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

A logo with text and symbols

Description automatically generated

**A dissertation submitted to the Shahjalal University of Science and Technology in partial fulfillment of the requirements for the Bachelor of Science (BSc) Degree in Biochemistry and Molecular Biology**

**Submitted By**

**Registration No:** 2019433077

**MS Session:** 2019-2020

**7th July 2025**

**Department of Biochemistry and Molecular Biology**

**School of Life Science**

**Shahjalal University of Science and Technology**

**Sylhet-3114, Bangladesh**

**Dedicated To,**

**My Beloved Parents**

**My Respected Teachers**

**Acknowledgments**

Foremost, I want to express my gratitude to Almighty Allah for the wisdom, strength, peace of mind, and good health. He bestowed upon me to complete my dissertation.

I place on record my sincere gratitude to my supervisor **Papia Rahman**, Lecturer, Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet-3114, for allowing me the opportunity to enroll in my project under her kind supervision and for her patience, motivation, and immense knowledge. Her guidance helped me throughout the research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis study. I am extremely thankful and indebted to her valuable guidance and encouragement.

Finally, I must express my very profound gratitude to my parents and friends for providing me with unwavering support and endless encouragement during my studies as well as during the process of researching and writing this thesis. This accomplishment would not have been possible without them.

**Joy Prokash Debnath**

7th July 2025

**To Whom It May Concern**

I hereby certify that in accordance with the laws of Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh, the project work entitled **“Beyond Differential Expression: Deep Neural Profiling Reveals RAP1GAP2 as a Latent Regulator of Tumor Invasion in Oropharyngeal Carcinoma”**described here is entirely own work of **Joy Prokash Debnath** bearing Registration No. **2019433077**, **Session:** 2019-2020. This project does not contain any materials which were previously published or written by another person, except duly referred. The work was conducted under my supervision and was enrolled in the degree of Bachelor of Science in Biochemistry and Molecular Biology at the Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh. All information provided in the project paper has been obtained and presented following academic rules and ethical guidelines.

I hereby endorse his project to be submitted for evaluation.

---------------------------------

Signature of the Supervisor

**Papia Rahman**

**Lecturer**

Department of Biochemistry and Molecular Biology (BMB)

Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

**Abbreviations**

|  |  |
| --- | --- |
| Abbreviation | Full Form |
| OC | **Oropharyngeal Carcinoma** |
| DGE | **Differential Gene Expression** |
| PCA | **Principal Component Analysis** |
| PEM | **Probabilistic Embedding Model** |
| SBS | **Sensitivity-Based Scoring** |
| GSEA | **Gene Set Enrichment Analysis** |
| IG | **Integrated Gradients** |
| GEO | **Gene Expression Omnibus** |
| NES | **Normalized Enrichment Score** |
| AUROC | **Area Under the Receiver Operating Characteristic curve** |
| AUPRC | **Area Under the Precision–Recall Curve** |
| KEGG | **Kyoto Encyclopedia of Genes and Genomes** |
| GO | **Gene Ontology** |

# **Abstract**

**Background:** Oropharyngeal carcinoma (OC) progression is driven by complex molecular changes that are not fully captured by traditional differential gene expression analysis. Variational autoencoders (VAEs) offer a deep learning approach to uncover hidden patterns in high-dimensional transcriptomic data, potentially revealing latent molecular drivers of cancer that conventional methods miss.

**Methods:** We aggregated multi-cohort OC gene expression datasets and trained a VAE to compress the data into a low-dimensional latent space. Integrated Gradients, an explainable AI attribution technique, was applied to quantify the contribution of each gene to each latent feature. Genes with consistently high attribution scores across latent dimensions were identified as candidate regulators. We further evaluated these genes through pathway enrichment and classification analyses to interpret their biological significance.

