**Beyond Differential Expression: Deep Neural Profiling Reveals RAP1GAP2 as a Latent Regulator of Tumor Invasion in Oropharyngeal Carcinoma**

**Chapter Two**

**Material and Methods**

**2.1 Overview of the Study**

The design of the overall study is illustrated in **Figure 2.1**

A diagram of a computer

AI-generated content may be incorrect.

***Figure 2.1 Overview of the study pipeline****. PCA-transformed multi-dataset gene expression is encoded via VAE to latent space, followed by Integrated Gradients-based gene attribution and supervised learning to identify molecular drivers and extract biological insights of the latent spaces.*

**2.2 Datasets Retrieval**

Publicly available gene-expression datasets of oral carcinoma (OC) generated using different platforms—including *[HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array, [HG-U133A] Affymetrix Human Genome U133A Array, Illumina NextSeq 500 (Homo sapiens)*—were downloaded. A total of 19 datasets were parsed from the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>) database for Oral Cancer types, where a python library ***GEOparse v2.0.0*** (<https://github.com/guma44/GEOparse>) was incorporated to extract the sequencing data with their phenotype data from the database server. All information about the datasets including sample size mentioned in **Table 2.1.**

# **Table 2.1 Expression Profiling Datasets for OC**

|  |  |  |  |
| --- | --- | --- | --- |
| **GEO\_Accession** | **Samples** | **Platform** | **Study\_Type** |
| GSE37991 | 80 (40 tumor + 40 normal) | GPL6883 (Illumina HumanRef‑8) | Expression profiling by array |
| GSE23558 | 31 (27 tumor + 4 normal) | GPL6480 (Agilent 44K) | Expression profiling by array |
| GSE25099 | 79 (57 tumor + 22 normal) | GPL5175 (Affymetrix Exon ST) | Expression profiling by array |
| GSE10121 | 41 (35 tumor + 6 normal) | Operon Oligoset 4.0 | Expression profiling by array |
| GSE31853 | 11 (8 tumor cell lines + 3 normal) | GPL96/570 (Affymetrix) | Expression profiling by array |
| GSE131182 | 12 (6 paired tumor + normal) | GPL20301 (Illumina HiSeq) | Expression profiling by RNA‑seq |
| GSE145272 | 10 (5 metastatic + 5 non-metastatic) | HiSeq 2500 RNA‑seq | Expression profiling by RNA‑seq |
| GSE217142 | 6 (primary + recurrent tumors) | NovaSeq 6000 RNA‑seq | Expression profiling by RNA‑seq |
| GSE85195 | 49 (34 OSCC + 15 OPL) | GPL6480 (Agilent 44K) | Expression profiling by array |
| GSE168227 | 6 paired tumor-normal samples | Agilent lncRNA microarray | Expression profiling by array |
| GSE84805 | 6 paired tumor-normal samples | Agilent lncRNA array | Expression profiling by array |
| GSE30784 | 229 total (167 tumor + others) | GPL570 (Affymetrix U133 Plus 2.0) | Expression profiling by array |
| GSE2280 | 32 (27 non-metastatic + 5 metastatic) | GPL96 (Affymetrix U133A) | Expression profiling by array |
| GSE3524 | 20 (16 tumor + 4 normal) | GPL96 (Affymetrix U133A) | Expression profiling by array |
| GSE6791 | 154 (119 tumor + 35 controls) | Affymetrix U133 Plus 2.0 | Expression profiling by array |
| GSE41442 | 55 (45 tumor + 10 normal) | GPL570 (Affymetrix) | Expression profiling by array |
| GSE37371 | 100 (50 tumor + 50 normal) | GPL96 (Affymetrix) | Expression profiling by array |
| GSE23030 | 30 metastatic tongue OSCC | GPL5175 (Affymetrix Exon ST) | Expression profiling by array |
| GSE29000 | 50 (40 tumor + 10 normal) | GPL570 (Affymetrix) | Expression profiling by array |

Extracted results according to the supplied GEO accession ids filtered out based on the treatment and condition of the samples. We got a total of 1001 samples from all the datasets combined, where sample number with OC positive was 754. Samples treated with radiation therapy, chemotherapy, targeted therapy, immunotherapy, hormonal therapy and drugs were excluded from the study manually.

