PERTURBBRIDGE: DIRECT PHENOTYPIC SUPERVISION FOR CELLULAR PERTURBATION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual cell models should capture phenotypes induced by specific perturbations. However, image-based models are typically trained and evaluated to match real-image distributions, which alone does not explicitly prioritize perturbation-specific biological responses. Recent approaches add biological signals through reinforcement-learning post-training or inference-time steering, raising a central question: **where should biological response information enter the generative pipeline?** We introduce PerturbBridge, a stochastic data-to-data bridge model that directly supervises perturbation-specific phenotypes in predicted perturbed images during training. PerturbBridge jointly optimizes bridge regression and differentiable biological response losses, with biological gradients passing through a frozen image encoder. We further introduce perturbation-response evaluations based on downstream classification and retrieval. Across three microscopy benchmarks for perturbation profiling and subcellular imaging, PerturbBridge surpasses CellFlux in biological fidelity, image quality, and sampling efficiency. On BBBC021, mechanism-of-action classification improves by 5.66 and 7.5 percentage points with morphological and learned visual features, respectively. On JUMP and AllenCell, retrieval Hit@10 and Hit@1 improve by 1.77 and 7.15 percentage points, respectively. On BBBC021 and JUMP, PerturbBridge achieves lower FID with 80% fewer model evaluations than CellFlux. Overall, these results show that directly supervising perturbation responses during training improves phenotypic fidelity while enabling efficient cellular image generation.

1 INTRODUCTION

Virtual cell models aim to predict how cells respond to interventions such as drug treatments and gene disruptions (Bunne et al., 2024). For image-based virtual cell modeling, phenotypic profiling characterizes these responses by measuring cell shape, staining intensity, and organelle organization from microscopy images. These measurements support high-throughput compound screening by comparing treated cells with controls. Imaging assays such as Cell Painting (Bray et al., 2016) capture multiple cellular components, providing complementary views of perturbation-induced changes. Recent methods such as PhenDiff (Bourou et al., 2024), IMPA (Palma et al., 2025), and CellFlux (Zhang et al., 2025) use image-based generative models to translate between cellular conditions or transport control populations to perturbed populations. For biological applications, however, generated images should capture the phenotype associated with the requested perturbation, not merely resemble the overall distribution of real perturbed data. Conditional generative objectives learn associations between images and perturbations from data, while explicit biological response supervision can further emphasize the features and labels relevant to downstream analysis.

Biological signals can be introduced at different stages of the generative pipeline. CellFluxRL (Fig. 1(a); Wu et al., 2026) uses biological evaluators as rewards for RL post-training, while SimuSDB (Fig. 1(b); Zhang et al., 2026a) uses phenotypic costs to steer candidate transitions during bridge sampling. These approaches motivate a central question: **where should biological response information enter the generative pipeline?** We investigate direct supervision of perturbation-specific responses during training of the virtual cell model. We introduce **PerturbBridge** (Fig. 1(c)),

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

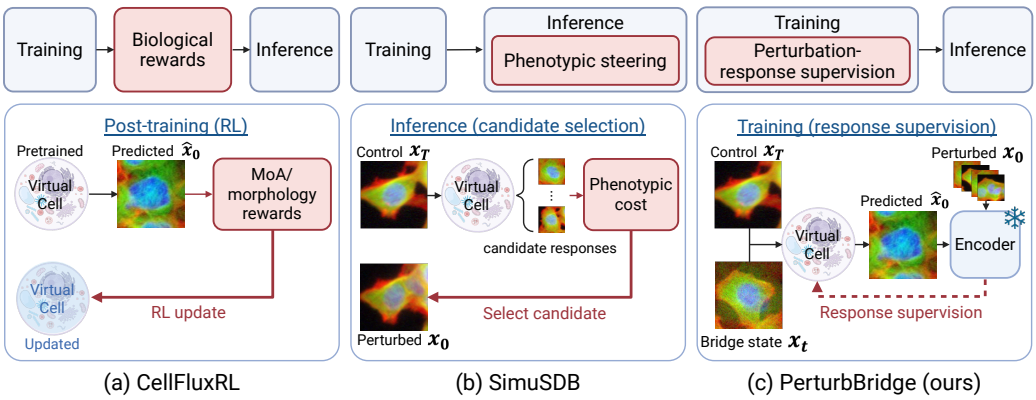

Figure 1: **Three mechanisms for incorporating biological response signals.** (a) CellFluxRL uses biological rewards for RL post-training. (b) SimuSDB uses phenotypic costs to select candidate transitions during sampling. (c) PerturbBridge jointly optimizes bridge regression and biological response losses on predicted perturbed images during generative training. Conditioning inputs and optional test-time selection in CellFluxRL are omitted for clarity.

a stochastic data-to-data bridge model that supervises phenotypic responses directly on predicted perturbed images during generative training.

Specifically, PerturbBridge jointly optimizes bridge regression and differentiable biological response losses on the features of predicted perturbed images. Gradients from the biological response losses pass through a frozen encoder to update the generative model. Targets come from condition-specific statistics of real perturbed training data and available biological labels. Across BBBC021 (Caie et al., 2010), JUMP (Chandrasekaran et al., 2023), and AllenCell (Allen Institute), we use condition-moment matching and InfoNCE-style contrastive objectives (Oord et al., 2018; Khosla et al., 2020), adapting the positive and negative sets to the conditions and annotations of each dataset. On BBBC021, the contrastive objective groups positive prototypes by mechanism of action (MoA) to encourage shared phenotypic features across related perturbations.

To test whether this supervision improves perturbation-response fidelity, we evaluate whether generated images preserve the requested biological mechanism and perturbation identity. We use downstream MoA classification and perturbation retrieval, respectively, alongside image-distribution metrics such as FID (Heusel et al., 2017). To assess gains beyond the supervising encoder, we use DINOv2 (Oquab et al., 2024) representations and CellProfiler (Carpenter et al., 2006) features, neither of which is used for biological response supervision.

Across BBBC021 (Caie et al., 2010), JUMP (Chandrasekaran et al., 2023), and AllenCell (Allen Institute) datasets, PerturbBridge improves biological evaluation scores and image-distribution metrics relative to CellFlux. On BBBC021, DINOv2-based MoA accuracy increases from 67.4% to 74.9%, with accuracy also increasing from 72.15% to 77.81% under CellProfiler features. Gains also extend to unseen-compound MoA classification under the CellFlux protocol and DINOv2-based perturbation retrieval on JUMP and AllenCell.

Ablations compare PerturbBridge with a Bridge-only model and with biological response supervision applied to perturbed-image predictions reconstructed from CellFlux velocity outputs. These comparisons validate the integrated bridge and supervision design. Using bridge-specific samplers (Zhang et al., 2026b; Zheng et al., 2025), PerturbBridge achieves lower FID on BBBC021 and JUMP with fewer model evaluations for generation than CellFlux.

We make three contributions:

1. **Direct phenotypic response supervision during generative training.** We introduce PerturbBridge, which jointly optimizes stochastic-bridge regression and perturbation-specific biological response losses on predicted perturbed images.
2. **Evaluation of perturbation-specific biological responses.** We assess generated images through downstream MoA classification and perturbation retrieval using representations not employed for

biological response supervision. Our analysis of FID provides theoretical support for evaluating biological responses alongside image-distribution metrics.

3. **Empirical validation across datasets and sampling efficiency.** Across BBBC021, JUMP, and AllenCell, PerturbBridge improves biological evaluation scores and image-distribution metrics over CellFlux, with a reduced sampling budget.

2 Related work and problem setting

2.1 Related work on cellular generation and biological response supervision

Virtual-cell modeling and image generation. The broader virtual-cell program spans molecular, transcriptional, and morphological responses (Slepchenko & Loew, 2010; Walker & Southgate, 2009; You, 2004; Bereket & Karaletsos, 2023; Hung et al., 2024; Lamiable et al., 2023; Rampášek et al., 2019). CellOT (Bunne et al., 2023) learns perturbation-response transport from unpaired cell measurements, while GEARS (Roohani et al., 2024) predicts transcriptional responses using gene relationships. Their biological goals are related, but their outputs and evaluation protocols differ from multichannel image generation. For cellular microscopy, PhenDiff (Bourou et al., 2024) uses conditional diffusion translation, IMPA (Palma et al., 2025) separates cellular content from perturbation-associated style, and CellFlux (Zhang et al., 2025) uses control-initialized flow matching. These methods are baselines in our experiments.

Biological response supervision and generative model design. CellFluxRL (Wu et al., 2026) and SimuSDB (Zhang et al., 2026a) incorporate biological signals through RL post-training and inference-time steering, respectively. Su et al. (2026) introduce stochastic training perturbations to alleviate sparse supervision in distribution-to-distribution learning. PerturbBridge complements these approaches with perturbation-specific objectives on predicted perturbed images.

2.2 Problem setting and generative modeling

Let B denote experimental context, C the requested perturbation, X_T a control-image random variable, and X_0 a perturbed-image random variable. Uppercase symbols denote random variables, and lowercase symbols denote their realizations, such as a specific control image x_T or perturbed image x_0. For a fixed context b and perturbation c, we seek a conditional generative distribution $K_\theta(x_0 \mid x_T, c, b)$ such that

$$\int K_\theta(x_0 \mid x_T, c, b)\, p_{\mathrm{ctrl}}(x_T \mid b)\, dx_T \approx p_{\mathrm{pert}}(x_0 \mid c, b). \tag{1}$$

b defines the sampling and evaluation strata, while the control image provides relevant information.

Training pairs perturbed and control images using a joint sampling distribution $\pi_{c,b}(x_0, x_T)$ for each perturbation c and experimental context b. Pairing within the same batch or plate keeps the experimental context consistent. One possible construction samples the two images independently from their respective populations within each stratum.

Stochastic bridge and image prediction. Diffusion bridges connect informative source and target distributions through stochastic intermediate states (Zhou et al., 2024; Liu et al., 2023). A Gaussian stochastic interpolant takes the form

$$X_t = \alpha_t X_0 + \beta_t X_T + \gamma_t Z, \qquad Z \sim \mathcal{N}(0, I), \tag{2}$$

where Z is independent of the image pair. We use a time coordinate t, with $t = 1$ at the control and $t = 0$ at the perturbed target. The coefficients satisfy $(\alpha_0, \beta_0, \gamma_0) = (1, 0, 0)$ and $(\alpha_1, \beta_1, \gamma_1) = (0, 1, 0)$. Hence, PerturbBridge predicts the perturbed image from an intermediate bridge state, allowing biological response objectives to act directly on that prediction. Section 3 specifies the path coefficients and training objective.

