A Standardized Framework for Evaluating Gene Expression Generative Models

Andrea Rubbi[1,2,†] **Andrea G. Di Francesco**[3,4,†] **Mohammad Lotfollahi**[2,5,6] **Pietro Liò**[1]

[†]Equal contribution.
[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom
[2]Wellcome Sanger Institute, Cambridge, United Kingdom
[3]Sapienza University of Rome, Rome, Italy
[4]ISTI-CNR, Institute of Information Science and Technologies, Pisa, Italy
[5]Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, United Kingdom
[6]Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom

Abstract

The rapid development of generative models for single-cell gene expression data has created an urgent need for standardised evaluation frameworks. Current evaluation practices suffer from inconsistent metric implementations, incomparable hyperparameter choices, and a lack of biologically-grounded metrics. We present *Generated Genetic Expression Evaluator* (*GGE*), an open-source Python framework that addresses these challenges by providing a comprehensive suite of distributional metrics with explicit computation space options and biologically-motivated evaluation through differentially expressed gene (DEG)-focused analysis and perturbation-effect correlation, enabling standardized reporting and reproducible benchmarking. Through extensive analysis of the single-cell generative modeling literature, we identify that no standardized evaluation protocol exists. Methods report incomparable metrics computed in different spaces with different hyperparameters. We demonstrate that metric values vary substantially depending on implementation choices, highlighting the critical need for standardization. *GGE* enables fair comparison across generative approaches and accelerates progress in perturbation response prediction, cellular identity modeling, and counterfactual inference.

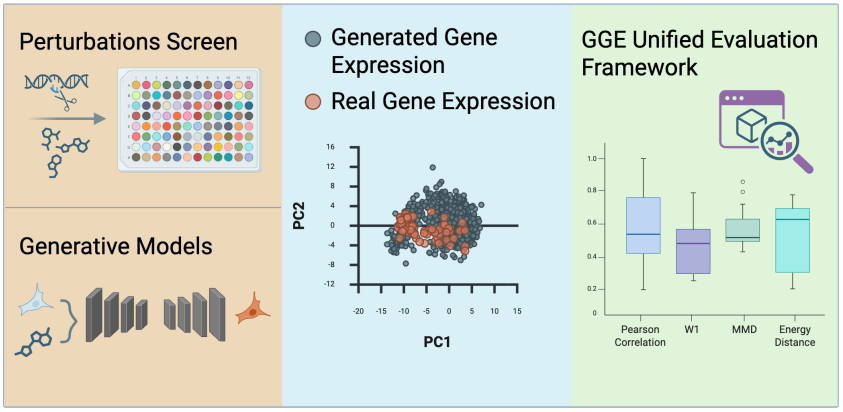

Figure 1: GGE is a unified framework for the evaluation of generated genetic expression profiles. GGE aims to standardise the otherwise heterogeneous set of evaluation protocols proposed by the multitude of generative models implemented for this task.

1 INTRODUCTION

Generative models for single-cell gene expression data have emerged as powerful tools for understanding cellular biology, with applications spanning perturbation response prediction, developmental trajectory modeling, and counterfactual inference for therapeutic discovery. The field has witnessed remarkable progress: variational autoencoders (VAEs) established early benchmarks for perturbation prediction (Lotfollahi et al., 2019), graph neural networks enabled reasoning over gene regulatory networks (Roohani et al., 2024), and optimal transport methods provided principled frameworks for modeling cellular transformations (Bunne et al., 2023). Most recently, flow matching approaches have demonstrated state-of-the-art performance on distribution matching tasks (Lipman et al., 2022; Tong et al., 2023), opening new possibilities for learning continuous transformations between cellular states.

Despite this rapid progress, a fundamental problem persists: the field lacks standardized evaluation practices. This absence of standardization has profound consequences for scientific progress. When different methods report "Wasserstein distance" using incompatible formulations–per-gene 1D distances averaged across genes, multivariate distances in raw $\mathbb{R}^{2000}$ space, or distances in PCA-reduced $\mathbb{R}^{50}$ space–meaningful comparison becomes impossible. When critical hyperparameters such as Sinkhorn regularization strength, kernel bandwidth for MMD, or DEG significance thresholds go unreported, results cannot be reproduced or compared. When aggregate metrics over all genes obscure biologically relevant differences concentrated in a small subset of differentially expressed genes, we may miss the most important signals.

The heterogeneity of evaluation practices in the current literature is striking. We surveyed 12 influential methods for single-cell generative modeling (Table 1) and found that no two papers use identical evaluation protocols. Some methods report only reconstruction metrics (MSE, R^2), while others focus on distributional metrics; some compute metrics in raw gene space, others in PCA space; some evaluate on all genes, others on DEGs alone. This inconsistency makes it virtually impossible to determine which methods represent genuine advances.

