Stability-Aware Relational kNN for Gene Expression Using Within-Sample Orderings

Izabela Justyna Kartowicz-Stolarska[0000−0003−2222−7452] and Marcin Czajkowski[0000−0002−3967−1819]

Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland
i.stolarska@pb.edu.pl, m.czajkowski@pb.edu.pl

Abstract. High-dimensional gene expression profiles are affected by technical variation such as measurement noise, missing features, and scale changes introduced by heterogeneous preprocessing. These effects can distort value-based distances, destabilize local neighborhoods, and lead to inconsistent predictions, particularly in k-nearest neighbors (kNN). An alternative is to compare samples through within-sample orderings between genes rather than absolute values, following the general lineage of relative expression ordering and gene-pair methods.
In this paper, we introduce a stability-aware relational kNN framework that operates in an ordering-based space using two complementary rank-derived distances: (i) an inversion-based distance that measures discordant gene-pair orderings between samples, and (ii) a displacement-based distance that aggregates absolute rank shifts.
The framework is evaluated against standard L_p norms (L_1, L_2, L_∞) on seven public gene expression datasets. Predictive quality is measured by macro-F1, while robustness is quantified by prediction flip rate and neighbor-set Jaccard similarity relative to an unperturbed baseline. Robustness is evaluated under five controlled perturbation scenarios with varying intensities: feature dropout, featurewise scaling, additive Gaussian noise, monotonic sample scaling (invariance check), and training-set instance dropout. The results characterize performance-stability trade-offs and show that ordering-based distances can improve neighborhood and prediction stability in several non-trivial perturbation regimes while maintaining competitive macro-F1.

Keywords: Gene Expression · Within-Sample Orderings · Rank-Based Distance · kNN · Neighborhood Stability · Robustness

1 Introduction

High-throughput gene expression profiling is a central modality in precision medicine, supporting molecular diagnosis, prognosis, and patient stratification [1]. At the same time, transcriptomic datasets are typically high-dimensional with limited sample sizes ($p \gg n$) and are affected by substantial technical variation. Measurement noise, missing features, and batch effects, including amplitude shifts introduced by heterogeneous laboratory protocols, platforms, and

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

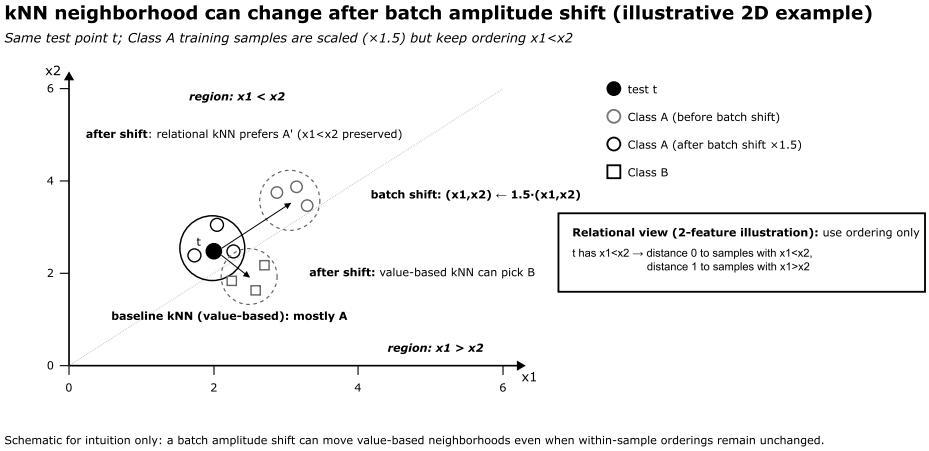

Fig. 1. Illustrative 2D example of how amplitude scaling can destabilize kNN under value-based distances. The test sample t lies in the region $x_1 < x_2$. Before perturbation, its nearest neighbors are mostly Class A. After scaling Class A training samples by a positive factor (here: $\times 1.5$), value-based distances (L_p) favor Class B neighbors, although the ordering $x_1 < x_2$ within Class A is preserved. A comparison based on within-sample orderings keeps the neighborhood consistent in this case.

preprocessing choices, can distort the observed signal [2, 3]. Related observations have also been reported for compositional sequencing data, where rank-based or relative representations can be more informative than absolute abundances and less sensitive to normalization choices [4]. As a result, models that appear accurate under a single pipeline or within-cohort validation may lose reproducibility under realistic technical perturbations or when applied across cohorts.

The k-nearest neighbors (kNN) classifier is a simple and widely used baseline for biomedical classification [5, 6]. Its prediction is fully determined by the neighborhood of the query sample, which makes the method conceptually interpretable: decisions can be traced to a small set of similar training instances. However, this reliance on the neighborhood also makes kNN highly sensitive to the choice of distance metric. In high-dimensional settings, value-based distances may behave counterintuitively and become unstable even under mild global shifts or feature-level perturbations [7]. As illustrated in Fig. 1, small technical artifacts can alter the neighbor set and flip the final prediction, even if the underlying biological ordering patterns are largely preserved.