3 Method

Fig. 2 summarizes PerturbBridge, which jointly optimizes bridge regression and biological response losses on predicted perturbed images. We instantiate the stochastic bridge in Equation (2) with a

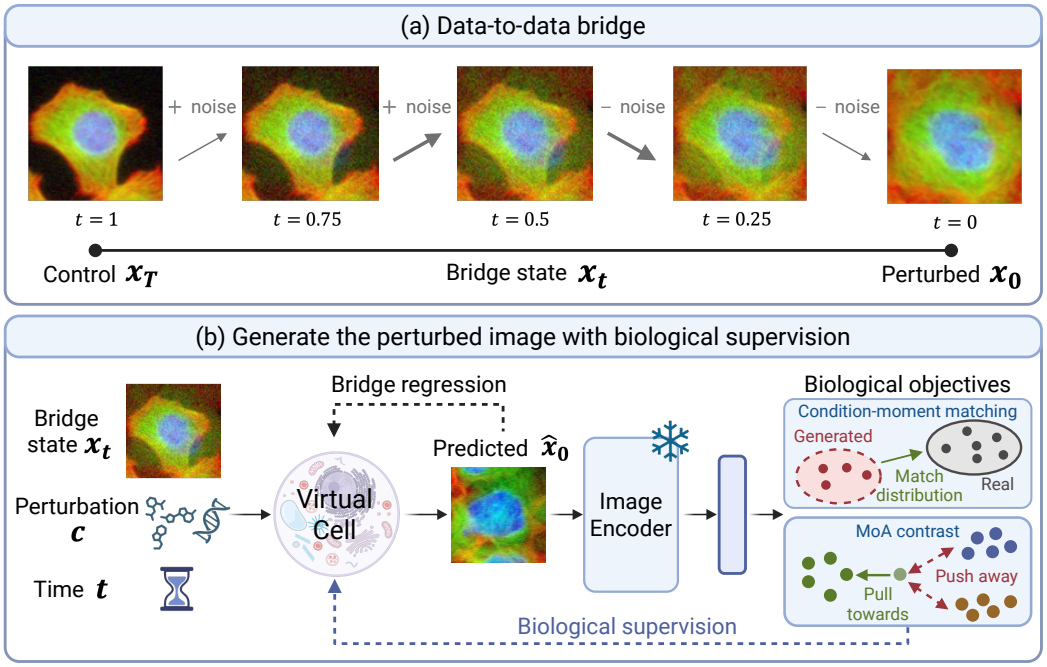

Figure 2: **Training framework of PerturbBridge.** (a) A stochastic data-to-data bridge connects the control image x_T at $t = 1$ to the perturbed image x_0 at $t = 0$, with noise strongest at the midpoint and vanishing at both ends. (b) Conditioned on the bridge state x_t, perturbation c, and time t, the generative model predicts the perturbed image $\hat{x}_0$. A frozen image encoder extracts features for MoA contrast and condition-moment matching. These objectives encourage separation across MoA classes and align condition-specific feature means and variances with real perturbed data. Biological gradients pass through the encoder to update the generative model jointly with bridge regression.

linear interpolation between control and perturbed images and added Gaussian noise (Zhou et al., 2024; Zhang et al., 2026b):

$$\alpha_t = 1 - t, \qquad \beta_t = t, \qquad \gamma_t = 2\gamma_{\max}\sqrt{t(1-t)}. \tag{3}$$

The added noise vanishes at both endpoints and is largest at $t = 1/2$. Setting $\gamma_{\max} = 0$ yields the deterministic linear path $X_t = (1-t)X_0 + tX_T$.

3.1 Predicting x_0 with biological response losses

Why predict the perturbed target image? JiT (Li & He, 2026) motivates direct data prediction through the manifold hypothesis: observed data can occupy a low-dimensional structure within a high-dimensional observation space, whereas noise-containing targets require modeling additional off-manifold variation. This motivates predicting the perturbed image in pixel space.

Image prediction also provides a shared output for pixel-wise regression and biological response supervision. Feature-based losses make errors in encoded phenotypic properties contribute directly to training; App. A.2 gives a local mathematical interpretation.

Perturbed-image prediction and bridge regression. A 1,024-dimensional perturbation embedding is added to the time embedding. The generative model predicts $\hat{X}_0 = D_\theta(X_t, t, c)$ using a preconditioned parameterization. We train this prediction with weighted pixel-wise mean squared error:

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}\left[\frac{w(t)}{d_x}\left\|\hat{X}_0 - X_0\right\|^2\right]. \tag{4}$$

Here, d_x is the number of image entries and $w(t)$ is the preconditioning-based regression weight. The expectation is over training image pairs and perturbation conditions, sampled times t, and Gaussian noise Z. As in Section 2.2, $\hat{X}_0$ denotes a random variable and $\hat{x}_0$ a realization. The control

X_T enters through the bridge state and sampling equations, rather than as a separate network input. App. B specifies the preconditioning, regression weight, and relation to the fully endpoint-conditioned bridge score.

Biological features and supervision targets. Let $f_i = \phi(\hat{x}_{0,i}) \in \mathbb{R}^{d_f}$, where ϕ is a frozen Inception-v3 encoder (Heusel et al., 2017) with $d_f = 2048$. The encoder parameters remain fixed while gradients from the biological response losses pass through it to update the generative model. The biological response supervision pipeline uses the first three image channels with differentiable preprocessing; the bridge regression uses all image channels. Reference statistics and prototypes are constructed from real perturbed training data. The biological response objectives use the condition statistics and biological annotations available in each dataset.

Condition-moment matching. When real perturbed training data are grouped by perturbation condition, their feature moments provide supervision targets. For each valid condition group c in a minibatch, let $\hat{\mu}_c, \hat{v}_c$ be the gate-weighted feature mean and diagonal variance of predicted perturbed images, and μ_c, v_c the corresponding statistics of real perturbed training data. We use

$$\mathcal{L}_{\mathrm{mom}} = \frac{1}{|\mathcal{C}_B| d_f} \sum_{c \in \mathcal{C}_B} \left[\|\hat{\mu}_c - \mu_c\|^2 + \lambda_{\mathrm{var}} \|\hat{v}_c - v_c\|^2 \right]. \tag{5}$$

Here, $\mathcal{C}_B$ is the set of valid condition groups in the minibatch, and λ_{var} balances mean and variance matching. We apply condition-moment matching on all three datasets; on BBBC021, $\lambda_{\mathrm{var}} = 0.5$.

Biological contrast. We compute probabilities q_{ij} by applying a softmax with temperature τ to cosine similarities between predicted-image features and condition prototypes from real perturbed training data. Let $\mathcal{P}_i$ contain the positive prototypes for sample i, defined by the requested condition and available biological annotations; the remaining prototypes are negatives. Following the multi-positive supervised contrastive reduction (Khosla et al., 2020),

$$\ell_i^{\mathrm{con}} = -\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log q_{ij}. \tag{6}$$

The batch loss $\mathcal{L}_{\mathrm{con}}$ is a time-gated weighted average of these per-image losses. The loss has the same form across datasets, while the positive and negative sets depend on their condition definitions and annotations. On BBBC021, we use 26 condition prototypes and $\tau = 0.1$, with positives sharing the requested MoA. This encourages shared MoA-associated features, while moment matching retains condition-specific constraints.

Joint training objective. Across all three datasets, we combine bridge regression with moment matching and biological contrast:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{bridge}} + \lambda_{\mathrm{mom}} \mathcal{L}_{\mathrm{mom}} + \lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}}. \tag{7}$$

Biological response losses use time-dependent gating and a gradual weight schedule during training. App. E details the implementation and weight settings. App. A decomposes the prototype loss.

Bridge sampling. We use an ECSI-based bridge sampler (Zhang et al., 2026b), followed by deterministic DBIM steps (Zheng et al., 2025). Each step combines the predicted perturbed image with the control image and analytic path coefficients. Sampling requires no biological evaluation. App. B provides the update equations and implementation details.

4 Experiments

4.1 Experimental setups

Datasets and backbone. We evaluate PerturbBridge on three microscopy benchmarks covering perturbation-induced cellular phenotypes and subcellular organization. BBBC021 (Caie et al., 2010) contains compound-treated cells with mechanism-of-action (MoA) annotations, enabling evaluation of whether generated images preserve features associated with the corresponding biological mechanism. JUMP (Chandrasekaran et al., 2023) provides perturbation-labeled images across experimental plates, supporting evaluation of perturbation identity through cross-plate retrieval. Allen-Cell (Allen Institute) captures cellular structures under compound treatments, allowing assessment of changes in subcellular organization and compound retrieval within each imaging target. Our

Table 1: BBBC021 MoA with CellProfiler features and DINOv2-L/14 features. Bridge-only omits biological response losses. All scores are percentages. Real perturbed data provide a reference. The CellProfiler-based classifier is trained on real perturbed data and tested on 5,120 generated images.

	CellProfiler features			DINOv2-L/14		
Method	Acc.↑	Macro-F1↑	W.-F1↑	Acc.↑	Macro-F1↑	W.-F1↑
IMPA (Palma et al., 2025)	57.9	53.9	60.7	57.8	36.2	60.6
CellFlux (Zhang et al., 2025)	72.2	65.8	73.2	67.4	43.4	66.6
CellFlux+bio	73.3	66.6	74.1	69.8	45.4	69.1
Bridge-only	73.1	67.6	74.9	70.6	45.8	70.1
PerturbBridge	**77.8**	**73.0**	**79.3**	**74.9**	**46.9**	**74.3**
Real Perturbed Data	80.1	76.3	81.8	78.6	57.6	79.7

standard BBBC021 comparison uses 63,781 perturbed training images and 5,120 matched test images. We adopt the 56.2M-parameter backbone and data preprocessing of CellFlux (Zhang et al., 2025), and set the bridge noise scale to $\gamma_{\max} = 0.1$ on BBBC021 and JUMP. PerturbBridge is trained for 150, 140, and 240 epochs on BBBC021, JUMP, and AllenCell, respectively. We compare against baselines with available implementations and additionally report published FID/KID scores for SimuSDB (Zhang et al., 2026a) (code not released).

Metrics. We assess biological responses through MoA classification on BBBC021 and perturbation retrieval on JUMP and AllenCell, alongside image-distribution quality. Distributional scores alone do not establish response correctness, as illustrated theoretically for pooled and conditional FID in App. A.1. MoA evaluation uses DINOv2 (Oquab et al., 2024) and CellProfiler (Carpenter et al., 2006) features, while retrieval uses DINOv2 features; none is used for biological response supervision, which relies on Inception-v3. We report classification accuracy, macro-F1, and weighted-F1; retrieval Hit@k, mean reciprocal rank (MRR), and mean average precision (mAP); median rank on JUMP; and mean rank on AllenCell. We report Fréchet and Kernel Distances in Inception-v3 (FID/KID) (Heusel et al., 2017; Bińkowski et al., 2018) and DINOv2 spaces (FDD/KDD) (Stein et al., 2023). Inception-based MoA scores and FID/KID follow the CellFlux protocol. See App. C.

4.2 Biological-level evaluation

PerturbBridge generates images that better preserve MoA-associated phenotypes. Table 1 shows improved MoA classification with PerturbBridge on BBBC021 using CellProfiler features and DINOv2-L/14. Relative to CellFlux, accuracy increases from 72.2% to 77.8% and from 67.4% to 74.9%, corresponding to gains of 5.7 and 7.5 percentage points, respectively. With CellProfiler features, macro-F1 increases from 65.8% to 73.0%, and weighted-F1 from 73.2% to 79.3%; both DINOv2 F1 scores also improve. These gains hold under both equal class weighting and weighting by class frequency. The two representations capture complementary image properties: CellProfiler measures morphology through features such as shape, intensity, and texture, while DINOv2 provides learned visual features. Neither is used for biological response supervision, which relies on Inception-v3 features. Their consistent improvements therefore support better preservation of MoA-associated phenotypes beyond the supervising representation.

PerturbBridge shows clearer separation of selected MoA classes in feature space. The shared UMAP projection in Fig. 3 provides qualitative evidence by visualizing the 512-dimensional hidden features of a frozen MoA classifier. Each point represents an image colored by its MoA label, and the projection allows comparison of class organization across real perturbed data, CellFlux, and PerturbBridge. In the highlighted regions, MoA classes form more distinct groups with less overlap in PerturbBridge samples than in CellFlux samples. This suggests better preservation of features that distinguish biological responses, consistent with the improved MoA classification scores.