**Results:** The VAE learned biologically meaningful latent features, and Integrated Gradients highlighted a panel of genes with strong contributions to these features. Notably, RAP1GAP2 consistently ranked among the top contributors across all 50 latent variables. Despite showing no significant differential expression in tumors versus normal samples, RAP1GAP2 demonstrated the highest latent-space importance and strong discriminative power for distinguishing OC, achieving an area under the precision–recall curve of 0.769. Biological interpretation suggests that RAP1GAP2, a Rap1 GTPase-activating protein, may promote tumor invasion by inactivating Rap1 and thereby influencing MAPK signaling and Golgi-mediated secretion.

**Conclusion:** Our deep learning framework uncovered RAP1GAP2 as a latent driver in oropharyngeal carcinoma, illustrating how VAEs combined with Integrated Gradients can identify hidden molecular regulators overlooked by traditional analyses. This approach provides new insights into OC tumor biology that could inform future research and therapy.

**Keywords:** Oropharyngeal carcinoma; transcriptomics; variational autoencoder; deep learning; integrated gradients; latent features; RAP1GAP2; Rap1 signaling; MAPK pathway; Golgi secretio

**Table of Contents**

**Contents Page No.**

[Abstract 6](#_Toc202657524)

[Keywords: 6](#_Toc202657525)

[List of Tables 8](#_Toc202657526)

[List of Figures 9](#_Toc202657527)

[Introduction 10](#_Toc202657528)

[1.1 Significance of Oropharyngeal Carcinoma 10](#_Toc202657529)

[1.2 Complexity of Cancer Biology and Analytical Gaps 11](#_Toc202657530)

[1.3 Deep Learning for Latent Feature Discovery 11](#_Toc202657531)

[1.4 Gene Attribution with Integrated Gradients 12](#_Toc202657532)

[1.5 Revealing Hidden Drivers: The Case of *RAP1GAP2* 13](#_Toc202657533)

[1.6 Hypothesis and Objective 14](#_Toc202657534)

[Material and Methods 16](#_Toc202657535)

[2.1 Overview of the Study 16](#_Toc202657536)

[2.2 Datasets Retrieval 17](#_Toc202657537)

[2.3 Data Integration, Batch Effect Removal and Preprocessing 19](#_Toc202657539)

[2.4 Training Deep Neural Network Models 19](#_Toc202657540)

[2.4.1 Datasets Merging and Standardization 19](#_Toc202657541)

[2.4.2 Traditional Deep Learning Model 20](#_Toc202657542)

[2.4.3 Additional Sample Distribution 20](#_Toc202657543)

[2.5 Neural Network Design and Hyperparameter Optimization 21](#_Toc202657544)

[2.5.1 Train Model with Adam Optimizer 21](#_Toc202657545)

[2.5.2 Cross validate and Extract Best Latent Dimension 21](#_Toc202657546)

[2.6 Learning Robust Latent Representations 21](#_Toc202657547)

[2.7 Gene Attribution and Pathway Analysis 22](#_Toc202657548)

[2.7.1 Sensitivity-Based Scoring (SBS) for Gene-to-Latent Attribution 22](#_Toc202657549)

[2.7.2 Pathway Enrichment Analysis of Latent Variable-Associated Genes 22](#_Toc202657550)

[2.7.3 Gene Set Enrichment Analysis (GSEA) 22](#_Toc202657551)

[2.8 Supervised Deep Learning Model Training 23](#_Toc202657552)

[2.8.1 Gene Selection and Data Collection 23](#_Toc202657553)

[2.8.2 Normalization and Batch Correction 23](#_Toc202657554)

[2.8.3 Model Development and Training 23](#_Toc202657555)

[2.8.4 Evaluation and Visualization 23](#_Toc202657556)

[2.9 Differential Gene Expression analysis 24](#_Toc202657557)

[Result 25](#_Toc202657558)

[3.1 Data Preprocessing and Quality Assessment 25](#_Toc202657559)

[3.2 Latent Space Extraction Using Deep Neural Network 26](#_Toc202657560)

[3.3 Latent Variables Capture Distinct Gene Programs and Biological Pathways 29](#_Toc202657561)

[3.4 Functional Characterization of Latent Variables via GSEA 31](#_Toc202657562)

[3.5 Deep Learning-Based Classification of Candidate Driver Genes in Oropharyngeal Carcinoma 33](#_Toc202657563)

