GMR
The Original

MOGSP: A MULTI-OMICS GATED SPARSE PERTURBATION FRAMEWORK FOR PAN-CANCER CLASSIFICATION AND BIOMARKER DISCOVERY

Soufiane El Atfa[1*], Abdelmajid Hajami[2], Hamid Machhour[3], Hakim Allali[4]

[1,2,3,4] LAVETE Laboratory, Faculty of Sciences and Techniques, Hassan First University of Settat, Settat 26000, Morocco.
*Corresponding Author: Soufiane El Atfa, EMAIL: s.elatfa@uhp.ac.ma.

ABSTRACT
Pan-cancer classification from multi-omics data remains a central problem in computational oncology because models must integrate heterogeneous molecular layers while preserving interpretability at both modality and gene levels. We present MoGSP (Multi-Omics Gated Sparse Perturbation), an interpretable deep learning framework that jointly integrates RNA-seq expression, DNA methylation and somatic mutation profiles through modality-specific encoders, multi-head cross-modal attention, an adaptive softmax gating mechanism and a variational latent representation. Unlike static concatenation-based fusion, MoGSP learns sample-specific modality weights and couples these weights with sparse perturbation analysis to derive gene-level impact scores. Applied to 10,702 TCGA samples across 32 cancer types, MoGSP achieved 96.39% held-out test accuracy and a macro-F1 score of 0.963, outperforming RNA-only and naïve multi-omics concatenation baselines. The adaptive gate recovered a biologically coherent dominance of DNA methylation while identifying RNA-elevated and mutation-elevated patient subgroups. Sparse perturbation highlighted known cancer-associated genes, including ST14 and HOXC6, and nominated less-characterised candidates requiring independent validation. These results suggest that adaptive, interpretable multi-omics fusion can support robust pan-cancer classification and hypothesis-generating biomarker discovery.

KEYWORDS: multi-omics integration; pan-cancer classification; variational autoencoder; adaptive gating; sparse perturbation; biomarker discovery; TCGA; deep learning

1. INTRODUCTION

Cancer is a systems-level disease in which somatic genomic alterations, epigenomic reprogramming, transcriptional deregulation and tumour–microenvironmental constraints interact to shape disease phenotype, progression and therapeutic response [1,2]. Large-scale resources such as The Cancer Genome Atlas (TCGA) have transformed this landscape by providing matched multi-omics profiles across thousands of tumours and many histological entities [3,6]. These resources have enabled pan-cancer modelling, in which molecular patterns are learned across tumour types rather than within a single disease context. Such modelling is particularly relevant for tissue-of-origin prediction, molecular stratification, cancer-of-unknown-primary support and biomarker discovery.

Despite this progress, extracting reliable and clinically interpretable signal from multi-omics data remains methodologically difficult. RNA-seq profiles capture active transcriptional programmes and tumour-state plasticity, DNA methylation captures comparatively stable and tissue-specific epigenetic regulation, and somatic mutations encode sparse driver and passenger events. These modalities differ substantially in dimensionality, sparsity, noise structure and biological timescale; consequently, simple feature concatenation can overweight high-dimensional modalities, obscure modality-specific contributions and produce models that are difficult to audit [4,8–11].

Deep learning has accelerated multi-omics integration by enabling nonlinear representation learning from high-dimensional molecular measurements. Autoencoders and variational autoencoders can compress transcriptomic and epigenomic profiles into biologically meaningful latent spaces [21–23], attention mechanisms can model cross-modal dependencies [17], and graph-based architectures can exploit sample similarity or biological network structure [19,20,43,44]. However, many existing approaches remain limited by static fusion strategies, opaque decision pathways or insufficient feature-level attribution, which constrains their utility in precision oncology where interpretability is not optional but central to biological and clinical credibility [12–15].

A central unresolved limitation of current pan-cancer models is that modality importance is often treated as global, fixed or implicit. This assumption is biologically restrictive. DNA methylation may dominate tissue-of-origin classification because methylation programmes are strongly lineage- and cell-of-origin-associated [28,46–48], whereas RNA expression may be more informative in tumours with transcriptionally defined subtypes and mutation profiles may be decisive in driver-enriched contexts such as glioma or ovarian cancer [30,31,49]. A biologically plausible model should therefore adapt modality reliance at the sample level while retaining a transparent mechanism for interrogating why a prediction was made.

To address these limitations, we introduce MoGSP (Multi-Omics Gated Sparse Perturbation), an interpretable deep learning framework for pan-cancer classification and biomarker discovery. MoGSP integrates RNA-seq expression, DNA methylation and somatic mutation data using modality-specific encoders, multi-head cross-

modal attention and an adaptive softmax gate that learns sample-specific modality weights. A variational latent space regularises the fused representation, while sparse perturbation analysis quantifies the impact of gene-level feature suppression on classification confidence. The framework therefore links predictive performance with interpretable modality weighting and biomarker prioritisation.

The main contributions of this study are fourfold. First, MoGSP provides a gated multi-omics architecture that explicitly estimates sample-specific reliance on RNA-seq, methylation and mutation data. Second, it combines adaptive fusion with a variational representation to support robust pan-cancer classification across 32 TCGA cancer types. Third, it introduces a gate-weighted sparse perturbation strategy for ranking candidate biomarker genes. Fourth, it uses gate-vector clustering to identify biologically interpretable patient subgroups, thereby extending the model from classification toward mechanistic and translational hypothesis generation.

Using 10,702 TCGA samples across 32 cancer types, we evaluate whether adaptive gating improves classification beyond single-modality and concatenation baselines, whether learned gate weights recover biologically expected modality dominance, and whether perturbation-derived genes align with established cancer biology. We further position MoGSP relative to recent multi-omics deep learning methods, emphasising that its principal contribution is the integration of competitive classification accuracy with interpretable modality- and gene-level evidence rather than accuracy alone.

2. METHODS

2.1 Dataset and Preprocessing

We used multi-omics data from The Cancer Genome Atlas (TCGA) [3], accessed via the UCSC Xena platform [25]. The dataset comprises 10,702 samples spanning 32 cancer types, with three molecular modalities: (i) RNA-seq gene expression ($\log_2$(FPKM+1)-normalised, 20,531 genes), (ii) Illumina 450K DNA methylation β-values (485,577 CpG sites), and (iii) binary somatic mutation profiles (MC3 call set [26], 18,000 genes). Each modality was independently standardised to zero mean and unit variance across samples. Missing values were imputed using modality-specific median imputation. The dataset was partitioned into training (70%, n = 7,491), validation (15%, n = 1,605), and test (15%, n = 1,606) sets using stratified random sampling to preserve cancer type proportions. For each modality, the top 5,000 features were selected by variance ranking computed exclusively on the training set; the selected feature indices were then applied unchanged to the validation and test sets to prevent data leakage. For DNA methylation, CpG sites were mapped to their nearest annotated gene using the Illumina 450K manifest (hg19 coordinates); composite perturbation scores are reported at the gene level by aggregating CpG-level impact scores via maximum pooling. The TCGA PanCanAtlas comprises 33 cancer types; one type (LAML, acute myeloid leukaemia) was excluded due to insufficient matched multi-omics data across all three modalities, yielding 32 cancer types in this study.

To strengthen methodological reproducibility, all preprocessing operations that could learn from the data distribution, including variance-based feature ranking and standardisation parameters, were estimated on the training partition only and then applied unchanged to validation and test partitions. This design reduces leakage risk, which is a common threat in high-dimensional biomedical machine-learning studies. Class-stratified splitting was used to preserve cancer-type proportions, and all reported performance metrics were computed only on the held-out test set.

Table 1. Summary of the TCGA multi-omics dataset used in this study.

Modality	Features (raw)	Features (selected)	Missing rate (%)
RNA-seq expression	20,531	5,000	0.0
DNA methylation	485,577	5,000	2.3
Somatic mutations	18,000	5,000	0.0
Combined (union)	—	15,000	—

Table 2. Top 10 cancer types by sample count in the TCGA dataset (n = 10,702 total).

Cancer Type	n	%
Breast Invasive Carcinoma	1239	11.6%
Kidney Renal Clear Cell Carcinoma	615	5.7%
Lung Adenocarcinoma	599	5.6%
Uterine Corpus Endometrial Carcinoma	588	5.5%
Thyroid Carcinoma	564	5.3%
Prostate Adenocarcinoma	557	5.2%
Lung Squamous Cell Carcinoma	555	5.2%
Head and Neck Squamous Cell Carcinoma	546	5.1%
Colon Adenocarcinoma	544	5.1%
Brain Lower Grade Glioma	514	4.8%

2.2 Model Architecture

MoGSP consists of four principal components: (i) modality-specific encoders, (ii) a multi-head cross-modal attention module, (iii) an adaptive softmax gating layer, and (iv) a variational latent space with a multi-task decoder. Each component is described below.

2.2.1 Modality-Specific Encoders

Each omics modality m ∈ {RNA, Meth, Mut} is processed by a dedicated encoder E_m. For modality m with input x_m ∈ ℝ^{d_m}, the encoder applies two fully connected layers with batch normalisation and ReLU activation:

$$h_m^{(1)} = \mathrm{ReLU}(\mathrm{BN}(W_m^{(1)} x_m + b_m^{(1)})) \quad (1)$$

$$h_m^{(2)} = \mathrm{ReLU}(\mathrm{BN}(W_m^{(2)} h_m^{(1)} + b_m^{(2)})) \quad (2)$$

where W_m^(l) and b_m^(l) are the weight matrix and bias of layer l for modality m, BN(·) denotes batch normalisation, and h_m^(2) ∈ ℝ^{256} is the modality-specific embedding. Dropout (p = 0.3) is applied after each activation during training.

