to larger datasets due to its runtime.

SINCERITIES: SINCERITIES [24] is a causality based method, that computes temporal changes in the expression of each gene. The GRN is inferred by solving a specifically formulated ridge regression problem.

PIDC: The PIDC inference scheme [22] is based on partial information decomposition, which is a multivariate information theoretic measure for triplets of random variables. Since it is symmetric, the resulting network is undirected.

PPCOR: PPCOR [23] calculates the partial and semi-partial correlation coefficients for every possible pair of genes. Edges are ranked according to these values. By using the correlation as sign, it is possible to assign a direction to the interactions but we found better topological performance in the undirected version.

4.3 Quantities used for benchmarking

4.3.1 Binary Edge Detection

In order to statistically benchmark the binary classification problem of correctly retrieving edges we followed previous studies [8], [9] by calculating the area

under the precision recall curve (AUPRC) and the area under the receiver operator curve (AUROC) on synthetic data. For the experimental datasets we use the early precision (EPr) among the top k edges, where k is the number of interactions in the corresponding gold standard. The EPr is better suited to classification accuracy on large datasets where the reference networks is not the entire ground truth.

4.3.2 Topological graph properties related to information exchange

For our topological analysis we select six different properties that allow for a meaningful structural characterisation of a network and group them with respect to their matter as either being related to the exchange of information or the presence of hubs. We assume a graph with n nodes and m edges.

Average shortest path length: The *Average Shortest Path Length* measures by how many links two random nodes are connected on average. $d(v, w)$ denotes the distance between two nodes v and w.

$$a = \sum_{v,w} \frac{d(v,w)}{n \cdot (n-1)} \ . \tag{2}$$

Global efficiency: The *Global efficiency* estimates how efficiently information is exchanged in the network on a global scale [26]. It is given by

$$E_{glob}(G) = \frac{E(G)}{E_{max}} \ , \tag{3}$$

where

$$E(G) = \frac{1}{n * (n-1)} \sum_{v \neq w} \frac{1}{d(v,w)} \tag{4}$$

and E_{max} is the maximum E_{glob} for a fully connected graph given a number of nodes. Since gene regulation can be interpreted as information exchange between nodes in GRN, E_{glob} is a meaningful quantity to estimate.

Local efficiency: The *Local efficiency* describes the resistance of the network to perturbation on a small scale [26]. It is defined as

$$E_{loc}(G) = \frac{1}{n} \sum_{v} E(G_v) \tag{5}$$

where G_v is the graph that only consists of v and its immediate neighbors. In practice there can be many factors that can perturb gene regulation, thus the local efficiency is an important score to be preserved, especially on experimental data.

4.3.3 Topological graph properties related to the presence of hubs

Assortativity: The preference for a network's nodes to attach to others that have a similar degree is captured by the Assortativity [38]. It is quantified by the *Assortativity Coefficient* r, which is the Pearson correlation coefficient between the degree k of a node and the average degree of its neighbors $\langle k_{nn} \rangle$ when treating them as random variables.

$$r = Cor(k, \langle k_{nn} \rangle) \ . \tag{6}$$

Networks with a negative Assortativity Coefficient are called disassortative, networks with a positive r are called assortative. Disassortative networks have a higher tendency to possess hubs, which is an important feature of GRN that we examine in section 2.3.

Centralization: The goal of the *Centralization* H is to provide an estimate of how centralized a graph is around the node v^* which has the highest degree [31]. It is defined as

$$H = \sum_{i=1}^{n} \frac{1}{n-1} |deg(v^*) - deg(v_i)| \ . \tag{7}$$

A highly centralized network is focused around a small number of nodes which could be identified as biologically important.

Clustering Coefficient: The Clustering Coefficient measures the extend to which nodes in a graph tend to cluster together. It is quantified by the *local clustering coefficient:*

$$C_v = \frac{L_v}{2 * k_v * (k_v - 1)} \tag{8}$$

and the *global clustering coefficient* [15]:

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_{v_i} \ . \tag{9}$$

where L_v represents the number of links between the k_v neighbors of node v. For our analysis we focused on the global clustering coefficient since it captures clustering on a global scale. A network with a larger clustering coefficient is more interconnected which can result in more complicated gene regulations.