[3.6 RAP1GAP2 Emerges as the Most Predictive Gene in Single-Feature Classification Models 33](#_Toc202657564)

[3.7 RAP1GAP2 Emerges as a Key Latent Driver Despite Non-Significance in Differential Expression Analysis 36](#_Toc202657565)

[Discussion 40](#_Toc202657566)

[4.1 Significance of Identified Latent Drivers 41](#_Toc202657567)

[4.2 Proposed Molecular Mechanism of RAP1GAP2 in Invasion 42](#_Toc202657568)

[4.3 Limitations of the Study 43](#_Toc202657569)

[4.4 Future Directions 45](#_Toc202657570)

[Conclusion 47](#_Toc202657571)

[References 49](#_Toc202657572)

[Appendices 50](#_Toc202657573)

# **List of Tables**

|  |  |  |
| --- | --- | --- |
| **SL. NO.** | **Table Captions** | **Page No.** |
| **Table 2.1** | Expression Profiling Datasets for OC | 5 |
| **Table 3.1** | Performance metrics for single-gene classification models | 32 |

# **List of Figures**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SL. NO.** | | **Figure Captions** | **Page No.** | |
| **Figure 2.1** | Overview of the study pipeline | | | 2 |
| **Figure 3.1** | Preprocessing and PCA of gene expression data | | | 3 |
| **Figure 3.2** | Model Performance and Gene Attribution | | | 4 |
| **Figure 3.3** | Interpretation of latent variables via gene attribution and enrichment | | | 7 |
| **Figure 3.4** | Pathway enrichment heatmap using GSEA | | | 10 |
| **Figure 3.5** | Identification of key driver genes and classification performance | | | 12 |
| **Figure 3.6** | RAP1GAP2 as a predictive gene for OC classification | | | 13 |
| **Figure 3.7** | RAP1GAP2 as latent driver despite non-significance in DGE | | | 15 |
| **Figure 4.1** | Mechanistic role of RAP1GAP2 in OC invasion and metastasis | | | 17 |

**Chapter One**

# **Introduction**

## **1.1 Significance of Oropharyngeal Carcinoma**

Oropharyngeal carcinoma (OC) is a form of head and neck cancer with considerable clinical importance. Its incidence has been rising in recent decades, driven in part by human papillomavirus (HPV) infection, making HPV-positive oropharyngeal squamous cell carcinoma one of the most rapidly increasing malignancies in many high-income countries(Lechner et al., 2022). OC often presents at advanced stages due to subtle early symptoms, resulting in significant morbidity and mortality. There is therefore a pressing need for improved understanding of the molecular underpinnings of OC to enable earlier detection, better patient stratification, and more effective, precision therapies [2]. Despite advances in surgery, radiation, and systemic treatments, outcomes for advanced OC remain guarded. A deeper insight into the tumor’s transcriptomic landscape could reveal novel molecular driver to improve patient management.

## **1.2 Complexity of Cancer Biology and Analytical Gaps**

The biology of cancer is extraordinarily complex, involving non-linear interactions among genes and pathways that drive tumor behavior. Traditional differential gene expression (DGE) analysis – which typically relies on linear models or statistical tests to find genes individually up- or down-regulated in tumors – has clear limitations when faced with this complexity. DGE methods excel at identifying genes with large average expression changes, but they may overlook hidden drivers that exert their effects through subtle or combinatorial patterns. In other words, patient subgroups or tumor phenotypes could be determined by gene sets that do not show obvious one-at-a-time differences and thus remain “invisible” to linear DGE approaches [3]. Indeed, recent work has cautioned that when nonlinear machine learning models identify patient groupings, the defining gene signatures might be missed by conventional DGE due to its linear nature [3]. This gap is problematic because such overlooked genes or gene interactions could be critical for cancer progression. There is a need for analytical strategies that move beyond linear assumptions and can capture the nonlinear dependencies in gene expression data. Explainable artificial intelligence (AI) offers a potential solution: by applying interpretability methods to complex models, one can uncover multivariate gene patterns that would otherwise escape notice [4]. In summary, overcoming the limitations of DGE requires methods capable of modeling and explaining the intricate, nonlinear relationships that characterize cancer biology.

