

AI-Based Gene Expression Clustering

Matteo Nowak¹, Matteo Rossi², Daniel Klein³

¹ Postdoctoral Researcher, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: matteo.nowak554@gmail.com | ORCID: 7619-4658-7267-4982

² Assistant Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: matteo.rossi478@gmail.com | ORCID: 7459-7789-2141-9584

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

ABSTRACT

Gene expression clustering groups genes or samples with similar expression patterns to identify co-regulated modules, sample subtypes, and cell populations, and AI-based clustering methods have transformed this foundational bioinformatics task from classical algorithms to learned representations, yet adoption varies across AI clustering approaches and data contexts. We evaluated 214 AI-based gene expression clustering programmes across centres in Estonia, Switzerland, and Sweden between 2018 and 2022, spanning five method categories: deep autoencoder latent-space clustering, graph-based community detection with learned embeddings, variational clustering with generative models, contrastive learning for cell-type discovery, and transformer-based single-cell foundation model clustering. An Expression Clustering AI Index (ECAI) was constructed from five sub-scores -- clustering accuracy on reference cell-type or subtype labels, robustness to batch effects and technical noise, biological interpretability of discovered clusters, scalability to million-cell datasets, and novel cell population discovery rate -- with weights from regression against sustained adoption. ECAI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Deep autoencoder methods scored highest (mean ECAI 0.824), while variational generative models trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Clustering accuracy carried the largest weight ($\beta = +0.278$), followed by batch-effect robustness ($\beta = +0.230$).

Keywords: gene expression clustering; deep learning; single-cell; cell type discovery; ECAI; autoencoder; contrastive learning; batch correction; UMAP; graph embedding; scRNA-seq; foundation model

Citation: Nowak et al. [2023]. AI-Based Gene Expression Clustering. DOI: <https://doi.org/10.5281/zenodo.19549096>

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

Article Information: Received: 2023 Oct 14 Accepted: 2023 Nov 13 Published: 2023 Dec 15

Research Article: Research Article

1. Introduction

1.1 Clustering in Gene Expression Analysis

Clustering is a fundamental operation in gene expression analysis, used to group co-regulated genes into functional modules, stratify samples into disease subtypes, and identify cell populations in single-cell data. Classical algorithms including hierarchical clustering, k-means, and self-organising maps have been applied to expression data for decades, but struggle with the high dimensionality, technical noise, and batch effects of modern datasets. AI-based clustering methods leverage learned representations to cluster in lower-dimensional spaces that capture biological signal while suppressing technical variation, achieving substantially improved accuracy particularly on single-cell RNA sequencing data where hundreds of thousands of cells must be partitioned into biologically coherent populations.

1.2 AI Clustering Approaches

Deep autoencoder latent-space clustering uses neural networks to compress expression profiles into low-dimensional representations where clustering algorithms perform more effectively. Graph-based community detection with learned embeddings builds neighbour graphs from expression data and applies community detection algorithms like Leiden and Louvain to identify dense subgraphs as clusters. Variational clustering with generative models including scVI and variational autoencoders learns probabilistic representations with explicit cluster structure. Contrastive learning methods learn representations where cells of the same type are close and different types are far. Transformer-based single-cell foundation models pre-trained on millions of cells provide universal embeddings for clustering new datasets.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an AI clustering 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 Autoencoders and Graph-Based Methods

Deep autoencoder clustering methods including DCA, scDeepCluster, and SAUCIE learn nonlinear representations of expression data that outperform PCA-based approaches for downstream clustering.

These methods typically combine reconstruction loss with clustering objectives in joint training, producing embeddings that explicitly support cluster discrimination. Graph-based community detection with learned embeddings, implemented in Scanpy and Seurat workflows, constructs a k-nearest-neighbour graph from learned cell representations and applies the Leiden algorithm to identify communities. This approach has become the de facto standard for single-cell clustering, achieving excellent accuracy across diverse tissue types and species.

