CS690 Final Project
Mosaic single-cell data integration using deep generative models

1st Aniruddh Pramod
Mathematics and Statistics
IIT Kanpur
210142

2nd Advaith Kannan
Biological Sciences and Bioengineering
IIT Kanpur
210072

3rd Debarpita Dash
Biological Sciences and Bioengineering
IIT Kanpur
220328

Abstract—**Integrating diverse single-cell datasets is a significant challenge due to the variability of molecular data modalities and the lack of overlapping features. Contemporary methods such as MaxFuse, MIDAS and scMODAL rely on identifying shared features to integrate different datasets. Among such methods scMODAL, a deep generative framework designed to align heterogeneous datasets in a shared latent space while preserving biological specificity, achieves significantly high label transfer accuracies and effective cross modality feature linking. While scMODAL demonstrates robust performance, it has a critical limitation—it relies on a predefined protein-gene conversion matrix to identify shared features, which constrains its adaptability.**

To address this, we integrated Canonical Correlation Analysis (CCA) into scMODAL to improve its capability to align datasets without predefined feature mappings. The enhanced model, CCA-integrated scMODAL, performed similarly to the original scMODAL and significantly outperformed competing approaches such as MaxFuse, as demonstrated through rigorous benchmarking on multiple datasets, including CITE-seq and TEA-seq. The effectiveness of the proposed approach was consistently evident across datasets, highlighting its potential to enhance single-cell multi-omics integration and enable deeper biological insights.

I. Introduction

The integration of diverse single-cell datasets poses significant challenges, particularly as datasets increase in size and complexity. Current methods typically depend on shared features between datasets, which restricts their scalability and effectiveness. Utilizing deep generative models, such as VAE-based or GAN-based architectures, offers a unified solution to overcome these limitations, enhancing the integration process and supporting more in-depth analyses. Mosaic integration refers to the computational process of aligning and combining datasets that capture distinct molecular modalities from single-cell or spatial omics studies.

II. Literature Survey

A. MaxFuse: Integration across modalities with weak linkage [1]

The input consists of two pairs of matrices. The first pair consists of all features from each modality, and the second pair consists of only the linked features. MaxFuse uses all features within each modality to create a nearest-neighbor graph (all-feature NN-graph) for cells in that modality. Fuzzy smoothing induced by the all-feature NN-graph is applied to the linked features in each modality. Cross-modal cell matching based on the smoothed linked features initializes the iterations.In each iteration, MaxFuse starts with a list of matched cell pairs. A cross-modal cell pair is called a pivot. MaxFuse learns CCA loadings over all features from both modalities based on these pivots. These CCA loadings allow the computation of CCA scores for each cell (including cells not in any pivot), which are used to obtain a joint embedding of all cells across both modalities. For each modality, the embedding coordinates then undergo fuzzy smoothing based on the modality-specific all-feature NN-graphs.The smoothed embedding coordinates are supplied to a linear assignment algorithm which produces an updated list of matched pairs to start the next iteration. After iterations end, MaxFuse screens the final list of pivots to remove low-quality matches. The retained pairs are called refined pivots. Within each modality, any cell that is not part of a refined pivot is connected to its nearest neighbor that belongs to a refined pivot and is matched to the cell from the other modality in this pivot. This propagation step results in a full matching. MaxFuse further learns the final CCA loadings over all features from both modalities based on the refined pivots. The resulting CCA scores give the final joint embedding coordinates.

B. MIDAS:Integration and Knowledge transfer of single cell data [4]

The MIDAS framework utilizes a variational autoencoder (VAE) to model the generation of cell count data and batch information from latent biological and technical noise variables. It employs self-supervised learning to align different data modalities in a shared latent space through joint posterior regularization while leveraging information-theoretic methods to disentangle these latent variables. MIDAS supports reference-to-query knowledge transfer through two strategies: model transfer, which applies a pretrained model for efficient integration of new data, and label transfer, which aligns reference and query datasets in the latent space to enable automatic cell annotation.

