Single-Cell Cross-Modality Prediction

Chris Chae[1], Zhangyi Sheng[1], Ji Zhang[1], Taishi Nishizawa[1] and Yiyang Nan[1]

[1]Department of Computer Science, Brown University

Abstract

Single-cell measurement technologies have enabled the simultaneous assessment of diverse cellular modalities such as DNA accessibility, RNA, and proteins within a single cell. This advancement offers a direct view into the intricate layers of gene regulation governing biological diversity and disease. The interconnection between these modalities presents challenges; DNA accessibility is fundamental for mRNA production, which, in turn, influences protein synthesis and phenotype. Understanding these intertwined regulatory processes is pivotal for advancements in synthetic biology and drug target discovery. Addressing the demand for predicting one modality from another requires accounting for these complex regulatory interactions, highlighting the crucial need for innovative approaches in analyzing cellular modalities. Our model Multi-modality prediction NET (mmpNET) can represent these modalities into a shared latent space using an autoencoder architecture. We obtained similar results to the baseline and winning models for each task from the OpenProblems 2021 competition using this approach.

Availability: https://github.com/nanyyyyyy/2952G-fall2023.git
Contact: N/A
Supplementary information: N/A

1 Introduction

Empowered by deep learning, more innovative models are emerging that enable the measurement of multiple modalities within the same single cell. Gaining insights into the flow of information in the cell holds immense significance for the field of synthetic biology and drug development. It is crucial to recognize that these cellular traits are not independent but intricately connected. Specifically, the production of mRNA, a critical step in gene expression, relies on the accessibility of chromatin DNA (captured by ATAC data). Moreover, this intricate process of genetic expression is often subject to regulation by the molecules it generates. For instance, certain proteins may exert control by binding to DNA, thus obstructing the generation of additional mRNA. The growing demand for comprehensive multi-modal data (Athayaet et al., 2023, Efremova et al., 2020, Gossi et al., 2023) underscores the critical importance of addressing these regulatory complexities to advance our understanding of cellular processes.

In biology, most measurements of cellular systems deal with populations of cells. For example, a tumor sample may contain cancerous cells as well as skin cells, blood cells, benign cells, and so forth. Although each of these cells provides different bodily functions, they all contain the same genome consisting of billions of nucleotides. The turning on and off of the genes coming from the genome is what differentiates the cells. Single-cell measurements examine specific individual cells rather than that of a culture consisting of many different cells, providing valuable information on the modalities of each cell. Traditional sequencing methods average the gene expression across multiple cells in a population, which can mask the traits of individual cells. Some modalities commonly measured are chromatin interactions, transcription of RNA, promoter-enhancer receptor binding, and 3D spatial dimension of DNA/RNA/proteins. However, because collecting single-cell measurements is a resource-intensive task, single-cell datasets tend to be more sparse compared to their counterparts, leading to large amounts of noise that interfere with analyses. In addition, these datasets are represented differently from each other, meaning it is difficult to relate the datasets together directly. The importance of single-cell measurements comes from gene expression of individual cells and understanding how the concentration of DNA/RNA/proteins affects health and the onset of disease. Understanding the flow of information between different cells that controls gene and protein production from the DNA sequence is vital to answering questions about genetic diseases.

Measuring modalities accurately means accounting for the effects of the gene regulatory process, where genes can be turned on or off from a protein coming from a faraway region of the chromosome, or even on the same sequence. There are even more factors that affect gene expression, such as transcription activity, chromatin accessibility and interactions, and spatial structure of DNA. At the single-cell level, researchers can obtain insights into individual cells rather than generalizing the data for entire cell populations.

Therefore, the paper's objective is to design a comprehensive model that predicts the flow of information from DNA to RNA and from RNA to protein by combining different datasets. The dataset has three types of input data: GEX, ATAC-seq, and ADT, which were acquired from samples of bone marrow mononuclear cells of donors of diverse backgrounds and genetic histories. The 2021 NeurIPS challenge (Lance et al., 2021) provides an updated comprehensive dataset for our task.

