

RESEARCH

Open Access



Cellograph: a semi-supervised approach to analyzing multi-condition single-cell RNA-sequencing data using graph neural networks

Jamshaid A. Shahir^{1,2,3}, Natalie Stanley^{2,3,4} and Jeremy E. Purvis^{1,2,3,5*}

*Correspondence:
purvisj@email.unc.edu

¹ Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

² Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

³ Computational Medicine Program, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁴ Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁵ Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

With the growing number of single-cell datasets collected under more complex experimental conditions, there is an opportunity to leverage single-cell variability to reveal deeper insights into how cells respond to perturbations. Many existing approaches rely on discretizing the data into clusters for differential gene expression (DGE), effectively ironing out any information unveiled by the single-cell variability across cell-types. In addition, DGE often assumes a statistical distribution that, if erroneous, can lead to false positive differentially expressed genes. Here, we present Cellograph: a semi-supervised framework that uses graph neural networks to quantify the effects of perturbations at single-cell granularity. Cellograph not only measures how prototypical cells are of each condition but also learns a latent space that is amenable to interpretable data visualization and clustering. The learned gene weight matrix from training reveals pertinent genes driving the differences between conditions. We demonstrate the utility of our approach on publicly-available datasets including cancer drug therapy, stem cell reprogramming, and organoid differentiation. Cellograph outperforms existing methods for quantifying the effects of experimental perturbations and offers a novel framework to analyze single-cell data using deep learning.

Keywords: Graph neural networks, Single-cell genomics, Semi-supervised learning

Background

The rapid progression of single-cell technologies [1] has enabled scientists to accumulate complex datasets to study differentiation and developmental trajectories in response to differing experimental perturbations, assess the efficacy of a drug in treating of disease, and evaluate the efficiency of different reprogramming protocols. Regardless of the preceding experimental design, many single-cell RNA sequencing (scRNA-seq) analyses follow the same pipeline [2]: pre-processing and quality control followed by clustering and differential gene expression. In the context of studying more continuous phenomena such as differentiation or cell reprogramming, trajectory analysis may also be employed



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

[3]. However, in the case of multiple experimental conditions such as different time points sampled for sequencing in cell reprogramming, or varying concentrations of a cancer drug, these methods may fall short in faithfully summarizing the underlying biology. In particular, clustering and differential gene expression give a bulk summary of the transcriptomic variation between computationally-inferred discrete populations, but do not explicitly consider the single-cell variability within treatment groups, such as how prototypical an individual cell is of its assigned treatment group.

Related methods

Differential abundance methods can rectify these challenges by quantifying differences between and within conditions at a finer resolution. Milo tests for differential abundance on k -nearest neighbor (k NN) graphs by aggregating cells into overlapping neighborhoods and performing a quasi-likelihood F test [4]. This returns a metric of the log-fold change of the differential abundance in each neighborhood. However, because Milo aggregates cells into neighborhoods, it does not provide single-cell resolution providing insight into the impact each perturbation has on an individual cell.

Covarying Neighbor Analysis (CNA) [5] performs association analysis agnostic of parameter tuning, making it an efficient method. Like Milo, it aggregates cells into neighborhoods, and calculates a neighborhood abundance matrix (NAM), where each entry $C_{n,m}$ is the relative abundance of cells from sample n in neighborhood m . From there, it derives principal components where positive loadings correspond to higher abundance while negative loadings correspond to lower abundance. This enables the characterization of transcriptional changes corresponding to maximal variation in neighborhood abundance across samples. Association testing is performed between transcriptional changes and attributes of interest using the first k NAM-PCs. It returns the Spearman correlation between the attribute and abundance of the neighborhood anchored at each cell, providing a single-cell metric. However, its performance falls short when considering more than two conditions.

MELD [6] sought out to quantify the effect of an experimental perturbation on individual cells in scRNA-seq data using graph signal-processing to infer a sample-associated density that is then normalized to give a probability of each cell belonging to a condition of interest defined as a relative likelihood. It uses all the class labels to derive these probabilities. The authors also introduced a novel clustering approach, called Vertex Frequency Clustering (VFC), which clusters data according to not just transcriptomic similarity but also how the MELD-derived relative likelihood scores, thereby identifying populations of cells similarly enriched or depleted in conditions according to the perturbation response. However, the original study restricted evaluation to datasets with two conditions to discriminate between: a control condition and a single perturbed condition, and therefore did not consider multiple treatment conditions, which are more prevalent and can provide more insight, for instance, the response of a drug at various time intervals, combining drugs, or administration of a differentiation stimulus at different time-points. Furthermore, robust calculation of the sample-associated likelihood relies on computationally-expensive parameter estimation that can take upwards of 12 h with 36 cores on a high-performance computing cluster for a dataset of 26,827 cells. In addition, VFC is memory-intensive, which limits its scalability to larger datasets.

Graph neural networks

In recent years, the rapidly-emerging field of deep learning has seen utility in scRNA-seq analysis [7–9]. More recently, graph neural networks (GNN) have demonstrated promise in capturing the structural information of scRNA-seq data via the graphical representation the high-dimensional assay naturally lends itself towards, with cells as vertices or nodes, and edges between them representing similarity in gene expression. This connectivity enables the model to naturally leverage the relationship between similar cells in a variety of tasks, most notably clustering and imputation. GNNs pass in graphical representations of data as input to perform a myriad of classification tasks, namely node classification, edge classification, and graph classification. Unlike convolutional neural networks (CNNs) which involve multiple layers and can take a long time to train depending on the size of the data, GNNs require only a few layers to achieve high performance in a fraction of the time. Furthermore, whereas CNNs require large amounts of training data, GNNs can learn patterns in data in a semi-supervised fashion: they take the entire data structure as input, but only a paucity of nodes are labeled; a larger portion is held out for validation and testing purposes. The applications of GNNs has been demonstrated in the case of graph classification, edge classification, and node classification. For example, scGNN [9] used GNNs and a Gaussian mixture model to perform imputation and cell clustering. Another study used Graph Attention Networks (GATs) [10, 11]—a subset of GNNs based on the self-attention mechanism commonly used in natural language processing—to predict disease state in scRNA-seq data from multiple sclerosis patients, followed by another study from the same group applied to COVID-19 patients [12]. GATs have also been used as part of variational graph autoencoders to facilitate clustering [13].

Moreover, GNNs have been used in conjunction with relational networks to predict breast cancer subtypes in bulk RNA-seq data [14]. However, their potential to ascertain the responsiveness of individual cells to perturbations in order to gauge the efficacy of the experimental stimulus, particularly in complex experimental designs that span multiple conditions or time points, has not been formally assessed.

Finally, a notable study introduced a GNN framework called single-cell Graph Convolutional Network (scGCN) which uses Graph Convolution Networks (GCNs) [15]—which are analogous to CNNs in that they both use convolution operators, but operate on different types of data structures—to transfer labels across diverse datasets and subsequently integrate the datasets, outperforming popular methods like Seurat v3 and Conos on these tasks [16]. However, the framework could not perform perturbation analysis, as its task was to predict cell type annotations of query data from the given reference data, illustrating room for the expansion of the novel applications of GNNs in single cell genomics, to which our work below seeks to contribute.

In this work, we introduce Cellograph: a novel computational framework using GCNs to perform node classification on scRNA-seq data collected from multiple conditions, treating the individual cells as nodes. Cellograph uses a two-layer GCN to learn a latent representation of the single-cell data according to how representative each cell is of its ground truth sample label. This latent space can be easily clustered to derive groups of cells associated with similar treatment response and transcriptomics, as well as projected into two dimensions for visualization purposes. Cellograph outperforms existing

approaches in quantifying the effects of perturbations and offers a novel GNN framework to cluster and visualize single-cell data. In addition, Cellograph is more scalable, performing at least an order of magnitude faster than MELD. In the following sections, we discuss the workflow of Cellograph, demonstrate its performance on three published scRNA-seq datasets, and benchmark it against previously published methods using cross-categorical entropy and normalized mutual information [17].