C. StabMap:Integration using unshared features [5]

StabMap is a methodology designed for integrating and visualizing data from multiple datasets with varying feature overlaps. It begins by summarizing each dataset using a multi-dimensional topology (MDT), which captures the relationships within the data. Cells from all datasets are projected onto a common reference space, with some cells directly projecting onto this space, while others undergo an intermediate projection before reaching the reference space. This projection process is repeated across selected reference datasets, with optional L2-norm reweighting to adjust the influence of each dataset. The reweighted embeddings are then concatenated to create a unified StabMap embedding, enabling consistent integration of diverse datasets for downstream analysis tasks.

D. UINMF:Integration using non negative matrix factorization [6]

UINMF (Unshared Integrated Non-negative Matrix Factorization) is a matrix factorization strategy designed for integrating heterogeneous datasets with both shared and unshared features. The approach decomposes each dataset into shared metagenes, dataset-specific metagenes derived from shared features, unshared metagenes, and cell factor loadings. The incorporation of the unshared factor matrix U_i allows for the inclusion of features that appear only in one dataset, enhancing the integration of disparate data types. UINMF can integrate various data modalities by leveraging both gene-centric and intergenic features.

E. MultiMAP:Dimensionality reduction and Integration [7]

MultiMAP integrates datasets with varying dimensions by recovering geodesic distances on a shared latent manifold, constructing a MultiGraph on this manifold, and projecting the data into a low-dimensional embedding for analysis and visualization. Key variables include the dataset X_i, points x_{ji}, the shared manifold M, and membership functions μ and ν for the fuzzy simple set. MultiMAP can be applied to integrate data across different omics modalities, species, individuals, batches, and states in cell atlas technologies.

F. scVI:Deep generative modeling for single-cell transcriptomics [8]

scVI is a model based on a hierarchical Bayesian framework, where conditional distributions are defined by deep neural networks, enabling efficient training even on large datasets. The transcriptome of each cell is encoded into a low-dimensional latent vector using a nonlinear transformation, representing normal random variables. This latent representation is then decoded through another nonlinear transformation to estimate the posterior distributional parameters for each gene in each cell. The model assumes a zero-inflated negative binomial distribution, which effectively accounts for overdispersion and limited sensitivity observed in gene expression data.

III. Method Overview

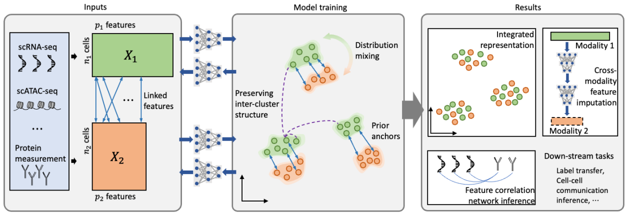

Fig. 1. scMODAL Framework

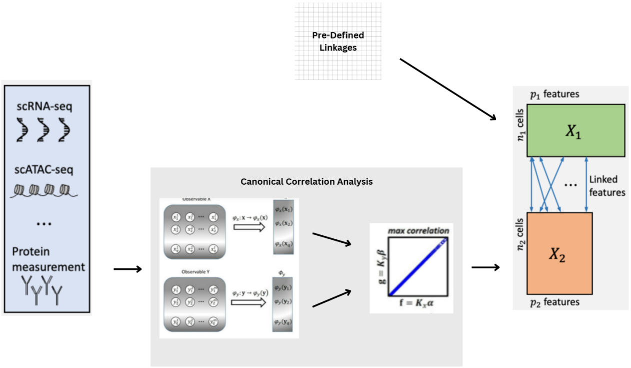

Fig. 2. Integration of CCA to enhance the functionality of scMODAL

scMODAL is a deep generative framework designed to learn integrated cell representations from single-cell multi-omics data. It takes as input cell-by-feature data matrices. Let's consider the case of two datasets, each with different numbers of cells and features. By leveraging prior knowledge about cross-modality relationships, the method identifies linked features and compiles them into new matrices, pairing related features like gene expression (from scRNA-seq) with gene activity scores (from scATAC-seq) or protein abundance with corresponding gene expression.

