life

MDPI

Article

MO-POISE-X: Multi-Objective Evolutionary Search with Assay-Aware Block-Sparse Initialization for Compact Multi-Omics Panels of Endometrial Cancer Molecular Subtypes

Yiqin Zhang †, Yan Ding †, Minyue Cao, Jiayu Yan, Yihan Sun, Genyi Jiang, Yanli Zhang, Luyao Kang, Jing Luo and Bilan Li *

Shanghai Key Laboratory of Maternal Fetal Medicine, Shanghai Institute of Maternal-Fetal Medicine and Gynecologic Oncology, Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai 200092, China; cathefrine811@163.com (Y.Z.); rdingyan@163.com (Y.D.); minlei0513@163.com (M.C.); yanjiayu@51mch.com (J.Y.); sunyh@tongji.edu.cn (Y.S.); jianggenyi@163.com (G.J.); zhangyanlia@51mch.com (Y.Z.); luyaokang@tongji.edu.cn (L.K.); 2532243@tongji.edu.cn (J.L.)

* Correspondence: libilan@51mch.com

† These authors contributed equally to this work.

Abstract

The four molecular subtypes of endometrial carcinoma defined by The Cancer Genome Atlas (TCGA) can be reconstructed from multi-omics data, but a practical panel should require few features and few assays. For 369 TCGA patients with five omics layers, we developed MO-POISE-X, a three-objective formulation of macro-averaged F_1, number of omics assays, and panel size within leakage-controlled nested cross-validation, with a block-sparse initialization that samples assays before features. Thirteen multi-objective evolutionary algorithms were plugged into MO-POISE-X and compared with filter fronts built by analysis of variance (ANOVA), minimum redundancy–maximum relevance (mRMR), and L1-penalized logistic regression. In hypervolume, which measures the whole trade-off, most algorithms did not differ significantly from the best. Without sampling assays first, the hypervolume of the ten algorithms with the highest hypervolume fell from 0.946–0.967 to 0.880–0.926, whereas denser initializations gave higher hypervolumes with larger panels. With at most five features, the evolutionary panels of these ten algorithms reached a macro-averaged F_1 of 0.746–0.804, compared with 0.674–0.752 for the filters, significantly more than the ANOVA and L1 panels but not the mRMR panels. The filter fronts, in turn, reached higher hypervolumes and a macro-averaged F_1 of about 0.90 with large panels. Two assays, somatic mutation plus methylation or expression, provided most of the attainable accuracy, and these findings held for other classifiers, candidate-pool sizes, and assay-specific costs. Built from single features, the evolutionary panels used the assays and markers of the clinical classification. A molecular surrogate of this classification, built from the same data, remained more accurate, with a Cohen's κ of 0.865; adding its readouts to the candidate pool raised the κ of the compact panels of three selected algorithms from 0.735 to 0.795. In this cohort, evolutionary search supplied compact panels with few features, and filter selection supplied the most accurate large panels.

Keywords: endometrial carcinoma; molecular subtyping; multi-omics; multi-objective evolutionary feature selection; assay cost

Check for updates

Academic Editors: Joon-Yong Chung and Lefteris Koumakis

Received: 23 October 2026
Revised: 14 November 2026
Accepted: 18 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The molecular classification of endometrial carcinoma introduced by The Cancer Genome Atlas (TCGA) divides tumors into four subtypes, *POLE*-ultramutated, microsatellite instability (MSI)-hypermutated, copy-number-low, and copy-number-high, that differ in their genomic drivers and in prognosis [1]. Surrogate classifiers built from three clinical assays reproduce these groups [2,3], molecular classification predicts the benefit of adjuvant therapy [4], and the guidelines of the European Society of Gynaecological Oncology, the European Society for Radiotherapy and Oncology, and the European Society of Pathology (ESGO/ESTRO/ESP) recommend it for risk stratification [5]. The subtypes can also be reconstructed from multi-omics profiles. In the TCGA uterine corpus endometrial carcinoma (TCGA-UCEC) cohort, however, five omics layers and a few clinical variables amount to about 460,000 features for fewer than 400 patients, and a classifier that uses all of them indicates neither which measurements carry the subtype signal nor which assays would have to be performed to apply it.

Feature selection addresses the first of these issues [6,7]. In a previous benchmark on this cohort [8], we compared 19 single-objective metaheuristic wrappers under leakage-controlled nested cross-validation [9]. Such wrappers, like most wrapper methods, combine classification error and panel size into a single weighted sum, whose weight is a convention rather than a calibrated trade-off [10,11], and they measure the cost of a panel only by the number of its features. For a clinical or translational application, the decisive cost is a different one. Each omics assay needs its own sample preparation, measurement platform, quality control, and analysis pipeline; it consumes part of a tumor sample that is often small, and every additional assay lengthens the time until a result is available. Once an assay has been performed, however, further features from it come at almost no extra cost: a methylation array reports hundreds of thousands of probes in a single run, and a sequencing panel reports all its genes together. The cost of applying a panel is therefore set mainly by the number of assays it requires and only marginally by the number of its features, so that a panel of ten features that draws on five different omics assays is far more expensive to apply than a panel of a hundred features from a single assay. The same reasoning drives current efforts to replace the three assays of the clinical classification by a single sequencing panel [12].

Multi-objective evolutionary algorithms (MOEAs) avoid the weighted sum by approximating the whole Pareto front of non-dominated trade-offs [13–15]. They also allow the cost of a panel to be expressed in the unit that matters, here the number of assays it requires. Two questions are then open. The first is methodological: under the same leakage controls as in the single-objective case, where along the trade-off between accuracy, assays, and panel size do the fronts found by MOEAs improve on what univariate and multivariate filter strategies achieve? The second is practical: which of the many available MOEAs should be used for this kind of problem, a binary search space of a few thousand variables with a noisy, expensive objective and a block structure? Algorithms for sparse and large-scale problems are mostly developed on synthetic benchmarks with cheap objectives and large evaluation budgets [16], whereas every evaluation of a panel is a cross-validation, so their ranking on such benchmarks need not carry over.

In this study, we present MO-POISE-X, a Multi-Objective Plug-and-play Optimizer for Imbalance-aware Subtyping of Endometrial cancer, which formulates the reconstruction of the TCGA subtypes as a three-objective problem: inner cross-validated macro-averaged F_1, number of omics assays, and panel size. The pipeline is plug-and-play in that any MOEA operates on the same problem, evaluation, and budget, with shared operators wherever the algorithm does not prescribe its own, and in that further candidate features can be added without changes to the pipeline. It is imbalance-aware in that the macro-

averaged F_1 weights the rare *POLE* class equally with the others. The X stands for the plugged-in MOEA, so that, for example, MO-POISE-NSGA-II denotes the pipeline with NSGA-II. We embed it in leakage-controlled nested cross-validation and compare 13 selected MOEAs. Each algorithm is judged against three filter fronts. Each front combines every subset of data blocks with panels of increasing size, whose features are ranked univariately by one-way analysis of variance (ANOVA), ranked with a redundancy penalty by minimum redundancy–maximum relevance (mRMR), or selected by L1-penalized logistic regression, and panels are chosen from a front by rules that use training data only. Our contributions are:

- We propose MO-POISE-X, an assay-aware, three-objective, leakage-controlled pipeline for multi-omics subtype reconstruction, with problem-specific operators for block-structured binary search spaces into which any MOEA can be plugged.
- We propose a block-sparse initialization that samples the omics assays of each initial panel before its features. An ablation with ten of the MOEAs shows that it raises the hypervolume compared with a sparse initialization without this block step, and an ablation with three of them shows that it yields panels with fewer assays and features than the usual uniform initialization.
- We present a comprehensive comparison of 13 MOEAs in the MO-POISE-X framework under an identical computational budget.
- We construct three filter reference fronts and training-side panel-selection rules, which locate where along the trade-off the evolutionary search improves on univariate and sparse-regression selection and where it matches redundancy-aware selection.
- We determine how many assays are needed to reconstruct the TCGA subtypes from multi-omics data, and which assays they are.
- We examine the robustness of these findings to the size of the candidate pool, to assay-specific costs, to the classifier, and across the outer folds, and compare the panels with a molecular surrogate of the clinical classification.

The remainder of the paper is organized as follows. Section 2 reviews the molecular classification of endometrial carcinoma, multi-objective evolutionary feature selection, and their intersection in the selection of multi-omics panels. Section 3 describes MO-POISE-X, outlined in Figure 1, with the data, the formulation, the algorithms, and the evaluation. Section 4 reports the results of the benchmark, Section 5 discusses them, and Section 6 concludes.

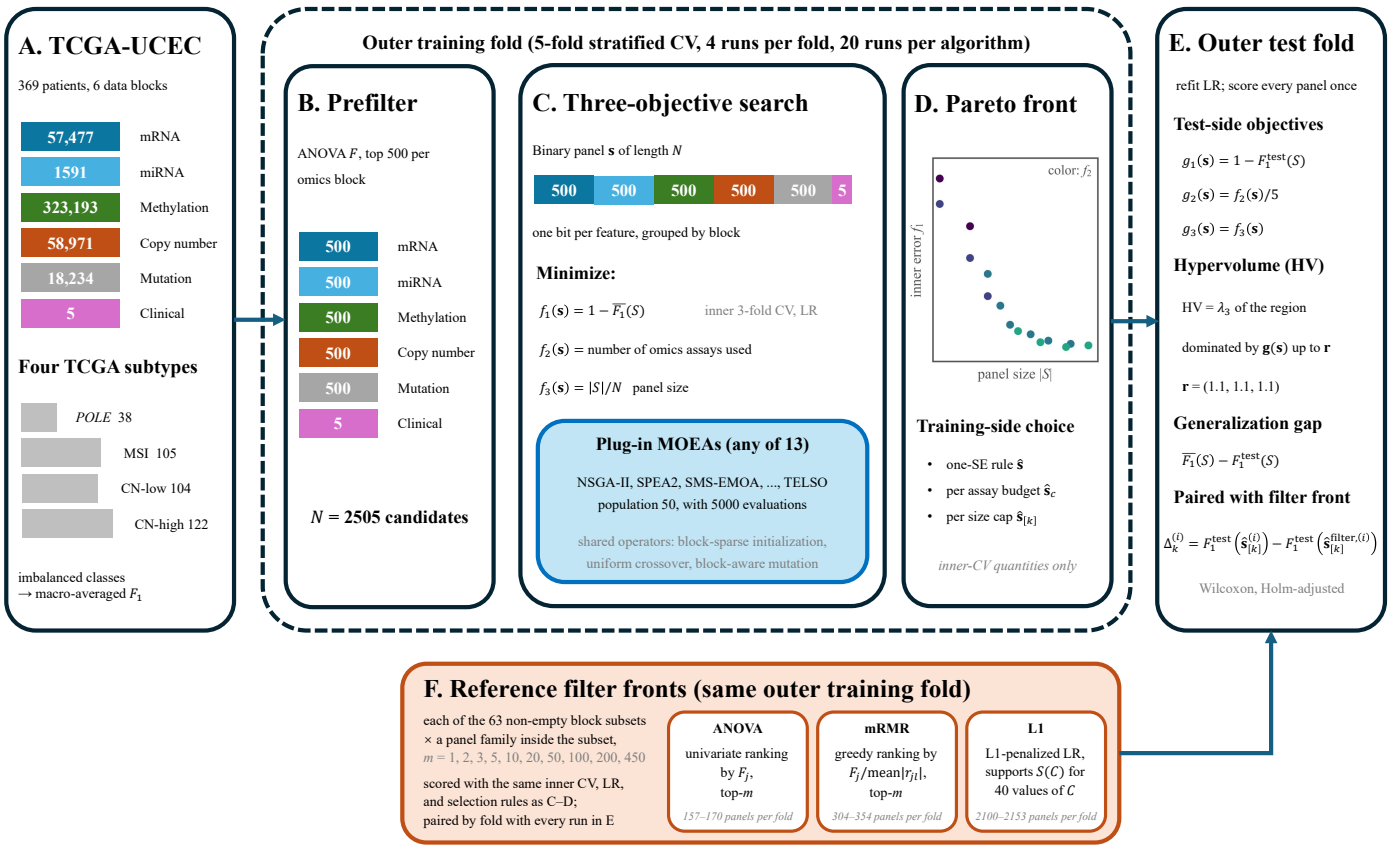

Figure 1. Overview of MO-POISE-X. (**A**) TCGA-UCEC cohort with six data blocks, with feature counts after label-free cleaning, and four imbalanced subtypes. (**B**) In-fold prefilter to $N = 2505$ candidate features. (**C**) Three-objective search over binary panels: any of the 13 MOEAs is plugged into the same problem, evaluation, and budget. (**D**) Training-side choice of panels from the final front. (**E**) Outer-test evaluation: test-side objectives, hypervolume, generalization gap, and paired comparison with the filter fronts. (**F**) Reference filter fronts built from every block subset on the same outer training fold and scored with the same inner cross-validation and selection rules. CV, cross-validation; LR, logistic regression; one-SE, one-standard-error; CN-low and CN-high, copy-number-low and copy-number-high.

2. Related Work

2.1. Molecular Classification of Endometrial Carcinoma

The four TCGA classes [1] have moved from a research finding into routine care: they are part of the ESGO/ESTRO/ESP guidelines [5] and of the 2023 staging of the International Federation of Gynecology and Obstetrics (FIGO) [17], and molecular classification predicts the benefit of adjuvant therapy [4]. In practice, the classes are assigned by a hierarchy of three assays: sequencing of the *POLE* exonuclease domain and immunohistochemistry for mismatch-repair proteins and p53, as in the Proactive Molecular Risk Classifier for Endometrial Cancer (ProMisE) [2,3]. The number of assays is a recognized barrier to adoption, and a single next-generation-sequencing panel has recently been shown to reproduce the three-assay algorithm with a Cohen's κ of 0.94 [12]. Reducing the number of assays has thus been a matter of expert design, guided by the biology of the classes, rather than a quantity that is optimized from data.

Computational work on the classes mostly predicts them from hematoxylin and eosin slides. Deep networks trained on slides from TCGA and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) predict the four classes and common mutations [18], but they do not indicate which molecular measurements carry the signal. Classifiers from

molecular data fix the assays and the number of features in advance. EMMA-STRAT combines mRNA, miRNA, and DNA-methylation data of TCGA-UCEC with 50 lasso-selected features per layer, without mutation data; its multi-omics models were more accurate than single-omics ones and were validated on CPTAC cohorts [19,20]. None of these studies searched jointly over assays and features to determine how few assays, and which, suffice to reconstruct the classes.

2.2. Multi-Objective Evolutionary Feature Selection

Feature selection trades classification error against subset size, two goals that conflict [6,7]. Single-objective metaheuristic wrappers combine them into a weighted sum [10,11], whereas multi-objective formulations approximate the front of non-dominated subsets, which can contain subsets that are both smaller and more accurate than those of single-objective searches [13]. Multi-objective evolutionary feature selection has been reviewed systematically [14,15]. The MOEAs used for it include dominance-based algorithms such as the non-dominated sorting genetic algorithm II (NSGA-II), the strength Pareto evolutionary algorithm 2 (SPEA2), and the Omni-optimizer [21–23], decomposition- and reference-direction-based algorithms such as the MOEA based on decomposition (MOEA/D) [24], NSGA-III [25], unified NSGA-III (U-NSGA-III) [26], and the reference vector guided evolutionary algorithm (RVEA) [27], the indicator-based S-metric selection evolutionary multi-objective optimization algorithm (SMS-EMOA) [28], and the adaptive geometry estimation-based MOEA (AGE-MOEA) and AGE-MOEA-II [29,30].

