SCOPES: Stability-Aware Cross-Platform Feature Selection for Matched TCGA Gene Expression and RNA-Seq Data

Abdullah Nayem Wasi Emran, Tanveer Rahman, Md. Shamsuzzoha Bayzid
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
abdullahnayem9274@gmail.com

September 2025

Abstract

Cross–platform reuse of legacy microarrays with modern RNA–Seq is attractive but challenging: the same gene can follow different measurement distributions, and feature selection can leak label information and inflate performance. We develop **SCOPES**, a leak–free, multi–objective feature–selection framework that balances three goals: predictive accuracy (AUC), selection stability (Kuncheva), and cross–platform alignment (Maximum Mean Discrepancy, MMD). On matched TCGA–BRCA Agilent/RNA–Seq data, an initial label–informed F–score slab produced an apparently perfect microarray model (AUC$\approx$1.0) but lost $\sim$ 0.30 AUC after transfer to RNA–Seq, revealing selection leakage and platform shift. Replacing the slab with an *unsupervised* MAD filter and enforcing patient–safe cross–validation exposed a clear trade–off on the Pareto front: an alignment–first solution with a single gene achieved modest source performance and slight transfer loss ($\text{AUC}_{\text{Agilent}} \approx 0.69$, $\text{AUC}_{\text{RNA-Seq}} \approx 0.61$, $\Delta\text{AUC} \approx -0.08$), whereas a richer 30–gene signature reached near–perfect source AUC but transferred poorly (ΔAUC≈ -0.38) with higher MMD. These results show that more genes often buy source accuracy at the expense of portability. SCOPES makes this trade–off explicit and suggests selecting near a Pareto "knee" under explicit size and alignment constraints. Reporting both ΔAUC and an alignment metric provides a simple, reproducible framework for building cross–platform gene signatures.

1 Introduction

Microarray technology fueled early cancer transcriptomic studies, producing thousands of publicly available expression profiles [1]. Today, RNA–Seq dominates owing to a greater dynamic range and lower noise [2]. Integrating historical microarrays with contemporary RNA–Seq could dramatically increase sample sizes and improve external validation, but differences between platforms make it risky to merge them without careful adjustment. Recent work shows that cross–platform normalization (e.g., quantile normalization, Training Distribution Matching) enables model transfer [3]. However, *feature–selection* pipelines remain largely platform–specific, threatening reproducibility and clinical translation. In an attempt to address this gap, we introduce SCOPES.

Cancer transcriptomics research critically depends on identifying reliable, reproducible biomarkers across studies and platforms. Larger training cohorts improve the statistical power of differential expression analysis and enable machine learning models to capture better the heterogeneity of cancer biology, which is crucial for applications such as prognosis prediction and therapeutic response modeling. While normalization strategies can reduce technical variation between microarray and RNA–Seq, they do not address the instability of feature selection, where small changes in data distribution often yield inconsistent gene sets. This instability reduces model interpretability and downstream clinical adoption.

Several studies have proposed domain adaptation and harmonization techniques to facilitate cross-platform integration, but most focus on adjusting distributions rather than ensuring robust gene selection. Furthermore, benchmark evaluations often emphasize classification accuracy while overlooking feature reproducibility, essential for identifying clinically actionable biomarkers. Thus, there remains a pressing need for a framework that jointly considers normalization, feature selection stability, and predictive performance.

To address these challenges, we propose SCOPES (*Stable Cross–Platform Expression Selection*), a framework designed to enable consistent and reproducible gene selection across microarray and RNA–Seq datasets. SCOPES leverages cross–platform harmonization techniques while incorporating stability–enhanced feature selection, yielding feature sets that are not only predictive but also robust across platforms.

The remainder of this paper is organized as follows. Section 2 reviews existing literature. Section 3 describes the SCOPES framework in detail. Section 4 presents experimental results. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Related Works

This section reviews feature selection methods and cross-platform techniques that improve the generalizability of predictive models in biomedical research.

Feature selection (FS), cross-platform normalization, and domain adaptation are widely used to enhance cancer genomics modeling. Several single-platform FS approaches have shown strong results. Liu et al.[4] introduced VEW, a three-stage method using variance filtering, Extremely Randomized Trees, and Whale Optimization to identify cancer-related gene subsets. Xu et al.[5] proposed FG-HFS, combining spectral clustering with group evolution, reaching 92–93% accuracy across multiple cancer datasets. Qu et al.[6] developed VNL-HHO, which integrates F-score filtering, Variable Neighborhood Learning, and Harris Hawks Optimization with mutation, achieving over 96% accuracy and up to 100% in some cases.

For cross-platform analysis, Feature-Specific Quantile Normalization (FSQN) has effectively aligned microarray and RNA-Seq data. Franks et al.[7], Foltz et al.[3], and Skubleny et al.[8] showed that FSQN outperforms global normalization and Training Distribution Matching, with Skubleny et al. further demonstrating that FSQN with iterative FS yields cross-platform performance close to within-platform models.

Domain adaptation methods also improve model transferability. Yuan et al.[9] proposed LogitDA and KNNDA, which identify domain-invariant features for drug response prediction with AUCs ranging from 0.70–1.00. Mourragui et al.[10] developed PRECISE to learn invariant predictors between preclinical and tumor data, though it may suffer from negative transfer under certain assumptions.

Recent works expand these strategies further. Krishna et al.[11] combined deep learning with sparsity-based FS to produce compact, predictive gene signatures. Kim and Jang[12] used scRNA-seq networks refined with XGBoost for pan-cancer tasks, outperforming bulk RNA-seq gene sets. Chowdhury et al.[13] applied multi-view FS with ensemble classifiers to classify 33 cancer types with 97.1% accuracy and near-perfect AUC. Thelagathoti et al.[14] used ensemble FS with nested validation to identify miRNA biomarkers for Usher Syndrome, achieving 97.7% accuracy. Tom et al.[15] applied a hybrid sequential FS pipeline for mRNA biomarkers in Usher Syndrome, reducing over 42,000 features to a small validated set.