PerturbBridge generates images with more identifiable perturbation-specific features. Table 2 shows improved DINOv2-based retrieval relative to CellFlux, using single-cell Mahalanobis retrieval across plates on JUMP and condition-profile retrieval within the same target on AllenCell. On JUMP, Hit@10 increases from 10.38% for CellFlux to 12.15% for PerturbBridge, while median rank decreases from 82 to 74. On AllenCell, Hit@1 increases from 62.14% to 69.29%. Further details of the retrieval protocols are provided in App. C.3.

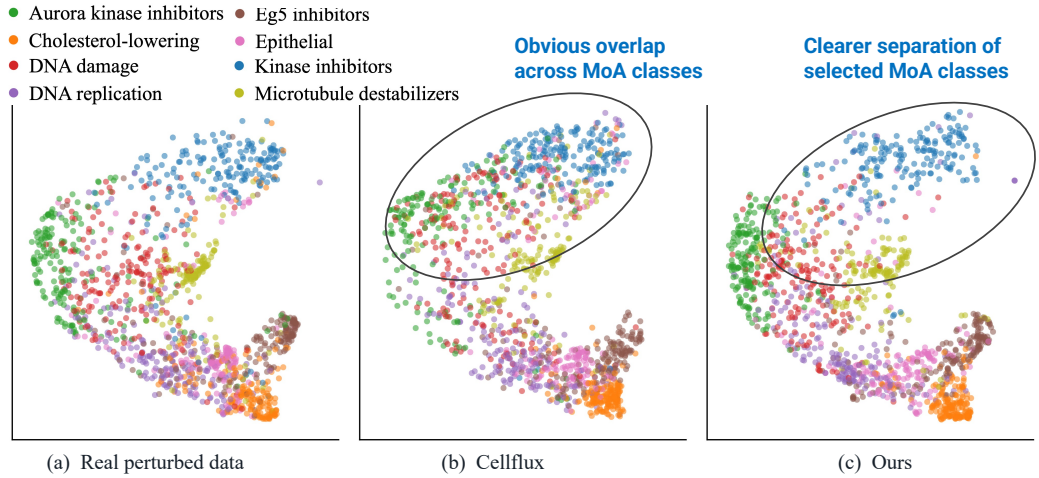

Figure 3: **MoA class separation on BBBC021.** (a) Real perturbed data; (b) CellFlux; (c) PerturbBridge (ours). Each point represents a cell image, colored by its MoA label. All panels use a shared UMAP projection of the 512-dimensional hidden features from a frozen MoA classifier. Ellipses highlight overlap across MoA classes in CellFlux and clearer separation in PerturbBridge.

Table 2: Perturbation retrieval with DINOv2 features. JUMP uses single-cell Mahalanobis retrieval across plates. AllenCell uses condition profiles to retrieve compounds within the same target from profiles of real perturbed training data. Bridge-only omits biological response losses and corresponds to the ECSI-only baseline on AllenCell. Hit rates are percentages; bold and underlined values mark the best and second-best generated-image results within each panel.

	JUMP						AllenCell			
Method	Hit @1↑	Hit @5↑	Hit @10↑	MRR ×100 ↑	mAP ×100 ↑	Median rank↓	Hit @1↑	MRR↑	mAP↑	Mean rank↓
IMPA (Palma et al., 2025)	1.50	5.41	9.52	4.65	0.88	84	52.14	0.73	0.73	1.64
CellFlux (Zhang et al., 2025)	1.56	5.78	10.38	4.92	<u>0.92</u>	82	<u>62.14</u>	0.78	0.78	1.58
Bridge-only	<u>1.75</u>	<u>5.98</u>	<u>10.42</u>	<u>5.08</u>	<u>0.92</u>	<u>79</u>	<u>62.14</u>	<u>0.79</u>	<u>0.79</u>	<u>1.54</u>
PerturbBridge	**2.04**	**7.27**	**12.15**	**5.94**	**0.97**	**74**	**69.29**	**0.82**	**0.82**	**1.48**
Real Perturbed Data	2.93	7.94	12.14	6.64	1.10	77	76.19	0.87	0.87	1.29

PerturbBridge captures characteristic perturbation-induced morphological changes. Fig. 4 compares control images, PerturbBridge samples, and real perturbed images on the AllenCell drug perturbation dataset. The selected examples show changes in ER organization, myosin bundles, actin distribution, tight junctions, and Golgi morphology, each corresponding to a phenotype reported for the same compound and cell line. PerturbBridge samples exhibit patterns resembling those in real perturbed images, and the resemblance is compound-specific rather than a generic treated-cell appearance: the Golgi breaks into discrete foci under brefeldin A but into larger, diffuse fragments under SNB, and both patterns are reproduced. This provides qualitative support for the preservation of perturbation-associated phenotypes.

PerturbBridge improves MoA classification on standard and unseen-compound splits. Table 3 compares MoA classification results on BBBC021 using the released CellFlux classifier with Inception-v3 features. For comparability, we follow the standard and unseen-compound experimental settings of CellFlux (Zhang et al., 2025), including its data splits, preprocessing, and evaluation protocol. Relative to CellFlux, PerturbBridge improves accuracy from 64.9% to 72.4% on the standard split and from 43.2% to 59.0% on unseen compounds. On the standard split, macro-F1 increases from 42.1% to 46.9%, and weighted-F1 from 63.8% to 72.0%. On unseen compounds, the corresponding scores increase from 36.6% to 47.3% and from 42.8% to 55.2%. These improvements show that the MoA classification gains also hold under the CellFlux evaluation protocol and extend to compounds excluded from training.

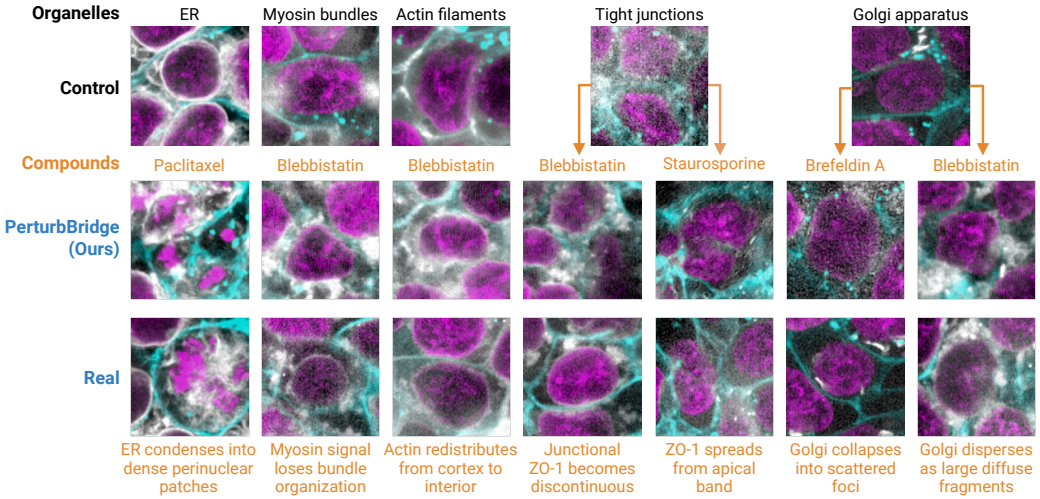

Figure 4: **Comparison of perturbation-induced changes on AllenCell.** Columns show seven compound–structure combinations involving the endoplasmic reticulum (ER), myosin bundles, actin filaments, tight junctions, and Golgi apparatus. Control images appear at the top, PerturbBridge samples in the middle (Ours), and real perturbed images at the bottom (Real). Cell nuclei are shown in purple and plasma membranes in blue. SNB denotes (S)-nitro-blebbistatin; annotations highlight subcellular changes. Examples are shown in App. F.

Table 3: BBBC021 MoA with Inception-v3 features on the standard and unseen-compound splits. Classification scores are percentages; bold and underlined values mark the best and second-best generated-image results within each split. Dashes indicate unavailable results. Bridge-only omits biological response losses. CellFlux+bio applies matched biological response losses to the predicted perturbed image reconstructed from velocity outputs. Gray rows denote ablation variants.

	Standard split			Unseen compounds		
Method	Acc.↑	Macro-F1↑	W.-F1↑	Acc.↑	Macro-F1↑	W.-F1↑
PhenDiff (Bourou et al., 2024)	52.6	33.6	52.1	9.6	9.3	7.4
IMPA (Palma et al., 2025)	63.7	40.2	64.8	16.0	10.0	13.1
CellFlux (Zhang et al., 2025)	64.9	42.1	63.8	43.2	36.6	42.8
CellFluxRL (Wu et al., 2026)	66.2	<u>42.9</u>	<u>67.9</u>	–	–	–
CellFlux+bio	66.3	42.6	65.1	44.7	<u>39.5</u>	45.2
Bridge-only	<u>67.7</u>	42.2	67.5	<u>50.7</u>	37.4	<u>46.1</u>
PerturbBridge	**72.4**	**46.9**	**72.0**	**59.0**	**47.3**	**55.2**
Real Perturbed Data	73.6	69.7	72.1	88.0	85.0	88.0

PerturbBridge improves image-distribution fidelity alongside biological task performance. Table 4 compares image-distribution quality across all three datasets in Inception-v3 and DINOv2 feature spaces. Relative to CellFlux, PerturbBridge reduces FID from 18.8 to 16.4 on BBBC021, from 9.0 to 8.3 on JUMP, and from 28.4 to 26.7 on AllenCell. KID also decreases on all three datasets. The improvements extend to DINOv2 features: FDD decreases from 207.2 to 173.8, from 123.1 to 108.8, and from 181.2 to 146.3, respectively, with lower KDD on each dataset as well. On unseen BBBC021 compounds, FID decreases from 42.0 to 29.5. Together, these results support improved distributional agreement with real perturbed data across datasets and feature representations.

4.3 Ablation Studies

Biological response supervision improves MoA preservation. Tables 3 and 4 compare PerturbBridge with the Bridge-only variant, which omits biological response supervision. On BBBC021, adding this supervision increases MoA accuracy from 67.7% to 72.4% on the standard split and from 50.7% to 59.0% on unseen compounds, with improvements in both macro-F1 and weighted-F1. On the standard split, FID decreases from 17.3 to 16.4 and KID from 1.02 to 0.91. These results show

Table 4: Image-distribution results in Inception-v3 (FID, KID) and DINOv2 (FDD, KDD) feature spaces. Lower is better. KID is multiplied by 100; KDD is unscaled. Bold and underlined values mark the best and second-best results within each dataset and metric. Dashes denote unreported values. CellFluxRL and SimuSDB FID/KID values are from Wu et al. (2026) and Zhang et al. (2026a), respectively. Gray rows denote ablation variants.