Feature selection on omics data belongs to the sparse large-scale multi-objective problems, in which most decision variables of Pareto-optimal solutions are zero. SparseEA introduced a two-layer encoding of a real vector and a binary mask, with variable scores estimated before the search, together with the sparse multi-objective optimization problem (SMOP) benchmark suite [16]. Several algorithms continue this line: the two-layer encoding learning swarm optimizer (TELSO) learns masks from frequent itemsets of good solutions [31], and the adjoint feature-selection-based evolutionary algorithm (AFSEA) adapts its mask operators with Relief-based importance [32]. General-purpose algorithms include the adaptive normal-vector guided MOEA (NRV-MOEA) [33], the multi-objective besiege-and-conquer algorithm (MOBCA) [34], and the inverse modeling constrained MOEA/D (IM-C-MOEA/D) with Gaussian-process inverse models [35]. These algorithms were developed on synthetic benchmarks and engineering problems with cheap objectives and large budgets, and several begin by scoring each variable separately. A wrapper objective is different: each evaluation is a cross-validation, and the estimate is noisy, so that the best subsets of a search are optimistically estimated [36], the same selection bias that affects gene selection and model selection outside nested cross-validation [9,37]. The initial population has also been adapted to feature selection: Xue et al. combined small and large initial feature subsets in particle swarm optimization [38], particle swarm optimization with information gain ratio-based subfeature grouping (ISPSO) partitions the features into groups by their information gain ratio to build the initial population [39], and SparseEA generates sparse initial masks guided by its variable scores [16]. These strategies control the number or the relevance of the initially selected features, not the data sources from which the features come.

2.3. Multi-Objective and Cost-Aware Selection of Multi-Omics Panels

At the intersection of the two fields, filter and embedded methods remain strong references. mRMR adds a redundancy penalty to univariate relevance [40], the lasso selects features jointly [41], and a benchmark on 15 cancer multi-omics data sets found mRMR among the best selectors, already strong at small panel sizes [42]. Methods for block-

structured data, such as lasso variants with a penalty factor per omics block [43], return one model per tuning parameter that usually uses every block, not a front over the number of blocks. Multi-objective evolutionary selection on omics data has been developed most consistently by Cattelani et al., who benchmarked seven multi-objective selectors on cancer transcriptomes with a hypervolume measure for cross-validation [44], designed three- and four-objective panels for TCGA renal-carcinoma subtypes with panel size as the proxy of clinical cost [45], and corrected the overestimate of wrapper scores during the search [46]. The multi-objective genetic algorithm for multi-view feature selection (MMFS-GA) selects features within and across omics views of TCGA data, without an objective at the level of views [47].

Measurement cost has entered feature selection in two forms. Cost-based multi-objective selection makes the total cost of the selected features an objective, as a sum of per-feature costs [48], and cost-aware clinical prediction trades accuracy against acquisition cost with features grouped under one cost per test [49]. In multi-omics, the dominant cost is of the second kind: an assay is paid once, however many of its features a panel uses. Our assay cost f_2 in Equation (3) applies this group-cost model with one cost per omics block and makes it a separate objective of the search, so that a front shows the attainable accuracy for every number of assays. To our knowledge, no previous study has made the number of omics assays an explicit objective of a multi-objective search for multi-omics panels, initialized such a search by sampling the assays of each panel before its features, applied it to the TCGA molecular subtypes of endometrial carcinoma, or compared a large set of MOEAs on this kind of problem against univariate, redundancy-aware, and sparse-regression filter fronts under leakage-controlled nested cross-validation.

3. Materials and Methods

Figure 1 gives an overview of MO-POISE-X. Within each outer training fold, the candidate pool is prefiltered, a multi-objective evolutionary algorithm searches panels under three objectives, and panels are chosen from the resulting front by rules that use inner cross-validation only. Every panel of the front is then scored once on the outer test fold and compared with three filter fronts built on the same training fold. Sections 3.2–3.8 describe these steps in turn, and Sections 3.9–3.11 the additional analyses.

3.1. Cohort and Data

Table 1 summarizes the data. Five omics blocks and the clinical table of TCGA-UCEC were obtained from the Genomic Data Commons (GDC) hub of the University of California, Santa Cruz (UCSC) Xena browser [50], and the four TCGA molecular subtypes [1] from the PanCancer Atlas endometrial study in cBioPortal [51]. One primary-tumor sample per patient was used, and only patients with all six blocks and a subtype label were kept, which gave 369 patients. Mutations were coded as a binary gene-by-patient matrix of non-silent variants. Omics features with missing values or constant across the cohort were removed before any analysis; this step used no labels and left 459,471 features.

Table 1. Data of the 369 TCGA-UCEC patients. Features: number of features after removing those with missing values or constant across the cohort. Candidates: features kept in each outer training fold by the in-fold prefilter of Section 3.2. FIGO stage is coded ordinally, and histology is one-hot over three types. All blocks are from the GDC hub of UCSC Xena; the subtype labels, 38 *POLE*, 105 MSI, 104 copy-number-low, and 122 copy-number-high, are from cBioPortal. STAR, Spliced Transcripts Alignment to a Reference; FPKM, fragments per kilobase of transcript per million mapped reads; RPM, reads per million mapped reads; ASCAT3, allele-specific copy-number analysis of tumors, version 3.

Data Block	Assay	Representation	Features	Candidates
mRNA expression	RNA sequencing, STAR	$\log_2(\text{FPKM}+1)$ per gene	57,477	500
miRNA expression	miRNA sequencing	$\log_2(\text{RPM}+1)$ per miRNA	1591	500
DNA methylation	Illumina 450K array	beta value per probe	323,193	500
Copy number	genotyping array, ASCAT3, gene level	copy number per gene	58,971	500
Somatic mutation	whole-exome sequencing	non-silent variant per gene, binary	18,234	500
Clinical	clinical record	age, FIGO stage, histology	5	5
Total			459,471	2505

3.2. Leakage-Controlled Nested Cross-Validation

Performance was estimated by five-fold stratified cross-validation with one fixed outer partition. Every data-dependent step was fitted on the outer training fold only, as required by nested cross-validation to avoid selection bias [9,37]: the prefilter, standardization, the multi-objective search, and the choice of the reported panel. Within each outer training fold, every omics block was reduced to its 500 features with the largest one-way ANOVA F statistic against the subtype label; in the mutation block, genes mutated in fewer than 3% of the training patients were discarded first. The five clinical variables were added without filtering, and missing ages were imputed with the training-fold median. For feature j and the n training patients in $K = 4$ classes, the F statistic is

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

where n_k is the number of training patients of class k, and $\bar{x}_{jk}$ and $\bar{x}_j$ are the class and overall means of feature j. The resulting candidate pool contains $N = 2505$ features. The limit of 500 candidates per block bounds the dimension of the search space and does not cut into weak features: in every outer training fold, the feature ranked 500th had an ANOVA p-value below 10^{-12} in the mRNA, methylation, copy-number, and mutation blocks, in each of which thousands of features passed a Bonferroni threshold of 0.05, and 134 to 187 miRNAs passed this threshold. Section 3.9 repeats the search with 250 and 1000 candidates per block. The prefilter uses the whole outer training fold, so the patients of the inner validation folds take part in the choice of the candidates. This shared step makes the inner estimate of a panel optimistic and contributes to the generalization gap reported in Section 4.8. It does not involve the outer test fold, so the outer-test estimates are unaffected. Refitting the prefilter inside every inner training fold would give the inner folds different candidate pools, so that a panel would no longer denote the same features in all inner folds of its own evaluation. Run i of every algorithm uses the same outer fold, the same inner partition, and the same seed, so that the runs of different algorithms are paired. Each algorithm was run 20 times, four times on each outer fold, with seed $2026 + 98i$ for run $i = 0, \ldots, 19$.

3.3. Three-Objective Problem Formulation

A candidate panel is a binary vector $\mathbf{s} \in \{0,1\}^N$, and S denotes the set of features it selects. Three objectives are minimized:

$$f_1(\mathbf{s}) = 1 - \overline{F_1}(S), \tag{2}$$

$$f_2(\mathbf{s}) = \sum_{b \in \mathcal{B}} c_b \, \mathbf{1}[\, S \cap B_b \neq \emptyset \,], \tag{3}$$

$$f_3(\mathbf{s}) = |S|/N. \tag{4}$$

Here $\overline{F_1}(S)$ is the macro-averaged F_1 score of a logistic-regression classifier with the default L2 penalty of scikit-learn [52], trained on the features S, averaged over a stratified three-fold cross-validation inside the outer training fold,

$$\overline{F_1}(S) = \frac{1}{V} \sum_{v=1}^{V} F_1^{(v)}(S), \qquad F_1^{(v)}(S) = \frac{1}{K} \sum_{k=1}^{K} \frac{2\,\mathrm{TP}_k^{(v)}}{2\,\mathrm{TP}_k^{(v)} + \mathrm{FP}_k^{(v)} + \mathrm{FN}_k^{(v)}}, \tag{5}$$

where $V = 3$, and $\mathrm{TP}_k^{(v)}$, $\mathrm{FP}_k^{(v)}$, and $\mathrm{FN}_k^{(v)}$ are counted for class k on validation fold v. Its standard error,

$$\sigma(S) = \frac{1}{\sqrt{V}} \left(\frac{1}{V-1} \sum_{v=1}^{V} \left(F_1^{(v)}(S) - \overline{F_1}(S) \right)^2 \right)^{1/2}, \tag{6}$$

is used by the selection rules of Section 3.6. In each inner fold, a z-score standardization is fitted on the inner training part only. The inner partition is drawn once per run and kept fixed, so that f_1 is a deterministic function of $\mathbf{s}$ within a run. $\mathcal{B}$ is the set of data blocks, B_b the candidate features of block b, and c_b the cost of acquiring block b; we set $c_b = 1$ for each of the five omics assays and $c_b = 0$ for the clinical variables, which are recorded in routine care. The assay cost f_2 therefore counts the number of omics assays a panel requires and ranges from 0 to 5. Equal costs were chosen because the number of assays determines the tissue consumed, the number of laboratory workflows, and the turnaround time, whereas the prices of the assays differ between laboratories, countries, and years. Section 3.9 repeats the search with assay-specific costs. We write $\mathcal{A}(\mathbf{s}) = \{b \in \mathcal{B} : S \cap B_b \neq \emptyset\}$ for the active blocks of a panel, so that $f_2(\mathbf{s}) = \sum_{b \in \mathcal{A}(\mathbf{s})} c_b$. Macro-averaged F_1 weights the four subtypes equally, so that the rare *POLE* class is not traded away for overall accuracy. The classifier is fixed rather than searched, so that the model evaluated on the outer test fold is always the model whose inner score was optimized. The empty panel is not a valid solution; depending on the algorithm, it is repaired by switching on one random feature or assigned the worst objective vector $(1, 5, 1)$.

3.4. Problem-Specific Search Operators

Three operators were shared by every algorithm that did not prescribe its own initialization and variation. First, we propose a block-sparse initialization for the block-structured search space: each initial individual uses a random, non-empty subset of the six blocks, and inside each chosen block every feature is selected with a probability drawn log-uniformly from $[0.005, 0.2]$,

$$q \sim \mathcal{U}\{1, \ldots, |\mathcal{B}|\}, \qquad A \sim \mathcal{U}\{A' \subseteq \mathcal{B} : |A'| = q\}, \qquad \ln \rho_b \sim \mathcal{U}(\ln 0.005, \ln 0.2),$$

$$s_j \sim \begin{cases} \mathrm{Bernoulli}(\rho_b), & j \in B_b,\ b \in A, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The initial population thus covers every number of assays from the start and contains compact panels for each of them. Uniform sampling with probability 0.5 would place every initial individual near $N/2$ features with all assays active, i.e., in a single corner of the objective space. A sparse sampling without the block step would still activate almost every omics assay: with 500 candidates per block, even the lowest density of 0.005 leaves a block empty with a probability of only $0.995^{500} \approx 0.08$. Second, uniform crossover was applied with probability 0.9. Third, block-aware mutation was applied: bit-flip mutation with rate $1/N$ followed, with probability $p_{\text{drop}} = 0.1$, by the removal of one whole active block,

$$\begin{aligned} s'_j = s_j \oplus z_j, \qquad z_j \sim \text{Bernoulli}(1/N), \\ \text{with probability } p_{\text{drop}}: \quad s'_j \leftarrow 0 \ \ \forall j \in B_{b^*}, \qquad b^* \sim \mathcal{U}\big(\mathcal{A}(\mathbf{s}')\big), \end{aligned} \tag{8}$$

where the block removal is skipped if $|\mathcal{A}(\mathbf{s}')| = 1$, so that a panel never becomes empty. A block removal lowers f_2 by c_{b^*} in a single step. Without the second step, a lower assay cost requires that every selected feature of a block be removed, by bit flips or by crossover with a parent that lacks the block, which rarely happens. The real-coded MOBCA and IM-C-MOEA/D receive the problem as $\mathbf{x} \in [0,1]^N$ with feature j selected if $x_j > 0.5$; their initial population is drawn by the block-sparse initialization and mapped to $(0.5, 1]$ for selected and $[0, 0.5)$ for unselected features.

3.5. Multi-Objective Evolutionary Algorithms

We selected 13 multi-objective evolutionary algorithms, which are listed in Table 2: NSGA-II [21], NSGA-III [25], U-NSGA-III [26], SPEA2 [22], RVEA [27], SMS-EMOA [28], AGE-MOEA [29], AGE-MOEA-II [30], the Omni-optimizer [23], NRV-MOEA [33], MOBCA [34], TELSO [31], and IM-C-MOEA/D [35]. Each was plugged into MO-POISE-X in place of X; for example, MO-POISE-NSGA-II denotes the instance with NSGA-II. The sparse-specific TELSO was run in its native binary mode with its own initialization and operators. The real-coded MOBCA and IM-C-MOEA/D used the shared block-sparse initialization and kept their own search operators on the threshold-decoded real encoding. All other algorithms used the operators of Section 3.4. Selection mechanisms were unchanged, duplicates were eliminated where the algorithm supports it, and all other parameters were left at their defaults; reference directions for NSGA-III, U-NSGA-III, and RVEA were 50 Riesz s-energy directions in three dimensions.

Table 2. The 13 multi-objective evolutionary algorithms compared and their parameter configurations. All algorithms used 5000 evaluations and a population of 50, except TELSO and IM-C-MOEA/D with 45. Shared: the operators of Section 3.4, i.e., block-sparse initialization, uniform crossover with $p_c = 0.9$, and bit-flip mutation with rate $1/N$ and block removal with $p_{\text{drop}} = 0.1$. Reference directions of NSGA-III, U-NSGA-III, and RVEA: Riesz s-energy directions; weight vectors of TELSO and IM-C-MOEA/D: uniformly spaced. δ: relative margin of the ϵ-dominance of the Omni-optimizer; α and f_r: rate of change of the penalty and frequency of reference-vector adaptation of RVEA; t and $t_{\max}$: current and maximum generation. APD, angle-penalized distance; BCB, besiege-and-conquer balance. Unlisted parameters are at the defaults of the reference implementation.

