UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FRANKLIN VINNY MEDINA NUNES

# A Framework for Model Selection and Interpretation in Single-Cell Transcriptomic Data

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. Márcio Dorn

Porto Alegre
June 2026

**CIP — CATALOGING-IN-PUBLICATION**

VINNY MEDINA NUNES, FRANKLIN

A Framework for Model Selection and Interpretation in Single-Cell Transcriptomic Data / FRANKLIN VINNY MEDINA NUNES. – Porto Alegre: PPGC da UFRGS, 2026.

137 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2026. Advisor: Márcio Dorn.

I. Dorn, Márcio. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Profª. Marcia Barbosa
Vice-Reitor: Prof. Pedro Costa
Pró-Reitor de Pós-Graduação: Profª. Claudia Wasserman
Diretor do Instituto de Informática: Prof. Luciano Paschoal Gaspary
Coordenador do PPGC: Prof. Gabriel Luca Nazar
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

## ABSTRACT

Traditional transcriptomic classification approaches may obscure cell-type-specific disease signals when heterogeneous cellular populations are treated as a single analytical unit. Single-cell RNA sequencing addresses this limitation by providing cellular resolution, but its datasets are typically high-dimensional, sparse, imbalanced, and susceptible to information leakage when feature selection or model optimization is performed outside the training data.

This study presents a computational framework for single-cell transcriptomic classification that integrates feature selection, supervised learning, multiobjective model selection, and SHAP-based interpretation. The framework was evaluated on a chronic obstructive pulmonary disease single-cell RNA-sequencing dataset, restricting the analysis to male samples to reduce sex-related transcriptional confounding. After global filtering, the dataset retained 69,366 cells and 31,902 genes, with 27 cell populations eligible for downstream modeling. For each cell population, Wilcoxon-based feature selection was performed within cross-validation folds, and Logistic Regression, Support Vector Machine, Random Forest, K-nearest neighbors, and Multilayer Perceptron classifiers were evaluated using accuracy, sensitivity, specificity, and F1-score. The proposed strategy was also compared with alternative feature-selection methods, including ANOVA, Random Forest importance, and recursive feature elimination.

The results showed that predictive performance varied across cell populations, indicating that no classifier was universally optimal. In the CD4-positive, alpha-beta T-cell case study, the Multilayer Perceptron achieved the most balanced metric profile, represented the only Pareto-optimal solution, and maintained comparable performance in the final hold-out evaluation. SHAP-based interpretation highlighted influential genes, while enrichment analysis associated the selected features with translation, stress-response, and regulatory processes.



**Keywords:** ScRNA-seq; feature selection; interpretable machine learning; multiobjective optimization; chronic obstructive pulmonary disease.

## RESUMO

Abordagens tradicionais de classificação transcriptômica podem ocultar sinais de doença específicos de cada tipo celular quando populações heterogêneas são tratadas como uma única unidade analítica. O sequenciamento de RNA de célula única (scRNA-seq) reduz essa limitação ao oferecer resolução celular, mas seus dados frequentemente apresentam alta dimensionalidade, esparsidade, desequilíbrio entre classes e risco de vazamento de informação quando a seleção de características ou a otimização de modelos é realizada fora dos dados de treinamento.

Este estudo apresenta uma estrutura computacional para classificação transcriptômica de célula única que integra seleção de características, aprendizado supervisionado, seleção multiobjetivo de modelos e interpretação baseada em SHAP. A estrutura foi avaliada em um conjunto de dados de scRNA-seq associado à doença pulmonar obstrutiva crônica, com análise restrita a amostras masculinas para reduzir efeitos transcricionais relacionados ao sexo. Após a filtragem global, o conjunto de dados reteve 69.366 células e 31.902 genes, com 27 populações celulares elegíveis para modelagem. Para cada população celular, a seleção de características baseada no teste de Wilcoxon foi realizada dentro das dobras de validação cruzada, e os classificadores Regressão Logística, Máquina de Vetores de Suporte, Floresta Aleatória, K-vizinhos mais próximos e Perceptron Multicamadas foram avaliados por acurácia, sensibilidade, especificidade e escore F1. A estratégia proposta também foi comparada com ANOVA, importância da Floresta Aleatória e eliminação recursiva de características.

Os resultados mostraram que o desempenho preditivo variou entre populações celulares, indicando que nenhum classificador foi universalmente superior. No estudo de caso das células T alfa-beta CD4-positivas, o Perceptron Multicamadas apresentou o perfil de métricas mais equilibrado, constituiu a única solução Pareto-ótima e manteve desempenho comparável na avaliação final por hold-out. A interpretação com SHAP destacou genes influentes, enquanto a análise de enriquecimento relacionou as características selecionadas a processos de tradução, resposta ao estresse e regulação.



**Palavras-chave:** scRNA-seq; seleção de atributos; aprendizado de máquina; inteligência artificial interpretável; otimização multiobjetivo; doença pulmonar obstrutiva crônica.

## LIST OF FIGURES

Figure 2.1 Pareto Front. ..... 23

Figure 3.1 The Central Dogma of Molecular Biology. ..... 30

Figure 4.1 Conceptual overview of the research design. ..... 37

Figure 5.1 Global UMAP colored by cell types. ..... 63
Figure 5.2 Global UMAP colored by condition. ..... 64
Figure 5.3 Relative abundance of each annotated cell type. ..... 64
Figure 5.4 Comparative macro-metrics for the evaluated classifiers in the *CD4-positive, alpha-beta T cell* subset. ..... 71
Figure 5.5 Cross-validation confusion matrices for the final MLP in the *CD4-positive, alpha-beta T cell* subset. ..... 71
Figure 5.6 Independent-test confusion matrix and ROC curve for the final MLP in the *CD4-positive, alpha-beta T cell* subset. ..... 72
Figure 5.7 Comparison of feature-selection overlap for CD4-positive, alpha-beta T cells. ..... 72
Figure 5.8 Comparison of ranked genes and computational cost for CD4-positive, alpha-beta T cells. ..... 73
Figure 5.9 SHAP-based feature importance and expression SHAP correlation analysis for CD4-positive, alpha-beta T cells. ..... 74
Figure 5.10 Comparison of low-dimensional structure and full-space class separation across feature selection methods. ..... 82
Figure 5.11 Functional enrichment and protein-protein interaction analysis. ..... 87
Figure 5.12 Class imbalance across cell types in the dataset. ..... 88

## LIST OF TABLES

Table 5.1 Summary of global quality-control and filtering steps applied to the male-restricted dataset. ..... 65
Table 5.2 Eligibility of cell types for downstream analysis. ..... 65
Table 5.3 Hyperparameters selected in each fold for the *CD4-positive, alpha-beta T cell* subset. ..... 68
Table 5.4 Fold-wise classification performance for the *CD4-positive, alpha-beta T cell* subset. ..... 69
Table 5.5 Aggregated classification performance across folds for the *CD4-positive, alpha-beta T cell* subset. ..... 70
Table 5.6 Comparative performance of the evaluated classifiers for the *CD4-positive, alpha-beta T cell* subset. The MLP was the only Pareto-optimal model and achieved the highest individual hypervolume score. ..... 70
Table 5.7 Hyperparameters of the final MLP selected for interpretability analysis. ..... 70
Table 5.8 Final train test performance metrics obtained by the MLP model. ..... 70
Table 5.9 Top 25 genes ranked by mean absolute SHAP value, indicating their contribution to the model predictions. ..... 76
Table 5.10 Comparison of protein protein interaction (PPI) network metrics and the number of enriched Gene Ontology Biological Process (GO BP) terms for the gene sets selected by each feature-selection method. Wilcoxon produced the most connected PPI network, whereas the proposed method recovered the largest number of enriched biological process terms, indicating broader functional coverage. ..... 84
Table 5.11 Class distribution across cell types in the male dataset. The table reports the numbers of normal and disease cells, the total number of cells, and the imbalance ratio for each cell type, where higher values indicate stronger class imbalance. ..... 89
Table 5.12 Comparative predictive performance across cell types based on stratified k-fold macro F1-score for each classifier, together with the hypervolume score and the best model selected by the multiobjective criterion. ..... 91
Table 5.13 Final train test performance of the selected classifier for each cell type. The table reports the held-out accuracy, sensitivity, specificity, macro F1-score, and the final model chosen for each cell population. ..... 92
Table 5.14 Global STRING-derived network statistics across the remaining cell types and feature-selection methods. Corrected floating-point values are reported for average node degree, average local clustering coefficient, and PPI enrichment p-values. ..... 94
Table 5.15 Comparative geometric separation across cell types and feature-selection methods. For each method, the table reports the separation ratio computed in the full standardized feature space, the centroid distance, the pooled within-class dispersion, the Silhouette score, the Davies Bouldin index, and the Calinski Harabasz index. Cell-type names are shown only once per block and are abbreviated for compactness. ..... 99

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| AI | Artificial Intelligence |
| ANOVA | Analysis of Variance |
| AUC | Area Under the Curve |
| cDNA | Complementary DNA |
| COPD | Chronic Obstructive Pulmonary Disease |
| CV | Cross-Validation |
| DNA | Deoxyribonucleic Acid |
| FASTQ | Fastq Format Sequence File |
| F1-score | Harmonic Mean of Precision and Recall |
| GO | Gene Ontology |
| GO BP | Gene Ontology Biological Process |
| H5AD | Hierarchical Data Format for AnnData |
| HV | Hypervolume |
| IML | Interpretable Machine Learning |
| KNN | K-nearest Neighbors |
| LR | Logistic Regression |
| ML | Machine Learning |
| MLP | Multilayer Perceptron |
| PCA | Principal Component Analysis |
| PPI | Protein Protein Interaction |
| QC | Quality Control |
| RF | Random Forest |
| RFE | Recursive Feature Elimination |
| RNA | Ribonucleic Acid |

| | |
|---|---|
| RNA-seq | RNA Sequencing |
| ROC | Receiver Operating Characteristic |
| scRNA-seq | Single-cell RNA Sequencing |
| SHAP | SHapley Additive exPlanations |
| snRNA-seq | Single-nucleus RNA Sequencing |
| SVM | Support Vector Machine |
| UMAP | Uniform Manifold Approximation and Projection |
| UMI | Unique Molecular Identifier |
| XAI | Explainable Artificial Intelligence |

**CONTENTS**

**1 INTRODUCTION ..... 12**
**2 MACHINE LEARNING ..... 14**
**2.1 Supervised learning ..... 14**
**2.2 Unsupervised learning ..... 14**
**2.3 Reinforcement learning ..... 15**
**2.4 Semi-supervised learning ..... 15**
**2.5 Linear Models ..... 16**
2.5.1 Logistic Regression ..... 16
2.5.2 Support Vector Machines ..... 17
**2.6 Nonlinear Models ..... 17**
2.6.1 Random Forest ..... 18
2.6.2 K-nearest neighbors ..... 18
2.6.3 Multilayer Perceptron ..... 19
**2.7 Interpretable Machine Learning ..... 20**
2.7.1 Game theory ..... 21
**2.8 Multiobjective optimization ..... 22**
2.8.1 Pareto front ..... 22
2.8.2 Hypervolume ..... 23
**2.9 Feature Selection ..... 24**
2.9.1 Filter ..... 25
2.9.2 Embedded ..... 26
2.9.3 Wrapper ..... 28
**2.10 Chapter summary ..... 29**
**3 CENTRAL DOGMA OF MOLECULAR BIOLOGY ..... 30**
**3.1 Transcriptomics ..... 31**
**3.2 Single-cell sequencing ..... 32**
**3.3 Single-cell data characteristics and processing steps ..... 33**
**3.4 Chapter Summary ..... 35**
**4 METHODOLOGY ..... 36**
**4.1 Introduction ..... 36**
**4.2 Research Design ..... 36**
4.2.1 Methodological perspective ..... 39
4.2.2 Data abstraction and formal definition ..... 39
4.2.3 Design principles ..... 40
4.2.4 Justification for cell type-specific modeling ..... 40
4.2.5 Justification for leakage-aware design ..... 41
**4.3 Overview of the framework ..... 41**
4.3.1 Active matrix definition and sanitization ..... 42
4.3.2 Preprocessing mode inference ..... 43
4.3.3 Metadata validation ..... 44
4.3.4 Global quality control ..... 44
4.3.5 Normalization and logarithmic transformation ..... 45
4.3.6 Global PCA and neighborhood graph ..... 45
4.3.7 Cell type-specific processing ..... 46
4.3.8 Outer cross-validation ..... 46
4.3.9 Training-fold balancing ..... 47
4.3.10 Fold-wise feature selection ..... 48
4.3.11 Alternative feature-selection methods ..... 48

4.3.12 Classifier framework construction ..... 49
4.3.13 Bayesian hyperparameter optimization ..... 49
4.3.14 Fold-wise evaluation ..... 50
4.3.15 Pareto analysis and hypervolume ranking ..... 51
4.3.16 SHAP-based interpretation ..... 52
4.3.17 SHAP aggregation and expression-attribution correlation ..... 52
4.3.18 Integrated output construction ..... 53
**4.4 Experimental Protocol ..... 53**
4.4.1 Parameterization of the experiments ..... 53
4.4.2 Comparative analysis of class separation induced by selected feature subsets ..... 53
4.4.3 Classifier search spaces ..... 57
4.4.4 Final hold-out evaluation ..... 58
**4.5 Complexity and Implementation Considerations ..... 58**
4.5.1 Computational complexity of the framework ..... 58
4.5.2 Memory considerations ..... 59
4.5.3 Scalability considerations ..... 59
4.5.4 Determinism and reproducibility ..... 60
4.5.5 Output organization ..... 60
**4.6 Chapter Summary ..... 61**
**5 RESULTS ..... 62**
**5.1 Introduction ..... 62**
**5.2 Chronic obstructive pulmonary disease dataset ..... 62**
5.2.1 Global quality control and filtering ..... 63
5.2.2 CD4-positive alpha-beta T cells ..... 65
5.2.2.1 Preprocessing outcomes ..... 65
5.2.2.2 Comparative classification performance, multiobjective selection, and final MLP evaluation ..... 66
**5.3 Interpretability results ..... 75**
5.3.1 SHAP global importance ..... 75
5.3.2 Class-specific SHAP interpretation ..... 75
5.3.3 Correlation between SHAP values and gene expression ..... 77
**5.4 Comparative low-dimensional structure preservation ..... 78**
5.4.1 Comparative low-dimensional structure preservation and full-space class separation ..... 78
**5.5 Functional enrichment and biological validation ..... 81**
5.5.1 Functional enrichment of selected genes by each method ..... 81
5.5.2 Functional enrichment of positively disease-associated genes selected by the proposed method ..... 84
**5.6 Global comparative analysis across the remaining cell types ..... 86**
5.6.1 Coverage of the analyzed cell types ..... 86
5.6.2 Predictive performance across cell types ..... 88
5.6.3 Global STRING-based network coherence across the remaining cell types by each feature selection method ..... 92
5.6.4 Comparative low-dimensional structure class separation for each cell type ..... 98
**5.7 Chapter summary ..... 107**
**6 DISCUSSION ..... 110**
**6.1 The chronic obstructive pulmonary disease dataset as a benchmark for feature selection ..... 110**
**6.2 Linear and nonlinear machine learning models for single-cell classification ..... 111**
6.2.1 Why no single classifier dominated all cell types ..... 111
6.2.2 Which model families were more appropriate under class imbalance ..... 112

**6.3 Multiobjective model selection in imbalanced single-cell classification ..... 113**
**6.4 Interpretability of machine learning in single-cell transcriptomics ..... 115**
**6.5 Biological meaning of the enriched GO terms in CD4-positive alpha-beta T cells ..... 116**
**6.6 What the top 25 features selected by each method reveal about feature-selection behavior ..... 118**
**7 CONCLUSION ..... 120**
**REFERENCES ..... 121**
**APPENDIX A - RESUMO EXPANDIDO EM PORTUGUÊS ..... 135**

# 1 INTRODUCTION

Bioinformatics is an interdisciplinary field that brings together biology, computer science, and statistics to interpret the vast complexity of life at the molecular level (BAYAT, 2002; MORRIS, 2017). Its primary goal is to develop and apply algorithms, computational tools, and databases that enable the storage, organization, and analysis of massive volumes of biological data, transforming raw sequences into functional knowledge (BAYAT, 2002; MCCUE; MCCOY, 2017). This convergence enables researchers to manage the data deluge generated by modern high-throughput technologies and establishes an essential bridge between experimental data collection and the systemic understanding of biological phenomena (VITORINO et al., 2024; PINU et al., 2019).

Bioinformatics supports omics research by providing methods for genomic variant analysis, protein-structure prediction, and the reconstruction of molecular and metabolic networks (OULAS et al., 2019; PINU et al., 2019; ROST; SCHMITT et al., 2018; AUSLANDER et al., 2021). In silico analysis and machine-learning-based approaches help researchers formulate testable hypotheses, prioritize experiments, and relate molecular observations to biological mechanisms (AUSLANDER et al., 2021; MARTORELL-MARUGÁN et al., 2019). Machine learning supports transcriptomic analysis by helping researchers identify informative variables and build predictive models for disease classification, cell-state discrimination, and biomarker discovery (NG et al., 2023). However, transcriptomic datasets present specific analytical constraints. They usually contain more measured genes than samples or cells, and this high dimensionality interacts with technical noise and biological variability (FELTES; POLONI; DORN, 2021).

Single-cell and single-nucleus transcriptomics intensify these constraints. These technologies measure gene expression at cellular resolution and therefore expose heterogeneity that bulk assays can average out. However, single-cell data also contain sparsity, dropout events, and technical variation (HWANG; LEE; BANG, 2018), which complicate downstream analysis and increase the risk of overfitting (CUI et al., 2025). Computational strategies for these data must therefore address predictive performance, scalability, and biological interpretability together.

Feature selection addresses part of this problem by reducing dimensionality, prioritizing informative genes, lowering computational cost, and improving interpretability (BARBIERI; GRISCI; DORN, 2024). Different feature-selection strategies define relevance in different ways. Filter, embedded, and wrapper methods rank genes according

to statistical association, model-based importance, or predictive contribution, and these choices can affect both classification outcomes and the biological conclusions derived from the analysis (PUDJIHARTONO et al., 2022).

The use of machine learning in bioinformatics also requires interpretable model outputs (RIZZO et al., 2025). In biomedical research, predictive performance does not by itself identify the molecular features that drive a model decision. Shapley-value-based methods estimate feature contributions and connect model outputs to biological interpretation (PHAM; MILAZZO, 2025). Model evaluation also requires more than one metric. In transcriptomic classification, accuracy, sensitivity, specificity, and F1-score capture different aspects of performance. Multiobjective optimization methods, including Pareto optimality and hypervolume, allow candidate models to be compared across these competing criteria (HORN; BISCHL, 2016).

The main objective of this research is to develop a computational framework for transcriptomic data analysis that integrates feature selection, machine learning, interpretability, and multiobjective optimization in order to identify informative genes and build predictive models. To achieve this objective, the study: (i) examines the biological and computational foundations of transcriptomics, single-cell sequencing, and bioinformatics analysis; (ii) investigates filter, embedded, and wrapper feature-selection strategies in high-dimensional gene expression data; (iii) evaluates linear and nonlinear machine learning models for transcriptomic classification tasks; (iv) employs multiobjective optimization concepts, including Pareto front analysis and hypervolume, to guide model selection; and (v) applies interpretable machine learning methods, with emphasis on SHAP-based approaches, to quantify feature contributions and support biological interpretation.

# 2 MACHINE LEARNING

Machine learning comprises a set of computational methods that learn patterns from data and use these patterns to support prediction, classification, or structure discovery (BADILLO et al., 2020). In contrast to rule-based programming, machine learning methods infer decision rules from empirical observations, which makes them especially useful in domains characterized by complex, high-dimensional data. In bioinformatics, these properties are particularly relevant because molecular datasets often contain thousands of variables, nontrivial noise structure, and nonlinear relationships that are difficult to model with conventional statistical approaches (LIBBRECHT; NOBLE, 2015).

## 2.1 Supervised learning

Supervised learning addresses tasks in which each training observation has an associated label or target value. Classification assigns an observation to a discrete category, whereas regression predicts a continuous numerical value (GÉRON, 2022). In biomedical applications, supervised models can classify medical images, support diagnostic screening, and analyze labeled clinical or molecular datasets (ALJUAID; ANWAR, 2022). In natural language processing, the same paradigm supports tasks such as sentiment classification, where annotated text provides the target labels used during training (NAGARHALLI; VAZE; RANA, 2021).

Across these applications, labeled data define the mapping between input variables and known outcomes. The reliability of the resulting model therefore depends on annotation quality, class representation, and the correspondence between the training data and the intended prediction task.

## 2.2 Unsupervised learning

Unsupervised learning analyzes data without predefined labels or target categories. Rather than predicting a known outcome, these methods identify structure in the data, including clusters, correlations, latent dimensions, and distributional patterns (GÉRON, 2022). This exploratory role makes unsupervised learning useful in exploratory data analysis, especially when researchers need to inspect structure before defining a supervised

task (HAFEN; CRITCHLOW, 2013).

Clustering represents a common application of unsupervised learning. Algorithms such as K-means (HARTIGAN; WONG, 1979) group observations according to similarity, which allows researchers to identify gene-expression modules, candidate disease subtypes, or cell populations with related transcriptional profiles (ANG et al., 2016). Dimensionality reduction provides a second major application by projecting high-dimensional data, including RNA-sequencing matrices, into lower-dimensional representations that support visualization and interpretation (LI; VARGHESE; RESSOM, 2024).

Unsupervised methods also support anomaly detection and association-rule learning. Anomaly detection estimates a reference pattern and flags observations that deviate from it, whereas association-rule methods identify dependencies among variables (AGYEMANG, 2024; YASSINE; MOHAMED; ACHHAB, 2021).

## 2.3 Reinforcement learning

Reinforcement learning models sequential decision-making through interactions between an agent and an environment. At each step, the agent observes a state, performs an action, and receives feedback through a reward signal. The learning objective is to identify a policy that maximizes cumulative reward over time (SUTTON; BARTO, 2018). Reinforcement learning supports applications that require sequential optimization, including autonomous navigation, robotics, resource allocation, and game-playing systems. More recent applications include reinforcement learning from human feedback, which uses preference signals to adjust language-model behavior toward target response criteria (YOUNG; LEE, 2023; LAMBERT, 2025).

## 2.4 Semi-supervised learning

Semi-supervised learning combines labeled and unlabeled data. This paradigm suits settings in which labeled observations are limited but unlabeled observations are abundant. The model uses the labeled subset to estimate the task structure and then incorporates information from unlabeled observations to refine the decision boundary (KASHYAP; MISHRA, 2022). Self-training represents one common strategy: the model assigns provisional labels to high-confidence unlabeled observations and adds them to the

training set. This strategy appears in bioinformatics and medical image analysis, where manual annotation of cell types or clinical images requires domain expertise and substantial time (CHEN; LU; RAO, 2024).

## 2.5 Linear Models

Linear models relate input variables to an output through a weighted linear combination. They estimate coefficients that, together with a bias term, minimize the discrepancy between predicted and observed values (WITTEN et al., 2017). This relationship can be expressed in simplified form by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where $\beta$ represents the weights learned by the model during training, and $\epsilon$ denotes the residual error. The simplicity of this structure allows researchers not only to make predictions, but also to understand the magnitude and direction of the influence of each variable on the final outcome, a characteristic known as model transparency.

Within this category, Linear Regression (KUCHIBHOTLA et al., 2019) and Logistic Regression (CHUNG, 2020) stand out. While the former predicts continuous values, the latter, despite its name, performs classification by using the sigmoid function to map any real value to the interval [0, 1], thereby representing the probability that a sample belongs to a given class. Researchers often use linear models as baselines in complex predictive studies because they are computationally efficient and less prone to overfitting in smaller datasets (SCHULZ et al., 2020). In addition, extensions such as regularization techniques, including Lasso and Ridge, allow these models to handle high-dimensional problems by selecting the most relevant variables and improving the system's generalization ability (MELKUMOVA; SHATSKIKH, 2017).

### 2.5.1 Logistic Regression

Logistic regression models categorical outcomes by linking a linear predictor to class probabilities. Multiple logistic regression incorporates several predictors into the same model, which allows researchers to adjust for covariates and estimate how biological variables, such as age, metabolite levels, or genetic markers, relate to the probability of an

outcome (BROWN, 1998; ELKAHWAGY; KIRIACOS; MANSOUR, 2024). In single-cell analysis, the outcome can include more than two categories. Multinomial logistic regression extends the binary formulation to assign each cell to one of several cell states or lineages according to estimated class probabilities. This extension supports classification tasks in tissues that contain multiple interacting cell populations (FISCHER et al., 2024).

### 2.5.2 Support Vector Machines

Support Vector Machines (SVM) (CORTES; VAPNIK, 1995) classify observations by estimating a separating hyperplane in feature space. With a linear kernel, the algorithm maximizes the margin between the hyperplane and the closest observations from each class, known as support vectors. This margin-based formulation focuses the decision boundary on the observations that define class separation.

In bioinformatics and genomics, linear SVMs are often applied to high-dimensional data, where the number of variables, such as genes, can exceed the number of samples (YANG, 2004). Gene-expression studies have used SVMs to classify tumor subtypes and distinguish experimental groups from molecular profiles (MASTROPIETRO; FELDMANN; BAJORATH, 2023). Although nonlinear kernels can model more complex boundaries, the linear formulation offers a direct balance between interpretability, computational cost, and predictive performance when the data support approximately linear separation (ZHANG et al., 2013).

## 2.6 Nonlinear Models

Nonlinear models address relationships that a straight line or simple hyperplane cannot represent. In biological systems, gene regulation and cellular responses can involve saturation, threshold effects, and interactions among molecular components (HEGDE; NGUYEN; CHENG, 2025). Algorithms such as neural networks and kernel-based methods can model curved decision boundaries, but this flexibility increases the risk of overfitting when sample size, noise, or dimensionality constrain the learning problem. For this reason, nonlinear modeling in bioinformatics usually requires cross-validation and regularization to evaluate whether predictive patterns generalize beyond the training data (ZHOU; YAN; ZHANG, 2024).

### 2.6.1 Random Forest

Random Forest (BREIMAN, 2001) is an ensemble method that combines multiple decision trees. During training, the algorithm builds each tree from a bootstrap sample of the data and evaluates a random subset of variables at each split. For classification, the forest assigns the class by majority vote across trees; for regression, it averages tree predictions. This aggregation reduces the instability associated with individual decision trees (ALTMAN; KRZYWINSKI, 2017).

In omics analysis, Random Forest can handle high-dimensional inputs and nonlinear relationships among variables. Transcriptomic studies use the method to classify cell types or distinguish healthy and diseased tissues (POUYAN; KOSTKA, 2018; FELTES et al., 2019). Because each tree evaluates different subsets of samples and genes, the ensemble can capture interactions that linear models may miss. Random Forest also provides feature-importance measures, which makes it useful for ranking candidate genes in predictive analyses (YUAN et al., 2023).

### 2.6.2 K-nearest neighbors

K-nearest neighbors (KNN) (LAAKSONEN; OJA, ) is an instance-based supervised learning method, also described as a lazy learner (HULETT; HALL; QU, 2012). KNN stores the training set and classifies a new observation by comparing it with labeled observations at inference time. The method assumes that observations with similar feature profiles lie close to one another in a multidimensional space. Common distance metrics include Euclidean, Manhattan, and Minkowski distances, and the choice of metric depends on the data structure (ALFEILAT et al., 2019). The Euclidean distance between two points $p$ and $q$ in an $n$-dimensional space is given by:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2.1}$$

After calculating distances, the algorithm identifies the $k$ nearest neighbors and assigns the new observation to the majority class among them. The value of $k$ controls the locality of the decision rule: small values increase sensitivity to noise, whereas large values smooth local decision boundaries and can mask class-specific structure.

In computational biology, KNN has been used for missing-value imputation in gene-expression matrices, where values are estimated from genes or samples with similar expression profiles (LIEW; LAW; YAN, 2010; TROYANSKAYA et al., 2001). KNN also supports cell-type classification in single-cell studies by comparing the transcriptional profile of an unknown cell with a labeled reference atlas (COTTRELL; HOZUMI; WEI, 2024). Its main limitation is computational cost at inference time, because the method requires distance calculations against the stored reference set.

### 2.6.3 Multilayer Perceptron

The multilayer perceptron (MLP) is a feed-forward neural network architecture (GONZáLEZ-CAMACHO et al., 2016; FENG et al., 2023). Unlike the single-layer perceptron (ROSENBLATT, 1958), which can solve only linearly separable problems, an MLP contains an input layer, one or more hidden layers, and an output layer. Each layer contains processing units, or neurons, connected to units in the next layer. During training, the algorithm updates the connection weights to map input features to target outputs.

The MLP uses nonlinear activation functions, such as the Rectified Linear Unit (ReLU), sigmoid, and hyperbolic tangent (tanh), to model nonlinear relationships (LEDERER, 2021). Without these functions, stacked linear layers would collapse into a single linear transformation. Mathematically, the computation performed by an individual neuron can be expressed as:

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.2}$$

where $x_i$ represents the input variables, $w_i$ denotes the corresponding weights, $b$ is the bias term, and $\phi$ is the activation function. In biological applications, this flexibility allows the network to model nonlinear relationships among genes or proteins, including interactions that linear models may not capture.

An MLP learns through the backpropagation algorithm (RUMELHART; HINTON; WILLIAMS, 1986) combined with optimization methods such as gradient descent. During training, the model performs a forward pass, computes the discrepancy between predicted and true values through a loss function, and propagates the error backward to update the weights. The model repeats this cycle until the optimization procedure reaches a stopping criterion. Because MLPs can overfit noisy and high-dimensional data, their

use in sequencing and molecular biology datasets requires careful architectural choices and regularization strategies, such as dropout and weight decay (BORHANI et al., 2022; HUANG et al., 2024).

## 2.7 Interpretable Machine Learning

Interpretable Machine Learning (IML) studies methods that explain how model inputs contribute to model outputs (AGARWAL; DAS, 2020). Complex architectures, such as deep neural networks and ensemble models, can obscure the relationship between input variables and final predictions. Interpretability methods help researchers identify variables that influence model behavior and evaluate whether a model relies on biologically plausible signal rather than noise or spurious correlations.

Model interpretability generates responses from the decision-making function that elucidate the connections between the input and output variables of the machine learning model. In various fields, stakeholders may face challenges in enforcing interpretability, either due to legal constraints or the risk of introducing bias into decision-making processes. Nevertheless, enhancing interpretability holds significant value for users, as it fosters greater trust in the system (MILLER, 2019).

Interpretability also supports model auditing in applications where automated decisions affect biomedical inference or clinical reasoning (DOSHI-VELEZ; KIM, 2017; RUDIN, 2019). Researchers generally distinguish between intrinsic interpretability, in which the model structure itself remains understandable, and post-hoc interpretability, in which external methods explain a trained model (DU; LIU; HU, 2018; MURDOCH et al., 2019). In scientific applications, these explanations help researchers evaluate whether the model uses variables that align with domain knowledge and experimental context (MURDOCH et al., 2019; ALLEN et al., 2023).

Black-box models contain internal structures that make the path from input variables to output predictions difficult to inspect directly. Deep neural networks and ensemble models often fall into this category because they combine many parameters with nonlinear transformations. Although these models can perform well on large-scale datasets, including single-cell sequencing atlases, limited transparency can obscure biases, technical artifacts, or spurious correlations (RUDIN, 2019; MURDOCH et al., 2019).

The accuracy - interpretability trade-off describes the practical tension between flexible predictive models and models whose decisions researchers can inspect directly.

Linear regression and simple decision trees provide transparent decision rules, but they may not capture nonlinear biological relationships. In genomic medicine and molecular biology, predictive accuracy alone does not establish biological validity; researchers also need explanations that make model decisions auditable and biologically plausible (RUDIN, 2019; MARCINKEVIcS; VOGT, 2023).

Explainable artificial intelligence (XAI) provides post-hoc methods for analyzing complex models after training. Methods such as SHAP estimate each feature's contribution to a prediction and help researchers examine whether genes or biomarkers drive model decisions in biologically plausible ways. In this dissertation, this function matters because model interpretation connects classification performance with downstream biological reasoning (LUNDBERG; LEE, 2017; MURDOCH et al., 2019; RYO; RILLIG, 2022).

### 2.7.1 Game theory

Game theory examines settings in which outcomes depend on the contributions of multiple agents. Machine-learning interpretability uses cooperative game theory through Shapley values, introduced by Lloyd Shapley (SHAPLEY, 1953). In this framework, a total payoff is distributed among players according to their marginal contributions across coalitions (LUNDBERG; LEE, 2017).

In model interpretation, input variables correspond to players and the model prediction corresponds to the value to be explained. To estimate the contribution of a variable, Shapley-based methods evaluate how the prediction changes when that variable enters different feature subsets. This formulation addresses feature interdependence by averaging contributions across multiple coalitions (LUNDBERG; LEE, 2017; BORGONOVO et al., 2024).

Shapley-based explanations satisfy mathematical axioms such as efficiency and additivity. These axioms connect individual feature contributions to the difference between the current prediction and a baseline prediction. In biological modeling, this structure allows researchers to audit model decisions, compare feature attributions with domain knowledge, and identify candidate variables for downstream interpretation (LUNDBERG; LEE, 2017; RYO; RILLIG, 2022).

## 2.8 Multiobjective optimization

Multiobjective optimization addresses problems in which two or more objectives must be optimized simultaneously. In most practical settings, these objectives conflict with one another, so improving one criterion may reduce performance in another. Instead of producing a single optimal solution, multiobjective optimization identifies a set of trade-off solutions. These solutions are commonly described as Pareto-optimal when no objective can be improved without degrading at least one other objective (ZHOU et al., 2011).

This formulation is particularly relevant when classifier selection cannot rely on a single performance metric, since different measures summarize different aspects of the confusion matrix (SOKOLOVA; LAPALME, 2009). In transcriptomic classification, accuracy, sensitivity, specificity, and F1-score may favor different classifiers, especially when class imbalance affects the interpretation of global accuracy and class-specific errors (HE; GARCIA, 2009; SUN; WONG; KAMEL, 2009). A multiobjective formulation allows these criteria to be evaluated jointly as competing objectives rather than collapsed into a single weighted score (MARLER; ARORA, 2004).

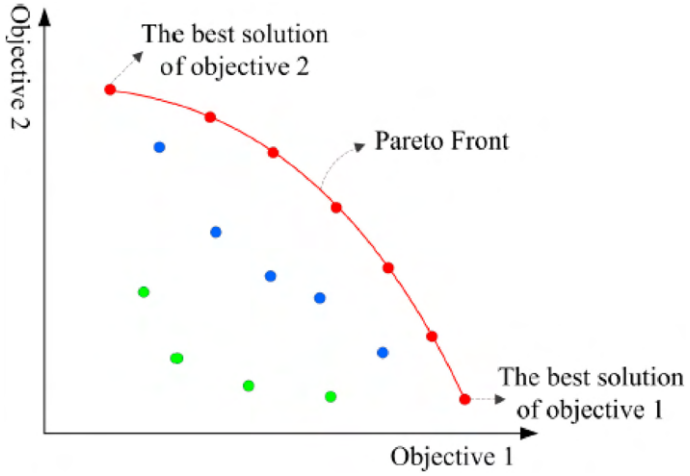
Multiobjective evolutionary algorithms (MOEAs) provide a suitable search framework for this type of problem because they evaluate a population of candidate solutions at each iteration. This population-based strategy allows MOEAs to approximate multiple non-dominated solutions in a single run. In contrast, many scalarization-based approaches require repeated optimizations with different weights to recover distinct regions of the Pareto front (ZITZLER; THIELE, 1999; ZHOU et al., 2011). Effective MOEAs must therefore satisfy two complementary requirements: they must move candidate solutions toward the Pareto front and preserve diversity across the front. Diversity remains important because it provides decision-makers with alternative trade-offs before selecting a final solution based on methodological or domain-specific criteria (SHUKLA; DEB, 2007).

### 2.8.1 Pareto front

The Pareto front is defined mathematically in terms of dominance. A solution $x^{(1)}$ dominates another solution $x^{(2)}$ if two conditions are met: first, $x^{(1)}$ is not worse than $x^{(2)}$ in all objectives; second, $x^{(1)}$ is strictly better than $x^{(2)}$ in at least one objective (DEB,

2011). In the entire search space, the points that no other point dominates form the Pareto front, representing the theoretical limit of system efficiency. At these points, optimization reaches a balance where improving one objective requires sacrificing at least one other objective (DEB, 2011). The figure 2.1 illustrates that the solutions on the Pareto front line exhibit different combinations of tradeoffs among the multiple conflict objectives.

Figure 2.1: A Pareto-optimal solution set or Pareto front from (YANG et al., 2018).




A property of the Pareto front is the distinction it draws between decision variable space and objective space. The decision variable space contains the problem's technical configurations, and the objective space provides a multidimensional representation of outcomes for each objective. The trade-off structure inherent to the Pareto front enhances the value of multiobjective optimization by revealing the conflicts among objectives (DEB, 2011).

The Pareto front supports downstream decision-making by exposing alternative trade-offs among objectives. This representation reduces dependence on arbitrary scalarization weights, which can favor one region of the objective space while ignoring other viable solutions. MOEAs therefore aim to approximate the Pareto front while preserving diversity among non-dominated solutions. A broad view of the front allows specialists to include additional criteria when selecting a final solution (COELLO, 2011; ZHOU et al., 2011).

### 2.8.2 Hypervolume

The hypervolume (HV) metric, or dominated space measure, is used in multiobjective optimization to capture both convergence and diversity in a single scalar value.