Methods

Overview of the Cellograph algorithm

Cellograph uses GCNs to perform node classification on cells from multiple samples to quantify how representative cells are of each sample. We found GCNs to be most apropos for our implementation as they explicitly draw upon neighborhood information to capture transcriptomic relationships between cells by considering the connections between neighboring cells (e.g., molecularly similar cells) in the graph. Furthermore, as scRNA-seq is prone to technical artifacts, such as dropouts or noise in gene expression measurements, GCNs can mitigate the impact of this noise by leveraging the collective information from neighboring cells in the graph as it maps the initial dataset to a latent embedding in the first layer. By propagating information through the graph structure, GCNs can capture more reliable and robust representations of cells, improving common downstream analysis tasks like clustering, dimensionality reduction, and classification, as we shall demonstrate in the results section. Finally, the GCNs offer interpretability by learning feature importance within the context of the graph structure. By examining the learned weights in the GCN layers, we can identify features (or genes, in this context) that contribute significantly to the model's predictions. This facilitates identification of biologically meaningful genes that drive cellular tendencies towards one experimental group versus another. This information can complement and corroborate findings from differential gene expression, but with an emphasis on group truth labels versus independently-inferred clusters.

In summary, Cellograph takes in a single-cell dataset $X \in \mathbb{R}^{n \times m}$ (where n denotes the number of cells or nodes, while m represents the number of genes or features) aggregated from multiple treatment conditions. We assume X has already been pre-processed and filtered according to typical pre-processing steps when working with scRNA-seq data. (see Fig. 1). X is then reduced to a PCA space, where a k -nearest neighbor graph is constructed using a select number of principal components (PCs), with a resulting adjacency matrix $A \in \mathbb{R}^{n \times n}$. The graph is then passed in as input into a two-layer graph neural network that uses a parameterized matrix weighed by the genes to encode each cell's transcriptome to lower dimensions that take into account the connectivity between cells. Specifically, we train a two-layer GCN on this derived graph.

In the first layer (Fig. 1D), we perform the following mathematical operation:

$$H^{(0)} = \sigma(\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} X W^{(0)}).$$

Here, \tilde{A} is calculated as $\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}}$, and \hat{A} is obtained by adding the identity matrix I to the adjacency matrix A . This adds self-loops to the adjacency matrix such that each cell is incorporating its own features in addition to its neighbors. $W^{(0)} \in \mathbb{R}^{m \times h}$ is a parameterized weight matrix that's updated throughout the training of the model. Each row

of the matrix corresponds to a gene, with a set of h weights per gene. Upon successful training, these weights can be summed up per row and ordered from highest to lowest, where genes with the highest weights denote biologically meaningful genes that distinguish the ground truth conditions. In other words, genes that more effectively distinguish conditions are given higher weights during the training. $H^{(0)} \in \mathbb{R}^{n \times h}$ is the output of the first layer in h latent dimensions. This matrix can be further reduced to 2 dimensions for visualization using a dimensionality reduction method like PHATE or UMAP [18]. This additional pre-processing step prior to visualization creates an embedding where cells are arranged not just according to transcriptomic similarity, but also how representative they are of each experimental condition. In the second and final layer, we have a very similar operation

$$H^{(1)} = \text{softmax}(\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} H^{(0)} W^{(1)}),$$

where now $W^{(1)} \in \mathbb{R}^{h \times c}$ with c as the number of conditions. Here, we take our latent embedding $H^{(0)}$ from the initial layer and apply the same operation, only this time we map it to a matrix of treatment probabilities for each cell, giving a single-cell metric of how responsive the cell is to each treatment. The output $H^1 \in \mathbb{R}^{n \times c}$ is a matrix of treatment probabilities. The softmax function is a nonlinear function that converts its inputs to a probability distribution proportional to the exponentials of the inputs as follows:

$$\text{softmax}(\hat{z}) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}.$$

Regarding the training process, as noted, GNNs learn in a semi-supervised manner. This means that during training, the entire graph is observed, but only a fraction of the nodes have labeled information. Specifically, we randomly select 1–3% of nodes from each condition as training nodes, while a larger fraction are held out for testing and validation. This random selection of nodes facilitates objective training. The quality of the training is assessed via a categorical cross-entropy loss function. By default, we train the GNN for 200 epochs and terminate training if there is no improvement after 30 epochs (patience). This is in contrast to MELD, which uses all the labels and is not holding out anything, leveraging the full cell-type information via these ground-truth labels to perform the calculations, instead.

Pre-processing the scRNA-seq data

We pre-process data as commonly done using Scanpy, unless specified otherwise [19, 20]. For the organoid dataset, we downloaded the publicly available, normalized dataset from <https://singlecell.broadinstitute.org> (study SCP1318) and filtered the most highly variable genes (using the default parameters in Scanpy: a minimum mean expression of 0.0125, maximum mean expression of 3, and minimum normalized dispersion of 0.5). Metadata was also included with cell type annotation and ground truth treatment groups. For the drug holiday dataset, we followed the pre-processing steps described in the original study, only implementing them in Scanpy over Seurat. For the myogenesis dataset, we followed the quality control steps described in the original paper, except implemented in Scanpy rather than Seurat (all cells with less than 300 genes expressed

were removed, as well as all genes expressed in less than 10 cells; furthermore, only cells with less than 20% percentage mitochondria expression were retained). The data was then normalized using Scanpy to 10,000 reads per cell, logarithmized, and filtered down to the top 2000 highly variable genes.

Results

We demonstrated the biological application of Cellograph on three published scRNA-seq datasets: a human organoid model of intestinal stem cells differentiating to Paneth cells with or without a stimulus to enhance the efficiency of the differentiation [21]; a non-small-cell lung carcinoma (NSCLC) cell line that was treated with a drug called Erlotinib at various time points and later temporarily withdrawn from the drug for several days [22]; and a myogenesis model of transdifferentiation and traditional cell reprogramming [23]. We benchmarked the performance of Cellograph against the aforementioned differential abundance methods, MELD, Milo, and CNA. Our results show robust performance of Cellograph on these distinct datasets, and provide valuable biological insights.

Cellograph captures shifts in cell type abundance during human intestinal organoid differentiation

We first applied Cellograph to an organoid model of intestinal stem cells differentiating to Paneth cells with or without KPT-330, an inhibitor of the nuclear exporter, Exportin 1, which was demonstrated in the original study [21] to enhance the abundance of Paneth cells following differentiation. Samples were collected from 6 donors for sequencing following 6 days of treatment with or without KPT-330. Cell type annotation revealed 9 prominent cell types: Stem cells transitioning from G1 to S phase of the cell cycle (G1/S), stem cells in G2 and M phase of the cell cycle, proliferative progenitor cells (Progenitor), enterocytes (Enterocyte), wound-associated epithelium cells (WAE), WAEs enriched in the well-characterized stress-associated gene *DUOX2* (*DUOX2+* WAE-like), quiescent progenitors (Quiescent progenitor), goblet cells (Goblet), and enteroendocrine cells (Enteroendocrine). We will refer to these two conditions as KPT and control cells, respectively. We trained Cellograph using a two-layer GCN with 80 out of the 2484 cells labeled, such that 40 were labeled for each condition. We projected the learned latent space to 2 dimensions with PHATE and colored cells according to the probability of belonging to the KPT-treated condition (Fig. 2A). We obtain a smooth gradient of cells along the PHATE plot, with cells arranged according to how impacted they are by KPT treatment. UMAP also captured the separation between conditions and gradient of probability scores [18] compared to traditional UMAPs on the high-dimensional PCA space (Additional file 1). To determine if this gradient reflected meaningful biology, we extracted the 25 top-weighted genes from the aforementioned learned weight matrix (discussed in “[Overview of the Cellograph algorithm](#)” section) and visualized them with a heatmap categorized by the two treatment groups (Fig. 2B), which corroborates existing findings for the source paper. This matrix is derived from the first layer of the GCN and parameterizes each gene, where the model upweights genes it finds most relevant in distinguishing between conditions. Among the top 25 genes is *GDF15*, a marker of *DUOX2+* WAE-like and WAE-like cells, which is highly expressed in KPT-treated cells,

where these cell types are more abundant due to the greater efficiency of Paneth cell differentiation [24, 25]. Conversely, KLK6 is highly expressed in the control-treated population, which has been shown to mediate the multipotency of intestinal stem cells [26, 27].

