# VAE-Based Multi-Modal Music Clustering with Hybrid Lyrics and Audio Representations

**Sajid Aryan Sami**
CSE425: Neural Networks Project
BRAC University
sajid.aryan.sami@g.bracu.ac.bd

## Abstract

We study unsupervised clustering of hybrid-language music tracks using Variational Autoencoders (VAEs) and multi-modal feature fusion (audio + lyrics). We implement three experimental tiers aligned with the course guidelines: (i) a Basic VAE trained on fixed-length audio features with K-Means clustering in latent space, (ii) a Convolutional VAE trained on mel-spectrograms and evaluated with multiple clustering algorithms plus an audio+lyrics hybrid representation, and (iii) a Hard setting with a disentanglement-oriented $\beta$-VAE, a regular VAE, and an Autoencoder baseline with both unimodal and audio+lyrics hybrid clustering. We report internal clustering metrics (Silhouette, Calinski–Harabasz, Davies–Bouldin) and external alignment with genre labels (ARI, NMI, purity), and include latent-space visualizations and reconstruction examples.

## 1 Introduction

Music similarity and clustering are core problems for organization, recommendation, and retrieval. Unsupervised clustering is challenging because raw audio is high-dimensional and perceptually complex. VAEs [1] offer a principled way to learn compact latent representations that can be clustered with standard algorithms such as K-Means.

This project implements an end-to-end pipeline for VAE-based music clustering, including preprocessing, model training, feature extraction, clustering, quantitative evaluation, and visualization. All experiments were run on a single GPU (NVIDIA GeForce RTX 3070 Ti).

## 2 Related Work

VAEs [1] and their variants are widely used for unsupervised representation learning. Disentanglement-oriented formulations such as $\beta$-VAE increase the KL regularization weight to encourage more factorized latent variables. Learned embeddings are commonly evaluated via downstream clustering (e.g., K-Means, agglomerative clustering) and visualized using dimensionality reduction techniques like t-SNE [2] and UMAP [3]. In music information retrieval, both engineered audio descriptors (e.g., MFCC statistics) and time–frequency representations (mel-spectrograms) are standard, and lyrics can provide complementary semantic cues for multi-modal analysis.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

# 3 Method

## 3.1 Representations

**Feature dataset (Easy/Hard).** We compute MFCC statistics (mean/std/max/min), spectral centroid/rolloff, zero-crossing rate, and chroma. Features are sanitized with `nan_to_num` and standardized.

**Spectrogram dataset (Medium).** We compute log mel-spectrograms and normalize them to $[0, 1]$. Because the ConvVAE downsamples/upsamples by a factor of 16, we pad the time axis to make the width divisible by 16.

**Lyrics dataset (Medium/Hard).** Lyrics are loaded from CSV files under `data/lyrics/`. Since the provided lyrics are not paired to each audio clip, we attach lyric samples by cycling deterministically. We encode lyrics with a bag-of-words vector and reduce its dimension with PCA.

**Caching.** CPU-bound preprocessing (especially librosa) is cached on disk under `data/cache/` to enable reproducible, tractable runs.

## 3.2 Models

**Basic VAE (Easy).** An MLP encoder/decoder VAE trained on fixed-length feature vectors using an ELBO objective with MSE reconstruction and KL regularization.

**ConvVAE (Medium).** A convolutional VAE trained on mel-spectrograms. Latents are extracted for clustering.

**$\beta$-VAE and Regular VAE (Hard).** We train a $\beta$-VAE ($\beta > 1$) to encourage more factorized latents, plus a regular VAE ($\beta = 1$) baseline.

**Autoencoder (Hard baseline).** A deterministic autoencoder with the same latent dimensionality as the VAEs.

## 3.3 Fusion, clustering, evaluation

For Medium and Hard, we form a hybrid representation by concatenating the audio embedding (ConvVAE/VAE/AE latent) with the reduced lyric embedding. We cluster with K-Means, agglomerative clustering, and DBSCAN.

We report internal metrics (Silhouette, Calinski–Harabasz, Davies–Bouldin) and external alignment with genre labels using ARI [4], NMI, and purity.

# 4 Experiments

**Splits and sizes.** Easy uses an 80/20 split over the dataset. Medium uses a random subset (150 items) for ConvVAE training, then 80/20 over that subset. Hard uses a subset of 200 items, then 80/20.

**Hyperparameters.** Easy: latent dim 32, hidden dim 256, Adam learning rate $10^{-4}$, 50 epochs. Medium: latent dim 64, Adam learning rate $10^{-3}$, 50 epochs. Hard: $\beta$-VAE latent dim 64, $\beta = 4$, 100 epochs; regular VAE $\beta = 1$, 50 epochs.