Algorithm	Year	Parameter Configurations
SPEA2 [22]	2001	shared operators; archive size 50
NSGA-II [21]	2002	shared operators
SMS-EMOA [28]	2007	shared operators; 50 offspring per generation
Omni-optimizer [23]	2008	shared operators; $\delta = 0.001$
NSGA-III [25]	2014	shared operators; 50 reference directions
U-NSGA-III [26]	2016	shared operators; 50 reference directions
RVEA [27]	2016	shared operators; 50 reference vectors; $\alpha = 2$; $f_r = 0.1$
AGE-MOEA [29]	2019	shared operators
AGE-MOEA-II [30]	2022	shared operators
NRV-MOEA [33]	2024	shared operators; normal-vector scale updated every 500 evaluations
MOBCA [34]	2024	shared initialization; real-coded, threshold 0.5; BCB = 0.2; 3 soldiers per army; 10 grid divisions
TELSO [31]	2024	own binary operators; single-variable scoring; APD penalty $(t/t_{\max})^2$
IM-C-MOEA/D [35]	2024	shared initialization; real-coded, threshold 0.5; 10 clusters; 3 resampled variables

All algorithms used a population of 50 and a budget of 5000 evaluations of the objective vector per run; TELSO and IM-C-MOEA/D round the population to 45, the largest number of uniformly spaced weight vectors in three objectives that does not exceed 50, as their reference implementations do.

3.6. Panel Selection from the Pareto Front

A run returns a front of panels, but a clinical or downstream application needs one panel, or one panel per assay budget. All rules were applied to the inner cross-validation estimates only, so that the panels they return can be evaluated on the outer test fold without selection bias. The one-standard-error rule [53] takes the panels whose inner error is within one standard error $\sigma(S^*)$ of the lowest inner error, where S^* is the feature set of the panel $\mathbf{s}^*$ with the lowest inner error, and among them the one with the lowest assay cost, then the smallest size, then the lowest error. For a set of panels $\mathcal{F}$,

$$\mathbf{s}^* = \arg\min_{\mathbf{s}\in\mathcal{F}} f_1(\mathbf{s}), \qquad \mathcal{F}_{\text{SE}} = \{\mathbf{s} \in \mathcal{F} : f_1(\mathbf{s}) \le f_1(\mathbf{s}^*) + \sigma(S^*)\}, \tag{9}$$

$$R(\mathcal{F}) = \operatorname{lex}\min_{\mathbf{s}\in\mathcal{F}_{\text{SE}}} \big(f_2(\mathbf{s}),\ |S|,\ f_1(\mathbf{s})\big). \tag{10}$$

For the final front $\mathcal{F}$ of a run, $\hat{\mathbf{s}} = R(\mathcal{F})$ is the panel chosen by the one-standard-error rule. The per-budget rule applies the one-standard-error rule to the panels with assay cost at most c, and the size-capped rule to the panels with at most k features,

$$\begin{aligned} \hat{\mathbf{s}}_c &= R\big(\{\mathbf{s} \in \mathcal{F} : f_2(\mathbf{s}) \le c\}\big), & c &= 0, \ldots, 5, \\ \hat{\mathbf{s}}_{[k]} &= R\big(\{\mathbf{s} \in \mathcal{F} : |S| \le k\}\big), & k &\in \{5, 10, 20, 50, 100\}. \end{aligned} \tag{11}$$

Thus $\hat{\mathbf{s}}_c$ is the best panel that uses at most c omics assays, and $\hat{\mathbf{s}}_{[k]}$ the best panel of at most k features; here and below, k denotes a size cap and not a class.

3.7. Filter-Based Baselines

As references against which every search was judged, three filter fronts were built in each outer training fold. Each combines every one of the 63 non-empty subsets of the six blocks with a family of panels inside the subset, and the three families differ in how they account for the joint behavior of features. The ANOVA front ranks the features of the subset by their univariate F statistic and takes the top m, $m \in \{1, 2, 3, 5, 10, 20, 50, 100, 200, 450\}$. The mRMR front takes the first m features, for the same m, of a greedy minimum redundancy–maximum relevance ranking [40], with the F statistic of Equation (1) as relevance, the mean absolute Pearson correlation r_{jl} with the features already chosen as redundancy, and their quotient as the criterion: starting from $S_1 = \{\arg\max_j F_j\}$,

$$S_{t+1} = S_t \cup \left\{ \underset{j \in B_U \setminus S_t}{\arg\max} \frac{F_j}{\frac{1}{t}\sum_{l \in S_t} |r_{jl}|} \right\}, \tag{12}$$

where B_U denotes the candidate features of the block subset U. The L1 front consists of the supports of one-vs-rest L1-penalized logistic regressions [41] on the standardized features $\mathbf{z}_i$ of the subset, with labels $y_{ik} = 1$ if patient i belongs to class k and $y_{ik} = -1$ otherwise,

$$(\hat{\mathbf{w}}_k, \hat{w}_{k0})(C) = \underset{\mathbf{w}, w_0}{\arg\min} \ \|\mathbf{w}\|_1 + C \sum_{i=1}^{n} \log\left(1 + e^{-y_{ik}(\mathbf{w}^\top \mathbf{z}_i + w_0)}\right),$$
$$S(C) = \bigcup_{k=1}^{K} \left\{ j : \hat{w}_{kj}(C) \neq 0 \right\}, \tag{13}$$

along the grid $C_\ell = C_{\min} \cdot 10^{2\ell/39}$, $\ell = 0, \ldots, 39$, where $C_{\min}$ is the smallest value of C at which a feature is selected. Each front also contains the panel of all candidate features of every subset. Identical panels reached from different subsets were counted once, which left 157–170, 304–354, and 2100–2153 panels per fold on the ANOVA, mRMR, and L1 fronts. Each panel was scored with the same inner cross-validation procedure and the same logistic regression as the searched panels, with the inner partition of each fold drawn with seed 2026, and the selection rules of Section 3.6 were applied to it in the same way. Although L1-penalized regression is an embedded method, we refer to all three as filter fronts for brevity. Because the rankings and penalized fits use the whole outer training fold, the inner error of a baseline panel is optimistic, as Section 4.8 shows. The panel that contains all 2505 candidate features of all blocks is the full-pool logistic-regression model.

3.8. Evaluation and Statistical Analysis

Every panel of a run's final non-dominated set was refit on the whole outer training fold and scored once on the outer test fold. The primary end point is the test-side hypervolume of the run [54]: with $\mathbf{g}(\mathbf{s}) = \left(1 - F_1^{\text{test}}(S),\ f_2(\mathbf{s})/5,\ f_3(\mathbf{s})\right)$ and the non-dominated subset $\mathcal{F}^{\text{nd}}$ of the final front with respect to $\mathbf{g}$,

$$\mathrm{HV}(\mathcal{F}) = \lambda_3 \left(\bigcup_{\mathbf{s} \in \mathcal{F}^{\text{nd}}} [g_1(\mathbf{s}), r_1] \times [g_2(\mathbf{s}), r_2] \times [g_3(\mathbf{s}), r_3] \right), \qquad \mathbf{r} = (1.1, 1.1, 1.1), \tag{14}$$

where λ_3 is the Lebesgue measure. Its upper bound is $1.1^3 = 1.331$. For a filter front, $\mathcal{F}$ is the set of its panels that are non-dominated in the inner objectives, as the final front of a search is, so that the hypervolume of the searched and of the filter fronts is computed in the same way. Because panel size enters g_3 on a linear scale relative to N, the hypervolume hardly distinguishes a panel of five features from one of 50; compact panels are therefore compared by the size-capped rule. Secondary end points are the inner hypervolume,

computed in the same way from the inner errors, the outer-test macro-averaged F_1 of the one-standard-error panel, of the per-budget panels, and of the size-capped panels, and the paired difference of the latter to the filter fronts.

Algorithms were compared with the Friedman test [55], once with the 20 paired runs as the blocking factor and once with the five-fold means as the blocking factor, and with two-sided Wilcoxon signed-rank tests of every algorithm against the one with the highest mean, adjusted by Holm's method [56]. The fold-level test accounts for the four runs of a fold sharing their patients. Each algorithm was compared with each filter front by the paired difference between each run and the filter front of its fold, for the test-side hypervolume and for the outer-test macro-averaged F_1 of the panels chosen under each size cap,

$$\Delta_k^{(i)} = F_1^{\text{test}}\left(\hat{\mathbf{s}}_{[k]}^{(i)}\right) - F_1^{\text{test}}\left(\hat{\mathbf{s}}_{[k]}^{\text{filter},(i)}\right), \qquad i = 0, \ldots, 19, \tag{15}$$

where $\hat{\mathbf{s}}_{[k]}^{\text{filter},(i)}$ is the panel chosen under the same size cap from the filter front of outer fold $\lfloor i/4 \rfloor$, the fold of run i, and the 20 differences were tested with the Wilcoxon signed-rank test, with Holm's adjustment across the algorithms for each filter front and size cap. Because the four runs of a fold share their patients and are compared with the same filter panel, the 20 differences are not independent; we therefore also report on how many of the five folds the mean difference of an algorithm had the same sign. To quantify how optimistic the inner estimates are, we report the generalization gap $\overline{F_1}(S) - F_1^{\text{test}}(S)$ of every panel.

3.9. Sensitivity and Ablation Analyses

Four sensitivity analyses tested whether the comparison with the ANOVA filter front depends on the initial density, the evaluation budget, the size of the candidate pool, or the assay costs. In the first, the upper end of the initial per-block density was raised from 0.2 to 1.0. In the second, the budget was raised from 5000 to 20,000 evaluations. In the third, the prefilter kept 250 or 1000 candidates per omics block in place of 500, which gives $N = 1255$ and $N = 5005$; the three filter fronts were rebuilt from the same candidate pools. In the fourth, the equal assay costs were replaced by assay-specific costs, $c_b = 1.9$ for somatic mutation, 1.1 for mRNA, 0.9 for DNA methylation, 0.8 for miRNA, and 0.3 for copy number. These are illustrative relative costs that follow the order of the per-sample prices listed by academic core facilities (e.g., the price lists of the Technology Center for Genomics and Bioinformatics, University of California, Los Angeles, https://www.uclahealth.org/departments/pathology/research-services/technology-center-genomics-bioinformatics-tcgb/pricing, and of the Genomics Facility, University of Chicago, https://voices.uchicago.edu/genomicsfacility/pricing-new; both accessed on 11 November 2026), with whole-exome sequencing the most expensive and a genotyping array the least expensive assay, and they are scaled to the same total of 5 as the equal costs. The hypervolume of the ANOVA filter front was recomputed with these costs, which change f_2 but not the panels. An ablation compared the block-sparse initialization with two alternatives, all other operators unchanged: a flat sparse initialization, which draws one density per individual from the same log-uniform distribution and applies it to all N candidate features without the block step, and the uniform initialization, which selects every feature with probability 0.5. The comparison with the flat sparse initialization, which isolates the block step, was run for all ten competitive algorithms of Section 4.1. The other analyses were run for MO-POISE-SPEA2, MO-POISE-SMS-EMOA, and MO-POISE-Omni-optimizer, referred to below as the three selected algorithms. They are three of the four algorithms with the highest mean test-side hypervolume, 0.964 to 0.967 in Table 3, and they represent three different environmental-selection principles: a strength-and-density archive in SPEA2, an indicator-based steady-state selection in SMS-EMOA, and ϵ-dominance with

crowding in both objective and decision space in the Omni-optimizer. The fourth, MO-POISE-NSGA-II, shares non-dominated sorting with crowding in objective space with the Omni-optimizer, which extends it. Every setting used the same 20 runs as the main setting and was compared with it by two-sided Wilcoxon signed-rank tests over the paired runs, without adjustment for multiple comparisons unless stated otherwise.

Table 3. Results of the 13 algorithms over 20 runs, mean (SD), sorted by the test-side hypervolume. Hypervolume: larger is better, with an upper bound of 1.331. *p*: Holm-adjusted Wilcoxon signed-rank test against the algorithm with the highest mean in the same column, which is marked –. One-SE panel: the panel chosen by the one-standard-error rule; test macro-F_1, outer-test macro-averaged F_1; cost, mean number of omics assays; size, median number of features. Front: mean number of panels on the final non-dominated set, or, for the filter fronts, number of panels evaluated. Filter fronts: mean (SD) over the five folds.

	Hypervolume			One-SE Panel				
Method	**Test**	**Inner**	***p***	**Test Macro-F_1**	***p***	**Cost**	**Size**	**Front**
MO-POISE-SPEA2	0.967 (0.047)	1.029 (0.045)	–	0.824 (0.053)	0.136	3.20	14	31.4
MO-POISE-SMS-EMOA	0.965 (0.058)	1.028 (0.047)	1.000	0.844 (0.061)	0.736	3.05	16	30.9
MO-POISE-NSGA-II	0.965 (0.055)	1.019 (0.042)	1.000	0.819 (0.054)	0.155	3.70	30	37.9
MO-POISE-Omni-optimizer	0.964 (0.053)	1.029 (0.048)	1.000	0.841 (0.041)	0.736	3.25	16	28.9
MO-POISE-U-NSGA-III	0.957 (0.055)	1.023 (0.042)	1.000	0.832 (0.059)	0.492	2.85	13	10.0
MO-POISE-AGE-MOEA-II	0.956 (0.063)	1.029 (0.048)	1.000	0.822 (0.067)	0.254	3.10	12	27.7
MO-POISE-NRV-MOEA	0.956 (0.056)	1.022 (0.048)	1.000	0.820 (0.058)	0.104	2.95	11	25.9
MO-POISE-AGE-MOEA	0.951 (0.059)	1.017 (0.047)	0.557	0.823 (0.057)	0.417	3.05	12	28.6
MO-POISE-RVEA	0.948 (0.059)	0.997 (0.043)	0.362	0.822 (0.050)	0.237	2.95	15	13.6
MO-POISE-NSGA-III	0.946 (0.073)	1.020 (0.052)	0.929	0.832 (0.063)	0.736	2.85	11	8.7
MO-POISE-TELSO	0.925 (0.065)	0.938 (0.050)	0.362	0.764 (0.084)	0.010	1.85	193	4.8
MO-POISE-IM-C-MOEA/D	0.873 (0.049)	0.903 (0.046)	<0.001	0.817 (0.071)	0.417	4.10	71	13.3
MO-POISE-MOBCA	0.857 (0.071)	0.873 (0.073)	<0.001	0.854 (0.035)	–	5.00	165	27.1
ANOVA filter front	1.055 (0.067)	1.083 (0.012)		0.901 (0.043)		2.20	1000	163
mRMR filter front	1.066 (0.035)	1.110 (0.005)		0.875 (0.042)		2.20	200	334
L1 filter front	1.086 (0.052)	1.165 (0.009)		0.872 (0.063)		2.60	144	2129

3.10. Additional Analyses of the Chosen Panels

Three further analyses required no new search. First, the one-standard-error panel, the panel $\hat{\mathbf{s}}_{[5]}$ with at most five features, and the panel $\hat{\mathbf{s}}_2$ for at most two assays of every run and of every filter front were refitted on the outer training fold with two other classifiers, a random forest with 100 trees and a support vector machine with a radial basis function kernel, both at the defaults of scikit-learn, and scored on the outer test fold. For the panels refitted with logistic regression, we also report per-class precision and recall and Cohen's κ with the TCGA labels. Second, the stability of the selection across the outer folds was measured by the share of one-standard-error panels per fold that contain a given marker and by the mean pairwise Jaccard index of these panels within and between folds. Third, to examine how the choice of a panel depends on the noise of the inner estimate, every panel of the final front of every run of the ten competitive algorithms was scored again with the same logistic regression on five repetitions of a stratified three-fold partition of the outer training fold, drawn with a seed that differs from the partition used in the search, and the one-standard-error rule and the size-capped rule were applied to these re-scored inner errors.