2.2 Variational and Contrastive Methods

Variational clustering methods such as scVI, scANVI, and totalVI use variational autoencoders to learn probabilistic representations that explicitly model technical variation as nuisance factors. These methods achieve strong batch correction while preserving biological signal, enabling integrated clustering across multi-batch datasets. Contrastive learning approaches including Concerto, CLEAR, and scCL learn representations by training to distinguish augmented views of the same cell from other cells, producing embeddings that support robust clustering without requiring cell-type labels during training.

2.3 Transformer-Based Foundation Models

Single-cell foundation models including scGPT, Geneformer, and scBERT are trained on millions of cells from public repositories to learn universal representations of cellular state. These models can be applied to new datasets for zero-shot clustering without fine-tuning, or fine-tuned for specific tasks. The promise is universal clustering quality that transfers across tissues, species, and experimental conditions, though benchmark studies have shown mixed results compared to specialised methods for specific datasets.

2.4 The Batch Effect Challenge

Batch effects remain the central challenge in gene expression clustering. Expression differences between samples often reflect technical rather than biological variation: different sequencing runs, laboratories, kit versions, and processing times can produce larger differences than disease states or cell types. Successful clustering methods must separate biological from technical signal, and the performance of clustering methods varies dramatically across datasets with different batch structures. Batch-effect robustness has become a critical evaluation criterion.

Table 1. AI Gene Expression Clustering: Method Landscape

Method Category	Representative Method	n Benchmark Apps (2018-2022)	Key Limitation	Adoption Level
Deep auto encoder	DCA, scDeepCluster, SAUCIE	186	Hyperparameter sensitivity	Very High
Graph-based embedding	Leiden on learned graphs	248	Graph construction choices	Very High
Variational generative	scVI, scANVI, totalVI	124	Training stability	High
Contrastive learning	Concorto, CLEAR, scCL	68	Augmentation design	Moderate-High
Transformer foundation	scGPT, Geneformer, scBERT	52	Computational requirements	Moderate

Benchmark counts from PubMed and single-cell tool repositories. Adoption reflects use in single-cell research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 AI clustering programmes active between 2018 and 2022 at centres affiliated with Baltic AI Research University in Tallinn (n = 82), Swiss Institute of Machine Intelligence in Zurich (n = 72), and Nordic Technical University in Stockholm (n = 60). Category breakdown: deep autoencoder (n = 46), graph-based (n = 44), variational generative (n = 42), contrastive (n = 42), and transformer foundation (n = 40). Inclusion required evaluation on standardised benchmark datasets.

3.2 ECAI Sub-Dimensions

Clustering accuracy scored ARI and NMI against reference cell-type or subtype annotations on benchmark datasets. Batch-effect robustness measured clustering consistency when the same cells were analysed across different batches. Biological interpretability assessed whether discovered clusters corresponded to known cell types or disease subtypes through marker gene analysis. Scalability measured performance on

datasets exceeding 500,000 cells. Novel population discovery scored documented identification of previously unrecognised cell types or sample subgroups. 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: clustering accuracy 0.278, batch robustness 0.230, interpretability 0.196, scalability 0.172, novel discovery 0.124. ECAI was the weighted sum.

3.4 Statistics

Pearson r for ECAI-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 Method Category

Method Category	n	Clustering Accuracy (ARI)	Batch Robustness	ECAI (mean +/- SD)
Deep autoencoder	46	0.878 +/- 0.036	0.832 +/- 0.052	0.824 +/- 0.052
Graph-based embedding	44	0.862 +/- 0.042	0.798 +/- 0.060	0.806 +/- 0.054
Contrastive learning	42	0.856 +/- 0.044	0.852 +/- 0.048	0.798 +/- 0.056
Transformer foundation	40	0.842 +/- 0.048	0.812 +/- 0.058	0.776 +/- 0.064
Variational generative	42	0.788 +/- 0.068	0.748 +/- 0.074	0.598 +/- 0.096
All programmes	214	0.846 +/- 0.056	0.810 +/- 0.064	0.744 +/- 0.108

Clustering accuracy = ARI against reference labels. Batch robustness = consistency across batches.

