DVPNet: A New XAI-Based Interpretable Genetic Profiling Framework Using Nucleotide Transformer and Probabilistic Circuits

Taishi Kusumoto

March 24, 2026

Taishi Kusumoto
Independent Researcher
Email: tai8242@gmail.com
LinkedIn: https://www.linkedin.com/in/tai-kus-39490gpt

Abstract

In this study, we present an XAI-based genetic profiling framework that quantifies gene importance for distinguishing cancer cells from normal cells based on an interpretable AI decision process. We propose a new explainable AI (XAI) classification model that combines probabilistic circuits with the Nucleotide Transformer. By leveraging the strong feature-extraction capability of the Nucleotide Transformer, we design a tractable classification framework based on probabilistic circuits while preserving probabilistic interpretability.

To demonstrate the capability of this framework, we used the GSE131907 single-cell lung cancer atlas and constructed a dataset consisting of cancer-cell and normal-cell classes. From each sample, 900 gene types were randomly selected and converted into embedding vectors using the Nucleotide Transformer, after which the classification model was trained. We then extracted class-specific probabilistic contributions from the tractable model and defined a contribution score for the cancer-cell class. Genetic profiling was performed based on these scores, providing insights into which genes and biological pathways are most important for the classification task. Notably, 1,524 of the 9,540 observed genes showed contribution scores that contradicted what would be expected from their class-wise occurrence frequencies, suggesting that the profiling goes beyond simple statistics by leveraging biological feature representations encoded by the Nucleotide Transformer. The top-ranked genes among these contradictory cases include several well-studied genes in cancer research (e.g., ITGA5, SIGLEC9, NOTUM, and TP73). Overall, these analyses go beyond traditional statistical or gene-expression-level approaches and provide new academic insights for genetic research.

1 Introduction

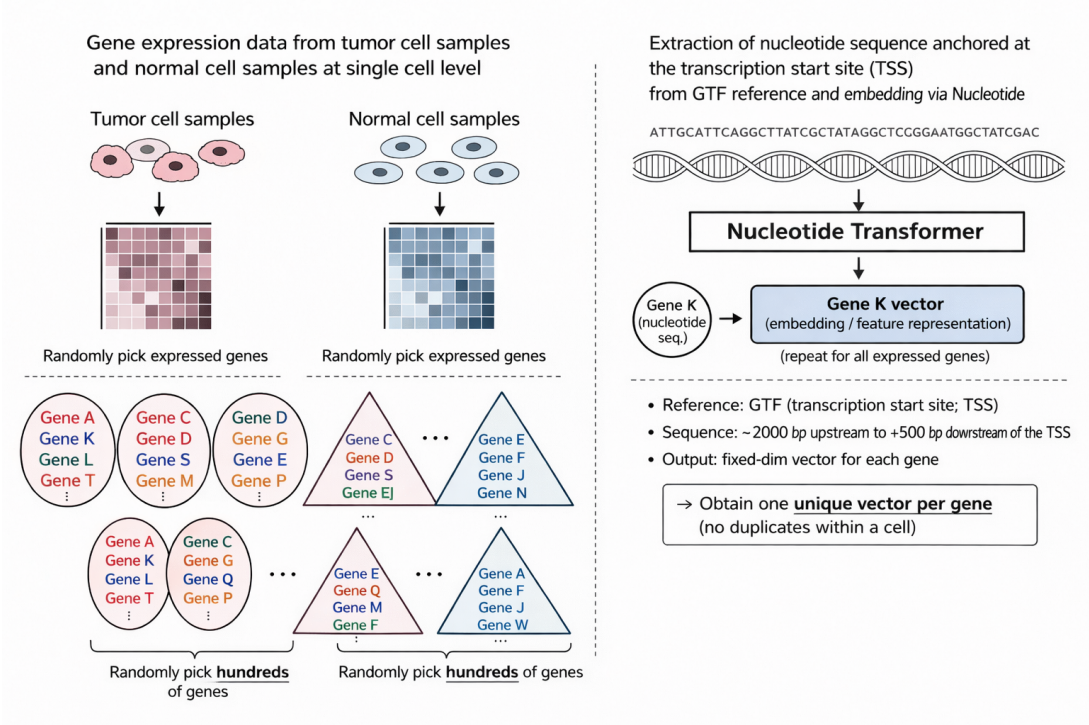
Gene co-expression networks constructed from RNA sequencing data have been used to infer gene functions and disease associations.[1] Indeed, previous studies have identified disease-associated gene modules and biologically meaningful pathways using co-expression network analysis, including in cancer [2] [3], neurological disorders [4], and complex traits.[5] However, gene co-expression networks, constructed from correlational relationships among RNA expression levels, primarily indicate which genes are active in the same biological processes but do not normally provide information about causality or distinguish between regulatory and regulated genes.[1] For instance, genes in the same biological pathway often do not show similar RNA expression patterns, which cannot be adequately treated in co-expression networks. [6] The fundamental reliance of co-expression networks on statistical correlations of expression levels restricts their ability to capture functional, regulatory, or context-dependent relationships among genes. Therefore, due to these limitations, further advancement of genetic research requires a new workflow to develop alternative genetic networks that provide different biological insights beyond statistical analyses of RNA expression levels.

The recent advancement of large language models clearly demonstrates that training transformer-based models on large amounts of text in a self-supervised manner enables strong contextual understanding of arbitrary input sequences, allowing models to solve diverse advanced tasks, including mathematical reasoning, scientific discussions, translation, and coding. [7] Similarly, a foundation model such as the Nucleotide Transformer [8], trained on large collections of nucleotide sequences for next-base prediction in a self-supervised manner, is expected to acquire contextual understanding of biological functions encoded in nucleotide sequences. Such genetic foundation models can encode expressed genes in RNA sequence data into meaningful feature representations that capture inherent biological

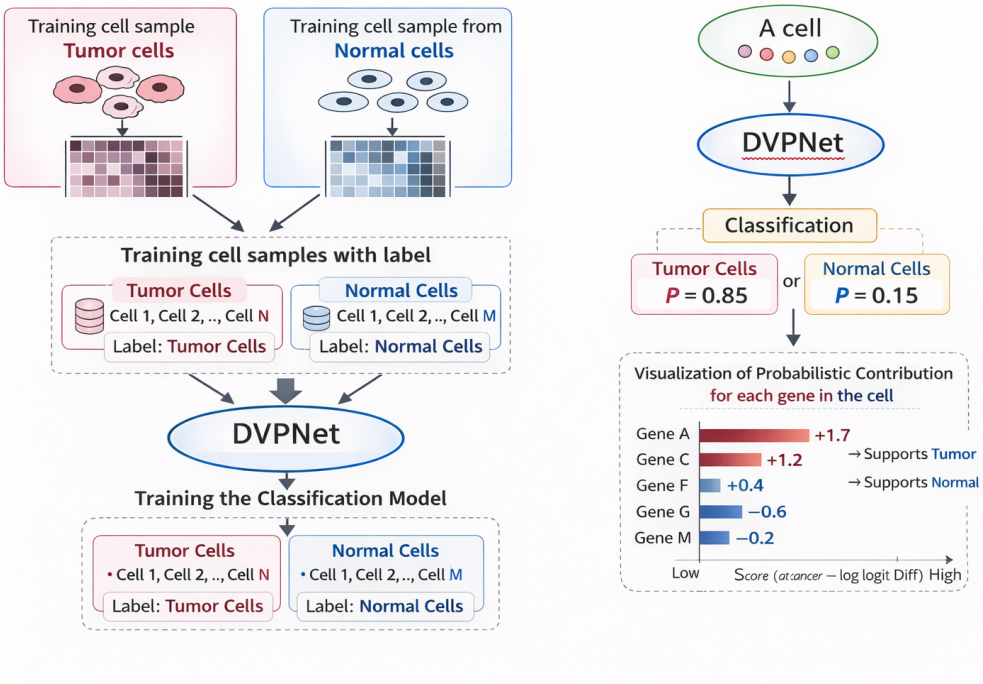
functions. Indeed, previous studies have shown their usefulness as encoders for downstream tasks.[8] Therefore, utilizing these representations can contribute to the construction of new types of genetic networks that provide biological insights distinct from those obtained through co-expression networks.

In this paper, we adopt the Nucleotide Transformer as an encoder for genes in RNA sequence data. We reconstruct nucleotide sequences encoding each gene by working backwards from transcription start sites (TSS). These nucleotide sequences are then input to obtain embedded feature representations for each expressed gene in a cell sample. Directly encoding mature RNA sequence data may lose information related to intronic regions; however, this approach avoids such information loss. The resulting embedded vectors are integrated into a novel explainable gene classification model, DVPNet, which is derived from VPNet, an explainable image classification model that combines embedded vectors from vision transformers with probabilistic circuits.

Gene classification models [9] can extract and select important features and internally identify genes that are important for specific classification tasks, such as distinguishing cancer cells from normal cells. However, the black-box nature of traditional classification models based on intractable neural networks, including convolutional neural networks [10] and transformers [11], complicates the interpretation of their decision processes. In contrast, DVPNet is designed to encode conditional probabilistic contributions for each gene to each class, given a sample, while leveraging the biological information encoded by the Nucleotide Transformer. These probabilistic contributions can be utilized for downstream analyses, including the construction of new genetic networks, enabling biological analyses from perspectives different from those provided by co-expression networks. In this paper, we demonstrate the construction of gene-level contribution landscapes and show that downstream biological analyses can be performed, supporting the validity of this workflow. This workflow can be applied to a wide range of biological classification tasks and has the potential to complement existing workflows in genetic research.



(a)



(b)

Figure 1: Overall experimental workflow: (a) Sample preparation using the Nucleotide Transformer, and (b) model training and extraction of gene-wise probabilistic contributions for each sample.

2 Methods

2.1 VPNet

VPNet [12] combines a tractable probabilistic circuit [13] with patch-embedding vectors from vision transformers.[14][15][16] The study demonstrated that, by leveraging the strong feature-extraction capability of ViTs, the model achieves significant robustness and representational capacity without losing tractability, by structurally preserving smoothness and decomposability.

VPNet is a novel image-classification algorithm that can probabilistically visualize the decision process of the classification model. Unlike most traditional classification models that adopt convolutional neural networks (CNNs) or transformers, the algorithm combines a vision transformer with probabilistic circuits, preserving tractability and explainability. This model can encode distributions over several independent sources of information. For instance, when the model treats images with two modalities, two encoders, and two classes, each probabilistic circuit in the model encodes the following class-conditional densities:

$$
\begin{aligned}
&P\big(z_{\text{mod1,enc1}} \mid \text{class}\,1\big), \quad P\big(z_{\text{mod1,enc1}} \mid \text{class}\,2\big),\\
&P\big(z_{\text{mod1,enc2}} \mid \text{class}\,1\big), \quad P\big(z_{\text{mod1,enc2}} \mid \text{class}\,2\big),\\
&P\big(z_{\text{mod2,enc1}} \mid \text{class}\,1\big), \quad P\big(z_{\text{mod2,enc1}} \mid \text{class}\,2\big),\\
&P\big(z_{\text{mod2,enc2}} \mid \text{class}\,1\big), \quad P\big(z_{\text{mod2,enc2}} \mid \text{class}\,2\big).
\end{aligned}
$$

Here, each class-conditional density can be combined (under the model's factorization) into the joint class-conditional density over all information:

$$
\begin{aligned}
&P\big(z_{\text{mod1,enc1}}, z_{\text{mod1,enc2}}, z_{\text{mod2,enc1}}, z_{\text{mod2,enc2}} \mid \text{class}\,1\big),\\
&P\big(z_{\text{mod1,enc1}}, z_{\text{mod1,enc2}}, z_{\text{mod2,enc1}}, z_{\text{mod2,enc2}} \mid \text{class}\,2\big).
\end{aligned}
$$

This means that VPNet can be optimized at the per-sample level, while at inference it can extract any partial information included in the sample. This enables visualization of the probabilistic contributions for each piece of information in the classification task. Figure 1 is one example of such a visualization. In the figure, the heatmaps clearly highlight the defective regions for 2D image classification between defective and non-defective products, demonstrating the validity of the VPNet algorithm. The more detailed description of the algorithm is described in [12].

2.2 Encoding gene-level vectors using Nucleotide Transformer

Nucleotide Transformer is a foundation model optimized for the masked genetic-sequence prediction task. Like other large models (LMs), including large language models and vision transformers, Nucleotide Transformer is trained on a large amount of nucleotide sequences from 3,202 human genomes and 850 genomes from diverse species in an unsupervised manner. This provides a guarantee of robust representational learning for structures, functions, and relationships. Therefore, this foundation model can be used to extract feature representations for downstream tasks.

From a single-cell–level RNA expression dataset, we extract N RNAs without any name overlap in each single cell. Then, for each RNA, the corresponding transcription start site (TSS) is determined using the reference genome annotation in Gene Transfer Format (GTF), and a nucleotide sequence of fixed length anchored at the TSS is extracted. In this paper, we extracted the nucleotide sequences from -2000 bp upstream to +500 bp downstream of the TSS. The extracted sequence is input to Nucleotide Transformer, and the embedding matrix containing each token vector from the second-to-last layer is obtained. Finally, the embedding matrix is converted to an embedding vector by average pooling of each scaler component of all token vectors. We call this vector a *gene vector*. The gene vector replaces the patch-embedding vector in VPNet; equivalently, we replace the encoder (vision transformer) in VPNet with Nucleotide Transformer.

2.3 Single cell level gene extraction

We assumed that gene regulation can be influenced by genes with low RNA expression levels as well as by genes with high RNA expression levels. Therefore, all expressed genes in a single cell were initially considered in this study, regardless of their expression levels. However, simply using all expressed genes in a cell may cause the model to focus primarily on statistical differences in the frequencies of gene types across classes. Because the model should capture intrinsic functional signals beyond such statistics, while still leveraging the meaningful embeddings produced by the Nucleotide Transformer, we randomly sampled a subset of 900 expressed genes per cell. This uniform random sampling treats genes with low RNA expression levels and genes with high RNA expression levels equally. Figure 1 summarizes the workflow.

