Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Graph-Regularised Multi-Omics Autoencoders Improve Stability and Clinical Coherence of Breast Cancer Patient Embeddings

Habiba El Keraby
Sultan Qaboos University
Hamza Zidoum
zidoum@squ.edu.om
Sultan Qaboos University
Maha Bouslimani
Université Clermont Auvergne

Research Article

Keywords: multi-omics integration, unsupervised representation learning, Laplacian graph regularisation, network-based modelling, representation stability, breast cancer

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10256267/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Graph-Regularised Multi-Omics Autoencoders Improve Stability and Clinical Coherence of Breast Cancer Patient Embeddings

Habiba El Keraby[1], Hamza Zidoum[1*], Maha Bouslimani[2]

[1*]Department of Computer Science, Sultan Qaboos University, Muscat, Oman.

[2*]Engineering Mathematics and Data Science, Polytech Clermont-Ferrand, Aubière, France.

*Corresponding author(s). E-mail(s): zidoum@squ.edu.om;
Contributing authors: s147635@student.squ.edu.om;

Abstract

Unsupervised patient stratification from multi-omics cancer data requires representations that are simultaneously stable across independent training runs, geometrically organised, and clinically informative, yet these properties are rarely evaluated jointly, and standard autoencoder frameworks discard the interaction topology encoded in protein–protein interaction (PPI) networks entirely. We perform a controlled comparison of three autoencoder variants trained on the TCGA-BRCA cohort (941 patients; four omics modalities; 13,001 genes) under matched latent dimensionality ($\boldsymbol{d = 64}$) and strictly unsupervised hyperparameter selection: a Baseline Autoencoder, a Denoising Autoencoder, and a Graph-Regularised Denoising Autoencoder augmented with a PPI-based Laplacian smoothness term. Relative to the unstructured DAE, introducing a weak Laplacian term ($\boldsymbol{\lambda = 10^{-6}}$) substantially improves representation quality: across-restart clustering stability increases from ARI $= \mathbf{0.861}$ to $\mathbf{0.980}$, latent geometry preservation from Spearman $\boldsymbol{\rho = 0.915}$ to $\mathbf{0.997}$, and effective dimensionality reduces from 13 to 3 principal components at a 90% variance threshold. Clinically, 61% of latent dimensions encode metastatic status after FDR correction (versus 0% in both unstructured baselines), and PAM50 subtypes are non-randomly distributed across clusters despite near-zero partition-level agreement ($\boldsymbol{\chi^2 = 108.7}$, $\boldsymbol{p = 7.9 \times 10^{-16}}$), indicating a complementary rather than redundant relationship with established subtyping. All results are cohort-specific and hypothesis-generating. These findings show that a single biologically motivated regularisation term, without architectural modification or supervision,

can substantially reshape latent geometry and clinical encoding in unsupervised multi-omics embeddings, making PPI topology a prominent organising principle of the learned representation.

Keywords: multi-omics integration, unsupervised representation learning, Laplacian graph regularisation, network-based modelling, representation stability, breast cancer

1 Introduction

Breast cancer exhibits profound molecular heterogeneity, with clinically distinct subtypes that differ in prognosis, receptor expression, and therapeutic response Perou et al. (2000). The simultaneous availability of multi-omics profiling (including transcriptomics, copy number alteration, epigenomics, and somatic mutation) offers the potential to characterise tumour biology at multiple complementary molecular layers. However, integrating these heterogeneous, high-dimensional data sources into patient-level representations that are reproducible, biologically coherent, and clinically informative remains a central unsolved challenge.

Unsupervised representation learning offers a principled route to patient stratification that does not require predefined clinical labels, which are costly, incomplete, and may reflect only a subset of the biologically relevant variation Bengio et al. (2013); LeCun et al. (2015). Autoencoder-based approaches have demonstrated utility in cancer subtype discovery, survival stratification, and biomarker identification Chaudhary et al. (2018); Tran et al. (2021). A fundamental limitation of standard architectures, however, is their treatment of molecular inputs as an *unstructured* flat vector: every gene is processed independently, and the rich relational biology encoded in gene interaction networks is entirely discarded.

Protein–protein interaction (PPI) networks provide a biologically grounded prior over gene feature spaces Barabási et al. (2011). Genes connected in PPI networks participate in shared biological processes, exhibit coordinated expression patterns, and are co-dysregulated in disease Cowen et al. (2017). Incorporating this interaction topology as an *inductive bias* (by constraining learned representations to respect known interaction structure, for instance) could in principle yield embeddings that are more reproducible across independent training runs, more geometrically organised around biologically meaningful axes, and more reflective of clinically relevant molecular variation Mitchell (1980); Shalev-Shwartz and Ben-David (2014). Existing graph-based multi-omics approaches have demonstrated biological plausibility Li et al. (2022); Schulte-Sasse et al. (2021), but typically couple the PPI prior with supervised objectives or architectural complexity that conflates multiple effects, making it impossible to isolate the specific contribution of the structural prior.

A parallel line of work introduces biological network structure not through the encoder architecture but through the training objective, via a graph Laplacian smoothness penalty. Let $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ be the symmetric normalised Laplacian

of a gene interaction graph. Adding the term

$$\lambda \operatorname{Tr}(\mathbf{Y}^\top \mathbf{L}_{\text{sym}} \mathbf{Y}) = \lambda \sum_{(g,g') \in \mathcal{E}} w_{gg'} \|\mathbf{y}_g - \mathbf{y}_{g'}\|^2$$

to any loss function penalises latent dissimilarity between PPI-connected genes, encoding the smoothness assumption that strongly interacting genes should occupy similar positions in the learned representation. This quadratic form is the graph signal processing analogue of Tikhonov regularisation on a manifold Anis et al. (2019), and has a long history in high-dimensional regression Sun and Li (2010); Wang et al. (2013) and network-constrained statistical learning Zhang et al. (2013); Chen et al. (2015) in genomics. Pathway-constrained autoencoders such as PathME Lemsara et al. (2020) and prior-informed variational autoencoders Liu et al. (2024) incorporate related structural constraints by restricting encoder connections to curated pathway annotations, yielding more interpretable representations; however, these methods either rely on discrete pathway membership rather than a continuous interaction-weighted graph or are evaluated primarily on clustering accuracy rather than representation stability. Network-smoothing methods such as netSmooth Ronen and Akalin (2018) apply similar Laplacian diffusion to gene expression matrices to reduce noise in single-cell RNA-seq, demonstrating that graph-based smoothing consistently improves biological coherence. Network-regularised PCA variants Feng et al. (2019); Cottrell et al. (2024) impose Laplacian smoothness on principal component scores, showing that prior interaction structure can reorganise variance concentration in biologically coherent directions, an effect directly analogous to the dimensionality collapse (13 PCs $\rightarrow$ 3 PCs) observed in our GR-DAE. Crucially, none of these approaches evaluate whether graph regularisation improves *across-restart representation stability* or study how the Laplacian term reshapes patient-level latent geometry as a primary endpoint. Our work addresses this gap by treating stability and geometry preservation as primary evaluation criteria and using a strictly unsupervised hyperparameter selection protocol.

In this work, we ask a focused question: *Does biologically structured inductive bias, introduced through the training objective, measurably improve the stability, geometry, and clinical relevance of unsupervised multi-omics patient embeddings?* To address this, we perform a controlled comparison of three autoencoder variants trained on TCGA-BRCA multi-omics data under an identical reconstruction objective, latent dimensionality, and optimisation protocol. The Baseline Autoencoder (BAE) provides an unstructured reference with no corruption and no biological prior. The Denoising Autoencoder (DAE) introduces statistical robustness through stochastic input corruption only, capturing the effect of denoising in the absence of explicit structure. The Graph-Regularised Denoising Autoencoder (GR-DAE) uses the same denoising objective but augments the loss with a PPI-based Laplacian smoothness term that enforces similarity of gene-level latent representations along the interaction network. Clinical labels are never observed during training or hyperparameter selection, so any differences in embedding behaviour can be attributed to the form and strength of structural inductive bias rather than supervision.

The goal of this work is not to introduce a fundamentally new graph-regularised architecture, but rather to provide a *systematic, controlled comparison* that quantifies how a weak PPI-based Laplacian term affects the stability, geometry, and clinical encoding of unsupervised multi-omics embeddings under matched optimisation protocols. Concretely, our contributions are:

- We design an unsupervised experimental framework that isolates the effect of a PPI-based Laplacian regulariser from denoising, label-driven objectives, and changes in patient-level latent dimensionality.
- We show that even very weak Laplacian regularisation can substantially improve across-restart stability and latent geometry preservation compared with unstructured baselines, under a label-free hyperparameter selection protocol.
- We quantify how PPI-informed structure reshapes the clinical information encoded in the latent space, with a particular focus on metastasis-related variation and its distribution across embedding dimensions.
- We characterise how PPI-informed multi-omics embeddings relate to PAM50 molecular subtypes, highlighting a complementary but non-redundant organisation of the same tumours.