3.11. Comparison with a Molecular Surrogate of the Clinical Classification

A molecular surrogate of the clinical classification was built from the same TCGA data, since immunohistochemistry for mismatch-repair proteins and p53, which the ProMisE classifier [2] and the clinical guidelines require, is not available for this cohort. It follows the integrated molecular classification of the ESGO/ESTRO/ESP guidelines [5], which assesses *POLE* before mismatch repair, unlike the original ProMisE classifier, and applies this hierarchy to molecular readouts: a pathogenic *POLE* exonuclease-domain hotspot mutation, one of P286R, V411L, S297F, A456P, S459F, F367S, L424I, M295R, P436R, M444K, and D368Y [57], assigns the *POLE* class; otherwise, *MLH1* promoter hypermethylation or a truncating mutation of *MLH1*, *MSH2*, *MSH6*, or *PMS2* assigns the MSI class; otherwise, a non-silent *TP53* mutation assigns the copy-number-high class; all remaining tumors are assigned the copy-number-low class. *MLH1* promoter methylation is the mean beta value of the probes in the promoter cytosine–phosphate–guanine (CpG) island shared by *MLH1* and *EPM2AIP1*, and its threshold was chosen on each outer training fold as the value that best separated MSI from other tumors by balanced accuracy. The surrogate uses the somatic-mutation and DNA-methylation assays and was scored on the same outer test folds.

In addition, the three readouts of the surrogate were added to the candidate pool as label-free features: *POLE* hotspot status and a truncating mismatch-repair mutation as binary features of the mutation block, and the mean beta value of the *MLH1* promoter probes as a feature of the methylation block. They were kept in every pool regardless of their ANOVA rank; the search was repeated for the three selected algorithms with the same 20 runs, and the three filter fronts were rebuilt from the same pools.

3.12. Implementation

MO-POISE-X was implemented in Python 3.14.6 with NumPy 2.5.3, SciPy 1.18.1, scikit-learn 1.9.1 [52], and pymoo 0.6.2 [58]. All runs were executed on a high-performance computing cluster, on compute nodes with two 32-core Intel Xeon Gold 6548Y+ processors (Intel Corporation, Santa Clara, CA, USA) at 2.5 GHz and 512 GB of memory under Rocky Linux 9. Each run used four threads, and eight runs shared one processor. The source code and all per-run results are available in the repository listed in the Data Availability Statement.

4. Results

4.1. Overall Performance Comparison

Table 3 and Figure 2 summarize the comparison. The test-side hypervolume differed between the algorithms in the Friedman test, with $p = 2.2 \times 10^{-10}$ over the 20 paired runs and $p = 3.9 \times 10^{-4}$ over the five-fold means. MO-POISE-SPEA2 had the highest mean, 0.967. After Holm adjustment, only MO-POISE-IM-C-MOEA/D and MO-POISE-MOBCA differed significantly from it, with means of 0.873 and 0.857 and $p < 0.001$; the other ten algorithms, with means between 0.925 and 0.965, had $p \geq 0.36$. The ranking by inner hypervolume was similar. We refer to the ten algorithms other than MO-POISE-TELSO, MO-POISE-IM-C-MOEA/D, and MO-POISE-MOBCA as the ten competitive algorithms; they are the ten algorithms with the highest mean test-side hypervolume and, as Section 4.2 shows, the ones that produced competitive compact panels. In practical terms, the differences among them were small. Their mean test-side hypervolumes spanned 0.021, from 0.946 to 0.967, and the mean outer-test macro-averaged F_1 of their one-standard-error panels spanned 0.025, from 0.819 to 0.844. Both ranges are smaller than the standard deviation over the runs of each of these algorithms, 0.047 to 0.073 for the hypervolume and 0.041 to

0.067 for the macro-averaged F_1, and the range of the macro-averaged F_1 is about the size of one inner standard error.

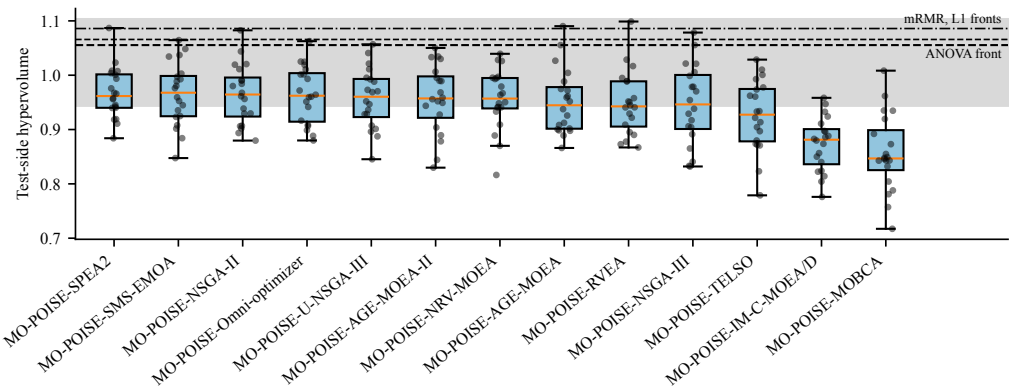

Figure 2. Test-side hypervolume of the 20 runs of each algorithm, sorted by the mean. Boxes: interquartile range with the median (orange line); whiskers: the most extreme runs within 1.5 times the interquartile range; dots: individual runs. The dashed line and the gray band show the mean and the range of the ANOVA filter front over the five folds; the thin dashed and dash-dotted lines above it show the means of the mRMR and L1 filter fronts.

Over the five folds, the ANOVA, mRMR, and L1 filter fronts reached mean test-side hypervolumes of 1.055, 1.066, and 1.086, with standard deviations (SDs) of 0.067, 0.035, and 0.052, higher than the mean of every algorithm. Sixteen of the 260 runs exceeded the ANOVA front of their fold, two exceeded the mRMR front, and none exceeded the L1 front. The mean paired differences to the ANOVA front ranged from -0.088 for MO-POISE-SPEA2 to -0.199 for MO-POISE-MOBCA, with $p < 0.001$ for every algorithm; for eight of the ten competitive algorithms, the mean difference was negative on all five folds, and for the other two on four. As the next subsection shows, these differences arise in the region of large panels.

4.2. Performance by Panel Size

Figure 3 shows where the hypervolume difference comes from. The fronts of most algorithms consist almost entirely of panels with fewer than about 30 features, whereas the filter fronts extend to panels of hundreds or thousands of features, whose outer-test macro-averaged F_1 approaches 0.9. Logistic regression on all 2505 candidate features reached a macro-averaged F_1 of 0.902. The panels chosen from the ANOVA, mRMR, and L1 fronts by the one-standard-error rule reached 0.901, 0.875, and 0.872, with median sizes of 1000, 200, and 144 features. The best panel on a run's front, identified with the outer test fold and therefore an upper bound, reached 0.852 to 0.892 on average for the algorithms other than MO-POISE-TELSO, MO-POISE-IM-C-MOEA/D, and MO-POISE-MOBCA, and 0.899 for MO-POISE-MOBCA. The panels chosen from the evolutionary fronts by the one-standard-error rule reached 0.764 to 0.854, as listed in Table 3.

For compact panels, the picture is reversed, as Table 4 and Figure 4 show. The grouping of the ten competitive algorithms in Section 4.1 rests on these compact panels; in hypervolume, MO-POISE-TELSO did not differ significantly from the best algorithm. When the choice was restricted to panels with at most k features, these ten algorithms were within 0.015 of the most accurate filter front for every $k \leq 20$ and within 0.025 at $k \leq 50$; none of them was significantly less accurate than the most accurate filter front at any $k \leq 50$. Their macro-averaged F_1 was 0.75 to 0.80 for $k \leq 5$, 0.81 to 0.83 for $k \leq 10$, and 0.82 to 0.84 for $k \leq 20$ and $k \leq 50$. Against the ANOVA front, the gain was 0.07 to 0.13 for $k \leq 5$ and about 0.10 for $k \leq 10$ and $k \leq 20$, significant after Holm adjustment for all ten algorithms at every $k \leq 50$; the mean difference

was positive on all five folds for every algorithm up to $k \leq 20$, and on four of the five folds for nine algorithms at $k \leq 50$; with at most five features, for example, MO-POISE-NSGA-II reached 0.804 and the ANOVA front 0.674. Against the L1 front, it was 0.06 to 0.12 for $k \leq 5$, 0.07 to 0.08 for $k \leq 10$, and 0.02 to 0.05 for $k \leq 20$, significant for all ten algorithms at $k \leq 5$, for six at $k \leq 10$, and for four at $k \leq 20$; it was positive on all five folds for every algorithm at $k \leq 5$ and on four of the five folds at $k \leq 10$. At $k \leq 50$, the L1 panels caught up, with 0.843. The mRMR front was the strongest filter for small panels. With at most five features, it reached 0.752, and the evolutionary panels exceeded it by up to 0.05 on average, but no difference was significant after Holm adjustment, the smallest adjusted p being 0.19. For $10 \leq k \leq 50$, the mRMR panels reached 0.82 to 0.83 and agreed with the evolutionary panels within 0.02. At $k \leq 100$, the mRMR and L1 panels reached 0.849 and 0.866 and became more accurate than the evolutionary ones, whose fronts contain few panels of that size; against the L1 front, the difference was significant for seven of the ten competitive algorithms. MO-POISE-TELSO, MO-POISE-IM-C-MOEA/D, and MO-POISE-MOBCA did not produce competitive compact panels.

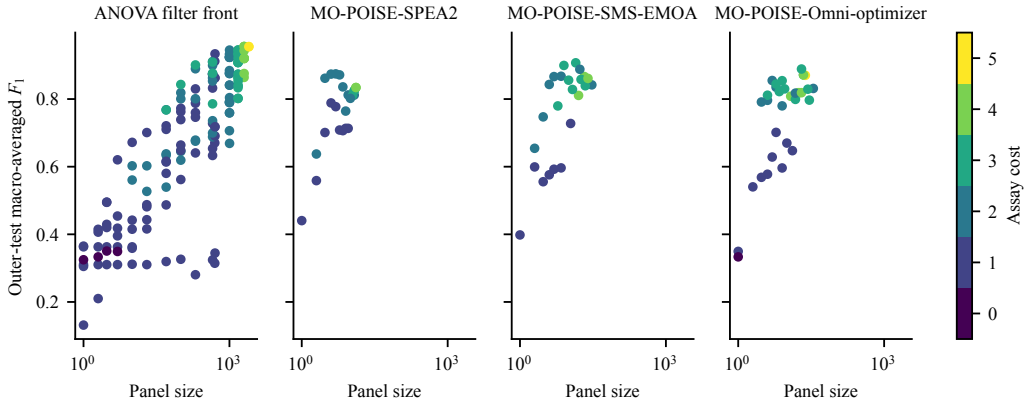

Figure 3. Outer-test macro-averaged F_1 against panel size for the ANOVA filter front and for the final fronts of the three selected algorithms of Section 3.9, MO-POISE-SPEA2, MO-POISE-SMS-EMOA, and MO-POISE-Omni-optimizer, in run 0 on outer fold 0, colored by the assay cost f_2, i.e., the number of omics assays the panel uses; panel size on a logarithmic scale.

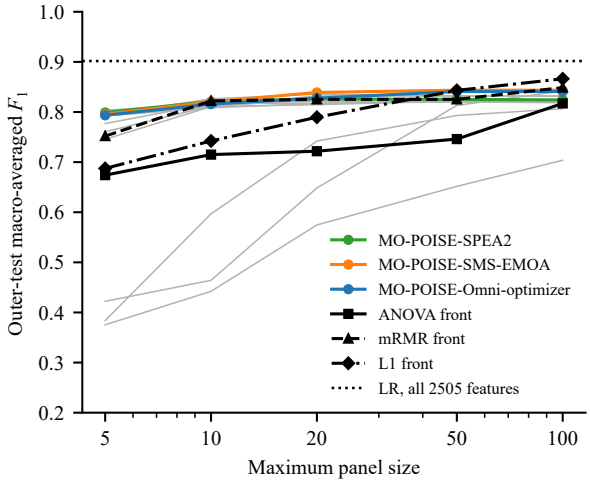

Figure 4. Mean outer-test macro-averaged F_1 over the runs of the panel chosen by the one-standard-error rule among the panels with at most k features. Colored lines: the three selected algorithms of Section 3.9; gray lines: the other algorithms; black lines: the three filter fronts; dotted line, labeled LR: logistic regression on all 2505 candidate features.

Table 4. Outer-test macro-averaged F_1 of the panel chosen by the one-standard-error rule among the panels with at most k features, and of the panel chosen by the per-budget rule for at most two omics assays, as means over the 20 runs of each algorithm and over the five folds of each filter front. Values in parentheses: Wilcoxon signed-rank p of the paired difference to the mRMR filter front of the same fold, Holm-adjusted across the algorithms for each k. For MO-POISE-IM-C-MOEA/D and MO-POISE-MOBCA, whose fronts lacked such panels in some runs, the means are over 12, 16, and 20 and over 12, 17, and 19 runs for $k \leq 5$, 10, and 20. Logistic regression on all 2505 candidate features reached 0.902.

Method	$k \leq 5$	$k \leq 10$	$k \leq 20$	$k \leq 50$	$k \leq 100$	2 Assays
MO-POISE-SPEA2	0.799 (0.455)	0.821 (1.000)	0.826 (1.000)	0.825 (1.000)	0.824 (0.193)	0.814
MO-POISE-SMS-EMOA	0.796 (0.352)	0.820 (1.000)	0.839 (0.897)	0.844 (1.000)	0.844 (1.000)	0.814
MO-POISE-NSGA-II	0.804 (0.192)	0.817 (1.000)	0.818 (1.000)	0.820 (1.000)	0.819 (0.038)	0.824
MO-POISE-Omni-optimizer	0.793 (0.352)	0.816 (1.000)	0.828 (1.000)	0.841 (1.000)	0.841 (1.000)	0.826
MO-POISE-U-NSGA-III	0.795 (0.266)	0.826 (1.000)	0.836 (1.000)	0.832 (1.000)	0.832 (1.000)	0.823
MO-POISE-AGE-MOEA-II	0.792 (0.455)	0.808 (1.000)	0.817 (1.000)	0.822 (1.000)	0.822 (0.207)	0.808
MO-POISE-NRV-MOEA	0.777 (0.832)	0.817 (1.000)	0.821 (1.000)	0.820 (1.000)	0.820 (0.193)	0.813
MO-POISE-AGE-MOEA	0.792 (0.455)	0.810 (1.000)	0.827 (1.000)	0.823 (1.000)	0.823 (0.350)	0.809
MO-POISE-RVEA	0.746 (1.000)	0.811 (1.000)	0.815 (1.000)	0.822 (1.000)	0.822 (0.172)	0.804
MO-POISE-NSGA-III	0.761 (0.832)	0.812 (1.000)	0.832 (1.000)	0.832 (1.000)	0.832 (1.000)	0.817
MO-POISE-TELSO	0.375 (0.001)	0.442 (0.001)	0.575 (<0.001)	0.652 (<0.001)	0.704 (<0.001)	0.729
MO-POISE-IM-C-MOEA/D	0.384 (0.006)	0.597 (0.003)	0.742 (0.008)	0.793 (0.765)	0.807 (0.046)	0.769
MO-POISE-MOBCA	0.422 (0.006)	0.464 (<0.001)	0.648 (<0.001)	0.813 (1.000)	0.842 (1.000)	0.752
ANOVA filter front	0.674	0.715	0.722	0.746	0.817	0.874
mRMR filter front	0.752	0.822	0.825	0.825	0.849	0.886
L1 filter front	0.687	0.742	0.790	0.843	0.866	0.879

Figure S1 shows the paired differences of Equation (15) for every algorithm and size cap. Against the ANOVA and L1 fronts, the ten competitive algorithms form a uniform block of positive differences up to $k \leq 20$, and against the ANOVA front up to $k \leq 50$, whereas against the mRMR front the differences are close to zero except for a small, non-significant positive column at $k \leq 5$.

