GRADIENT-BASED GENE SELECTION FOR MULTI-MODAL SCRNA-SEQ FOUNDATION MODELS

Pakaphol Thadawasin[1], **Farhan Khaodee**[1], **Rohola Zandie**[1], **Elazer R. Edelman**[1,2]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Medicine (Cardiovascular Medicine), Brigham and Women's Hospital, Boston, MA, USA
{big536, farhank, rohola, ere}@mit.edu

ABSTRACT

Foundation models have emerged as powerful tools for analyzing single-cell RNA sequencing (scRNA-seq) data. However, selecting informative gene features for both input to the model and analysis in the output remains a critical challenge. Traditional feature selection methods filter on the basis of highly variable genes and analyze them using differential distribution, but they often struggle with scalability and robustness in heterogeneous, high-dimensional datasets. In this study, we explore the limitations of conventional feature selection techniques in the context of a multimodal foundation model and propose alternative gradient-based attribution techniques on learned feature embeddings to improve feature selection. We demonstrate how our selection strategy enhances model performance, overcomes the limitations of traditional approaches, and holds the potential to reveal the inherent polygenicity of diseases. [1]

1 MOTIVATION

Single-cell RNA sequencing (scRNA-seq) has transformed our understanding of cellular heterogeneity by enabling the measurement of gene expression at the individual cell level. A key challenge in scRNA-seq analysis is feature selection: identifying genes that reduce dimensionality whilst preserving biological relevance. Traditional methods, such as selecting Highly Variable Genes (HVG) or Differentially Expressed Genes (DEG), have limitations in modern, large-scale datasets, particularly when integrated into foundation models trained on data spanning diverse tissues, conditions, and species.

Conventionally, scRNA-seq feature selection employs HVG selection, assuming these genes drive cellular heterogeneity. However, this approach has several drawbacks. Specifically, naive variance-based ranking can overestimate variability (Heumos et al., 2023), exclude contextually variable housekeeping genes (Yip et al., 2019), and overrepresent cell cycle or stress-related genes, obscuring other signals. Arbitrary variability thresholds further compromise consistency across studies (Chen et al., 2019). DEG selection between predefined experimental groups is another widely used approach, as it is designed to pinpoint genes that show significant expression changes between conditions. DEG analysis struggles with diverse scRNA-seq datasets, as it depends on predefined labels and may miss subtle, context-specific signals (seq, 2014; Corchete et al., 2020; Hoerbst et al., 2025). Technical noise, including dropouts and amplification bias, further distorts results, especially for lowly expressed genes (Conesa et al., 2016; McDermaid et al., 2019).

Transformer-based foundation models have emerged to integrate extensive scRNA-seq datasets (Cui et al., 2024), (Yang et al., 2022). Recently, Khodaee et al. (2025) introduced POLYGENE, a multimodal transformer-based scRNA-seq model that incorporates both scRNA-seq data and annotated phenotypes, enabling the model to learn complex relationships between genotypes and phenotypes. Building on POLYGENE, we propose a gradient-based gene selection algorithm that prioritizes genes in the models trained on complete gene sets, avoiding prior assumptions about variability. This method can be generalized to any foundation model. We identify the most influential genes contributing to the model's output, and validate our method across multiple downstream tasks, demon-

[1]Source Code: https://github.com/pakaphholbig/GradGeneSelection

strating its effectiveness and potential to overcome the limitations of traditional feature selection techniques in the era of large-scale single-cell analysis.

2 RELATED WORK

Feature selection in scRNA-seq traditionally relies on statistical models that identify genes exhibiting significant variability or differential expression across cell populations. HVG detection methods (Stuart et al., 2019), (O'Callaghan et al., 2024), (Buettner et al., 2017), play a crucial role in reducing dimensionality (Yip et al., 2019). Differential expression analysis, conducted using tools like DESeq2 (Love et al., 2014) and edgeR (Alessandrì et al., 2019), identifies genes that exhibit statistically significant expression differences between conditions (Rosati et al., 2024b). While effective, these methods often assume specific data distributions and struggle with highly heterogeneous single-cell datasets as their sizes increase. Deep learning methods have been also used as alternative approaches (Huang et al., 2023) and have shown promising results Huang et al. (2023). Nonetheless, feature selection methods haven't been studied for foundation models.

3 METHODOLOGY

The essence of the proposed method is to assess the attention of a multimodal foundation model, specifically POLYGENE[2] (Khodaee et al., 2025) places on each input feature when predicting phenotypic differences between two cells. This method is generalizable to any transformer-based foundation model, depending on its input format. The algorithm comprises two main phases: **Gene Blending** and **Gene Attributing**. Finally, a mapping of genes to their attribution values is generated, producing a ranked list for gene feature selection.

3.1 PRE-PROCESSING

We predefine a target phenotypic type and a set of controlled phenotypic types for analysis (see Appendix A.1 for a rigorous definition). The target phenotype serves as the independent variable, differentiating the input from the baseline, while both share the same controlled phenotypes. This defines a distribution from which the input is sampled; however, the necessity and choice of baseline will be explained in Section 4 and Appendix A.3. After this phase, we obtain an input and a baseline for further analysis.

3.2 GENE BLENDING

While scRNA-seq data are designed to share the same set of phenotypic categories, differences in gene sets between the input I and baseline B can introduce biases. To address this, we introduce **Gene Blending**, a preprocessing step that aligns shared genes in both datasets while preserving structural consistency. This ensures that genes occupy the same positions across the input and baseline. (see Figure 1). Genes are divided into three categories: shared, input-only, and baseline-only. They are then rearranged so that all shared genes are aligned, and padding tokens are used for input-only and baseline-only genes. This step is crucial for gradient-based attribution tools (e.g., DeepLIFT (Shrikumar et al., 2019)) to correctly compare gradients of the same features in both the input and baseline. A detailed formulation of the Gene Blending algorithm is provided in Appendix A.2.

3.3 GENE ATTRIBUTING

In this phase, our goal is to identify the genes that are most strongly associated with a target phenotype.[3] Genes, to which the model assigns significant focus, are likely to serve as markers for the target phenotype. We refer to this measure of focus as an **attribution value**. During the analysis,

[2]The current version of POLYGENE supports six phenotypic types as input: cell type, tissue, disease, developmental stage, sex, and assay. This flexibility allows us to systematically evaluate gene importance across different biological conditions, enabling a more nuanced understanding of how genetic features contribute to phenotypic variation.