	BBBC021				JUMP				AllenCell			
Method	FID↓	KID↓	FDD↓	KDD↓	FID↓	KID↓	FDD↓	KDD↓	FID↓	KID↓	FDD↓	KDD↓
PhenDiff (Bourou et al., 2024)	49.5	3.10	250.1	1.55	49.3	5.10	247.3	2.76	48.8	2.75	192.6	1.56
IMPA (Palma et al., 2025)	33.7	2.60	293.9	2.57	14.6	1.10	371.1	4.43	37.6	2.25	424.1	4.98
CellFlux (Zhang et al., 2025)	18.8	1.61	207.2	1.73	9.0	0.63	123.1	1.35	28.4	1.46	181.2	1.71
CellFluxRL (Wu et al., 2026)	24.0	1.40	–	–	–	–	–	–	–	–	–	–
SimuSDB (Zhang et al., 2026a)	16.5	1.32	–	–	8.7	0.58	–	–	–	–	–	–
CellFlux+bio	18.4	1.59	206.0	1.75	9.8	0.70	130.5	1.46	**25.9**	0.99	162.0	1.34
Bridge-only	17.3	1.02	**165.7**	**1.21**	**8.2**	0.47	139.1	1.52	27.0	1.27	171.0	1.58
PerturbBridge	**16.4**	**0.91**	173.8	1.30	8.3	**0.38**	**108.8**	**1.04**	26.7	**0.97**	**146.3**	**1.08**

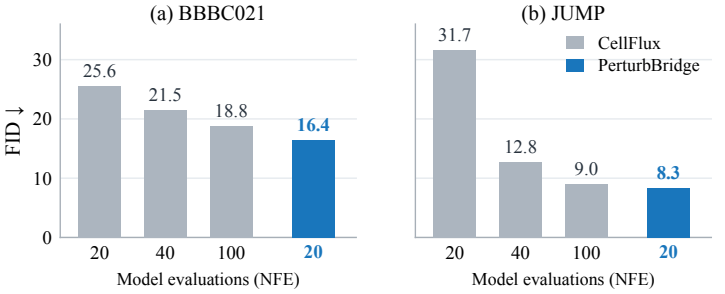

Figure 5: Sampling-budget comparison on BBBC021 and JUMP. Lower FID is better. Gray bars: CellFlux; blue bars: PerturbBridge. Bar labels report FID; horizontal axes indicate model evaluations (NFE).

that biological response supervision improves MoA preservation beyond bridge regression alone, including for compounds excluded from training.

PerturbBridge improves most evaluated metrics over CellFlux+bio. Tables 3 and 4 compare PerturbBridge with CellFlux+bio, which applies biological response supervision to perturbed images reconstructed from CellFlux velocity predictions. On BBBC021, CellFlux+bio reaches 66.3% MoA accuracy, compared with 72.4% for PerturbBridge. On JUMP, its FID is 9.8, compared with 8.3 for PerturbBridge. Together with the Bridge-only comparison, these results support the combined design of direct perturbed-image prediction and biological response supervision.

PerturbBridge achieves lower FID with fewer model evaluations. Fig. 5 compares CellFlux at 20, 40, and 100 model evaluations (NFE) with PerturbBridge at 20 NFE. At the matched budget of 20 NFE, PerturbBridge achieves FID values of 16.4 on BBBC021 and 8.3 on JUMP, compared with 25.6 and 31.7 for CellFlux. These scores also improve on CellFlux at 100 NFE, which reaches 18.8 and 9.0, respectively. Both methods use the same logical NFE convention; with classifier-free guidance, each logical evaluation uses two network forwards (App. B).

5 Conclusion

We introduce PerturbBridge, a stochastic data-to-data bridge model that directly supervises phenotypic responses in predicted perturbed images during generative training. By jointly optimizing bridge regression and differentiable biological response losses through a frozen encoder, PerturbBridge incorporates biological objectives without requiring RL post-training or biological evaluation during sampling. Across BBBC021, JUMP, and AllenCell, the model improves MoA classification, perturbation retrieval, and image-distribution metrics relative to CellFlux. Gains under CellProfiler and DINOv2 representations extend beyond the features used for supervision. Ablations further demonstrate the contribution of biological response supervision. With a bridge-specific sampler, PerturbBridge achieves lower FID than the 100-NFE CellFlux baseline using 20 model evaluations. These findings support direct biological response supervision as a practical approach to cellular perturbation generation and highlight the value of evaluating generated images through downstream biological tasks alongside distributional metrics.

AI USE STATEMENT

AI tools assisted with code implementation under human instructions and with language polishing of the manuscript. The authors remain responsible for the code, analyses, and final content of the paper.

REPRODUCIBILITY STATEMENT

All experiments use publicly available datasets: BBBC021, JUMP, and AllenCell. Our code is included in the supplementary materials. The appendices describe the model implementation, training configuration, biological response losses, sampling procedures, and evaluation protocols to support reproduction of the reported results.

REFERENCES

Allen Institute. Drug perturbation pilot study. `https://www.allencell.org/drug-perturbation-pilot.html`. Accessed: 2026-09-25.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Anis Bourou, Thomas Boyer, Marzieh Gheisari, Kévin Daupin, Véronique Dubreuil, Aurélie De Thonel, Valérie Mezger, and Auguste Genovesio. Phendiff: Revealing subtle phenotypes with diffusion models in real images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 358–367. Springer, 2024.

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.

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

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.

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*, 2010.

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, pp. 2023–03, 2023.

Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024.

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. *Advances in neural information processing systems*, 30, 2017.

Albert Z Hung, Charles J Zhang, Jonathan Z Sexton, Matthew James O'Meara, and Joshua D Welch. Lumic: Latent diffusion for multiplexed images of cells. *bioRxiv*, pp. 2024–11, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Alexis Lamiable, Tiphaine Champetier, Francesco Leonardi, Ethan Cohen, Peter Sommer, David Hardy, Nicolas Argy, Achille Massougbodji, Elaine Del Nery, Gilles Cottrell, et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 23–29 Jul 2023.

Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *Nature communications*, 15(1):1594, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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

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, 03 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz158.

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

Boris M Slepchenko and Leslie M Loew. Use of virtual cell in studies of cellular dynamics. *International review of cell and molecular biology*, 283:1–56, 2010.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shiye Su, Yuhui Zhang, Linqi Zhou, Rajesh Ranganath, and Serena Yeung-Levy. Stochastic perturbations improve distribution-to-distribution generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3965–3974, 2026.

Dawn C. Walker and Jennifer Southgate. The virtual cell—a candidate co-ordinator for 'middle-out' modelling of biological systems. *Briefings in Bioinformatics*, 10(4):450–461, 03 2009. ISSN 1467-5463. doi: 10.1093/bib/bbp010.

Dongxia Wu, Shiye Su, Yuhui Zhang, Elaine Sui, Emma Lundberg, Emily B. Fox, and Serena Yeung-Levy. CellFluxRL: Biologically-Constrained Virtual Cell Modeling via Reinforcement Learning. In *The 19th European Conference on Computer Vision*, 2026.

Lingchong You. Toward computational systems biology. *Cell Biochemistry and Biophysics*, 40(2): 167–184, 2004. doi: 10.1385/CBB:40:2:167.

Rongchao Zhang, Chengxin Li, Yiwei Lou, Yuling Shi, Hanpin Wang, and Yu Huang. Steering where to diffuse: Generative modeling of phenotypic response simulation with steered diffusion bridge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27367–27377, 2026a.

Shaorong Zhang, Yuanbin Cheng, and Greg Ver Steeg. Exploring the design space of diffusion bridge models. *Advances in Neural Information Processing Systems*, 38:48112–48154, 2026b.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey J Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating cellular morphology changes via flow matching. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76111–76138. PMLR, 13–19 Jul 2025.

Kaiwen Zheng, Guande He, Jianfei Chen, Fan Bao, and Jun Zhu. Diffusion bridge implicit models. In *International Conference on Learning Representations*, volume 2025, pp. 81857–81884, 2025.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, volume 2024, pp. 8160–8171, 2024.

A MATHEMATICAL SUPPORT FOR THE MOTIVATION AND METHOD

A.1 FID RANKING REVERSAL: STATEMENT AND PROOF

Shared appearance and experimental context can dominate image features, while perturbation-specific changes may be subtle (Chandrasekaran et al., 2024; Moshkov et al., 2024). Fig. 6 illustrates how matching non-target appearance can coexist with an incorrect phenotype. The construction below shows that both pooled and conditional FID can favor biologically incorrect outputs. This concerns what the metrics certify, rather than whether ideal conditional generative learning can recover the target distribution.

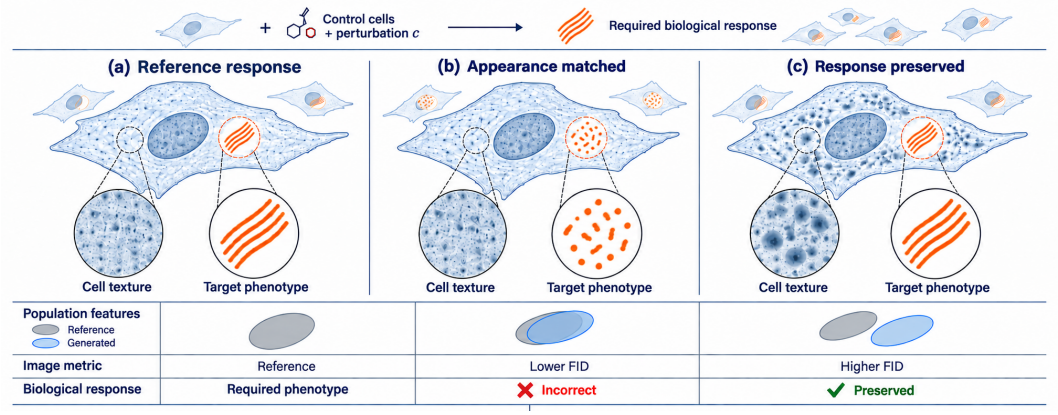

Figure 6: **Image fidelity and biological correctness.** Orange fibers depict the target phenotype; blue-gray texture depicts non-target intracellular appearance. Paired enlargements isolate both features. Relative to (a), (b) matches non-target appearance but disrupts the phenotype, whereas (c) preserves the phenotype despite a texture shift. Population-feature ellipses schematize the resulting disagreement between FID and biological correctness; the FID ordering follows Proposition 1. Cells and ellipses are conceptual illustrations.

Let P denote the real perturbed distribution, with equally likely conditions $C \in \{-1, +1\}$. Let $U \sim \mathcal{N}(0, I_d)$ be independent of C, $\delta > 0$, and $a \in \mathbb{R}^d \setminus \{0\}$. The final feature coordinate determines the biological label through $h(x) = \mathrm{sign}(x_{d+1})$. Compare

$$X = (U, \delta C), \qquad Q^{\mathrm{keep}} : \hat{X} = (U + a, \delta C), \qquad Q^{\mathrm{flip}} : \hat{X} = (U, -\delta C). \tag{8}$$

The first model preserves the requested label but shifts unrelated appearance; the second reverses the label while preserving the pooled image distribution.

Proposition 1 (FID and biological utility can rank models oppositely)**.** *Using identity features, and defining conditional FID as the average FID within each condition,*

$$\mathrm{FID}_{\mathrm{id}}(P_X, Q_X^{\mathrm{flip}}) = 0, \qquad \mathrm{FID}_{\mathrm{id}}(P_X, Q_X^{\mathrm{keep}}) = \|a\|^2, \tag{9}$$

$$\mathrm{cFID}_{\mathrm{id}}(P, Q^{\mathrm{flip}}) = 4\delta^2, \qquad \mathrm{cFID}_{\mathrm{id}}(P, Q^{\mathrm{keep}}) = \|a\|^2. \tag{10}$$

If $0 < 4\delta^2 < \|a\|^2$, both metrics prefer Q^{flip}, although its predicted biological labels are always wrong and those of Q^{keep} are always correct.

