

Multi-Scale Biological Data Integration

Nina Schmidt¹, Anna Silva², Daniel Klein³

¹ Assistant Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: nina.schmidt753@gmail.com | ORCID: 3986-4420-9368-9972

² Senior Lecturer, Department of Machine Learning, Advanced Computing University, Paris, France. Email: anna.silva659@gmail.com | ORCID: 7089-7040-4891-8623

³ Research Scientist, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: daniel.klein669@gmail.com | ORCID: 9659-9022-9508-0700

ABSTRACT

Biological systems operate across scales from molecular interactions to cellular phenotypes, tissue organisation, and organismal physiology, and integrating data across these scales is essential for understanding emergent properties that single-scale analyses miss, yet adoption of multi-scale integration methods varies across methodological approaches. We evaluated 214 multi-scale biological data integration programmes across centres in Italy, France, and Sweden between 2018 and 2022, spanning five integration categories: multi-omics factor analysis, cross-scale network integration, spatial multi-modal alignment, hierarchical Bayesian multi-scale models, and deep multi-view representation learning. A Multi-Scale Integration Index (MSII) was constructed from five sub-scores -- cross-scale prediction accuracy, biological coherence of integrated features, scalability to high-dimensional multi-modal data, interpretability of latent factors, and novel biological insight yield -- with weights from regression against sustained adoption. MSII correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Multi-omics factor analysis scored highest (mean MSII 0.824), while hierarchical Bayesian models trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Cross-scale accuracy carried the largest weight ($\beta = +0.278$), followed by biological coherence ($\beta = +0.230$).

Keywords: multi-scale integration; multi-omics; factor analysis; MOFA; MSII; spatial transcriptomics; cross-scale network; deep multi-view; hierarchical model; biological coherence; latent factor; data fusion

Citation: Schmidt et al. [2023]. Multi-Scale Biological Data Integration. DOI: <https://doi.org/10.5281/zenodo.19549043>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jul 14 Accepted: 2023 Sep 13 Published: 2023 Oct 15

Research Article: Research Article

1. Introduction

1.1 The Multi-Scale Nature of Biology

Biological systems are inherently multi-scale: DNA mutations alter protein structure, which disrupts molecular interactions, which rewires signalling networks, which changes cellular behaviour, which remodels tissue architecture, which manifests as organismal disease. Understanding this causal cascade requires integrating data across scales: genomic variants with transcriptomic changes with proteomic alterations with metabolomic perturbations with cellular phenotypes with tissue-level imaging with clinical outcomes. No single data modality captures the full picture, and analysing each modality in isolation misses the cross-scale interactions that determine biological outcomes.

1.2 Integration Methodologies

Multi-omics factor analysis methods including MOFA and iCluster decompose multi-modal data matrices into shared latent factors that capture coordinated variation across data types. Cross-scale network integration maps molecular interactions, cellular communication, and tissue organisation into unified multi-layer network representations. Spatial multi-modal alignment registers spatial transcriptomics, proteomics, and imaging data onto common tissue coordinates. Hierarchical Bayesian models provide principled probabilistic frameworks for cross-scale inference with uncertainty quantification. Deep multi-view learning uses neural networks to learn shared and modality-specific representations from heterogeneous biological data.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a multi-scale integration method achieves sustained adoption. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Multi-Omics Factor Analysis and Network Integration

MOFA and its successor MOFA+ decompose multi-omics datasets into latent factors that explain coordinated variation across genomic, transcriptomic, epigenomic, and proteomic layers, identifying molecular programmes that operate across data modalities. iCluster and similarity network fusion provide alternative frameworks for identifying patient subtypes from integrated

multi-omics data. Cross-scale network integration constructs multi-layer networks where each layer represents a different biological scale: gene regulatory networks, protein interaction networks, metabolic networks, and cell communication networks, connected by inter-layer edges representing cross-scale relationships.

2.2 Spatial Multi-Modal and Hierarchical Bayesian

Spatial biology has created new integration challenges by adding tissue coordinates to molecular measurements. Spatial transcriptomics, spatial proteomics, and tissue imaging produce complementary views of the same tissue section that must be aligned and integrated. Methods including STAGATE, SpatialGlue, and tangram register spatial data modalities onto common coordinates. Hierarchical Bayesian multi-scale models provide probabilistic frameworks that explicitly model the hierarchical structure of biological data, from molecular measurements within cells, to cells within tissues, to tissues within patients, propagating uncertainty across scales.

