Gaining insight into SARS-CoV-2 infection and COVID-19 severity using self-supervised edge features and Graph Neural Networks

Arijit Sehanobish

Yale University

July 17, 2020

Joint with: Neal G. Ravindra* and David van Dijk

Outline

1. Introduction

2. Experiments
   - Datasets used
   - Creating new edge features
   - Model Description
   - Experimental Results

3. Model Interpretation

4. References

Introduction

Single-cell transcriptome sequencing (scRNA-seq)

- Provides the expression profiles of individual cells and is considered the gold standard for defining cell states and phenotypes.
- Widely used across biological disciplines including Developmental biology, Neurology, Oncology, Immunology, and Infectious disease.
- However, identifying factors important for determining an individual cell's pathophysiological trajectory or response to viral insult remains a challenge as single-cell data is noisy, sparse, and multi-dimensional.
- After some standard data preprocessing, yields large sparse graphs.

Graph Neural Networks

Graph Neural Networks

- Graph Neural Networks (GNN) widely used in node classification, link prediction and graph classification.
- Lots of research in understanding the expressive power of GNNs.
- Typically use message passing or neighborhood aggregation to create new node feature vectors.
- Most graphs do not apriori come with edge features and thus most GNNs ignore edge features.

Our work

Our contributions

- Create new edge features in a self-supervised manner that are applicable to any single cell data.
- Show that these edge features improve the baseline GNNs for node classification tasks (in an inductive setting).
- Interpret our model to gain insight into SARS-CoV-2 infection dynamics and COVID-19 disease severity on an individual gene and cell level.

Datasets used

Table 1: Dataset description showing train/val/test splits.

Datasets	SARS-CoV-2 infected organoids	COVID-19 patients
# Nodes	54353/11646/11648	63486/13604/13605
# Node features	24714	25626
# Edges	1041226/230429/228630	2746280/703217/707529
# Edge features	18	18
# Classes	7	3

- 4 human bronchial epithelial cell cultures or “organoids” that were inoculated with SARS-CoV-2 and co-cultured for 1, 2, and 3 days post-infection [10].
- Bronchoalveolar lavage fluid samples from 12 patients enrolled in a study at Shenzen Third People’s Hospital in Guangdong Province, China of whom 3 were healthy controls, 3 had a mild or moderate form of COVID-19 and 6 had a severe or critical COVID-19 illness [9].

Unsupervised Clustering Method

Louvain Clustering

- Unsupervised clustering method to extract communities from large networks [2].
- Inspiration for this method of community detection is the optimization of modularity.
- Widely used in biological networks.

Using GAT to create edge features

- Train a 2-layer 8-head Graph Attention Network [11] to predict the cell types given by Louvain clustering.
- Extract and concatenate the (normalized) attention coefficients obtained from each head.

Equations for attention coefficients

The (normalized) self-attention α_{ij}^l between nodes i and j is given by :

$$e_{ij}^l = a^l(\mathbb{W}^l h_i, \mathbb{W}^l h_j)$$

$$\alpha_{ij}^l = \text{softmax}_j(e_{ij}^l) = \frac{\exp(e_{ij}^l)}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik}^l)}$$

Creating other edge features

Other edge features used

- Use GAT to predict batch ID, i.e. an unique identifier that keep track of what patient/lab the cells come from. Extract the learned attention coefficients as before.
- Use Forman-Ricci curvature [3] as an intrinsic topological invariant of the graph.
- Use node2vec [4] to embed the nodes in a 16-dimensional vector space and calculate similarity between nodes via dot product. Just as before we only calculate the dot product between 2 embedded nodes only if they share an edge.

Our model

Parts of our model

- For each node i, create a set $S_i := \{v_{ij} : j \in N_i\}$, where v_{ij} is the vector representing the edge features of the edge connecting nodes i and j.
- Encode this set S_i, which we call the edge feature set attached to the node i via any permutation invariant network like DeepSet [13], Set2Set [12] or Set Transformer [8].
- Use a message passing network like GCN [7], GAT [11] or GraphSage [5] to encode the node features.
- Concatenate the encoded node features with the encoded edge features and pass it through a linear layer for node classification.

