BioMedInformatics

MDPI

Article

Transcriptome-Based Subtype Discovery in Breast Cancer Using Variational Representation Learning

Al Hanouf Al Khawaldeh [1,*], Loai Alnemer [2] and Hazem Hiary [1]

[1] Department of Computer Science, The University of Jordan, Amman 11942, Jordan; hazemh@ju.edu.jo
[2] Department of Artificial Intelligence, The University of Jordan, Amman 11942, Jordan; nemer@ju.edu.jo
* Correspondence: h.khawaldeh@ju.edu.jo

Abstract

Breast cancer is a heterogeneous disease whose molecular complexity complicates therapeutic stratification. Although PAM50 is a widely used molecular subtyping framework, discrete subtype assignments may underrepresent transitional molecular states and within-subtype heterogeneity relevant to treatment. This study develops a biologically interpretable representation learning framework that projects bulk RNA-seq profiles into nonlinear latent spaces and applies unsupervised hierarchical clustering, cross-validated against *k*-means and Gaussian mixture models (GMMs), to investigate additional tumor structure. The pipeline integrates a variational autoencoder (VAE), t-SNE/UMAP visualization, and functional (differential expression, pathway enrichment) and immune (ESTIMATE, quanTIseq) characterization of the resulting clusters. From a source pool of 19,131 transcriptomic profiles, a primary analytical cohort of 8744 unique, non-duplicated, PAM50-labeled tumor and normal samples was used for VAE training and all downstream analyses, without class balancing or oversampling. Across this primary cohort, tissue of origin and malignant-versus-normal status—not PAM50 subtype—consistently dominated clustering structure across every representation (PCA-40; VAE $Z = 10, 20, 40$) and every clustering algorithm tested, a pattern reproduced even in the raw gene expression space. Isolating the resulting 835-sample malignant subpopulation and reclustering it in isolation revealed subtype-specific structure obscured at the whole-cohort level: Basal-like disease separated cleanly under every representation, algorithm, and gene-panel size tested (83–98% purity); Luminal A resolved into a distinct cluster plus a continuum bridging Luminal B; HER2-enriched concentrated but only inconsistently achieved cluster-majority status; and Luminal B never achieved outright cluster majority under any configuration tested, behaving as a shared boundary population rather than an independently separable subtype. Among malignant-subset representations, VAE $Z = 10$ achieved the strongest internal validation metrics, the highest concordance with PAM50 (Adjusted Rand Index = 0.377), and—together with VAE $Z = 40$—a concordance index exceeding the PAM50-only baseline (0.619 and 0.643 versus 0.606). Immune deconvolution showed that immune-hot and immune-cold status cut across, rather than align with, PAM50 boundaries, including a regulatory T-cell/M1-macrophage-skewed HER2-enriched/Luminal B cluster and a $CD8^{+}$-poor, B-cell-skewed Basal-like cluster. Bootstrap resampling and gene-panel-size sensitivity analyses confirmed that cluster stability tracked PAM50 purity, with Basal-like separation robust across all conditions and the HER2/Luminal B boundary consistently the least stable. These internally derived findings indicate that nonlinear representation learning can recover biologically coherent structure that refines, rather than replaces, PAM50 classification, though independent-cohort and spatial or single-cell validation are required before clinical interpretation.

Check for updates

Academic Editor: Joaquim Carreras

Received: 2 August 2026
Revised: 6 September 2026
Accepted: 23 September 2026
Published: 4 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: variationalautoencoders; breast cancer heterogeneity; PAM50 molecular subtypes; tumor immune microenvironment; unsupervised clustering; transcriptomics; robustness analysis

1. Introduction

Breast cancer remains a leading cause of oncology-related mortality worldwide, primarily due to its significant clinical and molecular heterogeneity [1]. In the past two decades, the shift from histopathological classification to molecular profiling has transformed the understanding of breast tumor biology [2]. The foundational study by Perou et al. introduced the concept of transcriptomic "molecular portraits," demonstrating that breast tumors segregate into distinct, reproducible intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, and Basal-like [2]. This finding established the basis for the clinical application of expression-based diagnostics, leading to the development of the PAM50 supervised risk predictor [3]. The PAM50 classifier employs a targeted 50-gene expression panel to assign tumors to specific intrinsic subtypes, offering prognostic information on recurrence risk and chemotherapy benefit beyond standard tumor staging [3,4]. Large-scale consortium initiatives, particularly The Cancer Genome Atlas (TCGA) network, have validated these principal axes of variation across multi-omics cohorts [5]. Subsequent efforts have provided comprehensive molecular subtyping for specific phenotypes, such as triple-negative breast cancer [6].

Although the PAM50 framework has been extensively validated, several limitations remain in breast cancer transcriptomic research. PAM50 uses a centroid-based approach to assign tumors to discrete categories [3,7]. A single-label assignment cannot fully represent within-subtype heterogeneity, transitional molecular states, or multiphenotypic expression gradients [8]. Patients sharing a PAM50 label may nevertheless have different clinical trajectories, immune microenvironment profiles, and therapeutic responses [8,9].

Second, classical unsupervised stratification approaches historically deployed to explore this biological complexity remain technically constrained. A vast majority of expression-based subtyping studies continue to rely heavily on linear dimensionality-reduction methods, such as Principal Component Analysis (PCA), or apply clustering algorithms directly within high-dimensional gene expression spaces [10,11]. High-dimensional genomic matrices are notoriously noisy and subject to multi-center technical batch effects [12,13]. More importantly, linear mathematical transformations are inherently incapable of capturing the complex, nonlinear gene–gene regulatory dependencies, feedback loops, and epistatic interactions that dictate actual tumor biology [14]. This limitation is compounded whenever the analyzed cohort spans multiple tissue types or a mixture of malignant and non-malignant samples, since the largest axis of linear variance in such a cohort need not coincide with the biological axis of interest. Consequently, classical unsupervised methods frequently yield unstable or biologically ambiguous clusters that lack translational value.

Third, deep representation learning methods can address some constraints of linear models, but their application to breast cancer remains fragmented. Variational autoencoders (VAEs) provide a probabilistic generative framework that compresses high-dimensional transcriptomes into low-dimensional latent variables while regularizing the representation through Kullback–Leibler Divergence (KLD) [15,16]. Way and Greene demonstrated that a VAE trained on pan-cancer TCGA RNA-seq data could extract biologically relevant latent features associated with known pathways [17]. Existing studies have largely focused on representation learning, reconstruction, visualization, or subtype discovery; fewer have

linked latent embeddings to a unified clinical, immunological, and functional characterization workflow, and fewer still have subjected the resulting cluster structure to systematic testing against resampling, model capacity, or feature-selection choices—despite clustering solutions in high-dimensional genomic data being generally known to be sensitive to exactly these choices [11,17–21].

Finally, the precise relationship between newly discovered deep learning clusters and the surrounding tumor immune microenvironment (TIME) has not been comprehensively explored through the lens of deep generative representations. Bulk tumor transcriptomic data represent an integrated mixture of malignant cells, stromal elements, and infiltrating immune cell types [22]. Algorithms such as ESTIMATE and CIBERSORT have enabled investigators to infer cellular purity, stromal abundance, and detailed immune cell-type proportions directly from bulk expression signatures [23,24]. While the configuration of the immune landscape—particularly the distinction between "immune-hot" (highly infiltrated) and "immune-cold" (deserted) phenotypes—is a critical determinant of survival and response to modern immune checkpoint therapies, current classification models fail to systematically link these microenvironmental landscapes with nonlinear transcriptomic representations or to test whether immune-hot and immune-cold status track PAM50 subtype boundaries at all [25,26]. This lack of integration limits both the biological interpretability and translational utility of deep representation networks within modern precision oncology workflows [27].

To bridge these interrelated gaps, this study proposes an integrative, biologically interpretable unsupervised deep learning pipeline that unifies probabilistic deep generative modeling with robust downstream statistical, functional, and cellular deconvolution workflows. Critically, this pipeline is applied in two stages. First, it is applied directly to a primary analytical cohort spanning both malignant and non-malignant, multi-tissue-origin samples to test whether unsupervised clustering of nonlinear latent representations recovers PAM50 subtype structure under realistic cohort composition. Having found that it instead recovers tissue of origin and malignancy status—a confound rather than a subtype signal—the pipeline isolates the resulting malignant subpopulation and repeats the full representation learning, clustering, and characterization workflow on this subset in isolation, where genuine subtype-associated structure can be assessed without interference from this larger, confounding axis of variation. The resulting subtype-level findings, together with their biological, immune, and clinical correlates, are then subjected to a dedicated multi-axis robustness analysis to establish which of them constitute stable, non-arbitrary structure rather than artifacts of a single modeling choice.

The primary contributions of this work are summarized as follows:

- **A Comprehensive, Cross-Validated Deep Generative Pipeline:** We propose and validate an integrated, unsupervised representation learning pipeline combining a variational autoencoder (VAE) architecture, nonlinear manifold visualization (t-SNE and UMAP), hierarchical clustering cross-validated against k-means and Gaussian mixture model clustering, and systematic functional and immune multi-omics characterization.
- **Resolution of a Confounding Tissue-of-Origin Signal:** We demonstrate that unsupervised clustering applied to a mixed malignant/non-malignant, multi-tissue-origin cohort—across every representation and clustering algorithm tested, including the raw gene expression space with no dimensionality reduction—consistently recovers tissue of origin and malignancy status rather than PAM50 subtype, and we isolate the resulting malignant subpopulation to reveal the subtype-associated structure this confound otherwise obscures.
- **A Precise, Cross-Algorithm Map of PAM50 Subtype Separability:** Within the isolated malignant population, we characterize exactly which PAM50 subtypes are, and are

not, separable by unsupervised clustering: Basal-like disease separates cleanly under every representation, algorithm, and gene-panel size tested; Luminal A resolves into a distinct cluster and a continuum bridging Luminal B, HER2-enriched concentrates but only inconsistently achieves cluster-majority status; and Luminal B never achieves outright cluster majority under any tested configuration—evidence for a continuous rather than strictly discrete boundary between these subtypes.

- **Characterization of Immune Heterogeneity Crossing PAM50 Boundaries:** We show that immune-hot and immune-cold status do not align with PAM50 subtype, including a Luminal A population that splits into an immune/stromal-rich cluster and immune-cold clusters, a Basal-like cluster that is immune-active yet CD8-poor and B-cell-skewed, and a HER2-enriched/Luminal B bridging cluster with a regulatory T-cell- and M1-macrophage-skewed profile suggestive of an immunosuppressive microenvironment.
- **A Multi-Axis Robustness Framework:** We subject the discovered clustering solution to four complementary robustness analyses—bootstrap resampling stability, stability of the primary-cohort clustering solution, sensitivity to latent space dimensionality, and sensitivity to highly variable gene-panel size—establishing which findings constitute stable, non-arbitrary structure rather than artifacts of a single modeling or feature-selection choice.

The remainder of this paper is structured as follows. Section 2 details our materials and methods, outlining dataset harmonization, VAE model architecture selection, hierarchical clustering parameters, and downstream microenvironmental deconvolution protocols. Section 3 contains our comprehensive experimental results, detailing model benchmarks, survival metrics, and functional enrichment readouts. Section 4 presents a detailed discussion of our findings, expanding on our structural contributions, biological interpretations, and clinical translational value. Finally, Section 5 concludes this paper with an overview of future research directions.

2. Materials and Methods

This paper describes an unsupervised deep learning pipeline for transcriptome-based subtype discovery in breast cancer. The framework combines nonlinear latent representation learning through a variational autoencoder (VAE), hierarchical clustering for molecular subtype discovery, and immune microenvironment characterization, evaluated through a multi-tier internal biological and clinical analysis. Transcriptomic data were drawn from harmonized public repositories and preprocessed into a standardized expression matrix. The primary analytical cohort comprised 8744 unique, PAM50-labeled or confirmed normal-tissue samples; the VAE was trained directly on this real, non-duplicated cohort and compressed each profile into a latent space across three bottleneck dimensionalities ($Z = 10, 20, 40$), each benchmarked against a linear PCA baseline. Discovered subtypes were then evaluated through internal geometric metrics, clinical concordance scoring, Kaplan–Meier survival analysis, multivariate Cox modeling, differential expression, functional enrichment, and immune deconvolution—all performed on the unique, non-duplicated sample cohort. Robustness of the clustering solution to resampling, latent dimensionality, and feature-selection thresholds was additionally assessed through a dedicated sensitivity analysis (Section 2.9).

2.1. Dataset Description and Preprocessing

We obtained transcriptomic data and corresponding metadata from the UCSC Xena platform (https://xenabrowser.net/ (accessed on 12 March 2026)). The total processed pool contained 19,131 tumor and normal tissue profiles, with expression quantified as

transcripts per million (TPM); this pool was not itself analyzed or reported on and served only as the source from which the primary analytical cohort below was derived. Samples were assigned a single label using the following priority order: GTEx-origin samples were labeled Normal-GTEx; TARGET-origin samples were labeled TARGET_Outlier and excluded as pediatric records ($n = 734$); samples with a metastatic sample-type annotation were labeled Metastatic and excluded ($n = 7$), since metastatic-site expression profiles are not comparable to primary-tumor or normal-tissue profiles and cannot be assigned a PAM50 primary-tumor subtype; samples with a solid tissue normal annotation were labeled Normal-TCGA; remaining samples with a non-missing `PAM50Call_RNAseq` values were labeled with that subtype; and all remaining samples were labeled Unknown_Tumor and excluded ($n = 9646$). This yielded a primary analytical cohort of 8744 unique, uniquely labeled samples (Table 1), used throughout this study for VAE training and all downstream biological and clinical characterizations.

Table 1. Primary analytical cohort ($n = 8744$) and those excluded because of missing subtype labels, pediatric origin, or metastatic sample type.

Category	Subtype Origin	Sample Count (*n*)
Included	Normal-GTEx	7792
Included	Normal-TCGA	113
Included	Luminal A	420
Included	Luminal B	192
Included	HER2-enriched	66
Included	Basal-like	139
Included	Normal-like	22
Excluded	Unknown_Tumor	9646
Excluded	TARGET_Outlier (pediatric)	734
Excluded	Metastatic	7
Total	All Processed Samples	19,131

Dimensionality reduction of the original 60,499 gene features was performed through a two-stage filtering strategy applied to the full 19,131-profile pool. First, genes with low expression (TPM < 1 in more than 90% of samples) were removed. The remaining genes were ranked by Median Absolute Deviation (MAD), and the 5000 most variable genes were retained. The resulting matrix was standardized using z-score normalization, fitted across the full pool. This standardized matrix was then subset to the 8744 samples comprising the primary analytical cohort; all downstream model training, clustering, and characterization used this 8744-sample matrix exclusively.

2.2. Training and Analytical Pipelines

The VAE architecture was implemented using PyTorch (v2.13.0+cpu) and trained directly on the primary analytical cohort of 8744 unique, non-duplicated samples—no class balancing or oversampling was applied in the primary pipeline. A stratified 80%/20% train/validation split (stratified by molecular or tissue label, fixed random seed = 42) produced 6995 training and 1749 validation samples; the validation partition was used solely to monitor convergence and drive early stopping, not as an external validation set. Training used a mini-batch size of 64, the Adam optimizer, and up to 300 epochs with early stopping (patience = 15 epochs on validation loss); weights from the best-performing epoch were restored before the encoder was applied to the full 8744-sample cohort to generate the latent representations used in all downstream analyses. Convergence was monitored via the Evidence Lower Bound (ELBO), applied identically across all three latent dimensionalities ($Z = 10, 20, 40$) and the PCA-40 baseline.

Malignant-only subset ($n = 835$): the subset of the primary analytical cohort assigned to the dominant malignant cluster by raw-feature-space hierarchical clustering (Section 3.4), used to independently train a parallel set of VAE models restricted to this smaller, tumor-enriched population. Because this subset is roughly one-tenth the size of the primary cohort, the training protocol was adjusted to guard against overfitting: batch size was reduced from 64 to 32, early-stopping patience was extended from 15 to 20 epochs (validation loss on 167 held-out samples is noisier than on the primary cohort's 1749), and L2 weight decay (1×10^{-4}) was added to the Adam optimizer. All other settings—optimizer, learning rate, logvar clamping, and gradient clipping—were retained unchanged from the primary-cohort protocol (Section 2.3). Because this subset's Normal-TCGA representation reduces to a single sample, the stratified split fell back to a plain random 80%/20% partition, producing 668 training and 167 validation samples.

2.3. Variational Autoencoder Architecture

Nonlinear transcriptomic features and tumor heterogeneity were modeled using a variational autoencoder (VAE). The architecture was constructed in PyTorch using a symmetric bottleneck layout. The encoder consists of two fully connected layers ($5000 \rightarrow 500 \rightarrow 100$ hidden units) with rectified linear unit (ReLU) activations and a dropout layer (rate = 0.2) after the first hidden layer. To permit end-to-end backpropagation, the latent vector was sampled via the reparameterization trick as $z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$ [15]. The decoder network mirrors the encoder geometry and concludes with a linear layer to reconstruct continuous gene expression profiles. Models were evaluated across three latent dimensionalities ($Z = 10, 20, 40$).

Models were trained directly on the primary analytical cohort ($n = 8744$), using the sample-level stratified 80/20 train/validation split described in Section 2.2, the Adam optimizer with a batch size of 64 and a learning rate of 1×10^{-3}, and early stopping (patience = 15 epochs on validation loss) for up to 300 epochs. To prevent the numerical instability of unconstrained variance estimates—observed during development as intermittent divergence of the reconstruction loss at higher latent dimensionalities—the encoder's log-variance output was clamped to $[-10, 10]$ and gradients were clipped to a maximum norm of 5.0 before each optimizer step; both are standard stabilization measures and do not alter the loss function itself. The joint loss function combined Mean Squared Error (MSE) for reconstruction accuracy and Kullback–Leibler Divergence (KLD) for latent space regularization. Model weights from the epoch with the lowest validation loss were restored before the trained encoder was applied to the full 8744-sample cohort to generate the latent representations used in all downstream analyses. Optimization outcomes across the three latent dimensionalities are detailed in Table 2.

Table 2. VAE optimization outcomes across latent dimensionalities for the primary analytical cohort ($n = 8744$).

