Benchmarking computational methods for multi-omics biomarker discovery in cancer

Athan Z. Li[1], Yuxuan Du[2], Yan Liu[1], Liang Chen[3], Ruishan Liu[1,3*]

[1]Department of Computer Science, University of Southern California, Los Angeles, 90089, CA, USA.
[2]Department of Electrical Engineering, University of Texas at San Antonio, San Antonio, 78249, TX, USA.
[3]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, 90089, CA, USA.

*Corresponding author(s). E-mail(s): ruishanl@usc.edu;

Abstract

Multi-omics profiling characterizes cancer biology and supports biomarker discovery for prognosis and therapy selection. Although numerous computational methods have been proposed, their ability to identify clinically relevant biomarkers has not been systematically evaluated. Here, we benchmark 20 representative statistical, machine learning (ML) and deep learning (DL) methods using curated gold-standard prognostic and therapeutic biomarkers across five real-world datasets. We evaluate performance in terms of both biomarker identification accuracy and stability. Pathway-informed transformer methods, including DeePathNet and DeepKEGG, achieve the best overall performance. Across methods, effective biomarker recovery is associated with the integration of biological knowledge, global feature interactions, multivariate feature attribution, and effective regularization. Analysis of omics-type contributions reveals method- and modality-specific biases, highlighting the importance of broader omics integration. We further evaluate methods on simulated datasets to probe sensitivity with controlled signal and noise. Finally, by aggregating results from top-performing methods, we construct consensus biomarker panels that nominate candidates for future investigations. All datasets, code, and evaluation pipelines are released as a public resource to support future studies in multi-omics biomarker discovery for cancer.

Introduction

High-throughput technologies have enabled large-scale profiling of human cancers across genomic, transcriptomic, epigenomic, proteomic, and metabolomic layers [1, 2]. Collectively, these assays constitute a comprehensive and complementary view of tumor biology, substantially expanding the landscape for discovering biomarkers that can inform prognosis and guide therapy selection from a vast pool of molecular candidates [3]. Nevertheless, the high dimensionality of the data, the heterogeneity of tumor samples, and the intricate interdependencies across omics make robust and clinically meaningful biomarker discovery inherently challenging [4].

Over the past several years, a diverse array of computational methods has been developed to address this need (Supplementary Table 1). Statistical and machine learning (ML) methods extend classical frameworks such as matrix factorization, canonical correlation analysis, or Uniform Manifold Approximation and Projection (UMAP) to extract shared patterns across omics and derive interpretable latent representations [5–7]. More recently, deep learning (DL) architectures, including feedforward neural networks (FNN), autoencoders, graph neural networks, and transformers, have been applied for multi-omics fusion by learning the nonlinear and hierarchical relationships across molecular layers [8–15]. These methods have achieved strong performance on a broad spectrum of cancer-related tasks, including prediction of cancer subtypes, metastatic status, pathological stages, survival outcomes, and therapeutic responses [8–10, 14–16]. Beyond predictions, many incorporate interpretability modules to discover biologically meaningful features, ranging from disease-associated subnetworks [17], dysregulated pathways [15], and most commonly, molecular biomarkers [7, 9–16, 18–21].

Despite substantial methodological progress, systematic evaluations of biomarker identification performance remain limited. Within individual method studies, evaluations of discovered biomarkers are typically proxy-based or qualitative, relying on functional enrichment analysis or literature support [7, 9–16, 18–21]. Such assessments are inferential and prone to selection bias, serving more as demonstrations of plausibility than as validation. Existing benchmark studies face similar limitations, as many rely on predictive performance as a proxy measure for assessing the selected biomarkers [22–32]. Also, although some utilize synthetic datasets with predefined discriminative features for direct and quantitative assessment, these datasets are often overly simplistic and fail to simulate the complexity and heterogeneity of real-world cancer data [6, 26, 28, 33, 34]. Moreover, although numerous multi-omics and DL-based methods have been developed for biomarker identification in recent years, they were often omitted from prior benchmark studies, leaving their biomarker identification capabilities largely unexamined [22, 24, 25, 29–33, 35, 36].

To address these gaps, we provide a systematic benchmark of computational methods for multi-omics biomarker identification. A major obstacle for previous studies has been the lack of comprehensive, task-specific biomarker sets for ground-truth evaluation. To address this, we collected Tier I biomarkers as defined by AMP/ASCO/CAP guidelines [37] from Clinical Interpretation of Variants in Cancer (CIViC) [38], MSK's Precision Oncology Knowledge Base (OncoKB) [39], and Cancer Genome Interpreter

(CGI) [40]. Using these gold-standard biomarkers collected from established cancer knowledge bases, our benchmarking enables a direct, quantitative, and clinically meaningful assessment of biomarker identification performance. We constructed five real-world task datasets from The Cancer Genome Atlas (TCGA) [41] using systematic filtering steps to ensure both computational feasibility and alignment between biomarkers and tasks.

In total, we benchmark 8 statistical and ML methods and 12 DL methods, covering a wide range of representative modeling and feature-identification approaches. Their performance was evaluated across diverse experimental settings and metrics. Several methods demonstrate strong capability to recover clinically validated biomarkers, frequently surfacing gold-standard biomarkers among the top ranks across experiments. Analysis of omics type contributions further characterizes biomarker- and method-specific biases. Transformer-based models integrating biological pathways with advanced *post hoc* feature-attribution methods achieve the highest accuracy and strong stability, whereas univariate feature-ablation methods underperform on both dimensions. Simulation experiments show that statistical and ML methods are more sensitive to shifts in the discriminative signals of ground-truth features, but discrepancies with real-data results underscore the indispensability of evaluating on real-world cancer cohorts. Additionally, we derive consensus panels of multi-omics cancer biomarkers by aggregating the results of top-performing methods, providing potential candidates for future investigations. Our open-source benchmarking datasets, results and pipelines constitute a valuable resource for future research in multi-omics biomarker discovery.

Results

Benchmarking design and evaluation framework

We conducted a systematic benchmark of 20 computational methods for multi-omics biomarker discovery, including 8 statistical and ML methods and 12 DL methods selected to represent the diversity of current algorithmic designs (Fig. 1c). These methods cover a diverse set of representative methodological choices, encompassing model architectures, feature selection or attribution, multi-omics fusion, and the incorporation of prior biological knowledge (Table 1, Methods).

To anchor evaluation in clinical relevance, we curated Tier I cancer biomarkers with Level A or B evidence, as defined by AMP/ASCO/CAP guidelines [37]. These biomarkers were harmonized across three established knowledge bases, including CIViC [38], OncoKB [39], and CGI [40] (Fig. 1a, Supplementary Tables 2, 3). After rigorous filtering, the curated biomarkers were mapped to five benchmark task datasets constructed from TCGA, including survival risk classification for Breast invasive carcinoma (BRCA), Lung adenocarcinoma (LUAD), and Colon and Rectum adenocarcinoma (COADREAD), and drug response prediction for Cisplatin with Bladder Urothelial Carcinoma (BLCA) and Temozolomide with Brain Lower Grade Glioma (LGG) (Fig. 1b, Methods). Summary statistics for all real-world datasets are provided in Supplementary Table 4.

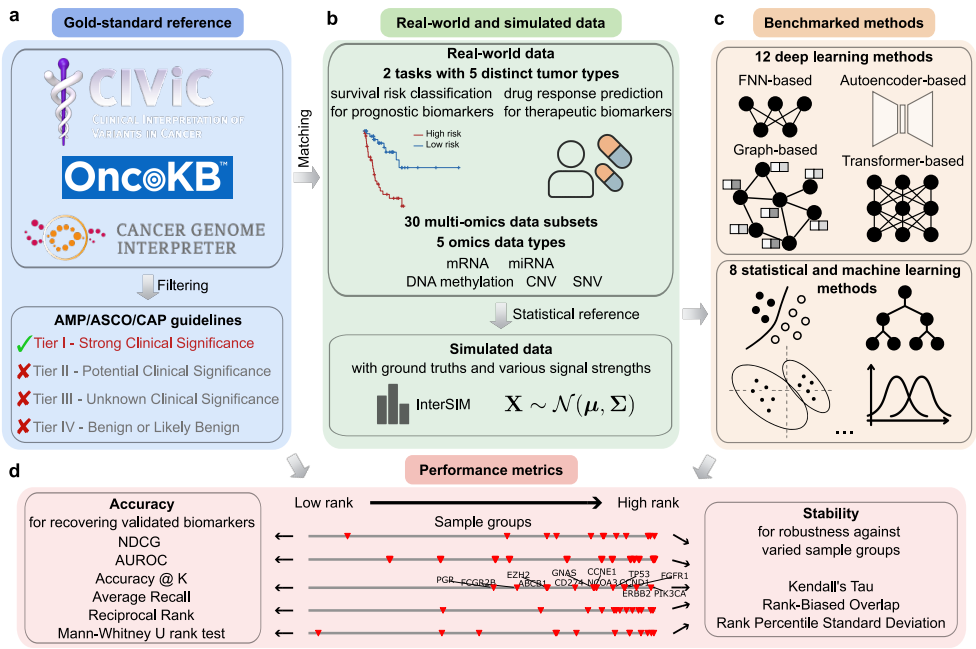

Fig. 1 Overview of benchmarking workflow. a, Biomarkers were collected from three oncology knowledge bases, including CIViC [38], OncoKB [39], and CGI [40], and those with Tier I evidence as determined by AMP/ASCO/CAP guidelines [37] were retained. **b,** Real-world data included 30 multi-omics data subsets encompassing five tasks and six omics combinations. Tumor types and drugs for prognostic and therapeutic tasks were matched to available biomarkers, and determined through a series of filtering rules (see Methods). Real-world data was used as statistical reference for simulated multi-omics data generation with InterSIM [42]. **c,** Twelve DL methods were benchmarked with different architectural choices including feedforward neural network (FNN), autoencoder, graph, and transformers. Eight statistical and ML methods were also benchmarked. **d,** Rank-based metrics for performance quantification. Six accuracy metrics and three stability metrics were used (Methods).

Because the benchmarked methods employed different omics combinations in their original case studies (Supplementary Table 5), we defined a unified set of omics combinations to ensure comparability across methods. For each task, we combined messenger RNA (mRNA) expression with two additional omics types from microRNA (miRNA), DNA methylation (DNAm), Copy Number Variation (CNV), and Single Nucleotide Variation (SNV). This produced six distinct omics combinations per task, leading to 30 multi-omics subdatasets that were each evaluated with five-fold cross-validation (Fig. 1b). To complement real-world data, we also generated simulated multi-omics data using InterSIM [42]. These simulations contain predefined discriminative features serving as ground truth, enabling controlled assessment of method sensitivity across a range of signal levels (Fig. 1b, Methods, Supplementary Table 6).

Biomarker identification performance was evaluated using nine complementary metrics that capture different aspects of accuracy and stability (Fig. 1d, Methods). For real cohorts with partial ground truth, rank-based metrics were applied to reward the prioritization of validated biomarkers without penalizing unvalidated candidates.

Table 1 Overview of the computational multi-omics biomarker identification methods benchmarked in this study.

Name	Model	Multi-omics fusion	Prior knowledge	Supervision	Biomarker identification	Ante hoc/ Post hoc	Year
			Deep learning				
P-Net [16]	Sparse neural network	Early	Biological pathways	Supervised	DeepLIFT	*Post hoc*	2021
GENIUS [20]	Convolutional neural network	Early	Physical genomic positioning	Supervised	Integrated Gradients	*Post hoc*	2023
TMO-Net [8]	Variational autoencoder	Intermediate	None	Supervised	Integrated Gradients	*Post hoc*	2024
CustOmics [9]	Variational autoencoder	Mixed	None	Supervised	SHAP	*Post hoc*	2023
MOGONET [10]	Graph convolutional network	Late	None	Supervised	Feature ablation	*Post hoc*	2021
MoAGL-SA [19]	Graph convolutional network and self-attention	Intermediate	None	Supervised	Feature ablation	*Post hoc*	2024
MORE [18]	Hypergraph and self-attention	Intermediate	None	Supervised	Feature ablation	*Post hoc*	2024
MOGLAM [12]	Graph convolutional network and self-attention	Intermediate	None	Supervised	Feature weights	*Ante hoc*	2023
GNN-SubNet [11]	Graph isomorphism network	Early	Protein-protein interaction (PPI) network	Supervised	GNNExplainer	*Post hoc*	2022
Pathformer [15]	Transformer encoder	Early	Biological pathways	Supervised	SHAP	*Post hoc*	2024
DeePathNet [14]	Transformer encoder	Early	Biological pathways	Supervised	SHAP	*Post hoc*	2024
DeepKEGG [13]	Transformer encoder	Intermediate	Biological pathways	Supervised	DeepLIFT	*Post hoc*	2024
			Statistical and machine learning				
MCIA [43]	Matrix decomposition	Intermediate	None	Unsupervised	Latent component weights	*Ante hoc*	2014
MOFA [5]	Matrix decomposition	Intermediate	None	Unsupervised	Latent component weights	*Ante hoc*	2018
GAUDI [7]	UMAP transformations and XGBoost/random forest	Intermediate	None	Unsupervised	SHAP	*Post hoc*	2025
DIABLO [6]	Sparse generalized canonical correlation analysis	Intermediate	None	Supervised	Latent component weights	*Ante hoc*	2019
asmbPLS-DA [44]	Sparse partial least squares discriminant analysis	Intermediate	None	Supervised	Latent component weights	*Ante hoc*	2023
Stabl [34]	Sparse regularized model	Intermediate	None	Supervised	Max selection frequency	*Ante hoc*	2024
GDF [17]	Greedy decision forest	Early	PPI network	Supervised	Gini gain	*Ante hoc*	2022
DPM [45]	Directional p-value merging	Intermediate	Inter-omics regulatory directions	Supervised	Merged p-values	*Ante hoc*	2024

For simulated datasets with complete ground truth, evaluation relied on hard metrics such as AUROC and accuracy. Stability was quantified as the similarity of ranked feature lists across cross-validation folds, using metrics that reflect top-rank consistency, global ranking agreement, and biomarker ranking variability (Methods).

Overview of benchmarked methods

The benchmarked methods encompass the representative methodological families adopted by multi-omics integration and biomarker identification approaches (Table 1, see Methods for details). Conventional statistical and ML methods (DIABLO [6], MOFA [5], MCIA [43], GAUDI [7], asmbPLS-DA [44], Stabl [34], GDF [17], DPM [45]) provide interpretable, often *ante hoc* feature scoring via latent components or sparse selection. DL architectures include feedforward and convolutional models (P-Net [16],

Table 2 The number of times a method ranks at least one gold-standard biomarker within the top 10, 50, or 100 out of all the 30 experiments[1].

