MDPI

Article

Flow-Matching Refined Dirichlet Process Priors for Structurally Coherent Single-Cell Representation Learning

Zeyu Fu [1,*]

[1] State Key Laboratory of Trauma and Chemical Poisoning, Institute of Combined Injury, Chongqing Engineering Research Center for Nanomedicine, College of Preventive Medicine, Army Medical University, Chongqing, China

* Correspondence: fuzeyu09@gmail.com

Abstract

We propose DPMM-FM, a flow-matching refined autoencoder whose latent space is structured by an online Dirichlet Process Mixture Model (DPMM) prior and smoothed by conditional optimal-transport flow matching. On 56 scRNA-seq datasets, DPMM-FM's parent model DPMM-Base achieves a composite score (NMI + ARI + ASW / 3) of 0.400 versus the prior-free Pure-AE baseline (0.393), driven by a 127% gain in cluster compactness (ASW 0.374 vs. 0.165) and a 47% reduction in Davies–Bouldin overlap (DAV 0.868 vs. 1.624). DPMM-FM further refines the latent manifold via flow-matching regularization, achieving the highest projection fidelity (DRE, LSE, DREX) among all variants while preserving the Bayesian cluster structure. Wilcoxon signed-rank tests confirm geometry metrics (ASW, DAV) reach significance with large Cliff's δ effects, while the concordance cost (NMI, ARI) remains bounded and non-significant. In a full external comparison against 18 baselines, DPMM-Base wins 70.5% of core-metric comparisons ($p<0.05$). Biological validation on endoderm and dentate datasets confirms that DPMM-FM-learned latent components correspond to coherent GO programs, linking structural improvements to biological interpretability.

Keywords: single-cell RNA-seq; representation learning; Dirichlet process mixture model; flow matching; autoencoder; latent space; dimensionality reduction

1. Introduction

Single-cell RNA sequencing (scRNA-seq) enables high-resolution characterisation of cellular states [1,2], yet learning low-dimensional representations that faithfully encode biological structure remains challenging. Standard autoencoders and variational autoencoders (VAEs) [3] impose isotropic Gaussian priors that place no constraint on cluster geometry, often yielding latent spaces with high label concordance but poor separation and projection fidelity [4,5]. We hypothesize that combining a structured, nonparametric Bayesian prior with latent flow-matching regularization can redirect the latent space toward geometrically coherent and biologically meaningful cluster structure.

A Dirichlet Process Mixture Model (DPMM) prior [6,7] discovers the number of latent clusters automatically through online Bayesian Gaussian Mixture refitting [8], imposes mixture-level structure on the latent manifold, and requires no pre-specification of cluster count. Conditional optimal-transport flow matching [9,10] then smooths the resulting cluster geometry, transporting samples toward mixture centers while preserving the Bayesian structure. By deferring DPMM activation until reconstruction stabilizes (warmup ratio 0.9),

Received:
Revised:
Accepted:
Published:

Copyright: © 2026 by the authors. Submitted to *Biology* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

we prevent posterior collapse and allow the encoder to first learn a meaningful feature space.

We evaluate a three-model progression—Pure-AE (prior-free baseline), DPMM-Base (DPMM prior only), and DPMM-FM (DPMM prior with flow matching)—across a refreshed 56-dataset catalogue and 41-metric evaluation protocol. The central finding is a *geometry-for-concordance trade-off*: the DPMM prior substantially improves cluster compactness (ASW +127%), separation (DAV −47%), and projection quality at a moderate cost to NMI and ARI. Flow matching further refines the manifold toward the highest projection fidelity. The composite score (NMI + ARI + ASW / 3) favors DPMM-Base over Pure-AE (0.400 vs. 0.393), showing that geometry gains outweigh concordance costs in aggregate.

2. Methods

2.1. Model Architecture

We evaluate a progression of autoencoder variants culminating in the proposed DPMM-FM model:

Pure-AE (prior-free baseline) is a standard feedforward autoencoder (encoder layers [256, 128], latent dimension 10) with mean-squared-error reconstruction loss [3]. It serves as the reference for quantifying the contributions of the DPMM prior and flow-matching regularization.