Latent Dim. (Z)	Best Epoch	Val. Total Loss	Val. MSE (Fidelity)	Val. KLD (Complexity)
$Z = 10$	51	1177.48	1145.05	32.42
$Z = 20$	87	1155.65	1109.35	46.30
$Z = 40$	97	1160.12	1111.54	48.59

For the malignant-only subset ($n = 835$; Section 3.4), the same encoder–decoder geometry ($5000 \rightarrow 500 \rightarrow 100 \rightarrow Z$) was retained for comparability, but regularization was strengthened given the smaller training set: dropout after the first encoder layer was increased from 0.2 to 0.3, and a second dropout layer (rate = 0.2) was added after the 100-unit encoder layer, which the primary-cohort architecture did not include. The train/validation

loss gap at the best epoch was computed for every latent dimensionality and flagged if it exceeded 20% of the validation loss; no configuration exceeded this threshold (Section 3.1).

2.4. Dimensionality Reduction and Visualization

To evaluate the quality of the learned representations, VAE latent features across all three dimensionalities ($Z = 10, 20, 40$) were benchmarked against a linear PCA baseline, computed on the same standardized 5000-gene matrix and retaining the top 40 principal components to match the maximum latent capacity of the VAE configurations. For visual assessment of cluster separation and global manifold structure, latent coordinates for each representation were projected independently into two dimensions using t-SNE [28] and UMAP [29]—the former emphasizing local neighborhood structure and subcluster detail, the latter preserving the balance between local and global topology. Consistent with their intended role, these projections were used strictly for visualization and were not used for cluster assignment, which was performed separately on the native-dimensional representations (Section 2.5).

t-SNE was run with a perplexity of 40 for the primary figures—perplexity governs the effective number of neighbors contributing to each point's local similarity structure, balancing sensitivity to fine local detail (low perplexity) against preservation of broader manifold shape (high perplexity). To confirm that the resulting structure was not sensitive to this choice, t-SNE was additionally run at perplexities of 30 and 50 for every representation; all three settings produced consistent cluster topology. UMAP was run with `n_neighbors` = 30 and `min_dist` = 0.3. Both methods used a fixed random seed (42) for reproducibility.

Together, these projections provided the basis for evaluating how cleanly each architecture separates healthy GTEx tissue from malignant molecular subtypes, and how compact the resulting clusters are across representations. To establish the biological interpretability of these low-dimensional landscapes before structural clustering, downstream differential expression testing was designated via the limma (v3.66.0) framework [30], while multi-omics functional annotations were mapped using clusterProfiler (v4.18.4) [31].

2.5. Unsupervised Subtype Discovery and Validation

Hierarchical agglomerative clustering was performed independently on each of the four representations under comparison—the PCA-40 baseline and the VAE latent spaces at $Z = 10$, 20, and 40—computed on the primary analytical cohort ($n = 8744$). For every representation, only the numeric latent or principal-component features were passed to the clustering algorithm; the reference label column and sample identifier were explicitly dropped, with the excluded columns printed and verified at runtime. This ensured that cluster assignment could not be informed, directly or indirectly, by the PAM50/tissue labels subsequently used for biological validation.

Euclidean distance and Ward's linkage method [32] were used to minimize within-cluster variance. The primary cluster resolution was set at $k = 7$ to match the number of reference molecular/tissue groups and facilitate downstream comparison; because this choice of k was informed by the reference partition, cluster stability was additionally examined across $k = 3$ to $k = 8$.

Cluster quality was quantified using three internal validation metrics, computed without reference to any label: the silhouette score (cohesion/separation), the Calinski–Harabasz index (between-/within-cluster variance ratio), and the Davies–Bouldin index (cluster overlap). To determine whether the resulting structure reflected Ward's specific merge criterion or a more general property of each representation's geometry, the same $k = 3$–8 sweep was repeated using two alternative unsupervised algorithms on the identical feature space: k-means (Lloyd's algorithm, 10 restarts) and a full-covariance Gaussian

mixture model fitted by expectation–maximization [32]. Agreement in the metric trend across all three algorithms was taken as evidence that the reported differences between representations reflect underlying geometric structure rather than the specific clustering heuristic used.

Biological plausibility was assessed only after cluster assignment, never as an input to it: at $k = 7$, the cluster labels produced by each of the three algorithms were cross-tabulated against the reference PAM50/tissue classes (Basal-like, HER2-enriched, Luminal A, Luminal B, Normal-like, Normal-GTEx, Normal-TCGA) to characterize cluster composition and to quantify the purity of the malignant/PAM50-associated population within its dominant cluster.

Because clustering the primary cohort's PCA-40 and VAE representations consistently resolved tissue-of-origin and malignancy status rather than PAM50 subtype structure (Sections 3.1–3.3), a raw-feature-space hierarchical clustering (Euclidean distance, Ward's linkage, applied directly to the standardized 5000-gene matrix with no PCA or VAE compression) was additionally performed to test whether this reflected a property of the representations themselves or of the transcriptome. The resulting dominant malignant cluster ($n = 835$; Section 3.4) was extracted using only this unsupervised cluster assignment, never the PAM50 reference label, and reclustered in isolation—first, directly on the raw gene expression features (hierarchical, k-means, and Gaussian mixture model clustering, $k = 3$–8; Section 3.4) and subsequently on a parallel set of PCA-40 and VAE ($Z = 10, 20, 40$) representations trained specifically on this malignant-only subset (Section 2.2) at the biologically motivated resolution of $k = 5$ (Sections 3.4.1 and 3.4.2). The external validation framework described in Section 2.6—PAM50 concordance, clinical association, and survival analysis—was applied to this malignant-only representation set using its hierarchical $k = 5$ cluster assignments, with results reported in Section 3.5.

2.6. Comparison with PAM50 Subtypes and Clinical Variables

External cluster validity was quantified using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). ARI measures agreement between unsupervised assignments and PAM50 labels beyond what would be expected by chance, while NMI captures the degree of shared information between the two partitions.

To verify that unsupervised clusters represent biologically grounded patient populations, association analysis against key clinicopathological features was performed. Pearson's chi-square tests were used for categorical variables (ER/PR/HER2 status and AJCC stage), and the Kruskal–Wallis H test was applied to the continuous variable (age at diagnosis). Because the group variable has five levels, a generalized exact test was not available; contingency tables with more than 20% of expected cell counts below 5 were flagged accordingly (Section 3.5.2). All analyses were implemented via the SciPy library [21].

The clinical relevance of the discovered subtypes was assessed through a multi-tier prognostic framework. Survival probabilities were estimated using Kaplan–Meier analysis [33], with the log-rank test used to evaluate differences between clusters. Prognostic accuracy was quantified using the concordance index (C-index) [34], allowing direct comparison against the PAM50 standard [3]. A multivariate Cox proportional hazards model [35], adjusting for age and AJCC stage, was used to determine whether the identified clusters carry independent prognostic value beyond established clinical variables.

2.7. Differential Expression and Pathway Enrichment

Differential expression (DE) analysis was performed using the limma framework [30,36]. By contrasting each VAE-derived cluster against the remainder of the dataset (one-vs.-rest approach), cluster-specific drivers were identified while avoiding the dilution inherent in

pairwise comparisons. Genes were ranked by moderated t-statistic within each cluster, and those meeting a dual threshold of magnitude ($|\log_2 \mathrm{FC}| \geq 1$) and statistical significance (adjusted p-value < 0.05) were designated as cluster-specific driver genes; the ten highest-ranked driver genes per cluster are reported in Supplementary Table S4.

To translate the driver-gene signatures into interpretable molecular programs, a three-tier functional annotation strategy was implemented. Over-representation analysis (GO Biological Process and KEGG) was conducted using clusterProfiler [31], applying a relaxed significance threshold (adjusted p-value < 0.05, $|\log_2 \mathrm{FC}| \geq 0.5$) to accommodate the comparatively modest effect sizes expected when contrasting tumor subclusters against one another rather than tumor against normal tissue; the full ranked term list was retained for each cluster, with the top 20 terms reported regardless of whether they individually reached corrected significance. Gene set enrichment analysis (GSEA) [37] against the MSigDB Hallmark collection [38] was conducted using the complete ranked gene list (by moderated t-statistic) via fgsea to detect coordinated pathway-level shifts not captured by individual gene-level thresholds. Functional interpretation was restricted to the VAE $Z = 10$ representation, established in Sections 3.4.1 and 3.5 as the strongest-performing malignant-subset representation on mathematical, biological, and PAM50-concordance grounds.

2.8. Immune Infiltration Analysis

Immune cell infiltration was estimated for each of the 835 malignant-subset samples to characterize the tumor microenvironment underlying each transcriptomic cluster, since PAM50 classification does not itself model immune context and immune infiltration is an established determinant of both prognosis and therapy response in breast cancer. Global stromal and immune content was quantified using ESTIMATE (v1.0.13) [23], which derives a stromal score and an immune score from single-sample gene set enrichment against curated stromal- and immune-specific signatures, a combined ESTIMATE score, and a derived tumor purity value via the platform-specific formula of Yoshihara et al. [23]. Because ESTIMATE's purity conversion is defined only for the Affymetrix platform, this formula was applied to the RNA-seq-derived scores following common practice in the absence of an RNA-seq-specific alternative; results should be interpreted with this caveat in mind. ESTIMATE was run on the same 5000-gene expression panel used throughout Sections 2.5–2.7, mapped from Ensembl identifiers to gene symbols; coverage of ESTIMATE's reference gene sets within this panel was 95 of 141 stromal signature genes (67.4%) and 52 of 141 immune signature genes (36.9%), reflecting the panel's fixed feature set rather than full transcriptome-wide coverage.

Immune cell composition was further resolved using quanTIseq [39], implemented via the immunedeconv package (v2.1.0) [40], which deconvolves ten immune cell types (B cells, M1 and M2 macrophages, monocytes, neutrophils, NK cells, CD4+ non-regulatory T cells, CD8+ T cells, regulatory T cells, and myeloid dendritic cells) plus an unclassified residual category from bulk expression. quanTIseq was selected over CIBERSORT because it requires no proprietary licensing (unlike CIBERSORT's LM22 signature matrix) while remaining methodologically comparable and is distributed within the same deconvolution framework. Input expression values were converted from $\log_2(\mathrm{TPM} + 1)$ back to linear TPM prior to deconvolution, as required by the method. Signature coverage within the 5000-gene panel was 70 of 138 TIL10 marker genes (50.7%); given this partial coverage, quanTIseq's residual "uncharacterized" category—which typically captures non-immune content in bulk tissue—was negligible across all samples, indicating that the reported cell-type fractions are best interpreted as relative proportions among the ten characterized immune populations rather than as absolute shares of the tumor microenvironment.

Differences in ESTIMATE scores, tumor purity, and quanTIseq cell-type fractions across the five clusters were assessed using the Kruskal–Wallis test, followed by Dunn's post hoc test with Bonferroni correction for pairwise comparisons where the omnibus test was significant; Dunn's test was used in place of pairwise Wilcoxon tests because it incorporates rank information from all five groups simultaneously rather than only the pair under comparison, providing a more statistically appropriate post hoc procedure following a Kruskal–Wallis omnibus result. Benjamini–Hochberg-adjusted p-values are additionally reported alongside the Bonferroni-corrected values. Because several cell-type fractions were zero-inflated and right-skewed, cluster-level summaries are reported as medians rather than means unless otherwise noted, as mean values for these variables were disproportionately influenced by a small number of high-expression outlier samples. Clusters were classified as immune-hot or immune-cold according to their combined ESTIMATE immune score and tumor purity relative to the cohort.

2.9. Robustness Analyses

To evaluate whether the clustering solutions and their biological and clinical associations reported above reflect stable, non-arbitrary structure rather than artifacts of a single analytical choice, four complementary robustness analyses were performed. First, cluster stability of the malignant-subset flagship solution (VAE $Z = 10$, hierarchical clustering, $k = 5$, $n = 835$) was assessed via bootstrap subsampling: fifty resamples, each retaining a random 85% of the 835 samples, were independently reclustered using the identical Ward linkage protocol, and agreement between each resampled partition and the original cluster assignments was quantified using the Jaccard similarity index [41] via the `clusterboot` function of the `fpc` package in R, with resampled clusters matched to their corresponding original cluster by maximum sample overlap. Mean Jaccard values above 0.85 were interpreted as highly stable, 0.60–0.85 as moderately stable, and below 0.60 as unstable, following the conventional thresholds proposed by Hennig (2007).

Second, an equivalent stability check was applied to the clustering solution obtained on the primary analytical cohort ($n = 8744$) prior to the pivot to the malignant-only subset (Section 3.4), to assess whether the whole-cohort failure to recover PAM50 subtype structure, rather than tissue of origin, is itself a stable, reproducible finding across resampling. Third, sensitivity of the malignant-subset clustering solution to latent space dimensionality was assessed by comparing internal validation metrics and survival/clinical concordance across four representations—PCA-40 and VAE latent spaces at $Z = 10$, 20, and 40—clustered under the identical hierarchical (Ward), $k = 5$ protocol (Sections 3.4.1, 3.5 and 3.5.2).

Fourth, sensitivity to the size of the highly variable gene panel used to construct the malignant-subset input matrix was assessed by reconstructing the panel at two additional sizes (3000 and 10,000 genes, versus the 5000-gene panel used throughout), retraining the VAE at $Z = 10$, and re-evaluating cluster–PAM50 concordance at each size. External validation of the discovered subtypes on an independent breast cancer cohort was not performed, as no publicly available dataset meeting this study's inclusion criteria (matched RNA-seq expression, PAM50 subtype calls, and survival follow-up) was identified; this is noted as a limitation in the Section 4.

3. Results

Performance was assessed across mathematical, biological, clinical, and immune microenvironmental dimensions using a hierarchical validation framework, applied to the primary analytical cohort of 8744 real, non-duplicated samples. The configuration that achieves the highest biological concordance, microenvironmental resolution, and prognostic discrimination is reported in the corresponding subsections, following clustering and

survival re-evaluation on this cohort. Robustness of these findings to resampling, latent dimensionality, and gene-panel size was assessed separately (Section 3.8).

3.1. VAE Configuration and Convergence

All three VAE architectures ($Z = 10, 20, 40$), trained directly on the 8744-sample primary analytical cohort, converged smoothly under early stopping, with none of the training instability observed during development at higher latent dimensionalities prior to the log-variance clamping and gradient-clipping measures described in Section 2.3. As shown in Figure 1, all three configurations show a rapid initial decline in training and validation loss followed by a long, stable plateau; early stopping (patience = 15 epochs) triggered at epochs 66 ($Z = 10$), 102 ($Z = 20$), and 112 ($Z = 40$), with the best-performing weights—from epochs 51, 87, and 97, respectively—restored for downstream use.

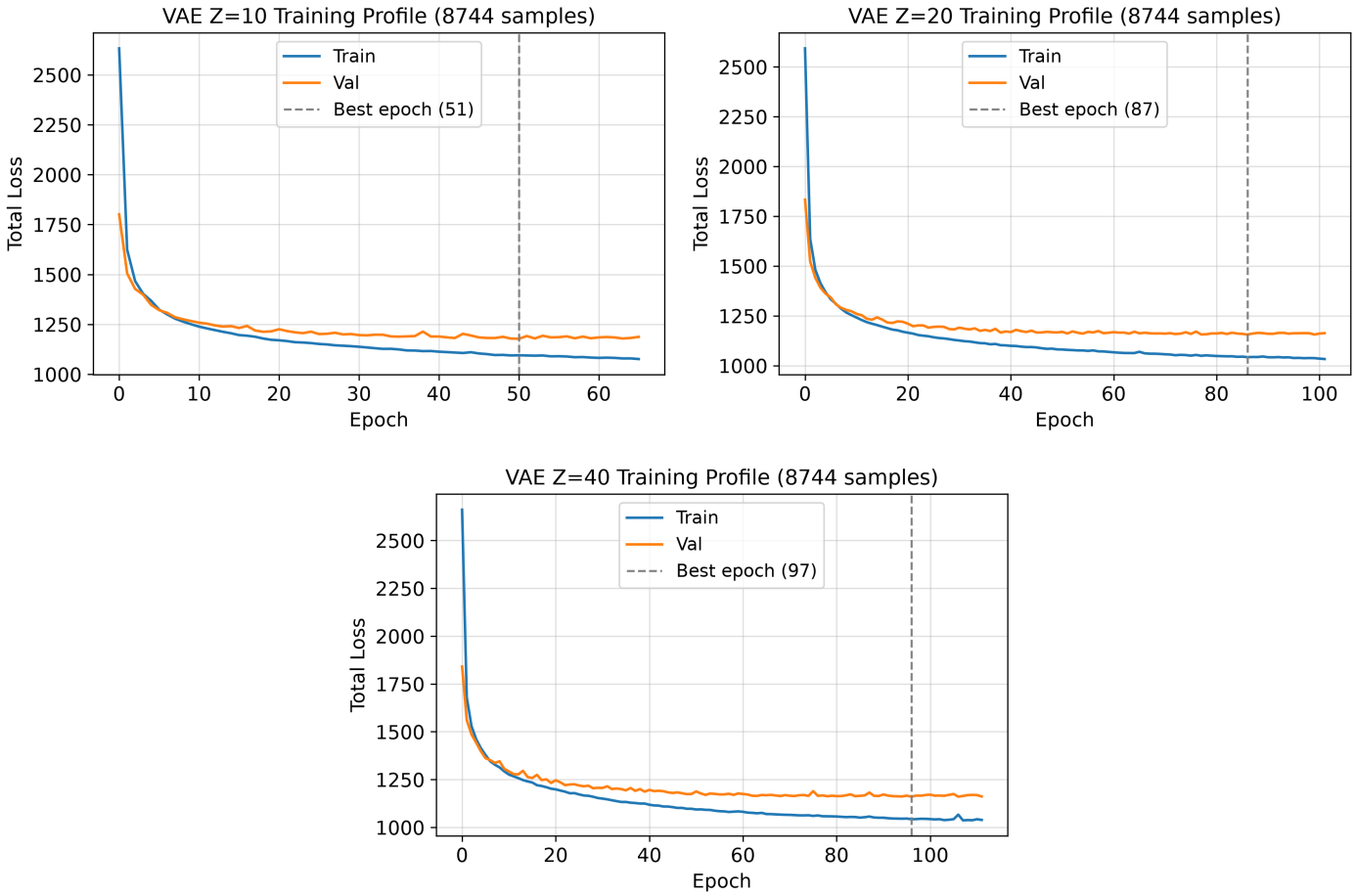


Figure 1. Training convergence of the VAE models trained directly on the primary analytical cohort ($n = 8744$; real, non-duplicated samples), across latent dimensionalities VAE $Z = 10, 20$, and 40. Early stopping (patience = 15 epochs) determined training duration for each configuration; the dashed line marks the best-validation-loss epoch, whose weights were restored for downstream use.

