WHAT MAKES A GENERATED CELL REALISTIC? DISSECTING BIOLOGICAL FIDELITY IN SINGLE-CELL GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell generative models are increasingly used to synthesize cellular profiles and predict unobserved biological states, but defining what makes a generated cell realistic remains an open problem. Existing evaluations often collapse realism into global statistical similarity, low-dimensional overlap, or classifier agreement, even though a generated cell may appear realistic in one sense while being biologically wrong in another. We propose CellFidelity, a multidimensional diagnostic framework that decomposes biological realism into four complementary dimensions: lexical fidelity of gene usage and expression magnitude, syntactic fidelity of gene relationships and functional programs, semantic fidelity of conditional cellular identity and identity-associated programs, and population fidelity of diversity and organization. CellFidelity operationalizes these dimensions using reference-calibrated measurements that distinguish deviations from the empirical variation observed among real cells. Across representative single-cell generators, CellFidelity exposes failure modes hidden by aggregate metrics: strong marker enrichment despite incomplete identity programs, plausible pathway activity despite sparse constituent-gene recovery, and globally similar populations despite distorted heterogeneity. These results show that generated-cell realism cannot be captured by a single notion of similarity. CellFidelity provides a diagnostic framework for revealing model-specific biological failures, guiding model development, and supporting more principled use of generated cells in downstream biology.

1 INTRODUCTION

Single-cell generative models are increasingly used to synthesize cellular states that are difficult, costly, or impossible to observe directly, with applications ranging from data augmentation (Lopez et al., 2018; Luo et al., 2024; Palma et al., 2025; Rizvi et al., 2026; Luo et al., 2026b), to predicting perturbation responses (Li et al., 2026; He et al., 2026; Zhang et al., 2026; Wang et al., 2026) and modeling cellular transitions across conditions and time (Yeo et al., 2021; Luo et al., 2026a). Generated cells are increasingly used as computational surrogates for unobserved experimental states, so errors in their expression profiles can propagate into differential expression, pathway analysis, cell-state characterization, and downstream biological conclusions. This makes a fundamental question increasingly important: **what makes a generated cell realistic?** Yet, the notion of biological realism itself remains poorly defined.

Current evaluations often reduce realism to a relatively small set of statistical criteria, including reconstruction error, correlation, distributional distances, classifier performance, or overlap in low-dimensional embeddings (Wei et al., 2026; Rubbi et al., 2026; Karkar & De Clerck, 2026). These measures are useful for detecting broad discrepancies, but provide only fragmented views of biological realism while implicitly treating it as a largely unified property. Recent biology-aware evaluations have broadened this scope to differential expression, pathways, co-expression, and perturbation-specific signals (Radig et al., 2026; Fu et al., 2026), but these readouts are still typically treated as separate endpoints rather than as complementary dimensions of biological fidelity. In practice, different aspects of a generated cell can agree with real biology to very different degrees. A profile may contain highly characteristic genes while omitting much of the broader transcriptional program; reproduce plausible pathway activity without reconstructing the coordinated genes that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

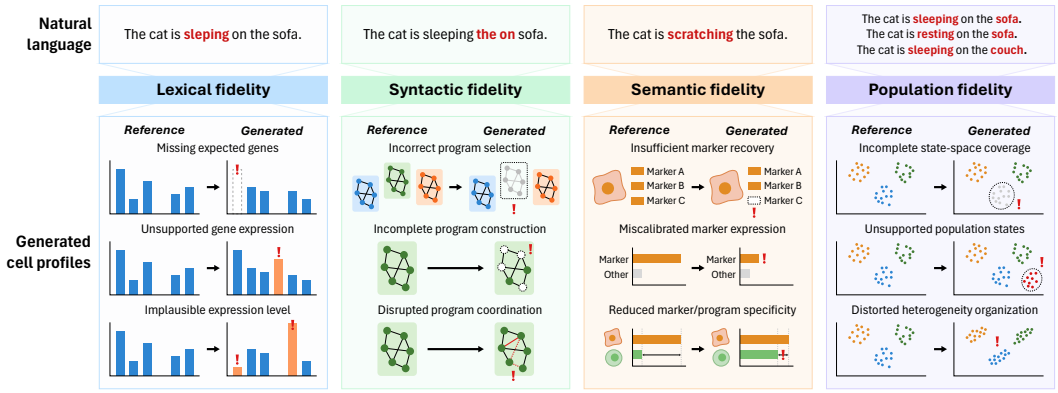

Figure 1: Overview of CellFidelity. The framework uses a natural-language analogy to formulate generated-cell realism across four complementary dimensions, including lexical, syntactic, semantic, and population fidelity.

give rise to it; or appear individually convincing while collectively collapsing the heterogeneity of real populations. Thus, generated cells can appear realistic for fundamentally different reasons, and can remain biologically flawed even when aggregate similarity is strong.

We draw an analogy to natural language, where realism is not a single property. Lexical, syntactic, and semantic correctness are related but distinct: a sentence may contain misspelled or inappropriate words yet preserve its intended meaning, while another may be lexically and grammatically flawless but convey the wrong message. Likewise, a generated cell may be faithful in some biological respects while failing in others. We therefore introduce CellFidelity to evaluate generated cells along four complementary dimensions: **lexical fidelity** of gene usage and expression magnitude, **syntactic fidelity** of gene relationships and functional programs, **semantic fidelity** of cellular identity and associated functions, and **population fidelity** of diversity and organization (Figure 1). Importantly, higher values on seemingly desirable metrics do not always indicate greater biological fidelity. Unusually strong marker enrichment, for example, may reflect concentration on a small set of canonical markers rather than faithful reconstruction of the broader cellular program. CellFidelity therefore evaluates generated cells against the empirical range of real-cell variation, so that deviations in either direction can indicate biological miscalibration.

We apply CellFidelity to representative single-cell generative models spanning distinct modeling paradigms and uncover markedly different fidelity profiles. Models that appear similar under aggregate metrics can fail in different biological dimensions: strong marker enrichment can coexist with incomplete identity programs, plausible pathway activity can be supported by only a small subset of pathway genes, and globally similar populations can still exhibit distorted heterogeneity. Together, these results establish generated-cell realism as a multidimensional property rather than a single notion of similarity. By making these dimensions explicit and measurable, CellFidelity provides a principled framework for evaluating generated cells and diagnosing model-specific failure modes, while offering concrete directions for the development of more biologically faithful single-cell generative models.

2 Related Work

2.1 Generative Modeling of Single-Cell Cellular States

Single-cell generative modeling has expanded from unconditional transcriptome synthesis to increasingly diverse forms of conditional and structured generation. Recent approaches span probabilistic latent-variable models, flow- and diffusion-based methods, and tokenized or language-like representations, and differ substantially in output space, gene coverage, modeling assumptions, and intended use (Luo et al., 2024; 2026b; Palma et al., 2025; Levine et al., 2024; Xiao et al., 2025). Beyond static transcriptome generation, these models are increasingly used for perturbation-response

prediction and temporal state forecasting (Li et al., 2026; He et al., 2026; Zhang et al., 2026; Wang et al., 2026; Yeo et al., 2021; Luo et al., 2026a). This growing diversity creates increasingly heterogeneous requirements for evaluating generated cells.

2.2 Evaluation of generated single-cell data

Evaluation of generated single-cell data has traditionally relied on statistical agreement between generated and real populations. Common metrics include reconstruction errors and correlations, agreement of gene-wise means and variances, distributional distances such as Wasserstein distance and maximum mean discrepancy, classifier-based distinguishability, and differential-expression consistency (Cao et al., 2021; Wei et al., 2026; Rubbi et al., 2026; Karkar & De Clerck, 2026). These metrics support standardized comparison and detect large discrepancies, but provide only a partial view of biological realism and offer limited resolution on which biological properties are distorted.

Recent studies have further exposed limitations of aggregate evaluation. Standard perturbation metrics can be dominated by systematic variation, evaluation design can inflate differential-expression agreement (Viñas Torné et al., 2026; Nicol et al., 2026), and strong global similarity does not necessarily imply recovery of perturbation-specific biological signals (Radig et al., 2026). These observations have motivated more biology-aware evaluations, such as pathway activity, gene–gene co-expression, and cell state transition (Li et al., 2024; Radig et al., 2026; Fu et al., 2026).