Together, these studies highlight that combining FS, normalization, and domain adaptation enables high accuracy and robust generalization across diverse biological contexts, supporting more reliable integrative genomic models. Table 1 summarizes the related works discussed in this section.

Table 1: Summary of feature selection, normalization, and domain adaptation methods in genomics.

Method / Study	**Key Approach**	**Application**	**Performance**
VEW [4]	Variance filter + Extra Trees + Whale Opt.	Cancer gene subset selection	Higher accuracy
FG-HFS [5]	Spectral clustering + group evolution	Gene expression FS	∼92–93% accuracy
VNL-HHO [6]	F-score + VNL + Harris Hawks Opt.	Cancer classification	>96%, up to 100%
FSQN [7, 3, 8]	Feature-specific quantile normalization	Microarray ↔ RNA-Seq	Cross-platform & within-platform
LogitDA / KNNDA [9]	Domain-invariant feature learning	Drug response prediction	AUC 0.70–1.00
PRECISE [10]	Invariant predictor transfer	Preclinical ↔ tumor	May risk negative transfer
Deep + Sparse FS [11]	Deep learning + sparsity FS	Survival prediction	Compact signatures
scRNA + XGBoost [12]	scRNA-seq networks + XGBoost	Pan-cancer gene sets	Outperforms bulk RNA-seq
Multi-view FS + Ensemble [13]	Partition + Boruta + ensemble	Pan-cancer (33 types)	97.1% acc., AUC 0.9996
Usher miRNA FS [14]	Ensemble FS + nested validation	Usher Syndrome miRNA	∼97.7% accuracy
Hybrid Seq. FS [15]	Variance thr. + RFE + Lasso	Usher Syndrome mRNA	Validated compact set

3 Methods

3.1 Cohort construction and data model

We assembled a matched breast cancer (TCGA-BRCA) cohort measured on two platforms: Agilent 244K microarray and Illumina HiSeq RNA-Seq V2. Let $\mathcal{P}$ denote patients and $\mathcal{G}$ denote genes.

Sample types. We restricted to primary tumor (code 01) and solid tissue normal (code 11) aliquots to pose a clean binary classification task (tumor vs. normal) while avoiding heterogeneity from recurrent/metastatic samples.

Patient de-duplication. Some patients have multiple aliquots of a given type. To avoid correlated replicates leaking across folds during cross-validation, we kept at most one tumor and one normal per patient, choosing the aliquot with the lowest (portion, vial) code ("first available").

Matched patients and genes. Let $\mathcal{P}_{\text{arr}}$ and $\mathcal{P}_{\text{rna}}$ be the patient sets available on microarray and RNA-Seq, respectively, and $\mathcal{G}_{\text{arr}}, \mathcal{G}_{\text{rna}}$ the corresponding gene sets. We retain only intersections

$$\mathcal{P} = \mathcal{P}_{\text{arr}} \cap \mathcal{P}_{\text{rna}}, \quad \mathcal{G} = \mathcal{G}_{\text{arr}} \cap \mathcal{G}_{\text{rna}},$$

so that *every* retained patient has both measurements and *every* retained gene is measured on both platforms.

Aligned matrices. After reindexing rows/columns to a common order, we obtain two matrices with identical shape

$$\mathbf{X}^{\text{arr}},\, \mathbf{X}^{\text{rna}} \in \mathbb{R}^{n\times p}, \quad n = |\mathcal{P}| = 530,\ p = |\mathcal{G}| = 16{,}146,$$

and a label vector $\mathbf{y} \in \{0,1\}^n$ with 1 for tumor and 0 for normal ($n_1 = 505$, $n_0 = 25$). Row i in both matrices corresponds to the *same* patient; column j corresponds to the *same* gene.

Groups for patient integrity. We derive a group identifier g_i from the first 12 characters of the TCGA barcode (patient ID). These groups are used to prevent any patient's samples from being split across training and validation folds.

Class imbalance. The extreme imbalance (505 tumors vs. 25 normals) biases accuracy. Throughout, we therefore optimize/evaluate with the area under the ROC curve (AUC), which is threshold-free and more informative under imbalance.

3.2 Light pre-filtering (consistency and leakage control)

Missingness filter. For gene j, let $\text{miss}_j^{(\cdot)}$ be the fraction of missing entries in $\mathbf{X}^{(\cdot)} \in \{\mathbf{X}^{\text{arr}}, \mathbf{X}^{\text{rna}}\}$. We drop any gene with $\max\big(\text{miss}_j^{\text{arr}}, \text{miss}_j^{\text{rna}}\big) \geq 0.5$.

Median imputation (platform-consistent). Let $\tilde{m}_j = \text{median}\big(\{X_{ij}^{\text{arr}} : i \in [1..n],\ X_{ij}^{\text{arr}} \text{ observed}\}\big)$. For any missing X_{ij}^{arr} or X_{ij}^{rna} we impute $X_{ij}^{(\cdot)} \leftarrow \tilde{m}_j$. Using the same imputation constant per gene across platforms prevents introducing platform-specific shifts.

Supervised slab for search (ANOVA F-score). To reduce the search space while preserving discriminative signal, we rank genes on the microarray matrix with the two-sample ANOVA F-statistic:

$$F_j \;=\; \frac{\displaystyle\sum_{c\in\{0,1\}} n_c\big(\bar{x}_{j,c}^{\text{arr}} - \bar{x}_{j,\cdot}^{\text{arr}}\big)^2/(C-1)}{\displaystyle\sum_{c\in\{0,1\}}\sum_{i:y_i=c} \big(x_{ij}^{\text{arr}} - \bar{x}_{j,c}^{\text{arr}}\big)^2/(n-C)}, \quad C = 2, \tag{1}$$

where $\bar{x}_{j,c}^{\text{arr}}$ is the class-c mean and $\bar{x}_{j,\cdot}^{\text{arr}}$ the global mean. We retain the top $p_0 = 1000$ genes to form a "gene slab" for downstream multi-objective selection. *Note:* computing (1) on *all* samples is supervised and can leak label information into cross-validation. We report and analyze this risk, and we later replace it with a leak-free, unsupervised MAD filter in follow-up experiments; here we document the initial pipeline used for the main optimization.