GEX data provides information about the gene expression levels in a particular cell or tissue at a specific time, which is crucial for understanding which genes are actively being transcribed into RNA molecules. ATAC-seq data provides information about the accessibility of chromatin, which is the complex of DNA and proteins that make up chromosomes, helping to identify regions of the genome that are open and accessible to transcription factors and other regulatory proteins. Finally, ADT data provides information about the abundance of specific proteins or protein modifications within a cell, which is crucial for understanding the protein products that are translated from RNA and for studying post-translational modifications that can affect protein function. With the combination of these datasets, researchers can better understand and be able to predict the process by which genotypes become phenotypes.

Specifically, we expect to optimize prediction from GEX to ATAC, ATAC to GEX, GEX to ADT, and ADT to GEX, in addition to maintaining performance on average across different modalities. The flow of information can be analogized to the task of machine translation, which requires translating information between different layers of gene regulation. The complex processes of transcription and translation in molecular biology are governed by specific rules, which, comparable to the rules of grammar and syntax, govern language translation in the machine translation task. For example, RNA translation to protein is comparable to translating a text from one language to another, where the sequence of codons (triplets of nucleotide bases) on the mRNA strand determines the sequence of amino acids in a protein. Similarly, in the translation task, the conversion of words from one language to another

must convey the intended meaning accurately. Therefore, we intend to apply the encoder-decoder architecture to predict the flow of information from DNA to RNA and then from RNA to protein, adapting the principles of sequence-to-sequence learning commonly used in machine translation tasks.

2 Related work

In this section, we will delve into the examination of different approaches for modality prediction. Our focus will be on scrutinizing the architectural aspects of these models and exploring potential enhancements for their effectiveness.

2.1 Graph Structure

The overall winning solution for predicting across all multimodal tasks in the competition uses a graph neural network (Scarselli et al., 2008). The method ran their bipartite graph constructed between modalities (ATAC, GEX, ADT) and the cells through neural networks that were implemented for single-cell modality prediction (ScMoGCN, SCMM, CMAE) (Wen et al., 2022). The graph is created so that each node can be either a cell or a gene and only gene-cell edges exist since the edge weights are determined by the gene counts. After graph creation and convolution, the embeddings of the nodes for each convolutional layer were concatenated and transformed linearly to represent the connections between modalities. DANCE (Python package *pydance*) (Ding et al., 2020) is a modality prediction pipeline that has many built-in deep learning models and performed the best overall at all subtasks in the competition. In our methods, we implement their model using DANCE, which predicts between scATAC-seq and GEX at the single cell level, (Wu et al., 2021).

2.2 Autoencoder Neural Network

BABEL (Wu et al., 2020) is made up of two autoencoder networks that project ATAC or RNA modalities onto a 16-dimensional latent space and infer the corresponding modality. The purpose of creating latent representations of modalities is to obtain as much significant cellular variation among single cells, better capturing the phenotypes that arise from certain patterns. After training on multiple human cell types combined, BABEL was found to have higher Pearson and Spearman correlation values than existing tools in addition to KNN clustering for both pairings of modalities. In addition, BABEL was evaluated on a withheld nonhuman cell line and a nonhuman single cell line. The model obtained an auROC score of 0.80 for RNA to ATAC and a Pearson correlation of 0.55 for ATAC to RNA, which are similar to the scores for the human datasets. In addition, BABEL identified the expression of biologically relevant gene markers present in specific cell types and used them for modality prediction for different cell types. In the reverse direction going from RNA gene expression to scATAC-seq, BABEL highlighted several DNA element regions important to the gene regulatory process.

The winning model for the ADT to GEX modality prediction was from team Novel which implemented an autoencoder architecture that transformed the input modalities ATAC and GEX using latent semantic indexing (LSI) while keeping ADT data as is for training. The Optuna framework was used for hyperparameter tuning search for each of the four tasks (GEX to ATAC, ATAC to GEX, ADT to GEX, GEX to ADT).