Proof. The pooled means are zero for P and Q^{flip} and $(a, 0)$ for Q^{keep}, with common covariance $\mathrm{diag}(I_d, \delta^2)$. The covariance contribution to FID therefore vanishes, leaving squared mean differences of zero and $\|a\|^2$. Conditional on $C = c$, the means are $(0, c\delta)$, $(a, c\delta)$, and $(0, -c\delta)$, with common covariance $\mathrm{diag}(I_d, 0)$, giving conditional FIDs $\|a\|^2$ and $4\delta^2$. The signs of the final coordinates establish label correctness. This is a feature-space illustration, not an empirical result on the cellular datasets.

A.2 Small image errors can carry large biological consequences

In Equation (8), removing the phenotype coordinate leaves normalized squared error $\delta^2/(d+1)$ but removes all condition information. A feature-sensitive loss can assign weight to this low-energy coordinate.

For image error $e = \hat{x}_0 - x_0$, a differentiable encoder ϕ with Jacobian J_ϕ at x_0, and $\lambda \geq 0$, the local expansion is

$$\|e\|^2 + \lambda \|\phi(x_0 + e) - \phi(x_0)\|^2 = e^\top \big(I + \lambda J_\phi^\top J_\phi\big) e + o(\|e\|^2). \tag{11}$$

This follows by substituting the first-order expansion of ϕ and squaring. The added term weights directions captured by the encoder. This local illustration explains the role of feature-sensitive supervision; the implemented losses use condition moments and prototypes rather than paired feature regression.

A.3 The implemented loss explicitly constrains label association

For a predicted-image feature Z and requested MoA label A, let $\mathcal{P}_A$ be the positive prototypes, $K_A = |\mathcal{P}_A|$, $q_A(Z) = \sum_{j \in \mathcal{P}_A} q_j(Z)$, and $r_j(Z) = q_j(Z)/q_A(Z)$ within this set.

Proposition 2 (Biological classification inside prototype contrast)**.** *For strictly positive prototype probabilities, the multi-positive loss satisfies*

$$\ell(Z, A) = -\frac{1}{K_A} \sum_{j \in \mathcal{P}_A} \log q_j(Z) = -\log q_A(Z) + \log K_A + D_{\mathrm{KL}}(U_A \| r_A(Z)), \tag{12}$$

where U_A is uniform on the positive set.

Proof and interpretation. Substitute $q_j = q_A r_j$ into the average negative log probability. The first term encourages probability mass on the requested MoA; the KL term distributes it across positive prototypes, while $\log K_A$ is constant for a given positive set. Condition-moment matching separately constrains compound-specific statistics. This decomposition describes the training objective and does not by itself guarantee classification performance on final generated images.

B Bridge training, design, and sampling

B.1 Path construction and the role of prior bridge models

We use the path and coefficients in Equations (2) and (3), with $t = 1$ at the control and $t = 0$ at the perturbed target. Sampling an image pair, an interior time, and Gaussian noise constructs a training state without simulating a trajectory.

I^{2}SB derives an analytic Gaussian bridge conditional on image pairs (Liu et al., 2023), while DDBM learns a bridge score with input/output preconditioning and weighted regression (Zhou et al., 2024). DBIM reuses a trained predictor for implicit sampling (Zheng et al., 2025), and ECSI separates path coefficients, preconditioning, and sampling stochasticity (Zhang et al., 2026b). We use the linear ECSI path and apply biological response supervision to predicted perturbed images.

B.2 Conditional distributions and the bridge score

At an interior time, the path has Gaussian density $q_t(x \mid x_0, x_T)$ with mean $\alpha_t x_0 + \beta_t x_T$ and covariance $\gamma_t^2 I$. Integrating over the target distribution induced by the training coupling gives $q_t(x \mid x_T, c)$. Differentiating this mixture expresses its score through the posterior mean of the perturbed target:

$$\begin{aligned} m_t(x, x_T, c) &= \mathbb{E}[X_0 \mid X_t = x, X_T = x_T, C = c], \\ \nabla_x \log q_t(x \mid x_T, c) &= -\frac{x - \alpha_t m_t(x, x_T, c) - \beta_t x_T}{\gamma_t^2}. \end{aligned} \tag{13}$$

This identity assumes positive interior noise and differentiation under the integral. It conditions explicitly on X_T, whereas the implemented predictor receives (X_t, t, c). App. B.4 explains this distinction and the condition under which the conditional means coincide.

B.3 THE IMPLEMENTED REGRESSION AND TRAINING CONFIGURATION

The preconditioned image prediction is

$$D_\theta(X_t, t, c) = c_{\text{skip}}(t)X_t + c_{\text{out}}(t)F_\theta(c_{\text{in}}(t)X_t, c_{\text{noise}}(t), c). \tag{14}$$

For our path, $\alpha_t = 1 - t$, $\beta_t = t$, and $\gamma_t = 2\gamma_{\max}\sqrt{t(1-t)}$. The BBBC021 preconditioner uses $S = \sigma^2_{\text{data}} = 0.25$ and $C_{xy} = 0.125$:

$$H_t = (\alpha_t^2 + \beta_t^2)S + 2\alpha_t\beta_t C_{xy} + \gamma_t^2, \qquad c_{\text{in}}(t) = H_t^{-1/2}, \tag{15}$$

$$c_{\text{skip}}(t) = \frac{\alpha_t S + \beta_t C_{xy}}{H_t}, \qquad c_{\text{out}}(t) = \sqrt{\frac{\beta_t^2(S^2 - C_{xy}^2) + S\gamma_t^2}{H_t}}, \tag{16}$$

$$c_{\text{noise}}(t) = 250\log(t + 10^{-44}), \qquad w(t) = c_{\text{out}}(t)^{-2}. \tag{17}$$

Here C_{xy} is a fixed preconditioning parameter. The raw U-Net output is combined with the skip term to produce D_θ, as in Equation (14); the regression averages squared error over image entries with weight $w(t)$. The bridge and sampler retain the control image explicitly. The implemented U-Net uses the inputs specified in Section 3.1; the distinction from a fully endpoint-conditioned regression is stated below.

Table 5 summarizes the BBBC021 architecture, optimization, and grouped-sampling settings. Condition groups provide multiple images for estimating the feature moments used by biological response supervision; classifier-free condition dropout also supplies unconditional training batches.

Table 5: BBBC021 training configuration.

Component	Setting
Backbone	56,201,219-parameter ADM/U-Net; RGB 96×96
Architecture	Base width 128; four residual blocks; multipliers $(2, 2, 2)$; attention resolution 48; one head; dropout 0.3
Blocks	Scale–shift normalization; new attention order; convolutional resampling and residual up/downsampling disabled
Perturbation	1,024-dimensional molecular embedding; local-batch condition dropout 0.2
Batching	Eight devices, 16 cells each; eight cells per compound; global batch 128
Optimizer	RAdam; learning rate 10^{-4}; weight decay 0; default betas; no learning-rate annealing
Precision and EMA	FP16 with gradient scaling; model EMA 0.999
Time and path noise	Uniform $t \in [0.0001, 0.9999)$; $\gamma_{\max} = 0.1$
Training budget	150 epochs, 499 grouped batches per epoch

PerturbBridge is trained for 140 epochs on JUMP and 240 epochs on AllenCell.

The 499 batches per epoch are a grouped-sampling budget, $\lceil 63781/128 \rceil$. Treated cells are combined with randomly sampled controls from the same experimental batch. Each image receives the original pixel dequantization $(\text{pixel} + U[0, 1))/255$, independent horizontal/vertical flips with probability 0.3, and mapping to approximately $[-1, 1]$. Target and control images pass through these transforms separately. Additional Gaussian noise augmentation of control images is disabled. App. E gives the biological response losses and their weight schedule.

B.4 SAMPLING DYNAMICS AND NUMERICAL IMPLEMENTATION

At an interior time, the predicted perturbed image determines the residual coordinate:

$$\hat{z}_t = \frac{x_t - \alpha_t\hat{x}_0 - \beta_t x_T}{\gamma_t}. \tag{18}$$

The deterministic DBIM update combines this residual with the control and predicted perturbed images:

$$x_r = \alpha_r\hat{x}_0 + \beta_r x_T + \gamma_r\hat{z}_t. \tag{19}$$

For decreasing time, the ECSI-style stochastic update uses

$$b_t = \dot{\alpha}_t \hat{x}_0 + \dot{\beta}_t x_T + \left(\dot{\gamma}_t + \frac{\epsilon_t}{\gamma_t}\right)\hat{z}_t,$$
$$x_{t-h} = x_t - h b_t + \sqrt{2\epsilon_t h}\,\xi, \qquad \xi \sim \mathcal{N}(0, I), \quad h > 0. \tag{20}$$

Dots denote time derivatives. Following ECSI (Zhang et al., 2026b), the noise strength for our linear path is $\epsilon_t = 2\eta\gamma_{\max}^2$ in the interior, where η controls sampling stochasticity. The score-based derivation uses the posterior mean in Equation (13); the implementation substitutes the learned prediction as discussed below. The final two steps set $\eta = 0$ and use the deterministic DBIM update in Equation (19), which returns $\hat{x}_0$ at $r = 0$.

Numerical grid and updates. The sampler uses $N = 20$ evaluations at times

$$t_i = \left[t_{\max}^{1/\rho} + \frac{i}{N-1}\left(t_{\min}^{1/\rho} - t_{\max}^{1/\rho}\right)\right]^{\rho}, \quad i = 0, \ldots, N-1, \tag{21}$$

followed by $t_N = 0$, with $t_{\max} = 0.9999$, $t_{\min} = 0.001$, and $\rho = 0.6$. It initializes the state with x_T at the first finite time. Residual-coordinate divisions are evaluated at interior times, where $\gamma_t > 0$. The first 18 transitions use the stochastic update, and the final two use Equation (19).

Guidance and counting convention. The CFG coefficient is fixed at $s = 0.2$ for PerturbBridge and all baselines on BBBC021 and JUMP. Under this convention, the guided denoiser returns $(1 + s)D_{\text{cond}} - sD_{\text{uncond}}$, using two separate network forwards. Thus 20 guided-denoiser evaluations correspond to 40 raw forwards. We count logical evaluations consistently with the CellFlux baseline sampling budget.

Conditioning in the implemented predictor. The endpoint-conditioned score above uses $\mathbb{E}[X_0 \mid X_t, t, c, X_T]$. Our network takes (X_t, t, c) as input and accesses the control through X_t. For pure weighted regression with a positive time-only weight, its population optimum is $\mathbb{E}[X_0 \mid X_t, t, c]$. These means coincide when (X_t, t, c) is sufficient for X_0 given X_T. The bridge identity motivates the sampling update. In implementation, we substitute the biologically supervised image prediction for the conditional mean, while retaining the control image explicitly in the update. This substitution is an approximation; exact equality is not established.

C Biological metrics and evaluation protocols

The protocols below distinguish MoA membership from exact perturbation identity. Evaluators use fixed references from real perturbed data; generated images are evaluated with their requested labels. The standard BBBC021 comparison uses 63,781 treated training images and 5,120 matched test images. The eight excluded compounds are docetaxel, AZ841, cytochalasin D, simvastatin, cycloheximide (listed as "cyclohexamide" in the dataset), latrunculin B, epothilone B, and lactacystin. The unseen-compound experiment follows the experimental setting of CellFlux (Zhang et al., 2025), using the same held-out compounds and evaluation protocol; its results are reported separately in Table 3.