A long-standing alternative in transcriptomics is to represent samples through *within-sample orderings* rather than absolute expression values. In gene-pair learning, Top Scoring Pairs (TSP) and its extensions encode each sample by ordinal relations, for example whether gene A exceeds gene B within the same sample, making the representation intrinsically resistant to monotonic transformations and certain normalization differences [8–12]. Closely related ideas appear in Relative Expression Ordering (REO) and, more broadly, Relative Ex-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

pression Analysis (RXA), where robustness is obtained by replacing values with comparisons and by reasoning in a space of within-sample relations. These developments naturally suggest extending the gene-pair viewpoint beyond discrete classifiers: rather than selecting a small set of pairs, one can define *sample-to-sample similarity* by the overall agreement of within-sample orderings.

Our recent work studied RXA-based within-sample relations embedded into interpretable tree-structured classifiers, including extensions toward multi-omics classification [13, 14]. In the present study, the same robustness-motivated viewpoint is applied in a complementary setting: instead of inducing explicit rules, ordering-derived distances are used inside kNN to study neighborhood and prediction stability under controlled perturbations.

We introduce a stability-aware *Relational kNN* framework for gene expression classification that operates in an ordering-based space. Value-based distances are replaced with ordering-derived distances that compare two samples by how consistently they preserve within-sample relations. Two complementary realizations are considered: (i) an inversion-based distance that counts discordant gene-pair orderings between samples, and (ii) a displacement-based distance that aggregates absolute rank shifts, capturing how far features move in the within-sample ordering. To handle missing features and varying effective comparisons across perturbations, an active-feature/active-pair normalization scheme keeps distances comparable when the set of usable relations changes. This paper extends our preliminary feasibility study [15]; here the emphasis shifts to stability, with a unified stress-test protocol contrasting inversion- and displacement-based variants under the same perturbations and reporting scheme. Rather than benchmarking a broad portfolio of classifiers, the goal is to isolate the effect of the distance representation within kNN.

The proposed framework is evaluated on seven public gene expression datasets under controlled perturbations reflecting common technical artifacts: feature dropout, featurewise scaling, additive Gaussian noise, monotonic sample scaling (invariance check), and training-set instance dropout. Predictive quality is measured by macro-F1, while robustness is quantified by prediction flip rate and neighbor-set Jaccard similarity relative to an unperturbed baseline. Beyond per-scenario analyses, robustness across the perturbation grid is summarized using winner maps, win counts, and a league-style aggregate score. Aggregates are reported both including and excluding monotonic scaling to avoid overstating invariances expected by design for ordering-based methods.

The main contributions of this paper are:

- A stability-centered evaluation protocol for kNN on gene expression data, quantifying both prediction stability (flip rate) and neighborhood stability (Jaccard overlap) under controlled perturbations.
- A relational kNN framework with two ordering-derived distance realizations (inversion-based and displacement-based) and an active-feature/active-pair normalization mechanism for fair comparison across perturbations with missing features.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

– An empirical characterization of the trade-off between predictive performance and stability on seven public datasets, including aggregate robustness summaries that control for monotonic invariance effects.

The remainder of the paper is organized as follows. Section 3 defines the relational distances and perturbation protocols. Section 4 describes datasets, evaluation measures, and experimental settings. Section 5 reports results. Section 6 concludes the study.

2 Background

This section briefly motivates ordering-based similarity and the stability measures used later.

2.1 kNN with value-based distances in high dimension

The kNN classifier assigns a label to a query sample by voting among its k closest training samples under a chosen distance metric [5, 6]. In gene expression analysis, common choices include Euclidean (L_2), Manhattan (L_1), and Chebyshev (L_∞) distances [16, 17]. Because these metrics operate on absolute expression values, they are sensitive to technical variation such as measurement noise, missing features, and batch effects, including amplitude shifts introduced by heterogeneous protocols, platforms, and preprocessing choices [2, 3].

This sensitivity is particularly problematic in transcriptomics for two reasons. First, high-dimensional spaces exhibit unintuitive distance behavior, which can make nearest-neighbor rankings fragile even under mild perturbations [7]. Second, preprocessing operations such as normalization, scaling, and batch correction may change the geometry of the value space and reorder nearest neighbors, directly affecting kNN predictions. As a result, evaluation based only on unperturbed accuracy may miss a practically important failure mode: good performance under a fixed pipeline but poor robustness under realistic technical artifacts.

