

UnCOT-AD: Unpaired Cross-Omics Translation Enables Multi-Omics Integration for Alzheimer's Disease Prediction

Abrar Rahman Abir¹, Sajib Acharjee Dip², and Liqing Zhang^{2,*}

¹ Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

² Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, 24060, VA, United States
1qzhang@cs.vt.edu

Abstract. Alzheimer's Disease (AD) is a progressive neurodegenerative disorder, posing a growing public health challenge. Traditional machine learning models for AD prediction have relied on single omics data or phenotypic assessments, limiting their ability to capture the disease's molecular complexity and resulting in poor performance. Recent advances in high-throughput multi-omics have provided deeper biological insights. However, due to the scarcity of paired omics datasets, existing multi-omics AD prediction models rely on unpaired omics data, where different omics profiles are combined without being derived from the same biological sample, leading to biologically less meaningful pairings and causing less accurate predictions. To address these issues, we propose **UnCOT-AD**, a novel deep learning framework for **Unpaired Cross-Omics Translation** enabling effective multi-omics integration for **AD** prediction. Our method introduces the first-ever cross-omics translation model trained on unpaired omics datasets, using two coupled Variational Autoencoders and a novel cycle consistency mechanism to ensure accurate bidirectional translation between omics types. We integrate adversarial training to ensure that the generated omics profiles are biologically realistic. Moreover, we employ contrastive learning to capture the disease specific patterns in latent space to make the cross-omics translation more accurate and biologically relevant. We rigorously validate UnCOT-AD on both cross-omics translation and AD prediction tasks. Results show that **UnCOT-AD** empowers multi-omics based AD prediction by combining real omics profiles with corresponding omics profiles generated by our cross-omics translation module and achieves state-of-the-art performance in accuracy and robustness. Source code is available at <https://github.com/abrarrahmanabir/UnCOT-AD>

Keywords: Alzheimer's Disease · Multi-omics Integration · Cross Omics Translation.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that primarily affects the elderly and is characterized by cognitive decline, memory loss, and ultimately, a loss of bodily functions. The number of AD cases is expected to increase significantly in the coming decades, creating a serious public health challenge globally [1]. Despite numerous studies attempting to identify molecular risk factors involved in AD pathogenesis, the precise mechanisms underlying AD occurrence and progression remain poorly understood. Current treatments for AD can only alleviate symptoms without addressing the root causes of the disease [2,3].

Research has primarily focused on analyzing phenotypic data, such as magnetic resonance imaging (MRI) and neuropsychological assessments [4]. Biomarkers associated with AD pathology, such as β -amyloid deposition and tau proteins, have been explored as well, with some recent studies incorporating these biomarkers for more accurate diagnoses [5,6]. However, a significant limitation in many existing AD prediction models is their reliance on single omics, rather than integrating multi-omics data to capture the complexity of AD. Recent advances have allowed for the collection of multi-omics data, which provide a more comprehensive view of biological systems [7,8,9,10,11]. A few studies proposed methods that utilize multi-omics data for prediction. For example, gene expression and DNA methylation data were combined to predict AD [12,13,6]. As paired multi-omics data from the same group of people is not available, the studies used all possible pairs of gene expression profiles from one group of people and DNA methylation profiles from another group of people as surrogate of paired data to predict AD. Using all possible pairs of gene expression and DNA methylation profiles can introduce biologically irrelevant and unrealistic combinations, which adversely affects prediction performance.

To address the challenges posed by the scarcity of paired multi-omics data and the limitations of current multi-omics based AD prediction methods, we propose a novel deep learning framework, **UnCOT-AD** for **Unpaired Cross-Omics Translation** and multi-omics integration for **AD** prediction. UnCOT-AD performs cross-omics translation using unpaired training datasets, unlike state-of-the-art models such as BABEL [14] and Polarbear [15] that require paired data. Our major contributions are:

1. **Cross-Omics Translation From Unpaired Data:** We introduce a novel Cross-Omics Translation module to perform cross-omics translation using unpaired omics datasets. To the best of our knowledge, this is the first work of cross-omics translation trained on unpaired data. This method allows us to map between different omics types, such as gene expression and DNA methylation, and generate one omics profile from another, even when direct pairings between the two omics types are not available while training. This approach addresses a significant gap in multi-omics integration by enabling biologically meaningful data generation across modalities.
2. **Multi-Omics Based AD Prediction Using Generated Paired Omics Data :** Using the cross-omics translation module, we are able to perform multi-omics-based AD prediction even in the absence of fully paired datasets. By generating a corresponding omics profile (e.g., DNA methylation) from real omics data (e.g., gene expression), we effectively create paired multi-omics data and then fuse the two modalities with our prediction module for AD prediction.

Additionally, as our cross-omics translation module is designed to be compatible with any two omics types and our translation and prediction modules are separate, our method can be applied solely for cross-omics translation between two modalities, even in the absence of paired data. We rigorously validate UnCOT-AD on both cross-omics translation performance and AD prediction tasks, achieving state-of-the-art results in accuracy and robustness.

2 Methodology

Our method is divided into two steps. In the first step, a Cross-Omics Translation Module is trained on unpaired omics data, meaning there is no one-to-one correspondence between samples from two different types of omics datasets. The translation module learns a bidirectional mapping such that, at inference, it can take one omics profile as input and translate it to the corresponding profile in the other omics. Next, a Prediction Module combines the real omics profile and the translated omics profile, predicted from the real omics using the pretrained translation module, to predict AD.