Comparative analysis of the best-epoch metrics (Table 2) shows that $Z = 20$ achieved both the lowest overall validation loss (1155.65) and the lowest reconstruction error (MSE = 1109.35) among the three configurations, followed closely by $Z = 40$ (validation loss = 1160.12; MSE = 1111.54), with $Z = 10$ clearly highest on both measures (validation loss = 1177.48; MSE = 1145.05). The $Z = 10$ model retained the lowest KLD (32.42), consistent with its more constrained bottleneck limiting latent complexity—although this constraint did not translate into a reconstruction advantage. The $Z = 40$ model's larger capacity increased KLD (48.59) without improving on $Z = 20$'s reconstruction fidelity,

suggesting the additional latent dimensions beyond $Z = 20$ captured comparatively little further structure in this cohort. Because training-loss trade-offs alone do not establish which latent dimensionality is most biologically informative, we deferred any designation of an optimal configuration to the clustering concordance and survival analyses below.

To test whether the latent space compression benefits observed for the whole cohort also held once the dominant tissue-of-origin signal was removed, the same three VAE architectures were independently retrained on the malignant-only subset ($n = 835$; Section 3.4), using the modified protocol. All three configurations converged smoothly, as shown in Figure 2, with no overfitting flagged at any latent dimensionality (training/validation gap 8.3%, 10.8%, and 12.9% of validation loss for $Z = 10$, 20, and 40, respectively—all under the 20% threshold, although widening steadily with latent capacity). Unlike the primary cohort, where $Z = 20$ achieved the lowest validation loss (Table 2), this subset showed the opposite ordering: $Z = 10$ achieved both the lowest validation loss (1506.98) and the lowest reconstruction error (MSE = 1480.75), with validation loss and KLD rising monotonically as latent capacity increased (Table 3). Reconstruction fidelity continued to improve marginally with capacity (MSE fell from 1480.75 to 1475.41 between $Z = 10$ and 40), but this gain was consistently outweighed by rising KLD, so total loss was minimized at the smallest latent dimensionality tested—a reversal that parallels the clustering-quality result for this same subset, where VAE $Z = 10$ also outperformed the wider configurations on every internal validation metric.

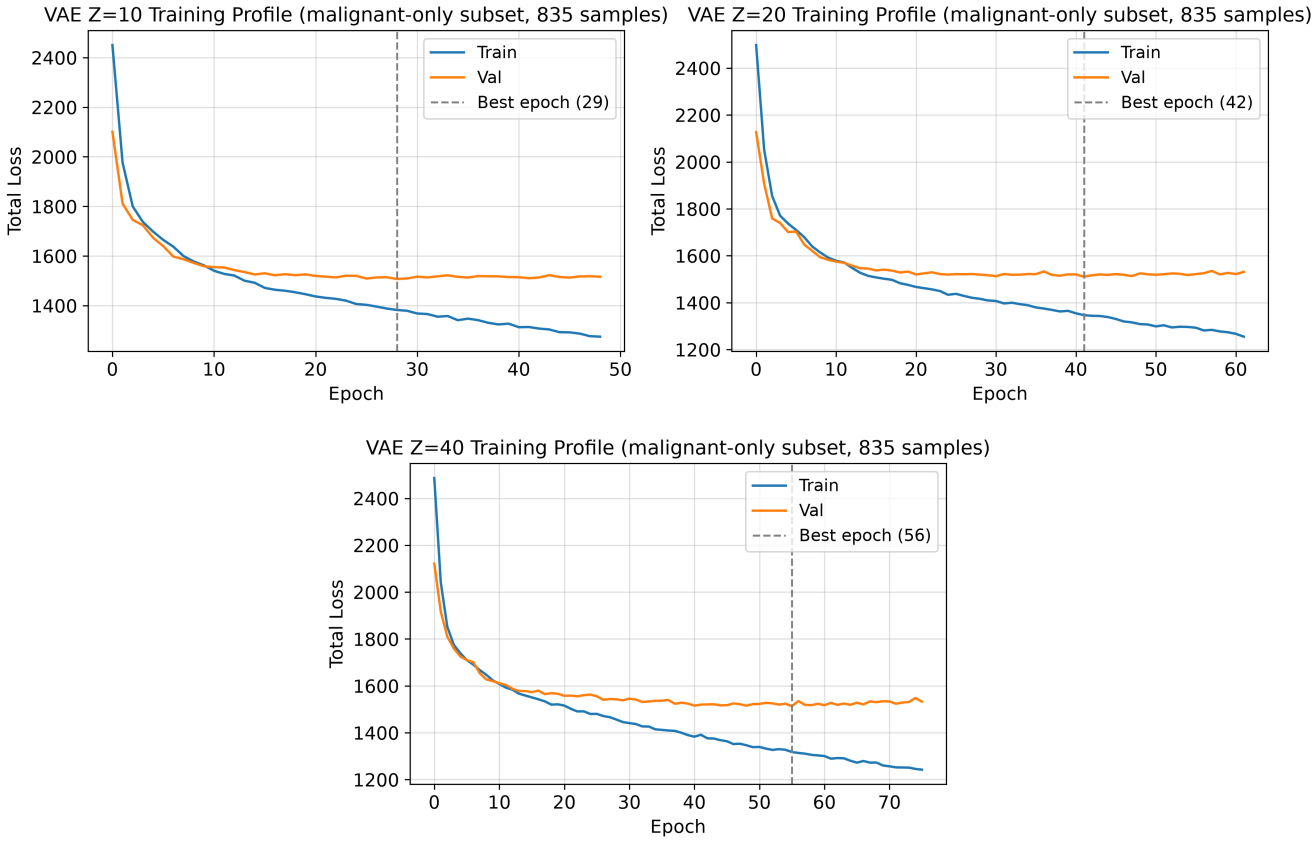


Figure 2. Training convergence of the VAE models trained on the malignant-only subset ($n = 835$; Section 3.4), across latent dimensionalities VAE $Z = 10$, 20, and 40. Early stopping (patience = 20 epochs) determined training duration for each configuration; the dashed line marks the best-validation-loss epoch, whose weights were restored for downstream use.

Table 3. VAE optimization outcomes across latent dimensionalities for the malignant-only subset ($n = 835$).

Latent Dim. (Z)	Best Epoch	Val. Total Loss	Val. MSE (Fidelity)	Val. KLD (Complexity)
$Z = 10$	29	1506.98	1480.75	26.22
$Z = 20$	42	1510.58	1476.28	34.30
$Z = 40$	56	1513.25	1475.41	37.84

3.2. Manifold Topology and Subtype Resolution

Two-dimensional UMAP projections, computed independently for each representation, provided a qualitative benchmark for the manifold structure recovered by the VAE relative to the linear PCA baseline, as shown in Figure 3. Across all four representations, Normal-GTEx samples separated into numerous distinct, well-isolated islands rather than forming a single mass—consistent with Normal-GTEx spanning multiple normal tissue types beyond breast, each retaining its own characteristic transcriptomic signature. The consistent recovery of this tissue-of-origin structure across every representation, including the linear PCA baseline, indicates that the projections captured a genuine biological signal rather than a representation-specific artifact.

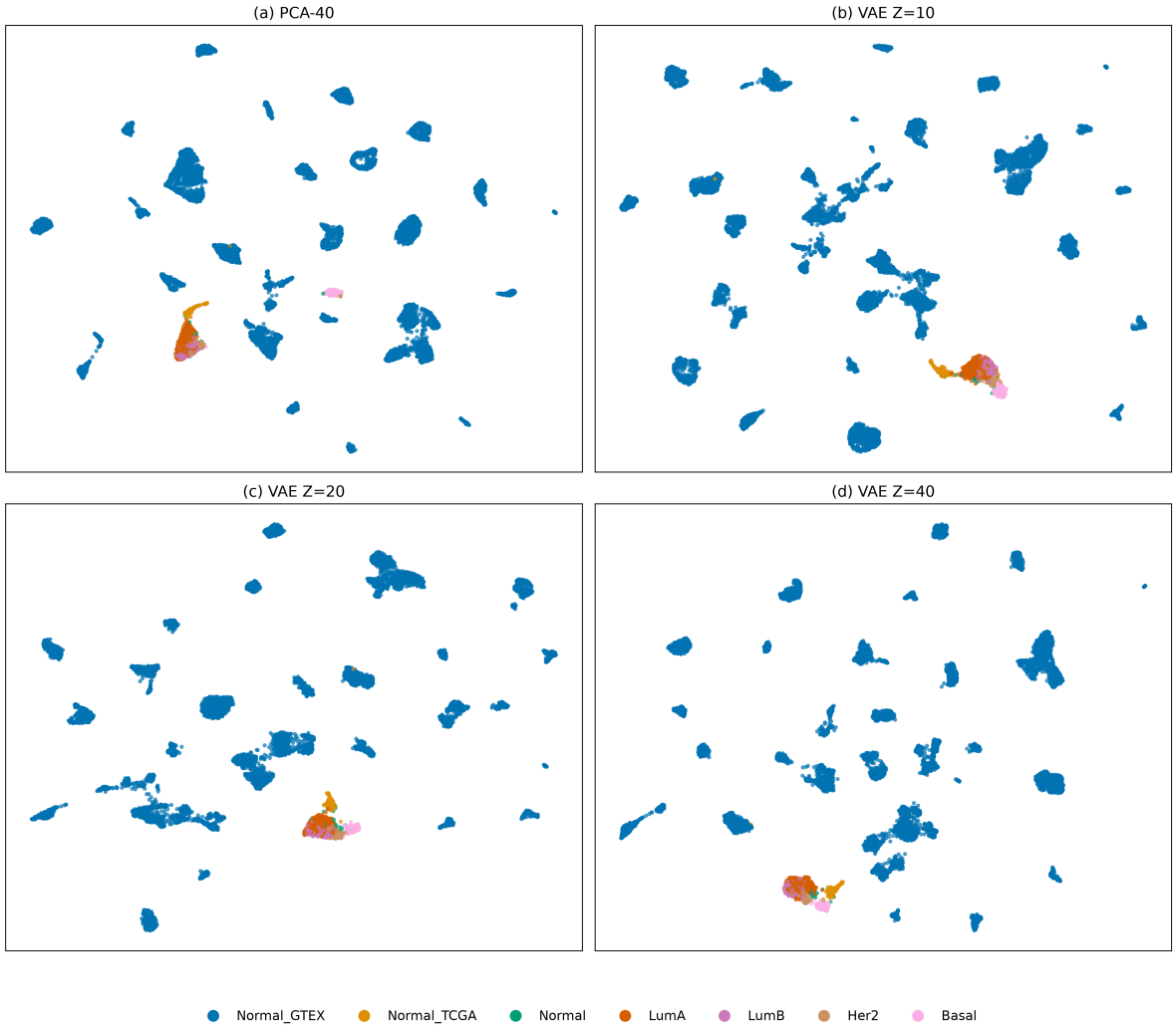


Figure 3. UMAP projections of the real 8744-sample cohort, colored by tissue origin and PAM50 molecular subtype, for (**a**) PCA-40, (**b**) VAE Z = 10, (**c**) VAE Z = 20, and (**d**) VAE Z = 40.

Tumor and Normal-TCGA samples occupy a region clearly set apart from the Normal-GTEx field in every representation. In the three VAE configurations, this region forms a single contiguous cluster, with Normal-TCGA samples forming a distinguishable tail extending from, but continuous with, the tumor mass. In the PCA baseline, by contrast, this region splits into two disconnected islands: a main cluster containing Luminal A, Luminal B, HER2-enriched, and most Normal-TCGA samples, and a fully separate island composed almost entirely of Basal-like tumors, indicating that, by simple visual separation, linear PCA isolates the Basal-like subtype more sharply than any of the three VAE configurations, where Basal-like samples instead sit at the periphery of, and remain contiguous with, the main tumor cluster. Within the main tumor cluster in all four representations, Luminal A and Luminal B samples remain substantially intermixed, with no representation producing a clean visual separation between them; the PAM50 Normal-like class is too sparsely represented in this cohort for its arrangement to be assessed visually.

This broader tissue-of-origin and tumor-region separation was consistent when re-examined using t-SNE instead of UMAP (Figure S1) and when t-SNE was re-run at perplexities of 30, 40, and 50 (Figure S2)—the latter check showing that both the coarse manifold structure and the finer Basal-like connectivity pattern for each representation held stably across the perplexity range. The finer-grained question of exactly how sharply Basal-like tumors separate from the rest of the tumor cluster, however, was not stable across dimensionality-reduction algorithms—the disconnected Basal-like island visible in the PCA-UMAP projection was not reproduced in the corresponding t-SNE projection, where it instead formed a connected peninsula, while the VAE projections showed the reverse pattern between the two methods. This instability is itself informative: it indicates that subcluster-level separation in a 2D projection is sensitive to the specific algorithm used and should not be over-interpreted, reinforcing that the quantitative cluster-validity and PAM50-concordance analyses in Sections 3.3 and 3.4, rather than this figure, are what should be relied on to assess how well each representation actually separates Basal-like tumors from the rest.

3.3. Comparative Internal Validation and Clustering Structure

Table 4 reports internal validation metrics at $k = 7$ for all four representations under all three clustering algorithms. Among representations, PCA-40 achieved the strongest internal geometry (highest silhouette and Calinski–Harabasz, lowest Davies–Bouldin), followed in order by VAE $Z = 10$, $Z = 20$, and $Z = 40$—a ranking that held consistently under hierarchical clustering, k-means, and GMM (full $k = 3$–8 hierarchical results are shown in Table S1), indicating that internal cluster tightness degrades monotonically as the VAE bottleneck widens, independent of which clustering algorithm is applied.

Table 4. Internal validation metrics at $k = 7$ for the primary analytical cohort ($n = 8744$).

Model	Silhouette Score (↑)	Calinski–Harabasz (↑)	Davies–Bouldin (↓)
PCA-40	0.2904	1992.2788	1.3445
$Z = 10$	0.1749	931.6753	1.7737
$Z = 20$	0.1189	578.4487	2.4245
$Z = 40$	0.0982	479.7356	2.4126

↑, higher values indicate better clustering; ↓, lower values indicate better clustering.

Biological composition told a more consistent story than the internal metrics alone would suggest (Table S2). Under Ward's hierarchical clustering, three of the four representations isolated an essentially pure malignant/PAM50-associated cluster: PCA-40 (99.8% pure, 2 stray Normal-TCGA samples out of 837), VAE $Z = 10$ (100% pure, 0 stray samples out of 830), and VAE $Z = 20$ (99.8% pure, 2 stray samples out of 834). VAE $Z = 40$ was

the exception, with 119 stray samples—the large majority being Normal-GTEx rather than Normal-TCGA—diluting purity to 87.5% (of 953 samples).

Across algorithms, hierarchical clustering produced the cleanest malignant-cluster boundary for three of the four representations. PCA-40 (99.8%) and VAE $Z = 10/20$ (100%/99.8%) all exceeded their k-means counterparts (90.2%, 89.3%, and 88.1%) and GMM counterparts (87.6%, 88.1%, and 87.9%). VAE $Z = 40$ reversed this ordering: k-means (89.5%) modestly outperformed hierarchical clustering (87.5%), with GMM again weakest (84.8%). In every case, the admixture into the malignant cluster came almost entirely from Normal-TCGA (bona fide solid-tissue-normal samples transcriptionally closest to tumor) rather than Normal-GTEx—except for VAE $Z = 40$ under hierarchical clustering, where GTEx contamination dominated instead. Full cluster-composition crosstabs for all representation–algorithm combinations are provided in Table S3.

Irrespective of representation or algorithm, no configuration fragmented the malignant/PAM50-associated cluster into subtype-resolved subclusters at $k = 7$: Basal-like, HER2-enriched, Luminal A, Luminal B, and Normal-like samples consistently co-occurred within the same cluster. Rather than separating PAM50 subtypes from one another, all three algorithms and all four representations consistently resolved the same larger-scale structure first—tissue of origin and malignant-versus-normal status.

3.4. Malignant-Only Clustering: Isolating Subtype Structure from Tissue of Origin

Section 3.3 established that none of the four representations fragmented the malignant/PAM50-associated cluster into subtype-resolved subclusters at any tested k; all four consistently resolved tissue-of-origin and malignant-versus-normal status first, with PAM50 subtypes remaining co-located within a single cluster. To rule out the possibility that this pattern reflects an artifact of dimensionality reduction, rather than an intrinsic property of the transcriptome itself, hierarchical clustering was additionally applied directly to the standardized 5000-gene matrix, with no PCA or VAE compression. This raw-feature-space clustering reproduced the identical pattern: a single, high-purity malignant cluster (99.9% pure, 834/835 samples) in which all five PAM50-associated classes co-occurred. These findings confirm that the tissue-of-origin/malignancy axis dominates the transcriptome's variance structure independently of representation, and that none of the methods tested so far was masking a latent subtype signal, as shown in Figure 4.

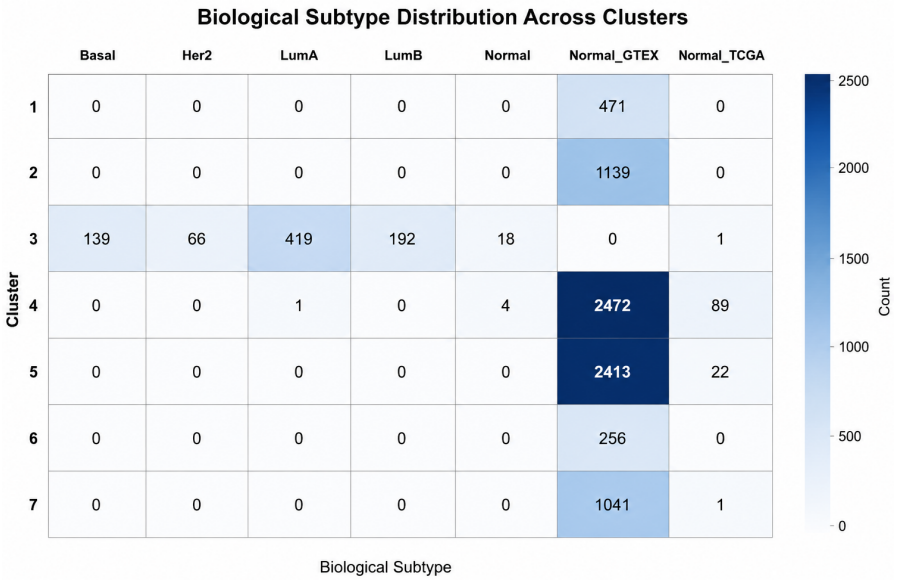


Figure 4. Biological subtype distribution across clusters ($k = 7$) from hierarchical clustering of the standardized 8744-sample × 5000-gene matrix.