As Figure 5 shows, the panels of at most five features also differ in the assays they draw on. The ten competitive algorithms combined DNA methylation and somatic mutation in 153 of their 200 runs, or 77%, and added a copy-number feature to this pair in another 17 runs. The ANOVA panels of this size consisted of mRNA features alone in every fold, the L1 panels of mutation features alone in three of the five folds, and the mRMR panels of mutation features alone in two folds and of methylation and mutation features in three.

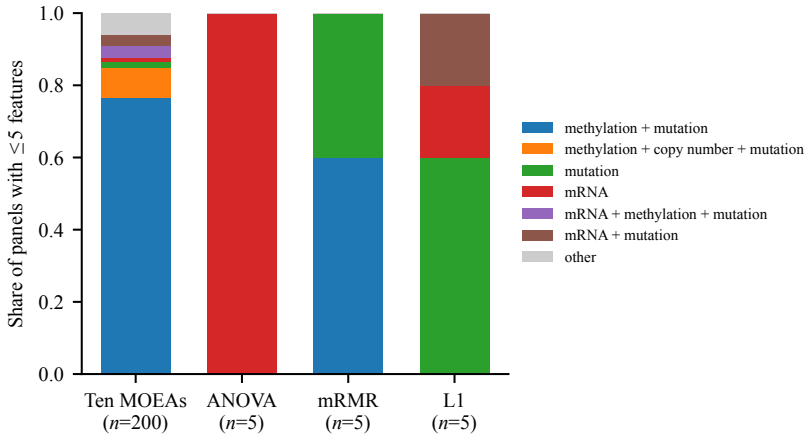

Figure 5. Data blocks used by the panel $\hat{\mathbf{s}}_{[5]}$ of Equation (11), chosen with at most five features, for the 200 runs of the ten competitive algorithms and for the five folds of each filter front. The six most frequent block combinations are shown; the remaining ones are pooled as "other".

4.3. Ablation of the Block-Sparse Initialization

Table 5 compares the block-sparse initialization with the flat sparse initialization, which draws the features without the block step, for all ten competitive algorithms. Without the block step, the test-side hypervolume fell for every one of them, from 0.946–0.967 to 0.880–0.926, significantly for nine, also after Holm adjustment across the ten algorithms; for MO-POISE-AGE-MOEA, the decrease from 0.951 to 0.926 had $p = 0.053$. The one-standard-error panels used more assays for all ten algorithms, 3.10 to 4.20 on average instead of 2.85 to 3.70, with an unadjusted $p < 0.05$ for four of them and no significant difference after Holm adjustment, and their outer-test macro-averaged F_1 did not differ significantly for any algorithm.

Table 5. Block-sparse versus flat sparse initialization for the ten competitive algorithms, as means over the same 20 runs. Assays: mean number of omics assays of the panel chosen by the one-standard-error rule; one-SE macro-F_1: outer-test macro-averaged F_1 of this panel. p: Wilcoxon signed-rank test of the paired difference, unadjusted; after Holm adjustment across the ten algorithms, nine of the differences in test HV and none of the differences in assays remain significant. The comparison with the uniform initialization is given in Table S1.

	Test HV			Assays			One-SE Macro-F_1		
Method	**Block**	**Flat**	p	**Block**	**Flat**	p	**Block**	**Flat**	p
MO-POISE-SPEA2	0.967	0.915	0.003	3.20	3.95	0.010	0.824	0.837	0.349
MO-POISE-SMS-EMOA	0.965	0.914	0.001	3.05	3.80	0.017	0.844	0.826	0.070
MO-POISE-NSGA-II	0.965	0.913	<0.001	3.70	4.20	0.119	0.819	0.826	0.546
MO-POISE-Omni-optimizer	0.964	0.920	0.014	3.25	3.80	0.064	0.841	0.825	0.261
MO-POISE-U-NSGA-III	0.957	0.902	<0.001	2.85	3.10	0.376	0.832	0.825	0.475
MO-POISE-AGE-MOEA-II	0.956	0.906	0.003	3.10	3.35	0.260	0.822	0.831	0.388
MO-POISE-NRV-MOEA	0.956	0.915	0.005	2.95	3.50	0.057	0.820	0.822	0.898
MO-POISE-AGE-MOEA	0.951	0.926	0.053	3.05	3.70	0.018	0.823	0.836	0.571
MO-POISE-RVEA	0.948	0.900	0.008	2.95	3.30	0.176	0.822	0.827	0.648
MO-POISE-NSGA-III	0.946	0.880	0.002	2.85	3.70	0.016	0.832	0.818	0.330

Table S1 adds the comparison with the uniform initialization for MO-POISE-SPEA2, MO-POISE-SMS-EMOA, and MO-POISE-Omni-optimizer. The uniform initialization had no consistent effect on the test-side hypervolume. It raised it significantly for MO-POISE-SPEA2, to 1.010, lowered it significantly for MO-POISE-SMS-EMOA, to 0.897, and did not change it significantly for MO-POISE-Omni-optimizer, at 0.962. Its one-standard-error panels used more assays for all three algorithms, 3.95 to 4.10 on average, and many more features, with medians of 70 to 143 instead of 14 to 16. They were more accurate for MO-POISE-SPEA2 and MO-POISE-Omni-optimizer, with 0.858 and 0.861 instead of 0.824 and 0.841, and not significantly different for MO-POISE-SMS-EMOA. The fronts obtained with the uniform initialization lacked panels of at most five features in 1, 7, and 3 of the 20 runs of the three algorithms. Where such panels existed, they reached a macro-averaged F_1 of 0.579 to 0.760, compared with 0.793 to 0.799 after the block-sparse initialization, a difference that was significant for MO-POISE-SMS-EMOA. The uniform initialization also doubled the search time, to 2.5 to 2.7 min per run, because larger panels take longer to evaluate.

4.4. Performance by Number of Assays

Figure 6 shows the outer-test macro-averaged F_1 of the panel chosen for each assay budget. With the clinical variables alone, the fronts reached 0.25 to 0.36. One omics assay raised this to 0.78 to 0.81 on the filter fronts and to 0.58 to 0.71 on the evolutionary fronts. Two assays raised it to 0.87 to 0.89 on the filter fronts and to 0.73 to 0.83 on the evolutionary fronts. On the filter fronts and for the ten competitive algorithms, a third, fourth, or fifth assay added at most 0.03. For a budget of two assays, the per-budget rule chose DNA

methylation and somatic mutation in four of the five folds of the ANOVA and L1 fronts and in 173 of the 200 runs of the ten competitive algorithms. Gene expression took the place of methylation in the remaining fold of the ANOVA and L1 fronts, in every fold of the mRMR front, and in 25 of the remaining 27 runs; in the other two runs, the rule chose a panel of somatic mutation alone. Somatic mutation was thus part of every panel chosen for this budget, and methylation and expression were equally effective partners in this cohort. Figure S2 shows the data blocks of the one-standard-error panels. Somatic mutation was part of this panel in every run of 12 of the 13 algorithms, and DNA methylation in at least 75% of the runs of every algorithm except MO-POISE-TELSO.

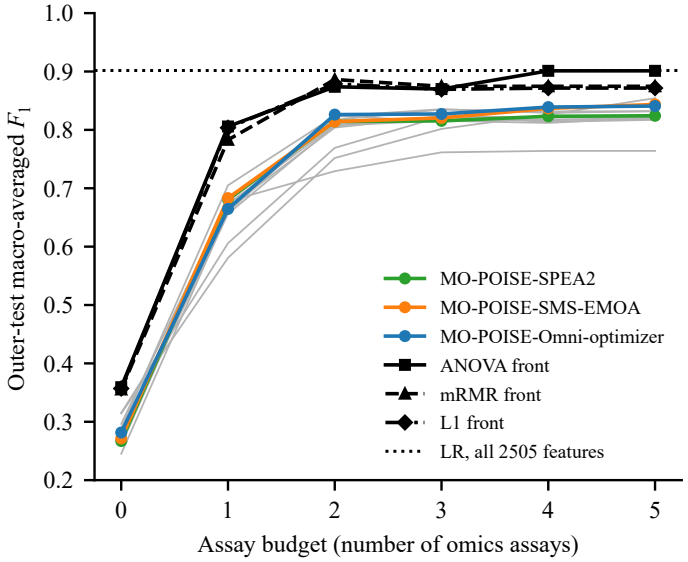

Figure 6. Mean outer-test macro-averaged F_1 over the runs of the panel chosen by the per-budget rule for each number of omics assays. Colored lines: the three selected algorithms of Section 3.9; gray lines: the other algorithms; black lines: the three filter fronts; dotted line, labeled LR: logistic regression on all 2505 candidate features.

4.5. Characteristics of the Selected Panels

The one-standard-error panels of the ten competitive algorithms had median sizes of 11 to 30 features and used 2.85 to 3.70 assays on average. Their outer-test macro-averaged F_1 ranged from 0.819 for MO-POISE-NSGA-II to 0.844 for MO-POISE-SMS-EMOA. Figure S3 shows the per-class F_1. For the rarest class, *POLE*, it was 0.73 to 0.82 for these algorithms, compared with 0.79 to 0.82 for the copy-number-low class and 0.84 to 0.88 for the MSI and copy-number-high classes. Table 6 adds the per-class precision and recall. For the *POLE* class, the one-standard-error panels of the ten competitive algorithms reached a precision of 0.839 and a recall of 0.765 on average; for the MSI class, 0.891 and 0.841; for the copy-number-low class, 0.781 and 0.841; and for the copy-number-high class, 0.875 and 0.859. The recall of the *POLE* class was thus the lowest of the four: about one in four *POLE* tumors was assigned to another class, whereas a tumor assigned to the *POLE* class belonged to it in most cases. The agreement of these panels with the TCGA labels reached a Cohen's κ of 0.763 to 0.790, compared with 0.826 to 0.861 for the one-standard-error panels of the filter fronts, which contain more features. For the panels with at most five features, the rare *POLE* class separated the approaches: its recall was 0.724 for the ten competitive algorithms, compared with 0.186, 0.632, and 0.325 for the ANOVA, mRMR, and L1 panels, and the κ of these panels was 0.725, compared with 0.669 to 0.695 for the filter panels.

Table 6. Per-class outer-test precision (P) and recall (R) and Cohen's κ with the TCGA labels of the panel chosen by the one-standard-error rule and of the panel chosen among those with at most five features, as means over the 20 runs of each algorithm and over the five folds of each filter front. Ten competitive: mean over the 200 runs of the ten competitive algorithms. CN-low and CN-high: copy-number-low and copy-number-high. Molecular surrogate: the rule-based surrogate of the clinical classification of Section 3.11, from somatic mutation and DNA methylation.

	POLE		MSI		CN-Low		CN-High		
Method	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	κ
One-standard-error panel									
MO-POISE-SPEA2	0.823	0.756	0.892	0.852	0.807	0.826	0.851	0.856	0.772
MO-POISE-SMS-EMOA	0.875	0.810	0.915	0.852	0.777	0.841	0.881	0.868	0.790
MO-POISE-Omni-optimizer	0.849	0.826	0.900	0.831	0.776	0.853	0.887	0.854	0.784
Ten competitive	0.839	0.765	0.891	0.841	0.781	0.841	0.875	0.859	0.775
ANOVA filter front	0.935	0.900	0.949	0.886	0.852	0.904	0.902	0.910	0.861
mRMR filter front	0.927	0.850	0.919	0.876	0.834	0.845	0.884	0.910	0.826
L1 filter front	0.931	0.768	0.926	0.905	0.841	0.875	0.900	0.909	0.837
Panel with at most five features									
MO-POISE-SPEA2	0.768	0.738	0.944	0.807	0.745	0.822	0.830	0.838	0.741
MO-POISE-SMS-EMOA	0.776	0.736	0.925	0.831	0.728	0.810	0.843	0.819	0.737
MO-POISE-Omni-optimizer	0.780	0.738	0.927	0.817	0.722	0.791	0.817	0.827	0.727
Ten competitive	0.743	0.724	0.919	0.808	0.722	0.801	0.836	0.823	0.725
ANOVA filter front	0.733	0.186	0.894	0.810	0.613	0.827	0.875	0.860	0.670
mRMR filter front	0.793	0.632	0.851	0.724	0.709	0.846	0.819	0.819	0.695
L1 filter front	0.453	0.325	0.849	0.714	0.657	0.893	0.880	0.835	0.669
Molecular surrogate	0.975	0.950	0.939	0.952	0.834	0.894	0.931	0.851	0.865

The features most often contained in the one-standard-error panels, listed in Table 7, were mutations of *RPL22*, *TP53*, and *POLE*, several DNA-methylation probes annotated to *MLH1*, and the expression of *MLH1*. They correspond to the defining properties of the classes: *POLE* mutation for the ultramutated class, silencing of *MLH1* by promoter methylation for the MSI class, and *TP53* mutation for the copy-number-high class. *RPL22* frameshift mutations are known to be frequent in MSI endometrial tumors [59]. In this cohort, *RPL22* was mutated in 55% of the MSI tumors and in none of the copy-number-low and copy-number-high tumors, and *TP53* in 86% of the copy-number-high tumors. The six *MLH1* probes of Table 7, cg14671526, cg27331401, cg25837710, cg12851504, cg19208331, and cg14598950, lie within 160 base pairs of each other in the promoter CpG island that *MLH1* shares with *EPM2AIP1*, at chr3:36,993,708–36,993,865 in the human reference genome assembly GRCh38 (hg38). Their mean beta value was 0.37 to 0.57 in the MSI tumors and at most 0.06 in the other classes, so that each of them records the same event, the epigenetic silencing of *MLH1*. The other mutated genes of the table, *TLR7*, *DNAH7*, *PAPPA*, *ZDBF2*, *NEB*, and *CMYA5*, were mutated in 68% to 87% of the *POLE* tumors, in at most 4% of the copy-number-low and copy-number-high tumors, and in at most 13% of the MSI tumors, except *DNAH7* and *NEB*, which were mutated in 35% and 28% of the MSI tumors. They are mostly long genes, and their mutation marks the extreme mutation burden of the *POLE* class, with a median of 4514 mutated genes per tumor, compared with 538 in the MSI class and 44 to 45 in the copy-number classes. *POLE* itself was mutated in every *POLE* tumor but also in 18% of the MSI tumors, so that a panel can identify the *POLE* class either by the gene or by such markers of ultramutation. The four methylation probes without an *MLH1* annotation, in *FABP6*, *ZBED9*, *USP7*, and *MYT1*, differed mainly between the copy-number-high class and the other classes.

Table 7. The features most often contained in the panels chosen by the one-standard-error rule, over all runs of the 13 algorithms. Share: fraction of runs. Methylation probes are given with the gene annotated in the GDC hg38 probe map.

