Report: A Geometric Deep Learning Framework for Single-cell Trancriptomics Data Analysis

Elyas Heidari[*1]***, Shayan Shekarforoush***[†2]***, Wolfgang Huber***[‡3]***, and Laleh Haghverdi***[§3]

[1]Department of Biological Systems Sciences and Engineering, ETH Zürich, Switzerland
[2]Sharif University of Technology, Iran
[3]EMBL Heidelberg, Germany

[*]eheidari@student.ethz.ch [†]shekshaa.75@gmail.com [‡]wolfgang.huber@embl.de
[§]laleh.haghverdi@embl.de

September 03, 2020

Abstract

Recent advancements in single-cell RNA sequencing (scRNAseq) have unraveled the transcriptional diversities underlying heterogeneous batches of cells. Current single-cell classification methods take genes as independent features to assign cell types disregarding their interactome which has been shown to be of great importance in cellular processes. As a result, existing methods are prone to dataset specific classification artifacts which prevents generalization of the model for cell type classification in new datasets. Here, we introduce single-cell Graph Convolutional Network (scGCN) which takes gene-gene interactions into account by constructing a co-expression network of a subset of genes and using a graph deep learning approach to classify single-cells. Using two different peripheral blood mononuclear cell (PBMC) datasets as train and test sets, we demonstrate the potentials of scGCN for transfer learning.

Contents

1 Introduction and motivation

Single cell transcriptomics have enabled resolution of transcriptional diversities of individual cells, hence the possibility to classify cells' identity and function via their transcriptional state. Noting that such identity classifications can elucidate organization and function of tissues in health and disease states, recently more investments have been devoted to this task, for example, in the Human Cell Atlas project (**???**). Current single-cell RNA sequencing (scRNA-seq) classification methods use genes expression as independent features to assign cell types, disregarding their interactome which is of great importance in consistency of different cell states. Cell type classification based on a few marker genes, for example, yields better results in intra data set settings where both train and test cells are sampled from the same dataset (Abdelaal et al. 2019) and may fail to correctly classify cell types under perturbations when expression of the predefined marker genes varies with respect to the control reference data set . Moreover, using genes expression as independent features can delude cell lineage relationships for cell states and cell types with subtle difference in their gene expression profiles. In (Fig. 1) we illustrate the two fold benifit of a graph represenation of cell states in removing unwanted and technical variations (batch effects) as well as providing a better resolution of transitory cell states and cell lineages.

Considering the modular architecture of gene regulatory networks (GRNs), we propose to further classify and use this information for cell type identification in new data sets. Whereas unwanted variations among different data sets makes any supervised learning method prone to overfitting and dataset specific model artifacts, we suggest that a supervised learning approach on GRNs imposes the proper inductive bias (Baxter (2000), Neyshabur, Tomioka, and Srebro (2014)) for 'transferable learning' (Levie, Isufi, and Kutyniok 2019), such that an available data set can be used to train a model which is able to predict cell types in new (unseen) data sets later on.

Recent advancements of deep learning models to capture modular structures in other applications such as image, audio and speech analysis as well as emergence of high-throughput scRNA-seq technologies which provide the large amounts of data required for such models motivated us to employ deep learning models applicable to non-regular graph data. Amongst such models, which are generally referred to as "geometric deep learning" (Bronstein et al. (2017), Monti et al. (2017)), Graph Convolutional Networks (GCNs) have significant power in extracting informative feature maps from signals defined over nodes of a graph while considering meaningful local structures and connectivities among them. This makes GCNs suitable for our purpose of analyzing non-regular gene networks (Zhou et al. 2018).

To meet the challenges described above, we implement single-cell Graph Convolutional Network (scGCN) which takes gene-gene interactions into account by constructing a regulatory network among a subset of genes and uses a graph deep learning approach to classify single cells. The pipeline is an integrated data flow, with tunable options to achieve the most accurate cell type annotation. Moreover, we provide multiple visualization tools including t-SNE, UMAP[1] as well as a few statistical measures at different stages of the model} to facilitate interpretation (Fig. 2[2]).

[1] **t**-distributed **S**tochastic **N**eighbor **E**mbedding (t-SNE) (Maaten and Hinton 2008) and **U**niform **M**anifold **A**pproximation and **P**rojection (UMAP) (McInnes, Healy, and Melville 2018) are non-linear methods for dimension reduction, widely used in the fields of machine learning and computatoinal biology.

