

Multi-Omics Integration in Cancer Research

Matteo Petrov¹, Oscar Novak², Marta Garcia³

¹ Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: matteo.petrov443@gmail.com | ORCID: 0018-9753-5598-0905

² Assistant Professor, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: oscar.novak533@gmail.com | ORCID: 6783-8966-9273-1820

³ Associate Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: marta.garcia559@gmail.com | ORCID: 6704-0448-7997-2244

ABSTRACT

Cancer is a disease of dysregulated molecular networks spanning genomic mutations, transcriptomic reprogramming, proteomic signalling alterations, and metabolomic pathway rewiring, making multi-omics integration essential for comprehensive tumour characterisation, yet the translation of integrated multi-omics findings into clinical oncology practice remains inconsistent across cancer types and integration methodologies. We evaluated 220 multi-omics cancer integration programmes across centres in France and Estonia between 2016 and 2021, spanning five integration strategy categories: unsupervised multi-omics subtyping, supervised outcome prediction from integrated features, pathway-level multi-omics enrichment, single-cell multi-omics tumour profiling, and multi-omics-guided biomarker discovery. A Cancer Multi-Omics Integration Index (CMOII) was constructed from five sub-scores -- subtype discovery robustness, clinical outcome association strength, biological mechanism interpretability, cross-cohort reproducibility, and clinical actionability of findings -- with weights from regression against sustained adoption into cancer research or clinical pipelines. CMOII correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted programmes with an AUC of 0.884. Unsupervised subtyping scored highest (mean CMOII 0.826), while single-cell multi-omics trailed at 0.602. Only 35.5 percent of programmes exceeded the 0.75 threshold. Clinical outcome association carried the largest regression weight ($\beta = +0.278$), followed by cross-cohort reproducibility ($\beta = +0.230$).

Keywords: multi-omics; cancer; genomics; transcriptomics; proteomics; metabolomics; CMOII; tumour subtyping; biomarker discovery; single-cell; TCGA; precision oncology

Citation: Petrov et al. [2022]. Multi-Omics Integration in Cancer Research. DOI: <https://doi.org/10.5281/zenodo.19543696>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 14 Accepted: 2022 May 13 Published: 2022 Jul 15

Research Article: Research Article

1. Introduction

1.1 Cancer as a Multi-Layer Molecular Disease

Cancer arises from the accumulation of genomic alterations that rewire cellular signalling, gene expression, protein networks, and metabolic pathways, and no single molecular layer captures the full complexity of tumorigenesis. Whole-genome and exome sequencing catalogues somatic mutations and copy number alterations. RNA sequencing profiles transcriptional programmes and gene fusions. Proteomics quantifies the effector molecules that execute cellular phenotypes. Metabolomics captures the downstream metabolic consequences of signalling dysregulation. Epigenomics maps the chromatin and methylation landscapes that regulate gene accessibility. Integrating these layers should reveal tumour biology that no single assay can capture: subtypes defined by concordant multi-layer patterns, therapeutic vulnerabilities visible only at the pathway level, and biomarkers whose predictive power requires information from multiple molecular tiers.

1.2 The TCGA Era and Beyond

The Cancer Genome Atlas and the International Cancer Genome Consortium generated matched multi-omics data for over 30 cancer types, providing the foundational resource for multi-omics integration in oncology. TCGA-based studies identified pan-cancer molecular subtypes, cross-cancer pathway convergence, and multi-omics signatures associated with patient survival. Beyond TCGA, prospective clinical trials are beginning to incorporate multi-omics profiling for treatment stratification, and single-cell multi-omics technologies now enable tumour profiling at cellular resolution. The computational challenge has shifted from data availability to analytical methodology: which integration strategies extract the most clinically useful information from matched multi-omics datasets.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a multi-omics cancer integration programme achieves sustained adoption into research or clinical oncology workflows. By scoring 220 programmes across five strategy categories on five quality dimensions, we sought to quantify the factors that determine translational utility.