2.2 Within-sample orderings in omics: TSP, RXA, and REO

A long-standing alternative in transcriptomics is to represent samples through *within-sample orderings* rather than absolute values. In supervised gene-pair learning, Top Scoring Pairs (TSP) and its extensions encode a sample by ordinal relations, for example whether gene A exceeds gene B within the same sample, where phenotype-discriminative pairs can be viewed as simple *biological switches* [8, 9]. More broadly, RXA and the REO line of work treat within-sample orderings as stable primitives for robust signatures and cross-sample comparisons; this broader relational perspective has also motivated subsequent methodological extensions of relative-expression algorithms beyond the original TSP setting [11, 18].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

These developments motivate extending the gene-pair viewpoint from rule-based classifiers to neighborhood-based learning: instead of selecting a small set of informative pairs, one can define *sample-to-sample similarity* by the overall agreement between within-sample orderings and use this similarity inside kNN.

2.3 Ordering-derived sample distances

To compare two samples in an ordering-based space, a natural strategy is to quantify how much their within-sample orderings disagree. We consider two complementary notions. First, *pairwise ordering discordance* counts how often an ordering relation between two genes is reversed between samples. Second, *rank-shift magnitude* measures how far genes move in the within-sample ranking between samples. These notions are related to classical inversion-counting and rank-disarray families [19, 20], but here they are used operationally as distances inside kNN. Formal definitions and normalization are given in Section 3.

In practice, feature dropout and ties change which relations are comparable. Therefore, ordering-derived distances should be normalized by the number of *active* usable features or comparisons to remain comparable across perturbations.

2.4 Stability measures for kNN

In this work, stability refers to how much neighbor sets and predictions change after perturbations, reflecting robustness of the kNN mechanism to technical artifacts and pipeline variability [21]. We use two measures aligned with kNN:

- **Prediction flip rate**: the fraction of test samples whose predicted label changes relative to the baseline (unperturbed) prediction.
- **Neighbor-set stability**: overlap between the baseline neighbor set N_0 and the perturbed neighbor set N_1, measured by Jaccard similarity:

$$J(N_0, N_1) = \frac{|N_0 \cap N_1|}{|N_0 \cup N_1|}.$$

Together with macro-F1, these measures operationalize the accuracy-stability trade-off that motivates ordering-based distances for kNN in transcriptomics.

3 Methodology

This section defines the stability-aware relational kNN framework. Samples are compared through *within-sample orderings* rather than absolute expression values. We introduce two ordering-derived distances, together with normalization that keeps them comparable under missing features and ties, and then specify the perturbation protocols and stability measures used in the evaluation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

3.1 Notation and within-sample ordering representation

A sample is a vector $x \in \mathbb{R}^p$ with p features (genes). Let $\mathcal{F} = \{1, \ldots, p\}$ denote the feature index set. Within-sample ordering can be represented either by (i) pairwise relations $\text{sign}(x_i - x_j)$ or (ii) a per-sample ranking $r_x(i)$.

Both views are used here: the inversion-based distance is defined on pairwise relations, whereas the displacement-based distance is defined on within-sample ranks. Ties are handled explicitly: tied pairs ($x_i = x_j$) are excluded from inversion counting, and ranks for the displacement distance are computed as *midranks* on the current feature subset.

In some perturbations, e.g. feature dropout, a subset of features is removed. We represent feature availability for sample x by a binary mask $m_x \in \{0,1\}^p$, where $m_x(i) = 1$ indicates that feature i is present.

3.2 Active-feature and active-pair normalization

When comparing two samples x and y, only features present in both are usable:

$$\mathcal{F}_{xy} = \{i \in \mathcal{F} : m_x(i) = 1 \ \wedge \ m_y(i) = 1\}, \qquad p_{xy} = |\mathcal{F}_{xy}|.$$

All ordering computations are performed after restriction to $\mathcal{F}_{xy}$, i.e., ranks are recomputed on the common feature subset.

For inversion-based distances, not all pairs in $\binom{p_{xy}}{2}$ are comparable because of ties; we therefore normalize by the number of *active pairs* (non-tied in both samples). This keeps distances comparable across perturbations that change p_{xy} or increase tie rates.

3.3 Ordering-derived distances

Two complementary distances are used. Both are defined on $\mathcal{F}_{xy}$ and map to $[0,1]$.

Pairwise discordance distance (inversion-based) For $i < j$ with $i, j \in \mathcal{F}_{xy}$, define the pair to be *active* if it is not tied in either sample:

$$\mathbb{I}_{\text{active}}(x, y; i, j) = \mathbf{1}(x_i \neq x_j) \cdot \mathbf{1}(y_i \neq y_j).$$

A discordance occurs if

$$(x_i - x_j)(y_i - y_j) < 0.$$

We count discordant active pairs and the total number of active pairs:

$$\text{disc}(x, y) = \sum_{\substack{i<j \\ i,j \in \mathcal{F}_{xy}}} \mathbb{I}_{\text{active}}(x, y; i, j) \cdot \mathbf{1}\left((x_i - x_j)(y_i - y_j) < 0\right),$$

$$\text{act}(x, y) = \sum_{\substack{i<j \\ i,j \in \mathcal{F}_{xy}}} \mathbb{I}_{\text{active}}(x, y; i, j).$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

The inversion-based distance is

$$D_{\mathrm{inv}}(x,y) = \begin{cases} \frac{\mathrm{disc}(x,y)}{\mathrm{act}(x,y)} & \text{if } \mathrm{act}(x,y) > 0, \\ 1.0 & \text{if } \mathrm{act}(x,y) = 0. \end{cases}$$

This is related to classical inversion counting [19]. The $\mathrm{act}(x,y) = 0$ case may occur under extreme ties or degenerate feature overlap; assigning distance 1.0 avoids treating such samples as artificially close.