4.3.4 Evaluation score

Since we are particularly interested in whether certain topological features are over- or underestimated we employ the Mean Signed Error (MSE) as evaluation quantity. For a property P which is being analyzed on networks $G_1, G_2, ..., G_n$ it is computed by

$$MSE(P) = \frac{1}{n} \sum_{i=1}^{n} P(G_i) - P(G_{i,\text{inferred}}) \ . \tag{10}$$

4.3.5 Hub identification

Degree centrality: The degree centrality purely evaluates the degree of a node in a network. In a directed graph there are two different degree centralities: the in-degree centrality and the out-degree centrality. In the directed case we use the out-degree centrality, in the undirected case we use the overall degree centrality.

Betweenness centrality: Betweenness centrality describes the extend to which nodes stand between each other. It was formally defined by Freeman [30] as

$$B_C(v) = \sum_{u \neq v \neq w} \frac{\eta_v(u,w)}{\eta(u,w)} \ , \tag{11}$$

where $\eta(u,w)$ is the number of shortest paths from u to w and $\eta_v(u,w)$ is the number of shortest paths from u to w passing through v. If the graph is not connected the betweenness centrality is evaluated for each connected component.

Page Rank centrality: The Page Rank centrality is the output of the Page Rank algorithm which is focused on link analysis. The output is a distribution that models the likelihood to reach any particular node when randomly moving along edges. Details of the algorithm can be found in [33].

Radiality: Radiality [32] considers the global structure of the networks and indicates how connected an individual is in the entire network structure. The definition makes use of the reverse distance matrix

$$R_D = (\max_{v,w \in G} d(v,w) + 1 - d(v_i, v_j))_{i,j} \ , \tag{12}$$

where $d(v,w)$ describes the shortest path length from v to w. Radiality is defined as

$$R(v_j) = \frac{1}{n-1} \sum_{i \neq j} R_D[i,j] \ . \tag{13}$$

Evaluation score: We rank the nodes of each network according to the centrality measures defined above and in networks with more than 50 nodes we classify the top 10% as hubs while in smaller networks we select the top 20%. Our goal is to analyze the set similarity between the hubs in the ground truth or gold standard Ω_{true} and the inferred networks Ω_{inf}. Therefore we compute the average Jaccard coefficient [34] for every network class, which is given by:

$$J(\Omega_{true}, \Omega_{inf}) = \frac{|\Omega_{true} \cap \Omega_{inf}|}{|\Omega_{true} \cup \Omega_{inf}|} \ . \tag{14}$$

To quantify the performance of the different centrality measures, we use as evaluation score the ratio between the Jaccard coefficient based on the inferred

networks and the expected coefficient for a random predictor which can be calculated explicitly [39]. If the fraction of nodes that is classified as hubs is given by $p \in (0.1]$, the expected value for this Jaccard coefficient is given by:

$$\mathbb{E}[J(\Omega_{true}, \Omega_{random})] = \frac{p^2}{2p - p^2} \tag{15}$$

4.4 Code availability

STREAMLINE is available at `github.com/ScialdoneLab/STREAMLINE`.

Acknowledgments

Jonathan Fiorentino and Marco Stock were supported by a Joachim Herz Stiftung Add-on Fellowship for Interdisciplinary Life Science. Marco Stock was supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS". Work in the Scialdone lab is funded by the Helmholtz Association.

References

[1] M. Hagemann-Jensen, C. Ziegenhain, P. Chen, *et al.*, "Single-cell rna counting at allele and isoform resolution using smart-seq3," *Nature Biotechnology*, vol. 38, no. 6, pp. 708–714, 2020.

[2] K. Akers and T. Murali, "Gene regulatory network inference in single-cell biology," *Current Opinion in Systems Biology*, vol. 26, pp. 87–97, 2021.

[3] M. M. Saint-Antoine and A. Singh, "Network inference in systems biology: Recent developments, challenges, and applications," *Current opinion in biotechnology*, vol. 63, pp. 89–98, 2020.

[4] M. P. Stumpf, "Inferring better gene regulation networks from single-cell data," *Current Opinion in Systems Biology*, vol. 27, p. 100 342, 2021.