**2.3 Data Integration, Batch Effect Removal and Preprocessing**

To amalgamate data from different platforms, a python data analysis library ***pandas******v1.5.3*** was incorporated. Data imputation was conducted by ***missForest v0.9***package in R to avoid the NA values in the datasets. For concatenating multiple datasets from multiple platforms with different techniques, a batch effect correction method based on python library was applied on the integrated data to combat the platform specific biases. A function called ***“ComBat”*** from python library ***pyComBat******v0.3.2*** was used to remove the technical biases that arose by the integration process. Expression data of merged dataset was log-transformed, Z-standardized on each gene to ensure that all features are on the same scale.

**2.3 Training Deep Neural Network Models**

**2.3.1 Datasets Merging and Standardization**

After manual selection and preprocessing, we had 663 cancer-positive samples, each containing 11020 genes—common in all datasets. Despite the high dimensional gene expression matrix, which was complex to interpret the samples with their condition, a principal component analysis was conducted with 500 PCs **(*n\_components=500*)** while preserving all important data and variance among the samples. PCA was performed in R using the following packages: ***stats v4.2.3, factoextra v1.0.7*** for extraction and display of PCA results, and ***dplyr v1.1.4*** for data manipulation.

**2.3.2 Traditional Deep Learning Model**

A probabilistic latent variable model was built on reduced PC data to learn a compact, non-linear delineation of the high-dimensional gene expression data. This is a type of neural network that contains an encoder and a decoder network with an entropy-limited latent mapping with *D* latent variables (here, *D* *≪* *M*, where *M=500PC*,represents the number of features) in the middle. This process generates an embedding ***Z***, which preserves the whole information of the input (*500PC*) into a lower dimensional space. Categorically, the encoder network, defined as , maps from the input space to latent embedding . Similarly, the decoder network, defined as maps the embedding **Z** back to input space. The main objective of the model is to minimize the anticipated squared Euclidean (L2) norm between the input and its reconstruction:

Here, and are the parameters of the encoder and decoder, respectively, and represents the reconstructed input for every sample. Where, L2 loss denoted by , captures the total reconstruction error across all dimensions of the input. Overtly, this corresponds to:

**2.3.3 Additional Sample Distribution**

Unlike conventional approach, we used probabilistic embedding model (PEM), which encodes each sample as a probability distribution—captures uncertainty and biological variability inherent in gene expression profiles. Samples with 500 principal components (PCs) were used to construct the input matrix , where is the number of samples and is the number of features. This matrix was passed to an encoder , which outputs a mean vector and a variance vector :

**.**

A decoder reconstructs the input from the sampled latent vector . The model is trained to minimize the following loss:

The first term ensures accurate reconstruction, while the KL divergence regularizes the latent space by encouraging it to resemble a standard Gaussian distribution. After training, the learned latent variables were used for gene importance analysis using Integrated Gradients, followed by pathway enrichment.

**2.4 Neural Network Design and Hyperparameter Optimization**

**2.4.1 Train Model with Adam Optimizer**

PEM models were trained to unite the PCs from the OC gene expression matrix as inputs. Three-layer encoder and decoder networks were designed as a mirror of each other. The model was trained in batches of 50 samples by using Adam optimizer, with a learning rate of 0.0005, with weight initialized randomly using the Glorot uniform method.

**2.4.2 Cross validate and Extract Best Latent Dimension**

To determine the best fitted latent space as per my study, we deliberately selected a set of sizes: 5, 10, 25, 50, 75, and 100. This comprehensive selection was made to give our models a broad scope to capture a wide range of information from the datasets. Hyperparameter tuning was performed to fine-tune hyperparameters including the dropout rate and the number of neurons per layer using 5-fold cross-validation, guided by validation reconstruction error. We tested dropout values including 0, 0.2, 0.4, and 0.6. For hidden layer configurations, we explored multiple settings such as (50, 5), (100, 25), (250, 50), (250, 100), and (300, 150), where the first and second values indicate the number of neurons in the first and second hidden layers, respectively. The model was implemented in Python using **Keras v2.2.4** and **TensorFlow v1.12.0.**

**2.5 Learning Robust Latent Representations**

To find out the stable and fruitful biological representation of the data, PEMs were trained with different random initializations and latent dimensionalities. For each latent size, training across multiple random seeds was repeated, resulting in a large collection of embeddings. To aggregate latent variables generated across multiple folds of different models, k-means clustering was applied to group (*I*) similar latent features together. To obtain the final ensemble latent dimension , G-means clustering was implemented, resulting in a fixed latent size *L*=50, which was used across all samples for downstream analysis. The final latent embedding for each sample was constructed by averaging all latent variables within each cluster.