We also performed k -means clustering on the latent space learned by Cellograph with $k = 3$ (Fig. 2A,C). Unlike clustering the original PCA-reduced data, which just focuses on differences in the transcriptome, Cellograph implicitly clusters according to how responsive cells are to the KPT-330 stimulus. This successfully groups together cells predicted to belong to the KPT-treated group (called the responsive cluster), a mixed population of cells predicted to be either control or KPT-treated cells (intermediate cluster), and a cluster of cells predicted to be prototypical of the control population (naive cluster). These predictions were determined using a threshold of 0.5 for ground truth assignment.

Based on the softmax probabilities learned by Cellograph ($z_{i,j} \geq 0.5$), we assigned cells to the control or KPT-treated populations independent of their ground truth labels, and created composition plots according to cluster assignment (Fig. 2D). We see that Cellograph's predictions corroborate the compositional changes in cell types abundance discussed in the original study, namely with decreases in dividing stem cell and progenitor populations, increases in quiescent progenitors, enterocytes, and DUOX2+ WAE-like cells.

Finally, we mapped the cell type annotations onto the clusters obtained by Cellograph (Fig. 2D) and observe a high abundance of cycling cells, progenitor cells, and WAE-like cells in the Naive cluster, followed by a decrease of WAE-like cells and progenitor cells in the intermediate population, and a high proportion of DUOX2+ WAE-like cells in the responsive cluster. Altogether, these results demonstrate Cellograph's ability to identify and visualize cells affected by KPT-330 stimulation. It corroborates existing findings and presents an interpretable framework for downstream tasks like visualizing and clustering the data.

Cellograph models heterogeneity in cancer drug response during a drug holiday

Encouraged by Cellograph's performance on the human intestinal organoid dataset, we next investigated how well it could capture heterogeneity in response to cancer drugs under complex treatment regimes. We trained Cellograph on the single-cell transcriptomes of 3042 PC9 cells treated with Erlotinib [22]—a tyrosine kinase inhibitor used to treat non-small cell lung cancer (NSCLC)—for 11 days, followed by withdrawal of the drug for 6 days, referred to as a drug holiday, where select cells were either retreated with Erlotinib or treated with DMSO as a control. This study examined the drug-tolerant states in a non-small-cell lung carcinoma (NSCLC) cell line, where the goal was to understand what cell populations would emerge from treatment and retreatment. Specifically, the authors treated the cell line with a drug called Erlotinib for 11 days, followed by a 6-day withdrawal period called a drug holiday as the cells developed resistance. A subset of cells was then reintroduced to Erlotinib for 2 days and cells were sequenced at each time point. The key takeaway from this paper was that they identified subpopulations of cells associated with genes that induced drug resistance, and those inhibiting drug resistance. However, this just considered transcriptomic variation and simple graph-based Leiden clustering, so we were interested if Cellograph could quantify the

effect of these temporal perturbations at single-cell resolution, corroborate these findings, and perhaps offer novel insights into these mechanisms of drug resistance. The cells were sequenced at 5 timepoints: 0 days with no Erlotinib treatment, 2 days of Erlotinib treatment, 11 days of Erlotinib treatment, at day 19 with or without re-exposure to Erlotinib on day 17, following 6 days of removal from the drug. We trained Cellograph on these PC9 cells with 30 cells labeled for each condition using a 2-layer GCN. We project the learned latent space into 2 dimensions with PHATE, which gives a clear temporal separation of the 6 treatment groups (Fig. 3A), comparable to UMAP (Additional file 1). Coloring cells according to the probability of belonging to each of the conditions provides a narrow distribution of scores in cells in the condition of interest, with the notable exception of Day 11 (Erlotinib before holiday) and Day 19 (Erlotinib after holiday), suggesting a non-uniform response to the drug in these cells both before and after the drug holiday (Fig. 3D). The heatmap of the top 25 weighted genes from training (Fig. 3B) implicates such genes as TUBA1B and CCDC80 in distinguishing the conditions, which are both markers of drug resistance, with CCDC80 highly expressed in D11 cells, corroborating the original study's observations of CCDC80, whereas TUBA1B expression is particularly elevated in D19 Erl cells. Almost all of these genes were previously identified through differential gene expression in the original paper, showcasing Cellograph's interpretability of the weight matrix in identifying pertinent genes defining molecular differences. However, MT-ND6, which was not among the differentially expressed genes to the best of our knowledge, is also strongly weighted and appears to uniformly define the population of cells that were treated with DMSO following the drug holiday. This is a mitochondrial gene which has been previously implicated in colorectal adenocarcinoma and associated with changing energy requirements due to cells aggressively proliferating [28]. Clustering the learned latent space identifies three clusters among these two conditions (Fig. 3A,D), one consisting of cells predicted to have a prototypical response after 11 days of Erlotinib treatment (cluster 3), and similarly for day 19 after re-exposure to the drug (cluster 5), followed by a mixed population of both cell types (cluster 2). Differential expression between the three clusters (Fig. 3C) identified high expression of TUBA1B in cluster 5, which is associated with poor prognosis in NSCLC, suggesting persisting drug tolerance after the holiday period. Similarly, we observe differential expression of INHBA in cluster 3, a senescence mediator that's associated with prognosis in many cancer types [29]. This suggests that there is drug resistance in both treatment regimes, yet seemingly stimulating different pathways of resistance as opposed to anti-resistance, highlighting the limitations of the treatment scheme. Interestingly, TUBA1B and INHBA expression are significantly reduced in the day 19 population that was not retreated with Erlotinib. Altogether, Cellograph captures clinically relevant genes driving heterogeneity in response to treatment, corroborates existing findings of pertinent genes driving treatment response, identifies an additional gene that was previously not described to the best of our knowledge, and suggests different modes of drug resistance.

Cellograph distinguishes between transdifferentiation and dedifferentiation in myogenesis

Finally, we assessed Cellograph's ability to distinguish cells undergoing distinct cell state transitions temporally on a scRNA-seq dataset of 33,380 mouse embryonic fibroblasts (MEF) undergoing either dedifferentiation to adult stem cells called induced myogenic

progenitor cells (iMPCs) or myogenic transdifferentiation to myotubes [23]. The original study was motivated to understand the transcriptional and epigenetic mechanisms of how over-expression of the MyoD transcription factor induced MEFs to undergo reprogramming to either myotubes or iMPCs with a MyoD-inducible transgenic model. The myotubes were induced by overexpression of MyoD, while the addition of small molecules produced Pax7⁺ iMPCs that were very similar to primary muscle stem cells. The authors used trajectory analysis via diffusion maps and UMAP embeddings of combined single-cell data of MEFs expressing MyoD or MyoD + a cocktail of small molecule inhibitors (forskolin, RepSox, and CHIR99021, collectively abbreviated as “FRC” in the original paper) to reveal that dedifferentiation and transdifferentiation follow two different trajectories.

We trained Cellograph on these differentiating cells with 200 cells per treatment group labeled for training for 400 epochs and obtained a single trajectory that starts with transdifferentiation and culminates in dedifferentiation to Pax7⁺ iMPCs (Fig. 4A; Additional file 1). Looking at the top-weighted genes from training the model (Fig. 4B), high expression of CRABP1 and LUM distinguished the transdifferentiating population, whereas dedifferentiation was weighted by high expression of cyclin D1, suggesting cell cycle entry is a necessary step to producing iMPCs. CRABP1 is known to promote stem cell proliferation by its downregulation [30]. However, it does not appear to inversely vary with cyclin D1. The original study revealed an overlap between the major fraction of day 4 MyoD-treated cells and day4/8 MyoD+FRC-treated cells in their UMAP and DPT embeddings. Interestingly, however, Cellograph detects no significant overlap (Fig. 4A), which is further supported by the derived probabilities of belonging to each of the experimental groups (Fig. 4C).