After filtering/imputation we have $\mathbf{X}_{\text{slab}}^{\text{arr}}, \mathbf{X}_{\text{slab}}^{\text{rna}} \in \mathbb{R}^{n\times p_0}$ with $p_0 = 1000$.

3.3 Leak-free variability prefilter (MAD)

To eliminate label leakage from the F-score slab while still shrinking the search space, we use an *unsupervised* median absolute deviation (MAD) filter.

Per-gene variability. For gene j on platform $q \in \{\text{arr}, \text{rna}\}$, define

$$\tilde{\mu}_j^{(q)} = \text{median}\{X_{ij}^{(q)} : i = 1, \ldots, n\}, \tag{2}$$

$$\text{MAD}_j^{(q)} = \text{median}\{\, |X_{ij}^{(q)} - \tilde{\mu}_j^{(q)}| \;:\; i = 1, \ldots, n\}. \tag{3}$$

MAD depends only on unlabeled values and is robust to outliers.

Platform-robust score. We aggregate the two platforms to favor genes that vary on both:

$$\mathrm{rMAD}_j \;=\; \frac{1}{2}\left(\frac{\mathrm{MAD}_j^{(\mathrm{arr})}}{\mathrm{median}_g\,\mathrm{MAD}_g^{(\mathrm{arr})}} \;+\; \frac{\mathrm{MAD}_j^{(\mathrm{rna})}}{\mathrm{median}_g\,\mathrm{MAD}_g^{(\mathrm{rna})}}\right). \tag{4}$$

(Using per-platform median scaling prevents one platform from dominating due to units.)

Leak-free slab. We rank genes by rMAD_j and retain the top $p_0 = 1000$ genes:

$$\mathcal{G}_{\mathrm{slab}} \;=\; \text{Top-}p_0 \text{ by } \mathrm{rMAD}_j\,.$$

This step uses no labels and is thus leak-free. Missingness filtering and platform-consistent median imputation are applied *before* computing MAD; imputation constants are Agilent medians as described in Sec. 3.2.

3.4 Evaluation protocol

We estimate within-source performance with stratified, patient-safe cross-validation.

StratifiedGroupKFold. We split indices $\{1,\ldots,n\}$ into K folds $\{\mathcal{I}_k\}_{k=1}^K$ such that: (i) class proportions are approximately preserved in each fold (stratification on y), and (ii) no group appears in multiple folds ($g_i = g_{i'} \Rightarrow i, i' \in \mathcal{I}_k$ for the same k). For each candidate feature set we train a logistic regression on $\bigcup_{k'\neq k}\mathcal{I}_{k'}$ and compute AUC on $\mathcal{I}_k$; we report the CV mean:

$$\mathrm{AUC}_{\mathrm{cv}} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{AUC}\left(\hat{f}_{-k}(\cdot),\, \mathbf{y}_{\mathcal{I}_k}\right).$$

3.5 Multi-objective feature selection (SCOPES)

Search space and subset size. On the p_0=1000-gene slab we search binary masks $\mathbf{s}\in\{0,1\}^{p_0}$ with $\|\mathbf{s}\|_0 = k$ selected genes. We target a practical panel size with $k \leq 120$; the algorithm may pick fewer genes if optimal.

Objectives. Given $\mathbf{s}$ (and the restricted matrices $\mathbf{X}_{\mathbf{s}}^{(\cdot)}$):

$$f_1(\mathbf{s}) := 1 - \mathrm{AUC}_{\mathrm{cv}}(\mathbf{s}) \qquad \text{(minimize: maximize AUC)}, \tag{5}$$

$$f_2(\mathbf{s}) := 1 - \mathrm{Kun}(\mathbf{s}) \qquad \text{(minimize: maximize selection stability)}, \tag{6}$$

$$f_3(\mathbf{s}) := \mathrm{MMD}_\gamma(\mathbf{X}_{\mathbf{s}}^{\mathrm{arr}}, \mathbf{X}_{\mathbf{s}}^{\mathrm{rna}}) \quad \text{(minimize: cross-platform alignment)}. \tag{7}$$

Stability (Kuncheva index). Let $S^{(a)}, S^{(b)} \subseteq \{1,\ldots,p_0\}$ be feature sets of size k obtained from two resamples (e.g., CV folds). The pairwise Kuncheva stability is

$$\mathrm{Kun}\big(S^{(a)}, S^{(b)}\big) = \frac{\left|S^{(a)} \cap S^{(b)}\right| - \frac{k^2}{p_0}}{k - \frac{k^2}{p_0}}, \qquad \in [-1, 1], \tag{8}$$

and we aggregate by averaging over all resample pairs:

$$\mathrm{Kun}_{\mathrm{avg}} = \frac{2}{K(K-1)}\sum_{1\leq a<b\leq K}\mathrm{Kun}\big(S^{(a)}, S^{(b)}\big).$$

Distribution alignment (MMD). Before computing MMD we z-score each selected gene within each platform $\tilde{X}_{ij}^{(q)} = \big(X_{ij}^{(q)} - \mu_j^{(q)}\big)/\sigma_j^{(q)}$ to remove scale units. For $(\mathbf{X}, \mathbf{Z}) = \big(\tilde{\mathbf{X}}_{\mathbf{s}}^{\mathrm{arr}}, \tilde{\mathbf{X}}_{\mathbf{s}}^{\mathrm{rna}}\big)$, with n_X and n_Z samples respectively, the squared maximum mean discrepancy with RBF kernel $k_\gamma(\mathbf{u}, \mathbf{v}) = \exp(-\|\mathbf{u} - \mathbf{v}\|_2^2/\gamma)$ is

$$\mathrm{MMD}_\gamma^2(\mathbf{X}, \mathbf{Z}) = \frac{1}{n_X^2} \sum_{i,i'} k_\gamma(\mathbf{x}_i, \mathbf{x}_{i'}) + \frac{1}{n_Z^2} \sum_{j,j'} k_\gamma(\mathbf{z}_j, \mathbf{z}_{j'}) - \frac{2}{n_X n_Z} \sum_{i,j} k_\gamma(\mathbf{x}_i, \mathbf{z}_j). \tag{9}$$

Evolutionary optimization (NSGA-II). We employ NSGA-II to minimize $\mathbf{f}(\mathbf{s}) = \big(f_1, f_2, f_3\big)$ subject to $\|\mathbf{s}\|_0 \leq 120$. The algorithm maintains a population of feasible masks (initialized with varying sizes, e.g., $5 \leq k \leq 120$), applies uniform crossover and bit-flip mutation, and selects the next generation by non-dominated sorting (Pareto fronts) with crowding-distance diversity.

