2023, 8-27
doi: tbd
Advance Access Publication Date: tbd
Research Article

Celltype Heterogeneity Extraction and Encoding in Single-cell Integrated Multimodalities

CHEESgrate

Alexander Halpin, Brendan Leahey, Micah Lessnick, and Winston Li

Abstract

Motivation: An abundance of genomic information exists in modes such as RNA-seq, ATAC-seq, and CITE-seq. Computational biologists have taken advantage of this information with the application of deep learning methods like convolutional neural networks (CNNs) and generative approaches (GANs, autoencoders). Recently, several state-of-the-art networks have obtained desirable results on single-cell analysis of COVID-19 disease factors through the alignment of relevant datasets in a latent space. These spaces are formed without a specific task in mind; we aim to incorporate cell-type specific information to improve the utility of the outputted space for tasks such as classification of case severity.

Results: By guiding alignment of data from different datasets with the use of cell-type-specific information, CHEESgrate is able to attain better classification results than comparable models with inputs from unguided training. Further, they offer improved interpretation of contributions from cell-type-specific factors, providing biologically valuable information.

Contact: *alexander_halpin@brown.edu, brendan_leahey@brown.edu, micah_lessnick@brown.edu, winston_li@brown.edu*

Supplementary information: Training data and supplementary figures will be available on our repository linked at *https://github.com/lwinstony8/DNAvigators*

1 Introduction

The advent of single-cell multimodal omics data, or multiomics, present unrivaled insights into physiology at the individual cellular level. Such data allows researchers to distinguish between cell-types, and moreover, the differential dynamics of different cells in response to similar stimuli as in a pathological state. Assays such as *Cellular indexing of transcriptomes and epitopes by sequencing* "CITE-seq" simultaneously quantifies surface protein levels and transcriptomic data of a single cell (Stoeckius et al. 2017), while SHARE-seq simultaneously capture chromatin accessibility and gene expression levels by combining ATAC- and RNA-seq data (Clyde 2021).

However, integrating these different multimodal assays at the single-cell level has remained challenging due to variations in statistical distributions, noise, sparsity, and increased dimensionality (Argelaguet et al. 2021). Furthermore, alignment is required when single-cell multimodal assays are performed at the tissue-level, since there are multiple cell-types in tissue. This is especially relevant when we consider the cellular heterogeneity hallmark of tumor microenvironments and immune responses to diseases such as COVID-19.

In this paper, we present CHEESgrate, which seeks first to align single-cell multimodal data to distinct cell-types. Then, using the integrated multimodal data that comprehensively "snapshots" each particular cell-type, we can predict disease severity and interpret influencing factors at the cell-type-specific level. Compared to past works, CHEESgrate explicitly simulates the contributions of particular cell-types and their interactions among each other.

2 Related Work

Fusing genomics with proteomics has allowed researchers to study a myriad of bioinformatics applications. Asgari and Mofrad (2015) pioneered a framework for feature extraction called BioVec (ProtVec and GeneVec for proteomics and genomics, respectively), which combined proteomics and genomics into a single unified embedding space. Unifying these omics in a single framework laid the groundwork for future deep learning models specializing in tasks such as protein family classification and protein structure prediction, now aided with additional genomics information. Complementary to this initial framework, Zhao et al. (2021) proposed a model called DeepOmix, which combines genomics, transcriptomics, and epigenomics to predict cancer survivability. Such tools show increasing promise in laboratory settings; DeepOmix outpaced other cutting-edge predictive models. In turn, this multiomics approach would increase the predictive capability and hasten effective treatments, ultimately improving patient outcomes.

Many multimodal methods are limited by the lack of parallel data between modes. To address this issue, alignment algorithmically determines cells with similar characteristics across different samples, enabling the utilization of multi-omics data that may not be directly paired. Stanojevic et al. introduces a variety of alignment methods used to generate single-cell multi-omic data. Neural networks may be applied to learn feature based representations of samples that may be aligned more easily. Similar to integration methods, generative methods such as autoencoders and generative adversarial networks may be applied to learn mappings between these spaces. While often limited by the assumption that these datasets contain some correspondence information, this expands the quantity of valuable omics information resulting in novel outputs such as multimodal reference atlases that suffice as inputs for learning multi omics information.

