Self-Supervised Contextual Representation Learning for Transcriptomic Generative AI

Anonymous Authors[1]

Abstract

Bulk RNA sequencing remains central to translational genomics, yet self-supervised foundation models for bulk data have lagged behind single-cell approaches. Existing bulk transformer models couple representation learning to expression magnitudes through discretization or reconstruction objectives, limiting portability across normalization schemes and cohorts. We introduce **SpecFormer**, a self-supervised framework that converts each unordered expression profile into a sample-specific gene sequence using term frequency–inverse document frequency (TF-IDF) ordering, then pretrains a transformer encoder via masked gene identity prediction rather than expression-value reconstruction. Pretrained on harmonized TCGA Pan-Cancer data spanning five normalization schemes, SpecFormer achieves 90.83% accuracy and macro AUC-ROC of 0.997 across 33 cancer types, captures pathway co-regulation structure with mean Pearson correlations of 0.754 and 0.762 across 1,387 PARADIGM pathways, and preserves tissue-level transcriptomic organization on independent GTEx healthy tissue data without retraining. Compared with BulkRNABert, SpecFormer produces markedly richer embedding geometry (effective rank 95.6 vs. 6.3) and more stable histological subtype discrimination, without requiring expression discretization or in-distribution pretraining exposure.

1. Introduction

Bulk RNA sequencing (RNA-seq) remains one of the most widely used molecular profiling technologies in translational genomics and clinical cohort studies (Smail & Montgomery, 2024; Divate et al., 2022). Yet, despite rapid progress in self-supervised representation learning for biological data, bulk RNA-seq lacks a broadly reusable encoder that operates at transcriptome scale, transfers across normalization schemes, and supports diverse downstream tasks without task-specific fine-tuning.

Transformer-based foundation models have shown substantial promise for single-cell transcriptomics: Geneformer uses rank-based gene ordering for transfer learning (Theodoris et al., 2023), and scGPT learns generative representations from millions of single-cell profiles (Cui et al., 2024). However, bulk RNA-seq aggregates expression across many cells and tissue compartments, producing a fundamentally different statistical object from single-cell profiles, with distinct normalization dependencies and lower within-cohort expression granularity. Single-cell strategies do not transfer directly.

Recent bulk transformer models, BulkRNABert (Gélard et al., 2024) and GexBERT (Jiang & Hassanpour, 2025), demonstrate the promise of masked pretraining for bulk data but couple representations to expression magnitudes through discretization or reconstruction objectives. This reduces portability across datasets with different preprocessing pipelines. We introduce **SpecFormer**, which addresses this limitation through two design choices: (1) specificity-weighted TF-IDF gene ordering that decouples sequence construction from expression scale, and (2) masked gene *identity* prediction that learns contextual co-occurrence structure without reconstructing numerical values.

2. Method

Figure 1 provides an overview of the SpecFormer framework. Each expression profile is converted to a TF-IDF-ordered gene sequence, divided into overlapping windows, and passed through a masked gene modeling objective during pretraining. The frozen encoder is then used to extract sample-level embeddings for downstream tasks.

2.1. TF-IDF Gene Ordering

Because bulk RNA-seq profiles are unordered vectors, SpecFormer converts each sample into a ranked gene sequence using TF-IDF scoring. For gene g_i in sample s, the term

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop (ICML 2026). Do not distribute.

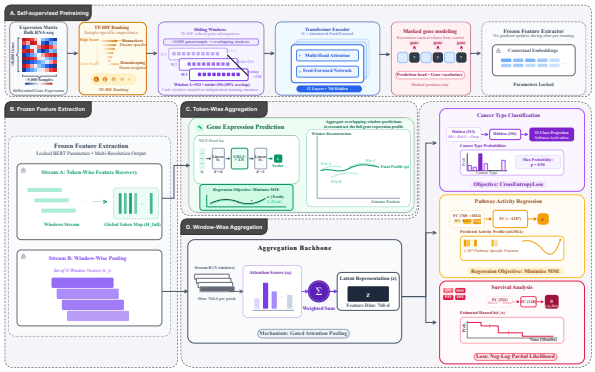

Figure 1. Overview of the SpecFormer framework. (A) TF-IDF ordering converts each unordered expression profile into a sample-specific ranked gene sequence. (B) Overlapping sliding windows of length L=512 (stride 256) divide the sequence into training instances. (C) Masked gene modeling pretrains the encoder to predict masked gene identities from transcriptomic context. (D) The frozen encoder produces sample-level embeddings via gated attention pooling over window representations, used across all downstream tasks without fine-tuning.