[3]Our question is: *If the target phenotype is masked in the cell data, how much focus does the model place on each gene and controlled variable when making its predictions?*

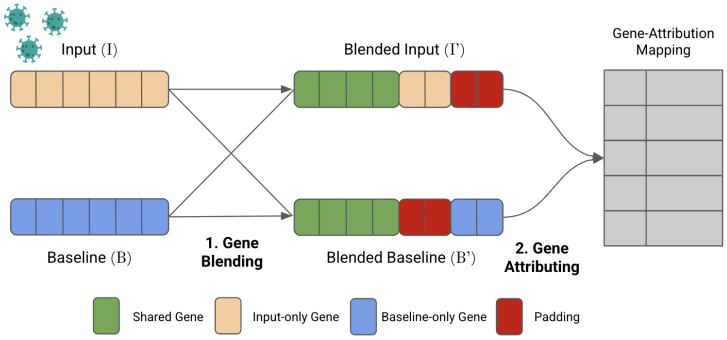


Figure 1: Overview of the proposed gene attribution pipeline. The method consists of two main phases: (1) Gene Blending, where the input and baseline gene expressions are blended to ensure that all genes are used during the analysis, and (2) Gene Attributing, where attributions are computed based on the model's predictions. The final output is a gene-attribution mapping used for feature selection and downstream analysis.

the target phenotype of the blended input and blended baseline will be masked and simultaneously fed into POLYGENE. DeepLIFT (Shrikumar et al., 2017)[4] and then estimates an attribution value for each gene by comparing the model's predictions of the target phenotype of the masked blended input and baseline. For better efficiency and convergence, we adopt Ancona et al. (2018) implementation of DeepLIFT. The raw attribution values will be normalized to ensure consistency across all cells and multiplied by the input to capture global attribution behavior (Ancona et al., 2018). Finally, the mean of normalized attribution values will be calculated for each feature, across all pairs of input and baseline.

4 CELL-STATE TRANSITION: BASELINE SELECTION

Motivated by differential gene expression (DGE) analysis for disease biomarker identification (Rosati et al., 2024a), we choose a healthy cell as the input and a diseased cell as the baseline, both exhibiting the same controlled phenotypic traits. While we use **diseased versus healthy as an example**, the same approach can be generalized to other phenotypic transitions depending on our choice of target phenotypic type, such as differentiating versus undifferentiated cells, drug-treated versus untreated cells, or activated versus resting immune cells. The full motivation and formulation can be found in Appendix A.3.

5 ATTRIBUTION IMPORTANCE: DATA ASYMMETRY

Data asymmetry arises during the gene blending process, as the input and baseline phenotypes typically do not share the exact same gene set. As discussed in Section 3.2, this is mitigated by applying padding; however, it is trivial that the attribution values of padding tokens in the input are zero (see Appendix A.4). To avoid this inherent bias, an additional pass is performed during the gene attribution step by swapping the input and baseline. Specifically, given an input I, a baseline B, and a function f which $f(I, B)$ assigns attribution values to each feature in the input-baseline pair, we also compute $f(B, I)$ in addition to $f(I, B)$. We denote the attribution values from $f(I, B)$ as **forward attributions** and those from $f(B, I)$ as **backward attributions**. Then, we compute the **attribution importance** for each gene, which is used for downstream analysis, by taking the arithmetic mean of the forward and backward attributions [5]. In Appendix A.7, we propose an alternative method for handling data asymmetry, which also yields a biologically meaningful set of genes for the biomarker identification task.

[4]We also experimented with other gradient-based attribution methods, such as Integrated Gradients (Sundararajan et al., 2017) and SHAP (Lundberg & Lee, 2017). However, due to the deep architecture of POLYGENE, the attribution values either failed to converge or required excessive computation time.

[5]Attribution Importance $= \frac{f(I,B)+f(B,I)}{2}$

6 EXPERIMENT

In this work, we use POLYGENE (Khodaee et al., 2025) as our choice of model, and utilize TABULA SAPIENS (Consortium, 2022), a manually annotated scRNA-seq dataset comprising nearly 500,000 cells across 24 different tissues and organs, as our validation set. For preprocessing and analysis, we employ the SCANPY package (Wolf et al., 2018), which facilitates the handling of scRNA-seq data, computation of HVGs, and differential gene expression (DEG) analysis. Total embeddings, the sum of input and token type embeddings, are the input reference for DeepLIFT (Shrikumar et al., 2019) analysis.

We select **disease** as the target phenotypic type, while **cell type**, **tissue**, or a combination of both serve as controlled phenotypic types. Three different selections of target phenotypes and controlled phenotypes are tested, as outlined in Table 1. In each set of experiments, 5,000 cells are sampled from the dataset. HVGs are identified by sampling 2 million cells from the entire dataset. To perform differential gene expression (DEG) analysis, only cells with the target disease and normal conditions from the specified cell type and tissue are sampled.

Experiment	Disease	Cell Type	Tissue
1	Alzheimer's Disease	Microglial Cell	Brain
2	Breast Cancer	Fibroblast	Breast
3	Dilated Cardiomyopathy	Any	Heart

Table 1: Selected cell types for downstream analysis across different diseases.

7 EVALUATION AND RESULTS

7.1 SPARSITY OF ATTRIBUTION IMPORTANCE

As discussed in Section 5, we expect both forward and backward attributions to be sparse. We found that the distribution of attribution importance is also sparse but centered around a non-zero mode, with a few outliers exhibiting extreme attribution values (see Appendix A.4). This suggests that certain genes strongly influence the model's prediction of the target phenotype.

7.2 IDENTIFICATION OF POTENTIAL DISEASE BIOMARKERS

Developed from Section 7.1, we identify genes that play a more significant role in the model's predictions. Since the model does not have access to the ground truth disease labels of the cells, attributions reflect the genes the model considers most important for prediction. Therefore, the idea of identifying positive outliers as potential disease biomarkers emerges naturally. Specifically, genes with exceptionally high attributions across multiple conditions are likely to serve as disease markers (A.5).

7.3 IMPORTANCE OF OUTLIERS: GENE PRUNING

We demonstrate how pruning genes based on different selection heuristics—attribution importance, forward attribution, HVGs, and random removal—affects the model's predictive accuracy across Alzheimer's disease, breast cancer, and dilated cardiomyopathy. The genes with low importance scores under each method are removed, and the model's classification metrics are evaluated. We refer to this evaluation framework as **Gene Pruning**. Figure 2 illustrates the relationship between accuracy and prune ratios across different heuristics, while plots for additional classification metrics (e.g., F1 score, precision, and recall) are provided in Appendix A.6. The setup of this experiment remains the same, and we remove a proportion of genes, determined by the prune ratio, that have the lowest importance scores. We hypothesize that the differing trends observed in each disease reflect the underlying causes and intrinsic nature of the diseases, as further discussed in Appendix A.6.