[2] **HVG:** highly variable gene, **TF:** transcription factor, **PGM:** probabilistic graphical models

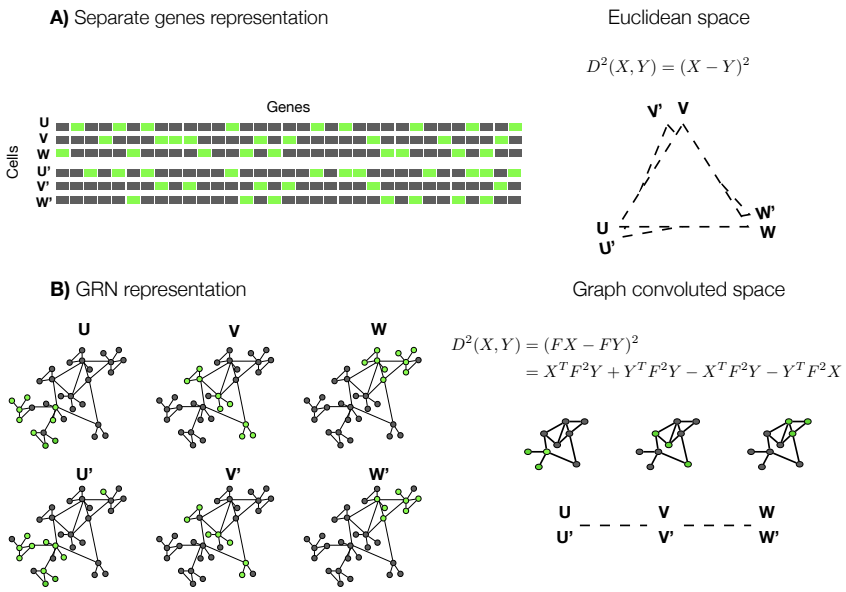

Figure 1: Graph convolution metric versus the Euclidean metric
(A) In separate genes representation and Euclidean space cell states X, Y, Z (and their corresponding cell states X‘, Y‘, Z‘ from another batch of data) appear as equidistant, whereas in (B) a GRN representation and the metric space defined by a graph kernel F, cell state Y (Y‘) appears between X (X‘) and Z (Z‘), marking the cell lineage trajectory between them.

2 The scGCN learning framework

We propose a deep learning framework for analysis of single-cell RNA-seq data. Our framework consists of three steps; 1) adaptation of a geometric deep learning model applicable to gene regulatory networks (GRNs), 2) construction of a modular GRN, which serves as the input graph structure for the geometric deep learning model, and 3) validation of the model on available scRNA-seq data sets. For each of these steps, we face a multitude of possible methods. We would like to systematically explore the solution space in order to find the best amongst the set of possible solutions. This, brought us to first integrating all steps into one pipeline, and then optimizing the whole pipeline to achieve the best solution.

We designed a pipeline of modules each of which represent a particular functionality. Each module is parametrized to ensure flexibility in trying different methods available for each step. This translates the whole solution space into a parameter grid, where parameters are methods and their arguments. Such an approach enables us to examine different solutions as paths of parameters on the parameter grid. The pipeline has three main modules including 1) preparation of GCN's input, 2) the InceptionGCN and 3) Analysis of results, each of which consist of multiple submodules. In the following we describe each of the modules in detail.

