Cell Painting Generates Single-Cell Transcriptomics via Conditional Diffusion

Reed Naidoo [1 2 3] **Jingyu Hu** [1 4] **Giuseppe Tripodi** [1 5] **Chris Bakal** [2 3] **Tapabrata Chakraborti** [1]

Abstract

Transcriptomic profiling resolves mechanism-of-action signal at single-cell resolution, but cannot match the scale or cost of morphological imaging. If the fingerprint of a treated cell population carries recoverable structure in transcriptomic space, every imaging experiment, spanning millions of cells at a fraction of the sequencing cost, becomes a latent source of molecular insight. We introduce PhenoSeq, a conditional diffusion model with a cross-attention denoising architecture that bridges two biological foundation models: a vision transformer for morphological features and a transcriptomic language model for single-cell gene expression. Operating under population-level supervision, it learns a conditional distribution over transcriptomic profiles from treatment-matched morphological observations. Evaluated on a 28-compound treatment-identification benchmark, PhenoSeq-generated embeddings outperform raw imaging in single-profile classification, and multimodal fusion recovers $\approx 29\%$ of the gap to the real-transcriptomics ceiling; in the multi-profile setting, synthetic fusion more than doubles imaging-only balanced accuracy. Embedding-space fidelity confirms correct treatment localisation for the majority of conditions. These results demonstrate that generative cross-modal modelling from imaging to transcriptomics is both architecturally feasible and downstream-useful in phenotypic drug discovery. Our code is available at: https://github.com/reednaidoo/PhenoSeq

[1]The Alan Turing Institute, London, United Kingdom [2]The Institute of Cancer Research, London, United Kingdom [3]Sentinal4D, London, United Kingdom [4]University of Bristol, Bristol, United Kingdom [5]University of Manchester, Manchester, United Kingdom. Correspondence to: Tapabrata Chakraborti <tchakraborty@turing.ac.uk>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

1. Introduction

High-content imaging assays such as Cell Painting (Bray et al., 2016) produce morphological fingerprints of cellular response to perturbation at scale (Haghighi et al., 2022), making them among the richest phenotypic resources in drug discovery (Kraus et al., 2024; De Vries et al., 2025). Through the spatial organisation of cellular substructures and nuclear geometry, these fingerprints encode the cumulative effect of a cell's molecular history and are broadly predictive of mechanism of action (Bakal et al., 2007; Lomakin et al., 2020). Yet morphology captures only one projection of cellular state. Single-cell RNA sequencing resolves the underlying molecular programme directly, including pathway activity, regulatory dynamics, and mechanism-of-action signal with access to mechanistic information that imaging alone cannot fully decode (Van de Sande et al., 2023; Dapello et al., 2026). These modalities are not in competition but in correspondence, both projections of the same biology onto different measurement axes (Way et al., 2022; Sailem et al., 2014).

This correspondence motivates a generative question: *can a model learn the mapping between them?* In digital pathology, cross-modal generation has produced clinically significant results: HE2RNA (Schmauch et al., 2020) showed bulk RNA-seq signal is recoverable from histology, and PathGen (Dey et al., 2025) demonstrated that diffusion-generated transcriptomics fused with imaging reaches performance indistinguishable from real molecular data on cancer grading and survival prediction. These results establish a clear principle: synthesised transcriptomics can provide functional signal genuinely complementary to imaging, even when generation is imperfect. The analogous challenge in drug discovery has remained unexplored, partly for lack of paired data. Dapello et al. (2026) recently introduced the first dataset pairing single-cell RNA-seq with Cell Painting images under matched perturbations, and explicitly identified joint generative modelling as an untested direction.

We introduce **PhenoSeq**, a conditional diffusion model (Ho et al., 2020; Song et al., 2022) that generates transcriptomic embeddings from Cell Painting imaging features without cell-paired supervision, trained on Dapello et al. (2026). Because RNA-seq is destructive, PhenoSeq models a population-level conditional distribution over transcrip-

tomic profiles given treatment-matched morphological observations, rather than learning a per-cell mapping. Overall, we propose PhenoSeq and make the following contributions:

- The first conditional diffusion model bridging Cell Painting imaging to single-cell transcriptomic embeddings, building on foundation models for morphological profiling (Kraus et al., 2024; Chandrasekaran et al., 2023), transcriptomic representation learning (Cui et al., 2024; Theodoris et al., 2023), and within-modality generative models (Zhang et al., 2025; Palma et al., 2025; Wang et al., 2025; Naidoo et al., 2025; Luo et al., 2024; He et al., 2026).
- A cross-attention denoising architecture paired with a self-attention imaging context encoder, enabling population-level conditioning without requiring per-cell paired supervision.
- Synthetic embeddings outperform imaging in single-profile classification (WE 0.293 vs. 0.270); multi-modal fusion recovers ≈29% of the gap to the real-transcriptomics ceiling (WE 0.315 vs. 0.425) and more than doubles imaging-only balanced accuracy in the multi-profile setting.