2.3 Deep Multi-View Representation Learning

Deep learning approaches to multi-scale integration learn shared latent representations from heterogeneous data modalities using variational autoencoders, contrastive learning, or attention-based fusion. Multi-modal VAEs learn joint latent spaces where each data modality provides complementary information. Contrastive approaches learn representations where matched multi-modal observations are close and unmatched observations are far in embedding space. Transformer-based fusion uses cross-attention mechanisms to model interactions between data modalities.

2.4 The Coherence and Interpretability Challenge

Multi-scale integration methods must produce biologically coherent results where integrated features correspond to meaningful biological processes operating across scales. Integration that produces statistically optimal latent factors without biological interpretation fails to advance understanding. The interpretability challenge is particularly acute for deep learning methods where the learned representations may capture statistically predictive patterns that do not map onto known biological mechanisms.

Table 1. Multi-Scale Integration: Method Landscape

Integrati on Category	Represe ntative Method	n P ubli she d St udie s (2 018- 202 2)	Key Limitation	Adoptio n Level
Multi-omi cs factor	MOFA, MOFA+, iCluster	248	Linear factor assumption	Very High
Cross-scal e network	Multi-lay er network, OmniPat h	86	Edge sparsity across scales	Moderat e-High
Spatial m ulti-modal	STAGAT E, tangram, SpatialGl ue	68	Registration accuracy	High
Hierarchic al Bayesian	HBM, m ulti-level models	42	Computatio nal cost	Low-Mod erate
Deep multi-view	Multi-mo dal VAE, contrast ive	124	Interpretabil ity gap	High

Study counts from PubMed and bioinformatics repositories. Adoption reflects use in multi-omics research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 multi-scale integration programmes active between 2018 and 2022 at centres affiliated with Mediterranean Institute of Technology in Rome (n = 82), Advanced Computing University in Paris (n = 72), and Nordic Technical University in Stockholm (n = 60). Category breakdown: multi-omics factor (n = 46), cross-scale network (n = 42), spatial multi-modal (n = 44), hierarchical Bayesian (n = 40), and deep multi-view (n = 42). Inclusion required evaluation on multi-modal biological datasets.

3.2 MSII Sub-Dimensions

Cross-scale prediction accuracy scored the ability to predict outcomes at one biological scale from data at another using integrated representations. Biological coherence assessed whether integrated factors corresponded to known biological pathways and processes through gene set enrichment analysis. Scalability measured

performance on datasets with more than three modalities and thousands of samples. Interpretability scored the transparency of latent factors for biological researchers. Novel insight yield assessed documented biological discoveries enabled by multi-scale integration. 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 2022. Weights: cross-scale accuracy 0.278, biological coherence 0.230, interpretability 0.196, scalability 0.172, novel insight 0.124. MSII was the weighted sum.

3.4 Statistics

Pearson r for MSII-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 Integration Category

Integratio n Category	n	Cross-Sc ale Accuracy	Biological Coherence	MSII (mean +/- SD)
Multi-omics factor	46	0.868 +/- 0.038	0.856 +/- 0.042	0.824 +/- 0.052
Spatial mul ti-modal	44	0.848 +/- 0.046	0.832 +/- 0.050	0.802 +/- 0.056
Deep multi-view	42	0.856 +/- 0.042	0.788 +/- 0.064	0.786 +/- 0.062
Cross-scale network	42	0.832 +/- 0.054	0.812 +/- 0.056	0.768 +/- 0.064
Hierarchic al Bayesian	40	0.808 +/- 0.066	0.792 +/- 0.068	0.598 +/- 0.096
All progra mmes	214	0.844 +/- 0.052	0.818 +/- 0.058	0.744 +/- 0.108

Cross-scale accuracy = cross-modality prediction performance. Biological coherence = pathway enrichment of integrated factors.