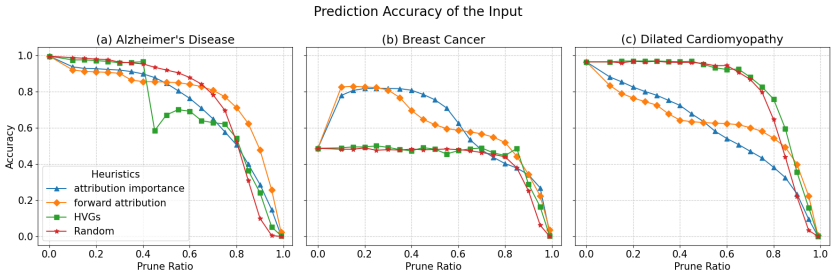
8 DISCUSSION

In this work, we present a gradient-based gene selection method for multimodal scRNA-seq foundation models. We mathematically and theoretically formalize a deep learning-based gene selection framework that can be generalized to other models. Motivated by cell-state transitions, our baseline selection strategy captures differential gene expression between an input and a reference. Attribution Importance is introduced to address data asymmetry, ensuring that all genes contribute meaningfully to the computed attribution values.

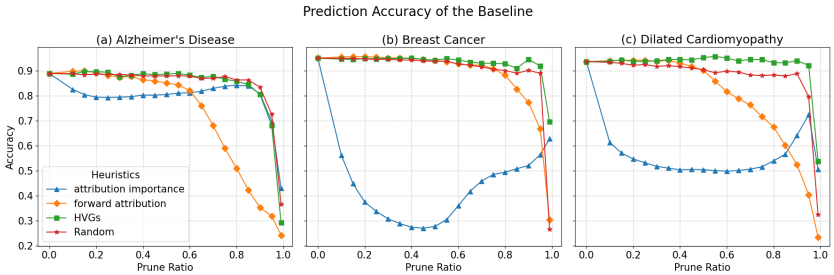
Upon evaluating our method across three different diseases, we found that attribution importance distributions are sparse, with notable outliers in both tails. This indicates that the model prioritizes certain genes over others. Nevertheless, only a few show known associations with the disease. We suspect this discrepancy may be due to experimental bias in the dataset (Megill et al., 2021). Additionally, some genes remain uncharacterized, motivating further experimental validation.

Although the primary goal of Gene Pruning is to benchmark the impact of gene removal on model performance, it also provides valuable insights into the underlying genetic interplay associated with each disease. As shown in Figure 2 and discussed in Appendix A.6, our method outperforms HVG-based selection in certain diseases. Moreover, the classification plots of the baseline (Figure 2) and reverse pruning experiments (Figure 13 and 14) further highlight the significance of genes with high attribution importance. We hypothesize that the differing trends observed reflect intrinsic disease characteristics, such as heterogeneity and polygenicity, though requires further investigation to validate.

Our work introduces a gradient-based gene selection framework tailored for multimodal scRNA-seq foundation models. However, it assumes gene independence in phenotype expression during the aggregation phase, as attribution values are averaged across all samples. Additionally, we observed convergence issues with gradient-based attribution tools, particularly when applied to large, deep transformer-based architectures.



(A) Prediction accuracy of the input after gene pruning.



(B) Prediction accuracy of the baseline after gene pruning .

Figure 2: Gene pruning accuracy comparisons using different heuristics for both input (A) and baseline (B) across three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

9 IMPACT STATEMENT

Our gene selection framework advances the field of single-cell transcriptomics by enabling more precise and biologically meaningful feature selection for foundation models. The integration of attribution-based gene selection into large-scale scRNA-seq analysis could lead to more effective disease classification and efficient biomarker discovery for early diagnosis and therapeutic targets.

REFERENCES

A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914, 2014.

Luca Alessandrì, Maddalena Arigoni, and Raffaele Calogero. Differential expression analysis in single-cell transcriptomics. *Single Cell Methods: Sequencing and Proteomics*, pp. 425–432, 2019.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018. URL https://arxiv.org/abs/1711.06104.

Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18:1–13, 2017.

Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17:1–19, 2016.

Tabula Sapiens Consortium. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896. URL https://www.science.org/doi/10.1126/science.abl4896.

Luis A Corchete, Elizabeta A Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C Gutiérrez, and Francisco J Burguillo. Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. *Scientific reports*, 10(1):19737, 2020.

Hao Cui, Cheng Wang, Hira Maan, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

Franziska Hoerbst, Gurpinder Singh Sidhu, Melissa Tomkins, and Richard J Morris. What is a differentially expressed gene? *bioRxiv*, pp. 2025–01, 2025.

Hao Huang, Chunlei Liu, Manoj M Wagle, and Pengyi Yang. Evaluation of deep learning-based feature selection for single-cell rna sequencing data analysis. *Genome Biology*, 24(1):259, 2023.

Farhan Khodaee, Rohola Zandie, and Elazer R. Edelman. Multimodal learning for mapping genotype–phenotype dynamics. *Nature Computational Science*, 2025. doi: 10.1038/s43588-024-00765-7.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Adam McDermaid, Brandon Monier, Jing Zhao, Bingqiang Liu, and Qin Ma. Interpretation of differential gene expression results of rna-seq data: review and integration. *Briefings in bioinformatics*, 20(6):2044–2054, 2019.

Colin Megill, Bruce Martin, Charlotte Weaver, Sidney Bell, Lia Prins, Seve Badajoz, Brian McCandless, Angela Oliveira Pisco, Marcus Kinsella, Fiona Griffin, et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *BioRxiv*, pp. 2021–04, 2021.

Alan O'Callaghan, Nils Eling, John C Marioni, and Catalina A Vallejos. Basics workflow: a step-by-step analysis of expression variability using single cell rna sequencing data. *F1000Research*, 11:59, 2024.

Diletta Rosati, Maria Palmieri, Giulia Brunelli, Andrea Morrione, Francesco Iannelli, Elisa Frullanti, and Antonio Giordano. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, 22:1154–1168, 2024a. doi: 10.1016/j.csbj.2024.02.018. URL https://doi.org/10.1016/j.csbj.2024.02.018.

Diletta Rosati, Maria Palmieri, Giulia Brunelli, Andrea Morrione, Francesco Iannelli, Elisa Frullanti, and Antonio Giordano. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and structural biotechnology journal*, 2024b.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019. URL https://arxiv.org/abs/1704.02685.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0. URL https://doi.org/10.1186/s13059-017-1382-0.

Fan Yang, Wenjia Wang, Fan Wang, et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4: 852–866, 2022. doi: 10.1038/s42256-022-00534-z. URL https://doi.org/10.1038/s42256-022-00534-z.

Shun H Yip, Pak Chung Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4):1583–1589, 2019.