2.4 DVPNet: gene-level explainable classification algorithm

By replacing the vision transformer with the Nucleotide Transformer, VPNet is remodeled into a gene-level classification algorithm, which we call *DVPNet*. Each sample comprises 900 gene vectors with no overlap:

$$\mathbf{G}_{\text{cell}} = \big[\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{900}\big], \qquad \text{where each } \mathbf{G}_i \text{ corresponds to a distinct gene.}$$

As we pick up these genes without overlap, this ensures that gene-vector scopes are disjoint, satisfying decomposability:

$$\text{scope}(\mathbf{G}_i) \cap \text{scope}(\mathbf{G}_j) = \quad \text{for } i \neq j.$$

Each embedding vector $\mathbf{G}$ is an input unit to the probabilistic circuit. Each $\mathbf{G}$ comprises 1024 scalar components,

$$\mathbf{G}_i = \big[g_{i,1},\, g_{i,2},\, \ldots,\, g_{i,1024}\big].$$

Additionally, during optimization, the parameters of the Nucleotide Transformer are frozen, so the scope of each scalar component is invariant. In particular,

$$\text{scope}(g_{i,k}) \cap \text{scope}(g_{j,\ell}) = \quad \text{for } (i,k) \neq (j,\ell).$$

This construction ensures decomposability for product nodes, and smoothness is enforced by defining sum-node children to have identical scopes.

Then, each scalar component of a vector is input to a leaf node of the probabilistic circuit, as in VPNet. The circuit prepares one leaf node for each scalar component. This input is encoded as a univariate probability density, such as a Gaussian,

$$\mathcal{N}\big(g_{i,k};\, \mu_{i,k}, \sigma_{i,k}^2\big) = \frac{1}{\sqrt{2\pi}\,\sigma_{i,k}} \exp\left(-\frac{(g_{i,k} - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right).$$

As in VPNet, smoothness and decomposability are strictly preserved in the structure of our probabilistic circuit. Consequently, the circuit encodes the joint probability distribution over all scalar components; because the Nucleotide Transformer is frozen during optimization, each gene vector has an immutable scope inherited from its corresponding gene. Thus,

$$P(\mathbf{G}_i) = P\big(g_{i,1}, g_{i,2}, \ldots, g_{i,1024}\big).$$

Then, we factorize the joint probability over all gene vectors included in the sample. Because gene-vector scopes are disjoint, this preserves decomposability, which means that the result of the factorization equals the joint probability of all genes in the sample. Additionally, the joint probability distribution over all genes equals the joint probability over the sample scope:

$$P(\text{sample}) \;=\; \prod_{i=1}^{900} P(\mathbf{G}_i) \;=\; P(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{900}).$$

In VPNet, the number of prepared probabilistic circuits is defined as

$$\#\text{circuits} \;=\; (\#\text{modalities}) \times (\#\text{encoders}) \times (\#\text{classes}).$$

Here, we use one modality (genes), one encoder (Nucleotide Transformer), and two classes in all experiments. Thus, the number of probabilistic circuits per experiment is 2. We treat the output of each probabilistic circuit as a probabilistic model encoding the *class-conditional* probability of the sample (class 1):

$$P(\text{sample} \mid \text{class 1}) \;=\; \prod_{i=1}^{900} P_{m,1}\big(\mathbf{G}_i\big) \;=\; \prod_{i=1}^{900} P_{m,1}\big(g_{i,1}, g_{i,2}, \ldots, g_{i,1024}\big),$$

where $P_{m,1}$ denotes the circuit (for modality m and class 1) that factorizes over the scalar components of each gene vector.

Similarly, the output of the other circuit is treated as the class-conditional probability for class 2:

$$P(\text{sample} \mid \text{class 2}) = \prod_{i=1}^{900} P_{m,2}\big(\mathbf{G}_i\big) = \prod_{i=1}^{900} P_{m,2}\big(g_{i,1}, g_{i,2}, \ldots, g_{i,1024}\big).$$

As in VPNet, we parameterize the class prior with a small neural network as a Bernoulli distribution:

$$P(y = \text{class 1}) = \pi, \qquad P(y = \text{class 2}) = 1 - \pi, \qquad 0 < \pi < 1,$$

where π is a learnable parameter.

Then, by Bayes' rule,

$$P(y = c \mid \text{sample}) \;=\; \frac{P(\text{sample} \mid y = c)\, P(y = c)}{\sum_{c' \in \{1,2\}} P(\text{sample} \mid y = c')\, P(y = c')}.$$

However, to stabilize the training process, we adopted the geometric mean of the likelihoods across the 900 feature vectors instead of the raw product likelihood. We define

$$\tilde{\mathcal{L}}_c = \left(\prod_{i=1}^{900} P_{m,c}\big(g_{i,1}, g_{i,2}, \ldots, g_{i,1024}\big) \right)^{\frac{1}{900}}.$$

Then, a power posterior is computed by applying Bayes' rule to the geometric mean likelihood:

$$P(y = c \mid \text{sample}) = \frac{\tilde{\mathcal{L}}_c\, P(y = c)}{\sum_{c' \in \{1,2\}} \tilde{\mathcal{L}}_{c'}\, P(y = c')}.$$

Although this departs from the exact Bayesian posterior under the full product likelihood, it preserves a valid probabilistic interpretation as a normalized power posterior [17].

As in VPNet, optimization is performed per sample. Let

$$p_c \;=\; P(y = c \mid \text{sample}), \qquad c \in \{1, 2\},$$

and let $\mathbf{y} = (y_1, y_2)$ be the one-hot target. The loss combines cross-entropy with a Shannon-entropy regularizer:

$$\mathcal{L} \;=\; \underbrace{-\sum_c y_c \log p_c}_{\text{cross-entropy}} + \lambda \underbrace{\sum_c p_c \log p_c}_{\text{Shannon-entropy loss (negative entropy)}}, \qquad \lambda \geq 0.$$

In the training dataset, some samples may share the same gene (i.e., the same scope), but no product node connects any samples; thus, decomposability of the model is preserved during optimization. Additionally, two gene vectors from the same gene can differ because standardization is computed over the vectors in each *sample*:

$$\tilde{\mathbf{G}}^{(A)} = \frac{\mathbf{G} - \boldsymbol{\mu}^{(A)}}{\boldsymbol{\sigma}^{(A)}}, \qquad \tilde{\mathbf{G}}^{(B)} = \frac{\mathbf{G} - \boldsymbol{\mu}^{(B)}}{\boldsymbol{\sigma}^{(B)}},$$

where $(\boldsymbol{\mu}^{(A)}, \boldsymbol{\sigma}^{(A)})$ and $(\boldsymbol{\mu}^{(B)}, \boldsymbol{\sigma}^{(B)})$ are samplewise statistics. In general, $\tilde{\mathbf{G}}^{(A)} \neq \tilde{\mathbf{G}}^{(B)}$. Due to samplewise standardization, the same raw vector can yield different normalized inputs across samples, acting as implicit data augmentation and inducing sample-dependent posteriors. This componentwise transformation preserves decomposability.

2.5 Gene-level probabilistic contributions

As mentioned above, the joint conditional distribution of a sample equals the joint conditional distribution of all gene vectors included in the sample:

$$P(\text{sample} \mid \text{class}) \;=\; P(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{900} \mid \text{class}).$$

This is computed by the product node connecting each conditional distribution:

$$P(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{900} \mid \text{class}) \;=\; \prod_{i=1}^{900} P(\mathbf{G}_i \mid \text{class}).$$

Therefore, at inference, we can extract any gene-level conditional distribution,

$$P(\mathbf{G}_i \mid \text{class}),$$

which represents the probabilistic contribution of that gene in the sample when all other genes are unobserved. For instance, in this paper we set cancer-cell and normal-cell classes; after optimization, the class-conditional distributions from a gene can be extracted as

$$P(\mathbf{G}_i \mid \text{cancer cell}), \qquad P(\mathbf{G}_i \mid \text{normal cell}).$$

Here, $P(\mathbf{G}_i \mid \text{cancer cell})$ represents the probabilistic contribution corresponding to the model's decision process indicating the extent to which the gene supports classifying the sample as cancer and $P(\mathbf{G}_i \mid \text{normal cell})$ is interpreted analogously. If the model is successfully optimized without overfitting or underfitting on the training domain, it is expected that these class-conditional distributions reflect functional contributions to the specific class beyond the statistical frequency of the gene on the training domain.

2.6 Comparison of probabilistic contributions between genes

In our experiment, the probabilistic circuit is implemented in log-space. The log joint class-conditional density is

$$\log P(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{900} \mid \text{class}) = \sum_{i=1}^{900} \log P(\mathbf{G}_i \mid \text{class}).$$

Here, we define the probabilistic contribution for a specific class as

$$S(\text{gene} \mid \text{class}) \;=\; \frac{1}{k} \sum \log P(\mathbf{G}_i \mid \text{class}),$$

where k is the number of appearances of the gene across the extracted samples. In this paper, we define the cancer-cell and normal-cell contribution scores as

$$S(\text{gene}) \;=\; S(\text{gene} \mid \text{cancer cell}) \;-\; S(\text{gene} \mid \text{normal cell}).$$

A high $S(\text{gene})$ indicates that the gene contributes more to the cancer class, whereas a low $S(\text{gene})$ indicates a stronger contribution to the normal class. By comparing this score across genes, we compare their probabilistic contributions.

2.7 Sample-level gene probabilistic contributions

Because of the sample-level standardization of the input vectors, the probabilistic circuit is trained to encode cancer-cell and normal-cell contribution scores for each gene under the genetic context of the sample. We therefore define the sample-level scores as

$$S(\text{gene} \mid \text{sample}) \;=\; S(\text{gene} \mid \text{cancer cell}, \text{sample}) \;-\; S(\text{gene} \mid \text{normal cell}, \text{sample}),$$

with

$$S(\text{gene} \mid \text{class}, \text{sample}) \;=\; \log P(\mathbf{G}_i \mid \text{class}, \text{sample}).$$

This sample-level gene score captures the sample dependency of gene function: for instance, genes with high $S(\text{gene} \mid \text{sample})$ in some samples but low $S(\text{gene} \mid \text{sample})$ in others suggest that the model considers those genes important for cancer classification only in specific gene contexts.

2.8 Datasets

We used the GSE131907 single-cell lung cancer atlas [18] including RNA-seq data from 208,506 cells spanning normal tissues and cancer from early to metastatic stages in 44 patients. Single-cell RNA sequencing was performed on whole cells derived from primary sites (tLung and tL/B), pleural effusions (PE), lymph nodes (mLN), and brain metastases (mBrain), as well as normal tissues from lung (nLung) and lymph nodes (nLN).

Two class labels—cancer cell and normal cell—are assigned to cells from tLung and nLung epithelium cells, respectively. The training, validation, and test sets were created by randomly splitting each subset into 3,420/570/570 cells per label at a fixed random seed. As described in Section 2.3, we randomly extracted 900 expressed genes from each cell at a fixed random seed; for each RNA, the corresponding transcription start site (TSS) was determined using the reference genome annotation in Gene Transfer Format (GTF). The selection was performed regardless of absolute RNA expression level.

In addition, to assess the robustness of the model across patients, we prepared additional training, validation, and test sets comprising data from different patients without any patient overlap. The numbers of cells in the training, validation, and test sets are 2,700, 500, and 1,000, respectively, including data from 7, 1, and 2 independent patients.

We adopted Nucleotide Transformer v2 from InstaDeep and NVIDIA [8], which has 500 million parameters, and converted nucleotide sequences from -2000 bp upstream to $+500$ bp downstream of the TSS into 1,024-dimensional vectors. As a result, each cell has 900 gene vectors, and each gene vector represents a specific RNA in the sample.

2.9 Experimental Setting

We set the hyperparameters of our circuits to 2, 16, 2, and 2 for depth, latents, repetitions, and pieces, respectively, for the patient-mixed model, and 3, 24, 2, and 2 for the patient-independent model [12]. We trained the model with a batch size of 20 using the Adam optimizer[19].

2.10 Statistical Filtering

A key concern is overfitting to the statistical frequencies of gene types. To address this, we filtered genes based on the following criteria:

Contradictory count–score pairs:

$$S(\text{gene}) > 0 \;\text{ and }\; \text{counts}_{\text{cancer}}(\text{gene}) < \text{counts}_{\text{normal}}(\text{gene}),$$

or

$$S(\text{gene}) < 0 \;\text{ and }\; \text{counts}_{\text{cancer}}(\text{gene}) > \text{counts}_{\text{normal}}(\text{gene}).$$

Such filtering reduces the influence of raw frequency in the dataset on the learned contributions.

3 Results and Discussions

3.1 Assessment of feature representations during optimization

As mentioned in the method section, DVPNet is designed so that its decision process or the class-conditional probabilistic distribution for each gene vector can be visualized. However, this does not guarantee that the model focuses on and captures meaningful feature representations for biological analysis. First, similar to most deep-learning-based classification models, the model needs to avoid the following situations:

- **Underfitting.** Here, the training task is too complex for the model to capture meaningful features for the classification task due to insufficient model capacity, inappropriate representations, or noisy supervision. In this situation, the model cannot achieve sufficient classification performance on the training dataset.
- **Overfitting.** Here, the model does not show robustness. The model can show sufficient performance only on the training dataset but not on the test dataset. This suggests that the model captures overly complex feature representations meaningful only for the training dataset. For meaningful biological analysis, at least the model needs to grasp generalized features applicable to the target domain (e.g., a test dataset split from the entire dataset).

The overall classification performance of the model on the training and test datasets is as follows. The patient-mixed model achieved AUROC, AUPRC, maximum accuracy, and F1 scores of 0.992, 0.994, 0.974, and 0.974 on the training set, and 0.975, 0.981, 0.940, and 0.939 on the test set, respectively. For the patient-independent model, it achieved AUROC, AUPRC, maximum accuracy, and F1 scores of 0.999, 0.999, 0.991, and 0.991 on the training set, and 0.976, 0.979, 0.920, and 0.918 on the test set.