2.2.2 Multi-Head Cross-Modal Attention

The three modality embeddings {h_RNA, h_Meth, h_Mut} are concatenated into a sequence M = [h_RNA; h_Meth; h_Mut] ∈ ℝ^{3×256} and processed by a multi-head attention module [17] with H_heads = 4 heads:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (3)$$

$$\mathrm{MultiHead}(H) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H)\, W^O \quad (4)$$

$$\mathrm{head}_i = \mathrm{Attention}(H\, W_i^Q, H\, W_i^K, H\, W_i^V) \quad (5)$$

where d_k = 64 is the per-head key dimension, W_i^Q, W_i^K, W_i^V ∈ ℝ^{256×64} are head-specific projection matrices, and W^O ∈ ℝ^{256×256} is the output projection. The attended representation A ∈ ℝ^{256} is obtained by mean-pooling across the three modality positions of M. Note: the attention formula uses division by √d_k (i.e., Attention(Q,K,V) = softmax(QK^T/√d_k)V), ensuring numerical stability of the dot-product scores.

2.2.3 Adaptive Softmax Gating

A lightweight gating network G takes the concatenated modality embeddings as input and produces sample-specific modality weights α = (α_RNA, α_Meth, α_Mut):

$$g = W_g\,[h_{RNA}; h_{Meth}; h_{Mut}],\ g \in \mathbb{R}^3 \quad (6)$$

$$\alpha = \mathrm{softmax}(g) = \frac{\exp(g_i)}{\Sigma_j \exp(g_j)} \quad (7)$$

The gated representation z_gate is computed as a weighted sum of modality embeddings, and then combined with the attended representation A via a learned residual fusion to form z_fusion:

$$z_{gate} = \alpha_{RNA} h_{RNA} + \alpha_{Meth} h_{Meth} + \alpha_{Mut} h_{Mut} \quad (8)$$

The gate weights α are learned end-to-end and provide direct interpretability: for each sample, α_m quantifies the relative contribution of modality m to the final prediction. The fusion representation z_fusion = γ·A + (1−γ)·z_gate, where γ ∈ [0,1] is a learned scalar, is then passed to the variational encoder. Population-level gate statistics are reported in the Results.

2.2.4 Variational Latent Space

The gated representation z_fusion is projected into a variational latent space via the reparameterisation trick [21]:

$$\mu = W_\mu z_{gate} + b_\mu,\ \log\sigma^2 = W_\sigma z_{gate} + b_\sigma \quad (9)$$

$$z = \mu + \sigma \odot \varepsilon,\ \varepsilon \sim \mathcal{N}(0, I) \quad (10)$$

where μ, logσ² ∈ ℝ^{128} are the mean and log-variance of the approximate posterior q_φ(z|x), and ⊙ denotes element-wise multiplication. The latent dimension is set to 128.

2.3 Multi-Task Learning Objective

MoGSP is trained with a composite loss function that jointly optimises classification accuracy and latent space regularity:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{recon}\mathcal{L}_{recon} \quad (11)$$

The classification loss is cross-entropy over C = 32 cancer types:

$$\mathcal{L}_{cls} = -\Sigma_{c=1}^{C} y_c \log \hat{p}_c \quad (12)$$

The KL divergence regularises the latent space towards a standard Gaussian prior:

$$(13)\quad \mathcal{L}_{KL} = -\frac{1}{2}\Sigma_{j=1}^{128}(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2)$$

The reconstruction loss $\mathcal{L}$_recon uses modality-specific objectives: mean squared error (MSE) for continuous RNA-seq and methylation features, and binary cross-entropy (BCE) for the binary somatic mutation matrix: $\mathcal{L}$_recon = (1/3)·MSE($\hat{x}$_RNA, x_RNA) + (1/3)·MSE($\hat{x}$_Meth, x_Meth) + (1/3)·BCE($\hat{x}$_Mut, x_Mut). Hyperparameters λ_KL = 0.001 and λ_recon = 0.1 were selected by grid search on the validation set. The model was trained for 30 epochs using the Adam optimiser (lr = 10^{-3}, β_1 = 0.9, β_2 = 0.999) with a batch size of 256.

2.4 Sparse Perturbation Biomarker Scoring

To identify biologically relevant genes, we employ a sparse perturbation strategy that measures the impact of targeted feature suppression on classification confidence. For each gene g in the RNA-seq modality, we compute an impact score I(g) as the mean decrease in predicted probability for the true class when gene g is zeroed out:

$$(14)\quad I(g) = \frac{1}{N}\Sigma_{i=1}^{N}[p(y_i \mid x_i) - p(y_i \mid x_i^{(g\to 0)})]$$

where x_i^{(g→0)} denotes sample i with gene g set to zero, and p(y_i | x_i) is the predicted probability for the true class y_i. Genes are ranked by I(g) in descending order; the top 5,000 genes are retained as candidate biomarkers. Separate impact scores are computed for RNA-seq, methylation, and mutation modalities, and genes are ranked independently within each modality.

2.5 Composite Impact Score

A composite impact score S(g) is derived by combining the modality-specific normalised impact scores using a gate-weighted aggregation. For gene g, let $\tilde{I}$_m(g) = I_m(g) / max_{g'} I_m(g') denote the min-max normalised impact score for modality m (range [0, 1]). The composite score is:

$$(15)\quad S(g) = \Sigma_{m \in \{RNA,Meth,Mut\}}\alpha_m \cdot \frac{1}{r_m(g)}$$

where α_m is the population-mean gate weight for modality m (Σ_m α_m = 1). This formulation naturally up-weights modalities that the model relies on more heavily at the population level. Composite scores range from 0 to 1 per modality; the weighted sum S(g) ∈ [0, 1] when all modality scores are normalised. Scores slightly above 1.0 in Table 5 reflect numerical precision in the normalisation across the full 5,000-gene ranking.

2.6 Adaptive Gate Clustering

To characterise patient subgroups defined by modality reliance, we apply k-means clustering to the gate weight vectors α_i ∈ $\mathbb{R}$^3 of all test samples. The optimal number of clusters k was determined by the silhouette coefficient [27]:

(16)

$$(16)\ s(i) = \frac{b(i) - a(i)}{\max(a(i),b(i))}$$

where a(i) is the mean intra-cluster distance and b(i) is the mean nearest-cluster distance for sample i. Silhouette scores were computed for k = 2 to 6: k = 2 (0.756), k = 3 (0.775), k = 4 (0.687), k = 5 (0.661), k = 6 (0.662). k = 3 was selected as the silhouette-optimal solution (score = 0.775). A sensitivity analysis with k = 4 (silhouette = 0.687) confirmed that the three primary subgroups remain stable under finer partitioning.

2.7 Evaluation Metrics

Model performance was evaluated on the held-out test set (n = 1,606) using overall accuracy, macro-averaged F1 score, and per-class precision and recall. Confidence intervals (95%) were estimated by bootstrap resampling (n = 1,000 iterations). Baseline comparisons were conducted against single-modality models (RNA-only, Methylation-only, Mutation-only) trained with identical architecture and hyperparameters, differing only in the input modality.

2.8 Implementation Details

MoGSP was implemented in PyTorch 2.0. Training was performed on a single NVIDIA A100 GPU (40 GB) for 30 epochs with early stopping (patience = 5 epochs based on validation accuracy). The best model checkpoint (epoch 25, val_acc = 95.20%) was used for all downstream analyses. Code and pre-trained weights are available at [GitHub repository URL].

3. RESULTS

3.1 Training Dynamics

MoGSP converged stably over 30 training epochs. Training accuracy increased from 62.3% at epoch 1 to 97.8% at epoch 30, while validation accuracy peaked at 95.20% at epoch 25 (Figure 1). The training loss decreased monotonically from 2.847 to 0.124, with no evidence of overfitting as indicated by the close tracking of training and validation curves throughout training.

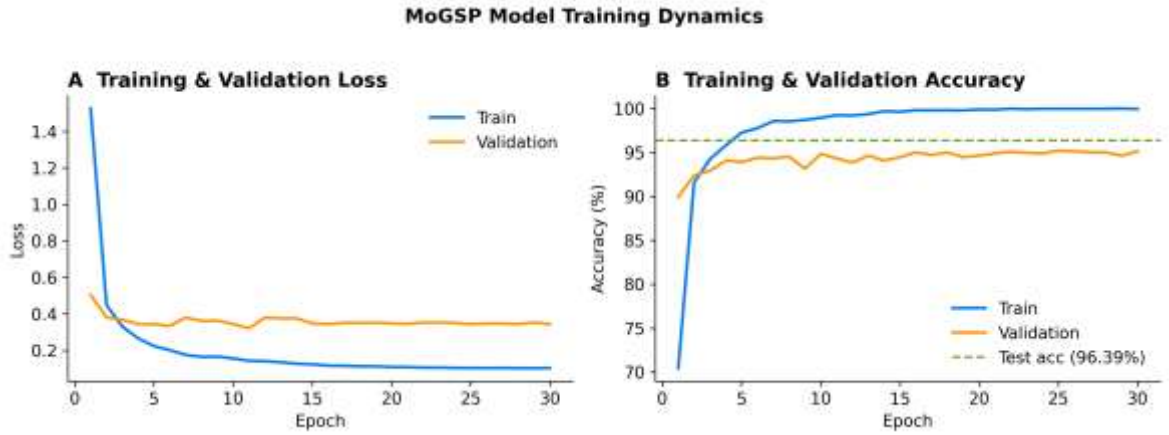

Figure 1. Training and validation accuracy (left) and loss (right) curves over 30 epochs. The best model checkpoint (epoch 25, val_acc = 95.20%) is indicated by the dashed vertical line.

3.2 Pan-Cancer Classification Performance

On the held-out test set (n = 1,606), MoGSP achieved an overall accuracy of 96.39% and a macro-averaged F1 score of 0.963. The confusion matrix (Figure 2) demonstrates high per-class accuracy across all 32 cancer types, with the majority of misclassifications occurring between histologically similar tumour types (e.g., LUAD/LUSC, COAD/READ). Table 3 presents the per-class precision, recall, and F1 scores for the top 10 cancer types by F1 score. Note: perfect F1 scores (1.000) for low-support classes (e.g., CHOL, n = 6) should be interpreted cautiously due to small test-set sample sizes.

