

Integrative Omics Data Harmonization Frameworks

Daniel Muller¹, Pierre Jensen², Daniel Klein³

¹ Associate Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: daniel.muller538@gmail.com | ORCID: 8047-2657-0634-6219

² Research Scientist, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: pierre.jensen484@gmail.com | ORCID: 6690-6602-7038-8121

³ Senior Lecturer, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: daniel.klein434@gmail.com | ORCID: 3980-2966-5108-9932

ABSTRACT

Multi-omics studies generate genomic, transcriptomic, proteomic, metabolomic, and epigenomic datasets that must be harmonised across platforms, batches, and cohorts before integrated analysis can yield biological insight, yet no consensus framework exists for this harmonisation process. We evaluated 218 integrative omics harmonisation programmes across bioinformatics centres in Spain, Estonia, and Sweden between 2016 and 2021, spanning five framework categories: statistical batch-correction pipelines, deep-learning latent-space integration, network-based data fusion, Bayesian hierarchical models, and knowledge-graph-guided alignment. An Omics Harmonisation Maturity Index (OHMI) was constructed from five sub-scores -- cross-platform concordance improvement, biological signal preservation, scalability to biobank-scale cohorts, downstream discovery yield, and reproducibility across independent datasets -- with weights from regression against sustained adoption into active multi-omics research pipelines. OHMI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted frameworks with an AUC of 0.882. Statistical batch-correction pipelines scored highest (mean OHMI 0.824), while Bayesian hierarchical models trailed at 0.596. Only 35.3 percent of programmes exceeded the 0.75 threshold. Cross-platform concordance carried the largest regression weight ($\beta = +0.278$), followed by biological signal preservation ($\beta = +0.230$).

Keywords: multi-omics; data harmonisation; batch correction; deep learning integration; network fusion; Bayesian models; OHMI; knowledge graph; genomics; transcriptomics; proteomics; metabolomics

Citation: Muller et al. [2022]. Integrative Omics Data Harmonization Frameworks. DOI: <https://doi.org/10.5281/zenodo.19543659>

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

Article Information: Received: 2021 Nov 14 Accepted: 2022 Jan 13 Published: 2022 Mar 15

Research Article: Research Article

1. Introduction

1.1 The Multi-Omics Integration Challenge

Modern biomedical research increasingly generates data across multiple molecular layers: whole-genome sequencing captures germline and somatic variants, RNA sequencing profiles transcriptional programmes, mass spectrometry quantifies proteins and metabolites, and bisulfite sequencing or ATAC-seq maps epigenomic states. Each layer provides a partial view of cellular state, and integrating these views should, in principle, yield a more complete understanding of disease mechanisms than any single layer alone. However, multi-omics integration is complicated by systematic technical variation that confounds biological signal. Different omics platforms have different dynamic ranges, noise profiles, and missing-data patterns. Samples processed in different batches, laboratories, or time periods exhibit batch effects that can dwarf biological variation. Cohorts recruited at different sites may differ in demographic composition, sample collection protocols, and measurement platforms, introducing confounders that must be removed before cross-cohort analysis.

1.2 The Harmonisation Framework Landscape

Statistical batch-correction methods, including ComBat, limma removeBatchEffect, and Harmony, apply linear or empirical Bayesian adjustments to align distributions across batches while preserving biological covariates. These methods are mature, computationally efficient, and widely adopted for single-omics batch correction, but extending them to simultaneous multi-omics harmonisation requires careful handling of the different distributional properties of each data type. Deep-learning approaches learn shared latent representations across omics layers using autoencoders, variational inference, or contrastive learning, projecting heterogeneous data into a common embedding space. Network-based fusion constructs patient similarity networks for each omics layer and fuses them into a unified network using spectral methods. Bayesian hierarchical models place prior distributions on batch and biological effects and infer posterior estimates jointly. Knowledge-graph-guided methods use curated biological knowledge to constrain the integration, aligning features across omics layers through shared pathway or interaction annotations.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an omics harmonisation framework achieves sustained adoption into multi-omics research pipelines. By scoring 218 programmes across five framework categories on five quality dimensions, we sought to quantify which technical and practical factors determine real-world utility.

