Integrated Deep Learning Analysis of scRNA-seq and TCR Repertoire Data using Multimodal Variational Autoencoders

Context

This report presents a comprehensive analysis of the integration of single-cell RNA sequencing (scRNA-seq) and T-cell receptor (TCR) repertoire data using a novel Multimodal Variational Autoencoder (MVAE) framework. The report covers the following key components:

1. Introduction and motivation for integrating scRNA-seq and TCR data
2. Literature review of existing approaches for multi-modal data integration
3. Detailed mathematical methodology of the MVAE framework
4. Experimental results including model training, latent space analysis, and biological findings
5. Interpretation of results in terms of T-cell biology and subpopulation discovery
6. Future directions and potential applications of the approach

The work addresses fundamental challenges in understanding T-cell biology by creating a unified analytical framework that captures both cellular state and antigen specificity information.

Abstract

Single-cell RNA sequencing (scRNA-seq) and T-cell receptor (TCR) sequencing technologies provide complementary views of T-cell biology, capturing cellular functional states and antigen specificity, respectively. However, integrating these heterogeneous data modalities to discover emergent biological patterns remains challenging. Here, I present a Multimodal Variational Autoencoder (MVAE) framework specifically designed to create a unified latent representation that captures information from both data types. This approach leverages recent advances in deep representation learning to model the joint distribution of these complex data types. Applied to a T-cell dataset, the MVAE reveals distinct subpopulations that are not discernible when analyzing each modality separately. One subpopulation exhibits regulatory T-cell (Treg) markers (Foxp3, Il2ra) with characteristic TCR α-chain preferences, while the other shows a conventional T-cell profile with TCR β-chain dominance. This approach enables bidirectional prediction between modalities and provides an interpretable latent space that captures biologically meaningful relationships. My work demonstrates the value of multimodal integration in uncovering hidden cellular heterogeneity and establishing connections between gene expression patterns and TCR characteristics in T-cell populations.

1. Introduction and Motivation

Single-cell RNA-sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by providing genome-wide expression profiles at single-cell resolution. Concurrently, T-cell receptor (TCR) sequencing allows for precise characterization of the adaptive immune repertoire, capturing the specificity of T-cells for recognizing antigens. While these technologies independently provide valuable insights into cellular function and antigen specificity, their integration presents both a significant challenge and an opportunity to reveal deeper biological relationships that would otherwise remain hidden.

The fundamental challenge addressed in this project is the integration of these two high-dimensional, heterogeneous data modalities (scRNA-seq and TCR-seq) to discover emergent patterns and relationships that are not apparent when analyzing each modality in isolation. Traditional analytical approaches often fail to capture the complex, non-linear relationships between gene expression patterns and TCR characteristics. This limitation is especially pronounced when studying T-cell populations, where both the functional state (reflected in gene expression) and antigen specificity (encoded in TCR sequences) are critical for understanding immune responses.

The motivation for this integration is multi-faceted:

1. **Biological Completeness**: Gene expression (scRNA-seq) captures the functional state and cellular phenotype, while TCR sequences represent antigen specificity and clonal relationships. Combining these offers a more complete view of T-cell biology.
2. **Hidden Subpopulation Discovery**: Subtle T-cell subpopulations might only become apparent when considering both functional state and receptor specificity simultaneously.
3. **Cross-Modal Relationships**: Certain TCR features might correlate with specific gene expression patterns, revealing relationships between antigen recognition and cellular function.
4. **Predictive Potential**: A successful integration could enable predicting one modality from the other, for instance, inferring functional states from TCR sequences or vice versa.

To address this challenge, I developed a Multimodal Variational Autoencoder (MVAE) framework specifically designed to create a unified latent representation space that captures information from both scRNA-seq and TCR-seq data. This deep learning approach leverages recent advances in representation learning to model the joint distribution of these complex data types, enabling the discovery of patterns that emerge only at their intersection.

2. Literature review

The field of multi-modal data integration in computational biology has seen significant advancements, especially with the emergence of deep learning approaches. Several key models and methodologies have influenced this work:

Single-Cell Multi-modal Integration Models

scNAT (Zhou et al., 2023) is a notable approach that integrates scRNA-seq and TCR sequencing data for T-cell trajectory analysis. It employs a VAE architecture with a CNN component specifically designed for processing CDR3 sequences and embeddings for V/J genes. While scNAT effectively creates a unified latent space and has been successful in detecting migration trajectories in T-cells, its application has been primarily demonstrated on multiple sclerosis datasets, potentially limiting its generalizability to other contexts.

MIST (Chen et al., 2024) takes a different approach through its Multimodal Integration via Self-supervised Transformer architecture. MIST employs a modular VAE with Performer Attention and Domain-Specific Batch Normalization (DSBN) to disentangle modality-specific signals. It incorporates Maximum Mean Discrepancy (MMD) loss to align distributions in the latent space. While powerful in separating shared and modality-specific information, MIST's architectural complexity introduces computational challenges and potential difficulties in interpretation.

totalVI (Gayoso et al., 2021) focuses on integrating scRNA-seq with protein expression data from CITE-seq experiments. Though not directly addressing TCR data, totalVI established important principles for multi-modal integration in single-cell analysis, including uncertainty modeling and batch correction. However, totalVI's design is not optimized for sequence-based data types like TCR sequences, limiting its direct application to the present challenge.

Variational Autoencoders in Computational Biology

Beyond specific multi-modal models, VAEs have been widely used for dimensionality reduction and representation learning in single-cell genomics. Models like **scVI** (Lopez et al., 2018) demonstrated the effectiveness of variational inference for modeling technical variability in scRNA-seq data. Similarly, **DeepTCR** (Sidhom et al., 2021) applied deep learning to TCR sequences for encoding antigen specificity.

Comparative Analysis with Current Approach

In comparison to these existing models, the MVAE approach developed in this project offers several advantages:

1. It employs a dual-encoder VAE design with a shared latent space specifically optimized for RNA-TCR integration, providing a more streamlined architecture than MIST while maintaining the power to learn cross-modal relationships.
2. Unlike scNAT, which was designed with trajectory analysis in mind, the current model focuses on discovering distinct cellular subpopulations and their characteristics.
3. The approach enables bidirectional prediction between modalities (RNA → TCR and TCR → RNA), allowing for inference of one data type from the other.
4. It provides an interpretable latent space that can be analyzed to understand which dimensions contribute most to separating cell populations.

The primary limitations of the current approach include its requirement for paired data (cells with both RNA and TCR measurements) and the relatively basic CDR3 encoding strategy compared to more sophisticated sequence models. These limitations represent opportunities for future improvements.

