TRIDENT: A Trimodal Cascade Generative Framework for Drug and RNA-Conditioned Cellular Morphology Synthesis

Rui Peng[1,2#] Ziru Liu[5#] Lingyuan Ye[6,7] Yuxing Lu[1] Boxin Shi[3,4*] Jinzhuo Wang[1*]

[1] Department of Big Data and Biomedical AI, College of Future Technology, Peking University

[2] Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University

[3] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[4] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[5] Yuanpei College, Peking University   [6] School of Life Sciences, Tsinghua University

[7] Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program (PTN), Tsinghua University

{pengrui, lzr, luyx}@stu.pku.edu.cn   yely23@mails.tsinghua.edu.cn

shiboxin@pku.edu.cn   wangjinzhuo@pku.edu.cn

Abstract

Accurately modeling the relationship between perturbations, transcriptional responses, and phenotypic changes is essential for building an AI Virtual Cell (AIVC). However, existing methods typically constrained to modeling direct associations, such as Perturbation → RNA or Perturbation → Morphology, overlook the crucial causal link from RNA to morphology. To bridge this gap, we propose TRIDENT, a cascade generative framework that synthesizes realistic cellular morphology by conditioning on both the perturbation and the corresponding gene expression profile. To train and evaluate this task, we construct MorphoGene, a new dataset pairing L1000 gene expression with Cell Painting images for 98 compounds. TRIDENT significantly outperforms state-of-the-art approaches, achieving up to 7-fold improvement with strong generalization to unseen compounds. In a case study on docetaxel, we validate that RNA-guided synthesis accurately produces the corresponding phenotype. An ablation study further confirms that this RNA conditioning is essential for the model's high fidelity. By explicitly modeling transcriptome–phenome mapping, TRIDENT provides a powerful in silico tool and moves us closer to a predictive virtual cell.

1. Introduction

High-throughput omics and artificial intelligence are advancing the vision of AI Virtual Cell (AIVC), a digital twin aiming to simulate cellular behavior across diverse states and scales [7]. Foundational to realizing this vision requires

---

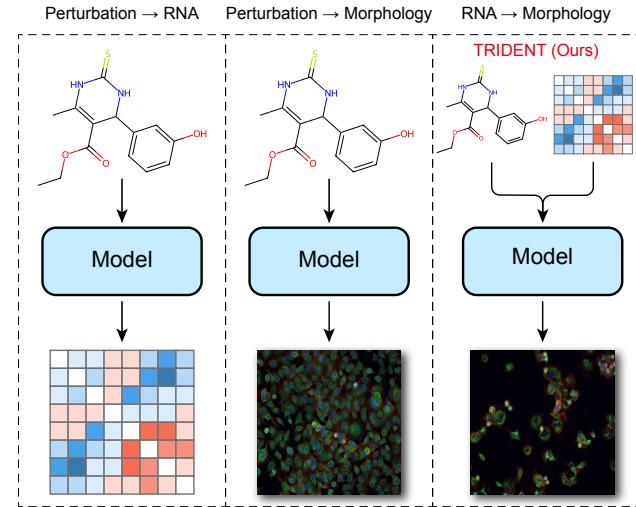# Equal contribution.   * Corresponding authors.

Figure 1. A comparison of cellular response modeling tasks. (Left) Predicting RNA from perturbation. (Middle) Predicting morphology from perturbation. (Right) Our model, TRIDENT, which integrates both perturbation and RNA to predict morphology, explicitly learning the *RNA → Morphology* relationship.

accurately modeling a cell's response to perturbations, which involves three key elements: the perturbation itself, the resulting gene expression (RNA) changes, and the ultimate phenotypic outcome in cellular morphology [49].

While technologies like L1000 profiling [43] and Cell Painting [5] have enabled models that predict *Perturbation → RNA* [1, 17, 36, 39] or *Perturbation → Morphology* [31, 34, 51], they overlook the fundamental *RNA → Morphology* (Fig. 1). This gap limits our ability to simulate the virtual cell as an integrated system where molecular events mechanistically drive phenotypic outcomes. Mean-

while, generative models, particularly diffusion models [12, 20, 21, 32, 33, 35, 41] and VAEs [3, 19, 23, 29, 47], have shown great potential in biological image synthesis and RNA reconstruction [13, 24, 25, 36, 37, 50]. These advances demonstrate the potential of deep generative models to capture intricate biological patterns. However, within the specific context of virtual cell modeling, their application has not yet addressed the critical challenge of explicitly learning the cross-modal mapping from transcriptome to phenome.

To bridge this critical gap, we introduce TRIDENT (**TR**anscription-drug **I**nformed latent **D**iffusion **E**mbedding **NeT**work for cellular morphology synthesis), a cascade generative framework that models the complete perturbation-RNA-morphology relationship (Fig. 1). TRIDENT first uses a VAE to integrate drug and RNA profiles into a unified latent embedding, which then conditions a Diffusion Transformer (DiT) [35] to synthesize high-fidelity cellular morphology. This explicit *(Perturbation + RNA) → Morphology* pathway enables a more mechanistic approach to virtual cell modeling. Our contributions are summarized as follows:

- We introduce TRIDENT, a novel cascade generative framework that, to our knowledge, is the first to model the complete tripartite relationship between perturbation, RNA, and cellular morphology by explicitly learning the fundamental RNA → Morphology mapping.
- TRIDENT generates cellular morphologies with state-of-the-art fidelity, significantly outperforming existing models in both in-distribution and challenging out-of-distribution settings.
- We construct MorphoGene, a new, paired trimodal dataset that integrates gene expression and cellular morphology data for 98 small-molecule drugs.

2. Related Work

Perturbation → RNA prediction. Significant effort addresses the prediction of transcriptomic responses to genetic or chemical perturbations. These approaches employ architectures like graph neural networks (GEARS [39]), encoder-decoders (chemCPA [17], PRnet [36]), and transformers (STATE [1]). A complementary line of VAE-based models (e.g., scGen [24], Dr.VAE [37], CoupleVAE [50], OntoVAE [13]) directly learns counterfactual gene expression. While effective at the *Perturbation → RNA* mapping, these models stop short of linking molecular changes to downstream phenotypes.

Perturbation → Morphology Prediction. Predicting morphological responses is a growing application of conditional image generation, often leveraging diffusion models. Pioneering works generated histopathology from RNA (RNA-CDM [9]) or synthesized cellular structures [30]. Current methods aim to simulate 2D or 3D post-perturbation phenotypes directly, including MorphoDiff [31], IMPA [34],

Mol2Image [51], and DISPR [48]. These approaches, however, typically bypass the intermediate molecular state. They do not explicitly model how transcriptional changes orchestrate morphological alterations, leaving the fundamental *RNA → Morphology* mapping as a black box.

3. Method

TRIDENT models the mapping from gene expression to cellular morphology, conditioned on a drug perturbation D. We formally define this as learning the conditional probability $p(I \mid G_{pre}, D)$, where I is the morphology and G_{pre} is the pre-perturbation gene expression. As shown in Fig. 2, we use a two-stage cascade architecture to model this complex, cross-modal relationship. To train this model, we construct MorphoGene, a novel trimodal dataset integrating Cell Painting data from BBBC021 [8] with L1000 gene expression profiles [43]. We will sequentially describe the dataset construction, the training and inference process in the following sections.

3.1. MorphoGene Dataset Construction

MorphoGene integrates morphology from BBBC021 (MCF7 breast cancer cell line) and gene expression from L1000, linked by 98 small-molecule perturbagens. For morphology data, we merge the DAPI (blue), tubulin (green), and actin (red) channels into RGB composites and cropping them to 512x512. For gene expression, we averaged all corresponding L1000 profiles for each of the 98 compounds into a single representative vector. We then augmented the image collection for each compound to 1,000 samples, creating a total corpus of 98,000 trimodal samples.

We partitioned these 98 compounds to create training, in-distribution (ID), and out-of-distribution (OOD) test sets to evaluate the model's generalization capabilities:
- **Training and ID Cohort:** 44 compounds present in both datasets. For each of these compounds, their samples were split 8:2 to form the training set and the ID test set.
- **OOD Cohort:** The remaining 54 compounds were held out entirely, with all associated samples forming the OOD test set.

Each sample in the final dataset contains the drug D, image I, and the corresponding averaged pre- and post-perturbation gene expression profiles G_{pre}, G_{post}.