To handle unwanted variations between modalities, scMODAL uses nonlinear neural network encoders (E1 and E2) to map cells into a shared latent space Z. Unlike methods that only rely on shared features, scMODAL feeds the entire feature matrices (X_1 and X_2) into the encoders to retain biological information. Decoders (G_1 and G_2) reconstruct cell features from the latent embeddings, ensuring autoencoding consistency. To align the latent distributions of the datasets, scMODAL employs a GAN-inspired approach with a discriminator to minimize Jensen-Shannon divergence.

However, blindly aligning distributions can lead to mismatches between distinct cell populations. Since real-world datasets often lack cells measured in both modalities to act as anchors, scMODAL uses mutual nearest neighbors (MNN) between minibatches as a way to guide integration. These MNN pairs, calculated from positively related features, act as priors during training. By adding an L2 penalty on the Euclidean distance between MNN pairs' embeddings, the method aligns datasets while preserving their biological context. This MNN-based alignment is further enhanced by regularizing the geometric structure of each dataset: scMODAL calculates Gaussian kernel distances for each cell relative to others in the minibatch and ensures these geometric representations are preserved.

Once trained, scMODAL provides aligned cell representations, enabling cross-modality analyses. It can map cells from one modality to another (using $E_1(G_2(\cdot))$ and $E_2(G_1(\cdot))$) and perform cross-modality feature imputation. This imputation can then be used to infer regulatory relationships between modalities, revealing deeper insights into the underlying biology.

IV. Results

A. Integration Results on CITESeq-PBMC Data - scMODAL

CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) is a multimodal dataset that integrates two complementary modalities—transcriptomics (scRNA-seq) and epitopes (protein markers)—to provide a detailed molecular profile of individual cells. Specifically:

- **Transcriptomics (scRNA-seq)**: Captures gene expression at the RNA level, providing insights into the actively transcribed genes in each cell.
- **Epitopes (protein markers)**: Measures protein abundance, either on the cell surface or intracellularly.

CITE-seq enables a comprehensive understanding of cellular states and functions by bridging the gap between transcriptomics and proteomics, facilitating detailed cell type annotation and functional characterization.

a) MaxFuse integration: achieved a label transfer accuracy of **86.9%** and a FOSCTTM score of **0.069**

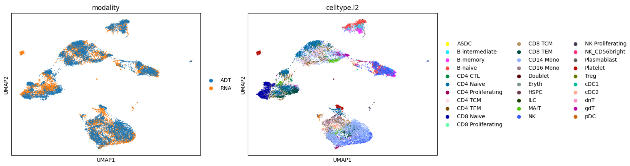

Fig. 3. Integration results on Cite-Seq data- MaxFuse

b) scMODAL integration: of the PBMC dataset resulted in a label transfer accuracy of **97.96%** and a FOSCTTM score of **0.00092**.

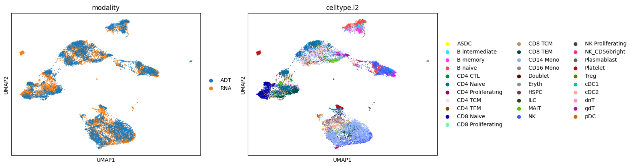

Fig. 4. Integration results on CITE-seq data-scMODAL

B. TEA-seq-PBMC Data Integration Results

TEA-seq is a trimodal dataset that integrates three complementary modalities namely transcriptomics (scRNA-seq), epitopes (protein markers), and chromatin accessibility (scATAC-seq)—to provide a comprehensive molecular profile of individual cells. Specifically:

- **Transcriptomics (scRNA-seq)** captures gene expression at the RNA level, identifying actively transcribed genes in each cell.
- **Epitopes (protein markers)** measure protein abundance, either on the cell surface or intracellularly, using antibodies tagged with unique oligonucleotides.
- **Chromatin Accessibility (scATAC-seq)** maps open chromatin regions, indicating potential regulatory elements such as enhancers and promoters that are accessible for transcription factor binding.