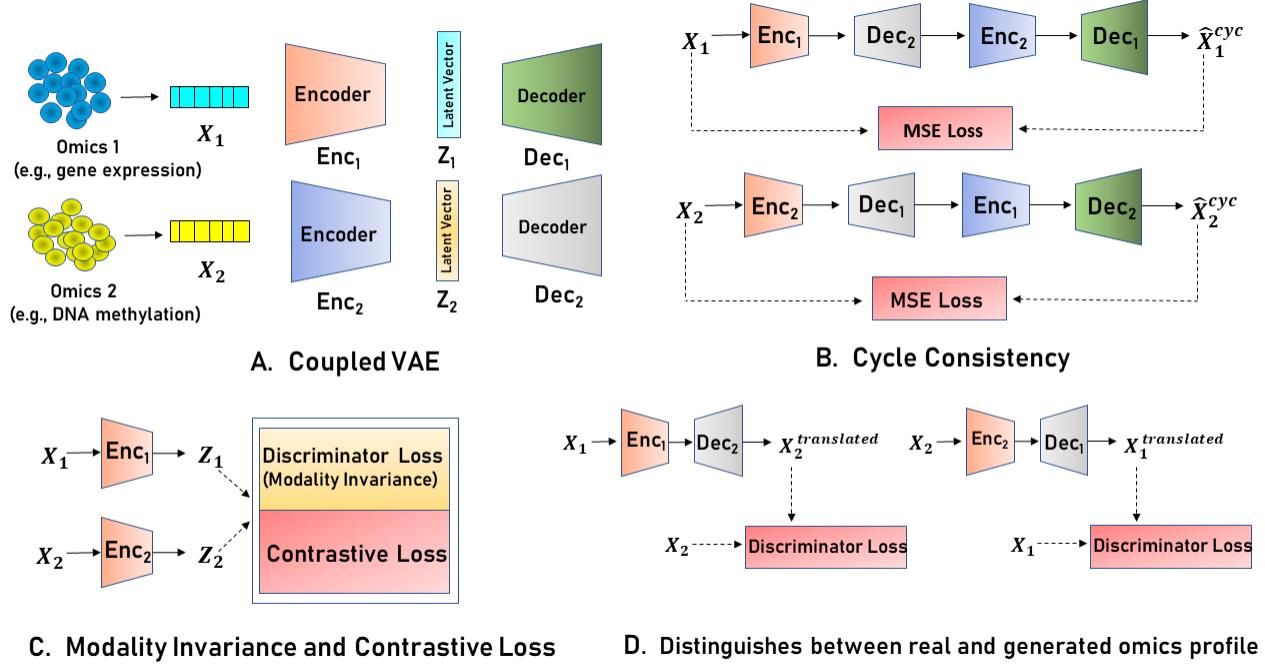


Fig. 1: A. Coupled VAE architecture that takes two different types of unpaired omics vectors denoted as X_1 and X_2 . B. Cycle Consistency Mechanism. C. Latent representations of both omics are trained to be modality invariant through a discriminator loss and an additional contrastive loss pushes the model to learn AD-specific patterns. D. Two additional discriminators attempt to distinguish between generated and real omics profiles, while the model aims to "fool" these discriminators, ensuring the generated profiles appear biologically realistic.

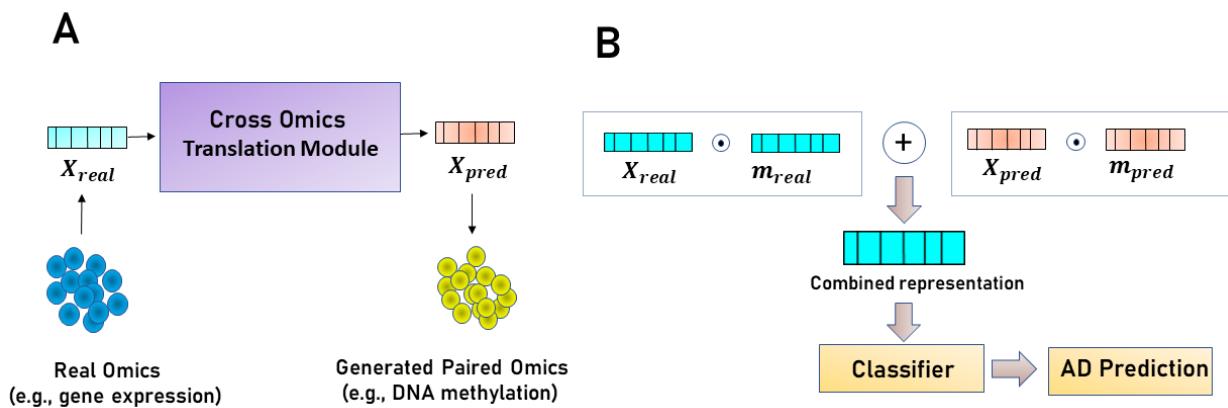


Fig. 2: A. Cross Omics Translation Module generates corresponding paired omics profile (e.g, gene expression) from real omics profile (e.g, DNA methylation or proteomics). B. Architecture of prediction module. X_{real} and X_{pred} are feature vectors of the real and generated omics and m_{real} and m_{pred} are learnable weight vectors. \odot is element-wise multiplication.

2.1 Cross-Omics Translation Module

The goal of Cross-Omics Translation Module is to learn bidirectional mappings between two omics data types, denoted as $\mathbf{x}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{d_2}$ where the given training samples are unpaired (Figure 1). d_1 and d_2 represent the respective dimensionalities of the two omics data types. The empirical data distributions of the two types are denoted as $\mathbf{x}_1 \sim P_{\text{data1}}(\mathbf{x}_1)$ and $\mathbf{x}_2 \sim P_{\text{data2}}(\mathbf{x}_2)$. We employ two separate Variational Autoencoders (VAEs)[16]. Despite being unpaired, to ensure consistent translations between the two modalities, we incorporate a cycle consistency mechanism. Additionally, three adversarial discriminators are utilized to enforce modality invariance in the latent space and generate biologically realistic omics profiles. To capture AD-specific features, a contrastive loss is introduced to encourage the separation of AD and control samples in the latent space.

Variational Autoencoder (VAE) : Each VAE models the latent variable distributions for the two omics types. The encoder for each VAE approximates the posterior distribution of the latent variable \mathbf{z} given the input data \mathbf{x} . Specifically, for omics type \mathbf{x}_1 , the encoder $E_{\theta_1} : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^l$ approximates $q_{\phi_1}(\mathbf{z}_1|\mathbf{x}_1)$, where l is the dimensionality of the latent space. The posterior is assumed to be a multivariate Gaussian distribution with a diagonal covariance structure: $q_{\phi_1}(\mathbf{z}_1|\mathbf{x}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ where the encoder outputs $\boldsymbol{\mu}_1$ (the mean) and $\log \boldsymbol{\sigma}_1^2$ (the log-variance). Similarly, for omics type \mathbf{x}_2 , the encoder $E_{\theta_2} : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^l$ approximates $q_{\phi_2}(\mathbf{z}_2|\mathbf{x}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$. To sample from the latent distribution, we employ the reparameterization trick, which ensures differentiability by expressing \mathbf{z} as a function of the mean, variance, and a noise term ϵ drawn from a standard normal distribution: $\mathbf{z}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\sigma}_1 \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$, where $\boldsymbol{\sigma}_1 = \exp(0.5 \log \boldsymbol{\sigma}_1^2)$. Similarly, for \mathbf{z}_2 : $\mathbf{z}_2 = \boldsymbol{\mu}_2 + \boldsymbol{\sigma}_2 \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$, where $\boldsymbol{\sigma}_2 = \exp(0.5 \log \boldsymbol{\sigma}_2^2)$. The latent variables \mathbf{z}_1 and \mathbf{z}_2 are then decoded by D_{ϕ_1} and D_{ϕ_2} to reconstruct the input omics data:

$$\hat{\mathbf{x}}_1 = D_{\phi_1}(\mathbf{z}_1), \quad \hat{\mathbf{x}}_2 = D_{\phi_2}(\mathbf{z}_2).$$