[5] P. Brennecke, S. Anders, J. K. Kim, *et al.*, "Accounting for technical noise in single-cell rna-seq experiments," *Nature methods*, vol. 10, no. 11, pp. 1093–1095, 2013.

[6] M. Banf and S. Y. Rhee, "Computational inference of gene regulatory networks: Approaches, limitations and opportunities," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1860, no. 1, pp. 41–52, 2017.

[7] T. Hillerton, D. Seçilmiş, S. Nelander, and E. L. Sonnhammer, "Fast and accurate gene regulatory network inference by normalized least squares regression," *Bioinformatics*, vol. 38, no. 8, pp. 2263–2268, 2022.

[8] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. Murali, "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data," *Nature methods*, vol. 17, no. 2, pp. 147–154, 2020.

[9] S. Chen and J. C. Mar, "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–21, 2018.

[10] Y. Kang, D. Thieffry, and L. Cantini, "Evaluating the reproducibility of single-cell gene regulatory network inference algorithms," *Frontiers in genetics*, p. 362, 2021.

[11] Y. Guo and A. Amir, "Exploring the effect of network topology, mrna and protein dynamics on gene regulatory network stability," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[12] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A guide to conquer the biological network era using graph theory," *Frontiers in bioengineering and biotechnology*, p. 34, 2020.

[13] N. A. Kiani, H. Zenil, J. Olczak, and J. Tegnér, "Evaluating network inference methods in preserving the topology and complexity of reconstructed genetic networks," *Semin. Cell Dev. Biol*, vol. 51, pp. 44–52, 2016.

[14] W. Z. Ouma, K. Pogacar, and E. Grotewold, "Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties," *PLoS computational biology*, vol. 14, no. 4, e1006098, 2018.

[15] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[16] M. Stone, J. Li, S. G. McCalla, *et al.*, "Identifying strengths and weaknesses of methods for computational network inference from single cell rna-seq data," *bioRxiv*, pp. 2021–06, 2022.

[17] P. Erdős and A. Rényi, "On random graphs i," *Publicationes mathematicae*, vol. 6, no. 1, pp. 290–297, 1959.

[18] A. P. Gasch, F. B. Yu, J. Hose, *et al.*, "Single-cell rna sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress," *PLoS biology*, vol. 15, no. 12, e2004050, 2017.

[19] A. K. Shalek, R. Satija, J. Shuga, *et al.*, "Single-cell rna-seq reveals dynamic paracrine control of cellular variation," *Nature*, vol. 510, no. 7505, pp. 363–369, 2014.

[20] K. A. Tran, S. J. Pietrzak, N. Z. Zaidan, *et al.*, "Defining reprogramming checkpoints from single-cell analyses of induced pluripotency," *Cell reports*, vol. 27, no. 6, pp. 1726–1741, 2019.

[21] X. Han, H. Chen, D. Huang, *et al.*, "Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell rna-sequencing," *Genome biology*, vol. 19, no. 1, pp. 1–19, 2018.

[22] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene regulatory network inference from single-cell data using multivariate information measures," *Cell systems*, vol. 5, no. 3, pp. 251–267, 2017.

[23] S. Kim, "Ppcor: An r package for a fast calculation to semi-partial correlation coefficients," *Communications for statistical applications and methods*, vol. 22, no. 6, p. 665, 2015.

[24] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan, "Sincerities: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles," *Bioinformatics*, vol. 34, no. 2, pp. 258–266, 2018.

[25] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, *et al.*, "Grnboost2 and arboreto: Efficient and scalable inference of gene regulatory networks," *Bioinformatics*, vol. 35, no. 12, pp. 2159–2161, 2019.

[26] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198 701, 2001.

[27] J. Kim, S. T. Jakobsen, K. N. Natarajan, and K.-J. Won, "Tenet: Gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data," *Nucleic acids research*, vol. 49, no. 1, e1–e1, 2021.

[28] J. Åkesson, Z. Lubovac-Pilav, R. Magnusson, and M. Gustafsson, "Comhub: Community predictions of hubs in gene regulatory networks," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–12, 2021.

[29] D. Koschützki and F. Schreiber, "Centrality analysis methods for biological networks and their application to gene regulatory networks," *Gene regulation and systems biology*, vol. 2, pp. 193–201, 2008.

