GENERATIVE MODELING REVEALS THE CONNECTION BETWEEN CELLULAR MORPHOLOGY AND GENE EXPRESSION

Shuo Wen, Ramon Viñas Torné, Maria Brbić
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{shuo.wen,ramon.vinastorne,mbrbic}@epfl.ch

Johannes Bues, Camille Lucie Lambert, Nadia Grenningloh, Timothée Ferrari, Elisa Bugani, Joern Pezoldt, Bart Deplancke
Institute of Bioengineering, School of Life Sciences
EPFL, Lausanne, Switzerland
{johannes.bues,camille.lambert,nadia.grenningloh,timothee.ferrari,elisa.bugani,jorn.pezoldt,bart.deplancke}@epfl.ch

Jillian Rose Love, Wouter Karthaus
Swiss Institute for Experimental Cancer Research, School of Life Sciences
EPFL, Lausanne, Switzerland
{jillian.love,wouter.karthaus}@epfl.ch

ABSTRACT

The understanding of how transcriptional programs give rise to cellular morphology, and how morphological features reflect and influence cell identity and function remains limited. This is due in part to the lack of large-scale datasets pairing the two modalities, as well as the absence of computational frameworks capable of modeling their cross-modal structure. Here, we introduce COSMIC, a bidirectional generative framework that enables quantitative decomposition of transcriptional variance reflected in morphology and morphological variance explained by gene expression. COSMIC builds on a foundation model trained on over 21 million segmented nuclei and couples it with existing transcriptomic embeddings. To enable cross-modal learning, we leveraged a newly generated multimodal dataset acquired using IRIS, a technology that captures high-resolution images and transcriptomes from the same single cells at scale. COSMIC accurately modeled cell type identity, establishing a quantitative link between morphological phenotypes and underlying gene expression. In prostate cancer cells, COSMIC identified morphological and transcriptomic differences between chemotherapy drug treatment-responsive and -resistant cells. Together, these results demonstrate that generative modeling powered by paired single-cell measurements can capture the bidirectional flow of information between cellular form and gene expression, opening new avenues for mechanistic discovery and predictive modeling in both basic and translational cell biology.

1 INTRODUCTION

Deciphering the extent to which transcriptional programs encode cellular form, and, conversely, how morphological variation reflects underlying gene expression remains an open question in cell biology. Single-cell RNA sequencing (scRNA-seq) (Macosko et al., 2015; Tang et al., 2009) provides high-resolution molecular profiles, while imaging technologies (Basiji et al., 2007; Bray et al., 2016; Nitta et al., 2018; Schraivogel et al., 2022) capture the detailed morphological and structural organization of individual cells. Bridging these complementary modalities is essential for deciphering

how transcriptional programs determine cellular form and, conversely, how morphological features affect cell identity, function, and responses to perturbations.

Multimodal computational frameworks could enable translation between transcriptomic and morphological information, revealing how gene-expression programs give rise to cellular phenotypes and how morphological variation reflects underlying molecular states. Progress toward this goal over the past decade has been driven by advances in both biotechnology (Cadwell et al., 2016; Pitino et al., 2025) and computational modeling (Yang et al., 2021; Lee & Welch, 2022). However, the biotechnologies remain limited either in their sensitivity and gene coverage or in the throughput needed for training AI models. At the same time, current computational approaches struggle to capture fine-grained morphological differences between cells. Due to these shortcomings, the quantitative decomposition of how transcriptional variance manifests in morphology and how morphological variance is shaped by gene expression remains unclear.

In this work, we developed COSMIC (**C**ross-m**O**dal generation between **S**ingle-cell RNA-seq and **MIC**roscopy images), a bidirectional generative framework that enables translation between single-cell transcriptomes and cellular morphology. To pair modalities, we built a collection of 17,109 human and 9,039 mouse single cells using the IRIS platform (Bues et al., 2025), a microfluidic system that simultaneously acquires high-resolution microscopy images and high-quality transcriptomic profiles from the same individual cells. We focus on nuclear morphology, which is known to be linked with transcriptional activity, chromatin organization, disease states, and can be used to identify pathologies (Skinner & Johnson, 2017). To capture distinct features of nuclear morphology, we build a foundation model pretrained on 21.8 million segmented nuclear images to capture diverse morphological representations. COSMIC couples this nuclear morphology encoder with transcriptomics encoders (Lopez et al., 2018; Rosen et al., 2023) to enable cross-modal generation using a diffusion model conditioned on the unimodal representations of transcriptome and nuclear morphology.

By enabling bidirectional translation between nuclear morphology and transcriptome, we show that discriminative morphological features can be generated from transcriptomic profiles with high accuracy, and that nuclear morphology, in turn, captures biologically meaningful variation in transcriptomic states. This bidirectional relationship spans multiple levels of biological information, capturing discrete differences between cell types as well as continuous variation associated with cell-cycle progression. We built a paired multi-modal dataset of prostate cancer cells treated with the chemotherapeutic agent Docetaxel and showed that COSMIC enables the identification of genes whose expression is associated with nuclear morphological differences between treatment-responsive and non-responsive cells.

2 Related work

2.1 Multimodal computational frameworks

Multimodal computational frameworks, particularly cross-modal generative models, offer a powerful approach for translating between different data modalities and uncovering the relationships that link them. Large-scale conditional generative models, such as diffusion-based architectures and models in the IMAGEN family, have demonstrated strong capability in synthesizing high-fidelity data in one modality conditioned on another by learning shared latent representations across modalities (Ho et al., 2020; Zhang et al., 2023; Wu et al., 2023).

In the single-cell biology domain, multi-domain translation (MDT) (Yang et al., 2021) leverages adversarial training to align feature spaces without paired data, extending applicability to settings lacking direct correspondence; however, this approach largely overlooks rich biological prior information. In contrast, MorphNet (Lee & Welch, 2022) seeks to predict cellular morphology from transcriptomic profiles using paired samples, but it struggles to capture fine-grained morphological variability across cells. In this work, leveraging paired measurements, we propose COSMIC, a deep generative framework that quantitatively decomposes transcriptional variance reflected in cellular morphology and, conversely, morphological variance explained by gene expression.

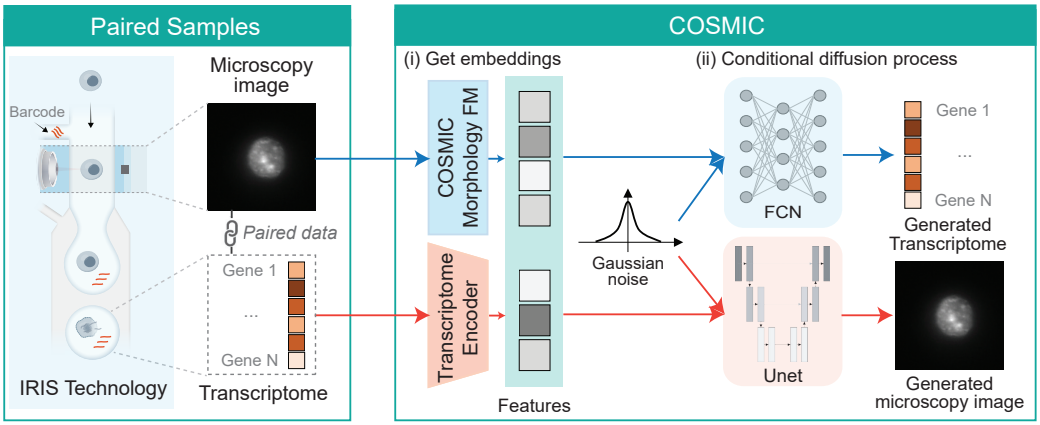

Figure 1: Training pipeline of COSMIC. Conditioned on cell features extracted from microscopy images and single-cell transcriptomes obtained from pretrained unimodal models, COSMIC performs conditional generation using two conditional diffusion models. The diffusion models are trained on paired samples generated with IRIS technology to learn cross-modal relationships between nuclear morphology and the transcriptome.