Prior studies have shown improving the fusion methods used prior to integration presents opportunities to both increase model performance and interpretability; Schreiber et al. (2020)'s AVOCADO provides example on how fusing different data sources into a latent representation presented better inputs for models to predict tasks compared to other models trained *directly* on aforementioned individual data sources. Given that we are working in latent spaces, we can also observe how perturbations in latent spaces, representing the *combined effects* on the

data inputs, can lead to potentially more biologically relevant interpretations.

It is worth noting that not all multimodal approaches yield superior results. An ablation experiment conducted by Boehm et al. raised concerns as their findings indicated that a late-fusion model, incorporating genomic, histopathological, radiological, and clinical information, performed worse than models that selected fewer data modalities. This emphasizes the need to approach multimodal deep learning with caution, as merely providing a greater number of input modes does not guarantee improved network performance. In light of this fact, we wish to explore how relevant inputs may be optimized for interpretation within a model's latent space.

We narrow our discussion on how the generation and application of multimodal inputs for COVID-19 severity classification may be streamlined using key relationships in our input. Several networks have carried out related multiomic tasks recently, including Multigrate, scMM, and Babel.

Introduced by Lotfollahi et al., Multigrate several peripheral blood mononuclear cell datasets into a shared representation that may be queried along key axes such as cell type and condition. Its autoencoder structure uses a shared decoder to impute values that are missing within the original dataset. As part of their efforts to build a COVID-19 atlas, the authors of Multigrate demonstrated how a smaller "query" atlas of sample diseased COVID-19 cells can be aligned to a "reference" atlas of healthy cells. In particular, cell-type annotations were added to the query atlas via random forest classifier. Visualizations of Multigrate's latent space through UMAP clustering demonstrate cell type to be very salient in the alignment process.

Even more recently, Zhou et al. (2023) further develops upon cell-type specificities in moETM. In this model, a product-of-experts framework is used to infer latent topics underlying single-cell multiomics data, which the authors call a "topic mixture membership" for the cell. These topics can then be mapped to cell-types based on the top gene signatures as revealed by the decoding of this topic latent space. The authors showcased moETM's ability to reveal immune cell-type signatures and identify cell-type-specific pathways and regulatory motifs.

In both cases, while cell-type specific information is mentioned, they come as an output of the model; they are inferred. Especially in Multigrate, cell-type signatures are a measure of quality; similarly in moETM, cell-type signatures are likewise used as a measure of quality in clustering and also biological interpretability of the latent topic space.

In contrast, CHEESgrate deliberately incorporates cell-type heterogeneity when making predictions, rather than merely using such information as an interpretation tool. We believe that exploiting the similarity between cells of the same type will allow us to more accurately classify other factors such as a patient's case severity. Furthermore, by explicitly forcing the model to account for cell-type information, we seek to emphasize factors at the cell-type specific level as well as interactions between different cell-types in predicting a disease severity. This in turn captures the diverse cellular landscape of a pathological state, and how multiple different cell-types contribute both uniquely and concertedly to disease. Consequently, this permits finer resolution interpretations that align better with biological phenomena, since we can directly analyze each individual cell-type's contribution to a prediction. More importantly, through perturbation analyses, we can "simulate" scenarios where particular cells are different from a "default" disease state; in such scenarios, we would portray a variant of a disease - a disease subtype - and thus observe how case severity may change.

3 Methods

To learn a meaningful combined representation of the multi-omics data we will utilize a variational autoencoder to construct a regularized latent space. The encoder structure of the model will be used for upstream tasks including survivability prediction and the decoder model can be used in combination with the encoding for imputation of missing modalities.

For data, we use several peripheral blood mononuclear cell datasets in a similar fashion to Multigrate. Dataset 1 is the 10X Single Cell Multiome ATAC + Gene Expression dataset (2021). Datasets 2-4 are CITE-seq datasets from Hao et al. (2020), Kotliarov et al. (2020), and Stephenson et al (2021).

To incorporate multiple modalities at the cell-type specific level, we begin with a variational autoencoder architecture to integrate together the multimodal data *for each cell-type*. The input for each "sub-VAE" will be from the datasets above with cell-type information extracted from metadata.

Having obtained a latent representation per cell-type, we can then fuse these representations together into a new latent space using a second VAE. From this latent space, we can imagine several architectures for making the final COVID severity prediction.