Despite this progress, these biological properties are still typically evaluated as separate endpoints rather than organized into complementary dimensions of realism. Moreover, apparently favorable metric values can themselves reflect distortion. This motivates an evaluation framework that decomposes biological fidelity and calibrates each dimension against variation observed in real cells.

3 Methods

3.1 CellFidelity: reference-calibrated decomposition of generated-cell realism

Let $\mathcal{D}^{\mathrm{ref}}$ denote a real-cell reference population and $\mathcal{D}^{\mathrm{gen}}$ generated cells under matched biological conditions. Rather than representing realism by a single similarity score, CellFidelity characterizes generated-cell realism as a multidimensional fidelity profile,

$$\mathcal{F}(\mathcal{D}^{\mathrm{gen}}; \mathcal{D}^{\mathrm{ref}}) = (\mathcal{F}_{\mathrm{lex}}, \mathcal{F}_{\mathrm{syn}}, \mathcal{F}_{\mathrm{sem}}, \mathcal{F}_{\mathrm{pop}}),$$

corresponding to lexical, syntactic, semantic, and population fidelity.

A central principle of CellFidelity is reference calibration. For each metric, we estimate the corresponding empirical behavior from matched real cells and interpret generated-cell values relative to this reference rather than assuming that larger or smaller values are intrinsically preferable. A subset of the reference pool is processed through the same evaluation pipeline as generated cells to define an internal real-validation group and an empirical calibration range. A separate non-overlapping real-cell control is used to assess the stability of these calibration patterns.

3.2 Lexical fidelity: gene usage and expression calibration

Lexical fidelity evaluates whether a generated cell expresses an appropriate set of genes at quantitatively plausible levels. Let X_{ig} denote library-size-normalized log-transformed expression; a gene is detected when $X_{ig} > 0$. We evaluate complementary properties: **transcriptomic complexity**, the number of detected genes; **expected-gene recall**, the recovery of genes recurrently detected in matched real cells; **expressed-gene support**, the mean reference detection rate of genes expressed by the evaluated cell; **fraction of infrequent genes in top 200 genes**; and **expression-magnitude calibration**.

For expression-magnitude calibration, each positive expression value is mapped to its matched gene- and cell-type-specific positive-expression reference distribution,

$$Q_{ig} = \widehat{F}^{+}_{gc}(X_{ig}).$$

Values near the reference median indicate typical expression magnitude, whereas systematic shifts toward either tail indicate attenuation or inflation. Together, these metrics distinguish transcriptome under-coverage, selective concentration on highly recurrent genes, expression of infrequently detected genes, and quantitative expression miscalibration. Full metric definitions are provided in Appendix C.1.

3.3 Syntactic fidelity: organization of functional programs

Syntactic fidelity evaluates whether plausible genes are assembled into biologically supported, sufficiently complete, and internally coordinated functional programs. We assess three properties: **program selection**, **program construction**, and **program coordination**.

Functional programs are defined from Gene Ontology gene sets. For program p with member genes G_p and expression-matched controls C_p, activity in cell i is

$$S_{ip} = \frac{1}{|G_p|} \sum_{g \in G_p} X_{ig} - \frac{1}{|C_p|} \sum_{g \in C_p} X_{ig}.$$

A program is active when $S_{ip} > 0$. Program selection measures whether activated programs are supported by matched real cells and whether recurrent reference programs are recovered. We also measure the fraction of programs active in each evaluated cell across bins of reference activity rate. This distinguishes activation of infrequent programs from failure to activate recurrent programs.

Program construction is evaluated from two complementary quantities: **member-gene recovery rate**, which measures recovery of constituent genes relative to matched real cells, and **member-gene expression magnitude**, which measures whether recovered members are expressed at reference-consistent magnitudes. This separates broad reconstruction of a program from apparently strong activity driven by only a few highly expressed members. Program coordination compares within-program gene-gene correlation structures between evaluated and reference populations, capturing both preservation of relative correlation patterns and absolute correlation errors. Detailed definitions are provided in Appendix C.2.

3.4 Semantic fidelity: preservation of cellular identity

Semantic fidelity evaluates whether generated cells preserve the biological identity specified by their conditioning labels. We assess identity at two levels. At the **marker level**, reference-derived cell-type markers are evaluated for enrichment among highly expressed genes, detection and expression calibration, and preservation of target-versus-nontarget specificity.

For marker g associated with target cell type c, we define its identity contrast as

$$\Delta_{gc}^{\text{marker}} = \overline{X}_{gc}^{\text{target}} - \overline{X}_{gc}^{\text{non-target}},$$

where the non-target group pools the other evaluated cell types.

At the **program level**, we analogously evaluate whether cell-type-associated functional programs preserve their expected target-versus-nontarget activity contrast. This tests whether identity is represented by broader distributed transcriptional programs rather than by a restricted set of canonical markers. Marker selection and full metric definitions are provided in Appendix C.3.

3.5 Population fidelity: preservation of population heterogeneity

Population fidelity evaluates whether generated cells collectively reproduce the distribution and heterogeneity of matched real-cell populations. We assess complementary properties: **global distributional similarity** using MMD; **manifold precision and coverage**, measuring unsupported generated states and missing reference states; **effective dimensionality**; **relative reference-PCA reconstruction residual**; and **gene-wise variance organization**.

MMD, precision, coverage, and effective dimensionality are evaluated in a fixed reference-derived PCA space using repeated equal-size subsampling. Effective dimensionality is quantified by the

entropy-based effective rank

$$r_{\text{eff}} = \exp\left(-\sum_{j:p_j>0} p_j \log p_j\right), \qquad p_j = \frac{\lambda_j}{\sum_k \lambda_k},$$

and compared between evaluated and reference populations. Normalized covariance-eigenvalue spectra additionally describe how variation is distributed across the fixed PCA space.

To assess variation outside the retained reference subspace, let $\boldsymbol{\mu}_{\text{ref}}$ the full-reference mean, and $U \in \mathbb{R}^{G\times 50}$ the matrix whose columns are the 50 fixed orthonormal reference principal components. The relative reference-PCA reconstruction residual was

$$r_i = \frac{\left\|(I - UU^\top)(\mathbf{x}_i - \boldsymbol{\mu}_{\text{ref}})\right\|_2^2}{\max\left(\|\mathbf{x}_i - \boldsymbol{\mu}_{\text{ref}}\|_2^2, 10^{-12}\right)}.$$

Gene-wise heterogeneity is assessed from both the correlation of reference and evaluated gene variances and their systematic log-scale bias. Together, these metrics distinguish global distributional mismatch, manifold under-coverage, dimensional collapse, altered residual variation, and distorted organization of cellular variability. Full implementation details are provided in Appendix C.4.

4 Results

4.1 Experimental Setup

We constructed a 396,972-cell real-cell reference from CELLxGENE and harmonized database-provided annotations into 18 primary cell categories. After removing genes sparsely detected across all reference cell types, 18,127 genes were retained for evaluation. 20% of the reference pool was designated as an internal real-validation subset for empirical calibration, while a separate non-overlapping set of 19,820 CELLxGENE cells served as an independent real-cell control.

We evaluated representative generators spanning distinct modeling paradigms: scDiffusion, CFGen, scVI, and C2S. scDiffusion was implemented using a count-output adaptation with classifier-guided latent diffusion and a negative-binomial decoder; CFGen combines latent flow matching with a count-aware decoder; scVI provides a probabilistic latent-variable baseline; and C2S generates rank-based gene lists with a restricted output space. For C2S, we evaluated both the default 200-gene setting and an expanded 1,000-gene setting to examine the effect of output-space restriction. Generator architectures, training procedures, and sampling configurations are described in Appendix B.

All real and generated profiles were mapped to the same evaluation representation using library-size normalization to 10,000 per cell followed by log1p transformation to ensure comparability. Main diagnostic figures focus on B/plasma cells, endothelial cells, and neurons, representing distinct immune, vascular, and tissue-specialized contexts. Reference-derived quantities were fixed before evaluation, and population-estimated metrics used repeated equal-size subsampling where required to control for sample-size effects. Details of data construction and evaluation procedures are provided in Appendix A and Appendix C.

4.2 Lexical fidelity reveals distinct failures in gene usage and expression magnitude

We first evaluated lexical fidelity by examining whether generated cells recovered a realistic breadth of gene expression. CFGen, scDiffusion, and scVI closely matched the real-cell range, whereas C2S showed pronounced transcriptome under-coverage. Both C2S configurations expressed substantially fewer genes per cell (transcriptomic complexity; Fig. 2A) and recovered fewer genes that are recurrently detected in the corresponding cell type (expected-gene recall; Fig. 2B). Expanding the C2S output space from 200 to 1,000 genes partially alleviated both effects.