This raised a distinct question: does subtype-specific structure genuinely not exist in this cohort, or is it present but obscured by the much larger tissue-of-origin and malignancy signal? To address this question without introducing label-based supervision, the 835 samples assigned by the raw-feature hierarchical clustering to its dominant malignant cluster were extracted using only the unsupervised cluster assignment, without reference to the PAM50 labels, and reclustered in isolation on the same standardized 5000-gene feature set. Hierarchical clustering (Ward), k-means, and Gaussian mixture model (GMM) clustering were evaluated across $k = 3$ to $k = 8$.

Silhouette score peaked sharply at $k = 3$ (0.0798) and then decreased by approximately an order of magnitude for every k from 4 to 8 (0.0023–0.0058), with no secondary peak across the evaluated range (Table 5). The Calinski–Harabasz index declined monotonically from 51.63 at $k = 3$ to 25.44 at $k = 8$, while the Davies–Bouldin index, for which lower values indicate better separation, was lowest at $k = 3$ (3.51) and worsened at every subsequent value of k (4.06–4.64). Thus, all three internal validation metrics converged on the same result: $k = 3$ represented the strongest mathematically supported partition of this population. Increasing k beyond 3 progressively reduced internal cluster cohesion and separation rather than revealing additional structure.

Table 5. Internal validation metrics (hierarchical, Ward linkage) across $k = 3$ to $k = 8$ for the malignant-only subset ($n = 835$).

k	Silhouette (↑)	Calinski–Harabasz (↑)	Davies–Bouldin (↓)
3	0.0798	51.6252	3.5121
4	0.0049	42.0772	4.0700
5	0.0039	36.0080	4.1205
6	0.0058	31.6343	4.0600
7	0.0023	28.0300	4.3778
8	0.0050	25.4401	4.6416

↑, higher values indicate better clustering; ↓, lower values indicate better clustering.

Because Ward's linkage nests cleanly across k (each additional k corresponds to a split of one previously defined cluster), comparing the biological subtype distributions across $k = 3$ through $k = 8$ revealed a specific order in which subtype-specific structure emerged. Basal-like separated first at $k = 3$, followed by the emergence of a distinct Luminal A cluster at $k = 4$, while HER2-enriched finally acquired its own dominant cluster at $k = 5$. Critically, the subsequent split at $k = 6$ did not reveal additional subtype-specific structure; instead, it simply divided the already-resolved Basal-like cluster into two Basal-like subgroups. Thus, $k = 5$ represents the most efficient biological resolution within the evaluated range: it is the smallest k at which three of the five reference subtypes each form a dominant cluster of their own, immediately before additional splits begin to re-partition structure that has already been resolved rather than revealing new subtype-specific organization.

The $k = 5$ hierarchical clustering solution (Table 6) makes this pattern concrete. Cluster 1 was 92.8% Basal-like (128/138), capturing 92.1% of all Basal-like samples in the subset. Cluster 3 was 89.6% Luminal A (129/144), capturing 30.8% of all Luminal A samples. HER2-enriched constituted the majority of Cluster 4, accounting for 52.9% of the cluster (37/70) and representing the only cluster across the entire k-sweep in which HER2-enriched constituted the plurality; it captured 56.1% of all HER2-enriched samples.

The remaining two clusters (Clusters 2 and 5) contained the unresolved Luminal A/Luminal B continuum. Luminal B never became the majority component of any cluster across the evaluated solutions; its highest within-cluster proportion was 39.7% in Cluster 2, where it remained second to Luminal A. Thus, $k = 5$ yielded three clearly interpretable subtype-dominant clusters and two clusters that retained the biologically

unresolved Luminal A/Luminal B continuum. Rather than forcing an artificial partition, this solution preserved the distinction between subtype structure that was clearly supported by the data and biological variation that did not separate sufficiently to justify an independent cluster.

Internal metrics and biological distributions both derive from the same clustering algorithm, so agreement between them alone does not establish that the observed structure is genuine; it could instead reflect an artifact of Ward's particular merge criterion. To address this concern, two structurally distinct clustering algorithms, k-means (centroid-based partitioning) and a Gaussian mixture model (GMM; probabilistic mixture model), were applied to the identical feature space at the identical k. This tests whether the same biological structure reappears independently of the clustering algorithm; agreement across all three methods provides stronger evidence that the observed pattern is a property of the data rather than a method-specific artifact.

Table 6. Biological subtype distribution across clusters at $k = 5$ for the malignant-only subset ($n = 835$), under three clustering algorithms.

Method	Cluster	Basal-like	HER2-enr	Luminal A	Luminal B	Normal-like	Normal-TCGA
Hierarchical	1	128	0	0	1	8	1
	2	0	8	112	79	0	0
	3	0	2	129	8	5	0
	4	5	37	7	20	1	0
	5	6	19	171	84	4	0
k-means	0	133	0	0	0	2	0
	1	0	3	165	58	0	0
	2	3	60	15	65	1	0
	3	3	1	166	7	13	1
	4	0	2	73	62	2	0
GMM	0	0	4	54	49	2	0
	1	3	55	17	69	1	0
	2	0	3	144	60	0	0
	3	135	1	0	0	3	0
	4	1	3	204	14	12	1

At $k = 5$, k-means assigned 98.5% of Cluster 0 (133/135) to the Basal-like subtype, capturing 95.7% of all Basal-like samples and producing a result at least as clean as the hierarchical solution. Clusters 1 and 3 were both Luminal A-dominant, with purities of 73.0% and 86.9%, respectively, and together captured 79.0% of all Luminal A samples, providing more complete Luminal A resolution than hierarchical clustering. Cluster 2 captured 90.9% of all HER2-enriched samples (60/66) in a single cluster, although Luminal B constituted the plurality of that cluster (45.1%), compared with 41.7% for HER2-enriched. Thus, k-means concentrated HER2-enriched samples even more tightly than hierarchical clustering, but did not produce a HER2-enriched-majority cluster. Luminal B again did not dominate any cluster; its highest within-cluster proportion was 45.1%, where it remained closely associated with HER2-enriched samples.

At $k = 5$, GMM assigned 97.1% of Cluster 3 (135/139) to Basal-like, capturing 97.1% of all Basal-like samples and closely matching the k-means and hierarchical results. Clusters 2 and 4 were both Luminal A-dominant, with purities of 69.6% and 86.8%, respectively, and together captured 83.1% of all Luminal A samples. This was consistent with the stronger Luminal A resolution observed under k-means compared with hierarchical clustering. Cluster 1 captured 83.3% of all HER2-enriched samples (55/66) in a single cluster; however, Luminal B constituted the plurality of this cluster (47.6%) rather than HER2-

enriched (37.9%). Luminal B therefore again failed to form a majority cluster, instead remaining closely associated with HER2-enriched samples.

All three algorithms, evaluated at the same biologically selected $k = 5$, agreed on every substantive finding. Basal-like separated cleanly and consistently, with 92–98% cluster purity and 92–97% sample capture across the three methods. Luminal A showed partial but reproducible resolution, forming one or two reasonably pure clusters, with more complete resolution under k-means and GMM than under hierarchical clustering. HER2-enriched consistently concentrated within a single cluster, becoming the majority component only under hierarchical clustering and remaining associated with Luminal B under k-means and GMM. Luminal B never formed the dominant component of any cluster under any algorithm at this k or at any other tested value.

Thus, the finding is not that clustering fails to recover PAM50 structure. Rather, the results provide a precise, cross-algorithm-validated description of which subtypes separate robustly and which remain unresolved. The consistent separation of Basal-like, partial resolution of Luminal A, concentration of HER2-enriched alongside Luminal B, and persistent lack of an independent Luminal B cluster indicate that the observed subtype structure is reproducible across fundamentally different clustering paradigms rather than being a method-specific artifact. This pattern is consistent with the independently documented degree of transcriptomic distinctiveness among the PAM50 subtypes.

3.4.1. Malignant-Only Representation Learning: PCA-40 and VAE Reclustering

Section 3.4 established, using the raw standardized 5000-gene feature space, that Basal-like separates cleanly under every algorithm tested, HER2-enriched concentrates but only becomes a cluster's majority under hierarchical clustering (52.9% purity), Luminal A resolves only partially (best capture, 30.8%), and Luminal B never leads any cluster at any k or under any algorithm. Because this study's central methodological contribution is nonlinear representation learning via a VAE, the natural next test is whether compressing this same 835-sample malignant-only population through PCA-40 and VAE ($Z = 10$, 20, 40) models, refit specifically on these samples rather than reused from the whole-cohort models, improves subtype resolution beyond what the raw gene space achieved. All four representations were clustered independently using the same hierarchical (Ward), k-means, and GMM protocol as in Section 3.4, across $k = 3$ to $k = 8$.

Table 7 reports internal validation metrics for all four malignant-subset-specific representations under hierarchical clustering, held at $k = 5$ to match the biologically motivated resolution already established in Section 3.4, for a direct comparison. VAE $Z = 10$ outperforms every other representation on all three metrics simultaneously: the highest silhouette (0.1401, vs. 0.0880 for PCA-40, 0.0750 for VAE $Z = 20$, and 0.0892 for VAE $Z = 40$), the highest Calinski–Harabasz (109.81, vs. 85.58, 77.44, and 70.34, respectively), and the lowest Davies–Bouldin (1.85, vs. 2.34, 2.13, and 2.56, respectively). Because all three metrics agree without exception at this shared k, VAE $Z = 10$ is the strongest-performing representation among the four malignant-subset representations—a reversal of the primary-cohort ranking (Table 4; Section 3.3), where PCA-40 led on every metric. This mirrors the training-loss result in Section 3.1, where VAE $Z = 10$ also achieved the lowest validation loss on this same subset: once separated from the dominant tissue-of-origin signal, the malignant-only population appears better served by a more constrained latent space than the broader, tissue-diverse primary cohort.

Table 7. Internal validation metrics (hierarchical, Ward linkage, $k = 5$) for all four malignant-subset-specific representations ($n = 835$).

Model	Silhouette (↑)	Calinski–Harabasz (↑)	Davies–Bouldin (↓)
PCA-40	0.0880	85.5786	2.3398
VAE $Z = 10$	0.1401	109.8056	1.8538
VAE $Z = 20$	0.0750	77.4355	2.1251
VAE $Z = 40$	0.0892	70.3447	2.5605

↑, higher values indicate better clustering; ↓, lower values indicate better clustering.

Table 8 reports the full biological subtype distribution at $k = 5$ under hierarchical clustering for each representation. Basal-like separates cleanly under every representation (83.4–95.2% purity, 85.6–96.4% capture), confirming this subtype's separation is a property of the underlying biology rather than any particular representation. HER2-enriched shows the sharpest representation-dependence of any subtype: VAE $Z = 20$ achieves 72.9% purity (43/59, Cluster 1)—the cleanest HER2-enriched resolution obtained anywhere in this study, exceeding both the raw gene result (52.9%; Section 3.4) and PCA-40 (57.3%). VAE $Z = 10$ and VAE $Z = 40$, by contrast, both lose HER2-enriched's cluster-majority status entirely: in each, the HER2-heavy cluster's plurality shifts to Luminal B (42.0% versus 36.2% for VAE $Z = 10$; 40.5% versus 37.4% for VAE $Z = 40$), though HER2-enriched capture remains high in both (75.8% and 92.4%, respectively). Luminal A achieves its best resolution anywhere in this study under VAE $Z = 10$ (85.2% purity, 67.1% capture, Cluster 2)—more than double the raw gene capture of 30.8%—with VAE $Z = 40$ also improving substantially on the raw gene and PCA-40 results (84.1% purity, 56.8% capture). Luminal B never becomes a cluster's majority under hierarchical clustering in any of the four representations at $k = 5$; its closest approach is a narrow plurality over HER2-enriched in VAE $Z = 10$ (42.0% versus 36.2%) and VAE $Z = 40$ (40.5% versus 37.4%), in both cases within a cluster that remains a near-even split between the two subtypes rather than a clean Luminal B resolution.

Table 8. Biological subtype distribution across clusters at $k = 5$ for all four malignant-subset-specific representations ($n = 835$), hierarchical clustering.

Model	Cluster	Basal-like	HER2-Enriched	Luminal A	Luminal B	Normal-like	Normal-TCGA
PCA-40	1	134	1	1	1	9	1
	2	0	1	180	7	5	0
	3	5	43	7	17	3	0
	4	0	0	30	14	0	0
	5	0	21	201	153	1	0
VAE $Z = 10$	1	121	5	0	0	9	1
	2	1	2	281	40	6	0
	3	16	50	13	58	1	0
	4	1	4	53	46	0	0
	5	0	5	72	48	2	0
VAE $Z = 20$	1	7	43	1	8	0	0
	2	121	6	5	8	5	0
	3	1	10	131	118	0	0
	4	1	2	117	29	5	0
	5	9	5	165	29	8	1
VAE $Z = 40$	1	5	61	30	66	1	0
	2	119	1	0	0	5	0
	3	5	2	238	28	9	1
	4	3	0	63	50	0	0
	5	7	2	88	48	3	0

Across both criteria, VAE $Z = 10$ ranks first overall for this malignant-only population: it is the unambiguous mathematical leader at $k = 5$ (Table 7), and it yields the single largest biological gain of any representation tested, more than doubling Luminal A's capture relative to the raw gene result (67.1% versus 30.8%). VAE $Z = 20$ ranks second: it does not lead on the internal validation metrics, and its HER2-enriched capture (65.2%) is lower than VAE $Z = 10$'s (75.8%), but it earns this position for what happens inside the HER2-enriched cluster itself. VAE $Z = 20$'s HER2-heavy cluster ($n = 59$) is 72.9% HER2-enriched—a genuine majority, and the only cluster across the raw genes, PCA-40, and all three VAE configurations tested in this study where HER2-enriched purity exceeds 70%. VAE $Z = 10$'s HER2-heavy cluster, by contrast, is more than twice the size ($n = 138$) and captures more HER2-enriched samples in absolute terms (50 versus 43), but Luminal B is actually the larger group within it (58 versus 50; 42.0% versus 36.2%), so despite the higher raw count that cluster cannot be read as "the HER2-enriched cluster" the way VAE $Z = 20$'s can. This is why VAE $Z = 20$ is retained as the second-ranked representation rather than discarded: it is the only configuration in which HER2-enriched, rather than another subtype, is actually the majority inside its own cluster. Neither representation, nor any other tested in this study, resolves Luminal B into its own cluster under hierarchical clustering at this k—consistent with the raw gene finding in Section 3.4 that Luminal B's separation is the weakest point of PAM50 subtype structure in this cohort, independent of representation.

3.4.2. Cross-Algorithm Validation of the VAE Z = 10 and VAE Z = 20 Ranking

The ranking established above rests on hierarchical clustering alone, which is this study's primary algorithm. To test how much of that ranking is a property of the data rather than of hierarchical clustering specifically, VAE $Z = 10$ and VAE $Z = 20$ were each reclustered with k-means and GMM at the same $k = 5$, and their internal metrics and biological subtype distributions were compared against the hierarchical results, as shown in Tables 8–10.

Table 9. Internal validation metrics at $k = 5$ for VAE $Z = 10$ and VAE $Z = 20$, compared across all three clustering algorithms ($n = 835$).

Model	Algorithm	Silhouette (↑)	Calinski–Harabasz (↑)	Davies–Bouldin (↓)
VAE $Z = 10$	Hierarchical	0.1401	109.8056	1.8538
	k-means	0.1307	131.6492	1.7730
	GMM	0.0887	103.4481	2.1996
VAE $Z = 20$	Hierarchical	0.0750	77.4355	2.1251
	k-means	0.1338	103.0694	1.9208
	GMM	0.1110	92.8105	2.1716

↑, higher values indicate better clustering; ↓, lower values indicate better clustering.

Table 9 compares VAE $Z = 10$ and VAE $Z = 20$ at $k = 5$ across all three algorithms. Under hierarchical clustering, VAE $Z = 10$ performs best across all three metrics (Table 7). Under k-means, the picture is mixed: VAE $Z = 20$ has the higher silhouette (0.1338 versus 0.1307), but VAE $Z = 10$ has the higher Calinski–Harabasz (131.65 versus 103.07) and the lower Davies–Bouldin (1.7730 versus 1.9208)—with two of the three metrics favoring VAE $Z = 10$. Under GMM, the split reverses direction: VAE $Z = 20$ has the higher silhouette (0.1110 versus 0.0887) and the lower Davies–Bouldin (2.1716 versus 2.1996), while VAE $Z = 10$ keeps the higher Calinski–Harabasz (103.45 versus 92.81)—with two of the three metrics favoring VAE $Z = 20$. VAE $Z = 10$'s mathematical advantage is therefore strongest and most consistent under the primary algorithm, and narrows to a near-even split once k-means and GMM are included.

Table 10. Biological subtype distribution across clusters at $k = 5$ for VAE $Z = 10$ and VAE $Z = 20$ under k-means and GMM clustering ($n = 835$).