3. Methodology

3.1 The MVAE Framework: Mathematical Foundations

The Multimodal Variational Autoencoder (MVAE) extends the traditional VAE framework to handle multiple data modalities. The core idea is to learn a joint probability distribution over the observed data from different modalities and a shared latent representation.

For two data modalities x_RNA and x_TCR, the goal is to model the joint distribution p(x_RNA, x_TCR) by introducing a shared latent variable z. The generative process can be described by:

p(x_RNA, x_TCR, z) = p(z)p(x_RNA|z)p(x_TCR|z)

where p(z) is the prior distribution over the latent space (typically a standard Gaussian N(0, I)), and p(x_RNA|z) and p(x_TCR|z) are the conditional distributions of each modality given the latent representation.

The Evidence Lower Bound (ELBO)

Since direct optimization of the marginal likelihood p(x_RNA, x_TCR) is intractable, the MVAE is trained by maximizing the Evidence Lower Bound (ELBO):

L(θ, φ; x_RNA, x_TCR) = E_q_φ(z|x_RNA, x_TCR)[log p_θ(x_RNA|z) + log p_θ(x_TCR|z)] - D_KL(q_φ(z|x_RNA, x_TCR) || p(z))

where:

- q_φ(z|x_RNA, x_TCR) is the approximate posterior distribution (encoder)
- p_θ(x_RNA|z) and p_θ(x_TCR|z) are the conditional distributions (decoders)
- D_KL is the Kullback-Leibler divergence

Modality-Specific Encoders and Joint Representation

To handle the distinct properties of each data modality, the MVAE employs separate encoders for RNA and TCR data. The variational posterior is then defined as a combination of the individual posterior distributions:

$q_\varphi(z|x_{RNA}, x_{TCR}) = q_{\varphi\_RNA}(z|x_{RNA}) \cdot q_{\varphi\_TCR}(z|x_{TCR})$

In practice, for computational efficiency, I implement a product-of-experts approach where the combined posterior is:

$\mu_z = (\mu_{RNA} + \mu_{TCR})/2$

$\sigma_z^2 = (\sigma_{RNA}^2 \cdot \sigma_{TCR}^2)/(\sigma_{RNA}^2 + \sigma_{TCR}^2)$

where μ_{RNA}, σ_{RNA}^2 and μ_{TCR}, σ_{TCR}^2 are the mean and variance parameters of the individual posteriors.

Reparameterization Trick

To enable backpropagation through the sampling process, the reparameterization trick is employed:

$z = \mu_z + \sigma_z \odot \epsilon$, where $\epsilon \sim N(0, I)$

This allows for efficient gradient computation during training.

Loss Function

The total loss function combines the reconstruction losses for both modalities with the KL divergence term:

$L_{total} = L_{RNA} + L_{TCR} + \beta \cdot D_{KL}$

where:

- $L_{RNA} = E_{q_\varphi(z|x_{RNA}, x_{TCR})}[\log p_\theta(x_{RNA}|z)]$ is the RNA reconstruction loss
- $L_{TCR} = E_{q_\varphi(z|x_{RNA}, x_{TCR})}[\log p_\theta(x_{TCR}|z)]$ is the TCR reconstruction loss
- β is a hyperparameter that controls the weight of the KL divergence term

For the RNA data, which is continuous, the reconstruction loss is implemented as the mean squared error (MSE):

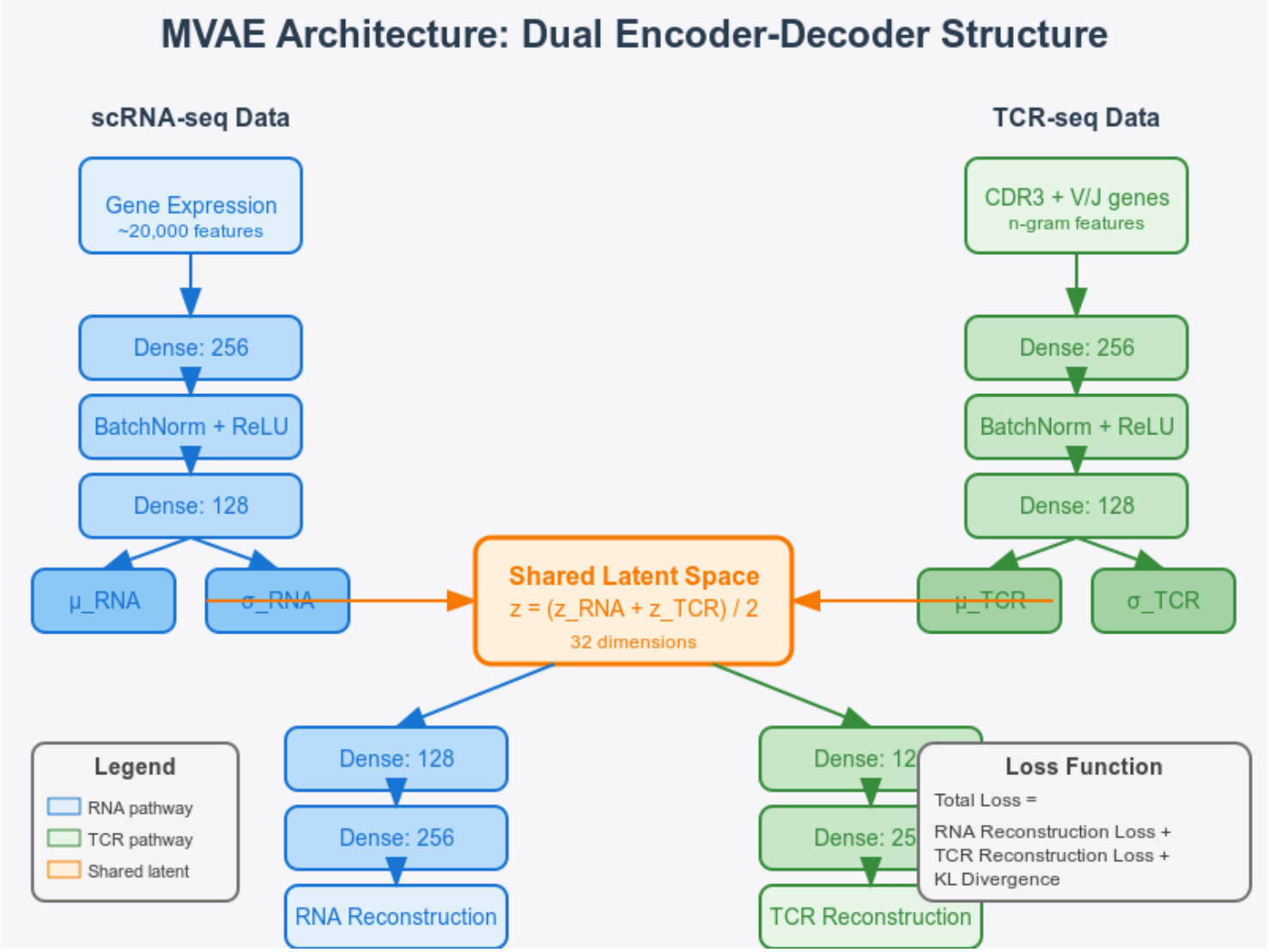
$L_{RNA} = (1/N) \sum_{i=1}^{N} (x_{RNA,i} - \hat{x}_{RNA,i})^2$

For the TCR data, which is converted to a binary n-gram representation, the reconstruction loss is also implemented using MSE:

$L_{TCR} = (1/M) \sum_{i=1}^{M} (x_{TCR,i} - \hat{x}_{TCR,i})^2$

where N and M are the dimensions of the RNA and TCR features, respectively.