frequency captures within-sample abundance,

$$\mathrm{tf}(g_i, s) = \frac{c_{g_i,s}}{\sum_{j=1}^{n} c_{g_j,s}},$$

while the inverse document frequency captures cohort-level specificity,

$$\mathrm{idf}(g_i) = \log\left(\frac{|\mathcal{S}|}{1 + |\{s \in \mathcal{S} : c_{g_i,s} > 0\}|}\right),$$

where the additive constant prevents division by zero for universally absent genes. Genes are ranked by $\mathrm{tfidf}(g_i, s) = \mathrm{tf}(g_i, s) \cdot \mathrm{idf}(g_i)$ in decreasing order. Unlike expression-magnitude ranking, TF-IDF ordering incorporates both within-sample abundance and across-sample specificity, downweighting broadly expressed housekeeping genes that dominate raw expression but carry limited disease-specific information.

2.2. Sliding-Window Sequence Construction

Because the full transcriptome ($\sim$10,000 genes) exceeds the input capacity of standard BERT-style encoders (Devlin et al., 2019), each TF-IDF-ordered sequence is divided into overlapping windows of length L=512 with stride 256. Each window is treated as an independent training instance during pretraining. The 50% overlap ensures most genes appear near the center of at least one window, mitigating boundary effects.

2.3. Masked Gene Modeling Pretraining

SpecFormer is pretrained using a masked gene modeling (MGM) objective. For each window, 15% of gene-token positions are selected at random; of these, 80% are replaced with `[MASK]`, 10% with a randomly sampled gene token, and 10% left unchanged, following the standard BERT protocol (Devlin et al., 2019). The encoder is trained to predict masked gene identities from context:

$$\mathcal{L}_{\mathrm{MGM}}(\theta) = -\mathbb{E}_{s\sim\mathcal{S}}\left[\sum_{i\in\mathcal{M}} \log p_\theta\left(x_i^{(s)} \mid \tilde{\mathbf{x}}^{(s)}\right)\right],$$

where $\mathcal{M}$ denotes masked positions, $x_i^{(s)}$ is the original gene token, and $\tilde{\mathbf{x}}^{(s)}$ is the masked input. Predicting gene identities rather than expression values decouples the pretrained representation from expression scale, discretization scheme, and normalization pipeline.

2.4. Encoder Architecture and Aggregation

SpecFormer uses a BERT-style encoder with 12 transformer layers, hidden dimension 768, 12 attention heads, and feed-forward dimension 3072. The encoder is frozen after pretraining. For downstream tasks, window-level embeddings are obtained by mean-pooling token embeddings within each window, then combined into a fixed-dimensional sample embedding $\mathbf{z}^{(s)} \in \mathbb{R}^{768}$ via a gated attention pooling module. This shared embedding is used across all downstream tasks without encoder fine-tuning. Full aggregation strategy and model capacity ablation results are provided in Appendix A.

3. Experiments

3.1. Data and Training

SpecFormer was pretrained on 8,315 harmonized TCGA Pan-Cancer samples spanning five RNA-seq normalization schemes: batch-effect normalized TCGA profiles and four UCSC Toil-derived quantifications (Weinstein et al., 2013; Vivian et al., 2017) (expected counts, FPKM, TPM, normalized counts). Gene identifiers were harmonized to HUGO symbols; low-information genes were removed by median-expression and variance filters, yielding $\sim$10,000 genes. A 90/5/5 train/validation/test split was used consistently across pretraining and all downstream evaluations. TF-IDF statistics were computed within each split to prevent leakage.