# 5 Results

We summarize quantitative outputs from the generated CSV files in `results/`.

## 5.1 Easy: Basic VAE + K-Means

Table 1 compares VAE+KMeans against a PCA+KMeans baseline.

Table 1: Easy task results (from `results/easy_task_results.csv`).

| Method | Silhouette | Calinski–Harabasz | ARI |
|---|---|---|---|
| VAE + K-Means | 0.3073 | 171.37 | 0.0486 |
| PCA + K-Means | 0.2112 | 55.58 | 0.1523 |

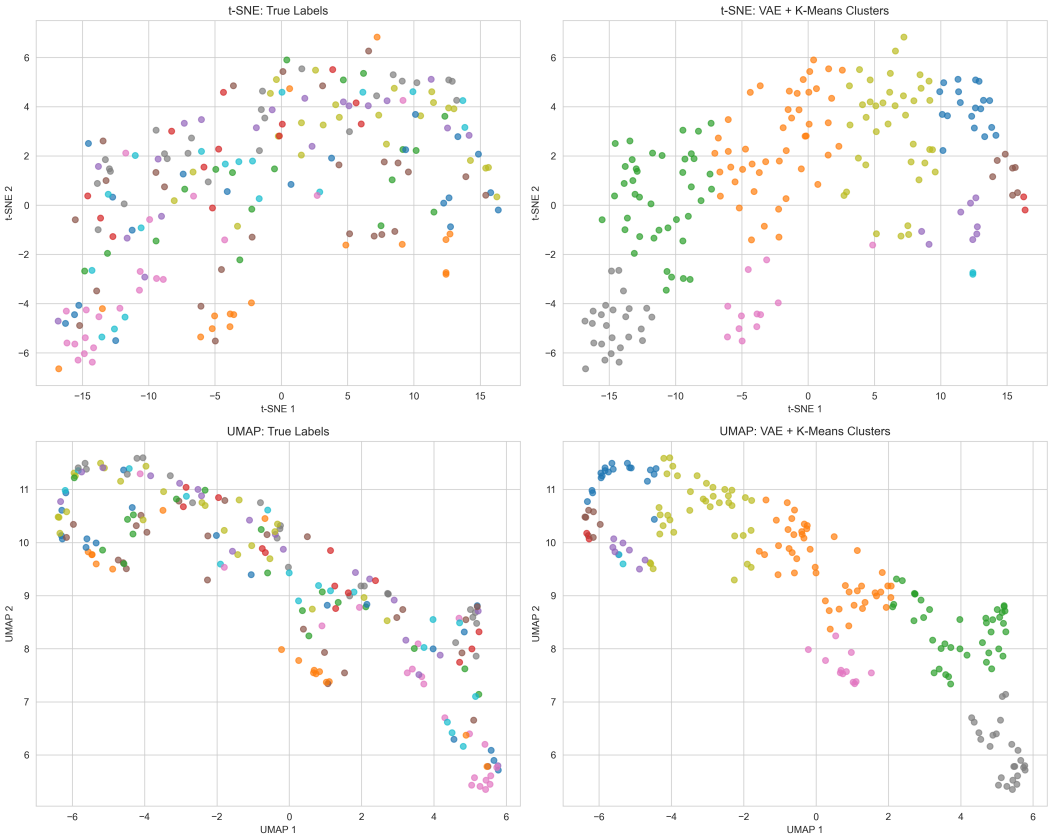


Figure 1: Easy task latent visualization (2D projection of VAE latents; points colored by genre).

### 5.2 Medium: ConvVAE + multiple clustering methods

Table 2 reports results across clustering methods, including the required hybrid audio+lyrics representation. DBSCAN can collapse into degenerate labelings on small samples; in that case, internal metrics can be undefined.

Table 2: Medium task results (from `results/medium_task_results.csv`).

| Method | Silhouette | CH | DB | ARI | NMI |
|---|---|---|---|---|---|
| Agglomerative (audio) | 0.4720 | 143.09 | 0.4708 | 0.0249 | 0.4939 |
| K-Means (audio) | 0.4412 | 135.84 | 0.5019 | 0.0008 | 0.4716 |
| Hybrid Agglomerative (audio+lyrics) | 0.0052 | 3.01 | 1.8434 | 0.1212 | 0.5691 |
| Hybrid K-Means (audio+lyrics) | 0.0051 | 3.06 | 1.8163 | 0.1212 | 0.5691 |
| PCA+KMeans (baseline) | 0.0625 | 4.60 | 1.3887 | 0.1067 | 0.5122 |
| Direct K-Means (baseline) | 0.0625 | 4.60 | 1.3887 | 0.1067 | 0.5122 |
| DBSCAN (audio) | 0.0000 | 0.00 | $\infty$ | 0.0000 | 0.0000 |