Rank-shift distance (displacement-based) Let $r_x^{xy}(i)$ denote the midrank of feature i within sample x computed only on the common subset $\mathcal{F}_{xy}$ (analogously for $r_y^{xy}(i)$). The normalized displacement distance is

$$D_{\mathrm{disp}}(x,y) = \begin{cases} \frac{1}{p_{xy}(p_{xy}-1)} \sum_{i \in \mathcal{F}_{xy}} \left| r_x^{xy}(i) - r_y^{xy}(i) \right| & \text{if } p_{xy} \geq 2, \\ 1.0 & \text{if } p_{xy} < 2. \end{cases}$$

The normalization maps the distance to $[0,1]$ because $r_x^{xy}(i) \in [1, p_{xy}]$, hence $|r_x^{xy}(i) - r_y^{xy}(i)| \leq p_{xy} - 1$ for each i. This construction follows the rank-displacement family [20].

3.4 kNN classifier and baselines

Given a distance $D(\cdot,\cdot)$, for each test sample t the k nearest training samples are selected and the class is predicted by majority vote [5, 6]. The same kNN procedure is applied to all distances. As value-based baselines, L_1, L_2, and L_∞ are computed on the same effective feature set used in a given experiment, i.e., after any fold-level feature selection and any scenario-specific masking. In high-dimensional settings, neighbor rankings can be fragile [7], motivating explicit stability evaluation under perturbations.

3.5 Perturbation protocols

Stability is evaluated by comparing predictions and neighbor sets from an unperturbed baseline with those obtained after controlled perturbations. Each perturbation uses fixed random seeds. An intensity parameter α controls perturbation strength and is evaluated over a grid of values (reported in Section 4).

- **Additive Gaussian noise (test-time).** For a test sample x, generate $\tilde{x} = x + \epsilon$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ independently and $\sigma_i = \alpha \cdot \mathrm{std}_{\mathrm{train}}(i)$.
- **Feature dropout (train and test).** Randomly remove an α fraction of features using a binary mask. The same mask is applied to all samples within a fold, so distances are computed on a consistent reduced feature space. For ordering-derived distances, rankings and relations are recomputed after restriction to the retained features.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

- **Featurewise scaling (train and test).** Model gene-specific amplitude distortions by $\tilde{x}_i = x_i \cdot s_i$, with $s_i \sim \mathcal{U}(1-\alpha, 1+\alpha)$ independently across features. This perturbation can change within-sample orderings locally.
- **Monotonic sample scaling (test-time invariance check).** Model sample-wise amplitude shifts by $\tilde{x} = s \cdot x$, where $s \sim \mathcal{U}(1-\alpha, 1+\alpha)$ is drawn independently per test sample. Ordering-based distances are invariant to strictly monotonic rescaling up to ties, whereas value-based distances are not.
- **Training-set instance dropout.** Randomly remove an α fraction of training samples using stratified sampling to preserve class ratios, then reclassify the fixed test set.

3.6 Stability measures

For each test sample $t \in T$, we compare the unperturbed baseline with the outcome under perturbation p using three complementary criteria: predictive quality (macro-F1), prediction stability, and neighborhood stability. Let $\hat{y}_0(t)$ and $\hat{y}_p(t)$ denote the predicted labels, and let $N_0(t)$ and $N_p(t)$ be the sets of k nearest neighbors, respectively. Prediction stability is quantified by the flip rate

$$\text{FlipRate} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}(\hat{y}_p(t) \neq \hat{y}_0(t)),$$

and neighborhood stability by the Jaccard overlap

$$\text{Jaccard} = \frac{1}{|T|} \sum_{t \in T} \frac{|N_0(t) \cap N_p(t)|}{|N_0(t) \cup N_p(t)|}.$$

Together with macro-F1, these measures distinguish stability of the decision from stability of the local neighborhood under controlled perturbations.

4 Experimental Setup

This section describes the datasets, preprocessing, compared distance metrics, parameter grids, and the evaluation protocol used to benchmark predictive quality and perturbation stability.

4.1 Datasets

We use seven public gene expression datasets from the Gene Expression Omnibus (GEO) [22]. Each dataset defines a binary classification task and matches the benchmark suite used in our earlier feasibility study [15]. Across datasets, the sample size ranges from 105 to 192, and the original dimensionality ranges from 22,215 to 54,676 features prior to fold-level feature selection. Table 1 reports dataset identifiers, sizes, and class ratios.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

Table 1. Summary of GEO gene expression datasets used in the benchmark.