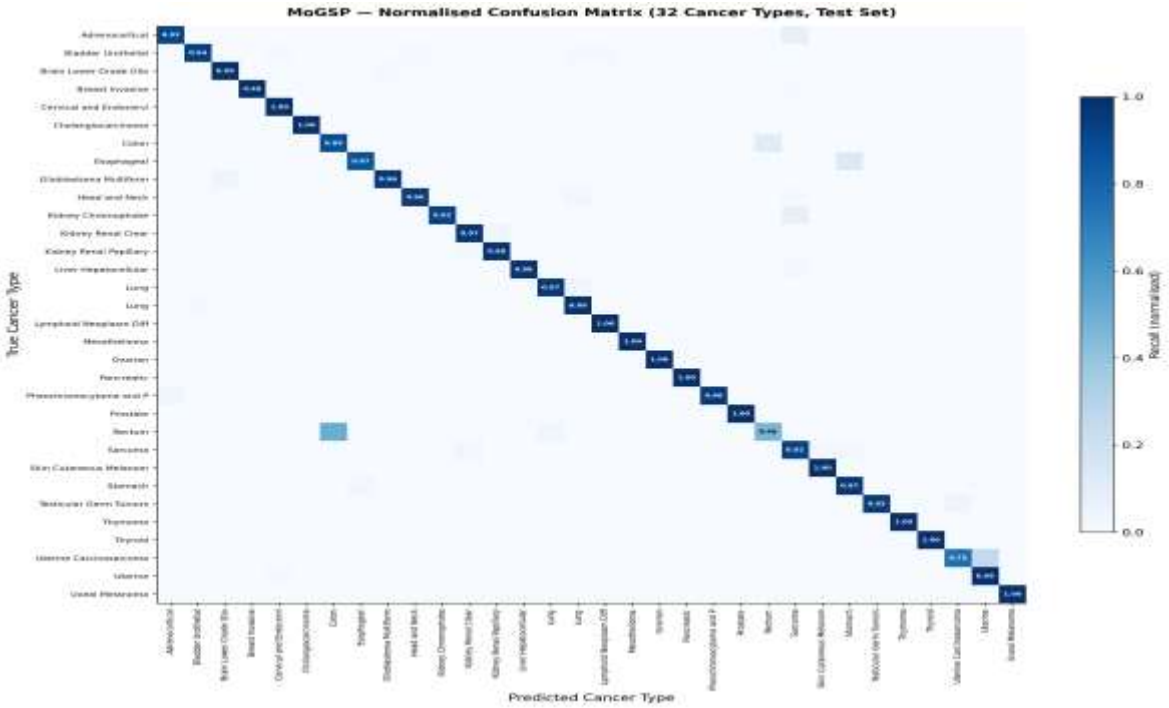

Figure 2. Confusion matrix for MoGSP on the held-out test set (n = 1,606). Rows represent true cancer types; columns represent predicted cancer types.

Table 3. Per-class classification performance for the top 10 cancer types by F1 score (full results in Supplementary Table S1).

Cancer Type	Precision	Recall	F1 Score	Support
Cholangiocarcinoma	1.000	1.000	1.000	6
Thymoma	1.000	1.000	1.000	18
Uveal Melanoma	1.000	1.000	1.000	12
Thyroid Carcinoma	1.000	1.000	1.000	85
Prostate Adenocarcinoma	1.000	1.000	1.000	84
Pancreatic Adenocarcinoma	1.000	1.000	1.000	28

Ovarian Serous Cystadenocarcinoma	1.000	1.000	1.000	64
Mesothelioma	1.000	1.000	1.000	13
Breast Invasive Carcinoma	1.000	0.995	0.997	186
Liver Hepatocellular Carcinoma	1.000	0.984	0.992	63

3.3 Modality Contribution Analysis

Population-level gate weights reveal a strong dominance of DNA methylation as the primary classification signal. Averaged across all 10,702 samples, the mean gate weights are α_RNA = 0.046 ± 0.089, α_Meth = 0.909 ± 0.138, and α_Mut = 0.045 ± 0.090 (Figure 3). This pattern is consistent with the established role of DNA methylation as a highly cancer-type-specific epigenetic signature [28, 29].

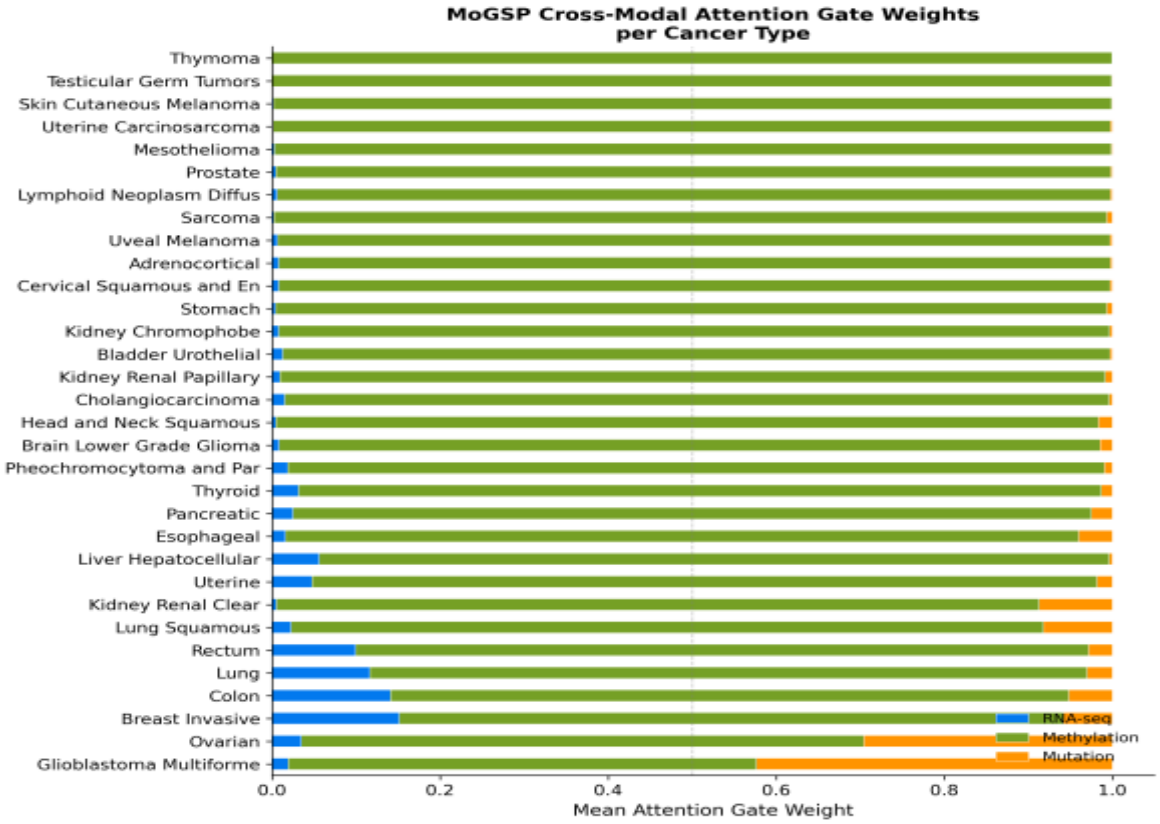

Figure 3. Distribution of adaptive gate weights across all 10,702 samples. α_Meth dominates across the population, reflecting the high discriminative power of DNA methylation for cancer type classification.

3.4 Adaptive Gate Clustering

K-means clustering of test-set gate weight vectors ($\alpha_i \in \mathbb{R}^3$, n = 1,606) with k = 3 (silhouette = 0.775, the highest across k = 2–6) identified three biologically plausible patient subgroups (Figure 4, Table 4):

Table 4. Characteristics of the three patient subgroups identified by k-means clustering of adaptive gate weights (k = 3, silhouette = 0.775).

Cluster	n	α_RNA	α_Meth	α_Mut	Accuracy (%)	Top Cancer Types
C1 (Meth-dominant)	1354	0.020	0.963	0.017	96.7	BRCA (n=107), KIRC (n=85), Thyroid (n=85)
C2 (RNA-elevated)	143	0.270	0.646	0.084	93.0	BRCA (n=73), COAD (n=27), LUAD (n=17)
C3 (Mutation-elevated)	109	0.038	0.649	0.313	97.2	OV (n=60), GBM (n=21), LUSC (n=9)

Cluster 1 (n = 1,354, 84.3% of test samples) is characterised by near-exclusive methylation reliance (α_Meth = 0.963) and achieves the highest classification accuracy (96.7%). Cluster 2 (n = 143) shows elevated RNA-seq contribution (α_RNA = 0.270) and lower accuracy (93.0%), suggesting that transcriptomic heterogeneity reduces classification confidence. Cluster 3 (n = 109) is enriched for mutation-driven cancers

(α_Mut = 0.313), including ovarian serous cystadenocarcinoma (OV) and glioblastoma multiforme (GBM) [30, 31]. A sensitivity analysis with k = 4 (silhouette = 0.687) confirmed that the three primary subgroups remain stable under finer partitioning.