2.2 Multimodal imaging and transcriptomics data

Over the past years, advances in imaging and transcriptomics technologies have enabled the large-scale acquisition of paired visual and molecular measurements. Large-scale programs such as GTEx (Lonsdale et al., 2013) and TCGA (Weinstein et al., 2013) provide extensive molecular and imaging data that have powered methods for inferring transcriptomic information from histology images (Fu et al., 2020; Schmauch et al., 2020; Alsaafin et al., 2023; Pizurica et al., 2024). At finer resolution, imaging-based spatial transcriptomics (Chen et al., 2015; Eng et al., 2019; Stickels et al., 2021; Jaume et al., 2024) and single-cell multimodal technologies, such as STAMP (Pitino et al., 2025) and Patch-seq (Cadwell et al., 2016), enable coupling of images with molecular readouts at the single-cell resolution. However, these approaches remain constrained by limited sensitivity and gene coverage, or by insufficient throughput for training large-scale AI models. More recently, the IRIS (Bues et al., 2025) platform enables simultaneous acquisition of high-resolution microscopy images and high-quality transcriptomic profiles from the same individual cells, opening new opportunities to study the connection between cellular morphology and gene expression. In this work, we construct a dataset using the IRIS technique and apply COSMIC to this data.

3 Method

We developed COSMIC, a bidirectional generative framework that enables the prediction of gene expression profiles from microscopy images of nuclei, and, conversely, the reconstruction of nuclear images from transcriptomic data. In each direction, COSMIC first encodes the input modality into a feature representation, then conditions the diffusion model on these features to generate the corresponding transcriptome or nuclear image (Figure 1). In this work, we focused on nuclear morphology, as it is widely available across microscopy imaging datasets and known to be associated with cellular function and disease states (Skinner & Johnson, 2017).

3.1 Encoding nuclear morphology and transcriptomic profiles

COSMIC builds on the morphology FM to encode information about nuclear morphology, and leverages existing embedding models to encode transcriptomic profile of cells (Lopez et al., 2018; Rosen et al., 2023; Cui et al., 2024).

Nuclear morphology encoder. To obtain robust single-cell nuclear image embeddings for conditioning the generative model, we train a large-scale model on 21,784,309 segmented nuclear images

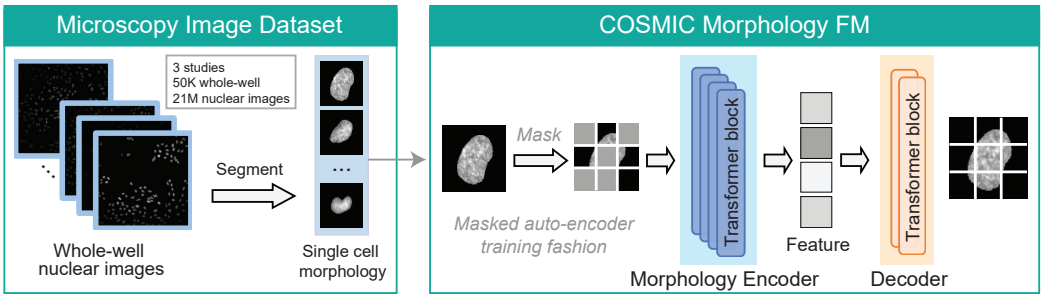

Figure 2: Training pipeline of COSMIC Morphology FM. We collected ∼50.3K whole-well microscopy images stained with Hoechst dyes from three studies. Whole-well images were segmented to obtain single-cell images, resulting in a total of ∼21.8M nuclear images. To learn meaningful morphological representations of cells, we pretrained a masked auto-encoder (MAE) on collected nuclear images. During pretraining, random patches of the input images were masked, and the model was trained to reconstruct the missing regions.

collected from three datasets (Weisbart et al., 2024; Stojic et al., 2020; Pascual-Vargas et al., 2017). We segmented individual cells from the whole-well images by applying an adaptive thresholding procedure (Otsu thresholding (Otsu, 1979)) followed by connected-component analysis to identify distinct cell regions. We then filtered out nuclei whose projected area lies outside the central range of the area distribution, discarding cells below the 5th percentile or above the 95th percentile in nuclear area, as well as nuclei that are severely occluded or truncated at image borders. After filtering, we crop each detected cell into a 128×128 patch centered on its nucleus. We mask out background pixels and neighboring cells to isolate single-cell morphology (Figure 2).

The model is trained in the encoder-decoder architecture. It uses a vision transformer (ViT-Large; 24 layers) as the backbone and consists of 307M parameters. We train the model in a masked autoencoder (MAE) (He et al., 2022) training fashion. We resize the input images to 224×224 and divide them into non-overlapping patches. In particular, during training, given an input image $\mathbf{x}$ from the segmented nuclear images, we apply a random binary mask $\mathbf{m}$ to produce $\mathbf{x}_{\text{masked}} = \mathbf{m} \odot \mathbf{x}$ with a 0.75 masked ratio per image, where $\odot$ denotes Hadamard product. The encoder Enc_{IMG} extracts a representation from the masked patches, and the decoder g_ϕ reconstructs the complete image, yielding $\hat{\mathbf{x}} = g_\phi(\text{Enc}_{\text{IMG}}(\mathbf{x}_{\text{masked}}))$. The training objective minimizes the reconstruction error on the masked regions:

$$\mathcal{L}_{\text{MAE}} = \|(\mathbf{1} - \mathbf{m}) \odot (\mathbf{x} - \hat{\mathbf{x}})\|_2^2.$$

This approach encourages the model to learn biologically meaningful features of nuclear morphology. To leverage these representations for COSMIC, we remove the decoder part and use the learned image embeddings from the last layer of the encoder, $\mathbf{z}_{\text{IMG}} = \text{Enc}_{\text{IMG}}(\mathbf{x})$ as conditioning inputs to the transcriptome generation model. The image embeddings are extracted from the full image without applying any masking during inference.

Representations of transcriptomic profiles. To obtain embeddings of transcriptomic profiles for conditioning the generative model, COSMIC is compatible with existing models for batch effect removal (Lopez et al., 2018; Lotfollahi et al., 2019), as well as single-cell foundation models (Rosen et al., 2023; Cui et al., 2024; Hao et al., 2024). In particular, we use scVI (Lopez et al., 2018), a probabilistic framework based on variational inference, as a default model for our experiments.