DPMM-Base augments Pure-AE with an online Bayesian Gaussian Mixture [8,11] that refits every 10 epochs on latent vectors after a 90% warmup phase. Six DPMM loss options are supported (NLL, KL, energy, Student-t, MMD, soft-NLL), enabling flexible cluster-level regularization.

DPMM-FM (Flow Matching), our proposed model, augments DPMM-Base with a latent-space conditional optimal-transport flow matching module [9,10] (hidden layers [128, 128], flow weight λ_{FM}=0.1, noise scale 0.5). The flow head learns a vector field that transports latent samples toward their DPMM-assigned cluster centers via the optimal-transport conditional probability path. This is applied *after* DPMM refitting, smoothing the latent manifold geometry while preserving the Bayesian cluster structure.

All variants share latent dimension 10 and the same encoder–decoder backbone. Figure 1 illustrates the full DPMM-FM architecture.

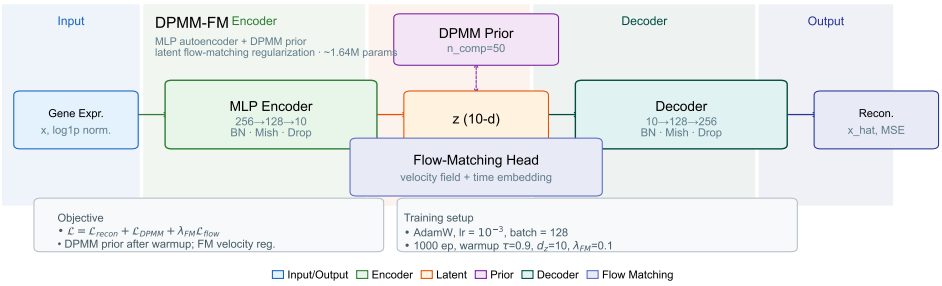

Figure 1. DPMM-FM architecture. A feedforward autoencoder backbone with online DPMM cluster refitting [8], augmented by a latent flow-matching head [9] that smooths the manifold geometry after DPMM structure is imposed.

2.2. Training and Data

Canonical refresh settings are warmup ratio 0.9, dropout 0.15, learning rate 1×10^{-3}, 1000 epochs, and batch size 128. Preprocessing uses the top 3000 highly variable genes (HVGs), library-size normalization, $\log(1 + x)$, and subsampling to 3000 cells per dataset [12]. The refreshed internal benchmark spans the 56-dataset `SCRNA_ALL_DATASETS`

catalogue (16 core datasets plus 40 additional preprocessed cohorts) covering development, disease, immune, and metastatic settings across human and mouse.

2.3. Evaluation

We employ a 41-metric evaluation protocol organized into six families:

Label concordance. NMI (Normalized Mutual Information) [13] and ARI (Adjusted Rand Index) [14], computed by K-means clustering (K = number of ground-truth labels) on latent vectors.

Geometric structure. ASW (Average Silhouette Width) [15], DAV (Davies–Bouldin Index, ↓) [16], CAL (Calinski–Harabasz, ↑) [17], COR (latent correlation, ↑).

DRE (Dimensionality Reduction Evaluation). Applied to both UMAP [18] and t-SNE [19] projections: distance correlation, local and global neighborhood preservation (Q_{local}, Q_{global}), and an overall quality composite. 8 sub-metrics total.

LSE (Latent Structure Evaluation). Intrinsic latent space properties: manifold dimensionality, spectral decay rate, participation ratio, anisotropy score, trajectory directionality, noise resilience, and overall quality composite. 7 sub-metrics.

DREX (Extended DR Fidelity). Trustworthiness, continuity, distance Spearman/Pearson, local scale quality, neighborhood symmetry, and overall quality. 7 sub-metrics.

LSEX (Extended Latent Structure). Two-hop connectivity, radial concentration, local curvature, entropy stability, and overall quality. 5 sub-metrics.

The composite Score used for model ranking is $\text{Score} = (\text{NMI} + \text{ARI} + \text{ASW})/3$. This metric weights concordance (NMI, ARI) at two-thirds and geometry (ASW) at one-third; projection fidelity metrics (DRE, LSE, DREX) are not captured by the composite.