3.2. Transcription-Drug Condition Module

The core function of this module is to compute a comprehensive conditional embedding z that encodes both the cell's molecular state and the applied perturbation. As depicted in Fig. 2c, this module is structured as a VAE architecture where the encoder maps the inputs (G_{pre}, D) to a latent space, and the decoder reconstructs the outcome (G_{post}). Formally, the inputs $G_{pre} \in \mathbb{R}^{N_{genes}}$ and the drug's molecular representation D (e.g., a SMILES string) are projected
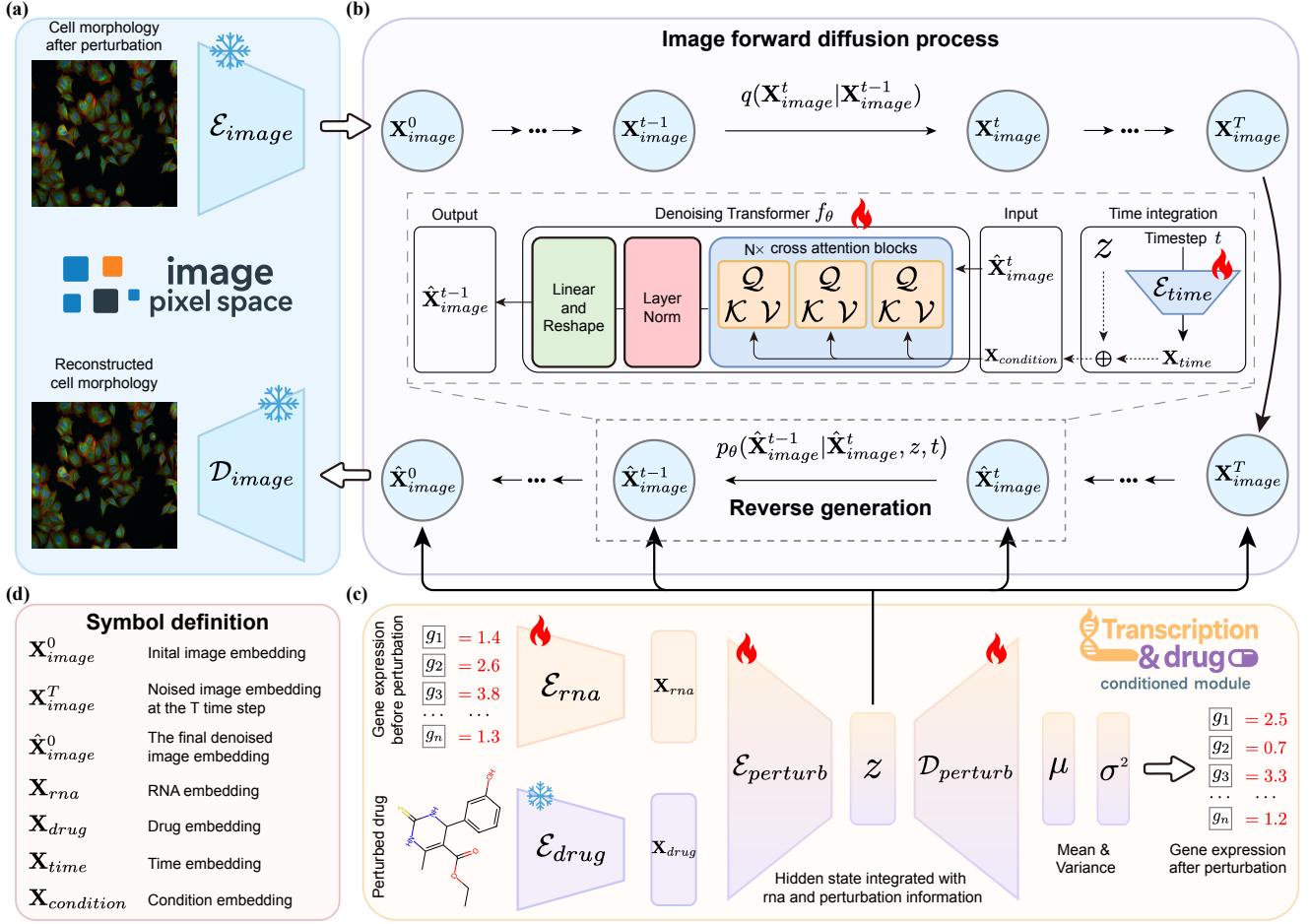
2

Figure 2. Overview of the TRIDENT framework. (a) A VAE maps the high-resolution morphology images from pixel space into a compressed latent representation. (b) The Morphology Generation Module. A denoising transformer learns to reverse a forward noising process, using cross-attention to integrate a guiding condition vector that combines RNA-drug latent and time information. (c) The Transcription-Drug Condition Module. A VAE-based module encodes pre-perturbation gene expression and drug information into a latent vector, which is used to guide image generation. (d) Symbol definitions.

into their respective embedding spaces via dedicated encoders: $\mathbf{X}_{rna} = \mathcal{E}_{rna}(G_{pre})$, $\mathbf{X}_{drug} = \mathcal{E}_{drug}(D)$.

These two embeddings are concatenated and processed by a perturbation encoder $\mathcal{E}_{perturb}$ to parameterize a posterior distribution $q_\phi(z|G_{pre}, D)$, which is modeled as a diagonal Gaussian:

$$[\mu_z, \log \sigma_z^2] = \mathcal{E}_{perturb}([\mathbf{X}_{rna}, \mathbf{X}_{drug}]) , \qquad (1)$$

The latent vector z is then sampled using the reparameterization trick: $z = \mu_z + \sigma_z \odot \epsilon_z$, where $\epsilon_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To ensure z captures biologically salient information regarding the perturbation's effect, we regularize the latent space by forcing it to predict the resulting post-perturbation gene expression G_{post}. A decoder $\mathcal{D}_{perturb}$ models the likelihood $p_\psi(G_{post}|z)$, also as a Gaussian, outputting its parameters μ and σ^2:

$$[\mu_{G_{post}}, \log \sigma_{G_{post}}^2] = \mathcal{D}_{perturb}(z) , \qquad (2)$$

Finally, this module is trained by maximizing the Evidence Lower Bound (ELBO), which consists of a reconstruction term and a Kullback-Leibler (KL) divergence term. The corresponding loss function $\mathcal{L}_{VAE}$ is:

$$\mathcal{L}_{VAE} = \underbrace{\mathbb{E}_{q_\phi(z|G_{pre}, D)}[-\log p_\psi(G_{post}|z)]}_{\text{Reconstruction Loss}} + \underbrace{D_{KL}(q_\phi(z|G_{pre}, D) \,\|\, p(z))}_{\text{Regularization (KL Divergence)}} , \qquad (3)$$

The prior $p(z)$ is a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Minimizing this loss forces z to be a compact representation that is predictive of the post-perturbation molecular state and retain the critical information linking the initial state (G_{pre}), the perturbation (D), and the resultant molecular outcome (G_{post}). This latent vector z, serving as the condition, is then passed to the Morphology Generation Module to guide the final image synthesis.

3

3.3. Morphology Generation Module

This module generates the high-resolution cell morphology $\hat{\mathbf{X}}^0_{image}$ conditioned on $\mathbf{X}_{condition}$. We employ the latent diffusion model (LDM) framework, performing the diffusion process within a compressed latent space for computational tractability.

Image Latent Space. As shown in Fig. 2a, a pre-trained VAE, consisting of an encoder $\mathcal{E}_{image}$ and a decoder $\mathcal{D}_{image}$, is utilized. A high-resolution image I is first encoded into a latent representation $\mathbf{X}^0_{image} = \mathcal{E}_{image}(I)$. The diffusion process operates entirely on $\mathbf{X}^0_{image}$. The final generated latent $\hat{\mathbf{X}}^0_{image}$ is then transformed back to pixel space via $\hat{I} = \mathcal{D}_{image}(\hat{\mathbf{X}}^0_{image})$.

Forward Diffusion Process (q). Following Fig. 2b, the training process is anchored by a fixed forward diffusion process q, which incrementally corrupts the initial image latent $\mathbf{X}^0_{image}$ over T discrete timesteps. This process is defined as a Markov chain that gradually adds Gaussian noise according to a pre-defined variance schedule $\{\beta_t \in (0, 1)\}^T_{t=1}$:

$$q(\mathbf{X}^t_{image} \mid \mathbf{X}^{t-1}_{image}) = \mathcal{N}(\mathbf{X}^t_{image}; \sqrt{1 - \beta_t}\mathbf{X}^{t-1}_{image}, \beta_t\mathbf{I}) , \quad (4)$$

A key property of this Markov chain is that we can sample the latent state $\mathbf{X}^t_{image}$ at any arbitrary timestep t in a closed form, conditioned only on the initial state $\mathbf{X}^0_{image}$. Using the notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod^t_{s=1} \alpha_s$, the distribution of $\mathbf{X}^t_{image}$ is given by:

$$q(\mathbf{X}^t_{image} \mid \mathbf{X}^0_{image}) = \mathcal{N}(\mathbf{X}^t_{image}; \sqrt{\bar{\alpha}_t}\mathbf{X}^0_{image}, (1 - \bar{\alpha}_t)\mathbf{I}) , \quad (5)$$

This allows us to directly generate a noisy sample for any timestep t by sampling a standard Gaussian noise variable $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the reparameterization:

$$\mathbf{X}^t_{image} = \sqrt{\bar{\alpha}_t}\mathbf{X}^0_{image} + \sqrt{1 - \bar{\alpha}_t}\epsilon , \quad (6)$$

This equation represents the explicit noising process, which provides the noisy inputs for training the model.

