Benchmarking computational methods for multi-omics biomarker discovery in cancer

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

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

Abstract

Multi-omics profiling characterizes cancer biology and supports biomarker discovery for prognosis and therapy selection. Although numerous computational multi-omics biomarker identification methods have been proposed, their ability to identify clinically relevant biomarkers has not been systematically evaluated, leaving it unclear whether the resulting biomarker nominations are reliable for downstream validation. Here we systematically benchmark 20 representative statistical, machine learning and deep learning 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. Overall, DeePathNet and DeepKEGG achieve the best 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. By aggregating results from top-performing methods, we construct consensus biomarker panels that nominate candidates for potential investigations. Finally, we provide user-friendly interfaces to allow researchers to benchmark new methods against the 20 baselines or apply selected methods for biomarker identification on custom multi-omics datasets. Our benchmark is publicly available at https://github.com/athanzli/CancerMOBI-Bench.

1 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 therapy selection from a vast space of molecular candidates [3]. Nevertheless, high dimensionality, tumor heterogeneity, and complex cross-omics dependencies 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 (CCA), or

*Corresponding author: ruishanl@usc.edu

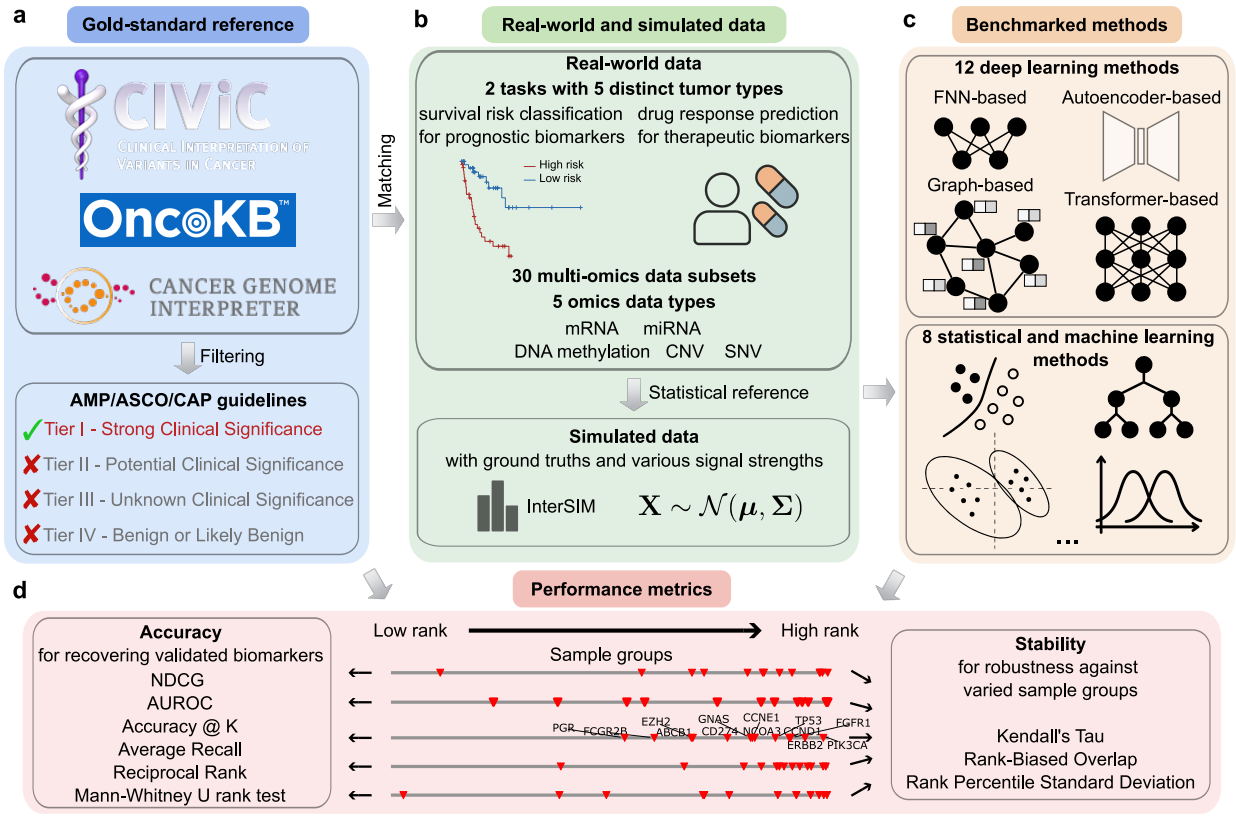

Figure 1: **Overview of benchmarking workflow.** **a,** Biomarkers were collected from three oncology knowledge bases, including CIViC [5], OncoKB [6], and CGI [7], and those with Tier I evidence determined by AMP/ASCO/CAP guidelines [8] were retained. **b,** Real-world data included 30 multi-omics data 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 (Supplementary Methods). Real-world data were used as statistical reference for simulated multi-omics data generation with InterSIM [9]. **c,** Twelve DL methods were benchmarked with different architectural backbones, including 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).

Uniform Manifold Approximation and Projection (UMAP) to extract shared patterns across omics and derive interpretable latent representations [10–12]. 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 [13–20]. 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 [13–15, 19–21]. Beyond predictions, many incorporate interpretability modules to identify biologically meaningful features, such as disease-associated subnetworks [22], dysregulated pathways [20], and most commonly, molecular biomarkers [12, 14–21, 23–26].

Despite the 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 [12, 14–21, 23–26]. Such assessments are inferential and prone to selection bias, pro-

viding limited quantitative validation. Also, existing benchmark studies often rely on predictive performance as a proxy for biomarker identification performance [27–37]. However, high predictive accuracy does not necessarily imply reliable biomarker identification, as models can achieve strong prediction using features that are statistically discriminative but not biologically meaningful, particularly in high-dimensional, correlated multi-omics data [38]. Although some utilize synthetic datasets with predefined discriminative features for direct and quantitative assessment, these datasets are often simplistic and fail to match the complexity and heterogeneity of real-world cancer data [11, 31, 33, 39, 40]. Moreover, many recent multi-omics and DL methods remain absent from prior benchmarks [27, 29, 30, 34–37, 39, 41, 42].

Translational implications can arise from these gaps. As biomarker development proceeds from early discovery toward tests intended for specific clinical applications, successful translation requires analytical validity, clinical validity, and clinical utility [43, 44]. Because less than 1% of published cancer biomarkers enter clinical practice, more rigorous evaluation of discovery-stage methods is needed to better prioritize approaches with downstream translational potential [45].

To address these limitations, we provide a systematic benchmark of computational methods for multi-omics biomarker identification. A major obstacle has been the lack of task-specific biomarker sets for ground-truth evaluation. To address this, we collected Tier I biomarkers as defined by AMP/ASCO/CAP guidelines [8] from Clinical Interpretation of Variants in Cancer (CIViC) [5], MSK's Precision Oncology Knowledge Base (OncoKB) [6], and Cancer Genome Interpreter (CGI) [7]. Using these gold-standard biomarkers, our benchmarking enables direct, quantitative, and clinically meaningful assessment of biomarker identification performance. Instead of relying on indirect evaluation through predictive performance, our approach directly quantifies each method's ability to recover curated, clinically relevant biomarkers via rank-based metrics. We constructed five real-world task datasets from The Cancer Genome Atlas (TCGA) [46] through systematic filtering to achieve 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. Performance was evaluated across diverse experimental settings and metrics. Several methods exhibit strong ability 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 methods 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 real-world cancer cohorts. Additionally, we derive consensus panels of multi-omics cancer biomarkers by aggregating the results of top-performing methods, providing candidates for potential downstream investigations. Our benchmark is publicly available with user-friendly evaluation pipelines that enable researchers to benchmark new methods against the 20 baselines on the curated task datasets, or to apply selected methods to other multi-omics data for biomarker identification.

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
DL							
P-Net [21]	Sparse neural network	Early	Biological pathways	Supervised	DeepLIFT	*Post hoc*	2021
GENIUS [25]	Convolutional neural network	Early	Physical genomic positioning	Supervised	Integrated Gradients	*Post hoc*	2023
TMO-Net [13]	Variational autoencoder	Intermediate	None	Supervised	Integrated Gradients	*Post hoc*	2024
CustOmics [14]	Variational autoencoder	Mixed	None	Supervised	SHAP	*Post hoc*	2023
MOGONET [15]	Graph convolutional network	Late	None	Supervised	Feature ablation	*Post hoc*	2021
MoAGL-SA [24]	Graph convolutional network and self-attention	Intermediate	None	Supervised	Feature ablation	*Post hoc*	2024
MORE [23]	Hypergraph and self-attention	Intermediate	None	Supervised	Feature ablation	*Post hoc*	2024
MOGLAM [17]	Graph convolutional network and self-attention	Intermediate	None	Supervised	Feature weights	*Ante hoc*	2023
GNN-SubNet [16]	Graph isomorphism network	Early	PPI network	Supervised	GNNExplainer	*Post hoc*	2022
Pathformer [20]	Transformer encoder	Early	Biological pathways	Supervised	SHAP	*Post hoc*	2024
DePathNet [19]	Transformer encoder	Early	Biological pathways	Supervised	SHAP	*Post hoc*	2024
DeepKEGG [18]	Transformer encoder	Intermediate	Biological pathways	Supervised	DeepLIFT	*Post hoc*	2024
Statistical and ML							
MCIA [47]	Matrix decomposition	Intermediate	None	Unsupervised	Latent component weights	*Ante hoc*	2014
MOFA [10]	Matrix decomposition	Intermediate	None	Unsupervised	Latent component weights	*Ante hoc*	2018
GAUDI [12]	UMAP transformations and XGBoost/RF	Intermediate	None	Unsupervised	SHAP	*Post hoc*	2025
DIABLO [11]	Sparse generalized CCA	Intermediate	None	Supervised	Latent component weights	*Ante hoc*	2019
asmbPLS-DA [48]	Sparse PLS-DA	Intermediate	None	Supervised	Latent component weights	*Ante hoc*	2023
Stabl [40]	Sparse regularized model	Intermediate	None	Supervised	Max selection frequency	*Ante hoc*	2024
GDF [22]	Greedy decision forest	Early	PPI network	Supervised	Gini gain	*Ante hoc*	2022
DPM [49]	Directional p-value merging	Intermediate	Inter-omics regulatory directions	Supervised	Merged p-values	*Ante hoc*	2024

PPI: protein-protein interaction; PLS-DA: partial least squares discriminant analysis; RF: random forest.