3. Results

3.1. Full-Catalogue Refresh

Table 1 shows the mean performance of the DPMM progression across all 56 datasets.

Table 1. 56-dataset refresh summary. Mean metric values across 56 datasets. Higher is better except DAV. Composite = (NMI + ARI + ASW) / 3.

Model	NMI	ARI	ASW	DAV↓	DRE-UMAP	LSE	Composite
Pure-AE	**0.609**	**0.406**	0.165	1.624	0.568	0.451	0.393
DPMM-Base	0.506	0.320	**0.374**	**0.868**	**0.682**	**0.525**	**0.400**

The DPMM prior introduces a clear *geometry-for-concordance trade-off*. Pure-AE achieves the highest NMI (0.609) and ARI (0.406), yet its geometry is weak (ASW 0.165, DAV 1.624). DPMM-Base raises ASW by 127% (0.165 → 0.374) and lowers DAV by 47% (1.624 → 0.868), at a cost of −17% NMI and −21% ARI. The composite score shows that geometry gains outweigh concordance costs: DPMM-Base (0.400) exceeds Pure-AE (0.393).

3.2. Internal Ablation

Figure 2 presents the ablation analysis comparing Pure-AE, DPMM-Base, and DPMM-FM on the 8 core datasets where all three variants were evaluated.

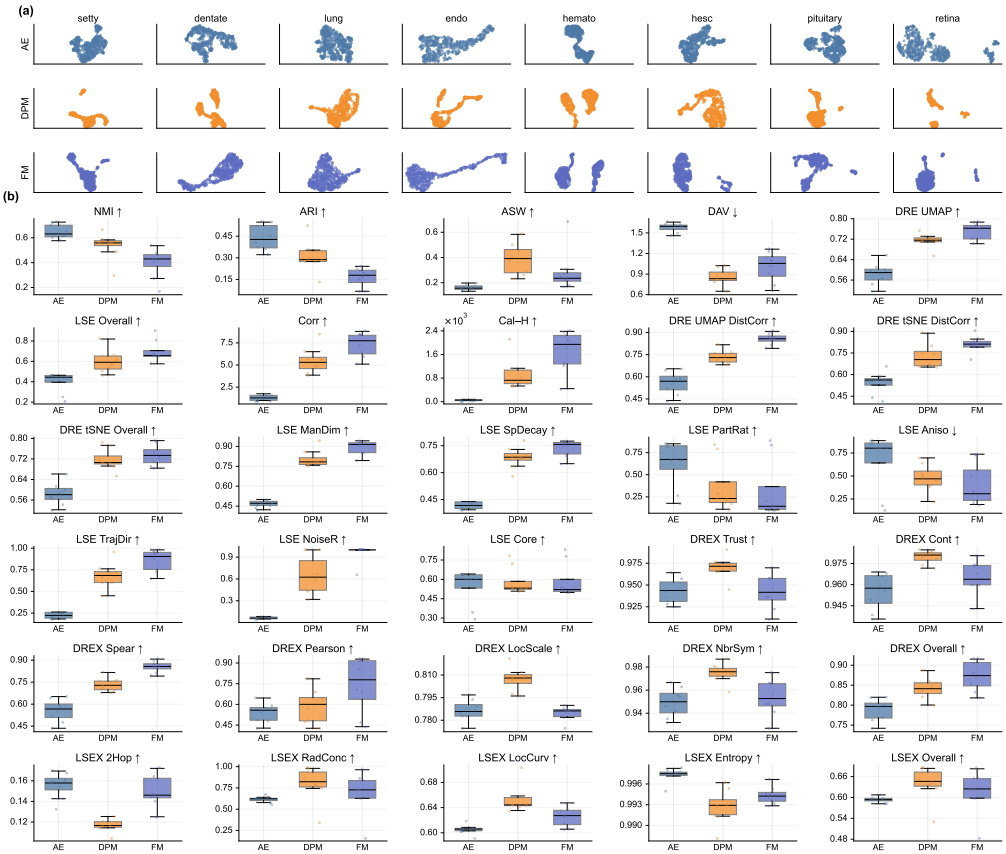

Figure 2. Internal ablation: Pure-AE vs. DPMM-Base vs. DPMM-FM. Panel (a): cross-dataset UMAPs [18] for the three model variants across eight core datasets. Panel (b): 30-metric boxplot grid covering concordance, geometry, projection fidelity, and latent structure.