Algorithm 1 SCOPES multi-objective feature selection (NSGA-II)

```
Require: $\mathbf{X}_{\mathrm{slab}}^{\mathrm{arr}}, \mathbf{X}_{\mathrm{slab}}^{\mathrm{rna}}, \mathbf{y}, \mathbf{g}$; pop. size $M$, generations $T$
Initialize population $\mathcal{P}_0 = \{\mathbf{s}^{(m)}\}_{m=1}^{M}$ with $5 \leq \|\mathbf{s}^{(m)}\|_0 \leq 120$
for $t = 0$ to $T - 1$ do
    for all $\mathbf{s} \in \mathcal{P}_t$ do
        Compute $f_1(\mathbf{s})$ via StratifiedGroupKFold AUC (Sec. 3.4)
        Compute $f_2(\mathbf{s})$ via averaged Kuncheva index (8)
        Compute $f_3(\mathbf{s})$ via MMD (9) with RBF kernel
    Rank $\mathcal{P}_t$ into Pareto fronts; compute crowding distances
    Create offspring by tournament selection, crossover, and mutation
    $\mathcal{Q}_t \leftarrow \mathcal{P}_t \cup$ offspring
    Form $\mathcal{P}_{t+1}$ by non-dominated sorting + crowding until size $M$
return final Pareto set $\mathcal{P}_T$
```

Final model selection. From the final Pareto set we select the subset by a lexicographic rule prioritizing cross-platform alignment:

$$\mathbf{s}^\star \in \arg\min_{\mathbf{s} \in \mathcal{P}_T} \ f_3(\mathbf{s}) \quad \text{with tie-break} \quad \arg\max \mathrm{AUC}_{\mathrm{cv}}(\mathbf{s}). \tag{10}$$

This "lowest-MMD wins" policy yields the most platform-consistent signature; we also later explore alternative picks (e.g., size/stability-constrained selections) to illustrate accuracy–alignment trade-offs.

3.6 Implementation details

We used logistic regression (liblinear/saga) for AUC evaluation, K=5 stratified group folds, population size M (e.g., $M \in [50, 100]$), and T generations (e.g., $T \in [80, 150]$). Kernel bandwidth γ in (9) followed the median-pairwise-distance heuristic. Hyperparameters were fixed across runs.

3.7 Leakage considerations

The F-score slab in Sec. 3.2 uses labels and, if computed on *all* samples, can inflate cross-validated AUC (selection bias). We therefore report results with this initial pipeline and, in additional experiments, replace the slab with an *unsupervised* median absolute deviation (MAD) filter (Sec. 3.3) to eliminate leakage; the rest of the SCOPES optimization is unchanged.

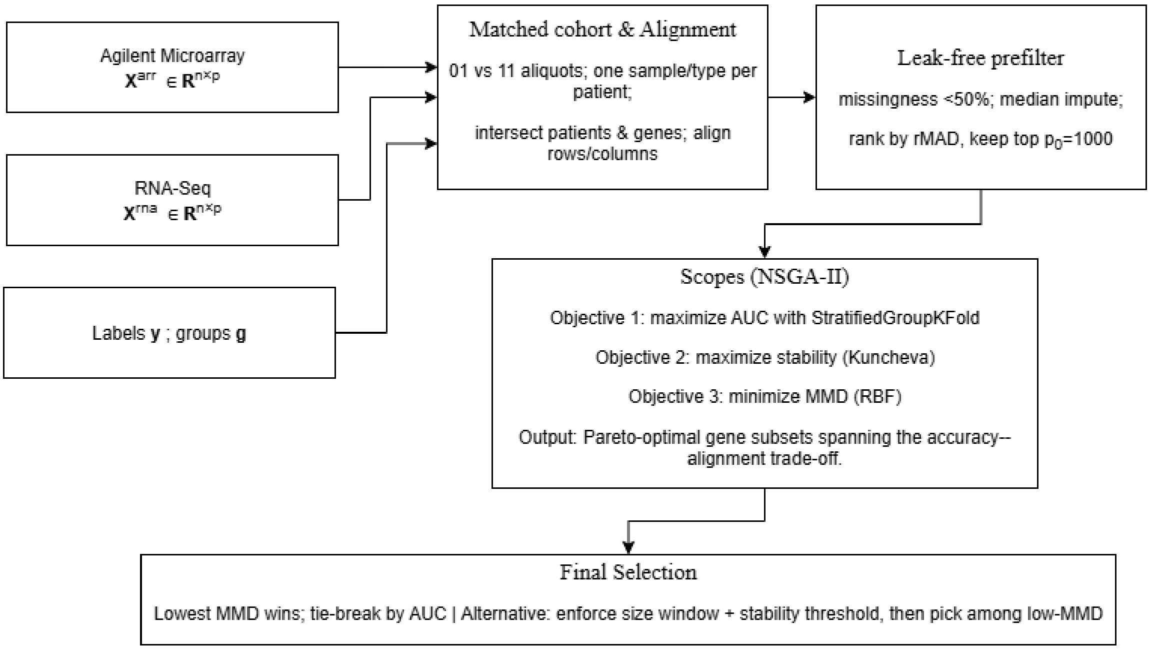

Agilent Microarray
X^arr ∈ R^nxp
RNA-Seq
X^rna ∈ R^nxp
Labels y ; groups g
Matched cohort & Alignment
01 vs 11 aliquots; one sample/type per patient;
intersect patients & genes; align rows/columns
Leak-free prefilter
missingness <50%; median impute;
rank by rMAD, keep top p0=1000
Scopes (NSGA-II)
Objective 1: maximize AUC with StratifiedGroupKFold
Objective 2: maximize stability (Kuncheva)
Objective 3: minimize MMD (RBF)
Output: Pareto-optimal gene subsets spanning the accuracy--alignment trade-off.
Final Selection
Lowest MMD wins; tie-break by AUC | Alternative: enforce size window + stability threshold, then pick among low-MMD

Figure 1: **SCOPES pipeline.** Matched cohort and alignment, leak-free rMAD prefilter, and NSGA-II multi-objective optimization with objectives: maximize AUC (patient-safe CV), maximize stability (Kuncheva), and minimize platform MMD (RBF). Final selection prioritizes lowest MMD with AUC tie-break; an alternative rule applies a size window and stability threshold before picking among low-MMD candidates.