Dataset	Original features	Samples	Ratio	Description
(a) GDS2771	22,215	192	102:90	Lung cancer
(b) GSE10072	22,284	107	58:49	Adenocarcinoma
(c) GSE17920	54,676	130	92:38	Classic Hodgkin's lymphoma
(d) GSE19804	54,613	120	60:60	NSCLC
(e) GSE27272	24,526	183	128:55	Tobacco effects on pregnancy
(f) GSE3365	22,284	127	85:42	PBMCs and Crohn's disease
(g) GSE6613	22,284	105	55:50	Parkinson disease

4.2 Preprocessing and fold-level feature selection

To maintain comparability with the IDS study [15], we replicate the same fold-level feature selection protocol: ReliefF [23] applied on the training fold only, retaining the top 1,000 genes, and then applied unchanged to the corresponding test fold. All distances (value-based and ordering-based) are computed on this fold-specific feature set; any scenario-specific feature masking (e.g., feature dropout) is applied on top of it. The same selected features are used for all compared distances, so observed stability differences reflect the distance definitions rather than different inputs.

4.3 Compared distances and parameter grids

Five kNN distance metrics are compared: (i) the pairwise-discordance ordering distance (RR_K; inversion-based), (ii) the rank-shift ordering distance (RR_F; displacement-based), and three value-based baselines: Manhattan (L_1), Euclidean (L_2), and Chebyshev (L_∞).

We evaluate neighborhood sizes $k \in \{3, 5, 7\}$ and perturbation intensities on a shared grid $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ for all perturbation scenarios. For ordering-derived distances, distances are normalized using the active-feature / active-pair scheme (Section 3) to ensure comparability when feature availability or tie patterns change.

4.4 Evaluation protocol

Stratified 10-fold cross-validation is used [24]. For each fold and each distance metric, baseline predictions and neighbor sets are computed on unperturbed test samples. Each perturbation scenario (A-E) is then applied at each α level, and predictions and neighbor sets are recomputed for the same test fold. Stability is measured relative to the baseline of the *same* method (i.e., method-specific baseline neighborhoods and predictions).

We report: (i) macro-F1 (predictive quality), (ii) prediction flip rate (predictive stability), and (iii) neighbor-set Jaccard similarity (neighborhood stability), as defined in Section 3. Robustness across perturbations and datasets is summarized using winner maps and win counts. In addition, a league-style aggregate

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

score is reported, defined as $z(\text{Jaccard}) - z(\text{FlipRate})$, where $z(\cdot)$ denotes standardization across methods within the same comparison set. Scenario D is used as an invariance sanity check; therefore, we also report aggregates excluding the monotonic sample-scaling scenario to avoid attributing expected rank-invariance effects to non-trivial robustness.

4.5 Implementation

The proposed solution was implemented in Python 3.12 and released as an open-source toolkit named *RelOmnica*, available in a public GitLab repository [25]. The toolkit provides a set of functions and scripts enabling reproduction of the experiments described in this article.

The project depends on the *pyRRM* library [26], which implements the proposed relational distance metrics used for ordering-based neighborhood construction. To accelerate and parallelize computationally intensive operations, the *numba* library [27] was employed.

To facilitate reproducibility and ease of deployment, the repository includes a Dockerfile as well as configuration files for creating a dedicated virtual environment. Additionally, example scripts for running the experimental pipelines are provided.

5 Results and Discussion

Perturbation robustness of kNN is evaluated for two ordering-derived distances (RR_F, RR_K) and three value-based baselines (L_1, L_2, L_∞). We first analyze the two relational variants across perturbation types, intensity α, and neighborhood size k.

5.1 Relational variants across scenarios, α, and k

Figure 2 reports macro-F1, prediction flip rate, and neighbor-set Jaccard stability for RR_F and RR_K across perturbation scenarios (rows A-E), intensities α (columns), and neighborhood sizes $k \in \{3, 5, 7\}$ (blocks).

Two patterns are consistent across both relational variants. First, monotonic sample scaling (D) acts as an invariance sanity check: positive sample-wise rescaling preserves within-sample orderings up to ties, so flip rates remain near zero and Jaccard overlap remains near one across the α grid. Second, perturbations that either disturb within-sample orderings (Gaussian noise, C) or modify the available reference set (training instance dropout, E) produce the clearest degradation with increasing α, visible as rising flip rates and reduced neighbor overlap. Featurewise scaling (B) typically yields intermediate effects, while feature dropout (A) stresses the active-feature/active-pair normalization by reducing the effective comparison set.