Reverse Denoising Process (p_θ) and Objective Function. The objective of training is to learn a reverse process $p_\theta(\hat{\mathbf{X}}^{t-1}_{image} \mid \hat{\mathbf{X}}^t_{image}, z, t)$ that can invert the diffusion, effectively learning to denoise the corrupted latents. This reverse process p_θ is parameterized as a Gaussian:

$$\mathcal{N}(\hat{\mathbf{X}}^{t-1}_{image}; \mu_\theta(\hat{\mathbf{X}}^t_{image}, z, t), \Sigma_\theta(\hat{\mathbf{X}}^t_{image}, z, t)) , \quad (7)$$

Following the DDPM framework [21], we set the covariance to untrained constants $\Sigma_\theta = \sigma^2_t\mathbf{I}$, where σ^2_t is typically set to β_t. The mean μ_θ is parameterized to predict the noise ϵ that was added during the forward process.

As shown in Fig. 2b, we implement this noise predictor as a denoising transformer f_θ. The predicted mean μ_θ is then derived from this noise prediction:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{X}}^t_{image} - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}f_\theta(\hat{\mathbf{X}}^t_{image}, \mathbf{X}_{condition})\right) , \quad (8)$$

The network f_θ takes two inputs: (1) The noisy image latent $\hat{\mathbf{X}}^t_{image}$ and (2) The latent embedding condition embedding $\mathbf{X}_{condition}$.

As depicted in the Time integration block of Fig. 2b, the condition vector $\mathbf{X}_{condition}$ is formed by the element-wise addition of the latent vector z (from the Transcription-Drug Condition module) and the time embedding $\mathbf{X}_{time}$, which is generated by passing the timestep t through an embedding layer $\mathcal{E}_{time}$.

The denoising transformer f_θ consists of N stacked blocks. Each block integrates the condition $\mathbf{X}_{condition}$ via a cross-attention mechanism, where the image representation provides the queries (Q) while $\mathbf{X}_{condition}$ provides the keys (K) and values (V). This repeated application of cross-attention throughout the network's depth is the critical mechanism that forces the model to learn the complex association between the RNA-drug condition and the morphological features. The model f_θ is then optimized via a simplified L_2 objective:

$$\mathcal{L}_{LDM} = \mathbb{E}_{t, \mathbf{X}^0_{image}, \epsilon, z}\left[||\epsilon - f_\theta(\hat{\mathbf{X}}^t_{image}, \mathbf{X}_{condition})||^2\right],$$
$$(9)$$

where $\hat{\mathbf{X}}^t_{image} = \sqrt{\bar{\alpha}_t}\mathbf{X}^0_{image} + \sqrt{1 - \bar{\alpha}_t}\epsilon$.

The final, end-to-end training objective for TRIDENT is the combined loss from both modules. We optimize the sum of the VAE loss and the LDM loss:

$$\mathcal{L}_{TRIDENT} = \mathcal{L}_{VAE} + \mathcal{L}_{LDM} , \quad (10)$$

Optimizing this total loss jointly trains the framework to learn a conditional latent space z that is both predictive of molecular outcomes (G_{post}) and informative for guiding the synthesis of high-fidelity morphological images (I). See the supplementary material for detailed pseudocode.

3.4. Inference Procedure

The inference process generates a novel cellular morphology $\hat{I}$ by reversing the diffusion process. The procedure begins by sampling an initial latent variable from the prior distribution $\hat{\mathbf{X}}^T_{image} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given the pre-perturbation gene expression G_{pre} and drug D, the Transcription-Drug Condition Module first computes the condition vector z.

The model then iteratively denoises the latent variable $\hat{\mathbf{X}}^T_{image}$ for $t = T, T - 1, ..., 1$, using the learned conditional distribution $p_\theta(\hat{\mathbf{X}}^{t-1}_{image} \mid \hat{\mathbf{X}}^t_{image}, z, t)$. Each step of this reverse Markov chain computes a slightly less noisy

Table 1. Quantitative comparison of TRIDENT against SOTA baselines on the in-distribution and out-of-distribution test sets. Performance is measured by FID, KID, and IS. Lower scores indicate better performance.

Methods	In-Distribution			Out-of-Distribution		
	FID↓	KID↓	IS↓	FID↓	KID↓	IS↓
MorphoDiff	250.290	0.248	2.614	387.135	0.436	2.747
Stable Diffusion	354.576	0.378	2.792	393.129	0.543	2.932
TRIDENT (ours)	**49.770**	**0.013**	**2.240**	**126.150**	**0.222**	**2.523**

latent $\hat{\mathbf{X}}_{image}^{t-1}$ from the previous $\hat{\mathbf{X}}_{image}^{t}$. Using our noise-prediction parameterization f_θ, the update step is given by:

$$\hat{\mathbf{X}}_{image}^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{\mathbf{X}}_{image}^{t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(\hat{\mathbf{X}}_{image}^{t}, \mathbf{X}_{condition}) \right) + \sigma_t \boldsymbol{\epsilon}' , \tag{11}$$

Here, the coefficients α_t and $\bar{\alpha}_t$ are derived from the fixed variance schedule $\{\beta_t\}_{t=1}^{T}$ of the forward process. ϵ' is a random Gaussian noise sample $\epsilon' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $t > 1$, and $\epsilon' = 0$ for $t = 1$.

After T steps, this process yields the final denoised latent representation $\hat{\mathbf{X}}_{image}^{0}$. Finally, this latent representation is passed through the image decoder $\mathcal{D}_{image}$ to reconstruct the final high-resolution morphology image: $\hat{I} = \mathcal{D}_{image}(\hat{\mathbf{X}}_{image}^{0})$.

4. Experiments

4.1. Comparison with State-of-the-Art Methods

We evaluate image quality generated by TRIDENT against two state-of-the-art (SOTA) diffusion methods: MorphoDiff and a fine-tuned unconditional Stable Diffusion. All models were trained on our MorphoGene dataset for 10,000 steps.

Qualitatively, TRIDENT captures complex, drug-specific cellular patterns across six distinct perturbations (Fig. A3). Many of these compounds are inhibitors that impede cell proliferation (e.g., by disrupting cytoskeletal dynamics), leading to phenotypes with reduced cell density. TRIDENT's generated images are virtually indistinguishable from ground-truth, replicating unique phenotypes like changes in cell shape and population density. For instance, it uniquely generates the characteristic low cell density for cytochalasin b [27, 46]. In contrast, both baselines fail to produce specific phenotypes, collapsing to a generic, high-density monolayer and thus failing to learn the conditional guidance.

Quantitatively, we benchmark using Fréchet Inception Distance (FID) [18], Kernel Inception Distance (KID) [4], and Inception Score (IS) [40] on ID and OOD test sets (Tab. 1). Lower scores are preferable for all three metrics.

FID and KID measure distributional similarity between generated and real images (lower = higher fidelity). For IS, in this context, a lower score is preferable as it indicates that the model has learned to consistently generate the specific, constrained phenotype for a given condition, rather than an overly broad range of morphologies.

TRIDENT demonstrates a profound improvement over both baselines (Tab. 1). On the ID test set, its FID (49.770) represents a 5- to 7-fold improvement over MorphoDiff (250.290) and Stable Diffusion (354.576). A large performance gap is also seen in the KID metric (0.013 vs. 0.248 and 0.378). Crucially, this superiority extends to the OOD task, which evaluates generalization to unseen compounds. Here, TRIDENT's FID (126.150) is more than 3-fold better than the SOTA baselines (387.135 and 393.129).

This quantitative and qualitative evidence confirms that TRIDENT's explicit modeling of the *(Perturbation + RNA) → Morphology* pathway enables superior fidelity and stronger alignment to the target perturbation's phenotypic space, for both known and unseen compounds.

4.2. High Dimensional Feature Analysis

Building upon the validation of high-fidelity image generation, we next assess if TRIDENT captures biologically interpretable morphological signatures in high-dimensional embedding and feature spaces. We project generated images of compounds with diverse Mechanism of Action (MOA) into a pre-trained Vision Transformer (ViT) [14] embedding space and visualize them via Linear Discriminant Analysis (LDA) [2] (Fig. 4a). The results reveal that generated images form distinct, tightly-clustered groups corresponding to their specific drug perturbations. For instance, the filamentous morphology generated for the kinase inhibitor staurosporine [6, 10, 22, 45] is clearly separated from the sparse, rounded-cell phenotype predicted for the protein synthesis inhibitor emetine [15, 16, 38], confirming the model's ability to capture distinct, MOA-specific biological programs.