A APPENDIX

A.1 PROBLEM SETUP

Formally, let P_i be the set of phenotypes for phenotypic type i, and G the set of genotypes. An annotated scRNA-seq dataset D with n phenotypes, and m genotypes is represented as a sequence $(p_1, \ldots, p_n, g_1, \ldots, g_m)$, where each $p_i \in P_i$ and each g_k denotes the expression value of gene $k \in G$. It is important to note that we keep n fixed for the entire dataset. In this work, we assume n is fixed, as phenotypes are manually annotated, while m varies with each cell across the dataset. Let $t \in \{1, \ldots, n\}$ denote a target phenotypic type. Then, the set of controlled phenotypic types C is defined as a subset of all phenotypic types except the target, specifically $C \subseteq \{1, \ldots, n\} \setminus \{t\}$.

A.2 GENE BLENDING: FORMALIZATION

Let G_I and G_B be the sets of gene indices that appear in the input and baseline data, respectively. For simplicity, assume $G_I \cup G_B = \{1, 2, \ldots, M\}$, where $M = |G_I \cup G_B|$. We denote the input as $I = \left(p^I, g^I\right)$ and the baseline as $B = \left(p^B, g^B\right)$, where:

- p^I and p^B are the phenotype parts (of length n)
- g^I and g^B are the genotype part (indexed by $\{1, \ldots, M\}$).
- If a particular index j is not in G_I, then g^I_j is not applicable (we will pad it as needed). Likewise for g^B_j when $j \notin G_B$.

We form blended version I' and B' by "sharing" genes where they overlap and padding otherwise (see Figure 1):

$$I' = \left(p^I, g'\right), \quad B' = \left(p^B, g''\right) \tag{1}$$

where, for each gene index $j \in \{1, \ldots, M\}$,

$$g'_j = \begin{cases} g^I_j, & \text{if } j \in G_I, \\ \texttt{[PAD]}, & \text{otherwise}, \end{cases} \quad g''_j = \begin{cases} g^B_j, & \text{if } j \in G_B, \\ \texttt{[PAD]}, & \text{otherwise}. \end{cases} \tag{2}$$

- **Shared genes** $(G_I \cap G_B)$ get real genotype values in both I' and B'.
- **Input-only genes** $(G_I \backslash G_B)$ appear normally in I' but are padded in B'
- **Baseline-only genes** $(G_B \backslash G_I)$ appear normally in B' but are padded in I'

This ensures that I' and B' remain aligned across relevant gene positions, with mismatched genes padded as `[PAD]`.

A.3 CELL-STATE TRANSITION-MOTIVATED BASELINE SELECTION

The choice of baseline determines the semantics of attribution values by assessing the *focus* the model places on each feature during prediction. Its primary functions are to prevent misinterpretation due to inherent attribution values of features and to address issues such as zero gradients and discontinuities in deep networks. However, selecting an appropriate baseline depends on the specific task and the underlying structure of the input domain. For example, Shrikumar et al. (2019) used an all-zero baseline for image data and an ACGT baseline sampled according to its natural frequency for genomic sequences.

Cell-state transitions are fundamental to cellular processes and serve as indicators of various phenotypic changes. Motivated by differential gene expression (DGE) analysis for disease biomarker identification (Rosati et al., 2024a), we leverage deep learning models to analyze these transitions. We define a **healthy cell** as our **baseline** and a **diseased cell** as our **input**. Both share the same value for certain controlled phenotypic types, while any other phenotypic types are left unspecified. Formally, let t be the **target phenotypic type** (diseased status), C be the set of **controlled phenotypic types**. u be the target phenotype for the input I and v for the baseline B, and finally

$I = (p_1^I, \ldots, p_n^I, g_1^I, \ldots, g_m^I)$ represents the input cell, $B = (p_1^B, \ldots, p_n^B, g_1^B, \ldots, g_{m'}^B)$ represents the baseline cell. We set the phenotypic value p_i^B in the baseline as follows:

$$p_i^B = \begin{cases} p_i^I, & \text{if } i \in C \\ u \neq v, & \text{if } i = t \\ \text{arbitrary}, & \text{otherwise} \end{cases} \tag{3}$$

using these definitions we can compare the relative importance the model assigns to each gene in the **diseased cell** (input) versus the **healthy cell** (baseline) thereby filtering out housekeeping genes. In this case, we define the phenotypic type as disease, as it provides a more intuitive basis for reasoning.

A.4 Sparsity of Attribution Values

To mitigate the gene asymmetry problem between the input and baseline, gene blending is introduced. As discussed in Section 3.2, padding is applied to the input and baseline to ensure proper alignment of common genes, with no attention mask imposed. Since we implement the gradient-based DeepLIFT method proposed by Ancona et al. (2018) to improve latency and convergence, we observe that the attribution values of genes present only in the input are zero. This occurs because attention to padding is masked, nullifying its contribution to the model's computations. As a result, the attribution value with respect to the masked attention value becomes zero. Thus, we aim to compare the distribution of forward attribution, backward distribution, and attribution importance.

We plotted the distributions of forward attributions, backward attributions, and attribution importance for the Alzheimer's disease experiment. Approximately 59% and 38% of genotypes have zero forward and backward attributions, respectively, which is expected, as discussed in Section 5. Attribution importance successfully addresses this issue, with only around 3% of genes having zero attribution importance. Despite this improvement, the distribution of attribution importance remains sparse (see Figure 3).

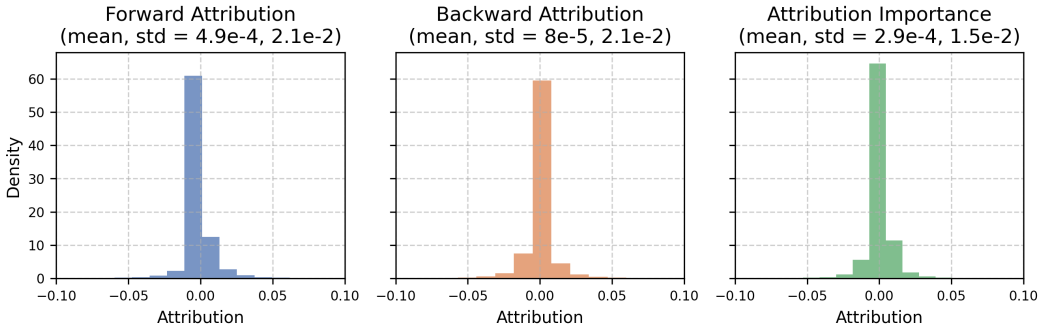


Figure 3: Distribution of each attribution heuristic in Alzheimer's disease.