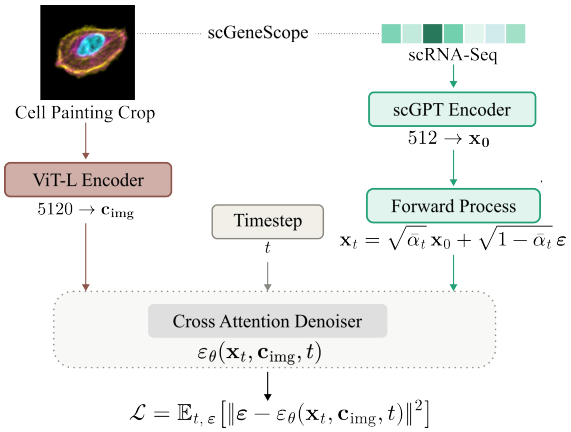

Figure 1. **PhenoSeq training schematic.** Cell Painting crops are encoded per-channel by a frozen ViT-L backbone and concatenated into a 5,120-dim imaging feature vector, which is then contextualised by a 2-layer self-attention encoder before serving as cross-attention conditioning context. scRNA-seq profiles are embedded by scGPT into a 512-dim vector $\mathbf{x}_0$. A timestep $t \sim \mathcal{U}\{1, T\}$ is sampled and noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to $\mathbf{x}_0$ via the cosine-schedule forward process to produce $\mathbf{x}_t$. The cross-attention denoiser ε_θ (6 blocks, AdaLN time conditioning) predicts the added noise, and the model is trained to minimise $\mathcal{L} = \mathbb{E}_{t,\boldsymbol{\varepsilon}}[\|\boldsymbol{\varepsilon} - \varepsilon_\theta(\mathbf{x}_t, \mathbf{c}_{\text{img}}, t)\|^2]$. Imaging and transcriptomic measurements are matched at the treatment-condition level rather than the individual cell level.

2. Methods and Evaluation Protocol

This section first introduces data processing pipelines (§2.1), followed by detailed description of our proposed PhenoSeq method (§2.2), and evaluation protocol (§2.3).

2.1. Dataset and Feature Extraction

Population-level pairing. Because RNA-seq is destructive, imaging and transcriptomic measurements are made on different cells from the same treated well; pairing exists at the treatment-condition level only. PhenoSeq therefore learns a population-level conditional generator, and evaluation is conducted at the treatment level.

The scGeneScope dataset (Dapello et al., 2026) pairs Cell Painting images with scRNA-seq profiles of U2-OS cells under 28 chemical perturbations, across five replicates in two experimental rounds. Imaging features are ViT-L ImageNet embeddings (Deng et al., 2009) (5 channels × 1024 = 5,120-dim). Transcriptomic targets are 512-dim scGPT embeddings (Cui et al., 2024), chosen so PhenoSeq inherits scGPT's biologically structured pre-training geometry as the target manifold. Embeddings are normalised to zero mean and unit variance using training-split statistics.

2.2. PhenoSeq: Conditional Diffusion Model

At each step, N=16 cells are sampled uniformly from the treatment pool (line 2); the self-attention context encoder (line 8) aggregates population-level structure before cross-attention conditioning in ε_θ.

Architecture. PhenoSeq is a denoising diffusion probabilistic model (Ho et al., 2020) that generates 512-dim transcriptomic embeddings conditioned on a set of per-cell imaging features, following the training procedure in Algorithm 1. The architecture has three components: an imaging context encoder, a time embedding module, and a cross-attention denoising stack.

The imaging context encoder projects N per-cell ViT-L embeddings (5,120-dim) to 1,024-dim and passes them through a 2-layer self-attention transformer (4 heads, GEGLU feedforward), capturing population-level covariation across cells before the context is consumed by the denoiser. The resulting sequence $\mathbf{c}_{\text{img}} \in \mathbb{R}^{N \times 1024}$ serves as the key-value input to all downstream cross-attention layers. The time embedding maps timestep t to a 256-dim vector via sinusoidal encoding and a 2-layer Mish MLP, consumed by adaptive layer norm (AdaLN) in every block to modulate scale and shift as a function of noise level.

The denoising stack comprises six transformer blocks, each operating on the noisy RNA embedding $\mathbf{x}_t$ projected to 1,024-dim. Within each block, 8-head cross-attention attends over $\mathbf{c}_{\text{img}}$ at every denoising step, followed by self-attention and a GEGLU feedforward layer ($4\times$ expansion), with AdaLN conditioning throughout. The final output is projected to 512-dim to produce $\varepsilon_\theta(\mathbf{x}_t, \mathbf{c}_{\text{img}}, t)$. Architec-

Algorithm 1 PhenoSeq Training

Require: Dataset $\mathcal{D} = \{(\mathbf{X}_{\text{img}}^{(c)}, \mathbf{x}_0^{(c)})\}$, noise schedule $\{\bar{\alpha}_t\}_{t=1}^{T}$, denoiser ε_θ

repeat
  Sample condition $c \sim \mathcal{D}$
  Sample N=16 imaging cells $\mathbf{X}_{\text{img}} \subset \mathbf{X}_{\text{img}}^{(c)}$
  Sample RNA embedding $\mathbf{x}_0 \sim p(\mathbf{x}_0 \mid c)$
  Sample timestep $t \sim \mathcal{U}\{1, T\}$
  Sample noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Compute $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}$
  Encode context $\mathbf{c}_{\text{img}} = \text{SelfAttn}(\text{Proj}(\mathbf{X}_{\text{img}}))$
  Take gradient step on $\nabla_\theta \|\boldsymbol{\varepsilon} - \varepsilon_\theta(\mathbf{x}_t, \mathbf{c}_{\text{img}}, t)\|^2$
  Update EMA weights $\bar{\theta} \leftarrow \lambda\bar{\theta} + (1-\lambda)\theta$
until convergence

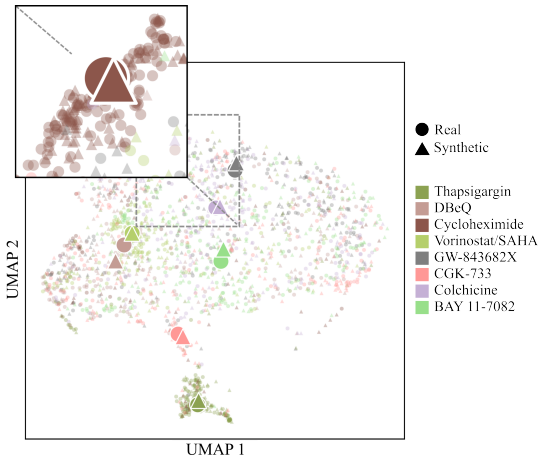

Figure 2. UMAP comparison of scGPT representations for the top 8 treatments. Circles (●) and triangles (▲) denote real and synthetic cells respectively; filled markers indicate cluster centroids. Inset shows a magnified view of the top-ranked treatment cluster.

ture hyperparameters are in Table 4.