3.2 DIFFUSION MODEL CONDITIONED ON CELL REPRESENTATIONS.

Building on the learned representations of nuclear morphology and transcriptomic profiles introduced above, COSMIC employs conditional denoising diffusion probabilistic models (Ho et al., 2020; Zhang et al., 2023) to generate one modality from the other by iteratively refining noise-corrupted inputs. In this setting, a clean sample $\mathbf{x}_0$ corresponds to either a nuclear morphology image or a transcriptomic profile of a single cell. COSMIC operates in two phases: *(i)* a forward diffusion process that gradually adds Gaussian noise to a data sample, and *(ii)* a learned reverse process that reconstructs the original data by denoising. Formally, given a clean sample $\mathbf{x}_0$, the forward process generates a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ by adding noise at each timestep:

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

where β_t is a fixed variance schedule. A closed-form expression allows sampling $\mathbf{x}_t$ directly from $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}),$$

with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

The reverse process is parameterized by a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x_t}, t, \mathbf{c})$, which predicts the noise component $\boldsymbol{\epsilon}$ given the current sample $\mathbf{x}_t$, the timestep t, and a conditioning embedding $\mathbf{c}$ derived from the opposite modality. The model is trained to minimize the denoising objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2\right].$$

COSMIC includes two such conditional diffusion models. The first is the *seq2img* module with an Attention U-Net (Oktay et al., 2018) architecture designed to synthesize nuclear morphology images from gene expression profiles. Here, the clean sample $\mathbf{x}_0$ is a nuclear morphology image and the conditioning input is the transcriptomic embedding $\mathbf{c} = \mathbf{z}_{\text{RNA}}$. The U-Net operates on 256×256 images with multiple resolution levels, and integrates the conditioning embedding at several layers through feature-wise affine transformations and attention. The second is the *img2seq* module with an MLP-based conditional diffusion architecture designed to generate transcriptomic profiles from nuclear morphology images. Here, the clean sample $\mathbf{x}_0$ is a transcriptomic profile, and the conditioning input is the morphology embedding $\mathbf{c} = \mathbf{z}_{\text{IMG}}$ extracted by the morphology FM.

By conditioning on $\mathbf{z}_{\text{IMG}}$ or $\mathbf{z}_{\text{RNA}}$, the diffusion models learn to capture biologically meaningful relationships between morphology and gene expression. Through iterative refinement, COSMIC generates high-fidelity outputs that reflect the underlying structure of both modalities. Appendix A shows the details about hyperparameters and model selection.

4 EXPERIMENT

4.1 SETUP

Dataset We built a dataset of a collection of 17,109 human and 9,039 mouse single cells using the IRIS platform (Bues et al., 2025). For mouse cells, we trained COSMIC on $4,520$ samples and evaluated it on an independent set of $4,519$ samples. We profiled mouse embryonic fibroblasts (3T3), macrophage-like cells (RAW), CAR-engineered T cells derived from the A20 B-cell lymphoma model (CAR_A20), and primary naive CD8+ T cells (naive CD8). The dataset comprises batches generated in our work and from Bues et al. (2025) (Appendix Table 1). For human cells, we trained COSMIC on $8,555$ samples and evaluated it on an independent set of $8,554$ samples. The human data comprises peripheral blood mononuclear cells (PBMCs, including lymphocytes and monocytes) and established cell lines: RPE1 (retinal pigment epithelial cells; named as RPE), C4-2B (prostate cancer cells from a lymph node metastasis; named as PCa-LN), and DU145 (prostate cancer cells from a brain metastasis; named as PCa-Br). Detailed description of the dataset and data pre-processing are shown in Appendix B.

Existing methods We compared COSMIC to two existing deep learning methods: *(i)* MorphNet (Lee & Welch, 2022) is a generative adversarial network based model trained to synthesize images from gene expression, and *(ii)* Multi-Domain Translation (MDT) (Yang et al., 2021) aligns unpaired samples across modalities via representation space matching, without relying on paired data.

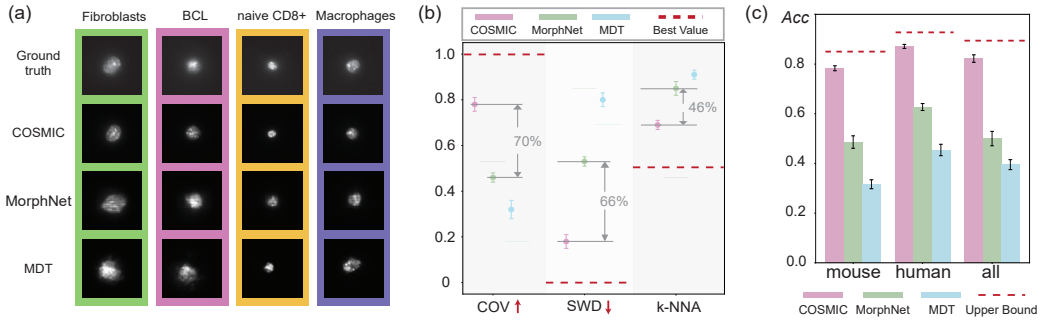

Figure 3: COSMIC generates high-quality nuclear images from transcriptomic profiles of cells. **(a)** Qualitative comparison of COSMIC with MorphNet (Lee & Welch, 2022) and MDT (Yang et al., 2021) on randomly selected mouse cell types. **(b)** Quantitative evaluation on IRIS mouse cells using COV, SWD, and k-NNA. Error bars represent the standard deviation across five independent rounds of generation. Red dashes indicate the best value; numbers denote COSMIC's improvement over the strongest baseline. **(c)** Cell type classification accuracy on generated images for IRIS mouse and human datasets using a classifier trained on ground-truth images. The red dashed line shows the ground-truth upper bound.

4.2 COSMIC GENERATES REALISTIC SINGLE-CELL NUCLEAR IMAGES FROM THE TRANSCRIPTOMIC PROFILES OF CELLS

We quantified the ability of COSMIC to generate nuclear images from transcriptomic profiles of cells. We first examined representative examples of generated images, selecting the cluster centers across diverse cell types. The COSMIC-generated images closely matched real microscopy images in nucleus size, texture, and shape variation, appearing nearly indistinguishable (Fig. 3a). These results demonstrate COSMIC's capacity to generate high-fidelity, cell type-specific nuclear morphologies from single-cell transcriptomes. Conversely, MorphNet produced lower-fidelity samples, while MDT generated outputs that failed to preserve cell type information. Similarly, we observe large improvements in the quality of the generated images on human cells (Appendix D Figure 6a).

To quantify the overall quality of the generated images, we use three metrics widely used in computer vision: coverage (COV) (Achlioptas et al., 2018), sliced Wasserstein distance (SWD) (Bonneel et al., 2015), and k-nearest neighbor accuracy (k-NNA) (Xu et al., 2018) (Appendix C). COV measures how well the generated samples span the real data distribution; SWD assesses statistical divergence between real and generated distributions; and k-NNA evaluates how well the real and generated samples are intermixed. Across all three metrics, COSMIC substantially outperformed both MorphNet and MDT (Fig. 3b), achieving 46% to 70% better performance compared to the best alternative method MorphNet. Similarly, we observe large improvements when applying COSMIC to human cells from the IRIS technology. Similar improvements are shown on human cells (Appendix D Figure 6b).

We next evaluated whether COSMIC-generated images preserve cell type-specific information. To this end, we trained a convolutional neural network (CNN) classifier to predict cell types using real images and applied it on synthetic images generated by COSMIC, MDT, and MorphNet. On the IRIS mouse dataset, we found that classifier applied to COSMIC's generated images achieved an average classification accuracy of 78%, close to the upper bound of 85% observed on the test set of real data (Fig. 3c). In contrast, images generated by alternative methods MorphNet and MDT achieved significantly lower performance (48% and 32%, respectively). We observed comparable performance on a human dataset, where the model achieved 88% accuracy, with an upper bound of 93% on real images. This demonstrates that the COSMIC-generated images retained a strong conditional signal and preserved discriminative features of cell types.

Together, these findings establish that COSMIC's transcriptome-to-image generation pipeline produces diverse nuclear morphologies that are quantitatively consistent with the input gene expression profiles, laying the foundation for a generative understanding of the relationship between cellular morphology and transcriptional state in single cells.

