**Project Report on**

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

A logo with text and symbols

Description automatically generated

**In partial fulfilment of the requirements of the BMB433 for the degree of Bachelor of Science in Biochemistry and Molecular Biology of the Shahjalal University of Science and Technology, Sylhet**

**Submitted By**

**Registration No: 2019433077**

**BSc Session: 2019-2020**

**Department of Biochemistry and Molecular Biology**

**School of Life Science**

**Shahjalal University of Science and Technology**

**Sylhet-3114, Bangladesh**

**Date of Submission: July 07, 2025**

**Acknowledgments**

Foremost, I want to express my gratitude to Almighty God 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 project 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**

July 07, 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 **“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 |
| 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 |
| FDR | False Discovery Rate |
| MLP | Multi-Layer Perceptron |
| ERK | Extracellular Signal-Regulated Kinase |
| MAPK | Mitogen-Activated Protein Kinase |
| MMP | Matrix Metalloproteinase |
| IG | Integrated Gradients |
| LFC | Log Fold Change |
| NES | Normalized Enrichment Score |

# **Abstract**

**Background:** Conventional differential gene expression (DGE) analysis inadequately captures the complex molecular changes that drive the progression of oropharyngeal carcinoma (OC). Probabilistic Embedding Models (PEMs) offer a deep learning approach to uncover hidden patterns in high-dimensional transcriptomic data, potentially

**Methods:** Gene expression datasets were combined, from multiple databases and trained a PEM to compress the data into a small, hidden space. Integrated Gradients was utilized, an automated attribution technique, to determine the contribution of each gene to each latent node (biological representation). Genes that consistently had high attribution scores across all latent dimensions were chosen as potential regulators (driver genes). Pathway enrichment analysis and classification analyses unveiled the biological significance of these genes.

**Results:** The PEM learned latent features that are biologically important, and Integrated Gradients showed a group of genes that have a big impact on these features. RAP1GAP2 was consistently one of the top contributors across all 50 latent variables, which is noteworthy. RAP1GAP2 had the highest latent-space importance and strong discriminative power for telling OC apart, with a performance of 0.769. This occurred despite the lack of substantial differential expression in tumors relative to normal samples. Biological interpretation suggests that RAP1GAP2, a protein that activates Rap1 GTPase, may help tumors invade by turning off Rap1 and changing MAPK signaling and Golgi-mediated secretion.

**Conclusion:** Our deep learning framework found RAP1GAP2 to be a hidden driver in oropharyngeal carcinoma. This demonstrates how PEMs and Integrated Gradients may discover molecular regulators overlooked by alternative approaches. This method delivers novel dimensions about the biology of OC tumors that could benefit future research and therapeutic approaches.

**Keywords:** Oropharyngeal carcinoma; transcriptomics; deep learning; latent features; RAP1GAP2; Rap1 signaling; MAPK pathway; Golgi secretion

**Table of Contents**

Contents Page No.

[Abstract IV](#_Toc202764840)

[Keywords: IV](#_Toc202764841)

[List of Tables VII](#_Toc202764842)

[List of Figures VII](#_Toc202764843)

[Introduction](#_Toc202764844)

[1.1 Significance of Oropharyngeal Carcinoma 1](#_Toc202764845)

[1.2 Complexity of Cancer Biology and Analytical Gaps 2](#_Toc202764846)

[1.3 Deep Learning for Latent Feature Discovery 2](#_Toc202764847)

[1.4 Gene Attribution with Integrated Gradients 4](#_Toc202764848)

[1.5 Revealing Hidden Driver: The Case of *RAP1GAP2* 5](#_Toc202764849)

[1.6 Hypothesis 5](#_Toc202764850)

[1.7 Significance of the Study 6](#_Toc202764851)

[1.8 Aims and Objective 6](#_Toc202764852)

Material and Methods

[2.1 Workflow of the Study 7](#_Toc202764854)

[2.2 Datasets Retrieval 7](#_Toc202764855)

[2.3 Data Integration, Batch Effect Removal and Preprocessing 9](#_Toc202764856)

[2.4 Training Deep Neural Network Models 10](#_Toc202764857)

[2.4.1 Datasets Merging and Standardization 10](#_Toc202764858)

[2.4.2 Traditional Deep Learning Model 10](#_Toc202764859)

[2.4.3 Additional Sample Distribution 10](#_Toc202764860)

[2.5 Neural Network Design and Hyperparameter Optimization 11](#_Toc202764861)

[2.5.1 Train Model with Adam Optimizer 11](#_Toc202764862)

[2.5.2 Cross validate and Extract Best Latent Dimension 11](#_Toc202764863)

[2.6 Learning Robust Latent Representations 12](#_Toc202764864)

[2.7 Gene Attribution and Pathway Analysis 12](#_Toc202764865)

[2.7.1 Sensitivity-Based Scoring (SBS) for Gene-to-Latent Attribution 12](#_Toc202764866)

[2.7.2 Pathway Enrichment Analysis of Latent Variable-Associated Genes 12](#_Toc202764867)

[2.7.3 Gene Set Enrichment Analysis (GSEA) 13](#_Toc202764868)

[2.8 Supervised Deep Learning Model Training 13](#_Toc202764869)

[2.8.1 Gene Selection and Data Collection 13](#_Toc202764870)

[2.8.2 Normalization and Batch Correction 13](#_Toc202764871)

[2.8.3 Model Development and Training 13](#_Toc202764872)

[2.8.4 Evaluation and Visualization 14](#_Toc202764873)

[2.9 Differential Gene Expression analysis 14](#_Toc202764874)

[Result](#_Toc202764875)

[3.1 Data Preprocessing and Quality Assessment 15](#_Toc202764876)

[3.2 Latent Space Extraction Using Deep Neural Network 16](#_Toc202764877)

[3.3 Latent Variables Capture Distinct Gene Programs and Biological Pathways 17](#_Toc202764878)

[3.4 Functional Characterization of Latent Variables via GSEA 19](#_Toc202764879)

[3.5 Deep Learning-Based Classification of Candidate Driver Genes in Oropharyngeal Carcinoma 20](#_Toc202764881)

[3.6 RAP1GAP2 Emerges as the Most Predictive Gene in Single-Feature Classification Models 22](#_Toc202764882)

[3.7 RAP1GAP2 Emerges as a Key Latent Driver Despite Non-Significance in Differential Expression Analysis 25](#_Toc202764883)

[Discussion](#_Toc202764884)

[4.1 Discussion 27](#_Toc202764885)

[4.2 Limitations of the Study 30](#_Toc202764886)

[4.3 Future Directions 31](#_Toc202764887)

[5. Conclusion 32](#_Toc202764888)

[References 34](#_Toc202764889)

[Appendices 38](#_Toc202764890)

# 

# **List of Tables**

|  |  |  |
| --- | --- | --- |
| **SL. NO.** | **Table Captions** | **Page No.** |
| **Table 2.1** | Expression Profiling Datasets for OC | 8 |
| **Table 3.1** | Performance metrics for single-gene classification models | 23 |

# **List of Figures**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SL. NO.** | | **Figure Captions** | **Page No.** | |
| **Figure 1.1** | Anatomical regions of the head and neck involved in cancer. | | | 1 |
| **Figure 1.2** | Basic architecture of a deep neural network. | | | 3 |
| **Figure 2.1** | Overview of the study pipeline | | | 7 |
| **Figure 3.1** | Preprocessing and PCA of gene expression data | | | 15 |
| **Figure 3.2** | Model Performance and Gene Attribution | | | 16 |
| **Figure 3.3** | Interpretation of latent variables via gene attribution and enrichment | | | 18 |
| **Figure 3.4** | Pathway enrichment heatmap using GSEA | | | 20 |
| **Figure 3.5** | Identification of key driver genes and classification performance | | | 21 |
| **Figure 3.6** | RAP1GAP2 as a predictive gene for OC classification | | | 24 |
| **Figure 3.7** | RAP1GAP2 as latent driver despite non-significance in DGE | | | 25 |
| **Figure 4.1** | Mechanistic role of RAP1GAP2 in OC invasion and metastasis | | | 29 |

**Chapter One**

# **Introduction**

## **1.1 Overview of Oropharyngeal Carcinoma**

One type of head and neck cancer that has significant clinical significance is oropharyngeal carcinoma (OC). Human papillomavirus (HPV) infection has contributed to the increase in its incidence in recent decades, making HPV-positive oropharyngeal squamous cell carcinoma one of the cancers that is growing the fastest in many high-income nations (Lechner et al. 2022). An anatomical illustration of the oropharynx and its neighboring regions is shown in **Figure 1.1** to highlight the tumor's location and clinical context.

A diagram of the human body

AI-generated content may be incorrect.

**Figure 1.1 Anatomical regions of the head and neck involved in cancer.** Figure created using Adobe Illustrator (v27.8.1).

Because of its subtle early symptoms, OC frequently manifests at advanced stages, leading to substantial morbidity and mortality. Therefore, a deeper comprehension of the molecular foundations of OC is urgently needed to facilitate earlier detection, better patient stratification, and more successful precision therapies (Sabbatini and Manganaro 2023). Results for advanced OC are still uncertain despite advancements in systemic treatments, radiation therapy, and surgery. Gaining a better understanding of the transcriptome landscape of the tumor may help identify new molecular drivers that could enhance patient care.

## **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 (Rampášek et al. 2019). 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 (Rampášek et al. 2019).

Such underappreciated genes or gene interactions may be crucial for the development of cancer, making this gap problematic. Analytical techniques that can capture the nonlinear dependencies in gene expression data and go beyond linear assumptions are required. One potential remedy is explainable algorithms (machine learning), which can reveal multivariate gene patterns that would otherwise go unnoticed by applying interpretability techniques to complex models (Abbas and El-Manzalawy 2020; Way et al. 2020). In conclusion, techniques that can model and explain the complex, nonlinear relationships that define cancer biology are necessary to overcome the shortcomings of DGE.

## **1.3 Deep Learning for Latent Feature Discovery**

We use deep learning—more especially, unsupervised deep neural networks—to learn biologically significant latent variables from transcriptomic data in order to overcome these difficulties. A class of deep generative models that are ideal for this task are Probabilistic Embedding Models (PEMs). A PEM preserves as much information as possible while compressing high-dimensional gene expression profiles into a lower-dimensional latent space. Complex gene expression patterns can be reduced by this method to a collection of latent features that capture patient variability and underlying biological signals. Figure 1.2 illustrates the basic architecture of a deep neural network, where an encoder maps gene expression into latent representations for downstream interpretation. High-dimensional gene expression data are processed through multiple layers of an encoder network to generate low-dimensional latent features. These latent variables represent condensed biological signals and are suitable for interpretation, classification, or further modeling.

A diagram of a deep learning neural network

AI-generated content may be incorrect.

**Figure 1.2 Basic architecture of a deep neural network.** Figure created using Adobe Illustrator (v27.8.1).

Largescale gene expression datasets have seen the successful application of PEMs and related autoencoder techniques, which have shown promise in modeling non-linear gene interactions and enhancing outcome predictions (Sundararajan et al. 2017). To illustrate the ability of deep learning to capture subtle transcriptomic effects of treatment, Rampášek et al. demonstrated that a PEM-based model ("Dr.PEM") could learn latent representations of cancer cell line expression data that improve drug response prediction (Zhang et al. 2006). Similar to this, Way et al. used PEM to compress pan cancer gene-expression data and discovered that different biological signals (like pathway activities and mutational status) emerged when the latent dimensionality was varied. This suggests that deep compression can learn complementary aspects of tumor biology that are not possible with a single linear compression or DGE analysis (Way et al. 2020). 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 PEM might yield abstract features that correlate with disease, but without interpretation we cannot translate those features into testable biological insights.