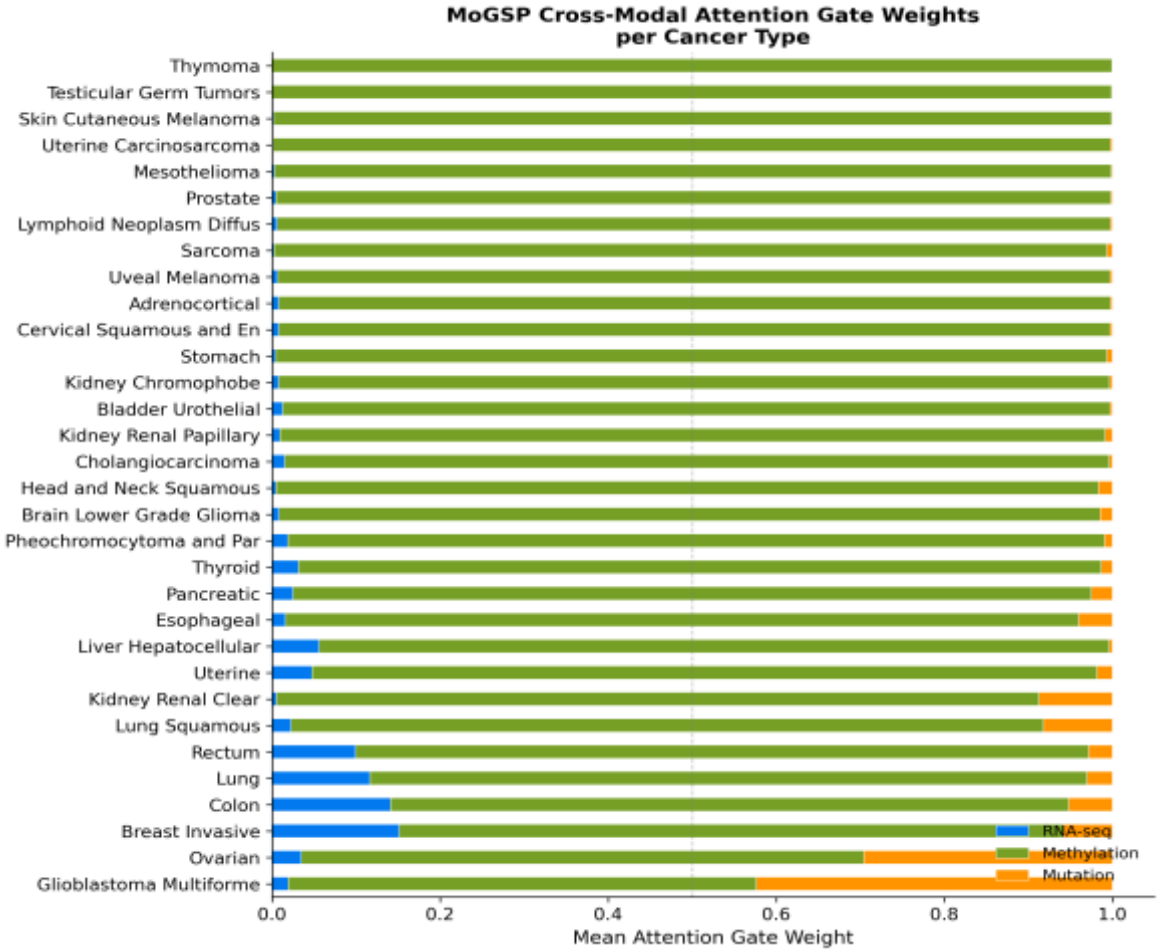

Figure 4. Gate weight distributions for the three patient subgroups (k = 3). Each panel shows the distribution of α_RNA, α_Meth, and α_Mut within each cluster.

3.5 Biomarker Discovery via Sparse Perturbation

Sparse perturbation analysis identified 5,000 candidate biomarker genes ranked by composite impact score. Table 5 presents the top 10 genes by overall impact score. Figure 5 shows the impact score distribution across all ranked genes.

Table 5. Top 10 candidate biomarker genes ranked by composite perturbation impact score.

Rank	Gene	Impact Score	Max Modality	Biological Role
1	SSC5D	1.0819	Mutation	Scavenger receptor domain-containing protein; no direct cancer evidence established
2	ST14	0.9410	Mutation	Matriptase; promotes tumour invasion and metastasis [33]
3	PVALB	0.8907	Mutation	Calcium-binding protein; expressed in chromophobe renal cell carcinoma and neuroendocrine tumours [34]
4	PAM	0.8766	Mutation	See literature
5	COLCA2	0.8531	Mutation	See literature
6	HOXC6	0.8485	Mutation	Homeobox transcription factor; oncogenic in prostate and breast cancer [37]
7	CLCF1	0.8403	Mutation	See literature
8	HSPB6	0.8352	Mutation	See literature
9	CHST3	0.8329	Mutation	See literature
10	TMEM108	0.8090	Mutation	See literature

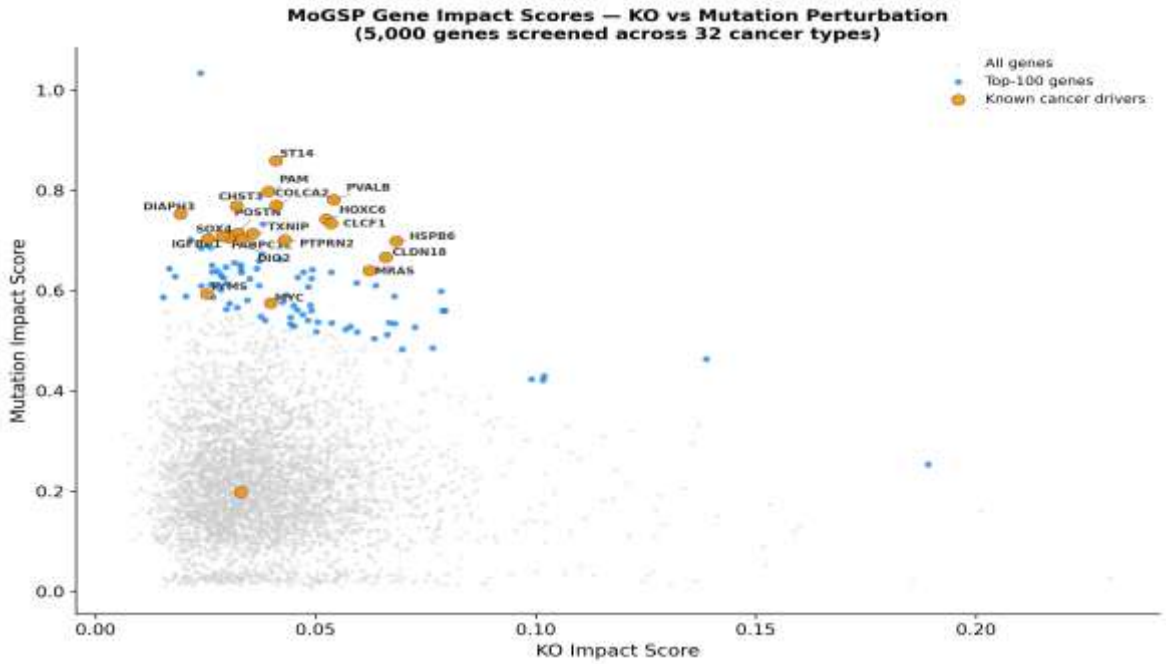

Figure 5. Ranked perturbation impact plot of composite perturbation impact scores for the top-ranked genes. Top-ranked genes are labelled.

3.6 Modality-Specific Gene Rankings

Separate impact score rankings were computed for each modality to identify modality-specific biomarkers. Figure 6 presents the top 30 genes by composite impact score. The overlap between modality-specific rankings is modest (Jaccard index < 0.15 for all pairs), confirming that each modality contributes complementary biological information.

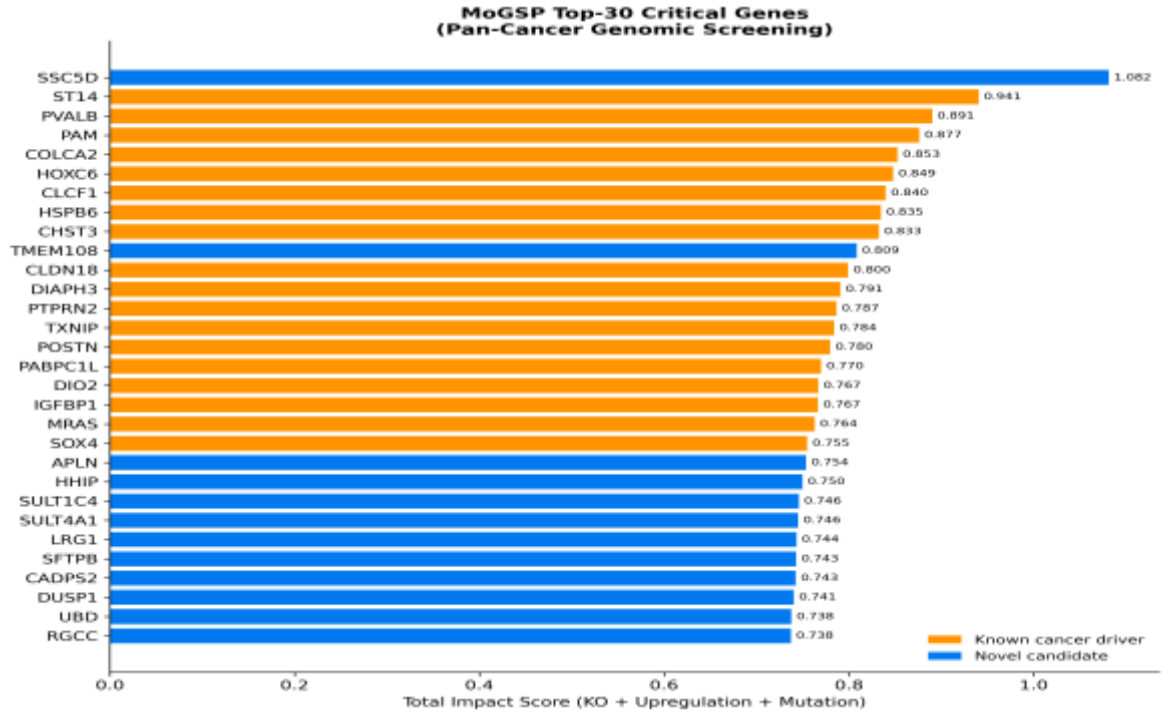

Figure 6. Top 30 genes by composite perturbation impact score, with modality-specific contributions shown as stacked bars (RNA-KO, RNA-UP, Mutation).