Model	Algorithm	Cluster	Basal-like	HER2-Enriched	Luminal A	Luminal B	Normal-like	Normal-TCGA
VAE $Z = 10$	k-means	0	15	2	169	8	14	1
		1	17	54	11	52	1	0
		2	0	10	180	86	0	0
		3	0	0	59	46	2	0
		4	107	0	0	0	1	0
	GMM	0	0	2	49	39	2	0
		1	1	2	191	16	9	0
		2	3	55	85	59	2	0
		3	0	5	94	78	0	0
		4	135	2	0	0	5	1
VAE $Z = 20$	k-means	0	122	0	0	0	2	0
		1	2	3	60	67	0	0
		2	9	2	214	16	16	1
		3	0	1	117	44	0	0
		4	6	60	28	65	0	0
	GMM	0	2	1	61	54	0	0
		1	1	2	200	13	9	0
		2	131	0	0	0	3	1
		3	0	3	112	30	2	0
		4	5	60	46	95	4	0

Basal-like and Luminal A both reproduce well under the alternative algorithms for both representations, confirming that these two results are properties of the data rather than artifacts of hierarchical clustering. Basal-like remains a clean, near-exclusive cluster majority in every case: 89.0% purity and 87.1% capture for VAE $Z = 10$ and 83.4% purity and 87.1% capture for VAE $Z = 20$ under hierarchical (Table 8); 99.1%/77.0% and 98.4%/87.8%, respectively, under k-means; and 94.4%/97.1% and 97.0%/94.2%, respectively, under GMM. Luminal A likewise keeps a majority cluster under every algorithm for both models, with capture generally improving under the alternative algorithms relative to hierarchical—for example, VAE $Z = 10$ reaches 83.3% combined Luminal A capture under k-means (Clusters 0 and 2, 169/419 and 180/419) and VAE $Z = 20$ reaches 79.0% combined capture under k-means (Clusters 2 and 3, 214/419 and 117/419).

HER2-enriched is where the two representations diverge most sharply once k-means and GMM are brought in. VAE $Z = 20$'s clean 72.9% HER2-enriched cluster majority (Table 8) is specific to hierarchical clustering and does not reproduce under either alternative algorithm: under k-means, the equivalent cluster ($n = 159$) becomes Luminal B-led instead (40.9% Luminal B versus 37.7% HER2-enriched), and under GMM the same reversal occurs in its equivalent cluster ($n = 210$; 45.2% Luminal B versus 28.6% HER2-enriched). In both cases, HER2-enriched capture inside that cluster stays high—60 of the cohort's 66 HER2-enriched samples (90.9%) are still concentrated there—so the algorithm keeps finding the same group of samples, but no longer returns HER2-enriched as that cluster's dominant label. VAE $Z = 10$ never achieves even a plurality for HER2-enriched under any of its three algorithms: Luminal B leads its HER2-heavy cluster under hierarchical (42.0% versus 36.2%, Table 8), HER2-enriched only barely edges Luminal B under k-means (40.0% versus 38.5%, Cluster 1, capturing 54 of 66 HER2-enriched samples), and Luminal A leads outright under GMM (41.7% versus 27.0% HER2-enriched, Cluster 2, which still captures 55 of 66). Taken together, VAE $Z = 20$'s HER2-enriched result is real but algorithm-dependent rather than a robust, cross-algorithm finding, while VAE $Z = 10$ does not produce a comparable HER2-enriched signal under any algorithm tested.

A new result appears only under k-means: VAE $Z = 20$'s Cluster 1 ($n = 132$) is 50.8% Luminal B (67/132)—an outright cluster majority. This is the only instance, across every representation, algorithm, and k tested in this study, in which Luminal B leads a cluster

outright rather than only a plurality. It does not reproduce under hierarchical or GMM for the same representation, nor does it appear for VAE $Z = 10$ under any algorithm, so it is reported here as an isolated, algorithm-specific observation rather than a validated resolution of Luminal B.

Cross-algorithm validation qualifies, rather than reverses, the ranking established under hierarchical clustering. VAE $Z = 10$'s mathematical lead is unanimous under the primary algorithm but narrows to a near-even 2–1 split under each alternative algorithm, so it should be read as the stronger representation on balance rather than a lead confirmed by every method. Basal-like and Luminal A resolution are robust findings that hold under all three algorithms for both representations. VAE $Z = 20$'s HER2-enriched cluster majority, by contrast, is specific to hierarchical clustering: k-means and GMM keep the same samples concentrated together (90.9% capture in both) but relabel the cluster's majority as Luminal B, and VAE $Z = 10$ never produces a comparable HER2-enriched signal under any algorithm. The one Luminal B majority observed anywhere in this study—VAE $Z = 20$ under k-means—is similarly algorithm-specific and does not reproduce elsewhere. Taken together, these results support retaining VAE $Z = 10$ as the primary malignant-subset representation and VAE $Z = 20$ as a secondary representation whose HER2-enriched resolution, while genuine under this study's primary clustering algorithm, should be reported as algorithm-dependent rather than as a fully consistent result.

3.5. Relationship to PAM50 and Clinical Features

Table 11 reports the Adjusted Rand Index and Normalized Mutual Information between each malignant-subset representation's hierarchical clustering ($k = 5$) and the PAM50 reference labels.

Table 11. PAM50 concordance (ARI, NMI) for all four malignant-subset-specific representations, hierarchical clustering ($k = 5$, $n = 835$).

Model	ARI	NMI
PCA-40	0.3122	0.4524
VAE $Z = 10$	0.3767	0.3962
VAE $Z = 20$	0.2435	0.3448
VAE $Z = 40$	0.3034	0.3716

VAE $Z = 10$ achieves the highest ARI of the four representations (0.3767), the strongest pairwise agreement with PAM50 of any malignant-subset model—consistent with its status as the strongest-performing representation mathematically and biologically already established in Section 3.4.1. PCA-40, however, achieves the highest NMI (0.4524) despite a lower ARI than VAE $Z = 10$ (0.3122 versus 0.3767): PCA-40 preserves more of PAM50's global information structure, but its cluster boundaries are comparatively less discrete, consistent with the mixed, near-even Luminal A/Luminal B cluster already identified in Section 3.4.1 (Cluster 5: 53.5% Luminal A, 40.7% Luminal B). All four values are well above the value of 0 expected under random labeling, confirming that clustering on these representations recovers genuine subtype-associated structure rather than noise, but the absolute magnitudes (ARI 0.24–0.38, NMI 0.34–0.45) are moderate rather than high—appropriately so, since the malignant-only subset was constructed specifically to strip out the dominant tissue-of-origin signal that drives higher concordance in the whole-cohort analysis (Section 3.4), leaving only the comparatively subtler inter-subtype transcriptomic differences for the clustering to recover in an unsupervised manner.

Notably, VAE $Z = 20$ has the lowest ARI (0.2435) and lowest NMI (0.3448) of all four representations, despite being retained in Section 3.4.2 as the secondary representation. This is not a contradiction: ARI and NMI are aggregate measures across all clusters and all

subtypes, whereas VAE $Z = 20$'s value lies in one specific, narrow signal—its unusually clean HER2-enriched cluster (72.9% purity, Section 3.4.1)—which a global concordance statistic does not specifically reward and which is diluted in the average by VAE $Z = 20$'s comparatively weaker performance elsewhere (Section 3.4.1, Cluster 3: 50.4% Luminal A, 45.4% Luminal B, its largest cluster). This reinforces the primary/secondary framing already established: VAE $Z = 10$ is the stronger representation on overall concordance with PAM50, while VAE $Z = 20$'s value is confined to a specific subtype resolution rather than broad agreement.

3.5.1. Refinement of PAM50 and Biological Bridging

The $k = 5$ contingency tables presented in Section 3.4.1 (Table 8) were examined for two specific patterns: subtypes resolved into more than one biologically interpretable cluster (refinement), and clusters not dominated by any single subtype but instead shared between two (bridging).

Refinement of the Luminal A subtype was observed under every representation tested (Table 8). Rather than occupying a single cluster, Luminal A samples were distributed across one predominantly pure cluster and one or more clusters co-dominated by Luminal B, a pattern that recurred independently across all four representations. Under VAE $Z = 10$, 67.1% of Luminal A samples were captured by a single 85.2%-pure cluster (Cluster 2), while the remaining samples were distributed across two smaller clusters in which Luminal A remained the plurality but Luminal B constituted a substantial fraction (Clusters 4 and 5; 51.0% and 56.7% Luminal A, respectively). A comparable pattern was observed under VAE $Z = 40$, where 56.8% of Luminal A samples were captured by an 84.1%-pure cluster (Cluster 3), with the remainder again distributed across mixed Luminal A/Luminal B clusters. The split was most pronounced under PCA-40, where Luminal A divided almost evenly between a pure cluster (Cluster 2; 93.3% purity, 43.0% capture) and a large mixed cluster of comparable size (Cluster 5; 53.5% Luminal A, 40.7% Luminal B, 48.0% capture). Because this bipartite structure recurred across independent representations rather than a single model, it is unlikely to reflect a representation-specific artifact; it instead suggests the presence of at least two transcriptomically distinguishable populations within the Luminal A subtype as currently defined by PAM50, one more distinct from Luminal B and one considerably closer to it.

Several clusters in Table 8 were not dominated by a single PAM50 subtype but were instead shared approximately equally between two subtypes, consistent with a transitional population rather than a misclassification. The clearest example occurred under VAE $Z = 20$, where the largest cluster (Cluster 3; $n = 260$) divided almost evenly between Luminal A (50.4%) and Luminal B (45.4%), with negligible representation of any other subtype. A structurally similar cluster was observed under VAE $Z = 10$ (Cluster 3; $n = 138$), in which HER2-enriched (36.2%) and Luminal B (42.0%) were jointly dominant; this is the same cluster discussed in Section 3.4.2, where it accounted for VAE $Z = 10$'s inability to resolve a distinct HER2-enriched cluster. Under PCA-40, the largest cluster overall (Cluster 5; $n = 376$) similarly mixed Luminal A (53.5%) and Luminal B (40.7%), but at a substantially larger scale, capturing close to half of all Luminal A and Luminal B samples in the cohort within a single, comparatively unresolved cluster. Across all four representations, the recurring bridge was specifically between Luminal A and Luminal B, with VAE $Z = 10$ additionally producing a HER2-enriched/Luminal B bridge; no cluster spanned three or more subtypes at $k = 5$. Basal-like was not represented in any mixed cluster under any representation, consistent with the raw gene finding reported in Section 3.4, and further supporting Basal-like as the most transcriptomically distinct subtype in this cohort independent of representation.

3.5.2. Statistical Significance Analysis

Table 12 reports the chi-square and Kruskal–Wallis statistics and associated p-values for the association between hierarchical cluster assignment ($k = 5$) and each clinical/demographic variable, across all four malignant-subset representations. Because degrees of freedom differ across variables (df = 4 for ER, PR, and HER2 status and for the Kruskal–Wallis test on age; df = 12 for AJCC stage), the chi-square/H statistic is comparable across models only within the same variable; comparisons across variables rely on the p-value alone.

Table 12. Clinical and demographic association between hierarchical clusters ($k = 5$) and patient-level variables for all four malignant-subset representations (χ^2 test, df = 4 for ER/PR/HER2 status and df = 12 for AJCC stage; Kruskal–Wallis H test df = 4 for age).

Model	ER χ^2	ER p	PR χ^2	PR p	HER2 χ^2	HER2 p	AJCC χ^2	AJCC p	Age H	Age p
PCA-40	470.88	1.33×10^{-100}	338.19	6.21×10^{-72}	87.46	4.55×10^{-18}	29.84	2.96×10^{-3}	38.58	8.50×10^{-8}
VAE $Z = 10$	431.17	5.11×10^{-92}	336.00	1.85×10^{-71}	109.05	1.16×10^{-22}	36.45	2.75×10^{-4}	20.51	3.96×10^{-4}
VAE $Z = 20$	416.11	9.17×10^{-89}	291.34	8.00×10^{-62}	83.44	3.24×10^{-17}	22.65 a	3.08×10^{-2}	23.88	8.44×10^{-5}
VAE $Z = 40$	387.89	1.15×10^{-82}	294.01	2.13×10^{-62}	124.00	7.48×10^{-26}	39.00 a	1.05×10^{-4}	29.40	6.48×10^{-6}

a: Expected cell counts sparse (25% of cells < 5).

ER, PR, and HER2 status were significantly associated with cluster assignment for all four representations ($p = 3.24 \times 10^{-17}$ to 1.33×10^{-100}; Table 12), consistent with these receptor statuses defining the PAM50 subtypes against which the clusters were validated in Section 3.5. VAE $Z = 20$ showed the lowest chi-square statistic among the four representations for both ER (416.11) and PR (291.34), consistent with its weaker ARI/NMI concordance reported in Section 3.5. HER2 status was most strongly associated with cluster assignment under VAE $Z = 40$ ($\chi^2 = 124.00$), followed by VAE $Z = 10$ (109.05), PCA-40 (87.46), and VAE $Z = 20$ (83.44).

AJCC stage was significantly associated with cluster assignment for all four representations, though less strongly than receptor status ($p = 1.05 \times 10^{-4}$ to 3.08×10^{-2}; Table 12). VAE $Z = 20$ produced both the weakest association ($p = 0.0308$) and the only result affected by sparse expected cell counts (25% of cells < 5); this result should be interpreted with that limitation noted.

Age at diagnosis differed significantly across clusters for all four representations ($p < 4 \times 10^{-4}$; Table 12). The association was strongest under PCA-40 ($H = 38.58$, $p = 8.50 \times 10^{-8}$) and weakest under VAE $Z = 10$ ($H = 20.51$, $p = 3.96 \times 10^{-4}$); this ordering does not correspond to the models' relative ranking on PAM50 concordance (Section 3.5).

To assess whether the malignant-subset clusters carry prognostic information beyond PAM50 subtyping, overall survival was analyzed for all four representations using the Cluster_Hier_k5 assignments (Table 13). For each representation, a multivariate log-rank test compared survival across the five clusters, a Cox proportional hazards model based on cluster identity alone was compared to an equivalent PAM50-only model on the identical sample subset via the concordance index (C-index), and a nested likelihood ratio test evaluated whether cluster membership remained prognostic after adjusting for age and AJCC stage.

Table 13. Survival analysis: log-rank test, concordance index (vs. PAM50), and age/stage-adjusted independent-marker test for the four malignant-subset representations (hierarchical clustering, $k = 5$).

Model	n (Survival)	Log-Rank χ^2	Log-Rank p	C-Index (AI Clusters)	C-Index (PAM50)	n (Adjusted)	Adjusted LR χ^2 ($df = 4$)	Adjusted LR p	Independent Marker
PCA-40	790	4.86	0.302	0.562	0.606	740	2.23	0.693	No
VAE $Z = 10$	790	9.67	0.046	0.619	0.606	740	5.12	0.275	No
VAE $Z = 20$	790	5.48	0.242	0.539	0.606	740	2.49	0.646	No
VAE $Z = 40$	790	15.58	0.0036	0.643	0.606	740	3.81	0.433	No

The four representations separated into two groups. VAE $Z = 10$ (log-rank $p = 0.046$) and VAE $Z = 40$ (log-rank $p = 0.0036$) produced clusters with significantly different survival curves, whereas PCA-40 ($p = 0.302$) and VAE $Z = 20$ ($p = 0.242$) did not. These same two representations were also the only ones whose cluster-only C-index (0.619 and 0.643, respectively) exceeded the PAM50-only C-index computed on the identical 790-sample subset (0.606), indicating that cluster membership ranks patient survival marginally better than PAM50 subtype alone. Figure 5 shows the corresponding Kaplan–Meier curves. For VAE $Z = 10$ (Figure 5a), Cluster 3 shows the earliest and most sustained decline in survival probability, while Clusters 1, 2, and 5 retain the highest survival probability through most of the follow-up period. For VAE $Z = 40$ (Figure 5b), Cluster 1 is the clearest poor-prognosis group, declining steadily to approximately 0.20 by day 4400, while Clusters 3 and 5 remain the highest-surviving groups, plateauing near 0.39–0.43.

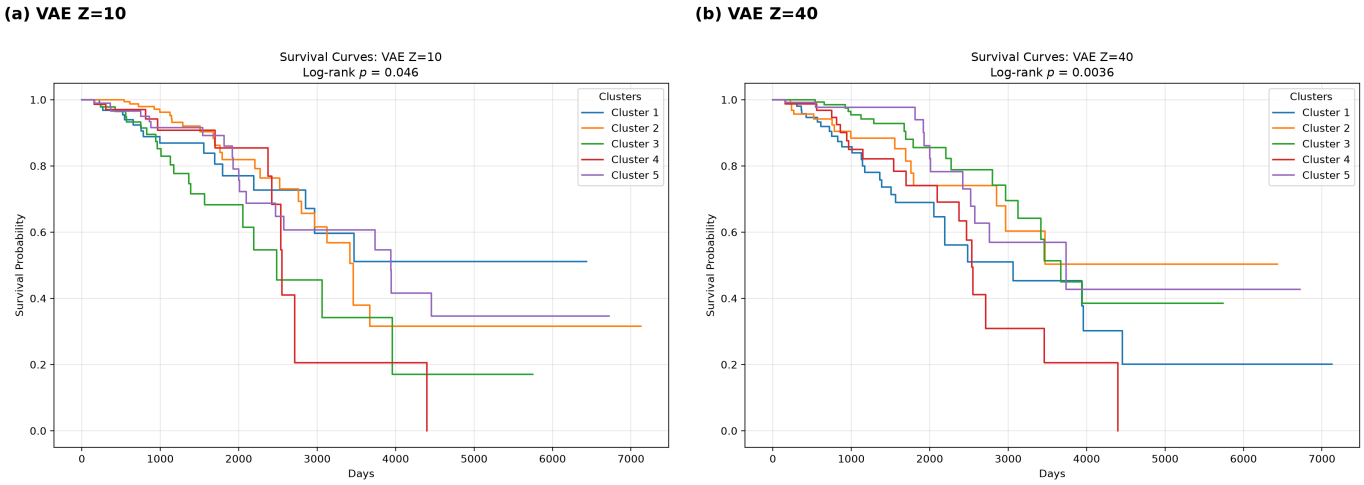


Figure 5. Kaplan–Meier overall survival curves by hierarchical cluster ($k = 5$) for the two malignant-subset representations whose concordance index exceeded PAM50's: (**a**) VAE $Z = 10$ (log-rank $p = 0.046$); (**b**) VAE $Z = 40$ (log-rank $p = 0.0036$).

Despite this univariate signal, none of the four representations reached significance on the age- and stage-adjusted likelihood ratio test (Independent Marker = No for all; Table 13); VAE $Z = 10$ came closest (adjusted LR $p = 0.275$), followed by VAE $Z = 40$ ($p = 0.433$). This does not contradict the log-rank and C-index findings above, but reflects a difference in statistical stringency: the log-rank test and C-index comparison ask whether cluster membership is associated with survival at all, which VAE $Z = 10$ and VAE $Z = 40$ answer affirmatively, whereas the adjusted test asks the stricter question of whether cluster membership adds prognostic information beyond age and stage, evaluated jointly across four correlated cluster-indicator coefficients. At the available sample size (740 patients with complete age, stage, and survival data), an effect of this magnitude does not clear that higher bar. Taken together, these results indicate that VAE $Z = 10$ and VAE $Z = 40$ capture a genuine, moderate survival signal that modestly outperforms PAM50 in ranking

patient risk, but do not yet establish the unsupervised subtypes as independent prognostic markers distinct from established clinicopathological staging—consistent with the strong concordance between these clusters and stage/receptor status already reported in Table 12.