## **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 (Janizek et al. 2023). 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 (Janizek et al. 2023). Introduced by Sundararajan et al. in 2017, this method has become a popular tool for explaining deep learning predictions in various domains (Janizek et al. 2023). In our study, we harness integrated gradients to attribute genes to latent variables learned by the PEM 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, earlier studies have shown how useful it is to combine feature attribution and deep generative models in genomics. For instance, Dincer et al. identified the top contributing genes for each latent dimension by applying integrated gradients to the latent features of a PEM trained on cancer gene expression data (Janizek et al., 2023). Researchers can anchor abstract features in concrete biology by using this post hoc interpretation of latent space. For example, based on the genes with the highest attributions, a latent dimension may end up representing a pathway or cell cycle signature. Building on these concepts, we derive gene-level importance scores for the learned latent factors by combining our PEM with integrated gradients. By doing this, we can identify the genes that are most important for differentiating oropharyngeal tumors from controls (or other tumor subtypes) and that drive the variations recorded in the latent space. In addition to maintaining interpretability, this combination of unsupervised, deep learning and explainability techniques enables us to find biologically significant patterns that would be missed by linear analysis alone.

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

By using this deep learning framework on OC transcriptomic data, new understandings of the molecular causes of the disease are revealed. Integrated gradients identify the genes that define the latent variables that the variational autoencoder extracts and that summarize gene expression patterns across tumors. Our analysis reveals that RAP1GAP2 is a crucial latent driver gene in oropharyngeal carcinoma, which is intriguing. With a high attribution score, RAP1GAP2 stands out in our model as one of the main contributors to a latent feature that is very predictive of the presence of OC. This finding is noteworthy because, according to standard differential expression analysis, RAP1GAP2 was not identified as significant; that is, its average expression levels between tumor and normal do not differ sufficiently to meet standard statistical thresholds. RAP1GAP2 would have been completely overlooked by traditional DGE, but our deep learning method revealed it to be a significant participant with a nonlinear contribution to the tumor transcriptome. The impact of RAP1GAP2 only becomes apparent when taking into account intricate interactions recorded in the latent space, demonstrating how deep learning can uncover "hidden" drivers that elude linear analysis.

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 (Zhang et al. 2006). 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**

We hypothesize that deep neural networks, particularly Probabilistic Embedding Model (PEM) models, can learn latent representations of transcriptomic data that capture complex, nonlinear biological signals associated with oropharyngeal carcinoma. These latent features are expected to reveal molecular regulators that conventional differential gene expression (DGE) analyses may overlook due to their reliance on linear assumptions. By integrating unsupervised deep learning with interpretability techniques such as integrated gradients, we anticipate uncovering key gene-level contributors—such as RAP1GAP2—that drive tumor invasion and progression despite showing no significant differential expression. This approach offers a novel avenue for identifying biologically relevant signals embedded in high-dimensional gene expression data.

## **1.7 Significance of the Study**

Comprehending the molecular pathways underlying oropharyngeal cancer (OC) is a significant challenge, especially due to the constraints of conventional gene expression analysis techniques that frequently depend on linear assumptions. This paper presents a deep learning system that may reveal nonlinear and concealed transcriptome signals, providing an innovative method for identifying genetic drivers of ovarian cancer. Utilizing variational autoencoders and integrated gradients, we discovered RAP1GAP2 as an unknown factor in tumor invasion and development, despite its absence of differential expression according to traditional statistical standards. This underscores the capability of modern computational modeling to not only augment but also exceed conventional analytical methods. The results of this study provide novel avenues for biological research and therapeutic development in ovarian cancer and create a framework for the application of interpretable deep learning to other intricate diseases.

## **1.8 Aims and Objective**

This study aims to uncover hidden transcriptomic patterns and identify novel gene-level drivers of oropharyngeal carcinoma (OC) by applying deep neural network-based methods—specifically variational autoencoders and integrated gradients—that go beyond the limitations of traditional differential expression analysis.

Objectives of the study are,

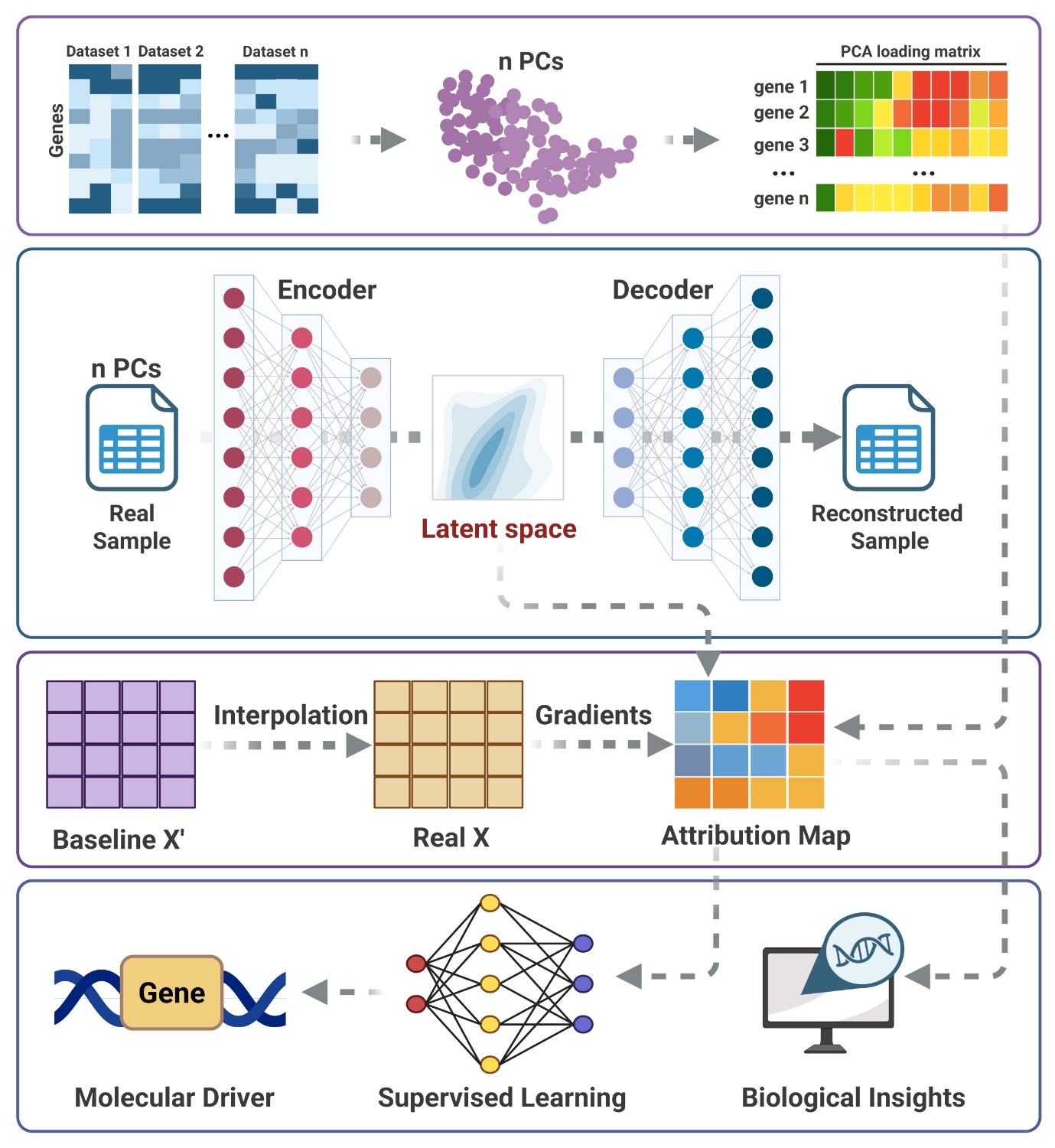
* To apply deep learning (PEM) for compressing gene expression into latent features.
* To detect complex, nonlinear gene patterns missed by standard tools.
* To interpret latent features using Integrated Gradients for gene attribution.
* To combine unsupervised modeling with supervised classification.
* To identify novel molecular drivers involved in OC progression.
* To compare the performance of this method with traditional differential gene expression approaches.

**Chapter Two**

# **Material and Methods**

## **2.1 Workflow of the Study**

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

**Figure 2.1: Workflow of the study pipeline.** *PCA-transformed multi-dataset gene expression is encoded via PEM 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. Figure generated using Adobe Illustrator (v27.8.1).*

## **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 ArrayExpress database (<https://www.ebi.ac.uk/arrayexpress/>), a public repository of functional genomics data maintained by the European Bioinformatics Institute (EBI) for Oral Cancer types. To extract both the gene expression data and the associated phenotype metadata, we utilized the Orange.bioinformatics.arrayexpress module v2.6.25, which provides the only available Python-based programmatic access to ArrayExpress. All information about the datasets including sample size mentioned in **Table 2.1.**

**Table 2.1 Expression Profiling Datasets for OC**

|  |  |  |
| --- | --- | --- |
| GEO\_Accession | Samples | Study\_Type |
| E-GEOD-37991 | 80 (40 tumor + 40 normal) | Expression profiling by array |
| E-GEOD-23558 | 31 (27 tumor + 4 normal) | Expression profiling by array |
| E-GEOD-25099 | 79 (57 tumor + 22 normal) | Expression profiling by array |
| E-GEOD-10121 | 41 (35 tumor + 6 normal) | Expression profiling by array |
| E-GEOD-31853 | 11 (8 tumor cell lines + 3 normal) | Expression profiling by array |
| E-GEOD-131182 | 12 (6 paired tumor + normal) | Expression profiling by RNA‑seq |
| E-GEOD-145272 | 10 (5 metastatic + 5 non-metastatic) | Expression profiling by RNA‑seq |
| E-GEOD-217142 | 6 (primary + recurrent tumors) | Expression profiling by RNA‑seq |
| E-GEOD-85195 | 49 (34 OSCC + 15 OPL) | Expression profiling by array |
| E-GEOD-168227 | 6 paired tumor-normal samples | Expression profiling by array |
| E-GEOD-84805 | 6 paired tumor-normal samples | Expression profiling by array |
| E-MTAB-30784 | 229 total (167 tumor + others) | Expression profiling by array |
| E-MTAB-2280 | 32 (27 non-metastatic + 5 metastatic) | Expression profiling by array |
| E-MTAB-3524 | 20 (16 tumor + 4 normal) | Expression profiling by array |
| E-MTAB-6791 | 154 (119 tumor + 35 controls) | Expression profiling by array |
| E-MTAB-41442 | 55 (45 tumor + 10 normal) | Expression profiling by array |
| E-MTAB-37371 | 100 (50 tumor + 50 normal) | Expression profiling by array |
| E-MTAB-23030 | 30 metastatic tongue OSCC | Expression profiling by array |
| E-MTAB-29000 | 50 (40 tumor + 10 normal) | Expression profiling by array |

Extracted results according to the supplied ArrayExpress 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 (McKinney 2011) was incorporated. Data imputation was conducted by missForest v0.9 (Stekhoven and Bühlmann 2012) 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 (Behdenna et al. 2023) 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 (Kassambara and Mundt 2020) for extraction and display of PCA results, and dplyr v1.1.4 (Hadley Wickham et al. 2020) 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 (Bro and Smilde 2014). 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 (Tian et al. 2017) between the input and its reconstruction:

**… ... … (i)**

Here in (i) equation, 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:

**… ... … (ii)**

### **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 :

**… ... … (iii)**

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

**… ... … (iv)**

The first term ensures accurate reconstruction, while the KL divergence regularizes the latent space by encouraging it to resemble a standard (Pan et al. 2020). 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 (Wang et al. 2022), 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 (Elgeldawi et al. 2021). 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 (Chollet 2015) and TensorFlow v1.12.0 (Filus and Domańska 2023).

## **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 (Sinaga and Yang 2020). 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 (Ri and Kim 2020).

## **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 (Peterson et al. 2020) 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 (Carbon et al. 2017; Jassal et al. 2020; Kanehisa et al. 2023). 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 (Ferreira and Zwinderman 2006). 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 (Balagopalan et al. 2009). 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 (Waskom 2021)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 Deep 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 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 (Wolf et al. 2018) 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 (Wazery et al. 2023). 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 (Tolstikhin et al. 2021). All models were trained using the Adam optimizer (Wang et al. 2022) (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 Linux-based computing environment.