We present *GGE*, an open-source Python framework that addresses these challenges through two key design principles. First, we make all implementation choices explicit through a unified API with a `space` parameter ("raw", "pca", or "deg") and standardised parameters. Second, we enable biologically meaningful evaluation through perturbation-effect correlation computed in DEG space, measuring whether models capture the direction and magnitude of perturbation effects rather than merely reconstructing expression levels. Our goal in this work is to standardize and expose evaluation choices that are often left implicit, rather than to claim a definitive benchmark across all model families.

2 RELATED WORK: THE EVALUATION LANDSCAPE

2.1 GENERATIVE MODELS FOR SINGLE-CELL PERTURBATION DATA

Generative modeling for single-cell gene expression has progressed from VAE-based likelihood models to distribution-matching and foundation-model approaches. Early work such as scVI (Lopez et al., 2018) and scGen (Lotfollahi et al., 2019) established VAE baselines for denoising and perturbation prediction, later extended to combinatorial perturbations by CPA (Lotfollahi et al., 2023). Biology-informed architectures incorporate prior knowledge, e.g. GEARS (Roohani et al., 2024) injects gene–gene interaction structure to improve generalization to unseen perturbations. In parallel, optimal transport formulations (e.g. CellOT (Bunne et al., 2023)) and flow matching methods (Lipman et al., 2022; Tong et al., 2023; Klein et al., 2025; Rubbi et al., 2026) explicitly target *conditional distributions* rather than conditional means. Finally, transformer-based foundation models (e.g. STATE (Adduri et al., 2025)) emphasize large-scale representation learning and multi-task generalization across contexts. A more detailed landscape review is provided in Appendix A.

Table 1: Survey of evaluation metrics used by single-cell generative models. "Space" indicates whether metrics are computed in raw gene space (Raw), PCA space (PCA-k), or on highly variable genes (HVG). Dashes indicate the metric was not reported. Recent methods (STATE, CellFlow) show a trend toward more comprehensive evaluation, yet protocols remain inconsistent.

Method	**MSE**	$\boldsymbol{R^2}$	$\boldsymbol{W_1/W_2}$	**MMD**	**Corr**	**DEG-specific**	**Space**
scVI (Lopez et al., 2018)	✓	✓	—	—	—	—	Raw
scGen (Lotfollahi et al., 2019)	✓	✓	✓	✓	✓	Top-100 DEG	Raw
Waddington-OT (Schiebinger et al., 2019)	—	—	✓	—	—	—	PCA-30
CellOT (Bunne et al., 2023)	✓	—	✓	—	—	—	PCA-50
CPA (Lotfollahi et al., 2023)	✓	✓	—	—	✓	Top-50 DEG	Raw
GEARS (Roohani et al., 2024)	✓	—	—	—	✓	Top-20 DEG	Raw/HVG
scVelo (Bergen et al., 2020)	—	✓	—	—	✓	—	Raw
PRESCIENT (Yeo et al., 2021)	—	—	✓	—	—	—	PCA-100
scDiff (Tang et al., 2023)	✓	—	✓	✓	—	—	PCA-50
CFGen (Palma et al., 2025)	—	—	✓	✓	—	—	PCA-50
STATE (Adduri et al., 2025)	✓	✓	✓	✓	✓	Top-20 DEG	PCA-50
CellFlow (Klein et al., 2025)	✓	—	✓	✓	✓	—	PCA-50

2.2 Survey of Evaluation Practices

To quantify the heterogeneity of evaluation practices in the literature, we surveyed 12 influential methods for single-cell generative modeling. Table 1 summarizes the metrics, computation spaces, and evaluation protocols used by each method. Several patterns emerge from this analysis.

First, there is no consensus on which metrics to report. While MSE or R^2 appears in 10/12 surveyed methods, distributional metrics (W_1, W_2, or MMD) appear in 8/12, and correlation-based metrics in 7/12. This heterogeneity makes direct cross-paper comparison difficult, as methods may appear superior by optimizing for mean accuracy while failing to capture the full distributional variance.

Second, implementation details for the same metric name differ substantially. "Wasserstein distance" is used inconsistently across the literature, referring variously to the 1D per-gene average W_1, a multivariate Sinkhorn approximation, or exact W_2 via linear programming on cell subsets. Furthermore, the number of PCA components used as the coordinate space varies from 30 to 100, which directly alters the sensitivity of distance-based evaluations.

Third, biological context is handled inconsistently. While early methods evaluated on all genes, newer perturbation models like GEARS focus on Pearson correlation and MSE calculated specifically on the top 20 differentially expressed genes (DEGs). This shift highlights the importance of capturing the perturbation signal itself rather than just the steady-state transcriptomic background.

These observations motivate the design of *GGE*: a framework that makes these choices explicit, provides standardized implementations of multivariate and per-gene metrics, and enables fair comparison across disparate modeling architectures.