**2.6 Gene Attribution and Pathway Analysis**

**2.6.1 Sensitivity-Based Scoring (SBS) for Gene-to-Latent Attribution**

To determine which gene contributed to what latent variables, a custom sensitivity-based scoring (SBS) approach was applied. SBS was first integrated into the method to calculate the importance of each PC for every latent variable. Then these attributions were scaled to gene level with the PC level weights, resulting in gene-level importance scores and by averaging we got global gene attributions for each latent.

**2.6.2 Pathway Enrichment Analysis of Latent Variable-Associated Genes**

To interpret the biological representation, top-ranked genes derived from every ensemble latent variable, we performed pathway enrichment analysis using the g:Profiler tool via the gprofiler2 v2.34 R package. Gene sets with the highest attribution scores were input into the gost() function, which maps genes to known functional categories including Gene Ontology (GO) terms (Biological Process, Molecular Function, Cellular Component), KEGG pathways, and Reactome pathways. We used the default settings for the organism (Homo sapiens), applied multiple testing correction via the Benjamini–Hochberg method (FDR < 0.05), and excluded electronic GO annotations to improve specificity. The results were visualized and ranked by adjusted p-values and term size to highlight the most enriched biological functions associated with each latent variable.

**2.6.3 Gene Set Enrichment Analysis (GSEA)**

To uncover the biological functions associated with each latent variable, we performed Gene Set Enrichment Analysis (GSEA) using pre-ranked gene lists derived from latent variable attributions. The enrichment results were obtained using a standardized pipeline and summarized across all latent variables. Pathways with a false discovery rate (FDR) < 0.05 were considered statistically significant. We calculated the normalized enrichment score (NES) for each term-latent pair and constructed a matrix of NES values. To focus on the most variable biological patterns, we selected the top 50 pathways based on the highest variance across latent variables. These were visualized as a heatmap using the ***seaborn v0.11.5*** library in Python, highlighting pathway–latent associations that may represent underlying biological signals.

**2.7 Supervised Deep Learning Model Training**

**2.7.1 Gene Selection and Data Collection**

To identify important driver genes for oropharyngeal carcinoma (OC), we analyzed gene attribution scores generated by the DeepProfile model across 50 latent variables. Based on this analysis, we selected 20 genes that consistently ranked among the top contributors across multiple latent dimensions. These candidate driver genes were validated using an independent dataset, which included both OC and non-tumor control samples profiled on Illumina HiSeq 4000 and NovaSeq 6000 sequencing platforms.

**2.7.2 Normalization and Batch Correction**

To address potential batch effects and platform-specific variability, we applied gene-wise Z-score normalization within each batch. Following normalization, batch correction was carried out using the empirical Bayes method implemented in the pycombat Python package (v0.3.5). All data manipulation and preprocessing were performed using the pandas (v2.2.1) and numpy (v1.24.4) libraries, with additional support from scanpy (v1.9.6) for annotation and matrix handling.

**2.7.3 Model Development and Training**

We developed and trained three types of deep learning models to classify samples into OC or control groups based on the expression of the 20 selected genes. These models were implemented using TensorFlow 2.12.0 with the Keras backend. Hyperparameter tuning was conducted using the kerastuner library (v1.3.5), and model performance was assessed through five-fold stratified cross-validation. The optimal MLP architecture consisted of two hidden layers with 128 and 64 neurons respectively, each followed by ReLU activation and dropout layers with a rate of 0.2. A final sigmoid-activated output layer was used for binary classification. All models were trained using the Adam optimizer (learning rate = 1e-4), binary cross-entropy loss, a batch size of 32, and early stopping based on validation loss with a patience of 10 epochs.

**2.7.4 Evaluation and Visualization**

Model performance was evaluated using two key metrics: area under the precision–recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC). Visualizations of model predictions, ROC curves, and PR curves were generated using matplotlib (v3.8.0) and seaborn (v0.13.2). All experiments were conducted in a secured Linux-based computing environment.