3.6. Biological Pathways Defining Each Cluster

Differential expression analysis identified cluster-specific driver genes for each of the five VAE $Z = 10$ (hierarchical, $k = 5$) clusters via a one-vs.-rest comparison against all other malignant-subset samples; the ten highest-ranked driver genes per cluster are reported in Supplementary Table S6. Table 14 summarizes the resulting provisional cluster identities. Four of the five clusters showed a single coherent direction of change among their driver genes. Cluster 1 was defined by coordinated down-regulation of core luminal-differentiation genes (*FOXA1*, *GATA3*, *AR*, *TFF3*, *AGR2*), consistent with its unmixed, non-luminal composition reported in Table 8. Cluster 2—the 85.2%-pure Luminal A population identified in Section 3.4.1—showed coordinated up-regulation of extracellular-matrix and stromal genes (*LAMA2*, *COL14A1*, *ECM2*, *SPARCL1*). Cluster 3, the HER2-enriched/Luminal B bridging cluster described in Section 3.4.2, showed down-regulation of luminal-lineage-restricted genes (*CBLC*, *LINC00472*). Cluster 5, the Luminal-B-adjacent Luminal A subset discussed in Section 3.5.1, showed the inverse pattern of Cluster 2: down-regulation of stromal/fibroblast-receptor genes (*PDGFRA*, *DDR2*, *CAV2*). Cluster 4 was the only cluster without a single dominant direction, combining up-regulated neuroendocrine-associated genes (*CPLX2*, *MYT1*) with down-regulation of the luminal epithelial marker *KRT7*. These gene-level patterns were tested formally against established biological pathways using over-representation and gene set enrichment analysis, reported below.

Table 14. Provisional biological identity of each VAE $Z = 10$ cluster (hierarchical, $k = 5$), based on cluster-specific driver genes (Supplementary Table S4).

Cluster	*n*	Provisional Identity	Anchor Genes	PAM50 Tie (Table 8)
1	136	Loss of luminal identity	*FOXA1* ↓, *GATA3* ↓, *AR* ↓, *TFF3* ↓, *AGR2* ↓	Basal-like
2	330	Stromal/ECM-relative signature	*LAMA2* ↑, *COL14A1* ↑, *ECM2* ↑, *SPARCL1* ↑	Luminal A (85.2% pure)
3	138	Reduced luminal-lineage markers	*CBLC* ↓, *LINC00472* ↓	HER2-enriched/Luminal B bridge
4	104	Neuroendocrine-leaning, loss of luminal cytokeratin	*CPLX2* ↑, *MYT1* ↑, *KRT7* ↓	Not yet resolved—pending
5	127	Reduced stromal signal; higher cellularity	*PDGFRA* ↓, *DDR2* ↓, *CAV2* ↓	Luminal A (Luminal-B-adjacent)

Formal over-representation and gene set enrichment analysis refined these provisional identities. Table 15 and Figure 6 summarize the refined biological programs across all five clusters. GSEA against the Hallmark collection provided a directionally resolved picture that over-representation testing alone could not: Cluster 1 combined strong proliferation (`E2F_TARGETS`, `G2M_CHECKPOINT`, `MITOTIC_SPINDLE` all significantly up-regulated) with loss of estrogen and androgen signaling (`ESTROGEN_RESPONSE_EARLY/LATE`, `ANDROGEN_RESPONSE` all down-regulated) and activated immune signaling (`ALLOGRAFT_REJECTION`, `IL6_JAK_STAT3_SIGNALING`, `TNFA_SIGNALING_VIA_NFKB`, `INTERFERON_GAMMA_RESPONSE` all up-regulated), consistent with the established immunogenicity of Basal-like disease. Cluster 2 showed the inverse proliferative profile (`MITOTIC_SPINDLE`, `MYC_TARGETS_V1/V2` all significantly down-regulated) alongside up-regulated EMT and stromal programs (`EPITHELIAL_MESENCHYMAL _TRANSITION`, `ADIPOGENESIS`, `MYOGENESIS`) and modest estro-

gen signaling, consistent with a low-proliferation Luminal A phenotype. Cluster 3 showed the strongest combined signature of the five clusters, with very high proliferation (`G2M_CHECKPOINT`, `E2F_TARGETS`, `MITOTIC_SPINDLE`) and broad immune activation (`ALLOGRAFT_REJECTION`, `INTERFERON _GAMMA_RESPONSE`, `INTERFERON_ALPHA_RESPONSE`, `COMPLEMENT`), a signature that over-representation analysis alone failed to detect (zero significant GO or KEGG terms), underscoring the value of the rank-based GSEA approach. Cluster 4, previously unresolved by driver-gene inspection alone, resolved into a high-proliferation, strongly estrogen-signaling cluster (`ESTROGEN_RESPONSE_EARLY` and `LATE` both up-regulated, alongside `E2F_TARGETS` and `G2M_CHECKPOINT`) that was comprehensively immune-cold (`ALLOGRAFT_REJECTION`, `INTERFERON _GAMMA_RESPONSE`, `INFLAMMATORY_RESPONSE`, `COMPLEMENT` all down-regulated), representing the high-proliferation counterpart to Cluster 2's low-proliferation luminal profile. Cluster 5's initial over-representation result (adaptive immune response, $p = 2.9 \times 10^{-47}$) proved, on directional inspection via GSEA, to reflect significantly reduced rather than elevated immune signaling (`ALLOGRAFT_REJECTION`, `INTERFERON_GAMMA_RESPONSE`, `INFLAMMATORY_RESPONSE` all down-regulated), together with reduced EMT and no significant proliferation signal in either direction, identifying it as a comparatively immune-quiescent cluster.

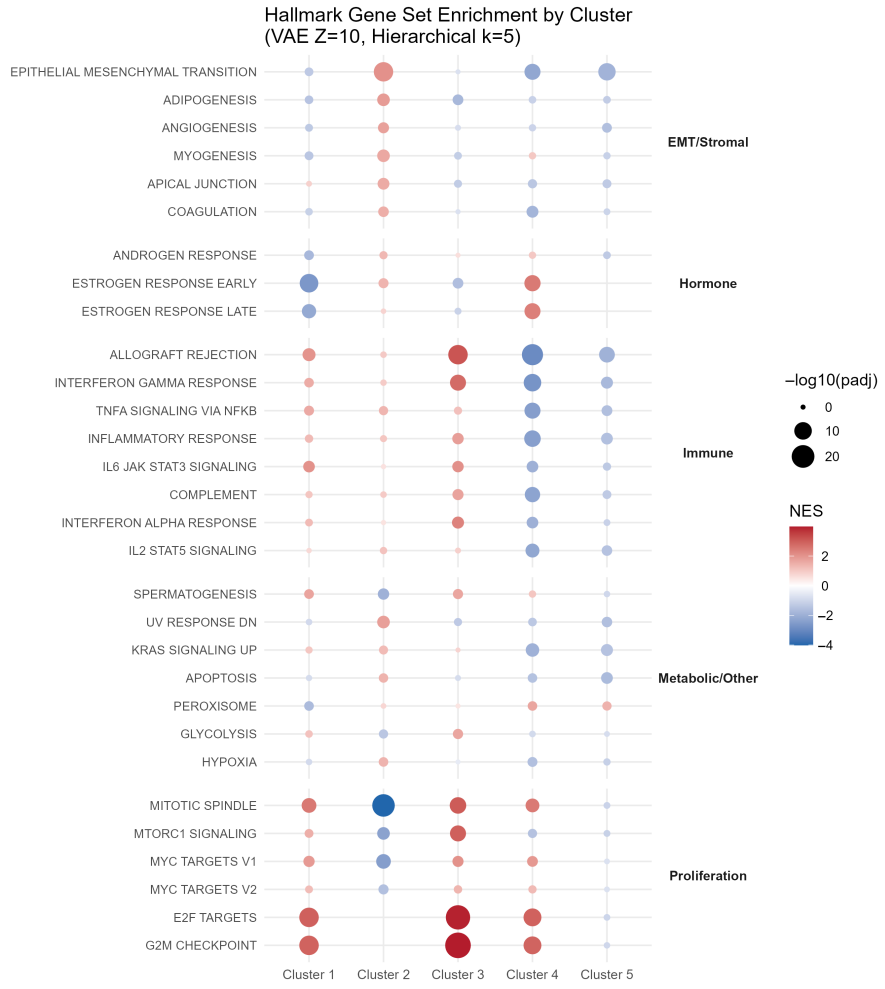


Figure 6. Hallmark gene set enrichment (GSEA) across the five VAE $Z = 10$ malignant-subset clusters (hierarchical, $k = 5$), grouped by biological theme. Dot color indicates normalized enrichment score (NES; red = up-regulated in that cluster relative to the rest, blue = down-regulated); dot size indicates statistical significance ($-\log_{10}$ adjusted p-value). Only Hallmark gene sets reaching significance (adjusted $p < 0.05$) in at least one cluster are shown.

DNA repair pathways (`HALLMARK_DNA_REPAIR`) showed no significant enrichment in either direction in any of the five clusters, indicating that DNA repair capacity does not distinguish these transcriptomic subgroups at the malignant-subset resolution examined here.

These results directly address whether Basal-like-enriched clusters show immune activation and whether luminal-associated clusters separate by proliferative state. Both patterns held: Cluster 1 (Basal-like) and Cluster 3 (HER2-enriched/Luminal B bridge) were the two most immune-active clusters, consistent with the established immunogenicity of non-luminal, high-grade breast cancer; and the two luminal-associated clusters separated cleanly by proliferative state, with Cluster 2 low-proliferation and Cluster 4 high-proliferation despite both retaining active estrogen signaling.

Table 15. Refined biological programs per VAE $Z = 10$ cluster (hierarchical, $k = 5$), based on GO/KEGG over-representation and Hallmark GSEA.

Cluster	*n*	Proliferation	Hormone/ER Signaling	Immune Signaling	EMT/Stromal	Refined Identity
1	136	High	Lost	Activated	—	Basal-like: proliferative, immune-active, hormone-null
2	330	Low	Modest, present	—	Up	Low-proliferation Luminal A
3	138	Very high	Lost	Strongly activated	—	HER2-enriched/Luminal B: most proliferative and immunogenic
4	104	High	Strong (early + late)	Cold	Reduced	High-proliferation, ER-active, immune-cold luminal-like
5	127	Not significant	—	Cold	Reduced	Immune-quiescent, epithelial

3.7. Immune Landscape of Discovered Subtypes

ESTIMATE and quanTIseq revealed a clear division in absolute immune infiltration among the five VAE $Z = 10$ clusters, orthogonal to the differential expression-based enrichment analysis of Section 3.6. Immune score, ESTIMATE score, stromal score, and tumor purity all differed significantly across clusters (Kruskal–Wallis, all $p < 5 \times 10^{-21}$; Table 16), separating Clusters 1, 2, and 3 as an immune/stromal-rich group from Clusters 4 and 5, which showed markedly lower immune infiltration and the highest tumor purity of the five clusters (medians 0.794 and 0.787, versus 0.746–0.751 in Clusters 1–3).

Table 16. Immune and stromal composition per VAE $Z = 10$ cluster (hierarchical, $k = 5$), based on ESTIMATE and quanTIseq (median values; $n = 835$).

Cluster	*n*	PAM50 Identity (Table 8)	Immune Score	Stromal Score	Tumor Purity	CD8+ T Cells	Tregs	NK Cells	Immune Character
1	136	Basal-like (89.0%)	449.3	286.8	0.751	0.0006	0.026	0	Hot; B-cell/leukocyte-skewed, CD8-poor
2	330	Luminal A (85.2%)	286.1	521.9	0.746	0.0262	0.096	0	Hot; stromal-dominant
3	138	HER2-enriched/Luminal B bridge	382.8	458.9	0.746	0.0051	0.121	0	Hot; regulatory/myeloid-skewed
4	104	Luminal A/B mixed	17.1	316.1	0.794	0.0232	0.082	0.060	Cold; residual cytotoxic/innate signal
5	127	Luminal A/B mixed	106.1	288.2	0.787	0.0308	0.101	0.063	Cold; highest residual CD8+/NK of all clusters

Full Kruskal–Wallis and Dunn's post hoc pairwise results for all fifteen variables tested—including the four variables (B cells, CD8+ T cells, M2 macrophages, myeloid dendritic cells) and one constant variable (monocyte) that did not survive Bonferroni-corrected pairwise comparison despite nominal omnibus significance in two cases—are reported in Supplementary Table S5.

NK-cell infiltration ($H = 130.8$, $p = 2.7 \times 10^{-27}$) and neutrophil fraction ($H = 34.9$, $p = 4.9 \times 10^{-7}$) both differed in the opposite direction, with Clusters 4 and 5 showing the higher values, indicating that low-infiltration clusters are not uniformly depleted of every immune population.

The three immune-rich clusters diverged sharply in composition once mapped onto their PAM50 identities (Tables 8 and 14). Cluster 1, the 89.0%-pure Basal-like cluster already identified by GSEA as immune-active (Section 3.6), showed the highest median immune score of all five clusters (449.3)—confirming that activation by an independent, composition-based method—but its infiltrate was essentially devoid of CD8+ T cells (median 0.0006, the lowest of any cluster) and had the lowest regulatory T-cell fraction of the cohort (median 0.026), with the highest B-cell fraction (median 0.291) of the five clusters. This indicates that Cluster 1's immune activation, consistent with the Hallmark interferon and inflammatory signaling identified in Section 3.6, is not accompanied by a corresponding cytotoxic T-cell compartment, distinguishing a transcriptionally inflamed but T-cell-poor Basal-like phenotype from a canonical CD8-infiltrated "hot" tumor. Cluster 3, the HER2-enriched/Luminal B bridging cluster identified by GSEA as the most proliferative and immunogenic of the five (Section 3.6), had the second-highest immune score (median 382.8) alongside the highest regulatory T-cell fraction of any cluster (median 0.121) and the lowest CD8+ T-cell fraction among the three immune-rich clusters (median 0.0051), together with elevated M1 macrophages (median 0.186, the highest of any cluster). This composition points to a regulatory- and myeloid-skewed immune infiltrate rather than a cytotoxic one, suggesting that Cluster 3's strong immune transcriptional signature may reflect an immunosuppressive rather than antitumor-active microenvironment. Cluster 2, the 85.2%-pure Luminal A cluster that showed no significant Hallmark immune signal in either direction in Section 3.6, nonetheless had the highest stromal score of all five clusters (median 521.9) and a moderate immune score (median 286.1)—indicating that Cluster 2 is not immune-cold, but that its microenvironment is stromal-dominant rather than lymphocyte-driven, consistent with its up-regulated extracellular-matrix and stromal driver genes (Section 3.6). This is a notable finding given that Luminal A disease is conventionally regarded as the least immunologically active PAM50 subtype.

Clusters 4 and 5—both predominantly Luminal A/B-mixed populations (Table 8) and independently flagged as immune-cold or immune-quiescent by GSEA (Section 3.6)—were confirmed as the two lowest-infiltration clusters by ESTIMATE, cross-validating the pathway-level finding with an entirely independent, composition-based method. Neither cluster was uniformly immune-depleted, however: Cluster 4 retained a measurable CD8+ T-cell fraction (median 0.023) and NK-cell fraction (median 0.060) despite its immune score being the lowest of the cohort by a wide margin (median 17.1), while Cluster 5 showed the highest CD8+ T-cell fraction (median 0.031) and highest NK-cell fraction (median 0.063) of all five clusters despite its own low overall infiltration (immune score median 106.1). This suggests that the residual immune content in these otherwise immune-cold, high-tumor-purity clusters is preferentially cytotoxic/innate in character rather than regulatory or humoral, in contrast to the B cell- and Treg-dominated profiles of Clusters 1 and 3.

These results address both interpretive questions posed for this analysis. Basal-like disease (Cluster 1) and the HER2-enriched/Luminal B bridge (Cluster 3) are indeed the two most immune-infiltrated clusters, consistent with the established immunogenicity of

non-luminal, high-grade breast cancer, but each achieves this through a distinct compositional route—B-cell/leukocyte-dominant and CD8-poor in Cluster 1, versus regulatory T-cell- and M1-macrophage-dominant in Cluster 3—rather than a shared cytotoxic profile. More notably, high immune infiltration was not confined to a single PAM50 background: Cluster 2 (85.2% Luminal A) is immune/stromal-rich, while Clusters 4 and 5—both also predominantly Luminal A/B-mixed populations—are immune-cold, demonstrating that immune-hot and immune-cold subclusters can co-exist within nominally the same PAM50 subtype. This heterogeneity, invisible to PAM50 classification alone, is summarized in Table 16 and Figures 7 and 8.



Figure 7. Distributions of ESTIMATE scores, tumor purity, and quanTIseq immune cell-type fractions by cluster (box plots), VAE $Z = 10$, hierarchical $k = 5$, malignant subset ($n = 835$). Boxes show the interquartile range, the horizontal line within each box marks the median, whiskers extend to the most extreme values within 1.5 times the interquartile range, and dots indicate outliers. Colors distinguish the five clusters.