Model Figure

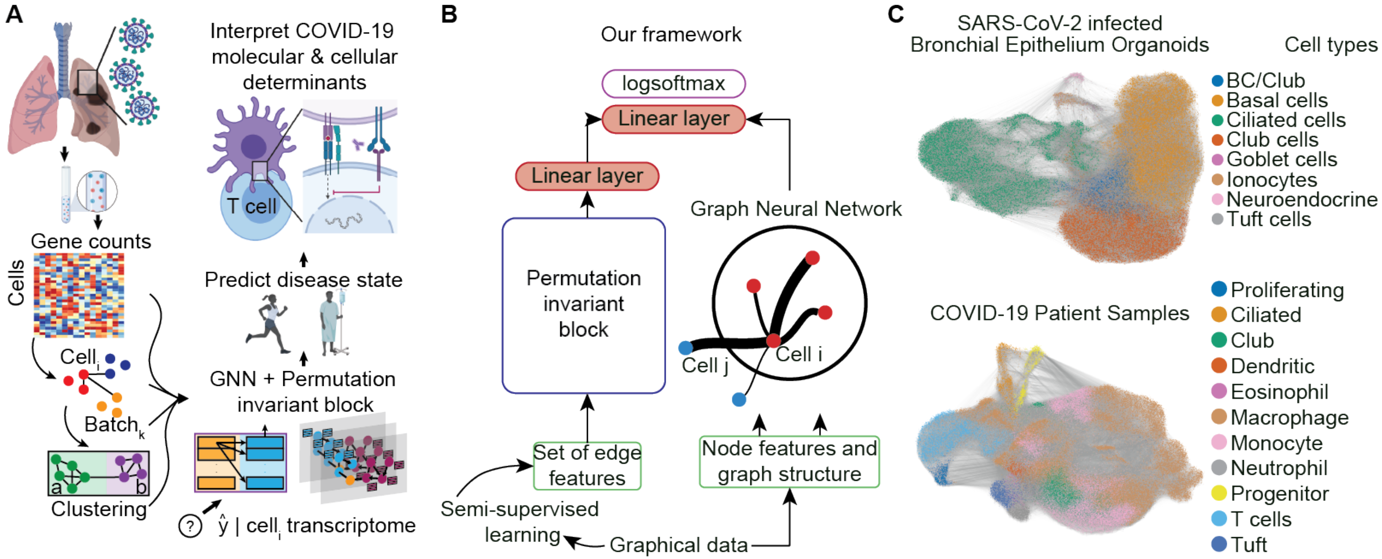

Figure 1: Our framework and datasets of interest. (**A**) Overview of our approach (**B**) Our framework (**C**) Graphical data used

Results

Table 2: Results on inductive tasks on single-cell datasets showing accuracy and 95% confidence intervals.

Models	SARS-CoV-2 infected organoids	COVID-19 patients
GCN (baseline)	65.43 (65.21-65.65)	89.26 (89.06-89.47)
GCN + DeepSet	79.75 (78.75-80.75)	87.2 (87.02-87.38)
GCN + Set2Set	71.65 (69.89-73.42)	88.34 (87.89-88.79)
GCN + Set Transformer	**81.61 (79.34-83.87)**	**92.84 (91.95-93.74)**
GAT (baseline)	73.10 (70.93-75.27)	92.25 (91.27-93.24)
GAT + DeepSet	79.45 (77.98-80.92)	89.8 (88.89-91.71)
GAT + Set2Set	75.99 (74.8-77.68)	92.87 (92.62-93.12)
GAT + Set Transformer	**82.95 (81.75-84.15)**	**95.12 (94.02-96.22)**

Interpreting our models to understand disease state predictions – I

For interpretability purposes will use GAT+Set Transformer model.

Understanding SARS-CoV-2 infected organoids

- Extract the learned weights from our models' first GAT layer to investigate feature saliency with respect to gene importance.
- Find saliency in counts of viral transcript, as well as genes that are involved in inflammatory response and cell death (NFKBIA) and signaling (IFI27, HCLS1, NDRG1, NR1D1, TF).
- Learned embedding shows that our model segregates infected ciliated cells, which is the reported SARS-CoV-2 cell tropism, validating our models' interpretability [10].

Interpreting our models to understand disease state predictions – II

Understanding COVID-19 severity from patient samples

- Extracting weights from GAT shows high weight to features (genes) involved in the innate immune system response to type I interferon (CCL2, CCL7, IFITM1), regulation of signaling (NUPR1, TAOK1, MTRNR2L12), a component of the major histocompatibility complex II (HLA-DQA2) important for developing immunity to infection, and a marker of eosinophil cells, which are cells involved in fighting parasites (RETN).
- Our model mixes macrophages and monocytes in a predominantly severe patient cell cluster while cells derived from mild and severe COVID-19 patients are mixed in a T cell cluster.
- Monocytes derived from macrophages are thought to be enriched in severe COVID-19 cases and T cells are proposed targets for immune checkpoint therapy of COVID-19 [1, 9, 6].