Since genes interact in phenotypic expression, we aggregate attributions by gene and compute their arithmetic mean to analyze their behavior across different cell states and conditions. The attributions are visualized in Figure 6, where genes are represented by gene IDs. Genes with exceptionally high or low attributions are classified as **positive** and **negative outliers**, respectively. A significant number of these outliers exhibit extreme attribution values beyond the 1st–99th percentile range. This sparsity pattern is also evident in both the breast cancer and dilated cardiomyopathy experiments. (see Figure 4, 5) We illustrate some possible use cases and interpretations of these outliers in the following findings.

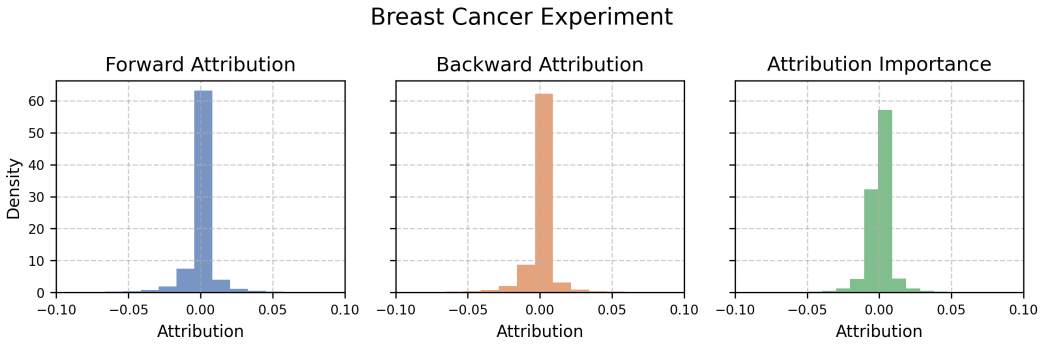


Figure 4: Distribution of each attribution heuristic in Breast Cancer disease.

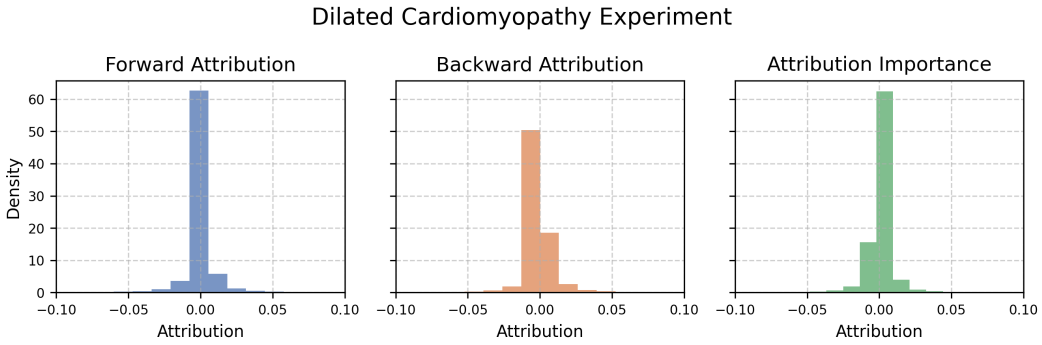


Figure 5: Distribution of each attribution heuristic in Dilated Cardiomyopathy disease.

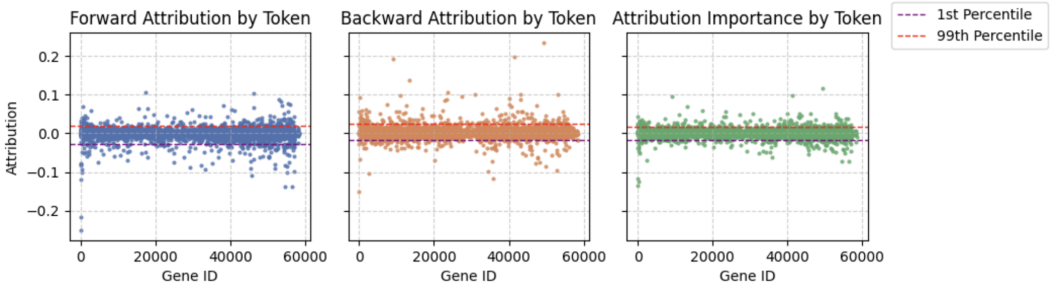


Figure 6: Attributions are grouped by Gene ID, where the purple line represents the 1st percentile and the red line represents the 99th percentile.

A.5 Top-30 Disease Biomarkers Comparison

Below is the list of the 30 genes with the highest attribution importance, as referenced in Section 7.2.

Rank	Alzheimer's Disease		Breast Cancer		Dilated Cardiomyopathy	
	Attribution	**DEG**	**Attribution**	**DEG**	**Attribution**	**DEG**
1	AC004540.4	ZFHX3	H2AC19	LINC02510	CTB-131B5.5	RP11-305P14.2
2	RP11-411B6.6	KCNIP4	LMOD2	AC131056.5	RP11-380M21.4	SNHG14
3	RGPD4-AS1	CSMD1	OR2A1	DRD3	KRT7	DISC2
4	RP11-422P24.15	PTPRD	MTRNR2L1	RP4-739H11.3	LINC00486	RP11-115H11.1
5	RP11-344F5.1	SNHG14	ALOX15B	CHRNA1	H2AC20	ZFPM2-AS1
6	AC005943.2	LINC00486	LINC02015	TOPAZ1	GRXCR2	POLR2J3
7	CTB-131B5.5	RP11-358F13.1	MTRNR2L10	U95743.1	SAA1	RP11-692P14.1
8	PARD6G-AS1	LCOR	MTRNR2L6	AC079135.1	KRT19	LINC01278
9	MIOX	NRG3	H4C14	AC144450.1	RP11-114H24.7	CHASERR
10	CIART	LRRTM4	OPRPN	UPK1A-AS1	CSF3	MIR100HG
11	RP11-958J22.1	DPP10	NPPA	C5orf67	RP11-446E24.4	TPT1-AS1
12	ALB	HNRNPU	S100A7	RP11-716H6.2	RP11-1012E15.2	DLEU1
13	CTC-359D24.6	SPATA13	ADIPOQ	RP4-534N18.4	KCNJ1	EIF1B-AS1
14	HPYR1	ERBB4	RHCG	OTX1	RP11-513H8.1	PPP3R1
15	AC116366.7	IL1RAPL1	LINC02211	GACAT2	RP11-108E14.1	PSMA3-AS1
16	RP11-1035H13.3	CCDC18-AS1	FCGR1B	GDPD4	AC114763.1	OIP5-AS1
17	KIF5C-AS1	RALYL	CCR1	RP11-61O1.1	RP11-745L13.2	ZNF544
18	RP4-777O23.3	CCSER1	MTRNR2L11	ELSPBP1	SLC5A3	GARS1-DT
19	CYP39A1	NLGN1	IGLC7	B4GALNT2	ADRA2C	CCDC18-AS1
20	FAM157B	RP11-452H21.1	U2AF1L5	RP11-434D12.1	SLPI	TSTD3
21	RP11-826N14.8	TRAF3IP2-AS1	RP1-125I3.2	AC006000.5	CTC-444N24.8	ATP5PO
22	PLA2G5	IQCJ-SCHIP1	MTRNR2L3	DAOA-AS1	CXCL9	HHATL
23	GNAT1	RP1-30E17.2	LVRN	CTB-91J4.1	KCNJ16	IRF1-AS1
24	DDIT4	GRIA2	MRPS9-AS2	RP11-425A23.1	TLR8	DDX39B
25	RP11-274G22.1	ABHD15-AS1	RP4-549L20.3	RP11-431M7.2	AC058791.1	MIR3936HG
26	NPPA	RYR2	CSNK2A3	IGLV3-1	PMP2	FXYD6
27	SLN	SNTG1	RP11-505K9.1	AC068491.2	SERPINE1	NPHP3
28	NUP210L	GARS1-DT	CH507-513H4.3	RP11-90E5.1	CTA-276F8.1	ATXN7
29	RP1-111C20.5	NCAM2	SLC28A3	ERP27	STARD13-AS	AC009120.6
30	PVALB	CNTN5	FKBP5	RP11-388K12.2	CIDEC	RP11-96H17.1

