Project Report for BIOINF576: Tool Development for Bioinformatics

adVAE: Alzheimer's Data Variational Autoencoder

Varnika Umashankar

Department of Computational Medicine and Bioinformatics, University of Michigan – Ann Arbor.

Abstract

Introduction: Alzheimer's research faces significant challenges due to limited access to high-quality, multimodal biomedical data. adVAE is a Python-based software package designed to generate synthetic gene expression and MRI data using Variational Autoencoders (VAEs), helping researchers augment datasets in low-data settings.

Methods: The tool introduces a modular pipeline for preprocessing, training, generation, and evaluation designed for both gene expression and MRI data. Models are implemented using PyTorch and evaluated using evaluation and statistical metrics.

Results: Trained on real datasets from AMP-AD and OASIS-1, adVAE effectively reconstructed and generated biologically plausible data with high correlation to ground truth, showing promising results for downstream augmentation.

Discussion: These results support the biological validity and usability of the generated data in Alzheimer's-specific research. This tool has the potential to assist with data augmentation, simulation, and exploratory analysis in low-data biomedical settings.

Conclusion: Moving forward, adVAE will incorporate validation/test splits, support EEG data, and explore more expressive generative models. The tool demonstrates clear potential as a flexible solution for synthetic biomedical data generation in the Alzheimer's research domain.

Availability: The software tool implementing the methods and its source code are freely available at https://github.com/varnikaumashankar/adVAE.git

Contact: varnika@umich.edu

1 Introduction

Alzheimer's disease is a progressive neurodegenerative disease that primarily affects memory, thinking and other cognitive abilities. Its increasing incidence among the older population necessitates a need for more comprehensive research to understand its underlying mechanisms and develop effective diagnostic and therapeutic strategies.

Amidst rapid technological development, the use of artificial intelligence has significantly expanded into the field of medical diagnostics, encompassing areas such as the analysis of medical images, drug development, design of personalized treatment plans, and disease prediction and treatment. Deep learning, which is an important branch in the field of artificial intelligence, is playing a key role in solving several medical challenges by creating models to aid the early detection, diagnosis, and treatment of Alzheimer's disease.

However, a major challenge in training deep learning models is their reliance on large datasets. Researchers aiming to model or predict disease progression, simulate patient data, or enhance diagnostic tools using deep learning models face a significant bottleneck due to the lack of sufficient training data.

While some datasets like Washington University's Open Access Series of Imaging Studies (OASIS) and Harvard Medical School's Harvard Aging Brain Study (HABS) are publicly accessible, richer resources like Alzheimer's Disease Neuroimaging Initiative (ADNI) and Accelerating Medicines Partnership Program for Alzheimer's Disease (AMP-AD) require extensive approval processes, including data use agreements and institutional background checks which can pose as barriers that can hinder smaller research efforts.

Moreover, while tools leveraging Variational Autoencoders (VAE) for data augmentation do exist, there remains a lack of streamlined solutions specifically tailored to handle the diverse data modalities associated with Alzheimer's disease. This highlights the need for a tool that can generate high-quality, synthetic multimodal data to support Alzheimer's research despite limited access to real-world samples.

To address this gap, adVAE (Alzheimer's Data Variational Autoencoder) - a Python-based software package designed to generate synthetic biomedical data using Variational Autoencoders was developed. It supports two pipelines - one for gene expression data and another for MRI brain scans, and is capable of preprocessing raw data, training specialized VAE models, evaluating reconstructions, visualizing latent representations, and generating new synthetic samples. The package is lightweight, modular, and easy to run via the command line interface, making it suitable for both researchers and developers working in data-scarce environments to integrate into existing bioinformatics and biomedical pipelines.

2 Methods

2.1 Overview

The adVAE software package implements a self-contained, modular, extensible framework built around Variational Autoencoders for the generation and evaluation of synthetic biomedical data. It currently supports two data modality pipelines - gene expression (RNA microarray) and structural brain MRI slices. The core objective is to aid data augmentation for Alzheimer's research, particularly under data-scarce conditions.

The package was developed in a Python 3.10.14 environment using a variety of scientific, machine learning, and domain-specific libraries. Core development was managed using the Conda package manager, with PyTorch 2.2.0 serving as the primary framework for building and training the VAE models. For data handling and preprocessing, libraries such as NumPy, pandas, and scikit-learn were used. Visualization and analysis were performed using Matplotlib and Seaborn. Domain-specific tools included NiBabel, which was used to load and process MRI images, among others.