2 Results

2.1 Benchmarking design and evaluation framework

We conducted a systematic benchmark of 20 computational methods for multi-omics biomarker identification, including 8 statistical and ML methods and 12 DL methods selected from literature (Supplementary Table 1) that encompass diverse methodological designs (Fig. 1c). Their model families, multi-omics fusion strategies, use of prior biological knowledge, and biomarker identification approaches are summarized in Table 1, with detailed descriptions provided in Supplementary Methods.

To anchor evaluation in clinical relevance, we curated Tier I cancer biomarkers, as defined by AMP/ASCO/CAP guidelines [8]. These biomarkers were harmonized across CIViC [5], OncoKB [6], and CGI [7] (Fig. 1a, Supplementary Tables 2, 3). After 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, Supplementary Methods, Table 2). The curated biomarker set spans multiple clinically established alteration classes, including sequence variants, copy-number alterations, expression-based markers, DNA methylation markers, gene fusions, and limited protein-level markers (Supplementary Table 3), capturing major mechanisms with established clinical significance such as oncogenic driver activation, tumor suppressor inactivation, DNA-repair deficiency, epigenetic silencing, and immune-related expression states [5–8].

Since the benchmarked methods employed different omics combinations in their original case studies (Supplementary Table 4), we defined a unified set of omics combinations to achieve 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 [9]. These simulations contain predefined discriminative features serving as ground truth, enabling controlled assessment of method sensitivity across a range of signal levels (Fig. 1b, Supplementary Table 5, Supplementary Methods).

Biomarker identification performance was evaluated using nine complementary metrics capturing different aspects of accuracy and stability (Fig. 1d, Methods). For real-world cohorts where ground truth biomarkers are partially known, rank-based metrics were applied to reward the prioritization of validated biomarkers without penalizing unvalidated candidates. For simulated datasets with complete ground truth, evaluation adopted metrics such as AUROC and accuracy. Stability was quantified as the similarity of ranked gene lists across cross-validation, using metrics that reflect top-rank consistency, global ranking agreement, and biomarker ranking variability (Methods). For comparability, all method output was converted to gene-level ranking lists prior to metric calculation (Methods).

Table 2: **Summary statistics of the real datasets used in this study.**

Task	Dataset	Sample size		#Features					#Biomarkers
		Low risk	High risk	CNV	DNAm	mRNA	miRNA	SNV	
Survival	BRCA	324	323	18 645	46 859	19 504	1 581	15 118	14
	LUAD	212	211	18 645	46 859	19 497	1 581	17 103	14
	COADREAD	187	185	18 645	46 859	19 471	1 555	18 394	12
		Response	Non-response						
Drug response	Cisplatin (BLCA)	40	20	18 645	46 859	19 288	1 404	10 502	1
	Temozolomide (LGG)	20	103	18 645	46 859	19 411	1 392	3 656	2

2.2 Accuracy on real-world data

Accurate identification of clinically relevant candidates from high-dimensional molecular space represents a hallmark of an effective biomarker discovery method. Therefore, we first examined the performance of each method for the recovery of clinically validated biomarkers in real-world 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 3, Methods). Detailed results of each omics combination are reported in Supplementary Figs. 1-5 and Supplementary Figs. 11-15.

We 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 3) as well as comparable overall accuracy (Fig. 2a,b). Within this group, the transformer-based, pathway-informed methods DeePathNet and DeepKEGG were top performers. 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 strong ability to surface 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 3). This divergence aligns with the 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, methods relying on feature ablation (MOGONET, MoAGL-SA, MORE) rarely recovered biomarkers across both success counts and accuracy metrics (Table 3, Fig. 2a, b). Their underperformance likely reflects the complementarity among correlated multi-omics features, where ablation of a single feature is compensated by others [38]. Among statistical and ML methods, MOFA, DPM, and MCIA consistently ranked lowest across tasks and metrics. DPM is intrinsically univariate, while MOFA and MCIA rank features via unsupervised matrix decomposition that may not align with clinical endpoints (Supplementary Methods). In contrast, GAUDI, despite also being unsupervised, achieved substantially better performance by combining tree-based models with SHAP-based attribution, highlighting the benefit of multivariate attribution even in unsupervised settings (Fig. 2a, b).

Deep neural networks are not intrinsically superior to conventional statistical and ML models. Although DL methods exhibit a higher performance ceiling, with DeePathNet and DeepKEGG achieving the highest accuracy and maintaining it across multiple cancers and omics combina-

Table 3: **Frequency of surfacing biomarkers.** 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, including six omics combinations with five-fold cross-validation.

	Survival BRCA			Survival LUAD			Survival COADREAD			Drug response Cisplatin (BLCA)			Drug response Temozolomide (LGG)			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

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

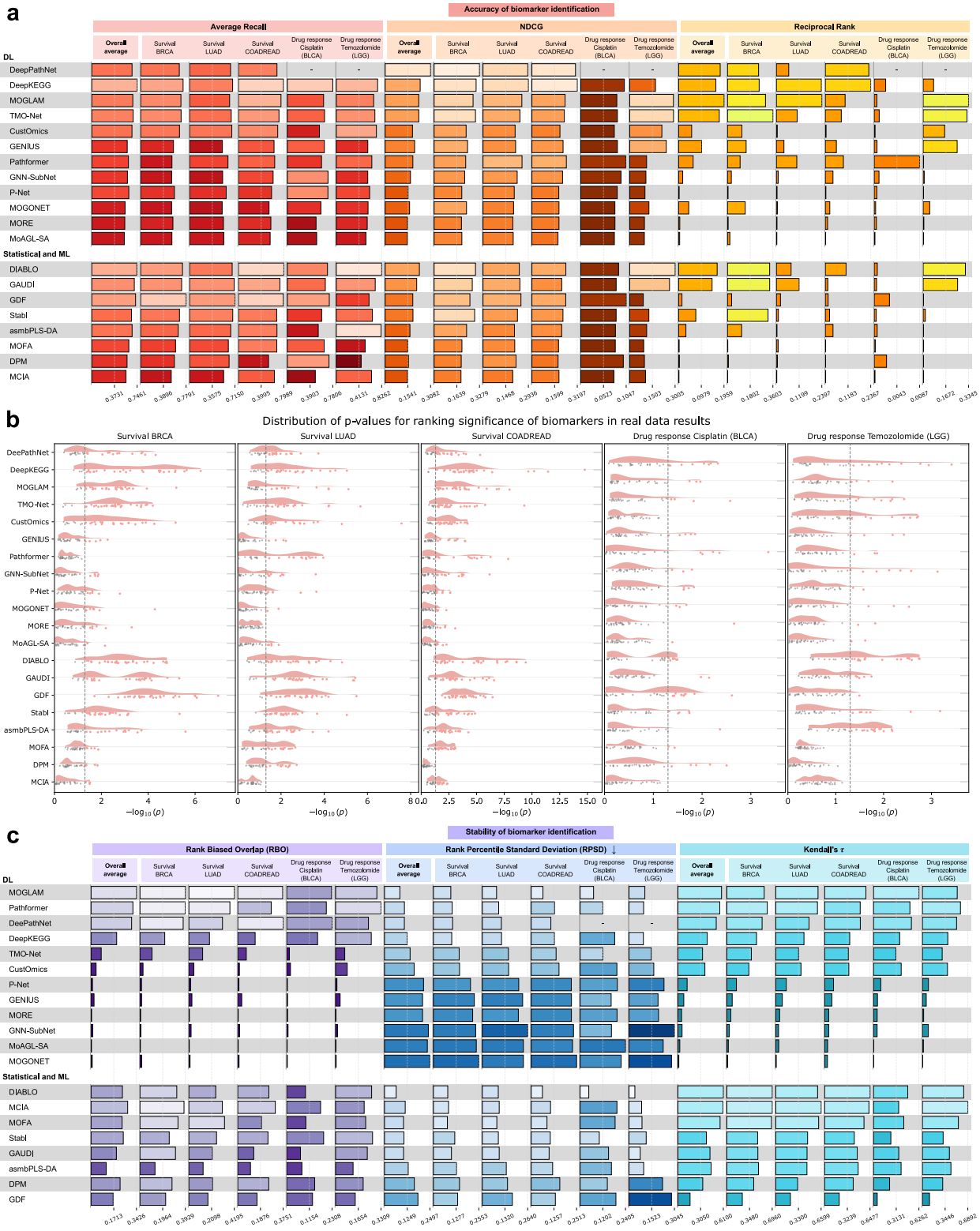

Figure 2: **Overall performance on real-world data. a,** Accuracy for all tasks and metrics. For a particular method and task, the height of the bar represents the averaged metric value of the cross-validation mean across all six omics combinations. **b,** Mann-Whitney U test p-values (negative $\log_{10}$-transformed). Each data point represents a single experiment corresponding to a specific omics combination and cross-validation fold. The vertical dashed line represents a p-value of 0.05. **c,** Stability for all tasks 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.

tions (Supplementary Figs. 1-5), well-designed statistical methods such as DIABLO and GAUDI achieved competitive accuracy. Conversely, Pathformer achieved only moderate accuracy despite using a transformer backbone, and MOGLAM was the strongest graph-based method despite most graph methods ranking low. These contrasts indicate that specific methodological designs, rather than model family, primarily determine success. The factors underlying these differences are systematically analyzed in the following sections.