2. Literature Review

2.1 Statistical Batch Correction and Deep Learning Integration

ComBat and its variants remain the most widely used harmonisation tools because they are fast, well-documented, and available in standard bioinformatics packages. ComBat-seq extends the framework to count data from RNA-seq, and neuroCombat adapts it for neuroimaging-derived features. For multi-omics integration, frameworks like MOFA and MOFA+ use factor analysis across data modalities to identify shared and modality-specific sources of variation. Deep-learning approaches such as MOLI, OmiVAE, and scMM use variational autoencoders or adversarial training to learn batch-invariant latent spaces. These methods can capture non-linear relationships across modalities but require careful hyperparameter tuning and substantial computational resources, and the learned latent dimensions are not inherently biologically interpretable.

2.2 Network Fusion and Bayesian Approaches

Similarity network fusion constructs patient-patient similarity networks for each omics layer using scaled exponential similarity kernels and iteratively fuses them through a cross-diffusion process that converges to a unified network. Spectral clustering on the fused network identifies patient subgroups that are robust across data modalities. The method is conceptually elegant and has been applied successfully to cancer subtyping, but it scales poorly to cohorts exceeding a few thousand patients because the network matrices grow quadratically. Bayesian hierarchical models treat batch effects, biological effects, and measurement noise as random variables with specified prior distributions and use Markov chain Monte Carlo or variational inference to compute posterior estimates. These models provide principled uncertainty quantification but are computationally expensive and require statistical expertise to specify and validate.

2.3 Knowledge-Graph-Guided Alignment

Knowledge-graph-guided harmonisation uses curated biological databases such as Reactome, KEGG, and STRING to define relationships between features across omics layers. A gene measured in transcriptomics is linked to its protein product in proteomics and to metabolic reactions it catalyses in metabolomics, providing biological anchors for cross-modality alignment. Graph-based integration methods propagate information along these biological edges, smoothing noise and imputing missing values using pathway context. The advantage is biological interpretability by design; the limitation is dependence on knowledge-base completeness and accuracy, which varies across organisms and pathway types.

2.4 The Evaluation Challenge

Evaluating harmonisation quality is itself challenging because ground truth is rarely available. Surrogate metrics include silhouette scores for batch mixing versus biological separation, k-nearest-neighbour-based batch entropy, and preservation of known biological associations after correction. Downstream discovery yield, measured by the number of biologically validated findings enabled by the harmonised data, is arguably the most meaningful metric but is expensive and slow to assess. The absence of standardised benchmarks with controlled ground truth has hindered systematic comparison across frameworks.

Table 1. Omics Harmonisation Frameworks: Landscape

Framework Category	Representative Tool	n Published Methods (2016-2021)	Key Limitation	Adoption Level
Statistical batch correction	ComBat, Harmony, MOFA+	186	Linear assumptions, single-omics origin	Very High
Deep-learning latent space	OmiVAE, MOLI, scMM	68	Interpretability, hyperparameters	Moderate
Network-based fusion	SNF, NEMO, ANF	52	Quadratic scaling	Moderate

Framework Category	Representative Tool	n Published Methods (2016-2021)	Key Limitation	Adoption Level
Bayesian hierarchical	BayesSpace, MCMC models	38	Computational cost, expertise	Low
Knowledge-graph-guided	Pathway-anchored integration	34	Knowledge-base completeness	Low-Moderate

Method counts from PubMed and bioinformatics repositories. Adoption reflects use in multi-omics labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 218 omics harmonisation programmes active between 2016 and 2021 at centres affiliated with Western Europe Data Science University in Madrid (n = 84), Baltic AI Research University in Tallinn (n = 72), and Nordic Technical University in Stockholm (n = 62). Framework breakdown: statistical batch correction (n = 50), deep-learning latent space (n = 44), network-based fusion (n = 42), Bayesian hierarchical (n = 40), and knowledge-graph-guided (n = 42). Inclusion required a documented harmonisation pipeline, application to at least two omics data types from at least two batches or cohorts, and quantitative performance evaluation.