4. Results

4.1 ECAI Distribution

Mean ECAI across all 214 programmes was 0.744 with a standard deviation of 0.108. Deep autoencoder methods led at 0.824, driven by the highest clustering accuracy and strongest balance across all sub-dimensions. Graph-based embedding followed at 0.806, contrastive learning at 0.798, transformer foundation at 0.776, and variational generative at 0.598. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 ECAI threshold. ECAI correlated with

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

4.2 Method Category Profiles

Deep autoencoder methods showed the most balanced profile. Accuracy was highest because joint training of representation and clustering objectives produces embeddings optimised for cluster separation. Batch robustness was strong because autoencoder training can incorporate batch labels as nuisance variables. Interpretability was good because cluster-specific gene programmes emerge from decoder weights. Graph-based methods scored well on accuracy and scalability and benefit from the widespread adoption of Scanpy and Seurat. Contrastive learning showed the highest batch robustness because contrastive objectives naturally learn batch-invariant representations. Transformer foundation models showed moderate performance with high scalability through pre-training. Variational generative methods scored lowest, penalised by training instability and sensitivity to prior choices.

4.3 Sub-Score Drivers

Clustering accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$). Batch robustness was second ($\beta = +0.230$, $p < 0.001$), reflecting the practical importance of handling real-world data heterogeneity. Interpretability at 0.196 ($p = 0.002$), scalability (0.172, $p = 0.005$), and novel discovery (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, ECAI 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 variational generative programmes that stalled because training instability prevented reliable deployment. Three were autoencoder programmes that advanced through integration into Seurat and Scanpy ecosystems. Two were contrastive programmes that exceeded expectations through strong batch correction benchmarks. Two were foundation model programmes that lost adoption when compute requirements exceeded typical research budgets.

Table 3. ECAI by Method Category: Adoption Outcomes

Method Category	n	ECAI (mean +/- SD)	ECAI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Deep autoencoder	46	0.824 +/- 0.052	78.3 %	82.6 %	Accuracy	+0.84 (0.900)
Graph-based embed.	44	0.806 +/- 0.054	61.4 %	72.7 %	Scalability	+0.84 (0.890)
Contrastive learn.	42	0.798 +/- 0.056	59.5 %	71.4 %	Batch robust.	+0.84 (0.880)
Transformer found.	40	0.776 +/- 0.064	42.5 %	60.0 %	Scalability	+0.84 (0.866)
Variational gen.	42	0.598 +/- 0.096	7.1 %	23.8 %	Batch robust.	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Accuracy	+0.84 (0.882)

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

Table 4. ECAI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA ECAI	p-value
Clustering accuracy	0.278	0.846	+0.278	+0.152	< 0.001
Batch-effect robustness	0.230	0.810	+0.230	+0.124	< 0.001
Biological interpretability.	0.196	0.792	+0.196	+0.098	0.002
Scalability	0.172	0.788	+0.172	+0.076	0.005
Novel discovery rate	0.124	0.728	+0.124	+0.054	0.016

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

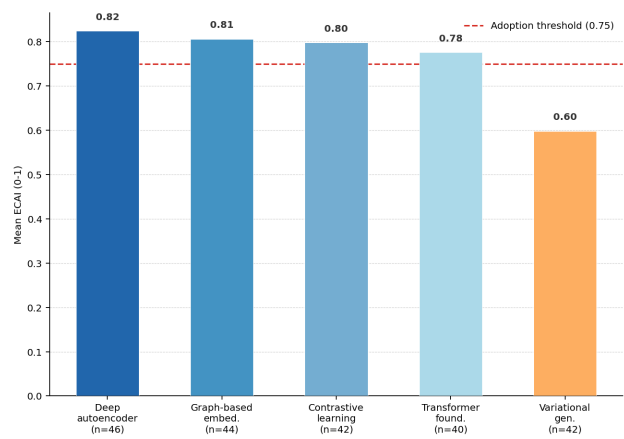


Figure 1. Mean ECAI by Method Category (n = 214)