Neighborhood size also matters. In this benchmark, smaller k tends to be more stable for ordering-derived distances: $k = 3$ most often provides the most

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

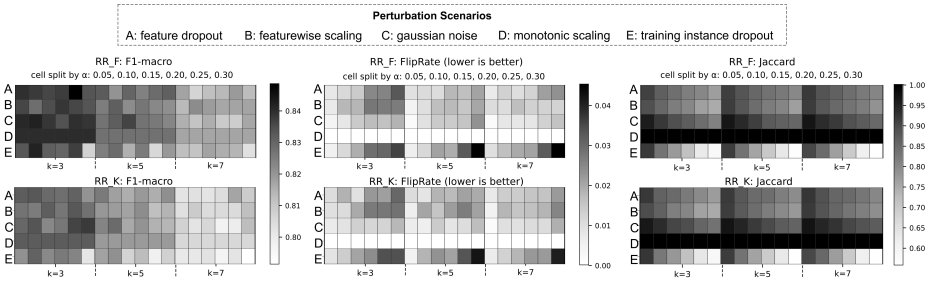

Fig. 2. Relational kNN variants across perturbation scenarios, α, and neighborhood size k. Rows correspond to scenarios (A: feature dropout, B: featurewise scaling, C: Gaussian noise, D: monotonic sample scaling, E: training instance dropout). Columns within each block correspond to $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. Metrics shown are macro-F1, flip rate (lower is better), and Jaccard similarity (higher is better).

favorable performance-stability balance, whereas larger k can include more distant neighbors and amplify instability under stronger perturbations. Based on this observation, and consistently with IDS [15], $k = 3$ is used for the main benchmark against value-based distances. Ordering-derived distances are computationally more demanding than standard L_p metrics because they operate on within-sample relations rather than direct value differences. In our implementation, a single 10-fold cross-validation pass on one dataset required approximately 0.3 s for an ordering-derived metric, compared with about 0.02 s for a standard value-based metric, indicating a moderate but practically manageable overhead at the benchmark scale considered here.

5.2 Benchmark against value-based distances for $k = 3$

Figure 3 compares RR_F and RR_K to L_1, L_2, and L_∞ for $k = 3$. The top row shows winner maps across scenarios and α, while the remaining panels aggregate results across datasets and perturbations, reported both including and excluding the monotonic-scaling sanity check (scenario D).

Three observations follow from Fig. 3. First, averaged over the perturbation grid, ordering-derived distances remain competitive in predictive quality relative to value-based baselines. Second, the league-style stability score consistently penalizes L_∞, indicating that Chebyshev distance is comparatively unstable under these perturbations. Third, excluding monotonic scaling (D) highlights the nontrivial robustness profile: ordering-derived distances frequently achieve a more favorable balance between prediction volatility and neighborhood consistency, often winning on flip rate while remaining competitive on macro-F1. At the same time, value-based distances can sometimes retain higher neighbor overlap (Jaccard) under mild perturbations, yet this does not necessarily translate into lower flip rates. This separation motivates reporting both FlipRate (semantic stability) and Jaccard (structural stability) alongside macro-F1.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

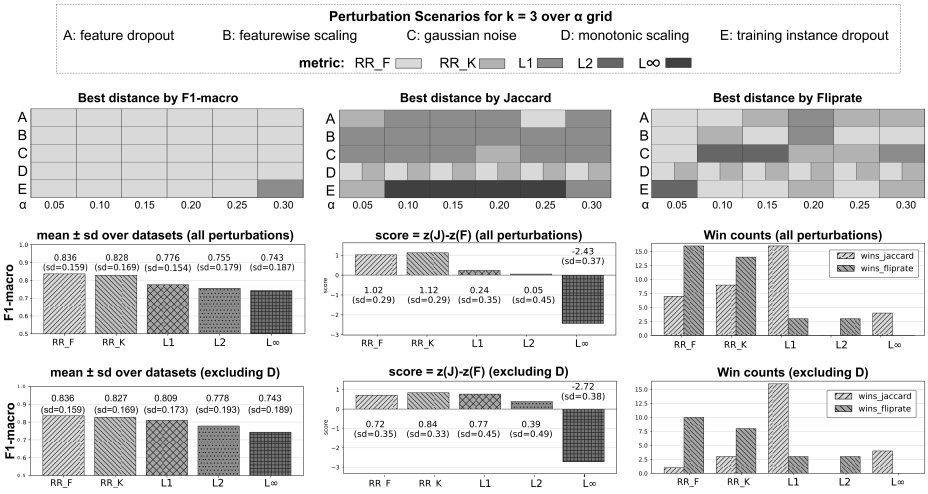

Fig. 3. Benchmark for $k = 3$ across the perturbation grid. **Top row:** winner maps indicating which distance achieves the best macro-F1 (left), best Jaccard stability (middle), and lowest flip rate (right) for each scenario (A–E) and α. **Middle row:** aggregates over all scenarios: mean±sd macro-F1 across datasets, league-style stability score $z(\text{Jaccard}) - z(\text{FlipRate})$ (higher is better), and win counts. **Bottom row:** the same aggregates excluding monotonic scaling (scenario D) to avoid inflating results due to trivial rank invariance.