Here, $D_{\phi_1} : \mathbb{R}^l \rightarrow \mathbb{R}^{d_1}$ and $D_{\phi_2} : \mathbb{R}^l \rightarrow \mathbb{R}^{d_2}$ represent the decoders that reconstruct the input. To ensure that the learned posterior $q_{\phi_1}(\mathbf{z}_1|\mathbf{x}_1)$ is close to the prior $p(\mathbf{z}_1) = \mathcal{N}(0, I)$, a KL divergence regularizer is applied. The KL divergence for each VAE is computed as:

$$KL(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^l (1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2)$$

Cycle Consistency Loss : Without having a direct paired mapping between the two modalities, the model might fail to properly align the features of one omics type with the other. To overcome this limitation, we incorporate a **cycle consistency loss** to enforce bidirectional mapping between \mathbf{x}_1 and \mathbf{x}_2 . We sample $\mathbf{x}_1 \sim P_{\text{data1}}(\mathbf{x}_1)$, encode it to obtain \mathbf{z}_1 , and then decode \mathbf{z}_1 into the other modality as : $\hat{\mathbf{x}}_2 = D_{\phi_2}(\mathbf{z}_1)$. The translated data $\hat{\mathbf{x}}_2$ is re-encoded using E_{θ_2} to obtain $\mathbf{z}'_2 = E_{\theta_2}(\hat{\mathbf{x}}_2)$, which is then decoded back to the original modality:

$$\hat{\mathbf{x}}_1^{(cyc)} = D_{\phi_1}(E_{\theta_2}(\hat{\mathbf{x}}_2)).$$

This process ensures that \mathbf{x}_1 can be recovered through the translation cycle. Similarly, the same process is applied to \mathbf{x}_2 . The **cycle consistency loss** is formulated as:

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}_{\mathbf{x}_1 \sim P_{\text{data1}}} \|\mathbf{x}_1 - \hat{\mathbf{x}}_1^{(cyc)}\|^2 + \mathbb{E}_{\mathbf{x}_2 \sim P_{\text{data2}}} \|\mathbf{x}_2 - \hat{\mathbf{x}}_2^{(cyc)}\|^2.$$

This loss encourages consistency between the two omics types, ensuring that even though the data is unpaired, the model can translate accurately in both directions and preserve the essential features of the original data. By enforcing this cycle consistency, we mitigate drift between the two modalities and effectively push the model to learn to generate paired data.

Adversarial Loss : In addition to cycle consistency, we utilize three adversarial discriminators to ensure modality invariance in the latent space and to enforce that the generated omics profiles are realistic. The first discriminator, D_{ψ} , operates in the latent space to ensure that \mathbf{z}_1 and \mathbf{z}_2 are indistinguishable, enforcing modality invariance. Modality invariance in latent space is crucial because we use the latent space of one omics type to reconstruct the other omics profile through the opposite decoder. The adversarial loss for modality invariance is:

$$\mathcal{L}_{D_{\text{mod}}} = -\mathbb{E}[\log D_{\psi}(\mathbf{z}_1)] - \mathbb{E}[\log(1 - D_{\psi}(\mathbf{z}_2))].$$

The second and third discriminators, D_{ψ_1} and D_{ψ_2} , ensure that the generated omics data from D_{ϕ_1} and D_{ϕ_2} , respectively, are indistinguishable from real data. For \mathbf{x}_1 , the adversarial loss is given by:

$$\mathcal{L}_{D_1} = -\mathbb{E}_{\mathbf{x}_1 \sim P_{\text{data1}}} [\log D_{\psi_1}(\mathbf{x}_1)] - \mathbb{E}_{\hat{\mathbf{x}}_1 \sim P_{\text{gen1}}} [\log(1 - D_{\psi_1}(\hat{\mathbf{x}}_1))].$$

Similarly, for \mathbf{x}_2 , the adversarial loss is:

$$\mathcal{L}_{D_2} = -\mathbb{E}_{\mathbf{x}_2 \sim P_{\text{data2}}} [\log D_{\psi_2}(\mathbf{x}_2)] - \mathbb{E}_{\hat{\mathbf{x}}_2 \sim P_{\text{gen2}}} [\log(1 - D_{\psi_2}(\hat{\mathbf{x}}_2))].$$

P_{gen1} and P_{gen2} represent the distributions of the generated omics profiles from the cross-omics translation. These adversarial losses ensure that the generated omics profiles remain biologically plausible and realistic, and that the latent space enforces modality invariance.

Contrastive Loss : To capture AD-specific features in the latent space, we employ a **contrastive loss** that encourages latent representations of omics data from the same class (AD or control) to be closer in the latent space. For a pair of latent variables \mathbf{z}_1 and \mathbf{z}_2 which represent positive pairs meaning samples with the same label $y_1 = y_2 \in \{0, 1\}$, the contrastive loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\mathbf{z}_1^\top \mathbf{z}_2 / \tau)}{\sum_j \exp(\mathbf{z}_1^\top \mathbf{z}_j / \tau)},$$

where τ is a temperature parameter that controls the concentration of the distribution and j runs over all samples in the batch, including both positive and negative samples.

Overall Objective : The total loss for this translation module is defined as :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{cycle}} \mathcal{L}_{\text{cycle}} + \lambda_{\text{adv}} (\mathcal{L}_{D_{\text{mod}}} + \mathcal{L}_{D_1} + \mathcal{L}_{D_2}) + \lambda_{\text{contrastive}} \mathcal{L}_{\text{contrastive}} + \lambda_{\text{KL}} (\text{KL}_1 + \text{KL}_2).$$

where $\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{x}_1 \sim P_{\text{data1}}} \|\mathbf{x}_1 - D_{\phi_1}(\mathbf{z}_1)\|^2 + \mathbb{E}_{\mathbf{x}_2 \sim P_{\text{data2}}} \|\mathbf{x}_2 - D_{\phi_2}(\mathbf{z}_2)\|^2$ and lambdas control the relative importance of each loss component. This formulation ensures that the translated omics data remains biologically realistic, AD-specific features are captured, and the unpaired nature of the data is properly handled through cycle consistency.