3.2. Cancer-Type Classification

Using frozen SpecFormer embeddings and a lightweight MLP head, the model achieved 90.83% accuracy, macro AUC-ROC of 0.9965, and MCC of 0.9036 across 33 TCGA cancer types on the held-out test set (Table 1). TF-IDF ordering substantially outperformed expression-value ordering (84.81%) and Z-score ordering (80.02%), demonstrating that cohort-level gene specificity provides an informative ordering signal beyond expression magnitude. Compared

with a non-contextual top-200 TF-IDF logistic regression baseline (65.00% accuracy), the large gap confirms that masked gene modeling captures gene-to-gene contextual relationships not available from gene identity alone.

Table 1. Cancer-type classification on the held-out TCGA test set (480 samples, 33 classes). CI = 95% bootstrap confidence interval.

Ordering	Accuracy	MCC	AUC
Z-score	0.800 [0.765, 0.835]	0.793	0.988
Expression value	0.848 [0.815, 0.879]	0.841	0.995
TF-IDF	**0.908** [0.881, 0.933]	**0.904**	**0.997**
Top-200 + LR (baseline)	0.650	0.648	0.980

3.3. Pathway Activity Regression

Frozen SpecFormer embeddings were used to predict 1,387 PARADIGM pathway activity scores (Subramanian et al., 2005). With TF-IDF ordering, the model achieved mean sample-wise Pearson correlation of 0.754 and pathway-wise Pearson correlation of 0.762 (mean MAE 0.494 on z-scored targets, corresponding to approximately half a standard deviation). The near-identical sample-wise and pathway-wise performance indicates that representations capture pathway co-regulation structure consistently across both axes rather than fitting one at the expense of the other. Figure 2 further confirms that predicted expression profiles preserve global cancer-type organization in a shared transcriptomic manifold, despite the absence of expression-value reconstruction during pretraining.

3.4. Survival Prediction

Separate Cox-style neural models were trained on frozen SpecFormer embeddings for four TCGA survival endpoints: overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI), and disease-free interval (DFI). Each model produced a scalar log-risk score optimized via negative log partial likelihood. With TF-IDF ordering, SpecFormer achieved C-index values of 0.646, 0.672, 0.612, and 0.671 for OS, DSS, PFI, and DFI respectively (Table 2), representing moderate discriminative ability consistent with transcriptomic survival models on pan-cancer cohorts. TF-IDF ordering achieved the highest C-index for DSS, PFI, and DFI, while expression-value ordering performed best for OS (0.684), potentially reflecting differences in the expression signals relevant to all-cause versus cancer-specific mortality. Together with classification and pathway results, these findings demonstrate that a single frozen encoder supports diverse prediction objectives spanning classification, regression, and time-to-event analysis.

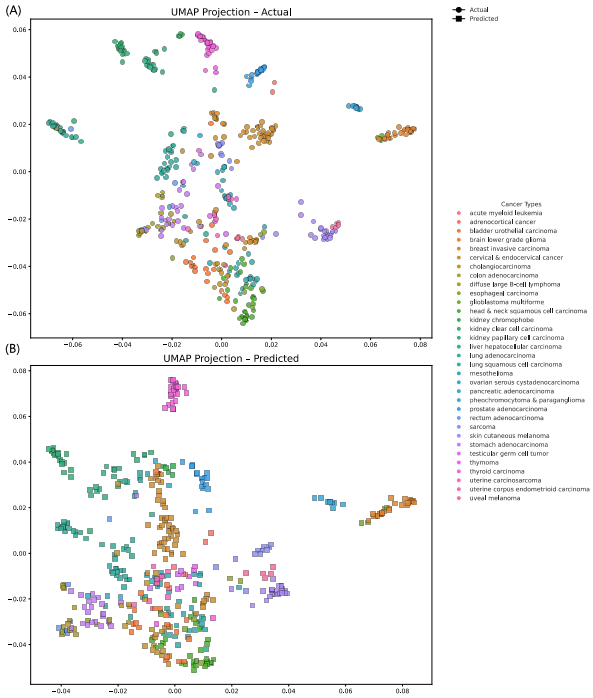

Figure 2. UMAP visualization of actual (A) and SpecFormer-predicted (B) gene-expression profiles for the held-out TCGA test set (n=480 samples), colored by cancer type. Despite predictions being generated from a frozen encoder never trained to reconstruct expression values, predicted profiles recapitulate the global transcriptomic organization of real data. Cancer-type clusters remain well separated and preserve relative neighborhood relationships across the manifold, indicating that SpecFormer representations encode higher-order biological structure beyond marginal expression statistics.