Having established that TRIDENT's representations are separable by MOA, we next evaluate their fidelity to the ground-truth. A UMAP visualization [28] (Fig. 4b) shows a striking alignment: embeddings from TRIDENT-generated and real images are tightly intermingled and occupy a shared manifold, separate from the control population. This demonstrates that generated morphologies are biologically correct and virtually indistinguishable from real data in this embedding space. This finding is further substantiated by a quantitative analysis of interpretable cytological measurements using CellProfiler [42] (Fig. 4c). The analysis of key features like *AreaOccupied* shows a remarkable alignment between the feature distributions of the generated and true images across the whole cell, cytoplasm, and nucleus. Both consistently exhibit a significant shift away from the control population, proving that TRIDENT accurately recapitulates
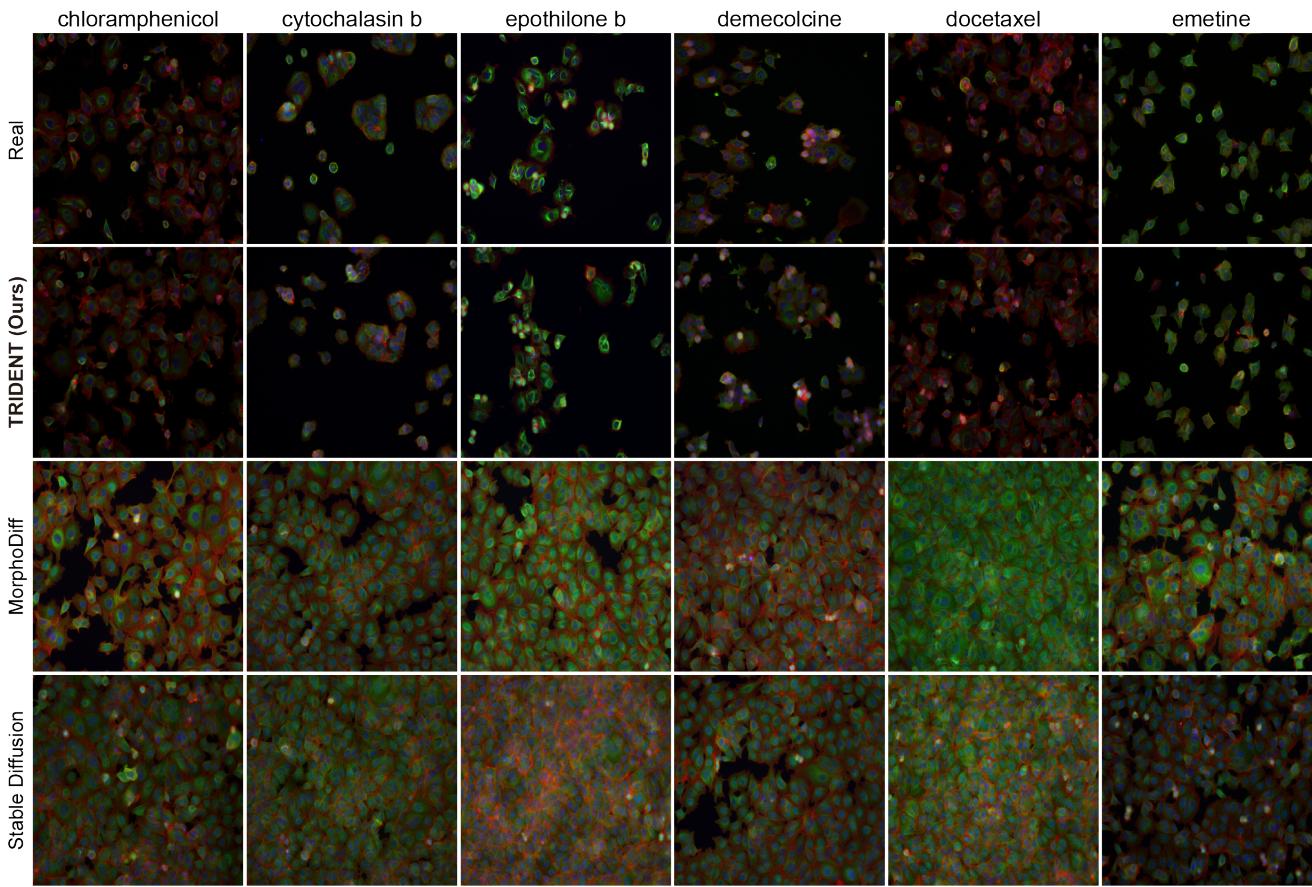
Figure 3. Visual comparison of generated cellular morphologies under six drug perturbations. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4). See supplementary material for more results.

the nuanced, multi-scale morphological changes induced by drug perturbations.
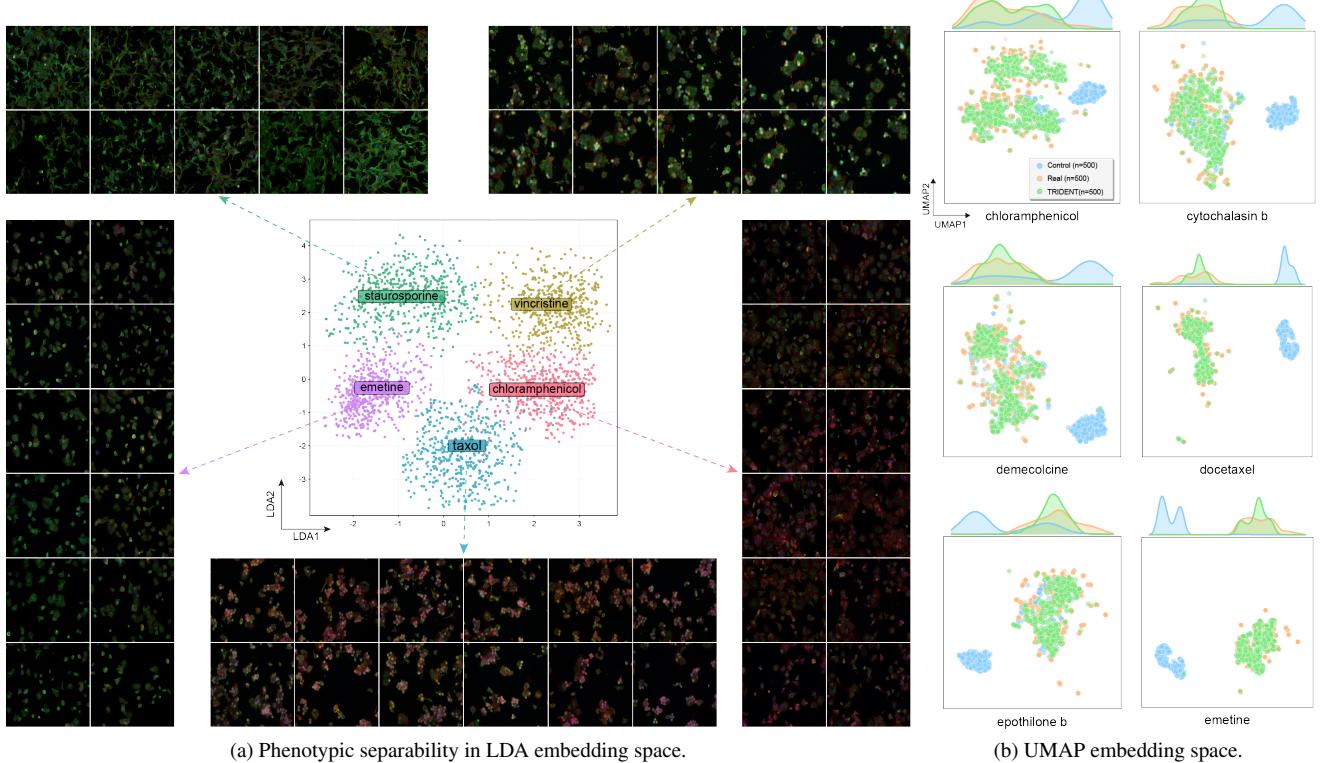
4.3. Cross-Modal Validation

We validate if TRIDENT learns the true biological mapping between the molecular state (RNA) and the physical phenotype (morphology). We first assess the accuracy of the post-perturbation transcriptome (G_{post}) predicted by the Transcription-Drug Condition Module, and then its consistency with the final generated morphology.

We evaluated the global accuracy of the predicted gene expression profiles. For 44 ID compounds, the Z-scored predicted log fold change (LFC) patterns demonstrate a striking similarity to the ground-truth patterns (Fig. 5a). This is confirmed quantitatively by a high Pearson correlation of 0.957, indicating the module generates highly accurate and biologically faithful transcriptional profiles. To further validate the biological relevance of these predictions, we perform a case study using docetaxel [11, 26, 44], a well-characterized compound. We partition the genes into upregulated (activated) and downregulated (suppressed) sets based on their LFC generated by TRIDENT. Functional enrichment analy-
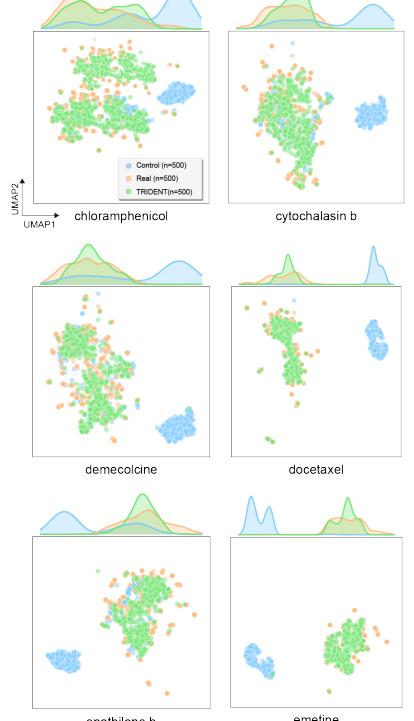
sis on these genes aligns perfectly with docetaxel's known MOA as a mitotic inhibitor that promotes cell cycle arrest and apoptosis (Fig. 5b). Suppressed genes are significantly enriched for terms like *regulation of cell growth*' and *DNA replication*, while activated genes are enriched for *regulation of apoptotic signaling pathway*. This demonstrates the predicted transcriptome is functionally and mechanistically correct.