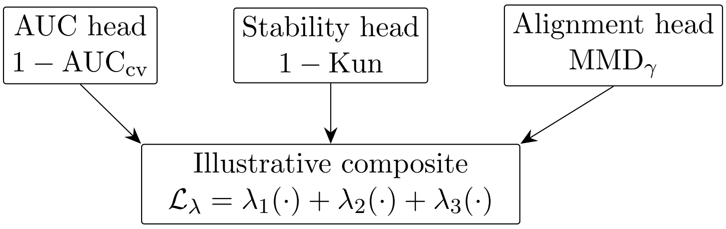

AUC head
$1 - \mathrm{AUC}_{\mathrm{cv}}$
Stability head
$1 - \mathrm{Kun}$
Alignment head
MMD_γ
Illustrative composite
$\mathcal{L}_\lambda = \lambda_1(\cdot) + \lambda_2(\cdot) + \lambda_3(\cdot)$

Figure 2: **Objective-head sketch** AUC, stability, and MMD heads shown in a loss-like diagram for intuition; the actual solver uses Pareto fronts (NSGA-II).

4 Results

We evaluate models by within-source AUC on Agilent ("source") and cross-platform AUC on RNA-Seq ("target"). To quantify transfer we report

$$\Delta\mathrm{AUC} \;=\; \mathrm{AUC}_{\mathrm{target}} - \mathrm{AUC}_{\mathrm{source}}, \tag{11}$$

so more negative values imply a larger platform shift. Selection stability is the mean pairwise Kuncheva index; alignment is the RBF-based MMD_γ between platforms on the chosen genes (smaller is better).

Label-informed baseline is over-optimistic. Using the initial, *label-informed* F-score slab (Sec. 3.2), NSGA-II returned a k=105-gene subset with $\mathrm{AUC}_{\mathrm{source}} \approx 1.00$, stability ≈ 1.00 and

$\mathrm{MMD}_\gamma \approx 0$. This is implausibly strong because the F-score was computed once on the full cohort, so each CV fold trained on features already tuned using its test fold (selection leakage). When transferred to RNA-Seq for the *same patients*, performance fell to $\mathrm{AUC}_{\mathrm{target}} \approx 0.701$ ($\Delta\mathrm{AUC} \approx -0.30$), consistent with known microarray–RNA-seq differences in dynamic range and quantitation.

Leak-free optimization exposes an accuracy–alignment trade-off. To remove leakage we replaced the slab with an *unsupervised* MAD filter and re-ran NSGA-II with the same three objectives. We then applied two deterministic pick rules on the final Pareto set: (i) **Run A** selects the *lowest-MMD* solution (ties → higher AUC); (ii) **Run B** enforces $60 \le k \le 120$ and stability ≥ 0.6, then picks among the low-MMD front by highest AUC. The Pareto landscape and the first pick is shown in Fig. 3. AUCs on source/target for the baseline and both picks are summarized in Fig. 4, and the transfer–alignment relationship is visualized in Fig. 5.

Run A yielded a *single* gene (k=1) with $\mathrm{AUC}_{\mathrm{source}} \approx 0.69$, $\mathrm{AUC}_{\mathrm{target}} \approx 0.61$ and $\Delta\mathrm{AUC} \approx -0.08$ at very low MMD_γ and near-perfect stability, i.e., minimal transfer loss at modest absolute AUC. By contrast, Run B selected k=30 genes with $\mathrm{AUC}_{\mathrm{source}} \in [0.996, 1.000]$ but $\mathrm{AUC}_{\mathrm{target}} \approx 0.615$ ($\Delta\mathrm{AUC} \approx -0.38$) and substantially larger MMD_γ. Together, Figs. 4 and 5 make the trade-off explicit: alignment-first (Run A) vs. accuracy-first (Run B). Gene-level platform agreement for the 30-gene Run B subset is shown in Figs. 6 and 7; several genes lie near the diagonal while others deviate substantially, consistent with the high MMD and large negative ΔAUC.

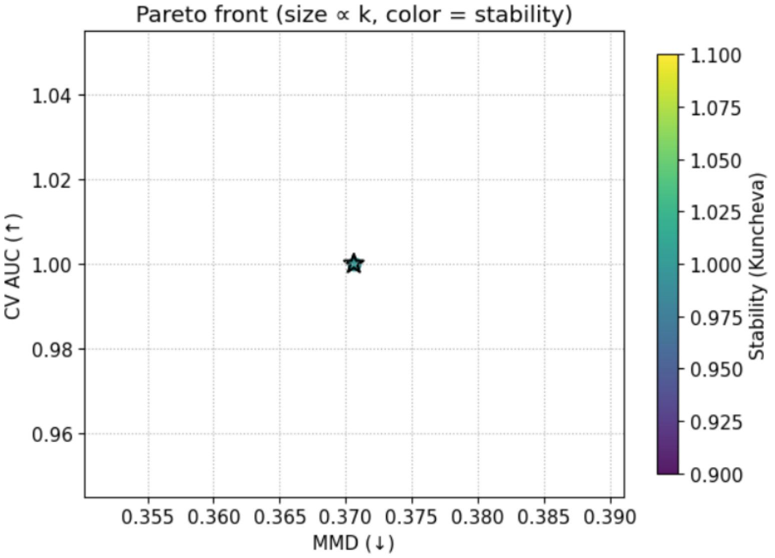


Figure 3: **Pareto front of SCOPES solutions (MAD slab).** Each point is a candidate gene set from NSGA-II; x-axis: alignment (*MMD*, lower is better); y-axis: *CV AUC* on Agilent (higher is better). Marker size encodes the number of genes k; color encodes stability (Kuncheva; lighter = higher). Starred points denote the deterministic picks used downstream: *Run A* (lowest MMD) and *Run B* (size/stability-constrained repick).