STMN2, an early neuronal marker, was also identified as a pertinent gene in distinguishing between these processes (Fig. 4B), with high expression in the transdifferentiation condition, perhaps owing to the instability and inefficiency of generating myotubes with MyoD alone. Clustering the latent space and mapping the clusters onto the PHATE embedding distinguished the different treatment conditions and heterogeneity in the MyoD+FRC day 8 condition. Notably, we observed differential expression of MYOG (Fig. 4D), which specifies the myotube fate, in the majority of cells, which corroborates observations from trajectory analysis in the original study where this gene is observed in both trajectories. Cell cycle differences underscored variability in the Pax7⁺ iMPCs (Fig. 4D,E). Altogether, Cellograph is able to successfully distinguish these biological processes, and identify additional gene programs explaining these differences.

Cellograph outperforms published differential abundance methods and popular single-cell clustering methods

Finally, we benchmarked Cellograph’s performance in identifying cells most impacted by perturbations against three MELD, Milo, and CNA. We used the Brier score for comparison between Cellograph and MELD as we believed a method quantifying experimental perturbations should capture a broad range of signals for each experimental label it is trying to predict. In particular, this metric quantifies the squared difference between predicted and true probabilities distributions by calculating the following sum,

$$\text{BrierScore}(y, p) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C (p_{ij} - \delta_{ij})^2$$

where y represents the true labels of the samples, with $y_i \in \{1, 2, \dots, C\}$ denoting the true class label of sample i , p represents the predicted probabilities of the samples, with p_{ij} denoting the predicted probability of sample i belonging to class j , N represents the total number of samples, and δ_{ij} is the Kronecker delta function defined as $\delta_{ij} = 1$ if $y_i = j$ and $\delta_{ij} = 0$ otherwise. Lower values reflect better quality performance. When applied to all the cells in our datasets, we obtain consistently lower scores than MELD (Table 1), despite MELD using all class labels during its learning process whereas Cellograph uses only a fraction.

When evaluating Cellograph relative to Milo and CNA, however, we could not perform direct quantitative comparison. As discussed in “[Related methods](#)” section, Milo gauges the presence of differential abundance on kNN graphs by aggregating cells into overlapping neighborhoods and performing a quasi-likelihood F test. This returns a metric of the log-fold change of the differential abundance in each neighbor, not a single-cell measurement giving the probability of that cell belonging to one treatment class versus another. Thus, we cannot perform a direct quantitative comparison and instead present a qualitative assessment of performance. Running Milo on the human organoid dataset, we observe a positive correlation between the Milo-derived log-fold changes in differential abundance and the probability of cells belonging to the KPT-treated group (Fig. 5A). However, when applied to the drug holiday and myogenesis datasets (Fig. 5B,C), which have more complex experimental designs with multiple conditions, Milo fails to yield clear, interpretable results, with low DA in the untreated population, high DA in the cells after one day of Erlotinib treatment, and minimal DA in all other conditions. Similarly, in the myogenesis dataset, we observe high DA in Pax7-treated cells, low DA in MEFs, and minimal DA everywhere else.

Applying CNA to the human organoid dataset with the KPT treatment status as the attribute of interest, we obtain similar results as our method, MELD, and Milo. Specifically, we observe high correlation in the KPT-treated cells, and low correlation in the untreated cells. This elevated correlation is on par with the high probability of observing cells in the KPT-treated group. However, on the Erlotinib and myogenesis datasets, like Milo, we obtain results incongruous with Cellograph or MELD’s performance. It is even at odds with Milo. High abundance is predicted for cells treated right before holiday and following the holiday, regardless of whether cells were retreated with Erlotinib, whereas low abundance is observed in both the untreated cells and cells treated with Erlotinib for one day, while cells with 11 days of treatment have zero correlation. Since this dataset

Table 1 Brier score of Cellograph versus MELD on all cells

Dataset	Cellograph	MELD (optimal settings)
Organoid	0.153	0.362
Drug holiday	0.133	0.222
Myogenesis	0.094	0.324

Table 2 Average NMI score from different clustering methods on the datasets (standard deviation of scores in parentheses) along with optimal resolution parameter or choice of k used in each method

Dataset	K-means on PCA space	Leiden	Louvain	K-means on Cellograph-derived latent space
Organoid	$k = 8$ 0.124 (0.019)	Resolution = 0.2 0.183 (0.071)	Resolution = 0.4 0.183 (0.021)	$k = 2$ 0.513 (6.695×10^{-16})
Drug holiday	$k = 4$ 0.706 (9.255×10^{-4})	Resolution = 0.2 0.817 (0.018)	Resolution = 0.2 0.824 (3.956×10^{-3})	$k = 5$ 0.805 (1.292×10^{-4})
Myogenesis	$k = 10$ 0.538 (0.013)	Resolution = 0.1 0.715 (0.016)	Resolution = 0.5 0.655 (0.018)	$k = 7$ 0.87 (5.174×10^{-5})

Cellograph uses GNNs to quantify experimental perturbations

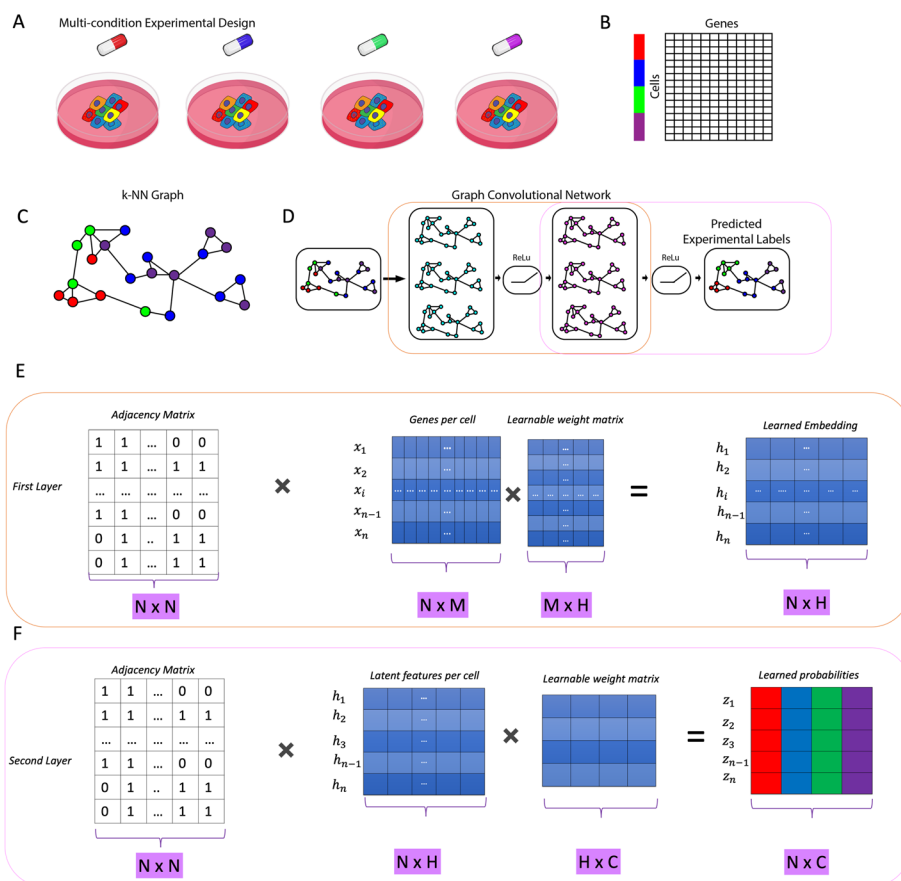


Fig. 1 Illustrative overview of Cellograph algorithm. Single-cell data collected from multiple sample drug treatments (**A**, **B**) is converted to a kNN graph (**C**), where cells are nodes, and edges denote connections between transcriptionally similar cells. The colored rectangles (**B**) correspond to the different samples represented by the drugs in **A**. This kNN is fed as input to a two-layer GCN (**D**) that quantitatively and visually learns how prototypical each cell is of its experimental label through the learned latent embedding. **E** A mathematical schematic of the first layer, where each cell's gene expression and its neighbors's gene expression is aggregated to produce a lower-dimensional representation of the cell in a latent space. **F** A mathematical schematic of the second layer respectively, where the output embedding of the first layer is mapped to softmax probabilities of cells belonging to each of the drug treatments