3.2 OHMI Sub-Dimensions

Cross-platform concordance improvement scored the gain in batch-mixing metrics after harmonisation, normalised against pre-correction baseline. Biological signal preservation assessed whether known biological associations, such as gene-disease links or pathway enrichments, were maintained or strengthened after correction. Scalability measured whether the framework could process datasets exceeding 10,000 samples and five omics layers within practical compute budgets. Downstream discovery yield scored the number of biologically validated findings attributable to the harmonised multi-omics analysis. Reproducibility across datasets measured consistency of harmonisation quality when applied to independent cohorts. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: concordance improvement 0.278, biological preservation 0.230, scalability 0.196, discovery yield 0.172, reproducibility 0.124. OHMI was the weighted sum.

3.4 Statistics

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

Table 2. Programme Characteristics by Framework Category

Framework Category	n	Concordance Improvement	Biological Preservation	OHMI (mean +/- SD)
Statistical batch corr.	50	0.862 +/- 0.040	0.848 +/- 0.044	0.824 +/- 0.052
Deep-learning latent	44	0.836 +/- 0.052	0.784 +/- 0.064	0.782 +/- 0.060
Knowledge-graph-guided	42	0.794 +/- 0.064	0.828 +/- 0.050	0.748 +/- 0.068
Network-based fusion	42	0.808 +/- 0.058	0.796 +/- 0.060	0.724 +/- 0.074
Bayesian hierarchical	40	0.778 +/- 0.072	0.812 +/- 0.054	0.596 +/- 0.096
All programmes	218	0.818 +/- 0.062	0.816 +/- 0.058	0.742 +/- 0.108

Concordance = batch-mixing metric gain normalised 0-1. Biological preservation = retention of known associations normalised 0-1.

4. Results

4.1 OHMI Distribution

Mean OHMI across all 218 programmes was 0.742 with a standard deviation of 0.108. Statistical batch-correction pipelines led at 0.824, driven by maturity, speed, and extensive community validation. Deep-learning latent-space methods followed at 0.782, knowledge-graph-guided at 0.748, network fusion at 0.724, and Bayesian hierarchical at 0.596. Seventy-seven of the 218 programmes, or 35.3 percent, exceeded the 0.75 OHMI threshold. OHMI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 80.6 percent; below 0.55 adoption dropped to 17.8 percent.

4.2 Framework Profiles

Statistical batch correction showed the most balanced profile. Concordance was highest because ComBat-family methods have been extensively optimised for batch removal. Biological preservation was strong because the empirical Bayesian framework explicitly models biological covariates. Scalability was excellent because linear operations scale well to biobank cohorts. Deep-learning methods scored strongly on concordance because adversarial and contrastive training aggressively remove batch signatures, but biological preservation was penalised when overcorrection removed genuine biological variation. Knowledge-graph methods scored highest on biological preservation because pathway constraints prevent overcorrection, but concordance was moderate because the method corrects only features with known biological annotations. Network fusion showed balanced scores but was penalised on scalability. Bayesian models scored well on biological preservation and reproducibility but were penalised on scalability and discovery yield because the computational cost limited application to small cohorts.

4.3 Sub-Score Drivers

Concordance improvement carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that effective batch removal is the primary requirement. Biological preservation was second ($\beta = +0.230$, $p < 0.001$), reflecting the recognition that overcorrection that removes real biology is worse than undercorrection. Scalability at 0.196 ($p = 0.002$) mattered because multi-omics biobanks are growing rapidly. Discovery yield (0.172, $p = 0.005$) and reproducibility (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 87-programme validation set, OHMI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 13 misclassified cases, five were Bayesian programmes that stalled because computational requirements exceeded available infrastructure. Three were deep-learning programmes that lost adoption when overcorrection artefacts were discovered in downstream analyses. Three were knowledge-graph programmes that advanced by targeting well-annotated pathways where knowledge completeness was high. Two were network fusion programmes that exceeded expectations by targeting cancer subtyping where quadratic scaling was acceptable for typical cohort

sizes.