**1.3 Deep Learning for Latent Feature Discovery**

To address these challenges, we turn to deep learning—specifically, unsupervised deep neural networks to learn biologically meaningful latent variables from transcriptomic data. Variational autoencoders (VAEs) are a class of deep generative models well-suited for this task. A VAE compresses high-dimensional gene expression profiles into a lower-dimensional latent space, while preserving as much information as possible. This process can distill complex gene expression patterns into a set of latent features that capture underlying biological signals and patient heterogeneity. VAEs and related autoencoder techniques have been successfully applied to large-scale gene expression datasets, where they have demonstrated the ability to model non-linear gene interactions and improve outcome predictions [5, 6]. For example, Rampášek et al. showed that a VAE-based model (“Dr.VAE”) could learn latent representations of cancer cell line expression data that improve drug response prediction, highlighting the power of deep learning to capture subtle transcriptomic effects of treatment [7]. Similarly, Way et al. used VAE to compress pan cancer gene expression data and found that varying the latent dimensionality allowed different biological signals (such as pathway activities and mutational status) to emerge, indicating that deep compression can learn complementary aspects of tumor biology beyond what a single linear compression or DGE analysis can achieve [8, 9]. These studies underscore that deep neural networks can extract non-linear features from gene expression data, potentially revealing patterns that are not evident to traditional methods.

However, a known drawback of deep learning models is their limited interpretability – the latent features or learned representations are “black boxes” without clear biological meaning. In the context of cancer transcriptomics, it is not enough to discover latent variables; we also need to understand which genes those variables represent or how they relate to known biology. Simply compressing data with a VAE might yield abstract features that correlate with disease, but without interpretation we cannot translate those features into testable biological insights. This motivates integrating explainable AI techniques into our deep learning pipeline.

## **1.4 Gene Attribution with Integrated Gradients**

To interpret the latent space and connect it back to gene-level biology, we employ integrated gradients, a robust feature attribution method for neural networks. Integrated gradients provide a way to quantify the contribution of each input feature (in this case, each gene’s expression) to a given output or latent variable in the model [10]. Formally, integrated gradients work by integrating the gradients of the model’s output with respect to inputs along a path from a baseline to the actual input, yielding an attribution score for every feature that satisfies desirable axioms of fairness and sensitivity [10]. Introduced by Sundararajan et al. in 2017, this method has become a popular tool for explaining deep learning predictions in various domains [11]. In our study, we harness integrated gradients to attribute genes to latent variables learned by the VAE and to any downstream predictive outputs. This approach effectively “opens the black box” of the autoencoder by highlighting which genes most strongly influence each latent dimension of the model.

Notably, prior research has demonstrated the utility of combining deep generative models with feature attribution in genomics. Dincer et al., for example, applied integrated gradients to the latent features of a VAE trained on cancer gene expression data, identifying the top contributing genes for each latent dimension [12]. Such post hoc interpretation of latent space enables researchers to anchor abstract features in concrete biology—for instance, a latent dimension might turn out to represent a pathway or cell cycle signature, based on the genes with highest attributions. Building on these ideas, we use integrated gradients in tandem with our VAE to derive gene-level importance scores for the learned latent factors. By doing so, we can determine which genes drive the variations captured in the latent space and are most relevant to distinguishing oropharyngeal tumors from controls (or different subtypes of tumors). This synergy between unsupervised deep learning and explainability techniques allows us to uncover biologically meaningful patterns that would be invisible to linear analysis alone, while also retaining interpretability.

