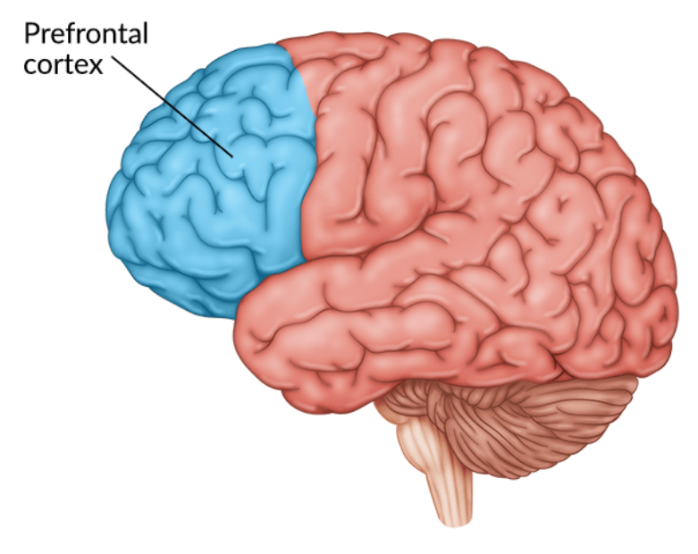
Gaussian Mixture Variational Autoencoder for Noise Reduction and Batch Correction of DNA Methylation Data

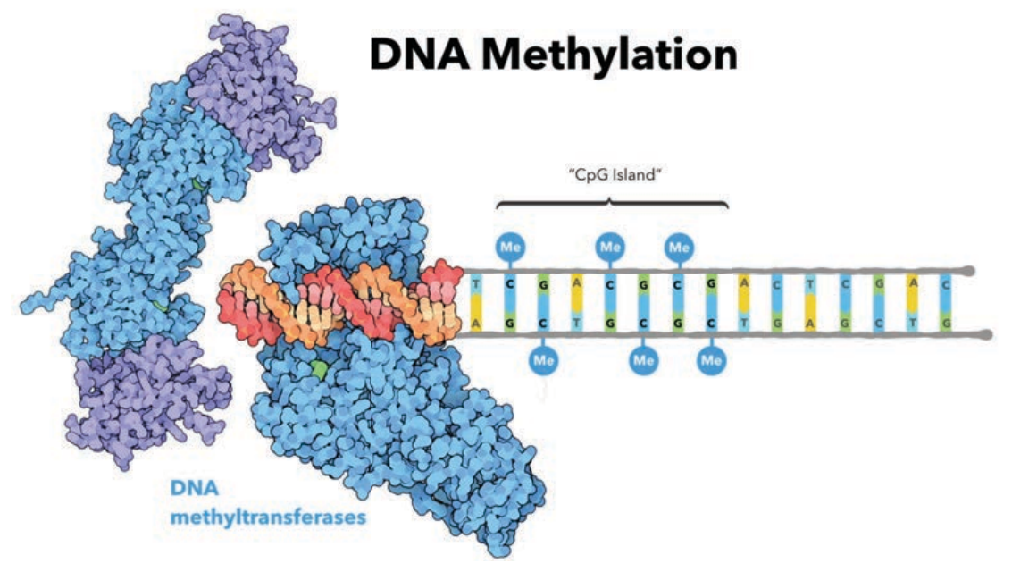
HUNTER CARROLL[1,2], Matthew Heffel[3,4], Cuining Liu[3], Chongyuan Luo[3]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 University of Colorado, Denver
3 Department of Human Genetics David Geffen School of Medicine, UCLA
4 Bioinformatics Interdepartmental Program, UCLA