2.2 Prediction Module

Let $\mathbf{x}_r \in \mathbb{R}^{d_r}$ represent the real omics profile of one type, and $\hat{\mathbf{x}}_p \in \mathbb{R}^{d_p}$ represent the predicted omics profile of another type, generated by the pretrained translation module. d_r and d_p represent the respective dimensionalities of the real and predicted omics types. The pretrained Cross-Omics Translation Module T maps the real omics profile to the predicted omics profile as follows: $\hat{\mathbf{x}}_p = T(\mathbf{x}_r)$, where T is the mapping function learned during the translation training phase. The prediction module combines these two profiles, \mathbf{x}_r and $\hat{\mathbf{x}}_p$, to perform AD classification (Figure 2). The fusion of the real and predicted omics profiles is performed by projecting each profile into a lower dimensional space of same dimension. The real omics profile \mathbf{x}_r and the predicted omics profile $\hat{\mathbf{x}}_p$ are separately projected using linear transformations: $\mathbf{g}_r = \mathbf{W}_r \mathbf{x}_r$ and $\mathbf{g}_p = \mathbf{W}_p \hat{\mathbf{x}}_p$, where $\mathbf{g}_r \in \mathbb{R}^{d'}$ and $\mathbf{g}_p \in \mathbb{R}^{d'}$ are the projected representations, and $\mathbf{W}_r \in \mathbb{R}^{d' \times d_r}$, $\mathbf{W}_p \in \mathbb{R}^{d' \times d_p}$ are learnable projection matrices. To dynamically weigh the contributions of each omics profile type, we apply learnable element-wise weights \mathbf{m}_r and \mathbf{m}_p : $\mathbf{h}_r = \mathbf{g}_r \odot \mathbf{m}_r$ and $\mathbf{h}_p = \mathbf{g}_p \odot \mathbf{m}_p$, where \odot denotes element-wise multiplication. Both \mathbf{m}_r and \mathbf{m}_p are of same dimension d' .

The fused latent representation is then computed as:

$$\mathbf{z}_{\text{fused}} = \mathbf{h}_r + \mathbf{h}_p$$

This fused representation $\mathbf{z}_{\text{fused}} \in \mathbb{R}^{d'}$ captures complementary information from both the real and predicted omics profiles of different types, and it will be used for AD classification. The fused representation $\mathbf{z}_{\text{fused}}$ is passed through a multi-layer classifier to predict the probability of Alzheimer's Disease. The classifier is composed of multiple fully connected layers with ReLU activations and batch normalization to ensure robust learning. The classifier computes the prediction as follows: $\hat{y} = \sigma(\mathcal{C}(\mathbf{z}_{\text{fused}}))$, where \mathcal{C} is the classifier's transformation and $\sigma(\cdot)$ is the sigmoid activation function, which outputs $\hat{y} \in [0, 1]$. The prediction module is trained to minimize the binary cross-entropy loss between the predicted probability \hat{y} and the true label $y \in \{0, 1\}$ for Alzheimer's Disease.

3 Experiments

We validated the performance and effectiveness of UnCOT-AD by conducting different experiments. We divided the experiments in two parts - one is AD prediction performance by integrating multi-omics data and the other is cross-omics translation performance analysis. For the experiments, we consider three different types of omics - gene expression, DNA methylation and proteomics. We performed 5 fold cross validation for each of the experiments.

3.1 Dataset

We collected the preprocessed gene expression and DNA methylation dataset from [6] and proteomics dataset from [17]. The gene expression data were obtained by integrating GSE33000 [18] and GSE44770 [2], containing 257 normal and 439 AD samples. DNA methylation data were collected from GSE80970 [19], comprising 68 normal and 74 AD samples. Proteomics dataset is comprised of 328 AD and 91 normal samples. Differentially expressed genes (DEGs) and differentially methylated positions (DMPs) were identified by filtering with P-value < 0.01 for DEGs, and P-value < 0.01 for DMPs and P-value < 0.05 for differentially expressed proteins(DEPs). Finally we got 200 DEGs, 500 DMPs and 696 DEPs. Detailed preprocessing steps can be found in [6,17].

3.2 UnCOT-AD Improves AD Prediction by Integrating Multi-Omics Data Over Single Omics

Initially, we trained individual classifiers on three separate omics datasets (gene expression, DNA methylation, and proteomics) to predict Alzheimer's Disease (AD). Next, we trained three Cross-Omics Translation Modules: one for gene-proteomics translation, one for gene-DNA methylation translation, and another for DNA methylation-proteomics translation, all in a bidirectional manner. After training the translation modules, we generated predictions by combining real and translated omics data. For example, we took the real gene expression data and used the pretrained gene-to-proteomics translation module to generate the corresponding proteomics profile. The real gene expression data and the predicted proteomics profile were then passed into the prediction module for AD prediction. We repeated this process for all possible pairs of three different omics.

Table 1: Performance Metrics for AD Prediction

Omics Type	Accuracy	Precision	Recall	F1 Score	MCC
Multi Omics					
Gene+Predicted DNA Methylation	0.9498	0.9743	0.9453	0.9594	0.8950
Gene+Predicted Protein	0.9427	0.9662	0.9429	0.9540	0.8792
Protein+Predicted Gene	0.9189	0.9487	0.9484	0.9474	0.7759
Protein+Predicted DNA Methylation	0.9429	0.9958	0.9303	0.9587	0.8842
DNA Methylation+Predicted Protein	0.8828	0.8857	0.8800	0.8828	0.7657
DNA Methylation+Predicted Gene	0.8759	0.8842	0.9067	0.8941	0.7469
Single Omics					
Gene	0.8765	0.8922	0.9157	0.9035	0.7335
DNA Methylation	0.8335	0.8312	0.8667	0.8474	0.6674
Protein	0.8904	0.9703	0.8879	0.9232	0.7636

Table 1 presents the results of AD prediction using both single omics and integrated multi-omics data. The performance of gene expression and proteomics, with accuracies of 0.8765 and 0.8904 respectively, suggests that these modalities offer substantial predictive power when used alone. However, DNA methylation, with an accuracy of 0.8335, performs less effectively on its own, indicating that it may not capture the full complexity

of AD on its own. Despite these reasonable results for single omics datasets, relying on a single data source limits the ability to fully capture the diverse biological signals related to AD.