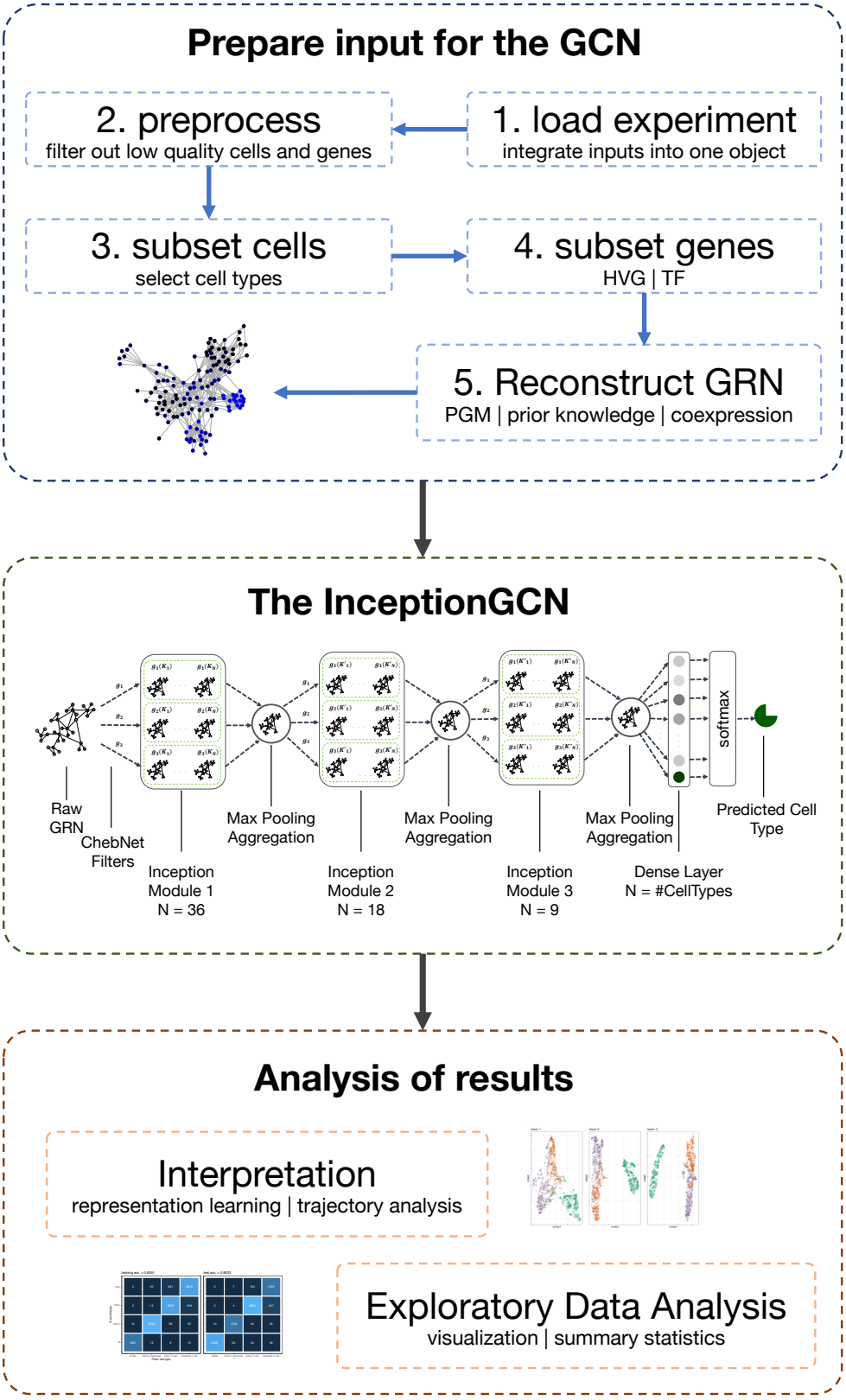

Figure 2: The graphical abstract of scGCN

Module 1: prepare GCN's input

After selection of training and test datasets, one has to preprocess them and prepare them to feed them to the GCN module. Preprocessing includes three steps, first, cell type selection and cell-wise data preprocessing, second, gene subset selection and lastly, GRN reconstruction. For each of these steps, we encountered some challenges and define our solution space as follows.

- **Cell type selection & preprocessing:** This stat is straightforward, in that one should select a subset of cell type in the intersection of cell types present in both training and test datasets. Afterwards, data for each cell should be preprocessed to rule out major technical noises. We used a very general method to perform preprocessing, that is cell-wise log scaling, with no tunable parameter.
- **Gene subset selection:** The main objective for gene subset selection is to select genes which are most relevant to cell type identification. There are multiple ideas discussed in the literature among which we examined networks built on transcription factors (TFs) and networks built on highly variable genes (HVGs). Because of computational limitations, feasible number of genes in the GRN as the input of GCN ranges between 100 and 1000 depending on biological specifications of the data.
- **GRN reconstruction:** There are numerous methods for GRN reconstruction proposed, some of which rely on knowledge bases, such as RegNetwork (Liu et al. 2015), and some simply use the data itself to estimate the underlying GRNs for the genes in hand, probabilistic graphical models and co-expression networks embody such an approach. With the aim of choosing the best method for our pipeline, we examined both approaches. The most important features of the GRN we were seeking were sparsity and modularity of the GRN.