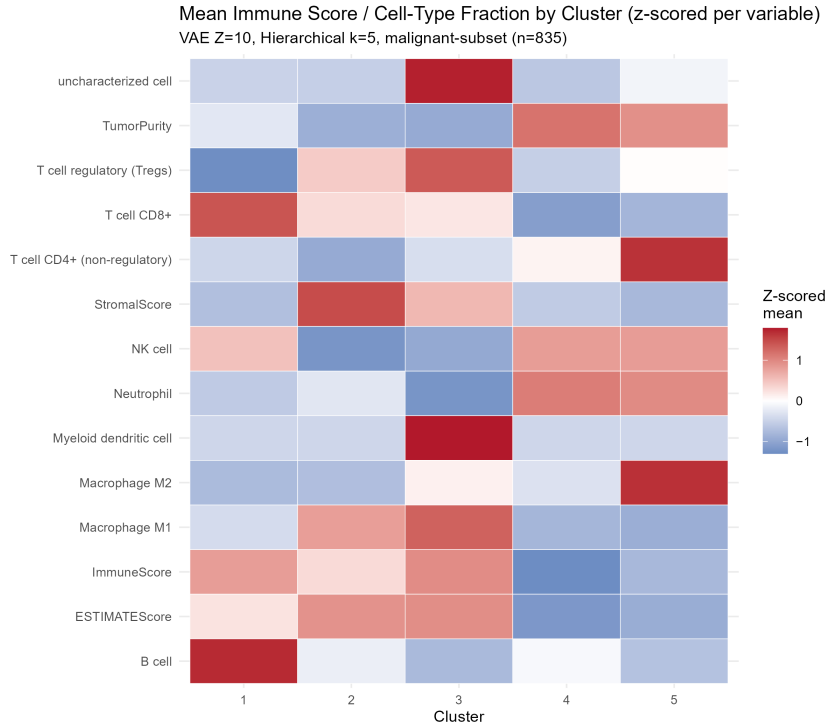


Figure 8. Mean immune score and cell-type fraction per cluster, z-scored per variable across clusters, VAE $Z = 10$, hierarchical $k = 5$, malignant subset ($n = 835$). Because means are sensitive to a small number of high-infiltration outlier samples, some rows (e.g., $CD8^+$ T cells and NK cells in Cluster 1) differ from the median-based summaries in Table 16 and Section 3.7; medians are the primary summary statistic (Section 2.8). Monocyte is not shown because its value was zero in all samples.

3.8. Robustness and Validation

Cluster stability of the malignant-subset flagship solution (VAE $Z = 10$, hierarchical, $k = 5$) was assessed via bootstrap subsampling (Table 17). The Basal-like-dominant Cluster 1 (89.0% purity, Table 8) and Luminal A-dominant Cluster 2 (85.2% purity) both achieved moderate-to-high stability under Hennig's (2007) [42] thresholds (mean Jaccard = 0.731 and 0.631, respectively), with Cluster 1 rarely dissolving under resampling (dissolved in 2 of 50 resamples, cleanly recovered in 29). By contrast, Clusters 3, 4, and 5—the HER2-enriched/Luminal B bridge and the two Luminal A/B-mixed populations (Table 8)—all fell below the conventional 0.60 stability threshold (mean Jaccard = 0.456, 0.456, and 0.383, respectively), with Cluster 5 dissolving in 43 of 50 resamples. This pattern closely tracks PAM50 purity: the two subtype-pure clusters are also the two most resampling-stable, while instability concentrates specifically in the clusters occupying the biologically continuous HER2/luminal boundary rather than being distributed uniformly across the five-cluster solution, consistent with genuine biological continuity at this subtype interface rather than a clustering artifact.

Table 17. Bootstrap cluster stability (50 resamples, 85% subsampling) for the VAE $Z = 10$, hierarchical $k = 5$ malignant-subset solution ($n = 835$).

Cluster	*n*	Mean Jaccard	Times Dissolved	Times Recovered
1	136	0.731	2	29
2	330	0.631	9	6
3	138	0.456	29	1
4	104	0.456	29	0
5	127	0.383	43	1

Sensitivity to latent space dimensionality was assessed previously rather than repeated here (Section 3.4.1, Table 7; Section 3.5.2, Table 13): VAE $Z = 10$ was the unambiguous mathematical optimum among the four malignant-subset representations, outperforming PCA-40 and VAE $Z = 20/40$ on all three internal validation metrics simultaneously (Table 7), and was one of only two representations—alongside VAE $Z = 40$—to produce clusters with both a significant log-rank survival separation and a cluster-only C-index exceeding the PAM50-only baseline (Table 13). Together, these results indicate that the choice of VAE $Z = 10$ as the flagship representation reflects a systematic, multi-criteria optimum rather than an arbitrary or post hoc selection.

Sensitivity of the malignant-subset clustering solution to the size of the highly variable gene input panel was assessed by reselecting the two-criterion gene panel (low-expression filter, MAD-based variance ranking; Section 2.2) at 3000 and 10,000 genes, retraining the VAE at $Z = 10$ on each panel using the identical architecture and training protocol as the reported 5000-gene model, and reclustering with hierarchical (Ward) $k = 5$ (Table 18).

Table 18. PAM50 concordance (Adjusted Rand Index, Normalized Mutual Information) for the VAE $Z = 10$, hierarchical $k = 5$ malignant-subset solution across three gene-panel sizes ($n = 835$).

Panel Size	*n*	ARI	NMI
3000 genes	835	0.315	0.401
5000 genes (reported)	835	0.377	0.396
10,000 genes	835	0.262	0.314

Overall PAM50 concordance was comparable between the 3000-gene and 5000-gene panels (ARI = 0.315 vs. 0.377; NMI = 0.401 vs. 0.396, marginally higher at 3000 genes) but declined at 10,000 genes (ARI = 0.262; NMI = 0.314), indicating that expanding the gene panel beyond the reported 5000-gene threshold introduces additional, less informative variance that dilutes rather than sharpens subtype-relevant structure—consistent with the latent-dimensionality result in Section 3.4.1 (Table 7), where expanding VAE capacity beyond $Z = 10$ similarly degraded internal cluster validation metrics on this same malignant-subset population.

The central biological conclusions held at every panel size tested. Basal-like disease separated into a distinct, high-purity cluster regardless of gene-panel size (89.0% at 5000 genes; 94.2% at 3000 genes; 95.2% at 10,000 genes), confirming that this subtype's separability reflects underlying tumor biology rather than a specific feature-selection threshold. Luminal A also consistently resolved into a cluster-level majority at every panel size, though its purity and cluster size varied (85.2% at 5000 genes [$n = 330$]; 86.7% at 3000 genes [$n = 278$]; a more diffuse 66.9% at 10,000 genes in a considerably larger, less-resolved cluster [$n = 423$])—this tendency of the 10,000-gene panel to merge rather than separate luminal samples is the primary driver of its lower overall concordance. HER2-enriched disease, consistent with the results reported throughout this study, never achieved cluster-majority status at any panel size, instead forming a mixed HER2/Luminal B cluster in all three cases (Supplementary Table S6)—the same cluster identified as the least stable under bootstrap resampling in the preceding stability analysis (Cluster 3, Table 17), providing convergent evidence from two independent robustness analyses that the HER2/Luminal B boundary in this cohort reflects genuine, continuous biological overlap rather than an artifact of either resampling or feature selection.

Full per-cluster PAM50 contingency tables for the 3000- and 10,000-gene panels are reported in Supplementary Table S6.

Stability of the primary analytical cohort's clustering solution (Section 3.3) was assessed by an equivalent bootstrap procedure applied at the whole-cohort level: fifty resam-

ples, each retaining a random 85% ($n = 7432$) of the 8744-sample cohort, were independently reclustered under the identical hierarchical (Ward), $k = 7$ protocol, and ARI/NMI were computed against PAM50 subtype identity, restricted to the 839 samples with a true subtype call. Reference ARI/NMI on the full, non-resampled cohort were near zero for every representation (ARI 0.008–0.013, NMI 0.024–0.038; Supplementary Table S7), directly corroborating Section 3.3's finding on a single-run basis.

The bootstrap procedure confirms this null result is reproducible on average for every representation (mean ARI 0.006–0.027, far below the malignant-subset's moderate concordance, Table 11), but this reproducibility was not uniform. The raw 5000-gene space, PCA-40, and VAE $Z = 10$ produced a uniformly near-zero result with no exceptions across all 50 resamples (sd 0.0035–0.0044; maximum ARI $\leq$ 0.016 in every case). VAE $Z = 20$ and VAE $Z = 40$, by contrast, each produced a small number of resamples that departed sharply from this pattern: 3 of 50 for VAE $Z = 20$ and 1 of 50 for VAE $Z = 40$ yielded ARI of 0.308–0.331 (NMI 0.432–0.472)—roughly thirty times the representation's typical value—driving up both the mean and, especially, the standard deviation for these two representations. Because these excursions occur in only a small minority of resamples and do not reproduce in the reference full-sample clustering, they most plausibly reflect an occasional alternative dendrogram-level partition available to Ward's linkage in these wider, less-compressed latent spaces under specific bootstrap draws, rather than a consistently recoverable subtype signal masked in the full cohort. This does not overturn Section 3.3's conclusion that no representation reliably resolves PAM50 subtype structure at the whole-cohort level; it adds a further, whole-cohort-level line of evidence that VAE $Z = 10$ produces a more stable, reproducible clustering solution under resampling than the wider VAE bottlenecks—extending to the primary cohort the same $Z = 10$ stability advantage already established for the malignant subset (Sections 3.1 and 3.4.1). Full per-resample bootstrap values are reported in Supplementary Table S7.

4. Discussion

The two-tier structure of this analysis reveals that the largest single axis of transcriptomic variation in this cohort is not PAM50 molecular subtype but tissue of origin and malignant-versus-normal status. This signal proved remarkably robust: it dominated every representation tested (PCA-40 and all three VAE bottlenecks, Table 4) and every clustering algorithm (hierarchical, k-means, and GMM, Section 3.3), and it persisted even when dimensionality reduction was removed entirely and clustering was applied directly to the raw 5000-gene matrix (Section 3.4)—consistent with the scale of transcriptomic separation originally documented between breast tumor and normal tissue and among tissue types more broadly [2,5]. Notably, where the malignant cluster's boundary blurred at all, the admixture came almost exclusively from Normal-TCGA rather than Normal-GTEx samples (Section 3.3)—solid-tissue-normal specimens drawn from cancer patients rather than healthy donors and therefore the tissue type transcriptionally closest to tumor, plausibly reflecting subtle field-level molecular alterations in histologically normal tissue adjacent to a malignancy [43]. This supports this study's central pivot—isolating a malignant-only subset before searching for subtype structure—as having targeted a genuine biological confound rather than a modeling artifact.

Once the tissue-of-origin axis was removed by restricting analysis to the 835-sample malignant subset, the representation ranking inverted. PCA-40 led on every internal validation metric for the whole cohort (Table 4), but VAE $Z = 10$ led on every internal validation metric for the malignant subset (Table 7). This inversion is consistent with the broader finding that nonlinear representation learning offers its clearest advantage once the coarsest axes of variation have been removed and the remaining signal is subtler and more entan-

gled [17,20]—a large, tissue-diverse cohort dominated by one comparatively coarse signal is well served by a simple linear projection, whereas a smaller, more homogeneous population benefits from a more constrained, nonlinear latent space regularized by a Kullback–Leibler penalty [15,42]. The malignant-subset training dynamics reproduced this same reversal independently: $Z = 20$ minimized validation loss for the whole cohort (Table 2), but $Z = 10$ minimized it for the malignant subset (Table 3), with loss rising monotonically as latent capacity increased—the expected behavior of a variational bottleneck once its capacity begins to exceed what the underlying signal requires [15,42].

VAE $Z = 10$'s status as the flagship malignant-subset representation rests on convergence across four independent criteria rather than a single favorable metric: it minimized training loss (Section 3.1), performed best on all three internal validation metrics simultaneously (Table 7), achieved the highest ARI against PAM50 of any representation (Table 11), and was one of only two representations—alongside $Z = 40$—to produce clusters with both significant log-rank survival separation and a C-index exceeding PAM50 (Table 13), consistent with recent benchmarking work showing that learned latent representations can modestly outperform simpler baselines on survival endpoints derived from bulk RNA-seq [20]. At the same time, PCA-40 achieved the highest NMI despite a lower ARI than VAE $Z = 10$ (Table 11)—a divergence consistent with the documented tendency of geometric and information-theoretic cluster-validity metrics to decouple from one another, and from downstream biological utility, in high-dimensional genomic embedding spaces [44]. This argues for reporting multiple complementary validity metrics jointly, as done throughout this study, rather than reducing cluster validity to a single number.

The malignant-subset clustering results are best read as a precise map of which PAM50 subtypes are, and are not, separable in unsupervised bulk transcriptomic space, rather than a simple pass/fail verdict on clustering. Basal-like disease was the one subtype that separated cleanly under every condition tested in this study—every representation, every clustering algorithm, and every gene-panel size (83–98% purity across Tables 6, 8, and 10; 89–95% across the three gene-panel sizes in Table 18)—corroborating the well-established transcriptomic distinctiveness of Basal-like/triple-negative disease relative to hormone-receptor-driven subtypes [6]. Luminal A showed a consistent bipartite structure across all four representations (Section 3.5.1): a substantial, comparatively pure cluster together with one or more clusters in which Luminal A remained the plurality but Luminal B made up a large minority. Because this same bipartite pattern recurred independently under PCA, VAE $Z = 10$, and VAE $Z = 40$, it is unlikely to be a representation-specific artifact, and instead aligns with prior clinical observations that patients sharing a single PAM50 label can nonetheless follow divergent molecular and clinical trajectories [8,9], and with the more general view of intrinsic-subtype boundaries as continuous rather than strictly discrete [9,10]. HER2-enriched disease told a more fragile story, concentrating into a single cluster under every representation but only becoming that cluster's outright majority under one specific combination (VAE $Z = 20$ with hierarchical clustering, 72.9% purity, Table 8) that did not reproduce under k-means or GMM (Section 3.4.2). Luminal B, meanwhile, never achieved outright majority status in a single cluster anywhere in this study—not under any of four representations, three clustering algorithms, six values of k, or three gene-panel sizes (Tables 6, 8, 10 and 18)—a consistent negative result in line with known difficulties in resolving closely related molecular states through unsupervised clustering even when real biological differences exist [11]. Taken together, these findings indicate that, within unsupervised bulk transcriptomic space, Luminal B behaves not as an independently separable entity but as a shared boundary population bridging Luminal A and HER2-enriched, reinforced independently by the robustness analyses discussed below.

Differential expression and pathway analysis (Section 3.6) gave the statistically defined VAE $Z = 10$ clusters coherent, directionally resolved biological identities (Tables 14 and 15) using established differential expression and enrichment frameworks [30,31,36]. GSEA was decisive where over-representation testing was not: Cluster 3, the HER2-enriched/Luminal B bridge, showed the strongest combined proliferative and immune signature of the five clusters under GSEA despite returning zero significant GO or KEGG terms, underscoring the value of rank-based enrichment testing when individual gene-level effect sizes are modest. A biologically notable secondary axis emerged within luminal-associated disease: Cluster 2 and Cluster 4 both retained active estrogen signaling yet separated cleanly by proliferative state (Table 15), consistent with proliferation being an established, partially independent determinant of prognosis and therapeutic behavior within hormone-receptor-positive disease [4,7]. Hallmark DNA repair pathways showed no significant enrichment in either direction in any cluster, indicating that DNA repair capacity is not a distinguishing axis among these five transcriptomic subgroups at the resolution examined here.

The immune microenvironment findings (Section 3.7) demonstrate that immune-hot and immune-cold status cuts across, rather than aligns with, PAM50 subtype boundaries [25,26]. Most notably, the two Luminal A-associated clusters diverged sharply: Cluster 2 was immune/stromal-rich while Clusters 4 and 5 were the cohort's two lowest-infiltration, highest-purity clusters—directly challenging the conventional assumption that Luminal A disease is uniformly immune-quiescent, and closely paralleling a recent multi-center finding that immune "crossover" profiles, rather than PAM50 label alone, stratify recurrence risk within nominally low-risk luminal breast cancer [45]. The composition of infiltration, not merely its magnitude, also diverged in clinically relevant ways. The Basal-like cluster carried the highest immune score in the cohort yet was nearly devoid of CD8+ T cells and was B-cell/leukocyte-skewed instead, while the HER2-enriched/Luminal B bridge cluster (Cluster 3) carried the second-highest immune score alongside the highest regulatory T-cell fraction and elevated M1 macrophages—a regulatory- and myeloid-skewed composition consistent with evidence that CD25-high effector regulatory T cells, specifically, rather than immune infiltration in aggregate, are associated with resistance to PD-1 blockade in breast cancer [46]. Because CD8+ infiltration, rather than aggregate immune score, is more directly linked to checkpoint-inhibitor response [25,27], these compositional distinctions are of direct translational interest, but remain hypothesis-generating: bulk RNA-seq deconvolution, as already noted in Section 3.7, cannot resolve regulatory T-cell activation state (e.g., CD25/PD-1 co-expression) or establish a direct mechanistic basis for immunosuppression; such mechanistic interference would require the kind of paired human-tissue and functional validation used in [46] before any therapeutic inference is drawn.

The survival results warrant a similarly measured interpretation. VAE $Z = 10$ and $Z = 40$ produced clusters with significant log-rank separation and a C-index modestly exceeding the PAM50-only baseline (0.619 and 0.643 versus 0.606, Table 13) [33,34], consistent with reports that learned representations can offer incremental prognostic value over baseline molecular classifiers on survival endpoints [20]. Neither representation, however, remained significant on the age- and stage-adjusted independent-marker test [35] (Table 13), and both cluster assignments were already strongly associated with ER, PR, HER2 status, and AJCC stage (Table 12). Taken together, these results indicate that the discovered clusters substantially recapitulate—and modestly sharpen—established clinicopathological risk factors [3,4,7], rather than constituting a wholly independent new prognostic axis at the sample sizes available here. The appropriate framing is therefore refinement of existing subtype-based risk stratification, not its replacement.

The robustness analyses (Section 3.8) converge on the same conclusion through three structurally different types of perturbation. Bootstrap resampling of the malignant-subset

solution (Table 17), following the cluster-stability framework of Hennig [41], showed that stability tracks PAM50 purity almost exactly: the Basal-like- and Luminal A-dominant clusters were moderately to highly stable, while the HER2-enriched/Luminal B bridge and the two Luminal A/B-mixed clusters all fell below the conventional stability threshold. Reconstructing the gene panel at 3000 and 10,000 genes and repeating the full pipeline (Table 18) reproduced this same pattern from an entirely different angle: Basal-like separated cleanly at every panel size, while HER2-enriched consistently re-formed the same HER2-enriched/Luminal B bridge identified as the least stable cluster under resampling. Two independent perturbations therefore point to the same conclusion: the HER2-enriched/Luminal B boundary reflects genuine, continuous biological overlap in this cohort [11], not an artifact of sample selection, feature-selection threshold, or any single modeling choice. The primary-cohort bootstrap stability check (Supplementary Table S7) provides exactly this third line of evidence, with an instructive nuance. Across all fifty resamples, mean ARI against PAM50 subtype remained far below even the malignant subset's moderate concordance (0.006–0.027 versus 0.24–0.38, Table 11) for every representation, confirming that the whole-cohort failure to resolve subtype structure at $k = 7$ is not an artifact of the single run reported in Section 3.3. This stability was itself representation-dependent, however: the raw gene space, PCA-40, and VAE $Z = 10$ produced a uniformly near-zero result with no exceptions across 50 resamples, while VAE $Z = 20$ and $Z = 40$ each produced a small number of resamples (3/50 and 1/50, respectively) in which ARI jumped to 0.31–0.33—an occasional non-reproducing departure from the otherwise consistent null result. Rather than complicating the pivot rationale, this adds a further, independent line of evidence for VAE $Z = 10$'s relative advantage: its whole-cohort clustering solution is not only null with respect to subtype recovery, as expected, but the most stably null of the three fully consistent representations—mirroring the same $Z = 10$ stability advantage already established within the malignant subset (Sections 3.1 and 3.4.1).