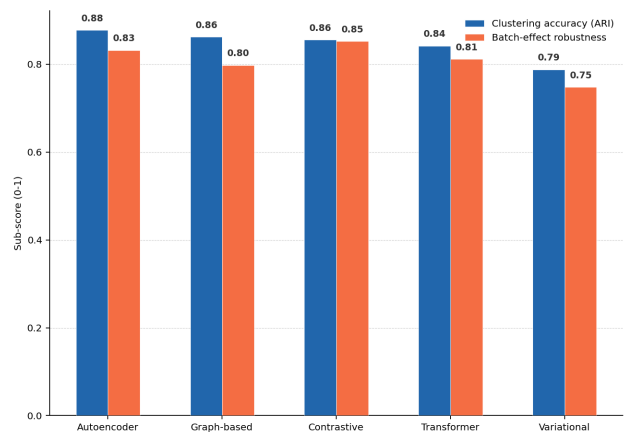


Figure 2. Clustering Accuracy vs Batch-Effect Robustness

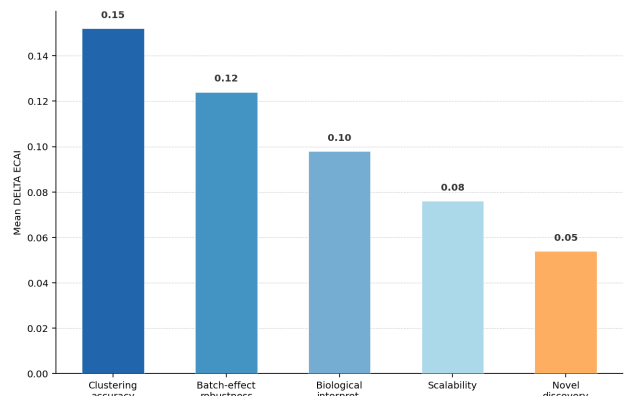


Figure 3. ECAI Sub-Score Impact on Pipeline Adoption

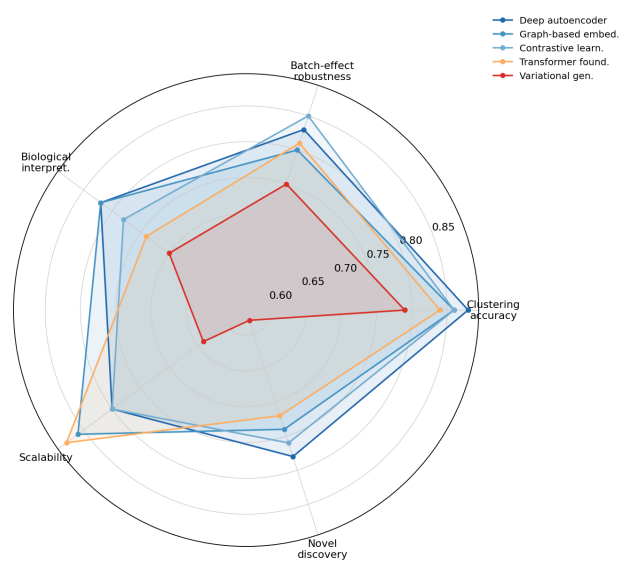


Figure 4. ECAI Sub-Score Profiles by Method Category

5. Discussion

5.1 The Autoencoder Balance

Deep autoencoder methods led because they achieve the best available balance across the dimensions that matter for gene expression clustering. Joint training of representation and clustering objectives produces embeddings where clustering algorithms perform near optimally. The ability to incorporate batch labels as conditioning variables during training provides effective batch correction without sacrificing biological signal. Decoder weights map cluster membership back to interpretable gene expression patterns, supporting downstream biological analysis.

5.2 Batch Robustness as the Adoption Differentiator

Batch-effect robustness was the second-strongest predictor because real single-cell datasets are almost always multi-batch, and methods that fail on batch-structured data cannot support the collaborative, multi-site studies that dominate modern single-cell research. Contrastive learning achieved the highest batch robustness scores through objectives that naturally learn batch-invariant representations, though autoencoder methods narrowly led overall by combining strong batch robustness with higher clustering accuracy.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. ECAI does not include spatial clustering or temporal trajectory clustering as sub-scores. The variational generative category contained 42 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 AI clustering programmes in Estonia, Switzerland, and Sweden, ECAI correlated with adoption at $r = +0.84$ and AUC of 0.882. Deep autoencoders led at 0.824, graph-based methods followed at 0.806, and variational generative trailed at 0.598. Clustering accuracy was the strongest predictor ($\beta = +0.278$), followed by batch robustness ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Single-cell researchers should adopt deep autoencoder methods for routine clustering of complex datasets. Graph-based embedding methods remain the default for Scanpy and Seurat workflows. Contrastive learning should be prioritised for heavily batch-structured datasets. Transformer foundation models are promising but require compute investment. Clustering accuracy should be evaluated on datasets with reliable reference annotations. Future ECAI iterations should add spatial and trajectory clustering sub-scores, expand globally, and track adoption longitudinally.