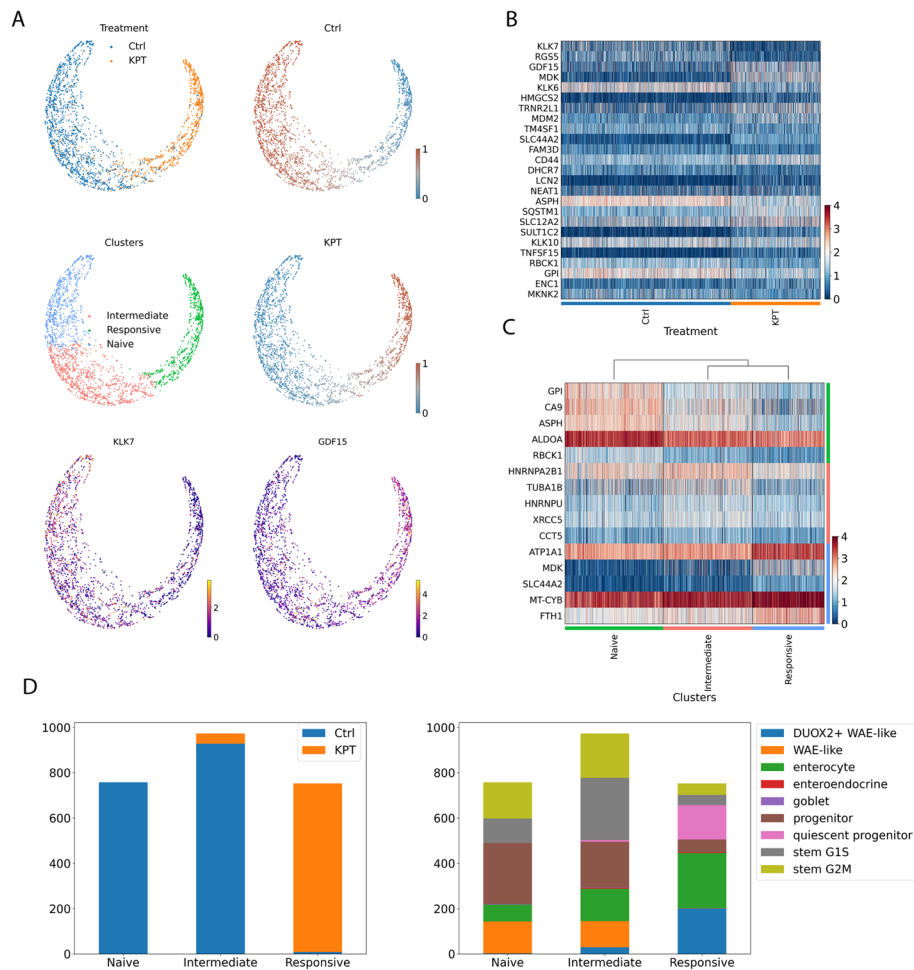


Fig. 2 Cellograph identifies treatment groups and distinguishes genes defining these groups on a human organoid dataset. **A** PHATE projection of learned latent space, with cells colored by treatment labels, probabilities of belonging to control or KPT-treated cells, clusters obtained by k-means clustering of the learned latent embedding with $k = 3$, and gene expression of GDF15 and KLK7. **B** Heatmap of top 25 weighted genes from parameterized gene weight matrix. **C** Heatmap of differentially expressed genes between clusters derived from Cellograph. **D** Compositional plot of predicted treatment groups from the softmax probabilities ($z_{ij} > 0.5$) (left) and cell types annotated by the original study (right) partitioned by clusters

spans multiple conditions and CNA just calculates one set of metrics, it was difficult to interpret these results in the context of the experiment. We obtained similarly incongruous results for the myogenesis dataset (Fig. 5). Altogether, Cellograph provides robust and interpretable results for more complex experimental designs with multiple treatment groups compared to CNA and Milo, and performs consistently better than MELD with a significantly lower runtime for optimal performance (Fig. 6).

When evaluating clustering performance with NMI, k -means clustering on the learned latent space yielded consistently high metrics compared to the Leiden and Louvain clustering algorithms, and k -means clustering on data in PCA space. 100 NMI values were calculated for each dataset by performing independent runs of the clustering algorithms. (Fig. 7). The treatment annotations given in the source papers

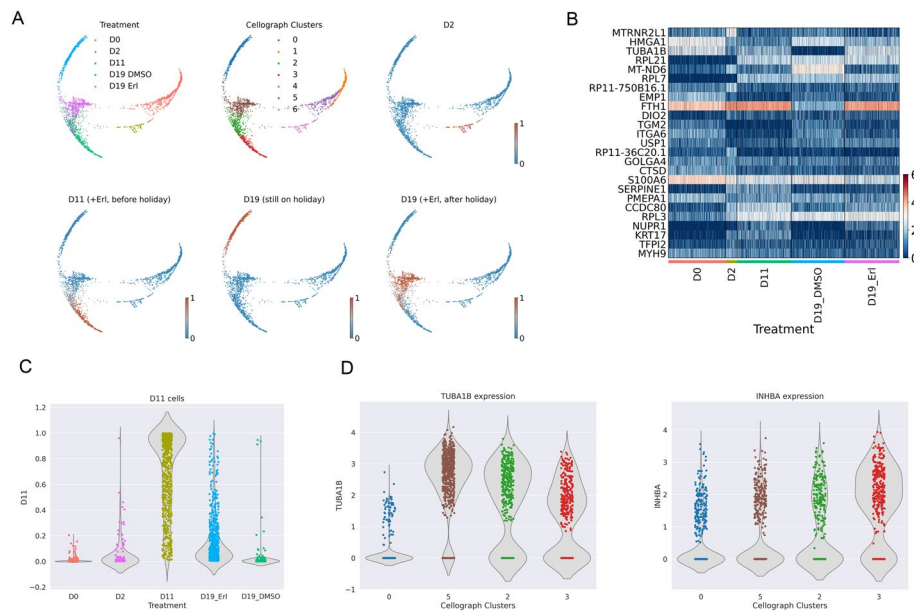


Fig. 3 Cellgraph defines genetic signatures of distinct drug holiday responses in the drug holiday dataset. **A** PHATE embeddings of the learned latent space colored according to the treatment labels, clusters, and treatment probabilities (day 0 not shown). **B** Heatmap of top 25 weighted genes from learned parameterized gene weight matrix. **C** The distribution of treatment probabilities for Day 11 cells partitioned by treatment groups. **D** The distribution of gene expression between clusters 0, 5, 3, and 2 of select differentially expressed genes (INHBA, TUBA1B)