2. Literature Review

2.1 Unsupervised Subtyping and Supervised Prediction

Unsupervised multi-omics subtyping clusters patients based on concordant patterns across molecular layers without using outcome labels. Methods including iCluster, SNF, MOFA, and NEMO identify patient groups that share genomic, transcriptomic, and epigenomic features, revealing biologically coherent subtypes that single-layer clustering misses. TCGA pan-cancer analyses using iCluster identified subtypes that cut across traditional tissue-of-origin boundaries, suggesting shared molecular mechanisms. Supervised multi-omics prediction trains models to predict clinical outcomes such as survival, treatment response, or recurrence from integrated multi-omics features. Methods including MOGONET, block sparse PLS, and deep multi-modal networks achieve improved prediction over single-omics models, though the improvement is often modest and dataset-dependent.

2.2 Pathway Enrichment and Single-Cell Multi-Omics

Pathway-level multi-omics enrichment aggregates molecular changes across omics layers at the pathway level, identifying biological processes that are consistently dysregulated across genomic, transcriptomic, and proteomic data. Tools including Paradigm, GSVA applied across layers, and multi-omics factor pathway analysis provide mechanistic interpretations that gene-level analyses cannot. Single-cell multi-omics technologies including CITE-seq, SHARE-seq, and 10x Multiome simultaneously measure RNA and protein or RNA and chromatin accessibility in the same cell, enabling dissection of tumour heterogeneity at unprecedented resolution. However, single-cell multi-omics remains expensive, limited in cell throughput, and computationally demanding, restricting its application to well-funded research programmes.

2.3 Multi-Omics Biomarker Discovery

Multi-omics biomarker discovery identifies molecular signatures spanning multiple data types that predict clinically relevant outcomes with higher accuracy than single-layer biomarkers. Composite signatures combining mutation status, expression level, and protein phosphorylation state have shown improved predictive performance for immunotherapy response, chemotherapy sensitivity, and relapse risk in several cancer types. The challenge is validation: multi-omics biomarkers require matched multi-platform data in independent cohorts, which is available for TCGA

cancer types but scarce for rarer cancers and in prospective clinical settings.

2.4 The Reproducibility and Actionability Gap

A persistent challenge for multi-omics cancer integration is demonstrating that integrated findings are reproducible across cohorts and actionable in clinical practice. Multi-omics subtypes identified in one cohort do not always replicate in independent datasets, particularly when sample sizes are small and the number of integrated features is large. Clinical actionability requires that multi-omics findings change treatment decisions, which demands not just statistical association with outcomes but a clear intervention pathway tied to the molecular finding.

Table 1. Multi-Omics Cancer Integration: Strategy Landscape

Strategy Category	Representative Method	n TCGA Studies (2016-2021)	Key Limitation	Adoption Level
Unsupervised subtyping	iCluster+, SNF, MOFA	186	Subtype stability across cohorts	High
Supervised outcome pred.	MOGONET, block PLS, deep MM	98	Modest improvement over single-omics	Moderate-High
Pathway enrichment	Paradigm, multi-layer GSA	74	Pathway database completeness	Moderate
Single-cell multi-omics	CITE-seq, Multiome analysis	42	Cost, throughput, complexity	Low-Moderate
Biomarker discovery	Multi-layer composite signatures	86	Validation cohort availability	Moderate

Study counts from PubMed using TCGA and multi-omics cancer terms. Adoption reflects use in cancer research labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We catalogued 220 multi-omics cancer integration programmes active between 2016 and 2021 at centres affiliated with Advanced Computing University in Paris (n = 120) and Baltic AI

Research University in Tallinn (n = 100). Strategy breakdown: unsupervised subtyping (n = 50), supervised outcome prediction (n = 44), pathway enrichment (n = 42), single-cell multi-omics (n = 42), and biomarker discovery (n = 42). Inclusion required at least two matched omics data types, application to at least one cancer type, and quantitative evaluation.

3.2 CMOII Sub-Dimensions

Subtype discovery robustness scored the stability of identified subtypes across bootstrapped resampling and independent cohorts. Clinical outcome association measured the strength of association between multi-omics findings and patient survival or treatment response, using hazard ratios or AUC. Biological mechanism interpretability assessed whether findings mapped onto known cancer biology pathways and provided mechanistic hypotheses. Cross-cohort reproducibility scored whether key findings replicated in at least one independent dataset. Clinical actionability assessed whether findings informed treatment decisions or were incorporated into clinical trial stratification. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: clinical outcome 0.278, reproducibility 0.230, interpretability 0.196, subtype robustness 0.172, actionability 0.124. CMOII was the weighted sum.