	Survival BRCA			Survival LUAD			Survival COADREAD			Drug response Cisplatin[2]			Drug response Temozolomide[3]			Total		
Method	Top 10	50	100	Top 10	50	100	Top 10	50	100	Top 10	50	100	Top 10	50	100	Top 10	50	100
DL																		
DeePathNet	21	29	29	6	23	28	17	22	26	-	-	-	-	-	-	44	74	83
DeepKEGG	10	20	22	15	22	23	17	26	26	0	0	2	4	7	9	46	75	82
MOGLAM	10	10	12	10	10	10	10	10	10	0	0	0	10	10	12	40	40	44
TMO-Net	15	15	15	6	10	15	7	11	12	0	0	1	11	12	13	39	48	56
CustOmics	7	13	15	0	1	3	0	1	3	0	0	0	9	13	13	16	28	34
GENIUS	7	11	11	3	6	8	6	11	14	0	0	0	8	8	8	24	36	41
Pathformer	6	10	10	7	12	14	8	11	11	1	2	3	0	2	2	22	37	40
GNN-SubNet	3	3	4	1	2	5	1	4	6	0	0	0	1	3	4	6	12	19
P-Net	0	1	1	0	0	0	0	0	2	0	0	0	0	0	0	0	1	3
MOGONET	7	12	15	0	2	3	1	4	4	0	0	0	2	3	3	10	21	25
MORE	0	2	2	0	1	1	1	2	4	0	2	2	0	0	0	1	7	9
MoAGL-SA	1	1	2	1	1	2	0	0	0	0	0	0	0	0	0	2	2	4
Statistical and ML																		
DIABLO	15	15	15	8	9	9	7	8	10	0	0	0	15	15	15	45	47	49
GAUDI	15	15	15	7	11	12	1	5	10	0	0	0	13	15	15	36	46	52
GDF	1	19	23	1	11	19	0	7	14	0	1	1	0	0	0	2	38	57
Stabl	18	23	24	0	6	12	0	11	18	0	0	0	1	3	7	19	43	61
asmbPLS-DA	6	12	16	0	0	1	2	11	16	0	0	1	0	1	1	8	24	35
MOFA	0	0	0	0	5	9	0	0	0	0	0	0	0	0	0	0	5	9
DPM	0	2	3	0	1	3	0	0	0	0	1	2	0	0	0	0	4	8
MCIA	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	2

[1] Six omics combinations with five-fold cross-validation.

[2] Within TCGA-BLCA.

[3] Within TCGA-LGG.

[4] "-" indicates unavailability due to lack of biomarkers in the method's prior knowledge gene set.

GENIUS [20]), autoencoders (TMO-Net [8], CustOmics [9]), graph neural networks (MOGONET [10], GNN-SubNet [11], MoAGL-SA [19], MORE [18], MOGLAM [12]), and transformers that incorporate biological pathways (Pathformer [15], DeePathNet [14], DeepKEGG [13]). Most DL methods rely on *post hoc*, model-agnostic feature attribution approaches (DeepLIFT [46], Integrated Gradients [47], SHAP [48], GNNExplainer [49]) for identifying biomarkers, while *ante hoc* DL methods such as MOGLAM are inherently designed to learn feature weights during training. For comparability, all model outputs were converted to gene-level ranking lists prior to metric calculation (Methods).

Accuracy on real-world data

Accurate identification of clinically relevant candidates from high-dimensional molecular space represents a hallmark of a successful biomarker discovery method. Therefore, we first examined the performance of each method for the recovery of clinically validated biomarkers in real cancer cohorts. Across five TCGA-derived tasks and six omics combinations, we compared methods using rank-based metrics that quantify different aspects of biomarker prioritization: Average Recall (AR) for overall prioritization, Reciprocal Rank (RR) for early retrieval of a single biomarker, and Normalized Discounted Cumulative Gain (NDCG) for early recovery of multiple biomarkers. These were complemented by a Mann-Whitney U test on global rankings and by success counts within the top 10, 50, and 100 ranks (Fig. 2a,b, Table 2, see Methods for

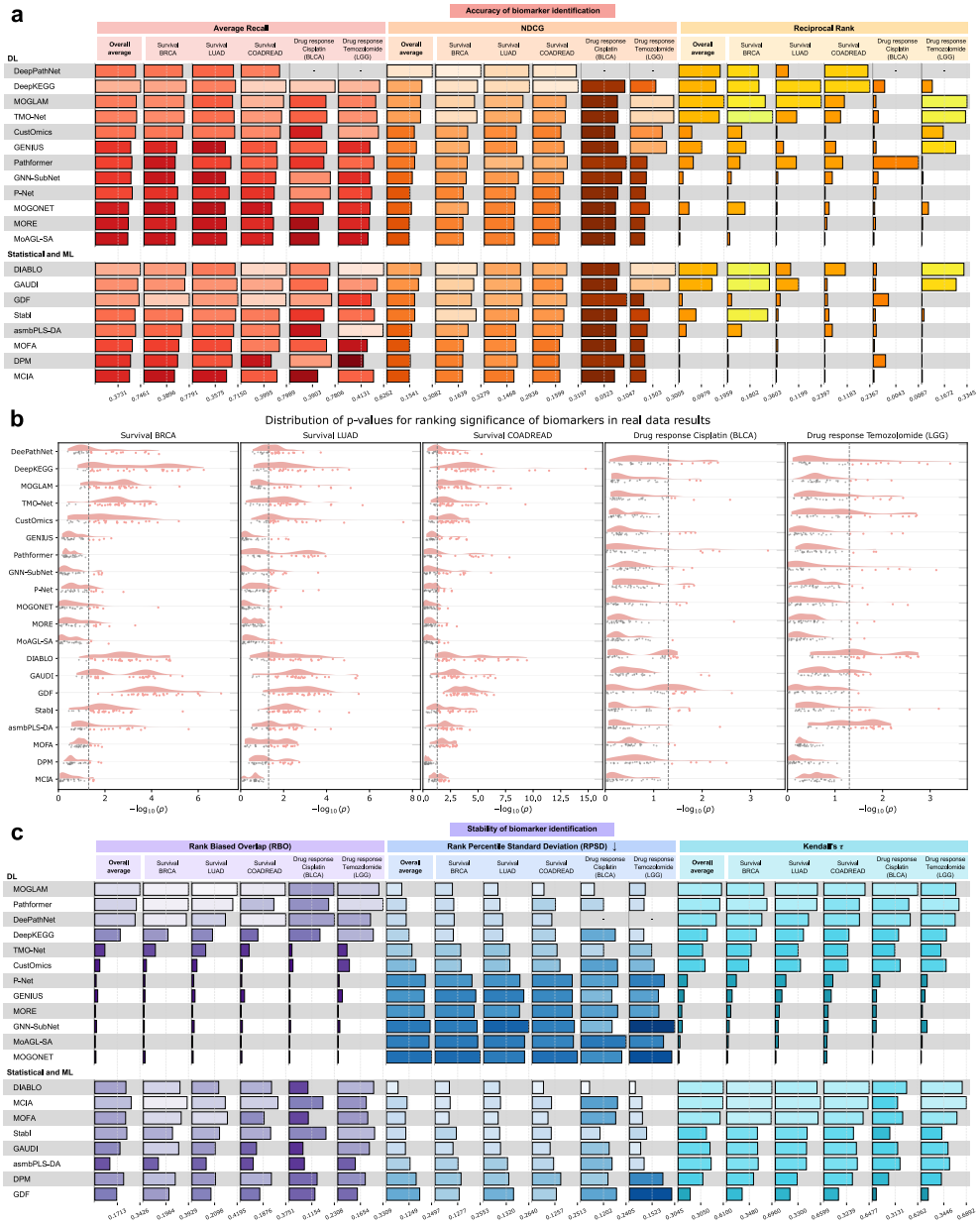

Fig. 2 Overall performance on real-world data. a, Accuracy of methods for all task datasets and metrics. For a particular method and task, the height of the bar represents the averaged metric value of the cross-validation mean across all six omics combinations. **b,** Mann-Whitney U test p-values (log10-transformed). Each dot represents a single experiment corresponding to a specific omics combination and cross-validation fold. **c,** Stability of methods for all task datasets and metrics. For a particular method and task, the height of the bar represents the averaged metric value across all six omics combinations. NDCG: Normalized Discounted Cumulative Gain. "-" indicates unavailability due to lack of biomarkers in the method's prior knowledge gene set.

details). Detailed results by each omics combination are reported in Supplementary Figs. 1-5 and Supplementary Figs. 11-15.

The results identified a group of relatively accurate methods, including DeePathNet, DeepKEGG, MOGLAM, TMO-Net, DIABLO, and GAUDI, which frequently recovered validated biomarkers among the top ranks, with a top-10 success rate of around 26.7% (about 40 out of 150 experiments; Table 2) as well as comparable overall accuracy (Fig. 2a,b). Within this group, the transformer-based, pathway-informed models DeePathNet and DeepKEGG formed a distinct top tier. In survival prediction tasks, DeePathNet and DeepKEGG recovered a validated biomarker within the top 10 in around half of all experiments (44/90 and 42/90, respectively) and achieved the highest NDCG scores, indicating a strong capacity to identify multiple biomarkers simultaneously. Expanding the ranking threshold from top-10 to top-100 further accentuates this advantage, increasing the success rate of DeePathNet and DeepKEGG to around 50% across all tasks and 90% in survival tasks, whereas MOGLAM, TMO-Net, DIABLO, and GAUDI showed minimal improvement (Table 2). This divergence is consistent with our omics contribution analysis (Fig. 3), suggesting that DeePathNet and DeepKEGG exploit a broader spectrum of omics combinations to retrieve relevant signals.

At the lower end of performance, several methods relying on feature ablation to estimate importance, including MOGONET, MoAGL-SA, and MORE, rarely recovered validated biomarkers and showed weak performance across both success counts and accuracy metrics (Fig. 2a,b, Table 2). Their underperformance likely arises from the complementarity induced by the inherent correlations between multi-omics features, where ablation of a single feature is compensated by correlated ones [50]. Among statistical and ML methods, MOFA, DPM, and MCIA consistently ranked lowest across tasks and metrics. DPM aggregates per-feature Mann-Whitney test statistics across omics and is intrinsically univariate, unable to capture multivariate structures in the omics feature space. MOFA and MCIA, both unsupervised frameworks without utilizing sample-level label information, rank features by latent component weights via variance-based matrix decomposition algorithms, prioritizing high-variance components that do not necessarily align with clinical endpoints (Supplementary Methods). In contrast, GAUDI provides a counterexample. Despite also being unsupervised, it combines tree-based models (XGBoost or random forest) with SHAP-based attribution, leading to substantially better performance and highlighting the benefit of more advanced attribution methods even in unsupervised settings (Fig. 2a,b, Supplementary Methods).

These comparisons show that deep neural networks are not intrinsically superior to conventional statistical and ML models for learning from complex multi-omics data. Well-designed multivariate statistical and ML methods such as DIABLO can achieve accuracy comparable to top-performing DL methods when optimized for correlated data. However, DL methods appear to have a higher performance ceiling. DeePathNet and DeepKEGG are not only among the most accurate methods, but also maintained this performance across multiple cancers and omics combinations (Supplementary Figs. 1-5), implying that encoding biological pathway structures into

transformers and utilizing modern feature attribution methods is a highly competitive design choice. Still, no single recipe guarantees optimal performance. Pathformer, which shares a transformer backbone with DeePathNet and DeepKEGG, ranked only mid-tier. Although most graph-based methods were suboptimal, MOGLAM showed strong performance. Most unsupervised methods exhibited weak performance, yet GAUDI outperformed several supervised approaches. The incorporation of biological knowledge does not necessarily induce high accuracy. Thus, methodological design choices, rather than model family, primarily determine success in recovering validated biomarkers.

Task complexity also affects performance. Almost every method underperformed on the BLCA cisplatin drug-response prediction task, particularly on top-rank-biased metrics such as NDCG and RR, indicating the increased difficulty of surfacing ERCC2 relative to other targets (Fig. 2a, Supplementary Fig. 4, 14). This likely stems from two factors. Clinically, ERCC2's predictive signal is specific to cisplatin-based neoadjuvant chemotherapy with pathologic downstaging endpoints, hence mixing metastatic settings, radiographic endpoints, or carboplatin treatments introduces label noise [51]. Biologically, only helicase-domain loss-of-function ERCC2 variants truly abrogate nucleotide-excision repair and confer cisplatin sensitivity, thus collapsing all ERCC2 mutations into a single feature further obscures the signal [52]. Nevertheless, Pathformer appeared as a notable exception, as it was the only method consistently identifying ERCC2 within the top ranks when using the mRNA+CNV+SNV omics combination (Supplementary Figs. 4, 14). Relative to other transformer-based methods, Pathformer employs a more expressive criss-cross attention mechanism and architectural design (Supplementary Methods). This suggests that more sophisticated architectures can sometimes extract weak signals from noisy data, although such advantages may trade off against performance on less difficult tasks.

Stability on real-world data

The reproducibility of biomarkers is essential for clinical translation. Thus, we next evaluated the stability of each method, defined as the consistency of its biomarker rankings across resampled data for the same task. Methods that are more stable are expected to produce biomarker sets with greater fidelity across multiple real-world cohorts. We quantified stability using the five cross-validation splits per task, per omics combination, computing average pairwise similarity between rankings. Rank-biased Overlap (RBO) assesses top-rank consistency, Rank Percentile Standard Deviation (RPSD) measures the dispersion of biomarker rankings across splits, and Kendall's τ quantifies the overall consistency across full rankings (Fig. 2c, see Methods for details). Stability results by each omics combination are reported in Supplementary Figs. 6-10.