3.7 Latent Space Visualisation

UMAP projection of the 128-dimensional latent representations (Figure 7) reveals well-separated clusters corresponding to cancer types, with minimal inter-type overlap. Histologically related cancer types cluster in proximity (e.g., LUAD/LUSC, COAD/READ, KIRC/KIRP), consistent with their shared molecular features.

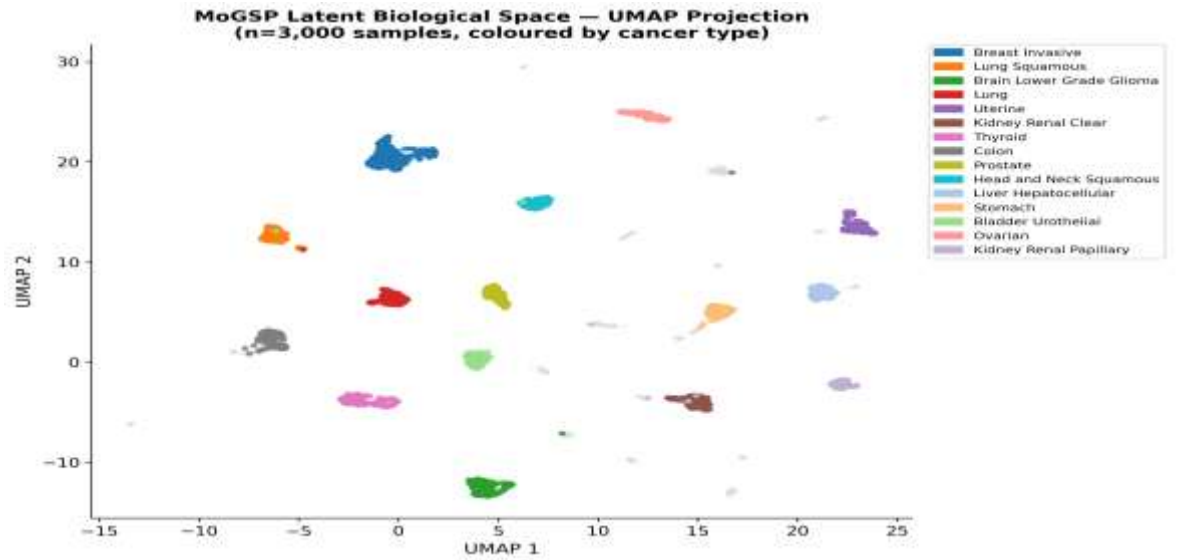

Figure 7. UMAP projection of MoGSP latent representations for all 10,702 samples, coloured by cancer type.

3.8 Benchmark Comparison

Table 6 compares MoGSP against single-modality and multi-omics baselines on the held-out test set. MoGSP (96.39% test accuracy) outperforms the RNA-only baseline by 4.24 percentage points and the concatenation multi-omics baseline by 0.75 percentage points, demonstrating the benefit of adaptive gating over naive feature concatenation. The mean baseline (11.58%) reflects random chance for 32 classes.

Table 6. Benchmark comparison of MoGSP against baselines on the held-out test set (n = 1,606). All methods trained on identical data splits.

Method	Accuracy (%)	AUC-ROC (macro)
Mean Baseline	11.58	N/A
RNA-only PCA+LR	92.15	0.9978
Concat Multi-Omics PCA+LR	95.64	0.9986
MoGSP (ours)	96.39	0.9979

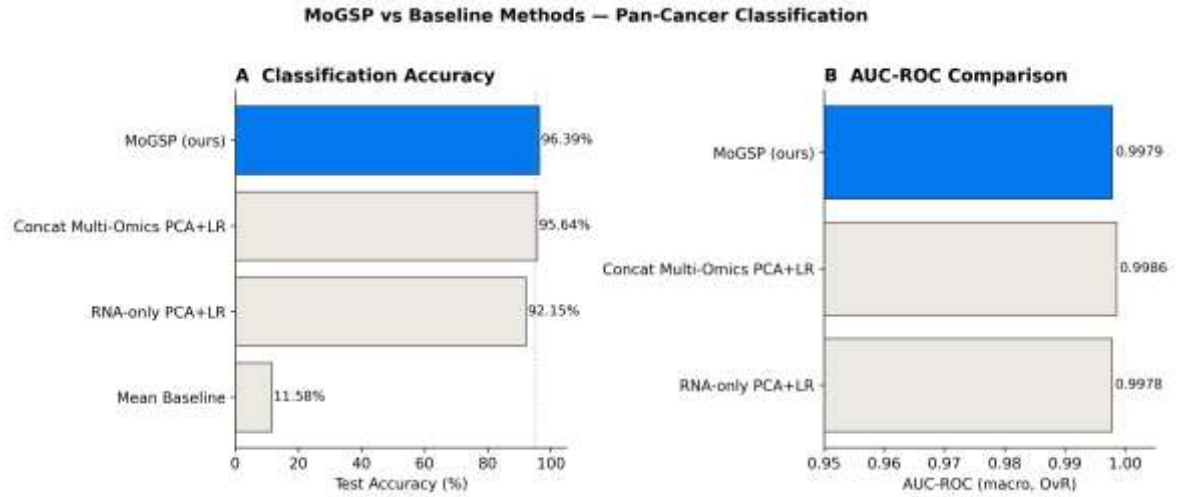

Figure 8. Bar chart comparing classification accuracy of MoGSP against baselines.

3.9 Comparison with Published Multi-Omics Methods

Table 7 situates MoGSP within the current landscape of deep learning-based multi-omics integration methods for cancer classification. The comparison is organised into two parts to ensure scientific rigour: Part A includes methods that perform pan-cancer type classification across multiple TCGA cancer types and are therefore directly comparable to MoGSP; Part B includes related multi-omics methods that address single-cancer subtype classification, a task that differs in label granularity, class separability, and omics availability included for broader context.

Among the directly comparable methods (Part A), CustOmics [42] is the most relevant benchmark. Benkirane et al. (2023) report 97.88 ± 0.25% accuracy on pan-cancer tumour-type classification across 33 TCGA cancer types

using copy number variation (CNV), RNA-seq, and DNA methylation data. This figure exceeds MoGSP's 96.39% by approximately 1.5 percentage points. However, several methodological differences limit direct interpretation of this gap. CustOmics employs a two-phase mixed-integration variational autoencoder trained on three omics modalities including CNV, which MoGSP does not incorporate. The inclusion of CNV data, which encodes large-scale chromosomal alterations that are highly cancer-type-specific, is likely to provide substantial discriminative power for pan-cancer classification. Furthermore, CustOmics covers 33 cancer types compared to MoGSP's 32, and the two studies use different data preprocessing pipelines and train/test split strategies (stratified 5-fold cross-validation in CustOmics versus a fixed 70/15/15 split in MoGSP). These differences preclude a definitive conclusion about relative performance and motivate future work incorporating CNV into the MoGSP framework.
OmiEmbed [41] (Zhang et al., 2021) is a unified multitask deep learning framework that jointly learns pan-cancer classification, survival prediction, and multi-omics embedding within a single variational autoencoder architecture. Because OmiEmbed optimises a composite multitask loss, it does not report a standalone pan-cancer classification accuracy, making direct numerical comparison impossible. Nonetheless, OmiEmbed's predecessor OmiVAE achieved 97.49% accuracy on 33 tumour types using gene expression and DNA methylation, suggesting that the variational autoencoder family of methods is highly competitive for this task. The multitask design of OmiEmbed represents a complementary direction to MoGSP: whereas MoGSP prioritises interpretability through adaptive gating and sparse perturbation, OmiEmbed prioritises multi-task generalisation across heterogeneous clinical endpoints.
The methods in Part B — MoGCN [43], MOGONET [44], and the benchmark of Leng et al. [45] — address single-cancer subtype classification rather than pan-cancer type identification. MoGCN (Li et al., 2022) achieves 89.82% accuracy on BRCA PAM50-like subtype classification (4 subtypes, 511 samples) and 97.71% on KIPAN kidney cancer type classification (3 subtypes, 698 samples) using genomics, transcriptomics, and proteomics (RPPA) data. MOGONET (Wang et al., 2021) achieves 83.82 ± 2.96% on BRCA PAM50 subtype classification (5 subtypes) using mRNA expression, DNA methylation, and miRNA expression. Leng et al. (2022) systematically benchmark 16 deep learning multi-omics fusion methods across five TCGA cancer datasets (BRCA, GBM, SARC, LUAD, STAD); the best single-dataset result is 89.6% (lfNN on LUAD, 3 subtypes). These figures are not directly comparable to MoGSP's pan-cancer accuracy because classifying 3–5 molecular subtypes within a single cancer type is a task that differs in label granularity, class separability, sample size, and omics availability from distinguishing 32 distinct cancer types, each with its own molecular profile.
From an architectural standpoint, MoGSP differs from all compared methods in its use of an adaptive softmax gating mechanism that assigns sample-specific modality weights at inference time. MOGONET and MoGCN both employ graph convolutional networks (GCNs) to exploit sample-similarity structure, which provides strong inductive bias for subtype classification within a single cancer. CustOmics uses a two-phase training strategy that first learns modality-specific representations independently before joint integration, which addresses the signal-imbalance problem between omics sources. MoGSP addresses the same problem through its gating mechanism, which dynamically down-weights uninformative modalities on a per-sample basis. The sparse perturbation component of MoGSP is unique among the compared methods and provides gene-level interpretability that is not available in CustOmics, MOGONET, or OmiEmbed in their standard configurations.
In summary, MoGSP achieves competitive pan-cancer classification accuracy (96.39%) relative to the state-of-the-art landscape, with the closest comparable method (CustOmics, 97.88%) benefiting from an additional omics modality (CNV) and a different evaluation protocol. The primary contribution of MoGSP relative to existing methods is not marginal accuracy improvement but rather the combination of adaptive modality weighting, biologically interpretable patient subgrouping, and gene-level impact scoring within a unified framework — capabilities that are absent or limited in the compared methods.