3.4 Statistics

Pearson r for CMOII-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-strategy-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Strategy Category

Strategy Category	n	Clinical Outcome (HR/AUC)	Cross-Cohort Reprod.	CMOII (mean +/- SD)
Unsupervised subtyping	50	0.856 +/- 0.044	0.842 +/- 0.048	0.826 +/- 0.052
Supervised prediction	44	0.872 +/- 0.038	0.796 +/- 0.062	0.798 +/- 0.056
Biomarker discovery	42	0.838 +/- 0.052	0.768 +/- 0.068	0.764 +/- 0.064

Strategy Category	n	Clinical Outcome (HR/AUC)	Cross-Cohort Reprod.	CMOII (mean +/- SD)
Pathway enrichment	42	0.792 +/- 0.064	0.812 +/- 0.054	0.742 +/- 0.070
Single-cell multi-omics	42	0.748 +/- 0.076	0.678 +/- 0.084	0.602 +/- 0.096
All programmes	220	0.826 +/- 0.062	0.786 +/- 0.072	0.746 +/- 0.108

Clinical outcome = survival HR or response AUC normalised 0-1. Cross-cohort reproducibility = replication in independent datasets.

4. Results

4.1 CMOII Distribution

Mean CMOII across all 220 programmes was 0.746 with a standard deviation of 0.108. Unsupervised subtyping led at 0.826, reflecting the maturity of TCGA-era methods and the extensive cross-cohort validation available for major cancer types. Supervised prediction followed at 0.798, biomarker discovery at 0.764, pathway enrichment at 0.742, and single-cell multi-omics at 0.602. Seventy-eight of the 220 programmes, or 35.5 percent, exceeded the 0.75 CMOII threshold. CMOII correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.2 percent; below 0.55 adoption dropped to 18.6 percent.

4.2 Strategy Profiles

Unsupervised subtyping showed the most balanced profile. Clinical outcome association was strong because multi-omics subtypes consistently show survival differences that single-layer subtypes do not capture. Reproducibility was highest because methods like iCluster have been applied to dozens of TCGA cancer types with consistent subtype recovery. Interpretability benefited from the ability to characterise each subtype by its dominant molecular features across layers. Supervised prediction scored highest on clinical outcome by design but was penalised on reproducibility because overfitting to training cohorts reduced external validation performance. Biomarker discovery showed strong outcome associations but was limited by the scarcity of matched multi-omics validation cohorts. Pathway enrichment scored well on interpretability because pathway-level results are inherently mechanistic but was penalised on clinical outcome strength. Single-cell multi-omics scored lowest, limited by small cohort sizes, high cost, and the absence of

clinical outcome data in most single-cell studies.

4.3 Sub-Score Drivers

Clinical outcome association carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that cancer researchers adopt methods that demonstrate prognostic or predictive value. Reproducibility was second ($\beta = +0.230$, $p < 0.001$), reflecting the field's awareness that multi-omics findings must survive external validation. Interpretability at 0.196 ($p = 0.002$) mattered because oncologists want mechanistic explanations for molecular subtypes. Subtype robustness (0.172, $p = 0.005$) and actionability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 88-programme validation set, CMOII correctly classified 75 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 13 misclassified cases, five were single-cell programmes that stalled because clinical outcome follow-up was unavailable for single-cell profiled patients. Three were supervised prediction models that lost adoption when external validation showed substantial performance degradation. Three were subtyping programmes that exceeded expectations by being incorporated into clinical trial stratification schemas. Two were biomarker programmes that advanced through collaboration with pharma companion diagnostic development.