3 Theoretical Foundations

Before describing *GGE*'s implementation, we establish the theoretical foundations of distributional metrics for gene expression data and analyze when different computation spaces are appropriate. Our goal is to formalize evaluation choices that are often treated implicitly in prior work, and to clarify their biological interpretation.

3.1 Problem Formulation

Let $P = \{x_1, \ldots, x_n\}$ and $Q = \{y_1, \ldots, y_m\}$ be samples from two distributions representing real and generated gene expression data, where $x_i, y_j \in \mathbb{R}^G$ (expression of G genes). We seek metrics $d(P, Q)$ that quantify distributional similarity. In the context of single-cell perturbation modeling, these distributions represent empirical samples from conditional data-generating processes (e.g. control vs. perturbed cells), and the choice of metric encodes which biological properties of the distribution we prioritize: reconstruction fidelity, preservation of population heterogeneity, recovery

of multimodality, or consistency of gene–gene dependency structure (Lopez et al., 2018; Bunne et al., 2023; Lipman et al., 2022).

Importantly, unlike bulk RNA-seq, single-cell data exhibits substantial intrinsic stochasticity, cell-state heterogeneity, and technical noise (Svensson et al., 2017). As a result, evaluating only pointwise errors (e.g. MSE between predicted and observed means) conflates distributional mismatch with biological variability. Distributional metrics are therefore essential for assessing whether generative models capture the full conditional distribution of cellular responses rather than only their average behavior (Bunne et al., 2023; Tong et al., 2023).

3.2 Distributional metrics: Optimal Transport, MMD, and Energy

Distributional metrics are essential for evaluating single-cell generative models because they assess whether a model recovers the *population-level response* to a perturbation, including heterogeneity and multimodality, rather than only matching mean expression levels. This is particularly important in perturbation settings, where interventions often induce continuous phenotypic shifts and heterogeneous outcomes across subpopulations.

Optimal transport (OT) metrics. Wasserstein distances (e.g., 1- and 2-Wasserstein) define a geometric notion of discrepancy between probability distributions by measuring the minimum cost required to transport mass from generated samples to real data under a chosen ground metric (Villani, 2008; Peyré & Cuturi, 2019). In biological applications, OT admits a natural interpretation as the minimal displacement of cell populations in expression space required to align perturbed and reference distributions, making it well-suited for modeling cellular state transitions. Recent generative modeling frameworks, including flow matching and OT-based approaches, explicitly optimize objectives related to Wasserstein distances, further motivating their use as evaluation criteria (Arjovsky et al., 2017; Lipman et al., 2022).

Kernel- and distance-based two-sample metrics. Maximum Mean Discrepancy (MMD) and Energy Distance provide complementary, nonparametric two-sample statistics for comparing generated and real distributions directly from samples, without assuming parametric density models (Gretton et al., 2012; Székely & Rizzo, 2013). Unlike OT, these metrics do not induce an explicit notion of mass transport or sample coupling and therefore lack a direct geometric interpretation in data space. However, they are typically more computationally efficient and stable to estimate, which has led to their widespread adoption as benchmarking metrics for generative models.

In *GGE*, we support OT-, MMD-, and Energy-based metrics under a unified API. Crucially, we expose all implementation choices (e.g. entropic regularization for OT, kernel bandwidth for MMD) and the representation space in which distances are computed, as these choices substantially affect reported values (formal definitions and computational details are provided in Appendix B).

3.3 The Space Question: A Theoretical Analysis

A critical choice in computing any distributional metric is the space in which distances are computed. This choice substantially affects both statistical behavior and biological interpretability (Bunne et al., 2023; Lotfollahi et al., 2023).

Raw gene space. Computing distances in $\mathbb{R}^G$ with $G \sim 5{,}000$–$20{,}000$ ($G \sim 2{,}000$–$5{,}000$ when only highly variable genes are selected) preserves gene-level interpretability, allowing direct attribution of discrepancies to specific genes. However, high-dimensional concentration of measure implies that pairwise distances become less discriminative as G grows (Beyer et al., 1999), and noisy lowly expressed genes contribute equally to the metric despite limited biological relevance. Moreover, technical variation and dropout inflate distances in raw space, potentially dominating biological signal (Svensson et al., 2017).

PCA space. Projecting to $\mathbb{R}^k$ using PCA mitigates both statistical and computational issues by restricting attention to dominant axes of variation (Pearson, 1901). In single-cell analysis, PCA is routinely used to denoise expression data and capture major biological programs (e.g. cell cycle, differentiation trajectories) (Luecken & Theis, 2019). However, PCA bases are fit globally and may underrepresent perturbation-specific gene programs that exhibit low variance in control conditions.

As a result, PCA-based distances may systematically underweight rare but biologically important responses.