It quantifies the volume of the region in objective space dominated by a set of non-dominated solutions, bounded by a predefined reference point(GUERREIRO et al., 2021). In scenarios with two objectives, such as minimizing cost while maximizing torque, the hypervolume is the area enclosed by the union of the rectangles formed by each solution on the Pareto front approximation and the reference point (AUGER et al., 2012; ZITZLER; BROCKHOFF; THIELE, 2007).

In Pareto front analysis, a higher HV value usually indicates that the solution set dominates a larger region of objective space. This increase can reflect better convergence toward the Pareto front, broader coverage of non-dominated solutions, or both. For this reason, researchers use the hypervolume indicator as a quality measure in evolutionary multiobjective optimization (GUERREIRO et al., 2021; BEUME et al., 2009; AUDET et al., 2021).

However, despite its advantages as a unary indicator, there are certain limitations. The HV value depends on the choice of reference point, and different reference points can affect algorithm comparisons. Additionally, the computational cost of precise hypervolume calculations increases significantly with dimensionality, making them more challenging to use in many-objective contexts. Nonetheless, due to its theoretical foundation and its dual ability to capture both convergence and diversity, hypervolume remains one of the primary metrics for assessing the progression of populations toward the Pareto-optimal front (ZITZLER; BROCKHOFF; THIELE, 2007; AUGER et al., 2012).

## 2.9 Feature Selection

Dimensionality reduction, particularly feature selection, actively identifies a subset of variables that retain the most relevant information for a specific downstream task while discarding noisy, redundant, or irrelevant features (GUYON; ELISSEEFF, 2003). When dealing with high-dimensional data, this approach enhances predictive performance, reduces computational costs, and improves interpretability by concentrating the analysis on features more likely to reflect the underlying data distribution. These benefits are especially crucial in transcriptomics data, where the number of measured genes often exceeds the number of samples or labeled cells. In these cases, factors such as sparsity, technical noise, and feature redundancy can significantly impact classification, clustering, and marker discovery (LUECKEN; THEIS, 2019; TOWNES et al., 2019).

Researchers categorize feature selection methods into three primary types: filter,

embedded, and wrapper methods. Filter methods independently evaluate each feature without considering the learning algorithm, while embedded methods select features during model training. In contrast, wrapper methods assess candidate feature subsets based on the predictive performance of a specific model (KOHAVI; JOHN, 1997). This classification remains significant in bioinformatics, as each type strikes a different balance between computational efficiency, model awareness, and the ability to capture feature interactions.

### 2.9.1 Filter

Filter methods rank features using statistical, correlation-based, or information-theoretic criteria before training models. Because these methods do not rely on the final classifier, they scale efficiently to high-dimensional omics datasets and provide an effective strategy for dimensionality reduction during preprocessing (GUYON; ELISSEEFF, 2003). However, filter methods typically evaluate each feature independently, potentially missing important feature interactions.

One widely used parametric filter is one-way analysis of variance (ANOVA). For a given gene $j$, ANOVA compares the variance between groups to the variance within groups. For $K$ groups and $N$ total observations, you can express the ANOVA statistic as follows:

$$F_j = \frac{\sum_{k=1}^{K} n_k \left(\bar{x}_{jk} - \bar{x}_j\right)^2 / (K-1)}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_{ijk} - \bar{x}_{jk}\right)^2 / (N-K)}, \tag{2.3}$$

Here, $x_{ijk}$ denotes for example the expression of gene $j$ in sample $i$ from group $k$, $\bar{x}_{jk}$ is the mean expression of gene $j$ in group $k$, and $\bar{x}_j$ is the global mean expression of gene $j$. Large $F_j$ values indicate that the gene discriminates among groups more strongly than expected from within-group variation alone. In transcriptomics, ANOVA offers a critical framework for identifying differentially expressed genes in microarray experiments, especially in studies involving multiple biological conditions or factorial designs (KERR; MARTIN; CHURCHILL, 2000; PAVLIDIS, 2003). This same rationale applies to labeled single-cell datasets, where researchers can rank genes based on their ability to separate predefined cell types or cellular states.

A commonly used nonparametric filter is the Wilcoxon rank-sum test, which compares two groups without assuming normality (WILCOXON, 1945). For a given gene $j$, this method pools the observations from both groups, assigns ranks to all observations, and computes the rank sum for one group:

$$W_j = \sum_{i=1}^{n_1} R_i, \tag{2.4}$$

In this equation, $n_1$ is the number of observations in the first group, and $R_i$ represents the rank of the $i$-th observation in that group. The equivalent Mann–Whitney statistic is:

$$U_j = W_j - \frac{n_1(n_1+1)}{2}. \tag{2.5}$$

Extreme values of $W_j$ or $U_j$ indicate substantial differences in the expression distribution of a gene between groups. This rank-based strategy proves particularly useful in bioinformatics due to the often skewed, heteroscedastic, outlier-prone, and zero-inflated nature of expression data. In single-cell data analysis, the Wilcoxon rank-sum test are used for marker-gene identification and differential expression analysis across cell populations (SONESON; ROBINSON, 2018). Recent benchmarking studies have revealed that Wilcoxon-based ranking remains highly competitive for identifying informative marker genes in single-cell data (PULLIN; MCCARTHY, 2024).

Overall, filter methods are appealing to bioinformatics researchers because they are fast, transparent, and easy to interpret. In bulk transcriptomics, these methods are used to rank candidate biomarkers before classification or pathway analysis. In single-cell transcriptomics, they facilitate marker discovery, cell-type annotation, and gene prioritization for subsequent dimensionality reduction and machine learning applications (PAVLIDIS, 2003; LUECKEN; THEIS, 2019).

### 2.9.2 Embedded

Embedded methods perform feature selection during model training, these methods estimate feature relevance during classifier optimization. As a result, embedded methods are usually more model-aware than filters and more computationally efficient than wrappers (GUYON; ELISSEEFF, 2003).

A random forest constructs an ensemble of decision trees and aggregates their pre-

dictions (BREIMAN, 2001). In classification, one common measure of feature relevance is the total impurity reduction a feature achieves across the forest. For a node $t$, the Gini impurity is:

$$G(t) = 1 - \sum_{c=1}^{C} p(c \mid t)^2, \tag{2.6}$$

where $p(c \mid t)$ denotes the proportion of observations from class $c$ at node $t$. If gene $j$ is used to split node $t$, the impurity reduction is

$$\Delta G_{j,t} = G(t) - p_L G(t_L) - p_R G(t_R), \tag{2.7}$$

where $t_L$ and $t_R$ are the left and right child nodes, and $p_L$ and $p_R$ are the proportions of observations assigned to them. The overall importance of gene $j$ can then be estimated by aggregating the impurity decrease across all relevant nodes and trees:

$$I_j = \frac{1}{T} \sum_{\tau=1}^{T} \sum_{t \in S_{j,\tau}} p(t)\, \Delta G_{j,t}, \tag{2.8}$$

where $T$ is the number of trees, $S_{j,\tau}$ is the set of nodes in tree $\tau$ split by gene $j$, and $p(t)$ weights the contribution of node $t$ by the proportion of samples it contains. Random forests can also compute permutation importance, which measures the increase in prediction error after randomly permuting the values of a feature (BREIMAN, 2001).

In transcriptomics, random forest importance measures are used to reduce data dimensionality, capture nonlinear relationships, and naturally provide rankings of informative genes. Early studies demonstrated that random forests could support both classification and gene selection in microarray datasets, including multiclass problems (DíAZ-URIARTE; ANDRéS, 2006). In single-cell analysis, researchers have employed random-forest-based strategies to select genes that enhance clustering, cell-type discrimination, and biological interpretability. For instance, RFCell utilizes random forest importance for gene selection in scRNA-seq clustering, while CellBRF combines class balancing with random forest-based ranking to improve feature selection for single-cell data (ZHAO et al., 2021; XU et al., 2023). These applications demonstrate that embedded methods can identify genes based on their direct contributions to the predictive task rather than on their marginal statistical significance alone.

### 2.9.3 Wrapper

Wrapper methods evaluate feature subsets by repeatedly training and validating a predictive model. Unlike filter methods, wrappers assess the quality of a feature subset based on the performance of a specific learning algorithm (KOHAVI; JOHN, 1997; GUYON; ELISSEEFF, 2003). This model-dependent strategy often yields highly predictive subsets, but it increases computational costs due to the need for multiple retrainings.

One of the most influential wrapper methods is Recursive Feature Elimination (RFE) (GUYON et al., 2002). RFE starts with the full feature set, fits a predictive model, and ranks features using a model-derived importance score. It removes the least informative feature or group of features and repeats the process until reaching the desired subset size (GUYON et al., 2002). In the classical support vector machine recursive feature elimination (SVM-RFE) framework, the ranking criterion for feature $j$ is defined as

$$r_j = w_j^2, \tag{2.9}$$

where $w_j$ represents the weight assigned to feature $j$ by a linear support vector machine. After ranking all features, the algorithm eliminates the feature with the smallest score:

$$S^{(t+1)} = S^{(t)} \setminus \arg\min_{j \in S^{(t)}} r_j, \tag{2.10}$$

where $S^{(t)}$ denotes the feature subset at iteration $t$. This backward elimination strategy progressively refines the feature set, preserving the variables that strongly contribute to the decision boundary.

In bioinformatics, RFE is a seminal wrapper-based method for dimensionality reduction that selects genes based on their predictive utility rather than isolated univariate significance(GUYON et al., 2002; SAEYS; INZA; nAGA, 2007). The original SVM-RFE study demonstrated strong performance in cancer classification with gene-expression data, establishing wrapper-based selection as an important strategy in transcriptomics (GUYON et al., 2002). More recent omics studies have built on this logic by using dynamic recursive feature elimination, which adjusts the elimination rate across iterations to reduce computational burden while maintaining predictive performance (BENJAMIN; KATIPALLI; PAQUOLA, 2023). In single-cell analysis, wrapper-style approaches have also facilitated the identification of compact, interpretable gene sets that differentiate cell

identities and biological states (PARK et al., 2020). The stochastic behavior of RFE results from its reliance on a predictive model with inherent randomness in its training process, stemming from factors such as initial parameter settings, random data sampling, or how certain algorithms handle ties in feature importance.

Filter, embedded, and wrapper methods address different practical goals in transcriptomics and single-cell analysis. Filter methods prioritize statistical simplicity and computational speed, embedded methods combine feature ranking with model training, and wrapper methods evaluate feature subsets through repeated model fitting. These distinctions matter in omics studies because each method defines gene relevance differently and can therefore select different feature sets.

## 2.10 Chapter summary

This chapter reviewed machine learning concepts used in computational biology and bioinformatics. It presented supervised, unsupervised, reinforcement, and semi-supervised learning as distinct analytical paradigms and related them to classification, regression, clustering, dimensionality reduction, anomaly detection, and sequential decision-making. The chapter then described linear and nonlinear models, with emphasis on logistic regression, support vector machines, random forests, k-nearest neighbors, and multilayer perceptrons. These models give bioinformatics researchers different ways to analyze molecular data, from transparent linear decision rules to nonlinear architectures that can represent interactions among genes, cells, and biological states. The discussion also connected interpretability to predictive modeling, because model explanations help researchers relate computational predictions to biological hypotheses.

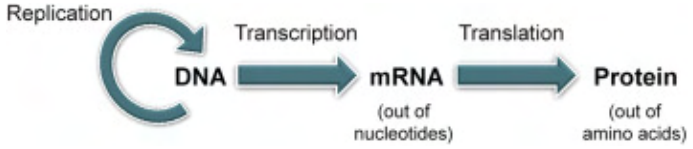
The chapter also introduced interpretable machine learning, game-theoretic attribution, multiobjective optimization, Pareto-front analysis, hypervolume, and feature selection as tools for model evaluation and biological interpretation. These topics provide the conceptual basis for understanding how predictive performance, transparency, feature relevance, and trade-off analysis guide machine learning workflows in high-dimensional omics studies.

The next chapter, *Central Dogma of Molecular Biology*, provides the molecular context for transcriptomic data. It explains how genetic information flows from DNA to RNA and protein and connects this biological framework to gene-expression measurements and the molecular features used in machine learning models.

# 3 CENTRAL DOGMA OF MOLECULAR BIOLOGY

Francis Crick proposed the Central Dogma of Molecular Biology in 1958 and later clarified its formulation in 1970 (CRICK, 1970; KOONIN, 2015). In its classical form, the dogma describes the transfer of genetic information from nucleic acids to proteins: DNA stores genetic information, RNA acts as an intermediate, and proteins perform structural and catalytic functions (CRICK, 1970; GUO; PRICE, 2014). Figure 3.1 illustrates the flow DNA $\rightarrow$ RNA $\rightarrow$ Protein, which provides the conceptual basis for relating gene expression to cellular function (CRICK, 1970; LOUTEN, 2016).

Figure 3.1: The Central Dogma of Molecular Biology from (LOUTEN, 2016).




The steps of the dogma involves three major processes: replication, transcription, and translation (LOUTEN, 2016; GUO; PRICE, 2014). During replication, cells copy DNA to preserve and transmit the genome across cellular generations (LOUTEN, 2016). During transcription, the information encoded in a gene is transcribed into a messenger RNA (mRNA) molecule that serves as a mobile template for gene expression (GUO; PRICE, 2014; ALBERTS et al., 2002). During translation, the ribosome interprets the nucleotide sequence of the mRNA and converts it into a chain of amino acids according to the genetic code, thereby synthesizing the protein specified by the original genetic instruction (ALBERTS et al., 2002; GUO; PRICE, 2014). Overall, these processes form the canonical pathway by which stored genetic information is expressed as biological function (CRICK, 1970; KOONIN, 2015).

Although the original concept shaped modern biology, subsequent discoveries revealed important exceptions and additional regulatory layers that made the system far more dynamic than initially envisioned. The discovery of reverse transcription in retroviruses demonstrated that genetic information can also flow from RNA to DNA, thereby expanding the classical interpretation of Crick's model (CRICK, 1970; COFFIN, 2021). In addition, the study of RNA viruses established that RNA can serve not only as an intermediate but also as a direct template for genome replication through RNA-dependent RNA synthesis (RAMPERSAD; TENNANT, 2018; NAGY; POGANY, 2011).

Further advances in epigenetics and non-coding RNA biology showed that gene-expression regulation also shapes phenotype. Epigenetic mechanisms modulate tran-

scriptional potential without altering the underlying DNA sequence, whereas non-coding RNAs participate in chromatin remodeling and in transcriptional and post-transcriptional regulation (KAIKKONEN; LAM; GLASS, 2011; HASELTINE, 2024; MATTICK et al., 2023). Bioinformatics and transcriptomic analyses therefore need to account for regulatory layers when linking molecular variation to phenotypic effects (HASELTINE, 2024; MATTICK et al., 2023).

## 3.1 Transcriptomics

Transcriptomics is the study of the complete set of RNA transcripts, known as the transcriptome, produced by the genome in a specific cell, tissue, or organism under defined conditions (HRDLIcKOVÁ; TOLOUE; TIAN, 2017). Whereas the genome is largely static, the transcriptome is highly dynamic and changes according to developmental stage, environmental stimuli, and pathological state (WANG; GERSTEIN; SNYDER, 2009). By capturing this snapshot of gene activity, transcriptomics allows researchers to determine which genes are actively expressed and to what extent, thereby providing a critical link between the information encoded in DNA and the resulting functional phenotype (OZSOLAK; MILOS, 2011).

Technical advances in this field were marked by the transition from microarrays to next-generation RNA sequencing (RNA-seq), which transformed both the precision and the scale of transcriptome analysis. RNA-seq not only enables sensitive quantification of known transcripts but also supports the discovery of novel splice isoforms, non-coding RNAs, and sequence variants present in transcribed regions (WANG; GERSTEIN; SNYDER, 2009; HRDLIcKOVÁ; TOLOUE; TIAN, 2017). At a finer resolution, technologies such as single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA sequencing (snRNA-seq) reveal cellular heterogeneity within complex tissues, enabling the identification of rare cell populations and differentiation trajectories that would remain hidden in bulk tissue analyses (HWANG; LEE; BANG, 2018; JOVIC et al., 2022; HABIB et al., 2016; KIM; KIM; KIM, 2023).

In addition to its descriptive function, transcriptomics serves as a powerful tool for systems biology and precision medicine (HASIN; SELDIN; LUSIS, 2017; OULAS et al., 2019). It enables the construction of gene co-expression networks and the identification of dysregulated pathways in diseases such as cancer, cardiovascular disorders, and neurodegenerative conditions (LANGFELDER; HORVATH, 2008; DAM et al., 2018;

MUBARAK et al., 2022). By comparing the transcriptional profiles of distinct biological states, researchers can identify diagnostic biomarkers and candidate therapeutic targets, thereby translating computational findings into clinically relevant insights (CHAKRABORTY et al., 2024; CHEN et al., 2023). Thus, transcriptomics remains at the forefront of genomic research by providing a detailed view of the molecular programs that sustain cellular and organismal function (WANG; GERSTEIN; SNYDER, 2009; OZSOLAK; MILOS, 2011).

## 3.2 Single-cell sequencing

Single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA sequencing (snRNA-seq) represent the current frontier of biological resolution by enabling the analysis of cellular heterogeneity in complex tissues (HWANG; LEE; BANG, 2018; JOVIC et al., 2022). Whereas scRNA-seq typically requires intact and viable cells to capture both cytoplasmic and nuclear RNA, snRNA-seq profiles RNA from isolated nuclei and is therefore especially useful for tissues that are difficult to dissociate, highly fibrotic, or available only as frozen material (HABIB et al., 2017; PIMPALWAR et al., 2021). This feature makes snRNA-seq particularly valuable for organs such as the heart and brain, where cell dissociation can bias recovery and compromise cell integrity, while nuclear material often remains preserved (PIMPALWAR et al., 2021; KIM; KIM; KIM, 2023).

The generation of these data relies predominantly on droplet-based microfluidic systems, including widely used commercial platforms such as 10x Genomics Chromium (ZHENG et al., 2017; CHOW et al., 2024). In these workflows, dissociated cells or nuclei are suspended in solution and co-encapsulated in oil droplets together with barcoded gel beads. Each bead carries oligonucleotides containing a poly(dT) capture sequence for polyadenylated RNA, a cell barcode that identifies the cell or nucleus of origin, and a unique molecular identifier (UMI) that enables molecule-level counting (PETUKHOV et al., 2018). Within each droplet, RNA molecules are reverse-transcribed into barcoded cDNA, which allows pooled sequencing while preserving the identity of the original cell and transcript molecule (ZHANG et al., 2019).

After sequencing on high-throughput Illumina platforms, fluorescence images are processed into base calls and represented as sequencing reads in FASTQ format. In droplet-based single-cell RNA-seq workflows, bioinformatics pipelines align reads to a reference genome or transcriptome, assign reads to cells using cell barcodes, collapse

transcript observations using unique molecular identifiers (UMIs), and generate a gene-by-cell count matrix. Each matrix entry represents the number of detected transcript molecules for a given gene in a given cell. This count matrix provides the quantitative basis for downstream analyses such as quality control, normalization, clustering, trajectory inference, and differential expression testing (COCK et al., 2010; ZHENG et al., 2017; LUECKEN; THEIS, 2019).

### 3.3 Single-cell data characteristics and processing steps

Single-cell sequencing data are primarily organized as a digital count matrix in which genes define one dimension and individual cells define the other, making the count table the basic input structure for most scRNA-seq analyses (AHLMANN-ELTZE; HUBER, 2023; HWANG; LEE; BANG, 2018). These data are often stored in efficient file formats such as H5AD, which support the AnnData framework for annotated matrix-based data and allow researchers to keep the core expression matrix together with cell-level metadata and gene-level annotations in a single compact object (WOLF; ANGERER; THEIS, 2018). The AnnData structure has become a central data container in modern bioinformatics frameworks because it supports scalable computation, standardized metadata handling, and rapid access to large single-cell datasets (WOLF; ANGERER; THEIS, 2018; VIRSHUP et al., 2021).

One of the most striking and analytically challenging properties of single-cell RNA sequencing data is their extreme sparsity, with a large proportion of matrix entries equal to zero (JIANG et al., 2022; KIM; ZHOU; CHEN, 2020). These zeros arise from a combination of biological factors, such as true absence or low abundance of gene expression, and technical effects, including incomplete capture and amplification inefficiencies that are often described as dropouts (JIANG et al., 2022; KIM; ZHOU; CHEN, 2020). As a result, a statistical modeling is essential to avoid confounding technical noise with genuine biological signal. In practice, negative binomial models are widely used for UMI-based single-cell data, whereas the use of zero-inflated models remains context-dependent and has been the subject of substantial debate, particularly because droplet-based UMI data are often adequately described without explicit zero inflation (SVENSSON, 2020; VIETH et al., 2019).

In addition to sparsity, single-cell transcriptomic data present a high-dimensional structure that complicates analysis and interpretation (HE et al., 2023; IMOTO et al.,

2022). Because the human genome contains on the order of tens of thousands of genes, each cell or nucleus can be represented as a point in a very high-dimensional space. However, biologically meaningful variation typically lies on a much lower-dimensional manifold defined by coordinated gene programs rather than by all genes independently (LUECKEN; THEIS, 2019). For this reason, the analysis of these data depends heavily on dimensionality reduction methods such as principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), which project complex transcriptional variation into an interpretable low-dimensional space and facilitate the visualization of cellular clusters and developmental trajectories (BECHT et al., 2019; CHAPPELL; RUSSELL; VOET, 2021).

These data are commonly processed with the Scanpy ecosystem, a Python library designed for scalable and efficient analysis of large single-cell datasets (WOLF; ANGERER; THEIS, 2018). A typical workflow begins with quality control (QC), in which researchers filter nuclei or cells with low gene counts, which may indicate empty droplets or poor-quality captures, and with high proportions of mitochondrial transcripts, which are often used as indicators of stressed or damaged cells (LUECKEN; THEIS, 2019). Scanpy supports this exploratory process through standard visualizations, including violin plots and scatter plots, that help define dataset-specific filtering thresholds before downstream analysis (BALZER et al., 2021).

After filtering, the data undergo normalization and logarithmic transformation. Because each cell or nucleus can exhibit a different sequencing depth for technical reasons, normalization adjusts counts to make observations more comparable across cells, often by rescaling each profile to a common library size before log-transformation (LUECKEN; THEIS, 2019; HEUMOS et al., 2023). Next, highly variable genes (HVGs) can be identified so that subsequent analyses can focus on the genes that contribute most to biological heterogeneity, while genes with relatively stable expression across cells contribute less to distinguishing between cell populations (STUART et al., 2019).

The final stage of basic processing involves clustering and annotation. Graph-based community detection algorithms, such as Louvain and Leiden, group cells according to the similarity of their transcriptional profiles after construction of a neighborhood graph (BLONDEL et al., 2008; TRAAG; WALTMAN; ECK, 2019). Researchers typically visualize these clusters with UMAP, where transcriptionally similar cells appear close together in a low-dimensional embedding (BECHT et al., 2019). They can then perform differential expression analysis to identify marker genes for each cluster, which

supports biological annotation, such as assigning clusters to cardiomyocytes, fibroblasts, or endothelial cells, and provides a starting point for deeper investigation of experimental effects (SONESON; ROBINSON, 2021; HEUMOS et al., 2023).

## 3.4 Chapter Summary

This chapter established the relationships among genetic information, gene expression, and transcriptomic analysis. It reviewed the Central Dogma of Molecular Biology, in which DNA stores genetic information, RNA acts as a molecular intermediate, and proteins perform structural and catalytic functions. It also described replication, transcription, and translation as linked processes that support genome maintenance, RNA synthesis, and protein production.

The chapter then extended this view by discussing reverse transcription, RNA-dependent RNA synthesis, epigenetic regulation, and non-coding RNA activity. These mechanisms show that gene expression depends not only on nucleotide sequence but also on regulatory processes that modulate transcriptional output and cellular phenotype.

The second part introduced transcriptomics and single-cell sequencing. It defined the transcriptome as a measurable representation of gene activity and explained why RNA-seq, scRNA-seq, and snRNA-seq support the analysis of biological variation across tissues, cell types, and disease states. The chapter also described droplet-based sequencing workflows, barcode-based cell identification, UMI-based transcript counting, and gene-by-cell count matrices as the quantitative basis for downstream analysis.

Finally, the chapter summarized the main analytical characteristics of single-cell data, including sparsity, high dimensionality, quality control, normalization, highly variable gene selection, dimensionality reduction, clustering, and annotation. The next chapter, *Methodology*, builds on this foundation to describe the computational framework used in this study, including preprocessing, feature selection, machine learning model training, multiobjective model selection, and interpretability analysis.

# 4 METHODOLOGY

## 4.1 Introduction

This chapter presents the methodological framework developed for the computational analysis of transcriptomic data. The methodology was organized around four requirements: reproducible processing of sparse expression matrices, prevention of information leakage, biological interpretability, and systematic comparison of feature-selection and classification strategies.

To satisfy these requirements, the framework was organized into four major components: *Research Design*, *Algorithmic framework*, *Experimental Protocol*, and *Complexity and Implementation Considerations*. The chapter first defines the analytical design and then describes the computational framework, experimental protocol, and implementation constraints.

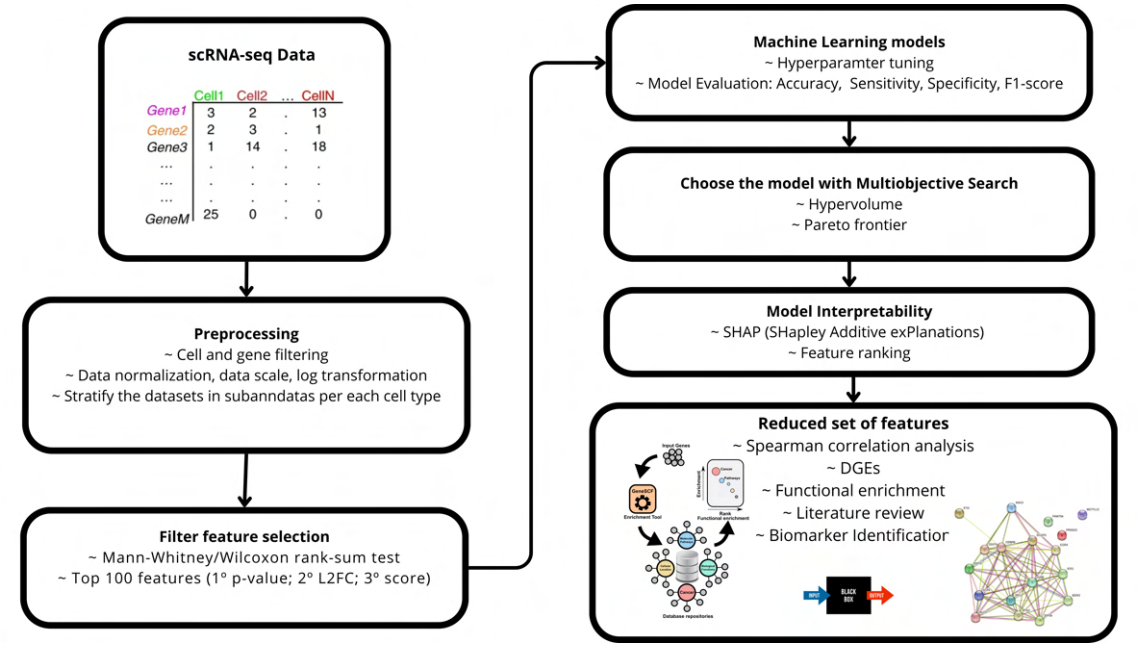
## 4.2 Research Design

Figure 4.1 presents the conceptual organization of the methodological framework adopted in this study. The framework was designed as an end-to-end analytical system for scRNA-seq classification, feature prioritization, and biological interpretation. Rather than treating preprocessing, learning, and interpretation as isolated procedures, the proposed design connects these stages through a leakage-aware, cell-type-specific analytical structure.

The framework begins with an annotated scRNA-seq dataset and an initial preprocessing stage. At this stage, the framework performs dataset sanitization, metadata validation, and preprocessing according to the numerical characteristics of the input matrix. When required, this stage includes quality control, filtering of cells and genes, normalization, logarithmic transformation, and low-dimensional embedding for diagnostic inspection. The design also incorporates stratification by cell type, since the supervised learning problem is defined separately within each cellular population. This decision reduces biological heterogeneity within each classification task and improves the interpretability of the selected transcriptomic signatures.

After preprocessing, the framework applies a filter-based feature selection strategy within each training fold. In the proposed method, genes are ranked using a Wilcoxon-

Figure 4.1: The framework begins with an annotated scRNA-seq dataset and proceeds through preprocessing, including data sanitization, quality control, filtering, normalization, transformation, and cell type-specific stratification when required. Within each cell type-specific learning problem, the proposed method performs fold-wise filter-based feature selection using Wilcoxon-based ranking to retain the top 100 genes. These features are used to train and tune machine learning classifiers under stratified 5-fold cross-validation, with performance assessed by accuracy, sensitivity, specificity, and F1-score. Model selection is then conducted through multiobjective analysis using Pareto frontier and hypervolume. Finally, SHAP-based interpretation is applied to quantify feature contributions and generate a reduced set of biologically informative genes for downstream analyses, including functional enrichment, literature-based interpretation, and potential biomarker discovery.



based differential analysis, and the top 100 are retained based on adjusted significance and effect-related criteria. This stage reduces dimensionality before classifier induction and ensures that feature selection remains consistent with the leakage-control principle adopted throughout the methodology, since no information from the test fold is used during ranking.

The selected features are then used as input for machine learning models. For each cell type-specific dataset, the methodology uses stratified 5-fold cross-validation. Within each fold, class balancing is applied only to the training partition when the disease and control classes are unequally represented, using random undersampling of the majority class to match the minority class. The validation and test partitions remain unbalanced to preserve the original data distribution. Within this structure, classifiers are trained, tuned, and evaluated using multiple performance criteria, namely accuracy, sensitivity, specificity, and F1-score. Because classifier quality cannot be adequately represented by a single metric, model comparison is conducted under a multiobjective framework.

The next stage of the framework performs model selection through Pareto frontier analysis and hypervolume. Instead of selecting a model from a single scalar measure, the framework identifies solutions that balance competing objectives. This step links predictive performance to a selection procedure that avoids favoring models that perform well only under a narrow criterion.

Once the best-performing solutions are identified, the framework applies SHAP-based interpretation to quantify each selected gene's contribution to model predictions. This interpretability stage allows the final classifier to be examined in terms of gene-level contributions, which are then evaluated through expression-attribution correlation and functional enrichment. The resulting ranked feature set is subsequently used in downstream analyses, including expression-attribution correlation, differential expression assessment, functional enrichment, literature-supported interpretation, and biomarker discovery.

Therefore, Figure 4.1 should be interpreted as a schematic representation of the main methodological blocks of the study rather than as an exhaustive algorithmic description. The figure summarizes the logical progression from annotated scRNA-seq data to biologically interpretable reduced feature sets. At the same time, the following sections formalize each step of the framework in mathematical and computational detail.

### 4.2.1 Methodological perspective

This work constructs a computational framework that identifies informative transcriptomic features and uses them for accurate and interpretable classification. The methodology evaluates how machine learning algorithms and feature-selection strategies behave when applied to scRNA-seq data under controlled comparison. The analysis therefore considers classification performance, stability, interpretability, and downstream functional enrichment. The enrichment step evaluates whether selected features define biologically coherent signatures and allows direct comparison between machine-learning-based and traditional feature-selection strategies.

The framework addresses three connected problems. First, the representation problem arises because transcriptomic data are high-dimensional, sparse, and heterogeneous; preprocessing and structured storage therefore define the numerical input for learning. Second, the learning problem requires training and evaluation procedures that prevent information leakage and estimate generalization reliably. Third, the interpretation problem requires feature rankings and model explanations that support downstream biological reasoning.

The framework therefore treats preprocessing, validation, feature selection, classification, and interpretation as linked stages. Preprocessing defines the valid numerical input space for learning. Cross-validation defines which information each stage may use. Feature selection reduces dimensionality before classification. Model interpretation occurs only after the classifier is trained under leakage-safe conditions. This sequential dependency defines the methodological logic of the chapter.

### 4.2.2 Data abstraction and formal definition

The input data are modeled as an annotated transcriptomic dataset, defined as

$$\mathcal{D} = \{X, \mathbf{y}, \mathbf{c}, \mathbf{m}\}, \tag{4.1}$$

where $X \in R^{n \times p}$ denotes the gene-expression matrix, $n$ denotes the number of observations, $p$ denotes the number of genes, $\mathbf{y}$ denotes the vector of target labels, $\mathbf{c}$ denotes the vector of cell-type annotations, and $\mathbf{m}$ denotes additional metadata. As shown in Eq. 4.1, the input data are modeled as an annotated transcriptomic dataset.

This abstraction was chosen because transcriptomic learning problems cannot be formulated using the expression matrix alone. The labels, cell identities, and metadata are not auxiliary details; they define the structure of the learning task. In particular, this methodology performs classification within cell-type-specific subsets, meaning that the annotation vector $\mathbf{c}$ plays a direct role in defining each supervised problem. Thus, the dataset is treated as a structured computational object rather than merely as a numeric table.

### 4.2.3 Design principles

The methodology is governed by five design principles:

- **Adaptivity:** datasets may arrive in raw or partially processed form. The framework must therefore infer how much preprocessing is still required instead of assuming homogeneous input data.
- **Leakage control:** no operation that depends on label structure or sample distribution is allowed to access test data during model training or feature selection.
- **Local biological context:** the same condition can induce different transcriptomic patterns across cell types. Therefore, the classification problem is defined separately for each cell population.
- **Multi-criterion evaluation:** classifier quality cannot be represented adequately by a single metric. The methodology therefore evaluates models through multiple objectives simultaneously.
- **Interpretability:** the framework output must include not only predictions, but also ranked biological features, attribution scores, and stability estimates.

These principles guided every technical choice described in the following sections.

### 4.2.4 Justification for cell type-specific modeling

A central design choice of the methodology is to analyze each cell type independently. Let $c$ denote a given cell type. The associated subset is defined as

$$\mathcal{D}_c = \{(x_i, y_i) \mid c_i = c\}. \tag{4.2}$$

As shown in Eq. 4.2, each cell type defines an independent supervised dataset.

This decision is justified on both biological and algorithmic grounds. Biologically, transcriptomic responses are highly context-dependent. A gene that is strongly discriminative in one cell type may be irrelevant in another. A global classifier trained across all cell types would tend to prioritize lineage-level variation and broad expression programs rather than the condition-specific effects within each cellular context.

Algorithmically, cell type-specific modeling reduces heterogeneity in the learning space. This improves feature selection, makes the classification boundary easier to interpret, and reduces the chance that the model learns trivial structure unrelated to the intended target. The result is a set of smaller, more coherent classification problems rather than one large but confounded problem (SQUAIR et al., 2021).

### 4.2.5 Justification for leakage-aware design

Information leakage is one of the most serious threats to validity in high-dimensional machine learning. In transcriptomic settings, where the number of genes is large and the number of observations is comparatively limited, even subtle leakage can produce inflated performance estimates.

For this reason, the methodology adopts a strict rule: any operation that can learn from the data distribution is applied only within the training split of a cross-validation fold. This includes class balancing, differential feature ranking, hyperparameter optimization, and final model fitting. The test fold is reserved exclusively for evaluation. This decision is not merely a technical precaution; it is a methodological necessity to ensure that the measured performance reflects genuine generalization rather than contamination from the evaluation data (VARMA; SIMON, 2006).

## 4.3 Overview of the framework

The computational framework is composed of the following major stages:

1. Ingestion of the annotated dataset and active matrix definition;
2. Matrix sanitization and metadata validation;
3. Automatic preprocessing mode inference;
4. Global quality control, normalization, and embedding;

5. Cell type-specific subsetting and local preprocessing;
6. Outer cross-validation partitioning;
7. Training-fold balancing;
8. Fold-wise feature selection;
9. Model training and hyperparameter optimization;
10. Fold-wise evaluation;
11. Multi-objective model comparison;
12. SHAP-based interpretation and correlation analysis;
13. Comparison with alternative feature-selection methods;
14. Final hold-out evaluation and export of results.

Each of these stages is described formally below.

### 4.3.1 Active matrix definition and sanitization

The framework begins by defining the active matrix $X$, which may correspond either to the primary expression matrix stored in the object or to a selected expression layer. This flexibility is necessary because single-cell transcriptomic objects commonly store multiple numerical representations of the same dataset, including raw counts, normalized values, log-transformed values, scaled matrices, or batch-corrected layers.

Before downstream analysis, the active matrix is inspected for non-finite and negative entries. This inspection is necessary because invalid numerical values may arise from file conversion, preprocessing artifacts, or previous transformations. However, missing or invalid values are not biologically equivalent to a true zero expression. In single-cell RNA-seq data, zero values already have a complex interpretation because they may reflect true absence of expression, limited molecular capture, or technical dropout (HICKS et al., 2018; LUN; BACH; MARIONI, 2016). Therefore, non-finite values are quantified globally and per gene/cell before correction.

When non-finite entries are rare and are interpreted as computational artifacts, they are replaced by zero only as a numerical safeguard to ensure compatibility with downstream algorithms that require finite input values. This replacement is not interpreted as a biological absence of expression. If non-finite values occur systematically across specific genes, cells, or groups, the affected entries are flagged for additional quality

control rather than silently converted. This precaution follows the general principle that simple deterministic imputation, including replacement by zero, can bias downstream estimates when missingness is not rare or random (LITTLE; RUBIN, 2019; ZHANG, 2016).