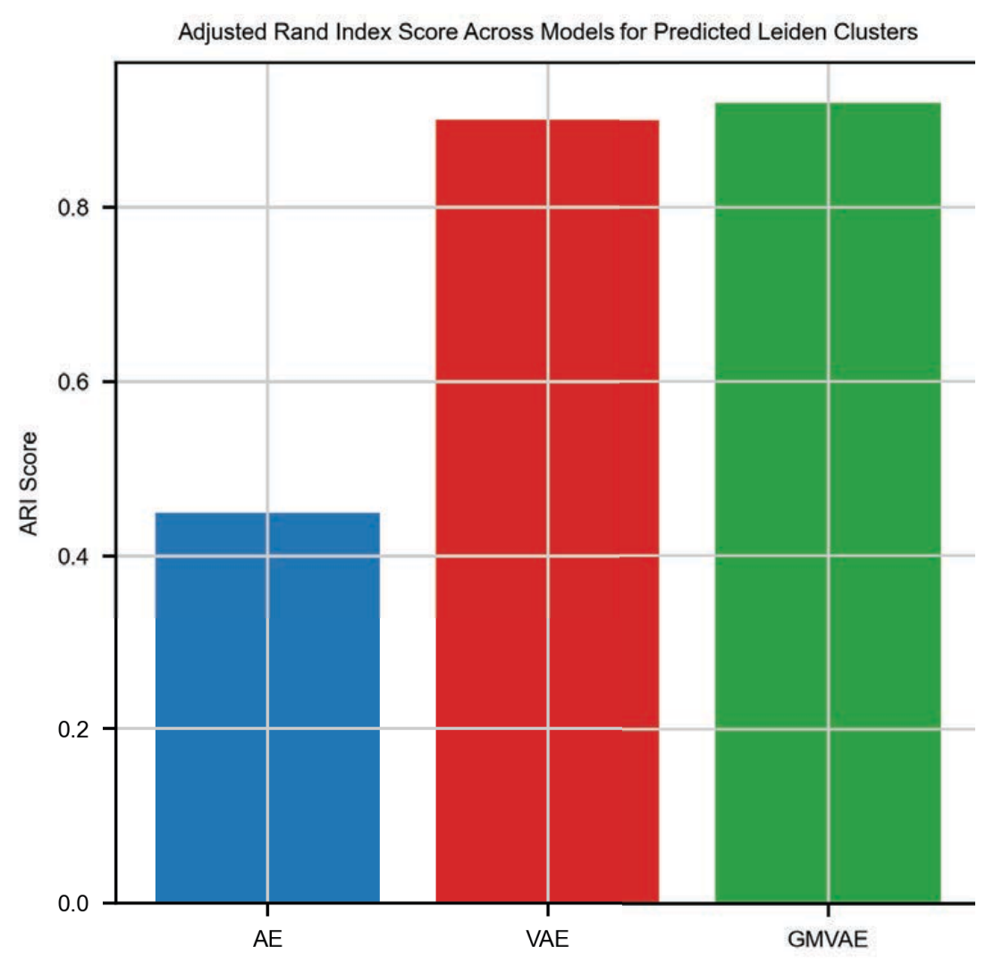
Abstract

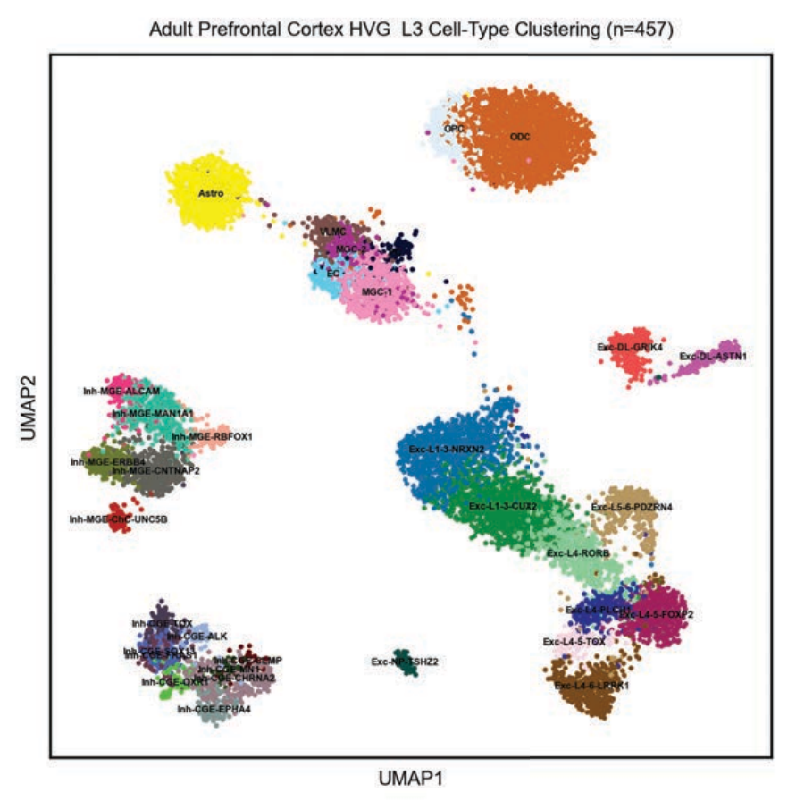
- Cytosine methylation (mC) is a crucial epigenetic modification regulating gene expression and cellular development. However, mC suffers from high dimensionality when studying multiple CpG sites astride the genome.
- Using the gene body counts from the adult prefrontal cortex we aim to reduce the in-group (cell-type) variation by using neural networks.
- Our data: single-nucleus methyl-3C sequencing (sn-m3C-seq) to capture chromatin organization and DNA methylation information
- We use three neural network models that are the Autoencoder (AE), variational Autoencoder (VAE), and a Gaussian Mixed VAE to interpret the adult prefrontal cortex in terms of in-group variance, reconstructive capabilities, and clustering.
- Our high-variance genes are determined by ranking gene groups and identifying n-chosen differentially expressed marker genes from each cell-type group.





ARI Score Across Models

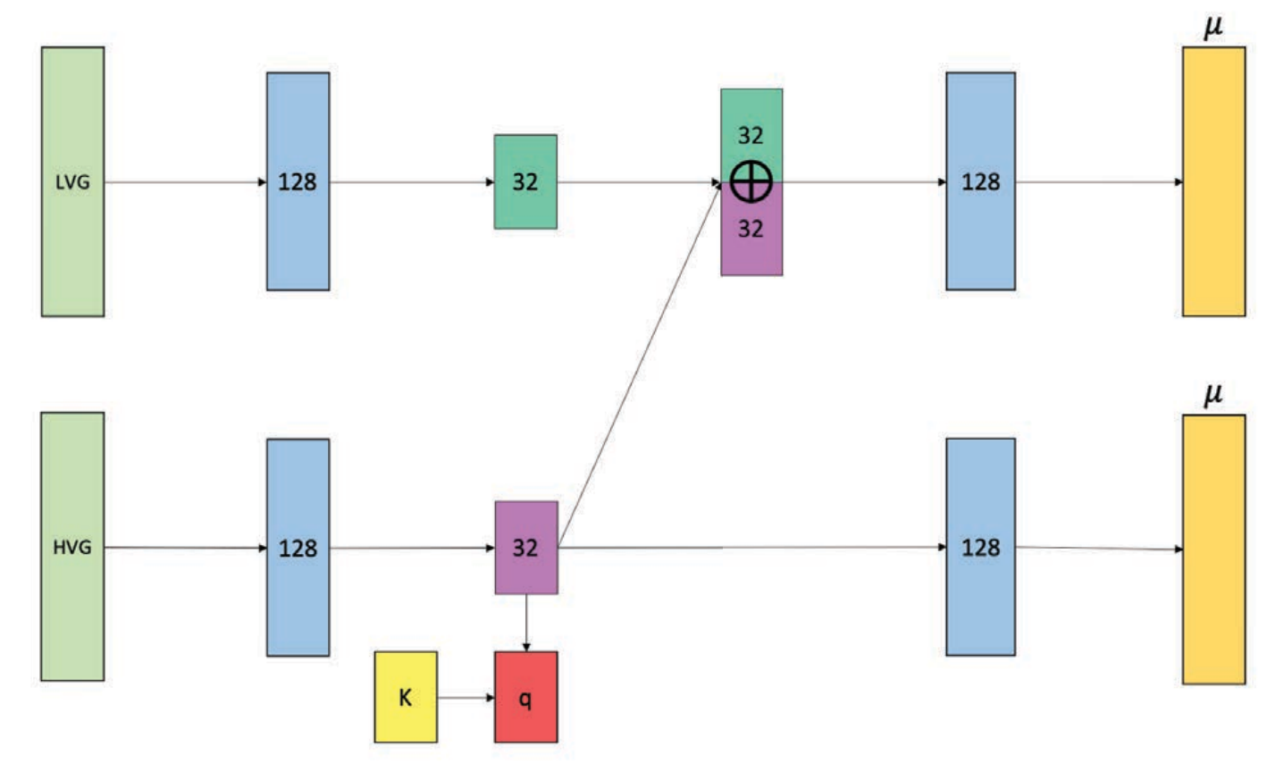




Original UMAP Space of Highly Variable Genes

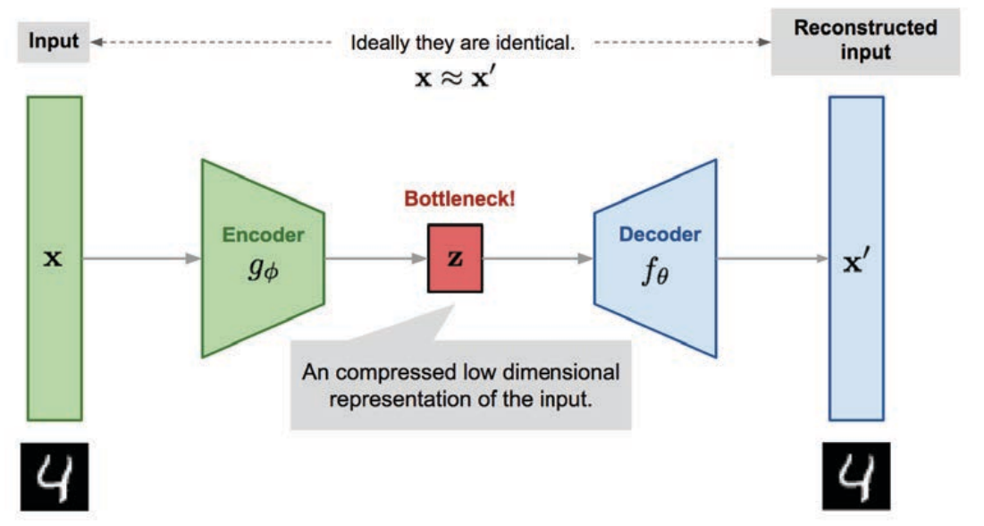
Conclusions

1. Our models demonstrate the reduction of noise within individual cell types and the accuracy of reconstruction by measuring the unsupervised Leiden cluster assignments between the input and reconstructed data.
2. We focus on the adult prefrontal cortex using 456 highly variable genes determined via differential expression in ranked gene groups. Our models demonstrate modest performance in terms of both our data reconstruction and clustering capabilities.
3. A noteworthy observation is that our Autoencoder performs best at noise reduction but our Gaussian Mixed Variational Autoencoder performs best in terms of clustering accuracy demonstrating an important trade-off to be considered for further analysis.
4. Our goal is to develop a model that can successfully balance these two objectives by creating a branched Gaussian Mixed VAE that employs the high variance genes to drive the reconstruction and clustering accuracy of the low variance genes by assigning Gaussian membership probabilities in our latent spaces as shown in the figure to the right.