[30] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[31] ——, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[32] T. W. Valente and R. K. Foreman, "Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network," *Social networks*, vol. 20, no. 1, pp. 89–105, 1998.

[33] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Stanford InfoLab, Tech. Rep., 1999.

[34] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[35] F. M. Lopes, D. C. Martins Jr, J. Barrera, and R. M. Cesar Jr, "A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks," *Information Sciences*, vol. 272, pp. 1–15, 2014.

[36] Y. S. Cho, J. S. Kim, J. Park, B. Kahng, and D. Kim, "Percolation transitions in scale-free networks under the achlioptas process," *Physical review letters*, vol. 103, no. 13, p. 135 702, 2009.

[37] J. Fairbanks, M. Besançon, S. Schölly, J. Hoffiman, N. Eubank, and S. Karpinski, *Juliagraphs/graphs. jl: An optimized graphs package for the julia programming language*, 2021. [Online]. Available: `https://github.com/JuliaGraphs/Graphs.jl/`.

[38] M. E. Newman, "Mixing patterns in networks," *Physical review E*, vol. 67, no. 2, p. 026 126, 2003.

[39] N. C. Chung, B. Miasojedow, M. Startek, and A. Gambin, "Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data," *BMC bioinformatics*, vol. 20, no. 15, pp. 1–11, 2019.

Supplementary Material

S1. Parameters of the reference networks

The following two tables summarize the properties of the ground truth and gold standard networks. For the synthetic networks we list the parameters that were used for sampling as well as the number of networks that resulted. For the experimental networks the number of cells and genes are shown as well as the number of reference databases that were used. A more elaborated discussion of the properties of the real datasets is given by Stone et al.[16] who collected them.

network type	#networks	#nodes	#edges	parameters
Random (ER)	60	[15,25]	[25,50]	none
Scale-Free (SF)	180	[15,25]	[25,50]	$\alpha_{in} \in [0.2, 0.5, 0.8]$
				$\alpha_{out} \in [0.2, 0.5, 0.8]$
Semi-Scale-Free (SSF)	60	[15,25]	[25,50]	$\alpha_{out} \in [0.2, 0.5, 0.8]$
Small-World (SW)	80	[15,25]	[30,50]	$\rho \in [0.1, 0.2, 0.5, 0.8]$

Table S1: Parameters used to sample synthetic networks. We used all possible combinations which resulted in over 300 sampled networks in total.

cell type	#cells	#genes	#reference databases
yeast [18]	163	3847	7
mDC [19]	1211	9411	3
mESC [20]	2369	6618	10
hESC [21]	5520	7468	8

Table S2: Properties of the experimental datasets and the number of reference databases used as gold standards.

S2. Binary Edge Detection

The results of our statistical benchmarking, shown in supplementary figure S1, reproduce the previous observation that inference accuracy in terms of AUPRC, AUROC and EPr is moderate at best [8], [9]. For synthetic data (Fig S1A-B), the highest prediction accuracy was achieved on Curated and Small-World networks, while on Scale-Free, Semi-Scale-Free and Random networks the performance is worse. Regarding the performance of the specific algorithms, GRNBoost2, PIDC and PPCOR perform similarly, while SINCERITIES identifies fewer correct interactions. This trend transfers to experimental data (Fig S1C). Additionally, the statistical performance on the yeast dataset is worse than on mDC, mESC and hESC.

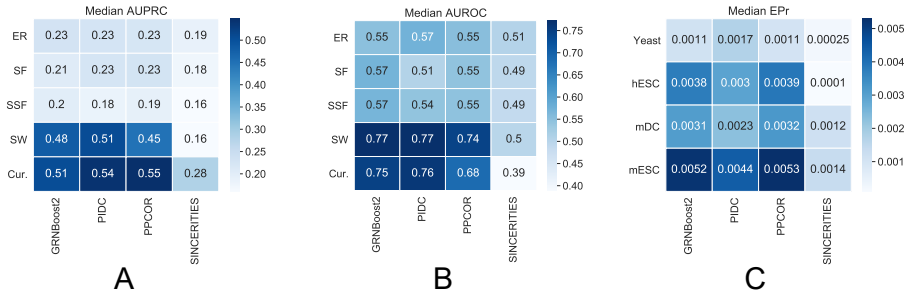

Figure S1: Benchmarking of edge detection of GRN inference algorithms on synthetic and experimental scRNA-seq datasets. The heatmaps report the median AUPRC and AUROC for synthetic data (panels A and B, respectively) and the median EPr for experimental data (panel C). Rows and columns correspond to GRN inference algorithms and network types, respectively.