Finally, we investigate the critical cross-modal link: whether this predicted biological program (i.e., suppressed growth, induced apoptosis) is reflected in the generated morphology. The biological functions identified in Fig. 5b strongly imply a phenotype characterized by reduced cell proliferation and increased cell death, which manifests as a significant decrease in cell density. Fig. 5c provides a direct visual comparison, showing that TRIDENT-generated images for docetaxel clearly exhibit a sparse cell population. This precisely matches the expected morphological outcome and aligns with our quantitative CellProfiler analysis (Fig. 4c).
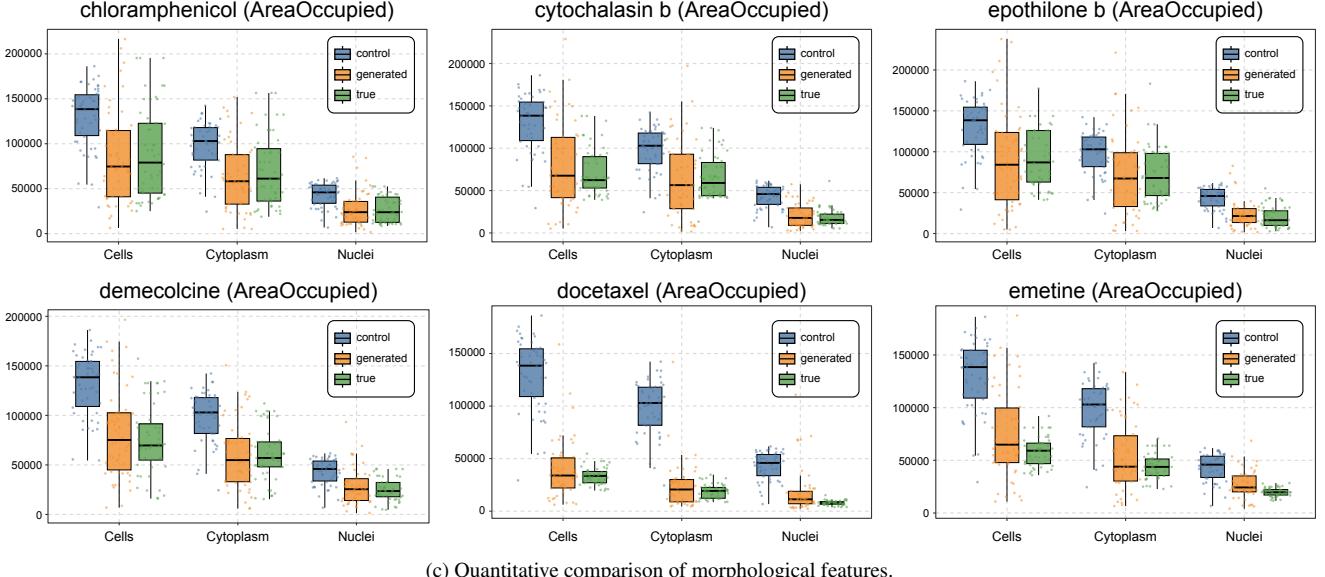
This evidence demonstrates that TRIDENT successfully learns the complex, cross-modal association between tran-

(a) Phenotypic separability in LDA embedding space.

(b) UMAP embedding space.

(c) Quantitative comparison of morphological features.

Figure 4. TRIDENT captures biologically interpretable signatures in embedding and feature space. (a) ViT embeddings of generated images form distinct, MOA-specific clusters in LDA space, with representative images shown for each cluster. (b) UMAP visualization confirms high distributional alignment between generated (green) and real (orange) images, which are both separate from the control (blue) population. (c) Quantitative CellProfiler analysis shows that *AreaOccupied* feature distributions of generated and real images are highly similar and distinct from control across all cellular compartments.

scriptome and morphology, where functionally-correct RNA predictions actively and accurately guide the synthesis of the correct corresponding cellular phenotype.

4.4. Ablation Studies

To demonstrate the importance of RNA expression as a critical intermediate for accurate morphology synthesis, we

(a) Prediction of transcriptomic log fold change.

(b) Functional enrichment analysis for docetaxel.

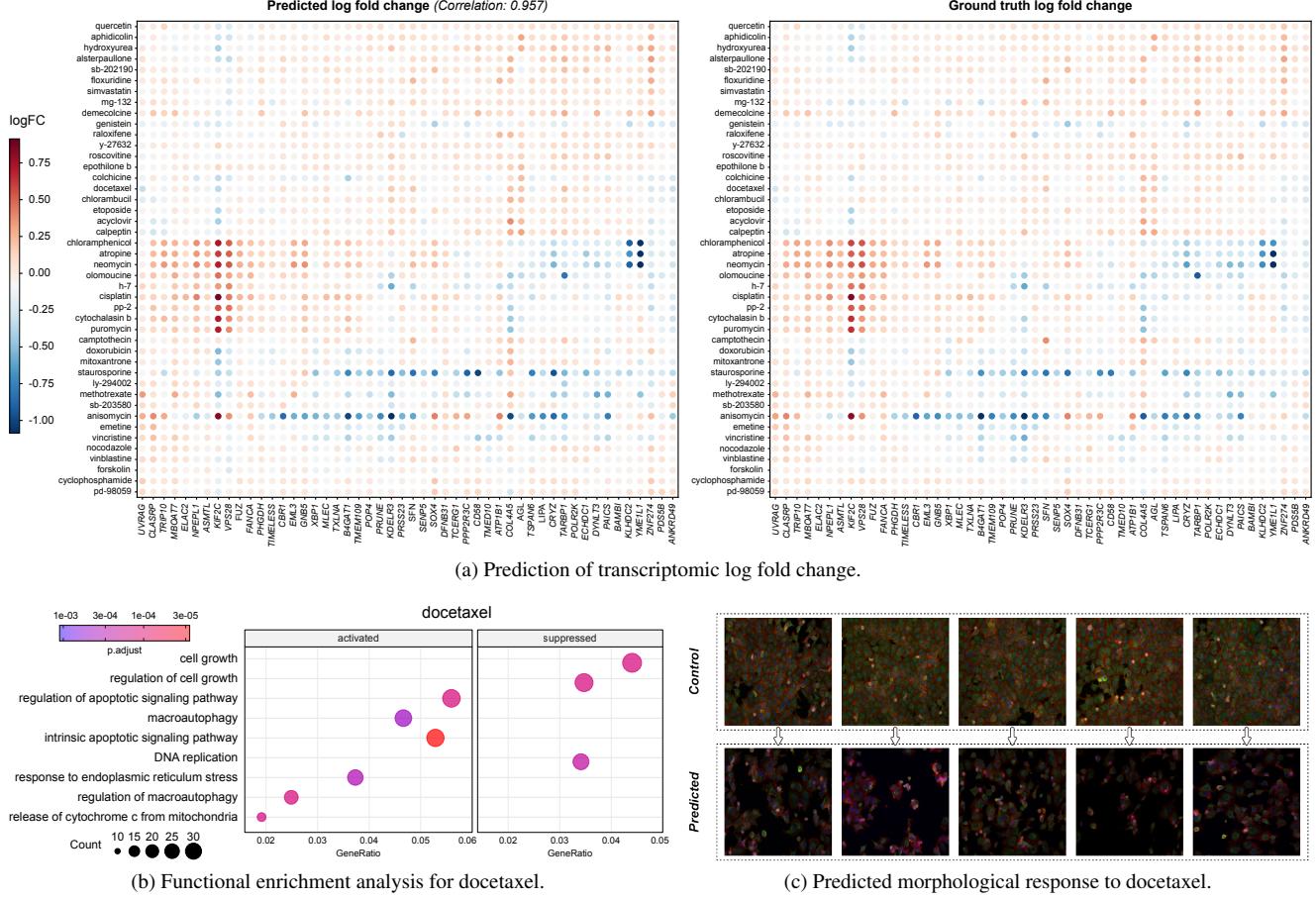(c) Predicted morphological response to docetaxel.

Figure 5. TRIDENT learns the association between transcriptome and morphology. (a) Heatmaps comparing predicted (left) versus ground-truth (right) gene expression log fold changes for 44 compounds. (b) Functional enrichment analysis of model-predicted genes for docetaxel identifies pathways consistent with its known MOA. (c) Visual comparison of TRIDENT-predicted morphology for docetaxel (bottom) versus control (top), correctly capturing the phenotype of reduced cell density.

Table 2. Ablation study results. Performance comparison of the full TRIDENT model against a variant without the RNA conditioning module on ID and OOD test sets.

Methods	In-Distribution			Out-of-Distribution		
	FID↓	KID↓	IS↓	FID↓	KID↓	IS↓
TRIDENT	49.770	0.013	2.240	126.150	0.222	2.523
TRIDENT (w/o RNA)	115.770	0.132	2.381	194.239	0.293	2.639

conduct an ablation study. We create a variant, TRIDENT (w/o RNA), which bypasses the **Transcription-Drug Condition Module** and uses only the drug embedding D to condition the diffusion model. This models a direct *Perturbation → Morphology* pathway. As shown in Tab. 2, the full TRIDENT framework significantly outperforms this ablated model. On the ID test set, TRIDENT's FID (49.770) is more than 2.3-fold better than the ablated model's (115.770). The performance gap is even more pronounced in the KID met-

ric (0.013 vs. 0.132), an order of magnitude improvement. This demonstrates that the drug information D alone is an insufficient condition. The latent vector z, which integrates both perturbation and transcriptional state, provides a far richer signal. The full model's superior performance on the OOD set (FID 126.150 vs. 194.239) further confirms that explicitly modeling the *(Perturbation + RNA) → Morphology* pathway is essential for high fidelity and accuracy.