2.3 Residual Neural Network

ResNet (He et al ., 2016) is another commonly used architecture for cell line cross-modality prediction and many other biomedical tasks. ResNet is used as a backbone in many winning solutions in the competition and it achieved the best performance results with the modality prediction of GEX to ATAC. It usually consists of multiple fully connected layers with Relu activation and batch normalization at each layer. At each layer, the model has residual connections by applying the fully connected layers interleaved with ReLU activation and batch normalization. The model employs a residual architecture, incorporating skip connections to help with the flow of gradients during training, and also includes a batch classifier for batch-wise classification.

3 Methods

mmpNet aims to employ a multi-modal prediction architecture to handle multiple inputs and output types. It has a single solid architecture that can predict ATAC, GEX, and ADT from one to the other (Figure 1 and Figure 7 in supplement) .

mmpNet consists of two parts: dimension reduction and model training and testing. In the preprocessing part, the model includes normalization, low-dimensional project, batch effect correction in a transductive setting, and autoencoder. The training part is a fully connected-layers neural network. The evaluation method for our model is the Root Mean Square Error (RMSE) function.

3.1 Data Dimension Reduction

The first component in the data preprocessing section is the LSI Transformer (Deerwester et al., 1990) with the TF-IDF matrix. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure widely employed in information retrieval and natural language processing to capture the significance of a term within a document in relation to a corpus of documents. The relationship of TF-IDF with modality prediction is explained in the supplementary part.

Upon obtaining the TF-IDF matrix, mmpNet integrates the TF-IDF methodology with Latent Semantic Indexing (LSI) to enable the extraction of underlying relationships and patterns from diverse gene data, particularly in instances where the gene data is both high-dimensional and sparse. Singular Value Decomposition (SVD) can be applied to the TF-IDF matrix to deconstruct it into three distinct matrices, with the middle matrix capable of capturing the latent connections between terms and documents. Retaining the most crucial dimensions can reduce data dimensionality while preserving the most noteworthy patterns and relationships, thereby mitigating challenges associated with high dimensionality and sparsity in the data. By leveraging semantic analysis and interpretation, hidden patterns associated with gene regulatory networks and chromatin accessibility can be unearthed. Consequently, the grouping of genes or samples based on their semantic similarities can facilitate a more comprehensive comprehension of the underlying biological mechanisms and relationships.

The second component in the data processing section is batch correction with Harmony (Korsunsky et al, 2019). Batch effects can happen in many cases like changes in experimental conditions, equipment variations, or other non-biological factors. Batch effects can significantly impact the accuracy and interpretability of downstream analyses of genomics data. In the absence of batch correction, batch effects can lead to misleading results, particularly in the identification of differentially expressed genes, clustering of samples, and other analyses.

Harmony is an established batch correction method designed specifically for single-cell RNA sequencing (GEX) data. The test data batches are independent of the training data batches, correcting the bath effects in a transductive setting minimizes the impact of the distribution shifts that may arise due to technical variations across different batches or acquisitions.

mmpNet uses an autoencoder neural network following with a fully connected layer to learn the relationship between modalities. The autoencoder layer's objective is to learn a compact and efficient representation of input data by comprising an encoder and a decoder. The encoder transforms the input through two linear layers, employing rectified linear unit (ReLU) activations and dropout for non-linearity and regularization, respectively. The first linear layer maps the input data to a hidden layer with 512 units. The second linear layer maps the hidden layer to the encoding layer with the specified dimensionality. The encoded representation is then decoded by a symmetrical structure, with the decoder aiming to reconstruct the original input. The first linear layer

maps the encoding layer back to a hidden layer with 512 units. The second linear layer maps the hidden layer to the output layer with the specified dimensionality. The entire network is trained to minimize the reconstruction error, encouraging the autoencoder to capture essential features of the input data in the reduced-dimensional encoding.