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 Figs. 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 [50]. 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 [51]. 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 extract weak signals from noisy data under certain omics combinations, though such advantages may trade off against performance on less difficult tasks.

2.3 Stability on real-world data

The reproducibility of biomarkers is essential for clinical translation, and high stability is a prerequisite for reproducible biomarker discovery across diverse cohorts. We quantified stability using the five cross-validation folds per omics combination per task, averaging the pairwise similarities between gene ranking lists (Methods). 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, Methods). Stability results of 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 the least stable 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 lower 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 network training, as small changes in the input samples can drive stochastic gradient-based optimization into distinct solution basins [52], producing substantially different top-ranked features identified by *post hoc* attribution methods.

Among the most stable DL methods, incorporating biological pathway structures (Pathformer,

DeePathNet, DeepKEGG) and employing explicit regularization (MOGLAM) appeared to improve stability, while DL methods lacking such design choices tended to be among the least stable (Fig. 2c).

Jointly considering all methods, those that fail to capture global feature dependencies (GNN-SubNet, P-Net, GENIUS) were generally less stable and also underperformed in accuracy, suggesting that insufficient modeling of long-range molecular interactions diminishes the stabilizing contribution of global context [53] (Fig. 2c).

2.4 Determinants of biomarker identification performance

Joint analysis of accuracy and stability revealed substantial variation across the benchmarked methods. Autoencoder-based methods such as TMO-Net and CustOmics achieved strong accuracy but were unstable, Pathformer exhibited high stability but only moderate accuracy, and conventional statistical methods including MCIA and MOFA were inaccurate but highly stable. Nevertheless, high accuracy and stability are achievable simultaneously: 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). These contrasts indicate that the decisive factor is specific methodological design rather than model family. Across the diverse tasks and omics combinations evaluated, four methodological factors consistently distinguished high- from low-performing methods.

Feature attribution. The approach used to infer feature importance emerged as a primary performance determinant. Methods employing multivariate attribution, including SHAP (DeePathNet, Pathformer, CustOmics, GAUDI), DeepLIFT (DeepKEGG, P-Net), Integrated Gradients (TMO-Net, GENIUS), GNNExplainer (GNN-SubNet), and *ante-hoc* feature weighting (MOGLAM, DIABLO, asmbPLS-DA, Stabl), consistently outperformed those relying on univariate feature ablation (MOGONET, MoAGL-SA, MORE) or per-feature statistical tests (DPM) (Fig. 2a, b, Table 3). In high-dimensional multi-omics data, features are correlated both within and across molecular layers. Univariate approaches underestimate the importance of features whose effects are distributed across correlated molecular partners, because removing or testing a single feature can be compensated by its correlated counterparts [38]. Multivariate methods evaluate each feature's contribution in the context of others, capturing the joint dependency structure essential for identifying clinically relevant biomarkers. The importance of this factor is also illustrated by GAUDI, which demonstrates that combining tree-based models with SHAP can compensate for the absence of supervised labels, outperforming several supervised DL methods that rely on weaker attribution approaches (Fig. 2a, b).

Integration of biological knowledge. Encoding biological pathway structures into model architectures was associated with high accuracy and stability. The two top-performing methods, DeePathNet and DeepKEGG, encode pathway knowledge into transformer backbones (Table 1, Supplementary Methods). Pathway-informed architectures constrain the hypothesis space to biologically plausible feature groups, reducing overfitting and directing the model toward functionally coherent molecular associations. However, knowledge encoding alone is insufficient. P-Net also incorporates pathways but underperforms due to sparse connectivity that suppresses inter-pathway communication, and GNN-SubNet uses PPI networks but is limited by local message-passing (Supplementary Methods). These contrasts indicate that the effectiveness of prior knowledge depends on the architecture's capacity to propagate learned representations across the encoded biological structures.

Scope of feature interactions. Methods modeling long-range dependencies across the feature space exhibited markedly higher accuracy and stability than those restricted to local interactions.

Transformer self-attention enables pathways to attend to one another, capturing cross-pathway dependencies that reflect the interconnected nature of biological processes. In contrast, CNNs (GENIUS) are constrained by local receptive fields, sparse networks (P-Net) suppress inter-node communication, and graph message-passing (GNN-SubNet) attenuates with distance [53]. Among statistical methods, DIABLO's generalized CCA models global cross-omics covariance, contributing to its competitive performance (Supplementary Methods). These observations indicate that when biomarker-associated features span multiple interacting molecular layers, architectures restricted to local interactions cannot adequately capture the cross-layer dependencies required for accurate and stable biomarker identification (Fig. 2).

Regularization. Regularization improved stability without compromising accuracy. MOGLAM learns graphs adaptively from the data and applies regularization on its feature-indicator matrix, yielding the highest stability among methods (Fig. 2c, Supplementary Figs. 6-10). DL methods without explicit regularization (P-Net, GENIUS, GNN-SubNet, MOGONET, MoAGL-SA, MORE) exhibited low stability, reflecting the sensitivity of stochastic optimization to training set perturbations [52]. The co-occurrence of high accuracy and stability in methods such as MOGLAM, DIABLO and Stabl demonstrates that these objectives are simultaneously achievable through appropriate regularization (Fig. 2).

These four factors interact synergistically. DeePathNet and DeepKEGG combine all four (pathway-informed transformer architectures with global self-attention, multivariate attribution, and implicit regularization via structured pathway encodings), achieving the highest overall performance (Fig. 2, Table 3). Methods lacking multiple factors (e.g., MOGONET, MoAGL-SA, MORE with local graph interactions, univariate ablation, and no prior knowledge) consistently show weak performance. This indicates that strong biomarker identification is not driven by a single design choice, but instead emerges from the combination of multiple favorable design principles.

2.5 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 one or two omics combinations in their original case studies (Supplementary Table 4), a systematic comparison across a broader set of omics combinations is essential for uncovering the underlying biological mechanisms and informing both method development and experimental design. Consistent with this need, we observed substantial differences in performance 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.

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 (Methods). For each such case, we recorded the omics type with the highest feature score (Methods) as dominant. Figure 3a provides a comprehensive summary for all experiments. Detailed results of each omics combination are reported 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),