## **2.9 Differential Gene Expression analysis**

Expression data were analyzed *using* DESeq2 v1.40.2 (Love et al. 2014). 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 (Blighe et al. 2021).

**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 A close-up of a graph

AI-generated content may be incorrect.

**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. Figure generated using Python (v3.12).*

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.

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

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)**.

A comparison of a graph

AI-generated content may be incorrect.

**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 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. Figure generated using Python (v3.12).*

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 the most relevant latent spaces, understand the core biology of OC from the complex environment of the data.

**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), HPCAL1 (0.00349), PICALM (0.00342), COMMD8 (0.0033), 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 PEM 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.

A diagram of a dna sequence

AI-generated content may be incorrect.

**Figure 3.3: Interpretation of PEM 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 padj). Figure generated using Python (v3.12).*

**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. A complete table of enriched pathways, including adjusted p-values, enrichment scores, and associated gene sets for all 50 latent, is provided in **Appendix II**.

## **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 III.**

## **A diagram of a genetic modification AI-generated content may be incorrect.**

**Figure 3.4:** **Pathway enrichment heatmap of PEM latent variables using GSEA.**  
*Heatmap shows 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 (NES < 0). Figure generated using Python (v3.12)*

## **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)** and boxplots in **Appendix IV**. Several genes A diagram of a graph

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 PEM 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. Figure generated using Python (v3.12)*

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.

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 PEM-derived genes, particularly RAP1GAP2, PDK3, and FABP4, may serve as effective driver markers or classifiers for oropharyngeal carcinoma in high-throughput transcriptomic data **(Table 3.1), (Appendix V)**.

## **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 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.

**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 graph and a chart of a number of numbers

AI-generated content may be incorrect.

**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 AUPRC of 0.769, indicating strong predictive capacity from this gene alone. Figure generated using Python (v3.12)*

## **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.

A diagram of a sample

AI-generated content may be incorrect.

**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. Figure generated using R (v4.3.2) with RStudio (v2023.09.1+494).*

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.

However, differential gene expression (DGE) analysis failed to identify RAP1GAP2 as significant in LFC values. As shown in **Figure 3.7C**, RAP1GAP2 resides within the "not significant" region of the volcano plot, indicating that it was not differentially expressed based on standard thresholds (log2 fold change and FDR-adjusted p-value).

**Chapter Four**

# **Discussion**

## **4.1 Discussion**

This study employed a deep learning framework to reveal novel molecular patterns and latent drivers overlooked by traditional methods, utilizing transcriptome data from oropharyngeal cancer (OC). We successfully reduced the data to 50 low-dimensional, biologically interpretable latent variables by training a variational autoencoder on high-dimensional gene expression matrices. By maintaining essential variation among samples, these latent traits enabled downstream modeling to reveal functional insights. A 50-dimensional embedding yielded the optimal balance between biological richness and training error when evaluating the reconstruction quality of the PEM across various latent dimensionalities **(Figure 3.1)**. The model's capacity to delineate the underlying illness structure was emphasized by the UMAP display of the acquired embeddings, which distinctly segregated the OC subgroups **(Figure 3.2)**.

Integrated Gradients were employed to quantify the contribution of each gene to each latent dimension, thereby enhancing the understanding of the biological relevance of these representations. The analysis revealed the presence of high-attribution gene sets that were not restricted to individual dimensions but were also enriched for key biological pathways, as identified through Gene Ontology and KEGG annotations **(Figure 3.4)**. Several latent variables were associated with biological processes such as cell adhesion, immune signaling, and extracellular matrix remodeling—mechanisms commonly implicated in tumor progression. In many cases, high-contribution genes appeared recurrently across multiple latent dimensions, indicating that shared biological programs may be embedded within distinct transcriptomic patterns. These results confirmed that the latent space captured by the model reflects physiologically meaningful signals and provided justification for further examination of genes contributing across dimensions.

This study aimed to investigate the molecular intricacies of oropharyngeal cancer (OC) via a deep learning analytical framework that transcends the limitations of conventional differential gene expression techniques. Utilizing a probabilistic embedding model (PEM) grounded on a neural network framework, and subsequently applying gene attribution through integrated gradients, we identified 50 latent dimensions that encapsulate compressed, physiologically significant transcriptome patterns. The latent dimensions were enriched for specific gene programs and biological pathways **(Figure 3.3, 3.4)**, uncovering concealed aspects of OC biology not addressed by conventional linear methods. RAP1GAP2 appeared as a notably consistent and discriminative component among the genes contributing to these latent traits (Figure 3.6). Despite its robust latent-space attribution and efficacy as a single-gene classifier (AUPRC ≈ 0.77), RAP1GAP2 was not deemed significant in LFC in the differential expression study (Figure 3.7C). The disparity between statistical insignificance and biological significance underscores the fundamental value of our approach—deep generative models can reveal non-linear molecular determinants that traditional methods may overlook.

Our results align with and contribute to the existing knowledge in the subject. Researchers have long recognized that the Ras-related GTPase Rap1 and its regulators influence the adhesion and motility of cancer cells (Zhang et al. 2017). Active Rap1 signaling has been demonstrated to enhance the invasiveness of head and neck malignancies by inducing the production of β-catenin and MMP7 (Zhang et al. 2017). Conversely, the established Rap1 inactivator Rap1GAP (a paralog of RAP1GAP2) is recognized for its ability to inhibit Rap1–ERK signaling and tumor proliferation (Zhang et al. 2006). Our identification of RAP1GAP2 enhances this paradigm while introducing a novel element. RAP1GAP2 functions as a pro-invasion factor, whereas Rap1GAP broadly inhibits HNSCC growth (Zhang et al. 2006). Upon examining the entirety of the situation, this seeming contradiction becomes comprehensible: Rap1 regulators frequently exert disparate effects on various cell types (Zhang et al. 2017). Research indicates that Rap1GAP often inhibits invasion in various malignancies; but, in certain instances, elevated levels of Rap1GAP may enhance cellular invasiveness (Zhang et al. 2017). Our findings indicate that oropharyngeal carcinoma exemplifies a scenario in which RAP1GAP2, functioning in a specific cellular region, promotes cancer proliferation. This constitutes a novel discovery, as RAP1GAP2 has not been previously examined in oropharyngeal cancer; it was essentially a "obscured" driver identified by our latent-space profiling.

We identified additional latent drivers, including PDK3 and FABP4, that corroborate the biological validity of our methodology. PDK3 (pyruvate dehydrogenase kinase 3) is a recognized mediator of the Warburg effect and is increased in hypoxic malignancies, resulting in metabolic reprogramming and aggressive behavior (Lu et al. 2011). FABP4 (fatty acid–binding protein 4) facilitates tumor metastasis and treatment resistance by accelerating lipid transport and signaling in cancer cells (Sun and Zhao 2022). Our model appears to have encapsulated significant characteristics of cancer, such as metabolic plasticity and microenvironmental adaptation, alongside the Rap1 signaling axis. The presence of PDK3 and FABP4 among our principal latent genes demonstrates this. The alignment of our data-driven discoveries with established cancer pathways corroborates the outcomes of our study. We have identified a novel driver (RAP1GAP2) and an accompanying array of genes implicated in oropharyngeal cancer invasion and demonstrated that deep neural profiling can uncover biologically significant targets overlooked by conventional techniques.

A diagram of a cell division

AI-generated content may be incorrect.

**Figure 4.1: Schematic model illustrates the proposed role of RAP1GAP2 in promoting invasion and metastasis in OC.** *Figure generated using Adobe Illustrator (v27.8.1).*

RAP1GAP2 is a GTPase-activating protein (GAP) for Rap1 (Johansen et al. 2023). It changes active GTP-bound Rap1 into an inactive GDP-bound state, which changes how cells stick together and send signals. Active Rap1 stabilizes integrins and E-cadherins, which helps cells stick together and keeps epithelial cells looking like epithelial cells (Price et al. 2004). RAP1GAP2 stops Rap1 from working, which breaks up these stable interactions and makes cells lose their ability to stick together. This is necessary for tumor cells to start moving and invading.

RAP1GAP2 inactivates Rap1, which not only stops adhesion but also stops Rap1 from stopping Ras–MAPK/ERK signaling. This makes the ERK pathway more active (Zhang et al. 2017). ERK signaling helps cells grow, move, and turn on invasive genes, such as matrix metalloproteinases (MMPs). This makes tumors even more aggressive (Mitra et al. 2008).

RAP1GAP2 also affects how tumors invade by changing how vesicles move around. It works with the synaptotagmin-like protein 1 (Slp1) and Rab27 complex to control secretory vesicles that come from the Golgi apparatus (Neumüller et al. 2009; Li et al. 2018). This interaction leads to the release of enzymes that break down the matrix, like MMP-2 and MMP-9, into the extracellular space. This makes it easier for tissues to break down and makes them more invasive (Mitra et al. 2008; Beroun et al. 2019).

So, RAP1GAP2 controls a coordinated, multi-dimensional invasion strategy: it weakens cellular adhesion, turns on pro-invasive ERK/MAPK signaling, and boosts Golgi’s ability to secrete proteases (Guo et al. 2020). This integrated mechanism shows how RAP1GAP2 can help metastasis even though it acts as a Rap1 inhibitor. Future experiments can test whether changing the expression of RAP1GAP2 affects the strength of cell adhesion, the levels of ERK activation, and the release of invasive factors. This would confirm its many roles in the progression of oropharyngeal carcinoma. Notably, this latent driver effect of RAP1GAP2 is captured by our model despite its lack of prominence in linear analysis, indicating that its contribution, while subtle at the expression level, is indeed biologically significant. Overall, the identification of RAP1GAP2 through latent-space analysis—supported by attribution, classifier performance, and mechanistic plausibility—highlights both the biological relevance of this gene and the power of our approach to reveal novel drivers in oropharyngeal carcinoma.

## **4.2 Limitations of the Study**

Based on integrative analyses of transcriptomic data, our study identifies RAP1GAP2 as a promising computationally predicted driver gene in oropharyngeal carcinoma (OC). To preserve a fair interpretation, a few restrictions must be noted.

First off, we didn't carry out functional tests to confirm RAP1GAP2's involvement in cellular functions like invasion and metastasis. Therefore, our results are still correlative, and there is no proof that RAP1GAP2 causes tumor behavior. Second, even with batch effect correction and gene harmonization, heterogeneity is introduced because we used retrospective integration of several public datasets from various platforms and clinical subgroups. Variations in tumor subsite, treatment history, and HPV status could affect the latent features that are extracted. Third, some candidate genes (such as RAP1GAP2) showed only slight expression changes and might contribute to false positives because our machine learning pipeline gave predictive power precedence over statistical significance. Although this risk was reduced by cross-validation, biological significance still needs to be ascertained through experimentation. Furthermore, we were unable to assess the prognostic significance of the identified drivers due to the restricted availability of comprehensive clinical endpoints, such as survival and metastasis data. Lastly, we only looked at the mRNA level, leaving out other regulatory mechanisms that could have a significant impact on RAP1GAP2's function, like mutations, epigenetic changes, and post-translational events. All of these drawbacks highlight the necessity of additional research that includes multi-omic integration and experimental validation in order to completely clarify the biological and therapeutic significance of our findings.

## **4.3 Future Directions**

Our results provide several avenues for additional research to confirm and broaden the biological significance of RAP1GAP2 in oropharyngeal carcinoma (OC). First and foremost, functional validation is essential. RAP1GAP2's function would be directly tested by knocking down or overexpressing it in OC cell lines and evaluating cell invasion, Rap1-GTP activity, and downstream signaling (such as ERK/MAPK and MMP secretion). Its pro-metastatic role may be further supported by in vivo models. RAP1GAP2's value as a biomarker may be defined clinically by assessing its expression in larger patient cohorts or tissue arrays, which may show associations with tumor stage, metastasis, HPV status, or prognosis.