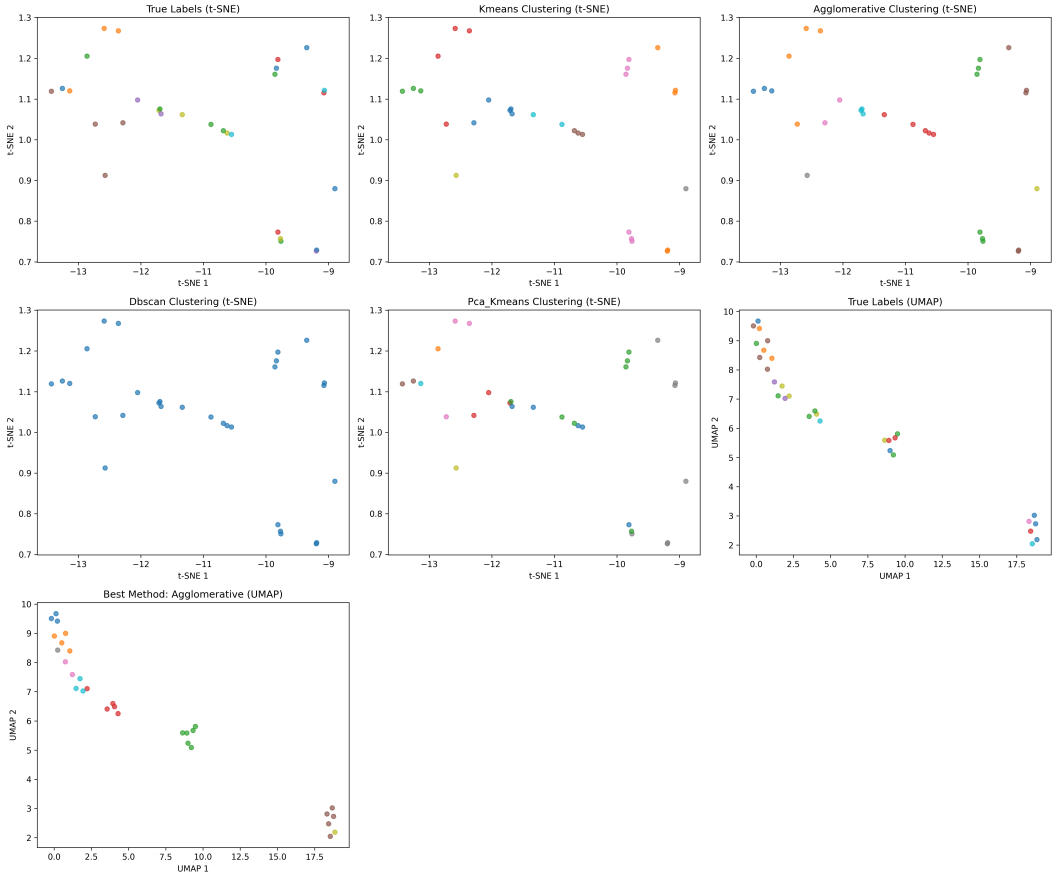


Figure 2: Medium task visualization of ConvVAE latents (2D projection; points colored by genre).

### 5.3 Hard: $\beta$-VAE, baselines, and comprehensive evaluation

Table 3 reports a summary of the strongest settings from the comprehensive run.

Table 3: Hard task summary (top methods) from `results/hard_task_comprehensive_results.csv`.

| Method | Silhouette | ARI | NMI | Purity |
|---|---|---|---|---|
| $\beta$-VAE + Agglomerative | 0.4469 | 0.2352 | 0.3716 | 0.9250 |
| Regular VAE + K-Means | 0.4454 | 0.1209 | 0.2489 | 0.8250 |
| $\beta$-VAE + K-Means | 0.4365 | 0.2348 | 0.3766 | 0.9250 |
| Autoencoder + Agglomerative | 0.4178 | 0.1749 | 0.3856 | 0.9250 |
| Autoencoder + K-Means | 0.3952 | 0.1749 | 0.3856 | 0.9250 |
| Autoencoder Hybrid (audio+lyrics) + K-Means | 0.3333 | 0.1465 | 0.3324 | 0.9000 |
| PCA baseline + K-Means | 0.0659 | 0.1823 | 0.4059 | 0.9500 |