2 Materials and Methods

2.1 Cohort and Multi-Omics Data

We used the TCGA-BRCA cohort ($N = 941$ tumour samples) with four omics modalities (Table 1), obtained from the GDC Data Portal[1]. The combined multi-omics representation for each patient is a gene feature tensor $\mathbf{X}_p \in \mathbb{R}^{G\times 4}$, where each of the $G = 13{,}001$ genes contributes a 4-dimensional feature vector (one value per modality). Clinical covariates (pathological stage, T/N/M classification, age at diagnosis) and PAM50 subtype labels were obtained from the TCGA clinical portal and used exclusively for post-hoc evaluation; they were never used during model training or hyperparameter selection.

For M classification, 865 of 941 patients had non-missing annotations (730 M0, 18 M1, 117 MX), reflecting the known predominance of non-metastatic and 'not assessable' cases in TCGA-BRCA.

2.2 PPI Graph Construction

The STRING v11 database Szklarczyk et al. (2021) was used as the source of protein–protein interaction information. Protein identifiers were mapped to HGNC gene symbols, combined confidence scores were normalised to $[0, 1]$, and edges with confidence ≥ 0.85 were retained. The resulting graph spans all 13,001 genes in the omics universe; 3,983 genes (30.6%) are isolated nodes with no edges at this threshold and are retained to preserve full alignment with the omics data, although they contribute no gradient signal through the regularisation term. The symmetric normalised graph

[1] https://portal.gdc.cancer.gov

Table 1: Multi-omics modalities and preprocessing (TCGA-BRCA, 941 patients). All modalities are aligned to the same gene universe of $G = 13{,}001$ genes defined by intersection with the filtered STRING PPI network.

Modality	Source	Preprocessing	Final features
RNA-seq	HTSeq FPKM-UQ	$\log_2(\text{TPM} + 1)$, z-score per gene	13,001 genes
Copy Number Variation	GISTIC2 gene-level	z-score per gene	13,001 genes
DNA Methylation	450k array beta values	Mean per gene; missing = 0	13,001 genes
Somatic Mutation	MuTect2 variant calls	Binary gene-level indicator	13,001 genes

Laplacian

$$\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \tag{1}$$

is precomputed once from the weighted adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$, and is used as the regularisation kernel throughout training.

2.3 Model Architectures

2.3.1 Baseline and Denoising Autoencoders

The Baseline Autoencoder (BAE) and Denoising Autoencoder (DAE) operate on the flattened 52,004-dimensional multi-omics input vector $\mathbf{x}_p \in \mathbb{R}^{52{,}004}$, obtained by concatenating all four omics modalities across the shared gene universe while preserving gene and modality ordering.

Both models use the same fully connected encoder–decoder architecture: an encoder MLP with layer dimensions $52{,}004 \to 1{,}024 \to 256 \to 64$ and ReLU activations, followed by a symmetric decoder $64 \to 256 \to 1{,}024 \to 52{,}004$ with ReLU activations and a linear output layer. The 64-dimensional encoder output is the patient embedding $\mathbf{z} \in \mathbb{R}^{64}$.

The BAE is trained to minimise mean squared error (MSE) between the clean input and its reconstruction:

$$\mathcal{L}_{\text{BAE}} = \frac{1}{N}\sum_{p=1}^{N} \|\mathbf{x}_p - \hat{\mathbf{x}}_p\|^2. \tag{2}$$

No input corruption or biological prior is used, so this model serves as an unstructured reference that treats all gene-level features as independent and unrelated. We intentionally use this simple MLP as a minimal unstructured reference baseline rather than as a state-of-the-art multi-omics architecture; our goal is to quantify the incremental impact of denoising and graph-based inductive bias under tightly controlled conditions.

The DAE shares the same architecture but introduces stochastic input corruption during training. A corrupted version $\tilde{\mathbf{x}}_p$ is obtained by adding element-wise Gaussian

noise with standard deviation $\sigma = 0.3$ to the clean input; clean inputs are used at inference. The training objective is MSE($\mathbf{x}$, decoder(encoder($\tilde{\mathbf{x}}$))).

The architecture of both models is illustrated in Figure 1.

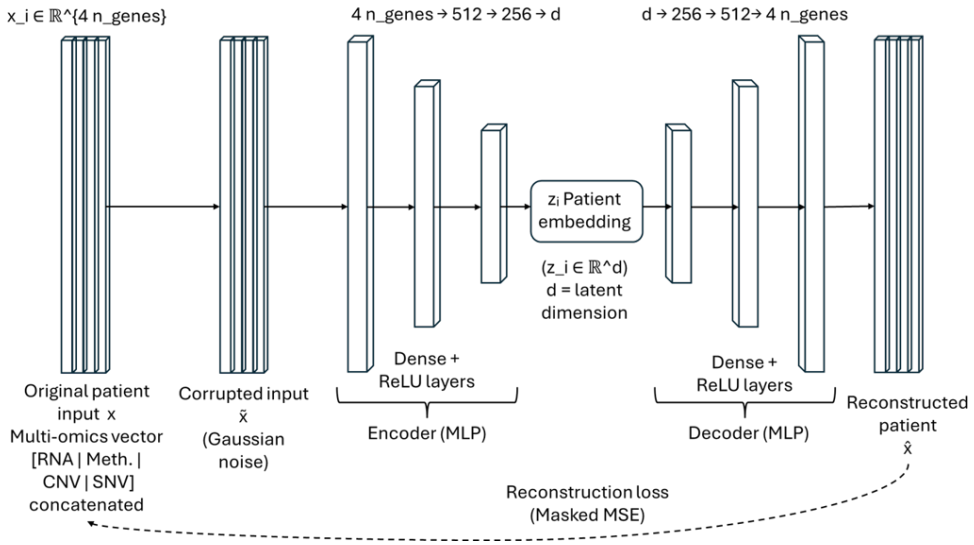


Fig. 1: Baseline and Denoising Autoencoder architecture. A corrupted version $\tilde{\mathbf{x}}$ of the flattened 52,004-dimensional multi-omics vector is passed through a patient-level MLP encoder (52,004 $\rightarrow$ 1,024 $\rightarrow$ 256 $\rightarrow$ 64) to produce a 64-dimensional patient embedding $\mathbf{z}$, then decoded to reconstruct the clean input $\mathbf{x}$ (MSE loss). All gene features are treated as an unstructured flat vector; no biological interaction structure is incorporated.

2.3.2 Graph-Regularised Denoising Autoencoder (GR-DAE)

The GR-DAE introduces PPI structure exclusively through the loss function while keeping the patient-level latent dimensionality and overall training protocol identical to the DAE.

Three design choices distinguish it from the flattened baselines; the first two are architectural prerequisites for the third. First, the input is restructured as a gene-feature tensor $\mathbf{X}_p \in \mathbb{R}^{G\times 4}$, and a shared-weight gene-wise MLP encoder ($4 \rightarrow 128 \rightarrow 64 \rightarrow 64$, ReLU) processes each gene's 4-dimensional omics vector independently, producing a gene-level latent matrix $\mathbf{Y} \in \mathbb{R}^{G\times 64}$. Second, the patient embedding $\mathbf{z}_p \in \mathbb{R}^{64}$ is obtained by mean pooling over all gene representations: $\mathbf{z}_p = \frac{1}{G}\sum_{g=1}^{G} \mathbf{Y}_{p,g}$. These two steps are necessary conditions for the third: the Laplacian smoothness penalty operates on the gene-level representations $\mathbf{Y}$, penalising latent dissimilarity between genes connected in the PPI graph. This penalty has no analogue in the flat patient-level MLP, where no intermediate gene-level representation exists and the encoder produces $\mathbf{z}_p$ directly from the 52,004-dimensional concatenated input.

Third, the training objective augments the per-gene reconstruction MSE with a Laplacian smoothness penalty:

$$\mathcal{L}_{\text{GR-DAE}} = \mathcal{L}_{\text{recon}} + \lambda \operatorname{Tr}\big(\mathbf{Y}^\top \mathbf{L}_{\text{sym}} \mathbf{Y}\big), \tag{3}$$

where

$$\mathcal{L}_{\text{recon}} = \frac{1}{N \cdot G} \sum_{p=1}^{N} \sum_{g=1}^{G} \|\mathbf{x}_{p,g} - \hat{\mathbf{x}}_{p,g}\|^2, \tag{4}$$

$\lambda \geq 0$ controls the strength of the biological prior (setting $\lambda = 0$ recovers the unregularised gene-wise DAE), and

$$\operatorname{Tr}\big(\mathbf{Y}^\top \mathbf{L}_{\text{sym}} \mathbf{Y}\big) = \sum_{(g,g') \in \mathcal{E}} w_{g,g'} \|\mathbf{y}_g - \mathbf{y}_{g'}\|^2 \tag{5}$$

is the graph smoothness term, which penalises large differences in latent representations between gene pairs (g, g') connected by an edge of weight $w_{g,g'}$ in the PPI network. A symmetric gene-wise decoder ($64 \to 64 \to 128 \to 4$) reconstructs the per-gene omics vectors from $\mathbf{Y}$. The selected regularisation strength $\lambda = 10^{-6}$ was determined by unsupervised hyperparameter search (Section 2.5). Input corruption ($\sigma = 0.3$) and clean-input inference are shared with the DAE.