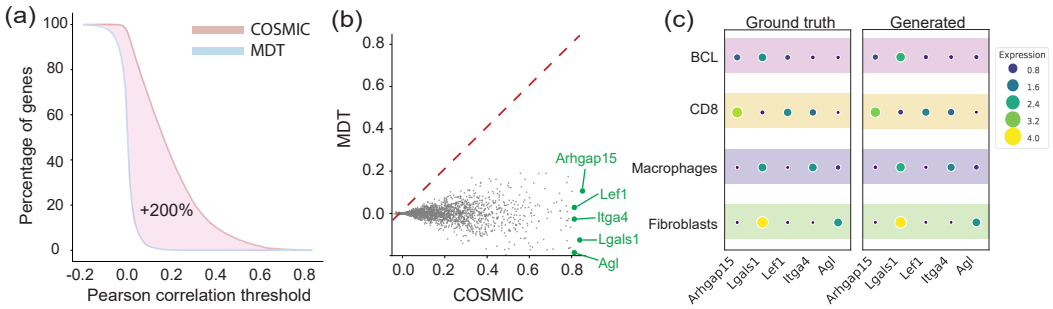

Figure 4: COSMIC generates transcriptomic profiles of cells from nuclear images. **(a)** Evaluation of the generated transcriptomic profiles using Pearson correlation coefficient (ρ) by comparing them to the ground-truth transcriptomes. Higher values indicate better performance. The curve shows the cumulative percentage of genes exceeding a given Pearson correlation threshold. We compare multi-domain translation (MDT) to COSMIC. **(b)** Comparison of Pearson correlation coefficient between COSMIC and MDT for individual genes. Five genes with the highest Pearson correlation are highlighted in green. **(c)** Average gene expression of the five genes with the highest Pearson correlation across different cell types. Ground truth expression (left) and COSMIC predictions (right).

4.3 COSMIC GENERATES TRANSCRIPTOMIC PROFILES OF SINGLE CELLS FROM NUCLEAR IMAGES.

We examined the performance of COSMIC in the reverse direction: predicting transcriptomic profiles from microscopy images of single cells. To evaluate gene-level fidelity, we computed the Pearson correlation coefficient between generated transcriptome profiles and their ground truth counterparts for all genes (details in Appendix C). We compared COSMIC to MDT, the only method designed for this task. Using COSMIC, 12.5% ($n = 1,814$ out of $14,455$ genes) had a correlation coefficient higher than 0.4 (Fig. 4a) compared to MDT, using which none of the genes satisfied this threshold.

When comparing individual genes, COSMIC was particularly predictive for cell type-specific marker genes, for example, some marker genes, including *Lef1*, *Itga4*, *Lgals1*, and *Arhgap15*, displayed correlation coefficients nearing 0.8 across all cells, suggesting that morphology encodes precise expression patterns for a subset of genes with strong phenotypic associations (Fig. 4b).

To understand the biological relevance of these predictions, we identified the top five most accurately predicted genes for each cell type and plotted their average expression across cell types (4c). These gene signatures respected the expected enrichment patterns, confirming that COSMIC could recover cell type-specific expression programs for a subset of genes from nuclear morphology alone. We observe similar results on human cells (Appendix D Figure 7).

Additionally, we show that our Morphology FM provides robust morphology representations that outperform other image foundation models like OpenPhenom (Kraus et al., 2024) and ImageNet MAE (He et al., 2022) (Appendix D Figure 8).

4.4 COSMIC IDENTIFIES MORPHOLOGY-ASSOCIATED GENES.

To evaluate whether COSMIC can uncover biologically meaningful morphology–transcriptome relationships in unstructured and clinically relevant contexts, we next applied it to prostate cancer cells. We treated DU145 cells with the chemotherapeutic agent Docetaxel, a taxane that blocks mitotic progression, induces G2/M arrest, and ultimately triggers apoptosis (Wall & Wani, 1995), and profiled all the treated and untreated cells using IRIS.

We implemented a cycle-consistency strategy that leverages COSMIC's bidirectional architecture: a gene is considered "morphology-associated" if its expression can be reconstructed through an entire inference cycle: from transcriptome to image back to transcriptome with minimal signal loss (Fig. 5a). Intuitively, only genes whose information is encoded in morphological features can pass through this bottleneck. While not implying causality, this criterion highlights genes whose expres-

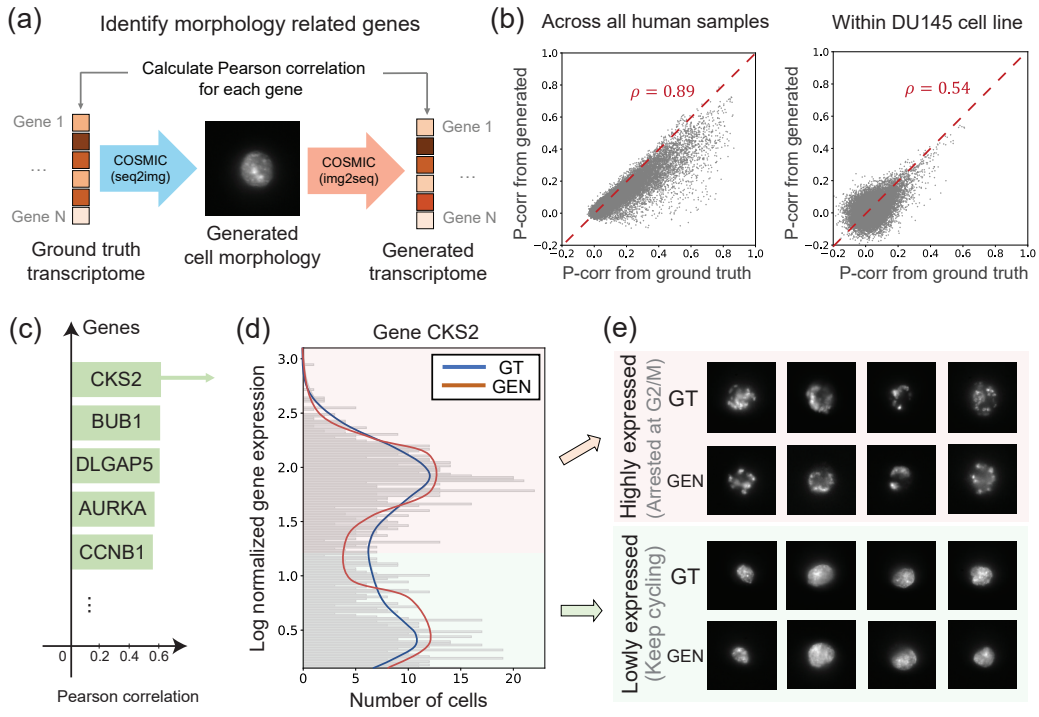

Figure 5: COSMIC identifies morphology-associated genes in prostate cancer cells. **(a)** Cycle generative framework using seq2img and img2seq models to identify genes associated with nuclear morphology. **(b)** Validation on IRIS human and DU145 cells showing gene-wise Pearson correlation between transcriptomes inferred from ground-truth images and COSMIC-generated images (cycle). **(c)** Top morphology-associated genes identified in DU145 cells: *CKS2*, *BUB1*, *DLGAP5*, *AURKA*, and *CCNB1*. **(d)** Distribution of *CKS2* expression in DU145 cells, showing agreement between COSMIC-generated (red) and ground-truth (blue) expressions. **(e)** Representative ground-truth and COSMIC-generated nuclei for cycling and G2/M-arrested cells.

sion strongly covaries with morphological structure. As a validation, we confirmed that COSMIC-generated images retained sufficient information to support complete inference cycles. Transcriptomes predicted from *(i)* real images and *(ii)* COSMIC-generated synthetic images showed strong Pearson correlations across all human cell types in the IRIS dataset (Fig. 5b).