For each entry $x_{ij}$, the corrected value is defined as

$$x_{ij}^{*} = \begin{cases} 0, & \text{if } x_{ij} < 0 \text{ and } X \text{ represents a non-negative abundance matrix,} \\ 0, & \text{if } x_{ij} \in \{\mathrm{NaN}, +\infty, -\infty\} \text{ and non-finite values are rare,} \\ x_{ij}, & \text{otherwise.} \end{cases} \tag{4.3}$$

Negative values are clipped to zero only when the active matrix represents raw counts, normalized counts, or log-transformed abundance-like values, for which non-negative expression values are expected. This rule is not applied to scaled, centered, residualized, or batch-corrected matrices, because negative values in these representations may carry valid analytical information. Thus, the sanitization step is treated as a conservative numerical correction rather than as a biological imputation procedure.

### 4.3.2 Preprocessing mode inference

The framework supports two preprocessing modes:

- **Full preprocessing mode**, used for raw or count-like matrices;
- **Curated-matrix mode**, used for previously processed matrices.

To decide between them, the framework estimates the fraction of integer-like values, the minimum and maximum value of a sampled subset, and the presence of previously computed embeddings. Let $\rho_{\mathrm{int}}$ be the fraction of sampled entries that are approximately equal to their rounded values. The matrix is considered count-like when

$$\rho_{\mathrm{int}} \geq 0.98, \qquad \max(X) \geq 20, \qquad \min(X) \geq 0. \tag{4.4}$$

This rule was designed to distinguish raw count matrices from normalized or integrated matrices while remaining tolerant to small numeric perturbations introduced by storage or type conversion.

### 4.3.3 Metadata validation

The framework removes all observations with missing or empty target labels or cell-type labels. The valid observation set is defined as

$$\mathcal{I}_{\text{valid}} = \{i : y_i \neq \wedge c_i \neq\}. \tag{4.5}$$

This step guarantees that every retained observation has a well-defined role in both the supervised and cell-type-specific stages of the framework. Observations without valid annotations cannot support classification and would create ambiguity in the downstream analysis.

### 4.3.4 Global quality control

If the matrix is count-like, cell-level quality control is applied. For each observation $i$, the following quantities are computed:

$$\begin{aligned} g_i &= \text{number of detected genes}, \\ u_i &= \sum_{j=1}^{p} x_{ij}, \\ m_i &= 100 \cdot \frac{\sum_{j \in G_{\text{MT}}} x_{ij}}{\sum_{j=1}^{p} x_{ij}}, \end{aligned} \tag{4.6}$$

where $G_{\text{MT}}$ denotes the set of mitochondrial genes.

The retention criteria are defined as

$$g_i \geq 500, \qquad u_i \geq 200, \qquad m_i \leq 10. \tag{4.7}$$

Genes are then filtered according to minimum prevalence:

$$\sum_{i=1}^{n} I(x_{ij} > 0) \geq 30. \tag{4.8}$$

In addition, cell types represented by fewer than $30$ observations were removed from the global dataset. This cutoff was adopted as an operational minimum to avoid modeling cell-type-specific subsets with insufficient observations for stratified cross-validation, class balancing, feature selection, and held-out evaluation. The quality-control thresholds

were used to remove technically poor observations and lowly detected genes, following common single-cell RNA-seq preprocessing practices (LUECKEN; THEIS, 2019). The minimum cell-type size criterion was motivated by the known instability of classification and validation procedures under very small sample sizes, where performance estimates and selected features can become highly sensitive to the data split (BELEITES et al., 2013; VABALAS et al., 2019).

### 4.3.5 Normalization and logarithmic transformation

In full preprocessing mode, each observation is normalized to a fixed library size:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sum_{k=1}^{p} x_{ik}} \cdot 10^4. \tag{4.9}$$

The normalized matrix is then log-transformed:

$$x_{ij}^{\log} = \log(1 + \tilde{x}_{ij}). \tag{4.10}$$

These operations are intended to reduce technical variability caused by unequal sequencing depth and to compress the heavy-tailed distribution of transcript counts. In curated-matrix mode, both operations are skipped to avoid damaging pre-existing transformations.

### 4.3.6 Global PCA and neighborhood graph

After preprocessing, PCA is applied to the global matrix. The number of components is defined as

$$d_{\text{global}} = \min(50, p - 1, n - 1). \tag{4.11}$$

The PCA scores are then used to construct a neighborhood graph, which in turn supports UMAP embedding. This global embedding serves as a diagnostic representation of the dataset, enabling inspection of large-scale structure before supervised analysis begins.

### 4.3.7 Cell type-specific processing

For each cell type $c$, the corresponding subset $\mathcal{D}_c$ is extracted. Classes with fewer than $30$ observations within that subset are removed, and genes with zero expression across all remaining observations are discarded.

If full preprocessing mode is active, normalization and log-transformation are repeated within the subset. PCA is then recomputed using

$$d_{\text{celltype}} = \min(30, p_c - 1, n_c - 1), \tag{4.12}$$

where $p_c$ and $n_c$ are the number of genes and observations in the subset. A local neighborhood graph and UMAP embedding are then computed.

The framework recomputes these structures locally because the geometry of a cell-type-specific subset can differ from the geometry of the global matrix. Local preprocessing therefore produces a representation aligned with the learning task defined for that cell type.

### 4.3.8 Outer cross-validation

For each valid cell type-specific dataset $\mathcal{D}_c$, the framework applies stratified $5$-fold cross-validation. The dataset is partitioned into five non-overlapping test folds:

$$\mathcal{D}_c = \bigsqcup_{k=1}^{5} \mathcal{D}_c^{\text{test},k}, \qquad \mathcal{D}_c^{\text{test},k} \cap \mathcal{D}_c^{\text{test},\ell} = \quad \text{for } k \neq \ell. \tag{4.13}$$

For each fold $k$, the corresponding training set is defined as the complement of the test fold:

$$\mathcal{D}_c^{\text{train},k} = \mathcal{D}_c \setminus \mathcal{D}_c^{\text{test},k}. \tag{4.14}$$

Stratification preserves class proportions as closely as possible in each test fold, which is important when the target classes are imbalanced. The use of five folds was adopted as an operational compromise between repeated evaluation and data availability. Because each retained class contains at least $30$ observations, a $5$-fold split leaves approximately six observations per class in each test fold and approximately twenty-four observations per class in each training fold before class balancing. This configuration pre-

serves enough training data for fold-wise feature selection, hyperparameter optimization, and model fitting, while still evaluating each observation once across the cross-validation procedure.

A larger number of folds, such as $10$, would increase the amount of training data per fold but would leave fewer observations from small classes in each test partition, increasing the instability of fold-level estimates. Conversely, fewer folds would reduce the number of repeated evaluations and provide a coarser estimate of model behavior. Thus, $5$-fold cross-validation was selected as a pragmatic choice for cell type-specific transcriptomic subsets, where each subset may contain limited and imbalanced class representation. This choice follows the general use of $k$-fold cross-validation for model assessment and model selection, while recognizing that validation estimates can become unstable when sample sizes are limited (KOHAVI, 1995; VABALAS et al., 2019).

### 4.3.9 Training-fold balancing

Within each outer training fold, class imbalance is corrected through random undersampling (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017; HE; GARCIA, 2009). If $n_r$ denotes the number of training observations in class $r$, and $n_{\min} = \min_r n_r$, the balanced training set contains

$$n_{\text{bal}} = C \cdot n_{\min} \tag{4.15}$$

observations, where $C$ is the number of classes.

This balancing strategy was chosen because it avoids generating artificial transcriptomic observations. Although majority-class information is partially discarded in a given fold, all retained examples remain biologically authentic. Random undersampling was preferred over synthetic balancing strategies because the objective of the study was not merely to equalize class frequencies, but to preserve the biological authenticity of the transcriptomic profiles used for training. In high-dimensional single-cell expression data, synthetic methods may interpolate sparse profiles in ways that are mathematically convenient but biologically difficult to justify, particularly when class structure is heterogeneous and cell-state boundaries are not well approximated by simple local interpolation. Although undersampling reduces the number of majority-class observations available in a given fold, it preserves the empirical character of the retained cells and avoids introducing artificial expression patterns that could complicate downstream interpretation.

### 4.3.10 Fold-wise feature selection

The primary feature-selection method is based on Wilcoxon rank-based differential analysis computed within each balanced training fold. This choice is supported by recent benchmarking evidence showing that label-guided marker-gene selection based on the Wilcoxon rank-sum test can perform competitively in single-cell feature-selection tasks. In a large benchmark of feature-selection methods for scRNA-seq integration and querying, Zappia et al. reported that Wilcoxon was the only method that explicitly used cell labels and achieved the highest average overall score, although with greater variability across datasets than several highly variable gene methods (ZAPPIA et al., 2025). This supports its use when reliable labels are available and when the objective is to identify discriminative genes associated with biological or phenotypic classes.

After ranking genes by adjusted significance and effect-related criteria, the top $100$ genes are retained:

$$G^{(k)} = \{g_1, g_2, \ldots, g_{100}\}. \tag{4.16}$$

These genes define the feature space for the fold-specific classification problem. The fixed cutoff of $100$ genes was chosen to balance dimensionality reduction, computational tractability, and biological interpretability. This compact feature set reduces the high dimensionality of single-cell transcriptomic data while preserving a sufficient number of discriminative genes for downstream classification.

### 4.3.11 Alternative feature-selection methods

To compare the proposed Wilcoxon-based strategy against alternative feature-selection principles, the framework also computes four reference selectors:

- ANOVA-based ranking (PAVLIDIS, 2003);
- Random Forest importance (BREIMAN, 2001; DíAZ-URIARTE; ANDRéS, 2006);
- Recursive Feature Elimination (RFE) (GUYON et al., 2002);
- Random ranking (ZAPPIA et al., 2025);

These methods were selected because they represent distinct methodological categories. ANOVA provides a univariate parametric filter and serves as a direct statistical comparator to the Wilcoxon rank-based strategy. Random Forest importance represents

an embedded non-linear approach, in which feature relevance is estimated during model fitting and may capture interactions not detected by univariate tests. RFE represents a wrapper-based strategy, where genes are iteratively removed according to their contribution to a predictive model. Random ranking is included as a null reference, allowing the analysis to determine whether informed feature-selection methods outperform arbitrary gene selection.

For stochastic selectors, $11$ replicates are executed to support stability analysis. This number was chosen as a pragmatic compromise between computational cost and the need to quantify variability in selected gene sets. Repeated runs allow the framework to estimate selection frequencies and pairwise overlap between feature sets. The use of an odd number of replicates also avoids ties when assessing majority recurrence across runs.

### 4.3.12 Classifier framework construction

For each feature-selection output, the framework constructs a two-stage classification pipeline:

$$\mathcal{P}_\theta = \mathcal{M}_\theta \circ \mathcal{S}, \tag{4.17}$$

where $\mathcal{S}$ is a min-max scaler and $\mathcal{M}_\theta$ is a classifier parameterized by $\theta$. For each gene $j$, min-max scaling is defined as

$$x'_{ij} = \frac{x_{ij} - \min(x_{\cdot j})}{\max(x_{\cdot j}) - \min(x_{\cdot j})}, \tag{4.18}$$

where $x_{ij}$ is the expression value of gene $j$ in cell $i$.

The evaluated classifiers are Logistic Regression, $k$-Nearest Neighbors, linear Support Vector Machine, Random Forest, and Multilayer Perceptron. These models were chosen to cover distinct learning paradigms: linear, local, margin-based, ensemble-based, and neural.

### 4.3.13 Bayesian hyperparameter optimization

Each classifier is optimized through Hyperopt-based Bayesian search, with an upper limit of $50$ candidate configurations per classifier (BERGSTRA; YAMINS; COX,

2013; BERGSTRA et al., 2015a). The search objective is defined as

$$\theta^* = \arg\max_{\theta \in \Theta} \Phi(\theta), \tag{4.19}$$

where

$$\Phi(\theta) = \frac{1}{4}\left[\text{Accuracy}(\theta) + \text{Sensitivity}(\theta) + \text{Specificity}(\theta) + \text{F1}(\theta)\right]. \tag{4.20}$$

This objective function was chosen to avoid selecting models according to a single performance criterion. The arithmetic mean of accuracy, sensitivity, specificity, and F1-score promotes classifiers that maintain balanced predictive behavior across complementary evaluation dimensions, rather than classifiers optimized for only one aspect of performance.

The limit of $50$ evaluations was defined as a fixed computational budget. Each candidate configuration requires model fitting and evaluation inside the training partition of an outer cross-validation fold, after class balancing and fold-specific feature selection. Therefore, increasing the number of evaluations would substantially increase the total cost across classifiers, folds, and cell type-specific tasks. The value of $50$ evaluations provides a pragmatic compromise between hyperparameter-space exploration and computational feasibility, while keeping the optimization protocol consistent across models.

The Bayesian search used only the training partition of each outer cross-validation fold. For every candidate configuration $\theta$, the framework estimated $\Phi(\theta)$ from training data after balancing and fold-specific feature selection. The corresponding outer test fold remained unused until final evaluation. This nesting kept hyperparameter selection, feature ranking, and model fitting inside the training portion of the outer fold, thereby preserving the leakage-aware structure of the framework.

### 4.3.14 Fold-wise evaluation

After hyperparameter selection, each classifier is retrained on the full balanced training fold and evaluated on the untouched test fold. The fold-level confusion matrix yields the following class-wise metrics:

$$\text{Sensitivity}_k = \frac{TP_k}{TP_k + FN_k}. \tag{4.21}$$

$$\text{Specificity}_k = \frac{TN_k}{TN_k + FP_k}. \tag{4.22}$$

$$\text{F1}_k = \frac{2TP_k}{2TP_k + FP_k + FN_k}. \tag{4.23}$$

Global accuracy is given by

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} TP_k}{N}. \tag{4.24}$$

Macro-averaged sensitivity, specificity, and F1-score are computed across classes to avoid dominance by larger classes.

### 4.3.15 Pareto analysis and hypervolume ranking

Each model is represented in the objective space:

$$\mathbf{z} = (\text{Accuracy}, \text{Sensitivity}, \text{Specificity}, \text{F1}). \tag{4.25}$$

A model $A$ dominates another model $B$ if it is at least as good in all objectives and strictly better in at least one objective:

$$\mathbf{z}_A \succeq \mathbf{z}_B \iff [z_{A,j} \geq z_{B,j} \quad \forall j \in \{1, \ldots, M\}] \wedge [\exists j \in \{1, \ldots, M\} \text{ such that } z_{A,j} > z_{B,j}], \tag{4.26}$$

where $M = 4$ denotes the number of objectives.

All non-dominated models form the empirical Pareto frontier. To obtain a scalar ranking, the framework computes the hypervolume indicator:

$$HV(S) = \lambda \left( \bigcup_{\mathbf{z} \in S} [r_1, z_1] \times [r_2, z_2] \times \cdots \times [r_M, z_M] \right), \tag{4.27}$$

where $S$ is the set of non-dominated solutions, $\mathbf{r}$ is the reference point, and $\lambda$ denotes the corresponding Lebesgue measure.

This stage allows the framework to identify models that achieve a balanced trade-off across multiple performance criteria rather than selecting models based on a single scalar metric.

### 4.3.16 SHAP-based interpretation

For the selected classifiers, the framework computes SHAP values (PHAM; MILAZZO, 2025). For an input $x$, the prediction is decomposed as

$$f(x) = \phi_0 + \sum_{j=1}^{q} \phi_j, \tag{4.28}$$

where $\phi_0$ is the baseline prediction and $\phi_j$ is the contribution of feature $j$.

The choice of SHAP was motivated by the need for additive explanations that can be applied across different classifier families. Different explainer types are used depending on the classifier family. All explanations are computed on held-out test data, preserving consistency with the leakage-aware evaluation framework.

### 4.3.17 SHAP aggregation and expression-attribution correlation

For each feature $j$, the framework computes the mean signed attribution:

$$\overline{\phi}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_{ij}, \tag{4.29}$$

and the mean absolute attribution:

$$\overline{|\phi|}_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_{ij}|. \tag{4.30}$$

To connect model explanations with transcript abundance, the framework also computes the Spearman correlation:

$$\rho_j = \mathrm{corr}_{\mathrm{Spearman}} \left( X_{\cdot j}, \Phi_{\cdot j} \right). \tag{4.31}$$

This step provides directional information on whether higher expression tends to increase or decrease model attribution for a given feature.

#### 4.3.18 Integrated output construction

At the end of the framework, all ranking sources are merged into a unified feature table. For each gene, the output may include differential rank, model-based rank, SHAP importance, correlation with attribution, replicate frequency, and stability measures. This integrated representation is necessary because different analytical views capture different aspects of feature relevance.

### 4.4 Experimental Protocol

#### 4.4.1 Parameterization of the experiments

The experiments were conducted under a fixed default configuration designed to ensure comparability across cell types and feature-selection methods. The main parameters are:

- number of selected genes: $100$;
- number of plotted top genes: $25$;
- number of replicates for stochastic baselines: $11$;
- number of outer cross-validation folds: $5$;
- maximum number of hyperparameter evaluations per model: $50$;
- final hold-out test proportion: $0.30$;
- global PCA upper bound: $50$;
- cell type-specific PCA upper bound: $30$;
- random seed: $42$.

These values were chosen to balance computational tractability, repeated evaluation, and interpretability.

#### 4.4.2 Comparative analysis of class separation induced by selected feature subsets

In addition to the predictive experiments described above, the methodology includes a complementary geometric analysis designed to compare how different feature-selection methods organize the disease groups in the reduced expression space. This

additional experiment was introduced to evaluate whether the selected features preserve a structured distinction between chronic obstructive pulmonary disease and normal cells beyond their utility for classification. Thus, this stage should be interpreted as a supplementary structural assessment of the selected feature subsets rather than as a replacement for the main predictive evaluation.

This analysis was performed under a fixed configuration consistent with the general parameterization of the experiments. In particular, each method was represented by a reduced subset of $25$ genes, and all outputs were generated under a fixed random seed and exported at publication-quality resolution. These choices maintain consistency with the broader experimental framework while allowing direct comparison between reduced feature spaces derived from different selection strategies.

The analysis was applied to a cell type-specific subset rather than to the full dataset. In the present configuration, only cells annotated as "alternatively activated macrophage" were retained. From this subset, the framework preserved only the observations assigned to the two disease groups of interest, namely chronic obstructive pulmonary disease and normal. Cells with missing disease labels were excluded before downstream computation. This design remains aligned with the methodological principle adopted throughout the chapter, according to which each biological learning problem is defined within a local cellular context rather than across all cell populations simultaneously.

Let

$$\mathcal{D}_{\text{sep}} = (X, \mathbf{y}) \tag{4.32}$$

denote the filtered dataset used in this experiment, where $X \in R^{n \times p}$ is the expression matrix of the retained cells and $\mathbf{y}$ is the corresponding vector of disease labels. The label space is restricted to

$$\mathcal{Y} = \{\text{COPD}, \text{Normal}\}\,. \tag{4.33}$$

Six feature-selection strategies were compared: the proposed method, ANOVA, Wilcoxon rank-sum test, Random Forest feature importance, Recursive Feature Elimination (RFE), and a random feature subset used as a negative baseline. These methods correspond to the same families already defined in the methodological framework and were included here to compare the structural behavior induced by their reduced feature subsets under a common geometric criterion.

For each strategy, the framework used a predefined subset of $25$ genes. These identifiers were matched to the variables stored in the AnnData object through an automatic

resolution procedure that removed Ensembl version suffixes when necessary and also searched alternative gene annotations such as `feature_name` and `gene_symbol`. Genes that could not be matched to the active dataset were recorded as missing and excluded from the corresponding method-specific analysis. This procedure ensured that all methods were compared only on valid genes effectively represented in the expression matrix.

After gene resolution, the method-specific expression matrix was extracted from the filtered AnnData object. When required, sparse matrices were converted to dense format for downstream numerical computation. In the current configuration, no additional `log1p` transformation was applied, since the active matrix used in this stage did not require further count-based transformation. Each extracted matrix was then standardized using `StandardScaler`. For gene $j$, the standardized expression value for observation $i$ is given by

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{4.34}$$

where $\mu_j$ and $\sigma_j$ denote the mean and standard deviation of gene $j$ within the reduced matrix. Standardization was required to ensure that all selected genes contributed on a comparable numerical scale to the subsequent distance-based analysis.

The main structural metrics were computed in the full standardized feature space rather than in a two-dimensional projection. Let

$$Z^{(m)} \in R^{n \times q} \tag{4.35}$$

denote the standardized matrix associated with method $m$, where $q$ is the number of valid genes retained for that method. For each disease group $r \in \mathcal{Y}$, the class centroid is defined as

$$\boldsymbol{\mu}_r = \frac{1}{|I_r|} \sum_{i \in I_r} Z_{i\cdot}^{(m)}, \tag{4.36}$$

where $I_r$ is the set of observations assigned to group $r$. The distance between the COPD and normal centroids is then computed as

$$d_{\text{centroid}} = \left\| \boldsymbol{\mu}_{\text{COPD}} - \boldsymbol{\mu}_{\text{Normal}} \right\|_2 . \tag{4.37}$$

To characterize within-group compactness, the methodology computes the mean Euclidean distance between each observation and the centroid of its own class. For group

$r$, the within-group dispersion is defined as

$$\delta_r = \frac{1}{|I_r|} \sum_{i \in I_r} \left\| Z_{i\cdot}^{(m)} - \boldsymbol{\mu}_r \right\|_2 . \tag{4.38}$$

The pooled within-group dispersion is then given by

$$\delta_{\text{pooled}} = \frac{|I_{\text{COPD}}|\, \delta_{\text{COPD}} + |I_{\text{Normal}}|\, \delta_{\text{Normal}}}{|I_{\text{COPD}}| + |I_{\text{Normal}}|}. \tag{4.39}$$

The principal comparison criterion is the separation ratio:

$$R_{\text{sep}} = \frac{d_{\text{centroid}}}{\delta_{\text{pooled}}}. \tag{4.40}$$

which normalizes the displacement between class centroids by the average within-group dispersion. This ratio follows the same general principle used in classical discriminant analysis, in which class separation is evaluated through the relation between between-group and within-group variability (FISHER, 1936). This ratio was adopted because centroid distance alone cannot distinguish between a well-separated structure and a situation in which large inter-centroid displacement coexists with broad class overlap. By contrast, the separation ratio jointly reflects class displacement and compactness, providing a more balanced description of the geometry induced by each feature subset.

To complement the centroid-based analysis, the methodology also computes three additional structural metrics in the same full standardized feature space: the Silhouette score (ROUSSEEUW, 1987), the Davies-Bouldin index (DAVIES; BOULDIN, 1979), and the Calinski-Harabasz index (CALIŃSKI; HARABASZ, 1974). The Silhouette score measures how well each observation matches its assigned class relative to the alternative class. The Davies-Bouldin index summarizes the relation between within-group scatter and between-group separation, with smaller values indicating more favorable structure. The Calinski-Harabasz index evaluates the ratio between between-group and within-group dispersion, with larger values indicating stronger class organization. These quantities are interpreted as complementary geometric descriptors and not as substitutes for the predictive metrics used in the main learning experiments.

Although the ranking of the methods is based on the full standardized feature space, low-dimensional embeddings were also generated for qualitative inspection. For each method-specific matrix, Principal Component Analysis was fitted independently

with an upper bound of $10$ components,

$$d_{\mathrm{PCA}} = \min(10, q, n-1). \tag{4.41}$$

and the first two principal components were used to visualize class overlap and centroid displacement. In parallel, UMAP was computed using $15$ neighbors, $\texttt{min_dist} = 0.30$, Euclidean distance, and fixed random seed $42$. These visualizations were used exclusively for interpretation and figure generation. They do not define the primary ranking criterion, which remains based on the metrics computed in the full standardized feature space.

For each feature-selection strategy, the framework stores the list of matched genes, the list of unmatched genes, the PCA and UMAP visualizations, and a summary table containing all structural metrics. The final comparison table ranks the methods according to the separation ratio obtained in the full standardized space. This output structure makes it possible to interpret the geometric behavior of the selected feature subsets alongside the predictive and interpretability analyses described in the remaining parts of the chapter.

This complementary experiment does not modify the classifier optimization procedure described in the next subsection. Instead, it provides an additional structural perspective from which the reduced feature subsets can be compared before returning to the specification of the supervised models and their associated hyperparameter search spaces.

### 4.4.3 Classifier search spaces

Each classifier is associated with a predefined search space. Logistic Regression searches over the regularization strength, since regularization controls model complexity and coefficient shrinkage in linear classifiers (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; PEDREGOSA et al., 2011). $k$-Nearest Neighbors searches over neighborhood size and weighting scheme, because the number of neighbors and the contribution assigned to nearby observations directly affect the decision boundary (COVER; HART, 1967; ALTMAN, 1992; PEDREGOSA et al., 2011). Support Vector Machine searches over the margin penalty parameter $C$, which controls the trade-off between margin maximization and training-error penalization (CORTES; VAPNIK, 1995). Random Forest searches over the number of trees and maximum tree depth, two parameters directly related to ensemble size, variance reduction, and tree complexity (BREIMAN, 2001; PROBST; WRIGHT; BOULESTEIX, 2019). Multilayer Perceptron searches over hidden-

layer configuration, activation function, solver, and regularization strength, because these parameters define the network architecture, non-linear transformation, optimization procedure, and weight penalization (GOODFELLOW; BENGIO; COURVILLE, 2016; PEDREGOSA et al., 2011).

The search spaces were intentionally bounded. Hyperparameter optimization requires repeated model fitting, and broader search spaces increase computational cost without necessarily improving model selection under a fixed evaluation budget (BERGSTRA; BENGIO, 2012; SNOEK; LAROCHELLE; ADAMS, 2012; BERGSTRA et al., 2015b). The selected ranges were therefore defined to cover materially different model behaviors while preserving computational feasibility across classifiers, cross-validation folds, and cell type-specific tasks.

### 4.4.4 Final hold-out evaluation

After cross-validation and model comparison, a final train-test evaluation is performed for the selected method:

$$\mathcal{D}_c = \mathcal{D}_c^{\text{train}} \cup \mathcal{D}_c^{\text{test}}, \qquad \mathcal{D}_c^{\text{train}} \cap \mathcal{D}_c^{\text{test}} =, \qquad \left|\mathcal{D}_c^{\text{test}}\right| = 0.30\,|\mathcal{D}_c|\,. \tag{4.42}$$

The $70/30$ partition was chosen to preserve enough data for final training while retaining a sufficiently large independent test set for evaluation. This final stage ensures that the model reported in the dissertation is not evaluated exclusively through the same procedure used for model comparison.

## 4.5 Complexity and Implementation Considerations

### 4.5.1 Computational complexity of the framework

The framework combines multiple computationally expensive stages: PCA has complexity dependent on the matrix dimensions and the decomposition method. Neighborhood graph construction depends on the number of observations and retained components. Cross-validation multiplies the cost of feature selection, balancing, optimization, and model fitting by the number of folds.

The total complexity is further increased by the number of cell types analyzed

independently. If $C_T$ is the number of valid cell types, $K$ is the number of folds, $M$ is the number of classifiers, and $E$ is the number of hyperparameter evaluations, the total number of optimization fits is on the order of

$$\mathcal{O}\left(C_T \cdot K \cdot M \cdot E\right). \tag{4.43}$$

This multiplicative structure justifies several practical choices in the methodology, such as limiting the number of PCA components, capping hyperparameter evaluations at $50$, and using $5$-fold rather than more expensive repeated or nested outer protocols.

### 4.5.2 Memory considerations

Transcriptomic matrices from single-cell assays are typically sparse, because many entries in cell-by-gene matrices are zero or near zero (JIANG et al., 2022). The framework therefore preserves sparse representations whenever possible and converts matrices to dense format only when required by specific downstream operations. This design follows the data-management principles used in the single-cell Python ecosystem, where AnnData and SciPy support sparse matrices to reduce memory usage and enable efficient handling of high-dimensional biological data (WOLF; ANGERER; THEIS, 2018; VIRSHUP et al., 2024; VIRTANEN et al., 2020). Because support for sparse input varies across estimators and preprocessing functions, dense conversion is restricted to the specific stages that require it, such as selected classifier-fitting or feature-ranking routines (PEDREGOSA et al., 2011).

Memory usage is particularly relevant during cell type-specific analysis. By processing one cell type at a time, the framework limits peak memory consumption compared with a design that would attempt to optimize models for all cell types simultaneously. This strategy is consistent with best-practice recommendations for scalable single-cell workflows, which emphasize memory-aware data structures, sparse matrices, and partial processing of large annotated matrices (HEUMOS et al., 2023; VIRSHUP et al., 2024).

### 4.5.3 Scalability considerations

The methodology was designed to scale across multiple cell types and multiple classifier families, but not to treat all cell types in a single global model. This was an

intentional scalability trade-off. A single monolithic model might appear simpler from an implementation perspective, but would be more difficult to interpret and more likely to confound cell identity with condition. The chosen design scales by decomposition: a large biological problem is split into smaller computational tasks.

This decomposition also enables selective re-execution. If a failure occurs for one cell type, the entire framework does not need to be restarted.

### 4.5.4 Determinism and reproducibility

A fixed random seed is used throughout the framework whenever possible. This affects fold generation, undersampling, hyperparameter optimization state initialization, classifier construction, and stochastic feature-selection baselines. The purpose of fixing the random seed is not to eliminate stochasticity entirely, but to ensure that the computational experiments are reproducible.

The framework also records execution logs, metadata, hyperparameters, feature rankings, model outputs, and summary tables. This level of traceability was considered essential because a Computer Science dissertation must allow the methodological chain to be inspected, repeated, and audited.

### 4.5.5 Output organization

The framework stores outputs at multiple levels:

- Global-level exploratory outputs;
- Cell type-specific preprocessing outputs;
- Fold-level feature-selection and evaluation outputs;
- Model-level hyperparameter and metric summaries;
- Interpretability outputs;
- Integrated summary tables.

This organization was chosen to support both debugging and scientific reporting. Intermediate artifacts are valuable not only for implementation verification, but also for methodological transparency. They allow each result in the final thesis to be traced back to the exact stage in which it was generated.

## 4.6 Chapter Summary

This chapter provided the study with its methodological framework. The *Research Design* provided the framework with analytical principles for leakage control, local biological context, and interpretability. The *Algorithmic framework* gave the analysis a computational sequence from the annotated matrix to the integrated feature tables. The *Experimental Protocol* provided the comparison with training, optimization, and evaluation procedures for feature-selection methods and classifiers. The *Complexity and Implementation Considerations* gave the implementation a description of computational trade-offs, memory constraints, reproducibility procedures, and output organization.

The next chapter, *Results*, applies this framework to the analyzed single-cell transcriptomic data and presents the outcomes of feature selection, model evaluation, multi-objective selection, and SHAP-based interpretation.

# 5 RESULTS

## 5.1 Introduction

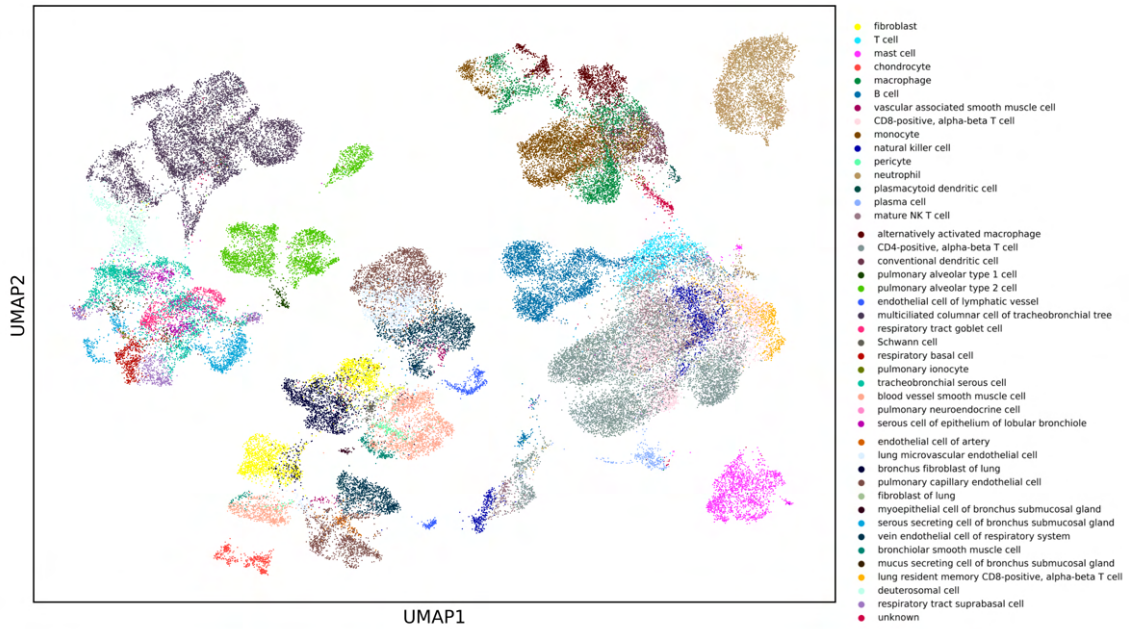
This chapter reports the results obtained from the proposed computational framework. The organization follows the main stages of the methodology and compares machine-learning-based feature selection with traditional strategies. The chapter first describes preprocessing and data-quality outcomes, then reports classification results for each evaluated model and feature-selection strategy. It then presents the multiobjective analysis used to identify compromise solutions across performance metrics. Finally, it examines interpretability, stability, and downstream biological validation through functional enrichment. This structure relates predictive performance to reproducibility and biological coherence.

## 5.2 Chronic obstructive pulmonary disease dataset

To evaluate the proposed framework, we used the single-cell RNA-sequencing dataset reported by (RUSTAM et al., 2023). In their study, the authors investigated the cellular organization of human distal airways and its disruption in chronic obstructive pulmonary disease (COPD). They generated a transcriptomic atlas comprising 111,412 single cells collected from anatomically defined airway regions, including proximal airways, pre-terminal airways, and terminal airway regions, from healthy lung donors and subjects with COPD. More specifically, the scRNA-seq dataset included cells isolated from the proximal airways of 5 healthy donors, pre-terminal airways of 12 healthy donors, terminal airways of 5 patients with COPD, and terminal airways of 7 healthy donors. The study focused on distal conducting airways, particularly preterminal and terminal bronchioles, and identified region-specific epithelial, structural, and immune cell populations across the proximal distal airway axis.

For the present work, this published dataset was used as a benchmark to test the proposed computational framework under a biologically relevant and clinically motivated setting. However, to reduce potential confounding effects associated with sex-specific transcriptional variation, only male samples were retained for the analyses performed in this study. This restriction was also motivated by sample availability, as male donors contributed more cells to the dataset subset used here. Therefore, the dataset analyzed in

Figure 5.1: Global UMAP projection of the male-restricted dataset colored by annotated cell type. The embedding preserves broad cellular diversity and shows that the retained subset spans the major transcriptional populations used in the downstream analyses.




this dissertation corresponds to a sex-restricted subset of the original publication, selected to improve comparability across samples and to minimize sex-related bias in downstream feature selection and classification analyses.

The UMAP projection colored by annotated cell types shows the cellular diversity retained in the dataset (Figure 5.1), whereas the UMAP colored by condition shows the distribution of normal and COPD-derived cells across the transcriptional landscape (Figure 5.2). Figure 5.3 summarizes the relative abundance of each annotated cell type.

### 5.2.1 Global quality control and filtering

The male-restricted dataset initially contained 69,366 cells and 36,397 genes. During object diagnosis, the active matrix was identified as a `float64` matrix with no additional layers, although a `raw` slot was available. Based on the automatic preprocessing diagnosis, the dataset was classified as `author_qc_mode`, indicating that the matrix was already curated or transformed. This interpretation was supported by the low fraction of approximately integer-valued entries (0.0121), the relatively small maximum observed value (6.5699), and the presence of precomputed embeddings. Therefore, no new `normalize_total` or `log1p` step was applied.

At the global filtering stage, no cells were removed due to missing annotations,

Figure 5.2: Global UMAP projection of the male-restricted dataset colored by condition. The embedding shows substantial overlap between normal and chronic obstructive pulmonary disease cells in the global transcriptional landscape, which reinforces the difficulty of the classification problem.