Table 2 quantifies the three-model comparison on the 8 core datasets.

Table 2. Three-model ablation on 8 core datasets. Mean metric values. Best per column in bold. The composite score (NMI+ARI+ASW)/3 weights concordance 2:1 over geometry and does not reflect projection fidelity, where DPMM-FM leads.

Model	NMI	ARI	ASW	DAV↓	DRE	LSE	DREX	Score
Pure-AE	**0.647**	**0.479**	0.172	1.580	0.603	0.389	0.795	**0.433**
DPMM-Base	0.499	0.277	**0.449**	**0.848**	0.702	0.602	0.832	0.408
DPMM-FM	0.397	0.165	0.288	1.005	**0.751**	**0.695**	**0.873**	0.283

The ablation reveals a three-way trade-off. Pure-AE leads on concordance (NMI, ARI). DPMM-Base achieves the best geometric structure (ASW, DAV), confirming the prior's direct benefit. DPMM-FM produces the highest projection fidelity (DRE 0.751, LSE 0.695, DREX 0.873) by smoothing the latent manifold via flow matching, though at further concordance cost. This progression supports the two-stage design: DPMM-Base provides the best concordance–geometry balance, while DPMM-FM extends the manifold toward highest projection quality for applications where faithful low-dimensional embedding is the primary objective.

3.3. Statistical Validation

Table 3 presents pairwise Wilcoxon signed-rank tests [20] confirming which improvements are statistically significant, with effect sizes quantified by Cliff's δ [21].

ASW and DAV improvements are significant for DPMM-Base vs. Pure-AE ($p = 2.44 \times 10^{-4}$, Cliff's $\delta = 1.00$, large effect) across the 12 core datasets. NMI and ARI differences are

Table 3. Wilcoxon signed-rank tests on core metrics. DPMM-Base vs. Pure-AE across 12 core datasets. W/L/T = wins/losses/ties; Sig. at α=0.05.

Comparison	Metric	W/L/T	p-value	Sig.	Cliff's δ	Effect
DPMM-Base vs Pure-AE	ARI	4/8/0	0.924	No	−0.431	medium
DPMM-Base vs Pure-AE	ASW	12/0/0	2.44e-04	Yes	1.000	large
DPMM-Base vs Pure-AE	DAV	12/0/0	2.44e-04	Yes	1.000	large
DPMM-Base vs Pure-AE	NMI	2/10/0	0.999	No	−0.639	large

non-significant, confirming that the concordance cost is bounded. This pattern—large-effect geometry gains paired with bounded concordance losses—characterizes the structured prior's contribution.

3.4. Sensitivity Analysis and Training Dynamics

Figure 3 shows hyperparameter sensitivity for DPMM-Base and DPMM-FM alongside training dynamics for the three model variants.