In general, we found that statistical and ML methods were markedly more stable than DL methods, whereas the latter exhibited greater variability, encompassing both the most and least stable methods (Fig. 2c). A subset of DL methods, including MOGLAM, Pathformer, and DeePathNet, surpassed all statistical and ML methods, achieving both high top-rank stability (RBO) and strong global stability (low RPSD and high Kendall's τ). By contrast, DIABLO, which is the most stable statistical and ML method, reached the best global stability but failed to attain comparable RBO,

suggesting that although its rankings are internally coherent, they are less consistent at top than those produced by the most stable DL methods. At the opposite extreme, several DL methods, including P-Net, GENIUS, GNN-SubNet, MOGONET, MoAGL-SA, and MORE, were less stable than even the weakest statistical and ML methods. Their RBO scores fall below that of asmbPLS-DA, the least stable statistical and ML baseline at the top ranks, and their global stability was weaker than that of GDF, which had the highest RPSD and lowest Kendall's τ among statistical and ML methods (Fig. 2c). These results reflect the sensitivity of deep neural networks training, as small changes in the input samples or random seed can drive stochastic gradient-based optimization into distinct solution basins [53], producing substantially different top-ranked features identified by *post hoc* explanation methods.

Two design choices suggested by the most stable DL methods appear to improve stability. First, incorporating biological pathway structure into model architecture improves stability, particularly for transformer-based models that explicitly encode pathway knowledge (Pathformer, DeePathNet, and DeepKEGG). Second, regularization techniques and adaptive learning can make models less sensitive to sample variances. For example, MOGLAM learns graphs adaptively from the data and employs explicit regularization, achieving high stability across tasks. In contrast, DL methods that lack such architectural or regularizing constraints tended to be among the least stable (Fig. 2c).

Considering statistical and ML methods and DL methods jointly, methods that fail to capture global dependencies were generally less stable and also underperformed in accuracy. GNN-SubNet, P-Net, and GENIUS exemplify this pattern, showing poor stability at both top and global rankings (Fig. 2c). In GNN-SubNet, genes are represented as graph nodes, but the message-passing does not effectively propagate information over long distances in the graph. P-Net uses a sparse neural network that largely suppresses communication across genes and pathways, and GENIUS relies on convolutional networks, which emphasize local receptive fields and inherently limit long-range context [54]. These observations indicate that insufficient modeling of long-range dependencies impair cross-signal compensation, thereby diminishing the stabilizing contribution of global molecular context.

Joint analysis of accuracy and stability revealed a heterogeneous landscape rather than an inherent trade-off. Autoencoder-based methods such as TMO-Net and CustOmics achieved reasonable accuracy but were comparatively unstable. Pathformer exhibited high stability but only moderate accuracy. Conventional statistical methods, including MCIA and MOFA, were inaccurate but highly stable. Nevertheless, the overall pattern does not support a universal accuracy-stability trade-off, as methods could succeed or fail along either metric. MOGLAM, DeePathNet, DeepKEGG, and DIABLO were both accurate and stable, whereas P-Net, GNN-SubNet, MOGONET, MORE, MoAGL-SA, and DPM underperformed on both dimensions (Fig. 2).

Altogether, these results illuminate several design principles for guiding future method development. Methods perform best when they (i) encode biological structure, especially pathways, (ii) enable global interactions among features, (iii) attribute feature importance in a multivariate way, and (iv) employ regularization or adaptive learning techniques to stabilize optimization. Approaches that rely on univariate

importance, whether through individual feature ablation or per-feature statistical tests as in DPM, systematically miss multivariate correlations and underperform. Likewise, sparse neural networks, convolutional neural networks, and gene-gene graph models, exemplified by P-Net, GENIUS, and GNN-SubNet, constrain long-range molecular communication that characterizes omics biology. In contrast, transformer-based models facilitate global pathway cross-talks via self-attention, achieving high accuracy and stability with modern feature attribution methods (Fig. 2, Supplementary Methods). Thus, accuracy and stability are attainable simultaneously, and the decisive factor is the specific methodological design rather than adherence to any particular model family.

Contribution of omics types to biomarker identification

Different molecular layers characterize complementary aspects of tumor biology, but their relative contributions to biomarker discovery remain unclear. Although most methods evaluated only one or two omics combinations in their original case studies (Supplementary Table 5), a systematic comparison across a broader set of omics combinations is essential for uncovering underlying biological mechanisms and informing both method development and experimental design. Consistent with this need, we observed substantial performance differences across omics combinations (Supplementary Figs. 1–15). In many cases, the omics combinations chosen in the original studies did not yield superior performance compared to untested alternatives, supporting our inclusion of an extended set of omics combinations. We also found that for almost all methods, high accuracy was confined to a specific omics combination, with DeePathNet and DeepKEGG being notable exceptions. In survival tasks, DeePathNet and DeepKEGG identified biomarkers within the top 10 ranks across nearly all omics combinations (RR ≥ 0.1; Supplementary Figs. 1c, 2c, 3c). To further dissect these results, we analyzed omics type contributions at both the biomarker and method level.

First, to pinpoint which omics types drive biomarker discovery, we examined the contributions of all five omics types for each identified biomarker (Fig. 3a). Here we define a biomarker as "identified" when it appeared within the top 1% of the gene-level ranking list. For each such case, we recorded the dominant omics type corresponding to the type with the highest feature score (Methods). Figure 3a provides a comprehensive summary for all experiments, while results per each omics combination are summarized in Supplementary Figs. 16-21.

We found that most identified biomarkers were driven by one predominant omics layer. For instance, for the temozolomide drug response prediction task within lower-grade glioma (LGG), the top-performing experiments consistently required SNV data. IDH1 was typically ranked at the top owing to SNV (Fig. 3a, Supplementary Figs. 17, 19, 21), consistent with the causal IDH1 R132 mutation and its co-occurrence with MGMT promoter methylation, the canonical predictor of temozolomide benefit [55]. In breast cancer (BRCA) survival, TP53 was almost always recovered through SNV (Fig. 3a, Supplementary Figs. 17, 19, 21), reflecting the prevalence of missense mutations in the DNA-binding domain that stratify poor outcomes, which is especially common in aggressive basal-like/TNBC [56, 57]. By contrast, FCGR2B was primarily identified through mRNA expression, particularly by Stabl and asmbPLS-DA (Fig. 3a,

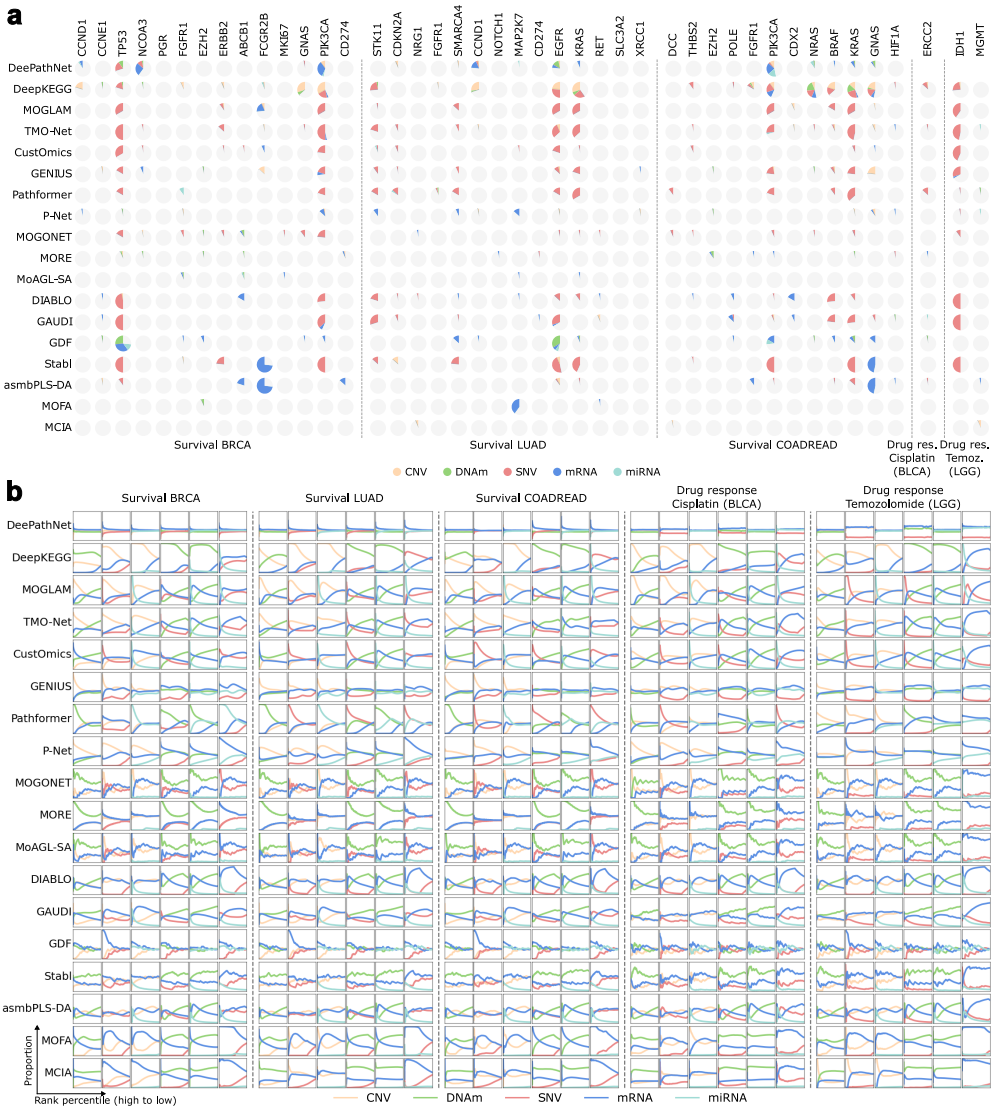

Fig. 3 Omics type contributions in biomarker identification. a, Pie charts showing the omics types driving an identification of a biomarker. Each pie has 30 portions, one for a specific experiment with a particular omics combination (six in total) and a cross-validation fold (five fold), with each portion colored by the dominant omics type when the biomarker is ranked in the top 1% (Methods), and left gray otherwise. **b**, omics type proportions (y-axis) from high (left) to low (right) ranks (x-axis). Each task contains six columns corresponding to the six omics combinations. Proportions were calculated at each of the 1000 rank percentile cutoffs. The ranks were based on the raw feature scores before gene-level conversion (see Methods for details), which were concatenated from the five score lists output by cross-validation.

Supplementary Figs. 16-21), where transcriptomic signatures reflect immune activation rather than direct tumor genetics in breast cancer prognosis [58]. These examples illustrate how the dominant layer responsible for identification often reflects underlying tumor biology.

Nevertheless, several biomarkers were occasionally identified through non-canonical omics layers. For example, IDH1 was identified by DeepKEGG through CNV (Supplementary Fig. 16), TP53 by DeePathNet and DeepKEGG through DNAm (Supplementary Figs. 16 and 19), and FCGR2B by DeepKEGG, MOGLAM, and TMO-Net through CNV or miRNA (Supplementary Figs. 16, 18, 21). These anomalous identifications underscore the complex interplay across omics modalities and the potential for cross-omics compensation.

Second, because methods can identify biomarkers through omics layers that do not commonly drive the underlying biology, we next interrogated whether methods have intrinsic preferences for specific omics types. To achieve this, we summarized the proportions of each omics type at different rank percentile cutoffs using each method's raw feature scores or rankings, for each omics combination and task (Fig. 3b). Some methods demonstrated clear biases. As an example, GDF relied almost exclusively on mRNA, DNAm, and miRNA, whereas CNV and SNV rarely appeared as dominant (Fig. 3). Consequently, whereas most methods identified TP53 primarily via SNV, GDF identified TP53 through mRNA, DNAm, and miRNA (Fig. 3a, Supplementary Figs. 16-21). Other methods exhibited more diverse patterns. Notably, DeePathNet, DeepKEGG, P-Net, and GAUDI were the only four methods for which all five omics types appeared as dominant contributors across biomarkers, and three of these (DeePathNet, DeepKEGG, and GAUDI) were among the top performers. DeePathNet and DeepKEGG, in particular, displayed especially diverse contributions. For a single biomarker, as many as four different omics types could be dominant (e.g., PIK3CA with DeePathNet; NRAS, KRAS, and GNAS with DeepKEGG in the COADREAD survival task). By contrast, most other methods typically have a single dominant omics type per biomarker (Fig. 3a).

As it is difficult to know *a priori* which omics combination will perform best, these findings indicate that including a broader set of omics types is beneficial for biomarker discovery from multi-omics data. Broad inclusion increases the likelihood of incorporating the mechanistic information relevant to specific biomarkers and allows methods to perform optimally, whereas restricting experiments to a single omics combination may overlook biomarkers whose key information resides in an untested molecular layer.

Performance on simulated data

A major challenge in benchmarking biomarker identification methods is the lack of complete ground truths. Our use of curated cancer biomarker knowledge bases, task matching, and rank-based metrics offers a practical solution to this for real data. In parallel, as a complement to real data, we generated multi-omics simulations with predefined discriminative features, enabling evaluation with standard metrics such as AUROC and accuracy, and varied the mean-shift of ground-truth features to modulate their discriminative strength (Methods). It was observed that simulations reproduced some trends from real-data results. Feature-ablation DL methods (MOGONET,

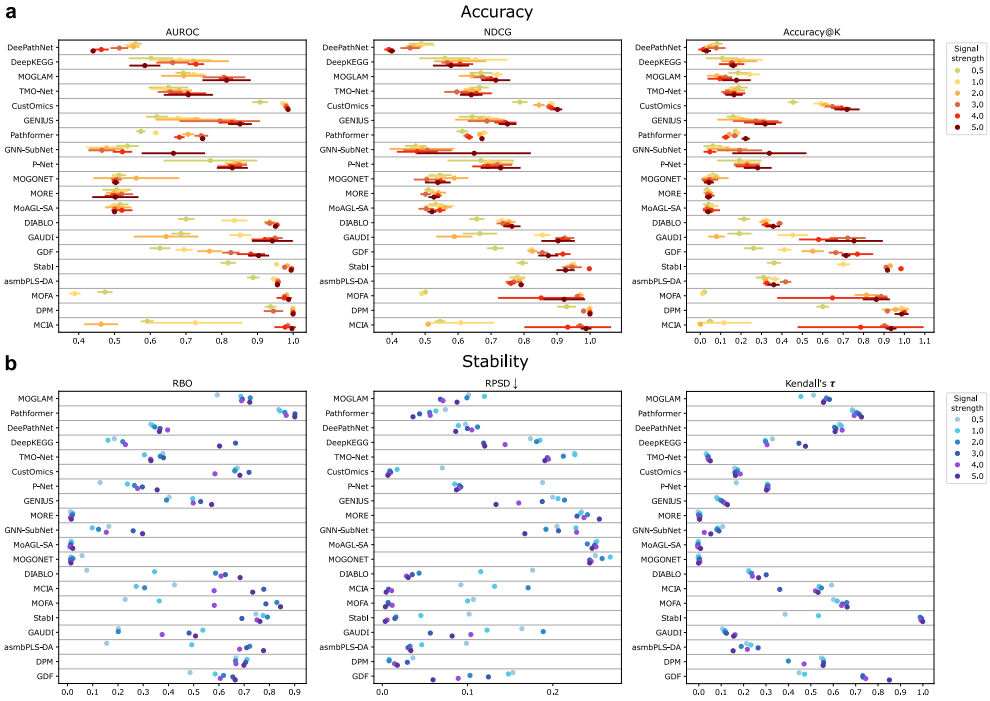

Fig. 4 Performance on simulated data. Cleveland dot plots showing biomarker identification accuracy (**a**) (bars represent standard deviation across five-fold cross-validation) and stability (**b**). Colors represent a particular biomarker signal strength (see Methods for details).