From a therapeutic standpoint, RAP1GAP2's downstream pathways, like MAPK signaling or Rab27-mediated secretion, provide actionable targets, even though directly targeting it may be challenging. Inhibitors of these effectors in RAP1GAP2-high models may be investigated in future research. To find cross-layer or context-specific drivers, our deep profiling framework can be methodologically extended to other cancers or combined with proteomic and epigenetic data. New patterns may be found by applying the pipeline to datasets related to head and neck cancer that are HPV-stratified.

Lastly, more research is necessary to fully understand the network of interactions between latent drivers such as RAP1GAP2, PDK3, and FABP4. Studies on gene perturbation and systems biology may shed light on whether these genes are linked by common regulators (like hypoxia) and provide combinatorial intervention points. Collectively, these avenues will enhance our comprehension of the function of RAP1GAP2 and facilitate the realization of our computational approach's translational potential.

**Chapter Five**

**Conclusion**

# **5. Conclusion**

In summary, this study demonstrates that new cancer-causing factors can be identified by combining deep learning with high-dimensional transcriptome data. We discovered that RAP1GAP2, a gene that is rarely observed to exhibit differential expression, might play a secret role in regulating the invasion of other tissues by oropharyngeal cancer. This new knowledge links biological mechanisms to data-driven modeling. It implies that these cancers become more aggressive due to dysregulation of Rap1 signaling (via RAP1GAP2), as well as modifications in metabolism and secretion. Our discussion demonstrates how this finding aligns with our current understanding of cancer pathways and provides a fresh perspective on and method for testing metastasis. We are one step closer to improved prognostic tools and customized treatments for oral cancers now that RAP1GAP2 is recognized as a molecular driver (Zhang et al. 2006). Ultimately, the study's methodology and findings highlight the significance of looking beyond conventional research to comprehend the intricate genetic elements influencing cancer behavior. This makes it possible to conduct cancer genomics research using more comprehensive and innovative methods.

# **References**

1. Abbas, M. and El-Manzalawy, Y. 2020. Machine learning based refined differential gene expression analysis of pediatric sepsis. BMC Medical Genomics 13(1). DOI: <https://doi.org/10.1186/s12920-020-00771-4>
2. Balagopalan, S., Ashok, S. and Mohandas, K.P. 2009. GSEA. Journal of Electrical Systems 5(4).
3. Behdenna, A., Colange, M., Haziza, J., Gema, A., Appé, G., Azencott, C.A. and Nordor, A. 2023. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. BMC Bioinformatics 24(1). DOI: <https://doi.org/10.1186/s12859-023-05578-5>
4. Beroun, A., Mitra, S., Michaluk, P., Pijet, B., Stefaniuk, M. and Kaczmarek, L. 2019. MMPs in learning and memory and neuropsychiatric disorders. Cellular and Molecular Life Sciences 76(16). DOI: <https://doi.org/10.1007/s00018-019-03180-8>
5. Blighe, K., Rana, S. and Lewis, M. 2021. EnhancedVolcano version 1.10.0: Publication-ready volcano plots with enhanced colouring and labeling. R-Package.
6. Bro, R. and Smilde, A.K. 2014. Principal component analysis. Analytical Methods 6(9). DOI: <https://doi.org/10.1039/c3ay41907j>
7. Carbon, S. et al. 2017. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. Nucleic Acids Research 45(D1). DOI: <https://doi.org/10.1093/nar/gkw1108>
8. Chollet, F. 2015. Keras: The Python Deep Learning library. Keras.Io.
9. Elgeldawi, E., Sayed, A., Galal, A.R. and Zaki, A.M. 2021. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. Informatics 8(4). DOI: <https://doi.org/10.3390/informatics8040079>
10. Ferreira, J.A. and Zwinderman, A.H. 2006. On the Benjamini-Hochberg method. Annals of Statistics 34(4). DOI: <https://doi.org/10.1214/009053606000000425>
11. Filus, K. and Domańska, J. 2023. Software vulnerabilities in TensorFlow-based deep learning applications. Computers and Security 124. DOI: <https://doi.org/10.1016/j.cose.2022.102948>
12. Guo, Y., Pan, W., Liu, S., Shen, Z., Xu, Y. and Hu, L. 2020. ERK/MAPK signalling pathway and tumorigenesis (Review). Experimental and Therapeutic Medicine. DOI: <https://doi.org/10.3892/etm.2020.8454>
13. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller. 2020. A grammar of data manipulation [R package dplyr version 1.0.0].
14. Janizek, J.D. et al. 2023. PAUSE: principled feature attribution for unsupervised gene expression analysis. Genome Biology 24(1). DOI: <https://doi.org/10.1186/s13059-023-02901-4>
15. Jassal, B. et al. 2020. The reactome pathway knowledgebase. Nucleic Acids Research 48(D1). DOI: <https://doi.org/10.1093/nar/gkz1031>
16. Johansen, K.H., Golec, D.P., Okkenhaug, K. and Schwartzberg, P.L. 2023. Mind the GAP: RASA2 and RASA3 GTPase-activating proteins as gatekeepers of T cell activation and adhesion. Trends in Immunology 44(11). DOI: <https://doi.org/10.1016/j.it.2023.09.002>
17. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. and Ishiguro-Watanabe, M. 2023. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research 51(D1). DOI: <https://doi.org/10.1093/nar/gkac963>
18. Kassambara, A. and Mundt, F. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Package Version 1.0.7. R package version.
19. Lechner, M., Liu, J., Masterson, L. and Fenton, T.R. 2022. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. Nature Reviews Clinical Oncology 19(5). DOI: <https://doi.org/10.1038/s41571-022-00603-7>
20. Li, Z., Fang, R., Fang, J., He, S. and Liu, T. 2018. Functional implications of Rab27 GTPases in Cancer. Cell Communication and Signaling 16(1). DOI: <https://doi.org/10.1186/s12964-018-0255-9>
21. Li, Z., Fang, R., Fang, J., He, S. and Liu, T. 2018. Functional implications of Rab27 GTPases in Cancer. Cell Communication and Signaling 16(1). DOI: <https://doi.org/10.1186/s12964-018-0255-9>
22. Love, M.I., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), p.550. <https://doi.org/10.1186/s13059-014-0550-8>
23. Lu, C.W. et al. 2011. Overexpression of pyruvate dehydrogenase kinase 3 increases drug resistance and early recurrence in colon cancer. American Journal of Pathology 179(3). DOI: <https://doi.org/10.1016/j.ajpath.2011.05.050>
24. McKinney, W. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing.
25. Mitra, R.S. et al. 2008. Rap1GAP promotes invasion via induction of matrix metalloproteinase 9 secretion, which is associated with poor survival in low N-stage squamous cell carcinoma. Cancer Research 68(10). DOI: <https://doi.org/10.1158/0008-5472.CAN-07-2755>
26. Neumüller, O., Hoffmeister, M., Babica, J., Prelle, C., Gegenbauer, K. and Smolenski, A.P. 2009. Synaptotagmin-like protein 1 interacts with the GTPase-activating protein Rap1GAP2 and regulates dense granule secretion in platelets. Blood 114(7). DOI: <https://doi.org/10.1182/blood-2008-05-155234>
27. Pan, T., Zhao, J., Wu, W. and Yang, J. 2020. Learning imbalanced datasets based on SMOTE and Gaussian distribution. Information Sciences 512. DOI: <https://doi.org/10.1016/j.ins.2019.10.048>
28. Peterson, H., Kolberg, L., Raudvere, U., Kuzmin, I. and Vilo, J. 2020. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. F1000Research 9. DOI: <https://doi.org/10.12688/f1000research.24956.2>
29. Price, L.S., Hajdo-Milasinovic, A., Zhao, J., Zwartkruis, F.J.T., Collard, J.G. and Bos, J.L. 2004. Rap1 regulates E-cadherin-mediated cell-cell adhesion. Journal of Biological Chemistry 279(34). DOI: <https://doi.org/10.1074/jbc.M404917200>
30. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. and Goldenberg, A. 2019. Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. Bioinformatics 35(19). DOI: <https://doi.org/10.1093/bioinformatics/btz158>
31. Ri, J.H. and Kim, H. 2020. G-mean based extreme learning machine for imbalance learning. Digital Signal Processing: A Review Journal 98. DOI: <https://doi.org/10.1016/j.dsp.2019.102637>
32. Sabbatini, G. and Manganaro, L. 2023. On potential limitations of differential expression analysis with non-linear machine learning models. EMBnet.journal 28. DOI: <https://doi.org/10.14806/ej.28.0.1035>
33. Sinaga, K.P. and Yang, M.S. 2020. Unsupervised K-means clustering algorithm. IEEE Access 8. DOI: <https://doi.org/10.1109/ACCESS.2020.2988796>
34. Stekhoven, D.J. and Bühlmann, P. 2012. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1). DOI: <https://doi.org/10.1093/bioinformatics/btr597>
35. Sun, N. and Zhao, X. 2022. Therapeutic Implications of FABP4 in Cancer: An Emerging Target to Tackle Cancer. Frontiers in Pharmacology 13. DOI: <https://doi.org/10.3389/fphar.2022.948610>
36. Sundararajan, M., Taly, A. and Yan, Q. 2017. Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning, ICML 2017.
37. Tian, Y., Fan, B. and Wu, F. 2017. L2-Net: Deep learning of discriminative patch descriptor in Euclidean space. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. DOI: <https://doi.org/10.1109/CVPR.2017.649>
38. Tolstikhin, I. et al. 2021. MLP-Mixer: An all-MLP Architecture for Vision. In: Advances in Neural Information Processing Systems.
39. Wang, Y., Xiao, Z. and Cao, G. 2022. A convolutional neural network method based on Adam optimizer with power-exponential learning rate for bearing fault diagnosis. Journal of Vibroengineering 24(4). DOI: <https://doi.org/10.21595/jve.2022.22271>
40. Waskom, M. 2021. seaborn: statistical data visualization. Journal of Open Source Software 6(60). DOI: <https://doi.org/10.21105/joss.03021>
41. Way, G.P., Zietz, M., Rubinetti, V., Himmelstein, D.S. and Greene, C.S. 2020. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biology 21(1). DOI: <https://doi.org/10.1186/s13059-020-02021-3>
42. Wazery, Y.M., Saleh, M.E. and Ali, A.A. 2023. An optimized hybrid deep learning model based on word embeddings and statistical features for extractive summarization. Journal of King Saud University - Computer and Information Sciences 35(7). DOI: <https://doi.org/10.1016/j.jksuci.2023.101614>
43. Wolf, F.A., Angerer, P. and Theis, F.J. 2018. SCANPY: Large-scale single-cell gene expression data analysis. Genome Biology 19(1). DOI: <https://doi.org/10.1186/s13059-017-1382-0>
44. Zhang, Y.L., Wang, R.C., Cheng, K., Ring, B.Z. and Su, L. 2017. Roles of Rap1 signaling in tumor cell migration and invasion. Cancer Biology and Medicine 14(1). DOI: <https://doi.org/10.20892/j.issn.2095-3941.2016.0086>
45. Zhang, Z. et al. 2006. Rap1GAP inhibits tumor growth in oropharyngeal squamous cell carcinoma. American Journal of Pathology 168(2). DOI: <https://doi.org/10.2353/ajpath.2006.050132>

# **Appendices**

**Appendix I**

A chart of bar code

AI-generated content may be incorrect.