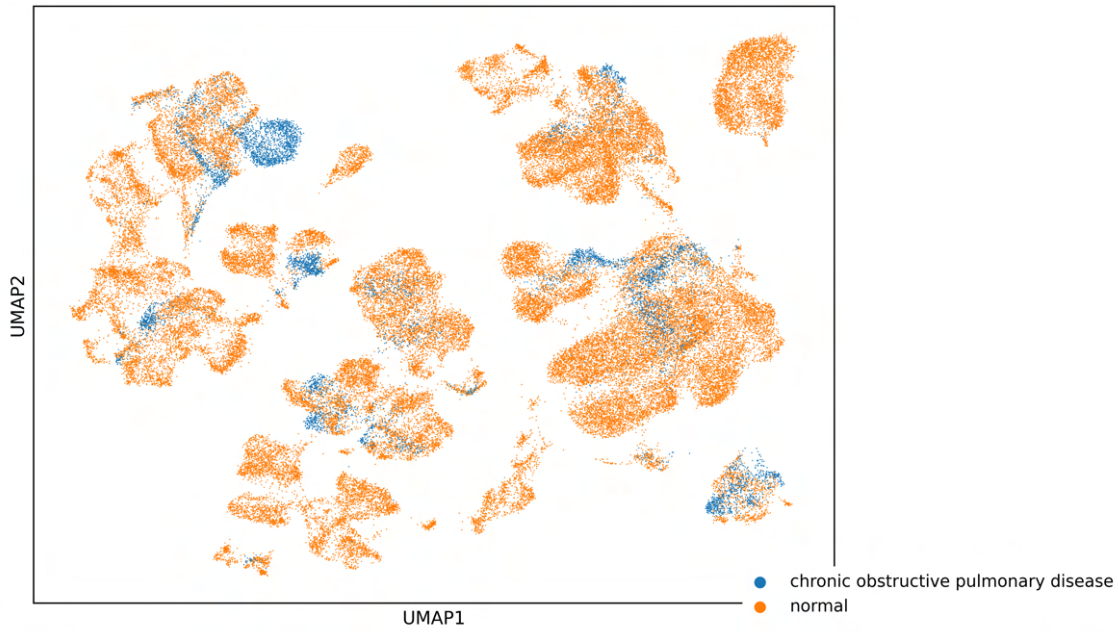
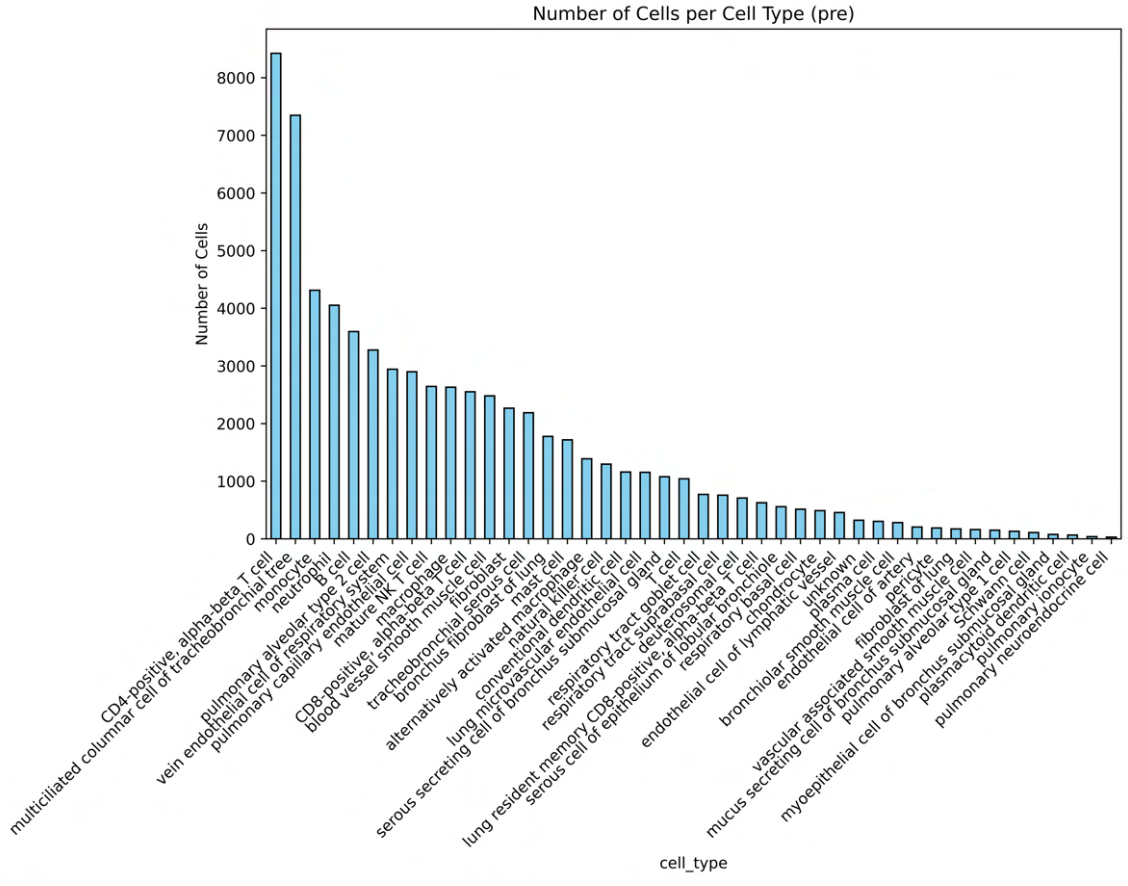



Figure 5.3: Relative abundance of each annotated cell type in the male-restricted dataset. The bar plot shows that the dataset retained broad cellular coverage, although cell-type abundance varied markedly across populations.



and no cell types were excluded by the minimum size threshold of 30 cells. In contrast, genes with zero expression across the active matrix were removed, reducing the feature space from 36,397 to 31,902 genes. Thus, the global quality-control stage preserved all cells while removing 4,495 uninformative genes. After filtering, PCA, neighborhood graph construction, and global UMAP embedding were completed.

Table 5.1: Summary of global quality-control and filtering steps applied to the male-restricted dataset.

| Step | Before | After | Removed |
|---|---|---|---|
| Cells after annotation validation | 69,366 | 69,366 | 0 |
| Genes after zero-only filtering | 36,397 | 31,902 | 4,495 |

At the cell-type-specific stage, 27 cell populations met the minimum class requirements. They were processed in the downstream framework, whereas 17 populations were excluded because they did not retain sufficient class representation after local filtering (see Table 5.2). The excluded cell types were *serous cell of epithelium of lobular bronchiole*, *respiratory tract suprabasal cell*, *deuterosomal cell*, *pulmonary neuroendocrine cell*, *endothelial cell of artery*, *unknown*, *lung resident memory CD8-positive, alpha-beta T cell*, *natural killer cell*, *vascular associated smooth muscle cell*, *Schwann cell*, *pulmonary alveolar type 1 cell*, *pulmonary ionocyte*, *serous secreting cell of bronchus submucosal gland*, *mucus secreting cell of bronchus submucosal gland*, *myoepithelial cell of bronchus submucosal gland*, *neutrophil*, and *plasmacytoid dendritic cell*.

Table 5.2: Eligibility of cell types for downstream analysis.

| **Category** | **Number of cell types** |
|---|---|
| Processed cell types | 27 |
| Skipped due to insufficient classes after filtering | 17 |

### 5.2.2 CD4-positive alpha-beta T cells

#### *5.2.2.1 Preprocessing outcomes*

The *CD4-positive, alpha-beta T cell* population satisfied the minimum requirements for downstream modeling and was therefore retained for cell type-specific analysis. After local filtering, both target classes remained represented, allowing this subset to proceed to cross-validation and supervised learning. No evidence of class collapse was

observed at this stage, confirming that this population constituted a valid classification problem within the proposed framework.

A marked class imbalance was observed in the training partitions of all cross-validation folds. In each fold, the training set contained 6,417 cells from the *normal* class and 320 cells from the *chronic obstructive pulmonary disease* class. To avoid bias toward the majority class during feature selection and model fitting, random undersampling was applied within each training fold. As a result, the balanced training set used in every fold contained 320 cells per class, yielding a final training distribution of 320 *normal* cells and 320 *chronic obstructive pulmonary disease* cells.

Thus, for the *CD4-positive, alpha-beta T cell* subset, preprocessing at the cell type-specific stage preserved class eligibility but revealed a strong imbalance between conditions, which was systematically corrected before downstream feature selection and classification.

#### *5.2.2.2 Comparative classification performance, multiobjective selection, and final MLP evaluation*

Table 5.6 summarizes the comparative performance of all evaluated classifiers for the *CD4-positive, alpha-beta T cell* subset. Among the tested models, the Multilayer Perceptron (MLP) produced the largest values for accuracy ($0.9006$), sensitivity ($0.9051$), specificity ($0.9051$), and macro F1-score ($0.7057$). These results indicate that, in this subset, the MLP represented the condition-associated structure more effectively than the other evaluated classifiers.

The Pareto frontier analysis reinforced this result. Figure 5.4 shows that the MLP occupied the most favorable region of the comparative performance profile and achieved the strongest joint balance across accuracy, sensitivity, specificity, and macro F1-score. Logistic Regression and SVM formed an intermediate-performance group. In contrast, KNN and Random Forest showed the weakest overall behavior, particularly in macro F1-score, indicating a less balanced trade-off between precision and recall under the imbalanced evaluation setting.

Based on this combined evidence, we selected the MLP as the final classifier for downstream interpretability analysis. Table 5.7 reports the hyperparameter configuration of the selected model. The final architecture consisted of two hidden layers with 50 neurons each, ReLU activation, Adam optimization, and $\alpha = 0.0893$, providing the best balance between discrimination capacity and generalization within the proposed leakage-

aware framework.

To obtain an independent estimate of predictive performance, we retrained the selected MLP on the training partition and evaluated it on the held-out test set. Table 5.8 shows that the final model achieved an accuracy of $0.8805$, macro sensitivity of $0.8977$, macro specificity of $0.8977$, and macro F1-score of $0.6774$. Although these values were slightly lower than the cross-validated averages, they remained consistent with the ranking observed during model comparison. They confirmed that the selected classifier retained strong predictive ability on unseen data.

The confusion matrices provided additional insight into model behavior. Across cross-validation folds, the MLP showed a stable predominance of correct classifications along the main diagonal, indicating consistent generalization across partitions (Figure 5.5). The final train test confusion matrix preserved this pattern, and the ROC analysis further supported this result (Figure 5.6). In the independent evaluation, the MLP achieved an AUC of $0.96$ for both classes, indicating excellent separability between *normal* and *chronic obstructive pulmonary disease* samples.

Because the dataset remained imbalanced, the F1-score represents the most robust metric for interpreting final performance, as it better captures the joint behavior of precision and recall than accuracy alone. Thus, although accuracy, sensitivity, and specificity were high, the lower macro F1-score indicates that minority-class prediction remained more challenging in the held-out test set. Even so, the final MLP demonstrated satisfactory generalization and provided a suitable basis for the subsequent SHAP-based interpretability analysis.

Beyond classifier performance, the comparison of feature-selection methods revealed both shared and method-specific structures within the CD4-positive, alpha-beta T-cell subset. Figure 5.7 shows that the top-25 gene lists overlapped only partially across methods, which indicates that each strategy captured a different notion of relevance. The proposed method shared 11 genes with RFE and 9 genes with both Wilcoxon and ANOVA, whereas Wilcoxon and RFE showed the smallest overlap. Figure 5.8 complements this result by showing that Random Forest, RFE, and ANOVA produced distinct score distributions for their top-ranked genes and differed sharply in computational cost. In particular, RFE was the most expensive method, whereas ANOVA and Wilcoxon remained computationally lightweight.

Table 5.3: Hyperparameters selected in each fold for the *CD4-positive, alpha-beta T cell* subset.

| Fold | Model | Selected hyperparameters |
|---|---|---|
| 1 | LogisticRegression | $C = 0.4862$, penalty=l2, solver=lbfgs, max_iter=1000 |
| 1 | KNN | $n_neighbors = 6$, weights=uniform |
| 1 | SVM | $C = 0.1440$, kernel=linear |
| 1 | RandomForest | $n_estimators = 200$, max_depth=20 |
| 1 | MLP | hidden_layer_sizes=(50,), activation=relu, alpha=0.055574, solver=adam |
| 2 | LogisticRegression | $C = 0.2965$, penalty=l2, solver=lbfgs, max_iter=1000 |
| 2 | KNN | $n_neighbors = 4$, weights=distance |
| 2 | SVM | $C = 0.4973$, kernel=linear |
| 2 | RandomForest | $n_estimators = 200$, max_depth=8 |
| 2 | MLP | hidden_layer_sizes=(50,50), activation=relu, alpha=0.000180, solver=adam |
| 3 | LogisticRegression | $C = 3.3757$, penalty=l2, solver=lbfgs, max_iter=1000 |
| 3 | KNN | $n_neighbors = 2$, weights=uniform |
| 3 | SVM | $C = 1.4089$, kernel=linear |
| 3 | RandomForest | $n_estimators = 100$, max_depth=11 |
| 3 | MLP | hidden_layer_sizes=(50,), activation=relu, alpha=0.020910, solver=adam |
| 4 | LogisticRegression | $C = 0.7981$, penalty=l2, solver=lbfgs, max_iter=1000 |
| 4 | KNN | $n_neighbors = 4$, weights=uniform |
| 4 | SVM | $C = 0.7071$, kernel=linear |
| 4 | RandomForest | $n_estimators = 100$, max_depth=11 |
| 4 | MLP | hidden_layer_sizes=(50,50), activation=relu, alpha=0.003023, solver=adam |
| 5 | LogisticRegression | $C = 0.8459$, penalty=l2, solver=lbfgs, max_iter=1000 |
| 5 | KNN | $n_neighbors = 10$, weights=distance |
| 5 | SVM | $C = 0.1724$, kernel=linear |
| 5 | RandomForest | $n_estimators = 300$, max_depth=8 |
| 5 | MLP | hidden_layer_sizes=(50,50), activation=relu, alpha=0.000193, solver=adam |

Table 5.4: Fold-wise classification performance for the *CD4-positive, alpha-beta T cell* subset.

| Fold | Model | Accuracy | Sensitivity | Specificity | F1-score | Time (s) |
|---|---|---|---|---|---|---|
| 1 | LogisticRegression | 0.8861 | 0.8749 | 0.8749 | 0.6775 | 3.1533 |
| 1 | KNN | 0.8582 | 0.8483 | 0.8483 | 0.6398 | 1.4785 |
| 1 | SVM | 0.8629 | 0.8508 | 0.8508 | 0.6451 | 7.0130 |
| 1 | RandomForest | 0.8190 | 0.8515 | 0.8515 | 0.6067 | 114.3384 |
| 1 | MLP | 0.9139 | 0.9073 | 0.9073 | 0.7256 | 150.4664 |
| 2 | LogisticRegression | 0.8558 | 0.9005 | 0.9005 | 0.6516 | 3.1173 |
| 2 | KNN | 0.8623 | 0.8624 | 0.8624 | 0.6478 | 1.7482 |
| 2 | SVM | 0.8558 | 0.8827 | 0.8827 | 0.6469 | 7.0920 |
| 2 | RandomForest | 0.8000 | 0.8594 | 0.8594 | 0.5942 | 105.5678 |
| 2 | MLP | 0.8837 | 0.9033 | 0.9033 | 0.6827 | 143.3167 |
| 3 | LogisticRegression | 0.8901 | 0.8770 | 0.8770 | 0.6832 | 3.5104 |
| 3 | KNN | 0.8135 | 0.8606 | 0.8606 | 0.6047 | 1.5918 |
| 3 | SVM | 0.8943 | 0.8792 | 0.8792 | 0.6892 | 6.3587 |
| 3 | RandomForest | 0.8355 | 0.8840 | 0.8840 | 0.6286 | 92.5012 |
| 3 | MLP | 0.8967 | 0.9042 | 0.9042 | 0.6996 | 140.0215 |
| 4 | LogisticRegression | 0.8925 | 0.9139 | 0.9139 | 0.6968 | 3.0881 |
| 4 | KNN | 0.8343 | 0.8715 | 0.8715 | 0.6244 | 1.6364 |
| 4 | SVM | 0.8854 | 0.8923 | 0.8923 | 0.6818 | 7.7385 |
| 4 | RandomForest | 0.8610 | 0.9211 | 0.9211 | 0.6622 | 92.6245 |
| 4 | MLP | 0.9026 | 0.9370 | 0.9370 | 0.7168 | 146.3889 |
| 5 | LogisticRegression | 0.9074 | 0.8682 | 0.8682 | 0.7038 | 3.0019 |
| 5 | KNN | 0.8729 | 0.8680 | 0.8680 | 0.6605 | 2.2285 |
| 5 | SVM | 0.8996 | 0.8820 | 0.8820 | 0.6972 | 7.5074 |
| 5 | RandomForest | 0.8605 | 0.8792 | 0.8792 | 0.6507 | 108.5508 |
| 5 | MLP | 0.9062 | 0.8736 | 0.8736 | 0.7038 | 150.4991 |

Table 5.5: Aggregated classification performance across folds for the *CD4-positive, alpha-beta T cell* subset.

| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **F1-score** |
|---|---|---|---|---|
| KNN | $0.8483 \pm 0.0240$ | $0.8621 \pm 0.0089$ | $0.8621 \pm 0.0089$ | $0.6354 \pm 0.0216$ |
| LogisticRegression | $0.8864 \pm 0.0189$ | $0.8869 \pm 0.0194$ | $0.8869 \pm 0.0194$ | $0.6826 \pm 0.0203$ |
| MLP | $0.9006 \pm 0.0113$ | $0.9051 \pm 0.0225$ | $0.9051 \pm 0.0225$ | $0.7057 \pm 0.0165$ |
| RandomForest | $0.8352 \pm 0.0265$ | $0.8790 \pm 0.0271$ | $0.8790 \pm 0.0271$ | $0.6285 \pm 0.0286$ |
| SVM | $0.8796 \pm 0.0193$ | $0.8774 \pm 0.0157$ | $0.8774 \pm 0.0157$ | $0.6720 \pm 0.0244$ |

Table 5.6: Comparative performance of the evaluated classifiers for the *CD4-positive, alpha-beta T cell* subset. The MLP was the only Pareto-optimal model and achieved the highest individual hypervolume score.

| Model | Accuracy | Sensitivity | Specificity | F1-score | HV score |
|---|---|---|---|---|---|
| KNN | 0.8483 | 0.8621 | 0.8621 | 0.6354 | 0.4006 |
| Logistic Regression | 0.8864 | 0.8869 | 0.8869 | 0.6826 | 0.4759 |
| MLP | **0.9006** | **0.9051** | **0.9051** | **0.7057** | **0.5206** |
| Random Forest | 0.8352 | 0.8790 | 0.8790 | 0.6285 | 0.4056 |
| SVM | 0.8796 | 0.8774 | 0.8774 | 0.6720 | 0.4551 |

Table 5.7: Hyperparameters of the final MLP selected for interpretability analysis.

| Hyperparameter | Value |
|---|---|
| Activation | ReLU |
| Alpha | 0.0893 |
| Hidden layer sizes | (50, 50) |
| Max iterations | 500 |
| Random state | 42 |
| Solver | Adam |

Table 5.8: Final train test performance metrics obtained by the MLP model.

| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **F1-score** |
|---|---|---|---|---|
| MLP | 0.8804 | 0.8976 | 0.8976 | 0.6774 |

Figure 5.4: Comparative macro-metrics of the evaluated classifiers in the *CD4-positive, alpha-beta T cell* subset. The MLP achieved the strongest overall balance across accuracy, sensitivity, specificity, and macro F1-score.

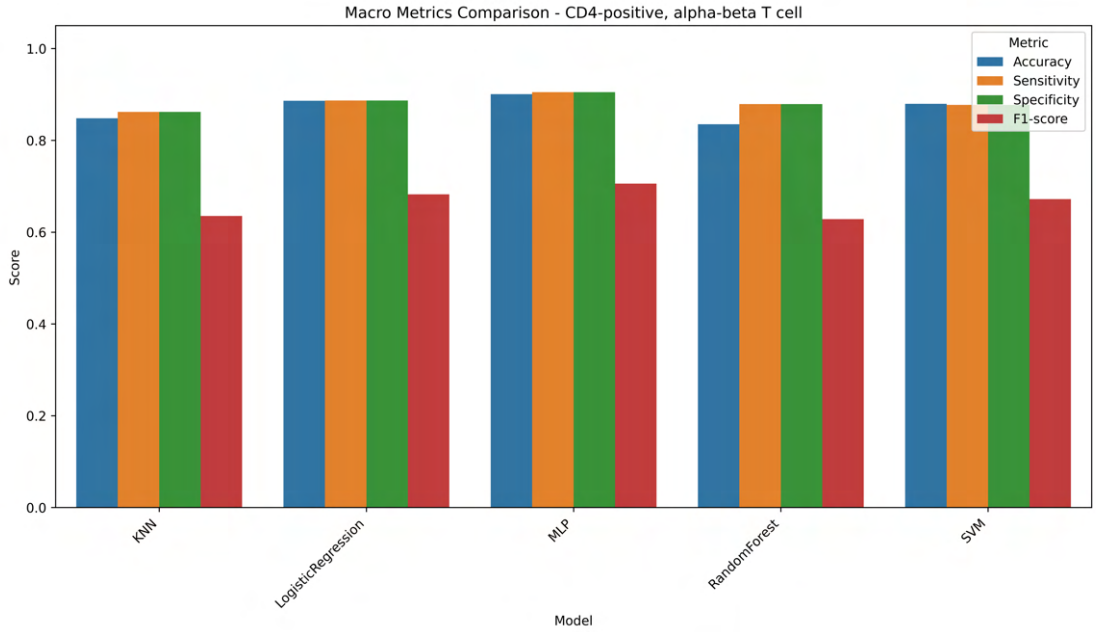



Figure 5.5: Confusion matrices obtained across the five cross-validation folds for the final MLP in the *CD4-positive, alpha-beta T cell* subset. The dominant main diagonal indicates stable classification across partitions, although minority-class errors persisted.

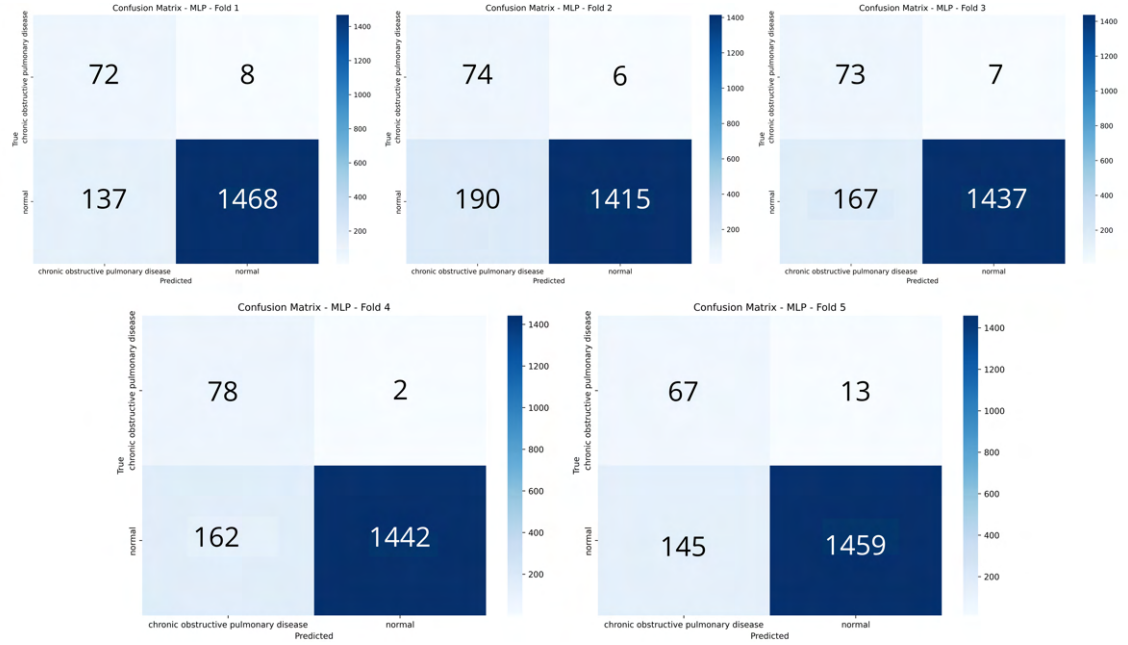


Figure 5.6: Independent train test evaluation of the final MLP in the *CD4-positive, alpha-beta T cell* subset. (A) Final confusion matrix of the selected MLP in the independent train test evaluation. (B) ROC curve of the final MLP model, showing the discriminative performance for both classes.

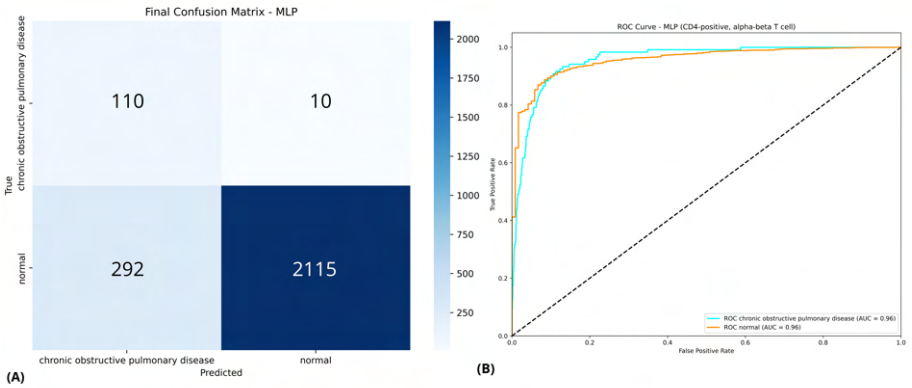



Figure 5.7: Comparison of feature-selection methods for CD4-positive, alpha-beta T cells. (A) UpSet plot of the top 25 genes selected by each method. (B) Venn diagram of overlaps among the top 25 genes from the various methods. (C) Heatmap of gene intersections shows ANOVA and random forest shared the most genes (15), followed by random forest and RFE (12), and the proposed method with RFE (11). The proposed method shared 9 genes with Wilcoxon and ANOVA, while Wilcoxon and RFE had the least overlap (4).

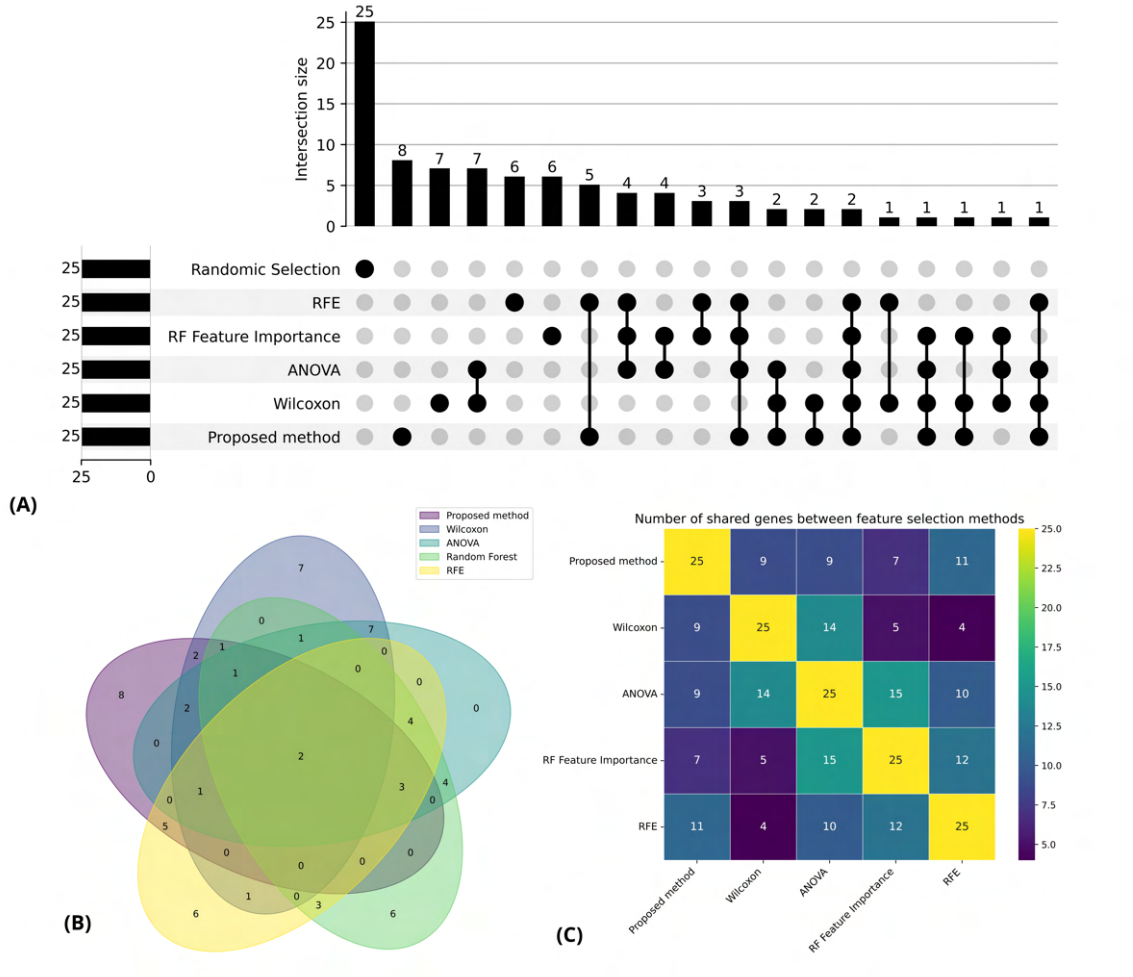


Figure 5.8: Comparison of feature-selection methods for CD4-positive, alpha-beta T cells. (A B C) Ranking of the top 25 genes based on Random Forest importance, RFE score, and ANOVA score, respectively. (D) Computational time across methods.

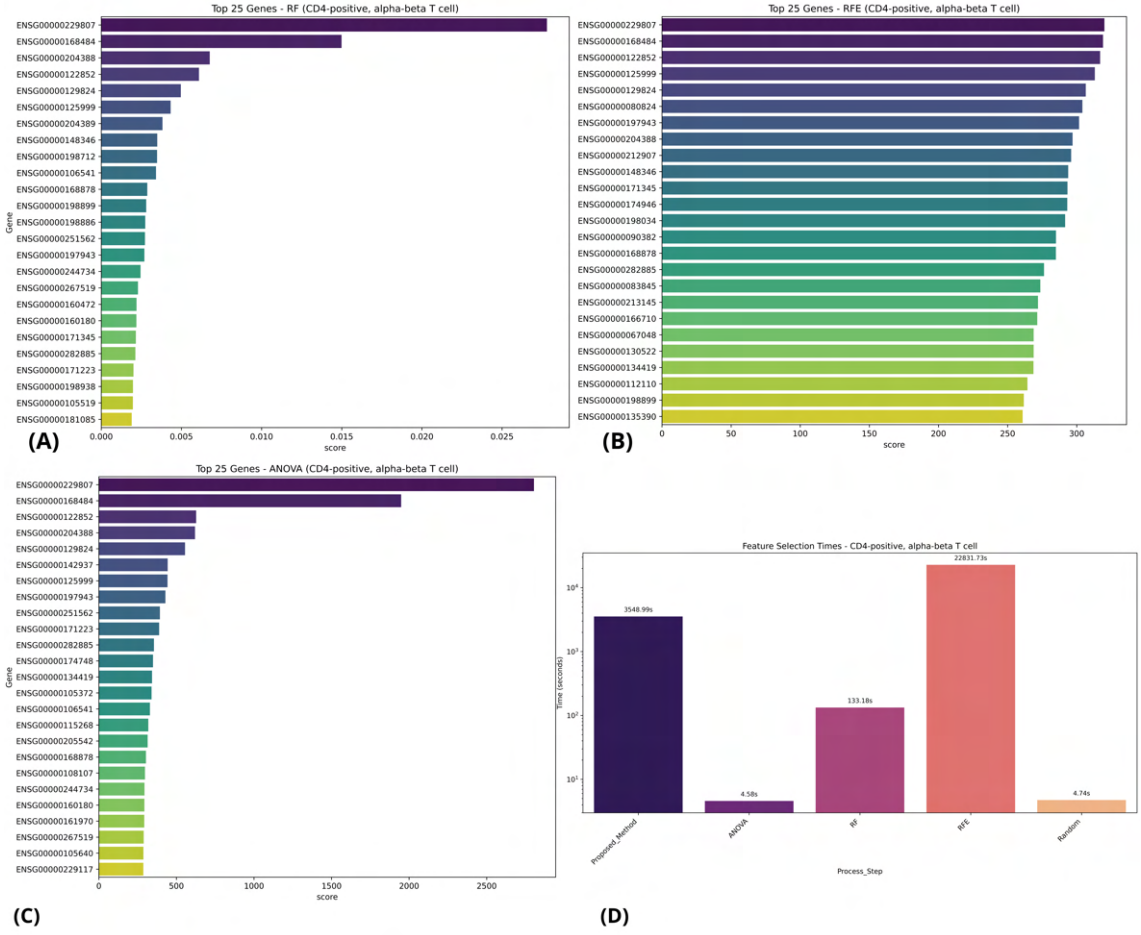
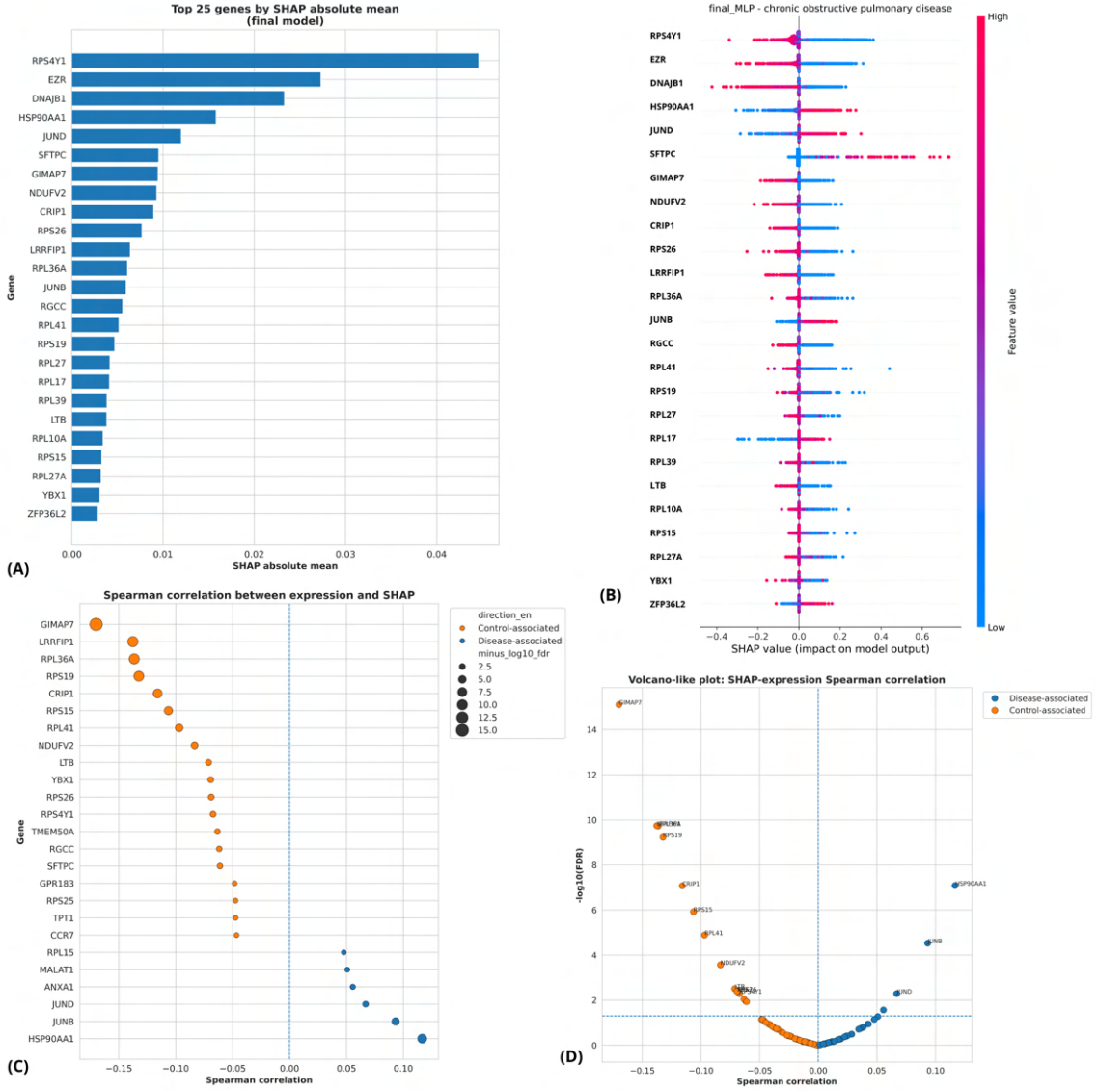


Figure 5.9: SHAP-based interpretability results for the final MLP in CD4-positive, alpha-beta T cells. (A) Mean absolute SHAP ranking of the top 25 genes in the final model. (B) SHAP beeswarm plot for the chronic obstructive pulmonary disease output. (C) Spearman correlation between gene expression and SHAP attribution. (D) Volcano-like representation of the SHAP expression Spearman correlations.



## 5.3 Interpretability results

### 5.3.1 SHAP global importance

This subsection presents the SHAP-based global feature-importance results for the best-performing classifier. The global SHAP ranking identified *RPS4Y1* as the most influential gene, followed by *DNAJB1*, *SFTPC*, *HSPA1B*, and *DDX3Y* (Table 5.9). Other highly ranked genes included *JUND*, *EZR*, *RPS26*, *PLCG2*, and *UBXN1*. The ranking also retained *XIST*, *RPS5*, *CRIP1*, *RPS8*, *MT-ND4L*, *GIMAP7*, *HSPA1A*, *JUNB*, *PTPRC*, *RPL41*, *RPL27*, *SMCHD1*, *YBX1*, *SLC25A6*, and *RPS15A* among the top 25 genes. These results show that the classifier relied on a compact set of genes with consistently high contributions to the model output.

Table 5.9 further shows that the top-ranked genes combined high mean absolute SHAP values with mean SHAP values close to zero. This pattern indicates that these genes strongly influenced predictions across samples but did not act in a uniform direction across all observations. Instead, the local contribution of each gene varied according to the molecular context of individual cells. Figure 5.9A supports this interpretation by showing the genes with the largest mean absolute SHAP values in the final model. In this panel, *RPS4Y1* remained the dominant feature, while *DNAJB1*, *SFTPC*, *JUND*, and *EZR* also showed strong global importance.

Because we computed SHAP values on the final selected feature matrix, the top SHAP genes overlapped with the genes retained during feature selection. Therefore, the global interpretability analysis confirmed that the classifier assigned the greatest attribution to genes already prioritized during the selection step. Taken together, these results demonstrate that the final model relied on a biologically coherent and internally consistent feature set.