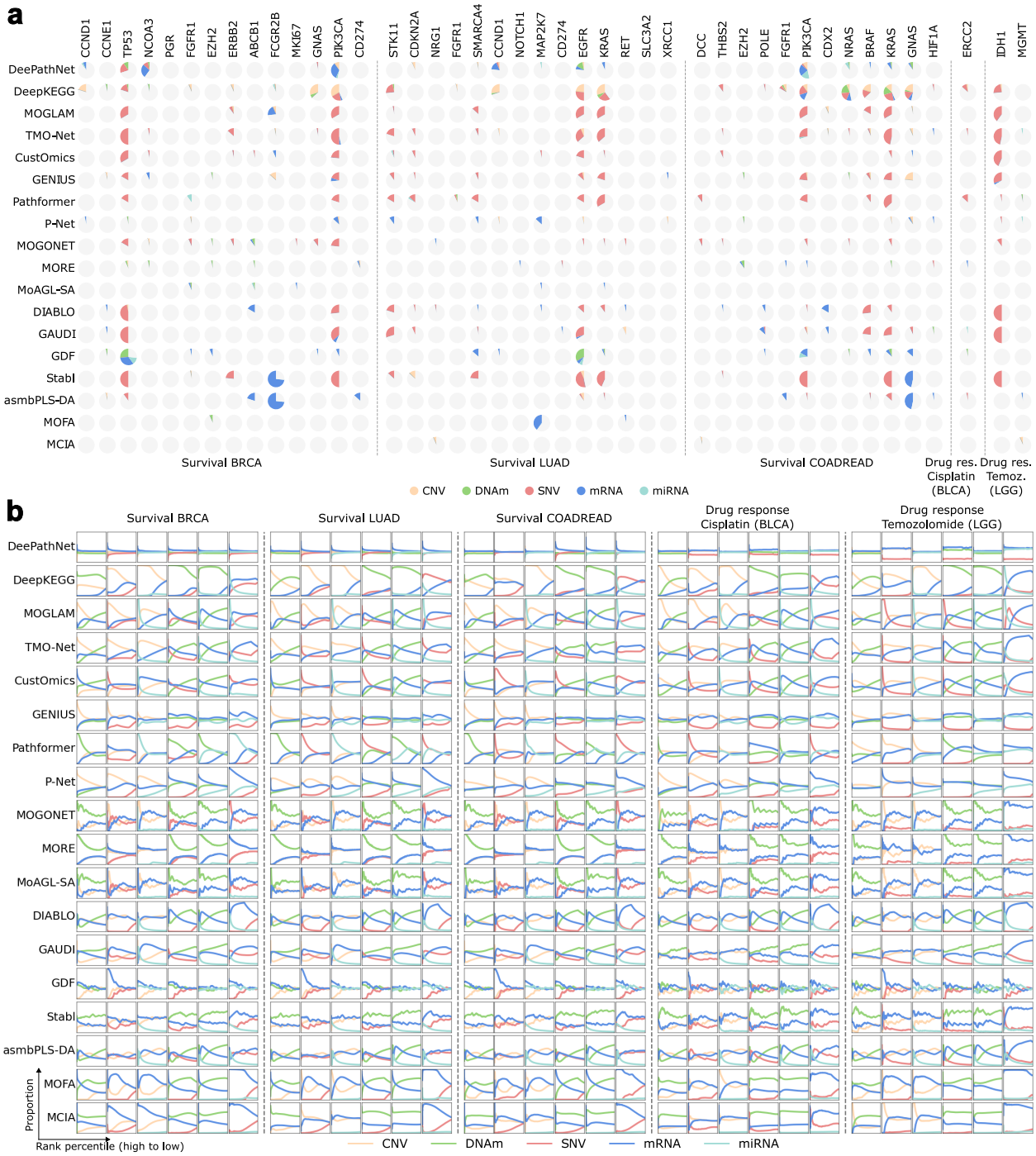

Figure 3: **Omics type contributions in biomarker identification.** **a**, Pie charts showing the omics types driving the identification of a biomarker. Each pie has 30 portions, with each representing a single experiment with a specific cross-validation fold (five in total) and omics combination (six in total). A portion is 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 (Methods), which were concatenated from the five score lists output by cross-validation.

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 [54]. 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 [55, 56]. By contrast, FCGR2B was primarily identified through mRNA expression, particularly by Stabl and asmbPLS-DA (Fig. 3a, Supplementary Figs. 16-21), where transcriptomic signatures reflect immune activation rather than direct tumor genetics in breast cancer prognosis [57]. 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.

Since 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. We summarized the proportions of each omics type at different rank percentile cutoffs using each method's raw feature scores or rankings, per 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 may be challenging 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.

2.6 Performance on simulated data

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

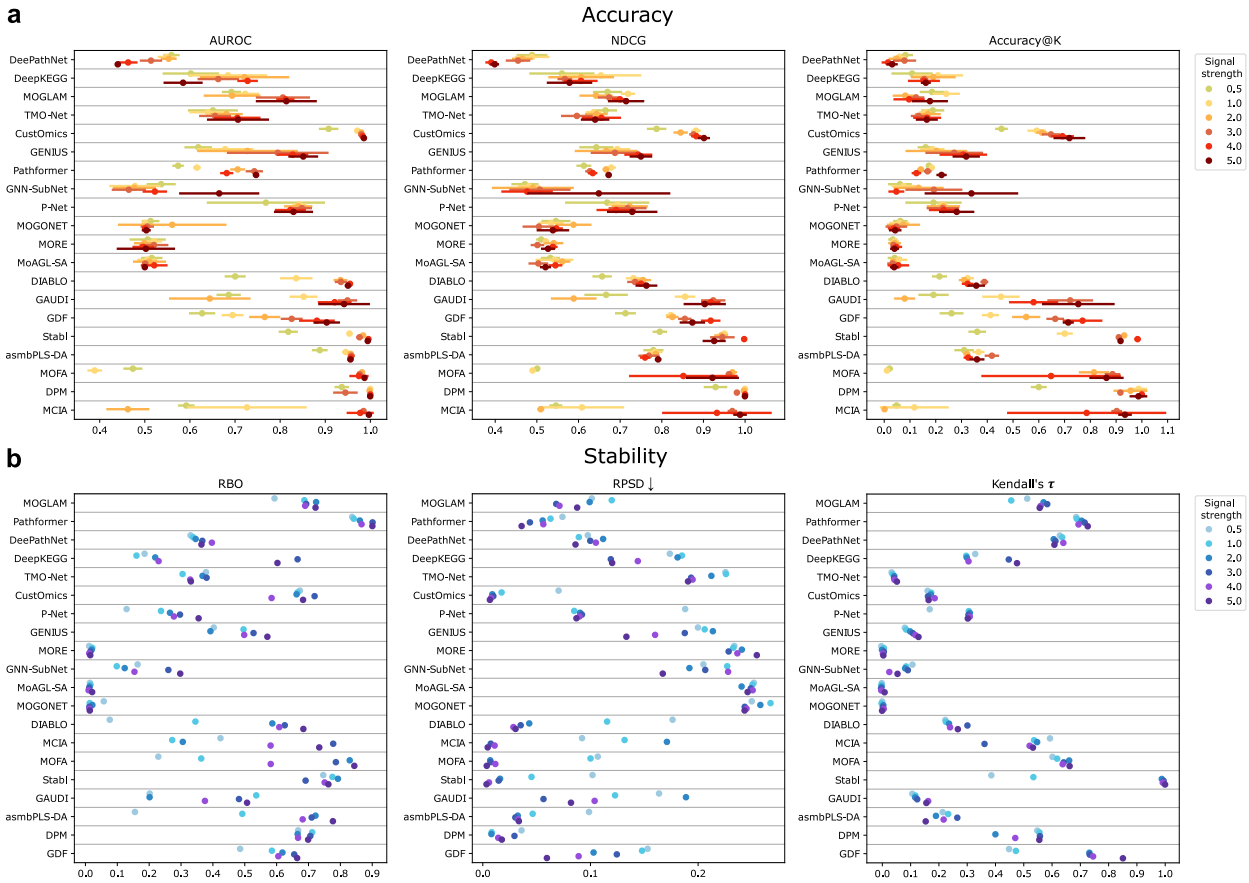

Figure 4: **Performance on simulated data.** Cleveland dot plots showing biomarker identification accuracy (**a**) (bars represent standard deviation across five-fold cross-validation) and stability (**b**). Colors represent a particular biomarker signal strength (Supplementary Methods).

ablation DL methods (MOGONET, 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. 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 methods intended to recover more complex relationships.

Across discriminative strengths, statistical and ML methods exhibited greater variability than DL methods (Fig. 4), indicating higher 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, thus simulations tend to favor conventional statistical and ML methods that assume simpler and less noisy underlying structure. In contrast, the capacity of DL models to utilize subtle or composite multivariate patterns becomes evident primarily in real-world datasets characterized by substantial heterogeneity and nonlinear cross-omics relationships. These discrepancies underscore the limitations of relying on synthetic discriminative features as bench-

marking ground truth. Although widely used in prior work [11, 31, 33, 39, 40], simulation-based evaluations did not align with assessments based on real-world cohorts and clinically established biomarkers.

Stability patterns on simulated data were more consistent with those from real cohorts than accuracy (Fig. 4b, Fig. 2c), suggesting that stability is driven more by methodological design than by the numerical properties of data. Nonetheless, statistical and ML methods showed larger stability shifts as discriminative strength varied, indicating higher sensitivity to the structures within simulated data.

2.7 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. Since design choices and omics-specific dependencies vary widely across methods, the resulting rankings show only modest concordance among related approaches and weak concordance across distinct methodological families (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 derived consensus rankings via Robust Rank Aggregation (RRA) [58] (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 distillation of sparse, accurate, and reproducible biomarker candidate panels (Table 4). As an example, in BRCA survival, the consensus identified 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 8 SNV biomarkers (FDR $< 10^{-6}$) (Table 4). Derived from diverse methods and omics combinations that 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.

Many consensus panel candidates are established cancer genes, including CDH1 and GATA3 in the BRCA survival SNV panel [55], APC in the COADREAD panel [59], KDM6A in the cisplatin (BLCA) panel [60], and ATRX in the temozolomide (LGG) panel [61]. The recovery of such well-characterized genes provides evidence that the top-performing methods and RRA aggregation identify biologically meaningful candidates. All candidates are defined on routinely profiled molecular readouts and can be tested in independent cohorts using standard assays. However, they represent discovery-stage, computationally prioritized candidates that would require independent analytical and clinical validation before translational use [43].

Table 4: **Consensus multi-omics biomarker panels by task and omics type.** Benjamini-Hochberg corrected p-values from RRA [58] (Methods) are shown. 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

2.8 Practical guidance for method selection

To facilitate practical use of the benchmark, we provide a method selection guide organized around three considerations: whether sample-level labels are available, whether GPU resources are accessible, and whether preservation of the original feature space is desired (Fig. 5). DeePathNet and DeepKEGG are the strongest general-purpose choices, achieving the highest accuracy with strong stability via pathway-informed transformer architectures with advanced attribution methods. When labels are unavailable, GAUDI is preferred; when GPU resources are unavailable, DIABLO offers strong performance with lightweight computation; when preservation of the original feature space is desired, MOGLAM is recommended. Methodological designs that are consistently unsuitable, including univariate feature ablation, univariate statistical scoring, and unsupervised matrix decomposition without multivariate attribution, are not recommended.