Binary Edge Detection

Network type	EPr	AUPRC	AUROC
ER		PIDC	PIDC
SF		PPCOR	GRNBoost2
SSF		GRNBoost2	GRNBoost2
SW		PIDC	GRNBoost2
Cur.		PPCOR	PIDC
yeast	PIDC		
mDC	PPCOR		
mESC	PPCOR		
hESC	PPCOR		

Table S3: The table above lists the best performing algorithms for the task of edge detection from the ground truth networks as a binary classification problem. The best performing algorithms were selected according to the highest EPr scores for the experimental datasets or the highest AUPRC and AUROC scores for the synthetic datasets.

S3. Topological properties of the reference networks

Figure S2 depicts the topological properties of the sampled ground truth networks as well as the gold standards for the experimental datasets.

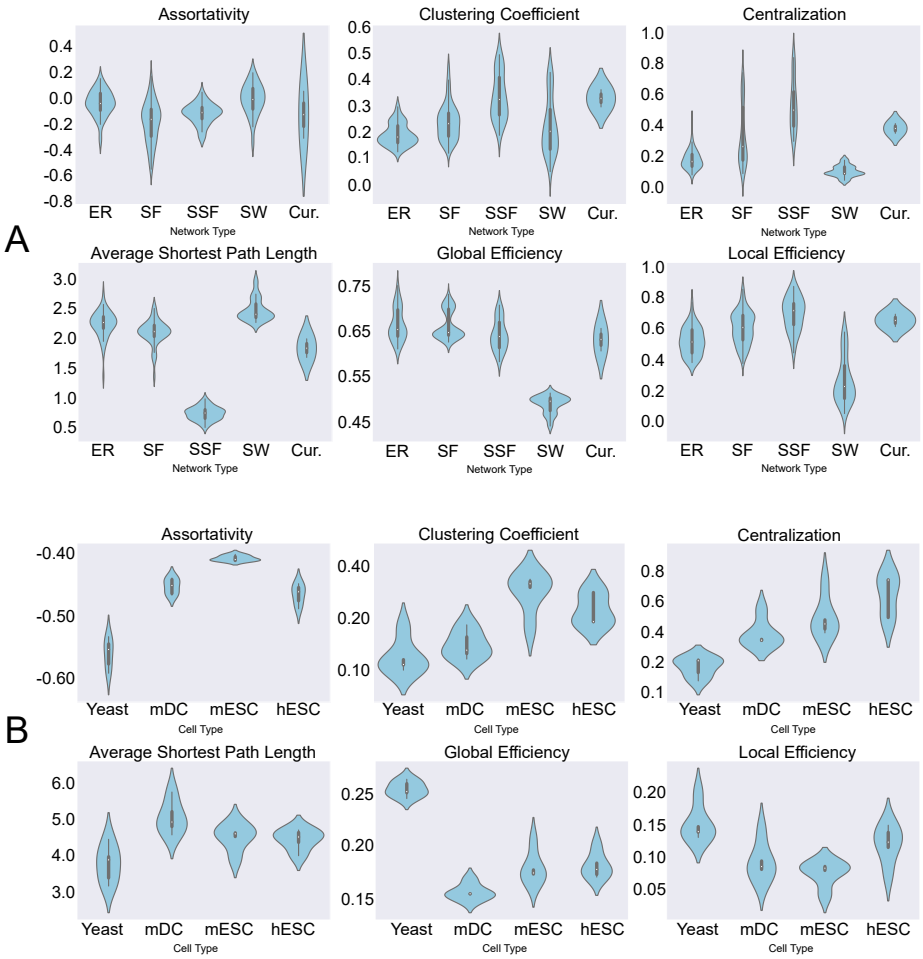

Figure S2: Distribution of the topological properties of the synthetic (panel A) and experimental (panel B) ground truth networks.