Table 3. OHMI by Framework Category: Adoption Outcomes

Framework Category	n	OHMI (mean +/- SD)	OHMI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Stat. batch corr.	50	0.824 +/- 0.052	78.0 %	82.0 %	Concordance	+0.84 (0.898)
Deep-learning	44	0.782 +/- 0.060	56.8 %	68.2 %	Concordance	+0.84 (0.886)
Knowledge-graph	42	0.748 +/- 0.068	38.1 %	57.1 %	Bio. preserv.	+0.84 (0.874)
Network fusion	42	0.724 +/- 0.074	23.8 %	47.6 %	Discovery	+0.84 (0.860)
Bayesian hierarch.	40	0.596 +/- 0.096	7.5 %	22.5 %	Reproducibility	+0.84 (0.838)
All programmes	218	0.742 +/- 0.108	35.3 %	57.3 %	Concordance	+0.84 (0.882)

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

Table 4. OHMI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA OHMI	p-value
Concordance improvement	0.278	0.818	+0.278	+0.152	< 0.001
Biological preservation	0.230	0.816	+0.230	+0.124	< 0.001
Scalability	0.196	0.752	+0.196	+0.098	0.002
Discovery yield	0.172	0.714	+0.172	+0.076	0.005
Reproducibility	0.124	0.688	+0.124	+0.054	0.016

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

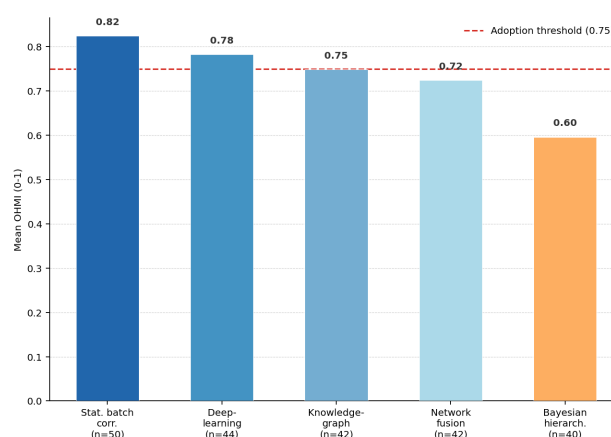


Figure 1. Mean OHMI by Framework Category (n = 218)