The benchmark is intentionally metric-centered: we compare neighborhood definitions within a fixed kNN classifier rather than a broad portfolio of learning algorithms. Broader classifier-level comparisons were reported in our preliminary IDS study [15]; here the goal is to isolate robustness effects attributable to the distance representation itself.

5.3 Dataset-level statistical analysis across seven datasets

For statistical testing, each dataset was treated as a matched block ($n = 7$). For each distance and perturbation level ($\alpha \in \{0.15, 0.30\}$), outcomes were aggregated by taking the median across scenarios, computed both including and excluding scenario D. Friedman tests were then applied across distances, followed by paired Wilcoxon signed-rank post-hoc tests with Benjamini-Hochberg correction.

For neighborhood stability (Jaccard), Friedman tests were significant at both perturbation levels, both including scenario D ($p < 10^{-4}$) and excluding scenario D ($p < 10^{-3}$). Including scenario D, post-hoc tests confirmed that RR_F and RR_K outperform the L_p baselines (all $p_{\text{adj}} < 0.05$). Excluding scenario D, relational distances consistently outperformed L_∞ ($p_{\text{adj}} < 0.05$), but did not differ significantly from L_1 and L_2 after correction; additional differences within the L_p baselines were observed.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

6 Conclusion and Future Work

This study benchmarked perturbation stability of kNN classification on high-dimensional gene expression data. We evaluated a stability-aware *Relational kNN* framework in which value-based distances were replaced with ordering-derived comparisons using two complementary realizations: an inversion-based distance measuring discordant within-sample relations (RR_K) and a rank-shift distance aggregating absolute rank changes (RR_F). To keep distances comparable when the effective feature set changes, e.g. under feature dropout, and when ties occur, normalization was applied based on the number of active features (for RR_F) and active non-tied pairs (for RR_K).

Three main conclusions follow from experiments on seven public datasets and five controlled perturbation scenarios. First, predictive quality alone is not a reliable proxy for robustness: distances with similar macro-F1 can differ substantially in neighborhood stability and in how often predictions change after perturbations. Second, ordering-derived distances often reduce prediction volatility under non-trivial perturbations such as additive noise and feature dropout while remaining competitive in macro-F1, although the advantage is scenario- and dataset-dependent. Reporting aggregates both including and excluding the monotonic-scaling sanity check provides a conservative robustness summary and avoids overstating invariances expected by design for ordering-based methods. Third, RR_K and RR_F show broadly similar behavior across the tested grid, suggesting that the main robustness effect comes from operating in an ordering-based space, while the specific form of inversion- or displacement-based disagreement plays a secondary role in this benchmark.

The present study is intentionally focused on robustness within a fixed kNN framework. Its aim is not to claim a universally best classifier, but to isolate how the choice of distance representation affects neighborhood stability and prediction consistency under controlled perturbations. This scope also explains two limitations of the current work. First, ordering-derived distances are computationally more expensive than L_p norms and depend on implementation efficiency, feature selection, and tie handling. Second, the perturbations used here are synthetic stress tests: they enable controlled sensitivity analysis, but do not fully capture the complexity of real cross-cohort and cross-platform shifts. In addition, while the experiments were performed on gene expression data, the present analysis remains primarily methodological and does not yet attempt biological interpretation of stable relations or pathways.

Future work will therefore proceed in three directions. First, efficiency can be improved through parallelization, approximate nearest-neighbor search for rank-based metrics, selective relation sampling, and, in the longer term, GPU acceleration. Second, evaluation should be extended to broader classifier-level comparisons and to real cross-cohort transfer settings, where training and test data come from different studies or platforms. Third, the relational representation itself can be made more task-adaptive by incorporating margin-aware variants that ignore near-ties, weighted variants that emphasize informative or stable gene pairs, and hybrid distances that interpolate between inversion- and

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

displacement-based disagreement. A further step will be to add interpretability layers that identify stable gene relations, compact within-sample rules, or ranked pair lists that can be linked to biologically meaningful processes and used for clinician-facing explanations.

Acknowledgments

This project was supported by the grant WI/WI-IIT/5/2025 (Application of a Relatively Relational Metric Between Features in GeneOmics Data) from Bialystok University of Technology, funded by the Polish Ministry of Science and Higher Education (first author), and by grant WZ/WI-IIT/4/2026 (second author).