Table 2: Top 30 genes identified by the Attribution-Based and DEG-Based methods across three diseases, using normal cells as the baseline.

A.6 GENE PRUNING: CLASSIFICATION METRICS

While forward attribution performs similarly to attribution importance in input prediction, it does not negatively impact baseline prediction accuracy. Since genes deemed least likely to cause disease are removed, a substantial drop in accuracy is expected, as seen in Figures 2 (A, b) and (A, c). This reasoning also extends to random gene removal. Given the hypothesis that only a small subset of genes serves as biomarkers, randomly removing genes is unlikely to cause a substantial decline in accuracy. Also, we

Our method still has several limitations. One key assumption is that all genes are independent, meaning that gene interactions are not explicitly considered. This assumption may overlook complex regulatory relationships that influence phenotypic expression and disease mechanisms. As a result, our attributions may be less reliable for diseases without well-defined biomarkers.

Another limitation is that certain diseases, such as dilated cardiomyopathy, are also shaped by environmental factors, which are inherently challenging to model. We believe this issue, combined with the independence assumption, contributes to the unexpected trends observed for both forward attribution and attribution importance in Figure 2 (A, c). Despite these limitations, attribution importance remains a reliable preliminary method for gene selection.

Additionally, we conducted an alternative experiment by removing genes with the highest importance value first, instead of the lowest to the highest (see Figure 13 and 14). We aim to demonstrate that the genes with high importance significantly impact the model's performance. In contrast to HVG and random selection, the accuracy of the model drops sharply as we remove genes with either higher attribution or forward importance. Notably, there exists a stable range of prune ratios, where further gene removal does not affect the model's performance, until the point where the model's accuracy degrades due to minimal remaining information.

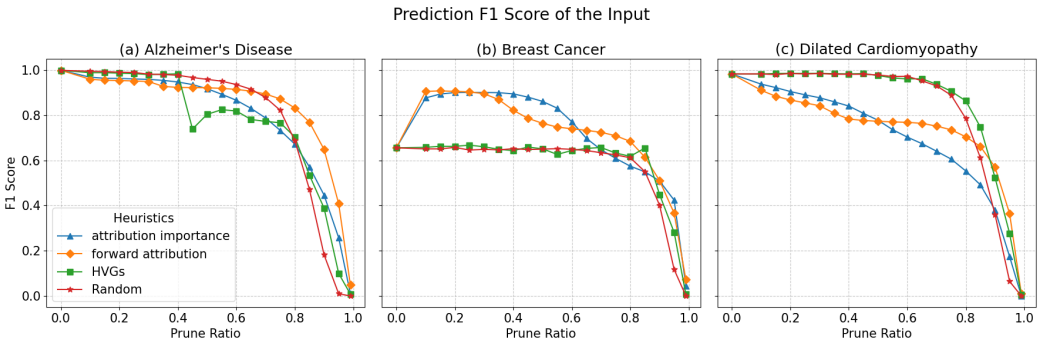


Figure 7: Prediction F1 Score of the input after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

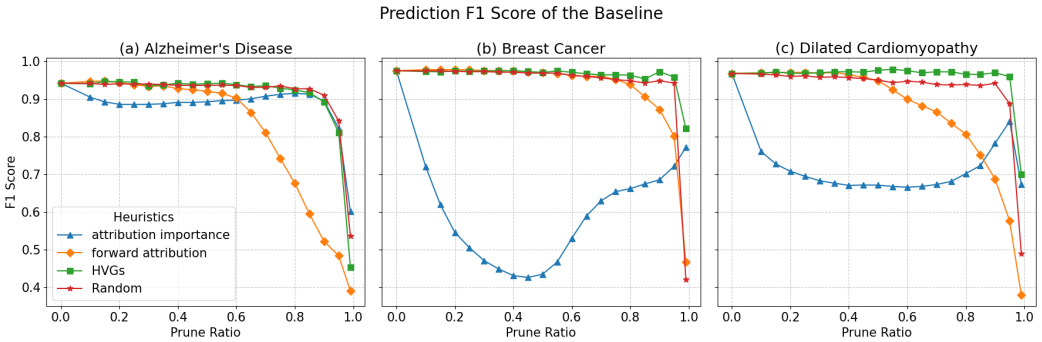


Figure 8: Prediction F1 score of the baseline after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

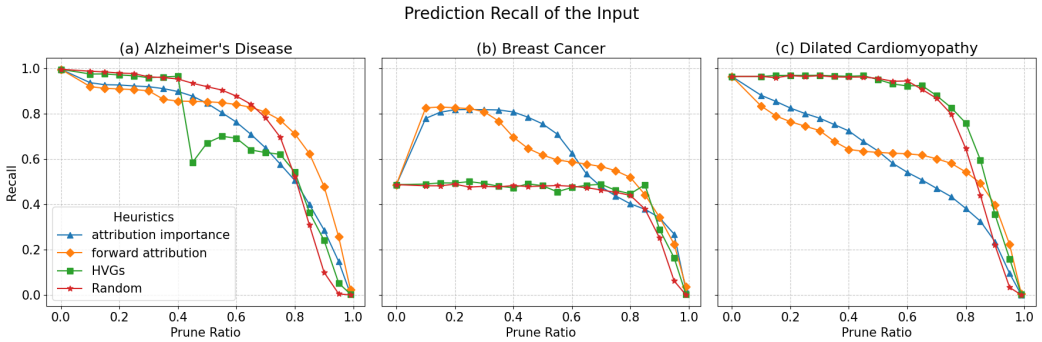


Figure 9: Prediction Recall of the input after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