Following the preprocessing of the entire dataset, the training, validation, and testing data have been transformed into binary format, incorporating extracted features. While alternative models, like Cajal, employ the binarize package from Sklearn for this conversion, our experimental findings indicate that our unique preprocessing approach yields superior results.

3.2 Model Training

mmpNet learns the output features from the autoencoder layer with a three-layered fully connected neural network. The architecture consists of linear layers with GELU activations and dropout layers, providing non-linearity and regularization. The overall architecture is the same across all modality predictions but with different layer unit sizes and different dropout probabilities. Take GEX to ATAC as an example, the first linear layer maps the input data to a hidden layer with 1024 units; followed by a second linear layer which maps the 1024 units to 256 units; the third layer maps the 256 units to 2048 units. Dropout layers with specified dropout probabilities of 0.3, 0.15, and 0.15, respectively.

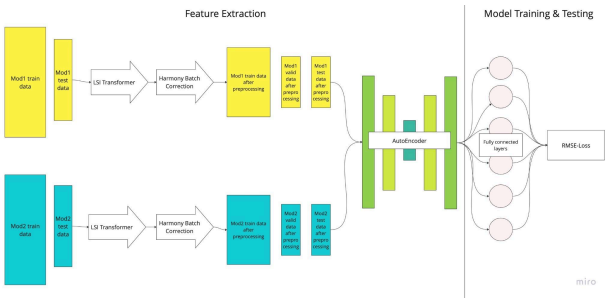

Fig.1 mmpNet Model Structure

3.3 Evaluation

To align with other models' evaluation method from the Open Problem in Single-Cell Analysis 2021 competition, the Root Mean Squared Error (RMSE) is applied to evaluate mmpNet. RMSE (Equation 1) measures the average magnitude of the errors between predicted and actual values, providing a comprehensive understanding of the model's performance. RMSE provides a unified evaluation metric for all four modality predictions in the same scale.

$$RMSE = \sqrt{\Sigma_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Equation 1: Root Mean Square Error

4 Experiments

The model input is the pre-processed dataset from the benchmarking dataset for the Open Problem in Single-Cell Analysis 2021 competition; the original dataset is available from NCBI GEO under accession GSE194122. The training data is in an AnnData h5ad file. The competition is organized in two phases. In the first phase, participants develop methods and submit their solutions. The challenge organizer then re-trains the submitted solutions via Docker, using a separate phase 2 dataset. We use the phase 2 private dataset to develop and tune our models for simplicity and comparison purposes. Therefore, it is unfair to directly compare them with the official solutions.

For each method, 5-fold cross-validation is used to reduce data splitting variance, assess and tune the model's parameters. By iteratively rotating the validation set through the folds, we get a comprehensive evaluation of our model's generalization capabilities. However, there isn't any major change during this process which indicates the data is evenly distributed.

In Figures 2 and 3, the visualization samples of the ATAC and GEX datasets respectively with UMAP after normalization and log1p transformation. The ATAC and GEX datasets both had 42492 total observations and 116490 and 13431 features, respectively.

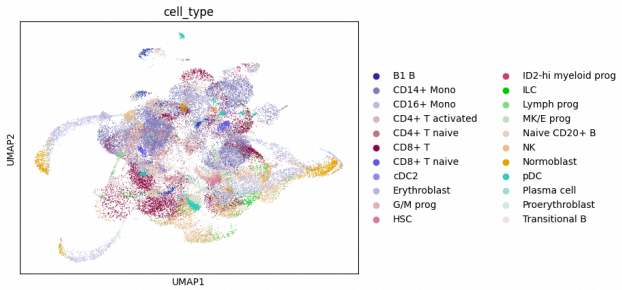

Figure 2: UMAP of a representative subset of the open chromatin (ATAC-seq) dataset clustered by cell type