Diffusion process. We use Gaussian diffusion with a cosine β schedule (Nichol & Dhariwal, 2021) over $T = 1{,}000$ steps. At each training step, a timestep $t \sim \mathcal{U}\{1, T\}$ is sampled, noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to the clean RNA embedding $\mathbf{x}_0$ according to the forward process, and the model is trained to predict the added noise: $\mathcal{L} = \mathbb{E}_{t,\,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t,\, \mathbf{c}_{\text{img}},\, t)\|^2\right]$. The cosine schedule is preferred over linear because it decays $\bar{\alpha}_t$ more gradually near the terminal end of the forward process, preserving signal-to-noise at high noise levels and improving the conditioning of the denoising task.

Training. The model is optimised with AdamW (lr=10^{-4}, weight decay 10^{-5}) with batch size 256, cosine schedule with 5-epoch linear warmup, early stopping (patience 50 on validation loss), and EMA (decay 0.9999) used exclusively at inference. Full training hyperparameters are in Table 5.

Inference. At inference, all available imaging cell embeddings are mean-pooled into a single context token (vs. N=16 random cells at training), removing sampling variance from the conditioning signal. Sampling follows DDIM (Song et al., 2022) with 50 steps under a fixed seed; outputs are de-normalised using training-split statistics before all downstream evaluation.

2.3. Evaluation Protocol

The effectiveness of PhenoSeq is evaluated in embedding-space fidelity and treatment classification performance.

Embedding-space fidelity. For each of the 29 conditions (28 compounds and DMSO, with DMSO retained as an additional reference point) we compute per-treatment mean embeddings for real and synthetic profiles separately. Because pairwise cosine similarities between real treatment centroids sit near 0.999 — a consequence of scGPT's pre-training geometry rather than model behaviour — absolute cosine values are not diagnostic of treatment specificity. The primary metrics are therefore retrieval rank, Accuracy@1, and Mean Reciprocal Rank (MRR). For each synthetic centroid $\mathbf{s}_T$ we rank all 29 real centroids by cosine similarity; rank 1 indicates the synthetic embedding is geometrically nearest to its own treatment's real centroid.

Treatment classification. We mirror the scGeneScope benchmark exactly (Dapello et al., 2026). The canonical split assigns replicate 3 to training, replicate 5 to validation, replicate 4 as the within-experiment (WE) test, and replicates 1+2 from the independent round 2 as the held-out-experiment (HE) test. DMSO is excluded from the prediction target, yielding a 28-class problem with chance level $1/28 \approx 0.036$ balanced accuracy. All results are averaged over 5 independent training seeds; we report mean ± std. Reference rows from Dapello et al. (2026) (marked †, reported as mean ± std error over 5 seeds on the same split) are included for calibration.

We evaluate two input settings. In the *single-profile* setting, all cells from a sample are averaged into one embedding before training and evaluation, removing subsampling stochasticity and providing the cleanest modality comparison. In the *multi-profile* (avgpool) setting, k=32 cells are drawn uniformly at random per sample per forward pass and averaged, simulating realistic inference where only a subset of profiled cells is available. For multimodal inputs, imaging and RNA embeddings are sampled independently from their respective cell pools, reflecting the experimental reality that the two modalities are measured on different cells.

3. Experimental Results

3.1. Embedding-Space Fidelity

To measure the fidelity of synthetic transcriptomic profiles relative to real ones, we compare the two qualitatively in Figure 2 via UMAP visualisation, and quantitatively in Appendix Table 3 via per-treatment retrieval rank and specificity. The visualization shows that the global cluster geometry of the real scGPT space is preserved in the synthetic space. Synthetic treatment centroids localise correctly in the real scGPT embedding space for the majority of conditions: 21 of 29 conditions achieve rank ≤ 4, and 15 achieve rank 1 (MRR $= 0.60$). Treatments with morphologically distinctive phenotypes rank first with the largest specificity margins: Thapsigargin (0.0109), DBeQ (0.008), Cycloheximide (0.0079). Simvastatin (rank 17), LY303511 (rank 18), and 12-O-tetradecanoylphorbol-13-acetate (rank 25) are three failure cases: they induce phenotypes that are insufficiently discriminative in U2-OS cells at the doses used, providing a natural ceiling: the model cannot place synthetic centroids in the correct neighbourhood when imaging features themselves do not separate the treatment from controls.

3.2. Treatment Classification: Single Profile

We further examine how much our synthetic embeddings can benefit treatment classification when used as a complementary modality. Table 1 reports single-profile results. The primary finding is that synthetic RNA improves over imaging both unimodally and in fusion; HE gains are more modest. Unimodally, synthetic RNA achieves WE balanced accuracy of 0.293 ± 0.002 against 0.270 ± 0.005 for imaging alone — a gain of $+0.023$ WE. Because PhenoSeq's output is a deterministic function of imaging features (fixed DDIM seed, mean-pool conditioning), this gain reflects feature distillation into biologically structured scGPT geometry rather than recovery of genuinely new transcriptomic signal: the same imaging information, transformed into a target space shaped by scGPT pre-training, is more useful to the downstream classifier under matched capacity.

Table 1. Treatment classification under single-profile setting (mean $\pm$ std over 5 seeds, canonical split: train=rep3 val=rep5 WE-test=rep4 HE-test=rep1+2). Reference rows (†) are reported results from Dapello et al. (2026) Table 1 using zero-shot embeddings.