### 5.3.2 Class-specific SHAP interpretation

The class-specific SHAP outputs showed that several genes contributed strongly to the chronic obstructive pulmonary disease output. Figure 5.9B presents the class-specific SHAP beeswarm plot for chronic obstructive pulmonary disease. In this panel, genes such as *RPS4Y1*, *EZR*, *DNAJB1*, *HSP90AA1*, *JUND*, and *SFTPC* displayed broad SHAP distributions, which indicates that these genes exerted strong and heterogeneous effects

Table 5.9: Top 25 genes ranked by mean absolute SHAP value, indicating their contribution to the model predictions.

| Gene | SHAP mean | SHAP std | Mean absolute SHAP | Rank |
|---|---|---|---|---|
| RPS4Y1 | $-3.5243 \times 10^{-19}$ | $2.3850 \times 10^{-18}$ | 0.0363211 | 1 |
| DNAJB1 | $3.7185 \times 10^{-19}$ | $1.4083 \times 10^{-18}$ | 0.0212272 | 2 |
| SFTPC | $4.3333 \times 10^{-20}$ | $6.6110 \times 10^{-19}$ | 0.0115264 | 3 |
| HSPA1B | $-6.3060 \times 10^{-20}$ | $1.6550 \times 10^{-18}$ | 0.0111047 | 4 |
| DDX3Y | $4.1907 \times 10^{-19}$ | 0.0 | 0.0104498 | 5 |
| JUND | $-1.6705 \times 10^{-19}$ | $1.1166 \times 10^{-18}$ | 0.0097059 | 6 |
| EZR | $3.1757 \times 10^{-19}$ | $1.2255 \times 10^{-18}$ | 0.0092657 | 7 |
| RPS26 | $-4.6467 \times 10^{-19}$ | $1.1991 \times 10^{-18}$ | 0.0083718 | 8 |
| PLCG2 | $2.5495 \times 10^{-19}$ | 0.0 | 0.0077492 | 9 |
| UBXN1 | $6.2825 \times 10^{-19}$ | $6.4192 \times 10^{-19}$ | 0.0071792 | 10 |
| XIST | $1.0362 \times 10^{-19}$ | $5.4864 \times 10^{-19}$ | 0.0062192 | 11 |
| RPS5 | $2.6427 \times 10^{-19}$ | $1.1429 \times 10^{-19}$ | 0.0061579 | 12 |
| CRIP1 | $2.9022 \times 10^{-19}$ | $9.5716 \times 10^{-19}$ | 0.0059109 | 13 |
| RPS8 | $-2.5533 \times 10^{-19}$ | $1.1048 \times 10^{-18}$ | 0.0054927 | 14 |
| MT-ND4L | $1.9592 \times 10^{-19}$ | $9.9325 \times 10^{-20}$ | 0.0051635 | 15 |
| GIMAP7 | $1.5790 \times 10^{-19}$ | $5.9510 \times 10^{-19}$ | 0.0049548 | 16 |
| HSPA1A | $4.4034 \times 10^{-19}$ | $5.6989 \times 10^{-19}$ | 0.0049546 | 17 |
| JUNB | $-3.1062 \times 10^{-19}$ | $8.2395 \times 10^{-19}$ | 0.0048691 | 18 |
| PTPRC | $-1.0834 \times 10^{-18}$ | 0.0 | 0.0046386 | 19 |
| RPL41 | $-2.1033 \times 10^{-19}$ | $7.7998 \times 10^{-19}$ | 0.0045583 | 20 |
| RPL27 | $5.8678 \times 10^{-20}$ | $5.8911 \times 10^{-19}$ | 0.0044920 | 21 |
| SMCHD1 | $5.5870 \times 10^{-19}$ | 0.0 | 0.0044080 | 22 |
| YBX1 | $4.2308 \times 10^{-19}$ | $7.0556 \times 10^{-19}$ | 0.0043728 | 23 |
| SLC25A6 | $-1.2540 \times 10^{-19}$ | $5.0597 \times 10^{-19}$ | 0.0039962 | 24 |
| RPS15A | $1.5132 \times 10^{-19}$ | $5.4844 \times 10^{-19}$ | 0.0037027 | 25 |

across observations.

Figure 5.9B further shows that high expression values of some genes shifted SHAP values toward the chronic obstructive pulmonary disease class, whereas negative SHAP values reduced support for that output and therefore favored the alternative class. This result indicates that the model used a shared core of discriminative genes, but modulated their local contributions according to the transcriptional context of each observation. In particular, genes such as *SFTPC*, *RPS4Y1*, and *DNAJB1* produced strong positive and negative attributions across cells, which confirms that the final model did not rely only on globally important features, but also on structured local attribution patterns that separated chronic obstructive pulmonary disease from normal samples.

### 5.3.3 Correlation between SHAP values and gene expression

To further interpret the model, we evaluated the Spearman correlation between gene expression and SHAP attribution. Figure 5.9C presents the dot plot of the strongest expression-attribution correlations, whereas Figure 5.9D presents the corresponding volcano-like representation. These panels show that several genes exhibited negative correlations, including *GIMAP7*, *LRRFIP1*, *RPL36A*, *RPS19*, *CRIP1*, and *RPS15*. In contrast, other genes showed positive correlations, including *HSP90AA1*, *JUNB*, *JUND*, and *ANXA1*.

Figure 5.9C shows that negative correlations predominated among the strongest associations, which suggests that higher expression of several genes reduced disease-associated attribution or reinforced control-associated prediction. In contrast, the positively correlated genes clustered on the right side of the plot, indicating that higher expression contributed more to disease prediction. Figure 5.9D reinforces this interpretation by showing that the most significant genes combined relatively large effect sizes with strong statistical support. Together, these results indicate that the classifier did not uniformly favor highly expressed genes or lowly expressed genes. Instead, it learned structured monotonic relationships in which the predictive contribution of a gene depended on both its expression level and the class context.

Overall, the SHAP analysis demonstrated that the final classifier relied on a stable group of globally important genes, assigned class-dependent effects to these genes, and captured biologically meaningful relationships between gene expression and predictive contribution. Together, Table 5.9 and Figure 5.9A D show that the model achieved not only strong predictive performance but also a coherent and interpretable decision struc-

ture.

## 5.4 Comparative low-dimensional structure preservation

### 5.4.1 Comparative low-dimensional structure preservation and full-space class separation

Figure 5.10 compares the geometric organization induced by the feature subsets selected by ANOVA, the proposed method, Random Forest, RFE, Wilcoxon, and random selection. Panels (A) (F) show the PCA projections reconstructed from the features selected by each method. In contrast, panel (G) summarizes the quantitative comparison of class separation in the full standardized feature space using the separation ratio, defined as the centroid distance normalized by the pooled within-class dispersion. This distinction is critical for interpretation. The PCA panels provide a qualitative two-dimensional representation of the data, while panel (G) reports the quantitative metric used for method comparison in the original multivariate space after standardization.

A first important result is that the visual arrangement of the classes in PCA did not fully reproduce the ranking obtained from the full-space separation ratio. This outcome is methodologically expected. PCA is a linear projection that compresses the multivariate structure of the selected features into two orthogonal components that maximize variance, not necessarily class discrimination. Therefore, visual separation in the PCA plane should be interpreted as an exploratory summary of the dominant low-dimensional structure, rather than as a direct measure of global class separability. In contrast, the separation ratio shown in panel (G) was computed in the full standardized feature space and incorporates both the distance between class centroids and the internal spread of the classes. For this reason, panel (G) provides a more rigorous basis for quantitative comparison across feature selection methods.

Among the evaluated approaches, ANOVA produced the highest separation ratio in the full standardized space ($1.402$), followed by Random Forest ($1.346$), RFE ($1.089$), the proposed method ($1.029$), and Wilcoxon ($1.003$). Random selection yielded a substantially lower ratio ($0.175$), indicating very limited structured class separation. These values show that ANOVA and Random Forest generated the feature subsets with the strongest balance between inter-class separation and intra-class compactness. In contrast, the proposed method and Wilcoxon produced more modest global discrimination under this spe-

cific geometric criterion. Importantly, the proposed method did not exhibit a collapse in the centroid distance itself. Rather, its lower ratio resulted from the combination of moderate centroid separation and relatively high pooled dispersion, which reduced the normalized separation score.

This point becomes clearer when panel (B) is compared with panel (G). In the PCA projection of the proposed method, the two classes appear visually displaced, and the centroids are clearly separated in the projected plane. At first sight, this could suggest stronger class discrimination than that reflected in the quantitative ranking. However, this apparent discrepancy does not indicate an error. Instead, PCA emphasizes the directions of the largest variance in two dimensions, whereas the separation ratio summarizes the relationship between centroid displacement and class spread across the entire standardized feature space. In the proposed method, the projected centroids are visually distinct, but the class clouds remain relatively broad and partially overlap. Consequently, once within-class dispersion is taken into account in the full feature space, the normalized separation becomes only slightly greater than 1.0, placing this method below ANOVA, Random Forest, and RFE.

Negative Silhouette values observed in some cell type method combinations should not be interpreted as a contradiction of the predictive results, but rather as evidence that class labels do not always coincide with a simple cluster structure in the analyzed feature space. In these cases, disease and control cells may remain locally intermixed even when the selected features still support useful decision boundaries for supervised classification. This distinction is important because Silhouette quantifies geometric cohesion and separation under a distance-based view of the data, whereas the predictive models evaluated in this study may exploit more complex class structure, including non-linear or distributed signal that is not fully captured by centroid-based or low-dimensional geometric summaries.

A similar reasoning applies to Wilcoxon in panel (E). Although the PCA projection shows an organized global structure with visible directional separation, the quantitative ratio in panel (G) remained close to unity ($1.003$). This indicates that the visual distinction observed in two dimensions was not accompanied by a strong gain in centroid separation relative to pooled dispersion in the multivariate space. Thus, the PCA representation alone would overestimate the effective class separation if interpreted without the full-space metric.

By contrast, ANOVA in panel (A) and Random Forest in panel (C) yielded the

highest quantitative separation ratios, even though their PCA panels do not necessarily produce the most immediately striking visual separation at first inspection. Their superior performance in panel (G) indicates that these methods selected feature subsets in which the class centroids were more favorably displaced relative to internal class variability in the full standardized space. In other words, they achieved a better geometric trade-off between separation and compactness. This is precisely why the ratio-based comparison is more informative than visual inspection alone. A method may appear visually separated in two dimensions but still perform worse quantitatively if the selected feature space also produces broad, diffuse, or heterogeneous class distributions.

RFE in panel (D) occupied an intermediate position. Its separation ratio ($1.089$) was lower than that of ANOVA and Random Forest, but still higher than those of the proposed method and Wilcoxon. This suggests that RFE retained features capable of producing some degree of centroid displacement with moderate control of class dispersion, although not enough to match the leading methods under this metric. The proposed method, in turn, remained close to RFE, indicating that its selected feature subset preserved non-random class structure but did not achieve the strongest global geometric discrimination among the tested strategies.

The random selection baseline, shown in panel (F), provides an important reference for interpretation. Its PCA projection showed a highly unstable, poorly structured geometry, and its separation ratio was markedly lower than those of all informed methods. This result confirms that the feature subsets selected by the biologically and statistically guided methods captured the true transcriptomic structure relevant to the classification problem. In contrast, randomly chosen genes failed to preserve meaningful class organization. Therefore, even methods with moderate ratios, such as the proposed method and the Wilcoxon method, still performed substantially better than a null feature-selection baseline.

Taken together, these results support three methodological conclusions. First, low-dimensional PCA projections are useful for qualitative inspection of the structure induced by each selected feature subset. Still, they should not be used as the sole basis for ranking feature selection methods. Second, the full standardized-space separation ratio provides a more appropriate quantitative summary because it jointly considers between-class centroid displacement and within-class dispersion. Third, under this criterion, the proposed method preserved a detectable and non-random disease-related structure. Still, its selected features did not achieve as effective global class separation as ANOVA and Ran-

dom Forest. The lower ranking of the proposed method in panel (G) should therefore be interpreted not as a contradiction of panel (B), but rather as evidence that visual separation in two dimensions was insufficient to compensate for the broader multivariate dispersion associated with its feature subset.

From a biological and analytical perspective, this finding is relevant because it shows that different feature selection methods may emphasize different properties of the data. Some methods favor a broader variance structure that produces visually expressive PCA organization. In contrast, others may prioritize features that yield tighter class compactness and stronger centroid-based discrimination in the original standardized space. Accordingly, the proposed method should not be judged solely based on centroid separation. Its full evaluation must also consider classification performance, stability, interpretability, and downstream biological coherence, since geometric separation represents only one aspect of feature subset quality.

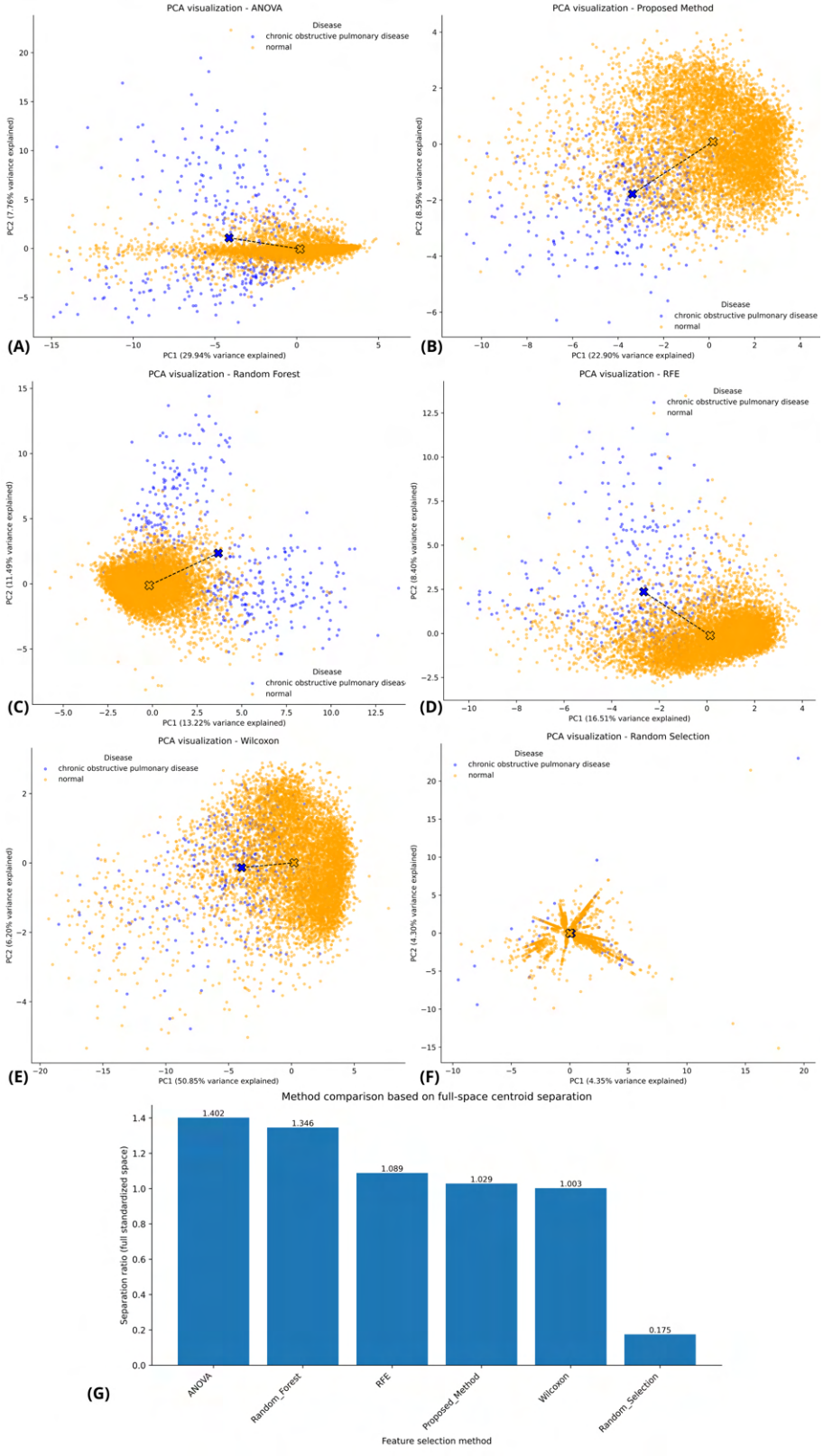
## 5.5 Functional enrichment and biological validation

### 5.5.1 Functional enrichment of selected genes by each method

To biologically validate the features selected by each method, we performed a functional enrichment analysis of the top 25 genes identified by the proposed method, Wilcoxon, ANOVA, random forest feature importance, and RFE. For the proposed method, we defined the final feature set as the genes ranked by aggregated SHAP feature importance, thereby selecting the genes with the highest aggregated SHAP importance across folds. We then submitted the top-ranked genes from each method to ShinyGO. We investigated Gene Ontology categories for Biological Process and Molecular Function, together with protein protein interaction analysis in STRINGdb. For the STRINGdb analysis, we set the minimum required interaction score to high confidence (0.700) to retain only strongly supported protein protein interactions. Before enrichment, we converted Ensembl gene identifiers into gene symbols using the BioTools web tool to ensure compatibility with downstream annotation resources. We further supported the biological interpretation of enriched terms using annotations from the Gene Ontology resource.

Figure 5.11 summarizes the interaction networks and enriched functional categories obtained for each feature-selection strategy. The proposed method produced a compact interaction network with a central cluster enriched for ribosomal and transla-

Figure 5.10: (A)-(F) show PCA projections reconstructed from the features selected by each method. Colored points represent samples from the two disease groups, and the markers indicate class centroids in the projected space. (G) reports the separation ratio computed in the full standardized feature space, defined as centroid distance normalized by pooled within-class dispersion.



tional genes, including *RPS4Y1*, *RPS5*, *RPS8*, *RPS26*, *RPL27*, *RPL41*, and *RPS15A*, together with smaller subnetworks involving stress-response and chaperone-related genes such as *HSPA1A*, *HSPA1B*, and *DNAJB1*. Consistent with this structure, Figure 5.11A shows that the proposed method significantly enriched biological processes related to *cytoplasmic translation*, *translation*, and *chaperone cofactor-dependent protein refolding*. It also identified terms linked to stress adaptation and regulatory programs, including *negative regulation of transcription from RNA polymerase II promoter in response to stress*, *regulation of mitotic spindle assembly*, and *cellular heat acclimation*. These findings indicate that the SHAP-derived feature set captured a biologically coherent combination of translational, stress-response, and regulatory signals.

Wilcoxon identified a denser, broader interaction network dominated by ribosomal proteins and heat-shock-related genes. Accordingly, Figure 5.11B shows that Wilcoxon strongly enriched broad biosynthetic and translational processes, including *cytoplasmic translation*, *translation*, *macromolecule biosynthetic process*, *cellular nitrogen compound biosynthetic process*, and *cellular macromolecule metabolic process*. This pattern suggests that Wilcoxon favored genes involved in general cellular biosynthesis and housekeeping activity. Although these processes remain biologically relevant, they provide less specific mechanistic resolution than the more structured enrichment profile observed for the proposed method.

ANOVA recovered an interaction network that also emphasized ribosomal genes, but included a more limited set of additional markers. Figure 5.11C shows that ANOVA enriched *cytoplasmic translation*, *translation*, *gene expression*, *cellular biosynthetic process*, and *organic substance biosynthetic process*. In addition, ANOVA uniquely highlighted *respiratory gaseous exchange by respiratory system*, which may reflect the pulmonary context of the dataset. However, the overall profile remained centered on broad biosynthetic functions, again indicating a stronger emphasis on general cellular activity than on integrated stress and regulatory programs.

Random forest feature importance yielded a more fragmented interaction network with smaller modules linked to mitochondrial and stress-associated genes. Figure 5.11D shows that this method enriched processes related to *regulation of spindle assembly*, *regulation of mitotic spindle assembly*, *positive regulation of endoribonuclease activity*, and *oxidative phosphorylation*. These results suggest that random forest prioritized features associated with cell-cycle regulation, post-transcriptional control, and mitochondrial metabolism. Compared with the proposed method, however, the resulting functional pro-

file appeared narrower and less connected to a dominant translational core.

RFE generated the sparsest enrichment profile among the evaluated strategies. Its interaction network included only a few small modules, and Figure 5.11E shows that it enriched only two molecular functions: *rRNA binding* and *structural constituent of ribosome*. This result indicates that RFE concentrated primarily on ribosome-related features and captured a more restricted functional space than the other methods.

Taken together, these results show that all methods identified genes involved in translation-related biology, but the proposed method combined this signal with stress-response, chaperone-mediated refolding, apoptotic regulation, and spindle-related processes. This enrichment pattern supports the biological relevance of the SHAP-aggregated features and suggests that the proposed method captured functional signals beyond generic housekeeping programs.

Table 5.10: Comparison of protein protein interaction (PPI) network metrics and the number of enriched Gene Ontology Biological Process (GO BP) terms for the gene sets selected by each feature-selection method. Wilcoxon produced the most connected PPI network, whereas the proposed method recovered the largest number of enriched biological process terms, indicating broader functional coverage.

| Method | Number of edges | Average node degree | Avg. local clustering coefficient | PPI enrichment p-value | GO-BP |
|---|---|---|---|---|---|
| Proposed method | 144 | 12.5 | 0.846 | $< 1.0 \times 10^{-16}$ | 47 |
| Wilcoxon | 166 | 13.8 | 0.895 | $< 1.0 \times 10^{-16}$ | 24 |
| ANOVA | 44 | 4.0 | 0.657 | $< 1.0 \times 10^{-16}$ | 7 |
| RF | 11 | 1.0 | 0.500 | $2.99 \times 10^{-7}$ | 4 |
| RFE | 15 | 1.25 | 0.562 | $7.13 \times 10^{-4}$ | 0 |

### 5.5.2 Functional enrichment of positively disease-associated genes selected by the proposed method

We performed functional enrichment analysis on the 25 disease-associated genes selected by the proposed method to investigate the biological processes represented by the positively contributing features. As shown in Figure 5.11, the protein protein interaction network formed a highly connected module dominated by ribosomal and translation-related genes, indicating strong functional coherence among the selected disease-associated features (Figure 5.11A). In particular, several ribosomal proteins clustered densely in the core of the network, whereas genes such as *HSP90AA1*, *HSPA1A*, *HSPA1B*, *ENO1*, *JUNB*, and *TAGLN2* appeared as functionally connected but more peripheral components. This organization suggests that the proposed method identified a coordinated disease-associated signature centered on genes involved in translational machinery and stress re-

sponse.

Gene Ontology biological process enrichment further supported this interpretation (Figure 5.11B). The most strongly enriched terms were *cytoplasmic translation* and *translation*, which showed the largest signal values and the lowest false discovery rates. We also identified significant enrichment for broader biosynthetic processes, including *macromolecule biosynthetic process* and *cellular nitrogen compound biosynthetic process*. In addition, the analysis highlighted more specific processes related to ribosome function and post-transcriptional regulation, such as *ribosomal large subunit biogenesis* and *rRNA processing*. Beyond these translation-centered terms, the enriched set also included pathways associated with protein homeostasis and stress adaptation, including *regulation of protein ubiquitination*, *negative regulation of protein ubiquitination*, and *positive regulation of nuclease activity*. We further observed enrichment for *regulation of intrinsic apoptotic signaling pathway*, suggesting that the disease-associated genes may also participate in survival- and stress-related cellular responses.

Taken together, these results show that the disease-associated genes selected by the proposed method converged on a biologically coherent program dominated by translation, ribosome-associated activity, and protein homeostasis. This pattern indicates that the proposed method did not merely capture predictive markers but also prioritized genes involved in coordinated molecular processes plausibly linked to disease-associated dysfunction in CD4-positive, alpha-beta T cells.

Table 5.10 presents a summary of the comparative biological profiles obtained from the evaluated feature-selection methods. The Wilcoxon method produced the most interconnected protein-protein interaction (PPI) network, featuring the highest number of edges (166), the greatest average node degree (13.8), and the highest local clustering coefficient (0.895), indicating superior network cohesion compared to all other methods. The proposed method ranked second in PPI connectivity, with 144 edges, an average node degree of 12.5, and a local clustering coefficient of 0.846. Notably, it also identified the largest number of enriched Gene Ontology (GO) Biological Process terms (47), surpassing Wilcoxon (24), ANOVA (7), RF (4), and RFE (0). ANOVA demonstrated intermediate performance, with 44 edges, an average node degree of 4.0, and a clustering coefficient of 0.657, whereas RF and RFE produced sparse interaction networks with limited functional enrichment. While all methods achieved statistically significant PPI enrichment, the network structure of Wilcoxon and the broader GO-BP coverage of the proposed method underscore their unique strengths. Overall, these results indicate that

Wilcoxon excels at maximizing interaction network cohesion, while the proposed method effectively captures a broader functional landscape.

## 5.6 Global comparative analysis across the remaining cell types

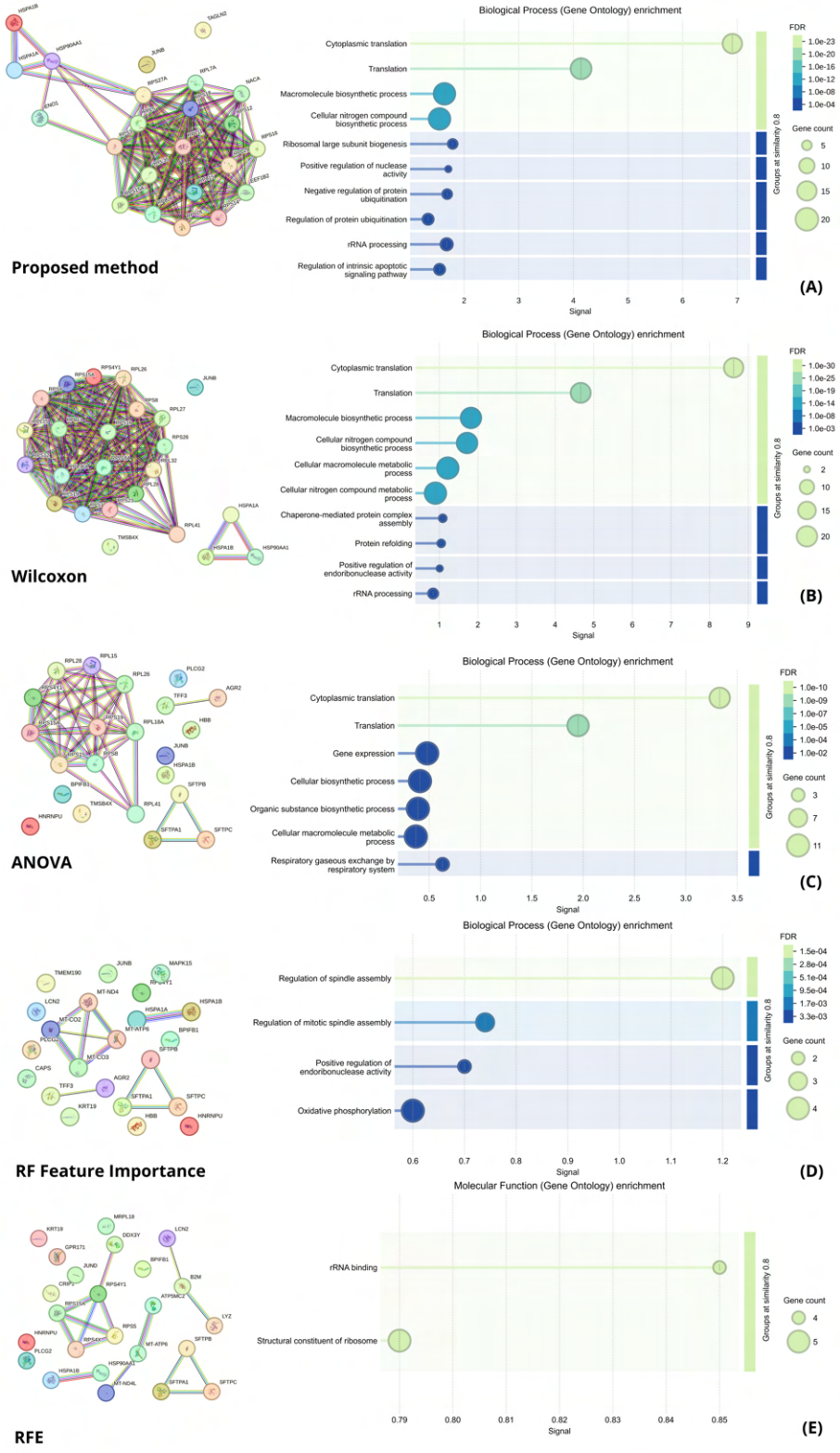
### 5.6.1 Coverage of the analyzed cell types

To contextualize the global comparative analysis across the remaining cell types, we first summarized the coverage and class composition of all cell populations retained for downstream evaluation. As shown in Table 5.11, the male dataset comprised 27 analyzed cell types and 58,987 cells, including 51,190 *normal* cells and 7,797 *disease* cells. Thus, the global class distribution was markedly imbalanced, with *normal* cells representing the large majority of observations. In addition to this overall imbalance, the dataset showed substantial heterogeneity in cell-type abundance. The largest subsets were *CD4-positive, alpha-beta T cell* (8,422 cells), *multiciliated columnar cell of tracheobronchial tree* (7,348 cells), and *monocyte* (4,312 cells). In contrast, the smallest subsets were *fibroblast of lung* (172 cells), *pericyte* (188 cells), and *bronchiolar smooth muscle cell* (283 cells).

Figure 5.12 provides a normalized view of class composition across cell types by showing the relative proportions of *normal* and *disease* cells in each subset. Most cell types displayed a clear predominance of *normal* cells, although the degree of imbalance varied considerably. This pattern is quantitatively confirmed in Table 5.11, where the imbalance ratio ranged from 1.201 in *mast cell*, the most balanced subset, to 38.782 in *tracheobronchial serous cell*, the most imbalanced subset. Strong imbalance was also observed for *macrophage* (34.080), *conventional dendritic cell* (31.194), *fibroblast* (27.350), *monocyte* (21.576), and *CD4-positive, alpha-beta T cell* (20.055). In contrast, *mast cell*, *respiratory tract goblet cell* (1.703), *multiciliated columnar cell of tracheobronchial tree* (2.156), and *T cell* (2.199) showed comparatively more balanced class distributions.

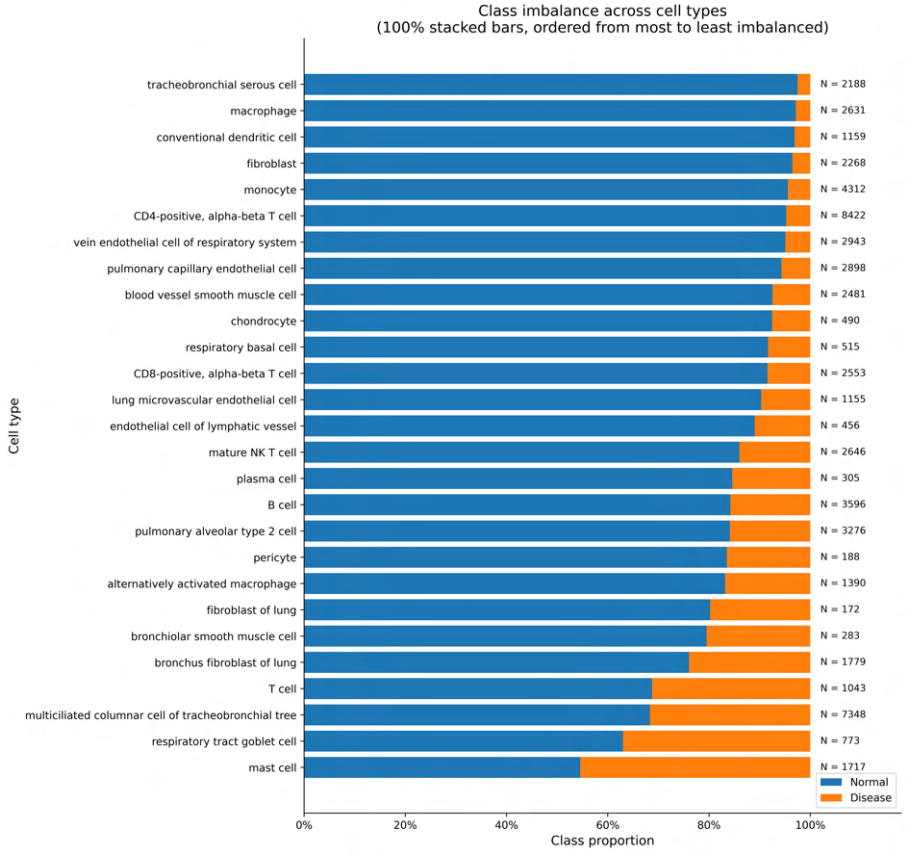
Taken together, Figure 5.12 and Table 5.11 show that the analyzed cell types differed not only in sample size but also in the severity of class imbalance. These differences are important for interpreting downstream classification results, because subsets with fewer cells or stronger imbalance are expected to pose greater difficulty for model training, yield more unstable estimates of predictive performance, and depend more on evaluation metrics such as the F1-score. For completeness, it is important to distinguish the scopes of the global result summaries. Table 5.11 reports the 27 cell types retained for

Figure 5.11: Functional enrichment and protein-protein interaction analysis were conducted on the top 25 genes from each feature-selection strategy. The left panels show STRINGdb interaction networks with a high-confidence threshold (0.700), while the right panels display significant Gene Ontology enrichment results. (A) refers to the proposed method using SHAP for feature importance, and (B-E) represent Wilcoxon, ANOVA, random forest, and RFE methods, respectively. The proposed method yielded a compact network enriched in translational, stress-response, and regulatory processes, unlike the other methods, which highlighted broader biosynthetic or ribosome-related functions.



descriptive coverage and class-composition analysis, whereas the predictive comparison tables reported in the next subsection include 25 cell types with complete classifier outputs. In this final predictive comparison, *pulmonary capillary endothelial cell* and *vein endothelial cell of respiratory system* are not included.

Figure 5.12: Class imbalance across cell types in the dataset represented as 100% stacked horizontal bar plots. Each bar shows the relative proportions of normal and disease cells within one cell type, ordered from the most imbalanced to the least imbalanced, with the total number of cells indicated on the right. Most cell types showed clear predominance of normal cells, although the magnitude of imbalance varied substantially. Mast cells were the most balanced group, whereas tracheobronchial serous cells, macrophages, fibroblasts, monocytes, and CD4-positive, alpha-beta T cells showed the strongest imbalance. These results demonstrate that class composition differed markedly across cell types and may influence downstream predictive analyses.




### 5.6.2 Predictive performance across cell types

Table 5.12 summarizes the stratified k-fold macro F1-scores obtained for all evaluated classifiers across the analyzed cell types, together with the hypervolume score and the final model selected by the multiobjective criterion. Overall, predictive performance varied markedly across cell types, indicating substantial heterogeneity in classification

Table 5.11: Class distribution across cell types in the male dataset. The table reports the numbers of normal and disease cells, the total number of cells, and the imbalance ratio for each cell type, where higher values indicate stronger class imbalance.

| **Cell type** | **Normal** | **Disease** | **Total cells** | **Imbalance ratio** |
|---|---|---|---|---|
| fibroblast of lung | 138 | 34 | 172 | 4.059 |
| pericyte | 157 | 31 | 188 | 5.065 |
| bronchiolar smooth muscle cell | 225 | 58 | 283 | 3.879 |
| plasma cell | 258 | 47 | 305 | 5.489 |
| endothelial cell of lymphatic vessel | 406 | 50 | 456 | 8.120 |
| chondrocyte | 453 | 37 | 490 | 12.243 |
| respiratory basal cell | 472 | 43 | 515 | 10.977 |
| respiratory tract goblet cell | 487 | 286 | 773 | 1.703 |
| T cell | 717 | 326 | 1043 | 2.199 |
| lung microvascular endothelial cell | 1043 | 112 | 1155 | 9.312 |
| conventional dendritic cell | 1123 | 36 | 1159 | 31.194 |
| alternatively activated macrophage | 1156 | 234 | 1390 | 4.940 |
| mast cell | 937 | 780 | 1717 | 1.201 |
| bronchus fibroblast of lung | 1353 | 426 | 1779 | 3.176 |
| tracheobronchial serous cell | 2133 | 55 | 2188 | 38.782 |
| fibroblast | 2188 | 80 | 2268 | 27.350 |
| blood vessel smooth muscle cell | 2296 | 185 | 2481 | 12.411 |
| CD8-positive, alpha-beta T cell | 2337 | 216 | 2553 | 10.819 |
| macrophage | 2556 | 75 | 2631 | 34.080 |
| mature NK T cell | 2276 | 370 | 2646 | 6.151 |
| pulmonary capillary endothelial cell | 2733 | 165 | 2898 | 16.564 |
| vein endothelial cell of respiratory system | 2798 | 145 | 2943 | 19.297 |
| pulmonary alveolar type 2 cell | 2756 | 520 | 3276 | 5.300 |
| B cell | 3029 | 567 | 3596 | 5.342 |
| monocyte | 4121 | 191 | 4312 | 21.576 |
| multiciliated columnar cell of tracheobronchial tree | 5020 | 2328 | 7348 | 2.156 |
| CD4-positive, alpha-beta T cell | 8022 | 320 | 8342 | 25.007 |

difficulty beyond the detailed CD4-positive, alpha-beta T-cell case study. The strongest results were observed for *pulmonary alveolar type 2 cell*, in which SVM achieved the highest F1-score ($0.9718$) and the largest hypervolume score ($0.9317$), followed by *alternatively activated macrophage*, where Logistic Regression reached an F1-score of $0.9737$ and an HV-score of $0.9288$. High-performance subsets also included *respiratory tract goblet cell*, *B cell*, *mast cell*, and *multiciliated columnar cell of tracheobronchial tree*, all of which showed consistently high macro F1-scores across multiple classifiers.