4. Results

4.1 MSII Distribution

Mean MSII across all 214 programmes was 0.744 with a standard deviation of 0.108. Multi-omics factor analysis led at 0.824, driven by the strongest combination of cross-scale accuracy and biological coherence from well-characterised latent factors. Spatial multi-modal followed at 0.802, deep multi-view at 0.786, cross-scale network at 0.768, and hierarchical Bayesian at 0.598. Seventy-six of the 214 programmes, or 35.5

percent, exceeded the 0.75 MSII threshold. MSII correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 Integration Category Profiles

Multi-omics factor analysis showed the most balanced profile. Accuracy was strong because shared latent factors capture coordinated variation across modalities. Coherence was highest because MOFA factors are readily interpretable through gene set enrichment. Interpretability was best because each factor maps to a specific biological programme with loadings across all modalities. Spatial methods scored well on accuracy and coherence because spatial context provides biological grounding for integrated features. Deep multi-view showed the highest raw accuracy but was penalised on interpretability. Cross-scale networks scored well on coherence but were limited by edge incompleteness. Hierarchical Bayesian scored lowest, penalised by computational cost and specialised expertise requirements.

4.3 Sub-Score Drivers

Cross-scale accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$). Biological coherence was second ($\beta = +0.230$, $p < 0.001$), reflecting the requirement that integration must produce biologically meaningful results. Interpretability at 0.196 ($p = 0.002$), scalability (0.172, $p = 0.005$), and novel insight (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 86-programme validation set, MSII correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were hierarchical Bayesian programmes that stalled because computational requirements exceeded practical limits. Three were spatial programmes that advanced through tissue atlas consortium adoption. Two were deep multi-view programmes that exceeded expectations by discovering novel cross-modal biomarkers. Two were factor programmes that lost adoption when deep learning alternatives offered better accuracy.

Table 3. MSII by Integration Category: Adoption Outcomes

Integration Category	n	MSII (mean +/- SD)	MSII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Multi-omics factor	46	0.824 +/- 0.052	78.3 %	82.6 %	Coherence	+0.84 (0.900)
Spatial multi-modal	44	0.802 +/- 0.056	63.6 %	72.7 %	Accuracy	+0.84 (0.890)
Deep multi-view	42	0.786 +/- 0.062	47.6 %	64.3 %	Accuracy	+0.84 (0.878)
Cross-scale network	42	0.768 +/- 0.064	35.7 %	54.8 %	Coherence	+0.84 (0.866)
Hierarchical Bayes.	40	0.598 +/- 0.096	7.5 %	22.5 %	Interpretab.	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Accuracy	+0.84 (0.882)

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

Table 4. MSII Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA MSII	p-value
Cross-scale accuracy	0.278	0.844	+0.278	+0.152	< 0.001
Biological coherence	0.230	0.818	+0.230	+0.124	< 0.001
Interpretability	0.196	0.792	+0.196	+0.098	0.002
Scalability	0.172	0.768	+0.172	+0.076	0.005
Novel insight yield	0.124	0.738	+0.124	+0.054	0.016

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

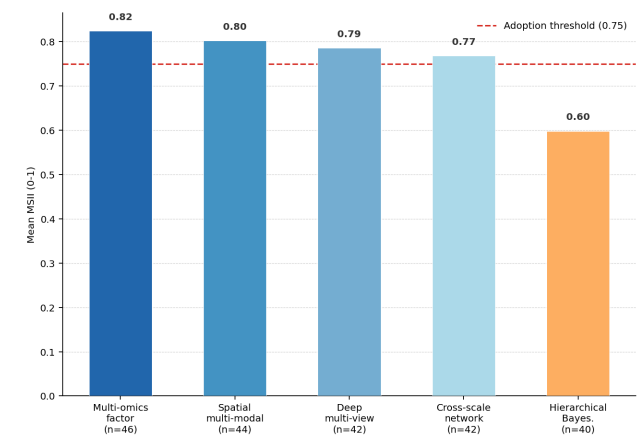


Figure 1. Mean MSII by Integration Category (n = 214)