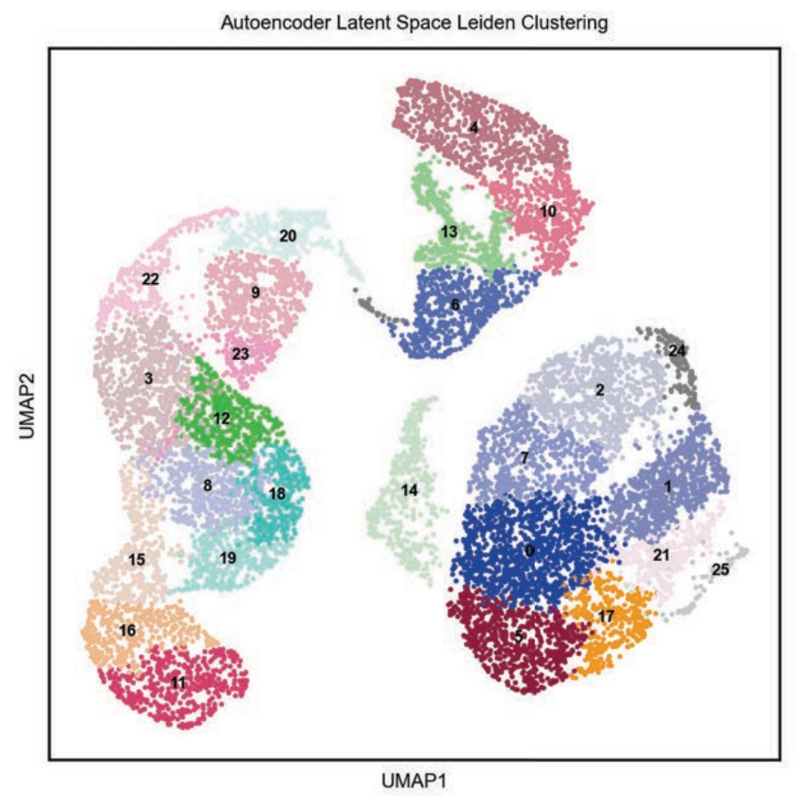


Concatenated high variance block and low variance block in the latent space conditioned on q - membership probabilities

Autoencoder

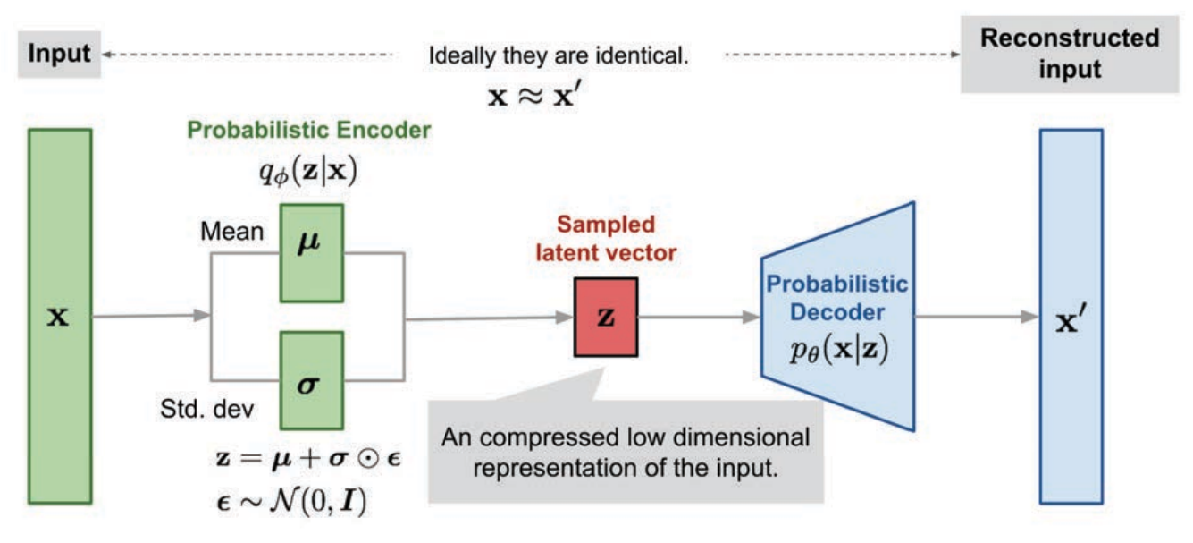

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
An compressed low dimensional representation of the input.
4
4

Autoencoder Architecture

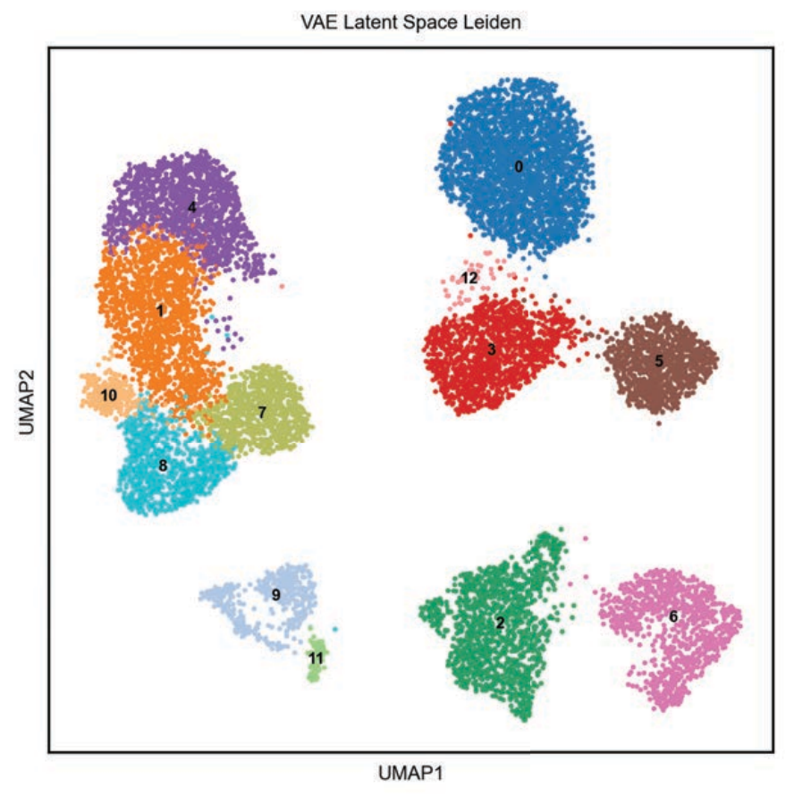


Latent Space of Autoencoder: Captures some general structure but needs more model complexity.

Variational Autoencoder (VAE)

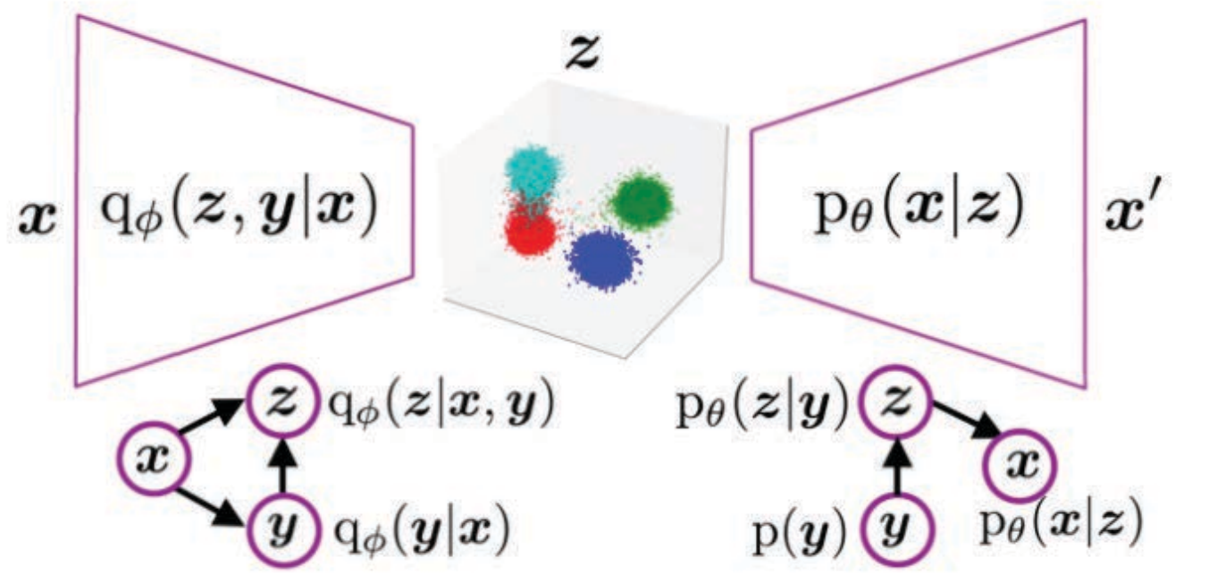

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
Mean
$\boldsymbol{\mu}$
Std. dev
$\boldsymbol{\sigma}$
$\mathbf{x}$
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon$
$\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