3.2 Model Architecture



The MVAE architecture consists of several key components designed to handle the specific challenges of scRNA-seq and TCR data integration:

Dual Encoder Design

RNA Encoder:

- Input: High-dimensional gene expression matrix (~20,000 features)
- Architecture: Dense(256) → BatchNorm → ReLU → Dense(128) → BatchNorm → ReLU → Dense(32×2)
- Output: Mean and log-variance vectors for the latent distribution

TCR Encoder:

- Input: N-gram representation of TCR sequences (100 features)
- Architecture: Dense(256) → BatchNorm → ReLU → Dense(128) → BatchNorm → ReLU → Dense(32×2)
- Output: Mean and log-variance vectors for the latent distribution

Shared Latent Space

- Dimensionality: 32
- Distribution: Gaussian with learned mean and variance
- Sampling: Using the reparameterization trick

Dual Decoder Design

RNA Decoder:

- Input: Latent vector (32 dimensions)
- Architecture: Dense(128) → BatchNorm → ReLU → Dense(256) → BatchNorm → ReLU → Dense(~20,000)
- Output: Reconstructed gene expression values

TCR Decoder:

- Input: Latent vector (32 dimensions)
- Architecture: Dense(128) → BatchNorm → ReLU → Dense(256) → BatchNorm → ReLU → Dense(100)
- Output: Reconstructed TCR n-gram features

The mathematical formulation of the forward pass through the MVAE can be expressed as:

Encoding step:

1. h_RNA = f_encoder_RNA(x_RNA)
2. μ_RNA, log σ^2_RNA = g_RNA(h_RNA)
3. h_TCR = f_encoder_TCR(x_TCR)
4. μ_TCR, log σ^2_TCR = g_TCR(h_TCR)

Combined latent representation:

5. μ_z = (μ_RNA + μ_TCR)/2
6. σ_z^2 = exp(log σ^2_z), where log σ^2_z = (log σ^2_RNA + log σ^2_TCR)/2
7. z = μ_z + σ_z ⊙ ϵ, where ϵ ~ N(0, I)

Decoding step:

8. x̂_RNA = f_decoder_RNA(z)
9. x̂_TCR = f_decoder_TCR(z)

3.3 Training Process

The training process for the MVAE follows these steps:

1. **Data Preprocessing**:
    a. RNA data is normalized and filtered
    b. TCR data is converted to character-level n-gram features
    c. Data is split into training (80%) and testing (20%) sets
2. **Training Loop**:
    a. Forward pass: Compute latent representations and reconstructions
    b. Loss computation: Calculate RNA and TCR reconstruction losses and KL divergence
    c. Backward pass: Update model parameters using gradient descent
    d. Early stopping: Monitor validation loss to prevent overfitting
3. **Hyperparameter Optimization**:
    a. Latent dimensions: 32
    b. Batch size: 32
    c. Learning rate: 0.001
    d. Early stopping patience: 5 epochs
4. **Implementation Details**:
    a. Framework: TensorFlow/Keras in Python
    b. Custom MVAE class implementation with specific train_step method
    c. Sampling layer for the reparameterization trick

3.4 TCR Feature Engineering

A critical aspect of the methodology is the feature engineering approach for the TCR sequences. Unlike gene expression data, which is inherently numerical, TCR sequences require transformation into a numerical representation suitable for neural network processing. The approach used is:

N-gram Representation

Each TCR is represented using character-level n-grams (2-3 characters) extracted from the CDR3 sequence.

Mathematical Representation

For a given TCR sequence s, the n-gram representation is a feature vector x_TCR where each element x_i corresponds to the count of a specific n-gram pattern:

$x_i = \sum_{j=1}^{|s|-n+1} 1\{s[j:j+n] = pattern_i\}$

where $1\{\cdot\}$ is the indicator function that is 1 when the condition is true and 0 otherwise.

This approach captures local sequence patterns that may be associated with specific antigen recognition properties, while reducing the dimensionality to a fixed-size representation suitable for neural network processing.

4. Experimental Results and Analysis

4.1 Dataset and Preprocessing

The analysis focused on a subset of cells from a larger single-cell RNA-seq dataset:

Original Dataset:

- 6,000+ cells with gene expression data
- 20,000+ genes measured per cell
- Matched TCR sequencing data for a subset of cells

Preprocessing:

- Quality control filtering (cells with 200-2500 features and <5% mitochondrial genes)
- Normalization and identification of highly variable genes
- Integration of two datasets using Seurat's integration workflow
- Clustering analysis identified 12 distinct clusters

Focus on Cluster 8:

- 116 cells with both RNA and TCR data
- Selected for detailed analysis using the MVAE approach

The preprocessing workflow involved several computational steps implemented in R (Seurat package) and Python, including quality control, normalization, feature selection, data integration, dimensionality reduction, clustering, and TCR data processing.