To assess the accuracy of our data integration approach, we calculated a label transfer accuracy of **80.96%**.

The **UMAP plot** in Fig. 2. illustrates the integrated multimodal data, providing a visual representation of the relationships between different cell populations across the transcriptomic, epitope, and chromatin accessibility layers.

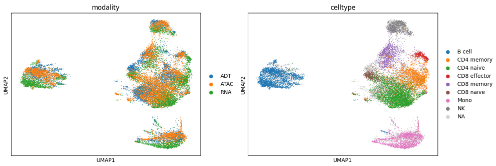

Fig. 5. Integration results on TEA-seq data

V. Method

A. Model Overview

scMODAL starts with two feature matrices, X_1 and X_2, representing single-cell data from two different modalities (e.g., proteomics and scRNA-seq). Since these features typically don't overlap, prior knowledge about cross-modality relationships is used to identify likely-correlated feature pairs. These pairs form new matrices, P_1 and P_2, where each column links related features (e.g., protein abundance with its coding gene expression).

To align the modalities, scMODAL introduces a shared latent space Z, where cells from both datasets are encoded using neural network encoders $E_1(\cdot)$ and $E_2(\cdot)$. The goal is to align the distributions of $E_1(X_1)$ and $E_2(X_2)$ in Z while preserving biological information.

B. Generative Adversarial Learning

Inspired by GANs, scMODAL aligns the distributions in Z using an auxiliary discriminator $D(\cdot)$, which learns to distinguish between embeddings from E_1 and E_2. The discriminator tries to maximize its ability to differentiate, while the encoders minimize this objective, effectively reducing the

Jensen-Shannon (JS) divergence between the two distributions. This is formulated as a minimax problem:

$$\min_{E_1,E_2} \max_{D} \mathcal{L}_{\text{GAN}}$$

C. Autoencoding Consistency

Two decoders, $G_1(\cdot)$ and $G_2(\cdot)$, are trained alongside the encoders to ensure that encoded features can reconstruct the original data. This is done by minimizing the autoencoder loss for both modalities:

$$\mathcal{L}_{\text{AE}} = ||X_1 - G_1(E_1(X_1))||^2 + ||X_2 - G_2(E_2(X_2))||^2$$

D. Aligning Anchors with Linked Features

To guide integration, scMODAL uses linked features P_1 and P_2 to find mutual nearest neighbors (MNNs) between minibatches during training. These MNN pairs act as anchors for alignment. The embeddings of these anchor pairs are encouraged to stay close by minimizing the distance between their latent representations.

E. Preserving Dataset Structure

To avoid losing dataset-specific information, scMODAL regularizes the geometric structure of each dataset. It calculates Gaussian kernel distances between cells in the original feature space and in the latent space Z. By preserving these geometric relationships, the model maintains relative distances between cells within each modality.

F. Training Procedure

The overall training process optimizes a combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{AE}}\mathcal{L}_{\text{AE}} + \lambda_{\text{Anchor}}\mathcal{L}_{\text{Anchor}} + \lambda_{\text{Geo}}\mathcal{L}_{\text{Geo}}$$

where $\lambda_{\text{AE}}, \lambda_{\text{Anchor}}, \lambda_{\text{Geo}}$ are coefficients that balance the contributions of each regularizer.

The networks are trained iteratively using this objective. Once training is complete, the embeddings in Z serve as integrated cell representations. Additionally, the mappings $G_2(E_1(\cdot))$ and $G_1(E_2(\cdot))$ can predict unmeasured features across modalities.

G. Implementation Details

scMODAL employs the Adam optimizer for stochastic optimization during model training. By default, the batch size is set to $\beta = 500$. The optimization process runs for 10,000 iterations with a learning rate of 0.001, coefficients for running averages $\beta_1, \beta_2 = (0.9, 0.999)$, and a weight decay parameter of $\lambda = 0.001$ across all networks. The latent space dimensionality is set to $q = 20$ and the neighborhood size is set to $k = 30$ for identifying MNNs. The regularization parameters are $\lambda_{AE} = 10.0, \lambda_{Anchor} = 1.0$ and $\lambda_{Geo} = 1.0$.