5. Conclusion

In this work, we address a critical gap in cellular modeling: the cross-modal mapping from transcriptome to phenotype. We introduce TRIDENT, a cascaded generative framework modeling the complete tripartite relationship between perturbation, gene expression, and cellular morphology. TRIDENT generates high-fidelity, MOA-specific morphologies, significantly outperforming SOTA methods quantitatively and generalizing better to unseen compounds. We validated that the high fidelity of the generated cellular morphology

is guided by functionally-correct transcriptomes. Ablation studies confirmed this intermediate RNA state is essential for high-fidelity generation. By mechanistically linking transcriptional state to morphological outcome, TRIDENT provides a foundational component essential for realizing the AIVC vision, enabling simulations where molecular events causally orchestrate cellular form.

Limitations. A primary limitation is our MorphoGene dataset, which uses a single cell line (MCF7) and bulk L1000 profiles. This limits cross-cell-type generalization and overlooks cell-to-cell heterogeneity. Future work should prioritize richer, multi-modal datasets, ideally pairing single-cell transcriptomics with imaging across diverse cell lines. Such data will enable more generalizable models, including zeroshot, cross-cell-line prediction and the pre-training of large-scale foundational models, marking the next step toward a predictive virtual cell.

References

[1] Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, pages 2025–06, 2025. 1, 2

[2] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998. 5

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 2

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 5

[5] 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. 1

[6] Silvia Bruno, Barbara Ardelt, Janusz S Skierski, Frank Traganos, and Zbigniew Darzynkiewicz. Different effects of staurosporine, an inhibitor of protein kinases, on the cell cycle and chromatin structure of normal and leukemic lymphocytes. *Cancer research*, 52(2):470–473, 1992. 5

[7] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. 1

[8] Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular cancer therapeutics*, 9(6):1913–1926, 2010. 2

[9] Francisco Carrillo-Perez, Marija Pizurica, Yuanning Zheng, Tarak Nath Nandi, Ravi Madduri, Jeanne Shen, and Olivier Gevaert. Generation of synthetic whole-slide image tiles of tumours from rna-sequencing data via cascaded diffusion models. *Nature Biomedical Engineering*, 9(3):320–332, 2025. 2

[10] Han-Jung Chae, Jang-Sook Kang, Jong-Ook Byun, Kyung-Soo Han, Dae-Up Kim, Se-Man Oh, Hyung-Min Kim, Soo-Wan Chae, and Hyung-Ryong Kim. Molecular mechanism of staurosporine-induced apoptosis in osteoblasts. *Pharmacological research*, 42(4):373–381, 2000. 5

[11] Stephen J Clarke and Laurent P Rivory. Clinical pharmacokinetics of docetaxel. *Clinical pharmacokinetics*, 36(2):99–114, 1999. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794, 2021. 2

[13] Daria Doncevic and Carl Herrmann. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. *Bioinformatics*, 39(6):btad387, 2023. 2

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[15] Radhey S Gupta and Louis Siminovitch. The molecular basis of emetine resistance in chinese hamster ovary cells: alteration in the 40s ribosomal subunit. *Cell*, 10(1):61–66, 1977. 5

[16] Radhey S Gupta, Jiri J Krepinsky, and Louis Siminovitch. Structural determinants responsible for the biological activity of (-)-emetine,(-)-cryptopleurine, and (-)-tylocrebrine: structure-activity relationship among related compounds. *Molecular Pharmacology*, 18(1):136–143, 1980. 5

[17] Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, pages 26711–26722, 2022. 1, 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 5

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 4

[22] Mazen W Karaman, Sanna Herrgard, Daniel K Treiber, Paul Gallant, Corey E Atteridge, Brian T Campbell, Katrina W Chan, Pietro Ciceri, Mindy I Davis, Philip T Edeen, et al. A

quantitative analysis of kinase inhibitor selectivity. *Nature biotechnology*, 26(1):127–132, 2008. 5

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019. 2

[25] Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J Theis, and F Alexander Wolf. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36 (Supplement_2):i610–i617, 2020. 2

[26] Katherine A Lyseng-Williamson and Caroline Fenton. Docetaxel: a review of its use in metastatic breast cancer. *Drugs*, 65(17):2513–2531, 2005. 6

[27] Susan MacLean-Fletcher and Thomas D Pollard. Mechanism of action of cytochalasin b on actin. *Cell*, 20(2):329–341, 1980. 5

[28] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5

[29] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017. 2

[30] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of IEEE/CVF winter conference on applications of computer vision*, pages 2000–2009, 2023. 2

[31] Zeinab Navidi, Jun Ma, Esteban Miglietta, Le Liu, Anne E Carpenter, Beth A Cimini, Benjamin Haibe-Kains, and Bo Wang. Morphodiff: Cellular morphology painting with diffusion models. In *International Conference on Learning Representations*, 2025. 1, 2

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[34] Alessandro Palma, Fabian J Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1): 505, 2025. 1, 2

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[36] Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional

[37] Ladislav Rampášek, Daniel Hidru, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Dr. vae: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, 35(19):3743–3751, 2019. 2

[38] Douglas D Rhoads and Donald J Roufa. Emetine resistance of chinese hamster cells: structures of wild-type and mutant ribosomal protein s14 mrnas. *Molecular and cellular biology*, 5(7):1655–1659, 1985. 5

[39] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024. 1, 2

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016. 5

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[42] David R Stirling, Madison J Swain-Bowden, Alice M Lucas, Anne E Carpenter, Beth A Cimini, and Allen Goodman. Cellprofiler 4: improvements in speed, utility and usability. *BMC bioinformatics*, 22(1):433, 2021. 5

[43] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017. 1, 2

[44] Ian F Tannock, Ronald De Wit, William R Berry, Jozsef Horti, Anna Pluzanska, Kim N Chi, Stephane Oudard, Christine Théodore, Nicholas D James, Ingela Turesson, et al. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *New England Journal of Medicine*, 351(15):1502–1512, 2004. 6

[45] Duangrudee Tanramluk, Adrian Schreyer, William R Pitt, and Tom L Blundell. On the origins of enzyme inhibitor selectivity and promiscuity: a case study of protein kinase binding to staurosporine. *Chemical biology & drug design*, 74(1):16–24, 2009. 5

[46] Panayotis A Theodoropoulos, Achille Gravanis, Anna Tsapara, Andrew N Margioris, Eva Papadogiorgaki, Vassilis Galanopoulos, and Christos Stournaras. Cytochalasin b may shorten actin filaments by a mechanism independent of barbed end capping. *Biochemical pharmacology*, 47(10):1875–1881, 1994. 5

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017. 2

[48] Dominik JE Waibel, Ernst Röell, Bastian Rieck, Raja Giryes, and Carsten Marr. A diffusion model predicts 3d shapes from 2d microscopy images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 2

[49] Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo,

Beth A Cimini, Kyle Karhohs, David J Logan, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell systems*, 13(11): 911–923, 2022. 1

[50] Yahao Wu, Jing Liu, Yanni Xiao, Shuqin Zhang, and Limin Li. Couplevae: coupled variational autoencoders for predicting perturbational single-cell rna sequencing data. *Briefings in Bioinformatics*, 26:2, 2025. 2

[51] Karren Yang, Samuel Goldman, Wengong Jin, Alex X Lu, Regina Barzilay, Tommi Jaakkola, and Caroline Uhler. Mol2image: improved conditional flow models for molecule to image synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6688–6698, 2021. 1, 2

Supplemental Material:
TRIDENT: A Trimodal Cascade Generative Framework for Drug and RNA-Conditioned Cellular Morphology Synthesis

Rui Peng[1,2#] Ziru Liu[5#] Lingyuan Ye[6,7] Yuxing Lu[1] Boxin Shi[3,4*] Jinzhuo Wang[1*]

[1] Department of Big Data and Biomedical AI, College of Future Technology, Peking University
[2] Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University
[3] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[4] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[5] Yuanpei College, Peking University   [6] School of Life Sciences, Tsinghua University
[7] Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program (PTN), Tsinghua University
{pengrui, lzr, luyx}@stu.pku.edu.cn   yely23@mails.tsinghua.edu.cn
shiboxin@pku.edu.cn   wangjinzhuo@pku.edu.cn

A. Implementation Details

All models are implemented in PyTorch and trained on a high-performance computing cluster equipped with eight NVIDIA A100 GPUs, each with 80GB of memory. The core of our Morphology Generation Module is a Diffusion Transformer architecture. This transformer is configured with 28 layers, a hidden dimension of 1152, and 16 attention heads. The complete TRIDENT framework is trained end-to-end for a total of 100,000 steps. We employ the AdamW optimizer with a constant learning rate of 1e-4. A global batch size of 32 is used, distributed across the eight GPUs. All Cell Painting images are processed at a resolution of 512×512 pixels. The total training process for the final model takes approximately four days.