DEG-restricted space. Restricting evaluation to differentially expressed genes focuses metrics on biologically salient perturbation effects, aligning evaluation with common biological validation practices in perturbation studies (Replogle et al., 2022). However, DEG selection introduces additional hyperparameters (log-fold-change and statistical significance thresholds), and DEG sets may be unstable under small sample sizes or noisy conditions, leading to high variance in downstream metrics.

Data normalization and transformation. In addition to the choice of computation space, the preprocessing applied to gene expression values substantially affects metric values. Single-cell RNA-seq data are typically processed through several transformations prior to downstream analysis, including library-size normalization (e.g. counts per million or size-factor normalization) and variance-stabilizing transformations such as $\log(1 + x)$ (Luecken & Theis, 2019). These transformations alter the geometry of the data: raw count space is dominated by library-size variation and highly expressed genes, whereas normalized and log-transformed representations compress dynamic range and emphasize relative expression differences across genes. Consequently, distributional metrics such as Wasserstein distance, MMD, and Energy distance may yield substantially different values depending on whether they are computed on raw counts, normalized counts, or log-transformed expression matrices.

In practice, most single-cell generative models operate on normalized and $\log(1 + x)$-transformed data during training and evaluation, and benchmarking studies often report metrics in this transformed space. However, this choice is rarely documented explicitly in published evaluation protocols, further complicating cross-paper comparison. For reproducibility, GGE assumes that input datasets are provided in the same representation used for model training (typically library-size-normalized and log-transformed expression values) and treats this preprocessing stage as part of the evaluation configuration that should be reported explicitly alongside the computation space.

Biological interpretation of distributional metrics. From a biological standpoint, distributional metrics answer a different question than pointwise errors: they quantify whether a generative model recovers the *population-level response* to a perturbation, including heterogeneity, multimodality, and the relative abundance of subpopulations. This is crucial in applications such as drug response modeling, CRISPR perturbation screens, and cell fate prediction, where interventions often induce heterogeneous outcomes rather than uniform shifts (Bunne et al., 2023; Lotfollahi et al., 2023). Evaluating only mean expression risks rewarding models that collapse diverse responses into averaged, biologically implausible predictions.

An effective framework. We recommend a multi-space evaluation strategy: PCA-50 for primary distributional metrics (statistical robustness and computational tractability), DEG-restricted space for biologically targeted evaluation, and raw gene space only when gene-level interpretability is explicitly required. This triangulation reflects the complementary roles of statistical fidelity and biological relevance in benchmarking generative models for single-cell perturbation prediction.

4 THE GGE FRAMEWORK

Building on the theoretical foundations established above, we now describe *GGE*'s architecture, API design, and key innovations.

4.1 ARCHITECTURE AND DESIGN PRINCIPLES

GGE is built around two core design principles that address the standardization challenges identified in our literature survey.

Explicit Configuration. Every implementation choice that affects metric values is exposed as an explicit parameter. The `space` parameter ("raw", "pca", or "deg") determines the computation space. For PCA space, `n_components` specifies the dimensionality. For DEG space, `deg_lfc`

and `deg_pval` control differential expression thresholds, while the `n_top_degs` parameter enables top-N DEG selection to match protocols from papers like scGen (top-100) or GEARS (top-20). For Sinkhorn approximations, `blur` sets the regularization strength. This explicitness enables reproducibility and fair comparison.

Universal Space Support. All metrics—distributional, correlation, and reconstruction—support all three computation spaces through a unified interface. A single call can evaluate in multiple spaces simultaneously, generating a comprehensive evaluation report. Internally, *GGE* manages PCA fitting, DEG computation, and space transformations transparently.

4.2 API DESIGN

The core *GGE* API makes the choice of computation space an explicit part of metric specification, while preserving a simple, unified interface for end users. Each metric is instantiated together with the representation space in which it is evaluated (e.g. PCA space for distributional metrics or DEG-restricted space for correlation-based metrics), along with any method-specific hyperparameters such as the number of principal components or DEG selection thresholds.

Evaluation is performed through a single high-level entry point that loads real and generated data lazily, computes metrics independently for each experimental condition (e.g. cell type and perturbation), and aggregates results across the dataset. This design enables stratified analysis of model performance across biological contexts while avoiding unnecessary memory overhead for large single-cell datasets. The API returns both per-condition scores and dataset-level summary statistics (mean and standard deviation across conditions), facilitating systematic comparison of generative models under heterogeneous perturbation regimes.

Concrete usage examples illustrating typical evaluation configurations are provided in the official documentation (https://docs.andrearubbi.com/gge).

4.3 DEG-SPACE EVALUATION WITH PERTURBATION EFFECTS

For perturbation prediction tasks, a subtle but critical issue arises when computing correlation metrics on expression means. If the control and perturbed conditions have similar mean expression levels, correlation on raw expression will be artificially high regardless of whether the model captures perturbation effects. This occurs because the correlation is dominated by genes that are similarly expressed across conditions, obscuring differences in the genes that actually respond to perturbation.