In contrast, a smaller group of cell types remained considerably more challenging. The lowest hypervolume scores were observed for *plasma cell* ($HV = 0.1096$), *pericyte* ($HV = 0.1990$), and *endothelial cell of lymphatic vessel* ($HV = 0.2753$), indicating weaker and less stable predictive performance across candidate models. Likewise, *respiratory basal cell*, *tracheobronchial serous cell*, and *macrophage* also showed relatively modest results, with best macro F1-scores below $0.70$. These patterns are consistent with the global heterogeneity in sample size and class imbalance described in the previous subsection and suggest that some cell populations provide a less separable signal for condition discrimination.

No single classifier dominated all cell types. Among the 25 cell types reported in Table 5.12, Random Forest was selected most frequently by the hypervolume criterion, being chosen for 9 cell types, followed by SVM for 7 cell types, MLP for 6, and Logistic Regression for 3, whereas KNN was not selected for any cell type. This result indicates that the optimal classification strategy depended on the specific structure of each cell population. At the same time, the hypervolume-based decision was highly consistent with the stratified k-fold macro F1-scores, since the selected model also achieved the highest F1-score in 23 of the 25 cell types. Only *T cell* and *bronchiolar smooth muscle cell* showed a small discrepancy between the model with the highest F1-score and the model selected by HV, which likely reflects the influence of the remaining performance metrics incorporated into the multiobjective selection procedure.

Taken together, these results show that predictive performance across cell types was heterogeneous but systematically captured by the proposed selection framework. Because the dataset remained imbalanced across most cell populations, macro F1-score provided the summary of classifier behavior, while the hypervolume criterion enabled a principled final choice by integrating performance across multiple objectives. This combined strategy therefore supported a cell-type-specific model selection process rather than assuming that a single classifier would be optimal for all biological subsets.

Table 5.12: Comparative predictive performance across cell types based on stratified k-fold macro F1-score for each classifier, together with the hypervolume score and the best model selected by the multiobjective criterion.

| Cell type | KNN | Logistic Regression | MLP | Random Forest | SVM | HV-score | Best model |
|---|---|---|---|---|---|---|---|
| CD4-positive, alpha-beta T cell | 0.6354 | 0.6826 | 0.7057 | 0.6285 | 0.6720 | 0.5206 | MLP |
| multiciliated columnar cell of tracheobronchial tree | 0.9226 | 0.9236 | 0.9609 | 0.9458 | 0.9277 | 0.7293 | MLP |
| monocyte | 0.7533 | 0.7508 | 0.7274 | 0.7296 | 0.7690 | 0.6268 | SVM |
| B cell | 0.8467 | 0.9196 | 0.9260 | 0.9421 | 0.9457 | 0.8446 | SVM |
| pulmonary alveolar type 2 cell | 0.9599 | 0.9692 | 0.9686 | 0.9248 | 0.9718 | 0.9317 | SVM |
| mature NK T cell | 0.6472 | 0.7509 | 0.7490 | 0.8109 | 0.7395 | 0.5672 | Random Forest |
| macrophage | 0.5900 | 0.6208 | 0.6145 | 0.5930 | 0.6247 | 0.4429 | SVM |
| CD8-positive, alpha-beta T cell | 0.7173 | 0.7448 | 0.7330 | 0.7591 | 0.7454 | 0.5401 | Random Forest |
| blood vessel smooth muscle cell | 0.5888 | 0.6930 | 0.6660 | 0.7146 | 0.6764 | 0.4701 | Random Forest |
| fibroblast | 0.8794 | 0.8902 | 0.8986 | 0.9186 | 0.8860 | 0.8528 | Random Forest |
| tracheobronchial serous cell | 0.6102 | 0.6129 | 0.5985 | 0.5855 | 0.6142 | 0.4401 | SVM |
| bronchus fibroblast of lung | 0.9005 | 0.8905 | 0.8940 | 0.9128 | 0.8966 | 0.7183 | Random Forest |
| mast cell | 0.9305 | 0.9241 | 0.9378 | 0.9467 | 0.9212 | 0.7995 | Random Forest |
| alternatively activated macrophage | 0.9421 | 0.9737 | 0.9702 | 0.9480 | 0.9689 | 0.9288 | Logistic Regression |
| conventional dendritic cell | 0.8794 | 0.8902 | 0.8986 | 0.9186 | 0.8860 | 0.8528 | Random Forest |
| lung microvascular endothelial cell | 0.8213 | 0.8505 | 0.8340 | 0.8410 | 0.8371 | 0.6907 | Logistic Regression |
| T cell | 0.8712 | 0.8962 | 0.8716 | 0.8857 | 0.8955 | 0.6680 | SVM |
| respiratory tract goblet cell | 0.9444 | 0.9546 | 0.9516 | 0.9599 | 0.9667 | 0.8753 | SVM |
| respiratory basal cell | 0.6129 | 0.6667 | 0.6965 | 0.6680 | 0.6505 | 0.4245 | MLP |
| chondrocyte | 0.7373 | 0.7824 | 0.7803 | 0.8644 | 0.7965 | 0.6915 | Random Forest |
| endothelial cell of lymphatic vessel | 0.5994 | 0.6416 | 0.6082 | 0.6143 | 0.6235 | 0.2753 | Logistic Regression |
| plasma cell | 0.5075 | 0.5032 | 0.5198 | 0.4471 | 0.5179 | 0.1096 | MLP |
| bronchiolar smooth muscle cell | 0.8861 | 0.9127 | 0.9251 | 0.8822 | 0.9257 | 0.7881 | MLP |
| pericyte | 0.4535 | 0.5851 | 0.6143 | 0.5211 | 0.5495 | 0.1990 | MLP |
| fibroblast of lung | 0.7669 | 0.7927 | 0.7469 | 0.8327 | 0.7633 | 0.5591 | Random Forest |

Table 5.13 summarizes the held-out train test performance of the final model selected for each cell type. Overall, predictive performance remained heterogeneous across cell populations, which is consistent with the differences in sample size, class imbalance, and biological separability observed in the previous analyses. The strongest results were obtained for *alternatively activated macrophage*, where Logistic Regression achieved an accuracy of $0.9928$ and a macro F1-score of $0.9871$, followed by *respiratory tract goblet cell* ($F1 = 0.9676$, SVM), *bronchiolar smooth muscle cell* ($F1 = 0.9632$, MLP), *multiciliated columnar cell of tracheobronchial tree* ($F1 = 0.9584$, MLP), and *pulmonary alveolar type 2 cell* ($F1 = 0.9510$, SVM). These results indicate that several cell populations supported highly accurate and well-balanced discrimination between conditions in the independent test set.

In contrast, a smaller subset of cell types remained more challenging. The lowest macro F1-scores were observed for *conventional dendritic cell* ($0.4717$, Logistic Regression), *pericyte* ($0.5266$, MLP), *tracheobronchial serous cell* ($0.5790$, SVM), *plasma cell* ($0.6077$, MLP), and *endothelial cell of lymphatic vessel* ($0.6202$, Logistic Regression). Notably, some of these cell types still showed relatively high accuracy despite lower F1-scores, which reinforces that macro F1-score provides the most robust interpretation under class imbalance because it better captures the joint behavior of precision and recall. Across all cell types, the selected final models included Random Forest for 8 subsets, SVM for 7, MLP for 6, and Logistic Regression for 4, further indicating that no single classifier generalized optimally across all biological populations. Taken together, the re-

sults in Table 5.13 confirm that the final train test evaluation preserved the same general pattern observed during model selection, while also highlighting the cell-type-specific variability in predictive difficulty.

Table 5.13: Final train test performance of the selected classifier for each cell type. The table reports the held-out accuracy, sensitivity, specificity, macro F1-score, and the final model chosen for each cell population.

| Cell type | Accuracy | Sensitivity | Specificity | F1-score | Final model |
|---|---|---|---|---|---|
| CD4-positive, alpha-beta T cell | 0.8805 | 0.8977 | 0.8977 | 0.6774 | MLP |
| multiciliated columnar cell of tracheobronchial tree | 0.9637 | 0.9631 | 0.9631 | 0.9584 | MLP |
| monocyte | 0.9189 | 0.8770 | 0.8770 | 0.7314 | SVM |
| B cell | 0.9583 | 0.9510 | 0.9510 | 0.9261 | SVM |
| pulmonary alveolar type 2 cell | 0.9725 | 0.9300 | 0.9300 | 0.9510 | SVM |
| mature NK T cell | 0.8515 | 0.8323 | 0.8323 | 0.8479 | Random Forest |
| macrophage | 0.9008 | 0.8378 | 0.8378 | 0.6483 | SVM |
| CD8-positive, alpha-beta T cell | 0.8971 | 0.8535 | 0.8535 | 0.8050 | Random Forest |
| blood vessel smooth muscle cell | 0.8977 | 0.8277 | 0.8277 | 0.6913 | Random Forest |
| fibroblast | 0.9573 | 0.7284 | 0.7284 | 0.8200 | Random Forest |
| tracheobronchial serous cell | 0.9648 | 0.5505 | 0.5505 | 0.5790 | SVM |
| bronchus fibroblast of lung | 0.9569 | 0.8025 | 0.8025 | 0.7848 | Random Forest |
| mast cell | 0.9467 | 0.9451 | 0.9451 | 0.9468 | Random Forest |
| alternatively activated macrophage | 0.9928 | 0.9843 | 0.9843 | 0.9871 | Logistic Regression |
| conventional dendritic cell | 0.6925 | 0.7093 | 0.7093 | 0.4717 | Logistic Regression |
| lung microvascular endothelial cell | 0.9100 | 0.9007 | 0.9007 | 0.8380 | Logistic Regression |
| T cell | 0.9467 | 0.8889 | 0.8889 | 0.9045 | SVM |
| respiratory tract goblet cell | 0.9677 | 0.9694 | 0.9694 | 0.9676 | SVM |
| respiratory basal cell | 0.9118 | 0.8570 | 0.8570 | 0.6765 | MLP |
| chondrocyte | 0.9449 | 0.8852 | 0.8852 | 0.8907 | Random Forest |
| endothelial cell of lymphatic vessel | 0.8587 | 0.7973 | 0.7973 | 0.6202 | Logistic Regression |
| plasma cell | 0.8413 | 0.7428 | 0.7428 | 0.6077 | MLP |
| bronchiolar smooth muscle cell | 0.9882 | 0.9662 | 0.9662 | 0.9632 | MLP |
| pericyte | 0.8947 | 0.8540 | 0.8540 | 0.5266 | MLP |
| fibroblast of lung | 0.9423 | 0.8984 | 0.8984 | 0.9034 | Random Forest |

### 5.6.3 Global STRING-based network coherence across the remaining cell types by each feature selection method

To extend the functional comparison beyond the detailed CD4-positive, alpha-beta T-cell case study, we summarized the STRING-derived network statistics for all remaining cell types and feature-selection methods in Table 5.14. This global analysis followed the same interaction-based framework adopted previously, in which the selected genes were evaluated through STRING-derived protein protein interaction metrics under a high-confidence threshold, allowing direct comparison of network cohesion across cell populations and feature-selection strategies. :contentReference[oaicite:0]index=0

Table 5.14 revealed substantial heterogeneity across cell types. The strongest interaction structures were observed in *pulmonary alveolar type 2 cell* and *CD4-positive, alpha-beta T cell*, which reached the highest numbers of edges overall (188 and 166, respectively), together with very high local clustering coefficients and extremely significant

PPI enrichment values. Strong network organization was also observed for *fibroblast*, *fibroblast of lung*, *tracheobronchial serous cell*, *blood vessel smooth muscle cell*, and *monocyte*, indicating that these cell populations retained highly connected and biologically coherent feature sets across at least one of the evaluated selection methods. In contrast, *conventional dendritic cell*, *endothelial cell of lymphatic vessel*, and *pericyte* showed much weaker interaction structures, with markedly lower edge counts and, in some cases, weaker enrichment significance, suggesting that their selected genes formed sparser or less functionally integrated networks.

At the method level, Wilcoxon showed the strongest overall network tendency across the remaining cell types. It achieved the highest mean number of edges, the highest mean average node degree, and the highest mean local clustering coefficient across the full comparison, indicating that this method most consistently recovered dense and cohesive interaction structures. The proposed method ranked second overall and showed competitive clustering behavior, supporting the interpretation that it also recovered biologically organized gene sets across multiple cell populations. Random Forest and ANOVA showed intermediate behavior, whereas RFE generally produced the sparsest networks and the weakest average clustering structure. This pattern is broadly consistent with the single-cell-type comparison described earlier, in which Wilcoxon tended to maximize network compactness while alternative methods sometimes captured complementary biological structure.

Taken together, Table 5.14 shows that network-level functional coherence was strongly cell-type dependent and method dependent. Some populations consistently supported dense and statistically enriched interaction networks, whereas others yielded only sparse structures regardless of the feature-selection strategy. These findings reinforce the view that the biological signal available for downstream interpretation varies substantially across cell populations and that global conclusions should therefore consider both the selected method and the intrinsic properties of each cell type.

A denser protein protein interaction network should not be interpreted automatically as evidence of greater disease specificity. High edge counts, higher average degree, or stronger overall connectivity may reflect the selection of broadly connected housekeeping or biosynthetic genes, which are central to general cellular function but not necessarily more informative about the condition under study. In contrast, the proposed method was considered biologically advantageous not because it maximized network density, but because it recovered a functionally coherent set of genes associated with processes that

were more directly consistent with the disease context under investigation. Therefore, in this study, network density and biological specificity were treated as related but non-equivalent criteria.

Table 5.14: Global STRING-derived network statistics across the remaining cell types and feature-selection methods. Corrected floating-point values are reported for average node degree, average local clustering coefficient, and PPI enrichment p-values.

| Cell type | Method | Number of edges | Average node degree | Avg. local clustering coefficient | PPI enrichment p-value |
|---|---|---|---|---|---|
| B cell | Proposed method | 25 | 2.17 | 0.510 | 0.013 |
| | Wilcoxon | 45 | 3.91 | 0.728 | $5.64 \times 10^{-5}$ |
| | ANOVA | 38 | 3.62 | 0.601 | $1.43 \times 10^{-6}$ |
| | RF | 62 | 5.64 | 0.795 | $2.33 \times 10^{-15}$ |
| | RFE | 61 | 5.30 | 0.653 | $1.11 \times 10^{-15}$ |
| CD4-positive alpha-beta T cell | Proposed method | 144 | 12.50 | 0.846 | $< 1.0 \times 10^{-16}$ |
| | Wilcoxon | 166 | 13.80 | 0.895 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 44 | 4.00 | 0.657 | $< 1.0 \times 10^{-16}$ |
| | RF | 11 | 1.00 | 0.500 | $2.99 \times 10^{-7}$ |
| | RFE | 15 | 1.25 | 0.562 | 0.00071 |
| CD8-positive alpha-beta T cell | Proposed method | 21 | 1.68 | 0.381 | 0.048 |
| | Wilcoxon | 34 | 2.72 | 0.729 | $2.23 \times 10^{-5}$ |
| | ANOVA | 27 | 2.25 | 0.600 | $< 1.0 \times 10^{-16}$ |
| | RF | 34 | 2.83 | 0.653 | $3.22 \times 10^{-15}$ |
| | RFE | 25 | 2.08 | 0.497 | $6.12 \times 10^{-8}$ |
| T cell | Proposed method | 54 | 4.91 | 0.593 | $8.94 \times 10^{-6}$ |
| | Wilcoxon | 22 | 2.10 | 0.641 | 0.00062 |
| | ANOVA | 20 | 1.90 | 0.390 | $1.43 \times 10^{-7}$ |
| | RF | 21 | 1.91 | 0.436 | 0.00073 |
| | RFE | 7 | 6.36 | 0.318 | 0.035 |
| alternatively activated macrophage | Proposed method | 40 | 3.20 | 0.696 | $2.35 \times 10^{-12}$ |
| | Wilcoxon | 37 | 3.08 | 0.495 | $5.55 \times 10^{-16}$ |
| | ANOVA | 17 | 1.42 | 0.407 | $2.17 \times 10^{-7}$ |
| | RF | 23 | 1.92 | 0.531 | $8.70 \times 10^{-11}$ |
| | RFE | 19 | 1.58 | 0.311 | 0.003 |
| blood vessel smooth muscle cell | Proposed method | 45 | 3.60 | 0.581 | $3.62 \times 10^{-13}$ |
| | Wilcoxon | 74 | 5.92 | 0.746 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 57 | 4.96 | 0.693 | $< 1.0 \times 10^{-16}$ |
| | RF | 54 | 4.70 | 0.571 | $< 1.0 \times 10^{-16}$ |

*Continued on next page*

| Cell type | Method | Number of edges | Average node degree | Avg. local clustering coefficient | PPI enrichment p-value |
|---|---|---|---|---|---|
| | RFE | 13 | 1.13 | 0.319 | 0.021 |
| bronchiolar smooth muscle cell | Proposed method | 32 | 2.56 | 0.497 | $3.91 \times 10^{-5}$ |
| | Wilcoxon | 58 | 4.83 | 0.740 | $1.11 \times 10^{-16}$ |
| | ANOVA | 56 | 4.48 | 0.716 | $1.33 \times 10^{-15}$ |
| | RF | 56 | 4.48 | 0.709 | $< 1.0 \times 10^{-16}$ |
| | RFE | 17 | 1.42 | 0.327 | 0.009 |
| bronchus fibroblast of lung | Proposed method | 18 | 1.57 | 0.262 | $7.70 \times 10^{-8}$ |
| | Wilcoxon | 37 | 3.08 | 0.475 | $2.07 \times 10^{-13}$ |
| | ANOVA | 9 | 8.18 | 0.291 | 0.069 |
| | RF | 22 | 1.91 | 0.418 | $7.81 \times 10^{-5}$ |
| | RFE | 15 | 1.25 | 0.308 | 0.00013 |
| chondrocyte | Proposed method | 19 | 1.65 | 0.430 | $2.51 \times 10^{-7}$ |
| | Wilcoxon | 32 | 2.67 | 0.601 | $7.87 \times 10^{-8}$ |
| | ANOVA | 23 | 2.30 | 0.450 | $1.19 \times 10^{-5}$ |
| | RF | 49 | 4.26 | 0.600 | $1.37 \times 10^{-10}$ |
| | RFE | 24 | 2.09 | 0.555 | $9.23 \times 10^{-6}$ |
| conventional dendritic cell | Proposed method | 1 | 0.09 | 0.091 | 0.603 |
| | Wilcoxon | 3 | 2.61 | 0.174 | 0.112 |
| | ANOVA | 0 | 0.00 | 0.000 | 1.000 |
| | RF | 6 | 0.60 | 0.367 | 0.048 |
| | RFE | 6 | 5.22 | 0.435 | 0.139 |
| endothelial cell of lymphatic vessel | Proposed method | 3 | 0.25 | 0.125 | 0.336 |
| | Wilcoxon | 5 | 5.26 | 0.368 | 0.271 |
| | ANOVA | 5 | 5.56 | 0.278 | 0.007 |
| | RF | 4 | 4.21 | 0.211 | 0.036 |
| | RFE | 8 | 7.27 | 0.455 | 0.004 |
| fibroblast | Proposed method | 86 | 6.88 | 0.715 | $< 1.0 \times 10^{-16}$ |
| | Wilcoxon | 99 | 7.92 | 0.672 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 20 | 1.67 | 0.375 | $4.98 \times 10^{-9}$ |
| | RF | 28 | 2.33 | 0.445 | $9.28 \times 10^{-11}$ |
| | RFE | 9 | 0.72 | 0.440 | 0.023 |
| fibroblast of lung | Proposed method | 86 | 6.88 | 0.715 | $< 1.0 \times 10^{-16}$ |
| | Wilcoxon | 99 | 7.92 | 0.672 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 20 | 1.67 | 0.375 | $4.98 \times 10^{-9}$ |
| | RF | 28 | 2.33 | 0.445 | $9.28 \times 10^{-11}$ |
| | RFE | 9 | 0.72 | 0.440 | 0.023 |

*Continued on next page*

| Cell type | Method | Number of edges | Average node degree | Avg. local clustering coefficient | PPI enrichment p-value |
|---|---|---|---|---|---|
| lung microvascular endothelial cell | Proposed method | 38 | 3.04 | 0.596 | $1.59 \times 10^{-11}$ |
| | Wilcoxon | 15 | 1.20 | 0.501 | 0.004 |
| | ANOVA | 11 | 9.17 | 0.472 | 0.008 |
| | RF | 18 | 1.44 | 0.459 | $3.36 \times 10^{-5}$ |
| | RFE | 13 | 1.13 | 0.312 | 0.010 |
| macrophage | Proposed method | 34 | 2.72 | 0.593 | $3.54 \times 10^{-10}$ |
| | Wilcoxon | 40 | 3.20 | 0.665 | $7.66 \times 10^{-15}$ |
| | ANOVA | 13 | 1.08 | 0.422 | $7.76 \times 10^{-7}$ |
| | RF | 9 | 8.18 | 0.364 | 0.003 |
| | RFE | 5 | 4.55 | 0.273 | 0.601 |
| mast cell | Proposed method | 8 | 6.67 | 0.347 | 0.002 |
| | Wilcoxon | 28 | 2.33 | 0.681 | $1.53 \times 10^{-7}$ |
| | ANOVA | 12 | 1.09 | 0.432 | 0.003 |
| | RF | 29 | 2.64 | 0.515 | $3.44 \times 10^{-8}$ |
| | RFE | 40 | 3.48 | 0.610 | $4.21 \times 10^{-6}$ |
| mature NK T cell | Proposed method | 20 | 1.67 | 0.546 | 0.002 |
| | Wilcoxon | 14 | 1.17 | 0.389 | 0.005 |
| | ANOVA | 18 | 1.71 | 0.357 | $3.44 \times 10^{-15}$ |
| | RF | 20 | 1.74 | 0.428 | $2.14 \times 10^{-14}$ |
| | RFE | 23 | 2.00 | 0.459 | $1.07 \times 10^{-7}$ |
| monocyte | Proposed method | 20 | 1.67 | 0.546 | 0.002 |
| | Wilcoxon | 73 | 5.84 | 0.603 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 43 | 3.44 | 0.568 | $4.65 \times 10^{-12}$ |
| | RF | 71 | 5.68 | 0.726 | $< 1.0 \times 10^{-16}$ |
| | RFE | 26 | 2.17 | 0.539 | 0.001 |
| multiciliated columnar cell of tracheobronchial tree | Proposed method | 31 | 2.58 | 0.643 | $7.44 \times 10^{-8}$ |
| | Wilcoxon | 20 | 1.67 | 0.362 | 0.001 |
| | ANOVA | 22 | 1.83 | 0.433 | $9.95 \times 10^{-5}$ |
| | RF | 28 | 2.33 | 0.675 | $5.52 \times 10^{-11}$ |
| | RFE | 12 | 1.04 | 0.458 | 0.001 |
| pericyte | Proposed method | 0 | 0.00 | 0.000 | 1.000 |
| | Wilcoxon | 11 | 9.57 | 0.348 | $1.16 \times 10^{-5}$ |
| | ANOVA | 12 | 1.00 | 0.361 | $5.27 \times 10^{-7}$ |
| | RF | 5 | 4.17 | 0.167 | 0.050 |
| | RFE | 7 | 5.83 | 0.296 | 0.020 |

*Continued on next page*

| Cell type | Method | Number of edges | Average node degree | Avg. local clustering coefficient | PPI enrichment p-value |
|---|---|---|---|---|---|
| plasma cell | Proposed method | 15 | 1.76 | 0.588 | $9.63 \times 10^{-5}$ |
| | Wilcoxon | 36 | 3.27 | 0.662 | $7.11 \times 10^{-11}$ |
| | ANOVA | 5 | 4.55 | 0.182 | 0.035 |
| | RF | 4 | 3.48 | 0.261 | 0.244 |
| | RFE | 10 | 8.33 | 0.290 | 0.025 |
| pulmonary alveolar type 2 cell | Proposed method | 150 | 12.00 | 0.907 | $< 1.0 \times 10^{-16}$ |
| | Wilcoxon | 188 | 15.00 | 0.888 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 151 | 12.10 | 0.797 | $< 1.0 \times 10^{-16}$ |
| | RF | 91 | 7.28 | 0.602 | $< 1.0 \times 10^{-16}$ |
| | RFE | 44 | 3.52 | 0.482 | $< 1.0 \times 10^{-16}$ |
| pulmonary capillary endothelial cell | Proposed method | 34 | 2.72 | 0.620 | $1.01 \times 10^{-14}$ |
| | Wilcoxon | 32 | 2.56 | 0.523 | $5.16 \times 10^{-10}$ |
| | ANOVA | 21 | 1.68 | 0.424 | $1.72 \times 10^{-5}$ |
| | RF | 23 | 1.84 | 0.307 | $6.62 \times 10^{-9}$ |
| | RFE | 13 | 1.08 | 0.378 | 0.002 |
| respiratory basal cell | Proposed method | 15 | 1.25 | 0.306 | 0.001 |
| | Wilcoxon | 9 | 0.75 | 0.347 | 0.043 |
| | ANOVA | 12 | 1.04 | 0.278 | 0.118 |
| | RF | 41 | 3.73 | 0.411 | $4.36 \times 10^{-6}$ |
| | RFE | 7 | 6.36 | 0.318 | 0.629 |
| respiratory tract goblet cell | Proposed method | 21 | 1.68 | 0.528 | 0.00012 |
| | Wilcoxon | 25 | 2.00 | 0.365 | $1.27 \times 10^{-6}$ |
| | ANOVA | 32 | 2.56 | 0.476 | $6.96 \times 10^{-14}$ |
| | RF | 41 | 3.28 | 0.637 | $2.17 \times 10^{-8}$ |
| | RFE | 16 | 1.28 | 0.486 | 0.00026 |
| tracheobronchial serous cell | Proposed method | 61 | 4.88 | 0.601 | $< 1.0 \times 10^{-16}$ |
| | Wilcoxon | 99 | 7.92 | 0.748 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 8 | 0.80 | 0.283 | 0.044 |
| | RF | 26 | 2.26 | 0.458 | $2.48 \times 10^{-9}$ |
| | RFE | 6 | 0.50 | 0.333 | 0.034 |
| vein endothelial cell of respiratory system | Proposed method | 49 | 3.92 | 0.600 | $3.97 \times 10^{-14}$ |
| | Wilcoxon | 55 | 4.40 | 0.627 | $< 1.0 \times 10^{-16}$ |
| | ANOVA | 11 | 1.16 | 0.368 | 0.001 |
| | RF | 27 | 2.35 | 0.457 | $1.71 \times 10^{-7}$ |
| | RFE | 19 | 1.73 | 0.311 | 0.001 |

### 5.6.4 Comparative low-dimensional structure class separation for each cell type

Table 5.15 summarizes the geometric comparison of the feature subsets across all evaluated cell types. Following the methodological definition adopted in this dissertation, the primary ranking criterion was the separation ratio computed in the full standardized feature space, whereas PCA served only as a qualitative low-dimensional representation of the dominant structure and not as the quantitative basis for model comparison. This distinction is important because visual separation in two dimensions does not necessarily preserve the multivariate relationship between centroid displacement and within-class dispersion established by the full-space metric.

Across the 27 analyzed cell types, ANOVA produced the highest separation ratio in 25 cases, Random Forest was the best method for mature NK T cell, and Wilcoxon was the best method for fibroblast of lung. At the global level, ANOVA also achieved the highest mean separation ratio across cell types ($\overline{R}_{sep} = 1.371$), followed by Random Forest ($\overline{R}_{sep} = 1.170$). In contrast, the proposed method showed a lower average geometric separation score under this specific criterion ($\overline{R}_{sep} = 0.800$), while random selection produced by far the weakest overall structure ($\overline{R}_{sep} = 0.237$).

The strongest class separation was observed for conventional dendritic cell, in which ANOVA reached a separation ratio of 4.914, indicating a very large centroid displacement relative to within-class dispersion. Other cell types with strong geometric separation included fibroblast (2.166), respiratory tract goblet cell (1.751), tracheobronchial serous cell (1.748), chondrocyte (1.608), and macrophage (1.584). These results indicate that, for these populations, the selected feature subsets were able to organize the disease groups into comparatively compact and well-separated structures in the standardized feature space.

In contrast, some cell types remained substantially more difficult from a geometric perspective. The lowest best-case separation ratios were observed for fibroblast of lung (0.573), mast cell (0.713), T cell (0.729), and bronchus fibroblast of lung (0.785). In these cases, even the best-performing feature subset did not produce a strongly separated full-space configuration, suggesting either greater overlap between normal and disease states or weaker condition-associated signal for the selected genes.

The complementary structural metrics reinforce this interpretation. Methods with larger separation ratios also tended to present higher Silhouette values and, in many cases, more favorable Davies Bouldin and Calinski Harabasz indices, although these secondary

criteria were not perfectly concordant for every cell type. Overall, the results support two main conclusions: first, the quality of geometric class separation was strongly cell-type dependent; second, under this specific structural criterion, ANOVA was the most consistently effective feature-selection strategy across the full comparison.

Table 5.15: Comparative geometric separation across cell types and feature-selection methods. For each method, the table reports the separation ratio computed in the full standardized feature space, the centroid distance, the pooled within-class dispersion, the Silhouette score, the Davies Bouldin index, and the Calinski Harabasz index. Cell-type names are shown only once per block and are abbreviated for compactness.