**Appendix II**

**Selected Latent Variable-Enriched Pathways**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Pathway Name | P-value | Term Size | Query Size | Intersection Size | Precision | Recall | Latent Node |
| regulation of nervous system development | 0.0236186136151001 | 457 | 16 | 5 | 0.3125 | 0.0109409190371991 | 7 |
| DNA integration | 0.0287922088818914 | 10 | 15 | 2 | 0.1333333333333333 | 0.2 | 9 |
| Glucagon signaling pathway | 0.0134641808035551 | 107 | 10 | 3 | 0.3 | 0.02803738317757 | 18 |
| Cushing syndrome | 0.0499569897009377 | 153 | 11 | 3 | 0.2727272727272727 | 0.0196078431372549 | 19 |
| Vibrio cholerae infection | 0.0246555455325596 | 50 | 6 | 2 | 0.3333333333333333 | 0.04 | 20 |
| Golgi lumen acidification | 0.0370534161963728 | 13 | 13 | 2 | 0.1538461538461538 | 0.1538461538461538 | 20 |
| Cushing syndrome | 0.0499569897009377 | 153 | 11 | 3 | 0.2727272727272727 | 0.0196078431372549 | 23 |
| tRNA metabolic process | 0.0414263692639973 | 210 | 18 | 4 | 0.2222222222222222 | 0.019047619047619 | 25 |
| ECM-receptor interaction | 0.010145054293642 | 89 | 11 | 3 | 0.2727272727272727 | 0.0337078651685393 | 31 |
| Glyoxylate and dicarboxylate metabolism | 0.012188180777088 | 30 | 7 | 2 | 0.2857142857142857 | 0.0666666666666666 | 34 |
| cellular response to 2,3,7,8-tetrachlorodibenzodioxine | 0.0038484693649248 | 4 | 15 | 2 | 0.1333333333333333 | 0.5 | 41 |
| response to 2,3,7,8-tetrachlorodibenzodioxine | 0.0179299300313183 | 8 | 15 | 2 | 0.1333333333333333 | 0.25 | 41 |
| Base excision repair | 0.0263601360727707 | 44 | 7 | 2 | 0.2857142857142857 | 0.0454545454545454 | 42 |
| RNA degradation | 0.0412133286559642 | 79 | 5 | 2 | 0.4 | 0.0253164556962025 | 43 |
| neuromuscular junction development | 0.0215300187657582 | 54 | 19 | 3 | 0.1578947368421052 | 0.0555555555555555 | 47 |
| regulation of response to external stimulus | 0.0390732084249658 | 1082 | 19 | 7 | 0.3684210526315789 | 0.0064695009242144 | 47 |