The Laplacian quadratic form $\operatorname{Tr}(\mathbf{Y}^\top \mathbf{L}_{\text{sym}} \mathbf{Y}) = \sum_{(g,g') \in \mathcal{E}} w_{gg'} \|\mathbf{y}_g - \mathbf{y}_{g'}\|^2$ can be interpreted as the total squared variation of gene-level latent representations across the PPI graph, weighted by interaction confidence. Its spectrum is non-negative (since $\mathbf{L}_{\text{sym}}$ is positive semi-definite) and its magnitude scales with the number and weight of edges; for our graph (9,018 connected genes, 159,602 edges, mean weight ≈ 0.90), the unweighted Frobenius norm of $\operatorname{Tr}(\mathbf{Y}^\top \mathbf{L}_{\text{sym}} \mathbf{Y})$ at initialisation is of order 10^4–10^5. Setting $\lambda = 10^{-6}$ therefore produces an initial regularisation contribution of order 10^{-2}–10^{-1}, comparable in scale to the reconstruction loss (approximately 0.01 at initialisation for the gene-wise model), ensuring that the structural constraint is present but does not override the reconstruction signal. This is consistent with the *gentle structural nudge* regime identified in our grid search, in which $\lambda \leq 10^{-4}$ reliably improves stability whereas $\lambda \geq 10^{-2}$ degrades it.

The GR-DAE architecture is illustrated in Figure 2.

Although the patient-level latent dimensionality and optimisation protocol are matched across all models, the GR-DAE introduces a different inductive structure through its gene-wise shared-weight encoder and mean-pooling aggregation. Consequently, some of the observed differences between the DAE and GR-DAE may reflect this architectural change in addition to the explicit Laplacian term. We partially address this by including $\lambda = 0$ in the GR-DAE hyperparameter grid (Section 2.5), but disentangling architectural and regularisation effects remains an important direction for future work.

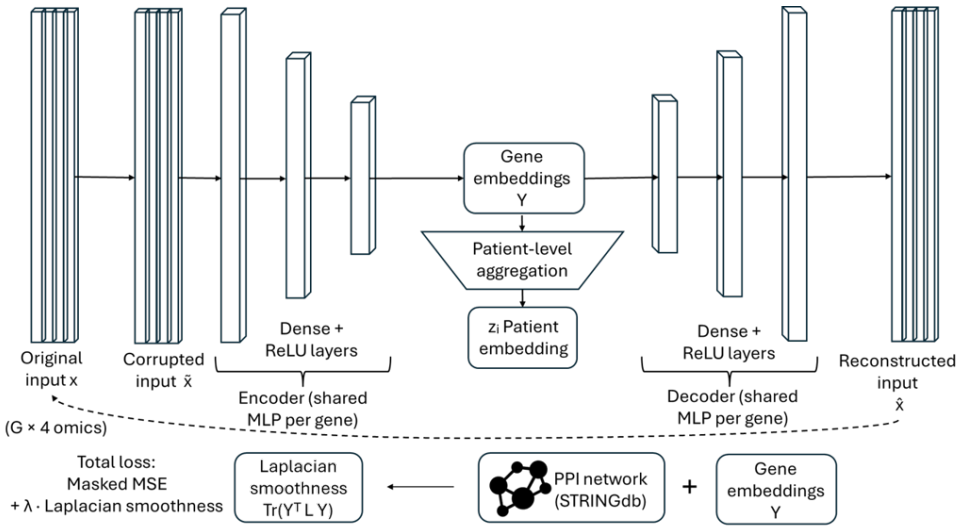


Fig. 2: Graph-Regularised Denoising Autoencoder (GR-DAE) architecture. The input is the gene feature tensor $\mathbf{X}_p \in \mathbb{R}^{G\times 4}$. A shared-weight gene-wise MLP processes each gene independently to produce gene-level latent representations $\mathbf{Y} \in \mathbb{R}^{G\times 64}$. The patient embedding $\mathbf{z} \in \mathbb{R}^{64}$ is obtained by mean pooling over genes. Biological prior knowledge enters exclusively through the Laplacian smoothness term $\lambda \cdot \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}_{\mathrm{sym}} \mathbf{Y})$, which penalises latent dissimilarity between PPI-connected genes, without modifying the reconstruction pathway or introducing graph propagation in the encoder.

2.4 Shared Training Protocol

All three models share the training configuration summarised in Table 2. Identical optimisation, early stopping, and latent dimensionality ensure that observed differences in representation behaviour can be attributed to the presence or absence of denoising and PPI-based graph regularisation rather than training differences.

2.5 Unsupervised Hyperparameter Selection

The DAE corruption parameters and the GR-DAE regularisation coefficient λ were selected using exclusively unsupervised stability criteria. For the DAE, a grid search over Gaussian noise levels and masking probabilities was performed; the selected configuration (Gaussian noise with $\sigma = 0.3$, no masking) maximised a composite score combining across-restart clustering stability and geometry preservation.

For the GR-DAE, the regularisation coefficient λ was selected by a grid search over ten values spanning multiple orders of magnitude: $\lambda \in \{0, 10^{-6}, 3\times 10^{-6}, 10^{-5}, 3\times 10^{-5}, 10^{-4}, 3\times 10^{-4}, 10^{-3}, 3\times 10^{-3}, 10^{-2}\}$.

Table 2: Shared training configuration applied identically to all models. Differences in evaluation results are primarily attributable to the presence of denoising and PPI-based Laplacian regularisation, together with the shift from a flattened patient-level MLP to a gene-wise shared-weight encoder in the GR-DAE.

Hyperparameter	**Value**
Optimiser	Adam
Learning rate	1×10^{-3} (fixed)
Batch size	64
Train / Validation split	90% / 10% (random, seed-controlled; 847 / 94 samples)
Maximum epochs	50
Early stopping	Patience 10 epochs; min. improvement $\delta = 10^{-4}$
Checkpoint	Epoch with lowest validation reconstruction MSE
Input corruption (training)	Gaussian noise $\sigma = 0.3$; evaluated on clean inputs
Latent dimensionality d	64
Number of training seeds	5 (seeds 0–4)
Weight initialisation	Kaiming uniform (PyTorch default)

Each configuration was trained across five independent random seeds, and the following composite criterion was computed:

$$\text{Score}(\lambda) = \text{ARI}_{\text{restarts}} + \rho_{\text{Spearman, restarts}},$$

where terms denote the mean pairwise ARI and mean pairwise Spearman geometry correlation across restarts, respectively. Clinical labels, PAM50 subtypes, and survival outcomes were explicitly excluded from this criterion. The no-regularisation variant ($\lambda = 0$) of the GR-DAE already achieves high restart stability (mean ARI = 0.934, geometry Spearman $\rho = 0.983$, composite score = 1.917), indicating that the gene-wise architecture alone provides a strong inductive bias. The selected value $\lambda = 10^{-6}$ further improves these metrics (mean ARI = 0.980, NMI = 0.964, geometry Spearman $\rho = 0.997$, composite score = 1.977), indicating that even very weak PPI regularisation improves unsupervised stability beyond the unregularised gene-wise denoising baseline.

2.6 Evaluation Protocol

All evaluations are strictly post hoc; clinical labels are used exclusively for association analysis and are never used for model training or hyperparameter selection. Unless otherwise specified, analyses are conducted on the full 941-sample cohort and applied uniformly to the BAE, DAE, and GR-DAE.

2.6.1 Representation Stability

To assess representation stability, k-means clustering ($k = 5$, `n_init` $= 20$) was applied at two levels: (i) within-representation stability, using 10 random k-means seeds on a fixed trained embedding, and (ii) across-restart stability, using 5 independent training seeds per model. We chose five training seeds as a compromise between computational

cost and the need to estimate variability across initialisations; preliminary experiments with fewer seeds showed qualitatively similar patterns but less stable estimates of restart metrics. Pairwise Adjusted Rand Index (ARI) and Normalised Mutual Information (NMI) were computed for all seed pairs. Latent geometry stability was quantified as Spearman and Pearson correlation between condensed pairwise Euclidean distance vectors (all patient–patient pairs) across independent training runs. The choice $k = 5$ is motivated by the five canonical PAM50 breast cancer subtypes; no assumption is made that learned clusters correspond one-to-one with these subtypes.

Choice of k.

The choice $k = 5$ is motivated by two complementary considerations. First, it enables post-hoc contextual alignment with the five canonical PAM50 molecular subtypes (Section 3.5), providing a meaningful external reference frame for the learned clusters. Second, internal cluster quality metrics (Silhouette score and Davies–Bouldin index) were evaluated across $k \in \{2, \ldots, 8\}$ for the BAE, DAE, and GR-DAE embeddings using 10 k-means random seeds on each reference-run embedding (Table 3). For the GR-DAE, the Silhouette score peaks at $k = 5$ (0.457) and the Davies–Bouldin index reaches its minimum in the range $k \in \{5, 6\}$ (0.891 and 0.890 respectively), jointly indicating that $k = 5$ is the data-supported optimum for this embedding. Importantly, the choice of $k = 5$ is therefore not only PAM50-motivated but independently validated by the cluster quality metrics. By contrast, BAE and DAE embeddings show flat, low Silhouette scores ($\approx$0.21–0.25) across all tested k, indicating poorly separated cluster structure regardless of the assumed number of clusters, a further reflection of their unstructured latent geometry.