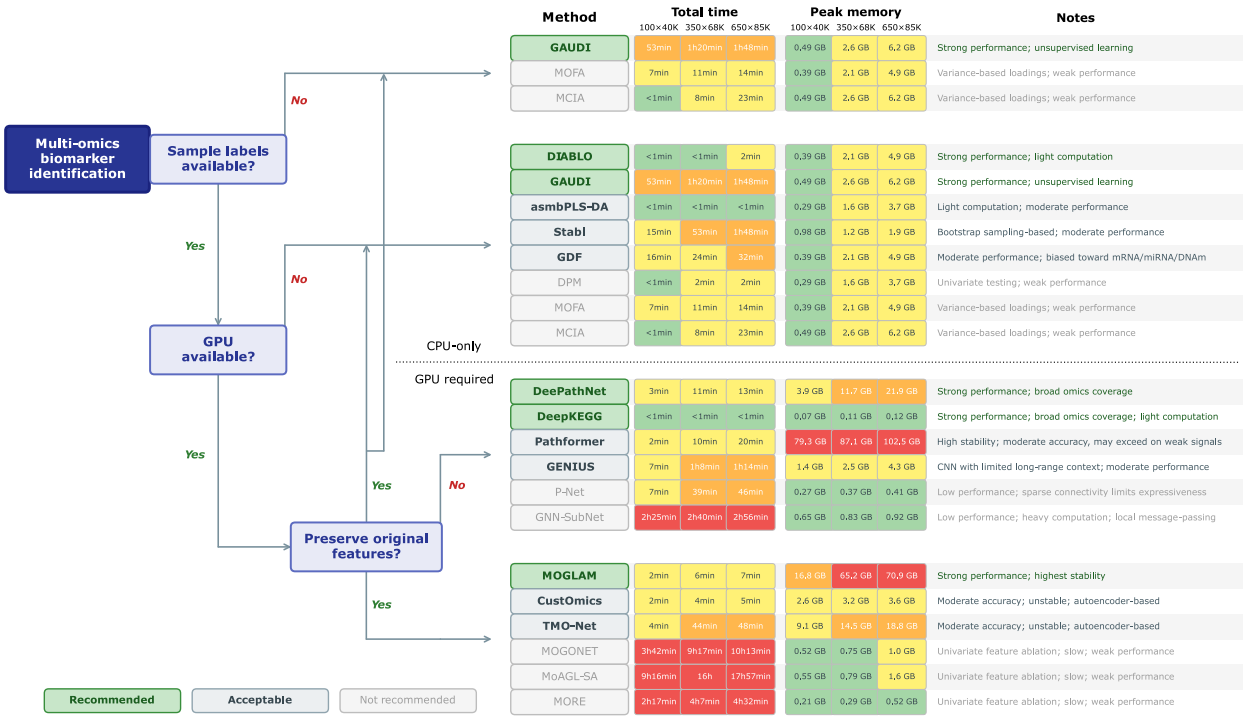

Figure 5: **Practical guidance for method selection.** Decision tree for selecting multi-omics biomarker identification methods based on sample-label availability, GPU accessibility, and whether preservation of the original feature space is desired. Methods are grouped into recommended, acceptable, and not recommended tiers according to overall benchmark performance. Representative runtime and peak memory across three data scales are provided for practical reference. GPU peak memory is provided for DL methods, and CPU peak memory is provided for non-DL methods.

Method selection does not depend on the specific omics types available, as all benchmarked methods support arbitrary omics combinations in our evaluation pipeline. We recommend including mRNA with two additional omics types to conform with our benchmark settings. For more robust biomarker prioritization, aggregating rankings from selected methods via RRA [58] can reduce method-specific biases.

3 Discussion

In this study, we presented a systematic benchmark for computational biomarker identification from multi-omics data, evaluated against clinically validated reference biomarkers. The results uncovered certain methodological designs that generally exhibited strong or weak accuracy and stability in biomarker identification, providing guiding principles for future method development and practical recommendations for method selection across different research scenarios. Analysis of omics type contributions indicates 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-world 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 computationally reproducible candidates that may inform potential biological and clinical investigations. Finally, the released evaluation framework enables researchers to benchmark new methods against the 20 baselines, or to apply selected methods for biomarker identification on other multi-omics data.

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 [5–7] and used rank-based metrics to mitigate the effects of missing annotations, biomarkers that have yet to be discovered or clinically characterized are naturally absent from evaluation. This is particularly relevant for therapeutic tasks, where the number of clinically validated biomarkers is smaller, 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 [8] 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.

Another limitation is that all benchmark tasks were derived from TCGA and therefore do not constitute external validation on independent non-TCGA cohorts. Such datasets remain scarce when requiring matched mRNA, miRNA, DNA methylation, CNV, and SNV profiles together with biomarker-aligned clinical annotations and sufficient sample size. Nevertheless, TCGA includes substantial cross-center heterogeneity, as samples within each task originate from multiple tissue source sites, and their compositions vary across both tasks and cross-validation folds (Supplementary Fig. 24). This provides a partial internal test of robustness, but dedicated benchmarking on more independent cohorts will be a valuable direction for future work.

Our conclusions should also be interpreted within the scope of gene-centered multi-omics biomarkers. The curated biomarker set spans multiple clinically established alteration classes, covering major mechanisms with established clinical significance (Supplementary Table 3). However, many therapeutic interventions and drug responses are mediated through protein abundance, post-translational regulation, and signal transduction pathways, and the associated biomarkers remain underrepresented in both current oncology knowledge bases and cancer cohorts. Incorporating such modalities with matched clinical reference standards will be a potential direction for extensions.

4 Methods

4.1 Overview of benchmarked methods

Here we give an overview of the benchmarked methods. A detailed description of each is further provided in Supplementary Methods.

4.1.1 Deep learning methods

Based on architectural backbones, DL methods can be broadly grouped into four categories: FNN-based (GENIUS [25], P-Net [21]), autoencoder-based (TMO-Net [13], CustOmics [14]), graph-based (MOGLAM [17], GNN-SubNet [16], MOGONET [15], MORE [23], MoAGL-SA [24]), and transformer-based (DePathNet [19], DeepKEGG [18], Pathformer [20]) (Fig. 1c).

For FNN-based methods, P-Net constructs a hierarchical sparse neural network based on gene-pathway and pathway-biological process connections [21]. 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 [25].

Autoencoder-based methods learn low-dimensional 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 [13]. CustOmics uses a hierarchical mixed-integration strategy, learning omics-specific sub-representations via separate autoencoders, which are then combined through a central variational autoencoder [14].

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 [15, 17, 23, 24]. The method using genes as graph nodes (GNN-SubNet) constructs the graph using protein-protein interaction (PPI) network as prior knowledge, learning graph feature representations followed by global-pooling for sample-level predictions [16].

Transformer-based methods integrate pathway knowledge into the model architecture. DePathNet 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 DePathNet and Pathformer use early fusion by integrating multi-omics features before model inputs, while DeepKEGG applies intermediate fusion via omics-specific encoders [18–20].

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 [62] and Integrated Gradients [63] (IG), as well as other types of methods such as SHAP [64] and GNNExplainer [65]. 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.

4.1.2 Statistical and machine learning methods

In addition to DL methods, multi-omics biomarker identification methods based on more conventional statistical or ML algorithms can be traced earlier in the research line. MCIA [47] was one

of the earliest models for multi-omics integration and biomarker identification, performing joint projections of omics into low-dimensional spaces by maximizing their shared covariance structure. MOFA [10] introduced a statistical framework using matrix factorization for factor analysis. More recently, GAUDI [12] applied UMAP-based transformations followed by XGBoost or random forest with SHAP [64] for biomarker identification. These methods are unsupervised without relying on sample-level labels.

In contrast, DIABLO [11] and asmbPLS-DA [48] 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 [40] 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 [22] applies a greedy decision forest over a PPI network using Gini importance for feature scoring. DPM [49] aims at pathway-level modeling but identifies gene-level biomarkers by integrating regulatory directionality (positive/negative associations) across omics with a p-value merging algorithm. Unlike DL methods, most statistical and ML methods achieve *ante hoc* biomarker identification via feature weights updated during model optimization.

4.2 Collection of gold-standard biomarkers

A critical challenge precluding 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 [5], OncoKB [6], and CGI [7]. 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 [8] (see Supplementary Table 2 for details) and retained only those with Tier I evidence. These biomarkers are supported by professional guidelines, FDA-approved therapies, or well-powered clinical trials [8]. 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), as were the gene names from the other two sources, to resolve aliases. Therapeutic agent names were harmonized using the standardized terms collected by Ding *et al.* [66], supplemented by a manually defined mapping to resolve investigational codes, salt forms, spelling errors, and formatting inconsistencies. This curation enables a direct and robust assessment of a method's ability to identify biomarkers that have a high likelihood of succeeding clinical trials and achieving widespread application.

4.3 Derivation of gene-level rankings

Since the collected biomarkers are at the 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

(Supplementary Methods), 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.

4.4 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 [58] 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; Supplementary Methods), 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.

4.5 Evaluation metrics

4.5.1 Accuracy metrics

We adopt rank-based metrics for evaluating 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.