Block	Feature	Share
Mutation	*RPL22*	0.34
Mutation	*TP53*	0.28
Mutation	*POLE*	0.14
Methylation	cg25837710 (*MLH1*)	0.14
Mutation	*TLR7*	0.13
Mutation	*DNAH7*	0.12
Methylation	cg17923945 (*FABP6*)	0.12
Methylation	cg03553715 (*USP7*)	0.11
Methylation	cg12851504 (*MLH1*)	0.11
Methylation	cg27331401 (*MLH1*)	0.11
Methylation	cg14671526 (*MLH1*)	0.11
Methylation	cg14598950 (*MLH1*)	0.10
Methylation	cg22152492 (*ZBED9*)	0.10
Mutation	*PAPPA*	0.10
mRNA	*MLH1* (ENSG00000076242)	0.09
Mutation	*ZDBF2*	0.09
Methylation	cg03102788 (*MYT1*)	0.09
Methylation	cg19208331 (*MLH1*)	0.09
Mutation	*NEB*	0.09
Mutation	*CMYA5*	0.08

Table S2 shows the panels refitted with the random forest and the support vector machine. The one-standard-error panels of the ten competitive algorithms reached a macro-averaged F_1 of 0.834 with the random forest and 0.830 with the support vector machine, compared with 0.828 with logistic regression; for the panels with at most five features, the corresponding values were 0.761, 0.783, and 0.785. The order between the evolutionary and the filter panels was the same for all three classifiers. With at most five features, the evolutionary panels were more accurate than the filter panels of the same fold by 0.053 to 0.127 against the ANOVA front, 0.033 to 0.057 against the mRMR front, and 0.074 to 0.098 against the L1 front, averaged over the folds, whereas the larger one-standard-error panels of the filter fronts remained more accurate than those of the evolutionary fronts, by 0.026 to 0.073. The panels found with logistic regression were thus not specific to this classifier.

4.6. Accuracy Relative to the Molecular Surrogate of the Clinical Classification

The molecular surrogate of the clinical classification, which uses the somatic-mutation and DNA-methylation assays, reached an outer-test macro-averaged F_1 of 0.913 over the five folds, with 0.878 to 0.932 on the individual folds, and a Cohen's κ of 0.865; Table 6 lists its per-class precision and recall. The panels of at most five features of the ten competitive algorithms, which consisted of somatic-mutation and DNA-methylation features alone in 77% of the runs, reached a macro-averaged F_1 of 0.785 and a κ of 0.725, and their one-standard-error panels reached 0.828 and 0.775. The κ of the surrogate exceeded that of the panels of at most five features in 199 of the 200 runs, and that of the one-standard-error panels in 190 of them, and it also exceeded the 0.826 to 0.861 of the one-standard-error panels of the filter fronts. The surrogate uses three aggregate readouts defined from clinical knowledge: hotspot status among the many *POLE* variants, one methylation call that summarizes the probes of the *MLH1* promoter, and truncating mutations of four mismatch-repair genes. None of them is among the candidates of MO-POISE-X, whose mutation features record any non-silent variant of a gene and whose methylation features are single

probes. These readouts are also close to the definitions of the TCGA classes, which were derived from *POLE* mutations with an extreme mutation rate, from microsatellite instability, and from copy-number clustering [1].

With these three readouts in the candidate pool, the search selected *POLE* hotspot status for 78% of the one-standard-error panels of the three selected algorithms and the truncating mismatch-repair mutation for 40%. The mean promoter methylation of *MLH1* entered no front, because single probes of the same promoter were selected in every run. The outer-test macro-averaged F_1 of the one-standard-error panels rose from 0.836 to 0.879, and their Cohen's κ from 0.782 to 0.822, and those of the panels with at most five features rose from 0.796 to 0.865 and from 0.735 to 0.795, all with $p < 0.001$. With the same readouts in their pools, the filter panels of at most five features reached 0.674 to 0.857 and a κ of 0.670 to 0.773, below the evolutionary means of 0.865 and 0.795, although the macro-averaged F_1 of MO-POISE-SPEA2 alone, 0.846, was below that of the mRMR panels, 0.857. The surrogate remained more accurate than the evolutionary panels on average, with $\kappa = 0.865$; the panels with at most five features reached or exceeded its κ in 11 of the 60 runs and the one-standard-error panels in 15. The larger one-standard-error panels of the filter fronts reached a κ of 0.845 to 0.868 with the same readouts, and that of the L1 front, 0.868, matched the surrogate. Table S3 lists the results.

4.7. Stability Across Outer Folds

Table S4 reports the results for each outer fold. The mean outer-test macro-averaged F_1 of the one-standard-error panels of the ten competitive algorithms ranged from 0.757 on fold 3 to 0.864 on fold 4, with a standard deviation of 0.043 between the fold means, and the ANOVA and L1 filter fronts also had their lowest values on fold 3. The partition thus shifts the evolutionary and the filter panels largely together. For a budget of two assays, the per-budget rule chose DNA methylation with somatic mutation in 83% to 93% of the runs of every fold. A readout of *MLH1* silencing, a methylation probe annotated to *MLH1*, or the expression of *MLH1*, was part of 94% to 98% of the one-standard-error panels in every fold. The individual features were less stable. The share of panels with an *RPL22* mutation ranged from 0% on fold 3 to 65% on fold 1, that with a *TP53* mutation from 8% to 40%, and that with a *POLE* mutation from 6% to 25%. The mean pairwise Jaccard index of the one-standard-error panels was 0.027 within a fold and 0.015 between folds over the 13 algorithms, and 0.036 and 0.019 over the ten competitive algorithms. The low values reflect that interchangeable features carry the same signal: any of the six adjacent *MLH1* promoter probes records *MLH1* silencing, and several mutated genes mark the ultramutated *POLE* class. The assays and the readout of *MLH1* silencing were therefore stable across folds, whereas the features that record the other events varied.

4.8. Inner Cross-Validation Versus Outer-Test Performance

Figure S4 shows that the inner and outer-test macro-averaged F_1 of a panel were closely related, with a Pearson r of 0.91 for the fronts of the three selected algorithms of Section 3.9, and of 0.95, 0.92, and 0.84 for the panels of the ANOVA, mRMR, and L1 fronts. These values pool panels of all sizes. Within a single front of the three algorithms, the median Spearman correlation was 0.79, with an interquartile range of 0.72 to 0.85 over the 60 runs, and 0.56 among the panels with at least five features. The inner estimates were, however, optimistic. The mean generalization gap was 0.017 for the ANOVA panels, 0.036 for the mRMR panels, 0.084 for the L1 panels, and 0.079 for the panels on the fronts of these three algorithms, and it reached 0.103 for the one-standard-error panels of the ten competitive algorithms. For the ANOVA panels with at most five features, it was 0.016. The gap thus grows with the degree to which the selection itself is fitted to the training

data: it is smallest for a univariate ranking, larger for greedy selection, similar for penalized selection and the evolutionary fronts, and largest for the panels that the one-standard-error rule picks from a search that optimizes the inner score directly, as the literature on wrapper selection predicts [36,46].

Re-scoring the final fronts on five independent repetitions of the inner partition separated two parts of this gap. For the ten competitive algorithms, the one-standard-error panels chosen with the search scores had an inner macro-averaged F_1 of 0.931 and an outer-test value of 0.828. Scored on the independent partitions, the same panels reached 0.906, so about a quarter of their gap of 0.103 came from the single inner partition on which the search had optimized, and a gap of 0.078 remained. Choosing the panels with the re-scored inner errors lowered the gap of the chosen panels to 0.085, significantly, with $p < 0.001$, but left their outer-test macro-averaged F_1 unchanged, at 0.827. The mean gap over all panels of the fronts of the ten competitive algorithms fell from 0.071 to 0.057. For the panels with at most five features, the gap fell from 0.065 to 0.060 and the outer-test macro-averaged F_1 was again unchanged, 0.787 compared with 0.785.

4.9. Sensitivity Analyses

Raising the upper end of the initial density from 0.2 to 1.0 increased the test-side hypervolume of MO-POISE-SPEA2 from 0.967 to 0.997, that of MO-POISE-SMS-EMOA from 0.965 to 1.000, and that of MO-POISE-Omni-optimizer from 0.964 to 0.997, significantly for all three, and enlarged the one-standard-error panels, with medians of 25 to 32 features instead of 14 to 16. Quadrupling the budget to 20,000 evaluations increased it to 0.970, 0.978, and 0.986, significantly only for MO-POISE-Omni-optimizer. With 250 or 1000 candidates per omics block in place of 500, the test-side hypervolume changed by -0.010 to $+0.021$, and no change was significant; the macro-averaged F_1 of the one-standard-error panels was 0.814 to 0.833, compared with 0.824 to 0.844, a decrease that was significant only for MO-POISE-SMS-EMOA with 1000 candidates, and the panels with at most five features reached 0.772 to 0.810, compared with 0.793 to 0.799. The difference to the ANOVA front rebuilt from the same candidate pool was -0.056 to -0.074 with 250 candidates and -0.073 to -0.092 with 1000, compared with -0.088 to -0.091 with 500; with 250 or 1000 candidates, the fronts remained below this front in 112 of the 120 runs. The hypervolume of the ANOVA front was 1.041, 1.055, and 1.047 with 250, 500, and 1000 candidates, and neither it nor that of the evolutionary fronts changed systematically with the size of the pool. The assays of the panels did not depend on the pool: somatic mutation was part of every one-standard-error panel and DNA methylation of 98% to 100% of them, counted over the 60 runs of each pool size, and the panel chosen for at most two omics assays combined somatic mutation with methylation or expression in all 60 runs of each pool size. With assay-specific costs, the one-standard-error panels of the three selected algorithms retained somatic mutation in every run, although it was the most expensive assay, and DNA methylation in 90% to 95% of the runs. The inexpensive copy-number assay entered 70% to 85% of these panels, compared with 35% to 50% under equal costs, and mRNA, the second most expensive assay, only 35% to 45%, compared with 60% to 70%. The mean assay-specific cost of the panels, 3.59 to 3.79, and their macro-averaged F_1, 0.820 to 0.834, did not differ significantly from those of the panels found under equal costs, evaluated with the same cost vector. On the cost-weighted scale, the difference to the ANOVA front, whose hypervolume was recomputed with the same costs, was -0.058 to -0.066, smaller than the difference under equal costs. Table S5 summarizes the five settings for the three selected algorithms. Across these 15 combinations, the fronts remained below the ANOVA filter front of their fold in 267 of the 300 runs, with mean paired differences between -0.055 and -0.092.

4.10. Convergence and Computational Cost

As Figure S5 shows, the inner hypervolume of the competitive algorithms rose quickly over the first 1000 evaluations and continued to rise slowly, without reaching a plateau, until the budget was exhausted. The curve of MO-POISE-TELSO starts late because its variable scoring consumes about half of the budget. Table S6 lists the run time of every algorithm in the computing environment of Section 3.12. The search took 66 to 83 s per run for the ten competitive algorithms, 136 s for MO-POISE-MOBCA, and 16 and 43 s for MO-POISE-IM-C-MOEA/D and MO-POISE-TELSO. This time was dominated by the inner cross-validation. Because a panel that has already been evaluated is taken from a cache, the run time depends mainly on the number of distinct panels an algorithm generates. MO-POISE-IM-C-MOEA/D evaluated only 732 distinct panels per run on average, MO-POISE-TELSO about 2700, and the other instances 4500 to 5000. The panels of MO-POISE-MOBCA were larger than those of the other instances and therefore more expensive to evaluate. Data loading, the in-fold prefilter, and the outer-test evaluation of the final front added about 5.5 s per run.

5. Discussion

5.1. Evolutionary Search Versus Filter Methods

The main result has two sides, and it assigns complementary roles to the two approaches: the evolutionary search supplies the compact panels, and filter selection supplies the large panels with the highest accuracy. For compact panels of at most 50 features, ten of the 13 algorithms were not significantly less accurate than the best of three filter strategies at any panel size, with mean differences of at most 0.025, and with at most five features, nine of them had a higher mean accuracy than every filter front. They significantly exceeded the univariate ANOVA ranking, by 0.07 to 0.13 in macro-averaged F_1, and L1-penalized regression, by up to 0.12, and they matched mRMR at every size up to 50 features and exceeded it at five features on average, though not significantly. The three filters show what makes a compact panel accurate. A univariate ranking fills a small panel with the features that individually separate the classes best, and these tend to be redundant. mRMR, which penalizes redundancy explicitly, recovers most of the gain, in line with a benchmark on cancer multi-omics data in which mRMR was among the strongest selectors at small panel sizes [42], whereas the L1 path, which selects jointly, recovers it only from about 50 features on. Avoiding redundancy across features and assays is thus a key ingredient of a compact panel. The evolutionary search arrives at such panels directly from the cross-validated error, without a hand-designed redundancy criterion and jointly with their assay cost. These best panels of at most five features combined somatic mutation and DNA-methylation features in 77% of the runs, a cross-assay pairing that the mRMR ranking reached in three of the five folds and the ANOVA ranking, which filled its five-feature panels with mRNA features, in none, as Figure 5 shows.

For large panels, the picture is reversed. On average, even the best panel on a front of the ten competitive algorithms stayed below the macro-averaged F_1 of about 0.90 that the filter fronts attain with hundreds to thousands of features, and the test-side hypervolume, which rewards the whole trade-off, was lower than that of the ANOVA front of the same fold in 244 of the 260 runs, lower than that of the mRMR front in 258, and lower than that of the L1 front in all of them. The hypervolume measures panel size on a linear scale relative to the 2505 candidates, so it hardly distinguishes a panel of five features from one of 50; the comparison of compact panels therefore rests on the size-capped analyses and not on the hypervolume. At the level of single features, the subtype signal in these data is spread over many features, and predictive performance keeps increasing with panel size well beyond the sizes at which the searches operate.

The sensitivity analyses show that this gap is partly a property of the search design and partly a property of the problem. Initializing with per-block densities up to 1.0 instead of 0.2 increased the test-side hypervolume of MO-POISE-SPEA2, MO-POISE-SMS-EMOA, and MO-POISE-Omni-optimizer by 0.030 to 0.035 but closed only about a third of the gap to the ANOVA front, and quadrupling the budget increased it by 0.003 to 0.021, significantly only for MO-POISE-Omni-optimizer, whereas with 250 or 1000 candidates per block the gap remained between 0.056 and 0.092. A front that spans panel sizes from a single feature to about 2000 features over a binary space of 2505 variables is hard to approximate with a population of 50 and a few thousand evaluations. Moreover, every additional feature must be paid for by a lower error, so that most non-dominated panels are small and the size objective keeps the population at small sizes. The two approaches are therefore complementary along the front: the search supplies the compact end, and the filters supply the large-panel end, which suggests seeding the population with mRMR- or L1-selected panels as a direct way to combine them.

5.2. Role of the Block-Sparse Initialization

The ablation separates the two parts of the block-sparse initialization. Its block step, which samples the assays of an initial panel before its features, raised the test-side hypervolume of all ten competitive algorithms, significantly for nine of them after Holm adjustment, and the chosen panels used fewer assays for all ten, although this difference was significant for only four of them before adjustment and for none after it. A sparse initialization without this step starts with almost every assay active, and the block-aware mutation removes at most one assay per offspring, so that the search must remove the surplus assays one at a time. Compared with the uniform initialization, the block-sparse initialization concentrates the search on the compact end of the trade-off. The uniform initialization produced larger panels with about four assays, which were more accurate for two of the three algorithms, but it changed the hypervolume in no consistent direction, left no panel of at most five features in 11 of 60 runs, and doubled the search time. Initial densities up to 1.0 instead of 0.2 also raised the hypervolume of all three algorithms, with larger one-standard-error panels. The upper limit of 0.2 is therefore a choice in favor of the compact end of the trade-off, at some cost in hypervolume. The block-sparse initialization thus directs the search to the region of the trade-off for which MO-POISE-X is designed, whereas the large-panel end is supplied by the filter fronts, as discussed above.