This package is modular, with components for data preprocessing, training, generating, and evaluating synthetic data. They are separated across folders like data_preprocessing, models, training, visualisation, metrics, and utils. A centralized main.py script enables users to execute any stage of either pipeline through command-line arguments. Configuration is controlled via a config.py module that sets hyperparameters and file paths. [Fig. 1]

2.2 Input and Output

For each supported modality, the tool takes as input:

- A real dataset - gene expression profiles (RNA microarray data) from AD individuals (in .tsv format) or 3D neuroimaging MRI brain scans (in .img format) preprocessed into a standardized format.
- Model configuration parameters (hidden dimensions, latent dimensions, learning rate, etc.).

The output includes:

- Trained model weights.
- Reconstructed samples and evaluation plots.
- Synthetic samples generated from the latent space.
- Evaluation metrics comparing the original samples with the reconstructed and synthetic samples.
- Distribution comparison plots between real and synthetic data.

2.3 Functionality

Refer to [Fig. 2 and 3] for an overview of the gene expression and MRI data pipelines.

Preprocessing

Gene expression: Data is loaded, missing values are handled, duplicate values are removed, and probe-level data is optionally aggregated to the gene level. Features are then scaled (standard/min-max) and reduced via PCA to retain 95% of the variance. The reduced matrix is saved alongside the PCA and scaler objects for inversion during synthesis.

MRI: Alzheimer's patient IDs are selected from metadata and corresponding 3D structural MRI volumes are loaded using NiBabel. 20 transverse slices are extracted within a target axial range, resized to 128×128 pixels, normalized to [-1, 1], and stored as grayscale tensors.

Model Training

A VAE is trained using reconstruction loss (MSE) and KL divergence with optional beta weighting. The model is trained on the entire dataset due to limited data availability.

GeneExpressionVAE: A feedforward encoder-decoder network.

MRIVAE: A convolutional architecture with residual blocks and transposed convolutions for decoding.

Evaluation

Gene expression: Reconstruction accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson's Correlation are used to assess how closely the VAE could replicate original gene expression profiles. Metrics obtained from distribution-based tests (Kolmogorov-Smirnov and T-test) are used to statistically compare real and reconstructed data distributions, while latent space plots and histograms help visualize how well the reconstructed data preserved the structure and spread of the real dataset.

MRI: Apart from the above-mentioned metrices, Root Mean Squared Error (RMSE), Structural Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR) are used to assess the spatial and perceptual quality of reconstructed and generated images. These image-specific metrics provided insight into how well the model retained anatomical features and visual realism.

Synthetic Data Generation

Gene expression: Samples are drawn from the standard normal latent space and decoded. Inverse PCA and scaling steps are applied to recover gene expression data, and the decoded samples are saved in .tsv format.

MRI: Latent samples are drawn from the learned posteriors of the real data. Decoded samples are saved as .gif for visual inspection.

Synthetic Data Validation

Synthetic and real data distributions are compared using histogram plots. The same evaluation metrics and summary statistics used to evaluate reconstructed samples are also used to validate realism of synthetic samples.

Grid Search (MRI only)

An optional grid search allows hyperparameter optimization for latent dimensionality, learning rate, and beta values, using reconstruction loss and SSIM as performance metrics.

3 Results

3.1 Gene Expression Pipeline

The GeneExpressionVAE was trained on PCA-reduced RNA microarray data obtained from 20 subjects from the AMP-AD MSBB (Mount Sinai Brain Bank) dataset, publicly available through the AD Knowledge Portal. This dataset is widely used in Alzheimer's research and provides access to well-annotated gene expression profiles relevant to disease progression. The data was processed to preserve 95% of the original variance across ~40,000 gene-level probes. Over 100 epochs, the model showed stable convergence, with the training loss settling at 0.5462 and reconstruction accuracy plateauing around 0.3826.

The reconstructed data demonstrated strong similarity to the original samples. Evaluation metrics reported a MAE of 3.3474, MSE of 34.8205, and a Pearson correlation of 0.9939. Statistical validation using T-test resulted in p = 0.5586, while the KS test showed a small but significant distributional deviation (KS stat = 0.0078, p = 1.41e-07).