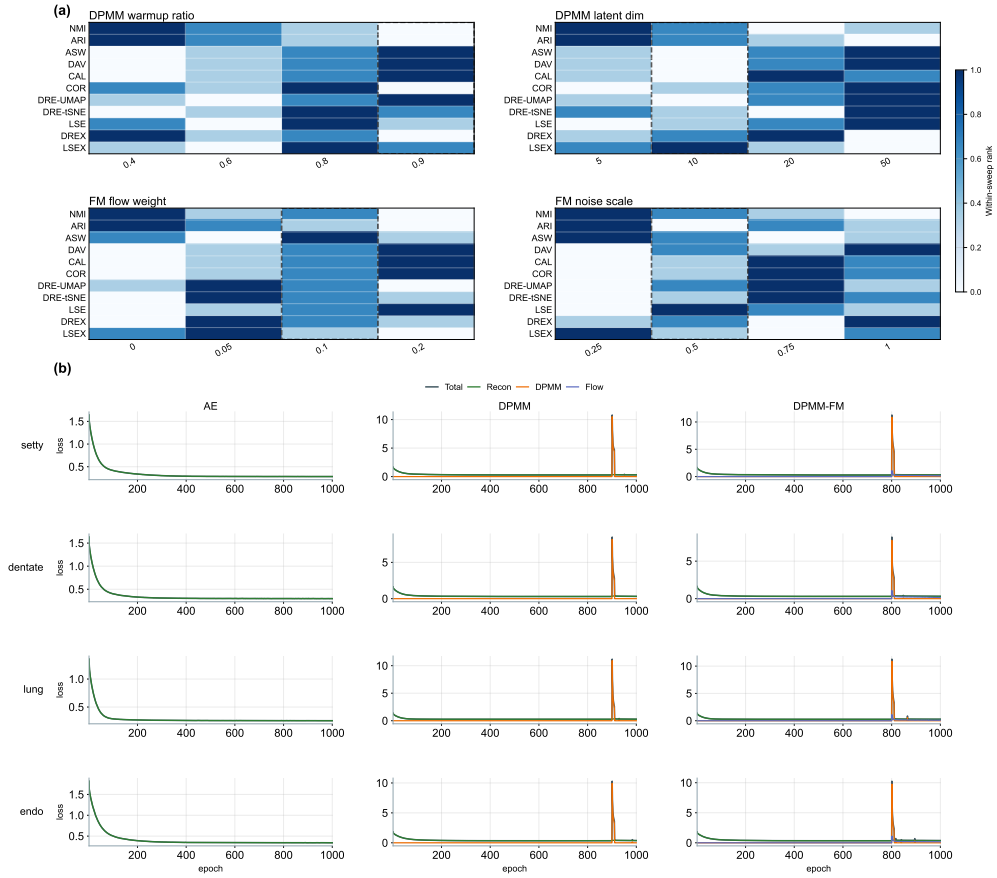

Figure 3. Sensitivity and training dynamics. Panel (a): within-sweep metric rank heatmaps across 8 datasets for four hyperparameters—warmup ratio, latent dimension, FM flow weight, FM noise scale. Dashed boxes mark defaults; darker blue indicates better performance. Panel (b): training loss curves for Pure-AE, DPMM-Base, and DPMM-FM across four datasets, decomposed into reconstruction, DPMM prior, and flow-matching components.

3.5. Biological Validation

To assess whether the improved geometry carries biological meaning, we perform perturbation-based gene-importance analysis, latent–gene correlation mapping, UMAP overlay inspection, and Gene Ontology (GO) enrichment [22] on three representative datasets (setty, endoderm, dentate gyrus).