We note that the across-restart stability advantage of the GR-DAE (Section 3.3) reflects properties of the latent geometry rather than sensitivity to this particular k choice: the substantially higher variance concentration in the GR-DAE (3 vs. 13 principal components at the 90% variance threshold) is a property of the embedding space itself, independent of any downstream clustering step.

2.6.2 Latent Geometry

To characterise latent geometry, per-dimension variance and the number of collapsed dimensions (variance $< 10^{-6}$) were computed for each model. Principal component analysis (PCA) was applied to the 64-dimensional embedding space, and the cumulative explained variance profile was summarised by the minimum number of components required to exceed 90% cumulative variance. Patient-level ℓ_2-norm distributions of the embedding vectors were computed across all 941 patients to assess scale and concentration.

2.6.3 Post-Hoc Clinical Association

Cluster-level enrichment analyses assessed whether unsupervised clusters capture known clinicopathological variation. Associations between k-means cluster membership and pathologic stage or age at diagnosis were tested using Kruskal–Wallis tests.

Table 3: Internal cluster quality across $k \in \{2, \ldots, 8\}$ for BAE, DAE, and GR-DAE embeddings (k-means, 10 random seeds per k, reference-run embedding). Silhouette score (↑ better) and Davies–Bouldin index (↓ better) are reported as mean ± s.d. For GR-DAE, the Silhouette score peaks at $k = 5$ (0.457) and the Davies–Bouldin index reaches its minimum in the range $k \in \{5, 6\}$, jointly supporting $k = 5$ as both data-driven and consistent with PAM50 contextual alignment. BAE and DAE show flat, low Silhouette scores across all k, indicating poorly separated cluster structure regardless of the number of clusters assumed.

Model	k	**Silhouette** (↑)	**Davies–Bouldin** (↓)
BAE	2	0.253 ±0.000	1.649 ±0.000
	3	0.233 ±0.000	1.749 ±0.000
	4	0.233 ±0.000	1.674 ±0.000
	5	0.221 ±0.000	1.645 ±0.002
	6	0.221 ±0.003	1.532 ±0.188
	7	0.222 ±0.002	1.581 ±0.087
	8	0.216 ±0.019	1.540 ±0.063
DAE	2	0.224 ±0.000	1.837 ±0.000
	3	0.217 ±0.000	1.794 ±0.002
	4	0.216 ±0.001	1.438 ±0.190
	5	0.210 ±0.013	1.476 ±0.057
	6	0.208 ±0.004	1.510 ±0.022
	7	0.212 ±0.001	1.541 ±0.021
	8	0.213 ±0.001	1.454 ±0.086
GR-DAE	2	0.357 ±0.000	1.247 ±0.000
	3	0.409 ±0.000	0.976 ±0.000
	4	0.441 ±0.000	0.920 ±0.002
	5	**0.457** ±0.000	**0.891** ±0.002
	6	0.435 ±0.025	0.890 ±0.087
	7	0.412 ±0.011	0.901 ±0.041
	8	0.404 ±0.003	0.937 ±0.012

Bold row indicates $k = 5$ for GR-DAE, at which the Silhouette score is maximised; $k = 5$ is used for all three models throughout the main analysis to enable PAM50 contextual alignment (Section 3.5). Standard deviations of 0.000 indicate that all 10 k-means seeds converged to identical cluster assignments at that k, reflecting a highly stable partition.

Associations with T, N, and M classifications were tested using chi-square tests, after collapsing rare category levels with fewer than 5 patients into an "Other" category.

Per-dimension clinical associations probed how broadly clinical information is encoded across the latent space. For each model, each of the 64 latent dimensions was tested individually using Spearman correlation (age, numeric pathologic stage) or Kruskal–Wallis tests (T, N, M classifications), with Benjamini–Hochberg FDR correction ($\alpha = 0.05$) applied across all 64 dimensions per variable. The primary summary statistic is the fraction of dimensions with a significant association after FDR correction.

2.6.4 PAM50 Contextual Alignment

Partition-level agreement between k-means cluster assignments and PAM50 labels was quantified using ARI and NMI. Statistical dependence between cluster membership and PAM50 subtype was assessed using a chi-square test on the cluster–subtype contingency table ($k = 5$ clusters, up to five PAM50 classes). PAM50 is treated as a reference framework rather than a ground-truth target: lack of cluster–subtype alignment is not interpreted as model failure and may reflect alternative biological organisation or finer-grained structure not captured by discrete PAM50 classes Parker et al. (2009).

3 Results

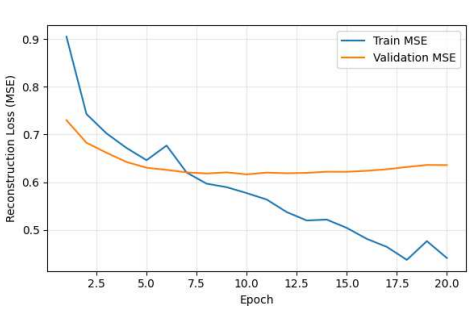
3.1 Training Convergence and Optimisation Consistency

All models converged within the 50-epoch training budget (Table 4; Figure 3). The BAE and DAE showed smooth, monotonic decreases in reconstruction loss from approximately 0.91 to 0.43, with early stopping selecting checkpoints around epochs 10–12 (validation MSE ≈ 0.62). Across-seed variance in final reconstruction MSE was modest but non-negligible for the unstructured models (standard deviation 0.022 for BAE and 0.012 for DAE), indicating some sensitivity to random weight initialisation and stochastic optimisation.

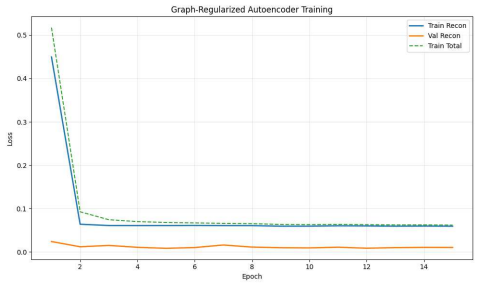
The GR-DAE exhibited qualitatively different but equally well-behaved convergence. Per-gene reconstruction MSE dropped rapidly to near zero within two to three epochs, after which the Laplacian smoothness term, itself decreasing from approximately 89,000 to near zero within the first five to seven epochs, dominated the early-training objective before plateauing. The selected checkpoint occurred at epoch 18 (validation MSE ≈ 0.009). Strikingly, the five independent restart curves for GR-DAE were nearly indistinguishable throughout training (reconstruction MSE standard deviation 0.0004), representing roughly a 30–50-fold reduction in across-seed variance compared with the unstructured DAE. This exceptional optimisation consistency is the first indication that the Laplacian regularisation narrows the effective hypothesis space and smooths the loss landscape, reducing the diversity of attainable local solutions.

Table 4: Training convergence statistics (reference run: seed 0) and across-restart reconstruction MSE. † Per-gene MSE for GR-DAE; values are not numerically comparable to BAE/DAE MSE because the output dimensionality differs (4 per gene vs. 52,004 flattened features).

Model	Stop epoch	Best epoch	Best val. MSE	Train MSE (mean)	Train MSE (std, 5 seeds)
BAE	21	11	0.622	0.430	0.022
DAE	20	10	0.6163	0.429	0.012
GR-DAE	28	18	0.009†	0.061†	0.0004



(a) DAE reconstruction MSE (seed 0).



(b) GR-DAE per-gene reconstruction MSE and total loss (seed 0).

Fig. 3: Training convergence. **(a)** DAE reconstruction MSE for the reference run (seed 0), with across-seed variance of final reconstruction loss (std ≈ 0.012). **(b)** GR-DAE per-gene reconstruction MSE (solid) and total loss (dashed; reconstruction $+\lambda\cdot$Laplacian smoothness) for the reference run; across-seed variance of final reconstruction MSE is 0.0004. Per-gene MSE values (b) are not numerically comparable to patient-level MSE values (a) due to the difference in output dimensionality.

Preliminary diagnostics showed that the BAE and DAE exhibit similar latent geometry and clinical null results, with the DAE providing slightly higher unsupervised stability. For this reason, detailed representation-level analyses focus on the comparison between the DAE (strongest unstructured baseline) and the GR-DAE (PPI-based inductive bias).

Effect of λ on stability.