Upon generating 5000 synthetic gene expression profiles, the model maintained high fidelity to real data. The synthetic and real samples had a MAE of 0.4746, a Pearson correlation of 1.0, and p-values p = 0.5209 from the T-test and p = 0.0546 from the KS test.

Table 1. Results of the Gene Expression Pipeline

Metric	Reconstructed Data	Synthetic Data
Reconstruction Accuracy	0.3929	
Mean Absolute Error	3.3474	0.4746
Mean Squared Error	34.8205	0.4918
Mean Pearson Correlation	0.9939	1.00
T-test p-value	0.5586	0.05461
KS test p-value	1.415e-07	0.5209

3.2 MRI Pipeline

The MRI VAE was trained on 2D transverse slices derived from Gain-field corrected, atlas-registered MRI scans from ~100 subjects from the OASIS-1 dataset (WashU OASIS Brains). This dataset includes brain scans of individuals across the Alzheimer's progression spectrum. Training over 10 epochs resulted in convergence, with the loss reaching 634.5533 and reconstruction accuracy plateauing at 0.6100.

The model achieved a MAE of 0.1037, MSE of 0.0217, and RMSE of 0.1469. Reconstructed images exhibited SSIM = 0.5869 and PSNR = 16.6856. However, both T-test (p ~ 0) and KS test (stat = 0.1654, p ~ 0) revealed statistically significant deviations in pixel distribution compared to the originals.

50 synthetic MRI images were also generated from the learned posterior distributions. These images had MAE = 0.0404, MSE = 0.0031, and RMSE = 0.0556. The PSNR dropped to 11.44 and SSIM fell to 0.3617. Strong distributional divergence was again confirmed by T-test (stat = 52.71, p ~ 0) and KS test (stat = 0.1654, p ~ 0).

Table 2. Results of the MRI Pipeline

Metric	Reconstructed Data	Synthetic Data
Reconstruction Accuracy	0.6138	
Mean Absolute Error	0.1037	0.0404
Mean Squared Error	0.0217	0.0031
Root Mean Squared Error	0.1469	0.0556
Structural Similarity Index Measure	0.5869	0.3617
Peak Signal to Noise Ratio	16.6856	11.4376
T-test p-value	0	0
KS test p-value	0	0

4 Discussion

In the gene expression pipeline, it can be seen that the model is learning the global patterns in the gene expression data well, as demonstrated by the strong similarity of reconstructed samples to the original samples. While reconstruction accuracy remains moderate, this is an anticipated challenge as we're dealing with high-dimensional biological data where small differences across thousands of features can accumulate. Still, the high correlation between the reconstructed and original data, along with the strong significance of the distribution-based tests shows promise. The generated gene expression data maintained high fidelity to the real data. The high correlation and the p-values from the distribution-based tests indicate strong alignment between the original and synthetic data distributions which suggest that adVAE can potentially serve as a reliable tool for gene expression data simulation and augmentation, especially when real data is scarce.

In the MRI pipeline, it can be seen that the model has captured coarse structural and intensity patterns of Alzheimer's-affected brain slices, though some fine-grained details were lost in reconstruction and synthesis. The fact that the perceptual quality of generated images was notably lower than that of reconstructions, and the statistically significant deviations in pixel distribution of reconstructions and generated images compared to the originals indicates room for improvement in preserving anatomical fidelity. Nonetheless, the model demonstrates stable training, reasonable image similarity metrics, and consistent generation of plausible 2D slices. These findings highlight the applicability of adVAE in medical imaging scenarios where acquiring large volumes of high-quality scans is difficult.

Future iterations of adVAE will aim to expand support for additional data types such as Electroencephalogram data (EEG), while incorporating proper validation

and test splits to improve generalizability. Further work will also explore more expressive generative architectures like Generative Adversarial Networks (GAN) and domain-specific priors to boost realism and interpretability across all data modalities. With continued development, and access to a larger and more comprehensive dataset, adVAE has the potential to become a valuable tool in biomedical research pipelines, enabling more robust model training from limited datasets.