Expected-gene recall asks how much of the reference-supported transcriptome is recovered; we next examined the complementary question of whether the genes that a model does express are supported by real cells. Despite their low coverage, both C2S variants preferentially expressed genes

with high reference detection rates, particularly under the 200-gene output restriction (Fig. 2C). Expanding the output space shifted this distribution toward the real-cell range, indicating that C2S-200 concentrates expression on a restricted set of highly recurrent genes. scVI showed the converse pattern: despite expressing a near-real number of genes, scVI allocated more expression to genes that were infrequently detected in matched reference cells, pointing to abnormal gene selection rather than transcriptome under-coverage (Fig. 2D).

We next examined expression magnitude by placing each positive expression value on its gene- and cell-type-matched real-cell percentile scale (Fig. 2E). The internal real-validation defined the expected calibration profile, and the independent real-cell control closely reproduced it, supporting the stability of the reference-calibrated scale. CFGen and scDiffusion broadly reproduced the real-cell relationship between gene detection rate and expression magnitude. In contrast, C2S placed its retained genes disproportionately toward the upper tail of their reference distributions, with partial correction in C2S-1000, while scVI showed a milder upward shift in several gene groups.

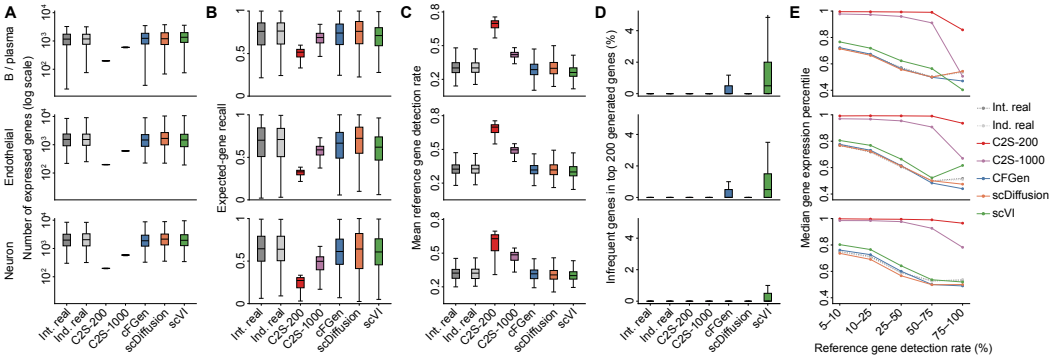

Figure 2: Lexical fidelity of generated cells. (A) Number of detected genes per cell. (B) Expected-gene recall, defined as the fraction of genes detected in at least 50% of matched reference cells that are recovered in each evaluated cell. (C) Mean reference detection rate of genes expressed in each evaluated cell. (D) Percentage of infrequently detected genes (reference detection rate $< 1\%$) among the top 200 expressed genes. (E) Median positive-expression percentile relative to the matched gene- and cell-type-specific reference distribution, stratified by reference detection-rate bins. Outliers are not shown in boxplots. Int. real: internal real; ind. real: independent real.

4.3 Syntactic fidelity reveals distinct failures in functional program selection, construction, and coordination

We next asked whether generated cells reconstructed biologically coherent functional programs. We first examined program selection. Models differed in both the number of active programs and the recovery of programs recurrently active in matched real cells (Fig. S1A–B), as well as in how activation depended on reference activity frequency (Fig. 3A–B). CFGen and scDiffusion were generally close to the real-cell pattern. C2S showed an overall activation deficit, with fewer active programs, lower expected-program recall, and a bias toward programs with high reference active rates. In contrast, scVI showed lower activity specificity, overactivating low-frequency programs while under-recovering highly recurrent ones. Expanding C2S to 1,000 genes partially shifted these properties toward the real-cell range. As in the lexical analysis, we also calibrated positive program scores against their matched program- and cell-type-specific reference distributions (Fig. S1C). Once active, program-level activity was broadly comparable across models and real cells, suggesting that the larger differences lay in program selection rather than aggregate activity magnitude.

We next examined program construction: whether program-level activity was supported by broad recovery of constituent genes or driven by only a small subset of members. For both ubiquitous and cell-type-associated programs, we jointly measured member-gene recovery and the expression magnitude of recovered members (Fig. 3C). scDiffusion remained close to the real-cell range, while CFGen showed mild reductions in member-gene recovery for some cell-type-associated programs. scVI performed well for ubiquitous programs but showed clearer recovery deficits for cell-type-

associated programs. C2S showed the strongest deviation, recovering fewer program members while expressing the retained genes at unusually high levels; expanding to 1,000 genes alleviated, but did not eliminate, this pattern. These results showed that similar program-level activity can arise from substantially different underlying gene configurations.

Finally, we examined program coordination by comparing within-program gene–gene correlation structure (Fig. 3D). The internal and independent real-cell controls occupied a consistent high-similarity, low-error regime, whereas generated models showed distinct deviations. C2S-200 exhibited the strongest coordination defects, and expanding to 1,000 genes reduced error without consistently restoring correlation-pattern similarity. CFGen and scDiffusion were generally closer to the real-cell regime, while scVI showed intermediate, cell-type-dependent deviations especially for cell-type-associated programs. Thus, improved program recovery does not necessarily restore internal gene–gene coordination.

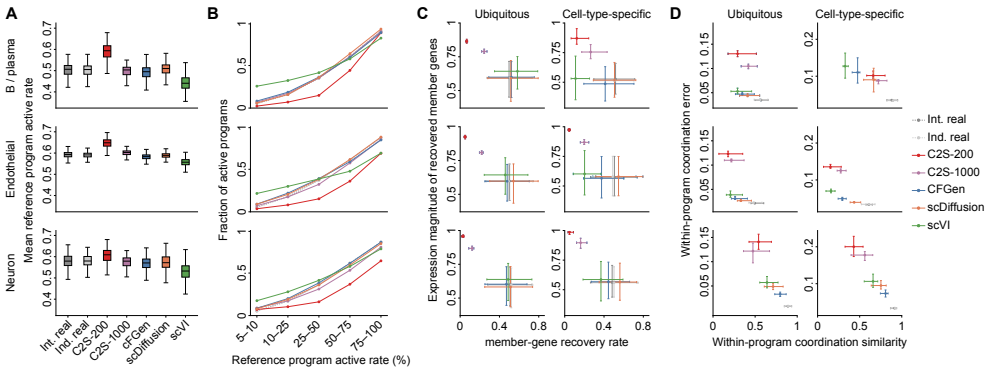

Figure 3: Syntactic fidelity of generated cells. (A) Mean reference active rate of programs activated in each cell, where a program is considered active when its score is > 0. Outliers are not shown. (B) Mean fraction of programs activated in each cell, stratified by their reference active-rate bins. (C) Program construction, jointly showing member-gene recovery and the reference-calibrated expression magnitude of recovered member genes for ubiquitous and cell-type-specific programs. (D) Program coordination, jointly showing similarity to the reference within-program gene–gene correlation structure and coordination error for ubiquitous and cell-type-specific programs. For (C–D), points show medians and bars show interquartile ranges.

4.4 Semantic fidelity reveals distinct failures in cellular identity preservation

We then asked whether generated profiles preserved the cellular identity specified by their conditioning label. We first examined whether highly expressed genes were enriched for canonical cell-type markers by counting marker genes among the top 200 expressed genes (Fig. 4A). scDiffusion and C2S reached or exceeded the real-cell range, whereas CFGen showed moderate attenuation and scVI showed pronounced depletion.

We then decomposed marker fidelity into marker recovery and expression magnitude (Fig. 4B). scDiffusion most closely matched real cells on both measures, whereas C2S recovered relatively few expected markers but expressed those retained at unusually high reference percentiles. Expanding from 200 to 1,000 genes alleviated both effects. CFGen showed milder deviations, while scVI showed clearer losses in both recovery and expression magnitude.