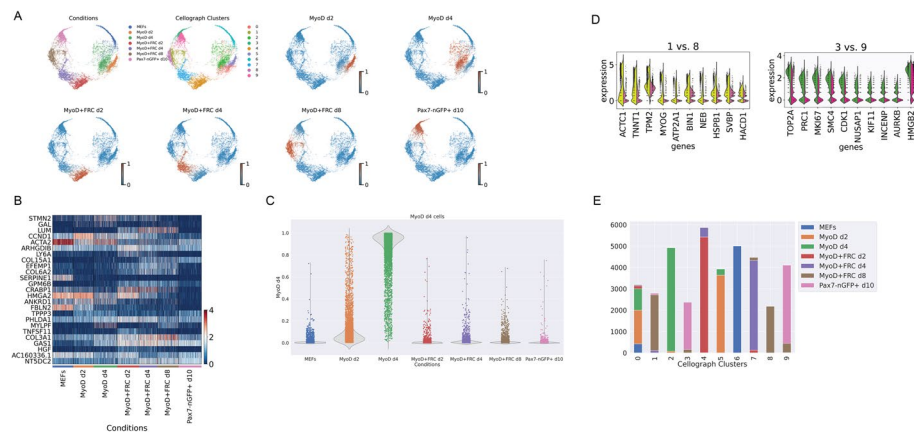


Fig. 4 Cellgraph distinguishes the molecular mechanisms of transdifferentiation and dedifferentiation in myogenesis. **A** PHATE embeddings of learned latent space annotated according to treatment conditions, clusters, and softmax probabilities of all conditions except for MEFs, defining the in-group variation. **B** Heatmap of top weighted genes from parameterized gene weight matrix, identifying pertinent genes such as cyclin D1 and CRABP1. **C** Violin plot of softmax probabilities of cells belonging to the MyoD/day 4 treatment group, showing similarities to the MyoD/day 2 population. **D** Violin plots of top 20 differentially expressed genes between clusters 1 and 8 and clusters 3 and 9, which define the Pax7⁺ cells and MyoD+FRC/day 8 treated cells, respectively. **E** Compositional plot of predicted cell types partitioned by cluster

were used as ground truth to derive the NMI values. Resolution parameters for the most optimal number of clusters in the Leiden and Louvain were chosen such based on the scib software [31] for more rigorous comparison (Table 2).

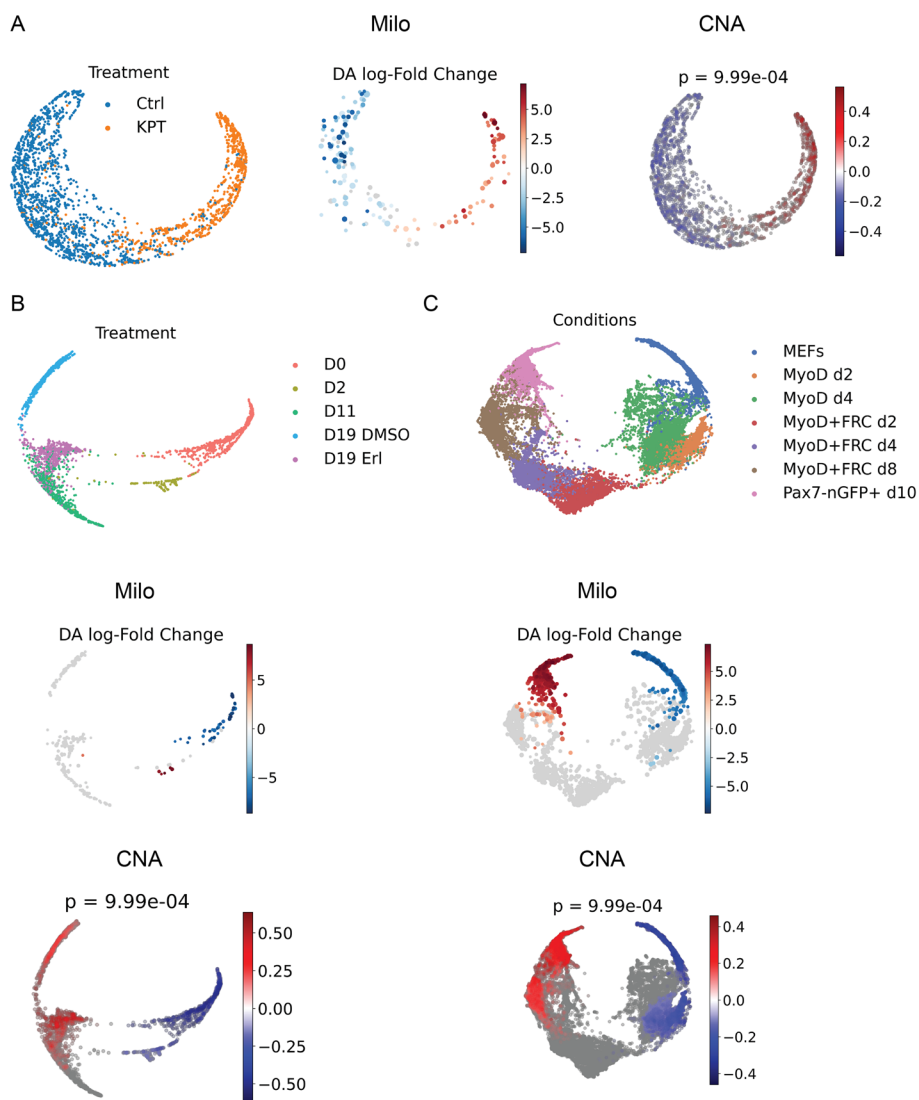


Fig. 5 Results of running Milo and CNA on the datasets evaluated. **A** Output of running Milo and CNA on the human organoid dataset. **B** Output of running Milo and CNA on the drug holiday dataset. **C** Output of running Milo and CNA on the myogenesis dataset

Resolution parameters were chosen such that the Leiden and Louvain algorithms generated the same number of clusters as k for k -means clustering for a more rigorous comparison (for the organoid dataset, resolution parameters of 0.3 and 0.2 were chosen for Louvain and Leiden clustering, respectively; for the drug holiday dataset, resolution parameters of 0.6 and 0.45 were chosen for Louvain and Leiden clustering, respectively; and for the myogenesis dataset, resolution parameters of 0.45 and 0.34 were chosen for Louvain and Leiden clustering, respectively).

However, we stress that this improvement in clustering is not a novel contribution of Cellograph. Ultimately, we are still performing k -means clustering, however, the input to the simple clustering algorithm is what impacts the performance. Traditional clustering methods like k -means perform clustering on a lower-dimensional

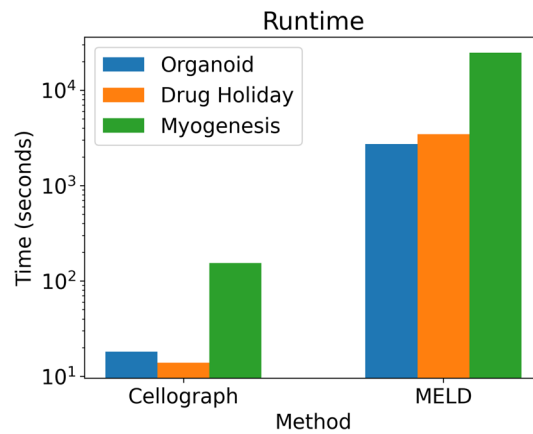


Fig. 6 Runtime of Cellograph's performance versus MELD's on optimal parameter settings. Cellograph consistently outperforms MELD on each dataset, while using fewer computing resources (y-axis is log-scaled)

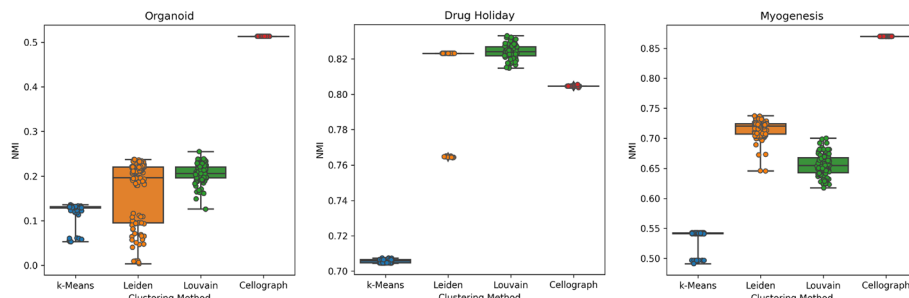


Fig. 7 Boxplots of NMI values per clustering algorithm. Distributions of 100 independent NMI calculations for each clustering algorithm for all three datasets evaluated, quantifying concordance between the cluster assignments and ground truth labels

PCA representation of the single-cell data. However, instead of a linear transformation of the data, we perform a non-linear transformation prior to clustering via the initial layer of the GCN. The clustering method itself is not novel, but the way the data is processed prior to clustering is. Instead, we emphasize that the novel contribution of Cellograph is a scalable means of ascertaining the effects of different experimental regimes at single-cell resolution.