5 Conclusion

adVAE presents an end-to-end modular, flexible, and integrable software framework for generating synthetic biomedical data using Variational Autoencoders - especially for contexts like Alzheimer's research where data is often limited. Even with modest datasets, the results show that the models can learn meaningful representations and produce biologically plausible outputs. This early version of adVAE lays the foundation for a tool that can support data augmentation, simulation, and exploratory analysis in low-data biomedical settings. With continued development, adVAE has the potential to become a practical and impactful tool in Alzheimer's research, supporting robust model development and data-driven discovery in settings where real-world data is scarce.

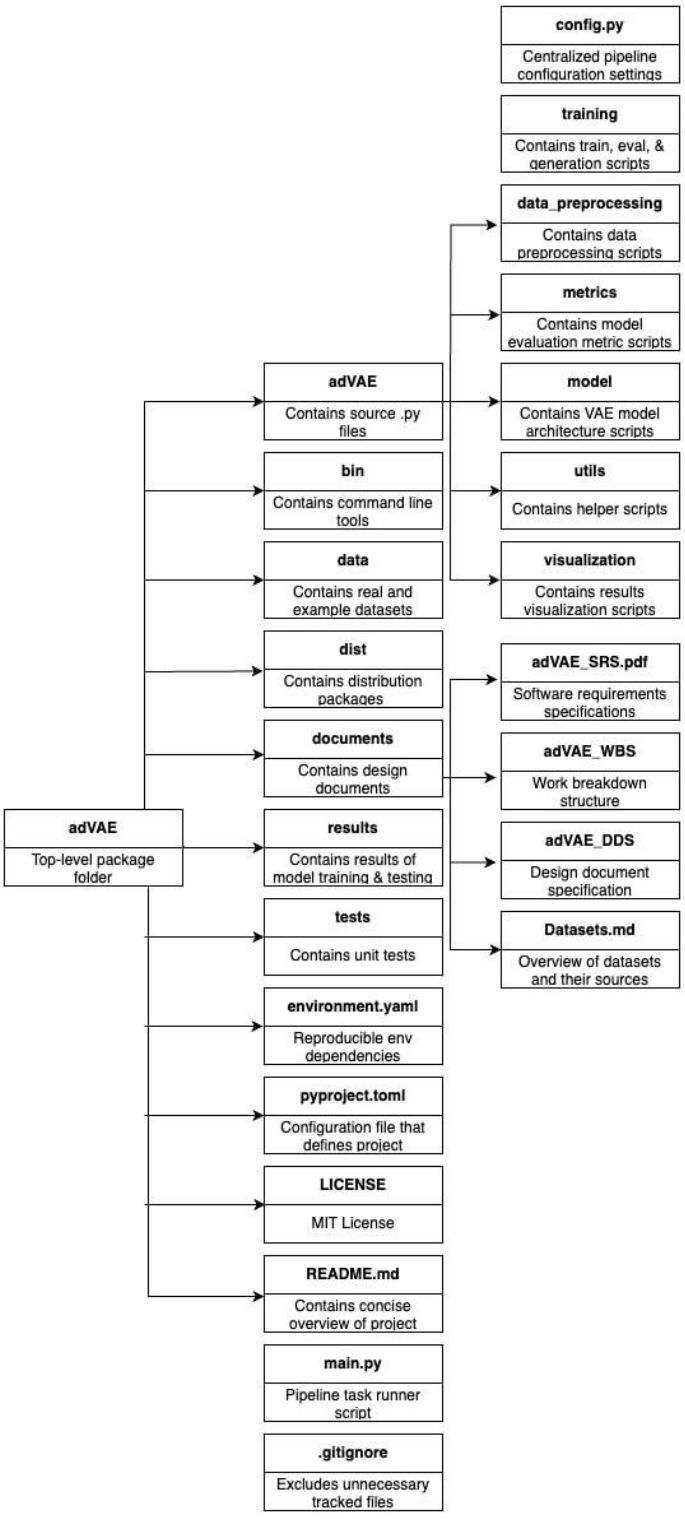
6 References

Gautier, V., Bousse, A., Sureau, F., Comtat, C., Maxim, V., & Sixou, B. (2024). Bimodal PET/MRI generative reconstruction based on VAE architectures. *Physics in Medicine & Biology*, *69*(24), 245019. doi:10.1088/1361-6560/ad9133