We sought a GRN representation of single-cell data which we could feed into a GCN to accurately annotate cells types generalizable to new datasets. Due to computational limitations, one has to reconstruct such a GRN of a subset of all expressed genes in the training dataset. Furthermore most of the methods we tried for GRN reconstruction, including knowledge-based or data-based, produced neither modular nor sparse networks and hence turned out not suitable for GCNs application that are inherently designed to exploit local (modular) structures of the input graphs to perform the learning task.

Overall, while seeking the best way to prepare input data for the GCN among multitude of solutions, we restricted our solution space to the ones which are computationally feasible and potentially enhance the learning efficiency of the GCN the most.

Module 2: the GCN

Among geometric deep learning strategies we decided to use the implementation of the InceptionGCN (Kazi et al. 2019), a Graph Convolutional Network (GCN) approach that captures a great range of distinct locality sizes on the network, hence suitable for the context of learning on modular gene networks. Moreover, as an instance of a regular GCN, InceptionGCN has a rigorous mathematical formulation, by spectral graph theory, hence, interpretable in contrast to most of other geometric deep learning models. The (hyper)parameters of InceptionGCN, e.g., number of layers, number of hidden nodes, and locality sizes should be tuned to find the optimal solution.

Module 3: validation and interpretation of results

The last module of the pipeline precesses the output of the GCN and tries to interpret them. This is utilized by proper visualizations and computations to evaluate specific measures, such as annotation accuracy, technical noise removal, and computation dynamics.

3 The `scGCNUtils` package

Once we parametrized our solution space, we implement the pipeline as a software package in R and python. The package `scGCNUtils` [ref] provides the user with the building blocks of the pipeline as well as performing pre-/post- exploratory data analysis (EDA). The user is able to run an end-to-end pipeline by defining their desired parameters, and decide on the ultimate parameter path using EDA tools provided for each module. The package also facilitates benchmarking, in that each step of the pipeline is implemented as an independent module and can be tweaked or substituted by another module with the same functionality. In the end, the performance and efficiency of the path can be validated by the package. Logging and caching are added to the package for the same purpose.

Ever since the beginning of the project, we reformatted the implementation from scratch quite a few times. For each dataset, we had to implement an ad-hoc pipeline, as different labs publish their studies in different formats, besides using different scRNA-seq platforms. We had to integrate new datasets into specific data objects in our programming environment, and pass them through a pipeline, specifically tailored for the dataset in hand. Eventually, we decided to develop a package to automate everything for the user. We integrated all of the tools we had exploited throughout the project into a single end-to-end pipeline, in which one can set specific parameters to validate the parameter path optimality. Also, each input experiment is stored in a specific data object in the environment with certain features and attributes which are modified in the pipeline data flow step by step. The whole pipeline is optimized in memory and performance by using appropriate data structures and computational resource optimization. Integrating everything into one single package also facilitates benchmarking, as now carrying out new experiments are made as convenient as reparametrization of the pipeline.

The package is designed to meet state of the art reproducibility standards and criteria of open science, such as coherence, integrity, documentation, readability, and testability.

4 Experiments and results

We examined the functionality of our framework on several datasets, e.g., (Paul et al. 2015), Nestorowa et al. (2016)], and (Cao et al. 2019). The following path turned out to be the optimal for all most all of our experiments.

1. **Cell subset selection:** Select a balanced subset of cells from the training dataset, with respect to the cell types (at least 500 cells from each cell type). Select cells from the same cell types, from the test dataset.
2. **Gene subset selection:** Subset both datasets to the genes which are highly variable in the training dataset (at most 1000 genes).
3. **GRN reconstruction:** Use the glasso algorithm[3] and tune its parameters to reconstruct the GRN on the training dataset.

[3] A PGM to construct sparse Graphical Models (Friedman, Hastie, and Tibshirani 2008).

4. **Cell type classification:** Employ InceptionGCN and tune its hyper parameters (dependent on the input dataset), to achieve the optimal classification.

However, the nature of each dataset essentially determines the best path. Therefore, for a new dataset multiple paths should be explored.