## **1.5 Revealing Hidden Drivers: The Case of *RAP1GAP2***

Applying this deep learning framework to OC transcriptomic data yields novel insights into the disease’s molecular drivers. The variational autoencoder extracts latent variables that summarize gene expression patterns across tumors, and integrated gradients pinpoint the genes that define those variables. Intriguingly, our analysis identifies RAP1GAP2 as a key latent driver gene in oropharyngeal carcinoma. RAP1GAP2 encodes a GTPase-activating protein that regulates the small GTPase Rap1, a signaling molecule involved in cell proliferation and adhesion. In our model, RAP1GAP2 stands out with a high attribution score: it is among the top contributors to a latent feature that is highly predictive of OC presence. This finding is remarkable because RAP1GAP2 was not flagged as significant in conventional differential expression analysis—in other words, its expression levels do not differ enough between tumor and normal on average to pass typical statistical thresholds. Traditional DGE would have missed RAP1GAP2 entirely, yet our deep learning approach uncovered it as an important player with a nonlinear contribution to the tumor transcriptome. This result highlights the advantage of using a non-linear model: RAP1GAP2’s impact emerges only when considering complex interactions captured in the latent space, illustrating how deep learning can reveal “hidden” drivers that evade linear scrutiny.

From a biological standpoint, the implication of RAP1GAP2 in OC is plausible and generates new hypotheses. Although RAP1GAP2 itself has not been well-studied in oropharyngeal cancer, it belongs to the same family as Rap1GAP (also known as RAP1GAP1), which has been reported to act as a tumor suppressor in squamous cell carcinoma. In fact, restoring Rap1GAP expression in OC cell lines was shown to reduce active Rap1 signaling and significantly slow tumor growth in vivo [13]. This prior evidence of the Rap1 pathway’s involvement in head and neck cancer provides context for our findings: it suggests that downregulation or dysregulation of Rap1-inhibitory proteins (like Rap1GAP or RAP1GAP2) could contribute to oncogenic processes in the oropharynx. Our discovery of RAP1GAP2 as a latent driver, despite its subtle expression changes, underscores how deep learning-based analysis can pinpoint functionally relevant genes that conventional analyses deem insignificant. Such genes might represent early changes or context specific vulnerabilities that are missed when focusing only on large fold-changes. Identifying RAP1GAP2 as highly predictive of OC opens the door to further experimental validation and investigation into its potential role in tumor suppression or as a biomarker for disease presence.

## **1.6 Hypothesis and Objective**

In light of the above considerations, we hypothesize that deep neural network models can expose latent transcriptomic features and gene drivers of oropharyngeal carcinoma that are not detectable through traditional linear methods. By capturing nonlinear combinations of gene expression changes, these models may reveal hidden biological signals critical to OC pathogenesis. To test this hypothesis, the objective of our study is to integrate unsupervised latent space modeling with supervised gene attribution to identify novel gene-level drivers of oropharyngeal carcinoma. Specifically, we employ a variational autoencoder to learn a compressed representation of OC transcriptomic data and then apply integrated gradients to attribute genes to the learned latent features (and to any predictive model built upon them). Through this approach, we aim to uncover previously unrecognized genes and pathways that drive oropharyngeal cancer, demonstrating the power of deep learning to advance our molecular understanding of this disease beyond the limits of differential expression analysis.

**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.

## **2.4 Training Deep Neural Network Models**

### **2.4.1 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.

### **2.4.2 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:

### **2.4.3 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.

## **2.5 Neural Network Design and Hyperparameter Optimization**

### **2.5.1 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.

### **2.5.2 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.**

## **2.6 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.

## **2.7 Gene Attribution and Pathway Analysis**

### **2.7.1 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.

### **2.7.2 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.

### **2.7.3 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.

## **2.8 Supervised Deep Learning Model Training**

### **2.8.1 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.

### **2.8.2 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.

### **2.8.3 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.

### **2.8.4 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.

## **2.9 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**

# **Result**

## **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**

## **4.1 Significance of Identified Latent Drivers**

Our integrative deep learning approach uncovered a set of latent driver genes that illuminate hidden aspects of oropharyngeal carcinoma biology. The most notable finding is the identification of RAP1GAP2 as a pivotal molecular driver of tumor invasion, despite this gene not being flagged by conventional differential expression analysis. This underscores a key insight: important regulatory genes can evade traditional statistical detection yet play outsized roles in cancer progression. By leveraging variational autoencoders and attribution techniques, we revealed that RAP1GAP2 consistently ranked among the top contributors to latent features associated with the tumor phenotype. In practical terms, RAP1GAP2 alone exhibited a strong ability to distinguish tumor samples from normals (single-gene classifier AUPRC ≈ 0.77), highlighting its potential as a latent oncogenic driver in oropharyngeal carcinoma. Notably, this discovery leads to a new understanding of our research question – it demonstrates that “beyond differential expression” lies a layer of latent genomic regulation where genes like RAP1GAP2 can orchestrate malignant behavior without dramatic expression changes.