The high performance on the training set indicates that the model does not exhibit underfitting during optimization. In addition, the relatively small performance gap between the training and test sets suggests that the model does not suffer from severe overfitting and generalizes well to unseen combinations of genes.

While the model avoids underfitting and overfitting, this is not sufficient to guarantee the advantage of our model for biological analysis compared to other statistical methods. The classification task should be solved using the feature representations of the embedding vectors from the Nucleotide Transformer as well as label-wise gene occurrence frequencies.

In Figure 2, the statistical frequency counts between two labels on the training set are compared with the mean $S(\text{gene})$ score provided by the model. The moderate Pearson ($r = 0.356$) and Spearman ($\rho = 0.408$) correlations between gene occurrence frequency differences and probabilistic contribution scores suggest that gene frequency alone does not fully explain the learned contributions. This observation is consistent with the model leveraging additional information encoded by the Nucleotide Transformer embeddings during classification.

In summary, the experimental results suggest that the model was successfully optimized without suffering from underfitting or overfitting, and the probabilistic contributions of genes were determined based on a combination of the statistical frequency and the feature representations from the Nucleotide Transformer. Assuming that the Nucleotide Transformer encodes inherent biological functions, these probabilistic contributions have the potential to enable meaningful biological analysis and provide new biological insights.

3.2 A WGCNA-based network analysis

Weighted gene co-expression network analysis (WGCNA) [20] [21]is a widely used method to identify clusters of genes whose RNA expression levels are highly correlated. This analysis contributes to understanding and exploring the

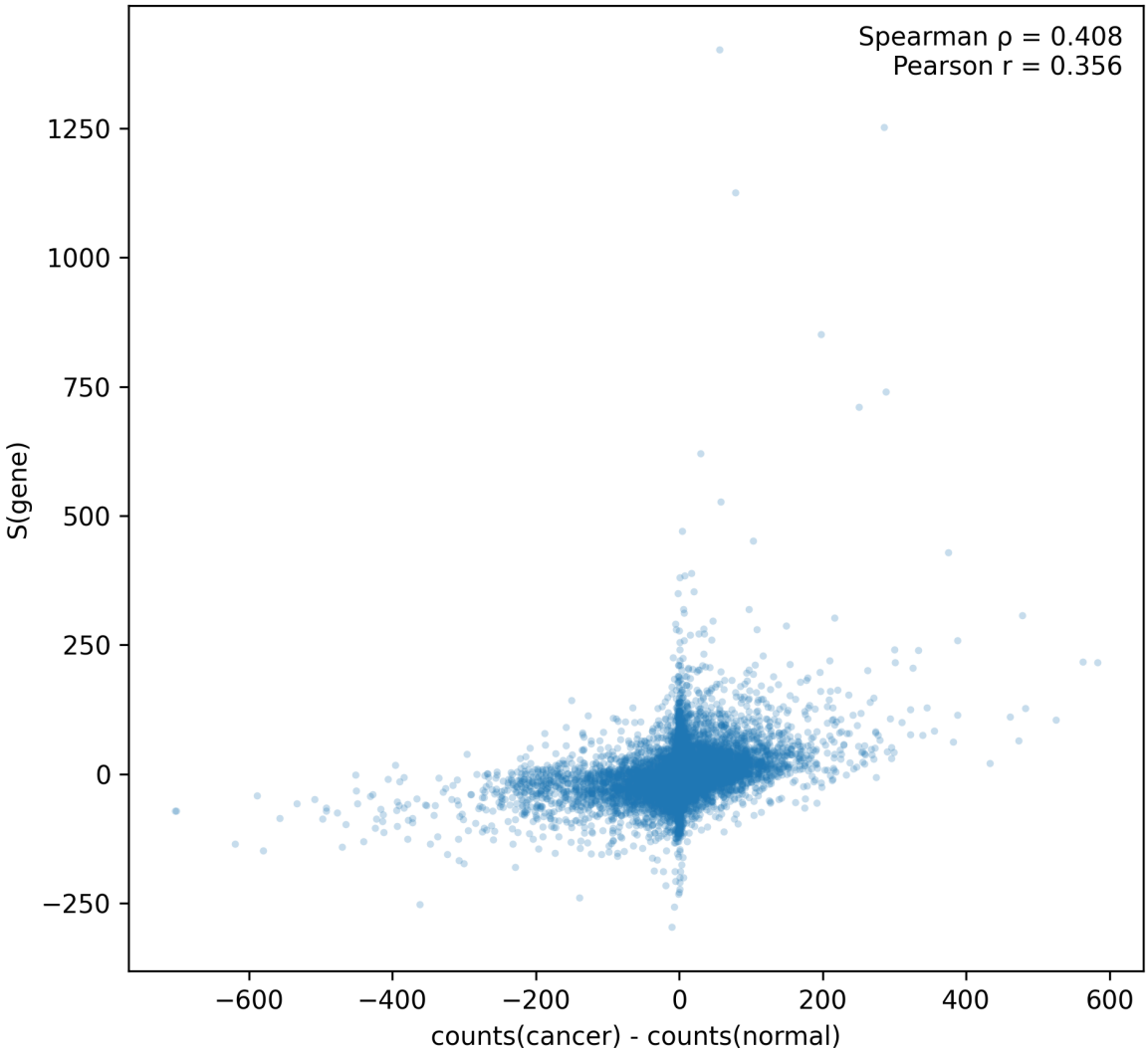


Figure 2: Relationship between label-wise gene occurrence count differences and the mean probabilistic contribution score $S(\text{gene})$. Pearson's r and Spearman's ρ are reported in the plot.

system-level functionality of genes [21]. Normally, WGCNA is constructed based on correlations in RNA expression variation across samples; however, instead of using RNA expression levels, we utilized the sample-dependent probabilistic contribution, $S(\text{gene} \mid \text{sample})$, to construct a new genetic network. Unlike traditional WGCNA, this network depends on the model's decisions regarding whether a given gene positively or negatively contributes to $S(\text{gene})$ for a specific sample.

Using all training samples and their corresponding $S(\text{gene} \mid \text{sample})$ values for 9,540 genes, an unsigned weighted network was constructed with a soft-thresholding power automatically selected according to the scale-free topology criterion. Gene modules were identified from the topological overlap matrix using hierarchical clustering followed by dynamic tree cutting (minimum module size = 2). This procedure identified 50 distinct modules comprising 1,812 genes, while the remaining genes were excluded as noise. Within each module, genes were ranked by intramodular connectivity, defined as the sum of connection strengths to other genes in the same module, and genes with the highest intramodular connectivity were considered hub genes.

To characterize the biological functions represented by each identified module, Gene Ontology (GO) enrichment analysis [22] [23] was performed separately for each module using the genes assigned to that module. Enriched GO terms were used to interpret the functional roles of individual modules and their corresponding hub genes.

Table 1 summarizes the module-level scores and representative hub genes, which are determined based on intramodular connectivity. Modules comprise between 11 and 301 genes and are ordered by $S(\text{genes})$ computed across all samples, representing the overall probabilistic contribution to the cancer cell class scored by the classification model. The module with the highest $S(\text{genes})$ value (45.71) is the orange module, whereas the royal blue module shows the lowest $S(\text{genes})$ value (-43.90). To illustrate condition-specific probabilistic contributions based on class

labels, S(genes) values were computed separately for samples with cancer cells as ground truth and for samples with normal cells as ground truth. The 0074ff module shows the largest difference between S(genes | cancer samples) and S(genes | normal samples) (38.62), suggesting that the model assigns a stronger contribution to this module for cancer classification in cancer samples, and conversely for normal samples.

Table 2 presents the five representative Gene Ontology enrichment terms for each module. These terms summarize the dominant biological functions associated with each module.

This explainable AI based network may provide biological insights distinct from those obtained by traditional WGCNA, which is constructed solely from correlation relationships of RNA expression levels. Because the proposed network is not derived directly from RNA expression correlations, the resulting S(genes) scores reflect probabilistic contributions inferred from the model's decision process. Consequently, the genetic modules and their associated biological pathways, as characterized by Gene Ontology analysis, may highlight functionally relevant gene sets that contribute to distinguishing cancer cells from normal cells.

Table 1: Module-level scores and hub genes. Modules are ordered by S(genes | all samples) (descending).

Module	n	S(genes | all samples)	S(genes | cancer samples)	S(genes | normal samples)	Hub1	Hub2	Hub3
orange	11	45.71	53.37	36.59	KRT83	TDRD1	IGKV1-12
darkgreen	11	41.51	41.24	41.87	IGHV4-59	AANAT	PATE2
salmon	14	39.31	38.56	40.05	OST4	FTH1	POLR2L
saddlebrown	10	27.10	19.54	38.05	ABHD1	SLC6A12	LINC01348
yellow	38	24.84	25.35	24.33	RPL19	GAPDH	RPL18
0074ff	48	18.54	36.89	-1.73	TNFRSF17	BOLA2	TRHDE
steelblue	9	15.57	-8.60	44.73	C1QTNF9	AQP10	LINC00566
84ff00	50	15.49	11.96	20.88	KC6	LINC01559	KLF1
00fff6	48	14.87	19.98	9.94	GJA5	LHB	MYBPC2
green	33	12.22	11.77	12.67	NEAT1	MUC1	PFN1
ff4700	53	10.35	10.49	10.21	PALD1	FAM90A1	TDRD5
00ffa9	48	6.13	9.13	2.45	LIPC	FAM177B	GRIK5
brown	62	5.59	5.44	5.74	S100A6	COX4I1	S100A10
cyan	14	2.63	1.92	3.35	NFKBIA	RPS15	TPI1
00c1ff	48	2.37	10.07	-5.37	LHX4	CEBPE	KCNH6
greenyellow	15	-0.59	-0.06	-1.12	RPLP1	ARPC3	TMEM59
ffe000	51	-1.26	-1.83	-0.69	C1orf220	CPEB1-AS1	ZNF781
ff005f	46	-2.02	14.64	-18.61	TNFRSF13B	CHIT1	WARS2-IT1
2500ff	46	-2.21	2.16	-7.39	SLC1A6	FCRL3	TCL1A
blue	76	-3.83	-3.16	-4.50	RPL27	RPL6	RPLP2
b900ff	46	-4.69	-4.43	-4.93	SPATA45	C10orf55	ASB16
skyblue	10	-5.73	-5.37	-5.98	HSPB6	ZNF256	KCNK3
violet	9	-5.78	-0.44	-11.96	SCN7A	ANGPTL1	ADAMTS8
ff9400	53	-6.53	-6.41	-6.65	B4GALNT2	RAPSN	CD5L
darkred	12	-6.63	1.15	-12.72	RTCA-AS1	SERPINA6	ZNF536
d1ff00	50	-7.24	-1.92	-12.93	SNX32	BCDIN3D-AS1	FOXI3
paleturquoise	9	-7.73	-0.95	-12.77	ITGA11	P2RX1	FBXW10
ff00ac	46	-7.73	-7.96	-7.51	NUCB1-AS1	MEIOB	RPLP0
red	26	-8.37	-8.38	-8.36	MYL6	RPL13	RPL26
lightgreen	9	-11.46	-4.06	-19.71	STX1B	GALNT16	LRP4-AS1
grey60	9	-11.79	-2.22	-38.43	PRG4	SLAMF9	ZNF285
turquoise	301	-12.58	-10.58	-15.11	ACRV1	PDE6H	PCAT18
00ff10	48	-12.95	-9.07	-17.68	LRRC55	CGB7	PRKG1-AS1
purple	17	-13.49	-13.81	-13.17	RPL31	RPS25	ATP1A1
tan	14	-14.12	-6.51	-21.43	FCRLA	YBX2	RBP5
midnightblue	13	-14.26	-14.11	-14.41	RPL27A	CD59	ANAPC16
lightcyan	13	-14.89	-15.53	-14.24	TMBIM6	FOS	CIRBP
3dff00	48	-17.59	-14.98	-19.87	TMEM266	HOXC4	HSD11B1
magenta	17	-17.73	-17.17	-18.28	RPS3	RPL11	GUK1
black	19	-19.05	-18.76	-19.35	FAU	APLP2	ANXA2
ff0000	54	-19.10	-18.92	-19.27	DPF1	ANP32D	RGS4
darkturquoise	11	-20.75	-20.55	-20.99	CHST4	NR2E3	NAV2-AS2
fc00f5	46	-21.05	-19.19	-22.78	IGKV4-1	HES5	GLIS2-AS1
lightyellow	9	-22.51	-29.15	-15.79	ZNF667	CDH2	ARHGAP22
00ff5c	48	-24.32	-18.27	-29.84	NPTX1	SCN2B	LRP1B
0028ff	48	-27.28	-27.71	-26.88	LINC01415	SZT2-AS1	ASCL1
darkorange	9	-28.89	-24.60	-32.47	DNMBP-AS1	CHAD	MYBPC3
darkgrey	11	-31.74	-43.79	5.17	RHOJ	SLC3A1	TPH1
7200ff	46	-35.67	-35.29	-36.03	CCL8	ZIC2	MADCAM1
pink	18	-35.83	-35.24	-36.41	RPS13	PRDX1	COX6A1
royalblue	12	-43.90	-49.37	-40.42	LINC00327	PAMR1	GSTM5

Table 2: Representative GO enrichment terms (top5) for each module (ordered by $S(\text{genes} \mid \text{all samples})$ descending).