A separate hyperparameter search evaluated GR-DAE stability across ten regularisation strengths $\lambda \in [0, 10^{-2}]$ using five independent restarts per value. The no-regularisation GR-DAE ($\lambda = 0$) already achieves high across-restart stability (mean ARI $= 0.934$, mean NMI $= 0.915$, geometry Spearman $\rho = 0.983$), and all weak-to-moderate regularisation values up to $\lambda = 10^{-4}$ increase the composite stability criterion relative to this baseline. The optimal configuration, $\lambda = 10^{-6}$, attains the highest composite score (mean ARI $= 0.980$, NMI $= 0.964$, geometry Spearman $\rho = 0.997$), while very strong regularisation ($\lambda = 10^{-2}$) degrades both geometry ($\rho = 0.856$) and stability (mean ARI $= 0.766$). These results support a non-monotone relationship between λ and stability, with a broad plateau of beneficial weak regularisation and clear deterioration once the smoothness penalty dominates reconstruction.

3.2 Latent Space Geometry

PCA analysis reveals a marked geometric reorganisation of the embedding space when PPI-based regularisation is introduced (Table 5; Figure 4). In the DAE, information is distributed broadly across the 64-dimensional latent space: the first principal component (PC1) explains 24.6% of latent variance, the first two components together

explain 44.7%, and 13 components are required to exceed the 90% cumulative variance threshold. No collapsed dimensions (variance $< 10^{-6}$) are detected, indicating that the encoder uses its full latent capacity but without a dominant set of organising axes.

The GR-DAE produces a substantially more concentrated structure. PC1 explains 59.5% of variance (a 2.42-fold increase relative to the DAE), the first two components together explain 82.9%, and the 90% threshold is reached by the third component (92.7% cumulative). This 4.3-fold reduction in effective dimensionality (13 PCs $\rightarrow$ 3 PCs) reflects the organising effect of the Laplacian regularisation: by penalising latent dissimilarity between PPI-connected genes, the encoder is biased toward configurations that align along a small number of globally coherent principal axes rather than spreading variance evenly across many weak directions.

While a more anisotropic latent space is not inherently preferable, in our setting the increased variance concentration coincides with improved restart stability and stronger clinical associations, suggesting that it reflects more structured rather than degenerate representations.

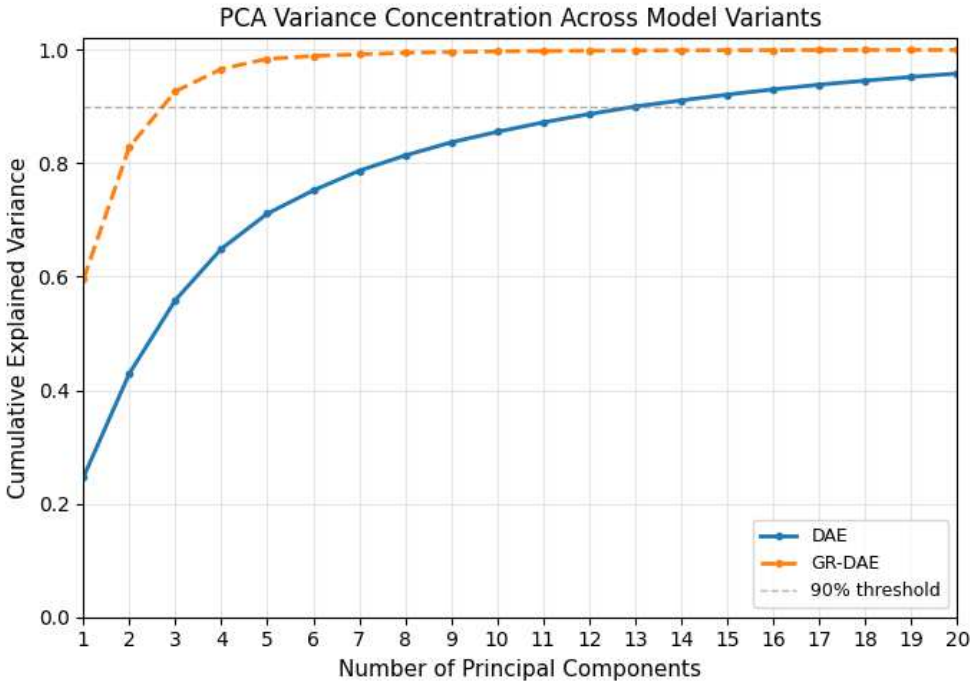


Fig. 4: PCA cumulative explained variance (reference run, seed 0). The DAE (blue) requires 13 principal components to exceed the 90% variance threshold (dashed line), whereas the GR-DAE (green dashes) reaches this threshold at 3 components. PC1 explains 24.6% vs. 59.5% of variance, respectively, a 2.42-fold increase. This geometric reorganisation reflects the PPI-based Laplacian regularisation aligning gene representations along a small number of biologically coherent principal directions.

Table 5: Latent geometry statistics (reference run, seed 0). ⋆ Substantially more concentrated structure in the GR-DAE. † Smaller norms reflect the gene-wise mean-pooling architecture and are not directly comparable to Baseline AE norms.

Model	Collapsed dims	PC1 var. (%)	PC1+2 var. (%)	PCs for 90%	Mean ℓ_2 norm
DAE	0/64	24.6	42.9	13	97.8 ± 25.3
GR-DAE	0/64	59.5⋆	82.9⋆	3⋆	$0.187 \pm 0.064^{\dagger}$

3.3 Representation Stability

A central result of this study is the substantial improvement in across-restart representation stability produced by PPI-based graph regularisation (Figure 5; Table 6). Across the 10 pairwise comparisons of the five independent training runs, the DAE achieves a mean pairwise ARI of 0.861 ± 0.028, indicating that different random initialisations lead to noticeably different patient partitions. The GR-DAE achieves mean pairwise ARI of 0.980 ± 0.009 (a 24.2% absolute improvement over the unstructured baseline) with all 10 pairwise comparisons exceeding ARI = 0.963. This near-perfect across-restart consistency shows that the PPI-regularised optimisation landscape converges to essentially the same patient partition regardless of initialisation.

Latent geometry stability, measured as Spearman correlation between condensed pairwise Euclidean distance matrices of all 941 patients across training seeds, exhibits a similar pattern. Geometry preservation improves from 0.915 ± 0.010 in the DAE to 0.997 ± 0.001 in the GR-DAE, indicating that patient–patient relationships in the latent space are effectively identical across independent training runs for the graph-regularised model.

Within-representation stability (ARI across k-means seeds for a fixed trained representation) is high for both models (mean ARI 0.9997 ± 0.0007 for the DAE and 0.9862 ± 0.0118 for the GR-DAE), confirming that cluster structure is a stable property of each learned embedding space rather than an artefact of k-means random initialisation.

Table 6: Representation stability metrics (mean ± s.d.). ⋆ Substantial improvement. Within-rep: 10 k-means seeds on the fixed reference-run representation (45 pairs). Restart: 5 independent training seeds (10 pairs). For GR-DAE, results are reported for the selected configuration with $\lambda = 10^{-6}$.

Model	Within-rep ARI	Restart ARI	Restart NMI	Geometry ρ
DAE	0.9952 ± 0.0037	0.861 ± 0.038	0.814 ± 0.041	0.915 ± 0.010
GR-DAE	0.9862 ± 0.0118	0.980 ± 0.009⋆	0.964 ± 0.015⋆	0.997 ± 0.001⋆
Absolute gain	−0.009	+0.119 (+13.8%)	+0.150	+0.082 (+9.0%)

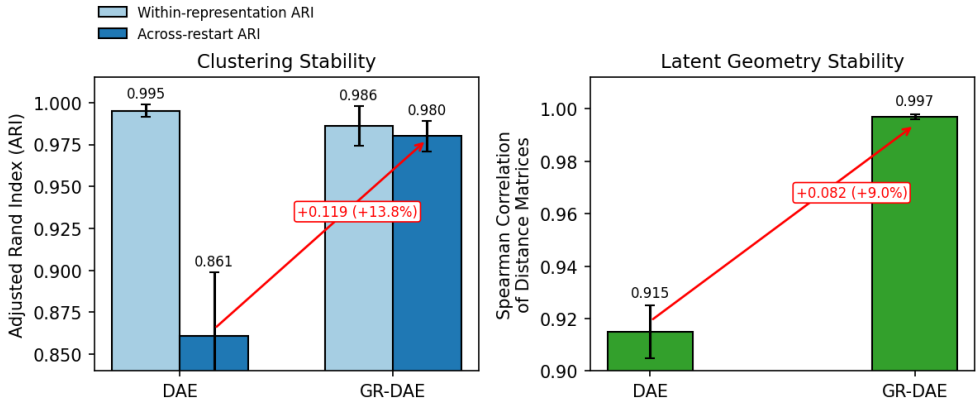


Fig. 5: Representation stability. **(a)** Within-representation ARI (light bars, 10 k-means seeds) and across-restart ARI (dark bars, 5 training seeds), mean $\pm$ s.d. **(b)** Across-restart Spearman correlation of pairwise patient distance matrices. Graph regularisation improves across-restart ARI from 0.861 to 0.980 (+13.8%) and geometry preservation from 0.915 to 0.997. Error bars show standard deviation across pairwise comparisons.