COSMIC identified a set of nuclear morphology-associated genes such as *CKS2*, *BUB1*, *DLGAP5*, *AURKA*, and *CCNB1*, which are all well-known regulators of mitosis and drivers of tumor progression (Toolabi et al., 2022) (Fig. 5c). Their enrichment for mitotic spindle and chromosomal segregation pathways aligns with the distorted G2/M nuclear shapes observed in Docetaxel-responsive cells, consistent with COSMIC detecting cross-modal structure linked to mitotic instability. For *CKS2*, which was the top morphology-associated gene and a regulator frequently upregulated in aggressive prostate cancers (Lai & Lin, 2024), COSMIC recovered two clear expression subpopulations (Fig. 5d). These corresponded to the "responsive" (G2/M-arrested, enlarged nuclei) and "non-responsive" (regular nuclei) Docetaxel groups. One plausible biological interpretation is that *CKS2*-high cells progress rapidly into the vulnerable G2/M window where Docetaxel exerts its effect, while *CKS2*-low cells cycle more slowly and therefore appear less affected at the measured timepoint. Alternatively, the *CKS2*-high cells might represent cells closest to progressing to G2/M in the cell cycle, which would explain that these are the first cells to arrest in G2/M following Docetaxel treatment. We observed similar results for other genes with high correlations (Appendix D Figure 9).

Additionally, COSMIC's bidirectional structure allowed us to generate synthetic nuclear images conditioned on transcriptomes. These generated images recapitulated the expected morphological differences. Specifically, *CKS2*-high cells exhibited enlarged, lobulated nuclei, whereas *CKS2*-low

cells showed compact, uniform nuclei (Fig. 5e). This demonstrates that COSMIC captures subtle but meaningful cross-modal structure and can resolve molecularly divergent subpopulations even within a homogeneous cancer line.

Together, these results show that COSMIC can robustly recover genes whose expression is tightly coupled to nuclear morphology in Docetaxel-treated prostate cancer cells. This highlights COSMIC's ability to uncover cross-modal structure that links transcriptomic state to continuous variation in nuclear architecture.

5 Conclusion

To model and quantify how transcriptional programs relate to nuclear morphology at single-cell resolution, we developed COSMIC, a bidirectional generative framework that quantifies and models the information shared between these two modalities. By training conditional diffusion models on paired transcriptomic and imaging data, COSMIC imposes strong biological priors on the mapping between modalities and enables generation in both directions: synthesizing realistic single-cell nuclear images from gene expression profiles and predicting transcriptomic states from single-cell nuclear images. Through this dual generative capability, COSMIC provides a way to characterize how molecular programs and nuclear morphology are coupled.

Acknowledgments

We are grateful to Duygu Koldere Vilain for making the professional illustration of the IRIS technique. We are also grateful to Natasha Samson for the data on the cell line CAR_A20. We gratefully acknowledge the support of the Swiss National Science Foundation (SNSF) starting grant TMSGI2_226252/1, SNSF grant IC00I0_231922, the Swiss AI Initiative, and the CIFAR Multiscale Human Catalyst. We gratefully acknowledge the support of the Peter und Traudl Engelhorn Foundation to R.V. Figure elements, including icons of species, were created with `BioRender.com`.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pp. 40–49. PMLR, 2018.

Areej Alsaafin, Amir Safarpoor, Milad Sikaroudi, Jason D Hipp, and HR Tizhoosh. Learning to predict RNA sequence expressions from whole slide images with applications for search and classification. *Communications Biology*, 6(1):304, 2023.

David A Basiji, William E Ortyn, Luchuan Liang, Vidya Venkatachalam, and Philip Morrissey. Cellular image analysis and imaging by flow cytometry. *Clinics in Laboratory Medicine*, 27(3): 653–670, 2007.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.

Johannes Bues, Joern Pezoldt, Camille Lucie Lambert, Benjamin David Hale, Elisa Bugani, Ramon Vinas Torne, Timothee Ferrari, Nadia Grenningloh, Vincent Gardeux, Shuo Wen, Caroline Wandinger, Maximilian Kohnen, Romina Augustin, Katharina Eckstein, Assia Ouanaya, Jillian Love, Sarthak Saha, Amirhossein Saba, Aviv Huttner, Maria Vittoria Impagliazzo, Jose Antonio Vasquez Porto Viso, Angel de Jesus Corria Osorio, Demetri Psaltis, Wouter Karthaus, Berend Snijder, Maria Brbic, and Bart Deplancke. Single-cell phenomics through integrated imaging and molecular profiling. *bioRxiv*, 2025. doi: 10.1101/2025.11.28.690954.

Cathryn R Cadwell, Athanasia Palasantza, Xiaolong Jiang, Philipp Berens, Qiaolin Deng, Marlene Yilmaz, Jacob Reimer, Shan Shen, Matthias Bethge, Kimberley F Tolias, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature Biotechnology*, 34(2):199–203, 2016.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Chiaowen Joyce Hsiao, PoYuan Tung, John D Blischak, Jonathan E Burnett, Kenneth A Barr, Kushal K Dey, Matthew Stephens, and Yoav Gilad. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Research*, 30(4):611–621, 2020.

Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

Yueliang Lai and Ye Lin. Biological functions and therapeutic potential of CKS2 in human cancer. *Frontiers in Oncology*, 14:1424569, 2024.

Hojae Lee and Joshua D Welch. MorphNet predicts cell morphology from single-cell gene expression. *bioRxiv*, pp. 2022–10, 2022.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

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

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

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Nao Nitta, Takeaki Sugimura, Akihiro Isozaki, Hideharu Mikami, Kei Hiraki, Shinya Sakuma, Takanori Iino, Fumihito Arai, Taichiro Endo, Yasuhiro Fujiwaki, et al. Intelligent image-activated cell sorting. *Cell*, 175(1):266–276, 2018.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-Net: Learning where to look for the pancreas. In *Conference on Medical Imaging with Deep Learning*, 2018.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Patricia Pascual-Vargas, Samuel Cooper, Julia Sero, Vicky Bousgouni, Mar Arias-Garcia, and Chris Bakal. RNAi screens for rho gtpase regulators of cell shape and YAP/TAZ localisation in triple negative breast cancer. *Scientific Data*, 4(1):1–13, 2017.

Emanuele Pitino, Anna Pascual-Reguant, Felipe Segato-Dezem, Kellie Wise, Irepan Salvador-Martinez, Helena Lucia Crowell, Maycon Marção, Max Ruiz, Elise Courtois, William F Flynn, et al. Stamp: Single-cell transcriptomics analysis and multimodal profiling through imaging. *Cell*, 2025.

Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15(1): 9886, 2024.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023–11, 2023.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, 2020.

Daniel Schraivogel, Terra M Kuhn, Benedikt Rauscher, Marta Rodríguez-Martínez, Malte Paulsen, Keegan Owsley, Aaron Middlebrook, Christian Tischer, Beáta Ramasz, Diana Ordoñez-Rueda, et al. High-speed fluorescence image–enabled cell sorting. *Science*, 375(6578):315–320, 2022.

Benjamin M Skinner and Emma EP Johnson. Nuclear morphologies: their diversity and functional relevance. *Chromosoma*, 126(2):195–212, 2017.

Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 39(3):313–319, 2021.