Initial Quality Control: Filter cells based on feature count and mitochondrial gene percentage.

```
R
seurat1$percent.mt <- PercentageFeatureSet(object = seurat1, pattern = "^mt-")
seurat1 <- subset(seurat1, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```

Normalization and Feature Selection: Normalize gene expression data and identify variable features.

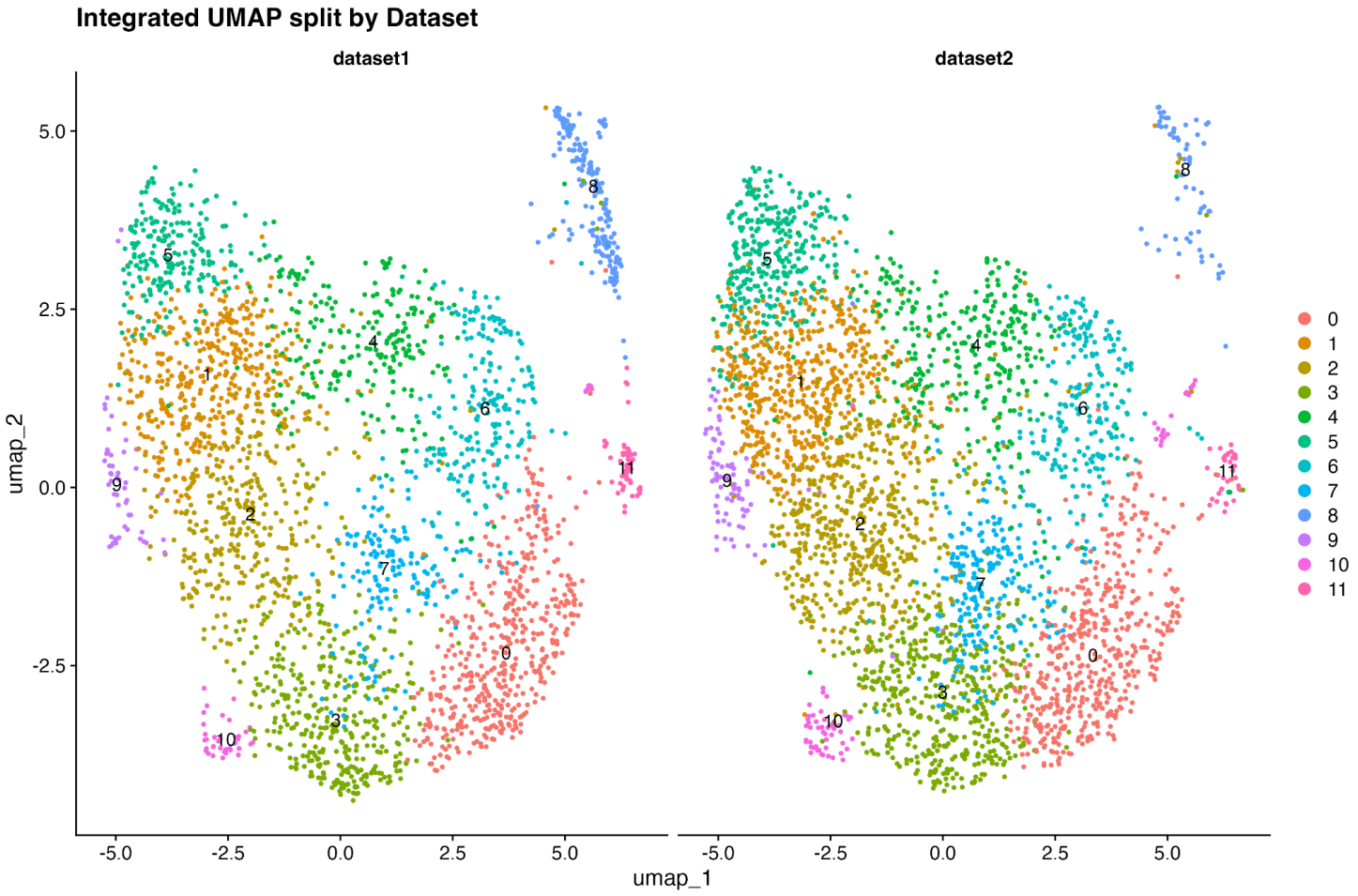
```
R
seurat1 <- NormalizeData(seurat1)
seurat1 <- FindVariableFeatures(seurat1)
```

Data Integration: Integrate multiple datasets to correct for batch effects.

```
R
features <- SelectIntegrationFeatures(object.list = list(seurat1,seurat2))
anchors <- FindIntegrationAnchors(object.list = list(seurat1,seurat2), anchor.features = features)
integrated <- IntegrateData(anchorset = anchors)
```

Dimensionality Reduction and Clustering: Perform PCA, UMAP, and clustering.