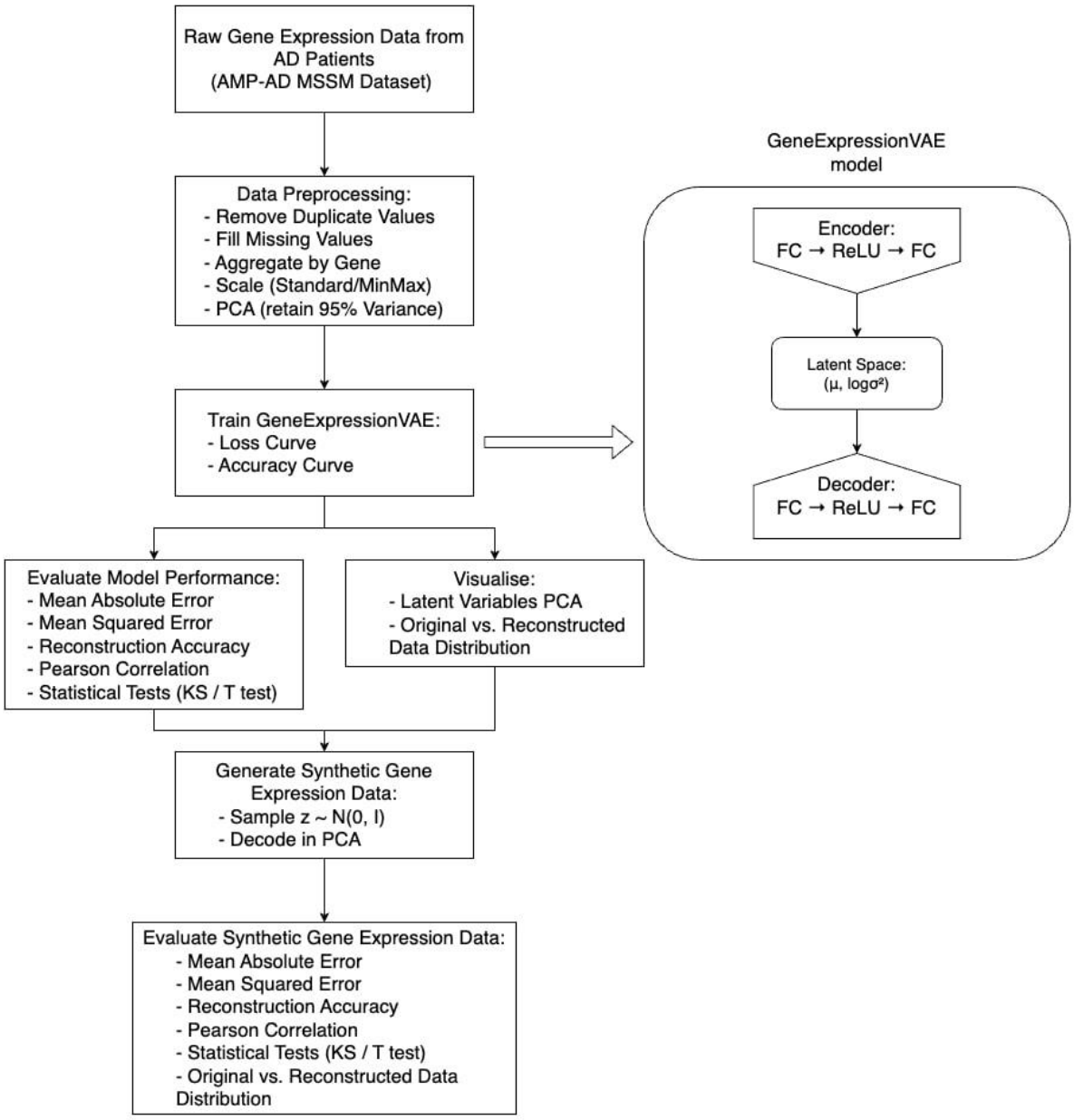
Karlberg, B., Kirchgaessner, R., Lee, J., Peterkort, M., Beckman, L., Goecks, J., & Ellrott, K. (2024). SyntheVAEiser: augmenting traditional machine learning methods with VAE-based gene expression sample generation for improved cancer subtype predictions. Genome biology, 25(1), 309. https://doi.org/10.1186/s13059-024-03431-3

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 23, 80–91.