We further examined identity specificity by comparing marker expression in the target cell type with that in non-target cell types (Fig. 4C). scDiffusion closely reproduced the real-cell contrast, CFGen showed moderate attenuation, and scVI substantially compressed cell-type differences. C2S preserved the strongest identity signals but lost many weaker and intermediate markers, with partial recovery at 1,000 genes. We also asked whether these specificity losses extended from canonical markers to broader cell-type-associated functional programs (Fig. 4D). scDiffusion again remained close to the real-cell relationship, whereas scVI showed pronounced compression; C2S and CFGen showed intermediate, program-dependent losses. Together, these results show that strong expression

of a restricted set of canonical markers can coexist with incomplete reconstruction of the broader transcriptional identity of a cell type.

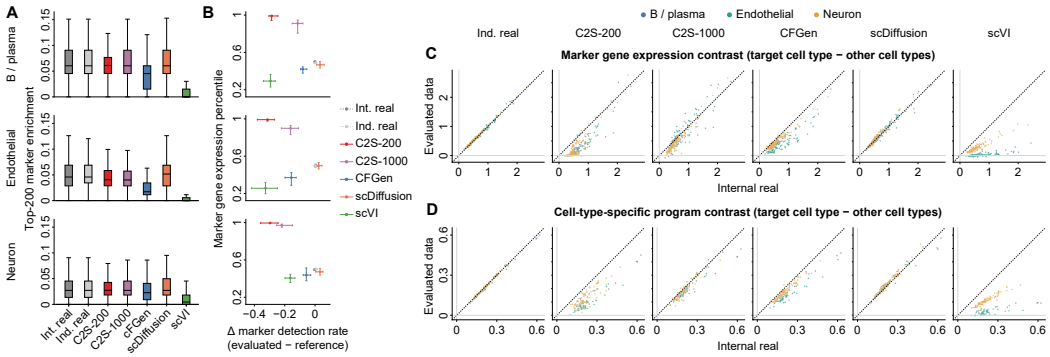

Figure 4: Semantic fidelity of generated cells. (A) Enrichment of cell-type marker genes among each cell's top 200 expressed genes, relative to the marker frequency in the common gene space. (B) Joint calibration of marker recovery and expression magnitude. The x-axis shows the difference in marker detection rate between evaluated and matched reference cells, and the y-axis shows the median positive-expression percentile of detected markers relative to their matched gene- and cell-type-specific reference distributions. Points show medians and bars show interquartile ranges. (C–D) Preservation of cell-type marker specificity (C) and cell-type-associated program specificity (D), measured by the expression/activity contrast between the target cell type and other cell types.

4.5 Population fidelity reveals distinct failures in state-space coverage and heterogeneity organization

Finally, we asked whether generated cells collectively reproduced the distribution and heterogeneity of matched real populations. We first examined state-space reconstruction. In a fixed reference-derived PCA space, MMD revealed substantial differences across generators (Fig. 5A). scDiffusion remained relatively close to the real-cell reference, whereas C2S-200 showed the largest discrepancy; expanding C2S to 1,000 genes markedly reduced this distance.

Global distributional distance does not reveal how a generated population differs from real cells. We therefore used manifold precision and coverage to distinguish two complementary failure modes (Fig. 5B): precision measures whether generated cells fall within regions supported by the reference population, whereas coverage measures how much of the reference state space is represented by generated cells. Real-cell controls were high on both measures. scDiffusion largely preserved precision, with only a modest reduction in coverage. C2S showed clear deficits in both precision and coverage, whereas C2S-1000, CFGen, and scVI occupied an intermediate regime, with both measures improved relative to C2S-200 but still below scDiffusion.

We next examined heterogeneity complexity. Effective dimensionality showed how broadly variation was distributed across reference-defined axes (Fig. 5C). scVI, CFGen, and scDiffusion showed moderate and cell-type-dependent reductions, whereas C2S showed the strongest loss of effective dimensionality, with no consistent improvement after expanding from 200 to 1,000 genes. Reduced effective dimensionality was accompanied by faster decay of higher-order covariance components (Fig. S2A). We additionally quantified the fraction of each cell's expression variation that fell outside the leading reference PCA subspace (PCA reconstruction residual; Fig. S2B). C2S-200 showed markedly less variation in these residual directions than real cells, indicating that its heterogeneity was overly concentrated along the dominant reference axes. In contrast, other generated models, including C2S-1000, often showed excess residual variation, suggesting additional variation outside the major reference-supported directions. Thus, matching the overall amount of variation does not guarantee that heterogeneity is organized along the same directions as in real populations.

Finally, we examined how variation was allocated across genes (Fig. 5D–E), asking both whether the same genes were variable and whether their variance magnitudes matched the reference. C2S showed strong variance compression, only partially alleviated at 1,000 genes. CFGen and scDiffu-

sion more closely preserved both the gene-wise variance pattern and its magnitude. In contrast, scVI maintained overall variance while less faithfully preserving which genes were variable. Together, these results show that population fidelity depends not only on the amount of heterogeneity, but also on whether the correct states are covered and whether variation is organized along the appropriate expression-space directions and genes.

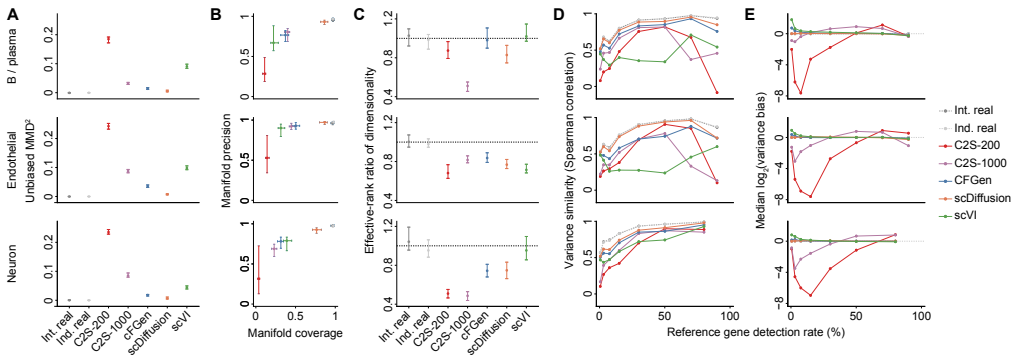

Figure 5: Population fidelity of generated cells. (A) Unbiased MMD^2 computed in a fixed 50-component reference PCA space. (B) k-nearest-neighbor manifold precision and coverage ($k = 5$). (C) Effective-rank ratio defined as the entropy-based effective rank of the generated population relative to that of matched real cells in the reference PCA space. (D) Spearman correlation between generated and reference per-gene variance, stratified by reference gene detection rate. (E) Median $\log_2$ variance bias relative to matched reference cells across the same detection-rate bins. In (A–C), points show medians and bars show empirical 2.5–97.5% intervals across repeats.

5 Discussion and Limitations

Our study frames generated-cell realism as a multidimensional biological property rather than a single notion of statistical similarity. CellFidelity separates fidelity into lexical, syntactic, semantic, and population dimensions, which were only partially coupled across models. A generator could preserve canonical markers while incompletely reconstructing broader identity, reproduce plausible program-level activity without recovering its gene-level structure, or generate individually plausible cells while distorting population heterogeneity. CellFidelity makes these failures explicit and interprets them relative to empirical real-cell variation rather than assuming that larger, smaller, or closer values are intrinsically better. This reference-calibrated view is important because apparent improvements, such as stronger marker enrichment, more expressed genes, or increased variability, can reflect either useful signal enhancement or biological miscalibration.

The C2S results illustrate this diagnostic role. Expanding its output space from 200 to 1,000 genes improved transcriptome, program, and marker coverage, showing that several fidelity deficits partly reflected the restricted output representation. Yet program coordination and population effective dimensionality remained distorted or showed little consistent improvement. CellFidelity can therefore distinguish output-space limitations from biological structures that remain poorly reconstructed after those constraints are relaxed.

This study is a proof of concept and evaluates only a subset of current single-cell generation models and cell types, while our semantic analysis focuses mainly on conditional cell-type identity. Future evaluations could extend the framework to foundation-model-based generators, perturbation models, temporal and within-cell-type state generation, and other generative settings. An important next step is to determine whether specific fidelity dimensions predict the reliability of downstream analyses such as differential expression, pathway inference, or data augmentation. Establishing such links would allow generated cells to be evaluated not only by how closely they resemble real data, but by which aspects of biological structure they preserve well enough to support a given scientific use.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refinement of evaluation metrics and methodology, code development and execution, literature search, manuscript editing, and preparation of the conceptual framework figure. The initial research question, overall study conception, core framework, and principal experimental design were developed by the authors. Generative AI tools were not used to generate or modify the real single-cell reference or validation data. The synthetic single-cell profiles evaluated in this study were generated by the single-cell generative models under investigation and constitute the objects of evaluation. All AI-assisted scientific suggestions and manuscript modifications were reviewed by at least 3 authors. AI-assisted and agent-executed code was manually inspected and checked for consistency with the intended analyses, and AI-assisted literature retrieval was verified against the original sources. We take responsibility for the final content of this work, including all text, methods, analyses, code, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All real single-cell datasets used in this study were obtained from publicly available resources, and were used in de-identified form. We did not recruit human participants, collect new biological specimens, access identifiable personal information, or generate new human-subject data. The synthetic single-cell profiles analyzed in this work were generated computationally from existing data for methodological evaluation.