MoAGL-SA, MORE), which were weak on real cohorts, also underperformed in simulations. DIABLO remained among the top performers (Fig. 4), consistent with its capacity to model coherent multivariate structure. These consistencies increase confidence in methods whose performance appears robust to changes in data generation complexity.

However, discrepancies were more common than agreements. DeePathNet, DeepKEGG, MOGLAM, and TMO-Net became less competitive, and many statistical and ML methods (e.g., MCIA, MOFA, DPM) improved substantially. Within either the DL or the statistical and ML group, accuracy rankings differed from those within real cohorts (Fig. 4a). These shifts indicate that simplified generative mechanisms introduce a bias toward methods optimized for less noisy discriminative effects, while attenuating the relative advantages of models intended to recover more complex relationships.

Across discriminative strengths, statistical and ML methods exhibited greater variability than DL methods (Fig. 4), indicating greater sensitivity to the magnitude of predefined effects. This is consistent with the fact that simulated datasets lack the heterogeneity and cross-omics dependencies observed in real tumors, which tends to favor more conventional statistical and ML methods that assume relatively clean

underlying structure. In contrast, the capacity of DL models to utilize subtle or composite multivariate patterns becomes evident primarily in real datasets characterized by substantial heterogeneity and nonlinear cross-omic relationships. These discrepancies underscore the limitations of relying solely on synthetic discriminative features as benchmarking ground truth. Although widely used in prior work [6, 26, 28, 33, 34], simulation-based evaluations did not align with assessments based on real cohorts and clinically established biomarkers. Simulated data remain useful for controlled tests beyond accuracy and stability, but they may not serve as a substitute for real-world benchmarking, which is required to assess clinical relevance and to avoid conclusions driven primarily by properties specific to the simulation framework.

By contrast, stability patterns on simulated data were more consistent with those from real cohorts (Fig. 4b, Fig. 2c), suggesting that stability is driven more by methodological design than by the numerical properties of the data. Even so, statistical and ML methods showed larger stability shifts as discriminative strength varied, indicating their sensitivity to the simple structures present in simulated datasets.

Consensus multi-omics biomarker panels

Our unified benchmarking framework, which evaluates multiple methods across shared tasks and omics combinations, provides a basis for deriving consensus biomarker panels. Therefore, we prioritized candidate biomarkers within each omics type by aggregating results from multiple top-performing methods. Because methodological choices and omics-specific dependencies vary widely across methods, rankings show only modest concordance (Supplementary Fig. 22). Integrating these rankings can therefore uncover robust biomarker candidates that are less susceptible to method-specific biases and more representative of integrative molecular signatures, producing an accurate and reproducible consensus.

For each benchmarked task, we selected top-performing method-omics combination pairs, and derive consensus rankings via Robust Rank Aggregation (RRA) [59] (Methods). RRA converts cross-method agreement into per-feature p-values, accommodates partial rankings, and alleviates outliers and method-specific biases. The consensus provides statistically robust significance levels with controlled false discovery rates (FDRs), enabling distillations of sparse, accurate and reproducible biomarker candidate panels (Table 3). As an example, in BRCA survival, the consensus identifies 10 mRNA biomarkers (FDR $< 10^{-3}$), 10 miRNA biomarkers (FDR $< 10^{-3}$), 14 DNA methylation (DNAm) biomarkers (FDR $< 10^{-3}$), 14 CNV biomarkers (FDR $< 10^{-4}$), and 10 SNV biomarkers (FDR $< 10^{-6}$) (Table 3). Derived from diverse methods and omics combinations that most accurately and consistently recover clinically validated biomarkers, these panels provide high-confidence candidates for biological and clinical investigation. Full consensus rankings for all tasks and multi-omics features are available in Supplementary Data 1.

Discussion

In this study, we presented a systematic benchmark for computational biomarker discovery from multi-omics data, evaluated against clinically validated reference

Table 3 Consensus multi-omics biomarker panels by task and omics type. Biomarkers already present in the gold-standard sets are omitted.

Task	Omics	p-value	Consensus panel
Survival BRCA	mRNA	$< 10^{-3}$	GNG5, MYC, TPGS1, NTRK3, GATA3, MAP2K6, MLPH, MTOR, PRM1, NPHS2
	miRNA	$< 10^{-3}$	hsa-let-7d, hsa-mir-16-1, hsa-mir-326, hsa-mir-502, hsa-mir-301b, hsa-mir-18a, hsa-mir-130b, hsa-mir-30e, hsa-let-7g, hsa-mir-331
	DNAm	$< 10^{-3}$	SEC22B, IL11RA, AL121900.1, NCAPG, DZIP3, PRUNE1, WDR12, BMPR2, LDHA, SMIM26, AL162231.3, TRMT11, ZNF641, ZNF331
	SNV	$< 10^{-6}$	CDH1, GATA3, TTN, SLITRK3, RYR2, HMCN1, RUNX1, KMT2C
	CNV	$< 10^{-4}$	SLC25A32, MYC, PIP4P2, SLAMF7, MTSS1, KLHL38, PLEKHO1, TRHR, RIPK2, C1orf54, TG, PABPC1, ARNT, DSCC1
Survival LUAD	mRNA	$< 10^{-6}$	SNRNP70, RHOT2, PIDD1, ZNF692, TEPSIN
	miRNA	$< 10^{-4}$	hsa-mir-143, hsa-mir-191, hsa-mir-210, hsa-mir-590, hsa-mir-93, hsa-mir-328, hsa-mir-1307
	DNAm	$< 10^{-5}$	RPL14, UCHL3, RPS8, AL356512.1, LIN7C, C6orf62, TRAF6, CEP350, BMT2, AL353708.1, RPP40, AP2S1, AC007036.5, IL27RA, ITGB3BP, MFNG, TTC8, BBS12, AC022098.2, FBXW11, AC051619.7, PLEKHF2, PLA2G12B
	SNV	$< 10^{-6}$	TLR4, TP53, APOB, TNR, RELN, COL11A1, TNN, LAMA2, RYR3, KEAP1
	CNV	$< 10^{-3}$	CREB3L4, ATP1B1, PKLR, CACNA1S, EFNA4, FCGR1A, SEC61G, CTNNA2, CRTC2, ASAP1, PHKG1, WDPCP, H4C15, PSPH, ATF2, CTSK, RGS2, RAB13, B3GNT2, MAPK13, THBS3
Survival COADREAD	mRNA	$< 10^{-5}$	MAPK1, GRB2, JAK1, ITGB1, NCOA3, MAPK8, GSK3A, MAP3K1, IKBKB
	miRNA	$< 10^{-4}$	hsa-mir-191, hsa-mir-17, hsa-mir-454, hsa-mir-93, hsa-mir-186, hsa-let-7d, hsa-mir-130b, hsa-mir-942, hsa-mir-33a
	DNAm	$< 10^{-4}$	WNT1, SDHC, AL359504.1, PRKACB, MIR618, TRAF6, U47924.1, MIR141, MIR200CHG, CALML3-AS1, CALML3, EGFEM1P, MAPK13, AC018521.2, CRLF3, RNF182, MIR200C, EGOT, AP000911.1, FGF5, CFAP299
	SNV	$< 10^{-6}$	APC, TP53, GRIK2, DAAM2, PCLO, KMT2C, TTN, TNRC6B, FLG, ARID1A, VCAN, OBSCN, ADGRV1, RYR1, KCNQ2, MUC16, NEB
	CNV	$< 10^{-4}$	PCK1, PEDS1, NCOA3, NFATC2, BCL2L1, MYLK2, COX4I2, PLCG1, GDF5, ACSS2, ITCH, PABPC1L, STK4, YWHAB, SDC4, RBPJL, ADA, SGK2
Drug response Cisplatin (BLCA)	mRNA	$< 10^{-4}$	PRF1, PRICKLE1, TAF11L11
	miRNA	$< 10^{-4}$	hsa-mir-133a-1, hsa-mir-133b, hsa-mir-1-2, hsa-mir-590, hsa-mir-16-2, hsa-mir-186, hsa-mir-1-1, hsa-mir-671, hsa-mir-133a-2, hsa-mir-92a-2, hsa-mir-454, hsa-mir-197, hsa-mir-103a-1, hsa-mir-19a, hsa-mir-423, hsa-mir-99a, hsa-let-7d
	DNAm	$< 10^{-4}$	UBXN7, AC104581.2, ESRRA, EXOSC5, BCKDHA, AC078916.1, ZNF195
	SNV	$< 10^{-5}$	SLC12A5, KDM6A, AKAP9, TTN, RYR3, TP53, ARID1A, TNFAIP3, KMT2C, TECTA, ERBB2, OBSCN, PTEN, DNAH3
	CNV	$< 10^{-5}$	CDV3, CLDN16, CLDN11, ANAPC13
Drug response Temozolomide (LGG)	mRNA	$< 10^{-4}$	GPR143, PGAP1, PRICKLE3, RASAL1, FBXO17, PCDH15, PDPN, FKBP9
	miRNA	$< 10^{-4}$	hsa-mir-324, hsa-mir-331, hsa-mir-107, hsa-mir-628, hsa-mir-185, hsa-mir-22, hsa-mir-27b, hsa-mir-23b, hsa-mir-29b-1, hsa-mir-874, hsa-mir-24-2, hsa-mir-24-1, hsa-mir-769
	DNAm	$< 10^{-4}$	EEF1B2P4, VCL, PTGER1, AC084026.2, CKMT2, PDE2A-AS2, MT1M, EPOR, BISPR, NEAT1, ARAP3, B3GNT5
	SNV	$< 10^{-4}$	ATRX, TP53, CIC, TTN, NOTCH1, FAT2, FLG, EGFR, OR51F2, ITPR3
	CNV	$< 10^{-4}$	BAMBI, ARMC4, DNAJC1, OTUD1, CCNY, ARHGAP21

biomarkers. The results uncovered certain methodological design choices that generally exhibited strong or weak accuracy and stability in biomarker identification, providing guiding principles for future method development. Analysis of omics type contributions indicate that biomarker identification is strongly influenced by the molecular modalities included and by model-specific biases, underscoring the importance of utilizing diverse omics layers for biomarker discovery. Results on simulated data showed that simplified generative assumptions can produce baseline results differing substantially from those observed in real cohorts, emphasizing the indispensability of real-world evaluations when assessing translational applicability. The consensus biomarker panels derived from the most accurate and stable methods and omics combinations provide reproducible candidates that may inform future biological and clinical investigations.

Challenges are identified that warrant future work. First, the incompleteness of available ground truth remains a fundamental limitation for real-world benchmarking. Although we compiled biomarkers from multiple curated cancer knowledge bases [38–40] and used rank-based metrics to mitigate the effects of missing annotations, biomarkers that have not yet been discovered or clinically characterized are naturally absent from evaluation. This is particularly relevant for therapeutic tasks, where the number of clinically validated biomarkers is more limited, which may lead to underestimation when models identify biologically relevant but undocumented candidates. As the coverage of cancer biomarker resources expands, future benchmarks will be able to incorporate broader reference sets. In addition, while our study focused on Tier I biomarkers [37] to maintain translational relevance, the inclusion of lower tiers in future work may help evaluate each method's capacity to detect plausible but currently unvalidated candidates.

Methods

Selection of benchmarked methods

We searched the literature through a combination of keywords regarding multi-omics integration and biomarker identification, and gathered 82 computational multi-omics biomarker identification methods into an extensive list in Supplementary Table 1. To select the methods for benchmarking, we performed strict filtering based on a number of essential criteria covering three aspects: methodological compatibility, accessibility/usability, and computational expense. Specifically, methods were included in benchmarking if they satisfy the following criteria:

Methodological compatibility. (1) The method should be a general-purpose approach for multi-omics data integration, capable of accepting other omics types rather than being limited to the omics types demonstrated in the original case studies. Such flexibility is an important characteristic of a well-developed multi-omics integration method, and is also essential for fair performance comparisons, as it ensures that all benchmarked methods are evaluated on the same set of omics combinations. (2) The biomarker identification module must output a global feature ranking or scoring list and should not produce an excessively sparse set of biomarkers. If only a small number of features receive non-zero scores, there is a high likelihood that these features will be unvalidated biomarkers absent from reference knowledge bases, while validated biomarkers cannot be evaluated using rank-based metrics. (3) For methods that generate sample-level predictions, classification must be supported.

Accessibility and usability. (1) Code required to run the method must be publicly available and complete, covering both the model training/inference and biomarker identification components. (2) The code should run without major bugs, and if provided as a package, it should install cleanly without critical errors. (3) Code should be implemented in Python or R and engineered in such a way that it can be integrated into our benchmarking pipeline either directly or after necessary modifications, without requiring prohibitive re-engineering efforts. (4) If the method relies on preprocessed prior knowledge data files, these files must be available for download. (5) Hyperparameters or guidance for setting them should be clearly documented, either in the manuscript, code, or accompanying documentation.

Computational expense. The method must complete execution within one day on a typical real-world dataset (approximately 400 samples $\times$ 40,000 features) for a single run.

After examining each gathered method in Supplementary Table 1, we selected 20 methods satisfying all the above criteria for benchmarking. The selected methods, along with their key characteristics, are summarized in Table 1.

Overview of benchmarked methods

The methods included in this benchmark span a diverse range of methodological designs (Table 1). Here, we provide an overview of the selected methods. More details for each are further provided in Supplementary Methods.