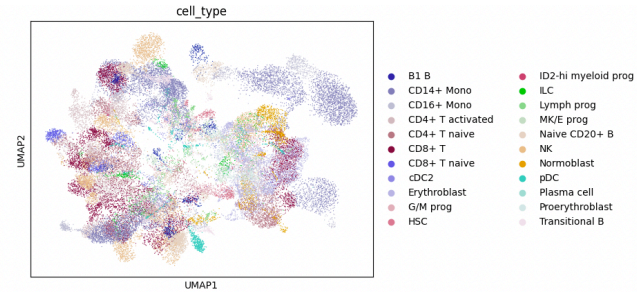

Figure 3: UMAP of a representative subset of the gene expression (GEX) dataset clustered by cell type

UMAPS are latent representations of the data where features are reduced to a lower dimension for clustering, where each point is a cell and the distance between two points indicates their similarity.e.g In Figures 2 and 3, it is difficult to distinguish between cell types from chromatin and RNA alone. This indicates that the models in the competition needed to employ a novel approach to predicting modalities without memorizing patterns among cell types.

The analysis part uses data grouped in batches, or donor sites where the DNA, RNA, and protein abundance is measured across the same cells. UMAPS can visualize the clusters in the training and testing datasets for RNA and protein. But, the GEX dataset is normalized in addition to performing log1p transformation.

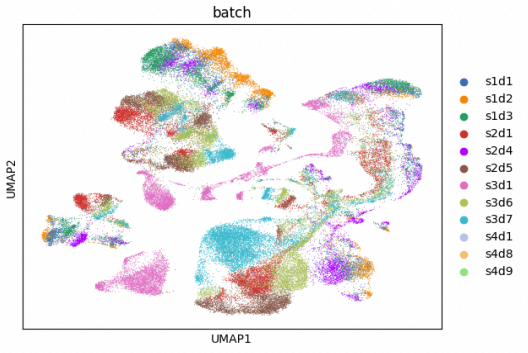

Figure 4: UMAP of full GEX training dataset clustered by batch/donor

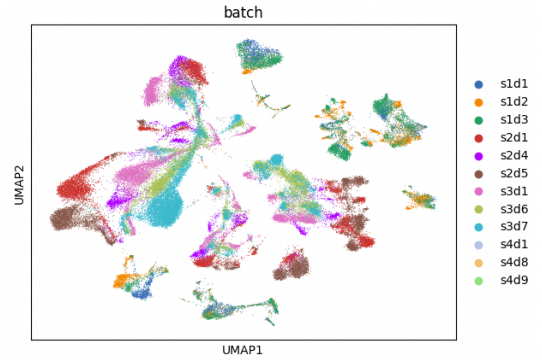

Figure 5: UMAP of full ADT training dataset clustered by batch/donor

In Figures 4 and 5, notice that most of the cells in each batch are somewhat similar to each other, which is what is expected since each batch comes from the same sample of cells. However, the batches are not equally represented in each cluster and each cluster is not clearly defined, meaning the training data may contain biases and cause the model to overgeneralize and make incorrect predictions.

5 Results

In this study, we employed two baseline methods that were also used in the 2021 competition. The first method returns the mean of modality 2 training data for all cells in the test set of modality 1. In this case, the prediction for each cell is identical and is solely based on the average behavior observed in the training data of modality 2. However, this method neglects any specific information provided by modality 1, as the prediction is the same for all cells in the test set. It serves as a straightforward baseline that does not take into account the individual characteristics of cells in modality 1. The second method employs linear regression on Principal Component Analysis (PCA)-transformed data from modality 1. In this approach, the input data from modality 1 is concatenated for both training and test sets. PCA is then applied to reduce the dimensionality of the data. The resulting PCA-transformed data is used to train a linear regression model, and predictions are made for the test set. The final step involves projecting these predictions back into the original feature space of modality 2. This method leverages the relationship between modality 1 and modality 2 through linear regression on the PCA-transformed space.