The integration of multi-omics data provides both notable quantitative improvements and biological insights into AD prediction. Combining gene expression with predicted DNA methylation achieves an 8.4% accuracy improvement over gene expression alone and a 13.9% improvement over DNA methylation alone, suggesting that transcriptional activity coupled with epigenetic information provides better understanding of AD mechanisms. Likewise, gene expression combined with predicted proteomics yields a 7.5% improvement over gene expression alone and 5.9% over proteomics alone, highlighting the complementary roles of transcriptional and protein-level data. Protein expression with predicted gene expression reaches an accuracy of 0.9189, underscoring how upstream genetic regulation benefits protein-level prediction. Integrating protein expression with predicted DNA methylation or DNA methylation with predicted proteomics improves accuracy by 5.9% and 5.1%, respectively, reinforcing that combining epigenetic information with proteomics leads to a more comprehensive model for AD prediction. We also report the Matthews Correlation Coefficient (MCC) to provide a more comprehensive evaluation of the model's performance [20]. For binary classification tasks, MCC is particularly useful as it takes into account the balance between true positives, false positives, true negatives, and false negatives. The highest MCC value is observed for the integration of gene expression with predicted DNA methylation (MCC of 0.8950) and the next is protein with predicted DNA methylation (MCC of 0.8842). These results indicate that the multi-omics models perform robustly across all confusion matrix categories. We observe that the trend of improvement after integrating multi-omics data over single omics is consistent across all pairs of omics in every evaluation metric.

Table 2: Comparison of UnCOT-AD with Baseline Models on Multi-Omics AD Prediction

	Gene + DM			Gene+Protein			Protein+DM		
	Acc.	F1	MCC	Acc.	F1	MCC	Acc.	F1	MCC
Abbas et al.	0.8689	0.8712	0.7153	0.8932	0.8890	0.7415	0.8307	0.8412	0.7153
Mahendran et al.	0.8402	0.8549	0.6913	0.8805	0.8548	0.6892	0.8091	0.8180	0.6483
Park et al.	0.8611	0.8710	0.7099	0.8896	0.8581	0.7452	0.8457	0.8310	0.7359
UnCOT-AD				Acc.	F1	MCC			
Gene + Predicted DNA Methylation				0.9498	0.9594	0.8950			
DNA Methylation + Predicted Gene				0.8759	0.8941	0.7469			
Gene + Predicted Protein				0.9427	0.9540	0.8792			
Protein + Predicted Gene				0.9189	0.9474	0.7759			
Protein + Predicted DNA Methylation				0.9429	0.9587	0.8842			
DNA Methylation + Predicted Protein				0.8828	0.8828	0.7657			

3.3 AD Prediction Performance Comparison with Baseline Models

We compared the performance of UnCOT-AD on multi-omics based AD prediction with three baseline models: Abbas et al., Mahendran et al., and Park et al.[12,13,6]. All of them used gene expression and DNA methylation data. So, we trained their models with proteomics data as well and conducted a thorough performance comparison. However, none of the baseline models utilized paired omics data. They used all possible pairs of gene expression and DNA methylation profile for each label - normal and AD. We address this limitation in UnCOT-AD by cross omics translation which gives us paired data. To show the effectiveness of UnCOT-AD, we compared the baseline's performance using gene expression and DNA methylation with our results from both gene expression + predicted DNA methylation and DNA methylation + predicted gene expression. From Table 2, we observe that UnCOT-AD gives better performance in both cases compared to the baseline's performance. This allowed us to show the added value of our cross-omics translation method. Similarly, we extended this comparison across all possible pairs of omics. We observe that UnCOT-AD consistently outperforms the baseline models across all evaluation metrics. The key factor behind this

improvement is the use of our cross-omics translation module, which effectively generates paired omics data from unpaired datasets.

3.4 UnCOT-AD Captures Alzheimer’s Disease-Specific Patterns

To validate that UnCOT-AD successfully captures Alzheimer’s Disease-specific patterns, we take a real omics profile (e.g., gene expression) and use the cross-omics translation module to generate the corresponding omics profile (e.g., DNA methylation or proteomics). The generated profile is then passed through its respective encoder, and the resulting latent representations are plotted using t-SNE. From Figure 3, we observe that AD and normal samples are well-separated on the latent space, as intended by the contrastive loss used during training. This approach shows us UnCOT-AD effectively captures disease-specific patterns in the latent space. However, we notice that the ones generated from DNA methylation (Figure 3d,3f) struggle to capture AD specific features resulting in more sparsed distribution of AD and normal samples. On the other hand, the ones generated from gene expression and proteomics show well clustered AD and normal samples.