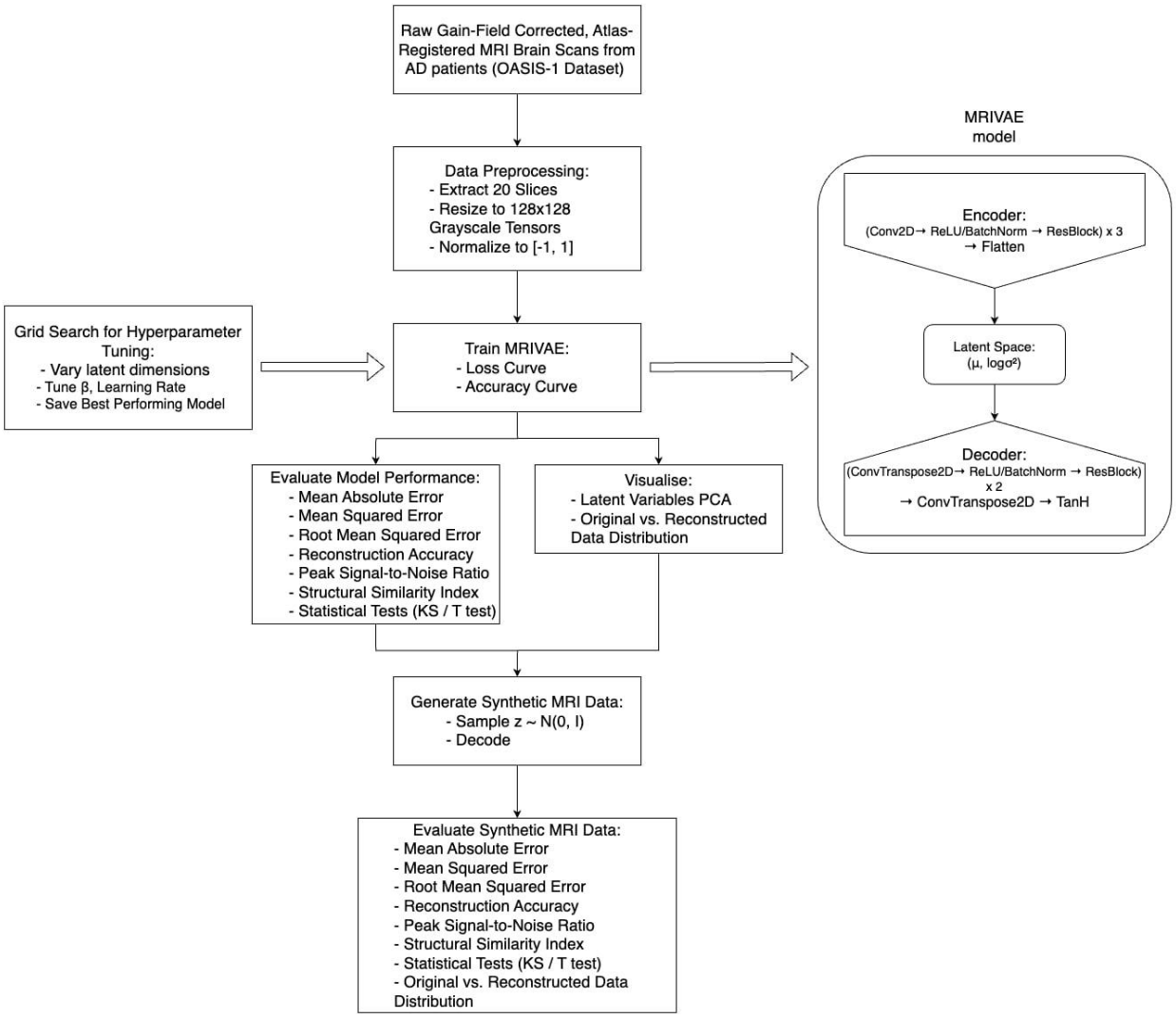
Way, G.P., Zietz, M., Rubinetti, V. et al. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol 21, 109 (2020). https://doi.org/10.1186/s13059-020-02021-3



[Fig. 1 adVAE Module Structure]



[Fig. 2 GeneExpressionVAE Model Architecture and Pipeline]



[Fig. 3 MRIVAE Model Architecture and Pipeline]