Method	WE Bal. Acc.	WE Macro-F1	HE Bal. Acc.	HE Macro-F1
Image (real)	0.270 ± 0.005	0.258 ± 0.002	0.206 ± 0.002	0.179 ± 0.003
RNA (real)	0.383 ± 0.002	0.375 ± 0.002	0.379 ± 0.002	0.340 ± 0.003
RNA (synthetic)	0.293 ± 0.002	0.243 ± 0.002	0.205 ± 0.001	0.152 ± 0.002
Image + RNA (real)	0.425 ± 0.003	0.403 ± 0.004	0.402 ± 0.003	0.375 ± 0.005
Image + RNA (synthetic)	0.315 ± 0.002	0.269 ± 0.002	0.226 ± 0.002	0.194 ± 0.003
Image (real)†	0.279 ± 0.009	0.264 ± 0.013	0.208 ± 0.003	0.184 ± 0.002
RNA (real, scGPT)†	0.387 ± 0.004	0.378 ± 0.003	0.381 ± 0.004	0.341 ± 0.005

[0]The scGeneScope benchmark reports results across multiple embedding backbones and input settings, including multimodal and multi-profile configurations not fully reproduced here. Our comparisons are scoped to the scGPT zero-shot rows, which match the embedding space PhenoSeq targets; higher-performing options exist in the full benchmark.

Fusing synthetic RNA with real imaging yields WE 0.315 ± 0.002 and HE 0.226 ± 0.002, recovering $\approx 29\%$ of the gap between imaging-only (WE 0.270) and the real-RNA multimodal ceiling (WE 0.425 ± 0.003, HE 0.402 ± 0.003). Our synthetic RNA and multimodal synthetic numbers sit close to but slightly below the scGeneScope reference scGPT rows (†), consistent with independently tuned hyperparameters on the same split. The HE generalisation of imaging-only (0.206) and synthetic RNA unimodal (0.205) are near-identical, confirming that the WE gain does not worsen out-of-distribution generalisation.

3.3. Treatment Classification: Multi Profile

Besides the single-profile setting, Table 2 further reports multi-profile performance (avgpool, k=32). The gain from synthetic RNA fusion is larger here in absolute terms than in the single-profile setting: multimodal fusion of synthetic RNA with imaging achieves WE 0.271 ± 0.029 against an imaging-only baseline of WE 0.157 ± 0.017, a $+0.114$ absolute gain on WE and $+0.125$ on HE (0.211 vs. 0.086).

Table 2. Treatment classification under multi-profile setting (avgpool, k=32, mean $\pm$ std over 5 seeds, canonical split).

Method	WE Bal. Acc.	WE Macro-F1	HE Bal. Acc.	HE Macro-F1
Image (real)	0.157 ± 0.017	0.126 ± 0.015	0.086 ± 0.013	0.055 ± 0.016
RNA (real)	0.429 ± 0.039	0.395 ± 0.039	0.307 ± 0.029	0.237 ± 0.037
RNA (synthetic)	0.300 ± 0.017	0.253 ± 0.033	0.286 ± 0.039	0.225 ± 0.050
Image + RNA (real)	0.436 ± 0.061	0.369 ± 0.070	0.336 ± 0.013	0.270 ± 0.032
Image + RNA (synthetic)	0.271 ± 0.029	0.232 ± 0.032	0.211 ± 0.031	0.179 ± 0.033

This pattern is consistent with the feature-distillation interpretation: in the multi-profile setting, the imaging-only classifier is weakened by stochastic 32-cell subsampling, making the representational improvement from generating into scGPT space more visible. Unimodal synthetic RNA achieves WE 0.300 ± 0.017, above the imaging-only baseline (WE 0.157); elevated standard deviation reflects the interaction between stochastic subsampling and treatment imbalance across replicate folds, and the single-profile setting remains the cleaner primary comparison. The real-RNA multimodal ceiling reaches WE 0.436 ± 0.061 and HE 0.336 ± 0.013.

4. Conclusion

PhenoSeq demonstrates that foundation model representations are complementary across modalities: Cell Painting imaging features, distilled into a biologically structured transcriptomic space via a conditional diffusion model, improve treatment identification over imaging alone. It recovers approximately 29% of the gap to the real-transcriptomics ceiling in single-profile fusion, and more than doubling

imaging-only balanced accuracy in the multi-profile setting. Embedding-space fidelity confirms correct localisation for 21 of 29 conditions. The population-level pairing imposed by destructive sequencing is a natural starting point; future work will incorporate cell-paired joint readouts, richer population-level conditioning via biological priors, and learned set-aggregation at inference to move toward per-cell supervision. In digital pathology, diffusion-generated transcriptomics has already produced clinically meaningful gains over imaging-only models (Schmauch et al., 2020; Dey et al., 2025) and PhenoSeq demonstrates the same is viable in phenotypic drug discovery.

Impact Statement

PhenoSeq generates synthetic transcriptomic embeddings from Cell Painting imaging as an auxiliary representation for downstream analysis in early-stage phenotypic drug discovery. Synthetic embeddings are intended to complement, not replace, empirical molecular measurement. The primary societal benefit is accelerating mechanism-of-action discovery at reduced experimental cost; we do not anticipate harms beyond those generally associated with machine-learning-driven compound prioritisation, such as the propagation of dataset biases into screening decisions. There are no concerns specific to this work that we feel must be highlighted here.

References

Bakal, C., Aach, J., Church, G., and Perrimon, N. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science*, 316(5832):1753–1756, 2007. doi: 10.1126/science.1140324. URL https://www.science.org/doi/abs/10.1126/science.1140324.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, Sep 2016. ISSN 1750-2799. doi: 10.1038/nprot.2016.105. URL https://doi.org/10.1038/nprot.2016.105.

Chandrasekaran, S. N. et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023. doi: 10.1101/2023.03.23.534023. URL https://www.biorxiv.org/content/early/2023/03/24/2023.03.23.534023.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, Aug 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Dapello, J. et al. scgenescope: A treatment-matched single cell imaging and transcriptomics dataset and benchmark for treatment response modeling. 2026. URL https://openreview.net/forum?id=918POZbZ50.