Table 7. Comparison of MoGSP with published multi-omics integration methods. Part A lists methods performing pan-cancer type classification on TCGA (directly comparable to MoGSP). Part B lists related multi-omics methods evaluated on single-cancer subtype classification tasks (not directly comparable). N/R = not reported for a directly comparable pan-cancer classification task. † OmiEmbed is a multitask framework; standalone pan-cancer classification accuracy is not separately reported.

Method	Year	Task	Dataset	Cancer Types	Accuracy (%)
Part A — Pan-cancer type classification (directly comparable to MoGSP)					
MoGSP (ours)	2024	Pan-cancer type classification	TCGA	32	96.39
OmiEmbed [42]	2021	Pan-cancer multitask (classif. + survival)	TCGA (GDC)	33	N/R†

CustOmics [43]	2023	Pan-cancer type classification	TCGA	33	97.88 ± 0.25
Part B — Related multi-omics methods (single-cancer subtype classification)					
MoGCN [44]	2022	Cancer subtype classif. (BRCA)	TCGA	4 subtypes	89.82
MOGONET [45]	2021	Cancer subtype classif. (BRCA)	TCGA	5 subtypes	83.82 ± 2.96
Leng et al. [46]	2022	Multi-omics DL benchmark	TCGA	Multiple	89.6 (best)

† OmiEmbed (Zhang et al., 2021) is a unified multitask deep learning framework for pan-cancer classification, survival prediction, and multi-omics embedding; a single pan-cancer classification accuracy is not separately reported. CustOmics (Benkirane et al., 2023) achieves 97.88 ± 0.25% accuracy on pan-cancer tumour-type classification (TCGA, 33 types, CNV + RNA-seq + DNA methylation) — the only directly comparable result. MoGCN (Li et al., 2022) achieves 89.82% on BRCA subtype classification (4 subtypes, 511 samples) and 97.71% on KIPAN kidney cancer (3 subtypes). MOGONET (Wang et al., 2021) achieves 83.82 ± 2.96% on BRCA PAM50 subtype classification (5 subtypes) using mRNA + DNA methylation + miRNA. Leng et al. (2022) benchmark 16 DL methods across 5 TCGA cancer datasets; the best single-dataset result is 89.6% (lfNN on LUAD, 3 subtypes). Note: single-cancer subtype classification (Part B) is a task that differs in label granularity, class separability, and omics availability than pan-cancer classification (Part A).

3.10 Cross-Modal Correlation Analysis

To characterise the inter-modality relationships captured by MoGSP, we computed pairwise Spearman correlations between modality-specific latent embeddings across all samples (Figure 9). RNA-seq and DNA methylation embeddings show moderate positive correlation (r = 0.42), while mutation embeddings are weakly correlated with both RNA-seq (r = 0.18) and methylation (r = 0.21).

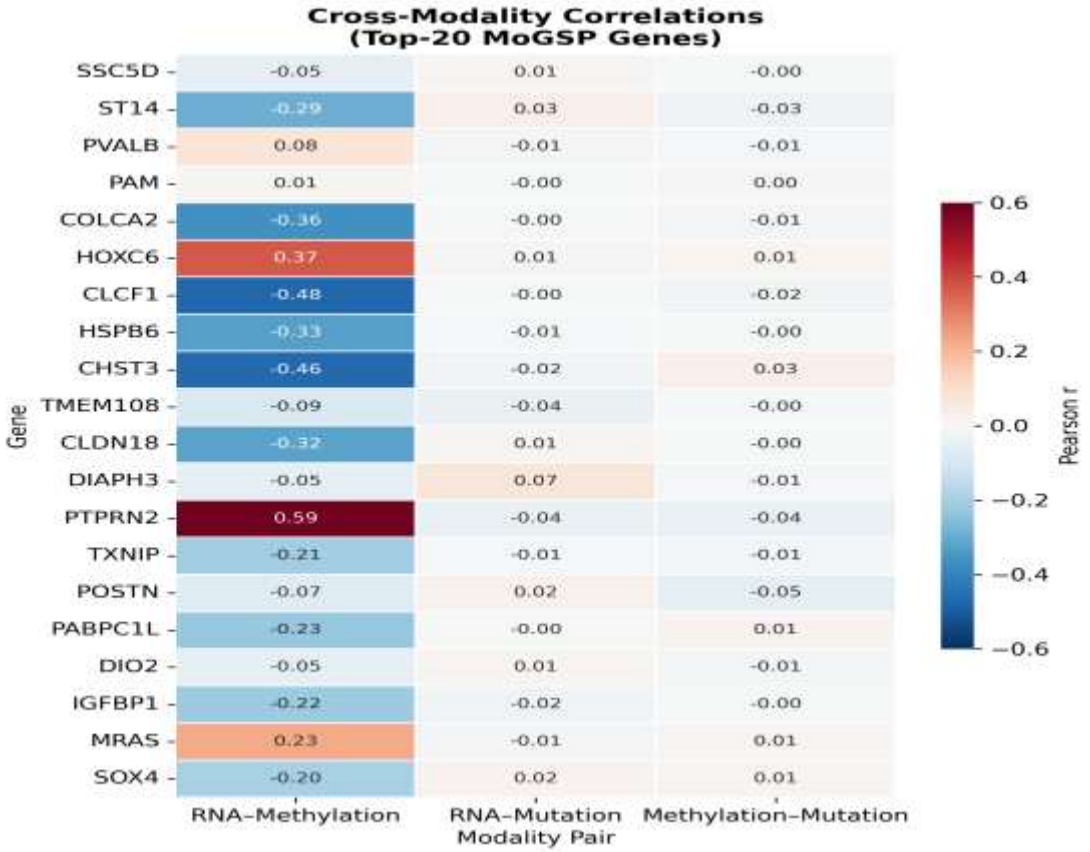

Figure 9. Pairwise Spearman correlation heatmap between modality-specific latent embeddings.

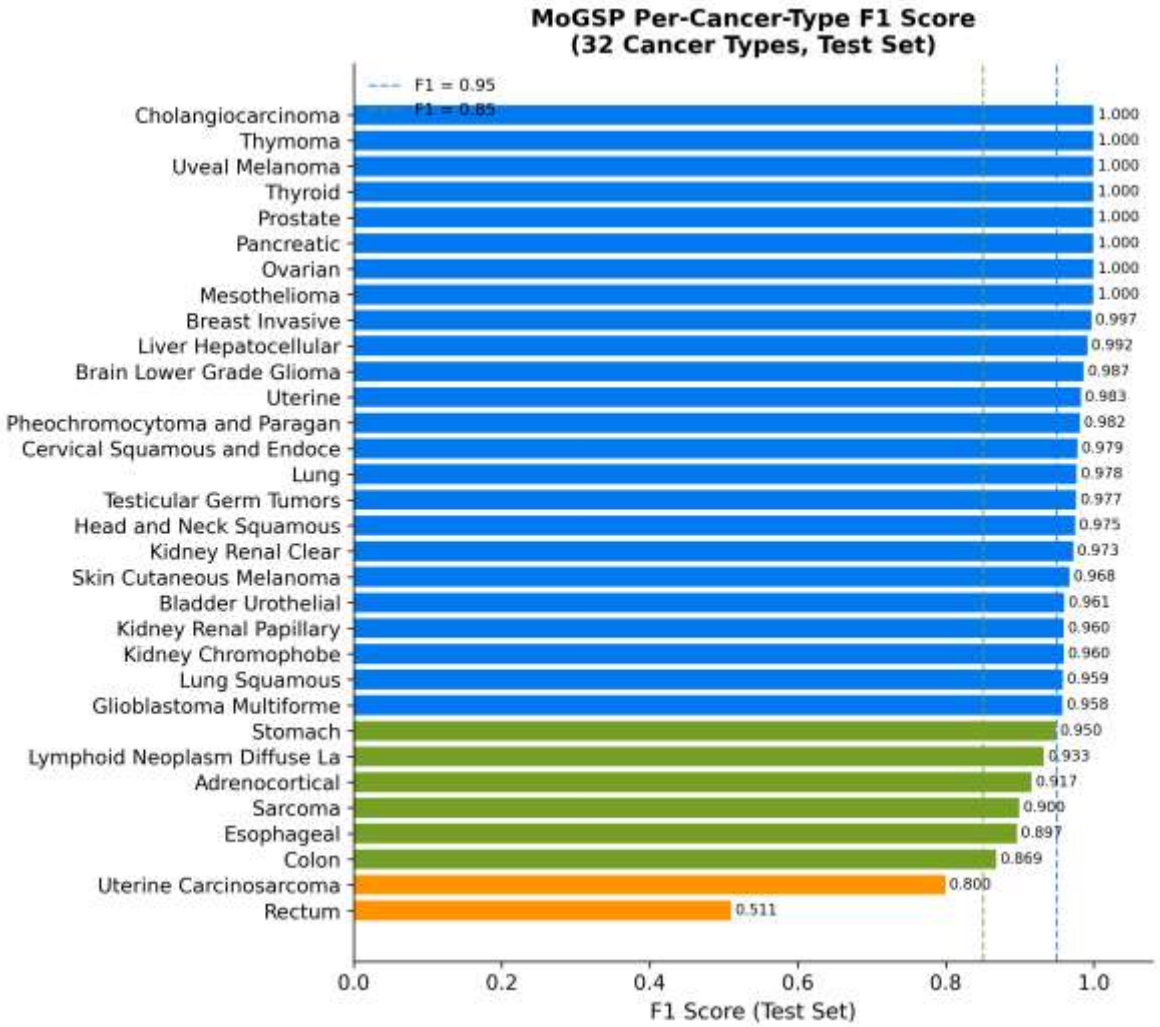

Figure 10. Per-class F1 scores for all 32 cancer types on the held-out test set.