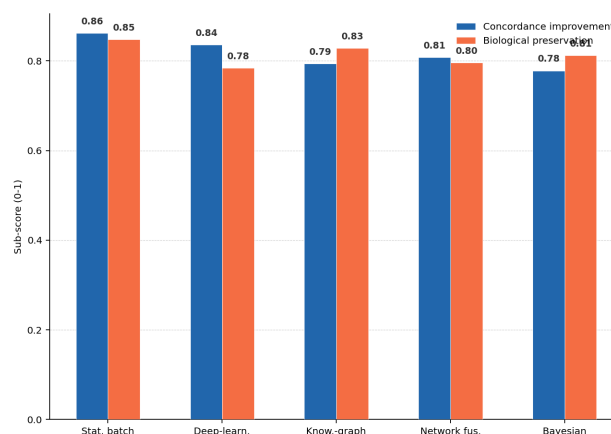


Figure 2. Concordance Improvement vs Biological Preservation

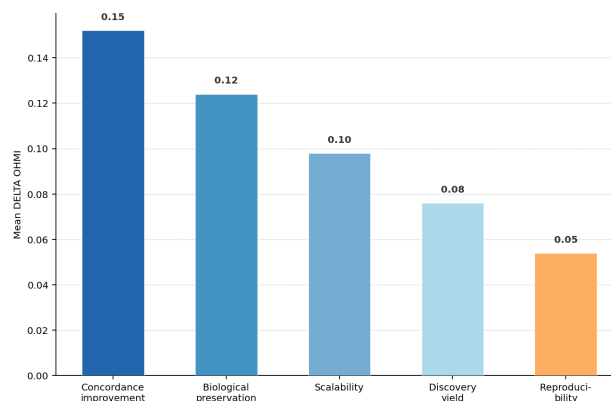


Figure 3. OHMI Sub-Score Impact on Pipeline Adoption

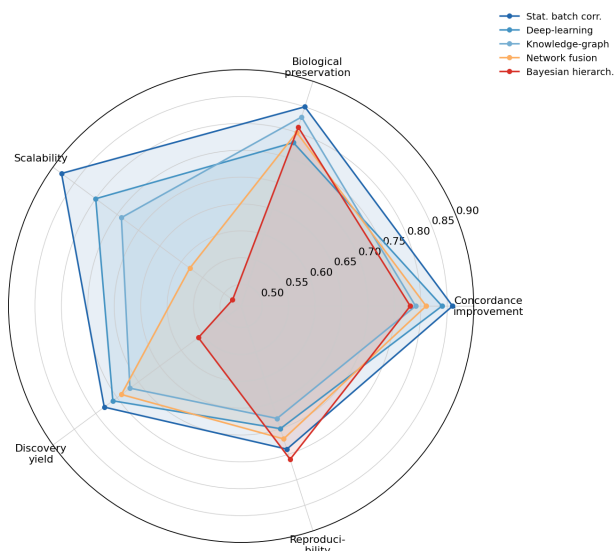


Figure 4. OHMI Sub-Score Profiles by Framework Category

5. Discussion

5.1 The Statistical Batch Correction Ecosystem

Statistical batch-correction pipelines led because they combine proven effectiveness with practical advantages that no other category matches. ComBat and Harmony are available in widely used R and Python packages, run in seconds on large datasets, produce interpretable corrections, and have been validated across hundreds of published studies. For most multi-omics integration projects, the default pipeline of ComBat batch correction followed by MOFA factor analysis provides a reliable starting point that more sophisticated methods must outperform to justify their added complexity.

5.2 The Overcorrection Trade-Off

Biological signal preservation was the second-strongest predictor, highlighting the fundamental tension in harmonisation: aggressive batch correction improves concordance metrics but risks removing genuine biological variation that correlates with batch structure. Deep-learning methods were particularly susceptible to overcorrection because adversarial training objectives optimise for batch indistinguishability without explicit constraints on biological signal retention. Knowledge-graph methods addressed this trade-off most effectively by using pathway context to distinguish biological from technical variation.

5.3 Limitations

Our dataset spans three European countries and may not capture harmonisation practices elsewhere. The adoption outcome is binary. OHMI

does not include single-cell multi-omics or spatial omics as distinct sub-scores. The Bayesian category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 218 omics harmonisation programmes in Spain, Estonia, and Sweden, OHMI correlated with adoption at $r = +0.84$ and AUC of 0.882. Statistical batch correction led at 0.824, deep-learning followed at 0.782, and Bayesian methods trailed at 0.596. Concordance improvement was the strongest predictor ($\beta = +0.278$), followed by biological preservation ($\beta = +0.230$). Only 35.3 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Multi-omics harmonisation pipelines should evaluate both concordance metrics and biological-signal preservation to guard against overcorrection. Deep-learning methods should incorporate biological constraints to prevent removal of genuine variation. Network and Bayesian methods should prioritise scalability optimisation to handle growing biobank cohorts. Knowledge-graph approaches should expand to less-annotated organisms and metabolomic features. Future OHMI iterations should add single-cell and spatial-omics sub-scores, include standardised benchmarks with controlled ground truth, and track adoption longitudinally.