H. Metrics

1) Label transfer accuracy: In the integrated cell embedding space, we transfer labels from one dataset to another dataset based on the nearest neighbor using the euclidean distance. Then we evaluate the ratio of correct transferred labels as the label transfer accuracy. A higher label transfer accuracy indicates a more accurate matching of corresponding cell states.

2) FOSCTTM: FOSCTTM (Fraction of Samples Closer Than the True Match) measures how well the embeddings from two datasets are aligned based on known ground truth pairs.

Given ground truth embeddings z_1^i and z_2^i, we calculate FOSCTTM as -

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\mathbb{1}\left(\|z_1^i - z_2^j\| < \|z_1^i - z_2^i\|\right)$$

$$+\mathbb{1}\left(\|z_1^j - z_2^i\| < \|z_1^i - z_2^i\|\right)\Big] \times \frac{1}{2n^2}$$

A lower FOSCTTM score indicates that the ground truth pairs are closer in the embedding space, reflecting better integration quality.

VI. Data and Code Availability

A. Datasets

a) CITE-seq PBMC dataset: The CITE-seq healthy human PBMC dataset with 228 antibody markers was obtained from Hao et al [3]. for benchmarking. Five batches of 10k cells each were sampled, and the first 15 components of the embedding vectors were used for benchmarking and UMAP visualization. Cell type annotations (8 types at lv1, 20 types at lv2) were taken from Hao et al.'s original data. For antibody dropping, a random forest model was used to predict cell types based on antibodies, and feature importance was calculated to rank the antibodies.

b) TEA- Seq PBMC dataset: The TEA-seq neutrophil-depleted human PBMC dataset (GSM4949911) from Swanson et al [10]. was analyzed using 46 antibody markers and chromatin accessibility data. Cell type annotation was performed with Seurat(v4) WNN clustering, using ADT PCA and ATAC LSI. Eight cell populations (e.g. Naive CD4,Monocyte,B) were identified from 7.4k cells. ADT expressions and gene activity scores were input for MaxFuse and other methods, with LSI features used for ATAC matching. The first 15 embedding components were used for benchmarking and UMAP visualization.

For convenience, code to download these datasets is included in the codebase.

B. Codebase

The codebase can be accessed on Github at https://github.com/atryt0ne/cs690-multimodal-integration. Instructions to reproduce the environments and Kaggle/Colab links are specified in the repository.

VII. Conclusion

In conclusion, the integration of Canonical Correlation Analysis (CCA) into the scMODAL framework significantly enhances its ability to align heterogeneous datasets without the need for predefined feature mappings. The CCA-integrated scMODAL outperforms traditional methods, such as MaxFuse, across various benchmark datasets, including CITE-seq and TEA-seq, demonstrating its robustness and scalability. This advancement in single-cell multi-omics integration paves the way for more accurate cross-modality analyses, offering deeper insights into complex biological systems. The approach holds great potential for furthering our understanding of cellular heterogeneity and multi-omics interactions. Future work will explore additional optimizations and the application of this method to other omics datasets.

VIII. Future Work

A. Robustness

Due to computational constraints, it was not possible to evaluate the method on a variety of datasets. Evaluation of an integration task generally requires loading a lot of embeddings into memory at once which was not feasible on a Kaggle or Colab environment. Additionally the CCA step utilises a Singular Value Decomposition scheme which too is extremely memory intensive for large datasets. Alternatives that are more memory efficient like Randomised SVD exist but come at a steep cost to performance and hence are not appropriate for evaluation. Keeping this in mind, it would be insightful to evaluate these methods on datasets such as **BMC [9]**, which includes human peripheral blood mononuclear cells (PBMCs) profiled using CITE-seq to capture single-cell transcriptomics and protein markers, along with bulk RNA-seq for gene expression validation. Additionally, testing on **CODEX [2]**, which features multiplex imaging data of human tonsil tissues using a panel of 46 antibodies, could provide further valuable insights, given the significantly larger and more complex dataset.