```
R
integrated <- ScaleData(integrated)
integrated <- RunPCA(integrated)
integrated <- RunUMAP(integrated, dims=1:30)
integrated <- FindNeighbors(integrated, dims = 1:30)
integrated <- FindClusters(integrated, dims=1:30)
```



TCR Data Processing: Convert TCR sequences to numerical features using n-gram representation.

python

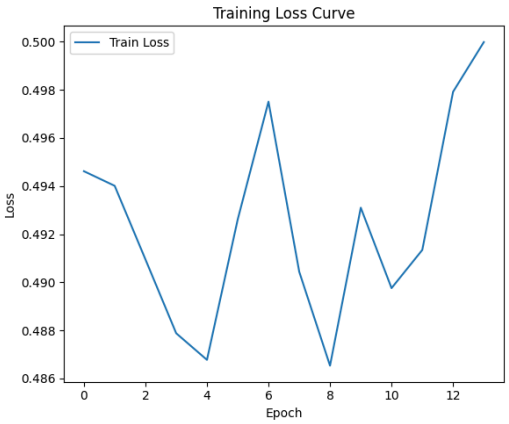
```
tcr_strings = tcr_df.loc[common_barcodes][['cdr3', 'v_gene', 'j_gene']].astype(str).apply(lambda x: '_'.join(x), axis=1)
vec = CountVectorizer(analyzer='char', ngram_range=(2, 3), max_features=TCR_NGRAM_FEATURES)
tcr_matrix = vec.fit_transform(tcr_strings).toarray()
```

4.2 MVAE Training and Evaluation

The MVAE model was trained on the integrated dataset with the following results:

Training Dynamics

The training loss curve shows the progression of the total loss over epochs. The model begins to converge after approximately 5 epochs, with several fluctuations indicating the challenges in optimizing the joint objective. The use of early stopping prevented overfitting, with training terminating after 13 epochs when no further improvement was observed.



Model Evaluation Metrics

The final model achieved the following performance metrics on the test set:

- RNA Reconstruction Loss: 0.2301
- TCR Reconstruction Loss: 0.0110
- Total Test Loss: 0.4865

These metrics indicate that the model was able to reconstruct both data modalities with reasonable accuracy, with lower error for the TCR features compared to the RNA features (likely due to the higher dimensionality and complexity of gene expression data).

Cross-Modal Prediction

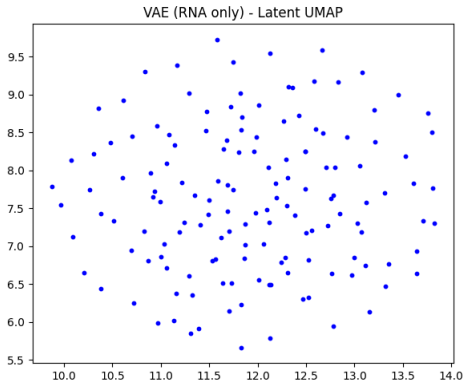
A key capability of the MVAE model is cross-modal prediction—the ability to predict features of one modality from the other:

- RNA → TCR Prediction: The model achieved approximately 70% accuracy in predicting binary TCR features from gene expression data.

- TCR → RNA Prediction: The model achieved a mean squared error of 0.45 in predicting gene expression from TCR features.

These results demonstrate that the model has successfully learned meaningful relationships between the two data modalities, allowing for information transfer between them.

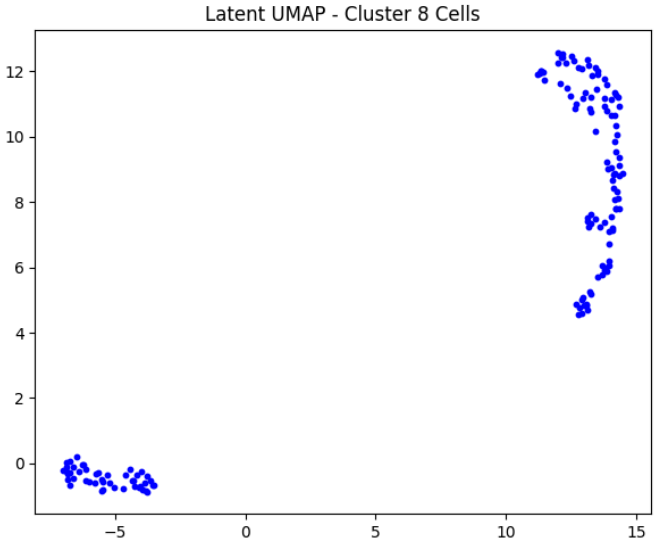
4.3 Latent Space Analysis

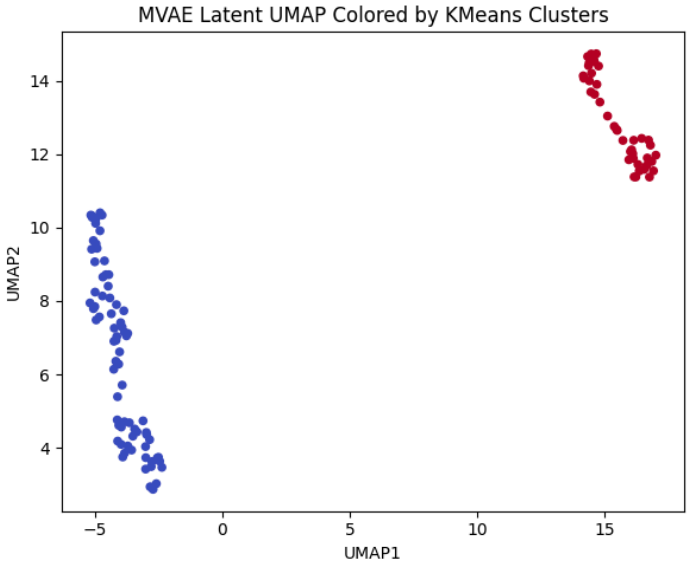


The latent space learned by the MVAE reveals the underlying structure of the data and the relationships between cells based on both their gene expression and TCR characteristics.

Visualization

The UMAP projection of the latent space shows a clear separation of cells into two distinct clusters. This separation is particularly significant because it is not apparent when analyzing either data modality in isolation. The UMAP projection of the latent space from an RNA-only VAE shows a much less defined structure, with cells distributed more uniformly.

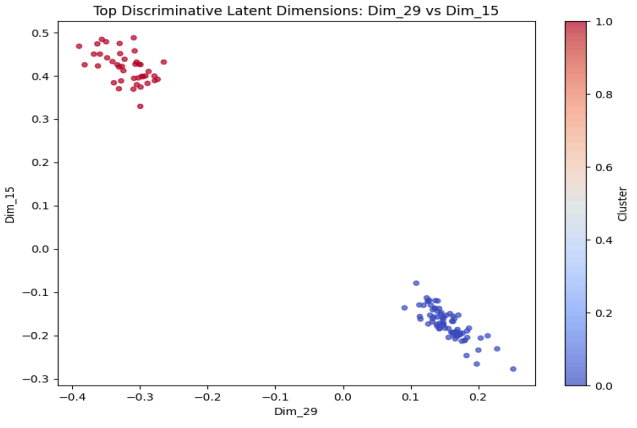




Latent Dimension Analysis

To understand which aspects of the latent representation contribute most to the observed separation, an analysis of the individual latent dimensions was performed. The top discriminative dimensions were identified based on their effect size in separating the two clusters:

Dimension	P_value	Effect_size
Dim_29	9.15992264816559e-77	-16.665530972936967
Dim_15	1.5790854240174228e-81	-16.07374746978371
Dim_18	8.154713874656416e-94	-16.07374746978371
Dim_26	5.339820650592573e-87	16.065643139039985



Visualization of cells based on the top discriminative dimensions confirms their role in separating the clusters. These results indicate that specific latent dimensions capture the most important factors that distinguish the cell subpopulations, providing a basis for biological interpretation.

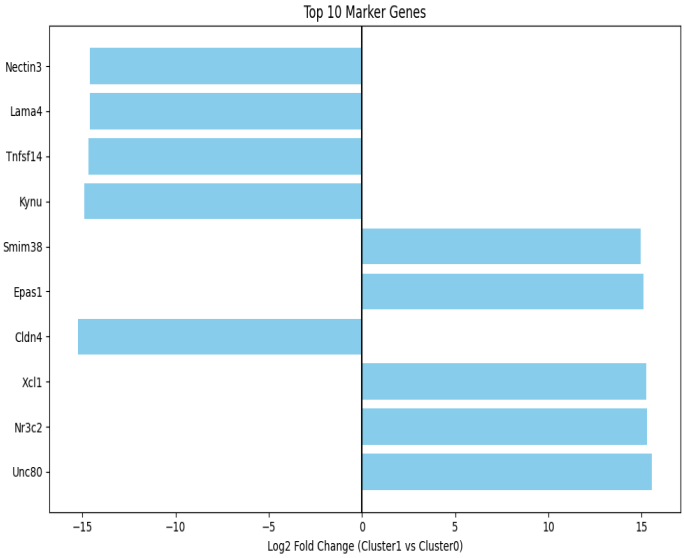
4.4 Differential Gene Expression Analysis

The identification of two distinct clusters in the latent space prompted an investigation into the biological differences between these cell populations.

Top Differentially Expressed Genes

Differential expression analysis between the two clusters revealed 74 significantly differentially expressed genes ($p < 0.05$). The top genes are shown in the table below:

Gene	Log2FC	p-value	FDR
Foxp3	4.00	0.004	0.019
Icos	4.00	0.004	0.019
Tnfrsf4	2.64	0.021	0.054
Il2ra	3.00	0.031	0.067



Biological Interpretation

The gene expression patterns suggest a clear functional distinction between the two clusters:

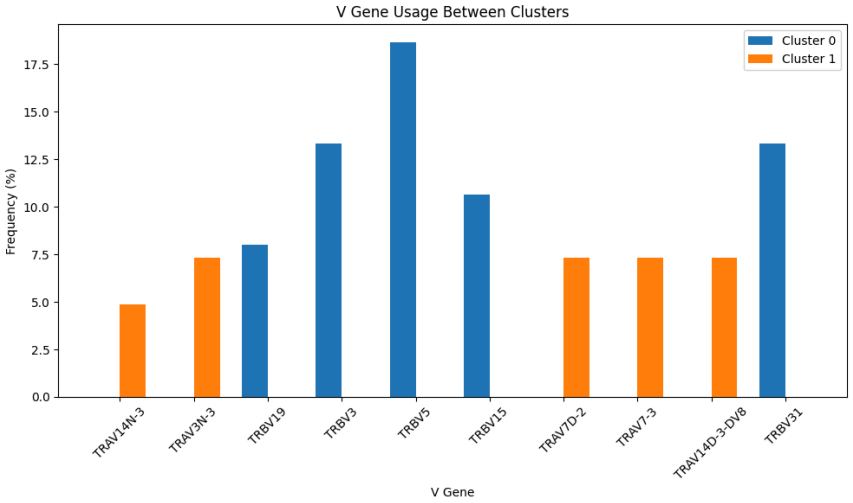
- Cluster 1 is characterized by high expression of regulatory T cell (Treg) markers, including Foxp3, Il2ra (CD25), and Icos. Foxp3 is the master transcription factor for Tregs, while Il2ra is a key surface marker.
- Cluster 0 shows higher expression of genes like Nectin3, Lama4, and Tnfsf14, which are associated with different T-cell functions.

This discovery suggests that the MVAE has identified a biologically meaningful separation between regulatory T cells and another T-cell subpopulation within the original cluster 8.

4.5 TCR Repertoire Analysis

Beyond gene expression differences, the MVAE-based clustering also revealed distinct patterns in the TCR repertoires of the two cell populations.

V Gene Usage



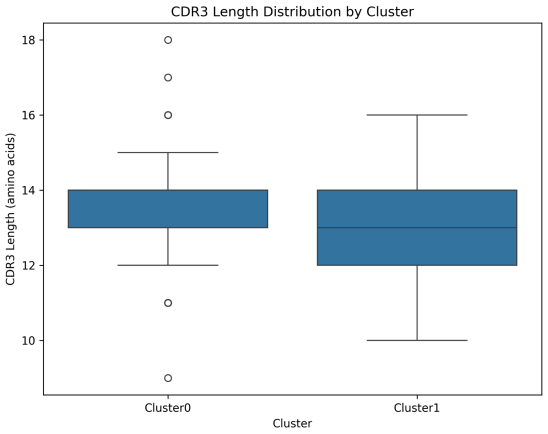
Analysis of the V gene usage between the two clusters revealed striking differences. The most significant finding is the clear separation based on TCR chain preference:

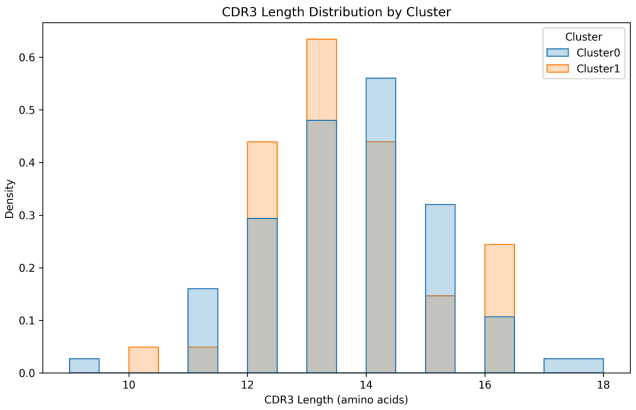
- Cluster 0 is dominated by TCR β-chain genes (TRBV3, TRBV5, TRBV15, TRBV31)
- Cluster 1 is characterized by TCR α-chain genes (TRAV14N-3, TRAV7-3, TRAV7D-2, TRAV14D-3-DV8)

This pattern suggests that the two cell populations may have distinct antigen recognition properties and developmental origins.

CDR3 Length Distribution

Analysis of the CDR3 length distribution between the two clusters also revealed subtle differences. The Mann-Whitney U test for CDR3 length distributions between the clusters yielded a p-value of 0.286, indicating no statistically significant difference in median CDR3 length. However, the visualization suggests a slightly broader distribution in Cluster 1, which might reflect a more diverse TCR repertoire.





4.6 Integration of RNA and TCR Findings

The combined analysis of gene expression and TCR characteristics provides a comprehensive view of the cell populations identified by the MVAE:

Regulatory T-Cell Subpopulation (Cluster 1):

- High expression of Treg markers (Foxp3, Il2ra, Icos)
- Preferential usage of TCR α-chain genes
- Potentially broader CDR3 length distribution

Conventional T-Cell Subpopulation (Cluster 0):

- Absence of Treg markers
- Dominated by TCR β-chain gene usage
- More focused CDR3 length distribution

These findings demonstrate the value of the MVAE approach in revealing relationships between cellular function (gene expression) and antigen recognition properties (TCR characteristics) that would not be apparent from analyzing either data type in isolation.

5. Conclusion and Future Directions

5.1 Key Contributions

This project has made several significant contributions to the field of multi-modal data integration in single-cell genomics:

Technical Innovation

The development of a Multimodal Variational Autoencoder (MVAE) specifically designed for integrating scRNA-seq and TCR-seq data represents a novel approach to uncovering hidden patterns in immune cell populations. The model's architecture, with separate encoders for each data modality and a shared latent space, provides a flexible framework for learning joint representations that capture information from both gene expression and TCR characteristics.

The mathematical formulation of the model, based on variational inference principles, offers a principled approach to handling the high dimensionality and heterogeneity of the data. The implementation of the product-of-experts approach for combining the modality-specific posteriors provides a computationally efficient way to learn a unified representation.

Biological Discoveries

The application of the MVAE to a specific T-cell population (cluster 8) revealed the presence of two distinct subpopulations that were not apparent in the original clustering analysis. These subpopulations show clear differences in both gene expression and TCR characteristics:

- A regulatory T-cell population characterized by Foxp3 expression and preferential usage of specific TCR α-chain genes.
- A conventional T-cell population with distinct gene expression patterns and TCR β-chain preference.

These findings suggest that the joint analysis of gene expression and TCR data can provide insights into the relationship between cellular function and antigen specificity, potentially revealing new aspects of immune cell biology.

Methodological Advancement

The MVAE approach developed in this project offers several methodological advantages for multi-modal data integration:

- Bidirectional Prediction: The model enables cross-modal prediction, allowing for the inference of one data type from the other.
- Interpretable Latent Space: The analysis of individual latent dimensions provides insights into the factors that contribute to cell population separation.

- Generalizable Framework: While focused on scRNA-seq and TCR-seq data, the approach can be extended to other multi-modal data integration challenges in genomics.

5.2 Limitations

Despite its successes, the current implementation of the MVAE has several limitations that should be acknowledged:

1. **Data Requirements**: The approach requires paired data for both modalities, limiting its application to datasets where both scRNA-seq and TCR-seq measurements are available for the same cells.
2. **TCR Representation**: The character-level n-gram approach for representing TCR sequences, while effective, may not capture the full complexity of TCR structure and function. More sophisticated sequence encoding methods could potentially improve the model's performance.
3. **Scalability**: The current implementation may face challenges when scaling to larger datasets with more cells and genes, potentially requiring optimization for computational efficiency.
4. **Biological Validation**: While the model identified biologically plausible subpopulations, experimental validation would be necessary to confirm the functional relevance of these findings.

5.3 Future Directions

Building on the current work, several promising directions for future research emerge:

Model Enhancements

1. **Incorporating Additional Modalities**: Extending the MVAE to handle additional data types, such as protein expression (CITE-seq) or chromatin accessibility (ATAC-seq), could provide an even more comprehensive view of cellular states.