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

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

Normalized Discounted Cumulative Gain. Normalized Discounted Cumulative Gain (NDCG) has been widely used in information retrieval [67, 68] studies and can measure the prioritization of relevant items in top rankings. The Discounted Cumulative Gain (DCG) is defined as

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

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

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

The Normalized DCG is then

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

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 [69]. Let $r_{(1)} := \min_{b \in B} r_b$ denote the rank of the highest-ranked biomarker, RR is defined as

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

assigning a score of 1 when a biomarker occupies the highest rank 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,

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

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.

4.5.2 Stability metrics

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

Average Kendall's τ. Kendall's τ [70] 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 lower-ranked genes, it may be significantly affected by the instability at lower ranks. We therefore further adopted Rank-Biased Overlap (RBO) [71], 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.* [71]. 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.

5 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.

6 Code availability

Our benchmark is publicly available at https://github.com/athanzli/CancerMOBI-Bench.

7 Data availability

Our benchmarking datasets and results are available at 10.5281/zenodo.17860662 [72]. 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 [73] database at https://string-db.org/ (accessed October 28 2024).

8 Author contributions statement

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.

9 Conflicts of interest

The authors declare that they have no competing interests.

10 Funding

This research received no external funding.

11 Author description

Athan Z. Li is a Ph.D. student in the department of Computer Science at the University of Southern California.

Yuxuan Du is an Assistant Professor of Electrical Engineering at The University of Texas at San Antonio.

Yan Liu is a Professor of Computer Science with a joint appointment with Electrical Engineering at the University of Southern California.

Liang Chen is a Professor of Quantitative and Computational Biology at the University of Southern California.

Ruishan Liu is an Assistant Professor of Computer Science with a joint appointment with Quantitative and Computational Biology at the University of Southern California.

References

[1] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 1A:68–77, 2015. ISSN 1428-2526. doi: 10.5114/wo.2014.47136. URL http://www.termedia.pl/doi/10.5114/wo.2014.47136. Publisher: Termedia Sp. z.o.o.

[2] Philipp Mertins, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang, Jana W. Qiao, Song Cao, Francesca Petralia, Emily Kawaler, Filip Mundt, Karsten Krug, Zhidong Tu, Jonathan T. Lei, Michael L. Gatza, Matthew Wilkerson, Charles M. Perou, Venkata Yellapantula, Kuan-lin Huang, Chenwei Lin, Michael D. McLellan, Ping Yan, Sherri R. Davies, R. Reid Townsend, Steven J. Skates, Jing Wang, Bing Zhang, Christopher R. Kinsinger, Mehdi Mesri, Henry Rodriguez, Li Ding, Amanda G. Paulovich, David Fenyö, Matthew J. Ellis, and Steven A. Carr. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, June 2016. ISSN 1476-4687. doi: 10.1038/nature18003. URL https://www.nature.com/articles/nature18003. Publisher: Nature Publishing Group.

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

[4] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1215-1. URL https://doi.org/10.1186/s13059-017-1215-1.

[5] Malachi Griffith, Nicholas C. Spies, Kilannin Krysiak, Joshua F. McMichael, Adam C. Coffman, Arpad M. Danos, Benjamin J. Ainscough, Cody A. Ramirez, Damian T. Rieke, Lynzey Kujan, Erica K. Barnell, Alex H. Wagner, Zachary L. Skidmore, Amber Wollam, Connor J. Liu, Martin R. Jones, Rachel L. Bilski, Robert Lesurf, Yan-Yang Feng, Nakul M. Shah, Melika Bonakdar, Lee Trani, Matthew Matlock, Avinash Ramu, Katie M. Campbell, Gregory C. Spies, Aaron P. Graubert, Karthik Gangavarapu, James M. Eldred, David E. Larson, Jason R. Walker, Benjamin M. Good, Chunlei Wu, Andrew I. Su, Rodrigo Dienstmann, Adam A. Margolin, David Tamborero, Nuria Lopez-Bigas, Steven J. M. Jones, Ron Bose, David H. Spencer, Lukas D. Wartman, Richard K. Wilson, Elaine R. Mardis, and Obi L. Griffith. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174, February 2017. ISSN 1546-1718. doi: 10.1038/ng.3774. URL https://www.nature.com/articles/ng.3774. Publisher: Nature Publishing Group.

[6] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana

Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, May 2017. doi: 10.1200/PO.17.00011. URL https://ascopubs.org/doi/10.1200/PO.17.00011. Publisher: Wolters Kluwer.

[7] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, Carmen de Torres, Rodrigo Dienstmann, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25, March 2018. ISSN 1756-994X. doi: 10.1186/s13073-018-0531-8. URL https://doi.org/10.1186/s13073-018-0531-8.

[8] Marilyn M. Li, Michael Datto, Eric J. Duncavage, Shashikant Kulkarni, Neal I. Lindeman, Somak Roy, Apostolia M. Tsimberidou, Cindy L. Vnencak-Jones, Daynna J. Wolff, Anas Younes, and Marina N. Nikiforova. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *The Journal of Molecular Diagnostics*, 19 (1):4–23, January 2017. ISSN 1525-1578. doi: 10.1016/j.jmoldx.2016.10.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1525157816302239. Publisher: Elsevier BV.

[9] Prabhakar Chalise, Rama R. Raghavan, and Brooke L. Fridley. InterSIM: Simulation tool for multiple integrative 'omic datasets'. *Computer Methods and Programs in Biomedicine*, 128: 69–74, 2016. doi: 10.1016/j.cmpb.2016.02.011. URL https://doi.org/10.1016/j.cmpb.2016.02.011.

[10] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, June 2018. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20178124. URL https://www.embopress.org/doi/10.15252/msb.20178124.

[11] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, September 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty1054. URL https://doi.org/10.1093/bioinformatics/bty1054.

[12] Pol Castellano-Escuder, Derek K. Zachman, Kevin Han, and Matthey D. Hirschey. GAUDI: interpretable multi-omics integration with UMAP embeddings and density-based clustering. *Nature Communications*, 16(1):5771, July 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60822-1. URL https://www.nature.com/articles/s41467-025-60822-1. Publisher: Nature Publishing Group.

[13] Feng-ao Wang, Zhenfeng Zhuang, Feng Gao, Ruikun He, Shaoting Zhang, Liansheng Wang, Junwei Liu, and Yixue Li. TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. *Genome Biology*, 25(1):149, June 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03293-9. URL https://doi.org/10.1186/s13059-024-03293-9.

[14] Hakim Benkirane, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, 19(3):e1010921, March 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010921. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010921. Publisher: Public Library of Science.

[15] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, June 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23774-w. URL https://www.nature.com/articles/s41467-021-23774-w. Number: 1 Publisher: Nature Publishing Group.

[16] Bastian Pfeifer, Anna Saranti, and Andreas Holzinger. GNN-SubNet: disease subnetwork detection with explainable graph neural networks. *Bioinformatics*, 38(Supplement_2):ii120–ii126, September 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac478. URL https://doi.org/10.1093/bioinformatics/btac478.

[17] Dong Ouyang, Yong Liang, Le Li, Ning Ai, Shanghui Lu, Mingkun Yu, Xiaoying Liu, and Shengli Xie. 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, September 2023. ISSN 00104825. doi: 10.1016/j.compbiomed.2023.107303. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482523007680.

[18] Wei Lan, Haibo Liao, Qingfeng Chen, Lingzhi Zhu, Yi Pan, and Yi-Ping Phoebe Chen. DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Briefings in Bioinformatics*, 25(3):bbae185, May 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae185. URL https://doi.org/10.1093/bib/bbae185.

[19] Zhaoxiang Cai, Rebecca C. Poulos, Adel Aref, Phillip J. Robinson, Roger R. Reddel, and Qing Zhong. DeePathNet: A Transformer-Based Deep Learning Model Integrating Multiomic Data with Cancer Pathways. *Cancer Research Communications*, 4(12):3151–3164, December 2024. ISSN 2767-9764. doi: 10.1158/2767-9764.CRC-24-0285. URL https://doi.org/10.1158/2767-9764.CRC-24-0285.

[20] Xiaofan Liu, Yuhuan Tao, Zilin Cai, Pengfei Bao, Hongli Ma, Kexing Li, Mengtao Li, Yunping Zhu, and Zhi John Lu. Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. *Bioinformatics*, 40(5):btae316, May 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae316. URL https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btae316/7671099.

[21] Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, Taylor E. Arnoff, Jihye Park, William C. Hahn, and Eliezer M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, October 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03922-4. URL https://www.nature.com/articles/s41586-021-03922-4. Publisher: Nature Publishing Group.

[22] Bastian Pfeifer, Hubert Baniecki, Anna Saranti, Przemyslaw Biecek, and Andreas Holzinger. Multi-omics disease module detection with an explainable Greedy Decision Forest. *Scientific Reports*, 12(1):16857, October 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-21417-8. URL

https://www.nature.com/articles/s41598-022-21417-8. Publisher: Nature Publishing Group.

[23] Yuhan Wang, Zhikang Wang, Xuan Yu, Xiaoyu Wang, Jiangning Song, Dong-Jun Yu, and Fang Ge. MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification. *Briefings in Bioinformatics*, 26(1):bbae658, January 2025. ISSN 1477-4054. doi: 10.1093/bib/bbae658. URL https://doi.org/10.1093/bib/bbae658.