Variational Autoencoder Architecture

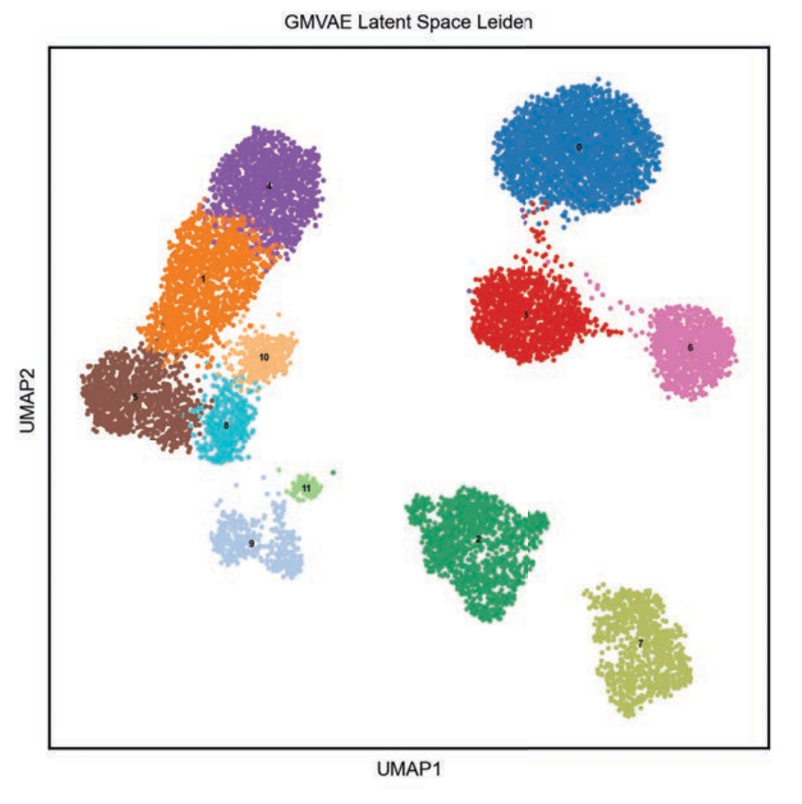


Latent Space of VAE: Capturing more detailed global features but we can improve this with GMVAE.

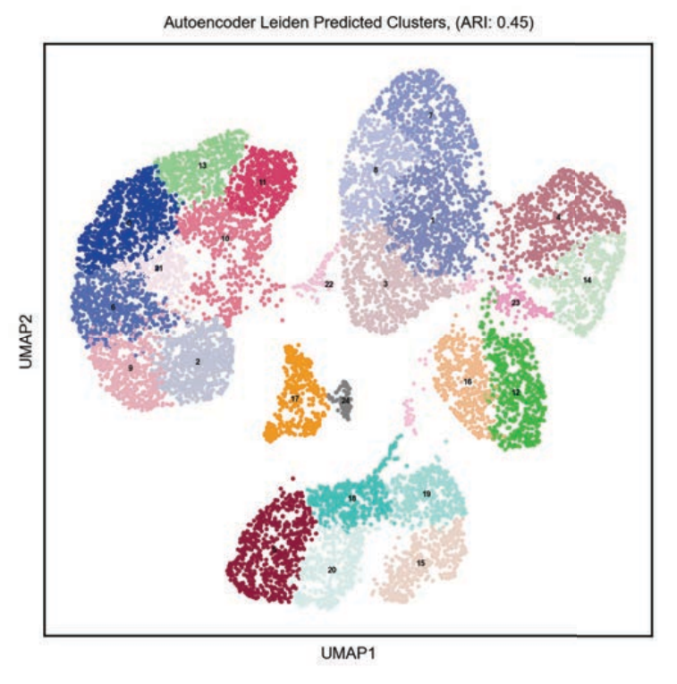
Gaussian Mixed VAE (GMVAE)


$\boldsymbol{x}$
$q_\phi(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{x})$
$\boldsymbol{z}$
$p_\theta(\boldsymbol{x}|\boldsymbol{z})$
$\boldsymbol{x}'$
$q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$
$q_\phi(\boldsymbol{y}|\boldsymbol{x})$
$p_\theta(\boldsymbol{z}|\boldsymbol{y})$
$p(\boldsymbol{y})$
$p_\theta(\boldsymbol{x}|\boldsymbol{z})$

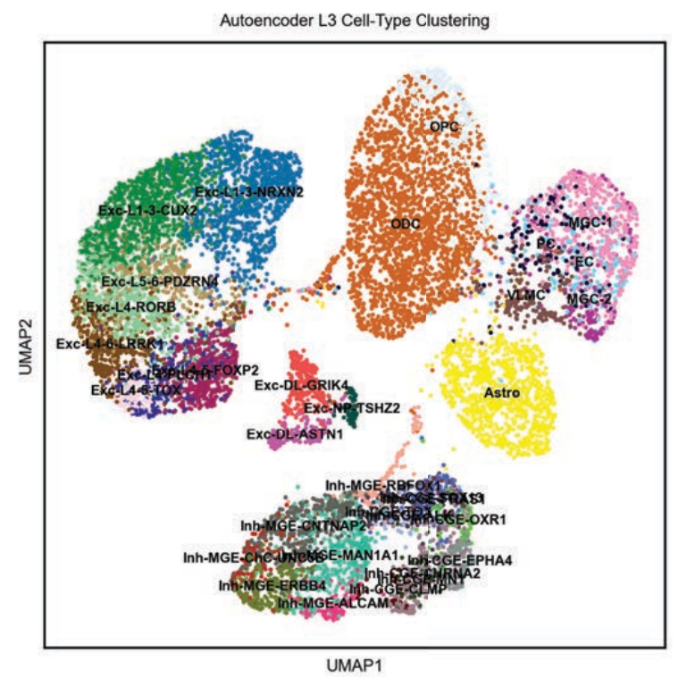
Gaussian Mixed Variational Autoencoder Architecture



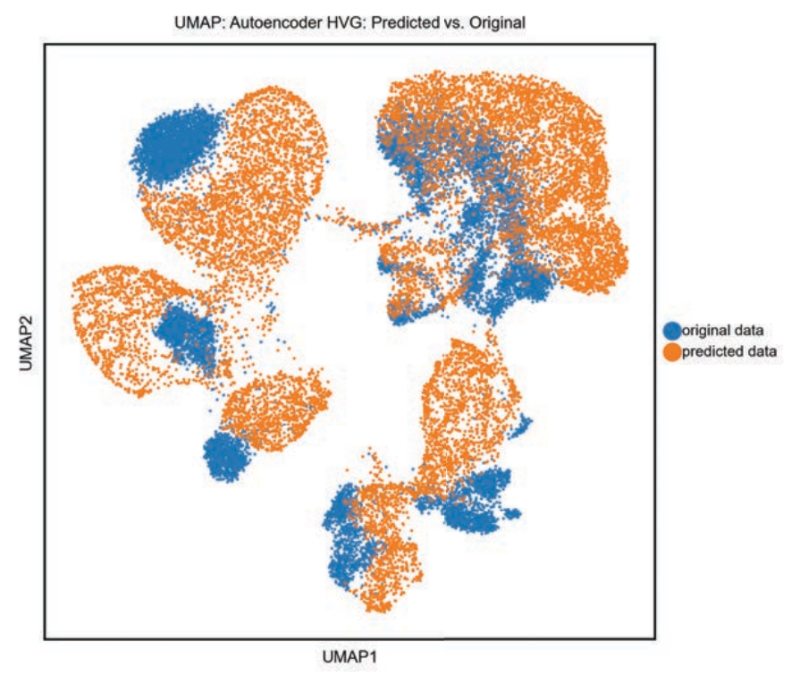
Latent Space of GMVAE: Our goal in order to eventually use to drive the low variance features