Lovorka Stojic, Aaron TL Lun, Patrice Mascalchi, Christina Ernst, Aisling M Redmond, Jasmin Mangei, Alexis R Barr, Vicky Bousgouni, Chris Bakal, John C Marioni, et al. A high-content RNAi screen reveals multiple roles for long noncoding RNAs in cell division. *Nature Communications*, 11(1):1851, 2020.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

Narges Toolabi, Fattane Sam Daliri, Amir Mokhlesi, and Mahmood Talkhabi. Identification of key regulators associated with colon cancer prognosis and pathogenesis. *Journal of Cell Communication and Signaling*, 16(1):115–127, 2022.

Monroe E Wall and Mansukh C Wani. Camptothecin and taxol: discovery to clinic—thirteenth bruce f. cain memorial award lecture. *Cancer Research*, 55(4):753–760, 1995.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Erin Weisbart, Ankur Kumar, John Arevalo, Anne E Carpenter, Beth A Cimini, and Shantanu Singh. Cell painting gallery: an open resource for image-based profiling. *Nature Methods*, 2024. doi: https://doi.org/10.1038/s41592-024-02399-z.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, GV Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(1):31, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pp. 3836–3847, October 2023.

A HYPERPARAMETERS AND MODEL SELECTION

Encoder of Gene Expression. In our transcriptomic encoder, we preprocess the dataset (mouse and human separately) by applying library-size normalization, log transformation, and selecting the top 1024 highly variable genes using `experiment_ID` as the batch key. The model is instantiated as an SCVI Lopez et al. (2018) variational autoencoder with a hidden dimension of 128 and a latent dimension of 64. Training is conducted for a maximum of 50 epochs while accounting for batch-specific effects through the experiment ID. The model is trained for a fixed budget of 50 epochs without early stopping or checkpoint-based selection. The final trained model is adopted directly, and its latent representation is used as the condition of the conditional diffusion model. When applying UCE Rosen et al. (2023) as the transcriptomic encoder, we directly use the pretrained 4-layer UCE model with its default configuration, without any further fine-tuning or retraining. For UCE, we perform inference only and use the resulting latent representation as an the conditioning signal for the diffusion model.

Encoder of Cell Morphology. In COSMIC, we pretrain the image encoder as a masked autoencoder built on a ViT-Large backbone with 16×16 patches and 224×224 inputs. We mask a fixed ratio of 0.75 tokens per image and reconstruct the corresponding pixel patches using an MAE decoder configured with a decoder embedding dimension of 512, a single transformer layer, 16 attention heads, an MLP ratio of 4.0, and no projection or attention dropout. Images are resized to 224 and augmented with independent horizontal and vertical flips (probability 0.5 each). Training uses mean-squared error between predicted and target masked patches, the AdamW optimizer with a learning rate of 1.5×10^{-4}, and mini-batches of size 256. We train the MAE for up to 10 epochs without early stopping or validation-based selection. We adopt the final-epoch checkpoint for downstream use.

Decoder of Gene Expression. In COSMIC, we map image-derived embeddings to gene-expression vectors in two training stages. Targets are constructed from the mouse dataset by applying library-size normalization followed by a $\log(1+x)$ transform. In Stage 1, a ViT-L/16 masked autoencoder backbone encodes each image into a 1024-dimensional feature vector (with only the last transformer block trainable), and a linear regressor maps this feature to the full gene-expression vector; we train this baseline with a mean squared error loss using Adam (learning rate 1×10^{-4}), a batch size of 32 for training (and 256 for testing), for 30 epochs. In Stage 2, we freeze the encoder and regressor and train a residual diffusion model on the prediction residuals $r_0 = x_0 - \hat{x}_{\text{base}}$, using a denoiser that is a 3-layer MLP (2048 of hidden dimension) conditioned on the 1024-dimensional image feature, We use a short diffusion horizon of $T = 10$ steps with a linear noise schedule from 5×10^{-5} to 5×10^{-4} and train the residual model with an MSE loss for 10 epochs. We use a fixed train/test split implemented by alternating indices (odd for training, even for testing), evaluate on the test split after each epoch, and adopt the final checkpoint from Stage 2 for downstream use.

Decoder of Cell Morphology. In COSMIC, we train a conditional diffusion image decoder using the Imagen framework to generate 256×256 images from transcriptomic embeddings. The architecture comprises a single UNET with base dimension 32, conditioning dimension 64, and multi-scale widths given by $(1, 2, 4, 8)$. Each scale uses one ResNet block; self- and cross-attention are enabled only at the final scale. Conditioning vectors are 64-dimensional with classifier-free guidance via a drop probability of 0.1. The diffusion process spans 1000 timesteps with prediction objective x_{start}. Inputs are resized to 256 without additional augmentations. Training optimizes the Imagen loss with AdamW at a learning rate of 1.5×10^{-4} and a mini-batch size of 16. We train for up to 100 epochs (iterating over mini-batches) without early stopping or validation-based checkpointing. We adopt the most recent checkpoint for downstream experiments.

B DATASETS AND DATA PRE-PROCESSING

For COSMIC training, we use paired single-cell nuclear images and transcriptomic profiles obtained from the IRIS platform, which enables the simultaneous capture of scRNA-seq data and fluorescence microscopy images from the same individual cells. The dataset includes both mouse and human samples. The mouse dataset contains 9,039 cells from 4 cell lines (3T3, RAW, CAR_A20, and naive CD8$^+$ T cells), spanning 8 batches. The human dataset consists of 17,109 cells drawn from 3 cell lines (DU145, RPE1, C4-2B) and PBMCs (Lymphocytes and Monocytes), spanning 15 batches. Table 1 shows the batches by source for each cell line.

Cell line	**This work (expIDs)**	**From IRIS work Bues et al. (2025) (expIDs)**
Human		
DU145	JP240	-
PBMC	JP247	NG037, NG039, NG050, NG042, JP250, JP268, NG045, NG027
RPE	JP270, NG046	-
C4-2B	JP273, NG047	-
Mouse		
3T3	-	JP241, JP277, NG015, NG026, NG029
Naive CD8+	JP263	-
CAR-A20	JP271	-
RAW	NG033	-

Table 1: Batches by source for each cell line.

Nuclear images are acquired via Hoechst staining and imaged under standard epifluorescence conditions. Raw whole-well images are first corrected for illumination heterogeneity and nuclei are segmented using a deep learning-based nucleus segmentation pipeline. Each segmented nucleus is cropped into a square patch centered on the nucleus centroid, with a fixed input size of 256×256 pixels. Cropped images are intensity-normalized to the $[0, 1]$ range per patch. The final preprocessed nuclear morphology images serve as input to the COSMIC morphology FM and as targets for the seq2img diffusion model.

We process transcriptomic data using standard single-cell RNA-seq workflows. Raw count matrices are filtered to remove low-quality cells (i.e., with fewer than 500 detected genes or high mitochondrial gene content) and genes not expressed in at least 5 cells. We normalize counts using total-count scaling to ensure that each cell has a total count equal to the median across cells, followed by a log1p transformation.