[24] Lei Cheng, Qian Huang, Zhengqun Zhu, Yanan Li, Shuguang Ge, Longzhen Zhang, and Ping Gong. MoAGL-SA: a multi-omics adaptive integration method with graph learning and self attention for cancer subtype classification. *BMC Bioinformatics*, 25(1):364, November 2024. ISSN 1471-2105. doi: 10.1186/s12859-024-05989-y. URL https://doi.org/10.1186/s12859-024-05989-y.

[25] Mateo Sokač, Asbjørn Kjær, Lars Dyrskjøt, Benjamin Haibe-Kains, Hugo J. W. L. Aerts, and Nicolai J. Birkbak. GENIUS: GEnome traNsformatIon and spatial representation of mUltiomicS data. *eLife*, 12, August 2023. doi: 10.7554/eLife.87133.2. URL https://elifesciences.org/reviewed-preprints/87133. Publisher: eLife Sciences Publications Limited.

[26] Yuxing Lu, Rui Peng, Lingkai Dong, Kun Xia, Renjie Wu, Shuai Xu, and Jinzhuo Wang. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Briefings in Bioinformatics*, 24(6):bbad378, November 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad378. URL https://doi.org/10.1093/bib/bbad378.

[27] Hryhorii Chereda, Andreas Leha, and Tim Beißbarth. 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, May 2024. ISSN 0933-3657. doi: 10.1016/j.artmed.2024.102840. URL https://www.sciencedirect.com/science/article/pii/S0933365724000824.

[28] Justine Labory, Evariste Njomgue-Fotso, and Silvia Bottini. 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, December 2024. ISSN 2001-0370. doi: 10.1016/j.csbj.2024.03.016. URL https://www.csbj.org/article/S2001-0370(24)00069-2/fulltext. Publisher: Elsevier.

[29] Qingxia Yang, Yaguo Gong, and Feng Zhu. Critical Assessment of the Biomarker Discovery and Classification Methods for Multiclass Metabolomics. *Analytical Chemistry*, 95(13):5542–5552, April 2023. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.2c04402. URL https://pubs.acs.org/doi/10.1021/acs.analchem.2c04402.

[30] Matej Petković, Ivica Slavkov, Dragi Kocev, and Sašo Džeroski. Biomarker discovery by feature ranking: Evaluation on a case study of embryonal tumors. *Computers in Biology and Medicine*, 128:104143, January 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.104143. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482520304741. Publisher: Elsevier BV.

[31] Animesh Acharjee, Joseph Larkman, Yuanwei Xu, Victor Roth Cardoso, and Georgios V. Gkoutos. A random forest based biomarker discovery and power analysis framework for diag-

nostics research. *BMC Medical Genomics*, 13(1):178, November 2020. ISSN 1755-8794. doi: 10.1186/s12920-020-00826-6. URL https://doi.org/10.1186/s12920-020-00826-6.

[32] Nicoletta Dessì, Emanuele Pascariello, and Barbara Pes. A Comparative Analysis of Biomarker Selection Techniques. *BioMed Research International*, 2013(1):387673, 2013. ISSN 2314-6141. doi: 10.1155/2013/387673. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2013/387673.

[33] Barbara Di Camillo, Tiziana Sanavia, Matteo Martini, Giuseppe Jurman, Francesco Sambo, Annalisa Barla, Margherita Squillario, Cesare Furlanello, Gianna Toffolo, and Claudio Cobelli. Effect of Size and Heterogeneity of Samples on Biomarker Discovery: Synthetic and Real Data Assessment. *PLOS ONE*, 7(3):e32200, March 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0032200. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032200. Publisher: Public Library of Science.

[34] Seppo Karrila, Julian Hock Ean Lee, and Greg Tucker-Kellogg. A comparison of Methods for Data-Driven Cancer Outlier Discovery, and An Application Scheme to Semisupervised Predictive Biomarker Discovery. *Cancer Informatics*, 10:CIN.S6868, January 2011. ISSN 1176-9351. doi: 10.4137/CIN.S6868. URL https://doi.org/10.4137/CIN.S6868. Publisher: SAGE Publications Ltd STM.

[35] JiYoon Park, Jae Won Lee, and Mira Park. Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. *BioData Mining*, 16(1):18, July 2023. ISSN 1756-0381. doi: 10.1186/s13040-023-00334-0. URL https://doi.org/10.1186/s13040-023-00334-0.

[36] Tapas Bhadra, Saurav Mallik, Neaj Hasan, and Zhongming Zhao. 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, April 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04678-y. URL https://doi.org/10.1186/s12859-022-04678-y.

[37] Nathaniel M. Crabtree, Jason H. Moore, John F. Bowyer, and Nysia I. George. Multi-class computational evolution: development, benchmark evaluation and application to RNA-Seq biomarker discovery. *BioData Mining*, 10(1):13, April 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0134-8. URL https://doi.org/10.1186/s13040-017-0134-8.

[38] Anna Halama, Shaza Zaghlool, Gaurav Thareja, Sara Kader, Wadha Al Muftah, Marjonneke Mook-Kanamori, Hina Sarwath, Yasmin Ali Mohamoud, Nisha Stephan, Sabine Ameling, Maja Pucic Baković, Jan Krumsiek, Cornelia Prehn, Jerzy Adamski, Jochen M. Schwenk, Nele Friedrich, Uwe Völker, Manfred Wuhrer, Gordan Lauc, S. Hani Najafi-Shoushtari, Joel A. Malek, Johannes Graumann, Dennis Mook-Kanamori, Frank Schmidt, and Karsten Suhre. A roadmap to the molecular human linking multiomics with population traits and diabetes subtypes. *Nature Communications*, 15(1):7111, August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51134-x. URL https://www.nature.com/articles/s41467-024-51134-x. Publisher: Nature Publishing Group.

[39] Ziyi Yang, Naiqi Wu, Yong Liang, Hui Zhang, and Yanqiong Ren. SMSPL: Robust Multimodal Approach to Integrative Analysis of Multiomics Data. *IEEE Transactions on Cybernetics*, 52(4):2082–2095, April 2022. ISSN 2168-2267, 2168-2275. doi: 10.1109/TCYB.2020.3006240. URL https://ieeexplore.ieee.org/document/9146338/.

[40] Julien Hédou, Ivana Marić, Grégoire Bellan, Jakob Einhaus, Dyani K. Gaudillière, Francois-Xavier Ladant, Franck Verdonk, Ina A. Stelzer, Dorien Feyaerts, Amy S. Tsai, Edward A. Ganio, Maximilian Sabayev, Joshua Gillard, Jonas Amar, Amelie Cambriel, Tomiko T. Oskotsky, Alennie Roldan, Jonathan L. Golob, Marina Sirota, Thomas A. Bonham, Masaki Sato, Maïgane Diop, Xavier Durand, Martin S. Angst, David K. Stevenson, Nima Aghaeepour, Andrea Montanari, and Brice Gaudillière. Discovery of sparse, reliable omic biomarkers with Stabl. *Nature Biotechnology*, 42(10):1581–1593, October 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-02033-x. URL https://www.nature.com/articles/s41587-023-02033-x. Publisher: Nature Publishing Group.

[41] Xiaokang Zhang, Inge Jonassen, and Anders Goksøyr. Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In Nakaya Helder I., editor, *Bioinformatics*. Exon Publications, Brisbane (AU), 2021. ISBN 978-0-6450017-1-6. URL http://www.ncbi.nlm.nih.gov/books/NBK569564/.

[42] Yingxia Li, Ulrich Mansmann, Shangming Du, and Roman Hornung. Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinformatics*, 23(1):412, October 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04962-x. URL https://doi.org/10.1186/s12859-022-04962-x.

[43] Daniel F. Hayes. Biomarker validation and testing. *Molecular Oncology*, 9(5):960–966, 2015. doi: https://doi.org/10.1016/j.molonc.2014.10.004. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/j.molonc.2014.10.004.

[44] LM McShane. In pursuit of greater reproducibility and credibility of early clinical biomarker research. *Clinical and Translational Science*, 10(2):58–60, 2017. doi: https://doi.org/10.1111/cts.12449. URL https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.12449.

[45] Scott E. Kern. Why Your New Cancer Biomarker May Never Work: Recurrent Patterns and Remarkable Diversity in Biomarker Failures. *Cancer Research*, 72(23):6097–6101, November 2012. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-12-3232. URL https://doi.org/10.1158/0008-5472.CAN-12-3232.

[46] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. ISSN 1546-1718. doi: 10.1038/ng.2764. URL https://www.nature.com/articles/ng.2764. Publisher: Nature Publishing Group.

[47] Chen Meng, Bernhard Kuster, Aedín C. Culhane, and Amin Moghaddas Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162, May 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-162. URL https://doi.org/10.1186/1471-2105-15-162.

[48] Runzhi Zhang and Susmita Datta. Adaptive Sparse Multi-Block PLS Discriminant Analysis: An Integrative Method for Identifying Key Biomarkers from Multi-Omics Data. *Genes*, 14(5):961, May 2023. ISSN 2073-4425. doi: 10.3390/genes14050961. URL https://www.mdpi.com/2073-4425/14/5/961. Publisher: Multidisciplinary Digital Publishing Institute.

[49] Mykhaylo Slobodyanyuk, Alexander T. Bahcheli, Zoe P. Klein, Masroor Bayati, Lisa J. Strug, and Jüri Reimand. Directional integration and pathway enrichment analysis for multi-omics

data. *Nature Communications*, 15(1):5690, July 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49986-4. URL https://www.nature.com/articles/s41467-024-49986-4. Publisher: Nature Publishing Group.