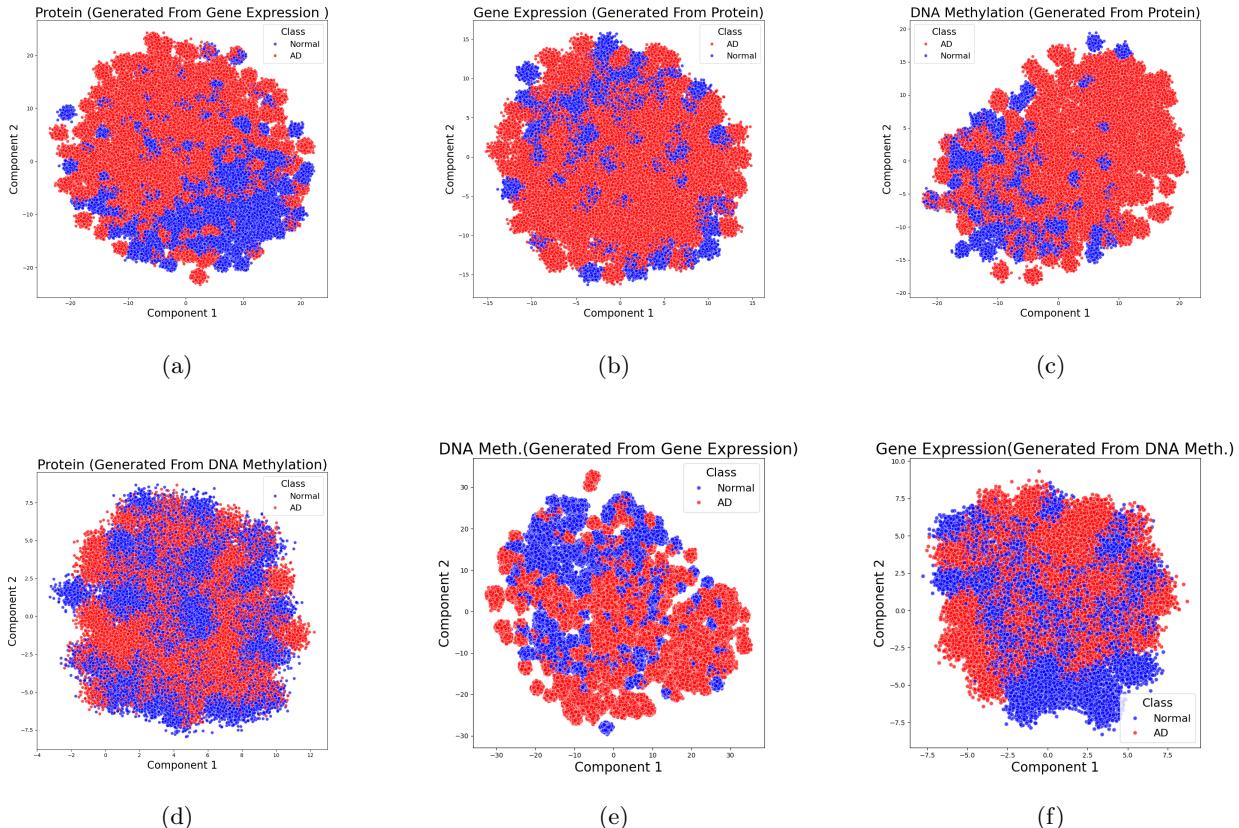


Fig. 3: t-SNE visualization of translated omics profiles generated from real omics, with AD and control samples shown in different colors.

3.5 Cross Omics Translation Performance

To assess the quality of cross-omics translation, researchers commonly use correlation-based metrics such as Pearson or Spearman correlations. However, these methods are typically applied in paired omics translation, where ground truth for cross-omics data is available for each sample. In our unpaired scenario, where no

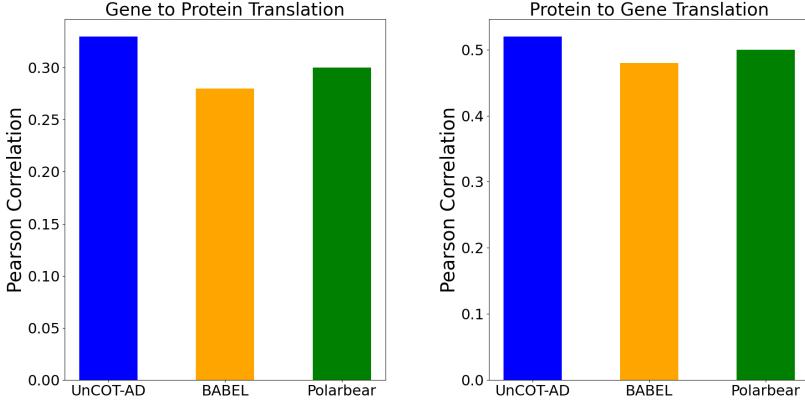


Fig. 4: Comparison of Pearson correlation between real and predicted omics profiles across UnCOT-AD, BABEL, and Polarbear for gene-to-protein and protein-to-gene translations.

such ground truth exists for the translated omics profiles, correlation-based evaluation is not applicable. To address this challenge and effectively assess the performance of unpaired cross-omics translation, we propose two evaluation metrics: Cycle Consistency Loss (CCL) and Fréchet Omics Distance (FOD). CCL measures how well the original omics data can be reconstructed after cross-translation between different omics modalities, serving as a key indicator of the preservation of relevant features during translation. A low CCL indicates that the translated omics profile successfully preserves the underlying structure of the original data, suggesting that the translation process generates omics data closely aligned with, or effectively "paired" to, the original modality. A lower FOD quantifies the lower distributional difference between real-world omics data and the omics profile generated through cross-omics translation, capturing how closely the generated omics data aligns with the real omics profile in the latent space.

Cycle Consistency Loss (CCL) : The Cycle Consistency Loss (CCL) of X_1 to X_2 ($X_1 \rightarrow X_2$) is computed as : $CCL = \|X_1 - \hat{X}_1\|^2$. where X_1 represents the input omics data from modality 1. First, X_1 is passed through the encoder Enc_1 of modality 1 to obtain its latent representation. This is then decoded using the decoder Dec_2 of modality 2 to generate the predicted omics \hat{X}_2 . The predicted omics \hat{X}_2 is passed through the encoder Enc_2 , followed by the decoder Dec_1 , resulting in the reconstructed data \hat{X}_1 . The CCL is then calculated as the mean squared error (MSE) between the original X_1 and the reconstructed \hat{X}_1 .

Fréchet Omics Distance (FOD) : Given two omics X_1 and X_2 , let Z_{X_1} represent the latent space of X_1 obtained via the encoder of the VAE corresponding to X_1 , denoted as $Z_{X_1} = Enc_{X_1}(X_1)$. The translated omics \hat{X}_2 (from X_1) is generated by decoding Z_{X_1} using the decoder of the VAE corresponding to X_2 , denoted $\hat{X}_2 = Dec_{X_2}(Z_{X_1})$. The real omics X_2 and The translated omics \hat{X}_2 are both passed through the encoder of the VAE of X_2 , denoted as Enc_{X_2} , to obtain their respective latent representations $Z_{X_2} = Enc_{X_2}(X_2)$, $Z_{\hat{X}_2} = Enc_{X_2}(\hat{X}_2)$. The Fréchet Omics Distance (FOD) between X_2 and \hat{X}_2 is then computed as:

$$FOD(X_2, \hat{X}_2) = \|\mu_{Z_{X_2}} - \mu_{Z_{\hat{X}_2}}\|^2 + \text{Tr}(\Sigma_{Z_{X_2}} + \Sigma_{Z_{\hat{X}_2}} - 2(\Sigma_{Z_{X_2}} \Sigma_{Z_{\hat{X}_2}})^{\frac{1}{2}})$$

where $\mu_{Z_{X_2}}$ and $\Sigma_{Z_{X_2}}$ are the mean and covariance of the latent representations of real X_2 , and $\mu_{Z_{\hat{X}_2}}$ and $\Sigma_{Z_{\hat{X}_2}}$ are the corresponding statistics for the predicted \hat{X}_2 .