Deep learning methods

Based on architectural backbones, deep-learning-based methods can be broadly grouped into four categories: feedforward neural network (FNN)-based (GENIUS [20], P-Net [16], autoencoder-based (TMO-Net [8], CustOmics [9]), graph-based (MOGLAM [12], GNN-SubNet [11], MOGONET [10], MORE [18], MoAGL-SA [19]), and transformer-based (DeePathNet [14], DeepKEGG [13], Pathformer [15]) (Fig. 1c).

For FNN-based methods, P-Net constructs a hierarchical sparse neural network based on gene-pathway and pathway-biological process connections [16]. GENIUS employs convolutional neural networks (CNNs) and integrates prior biological knowledge by converting multi-omics data into "gene images", using genomic coordinates to define spatial arrangements across omics layers as image channels [20].

Autoencoder-based methods learn low-dimensional joint representations across omics. Standard dense neural networks are typically employed as backbones. TMO-Net applies multiple variational autoencoders for intermediate fusion, capturing both self- and cross-modal associations [8]. CustOmics uses a hierarchical mixed-integration strategy, learning omics-specific sub-representations via separate autoencoders, which are then combined through a central variational autoencoder [9].

Graph-based methods can be generally divided into two types. Models using samples as graph nodes typically build a sample-level similarity graph for each omics type and perform fusion afterwards. MOGONET fuses the predictive output of each omic-specific graph convolutional network (GCN) in the label space, while MoAGL-SA, MORE, and MOGLAM use GCN or hypergraphs with self-attention mechanism to learn a joint graph representation of multi-omics features [10, 12, 18, 19]. The method using genes as graph nodes (GNN-SubNet) constructs the graph using protein-protein interactions (PPIs) network as prior knowledge, learning graph feature representations followed by global-pooling for sample-level predictions [11].

Transformer-based methods integrate pathway knowledge into the model architecture. DeePathNet groups features into pathways and applies self-attention over pathway-level embeddings. Pathformer transforms gene embeddings constructed from per-omic gene-level statistics into pathway embeddings, and uses criss-cross attention for pathway crosstalk modeling. Both DeePathNet and Pathformer use early fusion by integrating multi-omics features before model inputs, while DeepKEGG applies intermediate fusion via omics-specific encoders [13–15].

Most DL methods use *post hoc* feature attribution methods for biomarker identification. These model-agnostic algorithms are applied after training, computing feature scores using gradient-based methods represented by DeepLIFT [46] and Integrated Gradient [47] (IG), as well as other types of methods such as SHAP [48] and GNNExplainer [49]. Some methods use performance drop after zeroing out features (feature ablation) to attribute feature importance. Notably, MOGLAM is an *ante hoc* method, learning a sparsity-regularized feature-indicator matrix during training, which directly encodes feature importance.

Statistical and machine learning methods

In addition to DL-based methods, multi-omics biomarker identification methods based on more conventional statistical or ML algorithms can be traced earlier in the research line. MCIA [43] was one of the earliest models for multi-omics integration and biomarker identification, performing joint projections of omics into shared low-dimensional spaces by maximizing their shared covariance structure. MOFA [5] introduced a statistical framework using matrix factorization for factor analysis. More recently, GAUDI [7] applied UMAP-based transformations followed by XGBoost or random forest with SHAP [48] for biomarker identification. These methods are unsupervised without relying on sample-level labels.

In contrast, DIABLO [6] and asmbPLS-DA [44] are supervised learning methods based on sparse generalized canonical correlation analysis (sGCCA) and sparse partial least square discriminant analysis (sPLS-DA), respectively. Both impose sparsity constraints to identify biomarkers through non-zero component loadings. Stabl [34] wraps a sparse base learner (e.g., logistic regression with l_1 regularization) in bootstrap subsampling to compute per-feature selection frequencies, and identify those above a reliability threshold as biomarkers. GDF [17] and DPM [45] incorporate biological prior knowledge. GDF applies a greedy decision forest over a PPI network using Gini importance for feature scoring, while DPM focuses on pathway-level modeling but also identifies gene-level biomarkers by integrating regulatory directionality (positive/negative associations) across omics layers using a p-value merging algorithm. Unlike DL models, most statistical and ML methods use *ante hoc* biomarker identification via feature weights computed during model optimization.

Collection of gold-standard biomarkers

A critical challenge precluding the quantitative evaluation of biomarker identification performance lies in the lack of an off-the-shelf biomarker set that is both clinically validated and aligned with the task of interest. To address this, we collected biomarkers with strong clinical evidence from established oncology biomarker knowledge bases, including CIViC [38], OncoKB [39], and CGI [40]. To remove confounding effects introduced by biomarkers with weak evidence, such as those identified *in silico* by computational models without clinical validation, we harmonized the evidence levels of the three knowledge bases into a unified evidence system (Level A, B, C, and D) as defined by the AMP/ASCO/CAP guidelines [37] (see Supplementary Table 2 for details) and retained only those with Tier I evidence (Level A and B). These biomarkers are supported by professional guidelines, FDA-approved therapies, or well-powered clinical trials [37]. The curated dataset was stratified into prognostic markers (derived from OncoKB and CIViC) and predictive markers (aggregated from all three sources). To ensure consistency with TCGA data, cancer type names in each source were mapped to TCGA project identifiers. Molecular profiles in CIViC were mapped to HGNC gene symbols (Data availability), also were the gene names from the other two sources to resolve aliases. Finally, therapeutic agent names were harmonized using the standardized terms collected by Ding *et al.* [60], supplemented by a manually defined mapping

to resolve investigational codes, salt forms, spelling errors, and formatting inconsistencies. This rigorous curation allows for a faithful and straightforward assessment of a method's ability to identify biomarkers that have a high likelihood of succeeding clinical trials and achieving widespread application.

Data preprocessing

TCGA omics data

TCGA multi-omics data were downloaded from the Genomic Data Commons (GDC) Data Portal (see Data availability). Based on the omics type choices by the benchmarked methods (Supplementary Table 5) and their availability in TCGA, we opted for five omics data types, namely messenger RNA (mRNA), microRNA (miRNA), DNA methylation (DNAm), Copy Number Variation (CNV), and Single Nucleotide Variation (SNV). Proteomics data were excluded for real-world data benchmarking because TCGA RPPA quantifies only a pre-selected panel of high-quality antibodies (n=487) targeting canonical cancer-pathway proteins [61], which introduces bias towards identifying well-known targets. For mRNA-seq, we used TPM (unstranded) from STAR-Counts, retained protein-coding genes, and applied log2-transformation. For miRNA-seq, we used reads-per-million (RPM) miRNA expression quantifications and performed log2-transformation. For DNA methylation (Illumina HumanMethylation450), we used SeSAMe beta-values and removed CpGs without gene annotations in the Illumina manifest file (see Data availability). For SNVs, we used the masked somatic mutation data, and assigned 1 if a variant mapped to a gene, else 0. For CNVs (Affymetrix SNP 6.0), gene-level copy numbers were restricted to protein-coding genes, centered by subtracting the diploid baseline (2 copies), and capped at 2 for amplifications. Based on GISTIC2.0 [62], values were grouped into five levels of alterations, namely deep deletion (-2), shallow deletion (-1), diploid (0), gain (1) and amplification (2). For simulated data which requires proteomics as reference, we used RPPA protein expression profiles, removed features without a corresponding gene in the TCGA antibodies descriptions file (see Data availability), and excluded samples with more than 50% missing features. After each task dataset was constructed, features with missing values or zero variance were further removed. Finally, standard scaling was applied to each task dataset by fitting train sets and transforming validation and test sets.

TCGA survival and drug response data

Based on the clinical data file from TCGA (Data availability), for survival, we first excluded patient samples without available survival time information ('days_to_death' or 'days_to_last_follow-up'). Then in each task dataset, samples with survival times longer than the median were categorized as high-risk, and those below the median as low-risk. For drug response data, we first excluded patient samples without available 'therapeutic_agent' (for drug labels) or 'treatment_outcome' (for clinical response labels) annotations. Then, drug-patient pairs with inconsistent treatment outcomes were removed. Afterwards, clinical responses were binarized into response (complete response, partial response) and non-response (progressive disease, stable disease, no response, and persistent disease) according to the RECIST standard [63]. Drug names

were standardized using the TCGA drug name standardization information from Ding *et al.* [60], supplemented by a manually defined mapping to resolve investigational codes, salt forms and formatting inconsistencies.

Construction of task datasets

To construct task datasets for biomarker identification, we followed multiple filtering steps to achieve both computational feasibility and biomarker-task alignment, as detailed below.

First, we ensured that the candidate tasks align with the evidence types of biomarkers. Particularly, it was observed that the prognostic and therapeutic biomarkers constitute the primary portion of all the collected biomarker profile entries (Supplementary Fig. 23). Therefore, we opted for two major tasks: survival risk classification and drug response prediction, which were also common tasks chosen by the benchmarked methods [6–8, 14, 15].

Second, we filtered candidate tasks by sample and biomarker counts. For survival risk classification tasks, sample size and biomarker counts for each cancer type are summarized in Supplementary Table 7, and only those with ≥ 300 samples and ≥ 10 biomarkers were further considered. Here, only samples with all five omics data types (i.e., mRNA, miRNA, DNAm, CNV, and SNV) available were counted. Following this, we meticulously inspected each candidate cancer type's biomarker evidence statements in the source knowledge records, and excluded those with low rating or non-significant correlation with the corresponding cancer type's prognosis (Supplementary Table 8). After this, BRCA, LUAD, LUSC, and COADREAD remained as candidates with ≥ 10 biomarkers. For drug response prediction tasks, the sample and biomarker numbers for each drug-cancer pair are summarized into Supplementary Table 9 using TCGA clinical data file. Samples with missing omics data from the five omics types were further removed, and the drug-cancer-type pairs with ≥ 20 responders and ≥ 20 non-responders were selected. Manual inspections on biomarker evidence statements were also conducted to ensure alignment with drug efficacy, and no biomarker was excluded during this process (Supplementary Table 8). Finally, the following drug-cancer-type pairs were considered as candidates (#responders:#non-responders): Cisplatin-BLCA (40:20), Fluorouracil-STAD (51:29), Gemcitabine-PAAD (28:33), Temozolomide-LGG (20:103).

For each candidate task, we prepared its corresponding multi-omics dataset using samples with mRNA, miRNA, DNAm, CNV, and SNV data, along with their survival or drug response labels. Features with missing values or zero variance across samples were removed. Stratified splitting was then applied to generate five cross-validation folds, each dividing the data into an 80% train-validation set and a 20% test set. The train-validation portion was further split into training and validation subsets at a 7:1 ratio, leading to final proportions of 70% training, 10% validation, and 20% testing. For survival tasks, survival times were first grouped into 20 quantile bins before stratification for similar time distributions across the training, validation, and test sets.

Third, for each candidate task dataset, we used basic machine learning classifiers to test whether each dataset contained sufficient signal for classification. Datasets whose

classification performance is below the random baseline were discarded. This guarantees the benchmarked methods can capture distinguishable signals and thus provide trustworthy feature scores or ranks for biomarker identification. To achieve this, we used a random forest and a support vector machine as classifiers to run the 30 experiments (`sklearn.ensemble.RandomForestClassifier` and `sklearn.svm.LinearSVC` from scikit-learn v1.7.2, with default parameters), including 6 omics combinations across 5 cross-validation folds. Tasks with at least one result below the random baseline (mean cross-validation AUROC = 0.5) were excluded, and the remaining ones were selected as the final tasks (Supplementary Table 10). The final benchmarking consists of five multi-omics datasets: BRCA, LUAD, and COADREAD for survival risk classification, and BLCA treated with cisplatin and LGG treated with temozolomide for the drug response prediction. The 43 matched biomarkers are listed in Supplementary Table 3, the task dataset statistics are summarized in Supplementary Table 4, and methods' predictive performance is reported in Supplementary Tables 11-16.

Simulation of multi-omics data

We employed InterSIM [42] (v2.3.0), a multi-omics data generation tool that simulates DNA methylation, gene expression, and protein expression based on key statistics of real reference data for generating simulated multi-omics datasets. InterSIM has been applied in previous benchmark studies of multi-omics integration methods [7, 64], and is balanced between the complexity and interpretability of its generation for different sample groups, making it a suitable tool for our benchmarking.

InterSIM provides the option to use its built-in TCGA ovarian cancer dataset as reference for simulation, but it was limited to a few hundreds of CpGs, genes, and proteins. To better align with the statistical properties of real-world data such as high dimensionality, and to allow running prior-knowledge-based methods, we used real reference based on the BRCA survival task dataset due to its largest sample size (Supplementary Table 4). Specifically, we first included proteomics from RPPA as an additional omics type, and removed samples without proteomics data in the BRCA survival task dataset. Then, for a feasible generation time (less than one day per simulation), we removed CpG sites with less than 0.025 variance or without corresponding gene labels in the methylation array manifest file (Data availability). Following this, the genes regulated by the remaining CpGs were kept, and proteins without a corresponding gene in the RPPA antibodies description file were removed, leaving 6273 CpGs, 3862 genes, and 95 proteins as reference.

For simulation, InterSIM first generates methylation M-values $\boldsymbol{X}_1 \in \mathbb{R}^{N \times F_1}$ for N samples and F_1 CpGs via a multivariate Gaussian distribution

$$\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \delta_1 \boldsymbol{d}_1, \boldsymbol{\Sigma}_1),$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^{F_1}$ is the reference mean M-values across samples at each CpG site, and $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{F_1 \times F_1}$ is the reference covariance matrix, both pre-calculated from reference data. For simulating ground truth biomarkers, $\boldsymbol{d}_1 \in \{0, 1\}^{F_1}$ is an indicating vector generated by Bernoulli trials with $p = 0.01$ for differentially methylated CpGs, and δ_1 determines the power of mean shift.

Afterwards, the reference methylation data is converted into gene-level by taking the median M-values of CpGs mapping to the same gene, and the Pearson correlation coefficients $\boldsymbol{\rho_{12}} \in \mathbb{R}^{F_2}$ is computed between the median M-values and the corresponding gene expression for each gene, where F_2 denotes the number of genes. Then, gene expression data is simulated via

$$\boldsymbol{X}_2 \sim \mathcal{N}((\boldsymbol{\rho}_{12} \circ \boldsymbol{\mu}_1^{(g)} + \sqrt{1 - \boldsymbol{\rho}_{12}^2} \circ \boldsymbol{\mu}_2) + \delta_2 \boldsymbol{d}_2, \boldsymbol{\Sigma}_2),$$

where $\boldsymbol{\mu}_1^{(g)} \in \mathbb{R}^{F_2}$ is the reference vector for mean gene-level methylation, and $\boldsymbol{\mu}_2$ is the reference mean gene expression. The differentially expressed gene indicating vector $\boldsymbol{d}_2 \in \{0,1\}^{F_2}$ is determined by the regulatory mappings between CpGs and genes according to real reference data. Particularly, a gene is a biomarker if one of its regulating CpG is differentially methylated as in $\boldsymbol{d}_1$. $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{F_2 \times F_2}$ is the reference covariance matrix calculated from real gene expression data, and δ_2 determines the power of mean shifts.