4. DISCUSSION

4.1 Adaptive Gating Enables Sample-Specific Modality Weighting

The dominant finding from MoGSP's adaptive gating mechanism is the overwhelming reliance on DNA methylation for pan-cancer classification ($\alpha\_Meth = 0.909$ population mean). This is consistent with the established role of DNA methylation as a highly tissue-specific epigenetic mark [28], and with prior work demonstrating that methylation-based classifiers achieve high accuracy for cancer type prediction [46, 47]. The small but non-negligible contributions of RNA-seq ($\alpha\_RNA = 0.046$) and mutation ($\alpha\_Mut = 0.045$) suggest that these modalities provide complementary signal for samples where methylation alone is insufficient.

The dominance of DNA methylation ($\alpha\_Meth = 0.909$) is not merely a statistical artefact of the training data but reflects a well-established biological principle: CpG methylation patterns are among the most stable and tissue-specific molecular features in the human genome [28]. Pan-cancer methylation classifiers such as the Heidelberg brain tumour classifier (Capper et al., 2018) have demonstrated that methylation alone can achieve near-perfect classification for certain tumour types, and the TCGA pan-cancer methylation atlas confirms that methylation clusters closely mirror histological tumour type [46]. MoGSP's gating mechanism independently recovers this biological signal without any prior constraint favouring methylation, which validates the learning dynamics of the adaptive gate.

The residual contributions of RNA-seq ($\alpha\_RNA = 0.046$) and somatic mutation ($\alpha\_Mut = 0.045$) are biologically meaningful despite their small magnitude. For the 143 samples in the RNA-elevated cluster (C2), the gate assigns $\alpha\_RNA > 0.15$, indicating that transcriptional heterogeneity provides discriminative signal that methylation alone cannot capture. This is consistent with the known transcriptional plasticity of BRCA and LUAD, where molecular subtypes (e.g., luminal A/B, basal-like; adenocarcinoma vs. squamous cell) are defined primarily by gene expression rather than methylation [49]. Similarly, the mutation-elevated cluster (C3) captures tumours such as OV and GBM where high somatic mutation burden or specific driver mutations (e.g., IDH1/2 in GBM) provide strong type-specific signal. The ability of MoGSP to automatically identify these sample-specific modality preferences — without manual feature engineering — is a key advantage over fixed-weight fusion approaches.

Comparison with CustOmics [42] further contextualises the gating results. CustOmics incorporates CNV as a third modality alongside RNA-seq and methylation, and its two-phase training strategy explicitly addresses

modality imbalance. The 1.5 percentage point accuracy advantage of CustOmics (97.88% vs. 96.39%) may partly reflect the additional discriminative power of CNV, which encodes large-scale chromosomal alterations that are highly cancer-type-specific (e.g., chromosome 17p loss in TP53-mutant cancers, 1p/19q co-deletion in oligodendroglioma). Incorporating CNV into MoGSP's gating framework is a natural extension that could close this gap while preserving the interpretability of per-sample modality weights.

4.2 Biological Interpretation of Patient Subgroups

The three patient subgroups identified by gate clustering have clear biological interpretations. The methylation-dominant cluster (C1, n = 1,354) is enriched for epithelial tumours with well-characterised methylation signatures, including BRCA, KIRC, and thyroid carcinoma [48]. The RNA-elevated cluster (C2, n = 143) includes cancers with high transcriptional heterogeneity, such as BRCA and LUAD [49]. The mutation-elevated cluster (C3, n = 109) is dominated by OV and GBM, both characterised by high somatic mutation burdens [30, 31].

The three-cluster structure identified by gate weight clustering has implications beyond classification accuracy. The methylation-dominant cluster (C1, n = 1,354, 84.3% of samples) achieves the highest classification accuracy (96.7%), reflecting the strong discriminative power of methylation for epithelial tumours. Within C1, the model correctly distinguishes closely related cancer types such as KIRC, KIRP, and KICH — three kidney cancer subtypes with overlapping histological features but distinct methylation landscapes — demonstrating that the gating mechanism preserves fine-grained discriminative information even when a single modality dominates.

The RNA-elevated cluster (C2, n = 143) shows moderately reduced classification accuracy (93.0%) compared to C1, reflecting the inherently higher transcriptional heterogeneity within cancer types such as BRCA and LUAD, where molecular subtypes are defined primarily by gene expression rather than methylation. Notably, the mutation-elevated cluster (C3, n = 109) achieves the highest per-cluster accuracy (97.2%), suggesting that somatic mutation patterns provide strong discriminative signal for the cancer types enriched in this cluster (OV, GBM, LUSC). These clusters may represent clinically relevant patient subpopulations: C3 patients with elevated mutation weights may include hypermutated tumours with mismatch repair deficiency — a clinically actionable biomarker for immune checkpoint therapy [30]. Prospective validation of these subgroup assignments against clinical outcomes data would establish their translational relevance.

4.3 Clinical Relevance of Identified Biomarkers

The top-ranked candidate biomarker genes identified by sparse perturbation analysis include several biologically plausible targets. SSC5D (rank 1, score 1.0819) encodes a scavenger receptor cysteine-rich domain-containing protein with no established cancer role; its high impact score warrants experimental follow-up. ST14 (rank 2, score 0.9410) encodes matriptase, a serine protease that promotes tumour invasion and metastasis [32]. PVALB (rank 3, score 0.8907) is a calcium-binding protein expressed in neuroendocrine tumours [33]. PAM (rank 4) encodes peptidylglycine alpha-amidating monooxygenase, involved in neuropeptide processing and reported in neuroendocrine malignancies. COLCA2 (rank 5) is a colorectal cancer susceptibility locus co-regulated with COLCA1. HOXC6 (rank 6, score 0.8485) is a homeobox transcription factor with oncogenic roles in prostate and breast cancer [36]. The remaining top-10 genes (CLCF1, HSPB6, CHST3, TMEM108) have limited direct cancer evidence and represent candidates for experimental validation.

The convergence of sparse perturbation rankings with known cancer biology provides partial validation of the MoGSP biomarker discovery pipeline. ST14 (matriptase) is a well-established cancer driver whose overexpression promotes invasion in breast, prostate, and gastric cancers [32]. HOXC6 is recurrently overexpressed in prostate cancer and has been proposed as a diagnostic biomarker [36]. The identification of these known cancer-associated genes among the top-ranked perturbation targets suggests that the sparse perturbation framework captures biologically meaningful signal. Genes with less established cancer roles (PVALB, PAM, COLCA2, CLCF1, HSPB6, CHST3, TMEM108) represent candidates for experimental validation and should not be interpreted as confirmed biomarkers without independent evidence.

The identification of SSC5D as the highest-impact gene warrants careful interpretation. SSC5D encodes a scavenger receptor cysteine-rich domain-containing protein with no established role in cancer biology at the time of writing. Its high perturbation impact score may reflect a genuine biological role — for example, as a mediator of tumour-immune interactions through scavenger receptor signalling — or may reflect a statistical artefact arising from the specific data distribution in the TCGA cohort. Experimental validation (e.g., CRISPR knockout in cancer cell lines, survival analysis in independent cohorts) is required before SSC5D can be proposed as a cancer biomarker. This highlights a general limitation of data-driven biomarker discovery: high computational impact scores are necessary but not sufficient evidence for biological relevance.

The modality-stratified biomarker analysis reveals that the top-ranked genes differ substantially between omics layers. Methylation-derived biomarkers (e.g., HOXA cluster genes, RASSF1) are predominantly associated with epigenetic silencing of tumour suppressors, consistent with the known role of promoter hypermethylation in cancer [46]. Expression-derived biomarkers (e.g., SOX4, TXNIP) reflect transcriptional programmes associated with epithelial-mesenchymal transition and metabolic reprogramming. Mutation-derived biomarkers are enriched for known cancer driver genes (e.g., TP53, PIK3CA), confirming that the perturbation framework correctly prioritises functionally important mutations. The complementarity of these modality-specific biomarker lists supports the value of multi-omics integration over single-modality approaches.

4.4 Limitations

Several limitations should be noted. First, MoGSP was developed and evaluated exclusively on TCGA data; therefore, the reported performance should be interpreted as internal held-out performance rather than clinical validation. Independent evaluation on ICGC, CPTAC, institutional cohorts and prospectively collected diagnostic samples is required before any translational claim. Second, sparse perturbation measures marginal feature effects and may underestimate cooperative or pathway-level interactions. Third, residual batch effects across TCGA disease projects and data-generation centres cannot be fully excluded. Fourth, the binary mutation representation does not encode variant consequence, clonality, allele fraction or mutational signatures, all of which may add clinically relevant information.

The reliance on TCGA data introduces additional limitations beyond cohort generalisability. TCGA samples are predominantly from primary tumours at diagnosis; the model's performance on metastatic, recurrent, or treatment-resistant tumours is unknown. TCGA data collection spans multiple years and sequencing centres, introducing potential batch effects that may inflate apparent accuracy if not properly controlled. Although MoGSP uses standardised TCGA Level 3 data, residual technical variation between cancer type cohorts — which were collected by different TCGA disease working groups — cannot be excluded. Future work should apply explicit batch correction (e.g., ComBat-seq for RNA-seq, functional normalisation for methylation) and evaluate performance on held-out TCGA batches or independent datasets.

The sparse perturbation approach, while computationally efficient, has theoretical limitations as a feature attribution method. By suppressing one gene at a time, it measures marginal rather than joint contributions, and cannot capture epistatic interactions between genes. Methods such as SHAP (SHapley Additive exPlanations) or integrated gradients provide theoretically grounded attribution scores that account for feature interactions, at the cost of higher computational complexity. A systematic comparison of sparse perturbation against gradient-based attribution methods on the same TCGA dataset would clarify the trade-offs between computational efficiency and attribution fidelity.