Performance Analysis : We compared the bidirectional cross omics translation between gene expression, DNA methylation and proteomics. As baselines, we employed CycleGAN (adapted for omics data) [21], an autoencoder based architecture where we replace VAE with autoencoder, and a random baseline. From Table 3, we observe that UnCOT-AD significantly outperforms all the baselines for all pairs of omics which shows the effectiveness of our approach. Moreover, we notice the lowest CCL for translation between gene expression and proteomics. The cross modal translation involving DNA methylation with both gene expression and

Table 3: Performance on Unpaired Cross Omics Translation

	UnCOT-AD		CycleGAN		Autoencoder		Random Baseline	
	CCL	FOD	CCL	FOD	CCL	FOD	CCL	FOD
Gene → Protein	0.0476	0.5404	0.1254	1.2543	0.1804	2.2015	3.2831	4.7534
Protein → Gene	0.0396	1.3961	0.1209	2.1543	0.1754	2.8054	2.9103	3.6742
Gene → DM	0.1509	4.2545	0.2304	5.1230	0.2800	5.8056	4.5702	6.1223
DM → Gene	0.1070	4.0135	0.1953	4.8502	0.2456	5.6743	4.1209	6.0054
DM → Protein	0.1169	4.1921	0.2105	5.1344	0.2603	5.9801	4.3658	6.2917
Protein → DM	0.1513	3.8759	0.2356	4.7523	0.2901	5.3624	4.4503	5.8421

protein demonstrates higher CCL compared to gene and protein. This same pattern is observed with FOD as well where cross modal translation involving DNA methylation yields higher FOD compared to gene and protein. These results highlight the strong biological relationship between gene expression and proteins and the challenge of capturing complex pattern of DNA methylation. From Section 3.2, we also notice that DNA methylation shows less predictive power compared to the other two which aligns with the trend observed in this experiment.

3.6 UnCOT-AD Shows Superior Performance in Paired Cross-Omics Translation

To further demonstrate the effectiveness of UnCOT-AD, we evaluated its performance in traditional paired scenario with two state-of-the-art models, BABEL [14] and Polarbear [15], both designed specifically for paired data training. We evaluated on two paired gene expression and proteomics datasets, which were originally sourced from the TCGA BRCA cohort and made available in preprocessed form by [22]. The evaluation was conducted based on Pearson correlation between real and predicted omics profiles. UnCOT-AD consistently outperformed both BABEL and Polarbear across translation tasks in both directions: gene-to-protein and protein-to-gene (Figure 4). This result highlights UnCOT-AD’s robustness and generalizability in both paired and unpaired training scenarios.

3.7 Ablation Study

We performed an ablation study to evaluate the impact of key components in our framework: the Adversarial Loss and Cycle Consistency Mechanism on cross-omics translation, and the contribution of the Contrastive Loss in AD prediction. From Figure 5a, we observe that removing the cycle consistency mechanism significantly increases the CCL while affecting the FOD to a lesser extent. This is because the cycle consistency mechanism is crucial for generating omics profiles that can accurately reconstruct the input omics. The absence of this mechanism leads to poor alignment between the translated and original omics profiles, increasing the CCL. On the other hand, FOD is more related to generating realistic omics profiles. Thus, removing the adversarial loss leads to substantial increase in FOD, as the adversarial components helps ensure the generated omics data resemble the true distribution of omics profiles. Therefore, to achieve optimal performance in both metrics (CCL and FOD), both the cycle consistency mechanism and adversarial loss are necessary. Additionally, we validated the importance of learning AD-specific features in the cross-omics translation module. From Figure 5b, we find that when the contrastive loss is excluded, the MCC of AD prediction drops for every pair of omics data. This highlights the essential role of contrastive loss in encouraging the separation of AD and control samples in the latent space.

4 Conclusion

In this paper, we introduce UnCOT-AD, a novel approach to address the challenge of unpaired multi-omics data integration for Alzheimer’s Disease (AD) prediction. For the first time, we introduce a cross-omics translation module that allows for the generation of paired omics profiles, even when the training data from two modalities are unpaired. By combining the generated profiles with real omics data, we perform

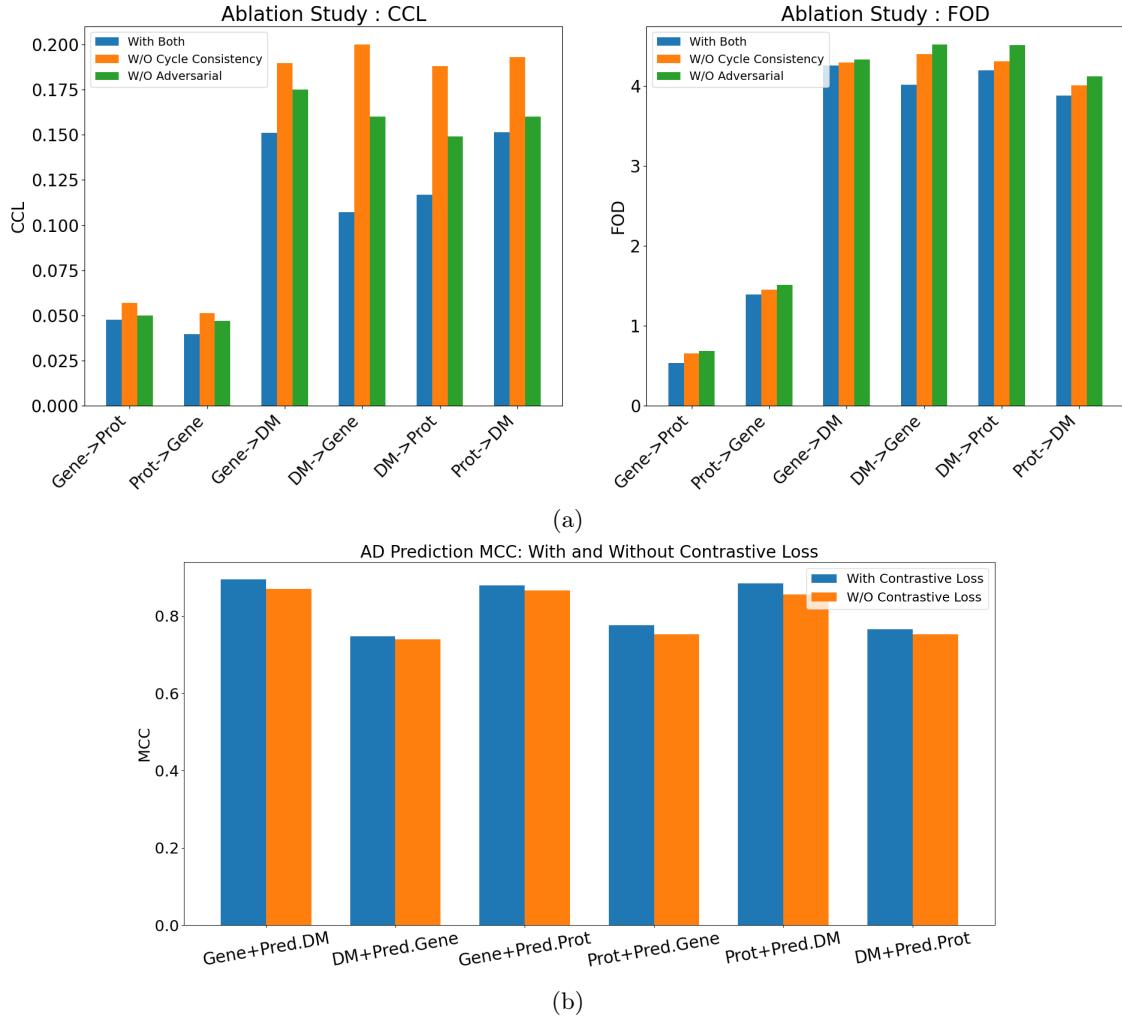


Fig. 5: Visualization of the Contribution of (a) Adversarial Loss, Cycle Consistency and (b) Contrastive Loss in Cross Omics Translation and AD Prediction.