**2.8 Differential Gene Expression analysis**

Expression data were analyzed using ***DESeq2 v1.40.2***. Low-expression entries were removed before normalization. Variance-stabilizing transformation was applied for visualization. Differential expression analysis was performed using negative binomial distribution, and significance was defined as adjusted p-value < 0.05 and absolute log₂ fold change > 1. Volcano plots were generated using ***EnhancedVolcano v1.20.0***.

**Chapter Three**

**Results**

**3.1 Data Preprocessing and Quality Assessment**

Highly expressive models such as deep neural networks tend to overfit when the sample size is small, we collected 19 available expression datasets from different platforms for human Oropharyngeal Cancer (OC). To remove the platform-specific biases, we preprocessed the datasets **(Figure 3.1A)**, manually excluded samples that did not satisfy the requirements, and finalized 643 samples for PCA, with 11020 genes common across all datasets. Standardized gene expression values were visualized using a boxplot **(Figure 3.1A)** among all the samples, showing consistent distribution across samples and confirming effective scalability. PCA was performed on the 643 samples expression to reduce the dimension of the features in 500 PCs for model training, where scatterplot **(Figure 3.1B)** showed no ostensible clustering or batch effect, indicating appropriateness for unsupervised modeling. The scree plot **(Figure 3.1C)** of the first 50 PCs shows uniformly low variance,confirmedthat the components are evenly distributed. Other 450 PCs are similarly contained the same proportion of variance around 0.002. A minor drop in ratio in PC9 was observed, which likely reflects numerical or structural variance fluctuations other than biological interpretation. Overall, all PCs supported further processing, used as inputs to train deep neural networks.

**3.2 Latent Space Extraction Using Deep Neural Network**

An unsupervised neural network, PEM, was introduced to the 500 PCs to train the models. A unique methodology implied to extract the optimal latent dimension, as this unswervingly affects the model’s performance and compatibility to balance the representation of the biological signals.

Multiple models trained using the latent dimensions, including 5, 10, 25, 50, 75, and 100, and evaluated their ability to reconstruct the same sample using the parameters based on reconstruction error in both training and validation sets **(Figure 3.2A)**. As the number of latent nodes increases, the reconstruction errors reduce as per the change, representing higher capacity of reconstruction. However, the improvement stops after 50 dimensions, which implies that higher nodes can increase the risk of overfitting the data as well as the complexity of the process.

Therefore, we selected 50 nodes of latent to finalize the PEM models and got multiple folds of latent from all the models in each fold. This hyperparameter tuning helped us to reach to the most

****

***Figure 3.1. Preprocessing and PCA of gene expression data.* *(A)*** *Boxplot of standardized expression values for 11,020 genes across 643 finalized samples. Each box represents one sample, where dots represent outliers.* ***(B)*** *PCA scatterplot, containing the first two principal components for all samples; X axis containing PC1 and Y axis containing PC2* ***(C)*** *Scree plot showing the proportion of variance explained by the first 50 principal components. The variance contribution is uniformly low, supporting their use in downstream neural network training.*

relevant latent spaces, understand the core biology of OC from the complex environment of the data. To extract the biological meaning of every latent, an integrated model was implemented to the trained PEM model. This attribution method quantified the contribution of each 500PCs (input features) to each latent variable. These attributions were traced back to the original gene space using PCA loadings, resulting in gene-level attributions scores for each latent dimension.

**Figure 3.2B**, a sample representation of the top 10 genes in the first latent dimension, showing the strong connection with the latent node 0, ranked by their importance score. These genes, including

MSRB1 (0.00416), TTI2 (0.00389), MPP5 (0.00358), ATF3 (0.00354), PYHIN1 (0.00349),



***Figure 3.2. Model Performance and Gene Attribution.* *(A)*** *Barplot showing reconstruction error for both training and validation sets across different latent dimensions. X axis represents the latent nodes selected for model training and Y axis showing the reconstruction error values. (B) Barplot showing the top 10 genes contributed to Latent Node 0, based on absolute Integrated Gradients (IG) scores from the ensemble attribution matrix. These top 10 genes are the strongest contributors to the biological signal captured by latent space 0.*

HPCAL1 (0.00349), PICALM (0.00342), COMMD8 (0.00331), SFT2D2 (0.00325), RNF130 (0.00320), are the primary drivers of the representation/signal captured by this latent space. Top 10 drivers of the representation from all 50 lanterns mentioned in **Appendix I**.