Module	Representative GO terms (top5)
orange	proton-transporting ATPase activity, rotational mechanism; regulation of macroautophagy; microvillus; zinc ion import into zymogen granule; protein C inhibitor-TMPRSS7 complex
darkgreen	aralkylamine N-acetyltransferase activity; indolalkylamine biosynthetic process; melatonin biosynthetic process; arylamine N-acetyltransferase activity; photoperiodism
salmon	extracellular exosome; protein binding; perinuclear region of cytoplasm; cytoplasm; cytosol
saddlebrown	amino acid transmembrane transport; amino acid transmembrane transporter activity; amino acid transport; protein binding; membrane
yellow	extracellular exosome; protein binding; cytoplasm; membrane; cytosol
0074ff	protein binding; metal ion binding; plasma membrane; signal transduction; cell adhesion
steelblue	regulation of acute inflammatory response to antigenic stimulus; CD4-positive, CD25-positive, alpha-beta regulatory T cell differentiation; negative regulation of bone trabecula formation; negative regulation of bone remodeling; sequestering of BMP from receptor via BMP binding
84ff00	cerebellar climbing fiber to Purkinje cell synapse; sperm-egg recognition; membrane; locomotory behavior; synapse
00fff6	SA node cell action potential; extracellular region; meiotic spindle organization; positive regulation of positive chemotaxis; t-circle formation
green	cytosol; protein binding; cytoplasm; regulation of muscle contraction; nucleus
ff4700	protein binding; attachment of mitotic spindle microtubules to kinetochore; metaphase chromosome alignment; cell differentiation; outer kinetochore
00ffa9	postsynaptic membrane; protein binding; regulation of postsynaptic membrane potential; cell surface receptor signaling pathway via JAK-STAT; plasma membrane
brown	protein binding; cytosol; extracellular exosome; RNA binding; mitochondrial inner membrane
cyan	cytosol; protein binding; glyceraldehyde-3-phosphate metabolic process; nucleus; extracellular exosome
00c1ff	coumarin metabolic process; regulation of transcription by RNA polymerase II; monoatomic ion transmembrane transport; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; arachidonate epoxygenase activity
greenyellow	protein binding; RNA binding; focal adhesion; cytosol; extracellular exosome
ffe000	protein binding; integrin alpha5-beta1 complex; negative regulation of myoblast fusion; plasma membrane; cytosolic large ribosomal subunit
ff005f	plasma membrane; protein binding; extracellular space; presynapse; azurophil granule lumen
2500ff	extracellular matrix binding; embryonic skeletal system morphogenesis; protein binding; sequence-specific double-stranded DNA binding; membrane
blue	protein binding; ribonucleoprotein complex; extracellular exosome; RNA binding; nucleus
b900ff	protein binding; plasma membrane; extracellular region; positive regulation of interleukin-13 production; viral translational frameshifting
skyblue	open rectifier potassium channel activity; regulation of myoblast proliferation; detection of hypoxic conditions in blood by carotid body chemoreceptor signaling; regulation of epithelial to mesenchymal transition involved in endocardial cushion formation; uterine wall breakdown
violet	extracellular matrix; low-affinity phosphate transmembrane transporter activity; negative regulation of cardiac chamber formation; negative regulation of heart looping; detection of hypoxic conditions in blood by carotid body chemoreceptor signaling
ff9400	protein binding; membrane; plasma membrane; organelle localization by membrane tethering; extracellular region
darkred	collagen fibril organization; positive regulation of hepatic stellate cell proliferation; protein tyrosine kinase collagen receptor activity; vesicle coat; negative regulation of cytoplasmic mRNA processing body assembly
d1ff00	membrane; protein binding; metal ion binding; extracellular space; cellular response to fatty acid
paleturquoise	protein kinase A binding; calcium-mediated signaling; suramin binding; integrin alpha11-beta1 complex; serotonin secretion by platelet
ff00ac	protein binding; insulin catabolic process; regulation of platelet aggregation; icosanoid biosynthetic process; cytoplasm
red	cytosol; RNA binding; ribonucleoprotein complex; protein binding; cytoplasmic translation
lightgreen	negative regulation of chronic inflammatory response to antigenic stimulus; positive regulation of spontaneous neurotransmitter secretion; negative regulation of synaptic vesicle recycling; regulation of response to wounding; negative regulation of macropinocytosis
grey60	Ndc80 complex; outer kinetochore; attachment of spindle microtubules to kinetochore; mitotic spindle assembly checkpoint signaling; mitotic spindle organization
turquoise	plasma membrane; protein binding; immunoglobulin complex; membrane; extracellular region
00ff10	DNA-binding transcription factor activity, RNA polymerase II-specific; RNA polymerase II cis-regulatory region sequence-specific DNA binding; plasma membrane; negative regulation of transcription by RNA polymerase II; metal ion binding
purple	extracellular exosome; protein binding; postsynaptic density; cytoplasm; melanosome
tan	protein binding; spindle; mitotic metaphase chromosome alignment; cell division; metaphase/anaphase transition of meiosis I
midnightblue	cell division; cytoplasm; protein binding; vesicle; MHC class II protein complex binding
lightcyan	protein binding; transcription factor AP-1 complex; nucleus; enzyme binding; response to muscle stretch
3dff00	plasma membrane; action potential; membrane; protein homodimerization activity; protein binding
magenta	cytoplasmic translation; ribonucleoprotein complex; cytosol; cytosolic ribosome; positive regulation of signal transduction by p53 class mediator
black	protein binding; cytosol; eukaryotic 48S preinitiation complex; eukaryotic 43S preinitiation complex; nucleus
ff0000	protein binding; intermediate filament bundle assembly; negative regulation of signal transduction; sialic acid binding; nucleus
darkturquoise	DNA-binding transcription activator activity, RNA polymerase II-specific; circadian sleep/wake cycle, REM sleep; reproductive behavior; noradrenergic neuron differentiation involved in brainstem development; positive regulation of convergent extension involved in gastrulation
fc00f5	protein binding; eosinophil degranulation; plasma membrane; glycoprotein metabolic process; ganglioside biosynthetic process
lightyellow	cell morphogenesis; negative regulation of canonical Wnt signaling pathway; negative regulation of retinal ganglion cell axon guidance; positive regulation of synaptic vesicle clustering; mesenchymal cell migration
00ff5c	plasma membrane; protein binding; extracellular region; signal transduction; sprouting angiogenesis
0028ff	plasma membrane; cell surface receptor signaling pathway; protein binding; locomotory behavior; immune response
darkorange	cell surface receptor signaling pathway; sarcomere; negative regulation of cytotoxic T cell differentiation; C zone; myofilament
darkgrey	serotonin biosynthetic process from tryptophan; basic amino acid transport; regulation of hemostasis; vascular endothelial growth factor receptor-3 signaling pathway; tryptophan 5-monooxygenase activity
7200ff	histone H3R26 arginine deiminase activity; histone arginine deiminase activity; metal ion binding; protein-arginine deiminase activity; protein binding
pink	protein binding; extracellular exosome; RNA binding; response to progesterone; azurophil granule membrane

Continued on next page

Module	Representative GO terms (top5)
royalblue	L-methionine catabolic process; 3'-tyrosyl-DNA phosphodiesterase activity; regulation of clathrin coat assembly; synaptic vesicle uncoating; methionine adenosyltransferase complex

3.3 Analysis of pathway level contribution

In this section, we applied GO enrichment analysis to 9540 genes on the training set without any clustering analysis. Then, the derived biological terms were evaluated based on mean S(genes) score averaged from S(genes) of all the genes comprising the pathway. This analysis can provide information about which biological pathways are more contributed to the distinguish cancer cells from normal cells for the classification model.

GO enrichment analysis identified a total of 4,000 biological terms with q-values less than 0.05. GO terms were ranked according to their mean S score (S_{mean}) calculated from the overlap genes.

The top 100 GO terms with the highest S_{mean} values are listed in Table 3, and the corresponding overlap genes for each term are shown in Table 4. These top-ranked GO terms include multiple entries related to immunoglobulin complexes, complement activation, antibody-dependent cellular cytotoxicity, and humoral immune response pathways.

The bottom 100 GO terms with the lowest S_{mean} values are presented in Table 5, and their corresponding overlap genes are summarized in Table 6. These GO terms span a wide range of functional categories, including immune regulation, complement-related processes, metabolic pathways, ion transport, and cellular structural components. All of these terms exhibit negative S_{mean} values relative to other enriched GO terms.

3.4 Analysis beyond statistics

In this section, we analyzed 1,524 genes from the training set that exhibited contradictory count–score pairs (Section 2.10), namely genes with $S(\text{gene}) > 0$ while $\text{counts}_{\text{cancer}} < \text{counts}_{\text{normal}}$, or $S(\text{gene}) < 0$ while $\text{counts}_{\text{cancer}} > \text{counts}_{\text{normal}}$. Table 7 and Table 8 display the top 100 and the bottom 100 genes based on $S(\text{gene})$ respectively.

The classification model is provided with only two sources of information: gene occurrence statistics in the training set and feature representations encoded by the Nucleotide Transformer. If the model relied solely on statistical gene counts, genes with higher occurrence in a given class would naturally receive higher contribution scores for that class. However, the existence of these contradictory genes demonstrates that this is not the case.

Because the statistical signal alone cannot explain the observed score assignments for these genes, the results indicate that the feature representations encoded by the Nucleotide Transformer dominate and override the frequency-based signal for the corresponding class decisions. Focusing on such genes enables further biological analysis that is not directly driven by gene occurrence statistics.

It is natural to consider that expressed genes appearing more frequently in cancer cell samples functionally contribute more to cancer cells. However, this result suggests that the explainable AI model evaluates the differentiation between cancer cells and normal cells as being regulated by more complex relationships beyond simple statistics; that is, the importance of genes for cancer cells cannot be determined solely by their occurrence statistics.

This result supports that our model can comprehensively evaluate probabilistic contributions by integrating both statistical information and feature representations encoding biological functions. As discussed above, this enables analyses beyond simple gene occurrence statistics and provides a complementary perspective to traditional statistics-based approaches, such as WGCNA and differential gene analysis [24].

3.5 Biological validation of the genetic profiling

Wet-lab experiments to test whether the genes prioritized by our probabilistic contribution scores exhibit functional relevance in cancer biology. If such experiments validate that top-ranked genes (e.g., TSPAN32, TPSD1, ANKRD2) are functionally involved in tumor progression or immune modulation, this would strongly support the biological relevance of our framework. However, such experimental validation was beyond the scope of the present study. Instead, we examined existing literature to assess whether previously reported clinical or biological findings are consistent with the genes prioritized by our framework.

Among the top-ranked genes in Table 7, multiple genes have been independently investigated as therapeutic targets in cancer research, including ITGA5[25][26], SIGLEC9[27][28], and NOTUM[29][30]. In addition, several well-characterized genes involved in tumor suppression and stress-response pathways, such as TP73[31], SFRP5[32], and NLRC3[33], are also present among the top-ranked genes. Furthermore, key regulators of immune and microenvironmental signaling including CD300A/C[34][35], PITPNM3[36], and SEMA6B[37] are included in the list.

The consistency between previously reported findings and our probabilistic gene ranking suggests that the proposed framework successfully prioritizes genes with established relevance in cancer biology. Importantly, the presence of less-