5.3. Number of Assays Required for Subtype Reconstruction

The assay objective answers a question that a single-objective search over features cannot ask. All three filter fronts and the fronts of the ten competitive algorithms show a steep increase in macro-averaged F_1 from the clinical variables alone to one omics assay and two assays, and little gain from a third, fourth, or fifth assay. For a budget of two assays, the per-budget rule chose DNA methylation and somatic mutation in most runs of the ten competitive algorithms and in four of the five folds of the ANOVA and L1 fronts, and somatic mutation with gene expression on the mRMR front, which reached the same accuracy. In this cohort and framework, somatic mutation was thus the assay that every two-assay panel retained, and it remained in every one-standard-error panel of the three selected algorithms rerun with assay-specific costs, in which it was the most expensive assay. Its best partner was a readout of *MLH1* silencing, either as promoter methylation or as loss of expression, and the two partners reached the same accuracy. This choice is biologically coherent and mirrors the logic of the clinical classification [2,3]. Somatic mutations identify the *POLE*-mutated and hypermutated MSI tumors and the *TP53*-mutated copy-number-high tumors. Promoter methylation of *MLH1*, the most common

cause of MSI in endometrial carcinoma [1], separates MSI tumors from the other classes. Copy-number data, which define the copy-number-high class, were added less often, because most copy-number-high tumors carry a *TP53* mutation [1] that the mutation assay already records. The per-budget rule thus turns an evolutionary or a filter front into one concrete panel per assay budget, which a weighted-sum wrapper cannot provide; at one and two assays, the filter fronts gave the more accurate panels. The one-standard-error panels of the filter fronts also used fewer assays, 2.20 to 2.60 on average, than those of the ten competitive algorithms, 2.85 to 3.70. The advantage of the search thus lies in the number of features, which matters where each feature is measured separately, for example by targeted sequencing or by methylation assays of single loci, whereas a panel that is read from complete omics assays is served at least as well by the filter fronts. These combinations describe where the subtype signal lies in the TCGA data. They are not assay recommendations for clinical practice, in which the subtypes are assigned by sequencing of *POLE* and immunohistochemistry [2,5].

5.4. Comparison with the Molecular Surrogate of the Clinical Classification

The molecular surrogate of the clinical classification agreed with the TCGA labels more closely, on average, than the panels of every algorithm and every filter front in the main analysis, with a Cohen's κ of 0.865, compared with 0.775 for the one-standard-error panels of the ten competitive algorithms. Its advantage has a clear source. Its three readouts, hotspot status among the many *POLE* variants, one methylation call for the *MLH1* promoter, and truncating mutations of the mismatch-repair genes, condense expert knowledge of the classes into aggregate features, and they are close to the definitions of the TCGA classes themselves [1].

MO-POISE-X received no such aggregate features. Its candidates are single genes and single probes, and it learned from the data which assays and which features matter. Its panels nevertheless reflected the structure of the clinical classification: the panels of at most five features consisted of somatic-mutation and DNA-methylation features alone in 77% of the runs of the ten competitive algorithms, the per-budget rule chose these two assays for a budget of two assays in 173 of their 200 runs, and the selected features record *POLE* mutation, *MLH1* silencing, and *TP53* mutation, the three readouts of the clinical hierarchy. This agreement is expected to a degree, because the TCGA classes were defined in part by *POLE* mutation and microsatellite instability, and the ANOVA and L1 fronts chose the same two assays for a budget of two assays in four of the five folds. The endometrial subtypes thus served as a test case with a known answer, in which panels built from single features agreed with expert knowledge. The comparison also shows how MO-POISE-X can be extended. Because the framework is plug-and-play, its candidate pool can take expert-defined aggregate features next to single features, and the search decides, jointly with their assay cost, whether they enter a panel. When the three readouts of the surrogate were added, the search took up *POLE* hotspot status in most panels, and the panels of at most five features gained 0.069 in macro-averaged F_1 and 0.060 in κ, remaining on average more accurate than the filter panels of the same size built with the same readouts, whereas the larger one-standard-error panels of the L1 front reached the κ of the surrogate. A gap between the compact panels and the surrogate remained. Logistic regression combines the features additively, whereas the surrogate applies them in a fixed order in which a *POLE* hotspot overrides every other readout, so a rule-based or tree-based classifier is a natural plug-in for such knowledge-augmented searches.

5.5. Optimism of the Inner Estimate

The generalization gap shifts the absolute values on a front and blurs the order of panels of similar size. Pooled over the fronts of the three selected algorithms, inner and outer-test macro-averaged F_1 were correlated with $r = 0.91$, but within a single front the median Spearman correlation was 0.79, and 0.56 among the panels with at least five features. The inner score thus separates small from large panels reliably but orders panels of similar size only roughly, and every comparison in this study uses outer-test values. Two sources contribute to the gap. The first is the prefilter, which ranks the candidates on the whole outer training fold, including the patients of the inner validation folds. The ANOVA filter front ranks its panels in the same way, and its small gap, 0.017 over all panels and 0.016 for those with at most five features, shows that this shared ranking makes the inner estimate only slightly optimistic. The second and larger source is the search itself. It scores thousands of panels on one fixed inner partition, and the one-standard-error panel is chosen from the resulting front by the same noisy score, so its inner value is the largest of many noisy estimates, a selection effect known from wrapper selection [36,46]. This explains why the gap grows from 0.071 for all panels on the fronts of the ten competitive algorithms to 0.103 for their one-standard-error panels. The re-scoring on independent inner partitions shows that only about a quarter of this gap is tied to the single partition of the search. The re-scoring keeps the prefilter, so the remaining 0.078 contains the share of the prefilter, which the ANOVA front suggests is small, and mainly reflects that the search adapts the panels to the patients of the outer training fold, whatever their partition.

The inner estimate of a searched panel should therefore never be reported as its performance. In an application, the panel should be chosen by training-side rules and its accuracy estimated on data not used in the search, as the outer test folds do here. Re-scoring the final front on repeated inner partitions before the one-standard-error rule is applied gives a less optimistic inner estimate, 0.085 instead of 0.103 above the outer-test value, at no cost in accuracy, and is therefore a simple step for any application. It did not make the chosen panels more accurate, because the remaining optimism comes from the search over the training patients; reducing it requires noise-aware dominance during the search or a wider selection margin, as discussed under future work.

5.6. Comparison of the Evolutionary Algorithms

The differences between the algorithms were small compared with the difference between any of them and the filter fronts in the region of large panels, and most algorithms could not be distinguished from the best one in the paired tests. They were also small in practical terms: among the ten competitive algorithms, the mean macro-averaged F_1 of the one-standard-error panels differed by at most 0.025, less than the variation between the runs of any single algorithm and about one inner standard error. For a user, the choice among these ten algorithms therefore matters less than the operators, the initialization, and the rule by which a panel is chosen. New multi-objective evolutionary algorithms continue to appear, and the plug-and-play design of MO-POISE-X lets each of them be assessed on the same problem, with the same operators, evaluation, and selection rules, and against the same filter fronts, without any change to the pipeline. MO-POISE-MOBCA and MO-POISE-IM-C-MOEA/D performed clearly worse in hypervolume and in compact panels. MOBCA is a real-coded algorithm designed for continuous search spaces, and its besiege-and-conquer moves translate poorly into threshold-decoded binary decisions over 2505 variables. Changes to the decoding could narrow this gap, for example, a threshold that adapts to the target number of features, or a transfer function that maps a continuous position to a selection probability. The difficulty, however, lies in the moves. In a sparse panel, almost all positions lie below the threshold, and a continuous step changes the

selection only where it crosses the threshold. Small steps therefore leave most of the panel unchanged, whereas large steps switch on many features at once, in line with the large one-standard-error panels of MO-POISE-MOBCA, with a median of 165 features. For sparse binary problems, binary or set-based operators are therefore the more suitable choice, unless the moves of such algorithms are redesigned for this encoding. MO-POISE-NRV-MOEA, which used the shared operators, behaved like the other instances with these operators. TELSO, designed for sparse large-scale problems, first scores each variable by evaluating single-variable solutions, which here consumed 2505 of the 5000 evaluations. In a four-class problem with a signal spread over many features, single features are weak guides: the single feature with the lowest inner error reached an outer-test macro-averaged F_1 of only 0.37 on average. TELSO keeps at most one solution per reference vector without refilling the population, so its swarm shrank to a few particles, and the fronts of MO-POISE-TELSO contained about five panels. IM-C-MOEA/D, like MOBCA, is real-coded; its inverse models resample only a few variables per objective, so that its offspring stay close to their parents in a space of 2505 variables. Good performance on synthetic or engineering benchmarks thus did not transfer to this problem, in which the budget is small relative to the dimension, the objective is noisy, and single variables carry little information.

5.7. Limitations

The study has limitations.

- All results come from the 369 patients of one cohort. The CPTAC endometrial cohort [20] differs from TCGA-UCEC in its methylation platform, in the derivation of copy number, in its processing batch, and in the procedure that assigns the subtypes, so the transfer of the specific panels to other cohorts was not examined, and the conclusions are restricted to TCGA-UCEC.
- The cohort is small for a four-class problem. It contains 38 *POLE* tumors, about eight per outer test fold, so the class-wise estimates for this class are uncertain.
- The inner estimate of macro-averaged F_1 has a standard error of about 0.02, so dominance decisions during the search partly follow noise; the one-standard-error rule accounts for this only when a panel is chosen.
- The main analysis assigns the same cost to every omics assay and none to the clinical variables. One assay-specific cost vector was examined, and laboratories with other prices may obtain other panels.
- The search used logistic regression. The panels kept their accuracy when refitted with a random forest or a support vector machine, but a search with another classifier may select other panels.
- Individual features varied between runs and folds; the stable units of the panels were the assays and the readout of *MLH1* silencing.
- A single outer partition was used, and the four runs of a fold share their patients, and the filter panels with which they are compared, so the run-level tests overstate the number of independent comparisons; the fold-level consistency is therefore reported with them.
- The algorithm parameters were fixed, and other settings may change the ranking of the algorithms.

5.8. Future Work

Three directions follow from these results. First, seeding the population with mRMR- or L1-selected panels, or alternating filter-based and evolutionary steps, could combine the compact end of the evolutionary fronts with the large-panel end of the filter fronts. Second, noise-aware dominance, for example, based on the inner standard error or on

repeated inner partitions, a selection rule with a wider margin than one standard error, or a correction of the overestimate during the search [46], would reduce the selection of noise. Third, the assay-cost formulation can be extended to the cost vectors of specific laboratories and to further aggregate features such as tumor mutational burden, combined with a classifier that can apply such readouts in a fixed order, as the clinical surrogate does.

6. Conclusions

We presented MO-POISE-X, which formulates the reconstruction of the TCGA molecular subtypes of endometrial carcinoma from multi-omics data as a three-objective problem of classification error, number of omics assays, and panel size. We compared 13 selected MOEAs on it under leakage-controlled nested cross-validation on the TCGA-UCEC cohort and judged every algorithm against univariate, redundancy-aware, and L1-penalized filter fronts. The two kinds of methods had complementary strengths. For compact panels, ten of the 13 algorithms were not significantly less accurate than the best filter at any size up to 50 features and, with at most five features, significantly more accurate than ANOVA-ranked and L1-selected panels, by up to 0.13 and 0.12 in outer-test macro-averaged F_1, and on average, though not significantly, more accurate than mRMR-ranked panels. The filter fronts, in turn, provide the large panels, with which logistic regression attains a macro-averaged F_1 of about 0.90, and they reach the higher hypervolume over the whole trade-off.

The proposed block-sparse initialization, which samples the assays of each initial panel before its features, raised the hypervolume of the ten algorithms with the highest hypervolume compared with a sparse initialization without this block step, and it led to panels with fewer assays and features than the uniform initialization, whereas denser initializations reached higher hypervolumes with larger panels. Two assays, somatic mutation together with DNA methylation or gene expression, captured most of the attainable performance in this cohort, in line with the clinical classification; somatic mutation remained in the panels when it was assigned the highest assay-specific cost, and the assays chosen were stable across the outer folds, although the individual features varied. The differences between most algorithms were small, and MO-POISE-TELSO, whose algorithm was designed for sparse benchmark problems, did not improve on the others.

In this cohort, multi-objective evolutionary search thus delivers, in a single run, compact panels with few features, which complement the large panels of filter selection. Its panels, built from single features, used the assays and recorded the biological events of the clinical classification, whose molecular surrogate, with its expert-defined readouts, remained more accurate. Because MO-POISE-X is plug-and-play, new algorithms and new candidate features, including such expert-defined readouts, can be added without changes to the pipeline; with the readouts of the surrogate in its candidate pool, the search selected them and its compact panels came closer to the surrogate.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/life16121986/s1, Figure S1: Mean paired difference Δ_k between the evolutionary and the filter panels; Figure S2: Data blocks of the one-standard-error panels; Figure S3: Per-class outer-test F_1; Figure S4: Inner cross-validated against outer-test macro-averaged F_1; Figure S5: Convergence of the inner hypervolume; Table S1: Ablation of the block-sparse initialization for the three selected algorithms; Table S2: Panels refitted with three classifiers; Table S3: Panels found with the readouts of the molecular surrogate in the candidate pool; Table S4: Results by outer fold; Table S5: Sensitivity analyses; Table S6: Run time of the algorithms.

Author Contributions: Conceptualization, Y.D., M.C. and B.L.; methodology and software, Y.Z. (Yiqin Zhang), Y.D. and M.C.; validation and data curation, Y.Z. (Yiqin Zhang), J.Y., Y.S. and G.J.; investigation, Y.Z. (Yanli Zhang), L.K. and J.L.; writing—original draft preparation, Y.Z. (Yiqin Zhang), Y.D. and M.C.; writing—review and editing, Y.Z. (Yiqin Zhang), Y.D., M.C., J.Y., Y.S., G.J., Y.Z. (Yanli

Zhang), L.K., J.L. and B.L.; supervision and funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers 82372925, 82172714 and 82674378 to Bilan Li), the Science and Technology Commission of Shanghai Municipality (Shanghai Pujiang Program, grant number 25PJD102 to Bilan Li), the Shanghai Hospital Development Center (grant number 2025SKMR-30 to Bilan Li), the Clinical Research Plan of SHDC (grant number SHDC2020CR4086 to Bilan Li), the Shanghai Oriental Talents Program (grant number QNWS2024042 to Bilan Li), the Shanghai Municipal Health Commission ("Medical Rising Star" Outstanding Youth Medical Talent Program, grant number 20264Z0005 to Bilan Li), the Shanghai Municipal Health System Youth Talent Award Foundation ("Silver Snake–Tailong" Special Technology Support Program, grant number YSTL2026048 to Bilan Li), the Shanghai Clinical Innovation and Translation Institute Co., Ltd. (grant number LZY20261RWS402 to Bilan Li), Tongji University ("Medicine + X" Interdisciplinary Research Project, grant number 2026-0555-YB-04 to Bilan Li), and Shanghai First Maternity and Infant Hospital (Key Project of Special Clinical Research, grant number 2025B02 to Bilan Li).

Institutional Review Board Statement: Not applicable; the study used only de-identified public data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and detailed experimental results can be downloaded at https://github.com/BilanLimch/MO-POISE-X (accessed on 17 November 2026). The omics data and clinical variables of TCGA-UCEC are available from the GDC hub of UCSC Xena (https://xenabrowser.net), and the molecular subtypes from the TCGA PanCancer Atlas endometrial study in cBioPortal (https://www.cbioportal.org).