C.1 BBBC021 MoA-Inception evaluation

Classifier and evaluation protocol. We reuse the released CellFlux MoA classifier and evaluator (Zhang et al., 2025). It extracts 2,048-dimensional features from uint8 cell images with the frozen Inception-v3 encoder used by the TorchMetrics FID implementation. A $2048 \to 512 \to K$ multilayer perceptron, with ReLU and dropout 0.5, predicts one of the K MoA labels in the CellFlux vocabulary. The classifier is trained on real perturbed training data using cross-entropy, Adam with learning rate 0.001, batch size 32, and 10 epochs. Preprocessing and augmentation follow the original CellFlux implementation. For each data split, the trained classifier is held fixed when evaluating real perturbed test data and the generated images from each method. In the standard-split comparison, generated RGB PNGs are decoded to $[0, 1]$ and converted by $(255x).\text{byte}()$ before the FID encoder. The released head uses 13 outputs including DMSO; the treated training and fixed test populations contain 12 and 9 labels. Its training uses independent horizontal and vertical flips with probability 0.3 and drops incomplete training batches.

During evaluation, the highest-scoring classifier output is compared with the biological label of the requested perturbation.

Accuracy and F1. Accuracy is the fraction of correctly classified images. Macro-F1 averages classwise F1 equally, while weighted-F1 weights each class by its true-label sample count. The class set is the union of true and predicted labels, following the original evaluator; undefined classwise F1 terms are set to zero. All three metrics are reported as percentages, using the same definitions for the CellProfiler-based classifier below.

C.2 BBBC021 MoA evaluation with cell profiling features

Images and morphology. We evaluate IMPA, CellFlux, Bridge-only, and PerturbBridge generated images. After excluding the eight held-out compounds in the CellFlux split, we use the first 5,120 treated test images in dataset order, with matched sample identifiers across real perturbed data and generated-image sets. Images remain native 96×96 uint8 RGB, without resizing or added contrast enhancement.

CellProfiler extraction and preprocessing. CellProfiler 4.2.8 native modules provide 2,106 candidate measurements of shape, intensity, texture, granularity, radial distribution, and colocalization, independently of the Inception-v3 encoder used for biological response supervision. Pixels are divided by 255 without channelwise percentile normalization. Native primary-object, secondary-object, and cytoplasm segmentation define the measured regions. Each image contributes measurements from the primary object closest to its center (within 24 pixels) and the corresponding regions, without averaging or taking medians across cells. Feature filtering on real perturbed training data yields $d_{\mathrm{CP}} \leq 2,106$ retained coordinates. Samples with failed segmentation or no qualifying object are retained, with missing measurements imputed by reference-feature medians; no central-circle fallback is used. The feature mask and imputation statistics are fixed from real perturbed training data and applied to all evaluation sets, giving $f \in \mathbb{R}^{d_{\mathrm{CP}}}$ per image.

Normalization and classification using only real perturbed training data. For feature j, let m_j, s_j, and d_j be the median, standard deviation, and scaled MAD on unflipped real perturbed training data:

$$d_j = 1.4826 \ \mathrm{median}_i \, |f_{ij} - m_j|, \qquad a_j = \max(0.001 s_j, 10^{-6}), \tag{22}$$

$$v_j = \begin{cases} d_j, & d_j > a_j, \\ \max(s_j, a_j), & \text{otherwise}, \end{cases} \qquad z_j = \mathrm{clip}\left(\frac{f_j - m_j}{v_j}, -12, 12 \right). \tag{23}$$

The normalization parameters fitted to real perturbed training data transform every evaluation set. A $d_{\mathrm{CP}} \to 512 \to 13$ MLP with ReLU and dropout 0.5 is trained with cross-entropy, Adam (learning rate 0.001), batch size 32 with the final incomplete batch dropped, and 10 epochs. Horizontal and vertical flips are sampled independently with probability 0.3 from precomputed original/H/V/HV features. The final classifier is frozen, with no test-time flips or test-based epoch selection. Its 13 outputs retain the original vocabulary including DMSO; treated training data contain 12 labels and this test subset contains 9. Accuracy is per-image correctness. Macro- and weighted-F1 use the union of true and predicted labels with undefined terms set to zero. No test or generated images are used to fit the scaler or classifier.

C.3 DINOv2 perturbation retrieval

Both evaluations use DINOv2 features. JUMP evaluates individual cells; AllenCell evaluates condition-level profiles. Their candidate sets and aggregation levels differ.

JUMP. There are 9,077 single-cell test queries. Candidates are ranked by increasing Mahalanobis distance in DINOv2 feature space. The gallery contains real perturbed test cells, with the query's entire plate excluded. Cells with the same perturbation identity are positives. The JUMP retrieval results use PerturbBridge trained for 140 epochs and the Bridge-only baseline. Table 2 uses the CellFlux outputs; the real perturbed data reference uses queries from real perturbed data under the same exclusion rule.

AllenCell. The evaluation contains 24 condition-profile queries. Each query retrieves among 2–5 compound candidates within the same target, using profiles from real perturbed training data as the gallery. The AllenCell retrieval results use PerturbBridge trained for 240 epochs, the Bridge-only

baseline, and CellFlux baseline. Table 2 compares these methods with real perturbed data. This is condition-profile retrieval within a target, without cross-plate exclusion or an unrestricted search over all target–compound combinations. The DINOv2 representation is separate from the Inception-v3 features used by the biological response losses.

Ranking metrics. Hit@k is the percentage of queries with at least one positive among the first k eligible candidates. MRR averages the reciprocal of the first-positive rank. For each query, average precision averages precision at the rank of every positive in the eligible gallery; mAP averages this quantity across queries. JUMP reports the median first-positive rank, whereas AllenCell reports the mean first-positive rank. These definitions use the dataset-specific galleries and positive sets described above.

Metric reporting. Unless otherwise noted, all quantitative results from our experiments are reported as the mean over three random seeds. FID and KID use Inception features; FDD and KDD apply Fréchet and kernel distribution comparisons in DINOv2-L/14 feature space. Table 4 reports KID multiplied by 100 and unscaled KDD. Dashes indicate unavailable values. Classification and retrieval hit rates are percentages. Table 6 reports the available standard deviations for PerturbBridge across three seeds. FID standard deviations range from 0.23 to 0.53 across datasets, while the reported BBBC021 MoA accuracy standard deviation is 0.350 percentage points. These values describe variation across runs, rather than uncertainty across individual cells. The displayed zero for AllenCell Hit@1 is at the reported precision and does not imply an absence of uncertainty. Standard deviations should be read in the units of each metric.

Table 6: Standard deviations across three random seeds for PerturbBridge on BBBC021, JUMP, and AllenCell; reported results are three-seed means. KID is multiplied by 100. MRR and mAP are multiplied by 100 on JUMP and remain unscaled on AllenCell, matching Table 2. MoA accuracy, F1 scores, and Hit@1 standard deviations are expressed in percentage points (p.p.). Standard deviations are rounded to three decimal places; dashes indicate unavailable statistics.

Metric	BBBC021	JUMP	AllenCell
FID	0.493	0.230	0.530
KID $\times 100$	0.072	0.039	0.126
FDD	6.855	3.410	7.940
KDD	0.089	0.065	0.095
MoA Accuracy (p.p.)	0.350	—	—
MoA Macro-F1 (p.p.)	0.887	—	—
MoA Weighted-F1 (p.p.)	0.450	—	—
Retrieval Hit@1 (p.p.)	—	0.110	0.000
MRR	—	0.192	0.001
mAP	—	0.009	0.001

D PSEUDOCODE

Algorithms 1 and 2 summarize training and generation. The predictor returns the preconditioned predicted perturbed image. Bridge coefficients and sampling settings are defined in App. B; biological response losses and their weights are defined in App. E.

`drop_condition` applies batchwise classifier-free dropout. The frozen encoder passes gradients to the predicted perturbed image. The weights of the two biological response losses specify the loss combination for each dataset. In sampling, `correct_prediction` applies the BBBC cached-prediction correction and otherwise returns the current prediction; `noise_strength` is zero for the final two transitions.

Training. Algorithm 1 constructs a noisy bridge state and predicts the perturbed image in a single training pass. Bridge regression is always active; batches retaining the perturbation condition additionally receive moment and MoA losses through the frozen encoder. The time gate and training-progress schedule control the contributions of these losses.

Algorithm 1 PerturbBridge: training objective.

```
# net: perturbed-image predictor; phi: frozen encoder
# x0, y, c: grouped target images, controls, conditions
# stats, prototypes: fixed references from real perturbed training data

def training_loss(net, x0, y, c, progress):
    t = sample_time(x0)
    z = randn_like(x0)
    alpha, beta, gamma = bridge_coefficients(t)
    xt = alpha * x0 + beta * y + gamma * z
    cond = drop_condition(c)
    xhat = net(xt, t, cond)
    loss = mean(weight(t) * (xhat - x0)**2)

    # Biological response supervision of the predicted image
    lm, la = bio_weights(progress)
    if cond is not None:
        f = phi(preprocess(xhat))
        g = time_gate(t)
        loss += lm * moment_loss(f, c, stats, g)
        loss += la * moa_loss(f, c, prototypes, g)
    return loss
```

Generation. Algorithm 2 initializes the state from a control image and applies classifier-free guidance to successive perturbed-image predictions. The known control and analytic bridge coefficients enter each update explicitly. Stochastic transitions are followed by deterministic finishing steps, without evaluating biological losses during generation.

Algorithm 2 PerturbBridge: sampling.

```
# net: trained perturbed-image predictor; y: control image
# times: decreasing grid; s: CFG coefficient

def generate(net, y, c, times, s):
    x, previous = y, None
    for t, r in adjacent_pairs(times):
        q = (1 + s) * net(x, t, c) - s * net(x, t, None)
        xhat = correct_prediction(q, previous, t, r)
        a, b, g = bridge_coefficients(t)
        zhat = (x - a * xhat - b * y) / g
        eps = noise_strength(t)

        if eps > 0:
            da, db, dg = bridge_derivatives(t)
            drift = da*xhat + db*y + (dg + eps/g)*zhat
            h = t - r
            x = x - h*drift + sqrt(2*eps*h)*randn_like(x)
        else:
            ar, br, gr = bridge_coefficients(r)
            x = ar*xhat + br*y + gr*zhat
        previous = (q, t)
    return x
```

E BIOLOGICAL RESPONSE LOSSES: IMPLEMENTATION AND SCHEDULES

BBBC021, JUMP, and AllenCell share the same two biological response objectives: condition-moment matching and InfoNCE-style contrastive supervision. Positive and negative sets are constructed according to each dataset's condition definitions and available biological annotations. On BBBC021, prototypes sharing the requested MoA are positives, while prototypes from other MoAs are negatives. The formulation below illustrates these shared objectives using BBBC021; both supervise the training-time predicted perturbed image $\hat{x}_0 = D_\theta(X_t, t, c)$.

E.1 DIFFERENTIABLE FEATURES AND CACHES OF REAL PERTURBED TRAINING DATA

The frozen Torch-Fidelity Inception-v3 encoder ϕ returns 2,048 features in evaluation mode. Its preprocessing takes the first three channels of $\hat{x}_0$ and uses straight-through clipping and centered one-bin dither. With sg denoting stop-gradient,

$$\mathrm{STclip}_{[a,b]}(v) = v + \mathrm{sg}(\mathrm{clip}_{[a,b]}(v) - v), \tag{24}$$

$$u_i = \mathrm{STclip}_{[0,1]}\left[\frac{\mathrm{STclip}_{[-1,1]}(\hat{x}_{0,i}) + 1}{2} + \frac{\delta_i}{255}\right], \quad \delta_i \sim U[-\tfrac{1}{2}, \tfrac{1}{2}], \tag{25}$$

$$f_i = \phi(255 u_i). \tag{26}$$

The encoder runs in FP32. Its parameters remain fixed, while gradients through f_i reach $\hat{x}_{0,i}$ and the generator. Offline MoA evaluation uses uint8 images as described in App. C.1; it does not add the dither used for biological response supervision.