GGE addresses this through *perturbation-effect correlation*. Instead of correlating raw expression means, we compute:

$$\rho_{\text{effect}} = \text{corr}(\mu_{\text{real}} - \mu_{\text{ctrl}}, \mu_{\text{gen}} - \mu_{\text{ctrl}}) \tag{1}$$

where μ_{real} is the mean expression of real perturbed cells, μ_{gen} is the mean expression of generated cells, and μ_{ctrl} is the mean expression of matched control cells. This measures whether the model captures the *direction and magnitude* of perturbation effects—the signal of biological interest.

When using DEG-space correlation, *GGE* first identifies DEGs between control and perturbed conditions using a Wilcoxon rank-sum test (or alternative methods), then computes perturbation-effect correlation on the subset of significant genes. This combination focuses evaluation on the biologically relevant signal while properly measuring perturbation effects.

4.4 CONDITION-AWARE EVALUATION

Single-cell perturbation datasets typically contain multiple cell types and perturbation conditions, each potentially exhibiting distinct DEG sets and response magnitudes. *GGE* evaluates metrics per condition (cell type $\times$ perturbation pair), enabling stratified analysis. This design choice surfaces heterogeneity across conditions that aggregate metrics would obscure.

For DEG-space metrics, this condition-aware design is essential: DEGs are computed separately for each condition based on comparison to matched control cells. A gene may be a DEG for perturbation A in cell type X but not for perturbation B in cell type Y. *GGE* automatically handles this complexity, computing condition-specific DEG sets and applying them to the appropriate subsets of data.

5 EXPERIMENTS

We present experiments demonstrating the importance of standardization and showcasing *GGE*'s capabilities across multiple benchmarks.

5.1 THE IMPORTANCE OF STANDARDIZATION

Our central claim is that metric values depend critically on implementation choices. To demonstrate this concretely, we compute distributional metrics on identical data under varying configurations. We benchmark on the Norman dataset (39k cells $\times$ 2000 genes in the test set, 138 perturbation conditions) against generated data comprising 1,000 cells per condition (138,000 cells in total) using a recent flow-matching model (MixFlow (Rubbi et al., 2026)), then systematically vary the computation space and measure resulting metric values.

Table 2: Distributional metric values under different computation spaces. The same data produces $\sim$5–10$\times$ different metric values depending on space and dimensionality. All metrics computed using *GGE* v0.6.2 defaults on MPS(M3).

Configuration	W_2	Energy	Time (s)
Raw ($G = 2000$)	104.3 ± 0.3	0.65 ± 0.01	1.9
PCA-100	53.8 ± 0.1	0.20 ± 0.00	3.0
PCA-50	33.6 ± 0.1	0.09 ± 0.00	2.7
PCA-25	17.2 ± 0.1	0.09 ± 0.00	2.4

Table 2 reports remarkable, yet expected, differences: W_2 distance varies by nearly an order of magnitude (17.2 to 104.3) depending solely on computation space. This variation is not random noise—it reflects the mathematical reality that distance metrics scale with dimensionality and feature selection. A paper reporting "$W_2 = 17.2$" computed in PCA-50 space cannot be compared to another reporting "$W_2 = 104.3$" in raw space, yet both might be described simply as "Wasserstein distance" without further specification.

5.2 ABLATION: EFFECT OF DEG THRESHOLDS

For DEG-space metrics, the choice of significance thresholds affects both the number of genes included and the resulting metric values. *GGE* supports both threshold-based DEG selection and top-N selection (e.g., top-100 DEGs as used by scGen, or top-20 as used by GEARS).

Table 3: Effect of DEG selection strategy on correlation metrics. Top-N selection provides consistent gene counts across conditions; threshold-based selection adapts to perturbation strength. We benchmark on the Norman dataset (39k cells $\times$ 5000 genes) with 138 perturbation conditions.

Selection Strategy	Avg # DEGs	Pearson	Spearman
Top-20 (like GEARS)	20	0.614 ± 0.066	0.600 ± 0.043
Top-100 (like scGen)	100	0.594 ± 0.024	0.619 ± 0.041
Strict (lfc>1, p<0.01)	15.3 ± 5.1	0.506 ± 0.217	0.370 ± 0.182
Relaxed (lfc>0.25, p<0.1)	71.7 ± 6.9	0.622 ± 0.079	0.609 ± 0.048

The choice between top-N and threshold-based selection involves a trade-off: top-N selection ensures consistent gene counts across conditions, enabling fair comparison, while threshold-based selection adapts to the biological signal strength of each perturbation. *GGE* supports both strategies through the `n_top_degs` parameter in `DEGSettings`. Table 3 reports the results of our ablation study.