De Vries, M., Dent, L. G., Curry, N., Rowe-Brown, L., Bousgouni, V., Fourkioti, O., Naidoo, R., Sparks, H., Tyson, A., Dunsby, C., and Bakal, C. Geometric deep learning and multiple-instance learning for 3d cell-shape profiling. *Cell Systems*, 16(3), Mar 2025. ISSN 2405-4712. doi: 10.1016/j.cels.2025.101229. URL https://doi.org/10.1016/j.cels.2025.101229.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dey, S., Banerji, C. R. S., Basuchowdhuri, P., Saha, S. K., Parashar, D., and Chakraborti, T. Generating crossmodal gene expression from cancer histopathology improves multimodal ai predictions. *Nature Communications*, 17 (1):259, Dec 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-66961-9. URL https://doi.org/10.1038/s41467-025-66961-9.

Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E., and Singh, S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 19(12): 1550–1557, Dec 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01667-0. URL https://doi.org/10.1038/s41592-022-01667-0.

He, S. et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 23(1):65–77, Jan 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02877-y. URL https://doi.org/10.1038/s41592-025-02877-y.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

Lomakin, A. J. et al. The nucleus acts as a ruler tailoring cell responses to spatial constraints. *Science*, 370(6514):eaba2894, 2020. doi: 10.1126/science.aba2894. URL https://www.science.org/doi/abs/10.1126/science.aba2894.

Luo, E., Hao, M., Wei, L., and Zhang, X. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9): btae518, 09 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae518. URL https://doi.org/10.1093/bioinformatics/btae518.

Naidoo, R., Vries, M. D., Fourkioti, O., Bousgouni, V., Arias-Garcia, M., Portillo-Malumbres, M., and Bakal, C. Morphologically intelligent perturbation prediction with form, 2025. URL https://arxiv.org/abs/2510.21337.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021. URL https://arxiv.org/abs/2102.09672.

Palma, A., Theis, F. J., and Lotfollahi, M. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16 (1):505, Jan 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55707-8. URL https://doi.org/10.1038/s41467-024-55707-8.

Sailem, H., Bousgouni, V., Cooper, S., and Bakal, C. Cross-talk between rho and rac gtpases drives deterministic exploration of cellular shape space and morphological heterogeneity. *Open Biology*, 4:130132, 2014. doi: 10.1098/rsob.130132.

Schmauch, B. et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, Aug 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17678-4. URL https://doi.org/10.1038/s41467-020-17678-4.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

Van de Sande, B. et al. Applications of single-cell rna sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6):496–520, Jun 2023. ISSN 1474-1784. doi: 10.1038/s41573-023-00688-4. URL https://doi.org/10.1038/s41573-023-00688-4.

Wang, Z., Chen, Y., Ma, P., Yu, Z., Wang, J., Liu, Y., Ye, X., Sakurai, T., and Zeng, X. Image-based generation for molecule design with sketchmol. *Nature Machine Intelligence*, 7(2):244–255, Feb 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00982-3. URL https://doi.org/10.1038/s42256-025-00982-3.

Way, G. P. et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Syst*, 13(11):911–923.e9, October 2022.

Zhang, Y., Su, Y., Wang, C., Li, T., Wefers, Z., Nirschl, J., Burgess, J., Ding, D., Lozano, A., Lundberg, E., and Yeung-Levy, S. Cellflux: Simulating cellular morphology changes via flow matching, 2025. URL https://arxiv.org/abs/2502.09775.

A. Detailed Methods

A.1. Diffusion Model

Forward process. PhenoSeq uses Gaussian diffusion (Ho et al., 2020) in the 512-dimensional scGPT embedding space. Given a clean RNA-seq embedding $\mathbf{x}_0 \in \mathbb{R}^{512}$, the forward process defines a fixed Markov chain that progressively *noises* $\mathbf{x}_0$ with Gaussian noise over $T = 1{,}000$ steps:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t\mathbf{I}\Big)\,, \tag{1}$$

where $\{\beta_t\}_{t=1}^{T}$ is the noise schedule. The marginal at any step admits a closed form:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s), \tag{2}$$

which is used to sample noisy embeddings directly at training time without iterating through the chain.

Noise schedule. We use the cosine schedule of Nichol & Dhariwal (2021), which defines:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos\left(\frac{t/T + s}{1+s}\cdot\frac{\pi}{2}\right)^2, \tag{3}$$

with offset $s = 0.008$ to avoid $\bar{\alpha}_t$ reaching zero too early near $t = T$.

Training objective. The denoiser ε_θ is trained with the simple noise-prediction objective (Ho et al., 2020):

$$\mathcal{L} = \mathbb{E}_{t,\,\mathbf{x}_0,\,\boldsymbol{\varepsilon}}\Big[\big\|\boldsymbol{\varepsilon} - \varepsilon_\theta(\mathbf{x}_t,\ \mathbf{c}_{\text{img}},\ t)\big\|^2\Big]\,, \tag{4}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_t$ is obtained from Equation 2, and $\mathbf{c}_{\text{img}}$ is the imaging context sequence described below. The timestep t is sampled uniformly from $\{1, \ldots, T\}$ at each training step.

Denoiser architecture. The denoiser ε_θ takes as input a noisy RNA embedding $\mathbf{x}_t$, an imaging context $\mathbf{c}_{\text{img}}$, and a scalar timestep t, and outputs a noise prediction of the same shape as $\mathbf{x}_0$. It has three components.