**Appendix III**

**GSEA Info for Top 50 Pathways**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Term | ES | NES | NOM p-val | FDR q-val | FWER p-val | Tag % | Gene % | latent |
| Ribosome | -0.2538932143878554 | -3.142324180841197 | 0.0 | 0.0 | 0.0 | 64/112 | 30.62% | 0 |
| Oxidative phosphorylation | -0.2407197097865659 | -2.611673921650873 | 0.0 | 0.0 | 0.0 | 70/86 | 56.60% | 0 |
| Thermogenesis | -0.158950904011502 | -2.405693731233716 | 0.0 | 0.0 | 0.0 | 87/159 | 37.50% | 0 |
| Taste transduction | -0.3048429308211678 | -2.350171914030135 | 0.0 | 0.0 | 0.0 | 26/32 | 50.10% | 0 |
| Epstein-Barr virus infection | 0.0556416038563949 | 0.8350148290636356 | 0.603448275862069 | 1.0 | 1.0 | 115/179 | 59.99% | 0 |
| Oxidative phosphorylation | 0.1142496981063722 | 1.2249093076286393 | 0.1525423728813559 | 0.7361948188317199 | 1.0 | 52/86 | 50.47% | 1 |
| Epstein-Barr virus infection | -0.0756836049783082 | -1.1357622653143702 | 0.3448275862068966 | 0.6369253352905851 | 1.0 | 62/179 | 25.54% | 1 |
| Thermogenesis | 0.0748845499516672 | 1.0927577331087608 | 0.2948717948717949 | 0.8631416491166021 | 1.0 | 134/159 | 77.74% | 1 |
| Ribosome | 0.0936709992384217 | 1.062166029297325 | 0.3787878787878788 | 0.8778851648488162 | 1.0 | 24/112 | 13.53% | 1 |
| Taste transduction | -0.1334816003429303 | -0.9706323778105916 | 0.4444444444444444 | 0.7495123224441576 | 1.0 | 23/32 | 57.23% | 1 |
| Ribosome | -0.3256476601525599 | -4.022551650465103 | 0.0 | 0.0 | 0.0 | 82/112 | 40.25% | 2 |
| Oxidative phosphorylation | -0.2609235596349908 | -2.837492429056554 | 0.0 | 0.0 | 0.0 | 53/86 | 34.95% | 2 |
| Epstein-Barr virus infection | -0.1484226108201942 | -2.3752365815471905 | 0.0 | 0.0057049714751426 | 0.06 | 88/179 | 33.48% | 2 |
| Thermogenesis | -0.0882268099454029 | -1.3402073272558992 | 0.0975609756097561 | 0.2926098272734442 | 1.0 | 113/159 | 61.71% | 2 |
| Taste transduction | -0.133293065125292 | -0.9230062434315842 | 0.4888888888888889 | 0.6908604815084033 | 1.0 | 13/32 | 26.30% | 2 |
| Oxidative phosphorylation | -0.3275733015959381 | -3.400119111390953 | 0.0 | 0.0 | 0.0 | 67/86 | 44.66% | 3 |
| Ribosome | -0.2617170297897189 | -3.2861659380353325 | 0.0 | 0.0 | 0.0 | 87/112 | 51.10% | 3 |
| Epstein-Barr virus infection | -0.1059999709738502 | -1.86209483044086 | 0.0 | 0.0579973024510487 | 0.56 | 88/179 | 37.69% | 3 |
| Thermogenesis | -0.1212049069795281 | -1.8205604074176376 | 0.0 | 0.0682293846797823 | 0.64 | 72/159 | 32.36% | 3 |
| Taste transduction | -0.2244361798488704 | -1.5172067052040372 | 0.0256410256410256 | 0.1659770744115658 | 1.0 | 14/32 | 19.96% | 3 |
| Ribosome | -0.2232951808963537 | -2.8364443956286745 | 0.0 | 0.0 | 0.0 | 79/112 | 47.23% | 4 |
| Oxidative phosphorylation | -0.1384100937509499 | -1.484898203180035 | 0.025 | 0.3228428314904074 | 0.99 | 72/86 | 69.16% | 4 |
| Taste transduction | -0.1599142914597349 | -1.1002297033883486 | 0.4193548387096774 | 0.6523335661482501 | 1.0 | 27/32 | 67.63% | 4 |
| Thermogenesis | 0.0627232772891256 | 0.8261903765388551 | 0.717948717948718 | 0.8583973003841902 | 1.0 | 139/159 | 82.01% | 4 |
| Epstein-Barr virus infection | 0.0475756092253786 | 0.6652826145585135 | 0.8552631578947368 | 0.951003837479759 | 1.0 | 101/179 | 53.24% | 4 |
| Epstein-Barr virus infection | -0.2321369172794 | -3.8752412064043895 | 0.0 | 0.0 | 0.0 | 76/179 | 18.53% | 5 |
| Ribosome | -0.2331264509898629 | -3.037331039310221 | 0.0 | 0.0 | 0.0 | 87/112 | 53.96% | 5 |
| Oxidative phosphorylation | -0.2113046485189271 | -2.5688251815532457 | 0.0 | 0.0008765821089591 | 0.01 | 68/86 | 57.49% | 5 |
| Thermogenesis | 0.0767866793683777 | 1.1101762451219783 | 0.3225806451612903 | 1.0 | 1.0 | 136/159 | 78.48% | 5 |
| Taste transduction | -0.1278100605290263 | -0.8108845295541842 | 0.6666666666666666 | 0.8161653728339712 | 1.0 | 21/32 | 51.73% | 5 |
| Ribosome | -0.3293229152882386 | -4.450933710867719 | 0.0 | 0.0 | 0.0 | 68/112 | 27.16% | 6 |
| Oxidative phosphorylation | -0.2803922514121083 | -2.9523157014033363 | 0.0 | 0.0 | 0.0 | 70/86 | 52.93% | 6 |
| Epstein-Barr virus infection | -0.1612168197723371 | -2.83990224960349 | 0.0 | 0.0 | 0.0 | 80/179 | 27.60% | 6 |
| Thermogenesis | -0.1045309528713468 | -1.6119871628753508 | 0.0526315789473684 | 0.1086496401862565 | 0.98 | 102/159 | 52.93% | 6 |
| Taste transduction | -0.0803663969469782 | -0.5944365699407909 | 0.9473684210526316 | 0.950402144772118 | 1.0 | 22/32 | 59.83% | 6 |
| Epstein-Barr virus infection | -0.1065109835343221 | -1.744106362733484 | 0.0 | 0.1176882575158902 | 0.78 | 49/179 | 15.52% | 7 |
| Ribosome | 0.1120774229219907 | 1.279890247034121 | 0.1076923076923077 | 0.7245521177194651 | 1.0 | 58/112 | 42.30% | 7 |
| Oxidative phosphorylation | 0.108161573283124 | 1.119584501050699 | 0.3442622950819672 | 0.9349297823532704 | 1.0 | 75/86 | 77.11% | 7 |
| Thermogenesis | 0.0715029801756083 | 1.0268194922167353 | 0.463768115942029 | 0.9815923459967548 | 1.0 | 129/159 | 75.05% | 7 |
| Taste transduction | 0.1249958881658193 | 0.8067070959874992 | 0.6229508196721312 | 1.0 | 1.0 | 23/32 | 61.00% | 7 |
| Thermogenesis | 0.1367508982822331 | 1.9891285890048445 | 0.0 | 0.1958251871161404 | 0.47 | 88/159 | 42.83% | 8 |
| Ribosome | 0.1454806205398361 | 1.7462592473255552 | 0.0408163265306122 | 0.3844675970986771 | 0.87 | 84/112 | 61.29% | 8 |
| Oxidative phosphorylation | 0.1309589399054042 | 1.3715003663606282 | 0.0666666666666666 | 0.7863250756787104 | 1.0 | 31/86 | 24.00% | 8 |
| Epstein-Barr virus infection | -0.0775275389761539 | -1.2408344861416116 | 0.1379310344827586 | 0.4750103126804719 | 1.0 | 105/179 | 50.05% | 8 |
| Taste transduction | -0.1302957666471551 | -0.876358542353001 | 0.5769230769230769 | 0.8278352697259158 | 1.0 | 12/32 | 23.35% | 8 |
| Thermogenesis | -0.1337794037804547 | -2.100302981011935 | 0.0 | 0.0240763092961898 | 0.22 | 50/159 | 16.67% | 9 |
| Oxidative phosphorylation | -0.1564861632838065 | -1.6899188229068458 | 0.0 | 0.1199516123863745 | 0.88 | 29/86 | 16.56% | 9 |
| Epstein-Barr virus infection | -0.0805443311691613 | -1.3345202650671202 | 0.125 | 0.3780101242756428 | 1.0 | 74/179 | 31.81% | 9 |
| Ribosome | -0.0944119478879425 | -1.2113882241551532 | 0.2222222222222222 | 0.5218758215459836 | 1.0 | 24/112 | 10.81% | 9 |
| Taste transduction | 0.142512319190756 | 0.8988740228825484 | 0.5555555555555556 | 1.0 | 1.0 | 8/32 | 12.37% | 9 |
| Ribosome | -0.1971230410161868 | -2.6788469240337296 | 0.0 | 0.0025584472871636 | 0.01 | 72/112 | 43.60% | 10 |
| Epstein-Barr virus infection | -0.1636372192150541 | -2.540005734409352 | 0.0 | 0.0031980591089545 | 0.02 | 85/179 | 29.82% | 10 |
| Oxidative phosphorylation | 0.1500924419982399 | 1.5510957038556674 | 0.0307692307692307 | 0.3611661477497877 | 0.99 | 52/86 | 46.85% | 10 |
| Thermogenesis | 0.1123031774550938 | 1.5077141535503211 | 0.088235294117647 | 0.4161262137117119 | 0.99 | 90/159 | 46.85% | 10 |
| Taste transduction | 0.1335782743774241 | 0.8638481365202263 | 0.65625 | 0.9752898173595196 | 1.0 | 21/32 | 53.91% | 10 |
| Epstein-Barr virus infection | -0.1548800515752707 | -2.431047289293832 | 0.0 | 0.0014575470250698 | 0.01 | 82/179 | 28.95% | 11 |
| Oxidative phosphorylation | 0.1182173254761076 | 1.1944193154577427 | 0.2622950819672131 | 0.6175221922037823 | 1.0 | 66/86 | 65.98% | 11 |
| Thermogenesis | 0.0840036749263786 | 1.118619955095944 | 0.28125 | 0.6860132448708288 | 1.0 | 95/159 | 52.77% | 11 |
| Ribosome | -0.0782791132880562 | -1.0479330673375782 | 0.3823529411764705 | 0.5760692258123907 | 1.0 | 71/112 | 54.18% | 11 |
| Taste transduction | -0.142353790073346 | -0.9466568114068772 | 0.4883720930232558 | 0.7009968306571455 | 1.0 | 12/32 | 21.73% | 11 |
| Ribosome | -0.1590080895545469 | -2.112035134935741 | 0.0 | 0.0280182897168985 | 0.25 | 82/112 | 56.27% | 12 |
| Oxidative phosphorylation | -0.1324742005360538 | -1.4618884159029404 | 0.081081081081081 | 0.2967726739750436 | 1.0 | 61/86 | 56.39% | 12 |
| Epstein-Barr virus infection | -0.0828798991794913 | -1.3064769028044512 | 0.1304347826086956 | 0.3912372091377831 | 1.0 | 61/179 | 24.35% | 12 |
| Taste transduction | -0.1640753539772809 | -1.2020766986467593 | 0.2285714285714285 | 0.4926908484063849 | 1.0 | 9/32 | 10.31% | 12 |
| Thermogenesis | 0.0430498109286673 | 0.6098984184263628 | 0.935064935064935 | 1.0 | 1.0 | 138/159 | 83.46% | 12 |
| Thermogenesis | 0.1106799758428105 | 1.552844450861132 | 0.1212121212121212 | 0.501035356316775 | 1.0 | 84/159 | 43.54% | 13 |
| Taste transduction | 0.1898271430862881 | 1.238439155252835 | 0.2 | 0.8120975144105445 | 1.0 | 8/32 | 7.70% | 13 |
| Ribosome | 0.101008067665986 | 1.158800145280816 | 0.2816901408450704 | 0.7224988615683561 | 1.0 | 57/112 | 42.36% | 13 |
| Oxidative phosphorylation | 0.0893802743756362 | 0.9178164713944112 | 0.5151515151515151 | 0.7740193224552736 | 1.0 | 75/86 | 79.11% | 13 |
| Epstein-Barr virus infection | -0.0369440849574122 | -0.5773223972275305 | 1.0 | 0.9895624758934112 | 1.0 | 162/179 | 86.23% | 13 |
| Ribosome | -0.2499602236106113 | -3.402811022966832 | 0.0 | 0.0 | 0.0 | 70/112 | 36.31% | 14 |
| Oxidative phosphorylation | -0.2533110757351272 | -2.8823733635911024 | 0.0 | 0.0 | 0.0 | 58/86 | 40.98% | 14 |
| Epstein-Barr virus infection | -0.1403322253608299 | -2.563615113060528 | 0.0 | 0.0026311111111111 | 0.01 | 89/179 | 34.09% | 14 |
| Thermogenesis | -0.0782485007267294 | -1.1257550331930222 | 0.2941176470588235 | 0.5246178861788618 | 1.0 | 110/159 | 60.17% | 14 |
| Taste transduction | 0.1120801633981848 | 0.7121301279549598 | 0.8653846153846154 | 1.0 | 1.0 | 29/32 | 79.91% | 14 |
| Epstein-Barr virus infection | -0.1564731165548349 | -2.6231667792912323 | 0.0 | 0.0 | 0.0 | 58/179 | 15.80% | 15 |
| Oxidative phosphorylation | -0.1427779365221995 | -1.5459902681911917 | 0.0714285714285714 | 0.1606630509590157 | 1.0 | 45/86 | 37.13% | 15 |
| Ribosome | -0.0835052583452637 | -1.0457258246958143 | 0.4473684210526316 | 0.6350407605087761 | 1.0 | 64/112 | 47.89% | 15 |
| Thermogenesis | 0.0685810193717143 | 0.9241792850269684 | 0.609375 | 0.9146829405107706 | 1.0 | 86/159 | 48.40% | 15 |
| Taste transduction | -0.1089751365711126 | -0.8046743421424567 | 0.7021276595744681 | 0.85876914142102 | 1.0 | 22/32 | 56.79% | 15 |
| Oxidative phosphorylation | -0.1934690155536926 | -2.25791714368282 | 0.0 | 0.013528491052921 | 0.1 | 70/86 | 61.20% | 16 |
| Thermogenesis | -0.119985713017665 | -2.059011466760391 | 0.0 | 0.0243252675663099 | 0.25 | 118/159 | 61.23% | 16 |
| Ribosome | -0.0986916187640376 | -1.2343055799625748 | 0.2413793103448276 | 0.5549282964592411 | 1.0 | 93/112 | 72.43% | 16 |
| Epstein-Barr virus infection | -0.0735439406826843 | -1.2010141513757822 | 0.25 | 0.5515253606328525 | 1.0 | 121/179 | 59.04% | 16 |