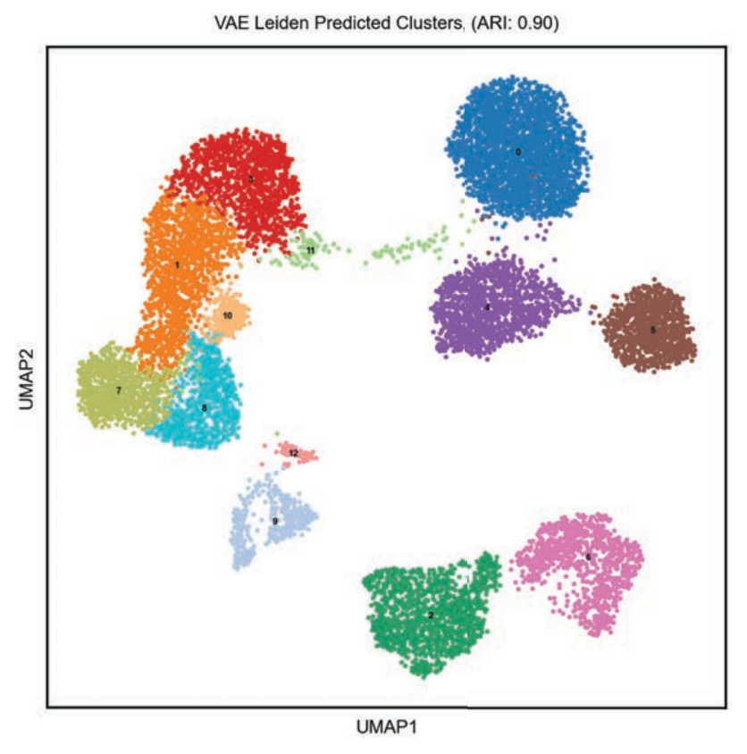
Autoencoder: N=457 Highly Variable Genes Predicted Leiden Clusters



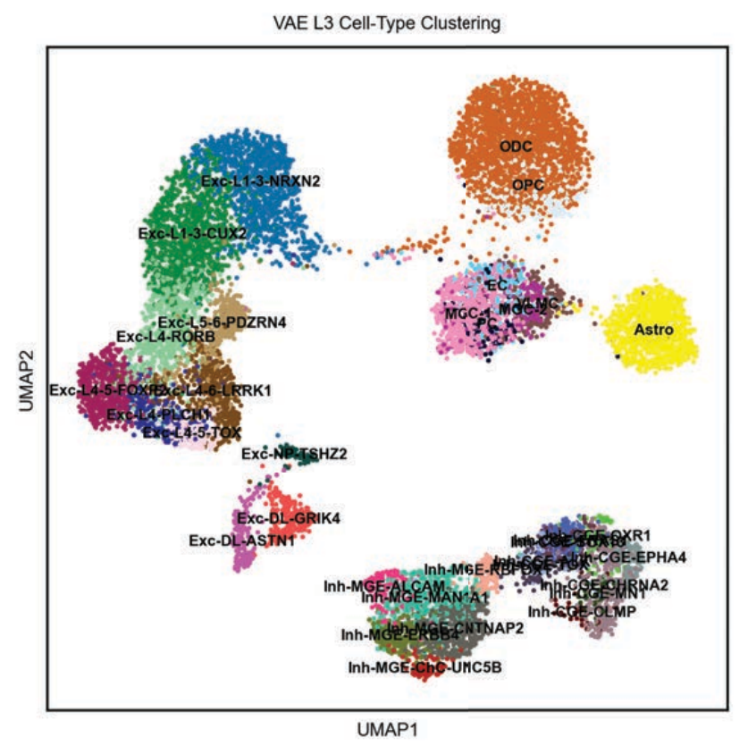
L3 Cell-Type Annotations



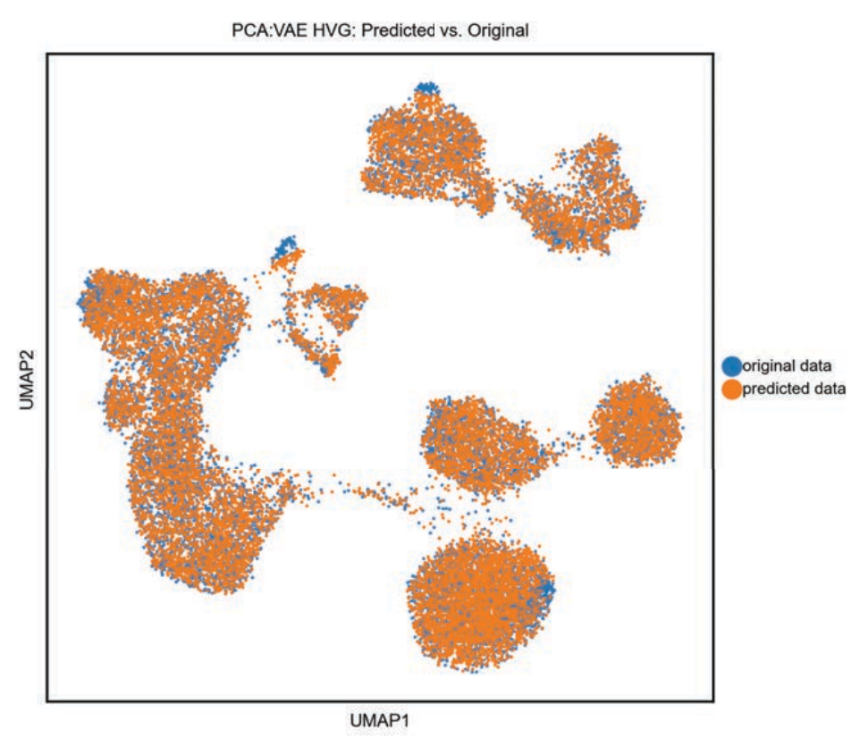
Predicted UMAP vs. Original UMAP



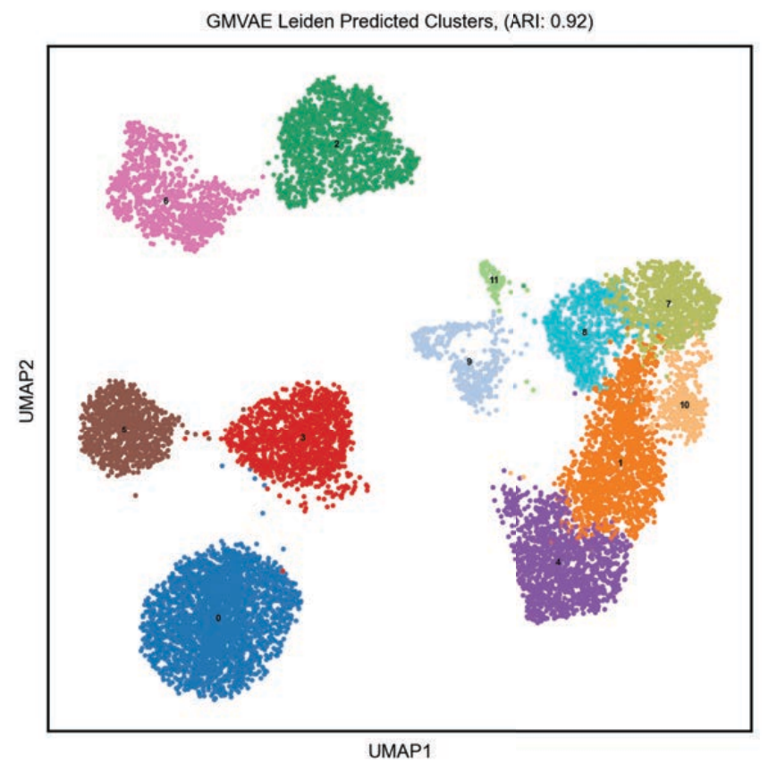
VAE: N=457 Highly Variable Genes Predicted Leiden Clusters