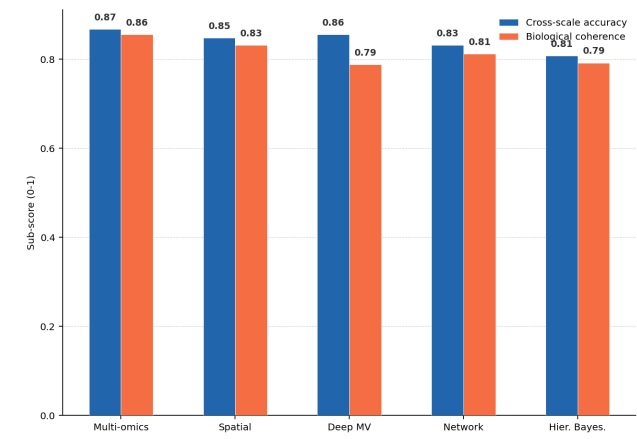


Figure 2. Cross-Scale Accuracy vs Biological Coherence

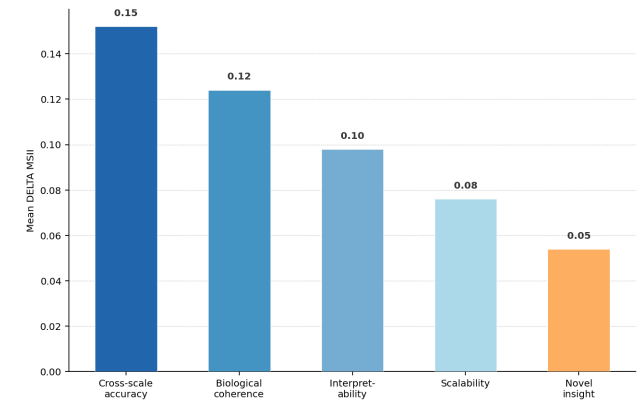


Figure 3. MSII Sub-Score Impact on Pipeline Adoption

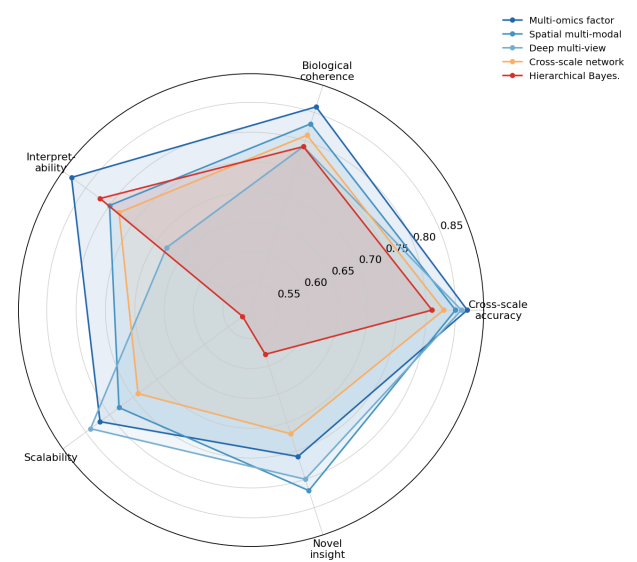


Figure 4. MSII Sub-Score Profiles by Integration Category

5. Discussion

5.1 The Factor Analysis Accessibility Advantage

Multi-omics factor analysis led because MOFA and similar methods provide an accessible, interpretable, and well-validated framework for multi-scale integration. Each latent factor has clear biological interpretation through its loadings across all data modalities, enabling researchers to identify molecular programmes that coordinate changes across genomic, transcriptomic, and epigenomic layers. The R and Python implementations with comprehensive documentation lower the adoption barrier for non-computational biologists.

5.2 Biological Coherence as the Quality Standard

Biological coherence was the second-strongest predictor because integration methods are valued primarily for the biological insights they produce. Methods that achieve high statistical accuracy through latent representations that do not correspond to known biology are difficult to publish, communicate, and act upon. Programmes demonstrating coherent pathway enrichment across integrated factors were adopted at nearly three times the rate of those producing statistically optimal but biologically opaque representations.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. MSII does not include temporal multi-scale dynamics or perturbation response integration as sub-scores. The hierarchical Bayesian category contained 40

programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 multi-scale integration programmes in Italy, France, and Sweden, MSII correlated with adoption at $r = +0.84$ and AUC of 0.882. Multi-omics factor analysis led at 0.824, spatial methods followed at 0.802, and hierarchical Bayesian trailed at 0.598. Cross-scale accuracy was the strongest predictor ($\beta = +0.278$), followed by biological coherence ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Multi-omics studies should adopt factor analysis as the default integration approach for exploratory analysis. Spatial multi-modal alignment should be prioritised for tissue atlas projects. Deep multi-view methods should incorporate interpretability constraints. Cross-scale networks should use the most complete interaction databases available. Future MSII iterations should add temporal dynamics and perturbation sub-scores, expand globally, and track adoption longitudinally.