6 COMPARISON WITH CELL-EVAL

The Arc Institute[1] recently released `cell-eval` as part of the STATE framework (Adduri et al., 2025), representing an important step toward standardized evaluation of single-cell generative models. While `cell-eval` provides a comprehensive benchmarking pipeline tailored to the STATE ecosystem, it makes several design choices that trade flexibility and transparency for workflow specialization.

In contrast, *GGE* is designed as a lightweight, model-agnostic evaluation layer that emphasizes (i) explicit control over computation space (raw, PCA, DEG-restricted), (ii) transparent metric configuration, and (iii) ease of integration into existing research pipelines. These differences reflect distinct design goals: `cell-eval` optimizes for a specific large-scale benchmarking workflow, whereas *GGE* prioritizes generality and reproducibility across heterogeneous experimental setups.

A detailed, implementation-level comparison is provided in Appendix C. Table 4 summarizes the main design differences. We emphasize that this comparison is conceptual rather than an empirical head-to-head benchmark, which we leave for future work.

6.1 LIMITATIONS AND FUTURE DIRECTIONS

GGE focuses on static evaluation of generated samples against ground truth. Future work could extend to temporal evaluation (trajectory fidelity), multi-modal evaluation, and counterfactual evaluation where no ground truth exists. Additionally, the choice of biologically meaningful metrics remains an active research area; while perturbation-effect correlation captures one aspect of biological relevance, other metrics measuring pathway enrichment or regulatory network fidelity may provide complementary information.

Finally, beyond metrics alone, standardised benchmarking also requires standardised datasets and splits. Integrating *GGE* with emerging community benchmarking suites (e.g., the curated perturbation datasets and predefined splits in `pertpy` (Heumos et al., 2025)) would enable fully reproducible, end-to-end evaluation pipelines, facilitating fair comparison across models and accelerating method development.

6.2 CONCLUSION

We have presented *GGE*, a standardized framework for evaluating gene expression generative models. Through a comprehensive survey of the literature, we identified that current evaluation practices are heterogeneous to the point of incomparability. Through theoretical analysis, we showed why metric values depend on computation space. Through experiments, we demonstrated that these differences are substantial and that standardization is essential for fair comparison. *GGE* provides a unified API with explicit configuration, hardware acceleration, and biologically-motivated DEG-space evaluation, enabling the field to move toward reproducible and comparable benchmarking.

CODE AVAILABILITY

GGE is available as an open-source Python package (https://pypi.org/project/gge-eval/), installable via `pip`, with the full source code hosted on GitHub (https://github.com/AndreaRubbi/GGE). Documentation is available at https://docs.andrearubbi.com/gge.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv*,

[1] https://arcinstitute.org/

2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1. Preprint.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *Proceedings of ICML*, 2017.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge–kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002. URL https://doi.org/10.1007/s002110050002.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414, 2020.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999. doi: 10.1007/3-540-49257-7_15. URL https://link.springer.com/chapter/10.1007/3-540-49257-7_15.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Saez-Rodriguez Del Castillo, Mitchell Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, 2013.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander J Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Lars Heumos, Yuhan Ji, Lukas May, et al. Pertpy: an end-to-end framework for perturbation analysis. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02909-7. URL https://doi.org/10.1038/s41592-025-02909-7.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrowskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1. Preprint.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2203–2213, 2017. doi: 10.48550/arXiv.1705.08584. URL https://arxiv.org/abs/1705.08584.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Brahim, Junyue Cao, Fabian J Theis, et al. Compositional perturbation autoencoder for single-cell response modeling. *Molecular Systems Biology*, 19(6):e11517, 2023.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746. URL https://www.embopress.org/doi/full/10.15252/msb.20188746.

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian Theis. Multi-modal and multi-attribute generation of single cells with CFGen. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2407.11734. URL https://arxiv.org/abs/2407.11734.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Marco Jost, Thomas M. Norman, Jonathan S. Weissman, et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.06.038. URL https://doi.org/10.1016/j.cell.2022.06.038.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Andrea Rubbi, Amir Akbarnejad, Mohammad Vali Sanian, Aryan Yazdan Parast, Hesam Asadollahzadeh, Arian Amani, Naveed Akhtar, Sarah Cooper, Andrew Bassett, Pietro Liò, Lassi Paavolainen, Sattar Vakili, and Mo Lotfollahi. Mixflow: Mixture-conditioned flow matching for out-of-distribution generalization. *arXiv preprint arXiv:2601.11827*, 2026. doi: 10.48550/arXiv.2601.11827. URL https://arxiv.org/abs/2601.11827.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berber, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Valentine Svensson, Kedar Natarajan, Long H. Ly, et al. Power analysis of single-cell rna-sequencing experiments. *Nature Methods*, 14(4):381–387, 2017. doi: 10.1038/nmeth.4220. URL https://doi.org/10.1038/nmeth.4220.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Wenzhuo Tang, Renming Liu, Hongzhi Wen, Xinnan Dai, Jiayuan Ding, Hang Li, Wenqi Fan, Yuying Xie, and Jiliang Tang. A general single-cell analysis framework via conditional diffusion generative models. *bioRxiv*, 2023. doi: 10.1101/2023.10.13.562243. URL https://www.biorxiv.org/content/10.1101/2023.10.13.562243v1. Preprint.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Cédric Villani. Optimal transport: old and new. *Springer Science & Business Media*, 338, 2008.