References

1. Cai, D., Qi, H., Yang, Q., et al.: Personalized risk stratification in colorectal cancer via PIANOS system. *Nature Communications* 16, 6561 (2025). doi:10.1038/s41467-025-61713-1
2. Leek, J.T., Scharpf, R.B., Bravo, H.C., et al.: Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* 11, 733-739 (2010). doi:10.1038/nrg2825
3. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1), 118-127 (2007). doi:10.1093/biostatistics/kxj037
4. Zheng, X., Jin, N., Wu, Q., et al.: Less is more: relative rank is more informative than absolute abundance for compositional NGS data. *Briefings in Functional Genomics* 24, elae045 (2025). doi:10.1093/bfgp/elae045
5. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1), 21-27 (1967). doi:10.1109/TIT.1967.1053964
6. Taunk, K., De, S., Verma, S., Swetapadma, A.: A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. pp. 1255-1260. IEEE (2019). doi:10.1109/ICCS45141.2019.9065747
7. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) Database Theory — ICDT 2001. LNCS, vol. 1973, pp. 420-434. Springer, Heidelberg (2001). doi:10.1007/3-540-44503-X_27
8. Geman, D., d'Avignon, C., Naiman, D.Q., Winslow, R.L.: Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical Applications in Genetics and Molecular Biology* 3(1) (2004). doi:10.2202/1544-6115.1071
9. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., Geman, D.: Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21(20), 3896-3904 (2005). doi:10.1093/bioinformatics/bti631
10. Czajkowski, M., Kretowski, M.: Novel extension of k-TSP algorithm for microarray classification. In: Nguyen, N.T., Borzemski, L., Grzech, A., Ali, M. (eds.) *New Frontiers in Applied Artificial Intelligence*. IEA/AIE 2008. LNCS, vol. 5027, pp. 420-429. Springer, Berlin, Heidelberg (2008). doi:10.1007/978-3-540-69052-8_48

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19

11. Eddy, J.A., Sung, J., Geman, D., Price, N.D.: Relative expression analysis for molecular cancer diagnosis and prognosis. *Technology in Cancer Research & Treatment* 9(2), 149-159 (2010). doi:10.1177/153303461000900204
12. Wu, C., Xie, X., Yang, X., Du, M., Lin, H., Huang, J.: Applications of gene pair methods in clinical research: advancing precision medicine. *Molecular Biomedicine* 6(1), 22 (2025). doi:10.1186/s43556-025-00263-w
13. Czajkowski, M., Jurczuk, K., Kretowski, M.: Enhancing transparency of omics data analysis with the Evolutionary Multi-Test Tree and Relative Expression. *Expert Systems with Applications* 276, 127131 (2025). doi:10.1016/j.eswa.2025.127131
14. Czajkowski, M., Jurczuk, K., Kretowski, M.: Enhancing multi-omics data classification with relative expression analysis and decision trees. *Journal of Computational Science* 84, 102460 (2025). doi:10.1016/j.jocs.2024.102460
15. Kartowicz-Stolarska, I.J., Czajkowski, M.: Relative relation in kNN classification for gene expression data: a preliminary study. In: *Proc. 32nd International Conference on Information Systems Development (ISD 2024)*. University of Gdańsk (2024). doi:10.62036/ISD.2024.94
16. Cunningham, P., Delany, S.J.: k-Nearest Neighbour Classifiers-A Tutorial. *ACM Computing Surveys* 54(6), Article 128, 25 pp. (2021). doi:10.1145/3459665
17. Parry, R.M., Jones, W., Stokes, T.H., Phan, J.H., Moffitt, R.A., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., Wang, M.D.: k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.* 10(4), 292-309 (2010). doi:10.1038/tpj.2010.56
18. Czajkowski, M., Kretowski, M.: Evolutionary approach for relative gene expression algorithms. *The Scientific World Journal* 2014, 593503 (2014). doi:10.1155/2014/593503
19. Kendall, M.G.: A New Measure of Rank Correlation. *Biometrika* 30(1-2), 81-93 (1938). doi:10.1093/biomet/30.1-2.81
20. Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(2), 262-268 (1977). doi:10.1111/j.2517-6161.1977.tb01624.x
21. McShane, L.M., Radmacher, M.D., Freidlin, B., Yu, R., Li, M.C., Simon, R.: Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18(11), 1462-1469 (2002). doi:10.1093/bioinformatics/18.11.1462
22. Barrett, T., Wilhite, S.E., Ledoux, P., et al.: NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research* 41(D1), D991-D995 (2013). doi:10.1093/nar/gks1193
23. Kononenko, I.: Estimating attributes: Analysis and extensions of RELIEF. In: De Raedt, L., Wrobel, S. (eds.) Machine Learning: ECML-94. LNCS, vol. 784, pp. 171-182. Springer (1994). doi:10.1007/3-540-57868-4_57
24. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of IJCAI 1995*, pp. 1137-1143 (1995).
25. RelOmnica project on GitLab. `https://gitlab.com/but-stolarska-czajkowski/relomnica`
26. PyRRM project on GitLab. `https://gitlab.com/izabeera/pyrrm`
27. Lam, S.K., Pitrou, A., Seibert, S.: Numba: A LLVM-based Python JIT Compiler. In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (LLVM-HPC 2015) (2015). doi:10.1145/2833157.2833162

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_19