REPRODUCIBILITY STATEMENT

We provide the main methodological definitions and evaluation procedures in Section 3, with detailed descriptions of dataset construction, generated-cell preparation, preprocessing, implementation settings, and additional computational procedures in Appendices A–C. All real datasets used in this study are publicly available, and their sources and processing procedures are documented. Code for reproducing the evaluation pipeline and analyses will be provided as anonymized supplementary material for review.

REFERENCES

Yue Cao, Pengyi Yang, and Jean Yee Hwa Yang. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nature Communications*, 12:6911, 2021. doi: 10.1038/s41467-021-27130-w.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Xiaotong Fu, Dominik Klein, Egor Antipov, Alessandro Palma, Alejandro Tejada-Lapuerta, Mojtaba Bahrami, Louis B. Kümmerle, Manuel Lubetzki, Francesco Paolo Casale, Malte D. Luecken, and Fabian J. Theis. Benchmarking gene expression reconstruction from single-cell latent representations. *bioRxiv*, 2026. doi: 10.64898/2026.06.15.731445. URL https://www.biorxiv.org/content/10.64898/2026.06.15.731445v1.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi,

and Kam W. Leong. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 23(1):65–77, 2026. doi: 10.1038/s41592-025-02877-y.

Skander Karkar and Olav De Clerck. Conditional single-cell RNA generation: A decoder model and a benchmark of generative models. In *Machine Learning for Genomics Explorations Workshop at ICLR 2026*, 2026. URL https://openreview.net/forum?id=fA0T9ujlQr.

Daniel Levine, Syed A. Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Chen Li, Haoxiang Gao, Yuli She, Haiyang Bian, Qing Chen, Kai Liu, Lei Wei, and Xuegong Zhang. Benchmarking ai models for in silico gene perturbation of cells. *BioRxiv*, pp. 2024–12, 2024.

Chen Li, Lei Wei, and Xuegong Zhang. Unified multimodal learning enables generalized cellular response prediction to diverse perturbations. *bioRxiv*, 2026. doi: 10.1101/2025.11.13.688367.

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. doi: 10.1038/s41592-018-0229-2.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024. doi: 10.1093/bioinformatics/btae518.

Erpai Luo, Haoxiang Gao, Haiyang Bian, Shuguang Peng, Yaru Li, Chen Li, Minsheng Hao, Mo Chen, Yuli She, Lei Wei, Kai Liu, and Xuegong Zhang. Forecasting cell state futures from static snapshots. *bioRxiv*, 2026a. doi: 10.64898/2026.02.08.704720.

Erpai Luo, Lei Wei, Minsheng Hao, Xuegong Zhang, and Qiao Liu. A multi-modal diffusion model with dual-cross-attention for multi-omics data generation and translation. *Nature Communications*, 17:5196, 2026b. doi: 10.1038/s41467-026-71744-x.

Phillip B. Nicol, Shriya Shivakumar, and Rafael A. Irizarry. Spurious correlation inflates performance in single-cell perturbation prediction. *bioRxiv*, 2026. doi: 10.64898/2026.05.07.723486.

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian J. Theis. Multi-modal and multi-attribute generation of single cells with CFGen. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3MnMGLctKb.

Jean Radig, Robin Droit, Daria Doncevic, Albert Li, Duc Thien Bui, Luis Herfurth, Thaddeus Kühn, and Carl Herrmann. scArchon: a scalable benchmarking framework for assessing single-cell perturbation models. *Genome Biology*, 27:162, 2026. doi: 10.1186/s13059-026-04104-z.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Hattie Chung, Rahul M. Dhodapkar, Paul Jaeger, Bryan Perozzi, Jeffrey Ishizuka, Shekoofeh Azizi, and David van Dijk. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, 2026. doi: 10.1101/2025.04.14.648850.

Andrea Rubbi, Andrea Giuseppe Di Francesco, Mohammad Lotfollahi, and Pietro Liò. A standardized framework for evaluating gene expression generative models. *arXiv preprint arXiv:2603.11244*, 2026. URL https://arxiv.org/abs/2603.11244.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Chloe Wang, Mehran Karimzadeh, Neal G. Ravindra, Lexi R. Bounds, Nader Alerasool, Ann C. Huang, Shihao Ma, Daniel R. Gulbranson, Haotian Cui, Yongju Lee, Anusuya Arjavalingam, Elliot J. MacKrell, Matthew S. Wilken, Jieming Chen, Benjamin W. Herken, Jesse A. Weber, Massimo M. Onesto, Barbara Gonzalez-Teran, Nicole F. Leung, Sally Yu Shi, Byron J. Smith, Sharon K. Lam, Adam Barner, Philip Wright, Elizabeth M. Rumsey, Soohong Kim, Rene V. Sit, Adam J. Litterman, Ci Chu, and Bo Wang. X-Cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.18.712807.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Chuxi Xiao, Haiyang Bian, Yixin Chen, Lei Wei, and Xuegong Zhang. CellTok: Early-fusion multimodal large language model for single-cell transcriptomics via tokenization. *bioRxiv*, 2025. doi: 10.1101/2025.10.22.684047.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications*, 12:3222, 2021. doi: 10.1038/s41467-021-23518-w.

Han Zhang, Guo-Hua Yuan, Chaohao Yuan, Tingyang Xu, Tian Bian, Hong Cheng, Wenbing Huang, Deli Zhao, and Yu Rong. Lingshu-Cell: A generative cellular world model for transcriptome modeling toward virtual cells. *arXiv preprint arXiv:2603.25240*, 2026. URL `https://arxiv.org/abs/2603.25240`.

A Data, Preprocessing, and Calibration

The real-cell reference pool contained 396,972 cells sampled from CELLxGENE (whole-dataset snapshot from January 2025; (CZI Cell Science Program et al., 2025)) and an initial expression matrix spanning 61,888 genes. Genes detected in fewer than 5% of cells in every reference cell type were removed, leaving 18,127 genes in the common evaluation space. For the common normalized representation, the filtered count matrix was library-size normalized to 10,000 counts per cell and transformed using $\log(1 + x)$. This representation was used for CellFidelity evaluation and for models requiring normalized expression as input. The corresponding raw counts were retained for models requiring count-based input.

Database-provided annotations were harmonized into 18 primary cell categories. Among the cell types shown in the main figures, the reference pool contained 27,940 B/plasma cells, 13,543 endothelial cells, and 22,309 neurons.

The reference pool was used to estimate gene-, program-, and population-level biological expectations. 20% of the pool was designated as the internal real-validation subset using NumPy `default_rng` and was evaluated through the same pipeline as generated cells to characterize the empirical calibration range of each metric. Because these cells belong to the reference pool, this group serves as a calibration baseline rather than an independent generalization test. A separate set of 19,820 CELLxGENE cells, non-overlapping with the reference pool, was used as an independent real-cell control to assess the stability of the resulting calibration patterns.

B Generator Implementations and Training Details

B.1 Common training corpus and harmonization

All generators were trained on the 396,972-cell reference dataset described in Appendix A. Both raw counts and library-size-normalized log1p expression were retained, allowing each model to use its native input representation. The 18 primary cell types were used as conditioning labels for conditional models, and the 303 CELLxGENE dataset identifiers were used as batch covariates where supported by the respective method.

We did not impose a common training representation across models: depending on the method, inputs consisted of raw counts, log-normalized expression, or ranked gene lists. Unless otherwise stated, training and generation hyperparameters followed the respective official implementations. For CellFidelity evaluation, all generated outputs were mapped to the common 18,127-gene space and converted to the same library-size-normalized log1p representation.