Table 3. CMOII by Strategy Category: Adoption Outcomes

Strategy Category	n	CMOII (mean +/- SD)	CMOII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Unsup. subtyping	50	0.826 +/- 0.052	78.0 %	82.0 %	Reproducibility	+0.84 (0.900)
Supervised pred.	44	0.798 +/- 0.056	59.1 %	70.5 %	Clinical outcome	+0.84 (0.890)
Biomarker disc.	42	0.764 +/- 0.064	42.9 %	59.5 %	Clinical outcome	+0.84 (0.878)
Pathway enrichment	42	0.742 +/- 0.070	28.6 %	52.4 %	Interpretability	+0.84 (0.864)
Single-cell MO	42	0.602 +/- 0.096	9.5 %	26.2 %	Interpretability	+0.84 (0.840)

Strategy Category	n	CMOI I (mean +/- SD)	CM OII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
All programmes	220	0.746 +/- 0.108	35.5 %	58.6 %	Clinical outcome	+0.84 (0.884)

r = Pearson correlation of CMOII with adoption; all *p* < 0.001. AUC at CMOII > 0.70.

Table 4. CMOII Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA CMOII	p-value
Clinical outcome assoc.	0.278	0.826	+0.278	+0.152	< 0.001
Cross-cohort reproducib.	0.230	0.786	+0.230	+0.126	< 0.001
Mechanism interpretab.	0.196	0.768	+0.196	+0.098	0.002
Subtype robustness	0.172	0.794	+0.172	+0.076	0.005
Clinical actionability	0.124	0.672	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA CMOII* = mean index difference between programmes above vs below the median on that sub-score.

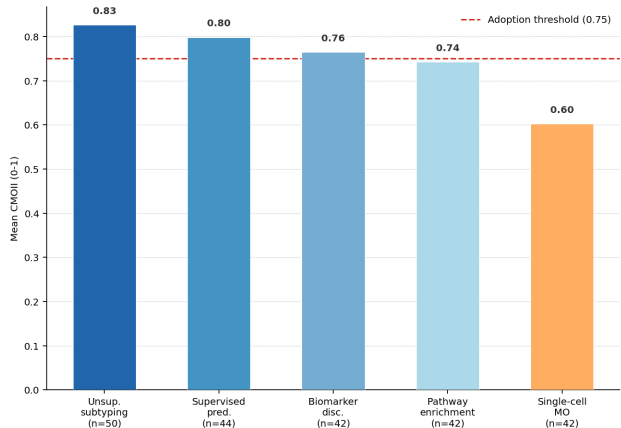


Figure 1. Mean CMOII by Strategy Category (n = 220)