**3.3 Latent Variables Capture Distinct Gene Programs and Biological Pathways**

To characterize the biological meaning of the latent space learned by the VAE model, we analyzed gene-level attributions using Integrated Gradients. We computed mean attribution scores for each gene across all 50 latent variables (latent nodes) and selected the top 20 genes with the highest overall contributions **(Figure 3.3A)**. These included genes such as DDX43, FABP4, RAP1GAP2, KCNK5, XIST, ZNF839, CTH, ERC2, and PDK3, among others. Mean attribution scores across latents ranged from 0.0035 to 0.0055, with FABP4 and CTH contributing strongly to Latent 24 and 25, and ERC2 and ZNF839 dominating Latent 28, indicating distinct gene modules regulating each latent.

Hierarchical clustering of latent variables based on gene attribution profiles revealed modular structures, where sets of genes co-regulated subsets of latent nodes. For instance, Latents 24, 25, and 28 clustered closely and shared top-contributing genes involved in lipid metabolism and oxidative stress response, such as FABP4, CTH, and SAA2-SAA4.

**Figure 3.3B** illustrates the g:Profiler enrichment analysis of top-ranking genes from individual latent variables. Each dot represents a significantly enriched biological process, mapped to its corresponding latent node. Several latent variables were linked to distinct and functionally relevant pathways. For example, Latent 9 showed strong enrichment for DNA integration, suggesting potential involvement in genomic stability or viral interaction processes. Latent 20 was enriched for Golgi lumen acidification and Golgi-associated signaling, indicating a role in intracellular trafficking and post-translational modification. Latent 5 was associated with regulation of nervous system development, while Latent 33 was enriched for ECM-receptor interaction, pointing toward microenvironmental and adhesion-related mechanisms. Pathways related to RNA degradation (Latent 39), base excision repair (Latent 34), and neuromuscular junction development (Latent 45) were also identified, reflecting the biological diversity embedded within the latent dimensions.

The distribution of these enriched processes across latent nodes highlights the ability of the VAE model to capture biologically meaningful and distinct regulatory programs. A complete table of

**A close-up of a chart

AI-generated content may be incorrect. *Figure 3.3. Interpretation of VAE latent variables through gene attribution and pathway enrichment. (A)*** *Heatmap showing the mean Integrated Gradients attribution scores of the top 20 genes across all 50 latent variables. Both rows (genes) and columns (latents) were hierarchically clustered, revealing modular structures among gene-latent relationships.* ***(B)*** *Dot plot summarizing the most significantly enriched biological pathways for selected latent variables. Each dot represents a latent-pathway pair, with dot size and color corresponding to the enrichment significance (−log₁₀ adjusted p-value).*

enriched pathways, including adjusted p-values, enrichment scores, and associated gene sets for all 50 latents, is provided in **Appendix X**.

These findings demonstrate that VAE’s latent variables are not only mathematically structured but also biologically meaningful, capturing diverse processes such as development, signaling, immune response, metabolism, and DNA repair. This confirms the interpretability and relevance of the latent representations in modeling the underlying biology of OC.

**3.4 Functional Characterization of Latent Variables via GSEA**

To further evaluate the functional relevance of the latent space, we performed Gene Set Enrichment Analysis (GSEA) using the ranked gene attributions for each of the 50 latent variables and visualized the results in a pathway–latent heatmap **(Figure 3.4A)**. The heatmap displays the Normalized Enrichment Scores (NES) across a curated panel of KEGG pathways, capturing the direction and magnitude of enrichment. Red tones indicate positive enrichment (NES > 0), whereas blue tones indicate negative enrichment (NES < 0).