3.5. Cross-Dataset Transfer to GTEx

The frozen encoder was evaluated on 500 independent GTEx healthy tissue samples not seen during pretraining. Procrustes alignment of tissue centroids derived from actual and predicted expression profiles demonstrated preservation of global tissue-level transcriptomic organization across 27 tissue types under a substantially shifted biological and experimental distribution (Figure 3).

3.6. Representation Quality vs. BulkRNABert

Table 3 compares SpecFormer against BulkRNABert (Gélard et al., 2024) on 480 samples spanning 11 cancer types. Note that these evaluation samples fall within BulkRNABert's pretraining distribution (TCGA data), whereas they are held out for SpecFormer, making this comparison conservative for SpecFormer.

BulkRNABert outperforms SpecFormer on linear probing

Table 2. Survival prediction C-index on the held-out TCGA test set across four endpoints and three gene ordering strategies. ↑ higher is better. Best per-column in **bold**.

Ordering	OS ↑	DSS ↑	PFI ↑	DFI ↑
Z-score	0.623	0.651	0.598	0.648
Expression value	**0.684**	0.663	0.608	0.662
TF-IDF	0.646	**0.672**	**0.612**	**0.671**

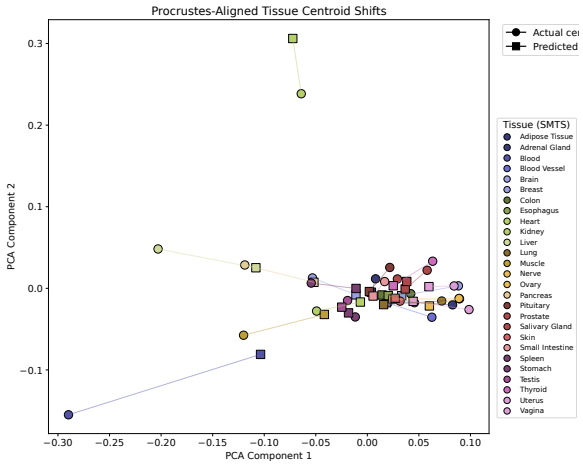

Figure 3. Procrustes-aligned tissue centroids from actual and predicted expression profiles for 500 independent GTEx samples. Short connecting distances indicate strong agreement between actual and predicted centroid locations across tissue types, demonstrating cross-dataset structural transferability without retraining.

and retrieval metrics, which is expected given in-distribution evaluation and a pretraining objective aligned with expression magnitude. However, SpecFormer produces a 15-fold higher effective rank and dramatically lower eigenvalue dominance, indicating a far more isotropic representation space. On lung adenocarcinoma vs. squamous cell carcinoma subtype discrimination, which is a task requiring regulatory rather than magnitude-based separation, both models achieve comparable mean accuracy, but SpecFormer exhibits threefold lower variance (6.5% vs. 19.2%), indicating more stable and generalizable subtype representations. These geometric properties are independent of the data-overlap confound.

4. Discussion

SpecFormer demonstrates that specificity-weighted TF-IDF ordering combined with masked gene identity prediction yields transferable whole-transcriptome representations without dependence on expression magnitudes. The consistent advantage of TF-IDF ordering across classification, pathway regression, and three of four survival endpoints supports the interpretation that the IDF component encodes a general-purpose biological signal by downweighting housekeeping genes and exposing disease-relevant transcriptional structure. The large gap over a non-contextual baseline confirms that masked modeling captures contextual gene-to-gene relationships not recoverable from gene identity alone. The divergence between OS and cancer-specific survival endpoints warrants further investigation and may reflect differences in the expression signals relevant to all-cause versus disease-specific mortality.

Table 3. Representation quality on 480 TCGA samples (11 cancer types). †Evaluation samples overlap with BulkRNABert's pretraining data. ↑ higher is better; ↓ lower is better.