For completeness, we also evaluated an unregularised gene-wise GR-DAE variant with $\lambda = 0$, which already achieves high restart stability (mean ARI = 0.934, NMI = 0.915, geometry Spearman $\rho = 0.983$). We do not include this model in Table 6 to keep the main comparison focused on the strongest flattened baseline (DAE) and the selected graph-regularised configuration ($\lambda = 10^{-6}$), but these values further illustrate that weak Laplacian regularisation improves stability on top of a strong architectural inductive bias.

An important methodological qualification concerns the relative contributions of the two design changes distinguishing the GR-DAE from the DAE: the shift to a gene-wise shared-weight encoder with mean-pooling, and the addition of the PPI-Laplacian regularisation term. Table 7 presents these as a two-step ablation using the $\lambda = 0$ variant already evaluated during hyperparameter search (Section 2.5). The gene-wise architecture alone improves across-restart ARI from 0.861 to 0.934, reflecting the statistical regularisation effect of parameter tying across 13,001 genes. Weak Laplacian regularisation ($\lambda = 10^{-6}$) provides a further improvement to ARI 0.980, with near-perfect geometry preservation ($\rho = 0.997$). Critically, the architectural contribution to stability does not imply that the clinical encoding results (the emergence of metastatic status as a dominant organising factor across 61% of embedding dimensions) can be attributed to architecture or input compression: this signal is absent in the unstructured DAE and emerges specifically in the graph-regularised configuration,

Table 7: Ablation of inductive bias sources. The GR-DAE ($\lambda = 0$) variant isolates the contribution of the gene-wise shared-weight architecture and mean-pooling aggregation, holding all other training conditions fixed; GR-DAE ($\lambda = 10^{-6}$) adds the PPI-Laplacian smoothness term on top. Both architectural and regularisation components contribute to across-restart stability, but the emergence of strong clinical encoding is observed exclusively in the graph-regularised configuration, and cannot be attributed to architectural changes or input compression alone.

Model	Architecture	Graph prior (λ)	Restart ARI	Restart NMI	Geometry ρ
DAE	Flat patient MLP	None	0.861 ± 0.038	0.814 ± 0.041	0.915 ± 0.010
GR-DAE ($\lambda = 0$)	Gene-wise MLP + pool	None (architecture only)	0.934 ± 0.027	0.915 ± 0.024	0.983 ± 0.005
GR-DAE ($\lambda = 10^{-6}$)	Gene-wise MLP + pool	PPI Laplacian	0.980 ± 0.009	0.964 ± 0.015	0.997 ± 0.001

The improvement from DAE to GR-DAE ($\lambda = 0$) reflects the statistical inductive bias of gene-wise parameter tying; the further improvement to GR-DAE ($\lambda = 10^{-6}$) reflects the additional biological inductive bias of PPI-Laplacian regularisation.

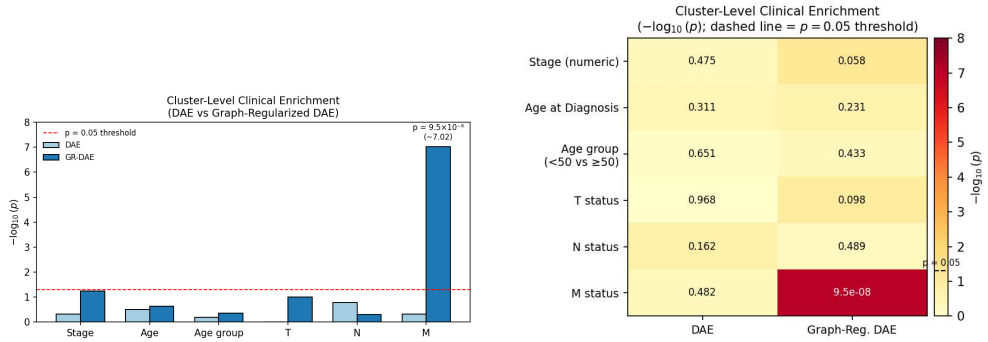
suggesting it is induced by the PPI-informed loss term. This term biases gene representations toward network-coherent configurations that happen to align with distributed, pathway-level metastatic variation (Section 4.2).

3.4 Post-Hoc Clinical Association

3.4.1 Cluster-Level Enrichment

Cluster-level association tests (Figure 6; Table 8) reveal a clear and systematic difference between the unstructured and graph-regularised embeddings. The DAE shows no significant association between k-means cluster membership and any of the clinical variables tested: pathologic stage (Kruskal–Wallis $p = 0.475$), age at diagnosis ($p = 0.311$), age group ($p = 0.651$), T classification (χ^2 $p = 0.968$), N classification ($p = 0.162$), and M classification ($p = 0.482$).

The GR-DAE shows a qualitatively different pattern. Metastatic status (M classification) is strongly and highly significantly associated with cluster membership (χ^2 $p = 9.5 \times 10^{-8}$), with additional trends toward significance for primary tumour status (T classification, $p = 0.098$), pathologic stage ($p = 0.058$), and N status ($p = 0.489$). Inspection of the M contingency table reveals a non-uniform distribution of M0, M1, and MX designations: GR-DAE Cluster 0 is almost entirely M0 (120 M0, 2 M1, 3 MX), whereas Cluster 1 shows substantially higher MX prevalence (84 M0, 0 M1, 25 MX). This structured separation of metastatic classification patterns is entirely absent in the DAE.



(a) $-\log_{10}(p)$ bar chart.

(b) $-\log_{10}(p)$ heatmap.

Fig. 6: Cluster-level clinical enrichment. **(a)** $-\log_{10}(p)$ bar chart for DAE and GR-DAE across six clinicopathological variables (stage, age at diagnosis, age group, T, N, M), with the dashed red line indicating the $p = 0.05$ significance threshold. **(b)** $-\log_{10}(p)$ heatmap summarising the same tests; numbers inside each cell are raw p-values. The DAE shows no significant association between cluster membership and any variable (all $p \geq 0.162$), whereas the GR-DAE exhibits a highly significant association with metastatic status (M classification, $p = 9.5 \times 10^{-8}$) and weaker trends for T status and stage.

Table 8: Cluster-level clinical enrichment p-values for $k = 5$ clusters (seed 0) in the DAE and GR-DAE embeddings. KW = Kruskal–Wallis; χ^2 = chi-square test. $\star$ Highly significant association present only in the GR-DAE, indicating that PPI-structured learning induces metastasis-related cluster separation absent in the unstructured DAE.

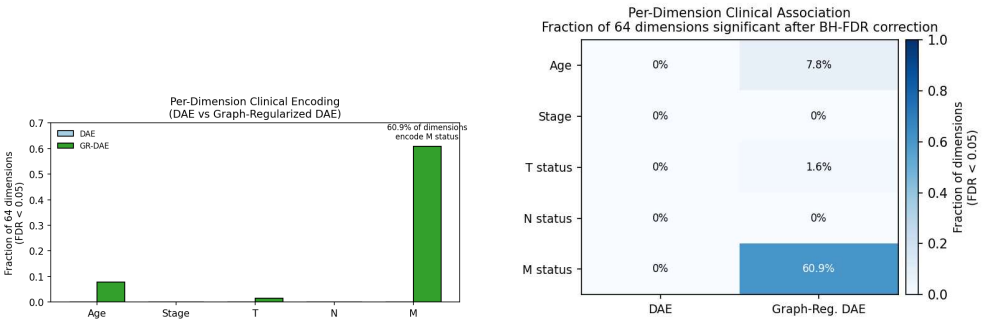
Clinical variable	**Test**	**DAE p-value**	**GR-DAE p-value**
Stage (numeric)	KW	0.475	0.058
Age at diagnosis	KW	0.311	0.231
Age group (<50 vs ≥50)	χ^2	0.651	0.433
T classification	χ^2	0.968	0.098
N classification	χ^2	0.162	0.489
M classification	χ^2	0.482	9.5×10^{-8} $\star$

3.4.2 Per-Dimension Clinical Encoding

FDR-corrected per-dimension association analysis (Benjamini–Hochberg, $\alpha = 0.05$) confirms and extends the cluster-level findings (Figure 7; Table 9). The DAE shows 0% significant dimensions for all clinical variables: age at diagnosis, numeric pathologic stage, T classification, N classification, and M classification. No individual latent dimension encodes any detectable clinicopathological signal after FDR correction.

The GR-DAE exhibits diffuse but highly specific clinical encoding. Metastatic status (M classification) is associated with 60.9% of the 64 embedding dimensions (39 of

64 dimensions pass FDR correction), suggesting that PPI-based graph regularisation elevates metastasis-related variation to a major organising factor of the latent space in this cohort. Age at diagnosis is encoded in 7.8% of dimensions (5 of 64), and T classification in 1.6% (1 of 64), while no dimensions pass FDR correction for N classification or numeric pathologic stage.



(a) Fraction of significant dimensions (bar chart).

(b) Fraction of significant dimensions (heatmap).