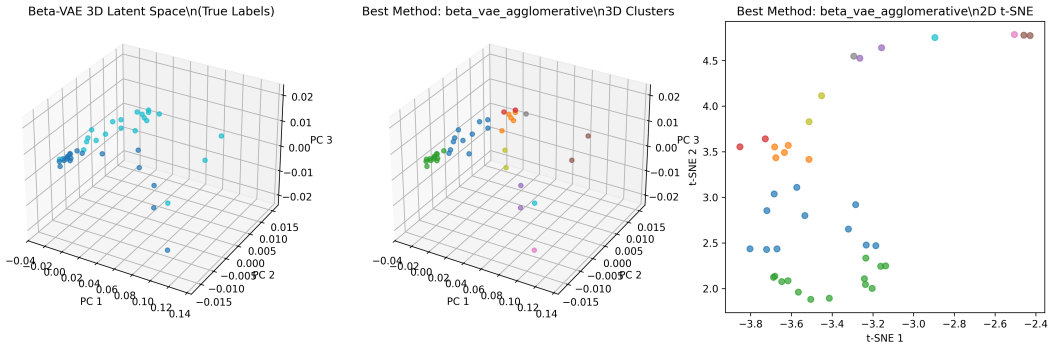


Figure 3: Hard task 3D latent visualization saved by the script.

## 6 Discussion

This section analyzes results for each task and highlights consistent patterns across metrics and modalities.

**Easy task (VAE vs PCA baseline).** In Table 1, the VAE representation improves internal structure (higher Silhouette and Calinski–Harabasz) compared to PCA, suggesting that the learned embedding forms tighter and more separated clusters under Euclidean distance. However, the PCA baseline attains higher ARI, indicating better alignment with the provided genre labels.

This discrepancy is expected when (i) genre labels do not perfectly correspond to geometric clusters in the feature space, (ii) K-Means' spherical-cluster bias mismatches the true class geometry, or (iii) the VAE learns factors of variation that are useful for reconstruction but not strictly label-aligned. The latent plot in Figure 1 is consistent with some structure without clean genre separation.

**Medium task (ConvVAE latents and hybrid fusion).** In Table 2, clustering directly on ConvVAE audio latents yields strong internal metrics (Agglomerative has the best Silhouette/CH and lowest DB). External metrics (ARI) remain low for pure-audio clustering, suggesting that the learned ConvVAE embedding captures acoustically coherent neighborhoods that are not strongly partitioned by genre on the sampled subset.

Hybrid fusion (audio+lyrics) increases ARI and NMI (0.1212 and 0.5691), but internal metrics collapse (Silhouette $\approx 0$, CH $\approx 3$, and high DB). This pattern indicates a scaling/geometry issue in the concatenated space: the fused embedding becomes poorly separated under Euclidean distance, even if it becomes more label-informed. A common mitigation is to re-scale audio vs lyric subspaces (e.g., z-score each block separately or learn a metric), but we keep the baseline fusion simple and reproducible.

DBSCAN degeneracy (all points in one cluster or mostly labeled as noise) yields undefined internal metrics and near-zero external scores; this is a known sensitivity to $\epsilon$, minPts, and feature scaling, especially with small sample sizes.

**Hard task ($\beta$-VAE, regular VAE, autoencoder, and baselines).** The Hard setting shows that a disentanglement-oriented VAE can improve external agreement while maintaining strong internal structure. In Table 3, $\beta$-VAE + Agglomerative achieves the best ARI among the listed top methods while preserving a high Silhouette, suggesting that its embedding better separates label-related factors while remaining cluster-friendly.

The autoencoder baseline is competitive in NMI/purity but generally lower in internal compactness compared to the best VAE variants; this is consistent with deterministic AEs sometimes producing less regularized latent geometry. The PCA baseline has very low Silhouette but high purity; purity can be inflated when the number of clusters matches the number of classes and clusters are imbalanced, so it should be interpreted alongside ARI/NMI.

Hybrid variants in the comprehensive CSV (not all shown in Table 3) can produce higher ARI/NMI while harming Silhouette, including negative Silhouette values, again pointing to metric mismatch in fused spaces without careful normalization.

**Limitations and practical notes.** Medium and Hard use subsets for tractable runtime, increasing variance in external metrics. Results also depend on clustering hyperparameters and feature scaling, especially for DBSCAN and hybrid fusion. Finally, the lyrics encoder is a simple bag-of-words baseline; stronger encoders (e.g., transformer embeddings) could improve fusion but were out of scope.

## 7 Conclusion

We implemented a complete VAE-based music clustering pipeline across Easy/Medium/Hard tiers. Learned latents consistently improve internal clustering structure over simple baselines, while external agreement with genre labels varies by representation and clustering choice. Multi-modal fusion can improve label alignment but requires careful scaling/metric design to preserve geometric separability.

## References


[1] Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[2] van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[3] McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[4] Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.