B.2 scDiffusion

We used a count-output adaptation of scDiffusion, retaining the original three-stage latent-diffusion framework. A count-based autoencoder was first trained from scratch to map scRNA-seq profiles into a latent space, using a negative-binomial RNA likelihood following the RNA decoder formulation of scDiffusion-X. The decoder parameterizes the negative-binomial mean as $\mu = s\rho$, where $\rho = \mathrm{softmax}(D(z))$ and s denotes the library-size factor, with gene-specific dispersion parameters learned during training. The diffusion backbone was then trained on the frozen latent representations, and an 18-way cell-type classifier was trained on the same latent space and used for classifier-guided conditional sampling, following the architecture and training configuration of the original scDiffusion implementation. Generated latent representations were obtained by classifier-guided reverse diffusion and decoded by sampling from the negative-binomial distribution, yielding raw integer counts.

B.3 CFGen

CFGen was trained in two stages on log-normalized counts, using cell type and dataset identifier as covariates. The first stage trained a latent autoencoder with a negative-binomial decoder. The second stage trained a flow-matching model on the frozen latent space, with an MLP time-step denoiser and classifier-free guidance applied to both covariates. Cells were generated by integrating the learned

flow, conditioned on a cell type and a dataset identifier sampled from the empirical per-cell-type distribution.

B.4 scVI

scVI was trained on raw counts with a zero-inflated negative-binomial likelihood, using the dataset identifier as the batch key and the cell type as the label key. Cells were generated de novo by sampling the latent prior and drawing counts from the decoded likelihood, with library sizes taken from the empirical per-cell-type distribution. No real cell was supplied as input, so this baseline isolates the model's learned prior.

B.5 C2S

C2S was trained to autoregressively generate cell sentences. Each cell was converted into a ranked list of its top K expressed genes, with a default of $K = 200$ and an expanded $K = 1{,}000$, and a Pythia-160M language model was trained from the base model on this task. Generation used the cell-type-conditioned prompt templates, after which non-gene tokens were removed, duplicate genes were collapsed by averaging their ranks, and each generated sentence was converted back into an expression vector through a rank-to-expression map fitted on real cells, so the expression magnitudes of C2S are imputed from a population-level rank–expression curve. Because only the listed genes receive non-zero values, the effective output space is restricted to K genes per cell.

C Detailed Methods and Metric Definitions

Throughout this section, c indexes a cell type, i a cell, g a gene, and p a functional program. Let X_{ig} denote library-size-normalized log-transformed expression in the common evaluation matrix. A gene is considered detected when $X_{ig} > 0$.

The real-cell reference pool was used to estimate gene-, program-, and population-level expectations. A subset of this same pool was designated as the internal real-validation group and evaluated using exactly the same pipeline as generated cells. This group therefore provides an empirical within-reference calibration range rather than an independent generalization test. A separate, non-overlapping real-cell control was used to assess the stability of the resulting calibration patterns.

For cell type c, let $\mathcal{R}_c$ denote its matched reference cells. The reference detection rate of gene g is

$$p_{gc}^{\mathrm{det}} = \frac{1}{|\mathcal{R}_c|} \sum_{j \in \mathcal{R}_c} \mathbf{1}(X_{jg} > 0). \tag{1}$$

C.1 Lexical fidelity

C.1.1 Gene usage and reference support

Transcriptomic complexity was measured by the number of detected genes in each evaluated cell,

$$K_i = \sum_g \mathbf{1}(X_{ig} > 0). \tag{2}$$

Genes detected in at least half of matched reference cells were defined as expected genes,

$$\mathcal{E}_c = \{g : p_{gc}^{\mathrm{det}} \geq 0.50\}. \tag{3}$$

Expected-gene recall was

$$\mathrm{Recall}_i^{\mathrm{gene}} = \frac{\sum_{g \in \mathcal{E}_c} \mathbf{1}(X_{ig} > 0)}{|\mathcal{E}_c|}. \tag{4}$$

An empty expected-gene set was treated as missing.

Expressed-gene support was defined as the mean reference detection rate of the genes detected in the evaluated cell,

$$\mathrm{Support}_i = \frac{\sum_{g : X_{ig} > 0} p_{gc}^{\mathrm{det}}}{\sum_g \mathbf{1}(X_{ig} > 0)}. \tag{5}$$

To quantify infrequent gene usage among highly expressed genes, detected genes were ranked by decreasing X_{ig} using a stable sort, with ties resolved by their order in the common gene matrix. Let $\mathcal{T}_i^{200}$ contain the first $\min(200, K_i)$ detected genes. The fraction of infrequently detected genes among this set was

$$\text{InfrequentGeneFraction}_i^{200} = \frac{\sum_{g \in \mathcal{T}_i^{200}} \mathbf{1}\left(p_{gc}^{\text{det}} < 0.01\right)}{|\mathcal{T}_i^{200}|}. \tag{6}$$

The threshold was strictly below 1%, and the denominator was the actual selected-gene count rather than a fixed 200 when fewer genes were detected.

C.1.2 Positive-expression calibration

Positive-expression distributions were estimated separately for each gene and cell type using positive reference expression values. Primary calibration required $p_{gc}^{\text{det}} \geq 0.05$ and at least 20 positive reference observations.

Each distribution was represented using 101 quantiles on an equally spaced probability grid between 0 and 1. Repeated quantile values were collapsed and assigned the mean probability of their corresponding grid positions.

For an evaluated positive expression value $X_{ig} > 0$, its reference-calibrated expression location was

$$Q_{ig} = \widehat{F}_{gc}^{+}(X_{ig}), \tag{7}$$

where $\widehat{F}_{gc}^{+}$ is the interpolated positive-expression reference CDF. Values below or above the represented reference range were assigned 0 and 1, respectively. For a constant reference quantile curve, a value equal to the constant was assigned 0.5.

For detection-stratified analyses, genes were grouped according to their matched-reference detection rates. Within each evaluated cell and group, the median Q_{ig} was reported across eligible genes with positive evaluated expression. Cell–group combinations containing no eligible positive genes were treated as missing.

C.2 Syntactic fidelity

C.2.1 Functional-program activity

Functional programs were defined using 6,218 Gene Ontology gene sets from Biological Process, Cellular Component, and Molecular Function after retaining sets with 10–500 genes overlapping the common expression space.

For each evaluated dataset, genes were ranked by their mean expression across that dataset and divided into 25 equally sized expression bins. Using NumPy `default_rng`, up to 50 control genes were sampled without replacement for each gene from the same expression bin. For each program, the control set was defined as the unique union of the gene-level control sets of its member genes after excluding program members. The same scoring procedure and random seed were applied independently to the reference and each generated dataset; consequently, expression-matched control genes could differ across datasets.

For program p with member-gene set G_p and matched control set C_p, activity in cell i was

$$S_{ip} = \frac{1}{|G_p|} \sum_{g \in G_p} X_{ig} - \frac{1}{|C_p|} \sum_{g \in C_p} X_{ig}. \tag{8}$$

A program was considered active when $S_{ip} > 0$. Zero and negative scores were classified as inactive. The number of active programs was $N_i^{\text{act}} = \sum_p \mathbf{1}(S_{ip} > 0)$.

The reference activity rate of program p in cell type c was

$$p_{pc}^{\text{act}} = \frac{1}{|\mathcal{R}_c|} \sum_{i \in \mathcal{R}_c} \mathbf{1}(S_{ip} > 0). \tag{9}$$

Program support for evaluated cell i was the mean reference activity rate of its active programs,

$$\mathrm{ProgSupport}_i = \frac{\sum_{p:S_{ip}>0} p_{pc}^{\mathrm{act}}}{\sum_p \mathbf{1}(S_{ip} > 0)}. \tag{10}$$

The implementation assigned zero support when no program was active.

Programs with $p_{pc}^{\mathrm{act}} \geq 0.50$ were defined as expected programs,

$$\mathcal{E}_c^{\mathrm{prog}} = \{p : p_{pc}^{\mathrm{act}} \geq 0.50\},$$

and expected-program recall was

$$\mathrm{Recall}_i^{\mathrm{prog}} = \frac{\sum_{p\in\mathcal{E}_c^{\mathrm{prog}}} \mathbf{1}(S_{ip} > 0)}{|\mathcal{E}_c^{\mathrm{prog}}|}. \tag{11}$$

To characterize activation across the range of reference activity rates, programs were assigned to intervals using p_{pc}^{act} computed from all matched reference cells. Let $\mathcal{P}_{bc}$ denote the programs in interval b for cell type c. The active fraction in cell i was

$$A_{ibc} = \frac{1}{|\mathcal{P}_{bc}|} \sum_{p\in\mathcal{P}_{bc}} \mathbf{1}(S_{ip} > 0) \tag{12}$$

An empty interval was treated as missing. The denominator includes all programs in the interval, not only those active in the evaluated cell. Programs with reference activity rate below 0.05 were omitted from these curves but remained eligible for the overall active-count, support, and recall calculations.