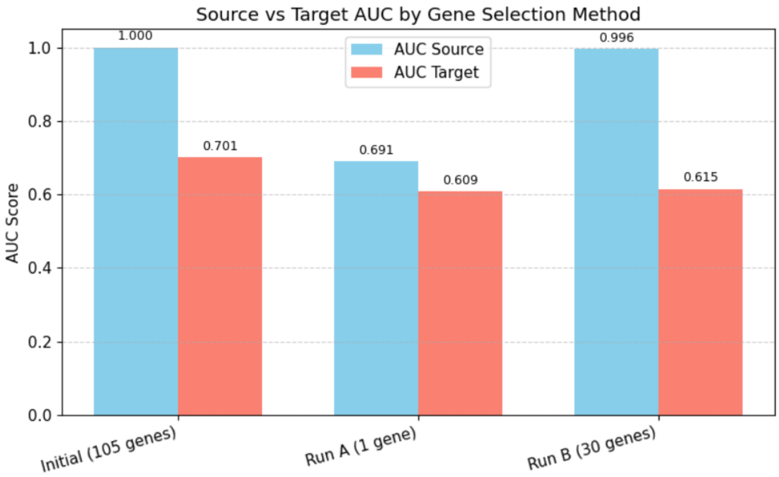


Figure 4: **Within-source vs. cross-platform AUC.** Bars show Agilent (source, blue) and RNA-Seq (target, orange) AUC for the *Initial* 105-gene baseline (label-informed slab), *Run A* (k=1), and *Run B* (k=30). Numeric labels give exact AUCs. The baseline exhibits a $\sim$30-point transfer drop; *Run A* keeps transfer loss small (ΔAUC ≈ -0.08) at moderate AUC, while *Run B* achieves near-perfect source AUC but transfers poorly.

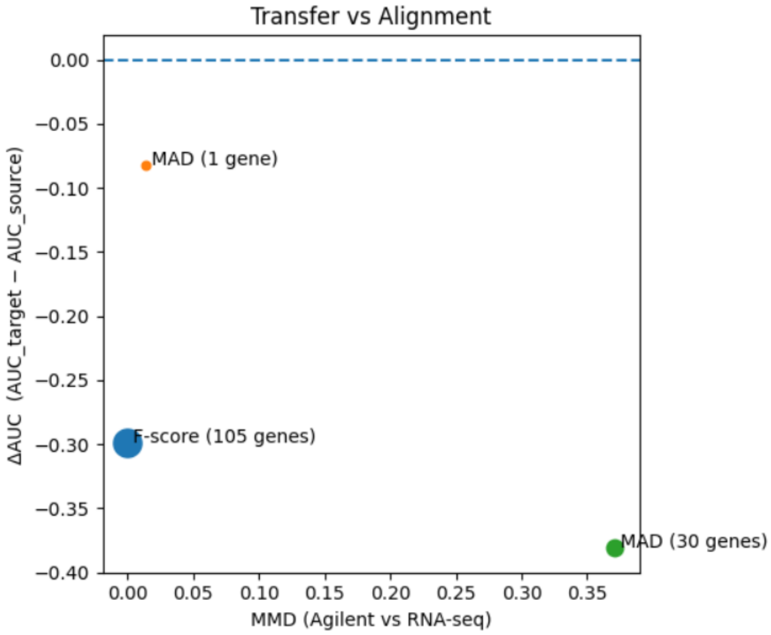


Figure 5: **Transfer vs. alignment.** Bubbles are gene subsets positioned by ($\text{MMD}_\gamma, \Delta\text{AUC}$); bubble area $\propto$ subset size k. Points for the two final picks highlight the trade-off: very small MMD with minimal transfer loss at k=1 (*Run A*) versus larger k with much higher MMD and large negative ΔAUC (*Run B*).

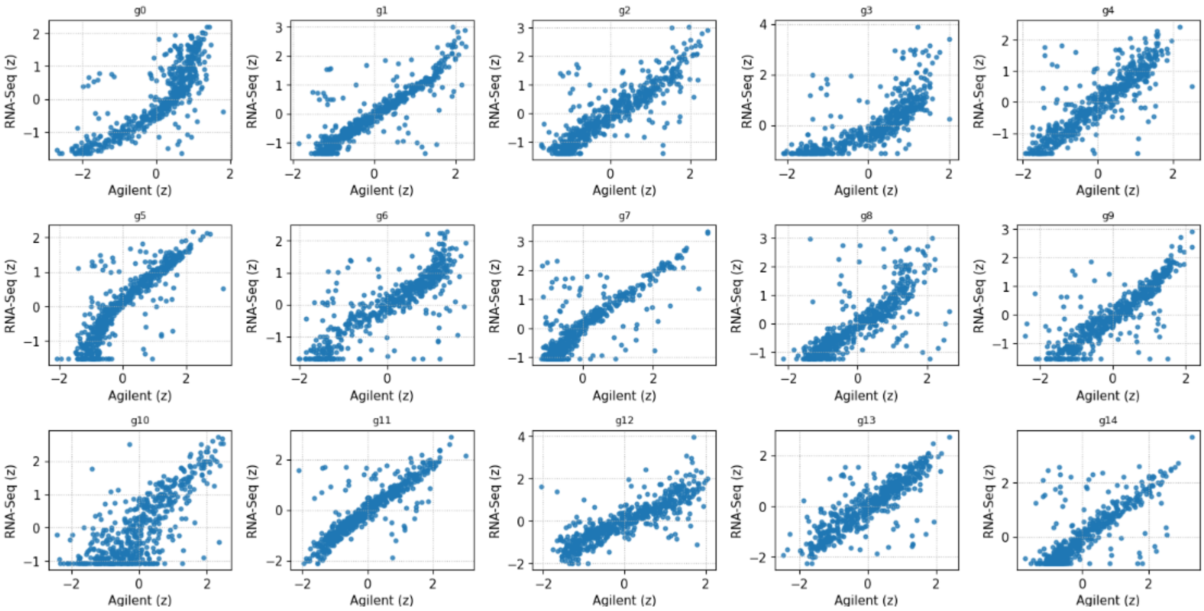


Figure 6: **Per-gene platform agreement for the Run B signature (g0–g14).** Each panel is a scatter of standardized expression for the same patients measured on **Agilent** (x-axis) and **RNA-Seq** (y-axis). Panels close to the diagonal indicate good cross-platform agreement; diffuse or curved clouds indicate mismatch. These patterns contribute to the larger alignment discrepancy (higher MMD) for Run B and help explain its reduced cross-platform AUC (cf. Figs. 4–5).