Several latent variables were significantly enriched for known cancer-related and immune-related pathways. Latent 6 and Latent 21 were positively enriched for *Ribosome* and *Oxidative Phosphorylation*, processes often upregulated in proliferative tumor cells. Latent 15 and Latent 24 showed strong positive enrichment in immune pathways such as *JAK-STAT signaling*, *Cytokine–cytokine receptor interaction*, and *Antigen processing and presentation*. Latent 36 and Latent 48 were associated with *Mismatch repair*, *Fanconi anemia*, and *Cell cycle*, indicating potential links to genomic instability. Negative enrichment was observed for several inflammation-related pathways (e.g., *Inflammatory bowel disease*, *Primary immunodeficiency*, *NF-kappa B signaling*), particularly in Latents 3, 9, and 18. Other Results of GSEA mentioned in **Appendix X.**

The diversity of enriched processes across the latent variables suggests that VAE captures a broad spectrum of biologically meaningful signatures, ranging from metabolic and proliferative programs to immune modulation and DNA repair pathways.

****

***Figure 3.4.******Pathway enrichment heatmap of VAE latent variables using GSEA.*** *Heatmap shows Normalized Enrichment Scores (NES) for pathways enriched across 50 latent variables. Each row represents a pathway and each column a latent node. Red shades indicate positive enrichment (NES > 0) and blue shades indicate negative enrichment (NES < 0).*

**3.5 Deep Learning-Based Classification of Candidate Driver Genes in Oropharyngeal Carcinoma**

To visualize the expression profiles of the 20 candidate driver genes across control and OC samples, we generated violin plots (Figure 3.5A). Several genes exhibited substantial differential expressions between the two groups. Notably, RAP1GAP2, CTH, and FABP4 were highly expressed in OC samples compared to controls, suggesting their potential role as diagnostic or functional markers. Conversely, genes like XIST and ERC2 displayed more variable patterns, hinting at subtype-specific or microenvironmental influences.

We then assessed the ability of the 20-gene panel to classify OC using supervised deep learning models. As shown in the performance plots (Figure 3.5B & C), the MLP model consistently outperformed CNN and LSTM across all evaluation folds. The MLP achieved a mean AUPRC of 0.86 and mean AUROC of 0.80, followed by the CNN with an AUPRC of 0.81 and AUROC of 0.79, and the LSTM with an AUPRC of 0.78 and AUROC of 0.70. Precision–recall curves showed a steeper and more stable shape for the MLP, indicating its superior ability to maintain high precision at varying recall thresholds. Similarly, the MLP's ROC curve demonstrated a better trade-off between sensitivity and specificity compared to the other models.

These results indicate that the MLP model is best suited for classifying OC based on the selected latent-informed gene set. The consistently high AUPRC and AUROC suggest that the VAE-derived genes, particularly RAP1GAP2, PDK3, and FABP4, may serve as effective driver markers or classifiers for oropharyngeal carcinoma in high-throughput transcriptomic data. This model framework provides a biologically informed, interpretable route to gene-based diagnosis or biomarker discovery, bridging unsupervised latent feature learning with supervised validation on independent clinical cohorts.

**3.6 RAP1GAP2 Emerges as the Most Predictive Gene in Single-Feature Classification Models**

To identify the most predictive gene within the consensus panel, we trained single-feature models for each of the 20 genes and computed their individual feature importances using the supervised MLP model described previously. The resulting importance scores are visualized in **Figure 3.6A**, where RAP1GAP2 ranked as the most informative gene, followed closely by XIST, SLC9A7, and