4.1 The PBMC datasets

The pivotal application of our method is transferring (generalizing) a learned model to new datasets in which overcoming intra-dataset variations, e.g., batch effects is of great importance. Thus, we sought to find two dataset with the same cell type profile, but acquired in separate experiments to inspect how well our method is able to overcome unwanted variations. Here, we present an application of the scGCN on two human peripheral blood mononuclear cell (PBMC) scRNA-seq experiments (Ding et al. 2019). The dataset contains two separate experiments, hereafter called PBMC 1 and PBMC 2, generated in a single center, but employing seven different scRNA-seq methods. This suits our goal to verify our framework on a simulated task of real transfer learning. As in potential applications we expect our method to learn annotation priorly on a complete dataset and transfer the learned model to annotate an unseen one.

4.1.1 Input

We take PBMC 1 as our training dataset and PBMC 2 as our test dataset. The task is to learn the model on PBMC 1, and validate it on PBMC 2, by predicting cell type annotations, and comparing them to true labels. We select a subset of size 6 of cell types, to carry out our computational experiment. Cells with less than 100 counts and cells with high percentage of mithocondrial gene counts ($>$0.01) are filtered out due to low sequencing/experimental quality. We subsample T cells, B cells, Natural killer cells, Megakaryocytes, Denderic cells, and monocytes, each of which with 150 cells in the training dataset. We do not subsample the test dataset for the aforementioned cell types. Afterwards, a subset of 300 most highly variable genes is selected from PBMC 1 and PBMC 2 is subset to the same set of genes. For GRN reconstruction, we use the glasso algorithm with tuned parameters, which gives us a modular as well as sparse GRN. The GRN is presented in Fig. 3.

Fig. 3(A) and Fig. 3(B) indicate that although our GRN reconstruction method is based on data and no prior knowledge, it is robust across datasets. As one finds in the figures, the network communities from PBMC 1, perfectly transfer to PBMC 2. Fig. 3(C) illustrates the same observation, as the density of expressions, i.e., activated modules are the same for both datasets, for each cell type. In fact, such an observation motivated us for the whole idea of exploiting a network topology to classify cell types.

4.1.2 Classification with InceptionGCN

We implement an InceptionGCN with three hidden layers with 36, 18, and 9 hidden nodes, respectively. At each layer, a complex, non-linear transformation (Chebnet filters) are applied to the inputs (see Fig. 2). One can describe such operations as diffusing information along the input GRN which further facilitates an accurate classification. Our InceptionGCN gives an accuracy of 90% on test dataset, which is rather a very high accuracy in a transfer learning task, specifically, when the classes are quite similar to each other (including two classes of T cells). Fig. 4 indicates the results of the classification on training and test datasets.