Grace Hui Ting Yeo, Sachit D Saksena, Sandeep S Bhate, Ling-Wei Wang, Akiko Marubeni, Jenna L Ruhl, Howard Y Chang, Li Shen, Jure Leskovec, Bradley E Bernstein, et al. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3510, 2021.

A Detailed landscape of generative models

A.1 Generative Models for Single-Cell Data

The landscape of generative models for single-cell gene expression has evolved substantially over the past five years, with distinct architectural paradigms emerging for different biological applications. Understanding this evolution provides essential context for the evaluation challenges that *GGE* addresses.

Variational Autoencoders (VAEs). VAEs established the foundation for single-cell generative modeling. scVI (Lopez et al., 2018) introduced deep generative models for single-cell data, using a VAE architecture with negative binomial likelihood to model count data. scGen (Lotfollahi et al., 2019) extended this approach to perturbation prediction by learning latent arithmetic operations that map control cells to perturbed states. CPA (Lotfollahi et al., 2023) further generalized to combinatorial perturbations through disentangled representations of cell state and perturbation effects. These VAE-based approaches typically evaluate using reconstruction MSE and R^2 between predicted and observed expression, computed either on all genes or a subset of highly variable genes.

Graph Neural Networks (GNNs). GNNs brought biological knowledge into model architectures. GEARS (Roohani et al., 2024) pioneered using gene regulatory networks as inductive biases, enabling prediction of unseen multi-gene perturbation combinations. By passing messages along known gene–gene interactions from STRING and co-expression networks, GEARS demonstrated that structure-aware models generalize better to novel perturbations. However, GEARS introduced a new evaluation metric—the fraction of DEGs correctly predicted in the top-20—that is not directly comparable to distributional metrics used by other methods.

Optimal Transport (OT). OT methods provided principled frameworks for modeling cellular transformations. CellOT (Bunne et al., 2023) formulated perturbation response as an optimal transport problem, learning a transport map that minimizes the 2-Wasserstein distance between control and perturbed distributions. This represented a shift from predicting mean expression to modeling full distributions. However, the reported "Wasserstein distance" values depend critically on whether computation occurs in raw gene space, PCA space, or on a subset of genes.

Flow Matching. Flow matching has recently emerged as a powerful paradigm for modeling distribution-to-distribution transformations. Conditional Flow Matching (Lipman et al., 2022) and minibatch optimal transport variants (Tong et al., 2023) learn continuous velocity fields that transport samples between source and target distributions. Flow-based approaches explicitly model full conditional distributions $p(x_1 \mid x_0, c)$ rather than only conditional means, making distributional metrics central to evaluation. Recent work such as CellFlow (Klein et al., 2025) emphasizes careful benchmarking and protocol design, yet reported gains remain difficult to compare across studies due to non-standardized evaluation pipelines.

Transformer-based foundation models. Transformer-based foundation models represent a complementary and increasingly influential direction. STATE (Adduri et al., 2025) from the Arc Institute introduces a large-scale transformer architecture for learning cellular dynamics and perturbation responses in a unified framework. Rather than learning explicit transport maps or continuous flows, STATE models transitions between cellular states using attention-based sequence modeling in learned latent spaces, enabling multi-task generalization across perturbations, trajectories, and experimental contexts. These models emphasize scalability, representation learning, and transfer across biological regimes, but introduce additional evaluation practices and metrics that are not directly aligned with those used in VAE-, OT-, or flow-based approaches.

B Distributional Metrics: Formal Definitions and Computation

B.1 Optimal Transport and Wasserstein Distances

Optimal transport provides a principled framework for comparing distributions by computing the minimum cost of transforming one distribution into another (Villani, 2008; Peyré & Cuturi, 2019).

The p-Wasserstein distance between distributions P and Q over $\mathbb{R}^d$ is defined as:

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int \|x - y\|^p \, d\gamma(x, y) \right)^{1/p}, \tag{2}$$

where $\Gamma(P, Q)$ denotes the set of couplings with marginals P and Q.

The 1-Wasserstein distance (W_1), also known as the Earth Mover's Distance, admits efficient computation in low dimensions (Rubner et al., 2000). The 2-Wasserstein distance (W_2) is particularly relevant for continuous-time generative modeling due to its connection to the Benamou–Brenier formulation of optimal transport as a fluid dynamics problem (Benamou & Brenier, 2000), which underlies recent flow matching and diffusion-based models (Lipman et al., 2022; Tong et al., 2023).