In addition to the competition baselines, we also compared mmpNet to simplified KAUST (ResNet) and vanilla GNN approaches, which were backbones of two winning solutions from the competitions that we found most related to our course materials. This comparison was conducted for each of the modality prediction tasks using the default hyperparameters. As shown in Figure 6 and Table 1 (sup. info.), our model performed similarly to the others in terms of RMSE loss across 3 of the 4 prediction tasks (GEX2ATAC, ATAC2GEX, ADT2GEX) and slightly underperformed in the GEX2ADT task compared to the winning models. RMSE loss was calculated by the equation shown in Equation 1.

The competition measured predictive performance by calculating the difference between the truth and predicted values for each modality task respectively. These values differed between each task due to the varying dimensionality of the DNA, RNA, and protein datasets. Each model's goal was to minimize the RMSE loss to zero.

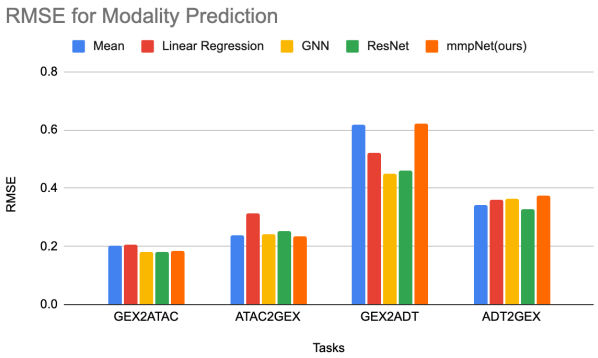

Figure 6: RMSE for the models across each modality prediction task

6 Discussion and Conclusion

Single-cell technology has led to immense breakthroughs in biology and furthered our understanding of cellular systems. Using deep learning techniques, we can align different modalities and help researchers find relationships between the transition between the transition of DNA to RNA and to protein. Although biologists understand some of the connections between these layers of the regulatory process, we use an encoder model to develop insights on the underlying characteristics of each modality to predict a different one.

mmpNet combines batch correction with an autoencoder architecture to predict one single cell modality from another by reducing the dimensionality of DNA, RNA, and protein datasets into a shared latent space. mmpNet not only performs as well as the winning models for some of the tasks, but also contains an interpretable latent space that can be used for further analysis on the biological importance of the elements of the gene regulatory process.

However, the results are lower than our initial expectations. We attribute this outcome to several factors. Firstly, the dataset from the competition appears to have inherent flaws. Additionally, the cell line data exhibits high dimensionality, and the signal strength or feature is notably weak. Consequently, dimension reduction becomes a necessary step for all methods employed. This circumstance partially explains why even traditional machine learning methods establish a robust baseline in this scenario. The inherent limitations exist in transferring information between certain modalities, such as the natural difficulty in transferring GEX data to ADT data. When we performed data exploration, we noticed that ATAC data is highly sparse and binarized and the scales of gene expression data differ by sites, which may limit our model's generalizability of predictions to unseen test data. In addition, further investigation revealed that some batches were not represented in the testing datasets, which could mean our model needed to be trained on more varied labels.

We are also aware of the limitations of deep learning methods in studying single cells. When the data is highly sparse and noisy, a nonparametric machine learning approach would provide more robust results. We realize the difficulty of creating a universal end-to-end method for all cross-modality predictions.

Contributions

All team members contributed equally to this final project. We updated our work distribution before each checkpoint. In the first stage, Chris implemented the DANCE workflow for multimodality prediction and created functions to obtain RMSE results from the BABEL model. Zhangyi created a subsampled dataset, explored the baseline linear regression model, and experimented with TF-IDF and SVD using linear regression and autoencoder structures. Anna attempted to reimplement the Novel architecture and tested the methods with different hyperparameters to observe their effects. Taishi assisted with initial experiments and baseline selection, while Yiyang focused on data preprocessing and exploring data analysis. In the second stage, Chris implemented DANCE for full datasets and collaborated with Taishi on data visualization. Taishi implemented the ResNet-based model for full datasets. Zhangyi constructed the LSI-autoencoder architecture with full datasets and experimented with different feature identification approaches on our model. Yiyang made modifications and improvements to both baselines and our models, running them on the GPU to obtain results. Anna introduced the Harmony approach to our model and generated graphs for the model architecture. Finally, we summarized all the feedback from peer reviewers and collaboratively made improvements to the final draft.