Fig. 7: Per-dimension clinical encoding. **(a)** Fraction of 64 embedding dimensions with significant clinical association (BH-FDR < 0.05) for each variable in the DAE and GR-DAE embeddings. **(b)** Heatmap of the same fractions; cell values are percentages of dimensions with significant association. The DAE shows 0% significant dimensions for all variables, whereas the GR-DAE encodes metastatic status (M classification) across 60.9% of dimensions and exhibits weaker, diffuse encoding for age and T status, indicating that PPI-based graph regularisation emphasises metastasis-related variation as a broad organising factor of the latent space in the studied cohort.

Table 9: Fraction of 64 embedding dimensions significantly associated with each clinical variable (BH-FDR $\alpha = 0.05$). Spearman correlation is used for age and stage; Kruskal–Wallis tests for T, N, and M classifications. $\star$ In the GR-DAE, 60.9% of dimensions encode metastatic status, compared with 0% in the DAE.

Model	**Age (%)**	**Stage (%)**	**T (%)**	**N (%)**	**M (%)**
DAE	0%	0%	0%	0%	0%
GR-DAE	7.8%	0%	1.6%	0%	60.9%*

3.5 PAM50 Contextual Alignment

Partition-level agreement between unsupervised clusters and PAM50 molecular subtypes is near zero for both models (DAE: ARI = −0.001, NMI = 0.004; GR-DAE: ARI = 0.031, NMI = 0.039; Table 10), as expected. PAM50 is a supervised transcriptional classifier derived from 50 RNA expression markers Parker et al. (2009), whereas our embeddings integrate four omics modalities through a PPI interaction lens without any clinical supervision.

Despite the near-zero ARI and NMI, chi-square testing on the 5×5 cluster $\times$ PAM50 contingency table reveals a striking difference in statistical dependence. The DAE shows no significant dependence between cluster membership and PAM50 subtype ($\chi^2 = 12.496$, $p = 0.709$). In contrast, the GR-DAE shows very strong dependence ($\chi^2 = 108.711$, $p = 7.9 \times 10^{-16}$): PAM50 subtypes are highly non-randomly distributed across the PPI-informed clusters, with specific subtypes systematically enriched in particular latent groups.

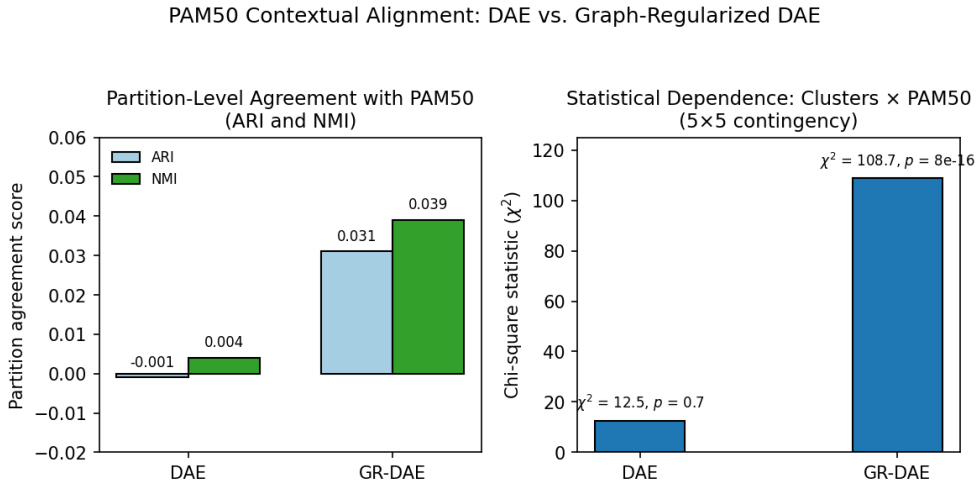


Fig. 8: PAM50 contextual alignment. **(a)** Partition-level agreement (ARI and NMI) between k-means clusters ($k = 5$) and PAM50 molecular subtypes for the DAE and GR-DAE embeddings. **(b)** Chi-square statistics for the 5×5 cluster $\times$ PAM50 contingency table. Both models show near-zero ARI and NMI, reflecting different supervision assumptions, but only the GR-DAE exhibits highly significant statistical dependence ($\chi^2 = 108.7$, $p = 7.9 \times 10^{-16}$), indicating that PPI-informed clusters capture complementary molecular structure relative to PAM50 subtypes.

4 Discussion

The central finding of this study is clear and experimentally grounded within the TCGA-BRCA cohort: incorporating PPI-based structural inductive bias through a

Table 10: PAM50 contextual alignment. Partition-level agreement (ARI, NMI) between k-means clusters ($k = 5$) and PAM50 molecular subtypes is near zero for both models, as expected given the different supervision assumptions, but only the GR-DAE shows highly significant chi-square dependence, indicating a non-random, complementary relationship with PAM50 subtyping. $\star$ Highly significant.

Model	**ARI**	**NMI**	χ^2	p**-value**
DAE	-0.001	0.004	12.496	0.709
GR-DAE	0.031	0.039	$108.711^{\star}$	7.9×10^{-16} *

single Laplacian smoothness regularisation term measurably improves several aspects of unsupervised multi-omics patient representation quality. The improvements in stability, latent geometry, clinical encoding, and contextual alignment with PAM50 are large, statistically robust, and mechanistically interpretable.

4.1 Why Laplacian Regularisation Stabilises Representations

The substantial improvement in across-restart ARI (0.861 $\rightarrow$ 0.980) and the marked reduction in reconstruction variance across seeds for the GR-DAE have a coherent mechanistic explanation grounded in the theory of inductive bias Mitchell (1980); Shalev-Shwartz and Ben-David (2014). The Laplacian smoothness term constrains the hypothesis space available to the encoder: rather than searching freely over all possible gene latent configurations, the optimiser is guided toward solutions in which PPI-connected genes occupy similar latent positions. This structural constraint narrows the loss landscape, reducing the number and diversity of degenerate local minima accessible to stochastic gradient descent and yielding a more deterministic mapping from molecular profiles to patient embeddings that converges to essentially the same configuration regardless of initialisation.

Importantly, this effect is achieved at $\lambda = 10^{-6}$, an extremely weak regularisation that adds negligible reconstruction overhead (validation MSE ≈ 0.0081 without regularisation vs. 0.0087 with). The grid search over λ reveals a non-monotone relationship between regularisation strength and stability: scores improve from the unregularised GR-DAE ($\lambda = 0$) up to $\lambda = 10^{-6}$ and remain high for $\lambda \leq 10^{-4}$, before degrading at larger values (e.g. $\lambda = 10^{-2}$). This pattern indicates that the PPI prior acts as a gentle structural nudge: strong enough to eliminate much of the landscape degeneracy, but too weak at the optimal value to override the reconstruction signal that drives meaningful feature learning.

4.2 Biological Interpretation: PPI Topology and Metastatic Status Encoding

The emergence of metastatic status (M classification) as the dominant organising variable, encoded across 60.9% of embedding dimensions in the GR-DAE vs. 0% in the DAE, is consistent with a mechanistic explanation rooted in PPI network biology. Metastatic processes in breast cancer involve coordinated dysregulation

of signalling pathways governing cell adhesion, invasion, epithelial-to-mesenchymal transition (EMT), and immune evasion Steeg (2006); Thiery et al. (2009), and these processes are characterised by network-wide molecular changes rather than isolated single-gene effects. A Laplacian regularisation term that penalises latent dissimilarity between PPI-connected genes can naturally amplify such distributed, network-coherent signals, providing a plausible biological mechanism linking the regularisation objective to the observed clinical associations.

Interpreting M-status associations in TCGA-BRCA requires appropriate caution. The cohort is predominantly M0 (730 of 865 patients with non-missing M status), and the MX designation (117 patients; not assessable) introduces classification ambiguity that may blur true metastatic signal. Future work should validate M-status encoding on cohorts with systematically assigned staging and more complete metastatic annotation. We also cannot rule out residual technical or cohort-specific factors that correlate with recorded M status and are amplified by the PPI prior, so the observed associations should be regarded as hypothesis-generating rather than definitive evidence of causal metastatic programmes.

4.3 Complementarity with PAM50 Subtyping

The combination of near-zero ARI (0.031) with extremely significant chi-square dependence ($p = 7.9 \times 10^{-16}$) between GR-DAE clusters and PAM50 subtypes reveals an important nuance in how these representations relate. PAM50 is a supervised RNA-based classifier trained to discriminate five transcriptional archetypes Parker et al. (2009), whereas our embeddings integrate four omics modalities through a PPI interaction lens without any supervision. The near-zero ARI confirms that these representations do not produce the same partition, which is expected given their substantially different information sources and objectives. The strong chi-square dependence establishes that PAM50 subtypes are non-randomly distributed across GR-DAE clusters, revealing shared but non-redundant molecular organisation. From a precision oncology perspective, this suggests that PPI-informed multi-omics embeddings capture a complementary layer of tumour biology that cross-cuts transcriptional subtype boundaries and may identify clinically relevant patient subgroups not captured by gene expression alone Cancer Genome Atlas Network (2012). In this sense, our goal is not to reproduce PAM50 partitions, but to provide an unsupervised representation that complements established expression-based subtyping while remaining statistically linked to it. The near-zero ARI and NMI should therefore be interpreted as a reflection of these differing modelling goals rather than as evidence that the learned representation fails to capture known biology.