C.2.2 Reference-calibrated program-score magnitude

For the supplementary program-score calibration curves, only programs active in the evaluated cell contributed a score percentile. Their scores were compared with the stored 101-point quantile distribution of all signed scores in matched reference cells, including zero and negative values.

For the ordered reference quantiles $q_{pc,0}, \ldots, q_{pc,100}$, define $L_{ipc} = \#\{k : q_{pc,k} < S_{ip}\}$ and $U_{ipc} = \#\{k : q_{pc,k} \leq S_{ip}\}$. For positive evaluated scores within the represented reference range, the bounded midrank percentile was

$$H_{ipc} = \frac{L_{ipc} + U_{ipc} - 1}{200}. \tag{13}$$

Scores below or above that range were assigned 0 or 1. Ties were assigned the midpoint of their positions on the reference probability grid, including 0.5 for equality with a constant quantile curve. Within each cell and reference activity-rate interval, the median percentile across active programs was reported; intervals with no active programs were treated as missing. Curves show medians of these per-cell summaries within each model and cell type.

C.2.3 Ubiquitous and cell-type-associated programs

Ubiquitous programs required a median reference activity rate of at least 0.75 across the 18 cell categories and an activity rate of at least 0.50 in at least 75% of categories. Candidate programs were ranked using their median activity rate, prevalence across categories, and median activity-score rank.

Cell-type-associated programs were identified from the 18-category reference population. For target cell type c, eligibility required: (i) target activity rate at least 0.25; (ii) target-minus-maximum-other activity-rate difference at least 0.10; (iii) target-minus-maximum-other median-score difference at least 0.05; (iv) target median-score percentile at least 0.50; and (v) minimum pairwise AUC of at least 0.65 against every alternative category.

Eligible programs were ranked by

$$R_{pc} = (\mathrm{AUC}_{pc}^{\min} - 0.5)\sqrt{p_{pc}^{\mathrm{act}}\,\Delta p_{pc}^{\mathrm{act}}}\,\Delta m_{pc}\,r_{pc}, \tag{14}$$

where $\Delta p_{pc}^{\mathrm{act}}$ is the target-minus-maximum-other activity-rate difference, Δm_{pc} the corresponding median-score difference, and r_{pc} the target score percentile. At most 100 cell-type-associated programs were retained per category. The resulting reference-derived sets were restricted to the three evaluated cell types for the displayed comparisons.

C.2.4 PROGRAM BREADTH AND MEMBER-GENE EXPRESSION CALIBRATION

For program p in cell type c, member genes were retained when $p_{gc}^{\mathrm{det}} \geq 0.05$. Program–cell-type combinations with fewer than 10 retained genes were excluded. Let G_{pc}^{*} denote the resulting member set.

The detected-member count in cell i was

$$K_{ip} = \sum_{g \in G_{pc}^{*}} \mathbf{1}(X_{ig} > 0). \tag{15}$$

Rather than interpreting the raw detected-member fraction directly, we calibrated this count against the corresponding matched-reference distribution,

$$B_{ip} = \widehat{F}_{K,pc}^{\mathrm{mid}}(K_{ip}) = \frac{\#\{j \in \mathcal{R}_c : K_{jp} < K_{ip}\} + \frac{1}{2}\#\{j \in \mathcal{R}_c : K_{jp} = K_{ip}\}}{|\mathcal{R}_c|}. \tag{16}$$

Thus, B_{ip} measures program breadth relative to its natural reference distribution and should not be interpreted as the raw fraction of member genes detected.

Expression calibration among recovered program members was

$$I_{ip} = \frac{\sum_{g \in G_{pc}^{*}} \mathbf{1}(X_{ig} > 0) Q_{ig}}{\sum_{g \in G_{pc}^{*}} \mathbf{1}(X_{ig} > 0)}. \tag{17}$$

The value was treated as missing when no retained program member was detected. Within each cell, breadth and member-gene expression calibration were summarized separately by the median across programs of the relevant program class.

C.2.5 WITHIN-PROGRAM COORDINATION

Internal coordination was evaluated by comparing gene–gene Spearman correlation structures between reference and evaluated populations, using the same member-gene detection threshold and minimum member count as above. The reference correlations were estimated from all matched reference cells. The supported pair set contained distinct gene pairs with finite reference correlation and $|\rho_{gh}^{\mathrm{ref}}| \geq 0.10$. Pairs with non-finite evaluated correlations were excluded within each comparison; at least 10 comparable pairs were required.

For valid pair set $\mathcal{Q}_{pc}$, coordination similarity was

$$\mathrm{CoordSim}_{pc} = \rho_{\mathrm{S}}\left(\{\rho_{gh}^{\mathrm{ref}}\}_{(g,h)\in\mathcal{Q}_{pc}}, \{\rho_{gh}^{\mathrm{eval}}\}_{(g,h)\in\mathcal{Q}_{pc}}\right), \tag{18}$$

and coordination error was

$$\mathrm{CoordErr}_{pc} = \frac{1}{|\mathcal{Q}_{pc}|} \sum_{(g,h)\in\mathcal{Q}_{pc}} \left|\rho_{gh}^{\mathrm{eval}} - \rho_{gh}^{\mathrm{ref}}\right|. \tag{19}$$

Because correlation estimates depend on sample size, exactly 600 cells were sampled without replacement from each evaluated group and cell type, and the analysis was repeated 50 times. The full-reference correlations and supported-pair definitions remained fixed across repeats. For each model, cell type, and program, similarity and error were first summarized by the median across valid repeats, then across programs with equal program weighting.

C.3 SEMANTIC FIDELITY

C.3.1 REFERENCE MARKER SELECTION

Marker sets were derived from expression statistics across the 18 reference cell categories. Approximate overall expression of gene g in cell type c was defined as

$$\widetilde{\mu}_{gc} = p_{gc}^{\mathrm{det}}\, \overline{q}_{gc}^{+}, \tag{20}$$

where $\overline{q}_{gc}^{+}$ is the mean of the stored positive-expression reference quantiles.

For each target cell type, detection difference was defined as the target detection rate minus the maximum detection rate among alternative cell types. Expression contrast was the base-2 logarithm of the ratio between target and maximum-other approximate overall expression, with 0.05 added to numerator and denominator.

Marker eligibility required target detection rate at least 0.25, detection-rate difference at least 0.10, expression contrast at least 1, and within-target overall-expression percentile rank at least 0.50. Eligible genes were ranked using expression contrast, target detection rate, detection difference, and within-target expression percentile. At most 100 markers were retained per cell type.

C.3.2 Marker enrichment and quantitative calibration

Marker enrichment was calculated from up to the top 200 detected genes in each cell, using the same selected-gene set as Equation 6. Let $\mathcal{M}_c$ denote the marker set for target cell type c. Enrichment was

$$\mathrm{MarkerEnrichment}_i = \frac{|\mathcal{T}_i^{200} \cap \mathcal{M}_c|/|\mathcal{T}_i^{200}|}{|\mathcal{M}_c|/|\mathcal{G}|}, \tag{21}$$

where $\mathcal{G}$ denotes the common gene space.

Marker detection error was

$$\Delta_{gc}^{\mathrm{det}} = p_{gc}^{\mathrm{det,eval}} - p_{gc}^{\mathrm{det,ref}}. \tag{22}$$

For expression-magnitude calibration, the median Q_{ig} of each eligible marker was first calculated across evaluated cells with positive expression. The displayed summary was the unweighted median across marker-level medians.

C.3.3 Marker and program identity contrast

For marker g associated with target cell type c, the non-target population was defined by pooling the other evaluated cell types. Mean expression included zero values. Marker identity contrast was

$$\Delta_{gc}^{\mathrm{marker}} = \overline{X}_{gc}^{\mathrm{target}} - \overline{X}_{gc}^{\mathrm{non-target}}. \tag{23}$$

Evaluated contrasts were compared with corresponding internal real-validation contrasts.