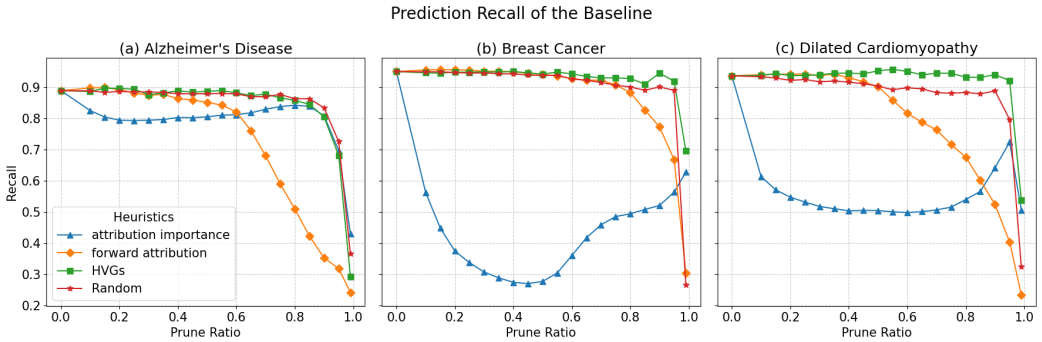


Figure 10: Prediction Recall of the baseline after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

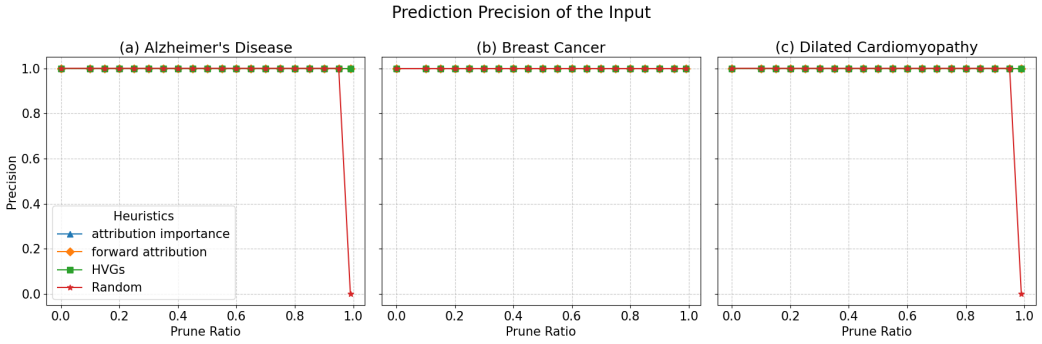


Figure 11: Prediction Precision of the input after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

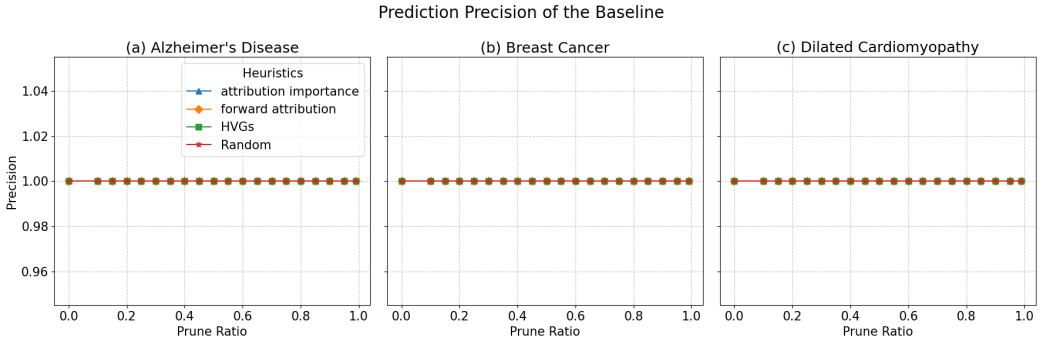


Figure 12: Prediction Precision of the baseline after gene pruning for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

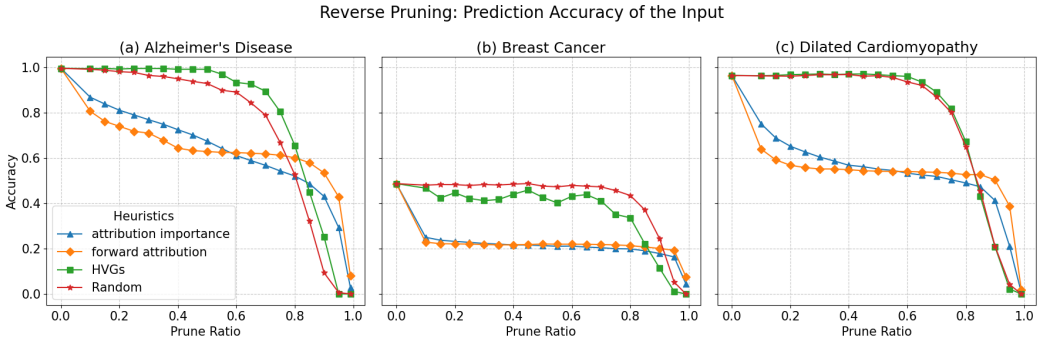


Figure 13: Prediction Accuracy of the input after reverse gene pruning (remove genes from high to low importance) for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

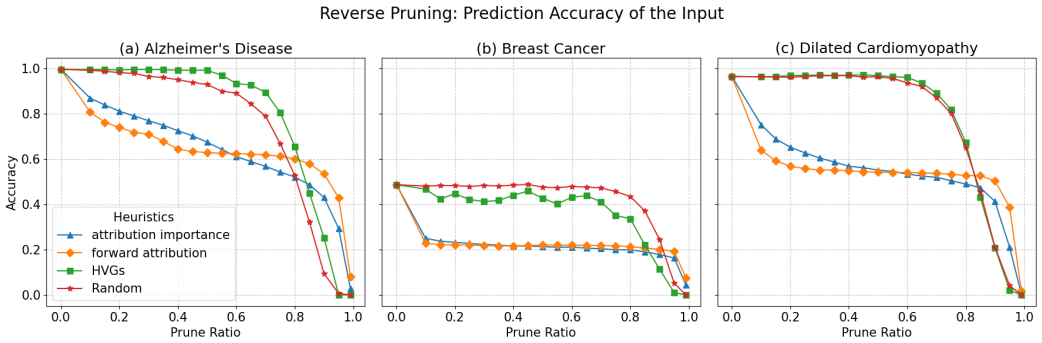


Figure 14: Prediction Accuracy of the baseline after reverse gene pruning (remove genes from high to low importance) for three diseases: (a) Alzheimer's Disease, (b) Breast Cancer, and (c) Dilated Cardiomyopathy. Different heuristics, including attribution importance, forward attribution, HVGs, and random selection, are compared across varying prune ratios.