**Table 3.1 Performance metrics for single-gene classification models**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Gene | AUROC | AUPRC | Accuracy | F1 | Precision | Recall |
| WDR41 | 0.640 | 0.650 | 0.556 | 0.711 | 0.560 | 0.971 |
| KCNK5 | 0.540 | 0.560 | 0.524 | 0.686 | 0.545 | 0.924 |
| GATA3 | 0.590 | 0.580 | 0.620 | 0.667 | 0.657 | 0.676 |
| DDX43 | 0.550 | 0.570 | 0.513 | 0.629 | 0.550 | 0.733 |
| DTWD1 | 0.650 | 0.640 | 0.535 | 0.679 | 0.554 | 0.876 |
| XIST | 0.594 | 0.708 | 0.540 | 0.688 | 0.556 | 0.905 |
| SAA2-SAA4 | 0.557 | 0.621 | 0.556 | 0.709 | 0.561 | 0.962 |
| ERC2 | 0.550 | 0.560 | 0.610 | 0.709 | 0.610 | 0.848 |
| ENO3 | 0.610 | 0.590 | 0.567 | 0.722 | 0.565 | 1.000 |
| SLC9A7 | 0.710 | 0.710 | 0.594 | 0.689 | 0.604 | 0.800 |
| CTH | 0.590 | 0.590 | 0.604 | 0.711 | 0.603 | 0.867 |
| PDK3 | 0.570 | 0.600 | 0.567 | 0.675 | 0.583 | 0.800 |
| GSTT2 | 0.643 | 0.664 | 0.642 | 0.735 | 0.628 | 0.886 |
| RPARP-AS1 | 0.528 | 0.619 | 0.540 | 0.699 | 0.552 | 0.952 |
| ZNF839 | 0.550 | 0.590 | 0.556 | 0.711 | 0.560 | 0.971 |
| FABP4 | 0.520 | 0.530 | 0.535 | 0.695 | 0.550 | 0.943 |
| RAP1GAP2 | 0.710 | 0.769 | 0.730 | 0.760 | 0.700 | 0.940 |
| HAPLN1 | 0.606 | 0.676 | 0.615 | 0.692 | 0.628 | 0.771 |

**A diagram of a gene

AI-generated content may be incorrect.**

***Figure 3.5. Identification of key driver genes and classification performance in oropharyngeal carcinoma. (A)*** *Violin plots showing the expression distributions of 20 consensus genes, derived from VAE latent space attribution scores, across control and oropharyngeal OC samples in an external high-throughput RNA-seq dataset.* ***(B)*** *Precision–Recall (PR) curves comparing three supervised deep learning models trained on the expression profiles of the 20 genes.* ***(C)*** *Receiver operating characteristic (ROC) curves for the same models.*

FABP4. This suggests that RAP1GAP2 holds strong discriminative power for separating oropharyngeal carcinoma from control samples, reinforcing its prominence in both latent attribution analysis and expression profiling.

To validate its predictive strength, we constructed a single-gene MLP classifier using only the expression values of RAP1GAP2. The resulting Precision–Recall curve, shown in **Figure 3.6B**, achieved a mean AUPRC of 0.769, indicating robust classification performance using this gene alone. This further supports the hypothesis that RAP1GAP2 may serve as a potent driver or biomarker of oropharyngeal carcinoma and warrants further experimental validation.

**3.7 RAP1GAP2 Emerges as a Key Latent Driver Despite Non-Significance in Differential Expression Analysis**

Figure 3.7A & B show the gene expression distributions of the RNA-seq datasets before and after normalization, respectively. **Figure 3.7A** illustrates the raw, unnormalized transcript counts, highlighting variability across samples. In contrast, **Figure 3.7B** demonstrates the effect of DEseq2 normalization, resulting in more comparable and standardized expression profiles across all samples, ensuring the reliability of downstream analyses.

To explore hidden regulatory signals not captured by traditional methods, we applied a deep learning framework trained on latent representations derived from transcriptomic data. Among the most notable findings was the gene RAP1GAP2, which consistently ranked as a top contributor across all 50 latent variables. Moreover, when used in supervised classification tasks, RAP1GAP2 exhibited the highest predictive performance among the 20 common genes shared across latent dimensions.

However, differential gene expression (DGE) analysis failed to identify RAP1GAP2 as statistically significant. As shown in **Figure 3.7C**, RAP1GAP2 resides within the "not significant" region of



***Figure 3.6. RAP1GAP2 identified as the top predictive gene for oropharyngeal carcinoma classification. (A)*** *Feature importance scores for each of the 20 genes in the supervised MLP model. RAP1GAP2 ranked highest, suggesting its dominant role in classification.* ***(B)*** *Precision–Recall curve for the single-gene classifier trained exclusively on RAP1GAP2 expression. The model achieved a mean area under the precision–recall curve (AUPRC) of 0.769, indicating strong predictive capacity from this gene alone.*

the volcano plot, indicating that it was not differentially expressed based on standard thresholds (log2 fold change and FDR-adjusted p-value). This contrast between latent-space importance and conventional differential expression underscores a critical disconnect between statistical significance and functional relevance.