| **Cell type** | **Method** | $R_{sep}$ | **Centroid** | **Pooled disp.** | **Silhouette** | **DB** | **CH** |
|---|---|---|---|---|---|---|---|
| B cell | ANOVA | 1.270 | 5.470 | 4.306 | 0.299 | 1.788 | 679.139 |
| | Random Forest | 1.201 | 5.128 | 4.269 | 0.275 | 1.814 | 583.555 |
| | Wilcoxon | 1.195 | 5.222 | 4.372 | 0.247 | 1.750 | 608.964 |
| | Proposed method | 1.094 | 4.896 | 4.475 | 0.233 | 1.947 | 524.415 |
| | RFE | 0.937 | 4.091 | 4.367 | 0.214 | 2.262 | 350.806 |
| | Randomic selection | 0.187 | 0.653 | 3.489 | 0.061 | 11.186 | 8.158 |
| CD4+ $\alpha\beta$ T | ANOVA | 1.423 | 5.521 | 3.880 | 0.527 | 2.456 | 491.519 |
| | Random Forest | 1.367 | 5.149 | 3.766 | 0.578 | 2.874 | 424.320 |
| | RFE | 1.089 | 4.763 | 4.375 | 0.442 | 2.828 | 360.431 |
| | Wilcoxon | 1.003 | 4.382 | 4.369 | 0.246 | 2.282 | 303.117 |
| | Proposed method | 0.877 | 3.891 | 4.439 | 0.220 | 2.599 | 237.234 |
| | Randomic selection | 0.184 | 0.427 | 2.321 | 0.180 | 12.566 | 3.152 |
| CD8+ $\alpha\beta$ T | ANOVA | 1.278 | 4.935 | 3.862 | 0.529 | 2.755 | 208.205 |
| | Random Forest | 1.221 | 4.799 | 3.929 | 0.512 | 2.803 | 195.990 |
| | RFE | 0.970 | 4.303 | 4.436 | 0.371 | 2.854 | 155.235 |
| | Proposed method | 0.778 | 3.697 | 4.755 | 0.182 | 2.832 | 112.801 |
| | Wilcoxon | 0.773 | 3.683 | 4.764 | 0.163 | 2.770 | 111.876 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Randomic selection | 0.155 | 0.507 | 3.262 | -0.079 | 12.044 | 2.031 |
| T cell | ANOVA | 0.729 | 3.347 | 4.588 | 0.138 | 2.866 | 110.881 |
| | Random Forest | 0.668 | 3.119 | 4.672 | 0.116 | 3.088 | 94.996 |
| | Wilcoxon | 0.640 | 3.021 | 4.717 | 0.092 | 3.147 | 88.602 |
| | RFE | 0.607 | 2.818 | 4.640 | 0.109 | 3.417 | 76.228 |
| | Proposed method | 0.498 | 2.381 | 4.780 | 0.053 | 4.004 | 53.317 |
| | Randomic selection | 0.152 | 0.428 | 2.805 | 0.025 | 13.341 | 1.640 |
| Alt. act. macro. | ANOVA | 1.121 | 4.785 | 4.268 | 0.242 | 1.899 | 214.003 |
| | Random Forest | 1.116 | 4.738 | 4.248 | 0.237 | 1.894 | 209.205 |
| | Wilcoxon | 1.079 | 4.580 | 4.243 | 0.179 | 1.803 | 193.484 |
| | Proposed method | 0.958 | 4.235 | 4.421 | 0.145 | 2.008 | 154.961 |
| | RFE | 0.902 | 4.003 | 4.438 | 0.154 | 2.209 | 143.103 |
| | Randomic selection | 0.309 | 1.155 | 3.735 | 0.035 | 6.558 | 11.363 |
| BV smooth muscle | ANOVA | 0.983 | 4.200 | 4.272 | 0.297 | 2.531 | 126.856 |
| | Random Forest | 0.954 | 4.112 | 4.310 | 0.309 | 2.659 | 121.387 |
| | Wilcoxon | 0.849 | 3.832 | 4.511 | 0.113 | 2.323 | 104.716 |
| | RFE | 0.714 | 3.275 | 4.585 | 0.221 | 3.260 | 75.633 |
| | Proposed method | 0.706 | 3.217 | 4.560 | 0.102 | 2.868 | 72.908 |
| | Randomic selection | 0.154 | 0.518 | 3.356 | 0.081 | 13.887 | 1.839 |
| Bronchiolar SMC | ANOVA | 1.528 | 6.090 | 3.985 | 0.304 | 1.340 | 89.571 |
| | Random Forest | 1.467 | 5.931 | 4.044 | 0.286 | 1.386 | 83.580 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Wilcoxon | 1.435 | 5.827 | 4.060 | 0.251 | 1.324 | 79.855 |
| | RFE | 1.222 | 5.292 | 4.329 | 0.239 | 1.655 | 62.735 |
| | Proposed method | 1.164 | 5.029 | 4.319 | 0.188 | 1.631 | 55.457 |
| | Randomic selection | 0.256 | 0.922 | 3.600 | -0.080 | 7.080 | 1.566 |
| Bronchial fibro. | ANOVA | 0.785 | 3.589 | 4.574 | 0.172 | 2.759 | 184.031 |
| | Random Forest | 0.744 | 3.438 | 4.622 | 0.157 | 2.876 | 167.475 |
| | Wilcoxon | 0.732 | 3.416 | 4.668 | 0.123 | 2.806 | 165.130 |
| | Proposed method | 0.696 | 3.255 | 4.678 | 0.127 | 2.987 | 148.587 |
| | RFE | 0.628 | 2.970 | 4.730 | 0.113 | 3.297 | 121.993 |
| | Randomic selection | 0.153 | 0.513 | 3.365 | 0.014 | 13.202 | 3.417 |
| Chondrocyte | ANOVA | 1.608 | 6.665 | 4.145 | 0.505 | 1.970 | 69.111 |
| | Random Forest | 1.426 | 6.274 | 4.399 | 0.412 | 1.877 | 60.260 |
| | RFE | 1.313 | 6.057 | 4.614 | 0.316 | 1.688 | 55.694 |
| | Wilcoxon | 1.208 | 5.654 | 4.681 | 0.230 | 1.627 | 47.823 |
| | Proposed method | 1.050 | 4.797 | 4.567 | 0.239 | 1.979 | 33.513 |
| | Randomic selection | 0.253 | 1.087 | 4.298 | -0.050 | 7.504 | 1.616 |
| cDC | ANOVA | 4.914 | 7.222 | 1.470 | 0.910 | 3.203 | 77.526 |
| | Random Forest | 1.927 | 6.404 | 3.323 | 0.734 | 3.134 | 60.084 |
| | RFE | 1.047 | 4.994 | 4.770 | 0.324 | 2.432 | 35.812 |
| | Wilcoxon | 0.897 | 3.947 | 4.399 | 0.275 | 2.798 | 22.110 |
| | Randomic selection | 0.544 | 2.125 | 3.906 | 0.329 | 5.131 | 6.324 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Proposed method | 0.396 | 1.754 | 4.433 | -0.189 | 3.646 | 4.304 |
| Lymphatic EC | ANOVA | 1.201 | 4.969 | 4.136 | 0.426 | 2.455 | 48.441 |
| | Random Forest | 1.082 | 4.718 | 4.361 | 0.367 | 2.512 | 43.223 |
| | RFE | 0.964 | 4.242 | 4.399 | 0.297 | 2.526 | 34.312 |
| | Wilcoxon | 0.948 | 3.945 | 4.161 | 0.350 | 2.871 | 29.381 |
| | Proposed method | 0.369 | 1.579 | 4.279 | -0.152 | 4.181 | 4.464 |
| | Randomic selection | 0.164 | 0.626 | 3.811 | -0.056 | 11.690 | 0.696 |
| Fibroblast | ANOVA | 2.166 | 9.161 | 4.230 | 0.550 | 1.480 | 292.260 |
| | Random Forest | 2.109 | 9.059 | 4.296 | 0.537 | 1.472 | 284.971 |
| | Wilcoxon | 1.775 | 7.981 | 4.496 | 0.397 | 1.277 | 215.140 |
| | RFE | 1.485 | 6.999 | 4.712 | 0.349 | 1.513 | 161.862 |
| | Proposed method | 1.358 | 6.300 | 4.640 | 0.275 | 1.508 | 129.431 |
| | Randomic selection | 0.351 | 1.270 | 3.616 | 0.199 | 6.823 | 4.988 |
| Lung fibro. | Wilcoxon | 0.573 | 2.595 | 4.527 | 0.070 | 3.524 | 7.585 |
| | Proposed method | 0.570 | 2.581 | 4.530 | 0.070 | 3.521 | 7.502 |
| | RFE | 0.550 | 2.551 | 4.637 | 0.063 | 3.623 | 7.321 |
| | ANOVA | 0.527 | 2.208 | 4.191 | 0.115 | 4.047 | 5.659 |
| | Random Forest | 0.505 | 2.191 | 4.341 | 0.067 | 4.035 | 5.570 |
| | Randomic selection | 0.343 | 1.036 | 3.020 | 0.179 | 6.789 | 1.951 |
| Lung microvasc. EC | ANOVA | 1.312 | 5.589 | 4.260 | 0.414 | 2.112 | 141.662 |
| | Random Forest | 1.240 | 5.401 | 4.357 | 0.372 | 2.119 | 131.189 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Wilcoxon | 1.039 | 4.786 | 4.606 | 0.168 | 1.873 | 100.579 |
| | RFE | 1.010 | 4.579 | 4.533 | 0.295 | 2.376 | 91.400 |
| | Proposed method | 0.825 | 3.854 | 4.670 | 0.072 | 2.213 | 63.268 |
| | Randomic selection | 0.225 | 0.709 | 3.146 | 0.151 | 10.087 | 2.035 |
| Macrophage | ANOVA | 1.584 | 6.276 | 3.963 | 0.599 | 2.562 | 119.960 |
| | Random Forest | 1.532 | 5.938 | 3.876 | 0.623 | 2.844 | 106.855 |
| | RFE | 1.103 | 4.624 | 4.191 | 0.430 | 2.707 | 63.776 |
| | Wilcoxon | 1.090 | 5.052 | 4.633 | 0.215 | 1.903 | 76.481 |
| | Proposed method | 1.022 | 4.837 | 4.733 | 0.207 | 2.059 | 69.936 |
| | Randomic selection | 0.255 | 0.908 | 3.569 | 0.130 | 8.731 | 2.405 |
| Mast cell | ANOVA | 0.713 | 3.282 | 4.601 | 0.107 | 2.818 | 205.156 |
| | Random Forest | 0.701 | 3.189 | 4.552 | 0.105 | 2.863 | 192.288 |
| | Wilcoxon | 0.698 | 3.194 | 4.579 | 0.103 | 2.870 | 193.088 |
| | Proposed method | 0.614 | 2.876 | 4.682 | 0.083 | 3.265 | 153.209 |
| | RFE | 0.596 | 2.723 | 4.572 | 0.087 | 3.373 | 136.156 |
| | Randomic selection | 0.126 | 0.489 | 3.871 | 0.001 | 15.779 | 4.082 |
| Mature NKT | Random Forest | 0.941 | 3.365 | 3.575 | 0.485 | 3.487 | 152.307 |
| | ANOVA | 0.906 | 3.444 | 3.800 | 0.450 | 3.432 | 160.036 |
| | RFE | 0.726 | 3.094 | 4.262 | 0.341 | 3.726 | 127.668 |
| | Wilcoxon | 0.544 | 2.622 | 4.822 | 0.110 | 3.891 | 90.423 |
| | Proposed method | 0.446 | 2.162 | 4.844 | 0.059 | 4.570 | 60.825 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Randomic selection | 0.114 | 0.450 | 3.944 | 0.001 | 17.469 | 2.578 |
| Monocyte | ANOVA | 1.028 | 4.589 | 4.461 | 0.320 | 2.446 | 159.343 |
| | Random Forest | 1.017 | 4.493 | 4.418 | 0.308 | 2.441 | 152.573 |
| | Wilcoxon | 0.880 | 4.119 | 4.681 | 0.102 | 2.137 | 127.494 |
| | RFE | 0.844 | 3.846 | 4.556 | 0.256 | 2.802 | 110.710 |
| | Proposed method | 0.252 | 1.169 | 4.644 | 0.067 | 8.324 | 10.004 |
| | Randomic selection | 0.175 | 0.571 | 3.264 | 0.126 | 12.704 | 2.383 |
| Multiciliated TB tree | ANOVA | 0.808 | 3.660 | 4.529 | 0.125 | 2.482 | 963.689 |
| | Wilcoxon | 0.801 | 3.599 | 4.496 | 0.122 | 2.487 | 927.769 |
| | Random Forest | 0.771 | 3.467 | 4.494 | 0.114 | 2.565 | 853.212 |
| | Proposed method | 0.729 | 3.277 | 4.494 | 0.105 | 2.711 | 753.118 |
| | RFE | 0.617 | 2.789 | 4.518 | 0.077 | 3.181 | 530.577 |
| | Randomic selection | 0.094 | 0.337 | 3.585 | 0.004 | 21.275 | 7.210 |
| Pericyte | ANOVA | 1.239 | 4.595 | 3.710 | 0.500 | 2.582 | 24.481 |
| | Random Forest | 1.009 | 4.207 | 4.169 | 0.403 | 2.793 | 20.088 |
| | RFE | 0.861 | 3.937 | 4.575 | 0.246 | 2.721 | 17.365 |
| | Wilcoxon | 0.811 | 3.349 | 4.128 | 0.312 | 3.195 | 12.249 |
| | Proposed method | 0.596 | 2.644 | 4.437 | 0.088 | 3.473 | 7.451 |
| | Randomic selection | 0.267 | 1.118 | 4.181 | 0.027 | 7.671 | 1.289 |
| Plasma cell | ANOVA | 1.003 | 3.641 | 3.629 | 0.495 | 3.293 | 22.498 |
| | Random Forest | 0.798 | 3.347 | 4.191 | 0.354 | 3.400 | 18.793 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | RFE | 0.648 | 3.028 | 4.670 | 0.132 | 3.277 | 15.213 |
| | Wilcoxon | 0.628 | 2.494 | 3.971 | 0.218 | 3.812 | 10.156 |
| | Proposed method | 0.452 | 1.718 | 3.802 | 0.148 | 5.001 | 4.734 |
| | Randomic selection | 0.290 | 1.000 | 3.444 | 0.100 | 7.502 | 1.589 |
| AT2 cell | ANOVA | 1.496 | 6.189 | 4.138 | 0.236 | 1.228 | 842.289 |
| | Wilcoxon | 1.494 | 6.138 | 4.108 | 0.223 | 1.208 | 824.692 |
| | Random Forest | 1.410 | 5.987 | 4.246 | 0.228 | 1.294 | 775.385 |
| | Proposed method | 1.375 | 5.790 | 4.209 | 0.208 | 1.347 | 714.066 |
| | RFE | 1.112 | 4.977 | 4.476 | 0.182 | 1.713 | 499.221 |
| | Randomic selection | 0.371 | 1.283 | 3.460 | 0.094 | 5.684 | 29.048 |
| Pulm. capillary EC | ANOVA | 1.352 | 5.858 | 4.333 | 0.447 | 2.186 | 230.459 |
| | Random Forest | 1.291 | 5.677 | 4.397 | 0.412 | 2.169 | 215.366 |
| | Wilcoxon | 1.031 | 4.817 | 4.674 | 0.194 | 1.976 | 151.910 |
| | RFE | 1.029 | 4.697 | 4.563 | 0.322 | 2.420 | 144.021 |
| | Proposed method | 0.967 | 4.545 | 4.702 | 0.181 | 2.109 | 134.455 |
| | Randomic selection | 0.189 | 0.611 | 3.230 | 0.003 | 10.571 | 2.320 |
| Resp. basal | ANOVA | 1.440 | 5.920 | 4.111 | 0.485 | 2.201 | 61.638 |
| | Random Forest | 1.202 | 5.204 | 4.330 | 0.405 | 2.355 | 46.360 |
| | RFE | 1.072 | 4.972 | 4.638 | 0.253 | 2.072 | 42.000 |
| | Wilcoxon | 1.056 | 4.829 | 4.571 | 0.188 | 1.931 | 39.433 |
| | Proposed method | 0.802 | 3.747 | 4.674 | 0.107 | 2.454 | 23.038 |

Continued on next page

| Cell type | Method | $R_{sep}$ | Centroid | Pooled disp. | Silhouette | DB | CH |
|---|---|---|---|---|---|---|---|
| | Randomic selection | 0.203 | 0.781 | 3.851 | 0.003 | 9.916 | 0.960 |
| Goblet cell | ANOVA | 1.751 | 6.491 | 3.707 | 0.319 | 1.097 | 498.851 |
| | Wilcoxon | 1.711 | 6.388 | 3.734 | 0.309 | 1.114 | 473.410 |
| | Random Forest | 1.638 | 6.198 | 3.784 | 0.299 | 1.167 | 430.217 |
| | Proposed method | 1.473 | 5.823 | 3.952 | 0.265 | 1.317 | 356.372 |
| | RFE | 1.077 | 4.641 | 4.309 | 0.192 | 1.815 | 193.765 |
| | Randomic selection | 0.391 | 1.417 | 3.628 | 0.054 | 5.110 | 14.718 |
| TB serous | ANOVA | 1.748 | 6.105 | 3.492 | 0.705 | 3.185 | 82.898 |
| | Random Forest | 1.282 | 5.249 | 4.096 | 0.587 | 3.191 | 60.682 |
| | Wilcoxon | 0.990 | 4.553 | 4.598 | 0.134 | 1.930 | 45.336 |
| | RFE | 0.942 | 4.413 | 4.683 | 0.342 | 2.843 | 42.542 |
| | Proposed method | 0.852 | 3.996 | 4.688 | 0.119 | 2.284 | 34.759 |
| | Randomic selection | 0.300 | 0.959 | 3.195 | 0.087 | 7.220 | 1.974 |
| Resp. venous EC | ANOVA | 1.115 | 4.693 | 4.209 | 0.423 | 2.715 | 126.576 |
| | Random Forest | 0.984 | 4.303 | 4.375 | 0.364 | 2.832 | 105.680 |
| | RFE | 0.805 | 3.757 | 4.668 | 0.229 | 2.907 | 79.902 |
| | Wilcoxon | 0.782 | 3.628 | 4.640 | 0.104 | 2.521 | 74.384 |
| | Proposed method | 0.694 | 3.245 | 4.677 | 0.081 | 2.820 | 59.177 |
| | Randomic selection | 0.200 | 0.650 | 3.256 | 0.151 | 11.383 | 2.333 |

*Abbreviations:* Alt. act. macro. = alternatively activated macrophage; BV smooth muscle = blood vessel smooth muscle cell; Bronchiolar SMC = bronchiolar smooth muscle cell; Bronchial fibro. = bronchus fibroblast of lung; cDC = conventional dendritic cell; EC = endothelial cell; Lung microvasc. EC = lung

microvascular endothelial cell; Mature NKT = mature NK T cell; Multiciliated TB tree = multiciliated columnar cell of tracheobronchial tree; AT2 cell = pulmonary alveolar type 2 cell; Pulm. capillary EC = pulmonary capillary endothelial cell; Resp. basal = respiratory basal cell; TB serous = tracheobronchial serous cell; Resp. venous EC = vein endothelial cell of respiratory system.

## 5.7 Chapter summary

This chapter showed that the proposed computational framework analyzed condition-associated signals in a single-cell RNA-sequencing dataset of chronic obstructive pulmonary disease. The preprocessing stage preserved the full set of annotated cells while removing uninformative genes, yielding a globally filtered matrix suitable for downstream analysis. At the cell type-specific level, the framework retained 27 eligible populations for modeling, which showed that the framework could operate across cellular contexts with heterogeneous abundance and class balance.

The detailed case study on *CD4-positive, alpha-beta T cell* showed that the framework addressed an imbalanced classification problem through leakage-aware preprocessing, feature selection, and model evaluation. In this subset, the MLP produced a balanced profile across performance metrics, emerged as the only Pareto-optimal solution, and maintained discrimination in the independent train test evaluation. These results showed that the proposed framework recovered predictive patterns associated with disease status despite the limited representation of the minority class.

The interpretability analyses extended this result. The SHAP-based global ranking identified a compact group of influential genes, and the class-specific analyses showed that the model used these genes in a direction-dependent manner to distinguish normal from disease samples. The correlation analysis between gene expression and SHAP attribution showed structured relationships between transcript abundance and model attribution. Together, these findings indicated that the final classifier combined satisfactory predictive performance with an interpretable decision structure.

The comparative feature-selection analyses also showed that predictive utility, geometric separation, and biological coherence did not always converge on the same method. In the CD4-positive, alpha-beta T-cell subset, the proposed method did not achieve the highest full-space geometric separation, as ANOVA and Random Forest yielded higher separation ratios under the centroid-dispersion criterion. This result showed that geometric compactness represents only one dimension of feature-subset quality and that

low-dimensional visual separation cannot substitute for multivariate quantitative evaluation. When classification, interpretability, and downstream biological validation were considered together, the proposed method remained analytically informative.

The functional enrichment analyses added biological validation. Although all evaluated methods recovered genes associated with translation-related biology, the proposed method combined translational, stress-response, chaperone-mediated refolding, apoptotic, and regulatory processes. The alternative methods tended to emphasize broader biosynthetic or ribosome-centered signatures. In addition, the disease-associated genes prioritized by the proposed method formed an interaction network enriched for translation, protein homeostasis, and stress-related pathways. These results indicated that the proposed method captured biologically coordinated programs rather than only generic housekeeping signals.

The global comparison across the remaining cell types showed that classification performance varied across cell populations. Some cell types supported accurate and balanced discrimination between conditions, whereas others showed weaker signals, smaller sample sizes, or stronger class imbalance. No single classifier dominated all subsets. Instead, the hypervolume-based multiobjective strategy selected distinct models for different cell types, indicating that cell population-specific model selection was more appropriate than imposing a universal classifier. In this setting, macro F1-score remained informative because it captured model behavior under class imbalance.

The structural analysis across cell types reinforced this heterogeneity. ANOVA produced the highest full-space separation ratios in most populations and therefore performed better under the geometric criterion alone. Nevertheless, the proposed method still preserved detectable and non-random disease-related structure and supported downstream modeling and interpretation. This result showed that a feature-selection strategy should not be judged exclusively by geometric separation, because predictive performance, stability, interpretability, and biological relevance provide complementary dimensions of quality.

This chapter showed that the proposed framework supported leakage-aware feature selection, model comparison, multiobjective selection, interpretable classification, and downstream biological validation in a single-cell disease setting. Although traditional methods such as ANOVA sometimes produced larger geometric separation, the proposed method integrated prediction with biological interpretability. These results support the proposed framework as a strategy for identifying disease-associated features and building

interpretable predictive models in single-cell transcriptomic studies.

This chapter presented the results of the proposed computational framework in a structured manner, covering preprocessing, classification, multiobjective analysis, feature selection, interpretability, stability, and biological validation. The chapter was organized to support both computational comparison and biological interpretation, thereby establishing the empirical basis for the discussion presented in the following chapter.

# 6 DISCUSSION

## 6.1 The chronic obstructive pulmonary disease dataset as a benchmark for feature selection

This study used the COPD dataset from Rustam et al. as a benchmark for feature-selection methods because the dataset links disease biology with a difficult classification setting. Rustam et al. mapped distal airway organization at single-cell resolution and reported COPD-associated changes in epithelial, structural, and immune compartments across defined airway regions (RUSTAM et al., 2023). The dataset therefore places feature selection in a setting where the disease signal does not belong to one lineage only. Different cell populations carry different abundances, transcriptional states, and degrees of overlap between normal and COPD cells. This structure limits the value of methods that work mainly in cleaner or more homogeneous data.

The results of this dissertation support this choice of benchmark. Figures 5.1 and 5.2 show that the retained male subset kept a diverse cell composition and still showed overlap between normal and COPD cells in the global transcriptomic space. Table 5.1 shows that preprocessing kept all annotated cells and removed only genes with zero counts. This preprocessing choice left the classification problem largely intact, instead of making the dataset easier through strict filtering. Table 5.2 also shows that some annotated populations did not enter the downstream models. These exclusions reflect a common problem in disease single-cell data: some cell types contain too few cells, or too uneven a class distribution, for stable supervised learning.

Class imbalance added another source of difficulty. Figure 5.12 and Table 5.11 show that several subsets contained many more normal cells than COPD cells, while only a smaller set showed a more even class distribution. Thus, the feature-selection methods faced conditions closer to biomedical data than to idealized benchmark data. High dimensionality, sparsity, uneven cell abundance, and class imbalance can affect both gene-ranking stability and model generalization (LUECKEN; THEIS, 2019; YANG; HUANG; LIU, 2021). In this dataset, the analysis therefore tested whether each method could select useful genes under these constraints, not only whether it could rank genes in a simplified setting.

The male-restricted design also shaped the interpretation. It reduced sex-associated transcriptional variation and made the cell-type comparisons more controlled during fea-

ture selection and model training. The same choice, however, limits the scope of the conclusions. The results refer most directly to the male subset analyzed here and should not be extended without caution to mixed-sex cohorts. This point matters for the SHAP interpretation because Table 5.9 includes Y-linked genes such as *RPS4Y1* and *DDX3Y*. These genes should be interpreted within the sampling design of this dissertation, rather than presented as general COPD biomarkers.

For these reasons, the Rustam et al. dataset provides a useful setting for evaluating feature selection in single-cell transcriptomic data. It includes disease-related variation, heterogeneous cell populations, high-dimensional gene expression, and uneven class structure. These properties make it appropriate for testing feature selection as part of the full classification and interpretation framework, rather than as a separate gene-ranking step.

## 6.2 Linear and nonlinear machine learning models for single-cell classification

### 6.2.1 Why no single classifier dominated all cell types

The results of this study do not support a single preference for linear or nonlinear classifiers in single-cell disease classification. Instead, the analysis linked model choice to the signal structure of each cell type. In the CD4-positive, alpha-beta T-cell subset, the MLP produced the most favorable result among the evaluated classifiers and occupied the only Pareto-optimal position. This model also yielded larger macro F1-score and hypervolume values than the remaining classifiers (Table 5.6; Figure 5.4). These results indicate that nonlinear decision boundaries represented the disease-associated structure of this subset more adequately than the linear alternatives.

The comparison across all modeled cell types gave a less uniform pattern. Table 5.12 shows that the selection procedure assigned Random Forest to more subsets during cross-validation than SVM, MLP, and Logistic Regression, while it assigned no subset to KNN. Table 5.13 retained the same general distribution in the held-out evaluation. This pattern indicates that each biological population imposed a distinct predictive setting. Some subsets favored classifiers with nonlinear capacity, whereas others favored linear margins or regularized linear decision functions.

This result agrees with single-cell machine-learning studies that reported competitive performance for linear classifiers, especially SVM and logistic regression, in au-

tomated cell identification. These models can handle high-dimensional gene-expression matrices when the number of measured genes exceeds the informative degrees of freedom (ABDELAAL et al., 2019; AEVERMANN et al., 2022). Other comparative studies also reported advantages for nonlinear methods when the dataset contains enough observations and the phenotype depends on interactions that additive models represent poorly (CAO et al., 2022). The present results fit this interpretation. They do not reject linear models; rather, they show that disease-state classification within a fixed cell type may require nonlinear modeling when the transcriptional boundary contains subtle or interaction-dependent variation.

### 6.2.2 Which model families were more appropriate under class imbalance

The global analysis gives model suitability under class imbalance a practical interpretation. In subsets where feature selection produced coherent disease-control separation, linear classifiers required no highly flexible boundary. Logistic Regression gave alternatively activated macrophages high held-out performance and gave lung microvascular endothelial cells competitive results (Table 5.13). SVM gave pulmonary alveolar type 2 cells, B cells, and respiratory tract goblet cells similarly favorable results. These cases indicate that linear models can classify cell types effectively when selected genes place disease-associated variation along a stable separable direction.

Nonlinear classifiers gave different subsets a better fit when disease status depended on richer expression patterns. The CD4-positive, alpha-beta T-cell subset provides the main example. In this subset, the MLP gave the evaluation larger macro F1-score and hypervolume values than Logistic Regression and SVM (Table 5.6), and the held-out test set gave it consistent independent performance (Figure 5.6). The MLP also gave multiciliated columnar cells of the tracheobronchial tree and bronchiolar smooth muscle cells favorable results (Table 5.13). These cases indicate that nonlinear models can represent disease-control separation when several genes jointly define the transcriptional boundary, rather than when one linear axis carries most of the signal.

Random Forest gave the global comparison an intermediate pattern. The model-selection procedure assigned it to several cell types, including fibroblasts, mast cells, and additional subsets (Tables 5.12 and 5.13). This pattern indicates that tree-based ensembles captured useful local rules when small groups of genes contributed jointly to prediction. The CD4-positive, alpha-beta T-cell analysis, however, gave Random Forest lower gen-

eralization than the MLP. Therefore, the results support a cell-type-specific interpretation of Random Forest performance, rather than a general preference for this model family.

KNN gave the weakest global pattern and received no final selection in the cross-validation comparison. Sparse high-dimensional single-cell matrices can give neighborhood relations unstable structure, and class imbalance can give local voting rules biased decision regions. These properties can disadvantage KNN even after feature reduction, especially when disease-associated cells occupy diffuse regions instead of compact manifolds. This interpretation agrees with discussions of high-dimensional classification and imbalance learning (HE; GARCIA, 2009; SAITO; REHMSMEIER, 2015).

In practical terms, the results give each model family a conditional role. Linear models suit cell types in which selected genes define stable disease-control separation, especially when limited sample size or class imbalance constrains model fitting. Nonlinear models suit cell types in which disease-associated transcriptional differences depend on gene interactions, local rules, or non-additive effects that remain after feature selection. The analysis therefore gives model choice an empirical basis: the data from each cell type, rather than the classifier family alone, should guide final model selection. This rationale supports the use of a flexible model-selection procedure across subsets instead of a fixed classifier strategy.

## 6.3 Multiobjective model selection in imbalanced single-cell classification

This study used multiobjective optimization to give final model selection a clearer methodological basis. In imbalanced transcriptomic classification, a single metric can give the analysis a distorted ranking. Accuracy can assign high scores to models that mainly predict the majority class, while sensitivity, specificity, and macro F1-score can expose errors in minority-class prediction (HE; GARCIA, 2009; SAITO; REHMSMEIER, 2015). A selection rule based on one metric can therefore favor classifiers that look adequate in numerical summaries but give the data poor disease-control discrimination.

The CD4-positive, alpha-beta T-cell analysis illustrates this problem directly. Table 5.6 and Figure 5.4 show that the MLP produced a balanced joint profile across the tracked metrics and occupied the only non-dominated position in the comparative analysis. The selection procedure therefore did not choose the model only because it maximized one isolated value. It selected the classifier that gave the dataset a balanced profile across accuracy, sensitivity, specificity, and macro F1-score. The hypervolume criterion

made this rationale explicit, whereas a single-metric ranking would have hidden part of the trade-off.

The global comparison gives the same argument additional support. Table 5.12 shows that the hypervolume-based selection matched the highest macro F1-score in 23 of the 25 analyzed cell types. This agreement indicates that the multiobjective procedure gave the analysis stable choices rather than arbitrary selections. In most subsets, it selected the same classifier favored by the metric with greater sensitivity to imbalance. In the remaining cases, it gave the comparison a way to account for accuracy, sensitivity, specificity, and macro F1-score together. These few differences strengthen the methodological argument because they show when F1-score alone did not describe the full performance balance.

This behavior agrees with the role of hypervolume in multiobjective optimization. Hypervolume gives the selection procedure an indicator that combines dominance and coverage relative to a reference point (AUGER et al., 2012). In this dissertation, this property gave model comparison a formal rule for rejecting classifiers that improved one metric by sacrificing others. This point matters in single-cell disease classification because each cell type gives the model a different class composition and error structure. A classifier can therefore differ from another not only in predictive strength, but also in the balance that it gives false-positive and false-negative errors. The hypervolume framework made these trade-offs visible during model selection.

The held-out evaluation indicates that this strategy also carried into independent testing. Table 5.13 shows that the selected models retained the same general performance pattern in the final train-test analysis. This result suggests that the multiobjective procedure did not only fit the internal cross-validation folds. It gave the framework a stable rule for moving from model comparison to final model commitment.

For this reason, the thesis does not present hypervolume as a replacement for macro F1-score. Macro F1-score remains necessary because class imbalance shaped the classification task. The combination of macro F1-score, Pareto ranking, and hypervolume gives model selection a more defensible criterion than any single metric alone. In single-cell disease classification, where each cell type can impose a different error profile, this multiobjective strategy gives the framework a practical and transparent basis for choosing classifiers.

## 6.4 Interpretability of machine learning in single-cell transcriptomics

Interpretability is not an optional layer in this study; it is part of the analytical objective. In single-cell transcriptomics, a classifier that predicts disease state accurately but fails to reveal which genes drive its decisions contributes limited biological value. This is particularly true in omics settings, where the same predictive performance can emerge from feature sets with very different mechanistic meanings (KARIM et al., 2023). The SHAP-based analysis used here addresses this issue by decomposing the output of the selected models into feature-level contributions and by connecting these contributions to expression behavior.

The CD4-positive, alpha-beta T-cell case illustrates this point. Table 5.9 and Figure 5.9 show that the final model relied on a compact set of influential genes rather than on diffuse low-amplitude contributions across the selected feature space. This result indicates that the model used a concentrated discriminative signature anchored in a limited number of genes with measurable attribution magnitudes.

The class-specific SHAP analysis adds an additional layer of biological insight. Figure 5.9B shows that several genes produced both positive and negative local attributions within the COPD output. This pattern is highly relevant in disease-state classification. It suggests that the classifier did not merely tag one class using one fixed set of unidirectional markers. Instead, it used a shared core of genes whose local effects depended on the surrounding transcriptional context. In biological terms, this is a more plausible decision structure than one based on rigid one-gene/one-class associations, especially in immune populations where activation, stress response, and homeostatic programs can be partially overlapping.

The expression-attribution correlation analysis further improved interpretability because it linked model importance to gene abundance rather than treating attribution as a purely abstract model quantity. This step is valuable in single-cell studies, where high importance does not necessarily mean high expression, and where the same expression level can contribute differently across cellular contexts. By correlating SHAP values with expression, the framework distinguished genes whose increased expression tended to strengthen disease prediction from genes whose expression tended to shift attribution toward the control class. This approach moves the analysis from feature ranking to feature behavior.

The interpretability results also require caution. Because the benchmark was re-

stricted to male samples, the prominence of Y-linked genes among the top SHAP features should not be overgeneralized. Their attribution values apply to this benchmark, but their biological interpretation must remain tied to the study design. This case illustrates a limitation of interpretable machine learning: an explanation inherits the sampling frame in which the model learned it. Interpretability therefore does not replace biological judgment; it gives that judgment a more explicit basis.

These points align with recent work arguing that explainable and interpretable machine learning supports bioinformatics when prediction must be linked to mechanistic interpretation (KARIM et al., 2023; GONZALEZ-FERRER et al., 2024). In this dissertation, interpretability showed that the selected model was not only predictive but also traceable to a biologically structured set of disease-relevant signals. The framework therefore connected prediction and explanation within the same analytical framework.

## 6.5 Biological meaning of the enriched GO terms in CD4-positive alpha-beta T cells

The enrichment results for the CD4-positive, alpha-beta T-cell subset support the biological interpretation of the proposed method. Figure 5.11 and Table 5.10 show that all evaluated feature-selection strategies recovered translation-related biology. However, the proposed method differed from the others by combining this translational core with stress-response, chaperone-mediated refolding, apoptotic regulation, and additional regulatory processes. This profile fits the biological context of COPD, a chronic inflammatory disease in which immune cells are repeatedly exposed to activation signals, proteostatic stress, and altered tissue microenvironments (FREEMAN et al., 2014; ROBERTS et al., 2015; PAATS et al., 2012).

This point deserves emphasis because the proposed method did not produce the densest network. Wilcoxon generated the most connected PPI graph and the highest average node degree (Table 5.10), but its enriched terms remained centered on broad biosynthetic and housekeeping-like programs. By contrast, the proposed method recovered the largest number of GO biological process terms and identified a disease-associated module that linked translation to heat-shock and stress-adaptation processes. For discussion purposes, this distinction is crucial. A denser network does not automatically imply a more disease-informative network. In this case, Wilcoxon appeared to emphasize strong general-expression differences, whereas the proposed method better captured a functionally integrated disease-related program.

The disease-oriented interpretation becomes stronger when the positively contributing genes selected by the proposed method are considered separately. The corresponding enrichment profile again highlighted translation, ribosome-associated activity, protein homeostasis, and stress-response processes (Figure 5.11). This pattern is consistent with the idea that activated or chronically stimulated T cells require tight control of protein synthesis and post-transcriptional regulation (JURGENS; POPOVIĆ; WOLKERS, 2021; SINCLAIR; CANTRELL, 2025). It is also compatible with broader evidence showing that stress-response pathways and integrated stress signaling are central to how cells adapt to proteostatic challenges (COSTA-MATTIOLI; WALTER, 2020; WILLIAMS et al., 2024). Thus, the enrichment signature recovered here is not only statistically coherent; it is also biologically plausible in the context of immune dysregulation.

Among the evaluated strategies, the proposed method showed disease-oriented biological coherence in the CD4-positive, alpha-beta T-cell subset. This interpretation does not imply that the proposed method alone identified disease-relevant biology or that the alternative methods were biologically uninformative. Rather, the proposed method combined three properties in the CD4 analysis: predictive contribution, disease-directional attribution, and enrichment specificity. Wilcoxon identified a connected network, but its enriched terms centered more on broad biosynthetic functions. ANOVA and RFE recovered narrower functional spaces. Random Forest produced more fragmented modules. The proposed method linked positively disease-associated genes to a program that combined translational, stress-response, and regulatory processes. Within the logic of this dissertation, that combination supports a disease-oriented interpretation.

This conclusion is also important for interpreting the SHAP results. Genes such as *HSP90AA1*, *HSPA1A*, *HSPA1B*, *JUNB*, and related factors did not appear as isolated markers. Instead, they participated in a network architecture that suggests coordinated adaptive responses. In chronic inflammatory conditions, such coordination is often more informative than isolated differential expression because it reflects how cells reorganize multiple regulatory layers rather than simply changing one transcript at a time. Therefore, the biological discussion favors the proposed method not because it maximized a single enrichment statistic, but because it produced the most interpretable disease-associated program.

## 6.6 What the top 25 features selected by each method reveal about feature-selection behavior

The comparison of the top 25 genes selected by each method shows that feature-selection strategies operationalize different definitions of relevance rather than simply reordering the same genes. Figure 5.7 shows partial overlap among gene lists together with method-specific components. Figure 5.8 shows that Random Forest, RFE, and ANOVA produced distinct score profiles for their top-ranked genes and differed in computational cost. Figure 5.10 and Table 5.10 show that these differences affected geometric organization and biological interpretation. The top 25 genes selected by each method should therefore be treated as different summaries of the signal space.

The proposed method selected genes according to predictive contribution aggregated through SHAP-based interpretation. Consequently, its top 25 genes reflected model-useful, class-relevant features rather than only marginal differential expression. This helps explain why the proposed method produced the most biologically integrated enrichment profile even though it did not maximize geometric separation in the full standardized space. It prioritized genes that contributed to disease discrimination in the chosen predictive setting, not genes that merely produced the strongest univariate contrasts.

Wilcoxon and ANOVA reflected a different notion of relevance. As filter methods, they emphasize marginal statistical separation between groups. In single-cell data, this behavior can strongly privilege genes with stable rank-based or variance-based differences, including genes associated with abundant translational machinery or other high-signal programs (SAEYS; INZA; nAGA, 2007; YANG; HUANG; LIU, 2021; PULLIN; MCCARTHY, 2024). This logic helps explain why ANOVA achieved the strongest full-space separation ratio in the CD4 analysis (Figure 5.10), and why both Wilcoxon and ANOVA tended to recover broader translation-centered signatures. These methods are therefore highly effective when the main goal is strong geometric discrimination or marker-like statistical contrast, but they may be less specific when the goal is disease-oriented biological interpretation.

Random Forest and RFE also reflected distinct perspectives on relevance. Random Forest selected genes that repeatedly improved tree splits, which made it sensitive to nonlinear interactions and local partitioning rules. This behavior likely contributed to its strong performance in several cell types, but it also produced more fragmented enrichment profiles in the CD4 subset. RFE, in turn, selected genes that best supported

a classifier-specific decision boundary through iterative elimination. This approach can produce compact predictive subsets, but it may also narrow the functional scope of the selected genes. The sparse enrichment profile observed for RFE in Figure 5.11 is consistent with this interpretation.

The random baseline plays an important role in this comparison because it shows what a non-informative top-25 set looks like. The poor geometric structure associated with random selection in Figure 5.10 confirms that the informed methods captured real disease-related organization. Therefore, even when the methods differed in biological specificity or geometric compactness, they remained clearly distinct from a null selection strategy.

This comparison aligns with a central argument in the feature-selection literature: no method optimizes all desirable properties at once (SAEYS; INZA; nAGA, 2007; YANG; HUANG; LIU, 2021; TOWNES et al., 2019). Some methods prioritize statistical contrast, others prioritize classifier utility, and others preserve interpretability directly. The results of this dissertation follow that pattern. ANOVA often produced larger geometric separation; Wilcoxon produced a more connected CD4 network; Random Forest performed well across several heterogeneous cell types; and the proposed method integrated prediction with biological interpretation. The top 25 genes selected by each method therefore reflect both the biology of the dataset and the bias introduced by the selection strategy.

For this reason, the proposed method should not be judged solely by whether its top 25 genes matched those from classical filters. Their partial divergence is expected and, in fact, informative. It shows that the proposed framework captures a different facet of relevance: the genes that the final predictive model actually uses to discriminate disease state and that remain biologically coherent after interpretation. In the context, that is precisely the kind of relevance that matters most.

# 7 CONCLUSION

This dissertation developed and evaluated a computational framework for single-cell transcriptomic analysis that integrates feature selection, machine learning, interpretable artificial intelligence, and multiobjective optimization. The main objective was to identify informative genes and construct predictive models that were both accurate and biologically interpretable. The evaluation supports the suitability of the framework for high-dimensional, sparse, and heterogeneous single-cell data. By combining leakage-aware preprocessing, cell type-specific modeling, fold-wise feature selection, Bayesian hyperparameter optimization, Pareto-guided model comparison, SHAP-based interpretation, and downstream biological validation, the framework provided an integrated strategy for transcriptomic classification in which predictive performance and biological interpretability were addressed jointly.

The evaluation on the COPD dataset showed that no single classifier generalized across all cell types. Predictive performance varied according to the local biological and statistical structure of each cell population. This finding supports one conclusion of the dissertation: model selection in single-cell transcriptomic studies should use cell type-specific and multiobjective evaluation rather than a universal modeling assumption. The detailed analysis of CD4-positive, alpha-beta T cells illustrated this point. In this subset, the Multilayer Perceptron produced a balanced profile across evaluation metrics, formed the only Pareto-optimal solution, and preserved comparable performance in the final hold-out evaluation. The interpretability analyses showed that the final model relied on a compact set of genes with biologically interpretable predictive contributions.

The comparative analyses across feature selection methods also showed that different strategies emphasize different notions of relevance. Although ANOVA and Wilcoxon performed well under specific criteria, the proposed method integrated predictive utility, interpretability, and biological coherence within the CD4 analysis. This result supports the conclusion that feature selection should not be judged only by geometric separation or statistical contrast, but by its combined impact on classification, explanation, and downstream biological validation. Overall, this dissertation presents the proposed framework as a leakage-aware and interpretable strategy for selecting disease-associated features and evaluating classifiers in single-cell transcriptomic data. Because the framework was evaluated on one COPD dataset, future work should test its generalizability in external cohorts and multiclass datasets.

# REFERENCES

ABDELAAL, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, v. 20, n. 1, p. 194, 2019.

AEVERMANN, B. D. et al. Machine learning for cell type classification from single nucleus RNA sequencing data. *PLOS ONE*, v. 17, n. 9, p. e0275070, 2022.

AGARWAL, N.; DAS, S. Interpretable machine learning tools: A survey. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020. p. 1528–1534. Available from Internet: <http://dx.doi.org/10.1109/SSCI47803.2020.9308260>.

AGYEMANG, E. F. Anomaly detection using unsupervised machine learning algorithms: A simulation study. *Scientific African*, v. 26, p. e02386, 2024. ISSN 2468-2276. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S2468227624003284>.

AHLMANN-ELTZE, C.; HUBER, W. Comparison of transformations for single-cell rna-seq data. *Nature Methods*, v. 20, p. 665–672, 2023.

ALBERTS, B. et al. From dna to rna. In: *Molecular Biology of the Cell*. 4. ed. [S.l.]: Garland Science, 2002.

ALFEILAT, H. A. A. et al. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, SAGE Publications, v. 7, n. 4, p. 221–248, dec. 2019. ISSN 2167-647X. Available from Internet: <http://dx.doi.org/10.1089/big.2018.0175>.

ALJUAID, A.; ANWAR, M. Survey of supervised learning for medical image processing. *SN Computer Science*, Springer Science and Business Media LLC, v. 3, n. 4, may 2022. ISSN 2661-8907. Available from Internet: <http://dx.doi.org/10.1007/s42979-022-01166-1>.

ALLEN, G. M. et al. Interpretable machine learning for discovery. *arXiv preprint arXiv:2308.01475*, 2023.