| Taste transduction | -0.1417973572809813 | -0.9239386376720244 | 0.6538461538461539 | 0.8495865697820095 | 1.0 | 15/32 | 31.04% | 16 |
| Ribosome | -0.1429843414461468 | -1.935132220388343 | 0.0 | 0.0402945619335347 | 0.44 | 86/112 | 61.80% | 17 |
| Epstein-Barr virus infection | -0.0989803694223771 | -1.6377537736470795 | 0.0 | 0.1405948784347576 | 0.95 | 86/179 | 37.03% | 17 |
| Oxidative phosphorylation | -0.1321516800386444 | -1.5503137660883377 | 0.0294117647058823 | 0.1873047842836362 | 0.98 | 47/86 | 40.21% | 17 |
| Thermogenesis | -0.0495010447489337 | -0.753272068712083 | 0.8709677419354839 | 0.9325568175837132 | 1.0 | 80/159 | 44.25% | 17 |
| Taste transduction | 0.0837504809450444 | 0.5391166734889357 | 1.0 | 0.986275292019886 | 1.0 | 18/32 | 49.33% | 17 |
| Ribosome | -0.2300643695538789 | -3.032637747261451 | 0.0 | 0.0 | 0.0 | 67/112 | 36.08% | 18 |
| Oxidative phosphorylation | -0.2258420027873637 | -2.426969664198208 | 0.0 | 0.0030002293168904 | 0.03 | 61/86 | 47.80% | 18 |
| Epstein-Barr virus infection | -0.163436713082365 | -2.312736020117996 | 0.0 | 0.0060671303963784 | 0.07 | 66/179 | 19.74% | 18 |
| Thermogenesis | -0.0540860555409017 | -0.8808549381934389 | 0.6428571428571429 | 0.7158308325876414 | 1.0 | 137/159 | 80.28% | 18 |
| Taste transduction | 0.0903422763299656 | 0.5715602775591783 | 0.9423076923076924 | 0.9854396948327132 | 1.0 | 12/32 | 29.38% | 18 |
| Ribosome | -0.2865315189205287 | -3.5741418374315392 | 0.0 | 0.0 | 0.0 | 80/112 | 41.74% | 19 |
| Oxidative phosphorylation | -0.228410490583011 | -2.4631593708013453 | 0.0 | 0.0 | 0.0 | 40/86 | 22.24% | 19 |
| Epstein-Barr virus infection | -0.1084580718550667 | -1.9171496093342697 | 0.0 | 0.0473985890652557 | 0.4 | 77/179 | 30.61% | 19 |
| Thermogenesis | -0.0895152851569006 | -1.32764710460203 | 0.1363636363636363 | 0.420344560588463 | 1.0 | 132/159 | 73.22% | 19 |
| Taste transduction | 0.1926586094746618 | 1.1079070707887402 | 0.3859649122807017 | 0.806349607678106 | 1.0 | 30/32 | 74.82% | 19 |
| Epstein-Barr virus infection | -0.0853369011310599 | -1.5479124740309584 | 0.0303030303030303 | 0.2999792957090357 | 0.97 | 46/179 | 15.94% | 20 |
| Ribosome | -0.1077871147797501 | -1.3732178053318371 | 0.0666666666666666 | 0.3707169657065273 | 1.0 | 59/112 | 40.58% | 20 |
| Taste transduction | -0.147339358915744 | -1.0153596695405562 | 0.4722222222222222 | 0.697960819046414 | 1.0 | 12/32 | 21.09% | 20 |
| Thermogenesis | 0.0542222308163235 | 0.7571389446423754 | 0.7866666666666666 | 0.9763971347870776 | 1.0 | 150/159 | 89.39% | 20 |
| Oxidative phosphorylation | 0.0704483502934994 | 0.7206262365493948 | 0.8235294117647058 | 0.9541908068839632 | 1.0 | 84/86 | 90.89% | 20 |
| Ribosome | 0.227543804362252 | 2.561665516577705 | 0.0 | 0.0061374558508482 | 0.01 | 72/112 | 42.71% | 21 |
| Epstein-Barr virus infection | -0.1441302848400509 | -2.324946604635544 | 0.0 | 0.0037351456909816 | 0.05 | 112/179 | 47.05% | 21 |
| Taste transduction | 0.2050893693918116 | 1.3499798123440814 | 0.1639344262295081 | 1.0 | 1.0 | 31/32 | 76.55% | 21 |
| Oxidative phosphorylation | -0.11407752268435 | -1.2261204071398637 | 0.2258064516129032 | 0.3921902975530735 | 1.0 | 28/86 | 20.10% | 21 |
| Thermogenesis | -0.0423312481418022 | -0.7048522598483786 | 0.935483870967742 | 0.919570897674442 | 1.0 | 152/159 | 91.18% | 21 |
| Epstein-Barr virus infection | -0.104099107261312 | -1.5395196576693624 | 0.08 | 0.137222527976518 | 0.99 | 38/179 | 9.71% | 22 |
| Ribosome | -0.1188469970281153 | -1.5179883663741862 | 0.0666666666666666 | 0.1478504771584324 | 0.99 | 32/112 | 15.52% | 22 |
| Oxidative phosphorylation | -0.1063207806531286 | -1.2234510409142016 | 0.1666666666666666 | 0.3499987708812442 | 1.0 | 64/86 | 62.78% | 22 |
| Taste transduction | -0.1162346220625998 | -0.8115098576982108 | 0.6571428571428571 | 0.7833040746558256 | 1.0 | 17/32 | 39.94% | 22 |
| Thermogenesis | 0.0308342620424767 | 0.4293938133054199 | 1.0 | 0.9970842273202398 | 1.0 | 29/159 | 16.15% | 22 |
| Epstein-Barr virus infection | -0.1490222125794351 | -2.4517198895361614 | 0.0 | 0.0081440791451449 | 0.04 | 94/179 | 36.77% | 23 |
| Oxidative phosphorylation | -0.1900501123715608 | -1.909310181176379 | 0.0 | 0.0470320570632122 | 0.42 | 71/86 | 63.12% | 23 |
| Taste transduction | -0.1815968266524949 | -1.3009534103446785 | 0.1428571428571428 | 0.3487975389103511 | 1.0 | 10/32 | 12.03% | 23 |
| Thermogenesis | -0.0755688546724218 | -1.0832749743737218 | 0.303030303030303 | 0.5208393115793502 | 1.0 | 61/159 | 29.92% | 23 |
| Ribosome | 0.0520079995085884 | 0.64139179465484 | 0.8679245283018868 | 0.955425070683007 | 1.0 | 53/112 | 43.18% | 23 |
| Oxidative phosphorylation | -0.2140088342006968 | -2.5581159053371945 | 0.0 | 0.0010142228443316 | 0.01 | 46/86 | 31.19% | 24 |
| Epstein-Barr virus infection | -0.1580411043735501 | -2.550278082209685 | 0.0 | 0.0009466079880428 | 0.01 | 60/179 | 16.72% | 24 |
| Taste transduction | -0.1907365919825491 | -1.293927415862764 | 0.2156862745098039 | 0.3142951241198439 | 1.0 | 17/32 | 32.98% | 24 |
| Ribosome | -0.0860341248910368 | -1.1004974235817673 | 0.3103448275862069 | 0.4815245370754795 | 1.0 | 106/112 | 85.90% | 24 |
| Thermogenesis | -0.0711413956891996 | -1.0751266704488225 | 0.3333333333333333 | 0.5077702335860602 | 1.0 | 117/159 | 65.75% | 24 |
| Thermogenesis | 0.1659622741611206 | 2.2467535634369544 | 0.0 | 0.041135225375626 | 0.17 | 116/159 | 57.68% | 25 |
| Oxidative phosphorylation | 0.188872351022407 | 2.166578649559165 | 0.0 | 0.0467445742904841 | 0.25 | 40/86 | 29.28% | 25 |
| Ribosome | -0.1565338198474687 | -1.860725280286564 | 0.0285714285714285 | 0.1290420066485343 | 0.64 | 69/112 | 44.77% | 25 |
| Taste transduction | -0.166541500257256 | -1.2417514094893585 | 0.1489361702127659 | 0.5870489573889394 | 1.0 | 11/32 | 16.31% | 25 |
| Epstein-Barr virus infection | -0.0471492858169478 | -0.753211572826966 | 0.8888888888888888 | 0.9572471118372036 | 1.0 | 148/179 | 77.11% | 25 |
| Epstein-Barr virus infection | -0.1120298340814006 | -2.0146072424615937 | 0.0 | 0.0252861368312757 | 0.23 | 103/179 | 44.77% | 26 |
| Oxidative phosphorylation | -0.1609719325626107 | -1.7046833340244654 | 0.03125 | 0.1078765707671957 | 0.75 | 44/86 | 33.41% | 26 |
| Ribosome | -0.1061645417419538 | -1.3642996693988605 | 0.15 | 0.3313058609825103 | 1.0 | 73/112 | 53.04% | 26 |
| Taste transduction | 0.1592193386692686 | 1.0458250483076112 | 0.3461538461538461 | 0.9616015093405912 | 1.0 | 10/32 | 17.21% | 26 |
| Thermogenesis | 0.0563497911713277 | 0.7360008639741472 | 0.8260869565217391 | 0.9576472894762056 | 1.0 | 90/159 | 52.59% | 26 |
| Ribosome | -0.1964814337453577 | -2.496369619941877 | 0.0 | 0.0 | 0.0 | 89/112 | 58.81% | 27 |
| Epstein-Barr virus infection | -0.1227396672517917 | -2.377325966460432 | 0.0 | 0.0035433331129767 | 0.02 | 102/179 | 42.99% | 27 |
| Taste transduction | -0.116205108934208 | -0.8434816132145562 | 0.6415094339622641 | 0.767727265474158 | 1.0 | 11/32 | 21.36% | 27 |
| Thermogenesis | 0.0549077516421173 | 0.7168894036733295 | 0.8412698412698413 | 0.9068752813784474 | 1.0 | 62/159 | 35.14% | 27 |
| Oxidative phosphorylation | -0.0615741373995872 | -0.6477804376991224 | 0.9117647058823528 | 0.9522364394893263 | 1.0 | 33/86 | 30.82% | 27 |
| Epstein-Barr virus infection | -0.1655189474382541 | -2.870958226917452 | 0.0 | 0.0 | 0.0 | 77/179 | 25.63% | 28 |
| Ribosome | -0.2069016676747544 | -2.800232676708195 | 0.0 | 0.0 | 0.0 | 80/112 | 50.28% | 28 |
| Thermogenesis | 0.1313157689829146 | 1.88494187185635 | 0.0 | 0.4625279304305816 | 0.67 | 110/159 | 56.99% | 28 |
| Taste transduction | -0.1471600495573211 | -0.9981835364480396 | 0.391304347826087 | 0.6522519625967902 | 1.0 | 30/32 | 78.87% | 28 |
| Oxidative phosphorylation | -0.0719395400172545 | -0.7948374235753343 | 0.7 | 0.858257819010995 | 1.0 | 85/86 | 91.65% | 28 |
| Ribosome | -0.3640114664957908 | -4.447733367143593 | 0.0 | 0.0 | 0.0 | 77/112 | 31.66% | 29 |
| Oxidative phosphorylation | -0.2743933145830459 | -3.267497189735176 | 0.0 | 0.0 | 0.0 | 57/86 | 37.99% | 29 |
| Thermogenesis | -0.1322448508758917 | -2.085066924759269 | 0.0227272727272727 | 0.0196933560477001 | 0.24 | 96/159 | 46.08% | 29 |
| Epstein-Barr virus infection | 0.0919162478505591 | 1.2637288151803443 | 0.1571428571428571 | 0.5393307349274904 | 1.0 | 58/179 | 24.48% | 29 |
| Taste transduction | -0.1409379951415683 | -0.9495072224306784 | 0.6 | 0.9666477380276356 | 1.0 | 15/32 | 31.53% | 29 |
| Oxidative phosphorylation | -0.2601385384155186 | -3.008403321908917 | 0.0 | 0.0 | 0.0 | 69/86 | 53.33% | 30 |
| Ribosome | -0.2225102822614033 | -2.9368797355344225 | 0.0 | 0.0 | 0.0 | 73/112 | 41.60% | 30 |
| Thermogenesis | -0.1091719810812835 | -1.7218181735948006 | 0.0 | 0.0964372318597276 | 0.76 | 114/159 | 59.59% | 30 |
| Taste transduction | -0.1681589880592587 | -1.1892261384806169 | 0.2325581395348837 | 0.4385299727053663 | 1.0 | 26/32 | 63.47% | 30 |
| Epstein-Barr virus infection | 0.0557083601545044 | 0.8433363336922206 | 0.625 | 0.7871948695747323 | 1.0 | 25/179 | 9.59% | 30 |
| Ribosome | -0.3088603817864165 | -4.5095090913599805 | 0.0 | 0.0 | 0.0 | 67/112 | 27.76% | 31 |
| Oxidative phosphorylation | -0.2280577290437473 | -2.633010961111174 | 0.0 | 0.0067149673087117 | 0.03 | 46/86 | 29.19% | 31 |
| Thermogenesis | -0.1089243720432984 | -1.6118959971372937 | 0.0357142857142857 | 0.1765077121147098 | 0.94 | 58/159 | 24.24% | 31 |
| Taste transduction | 0.1692150451803977 | 1.1212167825319377 | 0.2666666666666666 | 1.0 | 1.0 | 8/32 | 9.57% | 31 |
| Epstein-Barr virus infection | -0.0540938598818739 | -0.8876800724429279 | 0.5625 | 0.8275281530626998 | 1.0 | 108/179 | 53.43% | 31 |
| Ribosome | -0.2578530355840409 | -3.32931674505725 | 0.0 | 0.0 | 0.0 | 51/112 | 18.35% | 32 |
| Oxidative phosphorylation | -0.3010482068914579 | -3.251304941949133 | 0.0 | 0.0 | 0.0 | 55/86 | 32.65% | 32 |
| Thermogenesis | -0.195088115043457 | -3.1460939081379293 | 0.0 | 0.0 | 0.0 | 84/159 | 31.91% | 32 |
| Taste transduction | -0.1598448791793256 | -1.1499614807625744 | 0.2391304347826087 | 0.5808088388094853 | 1.0 | 9/32 | 10.34% | 32 |
| Epstein-Barr virus infection | 0.0430492200213103 | 0.6365137133053301 | 0.8985507246376812 | 0.9827247306416328 | 1.0 | 116/179 | 61.99% | 32 |
| Ribosome | -0.3630708010063278 | -4.739782665252094 | 0.0 | 0.0 | 0.0 | 85/112 | 39.11% | 33 |
| Oxidative phosphorylation | -0.3017187042155527 | -3.166046240333104 | 0.0 | 0.0 | 0.0 | 70/86 | 50.74% | 33 |
| Thermogenesis | -0.1383763649312712 | -2.030776858527007 | 0.0 | 0.0216223562388516 | 0.33 | 107/159 | 52.73% | 33 |
| Epstein-Barr virus infection | -0.1147388570848083 | -1.731023342971171 | 0.0344827586206896 | 0.0902363833497384 | 0.8 | 68/179 | 25.48% | 33 |
| Taste transduction | -0.2120107719433776 | -1.489015373061367 | 0.0980392156862745 | 0.2118406523401002 | 0.99 | 14/32 | 21.31% | 33 |