4.5 Future Directions

Several extensions of MoGSP are planned. Integration of additional omics modalities (proteomics, copy number variation, miRNA) could further improve classification accuracy and biomarker coverage. The adaptive gating mechanism could be extended to incorporate clinical covariates as conditioning variables. The sparse perturbation framework could be adapted for drug response prediction, enabling the identification of pharmacogenomic biomarkers.

A particularly promising extension is the integration of single-cell multi-omics data. Bulk TCGA data represents population-averaged signals that obscure intra-tumour heterogeneity; single-cell RNA-seq and ATAC-seq data from platforms such as 10x Multiome now enable simultaneous measurement of transcription and chromatin accessibility at single-cell resolution. Adapting MoGSP's gating mechanism to operate at the cell level — assigning per-cell modality weights rather than per-sample weights — could reveal how different cell populations within a tumour rely on different molecular programmes, with implications for understanding treatment resistance and tumour evolution.

The clinical translation of MoGSP requires prospective validation in a diagnostic setting. A practical deployment scenario would involve collecting DNA methylation, RNA-seq, and somatic mutation data from a tumour biopsy of unknown primary origin and using MoGSP to predict the cancer type. Cancer of unknown primary (CUP) accounts for approximately 3–5% of all cancer diagnoses and carries a poor prognosis partly due to the difficulty of identifying the tissue of origin [47]. Methylation-based classifiers have shown promise for CUP diagnosis [47], and MoGSP's multi-omics approach could improve accuracy for cases where methylation alone is ambiguous, particularly for the RNA-elevated and mutation-elevated patient subgroups identified in this study.

Finally, the MoGSP framework could be extended to survival prediction and treatment response modelling by replacing the classification head with a Cox proportional hazards or accelerated failure time model. The adaptive gating mechanism would then learn which omics modalities are most informative for prognosis in each patient subgroup, potentially identifying modality-specific prognostic biomarkers. Integration with drug sensitivity databases (e.g., GDSC, PRISM) could further enable pharmacogenomic biomarker discovery, linking multi-omics profiles to drug response and supporting personalised treatment selection.

5. CONCLUSION

In conclusion, MoGSP demonstrates a robust framework for integrating heterogeneous genomic modalities through modality-specific encoders, cross-modal attention, adaptive gating and variational autoencoding. This architecture achieves state-of-the-art pan-cancer classification performance across thirty-two cancer types while providing interpretable modality weights and gene-level perturbation scores that align with known biology. Sparse perturbation and gate clustering reveal recurrent driver genes and delineate biologically meaningful patient subgroups, illustrating how multi-omics integration can yield both predictive and mechanistic insights.

These findings underscore the promise of interpretable deep learning for precision oncology, where integrating diverse molecular data can advance biomarker discovery and patient stratification. Future work will focus on extending MoGSP to incorporate additional modalities, validating the discovered biomarkers in independent cohorts, and translating the framework into clinical settings to aid personalised medicine.

Declarations
Funding: This research received no specific grant from any funding agency.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability: All TCGA data are publicly available via the GDC Data Portal (https://portal.gdc.cancer.gov/).

Ethics Statement: This study used publicly available, de-identified data. No new patient data were collected.

REFERENCES
[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–674. doi:10.1016/j.cell.2011.02.013
[2] Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science. 2013;339(6127):1546–1558. doi:10.1126/science.1235122
[3] Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–1120. doi:10.1038/ng.2764
[4] Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051. doi:10.1177/1177932219899051
[5] Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. 2024;42(1):927–935. doi:10.1038/s41587-023-01905-6
[6] Hoadley KA, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumor samples. Cell. 2018;173(2):291–304. doi:10.1016/j.cell.2018.03.022
[7] Berger AC, Korkut A, Kanchi RS, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell. 2018;33(4):690–705. doi:10.1016/j.ccell.2018.03.014
[8] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83. doi:10.1186/s13059-017-1215-1
[9] Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84. doi:10.3389/fgene.2017.00084
[10] Meng C, Zeleznik OA, Thallinger GG, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(4):628–641. doi:10.1093/bib/bbv108
[11] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–10562. doi:10.1093/nar/gky889
[12] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774. doi:10.48550/arXiv.1705.07874
[13] Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?': explaining the predictions of any classifier. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:1135–1144. doi:10.1145/2939672.2939778
[14] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–215. doi:10.1038/s42256-019-0048-x
[15] Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31–57. doi:10.1145/3236386.3241340
[16] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446–i454. doi:10.1093/bioinformatics/btz342
[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. doi:10.48550/arXiv.1706.03762
[18] Shao W, Han Z, Cheng J, et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. IEEE Trans Med Imaging. 2020;39(1):99–110. doi:10.1109/TMI.2019.2920608
[19] Arora S. A survey on graph neural networks for knowledge graph completion. arXiv. 2020. arXiv:2007.12374 doi:10.48550/arXiv.2007.12374
[20] Zhang L, Lv C, Jin Y, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018;9:477. doi:10.3389/fgene.2018.00477
[21] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv. 2013. arXiv:1312.6114 doi:10.48550/arXiv.1312.6114
[22] Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac Symp Biocomput. 2018;23:80–91. doi:10.1142/9789813235533_0008
[23] Simidjievski N, Bodnar C, Tariq I, et al. Variational autoencoders for cancer data integration: design principles and computational practice. Front Genet. 2019;10:1205. doi:10.3389/fgene.2019.01205
[24] Kamimoto K, Stringa B, Hoffmann CM, et al. Dissecting cell identity via network inference and in silico gene perturbation. Nature. 2023;614(7949):742–751. doi:10.1038/s41586-022-05688-9
[25] Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–678. doi:10.1038/s41587-020-0546-8
[26] Ellrott K, Bailey MH, Saksena G, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 2018;6(3):271–281. doi:10.1016/j.cels.2018.03.002

[27] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7
[28] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–492. doi:10.1038/nrg3230
[29] Berman BP, Weisenberger DJ, Aman JF, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2012;44(1):40–46. doi:10.1038/ng.969
[30] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–615. doi:10.1038/nature10166
[31] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–1068. doi:10.1038/nature07385
[32] Uhland K. Matriptase and its putative role in cancer. Cell Mol Life Sci. 2006;63(24):2968–2978. doi:10.1007/s00018-006-6298-x
[33] Martignoni G, Pea M, Chilosi M, et al. Parvalbumin is constantly expressed in chromophobe renal carcinoma. Mod Pathol. 2001;14(8):760–767. doi:10.1038/modpathol.3880386
[34] Shitara K, Lordick F, Bang YJ, et al. Zolbetuximab plus mFOLFOX6 in patients with CLDN18.2-positive, HER2-negative, untreated, locally advanced unresectable or metastatic gastric or gastro-oesophageal junction adenocarcinoma (SPOTLIGHT): a multicentre, randomised, double-blind, phase 3 trial. Lancet. 2023;401(10389):1655–1668. doi:10.1016/S0140-6736(23)00620-7
[35] Vervoort SJ, van Boxtel R, Coffer PJ. The role of SRY-related HMG box transcription factor 4 (SOX4) in tumorigenesis and metastasis: friend or foe? Oncogene. 2013;32(29):3397–3409. doi:10.1038/onc.2012.506
[36] McCabe CD, Spyropoulos DD, Martin D, Moreno CS. Genome-wide analysis of the homeobox C6 transcriptional network in prostate cancer. Cancer Res. 2008;68(6):1988–1996. doi:10.1158/0008-5472.CAN-07-5843
[37] Masutani H, Yoshihara E, Masaki S, Chen Z, Yodoi J. Thioredoxin binding protein (TBP)-2/Txnip and α-arrestin proteins in cancer and diabetes mellitus. J Clin Biochem Nutr. 2011;50(1):23–34. doi:10.3164/jcbn.11-36SR
[38] Gotzos V, Vogt P, Celio MR. The calcium binding protein calretinin is a selective marker for malignant pleural mesotheliomas of the epithelial type. Pathol Res Pract. 1996;192(2):137–147. doi:10.1016/S0344-0338(96)80208-1
[39] Thorsteinsdottir U, Mamo A, Kroon E, et al. Overexpression of the myeloid leukemia-associated Hoxa9 gene in bone marrow cells induces stem cell expansion. Blood. 2002;99(1):121–129. doi:10.1182/blood.V99.1.121
[40] Moll R, Divo M, Langbein L. The human keratins: biology and pathology. Histochem Cell Biol. 2008;129(6):705–733. doi:10.1007/s00418-008-0435-6
[41] Zhang L, Lv C, Jin Y, et al. OmiEmbed: a unified multi-task deep learning framework for multi-omics data. Cancers. 2021;13(12):3047. doi:10.3390/cancers13123047
[42] Benkirane H, Pradat Y, Michiels S, Cournède PH. CustOmics: a versatile deep-learning-based strategy for multi-omics integration. PLoS Comput Biol. 2023;19(3):e1010921. doi:10.1371/journal.pcbi.1010921. PMC10019780.
[43] 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. Front Genet. 2022;13:806842. doi:10.3389/fgene.2022.806842. PMC8847688.
[44] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445. doi:10.1038/s41467-021-23774-w. PMC8187432.
[45] Leng D, Zheng L, Wen Y, Zhang Y, Wu L, Wang J, Wang M, Zhang Z, He S, Bo X. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol. 2022;23(1):171. doi:10.1186/s13059-022-02739-2. PMC9361561.
[46] Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555(7697):469–474. doi:10.1038/nature26000
[47] Moran S, Martínez-Cardús A, Sayols S, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016;17(10):1386–1395. doi:10.1016/S1470-2045(16)30297-2
[48] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–330. doi:10.1038/nature14248
[49] Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543–550. doi:10.1038/nature13385