The autoencoder is an encoder decoder architecture that learns to project high-dimensional data into a compressed latent space by reconstructing the original data after passing through a bottleneck. Variational autoencoders are similar to the traditional autoencoder model in that the information is still passed through as a relatively low dimensional latent vector however the latent vector is instead the parameters of a normal distribution. By forcing the model to learn to plot each dimension of the latent vector as a distribution and sampling from this vector's distribution before passing to the decoder, we can construct a more regularized latent space.

The variational auto-encoder will consist of a series of linear layers exhibiting the above mentioned bottleneck pattern. For each sample with m modalities with feature length k, the encoder will take as input m k-dimensional vectors Mi $\in$ Rk . For each modality vector Mi the encoder will compute a mean μi and standard deviation σι. The latent vector Z will be sampled from the distribution parameterized by the outputs of the encoder and passed to the decoder.

The model incorporates two different components into its loss function: a reconstruction loss and a regularization loss.

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

The reconstruction loss is a standard loss function that ensures that the decoder model accurately predicts the input data from the compressed latent vector. The regularization loss function ensures that the encoder outputs parametrize distributions similar to the prior distribution which we assume to be a normal distribution. The regularization loss is important because it provides the variational functionality to the model and prevents the model from "memorizing" the input data-points and instead encourages the different features of Z to exist as non-overlapping distributions in the latent space. The regularization loss is computed using the parameters μισι output by the encoder.

4 Expected Results

The purpose of CHEESgrate is two-fold: first, to provide better predictions of COVID severity compared to existing methods; second, to present a biologically relevant interpretation of the contributions of each cell-type.

4.1 Improved Prediction of COVID severity

Our results will depend on the assumption that CHEESgrate does in fact improve classification over single-modality models. Initially, we will treat other models' performance on single modality inputs as the most basic baseline. A summary of these results can be seen below in Table 1. We observe that current single-omic methods on RNA sequences perform extremely well, while multi-omics datasets lag behind slightly. Conversely, when combined proteomics and metabolomics outperform their individual counterparts, demonstrating our experiment to be worthwhile. Eventually, we seek to obtain the results of training on the latent output of models such as moETM or Multigrate for a "true" baseline within this task. This will be a more accurate representation of whether our experiment is an improvement on previous methods, and if it was worthwhile.

Table 1. Performance of similar networks over related input spaces

Model	Input	AUROC	AUPRC
1	RNA-SEQ (Korean)	1.0	.09
1	RNA-SEQ (Stanford)	0.86	0.89
2	RNA-SEQ+Spectometry	.93	-
3	Proteomics+Metablomics	.97	-
3	Proteomics	.90	-
3	Metablomics	.93	-

Table indicates that models trained on single input RNA-seq streams perform extremely well. However, existing multimodal methods currently lag behind, which we seek to improve

Further, our hope is that CHEESgrate does not sacrifice performance along related methods' key functions such as imputation, clustering, and reference building. A summary of clustering performance is displayed in Table 2. This illustrates that MoETM, the most recent rendition of a multiomic embedding model, typically performs the best at the clustering task. Giving up information like cell type will likely impact CHEESgrate's ability to perform supplementary tasks. Given that our focus is on post-processing classification, this is a cost we are willing to bear. Regardless, this experimentation will provide other valuable insights into the mechanisms of multimodal alignment.

Table 2. Comparison of existing methods' clustering effectiveness, obtained from moETM's supplementary data

Metrics	Methods	Genes + Peaks			
		BMMC	MSLAC	MKC	MBC
ARI	moETM	0.735	0.515	0.584	0.468
	SMILE	0.732	0.477	0.439	0.301
	scMM	0.693	0.412	0.420	0.333
	Cobolt	0.664	0.400	0.394	0.303
	MultiVI	0.697	0.413	0.403	0.502
	MOFA+	0.709	0.489	0.424	0.403
	Seurat V4	0.706	0.547	0.403	0.538
NMI	moETM	0.798	0.665	0.643	0.601
	SMILE	0.784	0.617	0.47	0.445
	scMM	0.744	0.583	0.488	0.524
	Cobolt	0.738	0.568	0.464	0.428
	MultiVI	0.755	0.580	0.470	0.669
	MOFA+	0.769	0.631	0.496	0.574
	Seurat V4	0.782	0.702	0.488	0.694

We aim for CHEESgrate to maintain performance along core functions such as clustering and imputation (not depicted, but the data is available alongside this table)

4.1.1 Choice of Evaluation Metrics

One of our supplementary goals is to develop balanced classes by imputing missing data points. For now, we choose to utilize area under the receiver operating characteristic (AUROC) and area under the precision recall curve (AUPRC), as these are the most resilient to class imbalance, and therefore most accurately represent a problem where this information is not available. However, we may also include raw accuracy if we are able to develop a balanced input as it is a simple and approachable metric for a wider audience.