Examining the sensitivity of the hyperparameters (the number of neighbors k , the latent dimension h , and principal components PCs) during our training by looking at the learning curves of accuracy and loss, we found Cellograph had fairly consistent performances with a minimum of $h = 16$ latent dimensions across all 3 datasets (Additional files 2, 3, and 4, for organoid, drug holiday, and myogenesis, respectively for tables of validation accuracy metrics for each combination of parameters).

Altogether, Cellograph outperforms MELD in estimating how prototypical cells are of their ground truth labels, and consistently ranks higher than standard algorithms for clustering.

Discussion

When designing single-cell experiments exploring the impacts of different treatments, it is vital to leverage the heterogeneity present at such resolution. The increasing complexity of the experimental design (e.g., multiple treatments, various timepoints, etc) can result in diminishing returns from standard differential gene expression and clustering approaches due to the biological and technical variability present at the single-cell level. Existing approaches like MELD and VFC are apt for studying the effects of one experimental treatment, but cannot be easily generalized to more complicated experimental programs. We designed Cellograph to address this challenge. Beyond just quantifying single-cell responses to perturbations analogously to MELD, Cellograph's primary innovation lies in its novel way of visualizing and clustering single-cell data by means of graph neural networks, which, through the parameterized gene weight matrix, provides an interpretable means of understanding which genes drive the difference between conditions. We have shown that our approach improves clustering on three diverse datasets compared to standard clustering approaches, as well as captures a stronger signal of the ground truth experimental label compared to MELD. Clustering agnostic of experimental conditions can fail to take into consideration the diversity of cellular responses to these perturbations and how those correspond to the transcriptomic variation. By applying simple k-means clustering to the latent space, we can obtain more informative clusters that enable deeper biological insight, especially in populations under the same experimental treatment. In addition to improved differential gene expression, we also obtain complementary information from the parameterized weight matrix after training, which reveals the most important genes in distinguishing between different treatments.

In a published dataset of donor-provided organoid samples, we were able to successfully corroborate original findings, while providing a visually informative view of the data, and revealed novel insights into drivers of KPT-mediated organoid differentiation. Similarly, in our drug holiday application, we identified additional markers of drug resistance using the parameterized gene weight matrix, and described heterogeneity of cells in response to Erlotinib after 11 days and post-holiday, while characterizing the population that was retreated after the holiday that could inform future experiments into druggable targets for NSCLC. Finally, in our myogenesis evaluation, we identified shared features between transdifferentiation and dedifferentiation, while capturing relevant markers that distinguished the two processes. We anticipate Cellograph will find a wide range of application to other biological contexts and different single-cell modalities as an all-in-one framework for facilitating visualization, clustering, and single-cell responses to perturbations, on top of its efficiency. For example, this work could find utility in clinical applications to studying heterogeneity in patient-treated samples in response to an experimental cancer drug. This could be also employed to study impacts of cancer drugs on cell cycle in protein immunofluorescence imaging data [32]. Potential extensions of our method could certainly explore the incorporation of batch effect corrections. While this could be a valuable avenue to address potential confounding factors and improve the robustness of the analysis, we want to emphasize that the primary objective of our method, as well as other methods in similar tasks, is not specifically focused on batch effect correction, and we would advise users to independently correct for any technical artifacts prior

to using Cellograph. If there are several replicates for a specific condition, the user may perform batch effect correction using approaches such as Harmony [33] or Seurat 3 [34], which have been independently shown to perform well in batch-effect correction [35].

The graph neural network architecture of node classification could even be extended to graph classification for looking at multiple patient samples, as is common in mass cytometry, or regression to predict continuous variables such as cellular pseudotime in the context of differentiation, cell cycle age [36], or gestational age in data from pregnant women [37], and may be further explored in future work. The choice of different graph constructions mechanisms could also warrant exploration in future studies. For example, CellVGAE uses variational graph autoencoders to reconstruct input graphs, adding additional, relevant edges, which can facilitate clustering and other downstream tasks in single-cell analysis [13]. scGNN is another GNN framework for single-cell analysis that selectively prunes edges in the initial k NN graph when pre-processing the data prior to training [9].

Concerning limitations of Cellograph, as a semi-supervised method, it requires labels to train on and make informed predictions. Consequently, in sparsely-labeled data or data with no labels at all, Cellograph's performance may fall short. In the case of sparse data, Cellograph could be used to impute the labels of other non-annotated cells. As for unlabeled data, one could pre-train Cellograph on a cellular atlas such as the Human Lung Cell Atlas [38] and apply the trained model to the unlabeled dataset to predict different cell types or distinguish diseased cells from healthy cells. Such endeavors could be the subject of future extensions of Cellograph. Because of the aforementioned dependency on labeled data, accurate labeling is imperative in achieving meaningful interpretation of Cellograph's results. In addition, data with multiple conditions that have similar phenotypes may present a challenge during Cellograph's learning due to the difficulties in separating conditions in the latent space.

Altogether, Cellograph provides a novel framework for perturbation analysis, data visualization, and feature importance in single-cell genomics. We anticipate it will find utility for testing drug efficacy in clinical samples, and the incorporation of other single-cell modalities, which may be explored in future studies.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-024-05641-9>.

Additional file 1. The additional file (SupplementalFile_cellograph.docx) provides supplementary figures of the corresponding UMAP embeddings of the respective single-cell transcriptomes on the Cellograph-derived latent space described in the main text (Figures 2, 3, and 4) and compares them to UMAPs of the multi-dimensional PCA coordinates, showcasing the visual gradient of treatment effects captured by the former visualizations.

Additional file 2. The additional file (S1.xlsx) provides tables examining the hyperparameter sensitivity used to evaluate Cellograph's robustness on the organoid dataset described in the main text.

Additional file 3. The additional file (S2.xlsx) provides tables examining the hyperparameter sensitivity used to evaluate Cellograph's robustness on the drug holiday dataset described in the main text.

Additional file 4. The additional file (S3.xlsx) provides tables examining the hyperparameter sensitivity used to evaluate Cellograph's robustness on the organoid dataset described in the main text.

Acknowledgements

The authors thank UNC Research Computing for providing the computational resources and technical support.

Author contributions

JAS and NS conceived the project. JAS derived the GNN methodology, developed the software package, and implemented it on various experimental datasets. NS provided useful input on the theoretical methodology and benchmarking. JEP provided critical insight into the biological interpretation, and supplied in-house, unpublished datasets to evaluate the method. JAS wrote the manuscript. JEP and NS critically read and commented on the manuscript. All authors read and approved the final manuscript.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650116 (to J.A.S.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. J.A.S. is supported by a fellowship from the Royster Society of Fellows at the University of North Carolina at Chapel Hill, and the aforementioned NSF Graduate Research Fellowship Program. J.E.P. is supported by NIH Grants R01-GM138834, NSF CAREER Award 1845796, and NSF 2242980.

Availability of data and materials

The metadata and digital gene expression data for the human organoid dataset was downloaded from <https://singlecell.broadinstitute.org> (study SCP1318). The Erlotinib drug holiday dataset was downloaded from the database Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo>) under the accession number GSE134841. The myogenesis dataset was downloaded from GEO under the accession number GSE171039. The code and installation instructions for Cellograph can be found at <https://github.com/jashahir/cellograph>.