GO id	termname	namespace	N_{bg}	K_{term}	k	S_{mean}
GO:0042571	immunoglobulin complex, circulating	cellular_component	42597	8	8	761.805
GO:0071751	secretory IgA immunoglobulin complex	cellular_component	42597	5	4	653.949
GO:0060267	positive regulation of respiratory burst	biological_process	42597	7	5	520.786
GO:0071755	IgM B cell receptor complex	cellular_component	42597	3	3	399.693
GO:0097278	complement-dependent cytotoxicity	biological_process	42597	3	3	254.109
GO:0006958	complement activation, classical pathway	biological_process	42597	39	25	206.421
GO:0022007	convergent extension involved in neural plate elongation	biological_process	42597	3	3	166.901
GO:1903131	mononuclear cell differentiation	biological_process	42597	3	3	150.981
GO:0001788	antibody-dependent cellular cytotoxicity	biological_process	42597	9	7	142.512
GO:0045004	DNA replication proofreading	biological_process	42597	3	3	129.122
GO:1901475	pyruvate transmembrane transport	biological_process	42597	4	4	108.484
GO:0150012	positive regulation of neuron projection arborization	biological_process	42597	8	5	102.228
GO:0046054	dGMP metabolic process	biological_process	42597	3	3	101.981
GO:0030172	troponin C binding	molecular_function	42597	3	3	101.147
GO:0021681	cerebellar granular layer development	biological_process	42597	3	3	98.7731
GO:0005861	troponin complex	cellular_component	42597	9	6	94.6442
GO:0070442	integrin alphaIIb-beta3 complex	cellular_component	42597	3	3	89.8362
GO:0060669	embryonic placenta morphogenesis	biological_process	42597	5	4	89.7589
GO:0001560	regulation of cell growth by extracellular stimulus	biological_process	42597	3	3	86.0928
GO:0008594	photoreceptor cell morphogenesis	biological_process	42597	6	5	82.4355
GO:0002455	humoral immune response mediated by circulating immunoglobulin	biological_process	42597	5	4	81.157
GO:0042631	cellular response to water deprivation	biological_process	42597	3	3	77.7187
GO:0044351	macropinocytosis	biological_process	42597	8	5	75.5463
GO:0008253	5'-nucleotidase activity	molecular_function	42597	13	7	69.5592
GO:0046850	regulation of bone remodeling	biological_process	42597	7	5	69.4121
GO:0006204	IMP catabolic process	biological_process	42597	5	5	69.395
GO:0072562	blood microparticle	cellular_component	42597	143	74	69.3481
GO:0034164	negative regulation of toll-like receptor 9 signaling pathway	biological_process	42597	5	4	68.4165
GO:0099173	postsynapse organization	biological_process	42597	12	9	66.6886
GO:0044388	small protein activating enzyme binding	molecular_function	42597	5	4	65.5027
GO:0043045	epigenetic programming of gene expression	biological_process	42597	3	3	64.1833
GO:0035372	protein localization to microtubule	biological_process	42597	10	6	63.2587
GO:1905368	peptidase complex	cellular_component	42597	13	8	62.9674
GO:0045098	type III intermediate filament	cellular_component	42597	3	3	62.2001
GO:0042867	pyruvate catabolic process	biological_process	42597	8	6	62.0691
GO:0000731	DNA synthesis involved in DNA repair	biological_process	42597	15	8	61.091
GO:0046942	carboxylic acid transport	biological_process	42597	5	4	60.6035
GO:0008381	mechanosensitive monoatomic ion channel activity	molecular_function	42597	12	8	59.963
GO:0005031	tumor necrosis factor receptor activity	molecular_function	42597	19	9	59.1672
GO:0090207	regulation of triglyceride metabolic process	biological_process	42597	8	5	58.9429
GO:1903829	positive regulation of protein localization	biological_process	42597	15	9	57.0475
GO:0043304	regulation of mast cell degranulation	biological_process	42597	8	5	56.8672
GO:0033058	directional locomotion	biological_process	42597	3	3	56.6716
GO:1903384	negative regulation of hydrogen peroxide-induced neuron intrinsic apoptotic signaling pathway	biological_process	42597	3	3	56.2217
GO:0048675	axon extension	biological_process	42597	30	12	55.695
GO:0098974	postsynaptic actin cytoskeleton organization	biological_process	42597	19	10	55.2975
GO:1905653	positive regulation of artery morphogenesis	biological_process	42597	5	4	53.2122
GO:0042088	T-helper 1 type immune response	biological_process	42597	10	7	53.1343
GO:0005114	type II transforming growth factor beta receptor binding	molecular_function	42597	11	8	53.0933
GO:1903307	positive regulation of regulated secretory pathway	biological_process	42597	5	5	52.2677
GO:0034457	Mpp10 complex	cellular_component	42597	3	3	52.1378
GO:0071321	cellular response to cGMP	biological_process	42597	7	5	51.472
GO:0060058	positive regulation of apoptotic process involved in mammary gland involution	biological_process	42597	3	3	50.8291
GO:0042058	regulation of epidermal growth factor receptor signaling pathway	biological_process	42597	12	7	50.3519
GO:0002690	positive regulation of leukocyte chemotaxis	biological_process	42597	13	7	50.2111
GO:0008537	proteasome activator complex	cellular_component	42597	3	3	49.9725
GO:0031641	regulation of myelination	biological_process	42597	18	12	49.5511
GO:0071716	leukotriene transport	biological_process	42597	7	5	49.1934
GO:0031210	phosphatidylcholine binding	molecular_function	42597	32	18	48.659
GO:1904036	negative regulation of epithelial cell apoptotic process	biological_process	42597	8	5	47.925
GO:0048535	lymph node development	biological_process	42597	21	11	46.8847
GO:0098843	postsynaptic endocytic zone	cellular_component	42597	10	6	46.7845
GO:0006937	regulation of muscle contraction	biological_process	42597	16	9	46.3207
GO:1902732	positive regulation of chondrocyte proliferation	biological_process	42597	8	5	46.0419
GO:0002232	leukocyte chemotaxis involved in inflammatory response	biological_process	42597	8	5	46.0157
GO:0051146	striated muscle cell differentiation	biological_process	42597	19	10	45.6931
GO:0022841	potassium ion leak channel activity	molecular_function	42597	16	8	45.6434
GO:2001224	positive regulation of neuron migration	biological_process	42597	15	8	44.9933
GO:0043117	positive regulation of vascular permeability	biological_process	42597	14	10	44.6318
GO:0071514	genomic imprinting	biological_process	42597	17	12	44.4369
GO:0098884	postsynaptic neurotransmitter receptor internalization	biological_process	42597	10	7	44.4017
GO:0106015	negative regulation of inflammatory response to wounding	biological_process	42597	10	6	44.3686
GO:0042116	macrophage activation	biological_process	42597	24	13	44.2837
GO:0048668	collateral sprouting	biological_process	42597	10	6	43.6274
GO:0048791	calcium ion-regulated exocytosis of neurotransmitter	biological_process	42597	5	4	43.3779
GO:0004536	DNA nuclease activity	molecular_function	42597	8	5	43.1822
GO:0017146	NMDA selective glutamate receptor complex	cellular_component	42597	9	6	43.0586
GO:0090103	cochlea morphogenesis	biological_process	42597	28	12	42.9738
GO:0010614	negative regulation of cardiac muscle hypertrophy	biological_process	42597	20	9	42.8447
GO:0016327	apicolateral plasma membrane	cellular_component	42597	22	10	42.0429
GO:0061298	retina vasculature development in camera-type eye	biological_process	42597	15	9	41.8306
GO:0032649	regulation of type II interferon production	biological_process	42597	13	8	41.6643
GO:0016310	phosphorylation	biological_process	42597	13	7	41.6266
GO:0042361	menaquinone catabolic process	biological_process	42597	5	4	41.5859
GO:0035249	synaptic transmission, glutamatergic	biological_process	42597	42	18	41.4424
GO:0003056	regulation of vascular associated smooth muscle contraction	biological_process	42597	5	4	41.3227
GO:2000300	regulation of synaptic vesicle exocytosis	biological_process	42597	38	19	41.0311
GO:0035435	phosphate ion transmembrane transport	biological_process	42597	14	8	40.705
GO:0032510	endosome to lysosome transport via multivesicular body sorting pathway	biological_process	42597	6	5	40.6655
GO:0009063	amino acid catabolic process	biological_process	42597	6	5	40.4728
GO:0050431	transforming growth factor beta binding	molecular_function	42597	23	13	40.3327
GO:0006287	base-excision repair, gap-filling	biological_process	42597	14	11	40.1353
GO:0019229	regulation of vasoconstriction	biological_process	42597	23	12	40.0241
GO:0071062	alphav-beta3 integrin-vitronectin complex	cellular_component	42597	3	3	39.7753
GO:0019086	late viral transcription	biological_process	42597	4	4	39.7481
GO:0005225	volume-sensitive anion channel activity	molecular_function	42597	8	6	39.6509
GO:0031460	glycine betaine transport	biological_process	42597	3	3	39.34
GO:0070245	positive regulation of thymocyte apoptotic process	biological_process	42597	5	4	38.7276
GO:2000463	positive regulation of excitatory postsynaptic potential	biological_process	42597	38	16	38.5168
GO:0022400	regulation of opsin-mediated signaling pathway	biological_process	42597	10	6	38.3407

Table 3: Top GO terms by S(mean)

GO id	overlap genes
GO:0042571	IGHA1;IGHA2;IGHD;IGHE;IGHG1;IGHG2;IGHG3;IGHG4
GO:0071751	IGHA1;IGHA2;IGKV3-20;PIGR
GO:0060267	CAMK1D;CLEC7A;IGHA1;IGHA2;INSR
GO:0071755	CD79A;CD79B;IGHM
GO:0097278	C3;IGHG1;RAB27A
GO:0006958	C1QA;C1QB;C1QBP;C1QC;C1R;C1RL;C1S;C3;C4BPA;C4BPB;C8A;C8B;CD46;CD55;CR1;CR2;IGHA1;IGHA2;IGHD;IGHE;IGHG1;IGHG2;IGHG3;IGHG4;SUSD4
GO:0022007	DVL1;DVL2;VANGL2
GO:1903131	AXL;FOS;IGHE
GO:0001788	FCGR1A;FCGR2A;FCGR2B;FCGR3A;FCGR3B;IGHE;IGHG1
GO:0045004	POLD1;POLE;POLG
GO:1901475	SLC16A1;SLC16A11;SLC16A3;SLC16A7
GO:0150012	DLG4;DVL1;DVL2;FZD4;GRIP1
GO:0046054	GUK1;NT5C2;NT5DC4
GO:0030172	TNNI3;TNNT2;TNNT3
GO:0021681	FAIM2;MDK;TTBK2
GO:0005861	TNNI1;TNNI2;TNNI3;TNNT1;TNNT2;TNNT3
GO:0070442	ITGA2B;ITGB3;TSPAN32
GO:0060669	CDKN1C;IGF2;TRIM28;ZNF568
GO:0001560	KIF26A;MYOCD;PPP1R9B
GO:0008594	BBS1;CABP4;CDHR1;MFSD2A;TMEM216
GO:0002455	BCL3;CD81;EXO1;FCMR
GO:0042631	AQP2;AVPR1A;SIPA1
GO:0044351	KCNN4;MAPKAPK2;PYCARD;RAB34;SNX33
GO:0008253	NT5C;NT5C1A;NT5C2;NT5C3B;NT5DC3;NT5DC4;NT5M
GO:0046850	DLK1;LEPR;MDK;SUCO;SYT7
GO:0006204	NT5C;NT5C1A;NT5C2;PNP;XDH
GO:0072562	A1BG;A2M;ACSM1;ACTA1;ACTC1;ACTG1;ACTG2;AGT;ANGPTL4;APOA1;APOA2;APOA4;APOE;C1QB;C1QC;C1R;C1RL;C1S;C3;C4BPA;C8A;CD5L;CFH;CFHR1;CFHR23;IGHV3-7;IGKC;IGKV1-17;IGKV1-39;IGKV1-5;IGKV2-30;IGKV3-11;IGKV3-15;IGKV3-20;IGKV4-1;ITGA2B;ITIH2;KDM4D;LGALS3BP;OAZ3;PFN1;POTEF;PSMC5;PZP;SERPINC1;SERPINF2;SERPING1;SLC2A1;SLC4A1;TGFB1;TMPRSS13;VTN;ZNF177
GO:0034164	LILRA4;PPT1;PTPRS;RAB7B
GO:0099173	ACTN1;ADAM10;DLG5;DVL1;HIP1R;ITGA3;NRP1;SPTBN2;WASF2
GO:0044388	PARK7;SAE1;UBA2;UBE2I
GO:0043045	DNMT1;DNMT3A;TRIM28
GO:0035372	CHAMP1;CRIPT;DVL1;KIF20B;MAPRE2;MAPRE3
GO:1905368	DMWD;FAP;TPSAB1;TPSB2;USP1;USP12;USP7;WDR20
GO:0045098	PRPH;UPP2;VMAC
GO:0042867	DLAT;LDHA;LDHB;LDHC;SLC16A1;SLC16A3
GO:0000731	CDKN2D;POLD1;POLD3;POLD4;POLE;RFC3;RRM1;SIRT1
GO:0046942	ABCC1;ABCC2;ABCC3;ABCC6
GO:0008381	PIEZO1;PIEZO2;TMC4;TMC5;TMC6;TMC7;TMC8;TMEM63A
GO:0005031	FAS;TNFRSF11A;TNFRSF14;TNFRSF18;TNFRSF19;TNFRSF1A;TNFRSF1B;TNFRSF25;TNFRSF4
GO:0090207	ATG14;CIDEB;GNB3;LMF1;MBOAT7
GO:1903829	CDH1;CYLD;MAP1A;MAPT;PKP3;PPP1R9B;PRKAA2;SFN;VPS35
GO:0043304	ADGRE2;FES;PDPK1;STXBP2;UNC13D
GO:0033058	ARPIN;GRIN2A;GRIN2C
GO:1903384	AKT1;PARK7;PINK1
GO:0048675	CYFIP1;DCLK1;DVL1;EDN2;ITGB1;NDN;PPP3CB;SEMA7A;SPG11;ULK1;ULK2;VCL
GO:0098974	ACTG1;ACTL8;FARP1;MYH10;NOS1AP;POTEF;POTEI;RAC3;SRCIN1;WASF2
GO:1905653	AKT3;MDK;NFATC3;RTN4
GO:0042088	BCL3;EBI3;HRAS;IL18;IL18BP;RELB;TRAF6
GO:0005114	AMH;BAMBI;LRG1;TGFB1;TGFB2;TGFB3;TGFBR3;TGFBR3L
GO:1903307	RAB15;RAB27A;RAB3A;RAB3D;UNC13D
GO:0034457	IMP3;IMP4;MPHOSPH10
GO:0071321	HCN2;HCN4;PDE2A;PDE3A;SLC6A4
GO:0060058	BAX;PML;VDR
GO:0042058	CDH13;CEACAM1;IFI6;MVB12A;RHBDF1;RHBDF2;WDR54
GO:0002690	ADAM17;EDN2;F7;IL6R;MDK;XCL1;ZNF580
GO:0008537	PSME1;PSME2;PSME3
GO:0031641	AKT1;HES5;KIF14;LGI4;MTOR;NCMAP;NKX6-2;RARA;SIRT2;TENM4;TNFRSF1B;ZPR1
GO:0071716	ABCC1;ABCC2;ABCC3;ABCC4;ABCC6
GO:0031210	APOA2;APOA4;APOA5;APOC1;CHMP2A;CHMP3;ESYT1;GPR12;IGHM;NF1;PCTP;PITPNA;PITPNM1;PITPNM2;RASGRP1;SERPINA5;SESTD1;VDAC2
GO:1904036	ITGB3BP;MDK;NPC1;NUPR1;YAP1
GO:0048535	CD248;CXCR5;FADD;PDPN;RC3H1;RCBTB2;RIPK3;RORC;SPNS2;TGFB1;TNFRSF11A
GO:0098843	AP2A1;AP2B1;CALY;EPS15;FCHO1;HIP1R
GO:0006937	ATP1A2;ATP2A2;KCNA1;PPP1R12B;TNNI3;TNNT1;TNNT2;TNNT3;TPM1
GO:1902732	COMP;PBXIP1;SIRT6;SIX2;SOX9
GO:0002232	ALOX5;FFAR2;MDK;SLAMF1;SLAMF8
GO:0051146	AKT1;BNIP2;CASP7;CDH2;CDON;FKRP;IGF2;KRAS;RB1;SPAG9
GO:0022841	KCNK1;KCNK12;KCNK13;KCNK2;KCNK3;KCNK4;KCNK6;KCNK7
GO:2001224	ARHGEF2;KIF20B;MDK;SHTN1;SRGAP2C;TBC1D24;WDR62;ZNF609
GO:0043117	FGFBP3;PDE2A;PDE3A;TACR1;TACR2;TGFB1;TRPV4;UCN;VEGFB;ZG16B
GO:0071514	ARID4A;ARID4B;CDKN1C;CTCF;DIRAS3;DNMT3A;EED;IGF2;MTA2;NDN;PRMT7;TRIM28
GO:0098884	AP2A1;AP2A2;AP2B1;AP2S1;AP3M1;CALY;EPS15
GO:0106015	DSG2;GIT1;MDK;PTPN6;SIGLEC10;STAT3
GO:0042116	ABCC1;C1QTNF4;CRTC3;EDN2;FCGR3A;HSPA12A;IGHE;JMJD6;NR4A1;PLA2G10;STK39;TMEM106A;WNK4
GO:0048668	BDNF;DVL1;NIN;PRKG1;SPG21;ZEB2
GO:0048791	DOC2B;STX1B;SYT1;SYT7
GO:0004536	DFFB;DNASE1;DNASE1L2;MAP1S;NME1
GO:0017146	EPS8;GRIN2A;GRIN2B;GRIN2C;GRIN2D;GRIN3B
GO:0090103	DVL1;DVL2;FZD2;GRHL3;HPN;MYO3A;MYO3B;SIX1;SOX9;TBX2;VANGL2;ZEB1
GO:0010614	CERS1;GLRX3;JUND;P2RX4;RGS2;SMAD4;STUB1;TNFRSF1A;TNFRSF1B
GO:0016327	CDH2;CLDN6;CLDN7;KRT19;KRT8;MTCL1;NEDD4;PALM;SLC28A2;ZG16B
GO:0061298	ACVRL1;CYP1B1;FZD4;HIF1A;LRP5;NRP1;ROM1;TGFB1;VSTM4
GO:0032649	CCR7;CD276;ISG15;RIPK3;TNFRSF14;TNFSF14;UFC1;ZNF683
GO:0016310	AKT1;ALK;EPHB2;MAPK3;NADK;SHPK;TOLLIP
GO:0042361	CYP4F11;CYP4F12;CYP4F3;CYP4F8
GO:0035249	CACNB4;CDH8;CLSTN3;CNIH2;CNIH3;GRID1;GRIK5;GRIN2A;GRIN2B;GRIN2C;GRIN2D;GRIN3B;NAPA;P2RX1;SLC17A7;SLC17A8;SLC1A4;VPS54
GO:0003056	ADRA2B;ATP2B1;MAP2K1;P2RX1
GO:2000300	CACNB4;CALM3;DVL1;FBXL20;GIT1;LRRK2;P2RX1;PREPL;PRKCB;PRKCG;RIMS3;SLC4A8;SNAPIN;SYT1;SYT13;SYT2;SYT5;SYT8;VPS18
GO:0035435	SLC17A7;SLC20A1;SLC25A10;SLC25A3;SLC37A2;SLC37A4;VDR;XPR1
GO:0032510	LYST;MVB12A;SNX32;UBXN6;VPS4B
GO:0009063	BCKDK;ENOSF1;ETFA;ETFB;IL4I1
GO:0050431	ACVR1;ACVRL1;LTBP1;LTBP3;LTBP4;TGFB3;TGFBR3;TGFBR3L;THBS1;TSKU;TWSG1;VASN;WFIKKN1
GO:0006287	APEX1;FEN1;LIG1;LIG3;NEIL1;PARG;PNKP;POLD1;POLE;POLG;POLL
GO:0019229	ACE;ADRA2A;ADRA2B;AGT;ASIC2;ATP1A2;BDKRB2;ECE1;EDN2;KCNMB4;P2RX1;TBXA2R
GO:0071062	ITGA5;ITGB3;VTN
GO:0019086	CCNT2;MON1B;USF1;USF2
GO:0005225	LRRC8B;LRRC8C;LRRC8D;LRRC8E;TTYH1;TTYH2
GO:0031460	SLC38A2;SLC6A12;SLC7A6
GO:0070245	ADAM8;BBC3;TP53;ZC3H8
GO:2000463	BAIAP2;CHRNA7;DLG4;DVL1;GRIN2A;GRIN2B;GRIN2C;GRIN2D;NLGN2;NRXN1;PRKCZ;PTEN;RGS4;SSH1;STX1B;TBC1D24
GO:0022400	AIPL1;CAMKMT;GRK1;GUCY2D;NMT1;NMT2