The same principle was applied to cell-type-associated functional programs. For program p associated with target cell type c,

$$\Delta_{pc}^{\mathrm{prog}} = \overline{S}_{pc}^{\mathrm{target}} - \overline{S}_{pc}^{\mathrm{non-target}}. \tag{24}$$

All signed program scores were retained when calculating population means.

C.4 Population fidelity

C.4.1 Reference PCA space and repeated sampling

PCA was fitted separately within each cell type using reference cells and 50 components. The implemented analysis used IncrementalPCA with a batch size of 512 on the common normalized log-transformed expression matrix. All evaluated cells were projected into this fixed reference-derived basis; the basis was not refitted for individual models or subsamples.

For each cell type, population analyses used 30 repeated comparisons with random seed 20260902. In each repeat, 600 reference cells and 600 cells from every evaluated group were sampled without replacement. The same sampled reference set was shared across evaluated groups within a repeat. Because the internal real-validation group belongs to the reference pool, its evaluated sample excluded cells used in that repeat's sampled reference set.

The sampled reference served as the comparison target for MMD, precision and coverage, effective-rank normalization, and gene-wise variance analysis. In contrast, each cell's PCA residual was defined by the fixed full-reference mean and basis; repeated evaluated-cell samples were used only to summarize that residual.

C.4.2 GLOBAL DISTRIBUTIONAL DISCREPANCY

Global discrepancy was quantified using unbiased squared maximum mean discrepancy in the fixed 50-component PCA space,

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(r_i, r_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(e_i, e_j) - \frac{2}{n^2} \sum_{i,j} k(r_i, e_j). \tag{25}$$

The kernel was the average of three radial-basis-function kernels with bandwidths 0.5, 1, and 2 times the median pairwise Euclidean distance among at most 2,000 reference cells. Negative finite-sample MMD estimates were retained.

C.4.3 REFERENCE-MANIFOLD PRECISION AND COVERAGE

For sampled reference cell r_j, let ρ_j denote the Euclidean distance to its fifth-nearest other reference cell. Precision was

$$\mathrm{Precision}(E, R) = \frac{1}{|E|} \sum_{e \in E} \mathbf{1}\left[\exists r_j \in R : \|e - r_j\|_2 \leq \rho_j\right], \tag{26}$$

and coverage was

$$\mathrm{Coverage}(E, R) = \frac{1}{|R|} \sum_{r_j \in R} \mathbf{1}\left[\min_{e \in E} \|r_j - e\|_2 \leq \rho_j\right]. \tag{27}$$

Thus, precision measures whether evaluated cells remain within reference-supported states, whereas coverage measures whether the evaluated population represents the diversity of reference states.

C.4.4 EFFECTIVE DIMENSIONALITY AND COVARIANCE SPECTRA

Effective dimensionality was calculated from covariance eigenvalues in the fixed PCA space. Negative numerical eigenvalues were clipped to zero. Let

$$p_j = \frac{\lambda_j}{\sum_k \lambda_k}. \tag{28}$$

The effective rank was

$$r_{\mathrm{eff}} = \exp\left(-\sum_{j : p_j > 0} p_j \log p_j\right). \tag{29}$$

Each repeated comparison reported

$$\mathrm{ER\ ratio} = \frac{r_{\mathrm{eff}}^{\mathrm{eval}}}{r_{\mathrm{eff}}^{\mathrm{ref}}}. \tag{30}$$

A zero-total-variance spectrum was assigned effective rank zero.

For the supplementary PCA spectra, covariance matrices were instead calculated using all available cells in each evaluated group and the full reference group, after projection into the same fixed 50-PC space. Their 50 eigenvalues were sorted in descending order and normalized by their own sum. These curves describe each group's covariance spectrum; they are not the explained-variance ratios of the reference PCA fit or averages of the repeated 600-cell spectra.

C.4.5 RELATIVE REFERENCE-PCA RECONSTRUCTION RESIDUAL

Let $\boldsymbol{\mu}_{\mathrm{ref}}$ the reference mean, and $U \in \mathbb{R}^{G \times 50}$ the matrix whose columns are the 50 fixed orthonormal reference principal components.

The relative reference-PCA reconstruction residual was

$$r_i = \frac{\left\|(I - UU^\top)(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ref}})\right\|_2^2}{\|\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{ref}}\|_2^2}.$$

For numerical stability, the denominator was lower-bounded at 10^{-12}, and residual values below zero due only to floating-point roundoff were clipped to zero.

C.4.6 GENE-WISE VARIANCE ORGANIZATION

Gene-wise variance analyses used sample variance with one degree-of-freedom correction on normalized expression. Genes were stratified according to reference detection rate using initial boundaries

$$0,\ 0.01,\ 0.05,\ 0.10,\ 0.20,\ 0.40,\ 0.60,\ 0.80,\ 1.$$

Intervals were left-closed and right-open, with the final interval including 1. Bins containing fewer than 50 reference genes were iteratively merged with the preceding bin, or with the following bin when the first bin was sparse. Merging preceded exclusion of genes with non-finite or zero full-reference variance. The displayed analyses used the full eligible gene space rather than restricting genes to those detected by each model.

Within each detection-rate bin and repeat, variance similarity was

$$\mathrm{VarSim} = \rho_{\mathrm{S}}\left(\{v_g^{\mathrm{ref}}\}, \{v_g^{\mathrm{eval}}\}\right), \tag{31}$$

and variance bias was

$$\mathrm{VarBias} = \mathrm{median}_g\left[\log_2\left(\frac{v_g^{\mathrm{eval}} + \epsilon}{v_g^{\mathrm{ref}} + \epsilon}\right)\right], \tag{32}$$

where

$$\epsilon = \max\left(10^{-8}, 0.01 \times \mathrm{median}\{v_g^{\text{full ref}} : v_g^{\text{full ref}} > 0\}\right). \tag{33}$$

Negative variance bias indicates variance compression and positive bias indicates variance inflation relative to the sampled reference.

C.5 STATISTICAL SUMMARIES

Unless otherwise specified, cell-level quantities were summarized using medians and interquartile ranges. Repeated population-level analyses report the median across repeats together with the empirical 2.5–97.5% interval. Program-level coordination summaries first aggregate across repeated equal-cell subsamples and then across programs, so that each retained program receives equal weight.

D SUPPLEMENTARY FIGURES

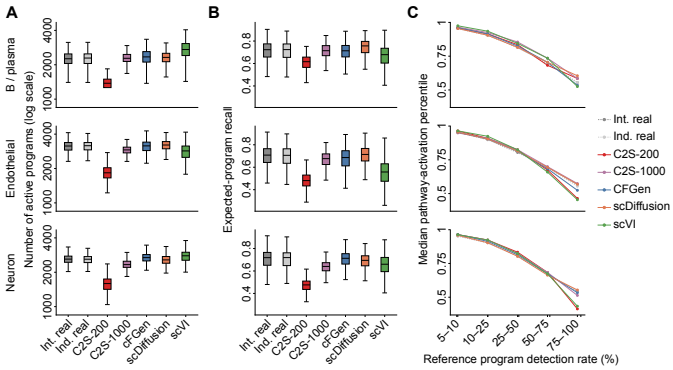

Figure S1: (A) Number of active programs per cell, where a program is considered active when its score is > 0. (B) Expected-program recall, defined as the fraction of programs active in at least 50% of matched real cells that are recovered as active in each evaluated cell. (C) Median positive program-score percentile relative to the matched program- and cell-type-specific reference distribution, stratified by reference active-rate bins. Outliers are not shown in boxplots.

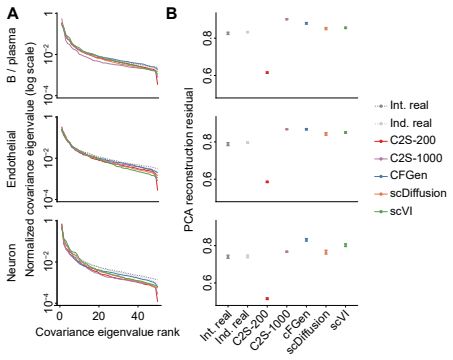

Figure S2: (A) Rank-ordered eigenvalues of the covariance matrix of each group's full 50-PC projections, normalized to sum to one. These spectra use all available cells in each group. (B) PCA reconstruction residual, measuring the fraction of cell-level variation lying outside the leading 50-component reference PCA subspace. Points show medians and bars show empirical 2.5–97.5% intervals across repeats.