B. Cellprofiler Feature Construction

To derive quantitative descriptors of cellular phenotype, we construct a bespoke analysis workflow using CellProfiler (v5.0). This pipeline is engineered to process each Cell Painting image and output a single, comprehensive feature vector summarizing its morphological characteristics.

The workflow's core is a three-step segmentation process to delineate cellular structures. First, nucleus are identified as primary objects from the blue (DNA) channel. Next, cell boundaries are segmented as secondary objects by propagating outwards from the identified nucleus, using the green channel to define the cell periphery. Finally, the cytoplasm is defined as a tertiary region, calculated by subtracting the nuclear mask from the corresponding cell mask. A quality control step is integrated to discard all objects touching the image border, ensuring that all downstream measurements are derived from complete, intact cells.

Following segmentation, a comprehensive suite of Cell-Profiler modules is executed to extract measurements from all three compartments (nucleus, cytoplasm, and cells) across all channels. The extracted features include morphological descriptors of size and shape, such as area, perimeter, major and minor axis lengths, eccentricity, and solidity. Furthermore, statistics on pixel intensity distribution like mean, median, standard deviation, median absolute deviation, and quartiles are computed. Finally, the pipeline captures relational metrics, such as spatial relationships between cells, inter-channel signal correlations, and radial intensity distributions.

To generate a single profile for each image, these per-object measurements are aggregated by calculating their mean, median, and standard deviation across all valid cells. This process yields a final, high-dimensional profile of 6,345 distinct morphological features for each image, providing a detailed quantitative fingerprint of the cellular phenotype in response to perturbation.

Equal contribution.   * Corresponding authors.

C. TRIDENT Algorithm

Algorithm 1 TRIDENT Framework: Training Procedure

Require: Training data $\mathcal{D} = \{(G_{pre}^{(i)}, D^{(i)}, I^{(i)}, G_{post}^{(i)})\}$
Require: Pre-trained image VAE $(\mathcal{E}_{image}, \mathcal{D}_{image})$
Require: Diffusion timesteps T, variance schedule $\{\beta_t\}_{t=1}^T$
  (and $\alpha_t, \bar{\alpha}_t$)
Ensure: Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$

1: Initialize VAE parameters ϕ (for $\mathcal{E}_{rna}, \mathcal{E}_{drug}$), ψ (for $\mathcal{D}_{perturb}$)
2: Initialize Denoising Transformer f_θ parameters θ
3: Initialize Time Embedding parameters γ for $\mathcal{E}_{time}$
4: **for** each epoch $e = 1, \ldots, E_{max}$ **do**
5:     **for** each batch $(G_{pre}, D, I, G_{post}) \sim \mathcal{D}$ **do**
6:         $\mathbf{X}_{rna} \leftarrow \mathcal{E}_{rna}(G_{pre})$, $\mathbf{X}_{drug} \leftarrow \mathcal{E}_{drug}(D)$
7:         $[\boldsymbol{\mu}_z, \log \boldsymbol{\sigma}_z^2] \leftarrow \mathcal{E}_{perturb}([\mathbf{X}_{rna}, \mathbf{X}_{drug}])$
8:         $\boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
9:         $\boldsymbol{z} \leftarrow \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}_z$
10:        $[\boldsymbol{\mu}_{G_{post}}, \log \boldsymbol{\sigma}_{G_{post}}^2] \leftarrow \mathcal{D}_{perturb}(\boldsymbol{z})$
11:        $\mathcal{L}_{recon} \leftarrow \mathbb{E}_{q_\phi}[-\log p_\psi(G_{post}|\boldsymbol{z})]$
12:        $\mathcal{L}_{KL} \leftarrow \mathrm{D}_{KL}(q_\phi(\boldsymbol{z}|G_{pre}, D) \,||\, p(\boldsymbol{z}))$
13:        $\mathcal{L}_{VAE} \leftarrow \mathcal{L}_{recon} + \mathcal{L}_{KL}$
14:        $\mathbf{X}_{image}^0 \leftarrow \mathcal{E}_{image}(I)$
15:        $t \sim \mathcal{U}(\{1, \ldots, T\})$
16:        $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
17:        $\mathbf{X}_{image}^t \leftarrow \sqrt{\bar{\alpha}_t}\mathbf{X}_{image}^0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
18:        $\mathbf{X}_{time} \leftarrow \mathcal{E}_{time}(t)$
19:        $\mathbf{X}_{condition} \leftarrow \boldsymbol{z} + \mathbf{X}_{time}$
20:        $\boldsymbol{\epsilon}_\theta \leftarrow f_\theta(\mathbf{X}_{image}^t, \mathbf{X}_{condition})$
21:        $\mathcal{L}_{LDM} \leftarrow ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta||^2$
22:        $\mathcal{L}_{TRIDENT} \leftarrow \mathcal{L}_{VAE} + \mathcal{L}_{LDM}$
23:        Update parameters $\phi, \psi, \theta, \gamma$ using $\nabla \mathcal{L}_{TRIDENT}$
24:     **end for**
25: **end for**
26: **return** Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$

Algorithm 2 TRIDENT Framework: Inference Procedure

Require: Input G_{pre}, D
Require: Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$
Require: Pre-trained image VAE $(\mathcal{E}_{image}, \mathcal{D}_{image})$
Require: Diffusion timesteps T and schedule $\{\beta_t\}_{t=1}^T$
Ensure: Generated Image $\hat{I}$

1: $\mathbf{X}_{rna} \leftarrow \mathcal{E}_{rna}(G_{pre})$
2: $\mathbf{X}_{drug} \leftarrow \mathcal{E}_{drug}(D)$
3: $[\boldsymbol{\mu}_z, \log \boldsymbol{\sigma}_z^2] \leftarrow \mathcal{E}_{perturb}([\mathbf{X}_{rna}, \mathbf{X}_{drug}])$
4: $\boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\boldsymbol{z} \leftarrow \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}_z$
6: $\hat{\mathbf{X}}_{image}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
7: **for** $t = T$ **down to** 1 **do**
8:     $\mathbf{X}_{time} \leftarrow \mathcal{E}_{time}(t)$
9:     $\mathbf{X}_{condition} \leftarrow \boldsymbol{z} + \mathbf{X}_{time}$
10:    $\boldsymbol{\epsilon}_\theta \leftarrow f_\theta(\hat{\mathbf{X}}_{image}^t, \mathbf{X}_{condition})$
11:    $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{\epsilon}' \leftarrow \mathbf{0}$
12:    $\hat{\mathbf{X}}_{image}^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{X}}_{image}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\right) + \sigma_t \boldsymbol{\epsilon}'$
13: **end for**
14: $\hat{I} \leftarrow \mathcal{D}_{image}(\hat{\mathbf{X}}_{image}^0)$
15: **return** $\hat{I}$
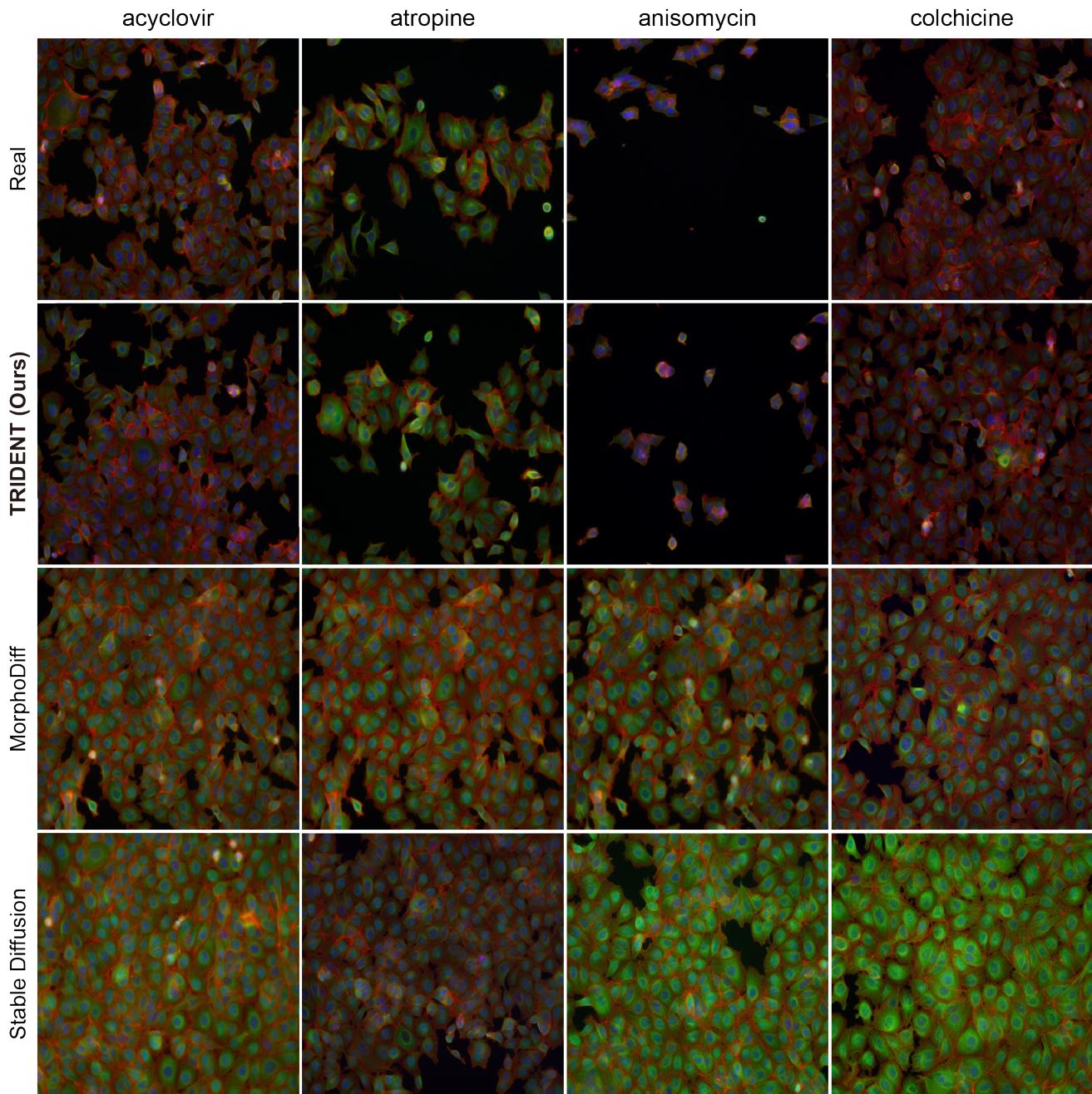