B. Variational CCA

Variational Canonical Correlation Analysis [2] is proposed as an improvement to classical CCA with probabilistic modelling. Traditional CCA is still a linear method and thus a probabilistic version like VCCA can improve on it since it relies on probabilistic modelling which allows modelling a non-linear relationship on the data. Unfortunately due to time constraints this could not be explored in detail.

C. Potential Improvements to Architecture

After further literature review a couple of different improvements to the main architecture of scMODAL can be explored —

The encoder being used in scMODAL can be replaced with a variational autoencoder. It is not trivial to decide what kind of probability distribution could be used for modality integration for a model that is trying to be as modality-independent as possible. Most tasks typically have Protein and RNA modalities so it is likely that we search for a probability distribution that is best for modelling these datasets in particular. Additionally, to investigate this one would need to take care to test on a variety of datasets otherwise we risk over-optimizing for a specific case and not actually improving the model itself.

Style Transfer is a CV technique that allows for the transfer of certain features or styles from one domain to another. This is most relevant to the Feature Imputation aspect of scMODAL.

An attention mechanism can be implemented for the encoder and generator networks. In our survey of the common datasets used in this task we found that datasets usually have far too many features and not quite as many useful. Attention mechanisms are perfect for such data since they can learn the important parts of the data that need to be focused on, especially when there is a lot of data available.

References

[1] Shuxiao Chen, Bokai Zhu, Sijia Huang, John W. Hickey, Kevin Z. Lin, Michael Snyder, William J. Greenleaf, Garry P. Nolan, Nancy R. Zhang, Zongming Ma bioRxiv 2023.01.12.523851; doi: https://doi.org/10.1101/2023.01.12.523851.

[2] W. Wang, X. Yan, H. Lee, and K. Livescu, Deep Variational Canonical Correlation Analysis. 2017. [Online]. Available: https://arxiv.org/abs/1610.03454.

[3] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III., Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integratedanalysis of multimodal single-cell data. Cell, 184(13):3573–3587, 2021. https://www.cell.com/cell/fulltext/S0092-8674(21)00583-3.

[4] He Z, Hu S, Chen Y, An S, Zhou J, Liu R, Shi J, Wang J, Dong G, Shi J, Zhao J, Ou-Yang L, Zhu Y, Bo X, Ying X. *Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS.* Nat Biotechnol. 2024 Oct;42(10):1594-1605. doi: 10.1038/s41587-023-02040-y. Epub 2024 Jan 23. PMID: 38263515; PMCID: PMC11471558. https://pubmed.ncbi.nlm.nih.gov/38263515/.

[5] Ghazanfar, S., Guibentif, C. & Marioni, J.C. Stabilized mosaic single-cell data integration using unshared features. *Nat Biotechnol* **42**, 284–292 (2024). https://doi.org/10.1038/s41587-023-01766-z.

[6] Kriebel, A.R., Welch, J.D. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat Commun* **13**, 780 (2022).https://doi.org/10.1038/s41467-022-28431-4.

[7] Jain, Mika Sarkin, et al. "MultiMAP: dimensionality reduction and integration of multimodal data."*Genome biology* 22.1 (2021): 1-26.

[8] Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15** , 1053–1058 (2018). https://doi.org/10.1038/s41592-018-0229-2.

[9] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128639

[10] Elliott Swanson, Cara Lord, Julian Reading, Alexander T Heubeck, Palak C Genge, Zachary Thomson, Morgan DA Weiss, Xiao-jun Li, Adam K Savage, Richard R Green, Troy R Torgerson, Thomas F Bumol, Lucas T Graybuck, Peter J Skene (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq eLife 10:e63632

[11] Kennedy-Darling, J. et al. Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. Eur. J. Immunol. 51, 1262–1277 (2021).