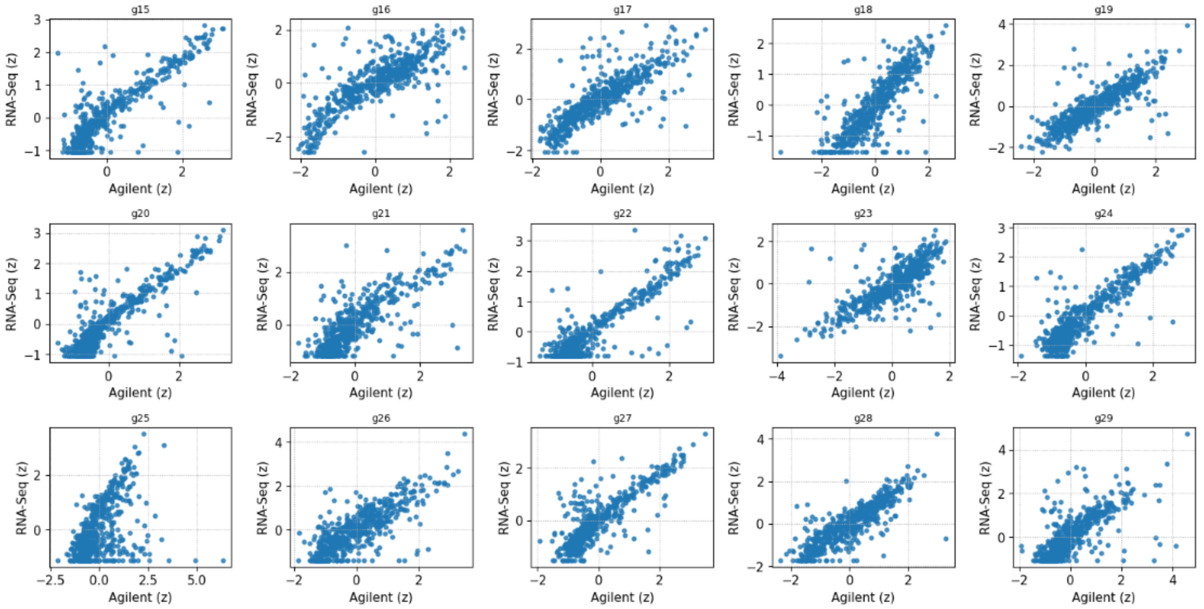
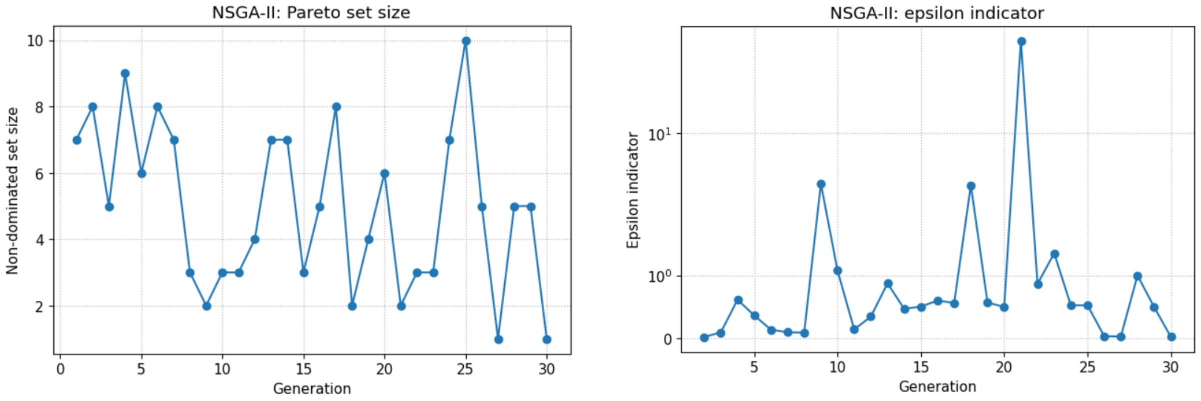


Figure 7: **Per-gene platform agreement for the Run B signature (g15–g29).** As in Fig. 6, several genes align well while others show pronounced mismatch. The mixture of well- and poorly-aligned genes collectively inflates MMD and drives the large negative ΔAUC observed for Run B.

Optimization diagnostics. For completeness, we report two standard traces of the evolutionary search. Figure 8a shows the number of non-dominated solutions per generation (diversity of the Pareto set), and Fig. 8b shows the additive ε-indicator on a symlog scale (coarse progress signal). Fluctuations are expected as variation operators explore new trade-offs; these diagnostics do not directly reflect biological performance but indicate healthy search dynamics.



(a) **Non-dominated set size per generation.** Larger fronts indicate richer diversity of trade-offs explored; transient dips/spikes arise from selection pressure and exploration.

(b) **ε-indicator per generation (symlog).** Lower values generally indicate improvement/convergence; isolated spikes correspond to exploratory moves opening new regions of the objective space.

Figure 8: **NSGA-II diagnostics across generations.** Left: Pareto front size; Right: ε-indicator. Together, they characterize convergence and diversity dynamics.

5 Discussion

Microarrays enabled the first large cancer transcriptome studies. RNA-Seq is now standard because it measures a wider dynamic range with less noise. If we could re-use old microarrays together with new RNA-Seq, we would gain power and stronger external validation. The problem is platform shift: the same gene can look different on the two technologies. Normalization helps, but feature selection itself can still have platform bias.