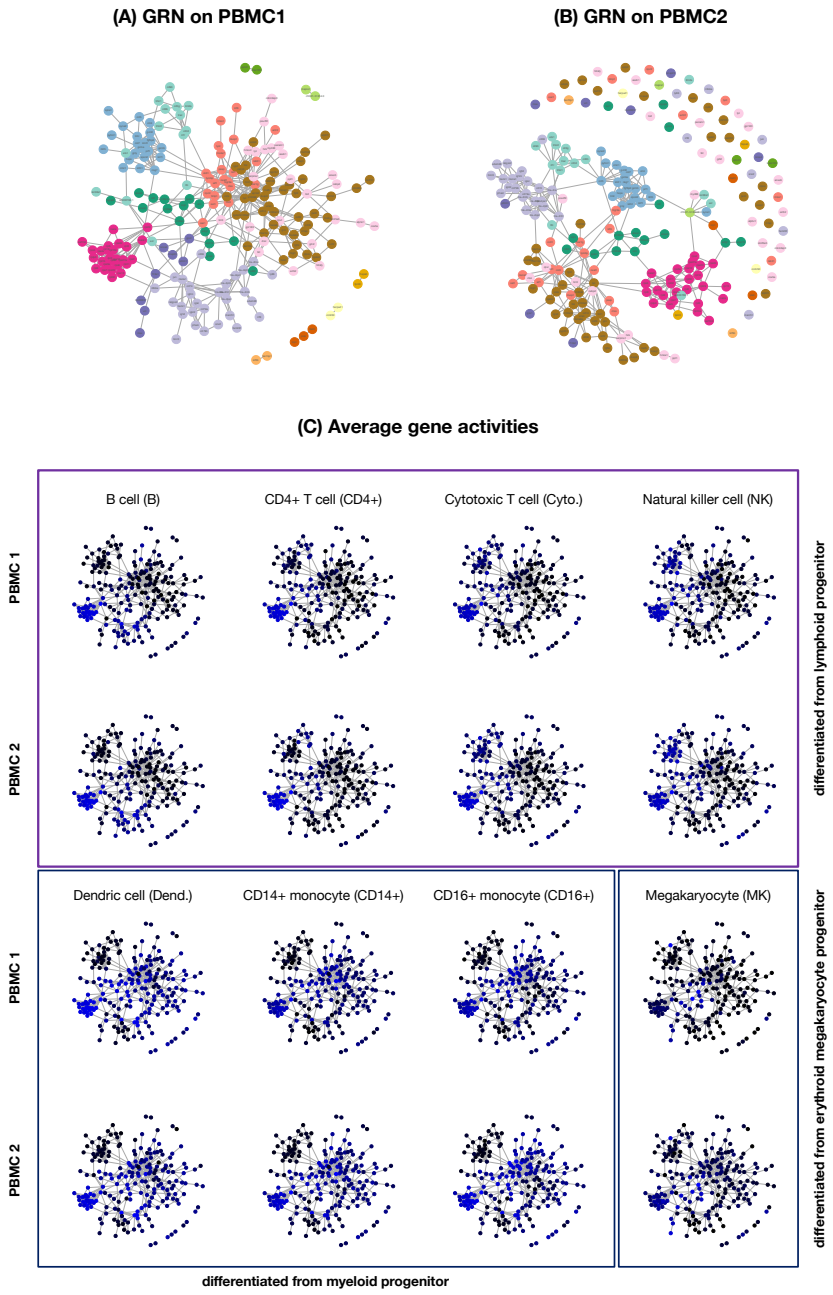

Figure 3: The learned GRNs and mean activation on GRNs
(A) and (B) GRNs learned on PBMC 1 and PBMC 2 respectively. Colors indicate network communities based on PBMC 1, with the louvain [@blondel2008fast] algorithm. (C) Average gene expressions for each class on the GRN learned from PBMC 1, visualized for PBMC 1 and PBMC 2.

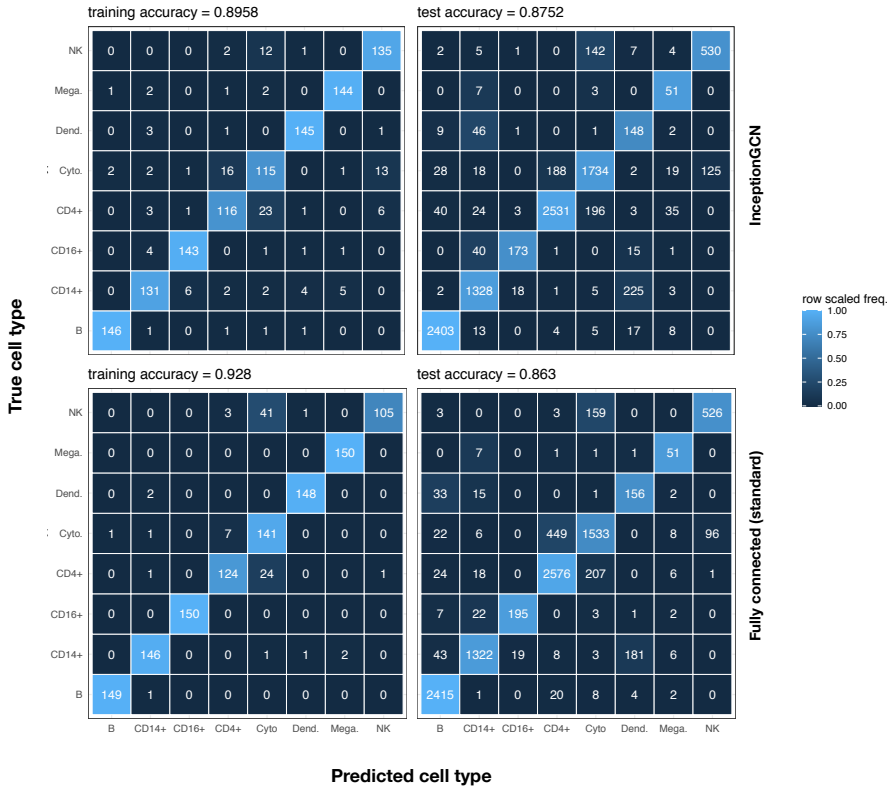

Figure 4: **The confusion matrices of the classification**
Each (table) cell indicates the number of respective (biological) cells.