Importantly, our findings align with and extend existing knowledge in the field. The Ras-related GTPase **Rap1** and its regulators have long been implicated in cancer cell adhesion and motility [1, 2]. In head and neck cancers, active Rap1 signaling has been shown to **enhance invasive behavior** (e.g. via β-catenin and MMP7 induction) [3], whereas the canonical Rap1 inactivator **Rap1GAP** (a paralog of RAP1GAP2) is known to act as a tumor suppressor by curbing Rap1–ERK signaling and tumor growth [4, 5]. Our identification of RAP1GAP2 thus builds upon this paradigm but also introduces a novel nuance. Unlike Rap1GAP, which broadly suppresses HNSCC progression [5], **RAP1GAP2 emerges here as a pro-invasion factor**. This apparent contradiction can be reconciled by context: Rap1 regulators often have pleiotropic, cell-type-specific effects [6]. Indeed, literature reports that while Rap1GAP generally inhibits invasion in many cancers, in certain contexts increased Rap1GAP activity can *paradoxically* promote invasive capacity [6]. Our results suggest that oropharyngeal carcinoma represents such a context where RAP1GAP2, operating in a specific subcellular niche, drives malignant progression. This is a **novel insight** because RAP1GAP2 has not been previously spotlighted in oropharyngeal cancer; it was essentially a “hidden” driver that our latent-space profiling was able to unmask. In addition to RAP1GAP2, we found other latent drivers like **PDK3** and **FABP4** that reinforce the biological validity of our approach. PDK3 (pyruvate dehydrogenase kinase 3) is a known mediator of the Warburg effect and is upregulated in hypoxic tumors, contributing to metabolic reprogramming and aggressive behavior [7]. Likewise, **FABP4** (fatty acid–binding protein 4) has been recognized as a promoter of tumor metastasis and therapy resistance by fueling lipid transport and signaling in cancer cells [8]. The presence of PDK3 and FABP4 among our top latent genes suggests our model captured key hallmarks of cancer – metabolic plasticity and microenvironmental adaptation – in addition to the Rap1 signaling axis. In summary, the convergence of our data-driven discoveries with known cancer pathways lends credence to the **outcomes of this study**: we have not only identified a novel driver (RAP1GAP2) and supporting cast of genes for oropharyngeal carcinoma invasion, but also demonstrated the power of deep neural profiling to reveal biologically meaningful targets that standard analyses would overlook.