(i) Imaging encoder. At each training step, $N = 16$ imaging cell embeddings are drawn uniformly at random from the sample's pool of ViT-L features (each 5,120-dimensional). These are projected to 1,024 dimensions via a linear layer and passed through a 2-layer self-attention transformer (4 heads, no positional encoding), producing a context sequence $\mathbf{c}_{\text{img}} \in \mathbb{R}^{N\times 1024}$. Self-attention across the N cells allows the encoder to capture population-level structure — heterogeneity and co-variation within the well — before serving as conditioning context.

(ii) Time embedding. The scalar timestep t is encoded as a 256-dimensional sinusoidal embedding, then passed through a two-layer MLP to produce a time vector $\mathbf{e}_t \in \mathbb{R}^{256}$. This vector is consumed by every adaptive layer norm (AdaLN) in the denoising stack, modulating both scale and shift of each sublayer's output as a function of the noise level.

(iii) Cross-attention denoising stack. Six transformer blocks progressively refine the noisy RNA query. In each block: (a) $\mathbf{x}_t$ is projected to 1,024 dimensions and treated as a single query token that attends over $\mathbf{c}_{\text{img}}$ via 8-head cross-attention, pulling information from the imaging context at every denoising step; (b) the output undergoes self-attention (residual refinement); (c) a feedforward layer with $4\times$ hidden expansion is applied. AdaLN with $\mathbf{e}_t$ conditions both the self-attention and feedforward sublayers. The final block output is projected from 1,024 to 512 dimensions to produce $\varepsilon_\theta(\mathbf{x}_t, \mathbf{c}_{\text{img}}, t)$. Full architecture hyperparameters are in Table 4.

Training details. The model is optimised with AdamW (lr $= 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-5}), batch size 256, gradient clipping at $\|\cdot\|_2 \le 1.0$, and a 5-epoch linear warm-up followed by cosine learning-rate decay. Training runs for up to 5,000 epochs with early stopping on validation loss (patience 50). An exponential moving average (EMA, decay 0.9999) of model weights is maintained throughout and used exclusively for inference. Training hyperparameters are summarised in Table 5.

Inference. Synthetic embeddings are generated per sample using DDIM (Song et al., 2022) with 50 deterministic denoising steps, starting from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At inference, the imaging context is the mean of all available imaging cell embeddings for the sample — a single 5,120-dim vector projected to a single context token — rather than a stochastic subset of $N = 16$ cells as at training time. This removes sampling variance from the conditioning signal and aligns with the population-level evaluation protocol. It introduces a mild train/inference asymmetry in the context encoder (trained on length-N sequences, evaluated on length-1), which we note as an avenue for future improvement. Under a fixed random seed, the DDIM sampler is deterministic: synthetic embeddings are therefore a deterministic function of the imaging input at inference time.

A.2. Downstream Classification: Evaluation Protocol

The treatment identification benchmark mirrors the scGeneScope protocol exactly (Dapello et al., 2026). This section documents every design decision in sufficient detail to allow independent replication.

Data splits. The scGeneScope dataset provides five pairs of biological replicates across two independent experimental rounds (Figure 2 of Dapello et al. 2026). We use the canonical split defined in that work:

- **Train**: Round 1, Replicate 3 (146,389 scImages / 103,865 scRNA-seq profiles).
- **Validation**: Round 1, Replicate 5 (138,842 / 113,446).
- **Within-Experiment (WE) test**: Round 1, Replicate 4 (138,842 / 113,446).
- **Held-out-Experiment (HE) test**: Round 2, Replicates 1 and 2 (231,710 / 179,357, pooled).

The WE test assesses in-distribution performance on a held-out replicate from the same experimental round. The HE test assesses generalisation to a fully independent experimental round collected under slightly different but functionally equivalent procedures (Dapello et al., 2026). Round 2 replicates are never included in the train or validation splits: treatment and batch are confounded in Round 2 (groups of seven treatments are processed consecutively), so splitting Round 2 across train/test would allow batch signal to inflate test accuracy.

Classification target. DMSO (CONTROL) is excluded from the prediction target, yielding a 28-class treatment identification problem. DMSO is retained in the embedding-space fidelity analysis (Appendix B) as an additional reference condition, but is never a classification target. This matches the scGeneScope benchmark definition and makes the chance level $1/28 \approx 0.036$ balanced accuracy.

Input settings. We evaluate two input settings, following Dapello et al. (2026):

Single-profile. Each input example is a single cell embedding $\mathbf{x}_i \in \mathbb{R}^D$. At training and evaluation time, all cells from a sample are averaged into one vector before being presented to the classifier. This removes stochasticity from the input and is the most stable setting for comparing modalities.

Multi-profile (avgpool). Each forward pass draws $k = 32$ cells uniformly at random from the sample's cell pool (without replacement if $n_{\text{cells}} \geq 32$, otherwise with replacement), computes their mean, and presents the result as the classifier input. This simulates the practical setting where a subset of profiled cells is used at inference, and tests whether classifiers can integrate evidence across a population. The stochastic subsampling introduces evaluation variance that compounds with treatment imbalance across the replicate-based CV folds; the single-profile setting is therefore the primary comparison.

For multimodal inputs, imaging and RNA-seq embeddings are drawn *independently* from their respective cell pools, following the scGeneScope multimodal sampling protocol. This reflects the experimental reality that the two modalities are measured on different cells.

Classifier architectures. *Unimodal MLP.* A feedforward network with n_{depth} hidden layers, each of width n_{hidden}, with LayerNorm and ReLU activations, followed by a linear classification head:

$$h(\mathbf{x}) = W_{\text{out}} \left[\text{ReLU} \circ \text{LN} \circ W_n \circ \cdots \circ \text{ReLU} \circ \text{LN} \circ W_1\right](\mathbf{x}), \tag{5}$$

where $W_1, \ldots, W_n$ are linear layers of width n_{hidden} and $W_{\text{out}} \in \mathbb{R}^{C \times n_{\text{hidden}}}$ maps to $C = 28$ treatment classes.