For training the COSMIC morphology FM, we construct a large-scale single-cell morphology dataset by aggregating Hoechst-stained nuclear images from three publicly available sources: the JUMP-Cell Painting (JUMP-CP) dataset Weisbart et al. (2024) (cell line: U2OS), the stojic-lncrnas dataset Stojic et al. (2020) (cell line: HeLa), and the pascualvargas-rhogtpases dataset Pascual-Vargas et al. (2017) (cell line: LM2). From each whole-well image, we perform automated segmentation using a deep learning-based pipeline to identify individual nuclei. Following segmentation, we extract cropped image patches centered on each nucleus using a fixed-size bounding box. Quality control steps included filtering out images with overlapping nuclei, debris, or segmentation errors, ensuring only well-centered and morphologically intact cells were retained. Altogether, this process yielded over 21.7 million high-quality single-nucleus images spanning distinct cell types and experimental perturbations. These curated image patches were used to train the COSMIC morphology FM from scratch with a masked autoencoder (MAE) objective, enabling the model to learn rich and generalizable representations of nuclear morphology.

C EVALUATION METRICS

We evaluated COSMIC using different metrics, including Coverage (COV) Achlioptas et al. (2018), Sliced Wasserstein Distance (SWD) Bonneel et al. (2015), k-Nearest Neighbor Accuracy (k-NNA) Xu et al. (2018), gene-level correlation, and classification accuracy. Unless stated otherwise, all distances are computed in the embedding spaces defined by the morphology FM or the transcriptomic encoder using the Euclidean distance.

Distributional fidelity and diversity. To assess the distributional fidelity and diversity of COSMIC's generated outputs relative to real data, we computed COV, SWD, and k-NNA. Coverage (COV) Achlioptas et al. (2018) measures the proportion of real samples that are the nearest neighbour of at least one generated sample. Formally, given real samples $\mathbf{X}_{\text{real}} = \{\mathbf{x}_i\}_{i=1}^{N}$ and generated samples $\mathbf{X}_{\text{gen}} = \{\hat{\mathbf{x}}_j\}_{j=1}^{M}$, we define

$$\text{COV}_d(\mathbf{X}_{\text{real}}, \mathbf{X}_{\text{gen}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\,\exists\, j \in \{1,\dots,M\} \text{ s.t. } i = \hat{\mathbf{x}}_j \mathbf{X}_{\text{real}}]\,.$$

where $\hat{\mathbf{x}}_j \mathbf{X}_{\text{real}} = \arg\min_{\mathbf{x}\in\mathbf{X}_{\text{real}}} d(\hat{\mathbf{x}}_j, x)$ is the (index of the) nearest real point to $\hat{\mathbf{x}}_j$. In all our experiments, we set $M = 10N$, sampling the model more densely than the data to obtain a more stable empirical estimate of coverage, in line with prior practice of using many more generated than real samples for evaluation Chen & Zhang (2019).

We measure fidelity with the Sliced Wasserstein distance (SWD) Bonneel et al. (2015). Intuitively, SWD compares two point sets by repeatedly *(i)* projecting both sets onto a random one-dimensional direction, *(ii)* sorting the projected points, and *(iii)* averaging the absolute gaps between the two sorted lists. Doing this over many random directions and averaging the results yields a stable distance: a small SWD means that the generated embeddings align well with the real ones.

Formally, let $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ be unit vectors drawn at random from the unit sphere in the embedding space. For each direction $\boldsymbol{\theta}_k$, we project the real and generated samples to obtain one-dimensional point sets

$$u^{(k)} = \{\langle \boldsymbol{\theta}_k, \mathbf{x}_i\rangle\}_{i=1}^{N}, \qquad v^{(k)} = \{\langle \boldsymbol{\theta}_k, \hat{\mathbf{x}}_j\rangle\}_{j=1}^{M}.$$

We then compute the one-dimensional Wasserstein–1 distance $W_1(u^{(k)}, v^{(k)})$ by sorting and pairing the projected points. The sliced Wasserstein distance between the two distributions is defined as the expectation of this quantity over random directions and is approximated in practice by the empirical average

$$\text{SWD}(\mathbf{X}_{\text{real}}, \mathbf{X}_{\text{gen}}) = \frac{1}{K}\sum_{k=1}^{K} W_1\big(u^{(k)}, v^{(k)}\big).$$

To normalize the scores to $[0, 1]$, we divide all SWD values by $\text{SWD}(\mathbf{X}_{\text{real}}, \mathbf{X}_{\text{gauss}})$, where $\mathbf{X}_{\text{gauss}}$ is sampled from a Gaussian distribution with the same mean and variance as $\mathbf{X}_{\text{real}}$.

To quantify how well real and generated samples mix in feature space, we use a k-nearest neighbor accuracy (k-NNA) metric extended from 1-NNA Xu et al. (2018). Let $\mathbf{X} = \mathbf{X}_{\text{real}} \cup \mathbf{X}_{\text{gen}}$ denote the union of real and generated samples, and let $y(x) \in \{0, 1\}$ be the domain label (real or generated). For each sample $\mathbf{x} \in \mathbf{X}$, we consider its k nearest neighbors $\{\mathbf{x}_{(i)}\}_{i=1}^{k}$ in $\mathbf{X} \setminus \{\mathbf{x}\}$ and compute the fraction of neighbors that come from the same domain. The k-NNA score is defined as

$$\text{k-NNA}_k = \mathbb{E}_{\mathbf{x}\in\mathbf{X}}\left[\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\big[y(\mathbf{x}_{(i)}) = y(\mathbf{x})\big]\right]. \tag{1}$$

For balanced real and generated sets with matching cardinality (that is, $|\mathbf{X}_{\text{real}}| = |\mathbf{X}_{\text{gen}}|$), an ideal generator yields k-NNA$_k \approx 0.5$, which indicates that local neighborhoods are well mixed across domains. Larger values of k-NNA$_k$ indicate that neighborhoods are dominated by a single domain and that the two distributions are more easily separable. In all experiments, we set $k = 0.05 \times (|\mathbf{X}_{\text{real}}| + |\mathbf{X}_{\text{gen}}|)$, that is, we use the 5% nearest neighbors of the pooled dataset.

Gene-level correlation. To evaluate transcriptome generation from images, we computed the Pearson correlation between predicted and ground-truth gene expression vectors. When evaluating across cell lines, for each gene g, we define

$$\text{P-corr}_g = \text{PearsonCorr}(\mathbf{x}_g^{\text{RNA}}, \hat{\mathbf{x}}_g^{\text{RNA}}),$$

where $\mathbf{x}_g^{\text{RNA}}$ and $\hat{\mathbf{x}}_g^{\text{RNA}}$ are the real and generated expression values for gene g across cells. The resulting distribution of correlation scores reflects gene-level accuracy and informs the extent to which morphological features constrain transcriptional variability.

When evaluating within each cell line, we calculate the Pearson correlation separately for each experimental batch and then average the correlation scores across batches. This reduces the influence of batch effects on the evaluation scores. Where appropriate, we assess statistical significance using the standard two-sided t-test for the Pearson correlation coefficient, with multiple-hypothesis correction applied to the resulting adjusted p-values.

Cell-type classification. To assess how well generated images capture cell type information, we trained a cell type classifier on real data and evaluated its performance on generated data. Specifically, we train an image cell type classifier (a four-layer convolutional backbone followed by a two-layer fully connected head) on ground-truth nuclear images and tested it on generated images to determine whether cell type discriminative visual features are preserved. Similarly, for transcriptome prediction, we train a two-layer multilayer perceptron (MLP) as the transcriptome cell type classifier and test it on generated transcriptomes to evaluate whether cell type discriminative molecular signatures are preserved.

Cell cycle angular score calculation. Following Hsiao et al. (2020), we assign each cell a continuous cell cycle phase on the interval $[0, 2\pi)$. We first apply a log1p transformation to FUCCI intensities, then project them onto the orthonormal basis given by the first two principal components (PC1 and PC2). For each cell, we compute the angle

$$\theta = \text{arctan2}(\text{PC2}, \text{PC1}),$$

which yields a continuous angular score along the cell cycle trajectory. The orientation of the angle is chosen such that phases assigned from FUCCI gating (for example G1, S, G2, M) align with the expected order as described in Hsiao et al. (2020).