The cache uses all 63,781 images in the real perturbed training data, without flips, with ordinary clipping and seeded dither (seed 20260810). For each compound condition c, it stores $\mu_c = \mathbb{E}[f \mid c]$ and diagonal population variance $v_c = \mathbb{E}[f^{\odot 2} \mid c] - \mu_c^{\odot 2}$. The map $A(c)$ supplies its MoA label. Conditions remain separate when they share a MoA.

E.2 TIME GATE AND CONDITION-MOMENT MATCHING

The two biological response losses share a smooth per-image gate:

$$g_i = \mathrm{sigmoid}\left(\frac{a - |c_{\mathrm{out}}(t_i)|}{b}\right), \quad a = \frac{1}{\sqrt{13}}, \quad b = \frac{a}{10}. \tag{27}$$

Smaller $|c_{\mathrm{out}}|$ gives a larger gate. For a local-batch condition group I_c, the weighted moments and valid groups are

$$G_c = \sum_{i \in I_c} g_i, \quad Z_c = \max(G_c, 10^{-6}), \tag{28}$$

$$\hat{\mu}_c = \frac{1}{Z_c}\sum_{i \in I_c} g_i f_i, \qquad \hat{v}_c = \max\left\{\frac{1}{Z_c}\sum_{i \in I_c} g_i f_i^{\odot 2} - \hat{\mu}_c^{\odot 2}, 0\right\}, \tag{29}$$

$$\mathcal{V} = \{c : |I_c| \geq 4,\ G_c > 10^{-6}\}. \tag{30}$$

We insert these moments into Equation (5), with $\mathcal{C}_B = \mathcal{V}$, $d_f = 2048$, and $\lambda_{\mathrm{var}} = 0.5$, averaging equally over valid conditions and feature dimensions. Each group contains eight cells; groups with fewer than four cells are excluded. The moment loss is zero when no group is valid.

E.3 MULTI-POSITIVE MOA PROTOTYPE CONTRAST

For BBBC021, we denote the general contrastive loss $\mathcal{L}_{\mathrm{con}}$ by $\mathcal{L}_{\mathrm{MoA}}$ and its weight λ_{con} by λ_{MoA}, reflecting the use of MoA labels to define positive prototypes. The cache supplies one unit-normalized prototype p_j per training condition, and $z_i = f_i / \|f_i\|_2$ is the normalized predicted-image feature. The 26 BBBC021 prototypes define

$$q_{ij} = \frac{\exp(z_i^\top p_j / \tau)}{\sum_{k=1}^{26} \exp(z_i^\top p_k / \tau)}, \qquad \tau = 0.1. \tag{31}$$

All prototypes sharing the requested MoA, including the prototype of the requested condition, are positives in Equation (6). The implementation uses equal temperature and base temperature, so their ratio is one. The batch reduction is

$$\mathcal{L}_{\mathrm{MoA}} = \frac{\sum_i g_i \ell_i^{\mathrm{MoA}}}{\max(\sum_i g_i, 10^{-12})}. \tag{32}$$

This retains the within-positive-set term in Proposition 2. Moment matching preserves compound-specific feature statistics, while MoA contrast supervises shared biological identity. Both pass gradients through the frozen encoder.

E.4 Biological response loss weight schedule

For BBBC021, biological response supervision is inactive for the first 40% of training. The moment and MoA weights then increase linearly from zero to 0.1 and 0.01, respectively, in the joint objective of Equation (7). The gate g_i weights samples within each loss, while this schedule controls their overall contribution. Batches with dropped perturbation conditions use bridge regression alone.

F Qualitative examples of generated cells

These examples complement the biological evaluations in Section 4. They compare selected real reference images with multiple PerturbBridge samples under the indicated conditions. The comparisons illustrate morphology and variation among displayed cells; they are not paired reconstructions or quantitative tests of population diversity.

Overview across seven structures. Fig. 7 organizes the comparison by compound and imaged structure, placing PerturbBridge samples and real references in parallel grids. The vehicle-control row provides context for the treatment rows, and highlighted entries direct attention to selected differences in subcellular organization. Blank entries denote combinations absent from the dataset. This overview extends the selected compound–structure comparisons in Fig. 4.

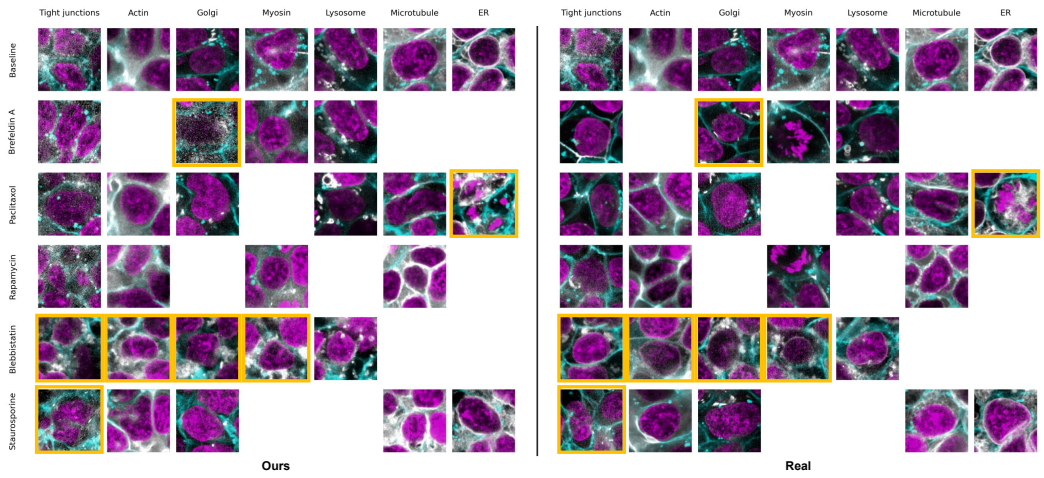

Figure 7: Drug-induced changes across seven cellular structures on AllenCell. PerturbBridge samples (left) and real references (right) are arranged by compound and structure. Vehicle denotes the control; SNB denotes (S)-nitro-blebbistatin. Yellow frames highlight selected comparisons; blank entries denote combinations absent from the dataset.

F.1 BBBC021: EXAMPLES ACROSS COMPOUND CONDITIONS

Cytochalasin B and AZ138. Fig. 8 contrasts two conditions with different MoA annotations. The cytochalasin B examples show irregular cell outlines and variation in peripheral staining, while the AZ138 examples include rounded or elongated cells with bright nuclear regions. Generated samples vary in shape and staining within each condition, allowing visual comparison with the real perturbed references. MoA preservation is assessed quantitatively in Tables 1 and 3.

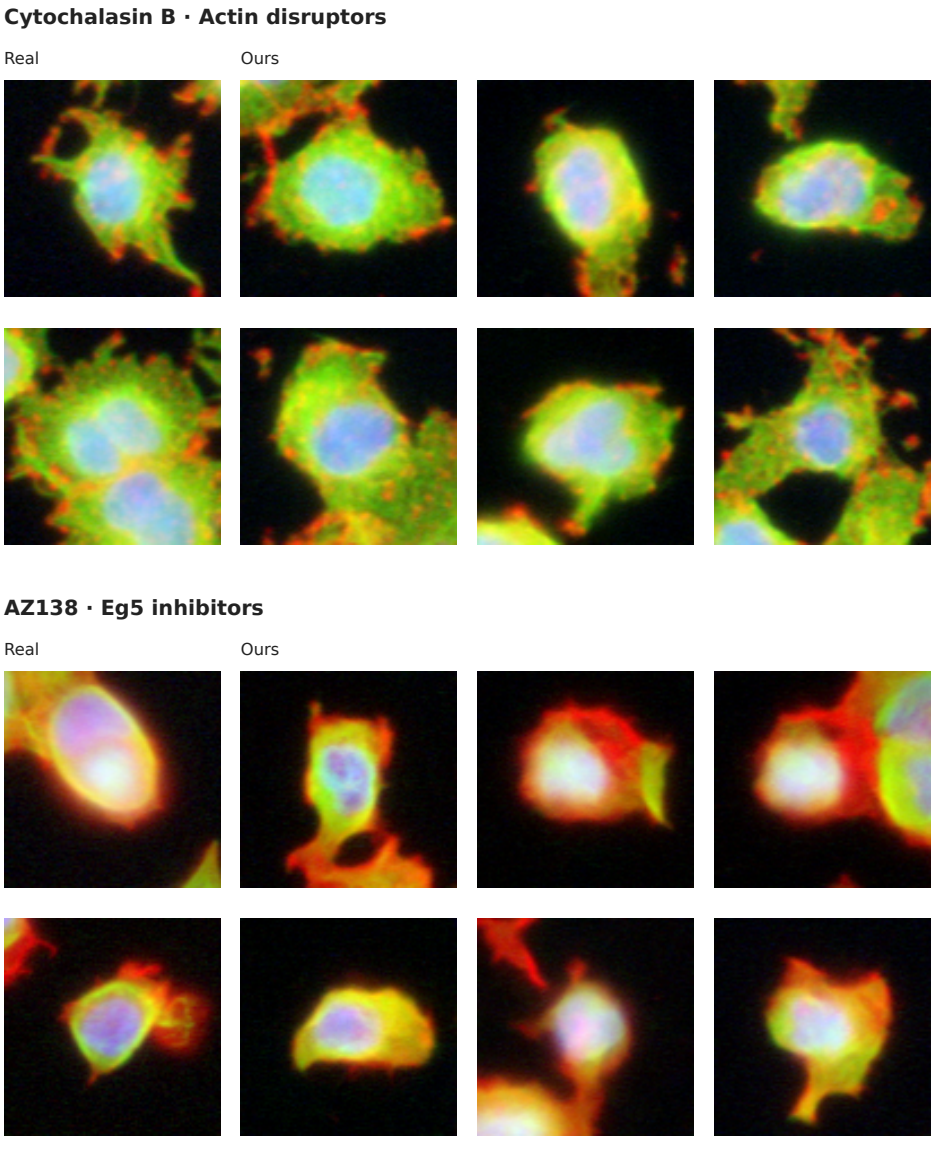

Figure 8: BBBC021 examples for cytochalasin B and AZ138. In each compound block, the left column contains real perturbed reference images and the remaining columns contain PerturbBridge samples.

AZ258 and PP-2. Fig. 9 extends the comparison to two additional compound conditions. The displayed cells differ in nuclear shape, cell extent, and the spatial distribution of channel intensities. Multiple generated images show how these properties vary under the same requested condition. Individual generated cells need not reproduce the geometry of a particular reference cell, because training and evaluation concern conditional cell populations.