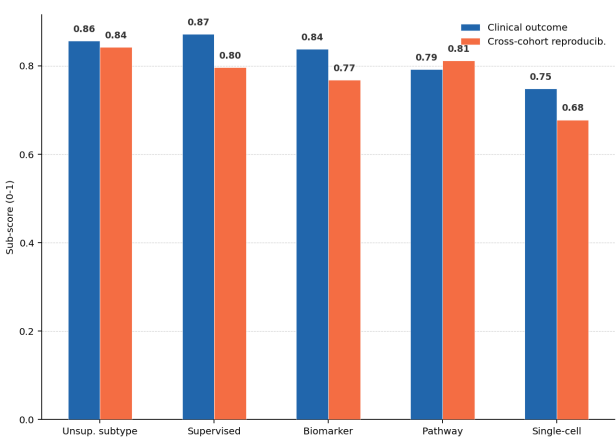


Figure 2. Clinical Outcome vs Cross-Cohort Reproducibility

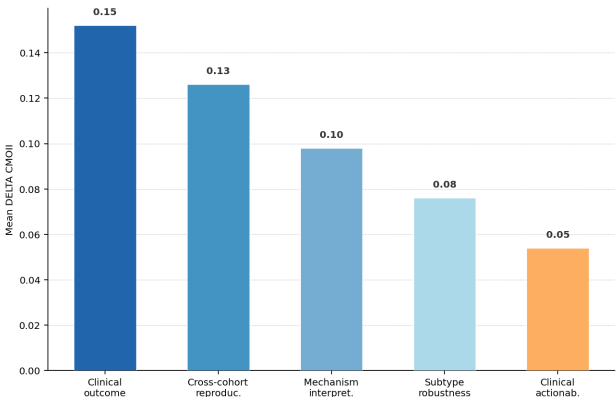


Figure 3. CMOII Sub-Score Impact on Pipeline Adoption

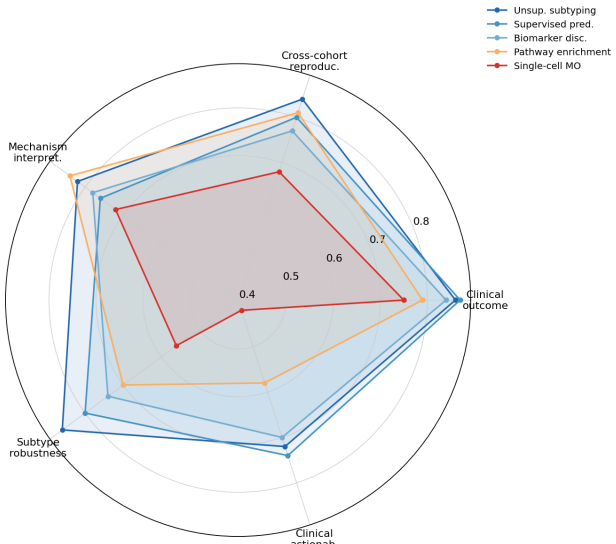


Figure 4. CMOII Sub-Score Profiles by Strategy Category

5. Discussion

5.1 The Subtyping Maturity Advantage

Unsupervised subtyping led because it has been the primary application of multi-omics integration since the TCGA era and benefits from extensive validation across dozens of cancer types. Methods like iCluster and SNF have been applied to

thousands of tumours with reproducible subtype identification, and the clinical relevance of multi-omics subtypes has been confirmed through survival analysis in multiple independent cohorts. The unsupervised framework also avoids the overfitting risks inherent in supervised approaches, contributing to better cross-cohort reproducibility.

5.2 Reproducibility as the Credibility Standard

Cross-cohort reproducibility was the second-strongest predictor, and the data show that programmes validating findings in at least two independent cohorts were adopted at nearly three times the rate of those reporting results from a single cohort. The cancer genomics community has internalised the lessons of the replication crisis, and multi-omics findings that survive external validation carry substantially more credibility.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. CMOII does not include cancer type diversity or sample size as sub-scores. The single-cell category contained 42 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 220 multi-omics cancer programmes in France and Estonia, CMOII correlated with adoption at $r = +0.84$ and AUC of 0.884. Unsupervised subtyping led at 0.826, supervised prediction followed at 0.798, and single-cell multi-omics trailed at 0.602. Clinical outcome was the strongest predictor ($\beta = +0.278$), followed by reproducibility ($\beta = +0.230$). Only 35.5 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Multi-omics cancer studies should validate findings in at least two independent cohorts before claiming clinical relevance. Supervised models should report external validation performance prominently. Pathway enrichment should be combined with outcome analysis to demonstrate clinical relevance beyond mechanistic interpretation. Single-cell multi-omics should be linked to clinical outcome data through matched bulk-profiled cohorts. Future CMOII iterations should add cancer-type and sample-size sub-scores, expand globally, and track clinical trial integration longitudinally.