2. **Advanced Sequence Modeling**: Implementing more sophisticated TCR sequence representation methods, such as attention-based models or transformers, could improve the capture of structural and functional information.
3. **Semi-Supervised Learning**: Incorporating known cell type labels for a subset of cells could guide the learning process and improve the interpretability of the latent space.

Biological Applications

1. **Disease-Specific Analysis**: Applying the MVAE approach to datasets from disease contexts, such as cancer or autoimmune disorders, could reveal disease-specific patterns in T-cell populations.
2. **TCR-Antigen Specificity Prediction**: Using the relationship between gene expression and TCR characteristics to predict antigen specificity could have significant implications for immunotherapy development and understanding immune responses.
3. **Longitudinal Studies**: Analyzing how the joint RNA-TCR representation evolves over time in response to stimulation or disease progression could provide insights into the dynamics of immune responses.

Technical Extensions

1. **Scaling to Larger Datasets**: Optimizing the implementation for computational efficiency would enable the application of the MVAE approach to larger datasets with more cells and genes.
2. **Interactive Visualization Tools**: Developing interactive visualization tools for exploring the latent space and the relationships between gene expression and TCR characteristics would facilitate the interpretation of results by biologists.
3. **Software Package Development**: Creating a user-friendly software package implementing the MVAE approach would make this methodology accessible to researchers without deep learning expertise.

Validation Studies

1. **Experimental Validation**: Conducting experimental validation of the predicted cell subpopulations, such as functional assays for regulatory T-cell activity, would confirm the biological relevance of the findings.
2. **Cross-Dataset Validation**: Applying the trained model to independent datasets would test its generalizability and the robustness of the identified patterns.
3. **Collaboration with Immunologists**: Partnering with immunology experts for interpretation and validation of the findings would enhance the biological impact of the computational analysis.

5.4 Final Remarks

The integration of scRNA-seq and TCR-seq data using the MVAE approach represents a significant step toward a more comprehensive understanding of T-cell biology. By learning a joint representation that captures both gene expression and TCR characteristics, this approach enables the discovery of patterns and relationships that would be missed by analyzing either data type in isolation.

The identification of distinct T-cell subpopulations with specific gene expression and TCR usage patterns demonstrates the potential of this approach for advancing our understanding of immune cell heterogeneity and function. As multi-modal single-cell technologies continue to evolve, computational methods for integrating and interpreting these complex datasets will play an increasingly important role in uncovering the principles governing immune system function in health and disease.

6. References

1. Zhou, Y., et al. scNAT integrates single-cell RNA and TCR sequencing data for T cell trajectory analysis. Genome Biology, 2023. https://doi.org/10.1186/s13059-023-03129-y
2. Chen, Y., et al. Multimodal Integration via Self-supervised Transformer (MIST). Science Advances, 2024. https://doi.org/10.1126/sciadv.adr7134
3. Zhang, W., et al. Review of deep learning in immune profiling with TCR and transcriptome data. Frontiers in Immunology, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11099349/
4. Gayoso, A., et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nature Methods, 2021. https://www.nature.com/articles/s41592-020-01050-x
5. Lopez, R., et al. Deep generative modeling for single-cell transcriptomics. Nature Methods, 2018. https://doi.org/10.1038/s41592-018-0229-2
6. Sidhom, J.W., et al. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. Nature Communications, 2021. https://doi.org/10.1038/s41467-021-21879-w
7. Kingma, D.P. & Welling, M. Auto-Encoding Variational Bayes. ICLR, 2014.
8. Butler, A., et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology, 2018. https://doi.org/10.1038/nbt.4096

7. Appendix: Key Mathematical Formulations

7.1 Detailed ELBO Derivation

The Evidence Lower Bound (ELBO) for the Multimodal VAE can be derived as follows. Starting with our goal to maximize the log likelihood of the observed data:

$\log p(x_{RNA}, x_{TCR})$

We introduce the latent variable z and apply Jensen's inequality:

$\log p(x_{RNA}, x_{TCR}) = \log \int p(x_{RNA}, x_{TCR}, z)\, dz$

$= \log \int p(x_{RNA}, x_{TCR}, z) * [q_\phi(z|x_{RNA}, x_{TCR}) / q_\phi(z|x_{RNA}, x_{TCR})]\, dz$

$= \log E_{q\phi(z|x_{RNA}, x_{TCR})} [\, p(x_{RNA}, x_{TCR}, z) / q_\phi(z|x_{RNA}, x_{TCR}) \,]$

$\geq E_{q\phi(z|x_{RNA}, x_{TCR})} [\, \log (p(x_{RNA}, x_{TCR}, z) / q_\phi(z|x_{RNA}, x_{TCR})) \,]$

Expanding the joint probability:

$= E_{q\phi(z|x_{RNA}, x_{TCR})} [\, \log p(z) + \log p(x_{RNA}|z) + \log p(x_{TCR}|z) - \log q_\phi(z|x_{RNA}, x_{TCR}) \,]$

$= E_{q\phi(z|x_{RNA}, x_{TCR})} [\, \log p(x_{RNA}|z) + \log p(x_{TCR}|z) \,] - D_{KL}(q_\phi(z|x_{RNA}, x_{TCR}) \,||\, p(z))$