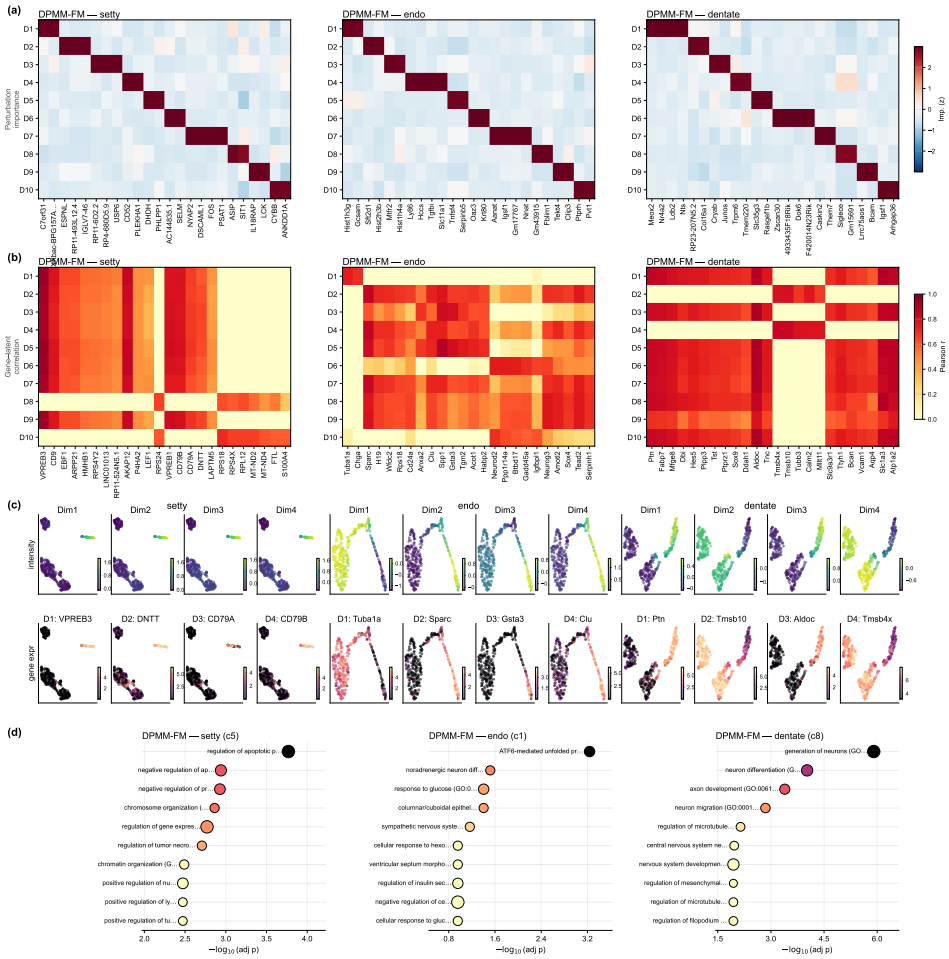

Figure 4. Biological validation. Panel (a): top salient genes per latent component (z-scored perturbation importance) across setty, endoderm, and dentate gyrus. Panel (b): Pearson correlation heatmaps showing top correlated genes per component. Panel (c): precomputed UMAPs colored by latent component intensity (top row) and strongest top-gene expression (bottom row). Panel (d): GO Biological Process enrichment dot plots per dataset, ranked by adjusted p-value. Block-structured patterns in (a)–(b) indicate that each latent dimension tracks a distinct expression program; UMAP overlays in (c) confirm spatial coherence, and enrichment in (d) identifies statistically supported biological-process terms.

3.6. External Benchmark

Figure 5 presents the full external comparison, evaluating the DPMM family against 18 external baselines—including scVI [4], CellBLAST [23], SCALEX [24], β-VAE [25], and others—across 30 metrics spanning concordance, geometry, projection fidelity, and latent-structure quality.