4.1.3 Analysis of the results

We achieve 93.67% and 91.47% label prediction accuracy on the training and test data sets respectively. This is a significant result demonstrating the transferable learning capabilities of scGCN on a test data set which has been collected in a different experiment than that used for training the model. For comparison, we include classification results from a fully connected neural network with 2 layers (200 and 100 hidden nodes respectively) which yields an accuracy of 90% on training data, but drops to 86.7 on the second PBMC dataset as the test set.} As mentioned before, one of our objectives in the design of the single-cell geometrical deep learning framework has been interpretability. Our framework enables us to explore the learning procedure in our deep neural network. Our design facilitates representation learning,[4] in that one can visualize the data at the hidden layers and observe the trajectory of data passing through the complex transformations, done by the InceptionGCN. This gives an overview of the learning procedure, inspecting batch-effect removal, and heterogeneous cell populations being separated. One can also use the same information to study the trajectory of expression density along the GRN, for each class. This gives one an idea of how information diffuses on the network and utilizes the classification task. Fig. 5 represents an embedding trajectory analysis, visualized as UMAPs. As indicated, we observe gradual batch-effect removal and cluster separation along the layers.

[4] **Representation learning** is learning representations of the data by transforming or projections into repetitively simpler spaces (Bengio, Courville, and Vincent 2013).

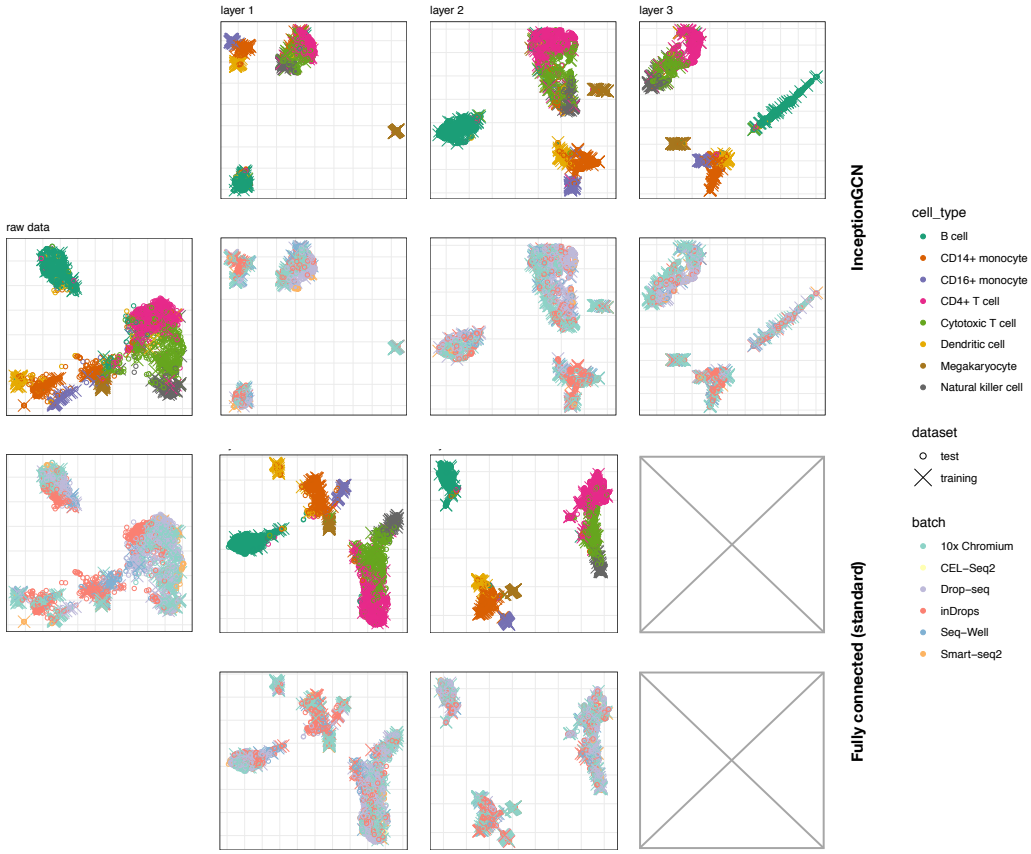

Figure 5: UMAPs of the raw and hidden embeddings
For each model, 3000 cells are sampled from the combined dataset of training and test datasets.