## **4.2 Proposed Molecular Mechanism of RAP1GAP2 in Invasion**

To translate these findings into a biological context, we propose a mechanistic model for RAP1GAP2’s role in promoting tumor invasion and metastasis (Figure 4.1). **RAP1GAP2 is a GTPase-activating protein** that targets Rap1, and our model posits that its upregulation in tumor cells leads to **inactivation of Rap1 signaling**. This has several downstream consequences conducive to invasion. First, turning **Rap1 “OFF”** can destabilize cell adhesions: active Rap1 normally strengthens integrin- and cadherin-mediated adhesion [9,1], so its inactivation may loosen tumor cells from each other and from the extracellular matrix, facilitating migration. Second, Rap1 inactivation by RAP1GAP2 might tilt the balance toward other oncogenic pathways such as Ras–MAPK signaling. There is crosstalk between Rap1 and Ras; for instance, loss of Rap1 activity in certain settings can lead to unchecked Ras/ERK activation [4, 2]. Consistent with this, our figure illustrates that **RAP1GAP2-driven Rap1 suppression could unleash MAPK/ERK signaling**, enhancing cell proliferation and motility. Third, and most uniquely, RAP1GAP2 appears to act at the **Golgi and vesicular trafficking level** to promote a pro-metastatic secretory phenotype. This protein is known to form a complex with synaptotagmin-like protein 1 (Slp1) and Rab27 on secretory vesicles, thereby regulating their release [10, 11]. In platelets, RAP1GAP2 binding to Slp1 actually **augments granule secretion** [12]. By analogy, in carcinoma cells RAP1GAP2 may stimulate the secretion of matrix-degrading enzymes (e.g., MMPs) and pro-invasive cytokines via Golgi-derived vesicles. The **Golgi-mediated secretion** of such factors would further enable tumor cells to remodel the surrounding stroma and invade adjacent tissues. Taken together, our model suggests that RAP1GAP2 exerts a multifaceted influence: it simultaneously *frees* the tumor cell from adhesion constraints, *activates* mitogenic signaling, and *triggers* the release of invasive effectors. This integrated mechanism offers a conceptual explanation for how RAP1GAP2, despite acting as a Rap1 inhibitor, ends up **promoting invasion and metastasis in oropharyngeal carcinoma**. This proposal not only interprets our results in light of molecular biology but also provides testable hypotheses (for example, that RAP1GAP2 high-expressing cells will show reduced Rap1-GTP levels, enhanced ERK phosphorylation, and increased exocytosis of pro-invasive granules). It is a stepping stone for future research to validate and refine the exact biochemical pathways involved.

A diagram of a cell line

AI-generated content may be incorrect.

***Figure 4.1. Schematic model illustrates the proposed role of RAP1GAP2 in promoting invasion and metastasis in oropharyngeal carcinoma through Rap1 inactivation, MAPK signaling, and Golgi-mediated secretion.***

## **4.3 Limitations of the Study**

While our study provides compelling computational evidence for RAP1GAP2 and other drivers in OC, several limitations must be acknowledged.

Lack of Functional Validation: Our conclusions are drawn from in silico analyses of transcriptomic data. We did not perform laboratory experiments to directly test RAP1GAP2’s effect on cell invasion or metastasis. This means causality remains inferential – for instance, whether elevating or silencing RAP1GAP2 actually alters invasive behavior in vitro is yet to be confirmed.

Retrospective Data Integration: We pooled multiple public datasets (spanning different platforms and sub-cohorts), which introduces heterogeneity. Even though we corrected for batch effects and focused on common genes, subtle biases or confounding variables (e.g. varying proportions of HPV-positive tumors, treatment status, or anatomic subsite differences) could influence the latent features. The term “oropharyngeal carcinoma” in our study encompasses samples that might come from slightly different contexts; thus, the generalizability to a specific clinical population needs caution.

Statistical Significance vs. Biological Significance: By design, our latent driver discovery looked beyond conventionally significant differential expression. This means some identified genes (like RAP1GAP2) showed only moderate expression changes. There is a risk that such genes could be false positives from modeling, or that their importance manifests only in combination with others. We mitigated this by cross-validating in a classifier, but the results still require biological interpretation – high predictive power does not always equate to being a solo driver of disease.

Clinical Data Limitations: Key clinical endpoints (e.g. actual metastasis occurrence, patient survival) were limited or not uniformly available across the GEO datasets. Therefore, we correlated our latent drivers mostly with sample class (tumor vs. normal, or metastatic vs. non-metastatic in a small subset) rather than long-term outcomes. This limits our ability to claim prognostic value.

Focus on mRNA Level: Our study centered on gene expression. We did not integrate other layers such as protein expression, mutations, or epigenetic modifications. It’s possible that post-transcriptional regulation of RAP1GAP2 or activity of Rap1 pathway components (through mutations or phosphorylation states) play a role in OC that our transcriptome-based method could not capture.

By recognizing these limitations, we ensure that the interpretations remain measured and avoid overstatement. They highlight the need for further research to deepen and validate our findings.