D. Additional Comparison Results

Figure A1. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).
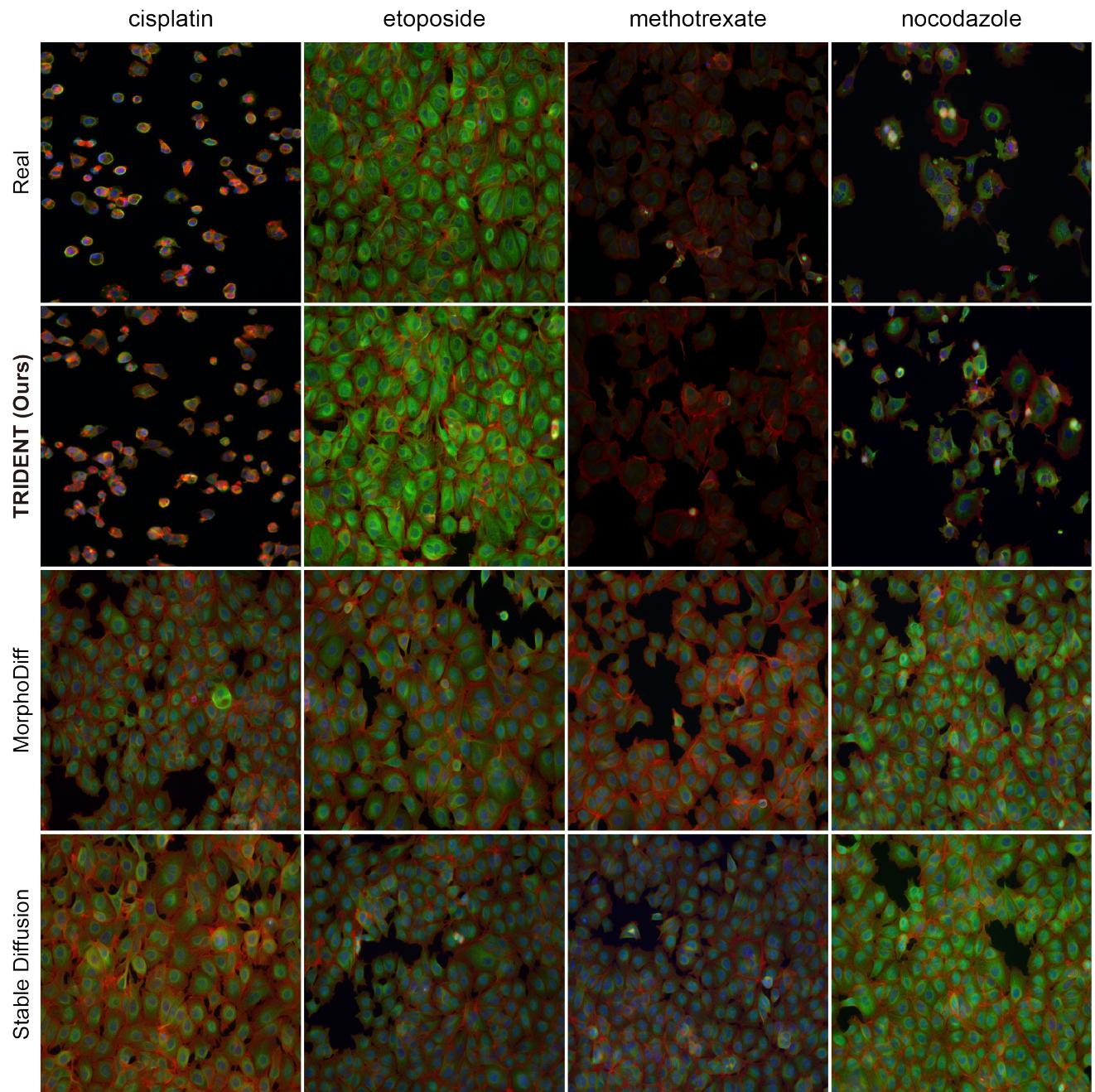
Figure A2. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).
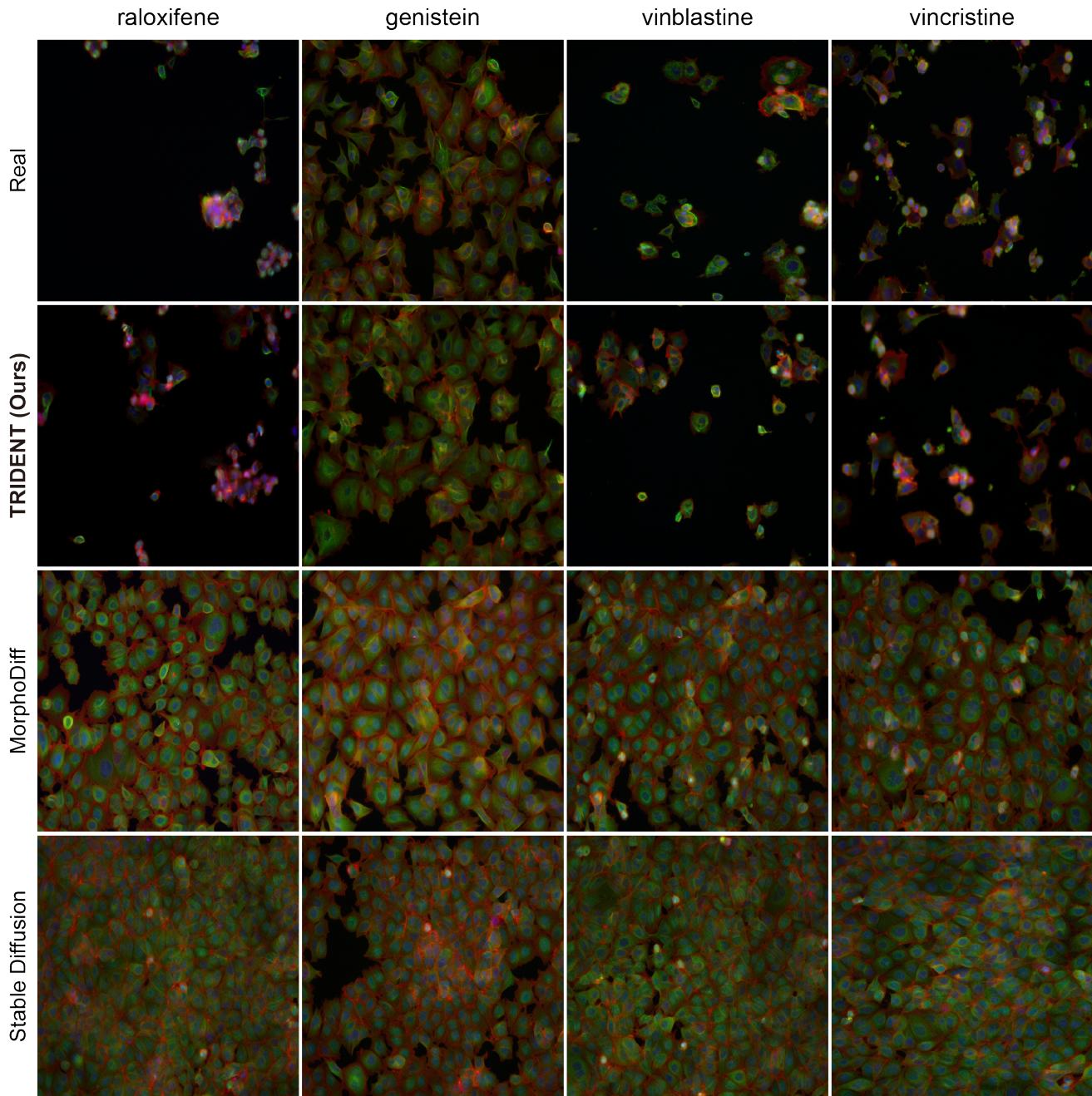
Figure A3. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).