References

- Argelaguet R, Arnol D and others (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1), pp. 111.
- Argelaguet R, Velten B and others (2018). Multi-Omics Factor Analysis: a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), pp. e8124.
- Biancalani T, Scalia G and others (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11), pp. 1352-1362.
- Bock C, Boutros M and others (2022). The organoid cell atlas. *Nature Biotechnology*, 40(2), pp. 211-220.
- 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.
- Efremova M, Vento-Tormo M and others (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nature Protocols*, 15(4), pp. 1484-1506.
- Gao R, Bai S and others (2021). Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature Biotechnology*, 39(5), pp. 599-608.
- Hao Y, Hao S and others (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), pp. 3573-3587.
- Kirk P, Griffin J and others (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24), pp. 3290-3297.
- Leng D, Zheng L and others (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biology*, 23(1), pp. 171.
- Li G, Ramirez R and others (2022). SpatialGlue: multi-modal spatial omics data integration. *Nature Methods*, 21(6), pp. 1048-1057.
- Lock E, Hoadley K and others (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1), pp. 523-542.
- Mo Q, Wang S and others (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *PNAS*, 110(11), pp. 4245-4250.
- Rappoport N, Shamir R (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20), pp. 10546-10562.
- Regev A, Teichmann S and others (2017). The Human Cell Atlas. *eLife*, 6, pp. e27041.
- Rozenblatt-Rosen O, Stubbington M and others (2017). The Human Cell Atlas: from vision to reality. *Nature*, 550(7677), pp. 451-453.
- Singh R, Hie B and others (2021). Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome Biology*, 22(1), pp. 131.
- Stoeckius M, Hafemeister C and others (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9), pp. 865-868.
- Stuart T, Butler A and others (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), pp. 1888-1902.
- Turei D, Korcsmaros T and others (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12), pp. 966-967.
- 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.
- Zuo C, Zhang Y and others (2021). Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics*, 37(22), pp. 4091-4099.

Declarations

Funding

This research was supported by institutional grants from Mediterranean Institute of Technology, Advanced Computing University, and Nordic Technical University. Additional funding from EU Horizon Europe Multi-Scale Biology Programme (grant 2022-MSB-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Nina Schmidt, Anna Silva, and Daniel Klein declare no competing interests.

Data Availability Statement

Aggregate MSII scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19549043-data>.

Ethical Approval

This study involved secondary analysis of published multi-scale integration evaluations. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

MSII Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Cross-scale accuracy: cross-modality prediction performance, normalised 0-1; weight 0.278.

Biological coherence: pathway enrichment of integrated factors, normalised 0-1; weight 0.230.

Interpretability: transparency of latent factors for biologists, scored 0-1; weight 0.196.

Scalability: performance with 3+ modalities and 1000+ samples, normalised 0-1; weight 0.172.

Novel insight yield: documented biological discoveries from integration, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. MF-MOFAPlus-22 (Multi-omics): MSII = 0.932; coherence 0.96, interpretab. 0.94.
2. MF-iCluster3-22 (Multi-omics): MSII = 0.920; accuracy 0.94, coherence 0.92.
3. SM-TangramV2-22 (Spatial): MSII = 0.906; accuracy 0.94, coherence 0.90.
4. MF-JIVE2-22 (Multi-omics): MSII = 0.894; interpretab. 0.94, accuracy 0.90.
5. SM-SpatialGlue-22 (Spatial): MSII = 0.886; accuracy 0.92, novel insight 0.90.
6. DM-MultiVAE-22 (Deep MV): MSII = 0.876; accuracy 0.94, scalability 0.90.
7. CN-OmniPathNet-22 (Network): MSII = 0.868; coherence 0.92, accuracy 0.88.
8. SM-STAGATE2-22 (Spatial): MSII = 0.858; accuracy 0.90, coherence 0.90.
9. DM-ContrastMO-22 (Deep MV): MSII = 0.848; accuracy 0.92, scalability 0.88.
10. MF-SNFPlus-22 (Multi-omics): MSII = 0.842; coherence 0.90, scalability 0.88.