Metric	SpecFormer	BulkRNABert†
Embedding Geometry		
Effective Rank ↑	**95.6**	6.3
Top Eigenvalue Dominance ↓	**0.114**	0.585
Lung Subtype Discrimination		
Accuracy ↑	$69.3 \pm 6.5\%$	$71.6 \pm 19.2\%$
Cross-Cancer Retrieval (Precision@k)		
P@1 ↑	0.644	**0.684**
P@10 ↑	0.460	**0.519**
Linear Probe (5-fold CV)		
Disease Accuracy ↑	$73.4 \pm 3.8\%$	$\mathbf{83.4 \pm 3.5}\%$

Compared with BulkRNABert, SpecFormer's richer embedding geometry and more stable subtype discrimination suggest that decoupling pretraining from expression reconstruction encourages distributed encoding of co-regulatory structure rather than axis-aligned magnitude compression. This property is likely to improve cross-cohort portability in settings where expression scales differ substantially across normalization pipelines, which is a common challenge in multi-cohort translational studies.

Limitations include validation primarily within TCGA, moderate survival prediction performance, and the absence of task-specific GTEx benchmarking. Comparison with standard generative baselines confirms that VAE and GAN approaches fail to preserve transcriptomic manifold structure at whole-transcriptome scale (Appendix A.3), further motivating the self-supervised discriminative approach introduced here. Future work should evaluate SpecFormer on independent disease cohorts and explore attention interpretability to characterize the transcriptional co-regulatory structures learned during pretraining. Normalization-robust bulk transcriptomic representations are a prerequisite for generative models and agentic systems that must reason across heterogeneous multi-cohort data.

References

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature*

methods, 21(8):1470–1480, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., and Nagaraj, S. H. Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers*, 14(5):1185, 2022.

Gélard, M., Richard, G., Pierrot, T., and Cournède, P.-H. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. *bioRxiv*, pp. 2024–06, 2024.

Jiang, S. and Hassanpour, S. Transformer-based representation learning for robust gene expression modeling and cancer prognosis. *Scientific Reports*, 15(1):37581, 2025.

Smail, C. and Montgomery, S. B. Rna sequencing in disease diagnosis. *Annual review of genomics and human genetics*, 25, 2024.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43):15545–15550, 2005.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., et al. Toil enables reproducible, open source, big biomedical data analyses. *Nature biotechnology*, 35(4):314–316, 2017.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

A. Ablation Studies

A.1. Aggregation Strategy

To justify the mean-attention aggregation used across all downstream tasks, six variants were evaluated on the 33-class cancer-type classification task, combining two window embedding types (CLS token vs. mean pooling) with three aggregation strategies (mean, attention, transformer). All variants share an identical classification head, ensuring that performance differences reflect embedding and aggregation design only (Table 4). Attention-based aggregation consistently outperforms both mean and transformer aggregation regardless of embedding type. The CLS-transformer configuration performs substantially worse than all other variants (70.03% accuracy), indicating that transformer-based aggregation is particularly sensitive to the quality of input window representations. Mean-pooled embeddings combined with attention-based aggregation achieved the highest performance across all metrics and was adopted for all downstream tasks.

A.2. Model Capacity

Three model capacity variants were compared under the mean-attention configuration (Table 5). Performance differences are modest, ranging from 0.17 to 2.08 percentage points across metrics. The Base-Large model shows slight degradation with early convergence indicative of mild overfitting given the training set size. The Base-2048 variant achieves the highest metrics at higher computational cost. These results confirm that the Base configuration offers the best trade-off between performance and efficiency and was adopted as the primary model.

A.3. Comparison with Generative Baselines

A variational autoencoder (VAE) and a generative adversarial network (GAN) were evaluated as generative baselines to assess whether standard latent-variable frameworks can capture the structure of high-dimensional bulk RNA-seq profiles. The VAE used encoder and decoder dimensions $1024 \to 128 \to 32$, trained with a β-annealed ELBO objective. The GAN generator mapped latent noise to expression space via linear layers with ReLU activations, with a symmetric LeakyReLU discriminator.