What we changed. Our first pipeline used an F-score slab built once on all samples. That leaks labels into cross-validation and makes AUC and stability look perfect. We then switched to a leak-free setup: an *unsupervised* MAD slab, patient-safe CV, and the same three SCOPES objectives (AUC, stability, MMD). From the Pareto set we applied two simple pick rules: *Run A* (lowest MMD) and *Run B* (size/stability constraints, then highest AUC). The optimization landscape and the final picks appear in Fig. 3. Source and target AUCs are in Fig. 4. The transfer–alignment trade-off is shown in Fig. 5.

What we found. The label-informed baseline scored AUC ≈ 1.0 on Agilent, yet lost about 0.30 AUC when applied to RNA-Seq, which is a classic sign of leakage plus platform shift. With the MAD slab, the trade-off became clear. *Run A* picked a single, platform-consistent gene: small transfer loss ($\Delta\text{AUC} \approx -0.08$) but only moderate AUC overall. *Run B* picked 30 genes: near-perfect AUC on Agilent, but poor transfer ($\Delta\text{AUC} \approx -0.38$) and high MMD. Per-gene scatter plots for Run B (Figs. 6–7) show why: some genes align well across platforms, others do not.

Practical meaning. If the goal is a portable signature, chasing maximum source AUC alone is risky. A small, stable set can travel better across platforms, even if its absolute AUC is lower. In practice, the best choice is likely a point near the Pareto "knee": a few genes, acceptable AUC, and low MMD.

Limitations. We have very few normal samples—only 25 compared to 505 tumor samples. Gene matching by symbol ignores probe design and transcript isoforms. MMD depends on kernel scale (we used a standard heuristic). We did not test on an independent cohort. These factors may affect absolute numbers, but the trade-off pattern is robust.

Future Work. Three straightforward improvements are worth testing: (i) add stronger alignment constraints (e.g., CORAL/whitening or moment matching) alongside MMD; (ii) calibrate platforms before selection (rank/quantile mapping or ComBat); (iii) enforce sparsity and a small size range to steer toward the Pareto knee. Independent-cohort tests, and the reverse transfer (RNA-Seq → microarray), would offer a fuller understanding of cross-platform generalization.

Conclusion. A leak-free, multi-objective view makes the trade-off visible: more genes usually buy source AUC at the expense of transfer. Reporting both ΔAUC and an alignment measure (like MMD), and selecting near the knee, are effective practices for enhancing cross-platform model robustness.

References

[1] G. Russo, C. Zegar, and A. Giordano, "Advantages and limitations of microarray technology in human cancer," *Oncogene*, vol. 22, no. 42, pp. 6497–6507, 2003.

[2] D. Spies and C. Ciaudo, "Dynamics in transcriptomics: advancements in rna-seq time course and downstream analysis," *Computational and structural biotechnology journal*, vol. 13, pp. 469–477, 2015.

[3] S. M. Foltz, C. S. Greene, and J. N. Taroni, "Cross-platform normalization enables machine learning model training on microarray and rna-seq data simultaneously," *Communications Biology*, vol. 6, no. 1, p. 222, 2023.

[4] J. Liu, C. Qu, L. Zhang, Y. Tang, J. Li, H. Feng, X. Zeng, and X. Peng, "A new hybrid algorithm for three-stage gene selection based on whale optimization," *Scientific Reports*, vol. 13, no. 1, p. 3783, 2023.

[5] Z. Xu, F. Yang, C. Tang, H. Wang, S. Wang, J. Sun, and Y. Zhang, "Fg-hfs: A feature filter and group evolution hybrid feature selection algorithm for high-dimensional gene expression data," *Expert Systems with Applications*, vol. 245, p. 123069, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423035716

[6] C. Qu, L. Zhang, J. Li, F. Deng, Y. Tang, X. Zeng, and X. Peng, "Improving feature selection performance for classification of gene expression data using harris hawks optimizer with variable neighborhood learning," *Briefings in bioinformatics*, vol. 22, no. 5, p. bbab097, 2021.

[7] J. M. Franks *et al.*, "Feature-specific quantile normalization for reducing cross-platform discrepancies in genomic data," *Bioinformatics*, 2018.

[8] D. Skubleny, S. Ghosh, J. Spratlin, D. E. Schiller, and G. R. Rayat, "Feature-specific quantile normalization and feature-specific mean–variance normalization deliver robust bi-directional classification and feature selection performance between microarray and rnaseq data," *BMC bioinformatics*, vol. 25, no. 1, p. 136, 2024.

[9] S. Yuan, Y.-C. Chen, C.-H. Tsai, H.-W. Chen, and G. S. Shieh, "Feature selection translates drug response predictors from cell lines to patients," *Frontiers in Genetics*, vol. 14, p. 1217414, 2023.

[10] S. Mourragui, M. Loog, M. A. van de Wiel, M. J. Reinders, and L. F. Wessels, “Precise: A domain adaptation framework for preclinical-to-clinical transfer of cancer genomics,” *Nature Communications*, vol. 10, no. 1, p. 2552, 2019.

[11] S. Krishna *et al.*, “Advancing gene selection in oncology: A fusion of deep learning and sparsity,” *arXiv preprint arXiv:2403.01927*, 2024.

[12] D. Kim and Y. Jang, “Pan-cancer gene set discovery via scrna-seq for optimal deep learning based downstream tasks,” *arXiv preprint arXiv:2408.07233*, 2024.

[13] A. Chowdhury *et al.*, “A pan-cancer classification model using multi-view feature selection method and ensemble classifier,” *arXiv preprint arXiv:2501.06805*, 2025.

[14] R. Thelagathoti *et al.*, “Ensemble feature selection and mirna biomarker discovery for usher syndrome,” *Bioengineering*, vol. 12, no. 5, p. 497, 2025.

[15] A. Tom *et al.*, “Hybrid sequential feature selection for mrna biomarker discovery in usher syndrome,” *Biomolecules*, vol. 15, no. 7, p. 963, 2025.