Similarly, adjusted random index (ARI) and normalized mutual information (NMI) are strong in balanced and imbalanced tasks respectively, and will be used as such. Evaluating clustering in this way will be essential to understand how the clustering task is affected by the modification of its inputs, and we will adopt similar, appropriate metrics for evaluating others tasks.

4.2 Improved Model Interpretability

We foresee two major methods for interpreting CheesGRATE. First, through the use of attention-maps, we can observe which cell-type's latent space was paid the most attention by the model. From these latent-spaces, we can also reconstruct the original multimodal data to see which features had changed (e.g. compared to a healthy sample, how has gene expression or chromatin accessibility changed in a cell-type?). This latter method provides insights into a possible underlying mechanism of COVID, as presumably, what the model pays the most attention to is that which holds the most predictive power in determining COVID severity, and therefore the most influential. By decoding and thus reconstructing the multimodal space, we can observe the biological analog of the aforementioned influential factor.

Alternatively, through selective perturbation analysis, we can input cell-types whose multimodal features are modified. For example, we can perturb a Helper T lymphocyte to express less Th2 cytokine production, we would conceptually expect reduced capacity of the COVID patient to mount an effective adaptive immune response, and thus worse disease severity. Conversely, we can consider perturbations associated with decreased COVID severity such as inactivation of inflammatory cells, and so downregulation of their metabolic products, to prevent the "cytokine storm" associated with severe COVID cases. Note, we emphasize selective when discussing perturbation analysis since it will be computationally infeasible to simulate every perturbation (not to mention the fact that these are continuous values). Thus, we will mainly investigate biologically relevant perturbations as found in the literature.u

Other methods will also be considered depending on our choice of implementation of the model. If we design our latent spaces to be biologically relevant, we can perform class optimization to see which combination of cell-types and their specific multimodal values lead to a particular latent space. Likewise, at a higher level, we can try to optimize for the allegedly "healthiest" patient and the "most severe" patient.

Ultimately, to verify the accuracy of these interpretive methods, we will have to reference the literature for biochemical underpinnings of COVID. For example, with regards to attention-maps, we can conduct a literature search on publication databases like PubMed to see if a particular feature the model is attentive on has been discussed before. Similarly for perturbation analysis, we can consider real, held-out cases of severe or healthy COVID to first compare its accuracy and second its biological relevance.

5 Proposed Timeline

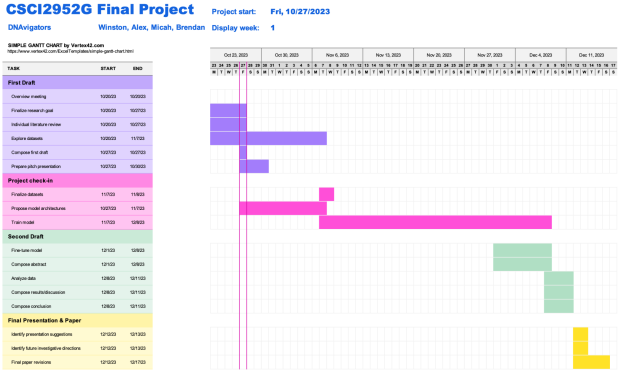

Figure 1. Proposed Timeline as a GANTT Chart.

6 Discussion

A discussion of the expected results is given in Section 4. Expected Results. To reiterate, we wish to focus on two aspects of CHEESgrate: model performance in predicting COVID severity and interpretability of the model.

In the former, we will compare CHEESgrate to existing models for predicting COVID severity. In the latter, we will perform several methods for interpreting the contribution of inputs as well as a survey of the biological landscape as "simulated" by our model. To confirm the real-life validity of our scenarios, we will reference the literature for biochemical underpinnings of COVID as to see whether 1) our model paid "attention" to such a factor, 2) our model correctly understands how such a factor affects COVID severity, 3) our model can make realistic replicate "default" patients that are either healthy or unhealthy via class optimization.