5 Discussion and future steps

We have demonstrated the potentials of learning on gene networks for transferable learning, i.e., learning models which are least affected by data set specific variations and hence are more general and can be used for predictions on new data as well.

On a test data of PBMC cells our scGCN outperforms a standard neural network which does not use the GRN information by about 86 percent (91 percent versus 86 percent). Regarding the GRN construction step, among all method that we examined so far, Gaussian Graphical Models (GGM) provide the most robust gene network which are compatible with the scGCN design as they are relatively sparse and constitute modular networks architecture. Sensibility of the used GRN is crucial for inferring correct cell type relationships, as it is through this network that we impose the inductive bias on our learning models. Thus, reliable knowledge about the GRN could boost the methods generalizability and transferable learning. Therefore, inclusion of available biological knowledge in our GRN stays of great interest for our future work and advancement of the framework. Upon availability of gene regulatory network architecture, we anticipate application of scGCN framework for transferable learning among different data modalities, e.g. learning on RNA-seq data and making reliable predictions on ATAC-seq data set of the same system.

Lastly, one of our motivations for creating the scGCN framework has been to obtain a more comprehensive understanding of cell type relationships and developmental trajectories resolution as shown in Fig. 1. In fact, cell-cell Euclidean distances found on the scGCN hidden layers (i.e. after application of graph convolutional filters) is such a desired distances as defined by kernel F in Fig. 1 as one can write:

$$(FX - FY)^2 = X^T F^2 X + Y^T F^2 Y - X^T F^2 Y - Y^T F^2 X$$

We can check improvement of cell lineage relationships resolution by comparing embedding (e.g. by UMAP or Diffusion map) of raw data with that of later (or the final) hidden layer of the scGCN, as in Fig. 5.

Code and data availability

- Code: GitHub repository
- PBMC dataset: link

References

Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. 2019. "A Comparison of Automatic Cell Identification Methods for Single-Cell Rna Sequencing Data." *Genome Biology* 20 (1). Springer: 194.

Baxter, Jonathan. 2000. "A Model of Inductive Bias Learning." *Journal of Artificial Intelligence Research* 12: 149–98.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8). IEEE: 1798–1828.

Bronstein, Michael M, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. "Geometric Deep Learning: Going Beyond Euclidean Data." *IEEE Signal Processing Magazine* 34 (4). IEEE: 18–42.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis." *Nature* 566 (7745). Nature Publishing Group: 496–502.

Ding, Jiarui, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, et al. 2019. "Systematic Comparative Analysis of Single Cell Rna-Sequencing Methods." *BioRxiv*. Cold Spring Harbor Laboratory, 632216.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* 9 (3). Oxford University Press: 432–41.

Kazi, Anees, Shayan Shekarforoush, S Arvind Krishna, Hendrik Burwinkel, Gerome Vivar, Benedict Wiestler, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab. 2019. "Graph Convolution Based Attention Model for Personalized Disease Prediction." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 122–30. Springer.

Levie, Ron, Elvin Isufi, and Gitta Kutyniok. 2019. "On the Transferability of Spectral Graph Filters." In *2019 13th International Conference on Sampling Theory and Applications (Sampta)*, 1–5. IEEE.

Liu, Zhi-Ping, Canglin Wu, Hongyu Miao, and Hulin Wu. 2015. "RegNetwork: An Integrated Database of Transcriptional and Post-Transcriptional Regulatory Networks in Human and Mouse." *Database* 2015. Narnia.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-Sne." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

McInnes, Leland, John Healy, and James Melville. 2018. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*.

Monti, Federico, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. 2017. "Geometric Deep Learning on Graphs and Manifolds Using Mixture Model Cnns." In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 5115–24.

Nestorowa, Sonia, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. 2016. "A Single-Cell Resolution Map of Mouse Hematopoietic Stem and Progenitor Cell Differentiation." *Blood, the Journal of the American Society of Hematology* 128 (8). American Society of Hematology Washington, DC: e20–e31.

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro. 2014. "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning." *arXiv Preprint arXiv:1412.6614*.

Paul, Franziska, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, et al. 2015. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." *Cell* 163 (7). Elsevier: 1663–77.

Zhou, Jie, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. "Graph Neural Networks: A Review of Methods and Applications." *arXiv Preprint arXiv:1812.08434*.