Acknowledgments: The results are based on data generated by the TCGA Research Network (https://www.cancer.gov/tcga (accessed on 17 November 2026)). During the preparation of this manuscript, the authors used DeepSeek (DeepSeek-V4.1-Flash) for language polishing of the text. The authors have reviewed and edited the output, and all authors take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AFSEA	adjoint feature-selection-based evolutionary algorithm
AGE-MOEA	adaptive geometry estimation-based multi-objective evolutionary algorithm
AGE-MOEA-II	second version of AGE-MOEA
ANOVA	analysis of variance
APD	angle-penalized distance
ASCAT3	allele-specific copy-number analysis of tumors, version 3
BCB	besiege-and-conquer balance
CN-high	copy-number-high
CN-low	copy-number-low
CpG	cytosine–phosphate–guanine dinucleotide
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CV	cross-validation
DNA	deoxyribonucleic acid
ESGO	European Society of Gynaecological Oncology
ESP	European Society of Pathology
ESTRO	European Society for Radiotherapy and Oncology
FIGO	International Federation of Gynecology and Obstetrics
FN	false negatives
FP	false positives

FPKM	fragments per kilobase of transcript per million mapped reads
GDC	Genomic Data Commons
hg38	human reference genome assembly GRCh38
HV	hypervolume
IM-C-MOEA/D	inverse modeling constrained multi-objective evolutionary algorithm based on decomposition
ISPSO	particle swarm optimization with information gain ratio-based subfeature grouping
L1	L1-penalized logistic regression
LR	logistic regression
miRNA	microRNA
MMFS-GA	multi-objective genetic algorithm for multi-view feature selection
MO-POISE-X	Multi-Objective Plug-and-play Optimizer for Imbalance-aware Subtyping of Endometrial cancer with the plugged-in MOEA X
MOBCA	multi-objective besiege-and-conquer algorithm
MOEA	multi-objective evolutionary algorithm
MOEA/D	multi-objective evolutionary algorithm based on decomposition
mRMR	minimum redundancy–maximum relevance
mRNA	messenger RNA
MSI	microsatellite instability
NRV-MOEA	adaptive normal-vector guided multi-objective evolutionary algorithm
NSGA-II	non-dominated sorting genetic algorithm II
NSGA-III	non-dominated sorting genetic algorithm III
one-SE	one-standard-error
ProMisE	Proactive Molecular Risk Classifier for Endometrial Cancer
RNA	ribonucleic acid
RPM	reads per million mapped reads
RVEA	reference vector guided evolutionary algorithm
SD	standard deviation
SE	standard error
SHDC	Shanghai Hospital Development Center
SMOP	sparse multi-objective optimization problem
SMS-EMOA	S-metric selection evolutionary multi-objective optimization algorithm
SPEA2	strength Pareto evolutionary algorithm 2
STAR	Spliced Transcripts Alignment to a Reference
TCGA	The Cancer Genome Atlas
TELSO	two-layer encoding learning swarm optimizer
TP	true positives
U-NSGA-III	unified NSGA-III
UCEC	uterine corpus endometrial carcinoma
UCSC	University of California, Santa Cruz

References

1. Levine, D.A.; The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **2013**, *497*, 67–73.
2. Talhouk, A.; McConechy, M.K.; Leung, S.; Li-Chang, H.H.; Kwon, J.S.; Melnyk, N.; Yang, W.; Senz, J.; Boyd, N.; Karnezis, A.N.; et al. A clinically applicable molecular-based classification for endometrial cancers. *Br. J. Cancer* **2015**, *113*, 299–310.
3. Kommoss, S.; McConechy, M.K.; Kommoss, F.; Leung, S.; Bunz, A.; Magrill, J.; Britton, H.; Kommoss, F.; Grevenkamp, F.; Karnezis, A.; et al. Final validation of the ProMisE molecular classifier for endometrial carcinoma in a large population-based case series. *Ann. Oncol.* **2018**, *29*, 1180–1188.
4. León-Castillo, A.; de Boer, S.M.; Powell, M.E.; Mileshkin, L.R.; Mackay, H.J.; Leary, A.; Nijman, H.W.; Singh, N.; Pollock, P.M.; Bessette, P.; et al. Molecular Classification of the PORTEC-3 Trial for High-Risk Endometrial Cancer: Impact on Prognosis and Benefit From Adjuvant Therapy. *J. Clin. Oncol.* **2020**, *38*, 3388–3397. https://doi.org/10.1200/JCO.20.00549.

5. Concin, N.; Matias-Guiu, X.; Vergote, I.; Cibula, D.; Mirza, M.R.; Marnitz, S.; Ledermann, J.; Bosse, T.; Chargari, C.; Fagotti, A.; et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. *Int. J. Gynecol. Cancer* **2021**, *31*, 12–39.
6. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
7. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324.
8. Cao, M.; Zhang, Y.; Ding, Y.; Jiang, G.; Yan, J.; Sun, Y.; Zhang, Y.; Kang, L.; Luo, J.; Li, B. A Leakage-Controlled Benchmark of Metaheuristic Multi-Omics Feature Selection for Reconstructing TCGA Molecular Subtypes in Endometrial Cancer. *Cancers* **2026**, *18*, 3353. https://doi.org/10.3390/cancers18203353.
9. Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6562–6566.
10. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Trans. Evol. Comput.* **2016**, *20*, 606–626. https://doi.org/10.1109/TEVC.2015.2504420.
11. Emary, E.; Zawbaa, H.M.; Hassanien, A.E. Binary grey wolf optimization approaches for feature selection. *Neurocomputing* **2016**, *172*, 371–381.
12. Valentine, M.C.; Wong, A.; Chen, L.; Du, F.; Hughes, A.E.; Spencer, D.H.; Duncavage, E.J.; Sun, L. A fully next-generation sequencing-based method of classifying molecular sub-types of endometrial cancer retains prognostic value and expands biomarker targets. *Int. J. Gynecol. Cancer* **2025**, *35*, 100060. https://doi.org/10.1016/j.ijgc.2024.100060.
13. Xue, B.; Zhang, M.; Browne, W.N. Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach. *IEEE Trans. Cybern.* **2013**, *43*, 1656–1671. https://doi.org/10.1109/tsmcb.2012.2227469.
14. Al-Tashi, Q.; Abdulkadir, S.J.; Rais, H.M.; Mirjalili, S.; Alhussian, H. Approaches to Multi-Objective Feature Selection: A Systematic Literature Review. *IEEE Access* **2020**, *8*, 125076–125096. https://doi.org/10.1109/access.2020.3007291.
15. Jiao, R.; Nguyen, B.H.; Xue, B.; Zhang, M. A Survey on Evolutionary Multiobjective Feature Selection in Classification: Approaches, Applications, and Challenges. *IEEE Trans. Evol. Comput.* **2024**, *28*, 1156–1176. https://doi.org/10.1109/tevc.2023.3292527.
16. Tian, Y.; Zhang, X.; Wang, C.; Jin, Y. An Evolutionary Algorithm for Large-Scale Sparse Multiobjective Optimization Problems. *IEEE Trans. Evol. Comput.* **2020**, *24*, 380–393. https://doi.org/10.1109/tevc.2019.2918140.
17. Berek, J.S.; Matias-Guiu, X.; Creutzberg, C.; Fotopoulou, C.; Gaffney, D.; Kehoe, S.; Lindemann, K.; Mutch, D.; Concin, N. FIGO staging of endometrial cancer: 2023. *Int. J. Gynecol. Obstet.* **2023**, *162*, 383–394. https://doi.org/10.1002/ijgo.14923.
18. Hong, R.; Liu, W.; DeLair, D.; Razavian, N.; Fenyö, D. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. *Cell Rep. Med.* **2021**, *2*, 100400. https://doi.org/10.1016/j.xcrm.2021.100400.
19. Patel, N.; Salumets, A.; Modhukur, V. EMMA-STRAT: A multi-omics based machine learning framework for stratification of endometrial carcinoma molecular subtypes and MSI status. *BioData Min.* **2026**, *19*, 63. https://doi.org/10.1186/s13040-026-00587-5.
20. Dou, Y.; Kawaler, E.A.; Cui Zhou, D.; Gritsenko, M.A.; Huang, C.; Blumenberg, L.; Karpova, A.; Petyuk, V.A.; Savage, S.R.; Satpathy, S.; et al. Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **2020**, *180*, 729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.
21. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. https://doi.org/10.1109/4235.996017.
22. Zitzler, E.; Laumanns, M.; Thiele, L. *SPEA2: Improving the Strength Pareto Evolutionary Algorithm*; Technical Report TIK-Report 103; Computer Engineering and Networks Laboratory (TIK), ETH Zurich: Zürich, Switzerland, 2001.
23. Deb, K.; Tiwari, S. Omni-optimizer: A generic evolutionary algorithm for single and multi-objective optimization. *Eur. J. Oper. Res.* **2008**, *185*, 1062–1087. https://doi.org/10.1016/j.ejor.2006.06.042.
24. Zhang, Q.; Li, H. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. Evol. Comput.* **2007**, *11*, 712–731. https://doi.org/10.1109/tevc.2007.892759.
25. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans. Evol. Comput.* **2014**, *18*, 577–601. https://doi.org/10.1109/tevc.2013.2281535.
26. Seada, H.; Deb, K. A Unified Evolutionary Optimization Procedure for Single, Multiple, and Many Objectives. *IEEE Trans. Evol. Comput.* **2016**, *20*, 358–369. https://doi.org/10.1109/tevc.2015.2459718.
27. Cheng, R.; Jin, Y.; Olhofer, M.; Sendhoff, B. A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization. *IEEE Trans. Evol. Comput.* **2016**, *20*, 773–791. https://doi.org/10.1109/tevc.2016.2519378.
28. Beume, N.; Naujoks, B.; Emmerich, M. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *Eur. J. Oper. Res.* **2007**, *181*, 1653–1669. https://doi.org/10.1016/j.ejor.2006.08.008.

29. Panichella, A. An adaptive evolutionary algorithm based on non-euclidean geometry for many-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*; ACM: New York, NY, USA, 2019; pp. 595–603. https://doi.org/10.1145/3321707.3321839.
30. Panichella, A. An improved Pareto front modeling algorithm for large-scale many-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*; ACM: New York, NY, USA, 2022; pp. 565–573. https://doi.org/10.1145/3512290.3528732.
31. Qi, S.; Wang, R.; Zhang, T.; Yang, X.; Sun, R.; Wang, L. A Two-Layer Encoding Learning Swarm Optimizer Based on Frequent Itemsets for Sparse Large-Scale Multi-Objective Optimization. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 1342–1357. https://doi.org/10.1109/jas.2024.124341.
32. Zhang, P.; Yin, H.; Tian, Y.; Zhang, X. An adjoint feature-selection-based evolutionary algorithm for sparse large-scale multiobjective optimization. *Complex Intell. Syst.* **2025**, *11*, 127. https://doi.org/10.1007/s40747-024-01752-1.
33. Hua, Y.; Liu, Q.; Hao, K. Adaptive normal vector guided evolutionary multi- and many-objective optimization. *Complex Intell. Syst.* **2024**, *10*, 3709–3726. https://doi.org/10.1007/s40747-024-01353-y.
34. Jiang, J.; Wu, J.; Luo, J.; Yang, X.; Huang, Z. MOBCA: Multi-Objective Besiege and Conquer Algorithm. *Biomimetics* **2024**, *9*, 316. https://doi.org/10.3390/biomimetics9060316.
35. Farias, L.R.C.; Araújo, A.F.R. An Inverse Modeling Constrained Multi-Objective Evolutionary Algorithm Based on Decomposition. In *Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*; IEEE: Piscataway, NJ, USA, 2024; pp. 3727–3732. https://doi.org/10.1109/SMC54092.2024.10831275.
36. Reunanen, J. Overfitting in Making Comparisons Between Variable Selection Methods. *J. Mach. Learn. Res.* **2003**, *3*, 1371–1382.
37. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. https://doi.org/10.1186/1471-2105-7-91.
38. Xue, B.; Zhang, M.; Browne, W.N. Particle Swarm Optimisation for Feature Selection in Classification: Novel Initialisation and Updating Mechanisms. *Appl. Soft Comput.* **2014**, *18*, 261–276. https://doi.org/10.1016/j.asoc.2013.09.018.
39. Gao, J.; Wang, Z.; Jin, T.; Cheng, J.; Lei, Z.; Gao, S. Information Gain Ratio-Based Subfeature Grouping Empowers Particle Swarm Optimization for Feature Selection. *Knowl.-Based Syst.* **2024**, *286*, 111380. https://doi.org/10.1016/j.knosys.2024.111380.
40. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.
41. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
42. Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinform.* **2022**, *23*, 412. https://doi.org/10.1186/s12859-022-04962-x.
43. Boulesteix, A.L.; De Bin, R.; Jiang, X.; Fuchs, M. IPF-LASSO: Integrative L_1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Comput. Math. Methods Med.* **2017**, *2017*, 7691937. https://doi.org/10.1155/2017/7691937.
44. Cattelani, L.; Ghosh, A.; Rintala, T.J.; Fortino, V. A Comprehensive Evaluation Framework for Benchmarking Multi-Objective Feature Selection in Omics-Based Biomarker Discovery. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2024**, *21*, 2432–2446. https://doi.org/10.1109/TCBB.2024.3480150.
45. Cattelani, L.; Fortino, V. Triple and quadruple optimization for feature selection in cancer biomarker discovery. *J. Biomed. Informatics* **2024**, *159*, 104736. https://doi.org/10.1016/j.jbi.2024.104736.
46. Cattelani, L.; Fortino, V. Dual-stage optimizer for systematic overestimation adjustment applied to multi-objective genetic algorithms for biomarker selection. *Briefings Bioinform.* **2025**, *26*, bbae674. https://doi.org/10.1093/bib/bbae674.
47. Imani, V.; Sevilla-Salcedo, C.; Moradi, E.; Fortino, V.; Tohka, J. Multi-objective genetic algorithm for multi-view feature selection. *Appl. Soft Comput.* **2024**, *167*, 112332. https://doi.org/10.1016/j.asoc.2024.112332.
48. Zhang, Y.; Gong, D.W.; Cheng, J. Multi-Objective Particle Swarm Optimization Approach for Cost-Based Feature Selection in Classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 64–75. https://doi.org/10.1109/TCBB.2015.2476796.
49. Erion, G.; Janizek, J.D.; Hudelson, C.; Utarnachitt, R.B.; McCoy, A.M.; Sayre, M.R.; White, N.J.; Lee, S.I. A cost-aware framework for the development of AI models for healthcare applications. *Nat. Biomed. Eng.* **2022**, *6*, 1384–1398. https://doi.org/10.1038/s41551-022-00872-8.
50. Goldman, M.J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N.; et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **2020**, *38*, 675–678.
51. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2012**, *2*, 401–404.
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

53. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009. https://doi.org/10.1007/978-0-387-84858-7.
54. Zitzler, E.; Thiele, L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **1999**, *3*, 257–271. https://doi.org/10.1109/4235.797969.
55. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
56. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
57. León-Castillo, A.; Britton, H.; McConechy, M.K.; McAlpine, J.N.; Nout, R.; Kommoss, S.; Brucker, S.Y.; Carlson, J.W.; Epstein, E.; Rau, T.T.; et al. Interpretation of Somatic *POLE* Mutations in Endometrial Carcinoma. *J. Pathol.* **2020**, *250*, 323–335. https://doi.org/10.1002/path.5372.
58. Blank, J.; Deb, K. Pymoo: Multi-Objective Optimization in Python. *IEEE Access* **2020**, *8*, 89497–89509. https://doi.org/10.1109/access.2020.2990567.
59. Ferreira, A.M.; Tuominen, I.; van Dijk-Bos, K.; Sanjabi, B.; van der Sluis, T.; van der Zee, A.G.; Hollema, H.; Zazula, M.; Sijmons, R.H.; Aaltonen, L.A.; et al. High Frequency of RPL22 Mutations in Microsatellite-Unstable Colorectal and Endometrial Tumors. *Hum. Mutat.* **2014**, *35*, 1442–1445. https://doi.org/10.1002/humu.22686.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.