Beyond these specific biological and clinical findings, this study's integrated, multi-algorithm design proved substantively necessary rather than redundant. Cross-algorithm validation was what revealed that VAE $Z = 20$'s HER2-enriched cluster majority is real but fragile—a conclusion invisible from hierarchical clustering alone (Section 3.4.2). Pairing rank-based GSEA with conventional over-representation testing was what recovered Cluster 3's strongest combined proliferative and immune signature of the five clusters, a signal that over-representation analysis alone missed entirely (Section 3.6). And pairing two independent immune deconvolution methods—pathway-level GSEA and composition-based ESTIMATE/quanTIseq—was what let the immune-cold designation of Clusters 4 and 5 be confirmed by an entirely separate analytical route rather than resting on a single method's assumptions (Section 3.7). In each case, a single-method pipeline would have produced a less complete or actively misleading picture; the comprehensiveness of the framework is therefore itself a finding, not merely a methodological convenience.

Several limitations qualify these findings. No independent external cohort was available that met this study's inclusion criteria, so every result reported here is internal to this dataset; external validation, ideally in a cohort such as METABRIC, is necessary before translation to practice, in line with recommended validation pathways for transcriptomic classifiers and emerging AI-based diagnostic tools in breast cancer [3,47]. Immune characterization relies entirely on bulk transcriptomic deconvolution [23,24]; as noted in Sections 2.8 and 3.7, this cannot establish cellular co-localization or confirm an immunosuppressive mechanism, and future work should pursue single-cell or spatial resolution of the kind that has recently clarified analogous immune niches in breast carcinoma [10,46]. Similarly, the differential expression and functional enrichment results in Section 3.6 are not statistically independent of the expression data used to define the clusters themselves, since

both clustering and cluster characterization operate on the same underlying transcriptomic measurements; as with any post hoc characterization of unsupervised structure, these results should be interpreted as descriptive of the discovered clusters rather than as an independent statistical confirmation of their biological reality. Survival analyses were retrospective, and the cluster-membership signal did not survive adjustment for age and stage at the available sample size—a negative result reported transparently rather than omitted. The primary cohort is heavily dominated by Normal-GTEx samples, and platform, tissue procurement, and batch differences between TCGA and GTEx may contribute to part of the tissue-of-origin separation reported in Sections 3.2 and 3.3, independent of any genuine biological signal—a well-documented risk whenever samples from different processing pipelines or repositories are analyzed jointly [12,13].

5. Conclusions

This study presents a two-stage representation learning framework that combines a variational autoencoder (VAE), nonlinear visualization (t-SNE and UMAP), multi-algorithm unsupervised clustering (hierarchical, k-means, and Gaussian mixture models), and multi-layer functional, clinical, and immune characterization, applied first to a primary analytical cohort of 8744 real, non-duplicated PAM50-labeled and normal-tissue samples and subsequently to an isolated malignant subpopulation extracted from it.

Across the primary cohort, tissue of origin and malignant-versus-normal status—not PAM50 molecular subtype—consistently dominated clustering structure across every representation and every clustering algorithm tested, including the raw gene expression space with no dimensionality reduction. Isolating the resulting 835-sample malignant subset and reclustering it in isolation revealed genuine subtype-specific structure that had been obscured at the whole-cohort level. Among the malignant-subset representations, VAE $Z = 10$ achieved the strongest internal validation metrics, the highest concordance with PAM50 (Adjusted Rand Index = 0.377), and—together with VAE $Z = 40$—a concordance index modestly exceeding the PAM50-only baseline (0.619 and 0.643 versus 0.606). Basal-like disease separated cleanly under every representation, clustering algorithm, and gene-panel size tested; Luminal A resolved into a distinct cluster and a continuum bridging Luminal B; HER2-enriched samples concentrated into a single cluster but only inconsistently achieved cluster-majority status; and Luminal B never achieved an outright cluster majority under any configuration tested, behaving as a shared boundary population rather than an independently separable subtype. Immune deconvolution further showed that immune-hot and immune-cold status cuts across, rather than aligns with, PAM50 subtype boundaries, including a regulatory T-cell-skewed and M1-macrophage-skewed HER2-enriched/Luminal B bridging cluster and a CD8-poor, B-cell-skewed Basal-like cluster. Bootstrap resampling and gene-panel-size sensitivity analyses confirmed that cluster stability tracked PAM50 purity, with Basal-like separation robust across every condition tested and the HER2-enriched/Luminal B boundary consistently the least stable.

These findings support the use of nonlinear representation learning as a research tool for refining, rather than replacing, PAM50-based molecular subtyping, revealing biologically coherent structure—particularly at the HER2-enriched/Luminal B boundary—that a discrete categorical framework does not capture. However, the cross-sectional, single-cohort, bulk-transcriptomic design of this study cannot establish external generalizability, cellular co-localization or mechanism within the immune microenvironment, or clinical actionability, and the discovered clusters did not constitute an independent prognostic marker beyond established clinicopathological staging at the sample sizes available here. External cohorts, patient-level validation, and spatial or single-cell assays are needed before this framework can inform risk stratification or treatment selection.

Future work will focus on several distinct directions:

1. **Cross-Platform Generalizability:** Evaluation of the pipeline, including the malignant-subset isolation step, on additional independent breast cancer transcriptomic cohorts, including METABRIC and GEO-hosted datasets, to assess whether the reported subtype-separability pattern and the HER2-enriched/Luminal B boundary reproduce outside this cohort.
2. **Multi-Omics Integration:** Incorporating complementary data streams, including DNA methylation, copy number variation (CNV), and proteomic profiles, within the VAE architecture to capture multilayered molecular heterogeneity beyond the transcriptome, particularly at the unresolved HER2-enriched/Luminal B boundary.
3. **Prospective Clinical Validation:** Evaluating the VAE $Z = 10$ malignant-subset representation in prospectively specified, independent treatment cohorts with sufficient sample size to test cluster membership as an independent prognostic marker beyond stage and receptor status.
4. **Spatial and Single-Cell Resolution:** Extending the pipeline to single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics to establish whether the regulatory T-cell/M1-macrophage-skewed and B-cell-skewed immune profiles identified here reflect genuine cellular co-localization and an immunosuppressive mechanism, rather than bulk-deconvolution-derived composition alone.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biomedinformatics6050086/s1, Figure S1: t-SNE projections of the primary analytical cohort (perplexity = 40); Figure S2: Perplexity-stability analysis of the t-SNE projections (perplexity = 30, 40 and 50); Table S1: Internal validation metrics across k = 3–8, hierarchical clustering, by representation; Table S2: Malignant/PAM50-associated cluster purity at k = 7, across representations and clustering algorithms; Table S3: Full biological subtype distribution at k = 7, all representations and algorithms; Table S4: Top 10 driver genes per VAE Z = 10 cluster; Table S5: Full Kruskal–Wallis and Dunn's post hoc test results for ESTIMATE scores and quanTIseq immune cell-type fractions; Table S6: Per-cluster PAM50 subtype contingency tables for the 3000- and 10,000-gene panels; Table S7: Bootstrap stability of the primary analytical cohort's clustering solution.

Author Contributions: Conceptualization, A.H.A.K. and L.A.; methodology, A.H.A.K. and L.A.; software, A.H.A.K.; validation, A.H.A.K., L.A., and H.H.; formal analysis, A.H.A.K.; investigation, A.H.A.K.; data curation, A.H.A.K.; writing—original draft preparation, A.H.A.K.; writing—review and editing, A.H.A.K., L.A., and H.H.; supervision, L.A. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The primary transcriptomic data and clinical metadata analyzed in this study are publicly available from the UCSC Xena functional genomics explorer (https://xenabrowser.net/ (accessed on 12 March 2026)). No new sequencing data were generated. The complete analysis pipeline (VAE training, clustering, differential expression, functional enrichment, immune deconvolution, and robustness-testing code), together with the derived outputs (cluster assignments for each representation, the 5000/3000/10,000-gene panels, and enrichment and immune-deconvolution results), are openly and permanently archived at https://doi.org/10.5281/zenodo.22541090.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AJCC	American Joint Committee on Cancer
ARI	Adjusted Rand Index
CH Index	Calinski–Harabasz Index
CIBERSORT	Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts
C-Index	Concordance Index
CNV	Copy Number Variation
DB	Davies–Bouldin Index
DE	Differential Expression
ELBO	Evidence Lower Bound
EMT	Epithelial-to-Mesenchymal Transition
ER	Estrogen Receptor
ESTIMATE	Estimation of STromal and Immune cells in MAlignant Tumors using Expression data
GMM	Gaussian Mixture Model
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GTEx	Genotype-Tissue Expression
HER2	Human Epidermal Growth Factor Receptor 2
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLD	Kullback–Leibler Divergence
LR	Likelihood Ratio
MAD	Median Absolute Deviation
MSE	Mean Squared Error
NES	Normalized Enrichment Score
NMI	Normalized Mutual Information
PAM50	Prediction Analysis of Microarray 50
PCA	Principal Component Analysis
PR	Progesterone Receptor
quanTIseq	Quantification of the Tumor Immune contexture from RNA-seq data
ReLU	Rectified Linear Unit
scRNA-seq	Single-cell RNA sequencing
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TIME	Tumor Immune Microenvironment
TPM	Transcripts Per Million
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Perou, C.M.; Sørlie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. Molecular Portraits of Human Breast Tumours. *Nature* **2000**, *406*, 747–752. [CrossRef] [PubMed]
3. Parker, J.S.; Mullins, M.; Cheang, M.C.U.; Perou, C.M.; Voduc, D.; Bhanchini, G.; Nielsen, T.O.; Bergh, J.; Chang, J.C.; Ellis, M.J.; et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **2009**, *27*, 1160–1167. [CrossRef] [PubMed]
4. Prat, A.; Cheang, M.C.; Spain, B.H.; Parker, J.S.; Perou, C.M. Towards Understanding the Prognostic and Therapeutic Value of Breast Cancer Intrinsic Subtypes. *Endocr.-Relat. Cancer* **2012**, *19*, R131–R145.
5. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* **2012**, *490*, 61–70. [PubMed]

6. Ciriello, G.; Gatza, M.L.; Beck, A.H.; Wilkerson, M.D.; Rhie, S.K.; Pastore, A.; Zhang, J.; Perou, C.M. Comprehensive Molecular Subtyping of Triple-Negative Breast Cancer. *Cell* **2015**, *163*, 506–519. [PubMed]
7. Wallden, B.; Storhoff, J.; Nielsen, T.; Dowidar, N.; Schaper, C.; Ferree, S.; Liu, S.; Leung, S.; Geiss, G.; Snider, J.; et al. Development and Verification of the PAM50-Based Prosigna Breast Cancer Prognostic Gene Signature Assay. *BMC Cancer* **2015**, *15*, 860. [CrossRef] [PubMed]
8. Prat, A.; Pineda, E.; Adamo, B.; Galván, P.; Fernández, A.; Gaba, L.; Perou, C.M. Clinical Implications of the Intrinsic Molecular Subtypes of Breast Cancer. *Breast* **2015**, *24*, S26–S35. [CrossRef] [PubMed]
9. Cejalvo, J.M.; Martínez de Dueñas, E.; Galván, P.; García-Recio, S.; Prat, A. Intrinsic Subtypes in Breast Cancer: From Tumor Biology to Clinical Management. *Cancer Treat. Rev.* **2019**, *73*, 44–54.
10. Saelens, W.; Cannoodt, R.; Todorov, H.; Saeys, Y. A Comparison of Single-Cell Trajectory Inference Methods. *Nat. Biotechnol.* **2019**, *37*, 547–554. [CrossRef] [PubMed]
11. Kiselev, V.Y.; Andrews, T.S.; Hemberg, M. Challenges in Unsupervised Clustering of Single-Cell RNA-seq Data. *Nat. Rev. Genet.* **2019**, *20*, 273–282. [CrossRef] [PubMed]
12. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef] [PubMed]
13. Leek, J.T.; Scharpf, R.B.; Bravo, H.C.; Simcha, D.; Langmead, B.; Johnson, W.E.; Geman, D.; Baggerly, K.; Irizarry, R.A. Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* **2010**, *11*, 733–739. [CrossRef] [PubMed]
14. Libbrecht, M.W.; Noble, W.S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
16. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
17. Way, G.P.; Greene, C.S. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. *Pac. Symp. Biocomput.* **2018**, *23*, 80–91. [CrossRef]
18. Rampášek, L.; Goldenberg, A. TensorFlow: Biology's Gateway to Deep Learning? *Cell Syst.* **2016**, *2*, 12–14. [CrossRef] [PubMed]
19. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep Generative Modeling for Single-Cell Transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef] [PubMed]
20. Gross, B.; Dauvin, A.; Cabeli, V.; Kmetzsch, V.; El Khoury, J.; Dissez, G.; Ouardini, K.; Grouard, S.; Davi, R.; Loeb, R.; et al. Robust Evaluation of Deep Learning-Based Representation Methods for Survival and Gene Essentiality Prediction on Bulk RNA-Seq Data. *Sci. Rep.* **2024**, *14*, 17098. [CrossRef] [PubMed]
21. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
22. Aran, D.; Sirota, M.; Butte, A.J. Systematic Pan-Cancer Analysis of Tumour Purity Using Gene Expression Data. *Nat. Commun.* **2015**, *6*, 8971. [PubMed]
23. Yoshihara, K.; Shahmoradgoli, M.; Martínez, E.; Vegesna, R.; Kim, H.; Torres-Garcia, W.; Treviño, V.; Verhaak, R.G.W. Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* **2013**, *4*, 2612. [CrossRef] [PubMed]
24. Newman, A.M.; Liu, C.L.; Green, M.R.; Gentles, A.J.; Feng, W.; Xu, Y.; Diehn, M.; Alizadeh, A.A. Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* **2015**, *12*, 453–457. [CrossRef] [PubMed]
25. Galon, J.; Bruni, D. Approaches to Treat Hot and Cold Tumours. *Nat. Rev. Drug Discov.* **2019**, *18*, 197–218. [CrossRef] [PubMed]
26. Thorsson, V.; Gibbs, D.L.; Brown, S.D.; Wolf, D.; Bortone, D.S.; Ou Yang, T.H.; Shmulevich, I. The Immune Landscape of Cancer. *Immunity* **2018**, *48*, 812–830. [CrossRef] [PubMed]
27. Bruni, D.; Angell, H.K.; Galon, J. The Immune Contexture and Immunoscore in Cancer Prognosis and Therapeutic Efficacy. *Nat. Rev. Cancer* **2020**, *20*, 662–680. [CrossRef] [PubMed]
28. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
29. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
30. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
31. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **2012**, *16*, 284–287. [CrossRef] [PubMed]
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Kaplan, E.L.; Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481. [CrossRef]

34. Harrell, F.E.; Califf, R.M.; Pryor, D.B.; Lee, K.L.; Rosati, R.A. Evaluating the Yield of Medical Tests. *JAMA* **1982**, *247*, 2543–2546. [CrossRef]
35. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–220. [CrossRef]
36. Smyth, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 3. [CrossRef] [PubMed]
37. 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. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
38. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell Syst.* **2015**, *1*, 417–425. [CrossRef] [PubMed]
39. Finotello, F.; Mayer, C.; Plattner, C.; Laschober, G.; Rieder, D.; Hackl, H.; Krogsdam, A.; Loncova, Z.; Posch, W.; Wilflingseder, D.; et al. Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* **2019**, *11*, 34. [CrossRef] [PubMed]
40. Sturm, G.; Finotello, F.; Petitprez, F.; Zhang, J.D.; Baumbach, J.; Fridman, W.H.; List, M.; Aneichyk, T. Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology. *Bioinformatics* **2019**, *35*, i436–i445. [CrossRef] [PubMed]
41. Hennig, C. Cluster-wise Assessment of Cluster Stability. *Comput. Stat. Data Anal.* **2007**, *52*, 258–271. [CrossRef]
42. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017
43. Aran, D.; Camarda, R.; Odegaard, J.; Paik, H.; Oskotsky, B.; Krings, G.; Goga, A.; Sirota, M.; Butte, A.J. Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes. *Nat. Commun.* **2017**, *8*, 1077. [CrossRef] [PubMed]
44. von Luxburg, U.; Williamson, R.C.; Guyon, I. Clustering: Science or Art? In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Edinburgh, Scotland, 26 June–1 July 2012; Volume 27, pp. 65–79.
45. Griguolo, G.; Dieci, M.V.; Paré, L.; Miglietta, F.; Generali, D.G.; Frassoldati, A.; Cavanna, L.; Bisagni, G.; Piacentini, F.; Tagliafico, E.; et al. Immune Microenvironment and Intrinsic Subtyping in Hormone Receptor-Positive/HER2-Negative Breast Cancer. *npj Breast Cancer* **2021**, *7*, 12. [CrossRef] [PubMed]
46. Fattori, S.; Le Roy, A.; Houacine, J.; Robert, L.; Abes, R.; Gorvel, L.; Granjeaud, S.; Rouvière, M.S.; Ben Amara, A.; Boucherit, N.; et al. CD25high Effector Regulatory T Cells Hamper Responses to PD-1 Blockade in Triple-Negative Breast Cancer. *Cancer Res.* **2023**, *83*, 3026–3044. [CrossRef] [PubMed]
47. Ivanov, V.; Khalid, U.; Gurung, J.; Dimov, R.; Chonov, V.; Uchikov, P.; Kostov, G.; Ivanov, S. Use of AI Histopathology in Breast Cancer Diagnosis. *Medicina* **2025**, *61*, 1878. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.