For protein expression, first, the Pearson correlation coefficients $\boldsymbol{\rho_{23}} \in \mathbb{R}^{F_3}$ is computed between the expression of each protein and the corresponding gene, where F_3 denotes the number of proteins. Then, protein expression is simulated via

$$\boldsymbol{X}_3 \sim \mathcal{N}((\boldsymbol{\rho}_{23} \circ \boldsymbol{\mu}_2^{(p)} + \sqrt{1 - \boldsymbol{\rho}_{23}^2} \circ \boldsymbol{\mu}_3) + \delta_3 \boldsymbol{d}_3, \boldsymbol{\Sigma}_3),$$

where $\boldsymbol{\mu}_2^{(p)} \in \mathbb{R}^{F_3}$ is the reference vector of mean gene expression corresponding to proteins, and $\boldsymbol{\mu}_3$ is the reference mean protein expression. The differentially expressed protein indicating vector $\boldsymbol{d}_3 \in \{0,1\}^{F_3}$ is determined by the regulatory mappings between genes and proteins, as in the real reference data. $\boldsymbol{\Sigma}_3 \in \mathbb{R}^{F_3 \times F_3}$ is the reference covariance matrix calculated from real protein expression data, and δ_3 determines the power of mean shifts.

Of note, since the numbers of reference genes (3862) and proteins (95) differ significantly, it is possible that $\boldsymbol{d}_3 = \boldsymbol{0}$ when the Bernoulli trials for $\boldsymbol{d}_1$ failed to succeed for any CpG regulating a protein-corresponding gene. Therefore, we separated the set of protein-corresponding genes for an additional Bernoulli trial with $p = 0.1$, and combined the resulting indicating vector with the original one as the final indicating vectors for both gene expression and protein expression data. To simulate datasets with a range of biomarker signal strengths, we varied the mean shifts parameter by $0.5, 1, 2, 3, 4, 5$ ($\delta_1 = \delta_2 = \delta_3$). All simulations generate datasets with 100 samples for two groups, with 50 samples in each. Finally, the generated methylation M-values were converted to beta-values via logit-transformation. Train, validation, and test sets splitting and standard scaling were applied in the same way as for real task datasets. The statistics of the simulated datasets are summarized in Supplementary Table 6.

Implementation and experimental setups of benchmarked methods

Here we describe the general implementation and experimental setups used across benchmarked methods. Method-specific details are provided in Supplementary Methods.

All DL methods were incorporated into unified Python pipelines with PyTorch, using their publicly available code. Unless noted otherwise (for example, MOGONET, MOGLAM, MORE, MoAGL-SA, and Pathformer, which employ fewer training epochs by default), each model was trained for up to 1,000 epochs. Early stopping with a patience of 100 epochs was applied to all models, and the checkpoint with the lowest validation loss was retained for both prediction and biomarker identification. Hyperparameters were set according to the default optimal values reported in the original publications or associated codebases. To account for class imbalance in drug response tasks, class weights were introduced into the cross-entropy loss. Concretely, for C classes with sample sizes $n_1, n_2, \ldots, n_C$, the weight for class i was computed as $w_i = \frac{N}{C\, n_i}$, for $i = 1, 2, \ldots, C$, where N is the total number of samples. All methods were evaluated using five-fold cross-validation within each task dataset. For *post hoc* feature attribution methods, including Integrated Gradients (IG), SHAP, and DeepLIFT, feature scoring was performed on the test sets, with the training set provided as background data for SHAP. To overcome memory constraints without affecting the attribution results, mini-batch processing was used for IG (step size = 50) and SHAP. Statistical and ML methods were incorporated using their publicly available code or documentation, with default optimal parameters. For methods implemented in R, we used the Python package `rpy2` (v3.5.16) to integrate into our unified Python pipeline.

Derivation of gene-level ranking from multi-omics feature scores

Since the collected biomarkers are at gene-level, we adopted a gene-centric approach to enable metric calculations. Specifically, for DNA methylation and miRNA, the average scores of the CpG sites or miRNA molecules regulating the same gene were used as the DNAm- or miRNA-type score for that gene, according to regulatory annotations provided by TCGA and miRTarBase (Data availability). For mRNA, CNV, and SNV, and methods with a gene-centric design, including GENIUS, DeePathNet, DeepKEGG, Pathformer, P-Net, GNN-SubNet, GDF, and DPM (see Supplementary Methods for details), this step was omitted since the output scores are already at gene-level. Following this, gene scores from different omics types were aggregated through max-pooling, resulting in a single score for each gene. For methods that output sample-specific feature scores, the gene scores were further averaged across samples within the same class, and then converted to absolute values and averaged across classes. Lastly, genes with zero scores were randomly permuted, and a final gene-level ranking was formed.

Evaluation metrics

Accuracy metrics

For quantifying biomarker identification accuracy, a key challenge lies in the lack of complete ground truth, since only discovered biomarkers are known. A method might correctly identify novel biomarkers that are unvalidated, making metrics like accuracy or F1-score misleading. To mitigate this, we adopt rank-based metrics by assuming that true biomarkers, whether discovered or not, should be ranked top by an ideal

method. Thus, the ranks of known biomarkers can be viewed as representative draws from the full distribution of true biomarkers, enabling a quantitative evaluation of biomarker identification accuracy.

Let G denote the gene set in a gene-level ranking list, and let $B \subseteq G$ be the subset of genes in G that are reference biomarkers. Let r_b be the rank of biomarker b ($b \in B, r_b \in \{1, 2, ..., |G|\}$.

Average Recall.

$$\mathrm{AR} = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{|R^{(k)} \cap B|}{|B|}$$

$\mathrm{AR} \in (0, 1)$, higher means more accurate.

Normalized Discounted Cumulative Gain. First, the Discounted Cumulative Gain (DCG) is defined as

$$\mathrm{DCG} := \sum_{b \in B} \frac{1}{\log_2(r_b + 1)},$$

The ideal DCG (IDCG) is obtained by placing all biomarkers at the top of the list:

$$\mathrm{IDCG} := \sum_{i=1}^{|B|} \frac{1}{\log_2(i + 1)}.$$

The Normalized DCG is then

$$\mathrm{NDCG} := \frac{\mathrm{DCG}}{\mathrm{IDCG}} \in (0, 1],$$

with 1 indicating a perfect ranking in which every biomarker precedes all non-biomarkers.

Reciprocal Rank. Reciprocal Rank (RR) is a metric commonly used in information retrieval and question answering to measure how highly the first relevant item is ranked [65]. Let $r_{(1)} := \min_{b \in B} r_b$ denote the rank of the highest-ranked biomarker, RR is defined as

$$\mathrm{RR} := \frac{1}{r_{(1)}} \in \left(\frac{1}{|G|}, 1 \right],$$

assigning a score of 1 when a biomarker occupies the top-most slot and decaying hyperbolically as the first biomarker appears deeper in the ranking list.

Mann-Whitney U test *p*-value. We compute a one-sided Mann-Whitney test comparing $\{r_g : g \in B\}$ to $\{r_g : g \notin B\}$ using `scipy.stats.mannwhitneyu(alternative='less', method='exact')` (SciPy v1.14.1). The reported quantity is the p-value $p \in (0, 1]$, where smaller p indicates stronger evidence that biomarkers rank higher than non-biomarkers.

Area Under the ROC Curve. For each $g \in G$, let s_g be its gene-level score (larger s_g = more biomarker-like) and $y_g := \mathbf{1}[g \in B]$. The AUROC is the area under the ROC curve obtained by thresholding s_g. Equivalently,

$$\mathrm{AUROC} = \frac{1}{|B| \cdot |G \setminus B|} \sum_{b \in B} \sum_{g \notin B} \left[\mathbf{1}(s_b > s_g) + \tfrac{1}{2} \mathbf{1}(s_b = s_g) \right]$$

We compute AUROC with `sklearn.metrics.roc_auc_score` (scikit-learn v1.7.2).

Accuracy@K. Let $R^{(K)}$ be the set of the top-K genes. Then

$$\text{Accuracy@}K := \frac{|R^{(K)} \cap B|}{K} \in [0, 1],$$

the fraction of the top-K predictions that are known biomarkers.

Stability metrics

For stability, we calculated the average similarities between each pair of the five-fold ranking lists.

Average Kendall's τ. τ Kendall's τ [66] is widely used for measuring the overall similarity between two rankings. Given rankings R_i and R_j, it is defined as

$$\tau(R_i, R_j) = \frac{c - d}{\binom{|G|}{2}}$$

where c is the number of concordant pairs (same relative order in both rankings) and d the number of discordant pairs. A higher value indicates higher similarity, with 1 denoting perfectly concordant and -1 denoting the least concordant ($\tau \in [-1, 1]$). We average τ over all pairs from the five-fold rankings as the final metric:

$$\text{Avg. Kendall's } \tau = \frac{1}{10} \sum_{i,j \in \{1,\ldots,5\}} \tau(R_i, R_j),\ i < j$$

Rank-Biased Overlap. Although Kendall's τ offers a global perspective on the overall ranking, including both top-ranked candidates and lowly ranked insignificant genes, it may be significantly affected by the instability at lower ranks. We therefore further adopted Rank-Biased Overlap (RBO) [67], a common metric to emphasize top-ranking similarity:

$$\text{RBO}(R_i, R_j; \lambda) = (1 - \lambda) \sum_{k=1}^{|G|} \lambda^{k-1} \frac{|R_i^{(k)} \cap R_j^{(k)}|}{k}.$$

where $R_i^{(k)}$ denotes the top-k genes in ranking i, and $\lambda \in (0, 1)$ is the decay rate. Throughout our experiments, λ was set to 0.98 to effectively account for the top 50 ranks, as evaluated and reported by Webber *et al.* [67]. RBO$\in [0, 1)$, with higher values indicating higher concordance. Similar to Kendall's τ, we average RBO over all pairs from the five-fold rankings as the final metric.

Rank Percentile Standard Deviation. We define Rank Percentile Standard Deviation (RPSD) to measure the variability of known biomarkers' ranks across cross-validation folds. For each biomarker b and fold i, let

$$p_{b,i} = \frac{r_{b,i} - 1}{|G| - 1} \in [0, 1]$$

be its rank percentile, and let

$$\bar{p}_b = \frac{1}{F}\sum_{i=1}^{F} p_{b,i}$$

be the mean rank percentile for b across F folds. We first compute the per-biomarker standard deviation

$$\sigma_b = \sqrt{\frac{1}{F-1}\sum_{i=1}^{F}\left(p_{b,i} - \bar{p}_b\right)^2},$$

and then average over all biomarkers

$$\text{RPSD} = \frac{1}{|B|}\sum_{b\in B}\sigma_b.$$

Lower RPSD indicates more consistent ranking of biomarkers across folds, with 0 denoting perfect stability.

Predictive performance metrics

We assessed binary predictive performance using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). For a given decision threshold t, predicted labels are summarized in terms of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Define

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t)+\text{FN}(t)}, \qquad \text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t)+\text{TN}(t)},$$

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t)+\text{FP}(t)}, \qquad \text{Recall}(t) = \text{TPR}(t).$$

The receiver operating characteristic (ROC) curve plots TPR(t) against FPR(t) as t varies over all possible thresholds, and AUROC is the area under this curve. The precision-recall (PR) curve plots Precision(t) against Recall(t), and the corresponding area under this curve defines AUPR, which is particularly informative in the presence of substantial class imbalance.

Derivation of consensus multi-omics biomarker panels

Performance varied substantially across different omics combinations (Supplementary Figs. 1-15). For this reason, we treated each five-fold experiment separately, where a five-fold experiment refers to a five-fold cross-validation performed within a specific method and omics combination. We first computed an overall performance score for each experiment,

$$\text{Overall Score} = \frac{(\text{AR} + \text{NDCG} + \text{RR}) + (\text{RBO} - \text{RPSD} + \text{Kendall's } \tau)}{6}$$

where each metric is averaged across the five folds. All experiments were then ranked in descending order according to this score.

Next, based on the dominant omics type within the top 1% rankings, for each of the five omics types (mRNA, miRNA, CNV, SNV, and DNAm), we selected the top three experiments that included that omics type. For each selected experiment, we applied the Robust Rank Aggregation (RRA) algorithm [59] using the RobustRankAggreg R package (v1.2.1) to integrate the rankings from the five folds. RRA produces an aggregated ranking with Benjamini-Hochberg corrected p-values that indicate consensus significance. After obtaining three aggregated rankings for a given omics type, we applied RRA again to integrate these three lists and derive a final consensus ranking for that omics type. During this process, for methods that do not require gene-level inputs (i.e., those except Pathformer, DeePathNet, DPM, GNN-SubNet, GENIUS, GDF, and P-Net; see Supplementary Methods for details), we converted CpG scores to gene level using mean aggregation for DNAm. For gene-centric methods, we converted miRNA scores to gene level. Full results are provided in Supplementary Data 1.

Computational resources

Experiments were primarily conducted on a Linux Ubuntu 22.04 system equipped with 8 NVIDIA RTX A6000 GPUs (48 GB each), a dual-socket AMD EPYC 7763 processor (2×64 cores, 2 threads per core), and 1008 GB system memory. Experiments requiring larger GPU memory were run on a Linux Ubuntu 24.04 system equipped with one NVIDIA H200 GPU (SXM, 141 GB), a dual-socket Intel Xeon Platinum 8570 processor (2×56 cores, 2 threads per core), and 2 TB system memory.