Table 4: Overlap genes for top GO terms

GO id	termname	namespace	N_{bg}	K_{term}	k	S_{mean}
GO:0062167	complement component C1q complex	cellular_component	42597	3	3	-131.714
GO:0150064	vertebrate eye-specific patterning	biological_process	42597	3	3	-119.347
GO:0002228	natural killer cell mediated immunity	biological_process	42597	6	5	-98.0437
GO:0098735	positive regulation of the force of heart contraction	biological_process	42597	5	4	-95.6033
GO:0050893	sensory processing	biological_process	42597	3	3	-86.1699
GO:0005602	complement component C1 complex	cellular_component	42597	5	5	-80.522
GO:0019864	IgG binding	molecular_function	42597	14	10	-80.4567
GO:0042269	regulation of natural killer cell mediated cytotoxicity	biological_process	42597	6	5	-79.4884
GO:0008404	arachidonate 14,15-epoxygenase activity	molecular_function	42597	3	3	-78.6832
GO:0045092	interleukin-18 receptor complex	cellular_component	42597	3	3	-78.2113
GO:0001798	positive regulation of type IIa hypersensitivity	biological_process	42597	3	3	-75.6974
GO:0031004	potassium ion-transporting ATPase complex	cellular_component	42597	3	3	-75.5071
GO:2000503	positive regulation of natural killer cell chemotaxis	biological_process	42597	5	5	-75.1167
GO:0051861	glycolipid binding	molecular_function	42597	7	5	-72.0352
GO:0051150	regulation of smooth muscle cell differentiation	biological_process	42597	5	4	-71.962
GO:0006082	organic acid metabolic process	biological_process	42597	8	6	-71.7391
GO:0160075	non-canonical inflammasome complex assembly	biological_process	42597	3	3	-70.9623
GO:0046456	icosanoid biosynthetic process	biological_process	42597	11	7	-69.927
GO:0051715	cytolysis in another organism	biological_process	42597	7	6	-69.533
GO:0004499	N,N-dimethylaniline monooxygenase activity	molecular_function	42597	6	5	-69.4782
GO:0001915	negative regulation of T cell mediated cytotoxicity	biological_process	42597	8	6	-67.8938
GO:0098883	synapse pruning	biological_process	42597	7	5	-67.7702
GO:0031726	CCR1 chemokine receptor binding	molecular_function	42597	6	5	-67.6924
GO:0002711	positive regulation of T cell mediated immunity	biological_process	42597	3	3	-66.7849
GO:1905036	positive regulation of antifungal innate immune response	biological_process	42597	3	3	-65.5416
GO:0019770	IgG receptor activity	molecular_function	42597	7	5	-64.7935
GO:0043420	anthranilate metabolic process	biological_process	42597	3	3	-64.5853
GO:0019805	quinolinate biosynthetic process	biological_process	42597	3	3	-64.5853
GO:0008611	ether lipid biosynthetic process	biological_process	42597	8	5	-63.3859
GO:0106139	symbiont cell surface	cellular_component	42597	20	10	-63.2861
GO:0060512	prostate gland morphogenesis	biological_process	42597	5	4	-62.8354
GO:0005149	interleukin-1 receptor binding	molecular_function	42597	14	8	-62.3051
GO:0031838	haptoglobin-hemoglobin complex	cellular_component	42597	10	6	-61.7766
GO:1902003	regulation of amyloid-beta formation	biological_process	42597	5	4	-61.6776
GO:0044546	NLRP3 inflammasome complex assembly	biological_process	42597	10	6	-61.2867
GO:0033001	Fc-gamma receptor III complex	cellular_component	42597	3	3	-61.1738
GO:0042590	antigen processing and presentation of exogenous peptide antigen via MHC class I	biological_process	42597	7	5	-60.8982
GO:0140639	positive regulation of pyroptotic inflammatory response	biological_process	42597	7	5	-59.4089
GO:0034115	negative regulation of heterotypic cell-cell adhesion	biological_process	42597	10	7	-59.0944
GO:0071144	heteromeric SMAD protein complex	cellular_component	42597	8	6	-58.9464
GO:0071236	cellular response to antibiotic	biological_process	42597	5	5	-58.3769
GO:0002223	stimulatory C-type lectin receptor signaling pathway	biological_process	42597	22	12	-58.2886
GO:0002283	neutrophil activation involved in immune response	biological_process	42597	9	6	-58.1554
GO:0090160	Golgi to lysosome transport	biological_process	42597	11	8	-57.6619
GO:0003149	membranous septum morphogenesis	biological_process	42597	10	7	-57.222
GO:0048143	astrocyte activation	biological_process	42597	17	9	-57.183
GO:0004875	complement receptor activity	molecular_function	42597	12	9	-56.6293
GO:0043615	astrocyte cell migration	biological_process	42597	5	4	-55.834
GO:0072558	NLRP1 inflammasome complex	cellular_component	42597	6	5	-54.6539
GO:0031102	neuron projection regeneration	biological_process	42597	4	4	-53.6753
GO:0023024	MHC class I protein complex binding	molecular_function	42597	9	6	-53.5911
GO:0001660	fever generation	biological_process	42597	3	3	-53.5758
GO:0030449	regulation of complement activation	biological_process	42597	11	10	-53.5381
GO:0019884	antigen processing and presentation of exogenous antigen	biological_process	42597	3	3	-53.3107
GO:0019370	leukotriene biosynthetic process	biological_process	42597	14	8	-53.3032
GO:0019863	IgE binding	molecular_function	42597	5	4	-52.7881
GO:0140058	neuron projection arborization	biological_process	42597	8	6	-52.5125
GO:0032787	monocarboxylic acid metabolic process	biological_process	42597	18	10	-51.695
GO:0002638	negative regulation of immunoglobulin production	biological_process	42597	8	5	-51.6233
GO:1901660	calcium ion export	biological_process	42597	3	3	-51.1563
GO:2000177	regulation of neural precursor cell proliferation	biological_process	42597	11	7	-50.7622
GO:0055100	adiponectin binding	molecular_function	42597	3	3	-50.3847
GO:0047044	androstan-3-alpha,17-beta-diol dehydrogenase (NAD+) activity	molecular_function	42597	9	7	-50.1162
GO:0047023	androsterone dehydrogenase [NAD(P)+] activity	molecular_function	42597	8	7	-50.1162
GO:0051058	negative regulation of small GTPase mediated signal transduction	biological_process	42597	15	8	-49.4716
GO:0014829	vascular associated smooth muscle contraction	biological_process	42597	7	5	-49.2375
GO:0045959	negative regulation of complement activation, classical pathway	biological_process	42597	7	7	-49.1381
GO:0030492	hemoglobin binding	molecular_function	42597	5	4	-48.6783
GO:0042759	long-chain fatty acid biosynthetic process	biological_process	42597	21	14	-48.6627
GO:0071141	SMAD protein complex	cellular_component	42597	7	5	-48.3592
GO:0002292	T cell differentiation involved in immune response	biological_process	42597	4	4	-47.3049
GO:0048006	antigen processing and presentation, endogenous lipid antigen via MHC class Ib	biological_process	42597	5	5	-47.1339
GO:0030884	exogenous lipid antigen binding	molecular_function	42597	6	5	-47.1339
GO:0030883	endogenous lipid antigen binding	molecular_function	42597	6	5	-47.1339
GO:0038094	Fc-gamma receptor signaling pathway	biological_process	42597	13	11	-46.8126
GO:0004376	GPI mannosyltransferase activity	molecular_function	42597	3	3	-46.5197
GO:0004878	complement component C5a receptor activity	molecular_function	42597	3	3	-46.4232
GO:0016701	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen	molecular_function	42597	4	4	-46.2557
GO:0044194	cytolytic granule	cellular_component	42597	13	11	-46.2019
GO:0033004	negative regulation of mast cell activation	biological_process	42597	5	4	-46.061
GO:0047115	trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity	molecular_function	42597	3	3	-45.8353
GO:1990405	protein antigen binding	molecular_function	42597	9	6	-45.8265
GO:0045103	intermediate filament-based process	biological_process	42597	3	3	-45.608
GO:0035516	broad specificity oxidative DNA demethylase activity	molecular_function	42597	5	4	-45.0572
GO:1990034	calcium ion export across plasma membrane	biological_process	42597	7	5	-45.0209
GO:0086036	regulation of cardiac muscle cell membrane potential	biological_process	42597	7	5	-44.9697
GO:0010959	regulation of metal ion transport	biological_process	42597	5	4	-44.9674
GO:0050700	CARD domain binding	molecular_function	42597	15	8	-44.0957
GO:0061535	glutamate secretion, neurotransmission	biological_process	42597	3	3	-43.9196
GO:0032909	regulation of transforming growth factor beta2 production	biological_process	42597	4	4	-43.5679
GO:0097242	amyloid-beta clearance	biological_process	42597	19	12	-43.535
GO:0010744	positive regulation of macrophage derived foam cell differentiation	biological_process	42597	21	10	-43.3167
GO:1990859	cellular response to endothelin	biological_process	42597	10	7	-42.8956
GO:0008392	arachidonate epoxygenase activity	molecular_function	42597	14	12	-42.7343
GO:0004556	alpha-amylase activity	molecular_function	42597	5	4	-42.7167
GO:0048020	CCR chemokine receptor binding	molecular_function	42597	26	16	-42.608
GO:0016344	meiotic chromosome movement towards spindle pole	biological_process	42597	3	3	-42.5197
GO:0002768	immune response-regulating cell surface receptor signaling pathway	biological_process	42597	13	7	-42.5031
GO:0097449	astrocyte projection	cellular_component	42597	13	7	-42.4728
GO:0006569	L-tryptophan catabolic process	biological_process	42597	9	6	-42.1785

Table 5: Bottom GO terms by S(mean)