Despite successful optimization convergence for both models, UMAP projections of generated samples reveal fundamental limitations (Figure 4). The VAE exhibits severe manifold collapse, producing overly compact clusters poorly aligned with the real data manifold. The GAN shows only partial alignment, with a substantial fraction of generated samples lying outside the support of the true TCGA distribution. These results highlight that standard generative models fail to faithfully reproduce transcriptomic structure at whole-transcriptome scale, motivating the discriminative self-supervised approach of SpecFormer.

Table 4. Aggregation strategy comparison for 33-class cancer-type classification on the held-out test set (480 samples). Results reported with 95% bootstrap confidence intervals. Bold indicates best per metric.

Configuration	Accuracy	F1 (Macro)	F1 (Weighted)	MCC	Top-3 Acc.	AUC-ROC
CLS-Transformer	0.700 [0.658, 0.744]	0.471 [0.426, 0.517]	0.654 [0.607, 0.701]	0.687 [0.645, 0.730]	0.858 [0.826, 0.888]	0.983 [0.975, 0.990]
CLS-Mean	0.807 [0.771, 0.842]	0.736 [0.679, 0.790]	0.804 [0.766, 0.841]	0.797 [0.759, 0.834]	0.944 [0.923, 0.965]	0.988 [0.980, 0.993]
Mean-Mean	0.825 [0.792, 0.858]	0.735 [0.686, 0.786]	0.821 [0.783, 0.855]	0.816 [0.780, 0.851]	0.951 [0.931, 0.969]	0.990 [0.985, 0.993]
Mean-Transformer	0.848 [0.815, 0.879]	0.727 [0.677, 0.778]	0.836 [0.799, 0.869]	0.840 [0.806, 0.873]	0.958 [0.940, 0.975]	0.990 [0.982, 0.996]
CLS-Attention	0.883 [0.854, 0.910]	0.834 [0.790, 0.874]	0.879 [0.849, 0.908]	0.877 [0.847, 0.906]	**0.975 [0.960, 0.988]**	0.995 [0.993, 0.997]
Mean-Attention	**0.908 [0.881, 0.933]**	**0.853 [0.808, 0.895]**	**0.904 [0.875, 0.929]**	**0.904 [0.876, 0.930]**	0.973 [0.958, 0.985]	**0.997 [0.994, 0.998]**

Table 5. Model capacity ablation under mean-attention aggregation. Results reported with 95% bootstrap confidence intervals on the held-out test set. Bold indicates best per metric.

Model	Accuracy	F1 (Macro)	F1 (Weighted)	MCC	Top-3 Acc.	AUC (Macro)
Base-Large	0.887 [0.858, 0.915]	0.831 [0.781, 0.872]	0.886 [0.856, 0.913]	0.881 [0.851, 0.910]	0.967 [0.952, 0.981]	0.995 [0.993, 0.998]
Base	0.908 [0.881, 0.933]	0.853 [0.808, 0.895]	0.904 [0.875, 0.929]	0.904 [0.876, 0.930]	0.973 [0.958, 0.985]	0.997 [0.994, 0.998]
Base-2048	**0.910 [0.892, 0.927]**	**0.868 [0.840, 0.896]**	**0.909 [0.891, 0.926]**	**0.905 [0.887, 0.923]**	**0.989 [0.982, 0.995]**	**0.997 [0.996, 0.998]**

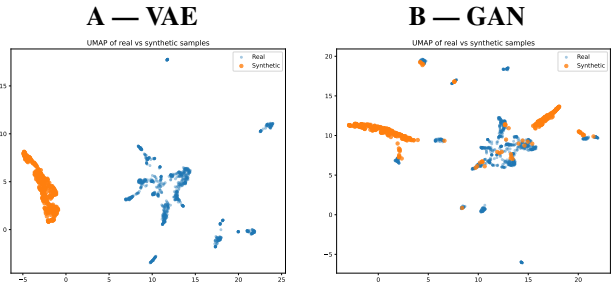

Figure 4. UMAP visualization of synthetic gene-expression profiles generated by (A) a VAE and (B) a GAN against real TCGA samples. The VAE exhibits manifold collapse with overly compact clusters misaligned with the real data distribution. The GAN shows partial alignment but with substantial out-of-support samples. Both results highlight fundamental limitations of standard generative approaches for whole-transcriptome bulk RNA-seq modeling.