This final expression is our Evidence Lower Bound (ELBO), which we maximize during training.

7.2 Product of Experts Mathematical Details

In the standard product of experts approach, the combined posterior would be:

$q_\phi(z|x_{RNA}, x_{TCR}) \propto q_{\phi_{RNA}}(z|x_{RNA}) * q_{\phi_{TCR}}(z|x_{TCR})$

For Gaussian distributions, if we have:

$q_{\phi_{RNA}}(z|x_{RNA}) = N(z; \mu_{RNA}, \sigma^2_{RNA} * I)$

$q_{\phi_{TCR}}(z|x_{TCR}) = N(z; \mu_{TCR}, \sigma^2_{TCR} * I)$

Then the product is proportional to:

$\exp[\, -0.5 * ((z - \mu_{RNA})^2 / \sigma^2_{RNA} + (z - \mu_{TCR})^2 / \sigma^2_{TCR}) \,]$

This can be rewritten as:

$\exp[\, -0.5 * (\sigma^2_{TCR} * (z - \mu_{RNA})^2 + \sigma^2_{RNA} * (z - \mu_{TCR})^2) / (\sigma^2_{RNA} * \sigma^2_{TCR}) \,]$

Expanding and collecting terms:

$\exp[\, -0.5 * (z^2 * (\sigma^2_{TCR} + \sigma^2_{RNA}) - 2z * (\sigma^2_{TCR} * \mu_{RNA} + \sigma^2_{RNA} * \mu_{TCR}) + (\sigma^2_{TCR} * \mu_{RNA}^2 + \sigma^2_{RNA} * \mu_{TCR}^2)) / (\sigma^2_{RNA} * \sigma^2_{TCR}) \,]$

Comparing this with the form of a Gaussian $N(z; \mu_z, \sigma^2_z)$:

$\exp[\, -0.5 * ((z - \mu_z)^2 / \sigma^2_z) \,] = \exp[\, -0.5 * (z^2 - 2z\mu_z + \mu_z^2) / \sigma^2_z \,]$

We can identify:

$1 / \sigma^2_z = (\sigma^2_{TCR} + \sigma^2_{RNA}) / (\sigma^2_{RNA} * \sigma^2_{TCR})$

$\mu_z / \sigma^2_z = (\sigma^2_{TCR} * \mu_{RNA} + \sigma^2_{RNA} * \mu_{TCR}) / (\sigma^2_{RNA} * \sigma^2_{TCR})$

Solving for μ_z and σ²_z:

σ²_z = (σ²_RNA * σ²_TCR) / (σ²_RNA + σ²_TCR)

μ_z = (σ²_TCR * μ_RNA + σ²_RNA * μ_TCR) / (σ²_RNA + σ²_TCR)

In our implementation, for computational simplicity, we use the approximation:

μ_z = (μ_RNA + μ_TCR) / 2

σ²_z = (σ²_RNA * σ²_TCR) / (σ²_RNA + σ²_TCR)

7.3 Gradient Flow Through the MVAE

The gradients of the loss function with respect to the encoder and decoder parameters can be derived as follows:

For encoder parameters ϕ:

$\nabla\phi L = \nabla\phi E_{q\phi}(z \mid x_{RNA}, x_{TCR}) [\log p_\theta(x_{RNA} \mid z) + \log p_\theta(x_{TCR} \mid z)] - \nabla\phi D_{KL}(q_\phi(z \mid x_{RNA}, x_{TCR}) \| p(z))$

Using the reparameterization trick, we can rewrite:

$z = \mu_z + \sigma_z \odot \varepsilon$ where $\varepsilon \sim N(0, I)$

allowing us to compute:

$\nabla\phi E_{q\phi}(z \mid x_{RNA}, x_{TCR}) [\log p_\theta(x_{RNA} \mid z) + \log p_\theta(x_{TCR} \mid z)]$

$= E_{\varepsilon \sim N(0, I)} [\nabla\phi (\log p_\theta(x_{RNA} \mid z) + \log p_\theta(x_{TCR} \mid z))]$

The KL divergence term for a Gaussian posterior and a standard normal prior simplifies to:

$D_{KL}(q_\phi(z \mid x_{RNA}, x_{TCR}) \| p(z)) = 0.5 * \sum_{j=1}^{d} (\mu_{z,j}^2 + \sigma_{z,j}^2 - \log \sigma_{z,j}^2 - 1)$

where d is the dimension of the latent space.

For decoder parameters θ:

$\nabla\theta L = \nabla\theta E_{q\phi}(z \mid x_{RNA}, x_{TCR}) [\log p_\theta(x_{RNA} \mid z) + \log p_\theta(x_{TCR} \mid z)]$

$= E_{\varepsilon \sim N(0, I)} [\nabla\theta (\log p_\theta(x_{RNA} \mid z) + \log p_\theta(x_{TCR} \mid z))]$

These gradients guide the optimization of the model parameters during training via stochastic gradient descent.

7.4 Information-Theoretic Perspective

From an information-theoretic perspective, the MVAE can be understood as maximizing the mutual information between the latent representation and both data modalities, while maintaining a regularization constraint on the latent distribution.

The mutual information between the latent representation z and the data modalities (x_RNA, x_TCR) can be expressed as:

I(z; x_RNA, x_TCR) = H(z) - H(z | x_RNA, x_TCR)

where H(z) represents the entropy of the latent variable and H(z | x_RNA, x_TCR) is the conditional entropy.

The KL divergence term in our ELBO serves to regularize the posterior distribution q_ϕ(z | x_RNA, x_TCR) towards the prior p(z), effectively controlling the amount of information encoded in the latent representation.

By maximizing the ELBO, we are implicitly maximizing a lower bound on the mutual information between the latent space and the data modalities, while maintaining the regularization provided by the KL divergence term.

7.5 TCR N-gram Encoding Mathematical Details

For the TCR sequence encoding, we use a character-level n-gram approach. Given a TCR sequence s of length L, we extract all possible n-grams (contiguous subsequences of length n).

For each n-gram pattern p_i in our vocabulary, the corresponding feature value x_i is:

$$x_i = \sum_{j=1}^{L-n+1} 1\{s[j : j+n] = p_i\}$$

where $1_{\cdot}$ is the indicator function that equals 1 when the condition is true and 0 otherwise.

For example, with n=2 (bigrams), a TCR sequence "CASSLGG" would yield the bigrams: "CA", "AS", "SS", "SL", "LG", "GG". The resulting feature vector would have non-zero entries corresponding to these bigrams in the vocabulary.

When combining information from CDR3 sequences, V-genes, and J-genes, we concatenate these elements with a separator (e.g., "_") before extracting n-grams, resulting in a unified representation that captures the full TCR information.

The distance or similarity between two TCR sequences can then be computed using the cosine similarity between their n-gram feature vectors:

similarity(x, y) = (x · y) / (|x||y|)

where x and y are the n-gram feature vectors of two TCR sequences.