Multimodal DualMLP. Separate MLP encoders for each modality:

$$\mathbf{z}_{\text{img}} = \text{MLP}_{\text{img}}(\mathbf{x}_{\text{img}}) \in \mathbb{R}^{913}, \tag{6}$$

$$\mathbf{z}_{\text{rna}} = \text{MLP}_{\text{rna}}(\mathbf{x}_{\text{rna}}) \in \mathbb{R}^{913}, \tag{7}$$

concatenated and passed through a linear head:

$$h(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{rna}}) = W_{\text{out}} \left[\mathbf{z}_{\text{img}} \parallel \mathbf{z}_{\text{rna}}\right], \quad W_{\text{out}} \in \mathbb{R}^{C \times 1826}. \tag{8}$$

The imaging encoder maps $\mathbb{R}^{5120} \to \mathbb{R}^{266\times3} \to \mathbb{R}^{913}$ and the RNA encoder maps $\mathbb{R}^{512} \to \mathbb{R}^{2748\times3} \to \mathbb{R}^{913}$, with LayerNorm-ReLU at each hidden layer. This architecture exactly mirrors the `MultiModalMultipleInputClassifier` used by Dapello et al. (2026).

Training procedure. All classifiers are trained with AdamW, batch size 4,096, cross-entropy loss, and early stopping on validation balanced accuracy (patience 25, mode max). Gradient clipping is applied at $\|\cdot\|_2 = 0.5$. The learning rate and weight decay for each modality are independently tuned (Table 6); all other hyperparameters are shared. Models are trained for up to 100 epochs; the checkpoint with the best validation balanced accuracy is used for test evaluation.

Metrics. We report balanced accuracy (macro-average recall across classes), macro-F1, and weighted-F1, averaged over the WE and HE test splits separately. Balanced accuracy is the primary metric, matching Dapello et al. (2026), and is robust to the mild class imbalance arising from variable cell yields per treatment replicate.

Canonical split: 5-seed evaluation. Primary results (Tables 1 and 2) are obtained by training 5 classifiers per modality on the canonical split above, each with a different random seed (`12345, 42, 1234, 99, 7`), and reporting mean $\pm$ standard deviation. The standard deviation over seeds captures training stochasticity (random weight initialisation and, in the multi-profile setting, stochastic 32-cell subsampling per forward pass) under a fixed data split.

3-fold cross-validation: robustness check. As an additional robustness check, we evaluate all modalities under 3-fold leave-one-replicate-out cross-validation over the three Round-1 replicates (3, 4, 5). In each fold, one replicate is held out as the WE test set, one as the validation set, and one as the training set:

Fold	Train	Val	WE test
1	Rep 3	Rep 5	Rep 4
2	Rep 4	Rep 3	Rep 5
3	Rep 5	Rep 4	Rep 3

The HE test (Round 2, Replicates 1+2) is evaluated at the end of every fold using the fold's best checkpoint. Mean $\pm$ std are reported over the three folds, with a single training seed (`12345`).

Comparison with scGeneScope published results. Tables 1 and 2 include reference rows from Dapello et al. (2026) Table 1 for ViT-L (zero-shot ImageNet) and scGPT (zero-shot). These are marked † and reported as mean $\pm$ standard error over 5 training seeds on their canonical split. Our classifiers for the same backbones differ from the published numbers because: (i) our hyperparameters were independently tuned on the validation split rather than using the scGeneScope search space, and (ii) our variant of the canonical split uses the same replicate assignment but a different random seed schedule. The directional ordering of modalities is consistent with the published benchmark.

A.3. Embedding-Space Fidelity Evaluation

For each of the 29 conditions (28 compounds and DMSO — DMSO retained here as an additional reference point), we compute:

1. **Treatment-mean centroids.** Let $\mathbf{r}_T = \frac{1}{N_T}\sum_{i=1}^{N_T} \mathbf{r}_{T,i}$ be the mean real scGPT embedding over all cells from treatment T, and $\mathbf{s}_T$ the analogous mean over synthetic embeddings.

2. **Cosine similarity** $\cos(\mathbf{s}_T, \mathbf{r}_T)$, Euclidean distance $\|\mathbf{s}_T - \mathbf{r}_T\|_2$, and Pearson correlation $r(\mathbf{s}_T, \mathbf{r}_T)$ between same-treatment centroids.

3. **Retrieval rank.** For each synthetic centroid $\mathbf{s}_T$, rank all 29 real centroids by cosine similarity to $\mathbf{s}_T$. The rank of $\mathbf{r}_T$ in this ordering (1 = nearest neighbour) measures whether PhenoSeq places synthetic embeddings in the correct region of scGPT space.

4. **Specificity score.**

$$\mathrm{spec}(T) = \cos(\mathbf{s}_T, \mathbf{r}_T) - \frac{1}{|\mathcal{T}| - 1} \sum_{T' \neq T} \cos(\mathbf{s}_T, \mathbf{r}_{T'}), \tag{9}$$

   the margin by which the correct-treatment cosine similarity exceeds the mean off-diagonal similarity. Positive specificity with rank 1 jointly confirm that the synthetic centroid is both near to and specifically near to the correct real centroid. Specificity alone is interpretable even when absolute cosine values are uninformative: pairwise cosine similarities between real treatment centroids in scGPT space sit near 0.999, reflecting the geometry of the pre-trained embedding space rather than model behaviour. The specificity score and retrieval rank are invariant to this global cosine bias.

5. **Summary statistics.** Accuracy@1 $= \frac{1}{|\mathcal{T}|} \sum_T \mathbf{1}[\mathrm{rank}(T) = 1]$ and Mean Reciprocal Rank $\mathrm{MRR} = \frac{1}{|\mathcal{T}|} \sum_T \frac{1}{\mathrm{rank}(T)}$, both computed over all 29 conditions. Full per-treatment results are in Table 3.