4.4 Practical Implications for Representation Design

These findings carry direct practical implications for the design of multi-omics patient embedding systems. First, PPI-based Laplacian regularisation can be considered a low-cost, high-impact default: no architectural modification is required, computational overhead is negligible, and stability gains are robust across training seeds. Second, representation stability across independent training runs (not just performance on a

single run) should be adopted as a mandatory evaluation criterion for unsupervised clinical representations, because unstable embeddings can lead to inconsistent clinical conclusions even when downstream models appear performant. Third, the gene-wise architecture of the GR-DAE, with shared weights across genes and mean pooling, provides a natural pathway to biological interpretability: individual gene-level latent representations can be inspected to identify which PPI subnetworks most strongly encode specific clinical variables such as metastatic status.

4.5 Limitations and Future Directions

Several limitations qualify the scope of these conclusions. All experiments were conducted on a single cohort (TCGA-BRCA, $N = 941$), so external validation on independent datasets such as METABRIC or ICGC-BRCA is needed to establish generalisability of the stability gains and clinical associations.

Second, the GR-DAE combines a gene-wise encoder architecture with Laplacian regularisation. Even the unregularised GR-DAE variant ($\lambda = 0$) shows substantial restart stability (mean ARI = 0.934, geometry Spearman $\rho = 0.983$), so our experiments do not fully disentangle the respective contributions of architectural inductive bias and the explicit graph prior; a systematic comparison of flattened and gene-wise architectures at $\lambda = 0$ and $\lambda > 0$ is an important direction for future work.

Third, the PPI graph is static and context-independent, aggregating interactions from multiple evidence sources without tissue-type or disease-state specificity; breast cancer-specific or context-aware interaction priors could provide more targeted inductive bias. Receptor status (ER, PR, HER2) and some other clinical variables were unavailable or incomplete for post-hoc evaluation, limiting the interpretability of clinical associations.

Future work should: (i) validate these findings on independent cohorts with longer follow-up and systematically assigned staging; (ii) investigate the effect of PPI confidence threshold and network density on stability and clinical encoding; (iii) combine Laplacian regularisation with breast cancer-specific or context-aware interaction networks; (iv) extend per-gene interpretability analyses to identify the PPI subnetworks most strongly encoding M-status variation; and (v) assess whether similar stability and encoding benefits generalise to other cancer types with available multi-omics profiles Reimand et al. (2019).

5 Conclusions

This study demonstrates that a single PPI-based Laplacian smoothness regularisation term, on top of a denoising autoencoder with matched patient-level latent dimensionality, is sufficient to substantially transform the quality of unsupervised multi-omics patient representations. On the TCGA-BRCA cohort, adding this term on top of a denoising autoencoder improves across-restart clustering stability from ARI 0.861 to 0.980, yields near-perfect latent geometry preservation (Spearman ρ from 0.915 to 0.997), reduces the effective dimensionality from 13 to 3 principal components at a 90% variance threshold, and induces M-status encoding across 60.9% of embedding

dimensions compared with 0% in the unstructured baseline. At the same time, GR-DAE clusters show highly significant non-random PAM50 alignment ($\chi^2 = 108.7$, $p = 7.9 \times 10^{-16}$ vs. $p = 0.709$ for the DAE) despite near-zero partition-level agreement, indicating that PPI-informed embeddings capture a complementary layer of tumour organisation. All of these gains are achieved at $\lambda = 10^{-6}$, selected using purely unsupervised criteria, with negligible computational overhead.

Within the scope of this cohort-specific analysis, these results suggest two directions for future representation design: (i) PPI-based Laplacian regularisation should be considered a low-cost, high-impact component of unsupervised representation learning pipelines, and (ii) across-restart representation stability should be treated as a mandatory evaluation criterion alongside downstream task performance. Stable, biologically structured patient embeddings are a necessary precondition for reproducible and clinically interpretable unsupervised stratification.

Acknowledgements. The authors thank the TCGA Research Network for making the BRCA multi-omics data publicly available. The authors also gratefully acknowledge Dr. Adhari Al-Zaabi, Dr. Abdul Rahman Aal Abdulsalam, and Dr. Sirin A. Yaeesh for their valuable discussions and feedback during the course of this work.

Declarations

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflicts of interest.

Ethics approval and consent to participate. This study uses publicly available, de-identified data from The Cancer Genome Atlas (TCGA). No ethics approval was required.

Consent for publication. Not applicable.

Data availability. TCGA-BRCA data are available at https://portal.gdc.cancer.gov (accessed 01 February 2026). STRING PPI data are available at https://string-db.org (v11, accessed 15 February 2026). Code will be deposited upon acceptance.

Author contribution. Conceptualisation, H.E.K. and H.Z.; methodology, H.E.K. and M.B.; formal analysis, H.E.K. and M.B.; data curation, H.E.K.; writing—original draft, H.E.K.; writing—review and editing, H.Z. and M.B.; supervision, H.Z. All authors have read and agreed to the published version of the manuscript.

References

Anis, A., El Gamal, A., Avestimehr, A.S., Ortega, A.: A sampling theory perspective of graph-based semi-supervised learning. IEEE Transactions on Information Theory **65**(4), 2322–2342 (2019)

Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013) https://doi.org/10.1109/TPAMI.2013.50

Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12**(1), 56–68 (2011) https://doi.org/10.1038/nrg2918

Chen, L., *et al.*: glmgraph: an R package for variable selection and predictive modelling of structured genomic data. Bioinformatics **31**(24), 3991–3993 (2015)

Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012) https://doi.org/10.1038/nature11412

Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics **18**(9), 551–562 (2017) https://doi.org/10.1038/nrg.2017.38

Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.X.: Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research **24**(6), 1248–1259 (2018) https://doi.org/10.1158/1078-0432.CCR-17-0853

Cottrell, S., Wang, R., Wei, G.-W.: PLPCA: Persistent Laplacian-enhanced PCA for microarray data analysis. Journal of Chemical Information and Modeling **64**(7), 2405–2420 (2024)

Feng, C.-M., *et al.*: PCA via joint graph Laplacian and sparse constraint. BMC Bioinformatics **20**(Suppl 22), 716 (2019)

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015) https://doi.org/10.1038/nature14539

Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., Zhu, Y.: MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. Frontiers in Genetics **13**, 806842 (2022) https://doi.org/10.3389/fgene.2022.806842

Lemsara, A., Ouadfel, S., Fröhlich, H.: PathME: Pathway based multimodal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinformatics **21**(1), 146 (2020)

Liu, B., Rosenhahn, B., Illig, T., DeLuca, D.S.: A variational autoencoder trained with priors from canonical pathways increases the interpretability of transcriptome data. PLOS Computational Biology **20**(7), 1011198 (2024)

Mitchell, T.M.: The need for biases in learning generalizations. Rutgers University Department of Computer Science Technical Report (CBM-TR-117) (1980)

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., Bernard, P.S.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal

of Clinical Oncology **27**(8), 1160–1167 (2009) https://doi.org/10.1200/JCO.2008.18.1370

Perou, C.M., Sørlie, T., Eisen, M.B., Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S., Lønning, P.E., Børresen-Dale, A.-L., Brown, P.O., Botstein, D.: Molecular portraits of human breast tumours. Nature **406**(6797), 747–752 (2000) https://doi.org/10.1038/35021093

Ronen, J., Akalin, A.: netSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Research **7**, 8 (2018)

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., Bader, G.D.: Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols **14**(2), 482–517 (2019) https://doi.org/10.1038/s41596-018-0103-9

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J., Mering, C.: The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research **49**(D1), 605–612 (2021) https://doi.org/10.1093/nar/gkaa1074

Sun, H., Li, H.: A bayesian approach for graph-constrained estimation for high-dimensional regression. International Journal of Systems and Synthetic Biology **1**(2), 255–272 (2010)

Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, ??? (2014)

Schulte-Sasse, R., Budach, S., Hnisz, D., Marsico, A.: Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nature Machine Intelligence **3**(6), 513–526 (2021) https://doi.org/10.1038/s42256-021-00333-z

Steeg, P.S.: Tumor metastasis: mechanistic insights and clinical challenges. Nature Medicine **12**(8), 895–904 (2006) https://doi.org/10.1038/nm1469

Thiery, J.P., Acloque, H., Huang, R.Y.J., Nieto, M.A.: Epithelial–mesenchymal transitions in development and disease. Cell **139**(5), 871–890 (2009) https://doi.org/10.1016/j.cell.2009.11.007

Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V., Waddell, N.: Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine **13**(1), 152 (2021) https://doi.org/10.1186/s13073-021-00966-x

Wang, Z., Xu, W., San Lucas, F.A., Liu, Y.: Incorporating prior knowledge into gene network study. Bioinformatics **29**(20), 2633–2640 (2013)

Zhang, W., *et al.*: Molecular pathway identification using biological network-regularized logistic models. BMC Genomics **14**(Suppl 8), 7 (2013)