multi-omics based AD prediction, overcoming the limitations of existing approaches that rely solely on single omics. Our method achieves state-of-the-art results in both cross-omics translation and AD prediction tasks, demonstrating significant improvements in all evaluation metrics compared to existing approaches. By generating biologically meaningful omics profiles and effectively integrating them, our framework has proven to be a robust solution for multi-omics integration in AD prediction. Thus, UnCOT-AD presents new potential by effectively combining multiple biological data modalities, enabling a more comprehensive understanding of complex diseases like Alzheimer's Disease. Our cross-omics translation module is designed to be compatible with any two omics types. Additionally, as our translation and prediction modules are separate, our method can be applied solely for cross-omics translation between two modalities, even in the absence of paired data.

References

1. Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H Michael Arrighi. Forecasting the global burden of alzheimer's disease. *Alzheimer's & dementia*, 3(3):186–191, 2007.
2. Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.
3. Chihyun Park, Youngmi Yoon, Min Oh, Seok Jong Yu, and Jaegyo Ahn. Systematic identification of differential gene network to elucidate alzheimer's disease. *Expert Systems with Applications*, 85:249–260, 2017.
4. Jin San Lee, Changsoo Kim, Jeong-Hyeon Shin, Hanna Cho, Dae-seock Shin, Nakyoung Kim, Hee Jin Kim, Yeshin Kim, Samuel N Lockhart, Duk L Na, et al. Machine learning-based individual assessment of cortical atrophy pattern in alzheimer's disease spectrum: development of the classifier and longitudinal evaluation. *Scientific reports*, 8(1):4161, 2018.
5. Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.
6. Chihyun Park, Jihwan Ha, and Sanghyun Park. Prediction of alzheimer's disease based on deep neural network by integrating gene expression and dna methylation dataset. *Expert Systems with Applications*, 140:112873, 2020.
7. Shichao Huang, Jianxin Mao, Kan Ding, Yue Zhou, Xianglu Zeng, Wenjuan Yang, Peipei Wang, Cun Zhao, Jian Yao, Peng Xia, et al. Polysaccharides from ganoderma lucidum promote cognitive function and neural progenitor proliferation in mouse model of alzheimer's disease. *Stem cell reports*, 8(1):84–94, 2017.
8. Pradeep Kodam, R Sai Swaroop, Sai Sanwid Pradhan, Venketesh Sivaramakrishnan, and Ramakrishna Vadrevu. Integrated multi-omics analysis of alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets. *Scientific reports*, 13(1):3695, 2023.
9. Sujith Kumar Pulukool, Sai Krishna Srimadh Bhagavatham, Vishnu Kannan, Brahmanaidu Parim, Sivateja Challa, Vidya Karnatam, Datta Darshan VM, Irfan Ahmad Mir, Piruthivi Sukumar, Vijayalakshmi Venkateshan, et al. Elevated atp, cytokines and potential microglial inflammation distinguish exfoliation glaucoma from exfoliation syndrome. *Cytokine*, 151:155807, 2022.
10. R Sai Swaroop, Sai Sanwid Pradhan, VM Datta Darshan, Kanikaram Sai Phalguna, and Venketesh Sivaramakrishnan. Integrated network pharmacology approach shows a potential role of ginseng catechins and ginsenosides in modulating protein aggregation in amyotrophic lateral sclerosis. *3 Biotech*, 12(12):333, 2022.
11. Sai Krishna Srimadh Bhagavatham, Prakash Khanchandani, Vishnu Kannan, Damodaram Potikuri, Divya Sridharan, Sujith Kumar Pulukool, Ashwin Ashok Naik, Rajesh Babu Dandamudi, Sai Mangala Divi, Ashish Paragaonkar, et al. Adenosine deaminase modulates metabolic remodeling and orchestrates joint destruction in rheumatoid arthritis. *Scientific Reports*, 11(1):15129, 2021.
12. Zeeshan Abbas, Hilal Tayara, and Kil To Chong. Alzheimer's disease prediction based on continuous feature representation using multi-omics data integration. *Chemometrics and Intelligent Laboratory Systems*, 223:104536, 2022.
13. Nivedhitha Mahendran and Durai Raj Vincent PM. Deep belief network-based approach for detecting alzheimer's disease using the multi-omics data. *Computational and Structural Biotechnology Journal*, 21:1651–1660, 2023.
14. Kevin E Wu, Kathryn E Yost, Howard Y Chang, and James Zou. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021.
15. Ran Zhang, Laetitia Meng-Papaxanthos, Jean-Philippe Vert, and William Stafford Noble. Semi-supervised single-cell cross-modality translation using polarbear. In *International Conference on Research in Computational Molecular Biology*, pages 20–35. Springer, 2022.
16. Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
17. Maxim N Shokhirev and Adiv A Johnson. An integrative machine-learning meta-analysis of high-throughput omics data identifies age-specific hallmarks of alzheimer's disease. *Ageing Research Reviews*, 81:101721, 2022.
18. Manikandan Narayanan, Jimmy L Huynh, Kai Wang, Xia Yang, Seungyeul Yoo, Joshua McElwee, Bin Zhang, Chunsheng Zhang, John R Lamb, Tao Xie, et al. Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Molecular systems biology*, 10(7):743, 2014.
19. Rebecca G Smith, Eilis Hannon, Philip L De Jager, Lori Chibnik, Simon J Lott, Daniel Condliffe, Adam R Smith, Vahram Haroutunian, Claire Troakes, Safa Al-Sarraj, et al. Elevated dna methylation across a 48-kb region spanning the hoxa gene cluster is associated with alzheimer's disease neuropathology. *Alzheimer's & Dementia*, 14(12):1580–1588, 2018.
20. Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

21. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
22. Xiao Li, Jie Ma, Ling Leng, Mingfei Han, Mansheng Li, Fuchu He, and Yunping Zhu. Mogcn: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Frontiers in Genetics*, 13:806842, 2022.