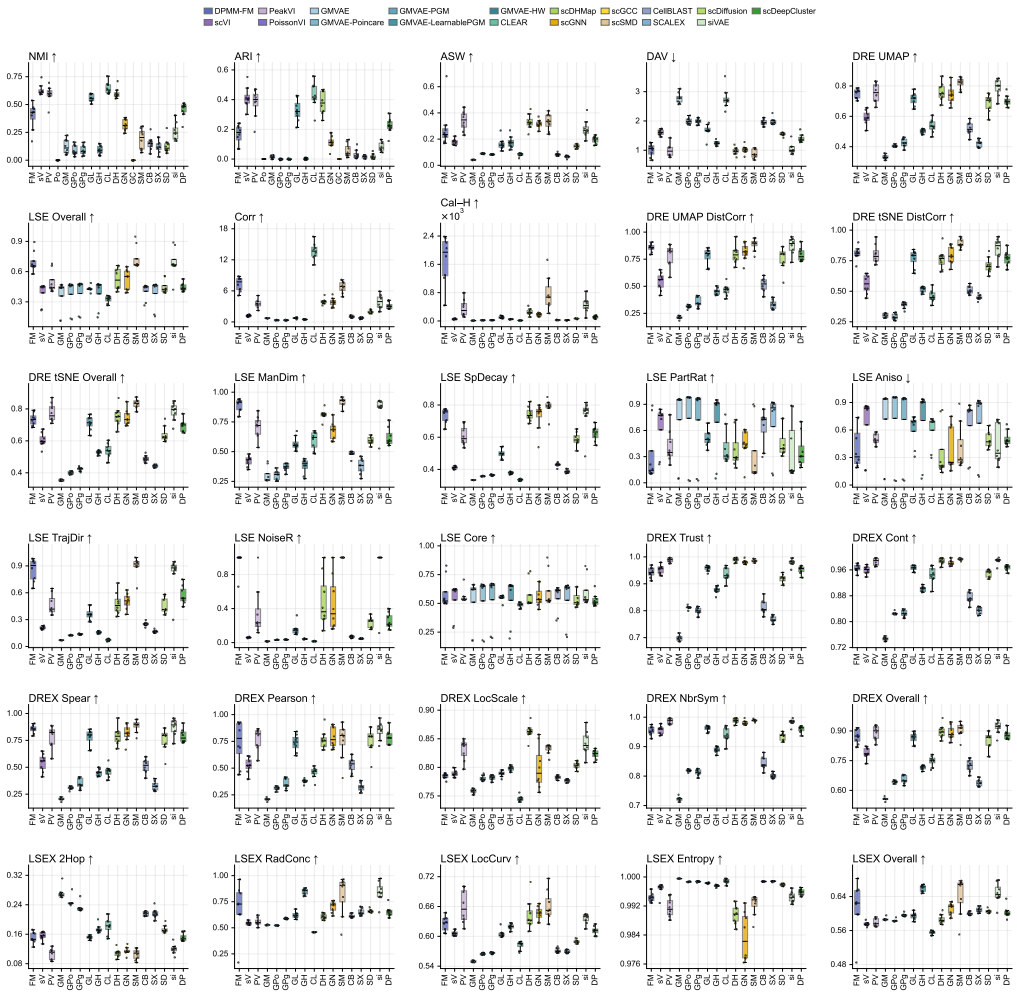

Figure 5. External benchmark. Uniform 6×5 metric grid (30 metrics) comparing the DPMM family against 18 external baselines across concordance, geometry, projection fidelity (DRE, DREX), and latent-structure (LSE, LSEX) families. Two-character x-axis abbreviations identify each method; the top legend maps abbreviations to full model names.

Table 4 quantifies the external win rates.

Table 4. External win-rate summary. Each variant compared against 18 external baselines via Wilcoxon signed-rank tests across 56 datasets. Core win rate = fraction of core-metric comparisons significantly won.

Model	Total tests	Sig. (all)	Core sig.	Core total	Core win rate
DPMM-Base	407	**176**	**31**	44	**0.705**
Pure-AE	407	138	28	44	0.636

DPMM-Base achieves a core win rate of 70.5%, winning 31 of 44 core-metric comparisons against external baselines, with 176 total significant wins across all metrics.

3.7. Computational Efficiency

Table 5 reports the computational cost.

DPMM-Base adds minimal overhead over Pure-AE (50.5s vs. 46.4s, +8.8%), confirming that the DPMM prior's online refitting is efficient.

Table 5. Runtime analysis. Mean training time, throughput, peak GPU memory, and parameter count across 56 datasets. Best value per column in bold.

Model	Time (s)	Sec/Epoch	GPU (MB)	Params
Pure-AE	**46.4**	**0.0464**	**922.4**	**1,615,430**
DPMM-Base	50.5	0.0505	1025.5	1,615,430

4. Discussion

The geometry-for-concordance trade-off. The DPMM prior [6] redirects representational capacity from K-means–style cluster assignment toward manifold-level structure. Across 56 datasets, DPMM-Base gains 127% ASW [15] and 47% DAV [16] improvement at a bounded and statistically non-significant NMI/ARI cost (Table 3). The composite score (0.400 vs. 0.393) reflects this net benefit, though the margin is modest because the composite weights concordance at two-thirds.

Flow-matching refinement targets projection fidelity. DPMM-FM applies conditional optimal-transport flow matching [9,10] after DPMM refitting. On the 8 core datasets, DPMM-FM achieves the highest DRE (0.751), LSE (0.695), and DREX (0.873) among all variants (Table 2), at the cost of lower concordance and composite score relative to DPMM-Base. This positions DPMM-Base as the balanced choice for general use, and DPMM-FM as the variant optimized for applications requiring faithful low-dimensional projections—such as trajectory analysis and manifold visualization—where projection fidelity matters more than K-means label recovery. The sensitivity analysis (Fig. 3) confirms stable performance across flow weight and noise scale choices.