(Optimistically) To conclude, CHEESgrate seeks to capture the heterogeneity of a disease landscape by considering multimodal features at the individual cell-type level. This is done by explicitly including cell-type information as an input. Consequently, our model casts its predictions based upon features defined at the cell-type-specific level, and so it will consider individual cell-type's contributions as well as their modular effects with other cell-types. This allows us to interpret our model in a more biologically relevant manner and verify the model's realistic validity via comparison to the literature.

Acknowledgements

We would like to acknowledge Ritambara Singh, the people who review our paper, etc.

Conflict of Interest: none declared.

References

Argelaguet, R., Cuomo, A. S. E., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, *39*(10), Article 10. https://doi.org/10.1038/s41587-021-00895-7

Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, *40*(8), 865-878.e6. https://doi.org/10.1016/j.ccell.2022.07.004

Chen, Z., Liu, Y., Zhang, Y., & Li, Q. (2023). Orthogonal latent space learning with feature weighting and graph learning for multimodal Alzheimer's disease diagnosis. *Medical Image Analysis*, *84*, 102698. https://doi.org/10.1016/j.media.2022.102698

Clyde, D. (2021). SHARE-seq reveals chromatin potential. *Nature Reviews Genetics*, *22*(1), Article 1. https://doi.org/10.1038/s41576-020-00308-6

Demetci, P., Tran, Q. H., Redko, I., & Singh, R. (2022). *Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport* (p. 2022.11.09.515883). bioRxiv. https://doi.org/10.1101/2022.11.09.515883

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., … Satija, R. (2020). Integrated analysis of multimodal single-cell data (p. 2020.10.12.335331). bioRxiv. https://doi.org/10.1101/2020.10.12.335331

Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. Foundations and Trends® in Machine Learning, 12(4), 307–392. https://doi.org/10.1561/2200000056

Kotliarov, Y., Sparks, R., Martins, A. J., Mulè, M. P., Lu, Y., Goswami, M., Kardava, L., Banchereau, R., Pascual, V., Biancotto, A., Chen, J., Schwartzberg, P. L., Bansal, N., Liu, C. C., Cheung, F., Moir, S., & Tsang, J. S. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, *26*(4), Article 4. https://doi.org/10.1038/s41591-020-0769-8

Lotfollahi, M., Litinetskaya, A., & Theis, F. J. (2022). *Multigrate: Single-cell multi-omic data integration* (p. 2022.03.16.484643). bioRxiv. https://doi.org/10.1101/2022.03.16.484643

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Avsec, Z., Misharin, A. V., & Theis, F. J. (2020). *Query to reference single-cell integration with transfer learning* (p. 2020.07.16.205997). bioRxiv. https://doi.org/10.1101/2020.07.16.205997

Madhumita, & Paul, S. (2022). Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping. *Computers in Biology and Medicine*, *148*, 105832. https://doi.org/10.1016/j.compbiomed.2022.105832

pbmc_granulocyte_sorted_10k -Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support. (n.d.). Retrieved October 27, 2023, from https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_10k

Schneider, L., Laiouar-Pedari, S., Kuntz, S., Krieghoff-Henning, E., Hekler, A., Kather, J. N., Gaiser, T., Fröhling, S., & Brinker, T. J. (2022). Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European Journal of Cancer*, *160*, 80–91. https://doi.org/10.1016/j.ejca.2021.10.007

Schreiber, J., Durham, T., Bilmes, J., & Noble, W. S. (2020). Avocado: A multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biology*, *21*(1), 81. https://doi.org/10.1186/s13059-020-01977-6

Stanojevic, S., Li, Y., Ristivojevic, A., & Garmire, L. X. (2022). Computational Methods for Single-cell Multi-omics Integration and Alignment. *Genomics, Proteomics & Bioinformatics*, *20*(5), 836–849. https://doi.org/10.1016/j.gpb.2022.11.013

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J. S., Wilson, N. K., Mende, N., Jardine, L., Gardner, L. C. S., … Haniffa, M. (2021a). Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, *27*(5), Article 5. https://doi.org/10.1038/s41591-021-01329-2

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J. S., Wilson, N. K., Mende, N., Jardine, L., Gardner, L. C., … Haniffa, M. (2021b). *The cellular immune response to COVID-19 deciphered by single cell multi-omics across three UK centres* (p. 2021.01.13.21249725). medRxiv. https://doi.org/10.1101/2021.01.13.21249725

Vandereyken, K., Sifrim, A., Thienpont, B., & Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, *24*(8), Article 8. https://doi.org/10.1038/s41576-023-00580-2