A.7 An Alternative Method: Empty Baseline

During experimentation, we also tested an alternative method that produces an alternative set of relevant genes for each disease. This method selects an empty baseline E, defined as a model input containing only phenotypic information without genotypic data.

Formally, let $I = (p^I, g^I)$ denote an input sample, where p^I represents the phenotype and g^I the genotype. The corresponding empty baseline for I is given by $E_I = (p^I, e)$, where e is a sequence of `[PAD]` tokens of size $|g^I|$.

Unlike in Section 3.2, where attributions are computed based on the probability of the model predicting the disease, this method calculates attributions with respect to the model's prediction of the **ground truth label**. We expect this approach to yield better convergence, as the model's high accuracy suggests that its most probable predictions are strongly associated with high confidence scores.

Let f' denote this modified prediction function. We define forward attribution as $f'(I, E_I)$, and backward attribution as $f'(B, E_B)$ where $B = (p^B, g^B)$ is a normal baseline cell as defined in Section 3.2. The attribution importance using the empty baseline is then computed as:

$$\text{Attribution Importance}_{\text{empty}} = \frac{f'(I, E_I) - f'(B, E_B)}{2}.$$

The reason for the average difference is that we use the ground truth label for attribution calculation. The interpretation of $f'(I, E_I)$ remains the same as in Section 3.2, while genes with high attribution in $f'(B, E_B)$ can be interpreted as genes that the model associates with normal cells. Conversely, genes with negative attributions are likely to indicate that the model does not associate them with normal cells, suggesting that they may be potential disease biomarkers.

We choose the proposed method over this alternative because genes with low attribution in $f'(B, E_B)$ do not necessarily imply an association with the target disease. It is possible that they are linked to other diseases, leading to misinterpretation. Additionally, the concept of biomarkers is inherently associated with disease states rather than normal cells, making the interpretation of attributions in $f'(B, E_B)$ less meaningful in the context of disease marker identification.

Furthermore, using the proposed method ensures that attribution importance is derived directly from the contrast between diseased and normal states, making it more robust for feature selection. While the empty baseline method offers an interesting perspective, its reliance on normal cell associations introduces ambiguity in determining disease-specific markers. As a result, we adopt the proposed attribution method as a more reliable approach for gene selection.

Nevertheless, we find that the list of genes generated by this method is both promising and biologically reasonable. Please refer to Table 3 for the complete list of identified genes.

Table 3: Top 30 genes for Alzheimer's Disease, Breast Cancer, and Dilated Cardiomyopathy given by the Empty Baseline Method.

Rank	Alzheimer's Disease		Breast Cancer		Dilated Cardiomyopathy	
	Attribution	**DEG**	**Attribution**	**DEG**	**Attribution**	**DEG**
1	PCSK1N	ZFHX3	ALOX15B	LINC02510	NPPA	RP11-305P14.2
2	RP11-701H24.9	KCNIP4	NPPA	AC131056.5	LINC01115	SNHG14
3	CIART	CSMD1	MTRNR2L6	DRD3	NPPB	DISC2
4	SOX9	PTPRD	RP11-46H11.3	RP4-739H11.3	SAA1	RP11-115H11.1
5	IGF2BP1	SNHG14	RP11-809O17.1	CHRNA1	H2AC20	ZFPM2-AS1
6	RP4-777O23.3	LINC00486	RNASE10	TOPAZ1	RP11-380D23.1	POLR2J3
7	DDIT4	RP11-358F13.1	S100A7	U95743.1	RP11-701H24.9	RP11-692P14.1
8	C7orf61	LCOR	ZNF229	AC079135.1	AC116366.7	LINC01278
9	UCHL1	NRG3	CTA-212A2.4	AC144450.1	PARD6G-AS1	CHASERR
10	HSPB8	LRRTM4	RP11-76P2.4	UPK1A-AS1	RP11-114H24.7	MIR100HG
11	RASD1	DPP10	CSNK2A3	C5orf67	OSGIN1	TPT1-AS1
12	ZBTB16	HNRNPU	LINC02019	RP11-716H6.2	EDN1	DLEU1
13	ADM	SPATA13	NECTIN3-AS1	RP4-534N18.4	COL3A1	EIF1B-AS1
14	GPX3	ERBB4	MTRNR2L10	OTX1	AP001055.8	PPP3R1
15	CTA-384D8.34	IL1RAPL1	MTRNR2L1	GACAT2	HSPA6	PSMA3-AS1
16	FXYD6	CCDC18-AS1	RP4-549L20.3	GDPD4	LINGO1	OIP5-AS1
17	LRRC37A	RALYL	OR4D9	RP11-61O1.1	RP11-315A19.1	ZNF544
18	IL6	CCSER1	SLC28A3	ELSPBP1	KIF25	GARS1-DT
19	SAP30-DT	NLGN1	U2AF1L5	B4GALNT2	RP11-343C2.9	CCDC18-AS1
20	MAS1	RP11-452H21.1	IQCN	RP11-434D12.1	ST3GAL3	TSTD3
21	HSPB1	TRAF3IP2-AS1	RP11-861E21.2	AC006000.5	RP11-813I20.2	ATP5PO
22	LINC02381	IQCJ-SCHIP1	TRIM72	DAOA-AS1	MALAT1	HHATL
23	EFNA3	RP1-30E17.2	SST	CTB-91J4.1	RP11-219A15.1	IRF1-AS1
24	ZNNT1	GRIA2	FGF22	RP11-425A23.1	CA4	DDX39B
25	HAP1	ABHD15-AS1	IL24	RP11-431M7.2	GRXCR2	MIR3936HG
26	H2BC21	RYR2	PRANCR	IGLV3-1	LVRN	FXYD6
27	RP11-138A9.1	SNTG1	ADIPOQ	AC068491.2	AC008746.3	NPHP3
28	RP11-513M16.8	GARS1-DT	RP11-577H5.5	RP11-90E5.1	LINC02641	ATXN7
29	OLIG1	NCAM2	XXbac-BPG252P9.9	ERP27	RP11-147H23.3	AC009120.6
30	HILPDA	CNTN5	LINC02381	RP11-388K12.2	RP11-687M24.8	RP11-96H17.1