For scalable computation, we employ entropy-regularized OT via the Sinkhorn algorithm (Cuturi, 2013):

$$W_{p,\epsilon}(P, Q) = \left(\min_{\gamma \in \Gamma(P,Q)} \int \|x - y\|^p \, d\gamma(x, y) + \epsilon H(\gamma) \right)^{1/p}, \tag{3}$$

where $H(\gamma)$ denotes the entropy of the coupling. The regularization parameter ϵ controls the bias–variance trade-off: larger ϵ yields smoother but more biased estimates, while smaller ϵ approximates true OT at higher computational cost and sensitivity to noise. In *GGE*, we use $\epsilon = 0.05$ by default, following common practice in large-scale OT-based evaluation (Feydy et al., 2019).

B.2 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) measures distributional discrepancy via kernel mean embeddings in a reproducing kernel Hilbert space (Gretton et al., 2012):

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P}[k(x, x')] - 2\mathbb{E}_{x \sim P, y \sim Q}[k(x, y)] + \mathbb{E}_{y,y' \sim Q}[k(y, y')], \tag{4}$$

where k is a characteristic positive-definite kernel. We use the Gaussian RBF kernel with bandwidth selected via the median heuristic (Gretton et al., 2012). MMD provides a consistent, nonparametric two-sample test without explicit density estimation and is widely used for benchmarking generative models (Li et al., 2017).

B.3 Energy Distance

Energy distance provides an alternative metric grounded in statistical energy theory (Székely & Rizzo, 2013):

$$E(P, Q) = 2\mathbb{E}_{x \sim P, y \sim Q}[\|x - y\|] - \mathbb{E}_{x,x' \sim P}[\|x - x'\|] - \mathbb{E}_{y,y' \sim Q}[\|y - y'\|]. \tag{5}$$

Energy distance equals zero if and only if $P = Q$, making it a proper metric on probability distributions. While it lacks the explicit geometric interpretation of OT, it provides a computationally efficient and statistically well-founded measure of global distributional similarity.

C Detailed Comparison with cell-eval

We provide a more granular comparison between *GGE* and the `cell-eval` framework (Adduri et al., 2025), focusing on API design, configurability, extensibility, and documentation.

API Design and Workflow. `cell-eval` adopts a multi-stage CLI-based workflow involving explicit preprocessing, execution, baseline generation, and scoring steps. While this design supports reproducible large-scale benchmarking, it requires users to manage intermediate files, directory structures, and pipeline orchestration. By contrast, *GGE* exposes a single high-level Python entry point that performs lazy data loading, per-condition metric computation, and aggregation within a unified interface, facilitating integration into custom training and evaluation loops.

Configuration Transparency. `cell-eval` relies on registry-based "profiles" (e.g. `full`, `minimal`, `vcc`) that bundle multiple metrics and preprocessing choices. While convenient for standardized benchmarks, these profiles obscure which metrics and spaces are being used without inspecting source code. In *GGE*, all metric choices and hyperparameters are specified explicitly via typed metric objects with documented arguments.

Control over Computation Space. `cell-eval` computes metrics in a fixed representation space, without user-level control over PCA dimensionality or DEG selection thresholds. In contrast, *GGE* treats representation space as a first-class parameter of each metric (e.g. raw gene space, PCA space with configurable dimensionality, or DEG-restricted space with tunable thresholds), enabling systematic sensitivity analyses.

Extensibility. Extending `cell-eval` with new metrics requires interacting with internal registries and enumerated types. *GGE* exposes a minimal abstract metric interface, allowing custom metrics to be implemented with minimal boilerplate, lowering the barrier to methodological extensions.

Documentation and Usability. `cell-eval` primarily documents usage through CLI help text and examples embedded in the STATE codebase. *GGE* provides a standalone documentation website with API references, tutorials, and reproducible examples.

Table 4: Comparison of *GGE* and `cell-eval` design choices.

Feature	**GGE**	**cell-eval**
Python API	✓ (single call)	✓ (multi-step)
CLI	✓	✓
Explicit space control	✓	—
Configurable DEG thresholds	✓	—
Top-N DEG selection	✓	—
GPU acceleration (CUDA)	✓	—
Apple MPS acceleration	✓	—
Extensive documentation	✓	Limited
Simple custom metrics	✓	Registry pattern
Visualization module	✓	—

D USE OF LARGE LANGUAGE MODELS

Large Language Models were used as writing assistants to improve clarity and presentation of the manuscript and software documentation. All scientific ideas, methods, experiments, and conclusions are solely those of the authors, and all LLM-assisted content was reviewed and edited for correctness prior to inclusion.