GO id	overlap genes
GO:0062167	C1QA;C1QB;C1QC
GO:0150064	C1QA;C3;ITGAM
GO:0002228	KLRC1;KLRC2;KLRC4;KLRD1;TYROBP
GO:0098735	NOS1;RYR2;SLC8A1;SLC9A1
GO:0050893	BBS4;CHRNA7;TTC8
GO:0005602	C1QA;C1QB;C1QC;C1R;C1S
GO:0019864	C1QA;C1QB;C1QC;FCER1G;FCGR1A;FCGR2A;FCGR2B;FCGR3A;FCGR3B;FCGRT
GO:0042269	KLRB1;KLRC1;KLRC2;NCR1;PIK3R6
GO:0008404	CYP2C9;CYP2J2;CYP4Z1
GO:0045092	IL18;IL18R1;IL18RAP
GO:0001798	C3;FCER1G;FCGR1A
GO:0031004	ABCC8;ABCC9;KCNJ8
GO:2000503	CCL3;CCL4;CCL5;CCL7;XCL1
GO:0051861	CD1C;CLEC4E;CLN3;GLTP;HSPA2
GO:0051150	BMP4;FGFR2;MYOCD;ZEB1
GO:0006082	ACADM;CYP2C8;CYP2J2;CYP2R1;FMO1;FMO2
GO:0160075	CASP4;GBP1;NR4A1
GO:0046456	CASP1;CYP2C8;CYP2C9;FADS1;NLRC4;PLA2G4A;PYCARD
GO:0051715	GBP1;GBP2;GBP3;GBP5;GBP7;RNASE7
GO:0004499	FMO1;FMO2;FMO3;FMO4;FMO5
GO:0001915	CEACAM1;KLRC1;KLRD1;LILRB1;PPP3CB;PTPRC
GO:0098883	C1QA;C1QB;C1QC;CX3CL1;DKK1
GO:0031726	CCL23;CCL3;CCL4;CCL5;CCL7
GO:0002711	IL1B;PTPRC;ZBTB1
GO:1905036	PLA2G5;SPI1;TRIM62
GO:0019770	FCGR1A;FCGR2A;FCGR2B;FCGR3A;FCGR3B
GO:0043420	HAAO;KMO;KYNU
GO:0019805	HAAO;KMO;KYNU
GO:0008611	AGPS;DHRS7B;FAR1;FASN;GNPAT
GO:0106139	C1QA;C1QB;C1QC;C1R;C1S;C3;CFH;FCN3;GBP1;GBP5
GO:0060512	BMP4;FGFR2;SERPINB5;SOX9
GO:0005149	IL1A;IL1B;IL1RN;IL36G;IL36RN;IL37;IRAK4;TLR5
GO:0031838	HBA1;HBB;HBE1;HBG2;HBZ;HP
GO:1902003	PICALM;ROCK1;ROCK2;TMED10
GO:0044546	CCDC92;GKN2;IRAK4;KCNJ8;NLRP3;PYCARD
GO:0033001	CD247;FCER1G;FCGR3A
GO:0042590	CLEC4A;FCER1G;FCGR1A;IFI30;MPEG1
GO:0140639	GBP1;GBP2;GBP3;GBP5;ZDHHC5
GO:0034115	APOA1;IL10;IL1RN;MAP2K5;MAPK7;MYADM;WNK1
GO:0071144	SMAD2;SMAD3;SMAD4;SMAD6;SMAD7;SMAD9
GO:0071236	CDKN1B;CRIP1;PLA2G4A;PLA2G4F;SLC9A1
GO:0002223	CLEC6A;CLEC7A;CREBBP;KLRC1;KLRC2;KLRC4;KLRD1;KLRK1;PAK1;PLCG2;TRAF6;TYROBP
GO:0002283	DNASE1;FCER1G;GKN2;LILRA2;SCNN1B;TYROBP
GO:0090160	ANKFY1;AP1G1;CCDC91;CLN3;EHD3;LAMP1;LAPTM5;SORT1
GO:0003149	BMP4;FGFR2;FZD2;ID2;NOG;TGFB2;VANGL2
GO:0048143	C1QA;C5AR1;FPR2;IFNG;IL1B;MAPT;NAGLU;PSEN1;ZEB2
GO:0004875	C3AR1;C5AR1;C5AR2;CMKLR1;CR1;CR2;FPR1;FPR2;FPR3
GO:0043615	APCDD1;CCL2;CCL3;MMP14
GO:0072558	CASP1;CASP4;CASP5;NLRP1;PYCARD
GO:0031102	GFAP;OMG;PTPRF;ULK1
GO:0023024	CD160;CD8A;KLRC1;KLRC2;KLRD1;TAPBPL
GO:0001660	IL1A;IL1B;TRPV1
GO:0030449	C1QBP;C3;CD55;CD59;CD5L;CFH;CR1;CR1L;PHB2;SUSD4
GO:0019884	FCGR1A;MR1;PSME1
GO:0019370	ABCC1;ALOX5;ALOX5AP;LTA4H;MGST3;PLA2G1B;PLA2G4A;PLA2G5
GO:0019863	FCER1A;FCER1G;LGALS3;MS4A2
GO:0140058	BCL7A;LRP2;LRRK2;PAK6;ROCK1;TUBA1A
GO:0032787	ACOT1;ACOT2;ACOT4;AKR1C1;AKR1C2;AKR1C3;CYP1A2;CYP2C18;CYP2C8;CYP2C9
GO:0002638	CD22;CR1;FCGR2B;FCRL3;TMBIM6
GO:1901660	ATP2B4;MCOLN1;SLC8A1
GO:2000177	FOXG1;FOXO1;ID2;NTRK3;PTBP2;SHCBP1;SIX3
GO:0055100	ADIPOR1;ADIPOR2;CDH13
GO:0047044	AKR1C1;AKR1C2;AKR1C3;DHRS9;HSD17B6;RDH16;RDH5
GO:0047023	AKR1C1;AKR1C2;AKR1C3;DHRS9;HSD17B6;RDH16;RDH5
GO:0051058	ARHGAP12;ARHGAP17;ARHGAP22;ARHGAP25;ARHGAP29;CGNL1;GMIP;WNK1
GO:0014829	ACTA2;COMP;EDNRB;PIK3C2A;SLC8A1
GO:0045959	C4BPA;C4BPB;CD46;CR1;CR1L;CR2;SUSD4
GO:0030492	HBB;HP;HPR;SLC4A1
GO:0042759	ACSBG1;ACSBG2;ALOX15;ALOX5;ASAH2;CYP1A1;CYP1A2;CYP2C8;CYP2C9;CYP2E1;FADS1;GPX4;GSTM4;PTGS2
GO:0071141	HMGA2;SMAD2;SMAD3;SMAD4;SMAD9
GO:0002292	CLEC4D;CLEC4E;FCER1G;SEMA4A
GO:0048006	CD1A;CD1B;CD1C;CD1D;CD1E
GO:0030884	CD1A;CD1B;CD1C;CD1D;CD1E
GO:0030883	CD1A;CD1B;CD1C;CD1D;CD1E
GO:0038094	CD247;CD33;CLEC4D;CLEC4E;FCER1G;FCGR1A;FCGR2A;FCGR2B;FCGR3A;FCGR3B;LCK
GO:0004376	PIGB;PIGM;PIGV
GO:0004878	C3AR1;C5AR1;C5AR2
GO:0016701	ALOX5;HPD;HPDL;PTGS2
GO:0044194	CALR;GNLY;GZMB;GZMH;LAMP1;NKG7;PRF1;RNF19B;SRGN;STXBP2;TIAL1
GO:0033004	CD300LF;CD84;CNR2;MILR1
GO:0047115	AKR1C1;AKR1C2;DHDH
GO:1990405	EP400;KLRC1;KLRC2;KLRD1;LCK;PPP2R1A
GO:0045103	GFAP;SYNC;VIM
GO:0035516	ALKBH1;ALKBH2;ALKBH3;FTO
GO:1990034	ATP2B1;SLC24A4;SLC35G1;SLC8A1;SLC8A2
GO:0086036	ATP2A2;EHD3;FXYD1;SLC8B1;SLC9A1
GO:0010959	AKAP6;NOS1;P2RX4;PRKCE
GO:0050700	BCL10;CARD14;CARD16;CARD18;CARD8;CASP1;CASP4;NOD2
GO:0061535	ABCC8;KCNJ8;NF1
GO:0032909	HIF1A;SMAD3;SMAD4;TGFB2
GO:0097242	ABCC1;C3;C5AR1;IDE;IGF1R;INSR;ITGAM;LDLR;LRP1;LRP2;MARCO;TYROBP
GO:0010744	AGT;ALOX15B;APOB;CSF1;IL18;IL1B;PLA2G10;PLA2G2A;PLA2G5;PRKCH
GO:1990859	CRK;IRS2;KLF1;KLF11;KLF2;KLF5;PRKD1
GO:0008392	CYP2A13;CYP2A6;CYP2A7;CYP2B6;CYP2C18;CYP2C8;CYP2C9;CYP2E1;CYP2F1;CYP2J2;CYP2S1;CYP4F12
GO:0004556	AMY1A;AMY1C;AMY2A;AMY2B
GO:0048020	CCL13;CCL14;CCL15;CCL16;CCL18;CCL2;CCL22;CCL23;CCL3;CCL4;CCL5;CCL7;CCL8;CX3CL1;XCL1;XCL2
GO:0016344	ACTR2;ACTR3;FMN2
GO:0002768	FGR;FPR2;KLRC1;KLRC2;TNFRSF14;VAV1;VAV3
GO:0097449	ATP1B2;GFAP;GRM5;KCNK2;MT3;PINK1;SLC1A2
GO:0006569	ACMSD;AFMID;HAAO;IL4I1;KMO;KYNU

Table 6: Overlap genes for bottom GO terms

characterized genes within the same ranked list indicates that our framework may also highlight novel candidates warranting further investigation.

3.6 Validation of Random Gene Selection

In this paper, we randomly selected 900 expressed genes for each cell sample and converted each gene into an embedded vector using the Nucleotide Transformer, while treating all genes equally regardless of their RNA expression levels. This assumes that genes with lower RNA expression levels may also contribute importantly to biological regulation or functionality. As discussed in the previous sections, our model avoids both overfitting and underfitting and shows that the probabilistic contribution scores are determined beyond simple statistics, supporting the validity of our approach.

However, this approach is not the only possible way to use DVPNet. Other approaches may provide different biological insights. For instance, we can select the top k genes for each sample based on RNA expression levels, or incorporate RNA expression information into the model by mathematically modifying the embedding vectors according to the expression levels, without affecting the smoothness and decomposability of the probabilistic circuits. Under these alternative approaches, the model would be optimized using different sources of information. Such optimization may assign probabilistic scores to genes from different biological perspectives by leveraging RNA expression level information. Exploring more effective approaches remains a task for future research.

3.7 Limitations

In this study, we used the GSE131907 single-cell lung cancer atlas for the explainable model to extract meaningful representations to differentiate between cancer cells and normal cells. However, this setting may involve some overestimation. First, we utilized only epithelial cell samples from primary sites labeled tLung and nLung to avoid situations in which the two classes have different cellular origins. For instance, if samples in the cancer-cell class are derived from epithelial cells, while samples in the normal-cell class are derived from lymph nodes, the model may focus on differences between epithelial cells and lymph cells rather than differences between cancer cells and normal cells. Our selection avoids this situation, but it also introduces a limitation.

Samples labeled as the cancer-cell class may include cells derived from the tumor microenvironment in addition to cancer cells. In this case, the classification task between cancer cells and normal cells may be implicitly shifted to a classification task between the tumor microenvironment and the normal tissue microenvironment. This interpretation is supported by our results in Section 3.3, where the top-ranked GO terms represent immune-related functions, including immunoglobulin complexes, complement activation, antibody-dependent cellular cytotoxicity, and humoral immune response pathways, suggesting that the model captured differences in immune responses between the tumor microenvironment and the normal tissue microenvironment.

Second, as our training dataset includes only one cancer type (lung cancer), drawing conclusions about biological differences between general cancer cells and normal cells would require a broader dataset encompassing multiple cancer types and corresponding normal cells.

gene	count_cancer	count_normal	S(gene)	count_total
TSPAN32	6	11	290.894	17
TPSD1	1	5	280.12	6
ANKRD2	5	7	188.399	12
MEIS3	1	3	179.817	4
CD300A	12	22	168.056	34
CYP4F35P	1	11	155.279	12
CDH16	14	34	140.037	48
CNIH2	15	39	131.022	54
CCDC144NL-AS1	2	5	130.409	7
LINC00304	1	3	120.81	4
WNK4	1	16	119.097	17
KCNK3	2	3	117.898	5
NLRC3	3	6	115.987	9
LRRC55	1	2	115.222	3
CD300C	3	14	112.126	17
SLC25A34	3	8	111.941	11
HSPB7	6	9	111.618	15
TMEM204	1	2	110.36	3
RHBDL3	2	3	108.622	5
KNDC1	32	124	107.854	156
FAM41C	30	50	107.7	80
HS3ST6	25	44	106.093	69
PKDCC	9	13	103.615	22
RASA3	10	39	103.509	49
CLCNKA	5	10	103.502	15
NHLRC4	8	93	102.023	101
RBFOX3	1	2	101.231	3
TEKT4	17	92	100.898	109
SERPINF2	70	132	99.887	202
SLC6A12	1	6	99.4355	7
LINC01135	2	3	97.2871	5
CEACAM4	16	43	96.5506	59
EFNB3	19	51	95.7778	70
ZNF213-AS1	28	44	94.7304	72
LYL1	34	46	94.6129	80
ZNF232	32	58	94.2527	90
TTC34	8	24	92.7959	32
FUOM	75	105	89.6725	180
RAMP2-AS1	3	4	88.1439	7
KRT79	1	2	86.8662	3
GNB3	1	2	86.4798	3
ASPRV1	3	4	85.5569	7
ZNF132	3	5	84.6041	8
GATA6-AS1	15	37	84.3283	52
PYY	1	3	84.0676	4
GAS8	115	153	83.1802	268
TTC39A	143	224	83.0518	367
HES2	13	51	82.2672	64
C11orf21	5	12	81.3029	17
INCA1	10	19	81.0351	29
CCDC78	23	211	79.0014	234
NAALADL1	5	17	76.1018	22
WDR90	83	114	75.1247	197
HAS1	2	4	74.349	6
MT1H	2	4	74.1198	6
SZT2-AS1	1	2	74.0155	3
SEMA6B	6	9	72.4902	15
SIGLEC9	2	13	72.4578	15
TNNT3	9	35	70.496	44
ATP2A3	32	62	70.281	94
ART5	3	4	70.0674	7
CIDEB	16	52	69.1473	68
LINC01089	31	50	68.6313	81
TMEM147-AS1	13	22	67.2793	35
SFRP5	15	97	66.6253	112
BRSK2	6	8	65.0268	14
DNAH2	53	102	64.3777	155
TP73	24	77	64.3066	101
ACAP1	35	106	64.1639	141
SCGB1B2P	11	16	63.6225	27
GALNS	77	125	63.0815	202
UTF1	1	7	62.9861	8
ARHGAP9	16	26	62.9793	42
ITGA5	18	28	62.6076	46
INAFM2	42	74	61.5643	116
SLFNL1	10	52	61.0375	62
PLIN4	6	11	60.6691	17
PSD	17	27	60.4447	44
ABCC6	201	271	59.3645	472
NOXO1	37	91	58.7282	128
CPAMD8	76	143	57.5733	219
YPEL4	2	3	55.4829	5
RRAD	348	506	55.0892	854
BAIAP3	29	76	54.561	105
CACNG8	2	9	54.0753	11
GPRIN2	42	103	53.8288	145
TNNI2	59	78	53.3674	137
PITPNM3	61	93	53.0017	154
NOTUM	8	49	52.0829	57
SLC7A6	80	129	51.6065	209
TNS2	127	195	51.5565	322
AJAP1	1	9	51.4228	10
KLHL29	25	55	51.353	80
PRDM11	22	37	50.6289	59
PRCD	2	5	50.4403	7
AP3B2	2	7	50.1222	9
IFFO1	2	13	49.2211	15
SLC4A8	27	64	48.8252	91
TRPV4	32	48	48.7152	80
C2orf81	30	86	48.4836	116