[50] Eliezer M. Van Allen, Kent W. Mouw, Philip Kim, Gopa Iyer, Nikhil Wagle, Hikmat Al-Ahmadie, Cong Zhu, Irina Ostrovnaya, Gregory V. Kryukov, Kevin W. O'Connor, John Sfakianos, Ilana Garcia-Grossman, Jaegil Kim, Elizabeth A. Guancial, Richard Bambury, Samira Bahl, Namrata Gupta, Deborah Farlow, Angela Qu, Sabina Signoretti, Justine A. Barletta, Victor Reuter, Jesse Boehm, Michael Lawrence, Gad Getz, Philip Kantoff, Bernard H. Bochner, Toni K. Choueiri, Dean F. Bajorin, David B. Solit, Stacey Gabriel, Alan D'Andrea, Levi A. Garraway, and Jonathan E. Rosenberg. Somatic ERCC2 Mutations Correlate with Cisplatin Sensitivity in Muscle-Invasive Urothelial Carcinoma. *Cancer Discovery*, 4(10):1140–1153, September 2014. ISSN 2159-8274. doi: 10.1158/2159-8290.CD-14-0623. URL https://doi.org/10.1158/2159-8290.CD-14-0623.

[51] Qiang Li, Alexis W. Damish, Zoë Frazier, David Liu, Elizaveta Reznichenko, Atanas Kamburov, Andrew Bell, Huiyong Zhao, Emmet J. Jordan, S. Paul Gao, Jennifer Ma, Philip H. Abbosh, Joaquim Bellmunt, Elizabeth R. Plimack, Jean-Bernard Lazaro, David B. Solit, Dean Bajorin, Jonathan E. Rosenberg, Alan D. D'Andrea, Nadeem Riaz, Eliezer M. Van Allen, Gopa Iyer, and Kent W. Mouw. 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, February 2019. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-18-1001. URL https://doi.org/10.1158/1078-0432.CCR-18-1001.

[52] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html. Submitted 11/20; Revised 12/21; Published 08/22.

[53] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29, Barcelona, Spain, 2016. 30th Conference on Neural Information Processing Systems (NIPS 2016).

[54] Lenny Dang, David W. White, Stefan Gross, Bryson D. Bennett, Mark A. Bittinger, Edward M. Driggers, Valeria R. Fantin, Hyun Gyung Jang, Shengfang Jin, Marie C. Keenan, Kevin M. Marks, Robert M. Prins, Patrick S. Ward, Katharine E. Yen, Linda M. Liau, Joshua D. Rabinowitz, Lewis C. Cantley, Craig B. Thompson, Matthew G. Vander Heiden, and Shinsan M. Su. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature*, 462(7274):739–744, December 2009. ISSN 1476-4687. doi: 10.1038/nature08617. URL https://www.nature.com/articles/nature08617.

[55] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Hye-Jung E. Chun, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Erin Pleasance, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, Nam H. Pho, Scott L. Carter, Steven E. Schumacher, Barbara Tabak, Bryan Hernandez, Jeff Gentry, Huy Nguyen, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Gad Getz, Stacey B. Gabriel, Matthew Meyerson, Lynda Chin, Peter J. Park, Raju Kucherlapati, Katherine A. Hoadley, J. Todd Auman, Cheng Fan, Yidi J. Turman, Yan Shi, Ling Li, Michael D. Topal, Xiaping He, Hann-Hsiang Chao, Aleix Prat, Grace O. Silva, Michael D. Iglesia, Wei Zhao, Jerry Usary, Jonathan S. Berg, Michael Adams, Jessica Booker, Junyuan Wu, Anisha Gulabani, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Simeen Malik, Swapna Mahurkar, Hui Shen, Daniel J. Weisenberger, Timothy Triche Jr, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Benjamin P. Berman, David J. Van Den Berg, Stephen B. Baylin, Peter W. Laird, Chad J. Creighton, Lawrence A. Donehower, Gad Getz, Michael Noble, Doug Voet, Gordon Saksena, Nils Gehlenborg, Daniel DiCara, Juinhua Zhang, Hailei Zhang, Chang-Jiun Wu, Spring Yingchun Liu, Michael S. Lawrence, Lihua Zou, Andrey Sivachenko, Pei Lin, Petar Stojanov, Rui Jing, Juok Cho, Raktim Sinha, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Peter J. Park, Lynda Chin, Sheila Reynolds, Richard B. Kreisberg, Brady Bernard, Ryan Bressler, Timo Erkkila, Jake Lin, Vesteinn Thorsson, Wei Zhang, Ilya Shmulevich, Giovanni Ciriello, Nils Weinhold, Nikolaus Schultz, Jianjiong Gao, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Marc Ladanyi, Chris Sander, Pavana Anur, Paul T. Spellman, Yiling Lu, Wenbin Liu, Roel R. G. Verhaak, Gordon B. Mills, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod D. Casasent, Chris Wakefield, Anna K. Unruh, Keith Baggerly, Kevin Coombes, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, Jingchun Zhu, Christopher C. Szeto, Gary K. Scott, Christina Yau, Evan O. Paull, Daniel Carlin, Christopher Wong, Artem Sokolov, Janita Thusberg, Sean Mooney, Sam Ng, Theodore C. Goldstein, Kyle Ellrott, Mia Grifford, Christopher Wilks, Singer Ma, Brian Craft, Chunhua Yan, Ying Hu, Daoud Meerzaman, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron D. Black, Robert E. Pyatt, Peter White, Erik J. Zmuda, Jessica Frick, Tara M. Lichtenberg, Robin Brookens, Myra M. George, Mark A. Gerken, Hollie A. Harper, Kristen M. Leraas, Lisa J. Wise, Teresa R. Tabler, Cynthia McAllister, Thomas Barr, Melissa Hart-Kothari, Katie Tarvin, Charles Saller, George Sandusky, Colleen Mitchell, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Oleg Dolzhansky, Mikhail Abramov, Olga Voronina, Olga Potapova, Jeffrey R. Marks, Wiktoria M. Suchorska, Dawid Murawa, Witold Kycler, Matthew Ibbs, Konstanty Korski, Arkadiusz Spychała, Paweł Murawa, Jacek J. Brzeziński, Hanna Perz, Radosław Łaźniak, Marek Teresiak, Honorata Tatka, Ewa Leporowska, Marta Bogusz-Czerniewicz, Julian Malicki, Andrzej Mackiewicz, Maciej Wiznerowicz, Xuan Van Le, Bernard Kohl, Nguyen Viet Tien, Richard Thorp, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Tran Viet The Phuong, Huynh Quyet Thang, Khurram Zaki Khan,

Robert Penny, David Mallery, Erin Curley, Candace Shelton, Peggy Yena, James N. Ingle, Fergus J. Couch, Wilma L. Lingle, Tari A. King, Ana Maria Gonzalez-Angulo, Gordon B. Mills, Mary D. Dyer, Shuying Liu, Xiaolong Meng, Modesto Patangan, 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, Chapel Hill University of North Carolina, 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, Santa Cruz/Buck Institute University of California, 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, and MD Anderson Cancer Center. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012. ISSN 1476-4687. doi: 10.1038/nature11412. URL https://doi.org/10.1038/nature11412.

[56] Ashkan Shahbandi, Hoang D. Nguyen, and James G. Jackson. TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. *Trends in Cancer*, 6(2):98–110, February 2020. ISSN 2405-8033. doi: 10.1016/j.trecan.2020.01.007. URL https://www.cell.com/trends/cancer/abstract/S2405-8033(20)30020-0. Publisher: Elsevier.

[57] Falk Nimmerjahn and Jeffrey V. Ravetch. Fc receptors as regulators of immune responses. *Nature Reviews Immunology*, 8(1):34–47, January 2008. ISSN 1474-1741. doi: 10.1038/nri2206. URL https://www.nature.com/articles/nri2206. Publisher: Nature Publishing Group.

[58] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, February 2012. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btr709. URL https://academic.oup.com/bioinformatics/article/28/4/573/213339.

[59] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012. doi: 10.1038/nature11252.

[60] A. Gordon Robertson, Jaegil Kim, Hikmat Al-Ahmadie, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556, 2017. doi: 10.1016/j.cell.2017.09.007.

[61] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015. doi: 10.1056/NEJMoa1402121.

[62] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3145–3153, Sydney, NSW, Australia, August 2017. JMLR.org.

[63] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, ,, 06–11 Aug 2017. PMLR. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[64] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, ,, 2017. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[65] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[66] Zijian Ding, Songpeng Zu, and Jin Gu. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, 32(19):2891–2895, 06 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw344. URL https://doi.org/10.1093/bioinformatics/btw344.

[67] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

[68] Tie-Yan Liu. Learning to rank for information retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 904, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301534. doi: 10.1145/1835449.1835676. URL https://doi.org/10.1145/1835449.1835676.

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

[70] M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, June 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

[71] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, November 2010. ISSN 1046-8188. doi: 10.1145/1852102.1852106. URL https://doi.org/10.1145/1852102.1852106.

[72] Athan Z. Li. Data and results for "benchmarking computational methods for multi-omics biomarker discovery in cancer", December 2025. URL https://zenodo.org/records/17860662.

[73] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1000. URL https://doi.org/10.1093/nar/gkac1000.