D ADDITIONAL RESULTS

D.1 COSMIC GENERATES REALISTIC SINGLE-CELL NUCLEAR IMAGES

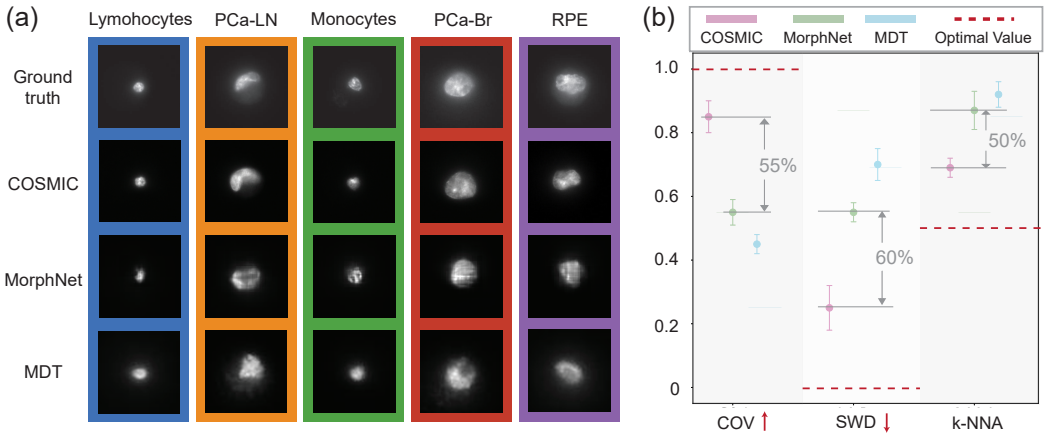

Figure 6: COSMIC generates high-quality nuclear images from transcriptomic profiles of human cells. **(a)** Qualitative comparison of COSMIC with MorphNet Lee & Welch (2022) and MDT Yang et al. (2021) on randomly selected mouse cell types. **(b)** Quantitative evaluation on IRIS mouse cells using COV, SWD, and k-NNA. Error bars represent the standard deviation across five independent rounds of generation. Red dashes indicate the best value; numbers denote COSMIC's improvement over the strongest baseline.

D.2 COSMIC GENERATES TRANSCRIPTOMIC PROFILES OF SINGLE CELLS

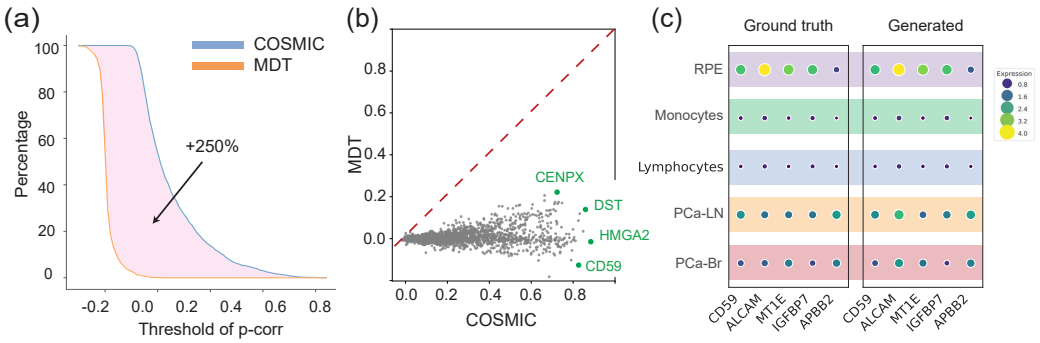

Figure 7: COSMIC generates transcriptomic profiles of cells from nuclear images on human data. **(a)** Cumulative distribution of per-gene Pearson correlations on human data comparing COSMIC to the MDT baseline; COSMIC yields a substantially larger fraction of positively correlated genes, corresponding to a 250% accumulated improvement from a correlation threshold of 0. **(b)** Per-gene performance comparison (MDT vs COSMIC); points below the diagonal indicate COSMIC outperforms MDT. Example genes are annotated (*CENPX*, *DST*, *HMGA2*, *CD59*). **(c)** Cell-type–level heat map of top predictable genes, showing that COSMIC preserves characteristic expression patterns relative to ground truth, consistent with accurate recovery of cell-type signatures.

D.3 MORPHOLOGY FM OUTPERFORMS THE OTHER IMAGE ENCODERS

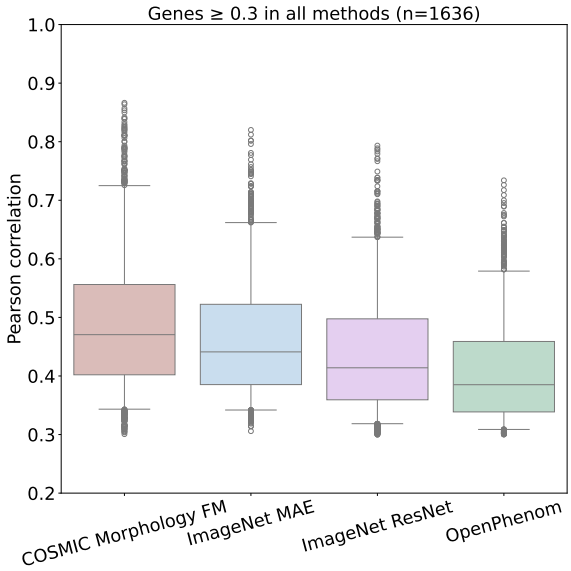

Figure 8: Comparison of four image encoders for generating nuclear images from transcriptomic profiles: COSMIC Morphology FM, ImageNet MAE He et al. (2022), ImageNet ResNet **?**, and OpenPhenom Kraus et al. (2024). Boxplots showing the distribution of Pearson correlations for genes consistently well-predicted across all four methods. Genes with correlation ≥ 0.3 in every method were retained (n = 1636). Each boxplot summarizes correlations for the same set of genes across different methods. Each box indicates the interquartile range (25th to 75th percentiles) with the central line marking the median. Whiskers extend to the 5th and 95th percentiles.

D.4 COSMIC IDENTIFIES MORPHOLOGY-ASSOCIATED GENES

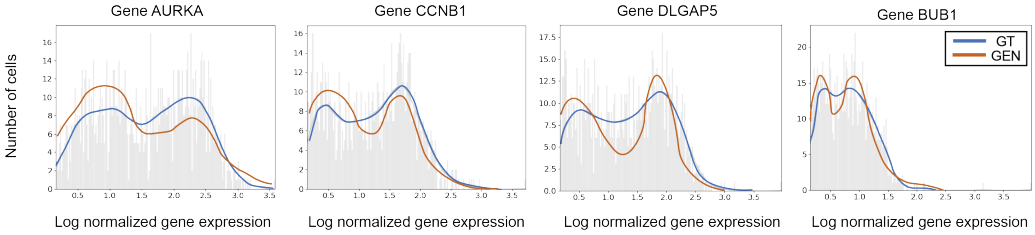

Figure 9: Expression histograms for additional top morphology-associated genes in DU145 cells, including *AURKA*, *CCNB1*, *DLGAP5*, and *BUB1*. For each gene, the histogram shows its normalized expression level (y-axis) as a function of the number of DU145 cells (x-axis), comparing COSMIC-generated expressions from nuclear morphology (red) with ground-truth expressions (blue).