Data availability

TCGA omics and clinical data were downloaded from the GDC data portal at https://portal.gdc.cancer.gov (accessed November 06 2024). Reference biomarkers were retrieved from https://www.oncokb.org/actionable-genes for OncoKB (accessed June 15 2025), https://civicdb.org/releases/main for CIViC (2025-06-01 release), and https://www.cancergenomeinterpreter.org/data/biomarkers for CGI (latest version, accessed June 15 2025). The HGNC complete gene set information was downloaded from https://storage.googleapis.com/public-download-files/hgnc/tsv/tsv/hgnc_complete_set.txt (accessed Feb 11 2025). miRNA target gene information was retrieved from miRTarBase at https://mirtarbase.cuhk.edu.cn/~miRTarBase/ (accessed April 21 2025). Methylation array manifest file (`HM450.hg38.manifest.gencode.v36.tsv.gz`) and TCGA antibodies descriptions file (`TCGA_antibodies_descriptions.gencode.v36.tsv`) were downloaded from https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files. All biological pathway data files were downloaded from the corresponding methods' data repositories. PPI network topology file was downloaded from the STRING [68] database at https://string-db.org/ (accessed October 28 2024).

Code availability

The code for the benchmarking pipeline and analysis is available at https://github.com/athanzli/benchmarking_multiomics_biomarker_identification_methods. Benchmarking datasets and results are available at 10.5281/zenodo.17860662 [69].

Author contributions

R.L. conceived and supervised the study. A.Z.L. conducted the literature review. A.Z.L. designed the methodology. A.Z.L. developed the software. A.Z.L. generated the visualizations. A.Z.L., Y.D., Y.L., L.C., and R.L. analyzed the results. A.Z.L. and R.L. drafted the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **1A**, 68–77 (2015) https://doi.org/10.5114/wo.2014.47136 . Publisher: Termedia Sp. z.o.o.

[2] Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J.T., Gatza, M.L., Wilkerson, M., Perou, C.M., Yellapantula, V., Huang, K.-l., Lin, C., McLellan, M.D., Yan, P., Davies, S.R., Townsend, R.R., Skates, S.J., Wang, J., Zhang, B., Kinsinger, C.R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A.G., Fenyö, D., Ellis, M.J., Carr, S.A.: Proteogenomics connects somatic mutations to signalling in breast cancer. Nature **534**(7605), 55–62 (2016) https://doi.org/10.1038/nature18003 . Publisher: Nature Publishing Group. Accessed 2025-07-22

[3] Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., al.: Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell **158**(4), 929–944 (2014) https://doi.org/10.1016/j.cell.2014.06.049

[4] Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biology **18**(1), 83 (2017) https://doi.org/10.1186/s13059-017-1215-1 . Accessed 2025-07-22

[5] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O.: Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems

Biology **14**(6), 8124 (2018) https://doi.org/10.15252/msb.20178124 . Accessed 2025-08-05

[6] Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., Lê Cao, K.-A.: DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics **35**(17), 3055–3062 (2019) https://doi.org/10.1093/bioinformatics/bty1054 . Accessed 2024-09-23

[7] Castellano-Escuder, P., Zachman, D.K., Han, K., Hirschey, M.D.: GAUDI: interpretable multi-omics integration with UMAP embeddings and density-based clustering. Nature Communications **16**(1), 5771 (2025) https://doi.org/10.1038/s41467-025-60822-1 . Publisher: Nature Publishing Group. Accessed 2025-08-17

[8] Wang, F.-a., Zhuang, Z., Gao, F., He, R., Zhang, S., Wang, L., Liu, J., Li, Y.: TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. Genome Biology **25**(1), 149 (2024) https://doi.org/10.1186/s13059-024-03293-9 . Accessed 2025-07-22

[9] Benkirane, H., Pradat, Y., Michiels, S., Cournède, P.-H.: CustOmics: A versatile deep-learning based strategy for multi-omics integration. PLOS Computational Biology **19**(3), 1010921 (2023) https://doi.org/10.1371/journal.pcbi.1010921 . Publisher: Public Library of Science. Accessed 2025-07-22

[10] Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K.: MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications **12**(1), 3445 (2021) https://doi.org/10.1038/s41467-021-23774-w . Number: 1 Publisher: Nature Publishing Group. Accessed 2023-09-28

[11] Pfeifer, B., Saranti, A., Holzinger, A.: GNN-SubNet: disease subnetwork detection with explainable graph neural networks. Bioinformatics **38**(Supplement_2), 120–126 (2022) https://doi.org/10.1093/bioinformatics/btac478 . Accessed 2025-07-22

[12] Ouyang, D., Liang, Y., Li, L., Ai, N., Lu, S., Yu, M., Liu, X., Xie, S.: Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Computers in Biology and Medicine **164**, 107303 (2023) https://doi.org/10.1016/j.compbiomed.2023.107303 . Accessed 2025-06-22

[13] Lan, W., Liao, H., Chen, Q., Zhu, L., Pan, Y., Chen, Y.-P.P.: DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. Briefings in Bioinformatics **25**(3), 185 (2024) https://doi.org/10.1093/bib/bbae185 . Accessed 2024-09-11

[14] Cai, Z., Poulos, R.C., Aref, A., Robinson, P.J., Reddel, R.R., Zhong, Q.: DeePathNet: A Transformer-Based Deep Learning Model Integrating Multiomic Data with

Cancer Pathways. Cancer Research Communications **4**(12), 3151–3164 (2024) https://doi.org/10.1158/2767-9764.CRC-24-0285 . Accessed 2025-07-22

[15] Liu, X., Tao, Y., Cai, Z., Bao, P., Ma, H., Li, K., Li, M., Zhu, Y., Lu, Z.J.: Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. Bioinformatics **40**(5), 316 (2024) https://doi.org/10.1093/bioinformatics/btae316 . Accessed 2024-09-03

[16] Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S.H., Salari, K., Kregel, S., Richter, C., Arnoff, T.E., Park, J., Hahn, W.C., Van Allen, E.M.: Biologically informed deep neural network for prostate cancer discovery. Nature **598**(7880), 348–352 (2021) https://doi.org/10.1038/s41586-021-03922-4 . Publisher: Nature Publishing Group. Accessed 2025-07-22

[17] Pfeifer, B., Baniecki, H., Saranti, A., Biecek, P., Holzinger, A.: Multi-omics disease module detection with an explainable Greedy Decision Forest. Scientific Reports **12**(1), 16857 (2022) https://doi.org/10.1038/s41598-022-21417-8 . Publisher: Nature Publishing Group. Accessed 2025-08-26

[18] Wang, Y., Wang, Z., Yu, X., Wang, X., Song, J., Yu, D.-J., Ge, F.: MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification. Briefings in Bioinformatics **26**(1), 658 (2025) https://doi.org/10.1093/bib/bbae658 . Accessed 2025-08-06

[19] Cheng, L., Huang, Q., Zhu, Z., Li, Y., Ge, S., Zhang, L., Gong, P.: MoAGL-SA: a multi-omics adaptive integration method with graph learning and self attention for cancer subtype classification. BMC Bioinformatics **25**(1), 364 (2024) https://doi.org/10.1186/s12859-024-05989-y . Accessed 2025-08-06

[20] Sokač, M., Kjær, A., Dyrskjøt, L., Haibe-Kains, B., Aerts, H.J.W.L., Birkbak, N.J.: GENIUS: GEnome traNsformatIon and spatial representation of mUltiomicS data. eLife **12** (2023) https://doi.org/10.7554/eLife.87133.2 . Publisher: eLife Sciences Publications Limited. Accessed 2025-07-22

[21] Lu, Y., Peng, R., Dong, L., Xia, K., Wu, R., Xu, S., Wang, J.: Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. Briefings in Bioinformatics **24**(6), 378 (2023) https://doi.org/10.1093/bib/bbad378 . Accessed 2025-08-06

[22] Chereda, H., Leha, A., Beißbarth, T.: Stable feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation for biomarker discovery in breast cancer. Artificial Intelligence in Medicine **151**, 102840 (2024) https://doi.org/10.1016/j.artmed.2024.102840 . Accessed 2025-04-18

[23] Labory, J., Njomgue-Fotso, E., Bottini, S.: Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning

algorithms for patient classification using metabolomics biomedical data. Computational and Structural Biotechnology Journal **23**, 1274–1287 (2024) https://doi.org/10.1016/j.csbj.2024.03.016 . Publisher: Elsevier. Accessed 2025-08-06

[24] Yang, Q., Gong, Y., Zhu, F.: Critical Assessment of the Biomarker Discovery and Classification Methods for Multiclass Metabolomics. Analytical Chemistry **95**(13), 5542–5552 (2023) https://doi.org/10.1021/acs.analchem.2c04402 . Accessed 2024-09-12

[25] Petković, M., Slavkov, I., Kocev, D., Džeroski, S.: Biomarker discovery by feature ranking: Evaluation on a case study of embryonal tumors. Computers in Biology and Medicine **128**, 104143 (2021) https://doi.org/10.1016/j.compbiomed.2020.104143 . Publisher: Elsevier BV. Accessed 2025-07-11

[26] Acharjee, A., Larkman, J., Xu, Y., Cardoso, V.R., Gkoutos, G.V.: A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Medical Genomics **13**(1), 178 (2020) https://doi.org/10.1186/s12920-020-00826-6 . Accessed 2025-08-06

[27] Dessì, N., Pascariello, E., Pes, B.: A Comparative Analysis of Biomarker Selection Techniques. BioMed Research International **2013**(1), 387673 (2013) https://doi.org/10.1155/2013/387673 . Accessed 2025-03-14

[28] Camillo, B.D., Sanavia, T., Martini, M., Jurman, G., Sambo, F., Barla, A., Squillario, M., Furlanello, C., Toffolo, G., Cobelli, C.: Effect of Size and Heterogeneity of Samples on Biomarker Discovery: Synthetic and Real Data Assessment. PLOS ONE **7**(3), 32200 (2012) https://doi.org/10.1371/journal.pone.0032200 . Publisher: Public Library of Science. Accessed 2025-08-06

[29] Karrila, S., Lee, J.H.E., Tucker-Kellogg, G.: A comparison of Methods for Data-Driven Cancer Outlier Discovery, and An Application Scheme to Semisupervised Predictive Biomarker Discovery. Cancer Informatics **10**, 6868 (2011) https://doi.org/10.4137/CIN.S6868 . Publisher: SAGE Publications Ltd STM. Accessed 2025-08-06

[30] Park, J., Lee, J.W., Park, M.: Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. BioData Mining **16**(1), 18 (2023) https://doi.org/10.1186/s13040-023-00334-0 . Accessed 2025-08-06

[31] Bhadra, T., Mallik, S., Hasan, N., Zhao, Z.: Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. BMC Bioinformatics **23**(3), 153 (2022) https://doi.org/10.1186/s12859-022-04678-y . Accessed 2025-08-06

[32] Crabtree, N.M., Moore, J.H., Bowyer, J.F., George, N.I.: Multi-class computational evolution: development, benchmark evaluation and application to RNA-Seq

biomarker discovery. BioData Mining **10**(1), 13 (2017) https://doi.org/10.1186/s13040-017-0134-8 . Accessed 2025-08-06

[33] Yang, Z., Wu, N., Liang, Y., Zhang, H., Ren, Y.: SMSPL: Robust Multimodal Approach to Integrative Analysis of Multiomics Data. IEEE Transactions on Cybernetics **52**(4), 2082–2095 (2022) https://doi.org/10.1109/TCYB.2020.3006240 . Accessed 2025-03-02

[34] Hédou, J., Marić, I., Bellan, G., Einhaus, J., Gaudillière, D.K., Ladant, F.-X., Verdonk, F., Stelzer, I.A., Feyaerts, D., Tsai, A.S., Ganio, E.A., Sabayev, M., Gillard, J., Amar, J., Cambriel, A., Oskotsky, T.T., Roldan, A., Golob, J.L., Sirota, M., Bonham, T.A., Sato, M., Diop, M., Durand, X., Angst, M.S., Stevenson, D.K., Aghaeepour, N., Montanari, A., Gaudillière, B.: Discovery of sparse, reliable omic biomarkers with Stabl. Nature Biotechnology **42**(10), 1581–1593 (2024) https://doi.org/10.1038/s41587-023-02033-x . Publisher: Nature Publishing Group. Accessed 2025-07-11

[35] Zhang, X., Jonassen, I., Goksøyr, A.: Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In: Helder I., N. (ed.) Bioinformatics. Exon Publications, Brisbane (AU) (2021). http://www.ncbi.nlm.nih.gov/books/NBK569564/ Accessed 2024-09-12

[36] Li, Y., Mansmann, U., Du, S., Hornung, R.: Benchmark study of feature selection strategies for multi-omics data. BMC Bioinformatics **23**(1), 412 (2022) https://doi.org/10.1186/s12859-022-04962-x . Accessed 2025-06-03

[37] Li, M.M., Datto, M., Duncavage, E.J., Kulkarni, S., Lindeman, N.I., Roy, S., Tsimberidou, A.M., Vnencak-Jones, C.L., Wolff, D.J., Younes, A., Nikiforova, M.N.: Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. The Journal of Molecular Diagnostics **19**(1), 4–23 (2017) https://doi.org/10.1016/j.jmoldx.2016.10.002 . Publisher: Elsevier BV. Accessed 2025-07-22

[38] Griffith, M., Spies, N.C., Krysiak, K., McMichael, J.F., Coffman, A.C., Danos, A.M., Ainscough, B.J., Ramirez, C.A., Rieke, D.T., Kujan, L., Barnell, E.K., Wagner, A.H., Skidmore, Z.L., Wollam, A., Liu, C.J., Jones, M.R., Bilski, R.L., Lesurf, R., Feng, Y.-Y., Shah, N.M., Bonakdar, M., Trani, L., Matlock, M., Ramu, A., Campbell, K.M., Spies, G.C., Graubert, A.P., Gangavarapu, K., Eldred, J.M., Larson, D.E., Walker, J.R., Good, B.M., Wu, C., Su, A.I., Dienstmann, R., Margolin, A.A., Tamborero, D., Lopez-Bigas, N., Jones, S.J.M., Bose, R., Spencer, D.H., Wartman, L.D., Wilson, R.K., Mardis, E.R., Griffith, O.L.: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nature Genetics **49**(2), 170–174 (2017) https://doi.org/10.1038/ng.3774 . Publisher: Nature Publishing Group. Accessed 2025-07-22