ALTMAN, N.; KRZYWINSKI, M. Ensemble methods: bagging and random forests. *Nature Methods*, Springer Science and Business Media LLC, v. 14, n. 10, p. 933–934, sep. 2017. ISSN 1548-7105. Available from Internet: <http://dx.doi.org/10.1038/nmeth.4438>.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, v. 46, n. 3, p. 175–185, 1992.

ANG, J. C. et al. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Institute of Electrical and Electronics Engineers (IEEE), v. 13, n. 5, p. 971–989, sep. 2016. ISSN 1545-5963. Available from Internet: <http://dx.doi.org/10.1109/TCBB.2015.2478454>.

AUDET, C. et al. Performance indicators in multiobjective optimization. *European Journal of Operational Research*, v. 292, n. 2, p. 397–422, 2021.

AUGER, A. et al. Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications. *Theoretical Computer Science*, v. 425, p. 75–103, 2012.

AUSLANDER, N. et al. Incorporating machine learning into established bioinformatics frameworks. *International Journal of Molecular Sciences*, v. 22, n. 6, p. 2903, 2021.

BADILLO, S. et al. An introduction to machine learning. *Clinical Pharmacology amp; Therapeutics*, Wiley, v. 107, n. 4, p. 871–885, mar. 2020. ISSN 1532-6535. Available from Internet: <http://dx.doi.org/10.1002/cpt.1796>.

BALZER, M. S. et al. How to get started with single-cell rna sequencing data analysis: a practical guide. *Kidney360*, v. 2, n. 10, p. 1681–1692, 2021.

BARBIERI, M. C.; GRISCI, B. I.; DORN, M. Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications*, Elsevier BV, v. 249, p. 123667, 2024. ISSN 0957-4174. Available from Internet: <http://dx.doi.org/10.1016/j.eswa.2024.123667>.

BAYAT, A. Bioinformatics. *BMJ*, v. 324, n. 7344, p. 1018–1022, 2002.

BECHT, E. et al. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, v. 37, n. 1, p. 38–44, 2019.

BELEITES, C. et al. Sample size planning for classification models. *Analytica Chimica Acta*, v. 760, p. 25–33, 2013.

BENJAMIN, K. J. M.; KATIPALLI, T.; PAQUOLA, A. C. M. drfetools: Dynamic recursive feature elimination for omics. *Bioinformatics*, v. 39, n. 8, p. btad513, 2023.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012.

BERGSTRA, J. et al. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, v. 8, n. 1, p. 014008, 2015.

BERGSTRA, J. et al. Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, v. 8, n. 1, p. 014008, 2015.

BERGSTRA, J.; YAMINS, D.; COX, D. D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: *Proceedings of the 30th International Conference on Machine Learning*. [S.l.]: PMLR, 2013. (Proceedings of Machine Learning Research, 1), p. 115–123.

BEUME, N. et al. On the complexity of computing the hypervolume indicator. *IEEE Transactions on Evolutionary Computation*, v. 13, n. 5, p. 1075–1082, 2009.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2008, n. 10, p. P10008, 2008.

BORGONOVO, E. et al. The many shapley values for explainable artificial intelligence. *European Journal of Operational Research*, 2024.

BORHANI, N. et al. A deep learning approach to predict inter-omics interactions in multi-layer networks. *BMC Bioinformatics*, Springer Science and Business Media LLC, v. 23, n. 1, jan. 2022. ISSN 1471-2105. Available from Internet: <http://dx.doi.org/10.1186/s12859-022-04569-2>.

BREIMAN, L. Random forests. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, oct. 2001. ISSN 1573-0565. Available from Internet: <http://dx.doi.org/10.1023/A:1010933404324>.

BROWN, C. E. Multiple logistic regression. In: ____. *Applied Multivariate Statistics in Geohydrology and Related Sciences*. Springer Berlin Heidelberg, 1998. p. 147–151. ISBN 9783642803284. Available from Internet: <http://dx.doi.org/10.1007/978-3-642-80328-$4_1$2>.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, v. 3, n. 1, p. 1–27, 1974.

CAO, X. et al. A systematic evaluation of methods for cell phenotype classification using single-cell RNA sequencing data. *Frontiers in Genetics*, v. 13, p. 836798, 2022.

CHAKRABORTY, S. et al. Multi-omics approaches in cancer biology: New era and future landscape. *Molecular and Cellular Probes*, v. 76, p. 101936, 2024.

CHAPPELL, L.; RUSSELL, A. J. C.; VOET, T. Single-cell (multi)omics technologies. *Annual Review of Genomics and Human Genetics*, v. 22, p. 15–41, 2021.

CHEN, C. et al. Applications of multi-omics analysis in human diseases. *MedComm*, v. 4, n. 1, p. e315, 2023.

CHEN, H.; LU, Y.; RAO, Y. A self-training interpretable cell type annotation framework using specific marker gene. *Bioinformatics*, v. 40, n. 10, p. btae569, 09 2024. ISSN 1367-4811. Available from Internet: <https://doi.org/10.1093/bioinformatics/btae569>.

CHOW, A. et al. Concepts and new developments in droplet-based single-cell sequencing. *Nature Reviews Bioengineering*, 2024.

CHUNG, M. K. *Introduction to logistic regression*. arXiv, 2020. Available from Internet: <https://arxiv.org/abs/2008.13567>.

COCK, P. J. A. et al. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, v. 38, n. 6, p. 1767–1771, 2010.

COELLO, C. A. C. Evolutionary multiobjective optimization: a short review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 1, n. 5, p. 444–454, 2011.

COFFIN, J. M. 50th anniversary of the discovery of reverse transcriptase. *Molecular Biology of the Cell*, v. 32, n. 7, p. 573–575, 2021.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, Springer Science and Business Media LLC, v. 20, n. 3, p. 273–297, sep. 1995. ISSN 1573-0565. Available from Internet: <http://dx.doi.org/10.1007/BF00994018>.

COSTA-MATTIOLI, M.; WALTER, P. The integrated stress response: From mechanism to disease. *Science*, v. 368, n. 6489, p. eaat5314, 2020.

COTTRELL, S.; HOZUMI, Y.; WEI, G.-W. K-nearest-neighbors induced topological pca for single cell rna-sequence data analysis. *Computers in Biology and Medicine*, Elsevier BV, v. 175, p. 108497, jun. 2024. ISSN 0010-4825. Available from Internet: <http://dx.doi.org/10.1016/j.compbiomed.2024.108497>.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.

CRICK, F. H. C. Central dogma of molecular biology. *Nature*, v. 227, p. 561–563, 1970.

CUI, J. et al. Exploring machine learning strategies for single-cell transcriptomic analysis in wound healing. *Burns amp; Trauma*, Oxford University Press (OUP), v. 13, 2025. ISSN 2321-3876. Available from Internet: <http://dx.doi.org/10.1093/burnst/tkaf032>.

DAM, S. van et al. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, v. 19, n. 4, p. 575–592, 2018.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

DEB, K. *Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction*. Kanpur, India, 2011. Available from Internet: <http://www.iitk.ac.in/kangal/deb.htm>.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

DU, M.; LIU, N.; HU, X. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*, 2018.

DíAZ-URIARTE, R.; ANDRéS, S. Alvarez de. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, v. 7, p. 3, 2006.

ELKAHWAGY, D. M. A. S.; KIRIACOS, C. J.; MANSOUR, M. Logistic regression and other statistical tools in diagnostic biomarker studies. *Clinical and Translational Oncology*, Springer Science and Business Media LLC, v. 26, n. 9, p. 2172–2180, mar. 2024. ISSN 1699-3055. Available from Internet: <http://dx.doi.org/10.1007/s12094-024-03413-8>.

FELTES, B.; POLONI, J.; DORN, M. Benchmarking and Testing Machine Learning Approaches with BARRA:CuRDa, a Curated RNA-Seq Database for Cancer Research. *Journal of Computational Biology*, v. 28, n. 9, p. 931–944, 2021. Available from Internet: <https://doi.org/10.1089/cmb.2020.0463>.

FELTES, B. C. et al. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, SAGE Publications, v. 26, n. 4, p. 376–386, abr. 2019. ISSN 1557-8666. Available from Internet: <http://dx.doi.org/10.1089/cmb.2018.0238>.

FENG, X. et al. Advancing single-cell rna-seq data analysis through the fusion of multi-layer perceptron and graph neural network. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 25, n. 1, nov. 2023. ISSN 1477-4054. Available from Internet: <http://dx.doi.org/10.1093/bib/bbad481>.

FISCHER, F. et al. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, Springer Science and Business Media LLC, v. 15, n. 1, aug. 2024. ISSN 2041-1723. Available from Internet: <http://dx.doi.org/10.1038/s41467-024-51059-5>.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936.

FREEMAN, C. M. et al. Basal gene expression by lung CD4+ t cells in chronic obstructive pulmonary disease identifies independent molecular correlates of airflow obstruction and emphysema extent. *PLOS ONE*, v. 9, n. 5, p. e96421, 2014.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Incorporated, 2022. ISBN 9781098125967. Available from Internet: <https://books.google.com.br/books?id=zpoc0AEACAAJ>.

GONZALEZ-FERRER, J. et al. SIMS: A deep-learning label transfer tool for single-cell RNA sequencing analysis. *Cell Genomics*, v. 4, n. 6, p. 100581, 2024.

GONZáLEZ-CAMACHO, J. M. et al. Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics*, Springer Science and Business Media LLC, v. 17, n. 1, mar. 2016. ISSN 1471-2164. Available from Internet: <http://dx.doi.org/10.1186/s12864-016-2553-1>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

GUERREIRO, A. P. et al. The hypervolume indicator: Computational problems and algorithms. *ACM Computing Surveys*, v. 54, n. 6, p. 1–42, 2021.

GUO, J.; PRICE, D. L. Transcription: the epicenter of gene expression. *Journal of Cellular Physiology*, v. 228, n. 3, p. 1–8, 2014.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, v. 3, p. 1157–1182, 2003.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, v. 46, p. 389–422, 2002.

HABIB, N. et al. Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods*, v. 14, n. 10, p. 955–958, 2017.

HABIB, N. et al. Div-seq: Single-nucleus rna-seq reveals dynamics of rare adult newborn neurons. *Science*, v. 353, n. 6302, p. 925–928, 2016.

HAFEN, R.; CRITCHLOW, T. Eda and ml – a perfect pair for large-scale data analysis. In: *2013 IEEE International Symposium on Parallel amp; Distributed Processing, Workshops and Phd Forum*. IEEE, 2013. p. 1894–1898. Available from Internet: <http://dx.doi.org/10.1109/IPDPSW.2013.118>.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, JSTOR, v. 28, n. 1, p. 100, 1979. ISSN 0035-9254. Available from Internet: <http://dx.doi.org/10.2307/2346830>.

HASELTINE, W. A. The rna revolution in the central dogma of molecular biology. *Journal of Biological Chemistry*, v. 300, n. 12, p. 108302, 2024.

HASIN, Y.; SELDIN, M.; LUSIS, A. Multi-omics approaches to disease. *Genome Biology*, v. 18, p. 83, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [S.l.]: Springer, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009.

HE, J. et al. Practical bioinformatics pipelines for single-cell rna-seq data analysis. *Briefings in Functional Genomics*, v. 22, n. 3, p. bqac017, 2023.

HEGDE, A.; NGUYEN, T.; CHENG, J. Machine learning methods for gene regulatory network inference. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 26, n. 5, aug. 2025. ISSN 1477-4054. Available from Internet: <http://dx.doi.org/10.1093/bib/bbaf470>.

HEUMOS, L. et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, v. 24, p. 550–572, 2023.

HICKS, S. C. et al. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, v. 19, n. 4, p. 562–578, 2018.

HORN, D.; BISCHL, B. Multi-objective parameter configuration of machine learning algorithms using model-based optimization. In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016. p. 1–8. Available from Internet: <http://dx.doi.org/10.1109/SSCI.2016.7850221>.

HRDLIcKOVÁ, R.; TOLOUE, M.; TIAN, B. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, v. 8, n. 1, p. e1364, 2017.

HUANG, H. et al. Identification and evaluation of plasma exosome rna biomarkers for non-invasive diagnosis of hepatocellular carcinoma using rna-seq. *BMC Cancer*, Springer Science and Business Media LLC, v. 24, n. 1, dec. 2024. ISSN 1471-2407. Available from Internet: <http://dx.doi.org/10.1186/s12885-024-13332-0>.

HULETT, C.; HALL, A.; QU, G. Dynamic selection of k nearest neighbors in instance-based learning. In: *2012 IEEE 13th International Conference on Information Reuse amp; Integration (IRI)*. IEEE, 2012. p. 85–92. Available from Internet: <http://dx.doi.org/10.1109/IRI.2012.6302995>.

HWANG, B.; LEE, J. H.; BANG, D. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, v. 50, n. 8, p. 1–14, 2018.

IMOTO, Y. et al. Resolution of the curse of dimensionality in single-cell rna sequencing data analysis. *Communications Biology*, v. 5, p. 790, 2022.

JIANG, R. et al. The zero-inflation controversy about scrna-seq data. *Genome Biology*, v. 23, p. 31, 2022.

JOVIC, D. et al. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, v. 12, n. 3, p. e694, 2022.

JURGENS, A. P.; POPOVIĆ, B.; WOLKERS, M. C. T cells at work: How post-transcriptional mechanisms control t cell homeostasis and activation. *European Journal of Immunology*, v. 51, n. 9, p. 2178–2187, 2021.

KAIKKONEN, M. U.; LAM, M. T. Y.; GLASS, C. K. Non-coding rnas as regulators of gene expression and epigenetics. *Cardiovascular Research*, v. 90, n. 3, p. 430–440, 2011.

KARIM, M. R. et al. Explainable AI for bioinformatics: Methods, tools and applications. *Briefings in Bioinformatics*, v. 24, n. 5, p. bbad236, 2023.

KASHYAP, N.; MISHRA, A. Chapter 1 - a discourse on metaheuristics techniques for solving clustering and semisupervised learning models. In: MISHRA, S. et al. (Ed.). *Cognitive Big Data Intelligence with a Metaheuristic Approach*. Academic Press, 2022, (Cognitive Data Science in Sustainable Computing). p. 1–19. ISBN 978-0-323-85117-6. Available from Internet: <https://www.sciencedirect.com/science/article/pii/B9780323851176000121>.

KERR, M. K.; MARTIN, M.; CHURCHILL, G. A. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, v. 7, n. 6, p. 819–837, 2000.

KIM, N.; KIM, S.; KIM, D. Perspectives on single-nucleus rna sequencing in different biological and clinical applications. *BMB Reports*, v. 56, n. 1, p. 10–18, 2023.

KIM, T. H.; ZHOU, X.; CHEN, M. Demystifying “drop-outs” in single-cell umi data. *Genome Biology*, v. 21, p. 196, 2020.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. [S.l.]: Morgan Kaufmann, 1995. v. 2, p. 1137–1143.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, n. 1–2, p. 273–324, 1997.

KOONIN, E. V. Why the central dogma: on the nature of the great biological exclusion principle. *Biology Direct*, v. 10, p. 52, 2015.

KUCHIBHOTLA, A. K. et al. *All of Linear Regression*. arXiv, 2019. Available from Internet: <https://arxiv.org/abs/1910.06386>.

LAAKSONEN, J.; OJA, E. Classification with learning k-nearest neighbors. In: *Proceedings of International Conference on Neural Networks (ICNN’96)*. IEEE. (ICNN-96, v. 3), p. 1480–1483. Available from Internet: <http://dx.doi.org/10.1109/ICNN.1996.549118>.

LAMBERT, N. *Reinforcement Learning from Human Feedback*. arXiv, 2025. Available from Internet: <https://arxiv.org/abs/2504.12501>.

LANGFELDER, P.; HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, v. 9, p. 559, 2008.

LEDERER, J. *Activation Functions in Artificial Neural Networks: A Systematic Overview*. arXiv, 2021. Available from Internet: <https://arxiv.org/abs/2101.09957>.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, n. 17, p. 1–5, 2017. Available from Internet: <http://jmlr.org/papers/v18/16-365.html>.

LI, J.; VARGHESE, R. S.; RESSOM, H. W. Rna-seq data analysis. In: ____. *RNA Amplification and Analysis*. Springer US, 2024. p. 263–290. ISBN 9781071639184. Available from Internet: <http://dx.doi.org/10.1007/978-1-0716-3918-4$_1$8>.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 16, n. 6, p. 321–332, may 2015. ISSN 1471-0064. Available from Internet: <http://dx.doi.org/10.1038/nrg3920>.

LIEW, A. W.-C.; LAW, N.-F.; YAN, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 12, n. 5, p. 498–513, dec. 2010. ISSN 1477-4054. Available from Internet: <http://dx.doi.org/10.1093/bib/bbq080>.

LITTLE, R. J. A.; RUBIN, D. B. *Statistical Analysis with Missing Data*. 3. ed. [S.l.]: Wiley, 2019.

LOUTEN, J. Features of host cells. In: ____. *Essential Human Virology*. Elsevier, 2016. p. 31–48. ISBN 9780128009475. Available from Internet: <http://dx.doi.org/10.1016/B978-0-12-800947-5.00003-X>.

LUECKEN, M. D.; THEIS, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, v. 15, n. 6, p. e8746, 2019.

LUN, A. T. L.; BACH, K.; MARIONI, J. C. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biology*, v. 17, n. 1, p. 75, 2016.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, 2017.

MARCINKEVIcS, R.; VOGT, J. E. Interpretable and explainable machine learning: A methods-centric overview with practical guidelines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 13, n. 6, p. e1493, 2023.

MARLER, R. T.; ARORA, J. S. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, v. 26, n. 6, p. 369–395, 2004.

MARTORELL-MARUGÁN, J. et al. Deep learning in omics data analysis and precision medicine. In: *Computational Biology*. [S.l.]: Springer, 2019. p. 37–53.

MASTROPIETRO, A.; FELDMANN, C.; BAJORATH, J. Calculation of exact shapley values for explaining support vector machine models using the radial basis function kernel. *Scientific Reports*, Springer Science and Business Media LLC, v. 13, n. 1, nov. 2023. ISSN 2045-2322. Available from Internet: <http://dx.doi.org/10.1038/s41598-023-46930-2>.

MATTICK, J. S. et al. Long non-coding rnas: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*, v. 24, p. 430–447, 2023.

MCCUE, M. E.; MCCOY, A. M. The scope of big data in one medicine: Unprecedented opportunities and challenges. *Frontiers in Veterinary Science*, v. 4, p. 194, 2017.

MELKUMOVA, L.; SHATSKIKH, S. Comparing ridge and lasso estimators for data analysis. *Procedia Engineering*, Elsevier BV, v. 201, p. 746–755, 2017. ISSN 1877-7058. Available from Internet: <http://dx.doi.org/10.1016/j.proeng.2017.09.615>.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Elsevier BV, v. 267, p. 1–38, feb. 2019. ISSN 0004-3702. Available from Internet: <http://dx.doi.org/10.1016/j.artint.2018.07.007>.

MORRIS, J. S. Statistical contributions to bioinformatics: Design, modeling, analysis, and inference. *Annual Review of Statistics and Its Application*, v. 4, p. 405–429, 2017.

MUBARAK, G. et al. Recent major transcriptomics and epitranscriptomics advances in cardiovascular, neurodevelopmental, and cancer diseases. *Molecules*, v. 27, n. 4, p. 1213, 2022.

MURDOCH, W. J. et al. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, v. 116, n. 44, p. 22071–22080, 2019.

NAGARHALLI, T. P.; VAZE, V.; RANA, N. K. Impact of machine learning in natural language processing: A review. In: *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, 2021. p. 1529–1534. Available from Internet: <http://dx.doi.org/10.1109/ICICV50876.2021.9388380>.

NAGY, P. D.; POGANY, J. The dependence of viral rna replication on co-opted host factors. *Nature Reviews Microbiology*, v. 10, n. 2, p. 137–149, 2011.

NG, S. et al. The benefits and pitfalls of machine learning for biomarker discovery. *Cell and Tissue Research*, Springer Science and Business Media LLC, v. 394, n. 1, p. 17–31, 2023. ISSN 1432-0878. Available from Internet: <http://dx.doi.org/10.1007/s00441-023-03816-z>.

OULAS, A. et al. Systems bioinformatics: Increasing precision of computational diagnostics and therapeutics through network-based approaches. *Briefings in Bioinformatics*, v. 20, n. 3, p. 806–824, 2019.

OZSOLAK, F.; MILOS, P. M. Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, v. 12, n. 2, p. 87–98, 2011.

PAATS, M. S. et al. Systemic CD4+ and CD8+ t-cell cytokine profiles correlate with GOLD stage in stable COPD. *European Respiratory Journal*, v. 40, n. 2, p. 330–337, 2012.

PARK, M. et al. Single-cell identity definition using random forests and recursive feature elimination. *bioRxiv*, 2020.

PAVLIDIS, P. Using anova for gene selection from microarray studies of the nervous system. *Methods*, v. 31, n. 4, p. 282–289, 2003.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PETUKHOV, V. et al. dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell rna-seq experiments. *Genome Biology*, v. 19, p. 78, 2018.

PHAM, G.; MILAZZO, P. A comprehensive review of the use of shapley value to assess node importance in the analysis of biological networks. *Computer Methods and Programs in Biomedicine Update*, v. 7, p. 100185, 2025. ISSN 2666-9900. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S2666990025000096>.

PIMPALWAR, N. et al. Methods for isolation and transcriptional profiling of individual cells from frozen human heart tissue. *BMC Biology*, v. 19, p. 1, 2021.

PINU, F. R. et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites*, v. 9, n. 4, p. 76, 2019.

POUYAN, M. B.; KOSTKA, D. Random forest based similarity learning for single cell rna sequencing data. *Bioinformatics*, Oxford University Press (OUP), v. 34, n. 13, p. i79–i88, jun. 2018. ISSN 1367-4811. Available from Internet: <http://dx.doi.org/10.1093/bioinformatics/bty260>.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 9, n. 3, p. e1301, 2019.

PUDJIHARTONO, N. et al. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, Frontiers Media SA, v. 2, 2022. ISSN 2673-7647. Available from Internet: <http://dx.doi.org/10.3389/fbinf.2022.927312>.

PULLIN, J. M.; MCCARTHY, D. J. A comparison of marker gene selection methods for single-cell rna sequencing data. *Genome Biology*, v. 25, n. 1, p. 56, 2024.

RAMPERSAD, S.; TENNANT, P. Replication and expression strategies of viruses. *Viruses*, v. 10, n. 4, p. 194, 2018.

RIZZO, M. et al. Machine learning models explanations as interpretations of evidence: a theoretical framework of explainability and its implications on high-stakes biomedical decision-making. *BMC Medical Research Methodology*, Springer Science and Business Media LLC, v. 25, n. S1, dec. 2025. ISSN 1471-2288. Available from Internet: <http://dx.doi.org/10.1186/s12874-025-02703-1>.

ROBERTS, M. E. P. et al. CD4+ t-cell profiles and peripheral blood ex-vivo responses to t-cell directed stimulation delineate COPD phenotypes. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, v. 2, n. 4, p. 268–280, 2015.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, American Psychological Association (APA), v. 65, n. 6, p. 386–408, 1958. ISSN 0033-295X. Available from Internet: <http://dx.doi.org/10.1037/h0042519>.

ROST, B.; SCHMITT, B. M. et al. Protein function in precision medicine: Deep understanding with machine learning. *FEBS Letters*, v. 592, n. 14, p. 2327–2337, 2018.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, n. 5, p. 206–215, 2019.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, Springer Science and Business Media LLC, v. 323, n. 6088, p. 533–536, oct. 1986. ISSN 1476-4687. Available from Internet: <http://dx.doi.org/10.1038/323533a0>.

RUSTAM, S. et al. A unique cellular organization of human distal airways and its disarray in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, Oxford University Press (OUP), v. 207, n. 9, p. 1171–1182, may 2023. ISSN 1535-4970. Available from Internet: <http://dx.doi.org/10.1164/rccm.202207-1384OC>.

RYO, M.; RILLIG, M. C. Explainable artificial intelligence and interpretable machine learning for the applied sciences. *Patterns*, v. 3, n. 8, p. 100508, 2022.

SAEYS, Y.; INZA, I. n.; nAGA, P. L. A review of feature selection techniques in bioinformatics. *Bioinformatics*, v. 23, n. 19, p. 2507–2517, 2007.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, v. 10, n. 3, p. e0118432, 2015.

SCHULZ, M.-A. et al. Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nature Communications*, Springer Science and Business Media LLC, v. 11, n. 1, aug. 2020. ISSN 2041-1723. Available from Internet: <http://dx.doi.org/10.1038/s41467-020-18037-z>.

SHAPLEY, L. S. A value for n-person games. In: KUHN, H. W.; TUCKER, A. W. (Ed.). *Contributions to the Theory of Games II*. [S.l.]: Princeton University Press, 1953. p. 307–317.

SHUKLA, P. K.; DEB, K. On finding multiple pareto-optimal solutions using classical and evolutionary generating methods. *European Journal of Operational Research*, v. 181, n. 3, p. 1630–1652, 2007.

SINCLAIR, L. V.; CANTRELL, D. A. Protein synthesis and metabolism in t cells. *Annual Review of Immunology*, v. 43, p. 343–366, 2025.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical bayesian optimization of machine learning algorithms. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2012. v. 25.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.

SONESON, C.; ROBINSON, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, v. 15, n. 4, p. 255–261, 2018.

SONESON, C.; ROBINSON, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, v. 18, p. 381–387, 2021.

SQUAIR, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, v. 12, p. 5692, 2021.

STUART, T. et al. Comprehensive integration of single-cell data. *Cell*, v. 177, n. 7, p. 1888–1902.e21, 2019.

SUN, Y.; WONG, A. K. C.; KAMEL, M. S. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 23, n. 4, p. 687–719, 2009.

SUTTON, R.; BARTO, A. *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018. (Adaptive Computation and Machine Learning series). ISBN 9780262039246. Available from Internet: <https://books.google.com.br/books?id=sWV0DwAAQBAJ>.

SVENSSON, V. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, v. 38, n. 2, p. 147–150, 2020.

TOWNES, F. W. et al. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biology*, v. 20, n. 1, p. 295, 2019.

TRAAG, V. A.; WALTMAN, L.; ECK, N. J. van. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, v. 9, p. 5233, 2019.

TROYANSKAYA, O. et al. Missing value estimation methods for dna microarrays. *Bioinformatics*, Oxford University Press (OUP), v. 17, n. 6, p. 520–525, jun. 2001. ISSN 1367-4803. Available from Internet: <http://dx.doi.org/10.1093/bioinformatics/17.6.520>.

VABALAS, A. et al. Machine learning algorithm validation with a limited sample size. *PLOS ONE*, v. 14, n. 11, p. e0224365, 2019.

VARMA, S.; SIMON, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, v. 7, p. 91, 2006.

VIETH, B. et al. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature Communications*, v. 10, p. 4667, 2019.

VIRSHUP, I. et al. anndata: Annotated data. *bioRxiv*, 2021.

VIRSHUP, I. et al. anndata: Access and store annotated data matrices. *Journal of Open Source Software*, v. 9, n. 101, p. 4371, 2024.

VIRTANEN, P. et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, v. 17, p. 261–272, 2020.

VITORINO, R. et al. The power of high-throughput omics integration. *Trends in Biotechnology*, 2024.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, v. 10, n. 1, p. 57–63, 2009.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80–83, 1945.

WILLIAMS, S. et al. Translation regulation in response to stress. *The FEBS Journal*, v. 291, n. 23, p. 5102–5122, 2024.

WITTEN, I. H. et al. Part ii. more advanced machine learning schemes. In: ____. *Data Mining*. Elsevier, 2017. p. 205–208. ISBN 9780128042915. Available from Internet: <http://dx.doi.org/10.1016/B978-0-12-804291-5.00021-0>.

WOLF, F. A.; ANGERER, P.; THEIS, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, v. 19, p. 15, 2018.

XU, Y. et al. Cellbrf: a feature selection method for single-cell clustering using cell balance and random forest. *Bioinformatics*, v. 39, n. Supplement$_1, p.i368 - -i376, 2023$.

YANG, L. et al. A knowledge-informed and pareto-based artificial bee colony optimization algorithm for multi-objective land-use allocation. *ISPRS International Journal of Geo-Information*, MDPI AG, v. 7, n. 2, p. 63, feb. 2018. ISSN 2220-9964. Available from Internet: <http://dx.doi.org/10.3390/ijgi7020063>.

YANG, P.; HUANG, H.; LIU, C. Feature selection revisited in the single-cell era. *Genome Biology*, v. 22, n. 1, p. 321, 2021.

YANG, Z. R. Biological applications of support vector machines. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 5, n. 4, p. 328–338, jan. 2004. ISSN 1477-4054. Available from Internet: <http://dx.doi.org/10.1093/bib/5.4.328>.

YASSINE, A.; MOHAMED, L.; ACHHAB, M. A. Intelligent recommender system based on unsupervised machine learning and demographic attributes. *Simulation Modelling Practice and Theory*, Elsevier BV, v. 107, p. 102198, feb. 2021. ISSN 1569-190X. Available from Internet: <http://dx.doi.org/10.1016/j.simpat.2020.102198>.

YOUNG, L.; LEE, B. K. A reinforcement learning approach to training chess engine neural networks. In: *2023 Congress in Computer Science, Computer Engineering, amp;amp; Applied Computing (CSCE)*. IEEE, 2023. p. 91–95. Available from Internet: <http://dx.doi.org/10.1109/CSCE60160.2023.00020>.

YUAN, X. et al. Feature importance ranking of random forest-based end-to-end learning algorithm. *Remote Sensing*, MDPI AG, v. 15, n. 21, p. 5203, nov. 2023. ISSN 2072-4292. Available from Internet: <http://dx.doi.org/10.3390/rs15215203>.

ZAPPIA, L. et al. Feature selection methods affect the performance of scrna-seq data integration and querying. *Nature Methods*, v. 22, p. 834–844, 2025.

ZHANG, F. et al. Recursive svm biomarker selection for early detection of breast cancer in peripheral blood. *BMC Medical Genomics*, Springer Science and Business Media LLC, v. 6, n. S1, jan. 2013. ISSN 1755-8794. Available from Internet: <http://dx.doi.org/10.1186/1755-8794-6-S1-S4>.

ZHANG, X. et al. Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. *Molecular Cell*, v. 73, n. 1, p. 130–142.e5, 2019.

ZHANG, Z. Missing data imputation: focusing on single imputation. *Annals of Translational Medicine*, v. 4, n. 1, p. 9, 2016.

ZHAO, Y. et al. Rfcell: A gene selection approach for scrna-seq clustering based on permutation and random forest. *Frontiers in Genetics*, v. 12, p. 665843, 2021.

ZHENG, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, v. 8, p. 14049, 2017.

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, v. 1, n. 1, p. 32–49, 2011.

ZHOU, W.; YAN, Z.; ZHANG, L. A comparative study of 11 non-linear regression models highlighting autoencoder, dbn, and svr, enhanced by shap importance analysis in soybean branching prediction. *Scientific Reports*, Springer Science and Business Media LLC, v. 14, n. 1, mar. 2024. ISSN 2045-2322. Available from Internet: <http://dx.doi.org/10.1038/s41598-024-55243-x>.

ZITZLER, E.; BROCKHOFF, D.; THIELE, L. The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In: *Evolutionary Multi-Criterion Optimization*. [S.l.: s.n.], 2007. p. 862–876.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, v. 3, n. 4, p. 257–271, 1999.

# APPENDIX A - RESUMO EXPANDIDO EM PORTUGUÊS

A transcriptômica de célula única permite analisar tecidos complexos em resolução celular, distinguindo programas transcricionais que poderiam ser mascarados em análises de populações heterogêneas. Essa resolução, entretanto, impõe desafios computacionais relevantes, como alta dimensionalidade, esparsidade, ruído técnico, desbalanceamento entre classes e risco de vazamento de informação durante a construção de modelos supervisionados. Esses aspectos são particularmente críticos quando etapas como seleção de atributos, balanceamento ou otimização de hiperparâmetros utilizam informações externas aos dados de treinamento, produzindo estimativas de desempenho artificialmente favoráveis.

Esta dissertação propôs e avaliou um framework computacional para classificação supervisionada em dados transcriptômicos de célula única. O método integra seleção de atributos, aprendizado de máquina, seleção multiobjetivo de modelos e interpretação baseada em SHAP. A principal premissa do trabalho é que a seleção de genes não deve ser tratada apenas como uma etapa preliminar de redução de dimensionalidade, mas como um componente central do processo analítico, diretamente relacionado ao desempenho preditivo, à estabilidade da interpretação e à plausibilidade biológica dos resultados.

O framework foi desenvolvido a partir de três decisões metodológicas. Primeiro, todas as etapas capazes de aprender padrões da distribuição dos dados foram restritas às partições de treinamento em cada dobra da validação cruzada, incluindo o balanceamento das classes, a seleção de atributos e a otimização dos classificadores. Segundo, cada tipo celular foi modelado como uma tarefa independente, evitando que diferenças entre identidades celulares fossem confundidas com o sinal associado à condição biológica. Terceiro, a interpretação dos modelos finais foi incorporada ao fluxo por meio de valores SHAP, permitindo avaliar a contribuição dos genes selecionados para as decisões preditivas.

A avaliação utilizou um conjunto de dados de RNA-seq de célula única em doença pulmonar obstrutiva crônica. Para reduzir efeitos de confusão associados ao sexo, a análise foi restrita a amostras masculinas. Após a filtragem global, a matriz analisada reteve 69.366 células e 31.902 genes. Na etapa específica por tipo celular, 27 populações atenderam aos critérios mínimos para modelagem. Para cada uma delas, o fluxo aplicou validação cruzada estratificada em cinco dobras, balanceamento por *random undersampling* nas partições de treino, seleção de atributos baseada em Wilcoxon dentro de cada

dobra, otimização Bayesiana de hiperparâmetros e comparação entre classificadores.

Foram avaliados cinco algoritmos: regressão logística, máquina de vetores de suporte, floresta aleatória, k-vizinhos mais próximos e perceptron multicamadas. A seleção final dos modelos não se baseou em uma única métrica. Em vez disso, o estudo considerou simultaneamente acurácia, sensibilidade, especificidade e F1-score por meio de fronteira de Pareto e *hypervolume*. Essa estratégia foi adotada porque o desbalanceamento entre classes poderia tornar métricas isoladas pouco informativas, especialmente em populações celulares com menor representação amostral ou sinal transcricional mais sutil.

O estudo também comparou a estratégia proposta com métodos alternativos de seleção de atributos, incluindo ANOVA, importância por floresta aleatória e eliminação recursiva de características. Essa comparação permitiu observar que diferentes métodos priorizam diferentes noções de relevância gênica. Enquanto métodos univariados tendem a destacar genes com contraste estatístico marginal, métodos embutidos e *wrapper* favorecem atributos associados ao comportamento de classificadores específicos. A estratégia proposta buscou equilibrar desempenho preditivo, interpretação do modelo e coerência funcional das assinaturas selecionadas.

Os resultados indicaram que nenhum classificador apresentou superioridade uniforme em todas as populações celulares. Em alguns tipos celulares, modelos lineares obtiveram desempenho competitivo, enquanto, em outros, modelos não lineares apresentaram melhores combinações de métricas. Esse resultado reforça a necessidade de tratar cada população celular como um problema de classificação próprio, em vez de impor um único classificador para todo o conjunto de dados.

O estudo de caso em células T alfa-beta CD4-positivas exemplificou essa variação. Nessa população, o perceptron multicamadas apresentou o melhor equilíbrio entre as métricas avaliadas, constituiu a única solução Pareto-ótima e manteve desempenho consistente na avaliação final por *hold-out*. O modelo final apresentou acurácia de 0,8805, sensibilidade de 0,8977, especificidade de 0,8977 e F1-score de 0,6774. A diferença entre o F1-score e as demais métricas também evidenciou o efeito do desbalanceamento ainda presente na avaliação final, reforçando a importância de analisar múltiplos critérios de desempenho.

A interpretação baseada em SHAP mostrou que o modelo selecionado utilizou uma assinatura compacta de genes com contribuição preditiva consistente. Entre os genes mais relevantes no estudo de caso destacaram-se *RPS4Y1*, *DNAJB1*, *SFTPC*, *HSPA1B* e

*DDX3Y*. A análise das atribuições indicou que a contribuição de um gene podia variar conforme a classe explicada, enquanto a correlação entre expressão gênica e valores SHAP mostrou relações estruturadas entre abundância transcricional e decisão do modelo.

A validação funcional indicou que os genes selecionados se associaram principalmente a processos de tradução, resposta ao estresse, *refolding* mediado por chaperonas, regulação apoptótica e homeostase proteica. Esses resultados sugerem que o framework não apenas selecionou genes úteis para a classificação, mas também recuperou sinais biológicos compatíveis com o contexto da doença. Na análise global, a dificuldade das tarefas variou entre os tipos celulares, refletindo diferenças de tamanho amostral, intensidade do sinal transcricional e grau de desbalanceamento entre as classes.

Em síntese, esta dissertação apresentou um framework para classificação supervisionada em transcriptômica de célula única que combina controle de vazamento, modelagem específica por tipo celular, seleção de atributos, otimização de classificadores, seleção multiobjetivo e interpretação baseada em SHAP. Os resultados indicam que essa integração favorece a construção de modelos preditivos mais interpretáveis e biologicamente informativos, especialmente em cenários nos quais a heterogeneidade celular e o desbalanceamento entre classes limitam abordagens baseadas em métricas únicas ou em análises globais.

**Palavras-chave: sequenciamento de RNA de célula única; seleção de atributos; aprendizado de máquina; inteligência artificial interpretável; otimização multiobjetivo; doença pulmonar obstrutiva crônica.**