Table 7: Top 100 genes among the 1,524 genes exhibiting contradictory count–score pairs.

gene	count_cancer	count_normal	S(gene)	count_total
RDH16	2	1	-200.206	3
CHRNA7	3	1	-189.008	4
BBOX1-AS1	6	3	-176.104	9
DPYD-AS1	14	3	-130.455	17
RGS18	2	1	-123.275	3
IGKV3-11	5	3	-117.894	8
FAP	17	3	-114.154	20
FCRL3	3	1	-114.051	4
GBP2	199	137	-111.949	336
TCHH	2	1	-111.708	3
LINC00539	4	2	-106.819	6
TAS2R30	3	2	-106.761	5
DPP10	47	8	-102.855	55
MMP3	6	3	-102.439	9
ACMSD	5	1	-102.061	6
FRMD6	25	4	-101.653	29
C1orf105	2	1	-101.606	3
LINC01553	8	1	-95.4427	9
LINC00239	25	19	-94.7855	44
NR2F2-AS1	20	9	-89.4794	29
NTS	103	26	-87.1156	129
C1orf112	28	10	-86.8001	38
CYSLTR2	3	2	-84.6883	5
MMP13	23	1	-83.6066	24
CR2	4	1	-83.4151	5
PTPN22	34	16	-82.4308	50
CDH11	36	12	-80.8751	48
FUT8	251	147	-79.7628	398
GNAO1	2	1	-78.6915	3
PICALM	342	263	-77.4564	605
FDXACB1	26	11	-77.1307	37
CA10	21	1	-75.0212	22
IGKV2D-29	3	2	-74.5871	5
FCRL2	4	2	-74.3567	6
PTGS2	258	144	-74.1761	402
PRG4	3	2	-73.0783	5
PCF11-AS1	8	6	-72.3468	14
SLC16A12-AS1	3	1	-72.3176	4
MED4-AS1	3	1	-71.861	4
PTPRB	24	10	-71.2826	34
DCTN1-AS1	2	1	-70.8279	3
DEPDC1	9	4	-70.3988	13
SLFN12L	5	2	-70.1853	7
XDH	25	12	-69.86	37
NEB	8	1	-69.7119	9
SPACA3	7	1	-69.5727	8
ANO3	3	2	-69.2607	5
GFI1	3	2	-69.2318	5
GPR183	95	57	-68.9089	152
ITGBL1	4	1	-68.6218	5
CD163L1	35	5	-67.8818	40
MEX3C	161	109	-67.8149	270
SAA4	15	6	-67.4342	21
NALCN	3	2	-66.6177	5
IPO4	87	64	-66.5498	151
AKR1E2	6	1	-66.2907	7
CD22	8	3	-65.9945	11
PRSS36	28	18	-65.8769	46
LMO7DN	31	21	-65.4171	52
TMEM100	28	5	-65.2854	33
SLC16A9	187	12	-65.2111	199
STON1	24	17	-64.9908	41
NOS1	3	1	-64.766	4
CTSK	49	34	-64.7598	83
LINC01559	3	1	-64.615	4
BCAR1	206	148	-64.3865	354
LINC01001	79	48	-64.2925	127
TAT	4	3	-64.063	7
TMEM86A	25	5	-64.0455	30
TGFB2	92	44	-63.1224	136
NOTCH2	239	173	-62.8125	412
DIO2	4	2	-62.0854	6
LINC00944	9	1	-61.7683	10
NEXN-AS1	13	7	-61.6923	20
ZNF583	22	16	-61.1687	38
C2CD4D	14	9	-61.1243	23
RASSF3	137	105	-61.0342	242
QPCT	75	13	-60.727	88
CHST4	3	1	-60.2852	4
CCL4L2	175	130	-60.1277	305
TMEM81	8	5	-60.0615	13
SYT12	90	1	-59.0335	91
EFCAB13	28	8	-58.7485	36
MIR181A1HG	35	13	-58.7383	48
IL18R1	7	2	-58.6987	9
LINC01588	34	23	-58.1008	57
IFIH1	107	70	-58.0466	177
STRIP1	116	76	-57.8226	192
WDR72	26	8	-57.0351	34
SLAMF7	25	10	-57.0199	35
DHRS2	5	1	-56.937	6
SSTR1	4	1	-56.8927	5
SPC25	10	6	-56.5381	16
FFAR2	3	2	-56.4511	5
IGHV3-23	5	1	-56.3772	6
TNFRSF9	31	13	-56.3454	44
IFIT3	167	105	-56.0038	272
OAS1	239	47	-56.0009	286
CLEC2D	91	63	-55.9954	154
FNDC7	3	1	-55.7327	4

Table 8: Bottom 100 genes among the 1,524 genes exhibiting contradictory count–score pairs. $

4 Conclusion

In this paper, we presented a new XAI-based genetic profiling approach in which an explainable classification model assigns a probabilistic contribution score to each gene based on optimization for the classification task between cancer cells and normal cells. The model achieved high accuracy on both the training and test sets without exhibiting underfitting or overfitting, suggesting that the proposed classification framework is valid and that the model was successfully optimized. In addition, this study shows that the model utilized both statistical signals and biological feature representations encoded by the Nucleotide Transformer. Some genes exhibited probabilistic contributions that were contradictory to their statistical frequencies, suggesting that our analysis, which integrates a biological foundation model with statistical information, may provide biological insights different from those obtained through traditional statistics-based analyses.

References

[1] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.

[2] L. Wang, X. Liu, M. Yue, Z. Liu, Y. Zhang, Y. Ma, J. Luo, W. Li, J. Bai, H. Yao, *et al.*, "Identification of hub genes in bladder cancer based on weighted gene co-expression network analysis from tcga database," *Cancer Reports*, vol. 5, no. 9, p. e1557, 2022.

[3] Y. Yuan, J. Chen, J. Wang, M. Xu, Y. Zhang, P. Sun, and L. Liang, "Identification hub genes in colorectal cancer by integrating weighted gene co-expression network analysis and clinical validation in vivo and vitro," *Frontiers in oncology*, vol. 10, p. 638, 2020.

[4] J. A. Miller, M. C. Oldham, and D. H. Geschwind, "A systems level analysis of transcriptional changes in alzheimer's disease and normal aging," *Journal of neuroscience*, vol. 28, no. 6, pp. 1410–1420, 2008.

[5] M. C. Oldham, G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. H. Geschwind, "Functional organization of the transcriptome in human brain," *Nature neuroscience*, vol. 11, no. 11, pp. 1271–1282, 2008.

[6] E. A. Serin, H. Nijveen, H. W. Hilhorst, and W. Ligterink, "Learning from co-expression networks: possibilities and challenges," *Frontiers in plant science*, vol. 7, p. 444, 2016.

[7] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[8] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, *et al.*, "Nucleotide transformer: building and evaluating robust foundation models for human genomics," *Nature Methods*, vol. 22, no. 2, pp. 287–297, 2025.

[9] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[10] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] T. Kusumoto, "Vpnet: A new explainable-ai paradigm for tractable multimodal, multidimensional classification with probabilistic circuits on vit features," *medRxiv*, 2025.

[13] Y. Choi, A. Vergari, and G. Van den Broeck, "Probabilistic circuits: A unifying framework for tractable probabilistic models." http://starai.cs.ucla.edu/papers/ProbCirc20.pdf, 2020. UCLA Technical Report.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020. arXiv preprint arXiv:2010.11929.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv preprint arXiv:2103.00020.

[16] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv preprint arXiv:2104.14294.

[17] N. Friel and A. N. Pettitt, "Marginal likelihood estimation via power posteriors," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, pp. 589–607, 04 2008.

[18] N. Kim, H. K. Kim, K. Lee, Y. Hong, J. H. Cho, J. W. Choi, J.-I. Lee, Y.-L. Suh, B. M. Ku, H. H. Eum, *et al.*, "Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma," *Nature communications*, vol. 11, no. 1, p. 2285, 2020.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[20] P. Nguyen and E. Zeng, "A protocol for weighted gene co-expression network analysis with module preservation and functional enrichment analysis for tumor and normal transcriptomic data," *Bio-protocol*, vol. 15, no. 18, 2025.

[21] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[22] "The gene ontology resource: enriching a gold mine," *Nucleic acids research*, vol. 49, no. D1, pp. D325–D334, 2021.

[23] T. Zhou, J. Yao, and Z. Liu, "Gene ontology, enrichment analysis, and pathway analysis," *Bioinformatics in aquaculture: Principles and methods*, pp. 150–168, 2017.

[24] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[25] K. M. Bell-McGuinn, C. M. Matthews, S. N. Ho, M. Barve, L. Gilbert, R. T. Penson, E. Lengyel, R. Palaparthy, K. Gilder, A. Vassos, *et al.*, "A phase ii, single-arm study of the anti-$\alpha 5\beta 1$ integrin antibody volociximab as monotherapy in patients with platinum-resistant advanced epithelial ovarian or primary peritoneal cancer," *Gynecologic oncology*, vol. 121, no. 2, pp. 273–279, 2011.

[26] J. S. Desgrosellier and D. A. Cheresh, "Integrins in cancer: biological implications and therapeutic opportunities," *Nature Reviews Cancer*, vol. 10, no. 1, pp. 9–22, 2010.

[27] H. Läubli, S. C. Nalle, and D. Maslyar, "Targeting the siglec–sialic acid immune axis in cancer: current and future approaches," *Cancer immunology research*, vol. 10, no. 12, pp. 1423–1432, 2022.

[28] H. Choi, M. Ho, O. S. Adeniji, L. Giron, D. Bordoloi, A. J. Kulkarni, A. P. Puchalt, M. Abdel-Mohsen, and K. Muthumani, "Development of siglec-9 blocking antibody to enhance anti-tumor immunity," *Frontiers in Oncology*, vol. 11, p. 778989, 2021.

[29] Y. Zhao, S. Jolly, S. Benvegnu, E. Y. Jones, and P. V. Fish, "Small-molecule inhibitors of carboxylesterase notum," *Future medicinal chemistry*, vol. 13, no. 11, pp. 1001–1015, 2021.

[30] E. D. Bayle, F. Svensson, B. N. Atkinson, D. Steadman, N. J. Willis, H. L. Woodward, P. Whiting, J.-P. Vincent, and P. V. Fish, "Carboxylesterase notum is a druggable target to modulate wnt signaling," *Journal of Medicinal Chemistry*, vol. 64, no. 8, pp. 4289–4311, 2021.

[31] J. M. Rozenberg, S. Zvereva, A. Dalina, I. Blatov, I. Zubarev, D. Luppov, A. Bessmertnyi, A. Romanishin, L. Alsoulaiman, V. Kumeiko, *et al.*, "Dual role of p73 in cancer microenvironment and dna damage response," *Cells*, vol. 10, no. 12, p. 3516, 2021.

[32] K. Kawakami, S. Yamamura, H. Hirata, K. Ueno, S. Saini, S. Majid, Y. Tanaka, K. Kawamoto, H. Enokida, M. Nakagawa, *et al.*, "Secreted frizzled-related protein-5 (sfrp-5) is epigenetically downregulated and functions as a tumor suppressor in renal cell cancer: 356," *Journal of Urology*, vol. 183, no. 4, p. e141, 2010.

[33] R. Karki, S. M. Man, R. S. Malireddi, S. Kesavardhana, Q. Zhu, A. R. Burton, B. R. Sharma, X. Qi, S. Pelletier, P. Vogel, *et al.*, "Nlrc3 is an inhibitory sensor of pi3k–mtor pathways in cancer," *Nature*, vol. 540, no. 7634, pp. 583–587, 2016.

[34] D. Lankry, T. L. Rovis, S. Jonjic, and O. Mandelboim, "The interaction between cd 300a and phosphatidylserine inhibits tumor cell killing by nk cells," *European journal of immunology*, vol. 43, no. 8, pp. 2151–2161, 2013.

[35] S. Kim, I.-H. Han, S. Lee, D. Park, H. Lee, J. Kim, J. Kim, J.-W. Jeon, and H. Bae, "The combination of cd300c antibody with pd-1 blockade suppresses tumor growth and metastasis by remodeling the tumor microenvironment in triple-negative breast cancer," *International Journal of Molecular Sciences*, vol. 26, no. 11, p. 5045, 2025.

[36] J. Chen, Y. Yao, C. Gong, F. Yu, S. Su, J. Chen, B. Liu, H. Deng, F. Wang, L. Lin, *et al.*, "Ccl18 from tumor-associated macrophages promotes breast cancer metastasis via pitpnm3," *Cancer cell*, vol. 19, no. 4, pp. 541–555, 2011.

[37] W. Wagner, B. Ochman, and W. Wagner, "Semaphorin 6 family—an important yet overlooked group of signaling proteins involved in cancerogenesis," *Cancers*, vol. 15, no. 23, p. 5536, 2023.