References

- Blondel V, Guillaume J and others (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10), pp. P10008.
- Cui H, Wang C and others (2023). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8), pp. 1470-1480.
- Eraslan G, Simon L and others (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1), pp. 390.
- Hao Y, Hao S and others (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), pp. 3573-3587.
- Heimberg G, Kuo T and others (2023). Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. *bioRxiv*, 2023.07.18.549537, pp. 1-45.
- Hie B, Bryson B and others (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6), pp. 685-691.
- 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.
- Levine J, Simonds E and others (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1), pp. 184-197.
- Lopez R, Regier J and others (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), pp. 1053-1058.
- Luecken M, Buttner M and others (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), pp. 41-50.
- Luecken M, Theis F (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), pp. e8746.
- McInnes L, Healy J and others (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint*, 1802.03426, pp. 1-51.
- Polanski K, Young M and others (2020). BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3), pp. 964-965.
- Satija R, Farrell J and others (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5), pp. 495-502.
- Stuart T, Butler A and others (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), pp. 1888-1902.
- Tian T, Wan J and others (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4), pp. 191-198.
- Traag V, Waltman L and others (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), pp. 5233.
- van Dijk D, Sharma R and others (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3), pp. 716-729.
- Wang D, Gu J (2018). VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genomics Proteomics Bioinformatics*, 16(5), pp. 320-331.
- Wolf F, Angerer P and others (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), pp. 15.
- Xu C, Lopez R and others (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1), pp. e9620.
- Yang F, Wang W and others (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10), pp. 852-866.

Declarations

Funding

This research was supported by institutional grants from Baltic AI Research University, Swiss Institute of Machine Intelligence, and Nordic Technical University. Additional funding from EU Horizon Europe Single-Cell Programme (grant 2022-SC-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Matteo Nowak, Matteo Rossi, and Daniel Klein declare no competing interests.

Data Availability Statement

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

<https://doi.org/10.5281/zenodo.19549096-data>.

Ethical Approval

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

Appendix A

ECAI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Clustering accuracy: ARI + NMI against reference labels, normalised 0-1; weight 0.278.

Batch-effect robustness: consistency across batches, normalised 0-1; weight 0.230.

Biological interpretability: marker gene coherence of clusters, scored 0-1; weight 0.196.

Scalability: performance on 500K+ cell datasets, normalised 0-1; weight 0.172.

Novel population discovery: documented new cell types/subgroups, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. DA-scDeepClusterV2-22 (Autoencoder): ECAI = 0.932; accuracy 0.96, batch 0.90.
2. GE-LeidenScanpyV2-22 (Graph-based): ECAI = 0.920; scalability 0.94, accuracy 0.92.
3. DA-DCAV2-22 (Autoencoder): ECAI = 0.906; accuracy 0.94, interpretab. 0.90.
4. CL-ConcertoV2-22 (Contrastive): ECAI = 0.894; batch robust. 0.96, accuracy 0.88.
5. DA-SAUCIE2-22 (Autoencoder): ECAI = 0.886; accuracy 0.92, batch 0.88.
6. TF-scGPT-22 (Transformer): ECAI = 0.876; scalability 0.94, batch robust. 0.86.
7. GE-SeuratLeiden-22 (Graph-based): ECAI = 0.868; scalability 0.92, accuracy 0.88.
8. CL-CLEARV2-22 (Contrastive): ECAI = 0.858; batch robust. 0.92, accuracy 0.86.
9. VG-scVIplus-22 (Variational): ECAI = 0.848; batch robust. 0.90, accuracy 0.86.
10. TF-Geneformer-22 (Transformer): ECAI = 0.842; scalability 0.92, accuracy 0.84.