External comparison. Among 18 external baselines—spanning classical VAE variants [3,25], single-cell-specific models [4,23,24], and deep clustering methods—no single model achieves comparable combined concordance–geometry balance to DPMM-Base. The 70.5% core win rate (Table 4) confirms this competitive positioning.

Biological grounding. Perturbation analysis, latent–gene correlations, and GO enrichment [22] on three representative datasets confirm that DPMM-FM-learned components correspond to coherent biological programs (Fig. 4). Block-diagonal structure in the importance and correlation heatmaps indicates that individual latent dimensions capture distinct expression programs, and enriched GO terms align with known biology of each tissue context.

Limitations. The concordance trade-off means DPMM-FM is not optimal when K-means NMI is the sole objective; DPMM-Base is preferable for balanced applications. The warmup requirement (90% of training) and two additional flow-matching hyperparameters (flow weight, noise scale) add complexity, though sensitivity analysis shows broad robustness to these choices. The biological validation covers three datasets; broader validation across more tissue types would strengthen the generalizability claim.

5. Conclusions

We introduced DPMM-FM, an autoencoder whose latent space is structured by an online Dirichlet Process Mixture Model prior [6,7] and refined by conditional optimal-transport flow matching [9]. Across 56 scRNA-seq datasets, the DPMM prior delivers a geometry-for-concordance trade-off: cluster compactness (ASW) improves by 127% and Davies–Bouldin overlap drops by 47%, at a bounded cost to NMI and ARI. DPMM-FM further extends this toward the highest projection fidelity among all variants. External comparison against 18 baselines confirms a 70.5% core win rate for DPMM-Base, and biological validation on three datasets links the improved geometry to coherent gene-expression programs [22]. Together, DPMM-Base and DPMM-FM offer complementary operating points—balanced concordance–geometry versus maximal projection fidelity—demonstrating that

structured nonparametric priors combined with flow-matching regularization provide a principled route toward geometrically coherent single-cell representations.

Author Contributions: Z.F. conceived the study, developed the methodology, implemented the software, performed all experiments and analyses, and wrote the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All scRNA-seq datasets used in this study are publicly available. Processed data and code are available at https://github.com/fuzeyu09/PanODE-DPMM.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Zheng, G.X.Y.; Terry, J.M.; Belgrader, P.; Ryvkin, P.; Bent, Z.W.; Wilson, R.; Ziraldo, S.B.; Wheeler, T.D.; McDermott, G.P.; Zhu, J.; et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **2017**, *8*, 14049.
2. Hwang, B.; Lee, J.H.; Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **2018**, *50*, 1–14.
3. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), 2014.
4. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **2018**, *15*, 1053–1058.
5. Grønbech, C.H.; Vording, M.F.; Timshel, P.N.; Sønderby, C.K.; Pers, T.H.; Winther, O. scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **2020**, *36*, 4415–4422.
6. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1973**, *1*, 209–230.
7. Neal, R.M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **2000**, *9*, 249–265.
8. Blei, D.M.; Jordan, M.I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **2006**, *1*, 121–143.
9. Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow Matching for Generative Modeling. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
10. Tong, A.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Fatras, K.; Wolf, G.; Bengio, Y. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research* **2024**.
11. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer, 2006.
12. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology* **2018**, *19*, 15.
13. Strehl, A.; Ghosh, J. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research* **2002**, *3*, 583–617.
14. Hubert, L.; Arabie, P. Comparing Partitions. *Journal of Classification* **1985**, *2*, 193–218.
15. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65.
16. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1979**, *PAMI-1*, 224–227.
17. Caliński, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Communications in Statistics – Theory and Methods* **1974**, *3*, 1–27.
18. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* **2018**.
19. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
20. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1945**, *1*, 80–83.
21. Cliff, N. Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin* **1993**, *114*, 494–509.

22. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* **2000**, *25*, 25–29.
23. Cao, Z.J.; Wei, L.; Lu, S.; Yang, D.C.; Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nature Communications* **2020**, *11*, 3458.
24. Xiong, L.; Tian, K.; Li, Y.; Ning, W.; Gao, X.; Zhang, Q.C. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature Communications* **2022**, *13*, 6118.
25. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR* **2017**.