***Figure 3.7. Identification of RAP1GAP2 as a latent driver despite non-significance in differential expression analysis. (A)*** *Raw gene expression across samples before normalization.* ***(B****) Normalized expression profiles of all samples.* ***(C)*** *Volcano plot of differential gene expression analysis: upregulated, downregulated, and non-significant genes are shown. RAP1GAP2, highlighted in red, was not significantly differentially expressed but was identified as a top contributor across all latent variables and showed the highest classification ability in the deep learning model, supporting its role as a hidden driver in oropharyngeal carcinoma.*

**Chapter Four**

**Discussion**

In this study, we applied the VAE framework to uncover biologically meaningful patterns in OC transcriptomic data using an unsupervised deep learning approach. Our goal was to move beyond traditional differential gene expression analysis and instead learn latent representations that reflect underlying biological programs. The results demonstrate that VAE effectively captures interpretable and functionally relevant gene modules, offering insights into cancer heterogeneity, immune activity, and DNA repair mechanisms.

**Chapter Two**

In Figure 1, we present the overall architecture of the VAE pipeline, which integrates a variational autoencoder (VAE) with Integrated Gradients attribution to produce interpretable latent features [1]. This unsupervised learning strategy enables the discovery of gene programs without requiring predefined phenotypes or differential expression thresholds, mitigating common biases and enhancing generalizability across datasets [2].

The attribution heatmap in Figure 2 shows that specific latent variables are strongly influenced by distinct sets of genes, with hierarchical clustering revealing modular structures in gene–latent interactions. Genes such as *FABP4*, *CTH*, and *ZNF839* contributed prominently to Latents 24, 25, and 28, suggesting that these latent features may reflect shared biological roles in metabolism, immune signaling, or redox processes [3].

To further validate the biological relevance of the latent variables, we conducted pathway enrichment analysis using both Gene Set Enrichment Analysis (GSEA) [4] and g:Profiler [5]. As shown in Figure 3, each latent variable was associated with distinct functional themes. Latent 1, for example, was enriched for *nervous system development*, while Latents 7 and 35 captured signaling and repair pathways such as *glucagon signaling* and *base excision repair*. These findings highlight the capacity of VAE to learn interpretable representations that correspond to coherent gene programs and known biological processes.

In Figure 4, we extended this analysis using a GSEA heatmap of KEGG pathways across all 50 latent variables. This global view revealed patterns consistent with cancer biology. Several latent variables were positively enriched for core oncogenic and immune pathways—including *JAK-STAT signaling*, *oxidative phosphorylation*, and *ribosome biogenesis*—while others were associated with *mismatch repair*, *Fanconi anemia*, and *cytokine–cytokine receptor interaction*. Interestingly, pathways related to viral infection and immune evasion (e.g., *Herpes simplex virus*, *Influenza A*) were also captured, suggesting a role of viral mimicry or immune dysregulation in OC progression [6].

Together, these results demonstrate that VAE can identify latent biological dimensions that align with known hallmarks of cancer, such as immune modulation, genomic instability, and altered metabolism [7]. Importantly, the use of both GSEA and g:Profiler enables cross-validation of pathway-level signals through complementary enrichment strategies—rank-based and over-representation—thereby increasing confidence in the functional interpretations [5].

**The observation that RAP1GAP2, a gene known to be involved in tumor invasion and progression, was not flagged by DGE but emerged as a key driver in our deep learning analysis provides compelling evidence for the added value of representation learning in transcriptomics. Conventional DGE methods are inherently limited to detecting genes with strong average shifts in expression between groups, often overlooking context-dependent regulatory hubs that drive phenotypic variation.**

**Our results demonstrate that deep learning models, by learning complex, non-linear patterns across the transcriptome, can uncover biologically significant features invisible to traditional tools. The identification of RAP1GAP2 as a latent driver—despite its lack of differential expression—validates the central hypothesis of this study: key molecular drivers may not exhibit overt expression differences but can still play pivotal roles in disease pathology.**

**This finding challenges the over-reliance on DGE as the primary tool for gene discovery in cancer research and supports the integration of deep representation learning into biomarker identification pipelines. Such approaches have the potential to reveal novel therapeutic targets that would otherwise remain undetected.**

**Chapter Two**

**A chart of bar code

AI-generated content may be incorrect.**

**A graph showing a line graph

AI-generated content may be incorrect.**