[39] Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph,

J.E., Yaeger, R., Soumerai, T., Nissan, M.H., Chang, M.T., Chandarlapaty, S., Traina, T.A., Paik, P.K., Ho, A.L., Hantash, F.M., Grupe, A., Baxi, S.S., Callahan, M.K., Snyder, A., Chi, P., Danila, D.C., Gounder, M., Harding, J.J., Hellmann, M.D., Iyer, G., Janjigian, Y.Y., Kaley, T., Levine, D.A., Lowery, M., Omuro, A., Postow, M.A., Rathkopf, D., Shoushtari, A.N., Shukla, N., Voss, M.H., Paraiso, E., Zehir, A., Berger, M.F., Taylor, B.S., Saltz, L.B., Riely, G.J., Ladanyi, M., Hyman, D.M., Baselga, J., Sabbatini, P., Solit, D.B., Schultz, N.: OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology (1), 1–16 (2017) https://doi.org/10.1200/PO.17.00011 . Publisher: Wolters Kluwer. Accessed 2025-07-22

[40] Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M.P., Vivancos, A., Rovira, A., Tusquets, I., Albanell, J., Rodon, J., Tabernero, J., Torres, C., Dienstmann, R., Gonzalez-Perez, A., Lopez-Bigas, N.: Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Medicine **10**(1), 25 (2018) https://doi.org/10.1186/s13073-018-0531-8 . Accessed 2025-06-08

[41] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics **45**(10), 1113–1120 (2013) https://doi.org/10.1038/ng.2764 . Publisher: Nature Publishing Group. Accessed 2025-10-20

[42] Chalise, P., Raghavan, R.R., Fridley, B.L.: InterSIM: Simulation tool for multiple integrative 'omic datasets'. Computer Methods and Programs in Biomedicine **128**, 69–74 (2016) https://doi.org/10.1016/j.cmpb.2016.02.011

[43] Meng, C., Kuster, B., Culhane, A.C., Gholami, A.M.: A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics **15**(1), 162 (2014) https://doi.org/10.1186/1471-2105-15-162 . Accessed 2025-09-10

[44] Zhang, R., Datta, S.: Adaptive Sparse Multi-Block PLS Discriminant Analysis: An Integrative Method for Identifying Key Biomarkers from Multi-Omics Data. Genes **14**(5), 961 (2023) https://doi.org/10.3390/genes14050961 . Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2025-09-10

[45] Slobodyanyuk, M., Bahcheli, A.T., Klein, Z.P., Bayati, M., Strug, L.J., Reimand, J.: Directional integration and pathway enrichment analysis for multi-omics data. Nature Communications **15**(1), 5690 (2024) https://doi.org/10.1038/s41467-024-49986-4 . Publisher: Nature Publishing Group. Accessed 2025-09-10

[46] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, pp. 3145–3153. JMLR.org, Sydney, NSW, Australia (2017)

[47] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR, , (2017). https://proceedings.mlr.press/v70/sundararajan17a.html

[48] Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., , (2017). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html Accessed 2023-04-16

[49] Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: GNNExplainer: Generating Explanations for Graph Neural Networks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

[50] Halama, A., Zaghlool, S., Thareja, G., Kader, S., Al Muftah, W., Mook-Kanamori, M., Sarwath, H., Mohamoud, Y.A., Stephan, N., Ameling, S., Pucic Baković, M., Krumsiek, J., Prehn, C., Adamski, J., Schwenk, J.M., Friedrich, N., Völker, U., Wuhrer, M., Lauc, G., Najafi-Shoushtari, S.H., Malek, J.A., Graumann, J., Mook-Kanamori, D., Schmidt, F., Suhre, K.: A roadmap to the molecular human linking multiomics with population traits and diabetes subtypes. Nature Communications **15**(1), 7111 (2024) https://doi.org/10.1038/s41467-024-51134-x . Publisher: Nature Publishing Group. Accessed 2025-10-27

[51] Van Allen, E.M., Mouw, K.W., Kim, P., Iyer, G., Wagle, N., Al-Ahmadie, H., Zhu, C., Ostrovnaya, I., Kryukov, G.V., O'Connor, K.W., Sfakianos, J., Garcia-Grossman, I., Kim, J., Guancial, E.A., Bambury, R., Bahl, S., Gupta, N., Farlow, D., Qu, A., Signoretti, S., Barletta, J.A., Reuter, V., Boehm, J., Lawrence, M., Getz, G., Kantoff, P., Bochner, B.H., Choueiri, T.K., Bajorin, D.F., Solit, D.B., Gabriel, S., D'Andrea, A., Garraway, L.A., Rosenberg, J.E.: Somatic ERCC2 Mutations Correlate with Cisplatin Sensitivity in Muscle-Invasive Urothelial Carcinoma. Cancer Discovery **4**(10), 1140–1153 (2014) https://doi.org/10.1158/2159-8290.CD-14-0623 . Accessed 2025-10-28

[52] Li, Q., Damish, A.W., Frazier, Z., Liu, D., Reznichenko, E., Kamburov, A., Bell, A., Zhao, H., Jordan, E.J., Gao, S.P., Ma, J., Abbosh, P.H., Bellmunt, J., Plimack, E.R., Lazaro, J.-B., Solit, D.B., Bajorin, D., Rosenberg, J.E., D'Andrea, A.D., Riaz, N., Van Allen, E.M., Iyer, G., Mouw, K.W.: ERCC2 Helicase Domain Mutations Confer Nucleotide Excision Repair Deficiency and Drive Cisplatin Sensitivity in Muscle-Invasive Bladder Cancer. Clinical Cancer Research **25**(3), 977–988 (2019) https://doi.org/10.1158/1078-0432.CCR-18-1001 . Accessed 2025-10-28

[53] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C.,

Osborne, T.F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., Sculley, D.: Underspecification Presents Challenges for Credibility in Modern Machine Learning. Journal of Machine Learning Research **23**(226), 1–61 (2022). Submitted 11/20; Revised 12/21; Published 08/22

[54] Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 29. Barcelona, Spain (2016). 30th Conference on Neural Information Processing Systems (NIPS 2016)

[55] Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., Marks, K.M., Prins, R.M., Ward, P.S., Yen, K.E., Liau, L.M., Rabinowitz, J.D., Cantley, L.C., Thompson, C.B., Vander Heiden, M.G., Su, S.M.: Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature **462**(7274), 739–744 (2009) https://doi.org/10.1038/nature08617 . Accessed 2025-09-22

[56] Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., Wilson, R.K., Ally, A., Balasundaram, M., Butterfield, Y.S.N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H.-J.E., Coope, R.J.N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R.A., Lee, D., Li, H.I., Mayo, M., Moore, R.A., Mungall, A.J., Pleasance, E., Gordon Robertson, A., Schein, J.E., Shafiei, A., Sipahimalani, P., Slobodan, J.R., Stoll, D., Tam, A., Thiessen, N., Varhol, R.J., Wye, N., Zeng, T., Zhao, Y., Birol, I., Jones, S.J.M., Marra, M.A., Cherniack, A.D., Saksena, G., Onofrio, R.C., Pho, N.H., Carter, S.L., Schumacher, S.E., Tabak, B., Hernandez, B., Gentry, J., Nguyen, H., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Gabriel, S.B., Meyerson, M., Chin, L., Park, P.J., Kucherlapati, R., Hoadley, K.A., Todd Auman, J., Fan, C., Turman, Y.J., Shi, Y., Li, L., Topal, M.D., He, X., Chao, H.-H., Prat, A., Silva, G.O., Iglesia, M.D., Zhao, W., Usary, J., Berg, J.S., Adams, M., Booker, J., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A.P., Simons, J.V., Soloway, M.G., Mose, L.E., Jefferys, S.R., Balu, S., Parker, J.S., Neil Hayes, D., Perou, C.M., Malik, S., Mahurkar, S., Shen, H., Weisenberger, D.J., Triche Jr, T., Lai, P.H., Bootwalla, M.S., Maglinte, D.T., Berman, B.P., Van Den Berg, D.J., Baylin, S.B., Laird, P.W., Creighton, C.J., Donehower, L.A., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M.S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Sinha, R., Park, R.W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P.J., Chin, L., Reynolds, S., Kreisberg, R.B., Bernard, B., Bressler, R., Erkkila, T., Lin, J., Thorsson, V., Zhang, W., Shmulevich, I., Ciriello, G., Weinhold, N., Schultz, N., Gao, J., Cerami, E., Gross, B., Jacobsen, A., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B.S., Ladanyi, M., Sander, C., Anur, P., Spellman, P.T., Lu, Y., Liu, W., Verhaak, R.R.G., Mills, G.B., Akbani, R.,

Zhang, N., Broom, B.M., Casasent, T.D., Wakefield, C., Unruh, A.K., Baggerly, K., Coombes, K., Weinstein, J.N., Haussler, D., Benz, C.C., Stuart, J.M., Benz, S.C., Zhu, J., Szeto, C.C., Scott, G.K., Yau, C., Paull, E.O., Carlin, D., Wong, C., Sokolov, A., Thusberg, J., Mooney, S., Ng, S., Goldstein, T.C., Ellrott, K., Grifford, M., Wilks, C., Ma, S., Craft, B., Yan, C., Hu, Y., Meerzaman, D., Gastier-Foster, J.M., Bowen, J., Ramirez, N.C., Black, A.D., Pyatt, R.E., White, P., Zmuda, E.J., Frick, J., Lichtenberg, T.M., Brookens, R., George, M.M., Gerken, M.A., Harper, H.A., Leraas, K.M., Wise, L.J., Tabler, T.R., McAllister, C., Barr, T., Hart-Kothari, M., Tarvin, K., Saller, C., Sandusky, G., Mitchell, C., Iacocca, M.V., Brown, J., Rabeno, B., Czerwinski, C., Petrelli, N., Dolzhansky, O., Abramov, M., Voronina, O., Potapova, O., Marks, J.R., Suchorska, W.M., Murawa, D., Kycler, W., Ibbs, M., Korski, K., Spychała, A., Murawa, P., Brzeziński, J.J., Perz, H., Łaźniak, R., Teresiak, M., Tatka, H., Leporowska, E., Bogusz-Czerniewicz, M., Malicki, J., Mackiewicz, A., Wiznerowicz, M., Van Le, X., Kohl, B., Viet Tien, N., Thorp, R., Van Bang, N., Sussman, H., Duc Phu, B., Hajek, R., Phi Hung, N., Viet The Phuong, T., Quyet Thang, H., Zaki Khan, K., Penny, R., Mallery, D., Curley, E., Shelton, C., Yena, P., Ingle, J.N., Couch, F.J., Lingle, W.L., King, T.A., Maria Gonzalez-Angulo, A., Mills, G.B., Dyer, M.D., Liu, S., Meng, X., Patangan, M., The Cancer Genome Atlas Network, Genome sequencing centres: Washington University in St Louis, Genome characterization centres: BC Cancer Agency, Broad Institute, Brigham & Women's Hospital & Harvard Medical School, North Carolina, C.H., University of Southern California/Johns Hopkins, Genome data analysis: Baylor College of Medicine, Institute for Systems Biology, Memorial Sloan-Kettering Cancer Center, Oregon Health & Science University, The University of Texas MD Anderson Cancer Center, California, S.C.I., NCI, Biospecimen core resource: Nationwide Children's Hospital Biospecimen Core Resource, Tissue source sites: ABS-IUPUI, Christiana, Cureline, Duke University Medical Center, The Greater Poland Cancer Centre, ILSBio, International Genomics Consortium, Mayo Clinic, MSKCC, MD Anderson Cancer Center: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012) https://doi.org/10.1038/nature11412

[57] Shahbandi, A., Nguyen, H.D., Jackson, J.G.: TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. Trends in Cancer **6**(2), 98–110 (2020) https://doi.org/10.1016/j.trecan.2020.01.007 . Publisher: Elsevier. Accessed 2025-10-28

[58] Nimmerjahn, F., Ravetch, J.V.: Fc receptors as regulators of immune responses. Nature Reviews Immunology **8**(1), 34–47 (2008) https://doi.org/10.1038/nri2206 . Publisher: Nature Publishing Group. Accessed 2025-09-22

[59] Kolde, R., Laur, S., Adler, P., Vilo, J.: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics **28**(4), 573–580 (2012) https://doi.org/10.1093/bioinformatics/btr709 . Accessed 2023-07-18

[60] Ding, Z., Zu, S., Gu, J.: Evaluating the molecule-based prediction of clinical drug

responses in cancer. Bioinformatics **32**(19), 2891–2895 (2016) https://doi.org/10.1093/bioinformatics/btw344 https://academic.oup.com/bioinformatics/article-pdf/32/19/2891/49021204/bioinformatics_32_19_2891.pdf

[61] Akbani, R., Ng, P.K.S., Werner, H.M.J., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-Y., Yoshihara, K., Li, J., Ling, S., Seviour, E.G., Ram, P.T., Minna, J.D., Diao, L., Tong, P., Heymach, J.V., Hill, S.M., Dondelinger, F., Städler, N., Byers, L.A., Meric-Bernstam, F., Weinstein, J.N., Broom, B.M., Verhaak, R.G.W., Liang, H., Mukherjee, S., Lu, Y., Mills, G.B.: A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nature Communications **5**(1), 3887 (2014) https://doi.org/10.1038/ncomms4887 . Publisher: Nature Publishing Group. Accessed 2025-08-06

[62] Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G.: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology **12**(4), 41 (2011) https://doi.org/10.1186/gb-2011-12-4-r41 . Accessed 2025-09-01

[63] Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., Verweij, J.: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer **45**(2), 228–247 (2009) https://doi.org/10.1016/j.ejca.2008.10.026

[64] Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., Baudot, A.: Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature Communications **12**(1), 124 (2021) https://doi.org/10.1038/s41467-020-20430-7 . Publisher: Nature Publishing Group. Accessed 2025-04-18

[65] Voorhees, E.M., Tice, D.M.: The TREC-8 question answering track. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00). European Language Resources Association (ELRA), Athens, Greece (2000). https://aclanthology.org/L00-1018/

[66] KENDALL, M.G.: A NEW MEASURE OF RANK CORRELATION. Biometrika **30**(1-2), 81–93 (1938) https://doi.org/10.1093/biomet/30.1-2.81

[67] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. **28**(4), 20–12038 (2010) https://doi.org/10.1145/1852102.1852106 . Accessed 2025-10-27

[68] Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N., Pyysalo, S., Bork, P., Jensen, L., von Mering, C.: The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic

Acids Research **51**(D1), 638–646 (2023) https://doi.org/10.1093/nar/gkac1000

[69] Li, A.Z.: Data and results for "Benchmarking computational methods for multi-omics biomarker discovery in cancer". Zenodo (2025). https://doi.org/10.5281/zenodo.17906214