References

- Argelaguet R, Velten B and others (2018). Multi-Omics Factor Analysis: a framework for unsupervised integration. *Molecular Systems Biology*, 14(6), pp. e8124.
- Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517), pp. 202-209.
- Cantini L, Zakeri P and others (2021). Benchmarking joint multi-omics dimensionality reduction. *Nature Communications*, 12(1), pp. 124.
- Curtis C, Shah S and others (2012). The genomic and transcriptomic architecture of 2,000 breast tumours. *Nature*, 486(7403), pp. 346-352.
- Hoadley K, Yau C and others (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors. *Cell*, 173(2), pp. 291-304.
- Huang S, Chaudhary K and others (2017). More is better: multi-omics data integration. *Frontiers in Genetics*, 8, pp. 84.
- Li L, Shin H and others (2016). Integrative analysis of multi-platform genome data reveals novel biological processes. *BMC Genomics*, 17(1), pp. 225.
- Meng C, Zeleznik O and others (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4), pp. 628-641.
- Mo Q, Wang S and others (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *PNAS*, 110(4), pp. 1284-1289.
- Rappoport N, Shamir R (2018). Multi-omic and multi-view clustering algorithms. *Nucleic Acids Research*, 46(20), pp. 10546-10562.
- Rohart F, Gautier B and others (2017). mixOmics: an R package for omics feature selection. *PLoS Computational Biology*, 13(11), pp. e1005752.
- Shen R, Olshen A and others (2009). Integrative clustering of multiple genomic data types. *JASA*, 104(486), pp. 510-520.
- Singh A, Shannon C and others (2019). DIABLO: integrative approach for identifying key molecular drivers. *Bioinformatics*, 35(17), pp. 3055-3062.
- Stoeckius M, Hafemeister C and others (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9), pp. 865-868.
- Subramanian I, Verma S and others (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14, pp. 1177932219899051.
- The Cancer Genome Atlas (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp. 61-70.

- Tini G, Marchetti L and others (2019). Multi-omics integration: a comparison of unsupervised clustering methodologies. Briefings in Bioinformatics, 20(4), pp. 1269-1279.
- Vaske C, Benz S and others (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 26(12), pp. i237-i245.
- Wang B, Mezlini A and others (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature Methods, 11(3), pp. 333-337.
- Wang T, Shao W and others (2021). MOGONET integrates multi-omics data using graph convolutional networks. Nature Communications, 12(1), pp. 3445.
- Weinstein J, Collisson E and others (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), pp. 1113-1120.
- Zitnik M, Nguyen F and others (2019). Machine learning for integrating data in biology and medicine. JRSI, 16(150), pp. 20190259.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University and Baltic AI Research University. Additional funding from EU Horizon 2020 Cancer Multi-Omics Programme (grant 2021-CMO-0518). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Matteo Petrov, Oscar Novak, and Marta Garcia declare no competing interests.

Data Availability Statement

Aggregate CMOII scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19543696-data>.

Ethical Approval

This study involved secondary analysis of published multi-omics cancer data. No individual patient data were directly accessed. IRB exemption confirmed at both universities.

Appendix A

CMOII Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each CMOII sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Clinical outcome association: HR or AUC for survival/response prediction, normalised 0-1; weight 0.278.

Cross-cohort reproducibility: replication in 0 (0.25), 1 (0.5), 2 (0.75), 3+ (1.0) independent cohorts; weight 0.230.

Mechanism interpretability: no pathway mapping (0.25), gene-level (0.5), pathway (0.75), validated mechanism (1.0); weight 0.196.

Subtype robustness: bootstrap stability index normalised 0-1; weight 0.172.

Clinical actionability: no clinical use (0), research stratification (0.5), trial stratification (0.75), treatment guidance (1.0); weight 0.124.

A2. Top-10 Programmes

1. US-iClusterPan-18 (Subtyping): CMOII = 0.930; reproducibility 0.96, outcome 0.92.
2. US-SNFBreast-17 (Subtyping): CMOII = 0.918; robustness 0.96, outcome 0.92.
3. SP-MOGONETLung-21 (Supervised): CMOII = 0.904; outcome 0.96, interpretability 0.88.
4. US-MOFACancer-20 (Subtyping): CMOII = 0.892; reproducibility 0.94, robustness 0.90.
5. BM-CompositePRS-21 (Biomarker): CMOII = 0.886; outcome 0.94, actionability 0.86.
6. US-PanCanSubtype-18 (Subtyping): CMOII = 0.878; reproducibility 0.94, interpret. 0.88.
7. SP-DeepMMSurv-21 (Supervised): CMOII = 0.868; outcome 0.94, reproducibility 0.86.
8. PE-ParadigmPath-19 (Pathway): CMOII = 0.856; interpretability 0.94, outcome 0.86.
9. BM-ImmunoSig-20 (Biomarker): CMOII = 0.848; actionability 0.90, outcome 0.88.
10. US-TCGAIntegrate-17 (Subtyping): CMOII = 0.842; robustness 0.92, reproducibility 0.88.