References

1. Athaya, T., Ripan, R., Li, X., Hu, H. (2023). Multimodal deep learning approaches for single-cell multi-omics data integration, Briefings in Bioinformatics, Volume 24, Issue 5, September 2023, bbad313.

2. Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0

3. Efremova, M., Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. Nature methods, 17(1), 14-17.

4. Gossi, F., Pati, P., Chouvardas, P., Martinelli, A. L., Kruithof-de Julio, M., Rapsomaniki, M. A. (2023). Matching single cells across modalities with contrastive learning and optimal transport. Briefings in bioinformatics, 24(3), bbad130.

5. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., ... Satija, R. (2020). Integrated analysis of multimodal single-cell data. bioRxiv.

6. Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., ... Krishnaswamy, S. (2022). Multimodal single cell data integration challenge: Results and lessons learned. Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track, 176, 162-176.

7. Liu, Q., Chen, S., Jiang, R., Wong, W. H. (2020). Simultaneous deep generative modeling and clustering of single cell genomic data. bioRxiv.

8. Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., ... Bloom, J. M. (2021). A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

9. Wen, H., Ding, J., Jin, W., Wang, Y., Xie, Y., Tang, J. (2022). Graph neural networks for multimodal single-cell data integration. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

10. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

11. Wu, K. E., Yost, K. E., Chang, H. Y., Zou, J. (2020). BABEL enables cross-modality translation between multi-omic profiles at single-cell resolution. bioRxiv.

12. Ding, J., Wen, H., Tang, W., Liu, R., Li, Z., Venegas, J., … Tang, J. (2022). DANCE: A Deep Learning Library and Benchmark for Single-Cell Analysis. bioRxiv.

13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

14. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks*, *20*(1), 61-80.

Supplementary Information

Model	ATAC2GEX	GEX2ATAC	GEX2ADT	ADT2GEX
Mean	0.2031	0.2394	0.6175	0.3430
Linear Regression	0.2406	0.3130	0.5213	0.3585
BABEL	0.1817	0.2409	0 0.4489	0.3654
KAUST	0.1793	0.2516	0.4613	0.3268
mmpNet	0.1850	0.2343	0.6204	0.3731

Table 1: RMSE results from testing

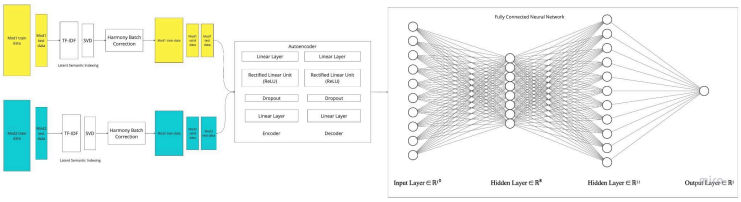

Figure 7: mmpNet Architecture

TF-IDF and Modality Prediction

In the realm of modality prediction, such as the prediction of ATAC data from gene expression (GEX), TF-IDF can serve as a feature engineering tool to capture the interplay between gene expression data and ATAC data. It helps in the identification of relevant genes likely to be associated with specific ATAC data patterns. Each sample or data point can be regarded as a 'document,' and the gene expression data can be considered as 'terms' within these 'documents.' The frequency of each gene in the gene expression data for a specific sample can be calculated to represent the degree to which a gene is expressed in that particular instance, thereby encapsulating the gene's importance within the sample. Subsequently, the inverse document frequency for each gene across the entire dataset can be computed to capture the uniqueness of a gene across all samples. Genes that exhibit lower prevalence across all samples but are highly expressed in specific samples are assigned higher importance. By multiplying the TF and IDF, a TF-IDF score for each gene in each sample can be derived, signifying the gene's significance for that specific sample while accounting for its relevance across the entire dataset. The TF-IDF methodology can yield valuable insights into the association between gene expression and ATAC data.