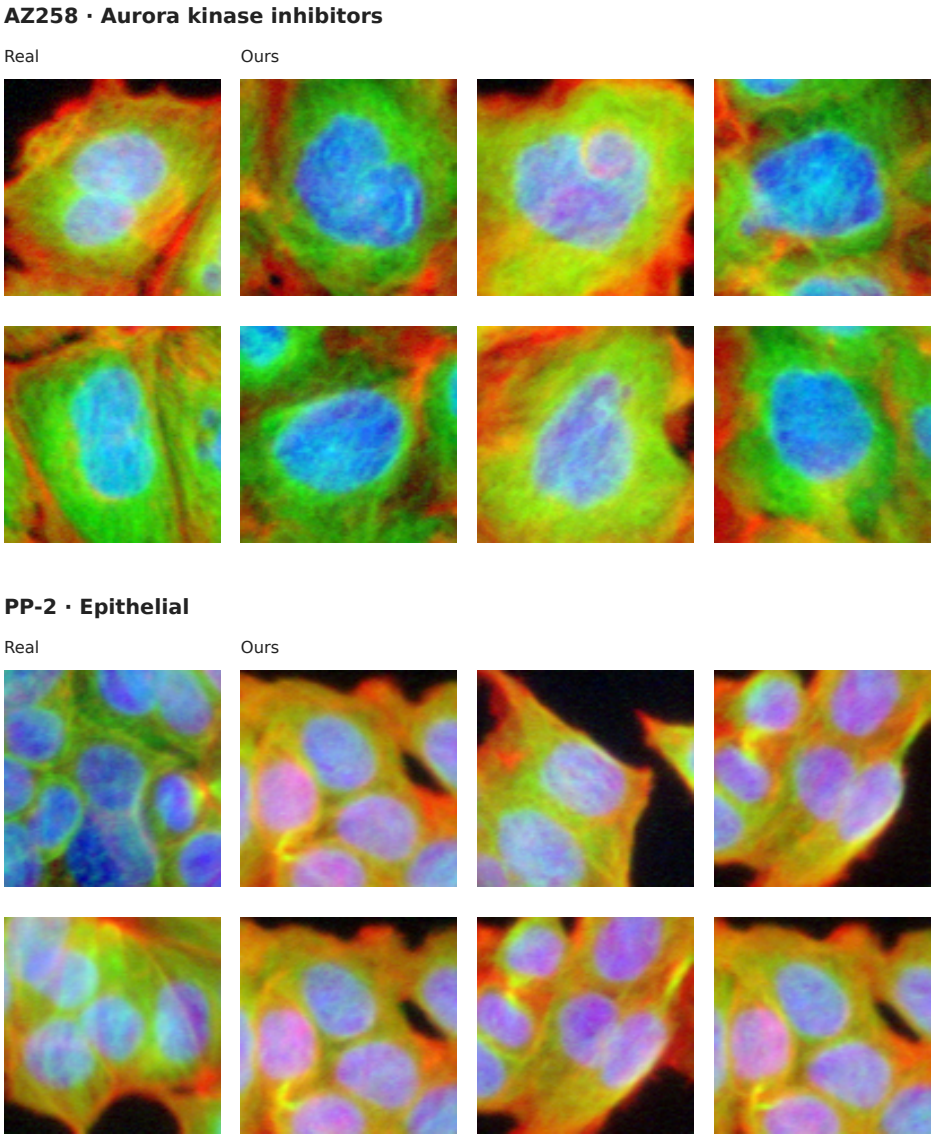

Figure 9: Additional BBBC021 examples for AZ258 and PP-2. Real perturbed references appear in the left column and PerturbBridge samples in the remaining columns.

F.2 ALLENCELL: STRUCTURE-SPECIFIC COMPARISONS

Brefeldin A and paclitaxel. Fig. 10 provides larger views of Golgi and microtubule imaging conditions. Vehicle and treated references allow changes in the location and organization of the fluorescent signal to be inspected alongside multiple generated samples. The samples exhibit variation in nuclear geometry and surrounding staining, rather than repeating a single cell appearance. These selected views complement the condition-profile retrieval results in Table 2.

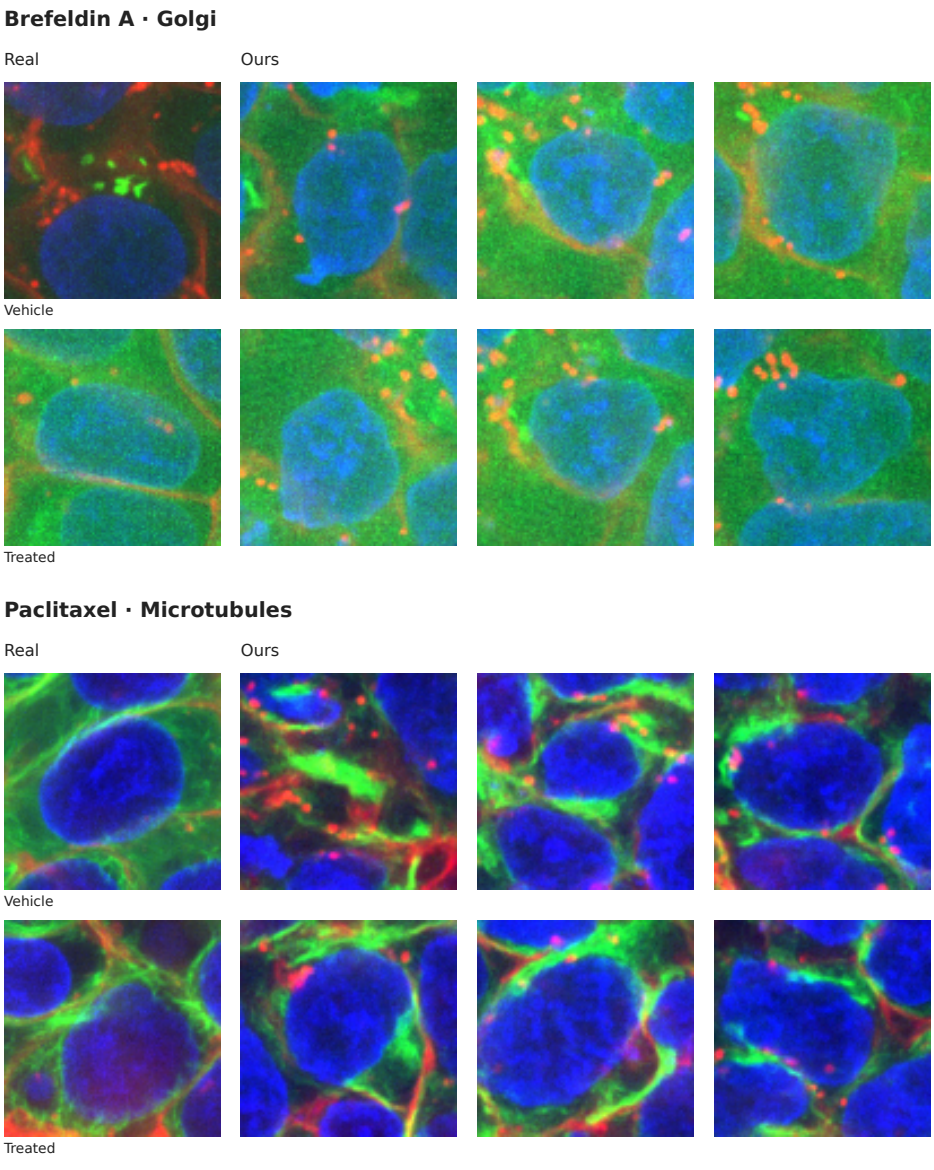

Figure 10: AllenCell examples for brefeldin A (Golgi) and paclitaxel (microtubules). The left column shows real vehicle and treated references; the remaining columns show PerturbBridge samples.

Two imaging targets under the same compound. Fig. 11 shows SNB examples for Golgi and tight-junction imaging. The two blocks emphasize different spatial features of the cellular response, so each should be compared with its corresponding vehicle and treated references. Together with the preceding examples, this comparison illustrates the distinct spatial appearances represented under different imaging targets.

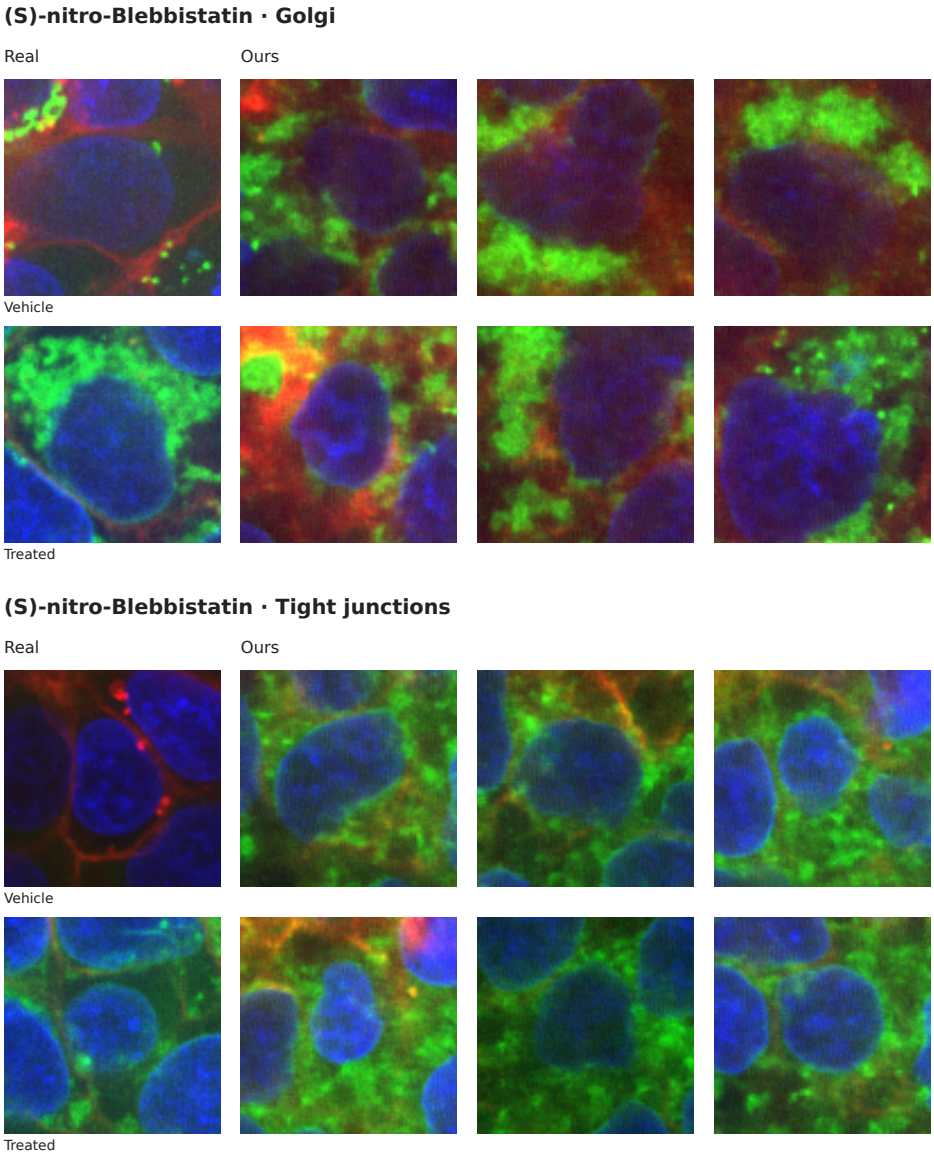

Figure 11: AllenCell examples for (S)-nitro-blebbistatin (SNB) under Golgi and tight-junction imaging. Real vehicle and treated references appear in the left column; PerturbBridge samples occupy the remaining columns.