| Oxidative phosphorylation | -0.2545196538658869 | -2.853425457553988 | 0.0 | 0.0 | 0.0 | 59/86 | 41.96% | 34 |
| Epstein-Barr virus infection | -0.1406276819473865 | -2.199380839754852 | 0.0 | 0.0069618490671122 | 0.08 | 101/179 | 40.64% | 34 |
| Ribosome | -0.1404734864995973 | -1.6958172669378744 | 0.0 | 0.0890740364424574 | 0.86 | 55/112 | 33.62% | 34 |
| Taste transduction | -0.1327015686283943 | -0.9523420525198376 | 0.4772727272727273 | 0.6694432158828444 | 1.0 | 8/32 | 10.41% | 34 |
| Thermogenesis | -0.0642275325875987 | -0.9438060795186816 | 0.5121951219512195 | 0.6694621168305378 | 1.0 | 46/159 | 21.19% | 34 |
| Epstein-Barr virus infection | -0.1337794321853186 | -2.1911135547340184 | 0.0 | 0.0103028113374597 | 0.14 | 66/179 | 22.07% | 35 |
| Thermogenesis | 0.1169191327605175 | 1.5391984192601724 | 0.0422535211267605 | 0.8218183832899328 | 1.0 | 88/159 | 45.25% | 35 |
| Oxidative phosphorylation | 0.0975026009723607 | 1.0324645616114798 | 0.4 | 0.9667967358698148 | 1.0 | 53/86 | 53.19% | 35 |
| Taste transduction | -0.1216536857868528 | -0.8727845560307057 | 0.5957446808510638 | 0.7538368738593555 | 1.0 | 19/32 | 45.99% | 35 |
| Ribosome | -0.0485747667164141 | -0.5931017707095675 | 0.9393939393939394 | 0.9753651291582288 | 1.0 | 84/112 | 69.03% | 35 |
| Oxidative phosphorylation | -0.2263274299523434 | -2.737231883792341 | 0.0 | 0.0 | 0.0 | 64/86 | 50.58% | 36 |
| Thermogenesis | -0.1201911536898513 | -1.9299521792350136 | 0.0 | 0.0538910872432636 | 0.44 | 93/159 | 44.78% | 36 |
| Taste transduction | 0.211778769747215 | 1.3365944361544937 | 0.173076923076923 | 0.5778762826788498 | 1.0 | 24/32 | 55.02% | 36 |
| Ribosome | -0.1034471847868074 | -1.2945477592320158 | 0.125 | 0.3413102192073362 | 1.0 | 49/112 | 31.63% | 36 |
| Epstein-Barr virus infection | -0.0576208425867244 | -0.9646737080773496 | 0.5 | 0.700175868349978 | 1.0 | 43/179 | 16.70% | 36 |
| Ribosome | -0.2780402190998905 | -3.7626702164566943 | 0.0 | 0.0 | 0.0 | 80/112 | 42.60% | 37 |
| Oxidative phosphorylation | -0.1833048188200864 | -2.050333616747916 | 0.0 | 0.0303655660377358 | 0.28 | 76/86 | 69.50% | 37 |
| Epstein-Barr virus infection | -0.0958088948657161 | -1.629439063271938 | 0.0 | 0.1462421909509686 | 0.86 | 102/179 | 46.10% | 37 |
| Thermogenesis | -0.0836067027722523 | -1.3190298606653903 | 0.1666666666666666 | 0.3955309627479438 | 1.0 | 152/159 | 87.09% | 37 |
| Taste transduction | 0.1273989683571168 | 0.8722501590968771 | 0.576271186440678 | 0.7863051117965717 | 1.0 | 6/32 | 7.30% | 37 |
| Thermogenesis | 0.1543315294185555 | 2.100021315152871 | 0.0149253731343283 | 0.2416159380188157 | 0.42 | 97/159 | 47.08% | 38 |
| Ribosome | -0.1395500689620349 | -2.0428100405841603 | 0.0 | 0.0313967673071058 | 0.19 | 88/112 | 63.77% | 38 |
| Epstein-Barr virus infection | -0.0968682231117353 | -1.5861196199728786 | 0.0 | 0.163543404553038 | 0.91 | 48/179 | 15.98% | 38 |
| Oxidative phosphorylation | 0.1404503366589868 | 1.4065092972713316 | 0.1166666666666666 | 0.7813613724405091 | 1.0 | 51/86 | 46.79% | 38 |
| Taste transduction | -0.1333043674603921 | -0.933825652743302 | 0.5853658536585366 | 0.7637583520474647 | 1.0 | 23/32 | 57.40% | 38 |
| Epstein-Barr virus infection | -0.1208544640341385 | -1.8792193758859483 | 0.0 | 0.0779192913084126 | 0.6 | 71/179 | 26.75% | 39 |
| Ribosome | -0.1214087761786404 | -1.6738897666231272 | 0.0277777777777777 | 0.1330911766058209 | 0.86 | 64/112 | 44.26% | 39 |
| Oxidative phosphorylation | -0.1337418025070802 | -1.4844452740123826 | 0.0909090909090909 | 0.2296105568072095 | 0.98 | 66/86 | 62.70% | 39 |
| Taste transduction | -0.1912523199587957 | -1.3816456726391608 | 0.1333333333333333 | 0.3064158024799114 | 1.0 | 13/32 | 20.46% | 39 |
| Thermogenesis | 0.0383834139505656 | 0.5746257618987081 | 0.9696969696969696 | 1.0 | 1.0 | 140/159 | 84.73% | 39 |
| Ribosome | -0.1330614426196696 | -1.6552021458471902 | 0.0476190476190476 | 0.1990882759126407 | 0.9 | 96/112 | 71.95% | 40 |
| Taste transduction | -0.1995615584775341 | -1.532337126571576 | 0.0465116279069767 | 0.2384375715795661 | 1.0 | 9/32 | 6.79% | 40 |
| Oxidative phosphorylation | -0.1128483370623297 | -1.2799354987414728 | 0.1707317073170731 | 0.3859904273473485 | 1.0 | 50/86 | 45.85% | 40 |
| Thermogenesis | -0.069802249076307 | -0.9963950374822887 | 0.4615384615384615 | 0.6727154407149333 | 1.0 | 47/159 | 21.68% | 40 |
| Epstein-Barr virus infection | -0.0423759172324384 | -0.6557109849000161 | 0.9333333333333332 | 0.9560671357686714 | 1.0 | 36/179 | 15.15% | 40 |
| Oxidative phosphorylation | -0.3178452567996256 | -3.2308443138546794 | 0.0 | 0.0 | 0.0 | 60/86 | 36.90% | 41 |
| Ribosome | -0.2398031423342298 | -3.2183620427351154 | 0.0 | 0.0 | 0.0 | 62/112 | 29.85% | 41 |
| Epstein-Barr virus infection | -0.0925176446724106 | -1.5139748150739774 | 0.0714285714285714 | 0.1867903688977247 | 1.0 | 94/179 | 41.80% | 41 |
| Thermogenesis | -0.0858185058777302 | -1.2891483325508557 | 0.1388888888888889 | 0.362055310413939 | 1.0 | 73/159 | 35.91% | 41 |
| Taste transduction | 0.1355329698643851 | 0.9167825200406078 | 0.6101694915254238 | 0.8683607535453032 | 1.0 | 11/32 | 22.45% | 41 |
| Ribosome | -0.1640123151182866 | -2.203301450185885 | 0.0 | 0.0058554195650259 | 0.03 | 47/112 | 24.40% | 42 |
| Epstein-Barr virus infection | -0.1240266084660245 | -2.058553522884664 | 0.0 | 0.0156144521734026 | 0.12 | 58/179 | 18.85% | 42 |
| Oxidative phosphorylation | -0.1371541527892488 | -1.525522447158714 | 0.0789473684210526 | 0.164211988690284 | 0.99 | 31/86 | 21.18% | 42 |
| Thermogenesis | 0.0720718234332094 | 1.0664691857698492 | 0.4 | 0.6857369966176559 | 1.0 | 76/159 | 41.91% | 42 |
| Taste transduction | 0.1515671345565241 | 1.030931393872168 | 0.4230769230769231 | 0.7313310413885615 | 1.0 | 18/32 | 42.56% | 42 |
| Oxidative phosphorylation | -0.2727001833743558 | -3.240032787152209 | 0.0 | 0.0 | 0.0 | 61/86 | 42.59% | 43 |
| Ribosome | -0.2035206597539268 | -2.6583790691907123 | 0.0 | 0.0 | 0.0 | 54/112 | 26.37% | 43 |
| Epstein-Barr virus infection | -0.1166104695400355 | -1.923537561138284 | 0.0 | 0.0455769077000986 | 0.48 | 63/179 | 22.09% | 43 |
| Thermogenesis | -0.1254676352054362 | -1.845754742062561 | 0.0 | 0.0558676937018577 | 0.64 | 118/159 | 60.51% | 43 |
| Taste transduction | -0.0841319508329352 | -0.5943039962466512 | 0.9310344827586208 | 0.9839238785681922 | 1.0 | 5/32 | 6.19% | 43 |
| Oxidative phosphorylation | -0.2321417570446411 | -2.608379742905366 | 0.0 | 0.0050045703839122 | 0.03 | 45/86 | 27.77% | 44 |
| Thermogenesis | -0.1553811792791847 | -2.413857190150817 | 0.0 | 0.0033363802559414 | 0.03 | 73/159 | 29.10% | 44 |
| Epstein-Barr virus infection | -0.0995353249186391 | -1.777463698855575 | 0.0 | 0.0973110907982937 | 0.78 | 65/179 | 24.97% | 44 |
| Ribosome | -0.082706191996807 | -1.1703424936817148 | 0.1891891891891892 | 0.5609355503322596 | 1.0 | 94/112 | 74.95% | 44 |
| Taste transduction | -0.115225257303539 | -0.7379242450079108 | 0.8333333333333334 | 0.9559267559120128 | 1.0 | 18/32 | 43.30% | 44 |
| Epstein-Barr virus infection | -0.1483740711922618 | -2.5401627938768514 | 0.0 | 0.0028021015761821 | 0.01 | 94/179 | 36.32% | 45 |
| Ribosome | -0.1752683634281028 | -2.1107840486152813 | 0.0 | 0.0196147110332749 | 0.12 | 104/112 | 75.04% | 45 |
| Thermogenesis | -0.0971470308528347 | -1.6584616343711849 | 0.0344827586206896 | 0.1473697865992086 | 0.9 | 91/159 | 46.24% | 45 |
| Oxidative phosphorylation | -0.109608407267224 | -1.2988032557097435 | 0.1481481481481481 | 0.4687428614939465 | 1.0 | 33/86 | 26.02% | 45 |
| Taste transduction | -0.1551495073391745 | -1.0785825390232742 | 0.3076923076923077 | 0.6719363847209733 | 1.0 | 20/32 | 45.38% | 45 |
| Ribosome | -0.1431119957761128 | -1.872286502984436 | 0.0 | 0.0499218630888855 | 0.55 | 73/112 | 49.78% | 46 |
| Epstein-Barr virus infection | -0.0793189315278572 | -1.3118898412789644 | 0.1481481481481481 | 0.2911412835816963 | 1.0 | 123/179 | 59.66% | 46 |
| Thermogenesis | 0.0936037460391565 | 1.2548579471203485 | 0.1969696969696969 | 1.0 | 1.0 | 111/159 | 61.67% | 46 |
| Oxidative phosphorylation | 0.0813586157156401 | 0.8352347040497169 | 0.6470588235294118 | 0.9734071341051324 | 1.0 | 59/86 | 61.67% | 46 |
| Taste transduction | -0.0950583728437162 | -0.6570628315004804 | 0.8235294117647058 | 0.918104559291236 | 1.0 | 26/32 | 70.77% | 46 |
| Epstein-Barr virus infection | -0.1052548510684108 | -1.6942473121961907 | 0.0 | 0.1527238882643749 | 0.83 | 70/179 | 27.11% | 47 |
| Taste transduction | -0.1665622841165926 | -1.1589575677115576 | 0.1578947368421052 | 0.5307371527555014 | 1.0 | 27/32 | 66.92% | 47 |
| Ribosome | 0.0928670008454053 | 1.0481938467277976 | 0.3174603174603174 | 0.9726296283829868 | 1.0 | 60/112 | 46.00% | 47 |
| Oxidative phosphorylation | 0.0764689987773298 | 0.8349208881250927 | 0.6190476190476191 | 0.9295712285133726 | 1.0 | 27/86 | 25.32% | 47 |
| Thermogenesis | 0.0516962539223096 | 0.6632178111894441 | 0.8947368421052632 | 0.9660380807615692 | 1.0 | 77/159 | 44.90% | 47 |
| Ribosome | -0.2086899163590077 | -2.8344373214639305 | 0.0 | 0.0 | 0.0 | 92/112 | 60.70% | 48 |
| Oxidative phosphorylation | -0.1847751109308437 | -2.06696296442626 | 0.0 | 0.0289672920968369 | 0.27 | 57/86 | 46.54% | 48 |
| Taste transduction | -0.2350461769872571 | -1.5705780859429173 | 0.0344827586206896 | 0.229022653140617 | 0.97 | 17/32 | 27.56% | 48 |
| Epstein-Barr virus infection | -0.087385234757465 | -1.4178785746948572 | 0.0625 | 0.3207701610135031 | 1.0 | 53/179 | 19.58% | 48 |
| Thermogenesis | 0.0805316014676845 | 1.1124624427720406 | 0.2876712328767123 | 0.7825594377496813 | 1.0 | 56/159 | 28.71% | 48 |
| Oxidative phosphorylation | -0.2510168501713146 | -2.7585729333414672 | 0.0 | 0.0094270547318323 | 0.01 | 61/86 | 44.79% | 49 |
| Thermogenesis | -0.0826014889987527 | -1.424594202281424 | 0.0 | 0.3296102350879946 | 1.0 | 85/159 | 43.73% | 49 |
| Epstein-Barr virus infection | 0.1011614267036261 | 1.4215808102880614 | 0.1111111111111111 | 0.4581956717621208 | 1.0 | 46/179 | 17.31% | 49 |
| Taste transduction | -0.1487174195437644 | -1.1160397511139348 | 0.35 | 0.6191689547867472 | 1.0 | 20/32 | 46.14% | 49 |
| Ribosome | -0.0626830455571333 | -0.8321608130739544 | 0.6756756756756757 | 0.9056015526361488 | 1.0 | 109/112 | 90.87% | 49 |

**Appendix IV**

**A graph showing a line graph

AI-generated content may be incorrect.**

**Figure** **Appendix IV: Precision–recall curves for the top 15 latent-space-contributing genes.**

**Appendix V**

**A diagram of a gene expression

AI-generated content may be incorrect.**

**Figure Appendix V: Boxplot of gene expression levels for the top latent contributors across tumor (1) and control (0) conditions**