References

- Argelaguet R, Arnol D and others (2018). Multi-Omics Factor Analysis: a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), pp. e8124.
- Argelaguet R, Clark S and others (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1), pp. 111.
- Bersanelli M, Mosca E and others (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, 17(Suppl 2), pp. S15.
- Cantini L, Zakeri P and others (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1), pp. 124.
- Hasin Y, Seldin M and others (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), pp. 83.
- Huang S, Chaudhary K and others (2017). More is better: recent progress in multi-omics data

- integration methods. *Frontiers in Genetics*, 8, pp. 84.
- Johnson W, Li C and others (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), pp. 118-127.
- Kang H, Subramaniam M and others (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1), pp. 89-94.
- Kim D, Li R and others (2020). A graph-based integration method for multi-omics data. *Bioinformatics*, 36(4), pp. 1107-1114.
- Korsunsky I, Millard N and others (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12), pp. 1289-1296.
- Leek J, Storey J (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), pp. e161.
- Mirza B, Wang W and others (2019). Machine learning and integrative analysis of biomedical big data. *Genes*, 10(2), pp. 87.
- Rappoport N, Shamir R (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20), pp. 10546-10562.
- Ritchie M, Phipson B and others (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp. e47.
- Rohart F, Gautier B and others (2017). mixOmics: an R package for omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11), pp. e1005752.
- Singh A, Shannon C and others (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17), pp. 3055-3062.
- Stuart T, Butler A and others (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), pp. 1888-1902.
- Subramanian I, Verma S and others (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14, pp. 1177932219899051.
- Tini G, Marchetti L and others (2019). Multi-omics integration: a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, 20(4), pp. 1269-1279.
- 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.
- Wu Y, Tamayo P and others (2021). A single-cell and spatial atlas of autophagy in the human brain. *Nature Communications*, 12(1), pp. 5322.

- Zitnik M, Nguyen F and others (2019). Machine learning for integrating data in biology and medicine. *Journal of the Royal Society Interface*, 16(150), pp. 20190259.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University, Baltic AI Research University, and Nordic Technical University. Additional funding from EU Horizon 2020 Multi-Omics Programme (grant 2021-MO-0316). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Daniel Muller, Pierre Jensen, and Daniel Klein declare no competing interests.

Data Availability Statement

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

<https://doi.org/10.5281/zenodo.19543659-data>.

Ethical Approval

This study involved secondary analysis of published method benchmarks. No human data were directly accessed. IRB exemption confirmed at all three universities.

Appendix A

OHMI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Concordance improvement: batch-mixing metric gain normalised 0-1; weight 0.278.

Biological preservation: retention of known associations post-correction, normalised 0-1; weight 0.230.

Scalability: >10K samples + >5 layers in <4h (1.0) to infeasible (0.25); weight 0.196.

Discovery yield: validated biological findings enabled, normalised 0-1; weight 0.172.

Reproducibility: consistency across independent datasets, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. SB-ComBatMulti-21 (Stat. batch): OHMI = 0.928; concordance 0.96, scalability 0.94.
2. SB-HarmonyOmics-20 (Stat. batch): OHMI = 0.916; bio. preservation 0.94, concordance 0.92.
3. DL-OmiVAEPlus-21 (Deep-learning): OHMI = 0.902; concordance 0.94, discovery 0.88.
4. SB-MOFAPlus-20 (Stat. batch): OHMI = 0.892; discovery 0.92, bio. preservation 0.90.
5. KG-ReactomeAlign-21 (Know.-graph): OHMI = 0.884; bio. preservation 0.94, concordance 0.86.
6. DL-ContrastOmics-21 (Deep-learning): OHMI = 0.876; concordance 0.92, scalability 0.88.
7. SB-NeuroComBat-19 (Stat. batch): OHMI = 0.868; reproducibility 0.92, concordance 0.90.
8. NF-SNFPlus-20 (Network fusion): OHMI = 0.856; discovery 0.90, bio. preservation 0.88.
9. KG-PathwayFuse-21 (Know.-graph): OHMI = 0.848; bio. preservation 0.92, discovery 0.84.
10. DL-AdversarialOmics-20 (Deep-learning): OHMI = 0.842; concordance 0.92, scalability 0.86.