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: 31 October 2023 Accepted: 4 January 2024

Published online: 15 January 2024

References

- Klein A, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz D, Kirschner M. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161(5):1187–201. <https://doi.org/10.1016/j.cell.2015.04.044>.
- Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019. <https://doi.org/10.15252/msb.20188746>.
- Haghverdi L, Büttner M, Wolf FA, Büttner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods*. 2016. <https://doi.org/10.1101/041384>.
- Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol*. 2021;40(2):245–53. <https://doi.org/10.1038/s41587-021-01033-z>.
- Reshef YA, Runkel L, Kang JB, Nathan A, Korsunsky I, Asgari S, Murray MB, Moody DB, Raychaudhuri S. Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat Biotechnol*. 2021;40(3):355–63. <https://doi.org/10.1038/s41587-021-01066-4>.
- Burkhardt DB, Stanley JS, Tong A, Perdigoto AL, Gigante SA, Herold KC, Wolf G, Giraldez AJ, Dijk D, Krishnaswamy S, et al. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat Biotechnol*. 2021;39(5):619–29. <https://doi.org/10.1038/s41587-020-00803-5>.
- Amodio M, Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, et al. Exploring single-cell data with deep multitasking neural networks. *Nat Methods*. 2019;16(11):1139–45. <https://doi.org/10.1038/s41592-019-0576-7>.
- Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15(12):1053–8. <https://doi.org/10.1038/s41592-018-0229-2>.
- Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, Wang C, Fu H, Ma Q, Xu D, et al. scGNN is a novel graph neural network framework for single-cell RNA-seq analyses. *Nat Commun*. 2021. <https://doi.org/10.1038/s41467-021-22197-x>.
- Ravindra N, Sehanobish A, Pappalardo JL, Hafner DA, Dijk D. Disease state prediction from single-cell data using graph attention networks. In: Proceedings of the ACM conference on health, inference, and learning 2020. <https://doi.org/10.1145/3368555.3384449>
- Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv; 2017. <https://doi.org/10.48550/ARXIV.1710.10903>.
- Sehanobish A, Ravindra NG, Dijk D. Gaining insight into SARS-CoV-2 infection and COVID-19 severity using self-supervised edge features and graph neural networks. arXiv; 2020. <https://doi.org/10.48550/ARXIV.2006.12971>.
- Buterez D, Bica I, Tariq I, Andrés-Terré H, Liò P. CellVGA: an unsupervised scRNA-seq analysis workflow with graph attention networks. *Bioinformatics*. 2021;38(5):1277–86. <https://doi.org/10.1093/bioinformatics/btab804>.

14. Rhee S, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence; 2018. <https://doi.org/10.24963/ijcai.2018/490>.
15. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *CoRR* **abs/1609.02907**; 2016. [arXiv:1609.02907](https://arxiv.org/abs/1609.02907).
16. Song Q, Su J, Zhang W. scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat Commun*. 2021;12(1):3826.
17. McDaid AF, Greene D, Hurley N. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv*; 2011. <https://doi.org/10.48550/ARXIV.1110.2515>.
18. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction; 2018. [arXiv:1802.03426](https://arxiv.org/abs/1802.03426).
19. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc*. 2021;16(1):1–9.
20. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19:1–5.
21. Mead BE, Hattori K, Levy L, Imada S, Goto N, Vukovic M, Sze D, Kummerlowe C, Matute JD, Duan J, et al. Screening for modulators of the cellular composition of gut epithelia via organoid models of intestinal stem cell differentiation. *Nat Biomed Eng*. 2022;6(4):476–94. <https://doi.org/10.1038/s41551-022-00863-9>.
22. Aissa AF, Islam AB, Ariss MM, Go CC, Rader AE, Conrardy RD, Gajda AM, Rubio-Perez C, Valyi-Nagy K, Pasquinelli M, et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun*. 2021. <https://doi.org/10.1038/s41467-021-21884-z>.
23. Yagi M, Ji F, Charlton J, Cristea S, Messemer K, Horwitz N, Di Stefano B, Tsopoulidis N, Hoetker MS, Huebner AJ, et al. Dissecting dual roles of MyoD during lineage conversion to mature myocytes and myogenic stem cells. *Genes Dev*. 2021;35(17–18):1209–28. <https://doi.org/10.1101/gad.348678.121>.
24. Zhang Y, Jiang M, Nouraié M, Roth MG, Tabib T, Winters S, Chen X, Sembrat J, Chu Y, Cardenes N, et al. Gdf15 is an epithelial-derived biomarker of idiopathic pulmonary fibrosis. *Am J Physiol-Lung Cell Mol Physiol*. 2019. <https://doi.org/10.1152/ajplung.00062.2019>.
25. Yang M, Darwish T, Larraufie P, Rimmington D, Cimino I, Goldspink DA, Jenkins B, Koulman A, Brighton CA, Ma M, et al. Inhibition of mitochondrial function by metformin increases glucose uptake, glycolysis and GDF-15 release from intestinal cells. *Sci Rep*. 2021. <https://doi.org/10.1038/s41598-021-81349-7>.
26. Schrader CH, Kolb M, Zaoui K, Flechtenmacher C, Grabe N, Weber K-J, Hielscher T, Plinkert PK, Hess J. Kallikrein-related peptidase 6 regulates epithelial-to-mesenchymal transition and serves as prognostic biomarker for head and neck squamous cell carcinoma patients. *Mol Cancer*. 2015. <https://doi.org/10.1186/s12943-015-0381-6>.
27. Zhou D, He Y, Li H, Huang W. KLK6 mediates stemness and metabolism of gastric carcinoma cells via the PI3K/AKT/mTOR signaling pathway. *Oncol Lett*. 2021. <https://doi.org/10.3892/ol.2021.13085>.
28. Wallace L, Mehrabi S, Bacanamwo M, Yao X, Aikhionbare FO. Expression of mitochondrial genes MT-ND1, MT-ND6, MT-CYB, MT-COI, MT-ATP6, and 12S/MT-RNR1 in colorectal adenopolyps. *Tumor Biol*. 2016;37(9):12465–75. <https://doi.org/10.1007/s13277-016-5101-3>.
29. Zhao K, Yi Y, Ma Z, Zhang W. INHBA is a prognostic biomarker and correlated with immune cell infiltration in cervical cancer. *Front Genet*. 2022. <https://doi.org/10.3389/fgene.2021.705512>.
30. Lin Y-L, Persaud SD, Nhieu J, Wei L-N. Cellular retinoic acid-binding protein 1 modulates stem cell proliferation to affect learning and memory in male mice. *Endocrinology*. 2017;158(9):3004–14. <https://doi.org/10.1210/en.2017-00353>.
31. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods*. 2022;19(1):41–50.
32. Gut G, Herrmann MD, Pelkmans L. Multiplexed protein maps link subcellular organization to cellular states. *Science*. 2018. <https://doi.org/10.1126/science.aar7042>.
33. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods*. 2019;16(12):1289–96.
34. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. *Cell*. 2019;177(7):1888–902.
35. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;21:1–32.
36. Stallaert W, Kedziora KM, Taylor CD, Zikry TM, Ranek JS, Sobon HK, Taylor SR, Young CL, Cook JG, Purvis JE, et al. The structure of the human cell cycle. *Cell Syst*. 2022;13(1):103. <https://doi.org/10.1016/j.cels.2021.12.006>.
37. Aghaepour N, Ganio EA, Mcilwain D, Tsai AS, Tingle M, Van Gassen S, Gaudilliere DK, Baca Q, McNeil L, Okada R, et al. An immune clock of human pregnancy. *Sci Immunol*. 2017. <https://doi.org/10.1126/sciimmunol.aan2946>.
38. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madisson E, Markov NS, Zaragosi L-E, Ji Y, Ansari M, et al. An integrated cell atlas of the lung in health and disease. *Nat Med*. 2023;29:1–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.