## **4.4 Future Directions**

This work opens several avenues for future investigation to build on our findings and address the above limitations:

* **Experimental Validation:** A top priority is to validate RAP1GAP2’s function in oropharyngeal cancer cells. For example, **knockdown or overexpression** of RAP1GAP2 in OC cell lines (with appropriate controls) can test its effect on cell invasion, migration, and colony formation. Assessing downstream signaling changes – such as Rap1-GTP levels, ERK/MAPK activation, and secretion of MMPs – would directly probe the proposed mechanism. Confirmation of RAP1GAP2’s pro-invasive role in vitro (and ideally in vivo using xenograft or metastasis models) would solidify it as a bona fide molecular driver.
* **Investigation of Clinical Correlates:** Using tissue microarrays or larger patient cohorts, one could evaluate **RAP1GAP2 expression in patient tumor samples** and correlate it with clinical outcomes. Is RAP1GAP2 higher in tumors that went on to metastasize or in advanced stages? Does its expression correlate with HPV status or other clinicopathologic features? Answers to these questions would clarify the clinical relevance of RAP1GAP2 as a potential prognostic indicator or therapeutic target.
* **Exploring Therapeutic Targeting:** If RAP1GAP2 is validated as promoting metastasis, it becomes interesting to ask how to counteract its effects. While targeting a GAP protein directly is challenging, we could exploit its pathway. For instance, **inhibiting downstream effectors** (such as using MAPK pathway inhibitors, or blocking secreted factors like MMPs) in RAP1GAP2-high models might reduce invasion, suggesting a strategy for therapy. Additionally, disrupting the RAP1GAP2–Slp1 interaction or Rab27-mediated secretion is another angle that future pharmacological research could explore [12].
* **Expanding the Deep Profiling Framework:** Methodologically, our deep neural profiling framework can be extended to other cancers or integrated with multi-omics. Future work might incorporate **protein expression or phosphoproteomic data** to see if latent drivers at the mRNA level (like RAP1GAP2) correspond to changes in protein activity. Moreover, applying our pipeline to related cancers (e.g., other head and neck cancer subsites or HPV-positive vs HPV-negative tumors separately) could uncover whether the same drivers emerge or new context-specific ones appear. This could help generalize the concept of latent drivers across oncologic settings.
* **Elucidating Network Interactions:** We identified a panel of genes including RAP1GAP2, PDK3, FABP4, and others. Future studies could map the **interaction network** of these latent drivers. Do they converge on common pathways or interact indirectly? For example, hypoxia (via HIF-1) might induce both PDK3 and possibly RAP1GAP2, linking metabolism and invasion. Systems biology approaches and gene set perturbation experiments can be employed to understand how these drivers collectively shape the tumor phenotype. Such insight may reveal combinatorial intervention points (for instance, targeting metabolic and signaling pathways together).

By pursuing these directions, we can transition our findings from computational prediction to tangible biological and clinical advances. In doing so, we aim to fully leverage the “deep profiling” strategy to impact understanding and management of oropharyngeal carcinoma.

**Chapter Five**

# **Conclusion**

In summary, this study demonstrates that integrating high-dimensional transcriptomic data with deep learning can unearth **previously hidden drivers of cancer**. We discovered that RAP1GAP2 – a gene not evident through standard differential expression – plays a potential latent role in orchestrating tumor invasion in oropharyngeal carcinoma. This insight bridges a gap between data-driven modeling and biological mechanism, suggesting that Rap1 signaling dysregulation (via RAP1GAP2) coupled with metabolic and secretory reprogramming drives aggressiveness in these tumors. Our discussion highlights how this finding correlates with known cancer pathways while offering a novel perspective and testable model for metastasis. Although further validation is required, recognizing RAP1GAP2 as a molecular driver provides a new **direction for research and therapy**, moving us a step closer to improved prognostic tools and targeted interventions for head and neck cancers [5, 8]. Ultimately, the approach and outcomes of this study reinforce the value of looking beyond conventional analyses to capture the complex genetic underpinnings of cancer behavior, paving the way for more comprehensive and **innovative strategies** in cancer genomics research.

**Chapter Six**

# **References**

**Chapter Seven**

# **Appendices**