Interpretability

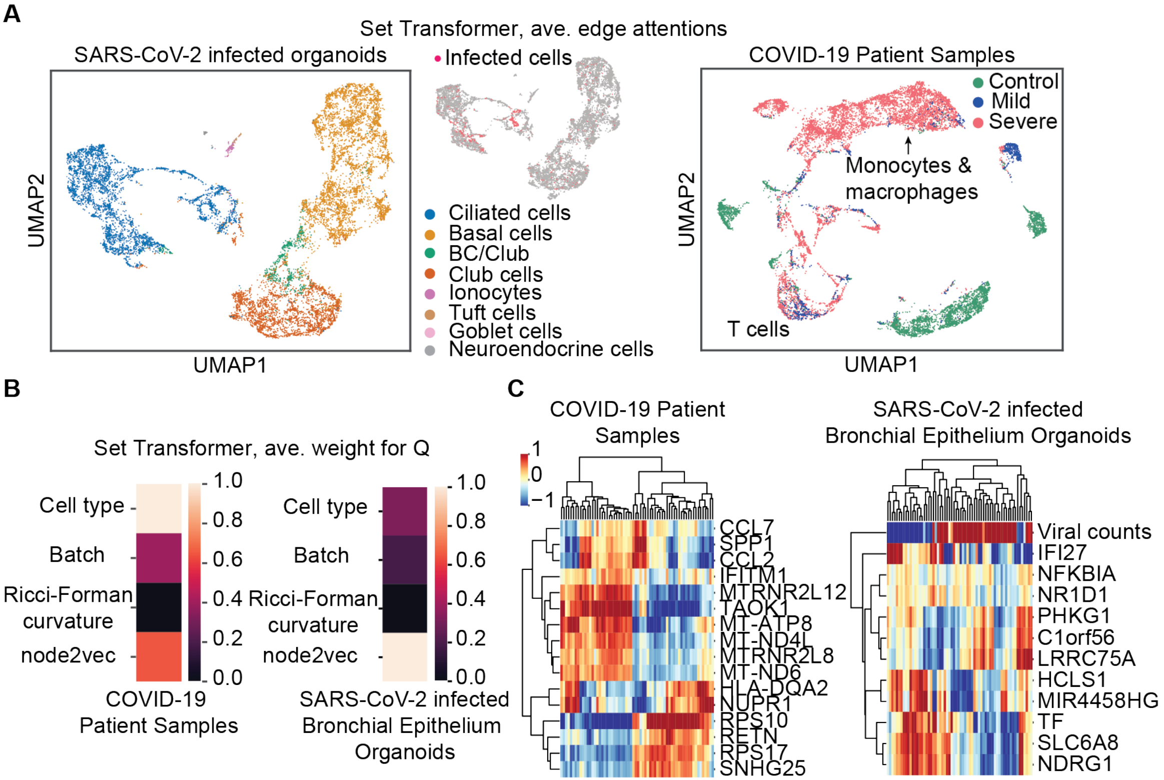

Figure 2: Model interpretability to understand genes and cells important to COVID-19 severity. (**A**) Learned graphical representations using attention from Set Transformer (**B**) Relative importance of edge features in prediction tasks. (**C**) Highly weighted gene features for infection and severity prediction

Concluding Remarks

Important Note

- However, we are not medical professionals so we do *NOT* claim that interpretation of our model will bear any fruit.
- Hope that the approach of seeking state-of-the-art results on predicting disease states at single-cell resolution will enhance study of biology and medicine and potentially accelerate our understanding of critical diseases.
- Further study into the interaction partners and the subtle transcriptional differences between the cells and cell types that we identified may provide complementary hypotheses or avenues for therapeutic intervention to mitigate the impacts of COVID-19.

Thank You!!

Bibliography I

[1] M. Bersanelli. Controversies about covid-19 and anticancer treatment with immune checkpoint inhibitors. *Immunotherapy*, 12(5):269–273, 2020. doi: 10.2217/imt-2020-0067.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008. doi: 10.1088/1742-5468/2008/10/p10008.

[3] R. Forman. Bochner's method for cell complexes and combinatorial ricci curvature. *Discrete and Computational Geometry*, 29:323–374, 2003.

[4] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks, 2016.

[5] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems (NIPS)*, pages 1025–1035, 2017.

[6] B. Israelow, E. Song, T. Mao, P. Lu, A. Meir, F. Liu, M. Madel Alfajaro, J. Wei, H. Dong, R. J. Homer, A. Ring, C. B. Wilen, and A. Iwasaki. Mouse model of sars-cov-2 reveals inflammatory role of type i interferon signaling. *bioRxiv*, 2020. doi: 10.1101/2020.05.27.118893. URL `https://www.biorxiv.org/content/early/2020/05/27/2020.05.27.118893`.

[7] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2017. URL `https://openreview.net/pdf?id=SJU4ayYgl`.

[8] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks, 2018.

[9] M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, L. Cheng, J. Li, X. Wang, F. Wang, L. Liu, I. Amit, S. Zhang, and Z. Zhang. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine*, 2020. doi: 10.1038/s41591-020-0901-9.

[10] N. G. Ravindra, M. M. Alfajaro, V. Gasque, J. Wei, R. B. Filler, N. C. Huston, H. Wan, K. Szigeti-Buck, B. Wang, R. R. Montgomery, S. C. Eisenbarth, A. Williams, A. M. Pyle, A. Iwasaki, T. L. Horvath, E. F. Foxman, D. van Dijk, and C. B. Wilen. Single-cell longitudinal analysis of sars-cov-2 infection in human bronchial epithelial cells. *bioRxiv*, 2020. doi: 10.1101/2020.05.06.081695.

[11] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks, 2018. URL `https://openreview.net/forum?id=rJXMpikCZ`.

[12] O. Vinyals, S. Bengio, and M. Kudlur. Order matters: Sequence to sequence for sets, 2015.

[13] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep sets, 2017.