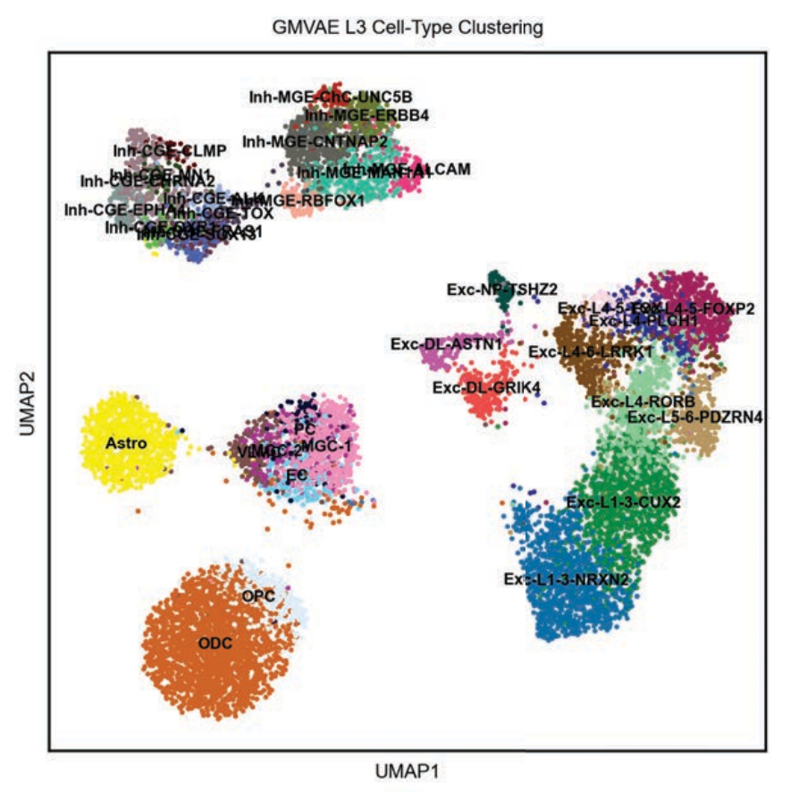
L3 Cell-Type Annotations



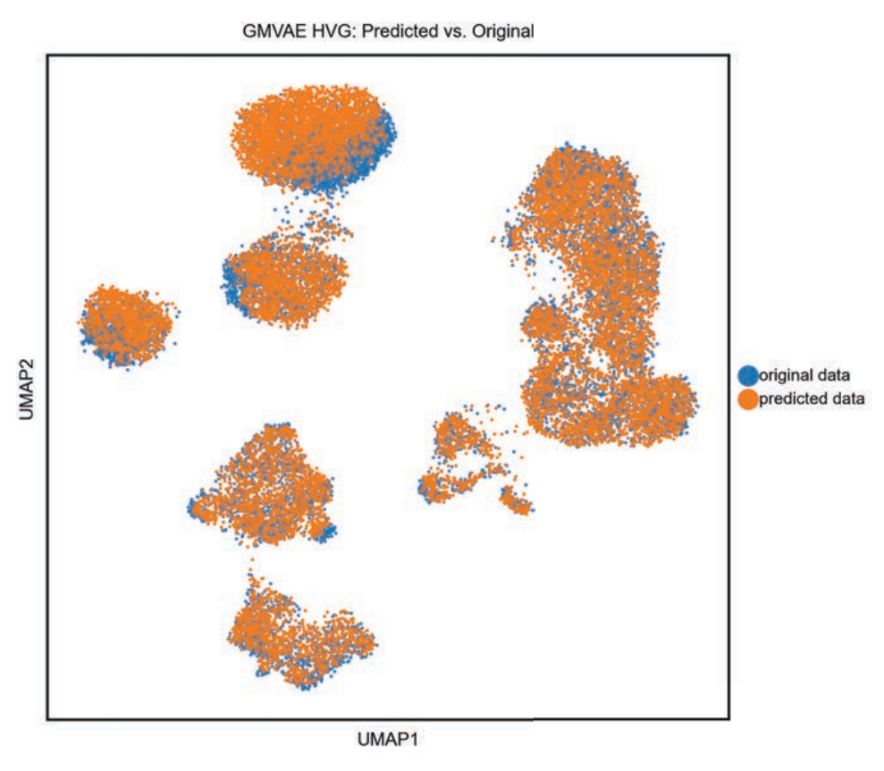
Predicted UMAP vs. Original UMAP



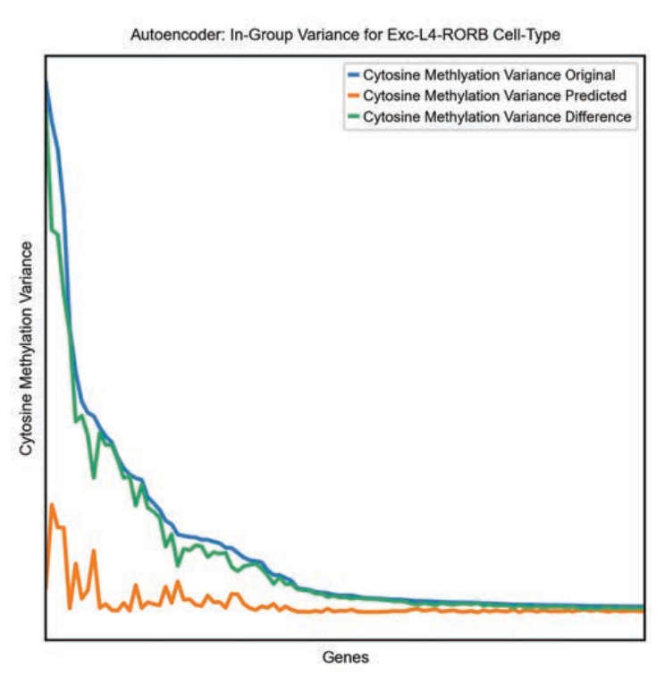
GMVAE: N=457 Highly Variable Genes Predicted Leiden Clusters