B. Embedding-Space Fidelity Details

C. Architecture and Hyperparameter Details

C.1. Diffusion Model Architecture

Table 4 summarises the hyperparameters of the PhenoSeq denoiser. The cross-attention transformer uses imaging features as keys/values while the noisy RNA embedding query attends across them.

C.2. Diffusion Model Training

Table 5 summarises the training hyperparameters.

C.3. Classifier Hyperparameters

Table 6 summarises the hyperparameters for all ten evaluation settings. All settings share AdamW optimisation, batch size 4096, maximum 100 epochs, early stopping (patience 25 on validation balanced accuracy), and gradient clipping ($\ell_2 = 0.5$).

Table 3. Per-treatment fidelity of synthetic scGPT embeddings. Rank = retrieval rank of the correct treatment when synthetic mean vectors are matched against all 29 real treatment mean vectors by cosine similarity (1 = nearest neighbour). Specificity $= \cos(\mathbf{s}_T, \mathbf{r}_T) - \overline{\cos(\mathbf{s}_T, \mathbf{r}_{T' \neq T})}$.

Treatment	Rank	Specificity	Euclidean Dist.
Thapsigargin	1	0.0109	0.0478
DBeQ	1	0.0080	0.0258
Cycloheximide	1	0.0079	0.0562
Vorinostat / SAHA	1	0.0074	0.0166
GW-843682X	1	0.0071	0.0200
CGK-733	1	0.0068	0.0150
Colchicine	1	0.0058	0.0175
BAY 11-7082	1	0.0054	0.0224
(R)-Roscovitine	1	0.0042	0.0356
AMG-900	1	0.0039	0.0228
PQ401	1	0.0031	0.0196
Caffeine	1	0.0029	0.0187
Phenacetin	1	0.0029	0.0122
Quinidine	1	0.0028	0.0157
Daporinad / FK-866	1	0.0028	0.0130
Aloxistatin / E-64d	3	0.0029	0.0137
Splitomicin	3	0.0027	0.0167
Benzbromarone	3	0.0028	0.0153
HARMAN	3	0.0028	0.0142
Wy 14643 / Pirinixic Acid	4	0.0029	0.0174
Fluocinonide	4	0.0028	0.0184
DMSO	6	0.0028	0.0193
SKII	11	0.0026	0.0298
Pantoprazole	12	0.0026	0.0247
PD-98059	14	0.0025	0.0284
(R)-MG132	14	0.0024	0.0308
Simvastatin	17	0.0021	0.0405
LY303511 (hydrochloride)	18	0.0018	0.0457
12-O-tetradecanoylphorbol-13-acetate	25	-0.0021	0.1023

Table 4. Diffusion model architecture hyperparameters. The denoiser is a cross-attention transformer: imaging features serve as context (keys/values) while the noisy RNA-seq query attends across them in each of the 6 transformer blocks.

Parameter	Description	Value	Notes
`img_dim`	Input (imaging)	5120	ViT-L features (5 channels $\times$ 1024)
`rna_dim`	Target (RNA)	512	scGPT embedding dimension
`model_dim`	Model dimension d	1024	Internal hidden dimension
`num_heads`	Attention heads	8	Multi-head attention
`num_layers`	Transformer blocks	6	Cross-attention denoising blocks
`time_dim`	Time embedding dim	256	Sinusoidal + MLP time embedding
`ff_mult`	FF expansion factor	4	Feedforward hidden $= 4d$
`dropout`	Dropout	0.10	Applied in attention & FF layers

Table 5. Diffusion model training hyperparameters.

Parameter	Description	Value
`num_steps`	Diffusion steps T	1000
`schedule`	Noise schedule	cosine
`batch_size`	Batch size	256
`lr`	Learning rate	10^{-4}
`weight_decay`	Weight decay	10^{-5}
`epochs`	Max epochs	5000
`warmup_epochs`	LR warm-up	5
`scheduler`	LR schedule	cosine
`max_grad_norm`	Gradient clipping	1.0
`ema_decay`	EMA decay	0.9999
`mixed_precision`	Mixed precision	float16
`seed`	Random seed	42

Table 6. Classifier hyperparameters for all ten evaluation settings. MLP: Linear–LayerNorm–ReLU blocks of width Hidden and depth Depth, followed by a linear output head. DualMLP: separate image ($5120 \to 266 \times 3 \to 913$) and RNA ($512 \to 2748 \times 3 \to 913$) encoders, concatenated before a $1826 \to C$ classification head. Shared across all settings: AdamW, batch size 4096, max 100 epochs, early stopping patience 25 (val balanced accuracy), gradient clipping $\ell_2 = 0.5$, seed 12345.

Modality	Profile	Input dim	Arch.	Hidden	Depth	LR	Weight decay
Image (real)	single	5120	MLP	512	5	10^{-3}	10^{-4}
RNA (real)	single	512	MLP	3934	3	1.49×10^{-4}	9.24×10^{-5}
RNA (synthetic)	single	512	MLP	3934	3	1.49×10^{-4}	9.24×10^{-5}
Image + RNA (real)	single	5120+512	DualMLP	266/2748	3	1.37×10^{-5}	4.52×10^{-5}
Image + RNA (synth.)	single	5120+512	DualMLP	266/2748	3	1.37×10^{-5}	4.52×10^{-5}
Image (real)	multi	5120	MLP	279	2	5.20×10^{-5}	5.79×10^{-3}
RNA (real)	multi	512	MLP	1922	4	1.63×10^{-4}	1.90×10^{-6}
RNA (synthetic)	multi	512	MLP	1922	4	1.63×10^{-4}	1.90×10^{-6}
Image + RNA (real)	multi	5120+512	DualMLP	266/2748	3	1.63×10^{-4}	4.52×10^{-5}
Image + RNA (synth.)	multi	5120+512	DualMLP	266/2748	3	1.63×10^{-4}	4.52×10^{-5}