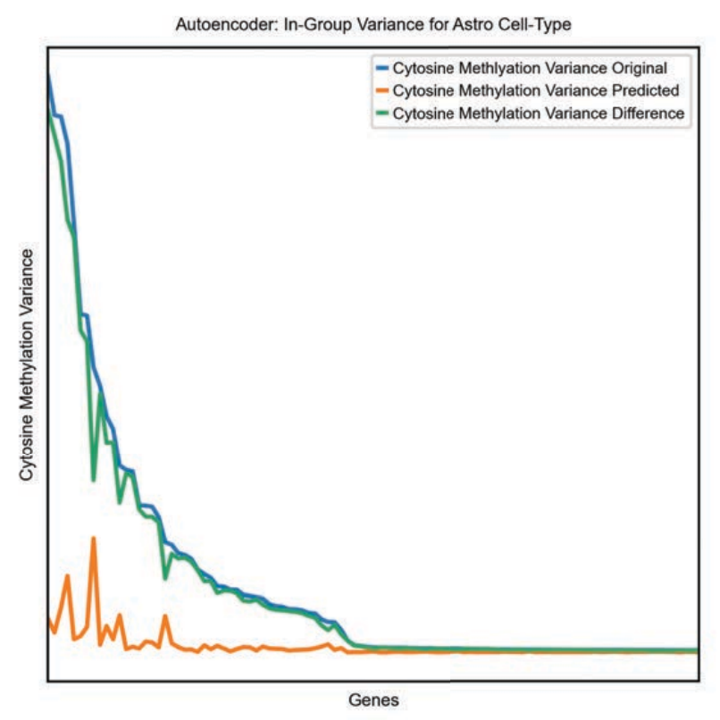
L3 Cell-Type Annotations



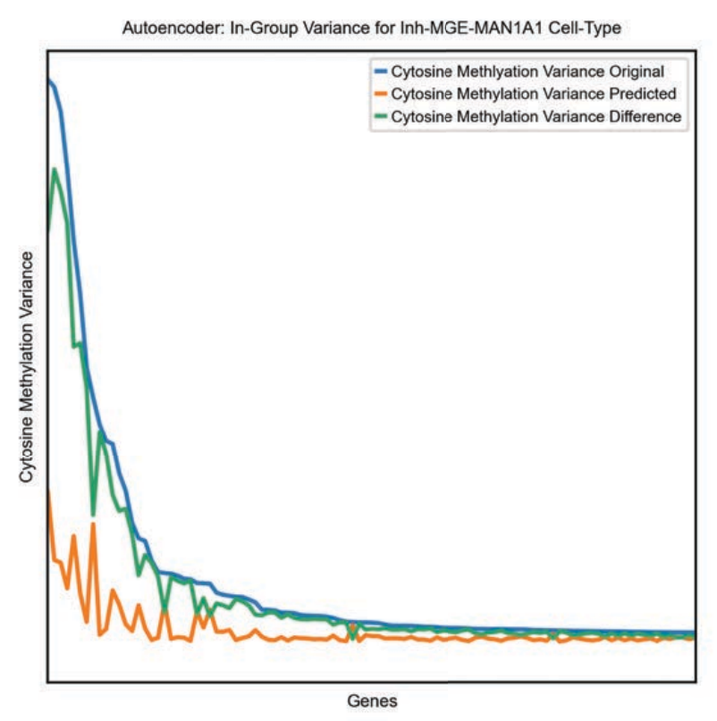
Predicted UMAP vs. Original UMAP



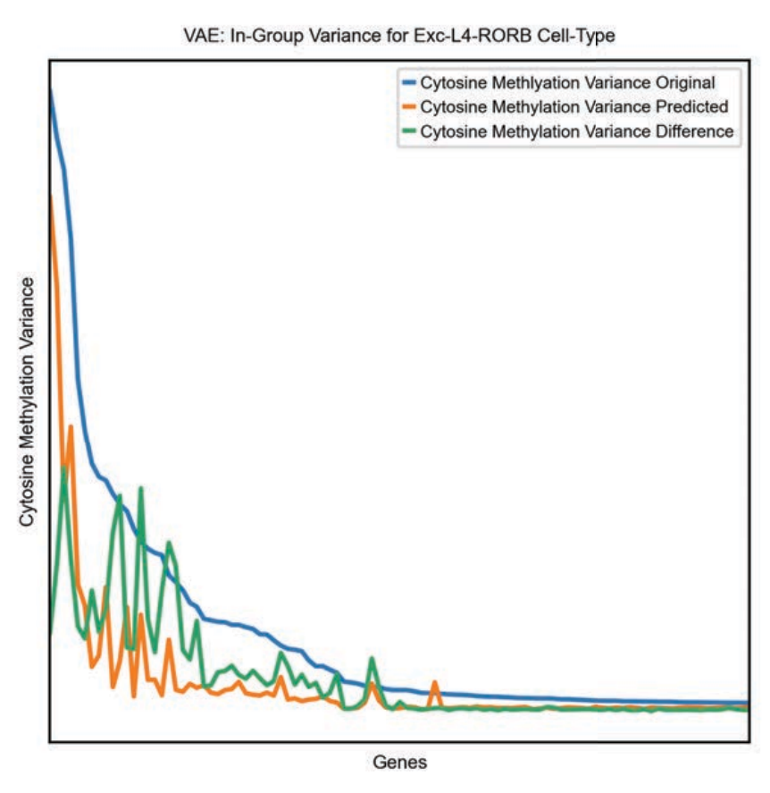
EXC-L4-RORB Cell-Type Variation



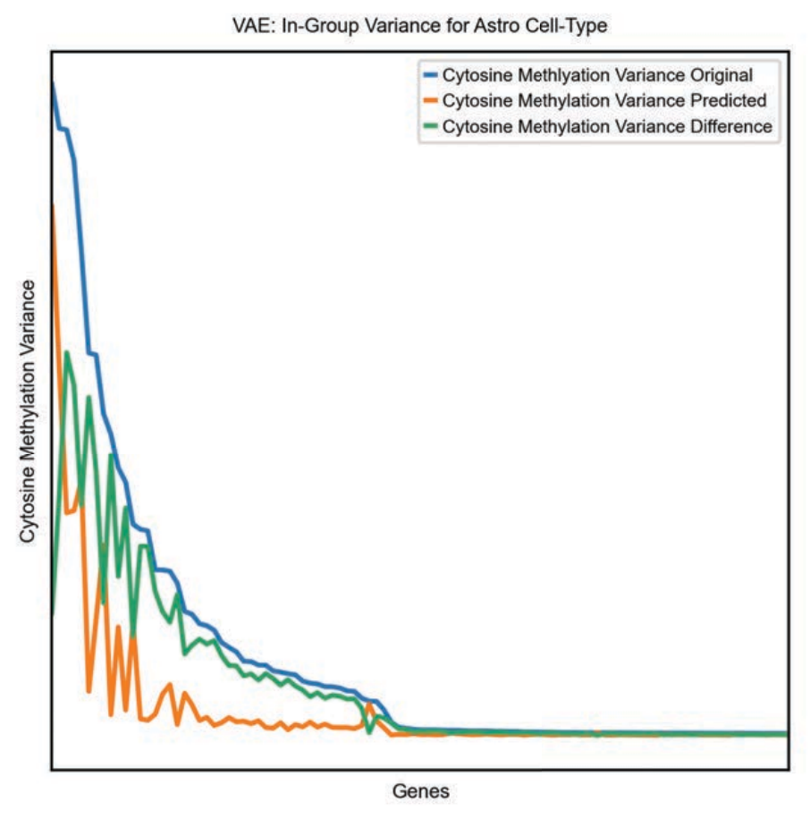
Astro Cell-Type Variation



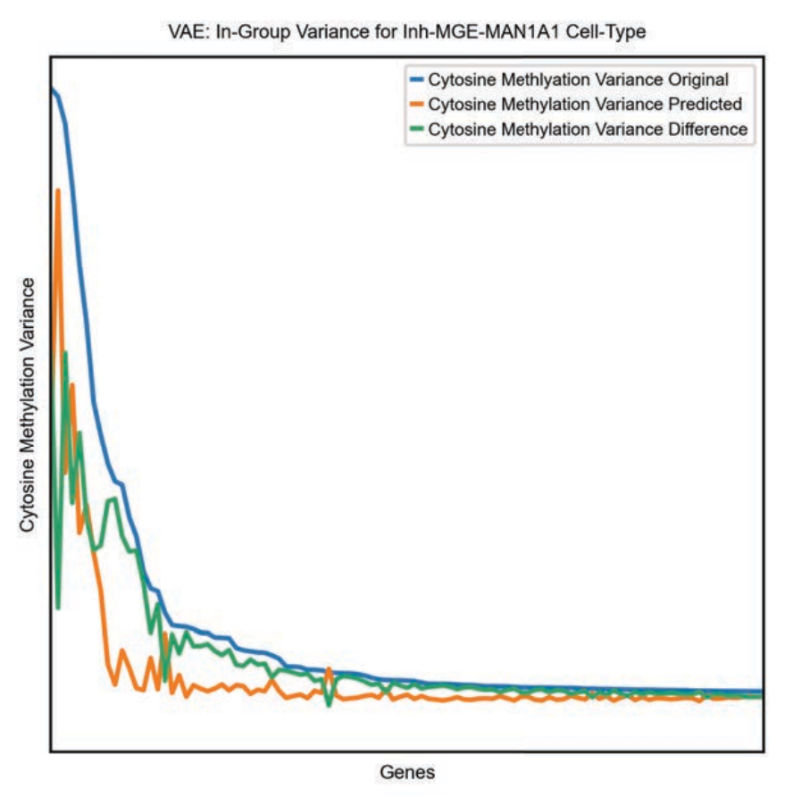
Inh-MGE-MAN1A1 Cell-Type Variation



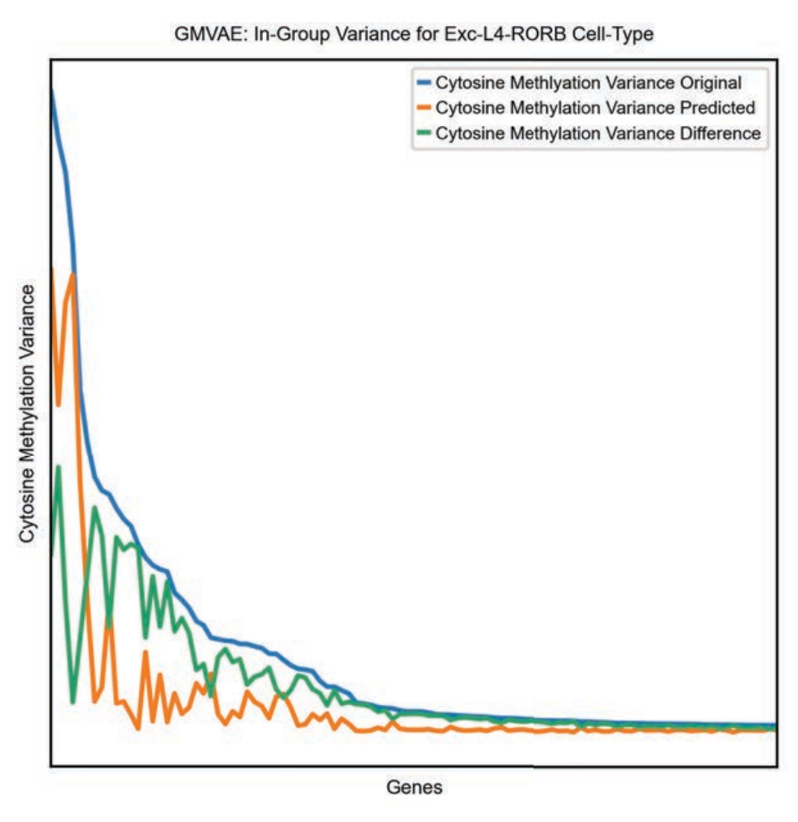
EXC-L4-RORB Cell-Type Variation



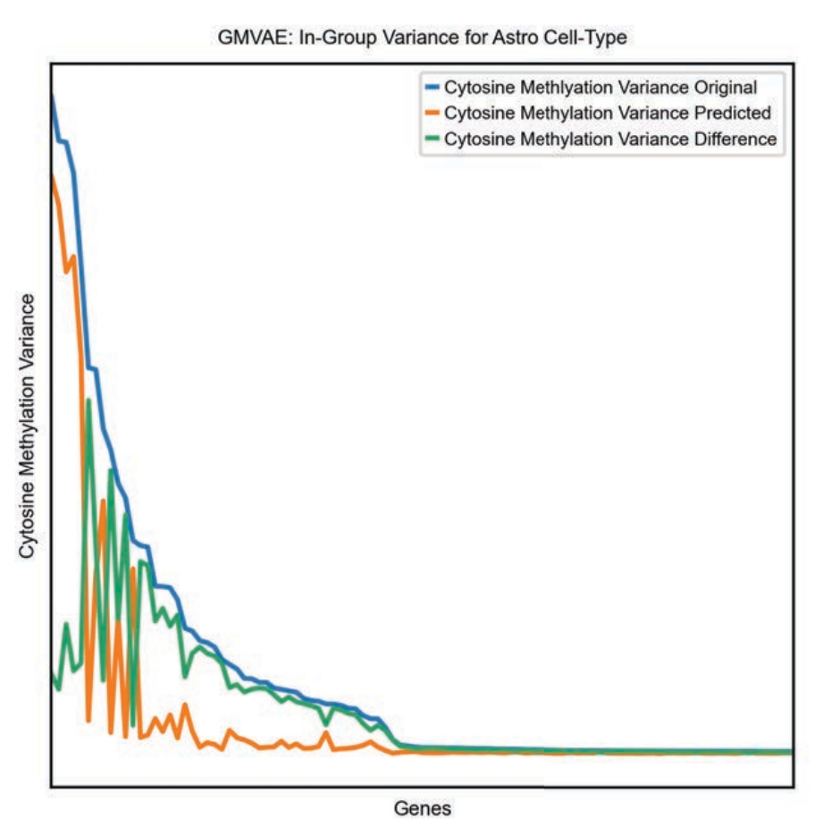
Astro Cell-Type Variation



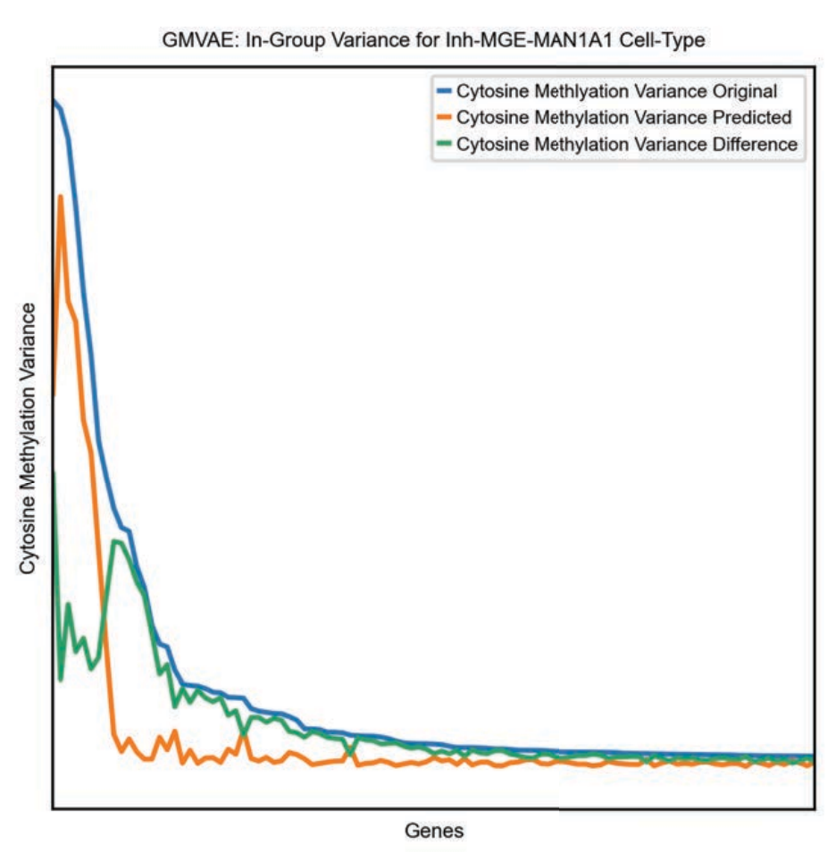
Inh-MGE-MAN1A1 Cell-Type Variation



EXC-L4-RORB Cell-Type Variation



Astro Cell-Type Variation



Inh-MGE-MAN1A1 Cell-Type Variation

Notes for Improvement

1. ***Priority Number One:*** UMAP is not always a meaningful method for determining biologically relevant information for it tends to distort the original manifold.UMAP suffers from noise and outliers, parameter sensitivity, and most importantly the balance tradeoff between local and global features. As shown in the analysis our models perform modestly on the high-variance features but when it comes to the low-variance features our methods suffer from extremely high dimensionality and sparse data density. Our goal is to be able to take advantage of neural networks as a means to preserve global features displayed by high-variance features and use them to condition the low-variance features to limit the distortion exhibited by other dimensionality reduction techniques.
2. ***Priority Number Two:*** Validating batch correction via marker gene expression and imputation accuracy.
3. ***Priority Number Three:*** Determining highly variable features and lowly variable features plays a substantial role in the performance of our models. For further analysis, we would work towards using different thresholds and methods for determining our high variance features.

References

"Introduction to Epigenetics - Learn.Omicslogic.Com." *YouTube*, 5 Dec. 2018, www.youtube.com/watch?v=IAu44BkOaSs&ab_channel=OmicsLogic.

Weng, Lilian. "From Autoencoder to Beta-Vae." *Lil'Log (Alt + H)*, 12 Aug. 2018, lilianweng.github.io/posts/2018-08-12-vae/.

Semi-Supervised Gaussian Mixture Variational Autoencoder for Pulse ..., www.researchgate.net/publication/360773339_Semi-Supervised_Gaussian_Mixture_Variational_Autoencoder_for_Pulse_Shape Accessed 7 Aug. 2023.

Github: https://github.com/HUNTERJCARROLL