RESPONDER-WEIGHTED FLOW MATCHING FOR CELL MORPHOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of cell morphology, such as CellFlux, learn to turn images of control cells into images of treated cells. They are trained on every imaged cell of a perturbation, although often only some of these cells respond. We show that a perfectly trained flow then reproduces this mixture, so its predicted change of any feature equals the responder effect times the fraction of responding cells. No single classifier-free guidance strength undoes this, because the required strength grows as one over that fraction. We therefore estimate for every treated cell the probability that it responded. A classifier compares the cell with control cells of the same plate and never sees the cell's own well during training. We then weight each cell's flow-matching loss by this probability. Estimated responder fractions in public datasets are often significantly smaller than one, especially CRISPR knockouts. In a benchmark where only one in ten training cells is a real treated cell, a CellFlux-style model recovers none of the perturbation effect on morphology. Responder weighting recovers 83% of it, close to the 89% of an oracle given the true labels. On real data, the proposed weighting mechanism lowers the energy distance to real treated cells by 29% on BBBC021 and 49% on RxRx1. On JUMP, where by our estimate the median perturbation changes fewer than one in five cells, it raises the distance, and only milder weighting improves the fit.

1 INTRODUCTION

Image-based profiling measures how chemical and genetic perturbations change cell morphology (Caicedo et al., 2017), and public Cell Painting datasets (Bray et al., 2016) now cover more than a hundred thousand perturbations (Chandrasekaran et al., 2023). Predicting these changes is one route toward a virtual cell (Bunne et al., 2024). In such experiments, cells grow in the wells of multi-well plates. Each well receives one perturbation or serves as a control. Single cell crops are cut around each nucleus in several imaged fields per well. Staining, imaging, and other experimental conditions vary across wells, plates, and experiments, introducing batch effects and making the controls specific to each plate and experiment. Imaging destroys the cells, so models like CellFlux (Zhang et al., 2025b) and CellFluxV2 (Zhang et al., 2026) learn a flow (Lipman et al., 2023) that transforms the images of control cells of a plate into the treated cells of the same plate.

These models use every treated cell as an example of the treated state, although treated cells rarely respond alike. Drug responses vary from cell to cell (Slack et al., 2008), and in CRISPR screens some cells escape the edit (Papalexi et al., 2021). In single-cell transcriptomics, the same problem motivated HiDDEN (Goeva et al., 2024): a label that marks the whole sample as treated is wrong for the cells that did not respond. CellCLIP (Lu et al., 2025) notes that many Cell Painting images of a perturbation resemble controls. In our measurements, fewer than half of the cells respond in 60% of the JUMP perturbations with a detectable effect, including 20 of the 23 CRISPR knockouts. Partial response is the norm rather than a corner case, and it is the problem this paper addresses.

A flow trained on such data learns this mixture. If a fraction ϱ of the treated cells responds and the rest looks like controls, even a perfectly trained flow predicts, for every feature, only ϱ times the change seen in the responders. For the treated population as a whole this is correct. A virtual cell, however, is usually asked what a perturbation does to a cell and how it acts on it. Most readouts average over the generated cells, and in such an average a strong effect on a few cells looks like a

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.

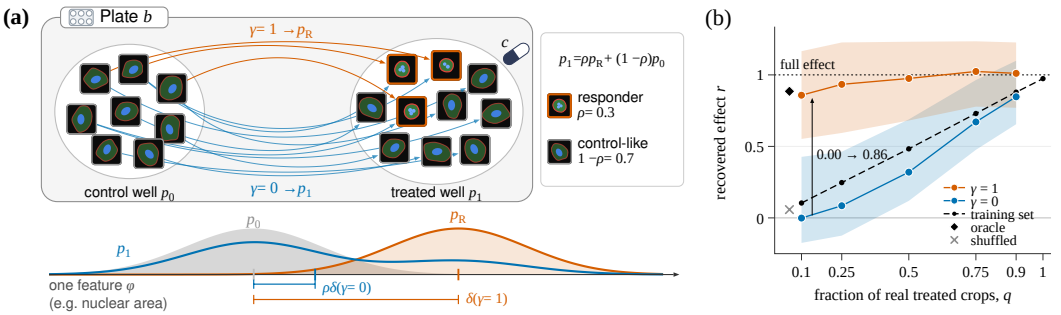

Figure 1: Training on all treated cells shrinks the effect. **(a)** A treated well mixes responders (p_R, fraction ϱ) with control-like non-responders. A CellFlux-style flow ($\gamma = 0$) reproduces only $\varrho\delta$ of a feature effect δ, and responder weighting ($\gamma = 1$) reproduces δ. **(b)** Recovered effect r (one is the full effect) against the fraction q of real treated crops in training, BBBC021 dilution benchmark.

weak effect on many. Classifier-free guidance (Ho & Salimans, 2022) cannot fix this with one setting, because the strength needed grows as $1/\varrho$, and ϱ differs between perturbations (Section 5.4).

We therefore want to encourage the generative model to focus on the responding cells rather than on all treated cells. Before training, we estimate for every treated cell the probability s that it responded. A classifier learns to distinguish the treated cells and control cells of the same plate in a cross-validation mode stratified by well. We then multiply each cell's flow-matching loss by s^γ (Figure 2). With $\gamma = 0$ this is the CellFlux objective, and with $\gamma = 1$ and calibrated scores it targets exactly the responders.

In summary, we explain why training on all treated cells shrinks every predicted effect and why guidance cannot undo it. We propose a scorer that turns frozen image embeddings into responder fractions and per-cell weights for flow matching. We test the method on a dilution benchmark in which the true effect is known, and on three public datasets.

2 Related work

Generative models of perturbed morphology. Mol2Image (Yang et al., 2021), PhenDiff (Bourou et al., 2024), IMPA (Palma et al., 2025) and MorphoDiff (Navidi et al., 2025) generate perturbed cells with flows, diffusion models or GANs. CellFlux (Zhang et al., 2025b) formulates the task as flow matching from same-plate controls, and CellFluxV2 (Zhang et al., 2026) moves it to the latent space of an autoencoder. Recent work scales and simplifies these flows (Jones et al., 2026), adds uncertainty estimates (Wu et al., 2026b), fine-tunes them with rewards for biological plausibility (CellFluxRL, Wu et al., 2026a) or learns a condition-dependent source distribution for unseen perturbations (SP-FM, released as MixFlow, Rubbi et al., 2026), and CellFlow (Klein et al., 2025) applies flow matching to single-cell phenotypes more broadly. All of them treat every treated cell as a sample of the perturbed state. Our weighting changes only the training target, so it applies to any of them.

Responders in single-cell data. CellOT (Bunne et al., 2023) maps whole populations by optimal transport. Mixscape (Papalexi et al., 2021) separates perturbed from escaping cells in CRISPR screens, MELD (Burkhardt et al., 2021) estimates how likely each cell is under the treated condition, and HiDDEN (Goeva et al., 2024) turns the outputs of a treated-versus-control classifier into perturbation scores. Our scorer follows HiDDEN, but it is cross-fitted over wells, and it converts the classifier output into a calibrated probability of response by first estimating the fraction of responders.

Mixture proportions and weighted objectives. Estimating the fraction of responders is a mixture proportion problem (Blanchard et al., 2010), related to learning from positive and unlabeled data (Elkan & Noto, 2008). Weighting a loss so that a model samples from a reweighted distribution goes back to reward-weighted regression (Peters & Schaal, 2007) and has been analyzed for flow matching (Zhang et al., 2025a). Guidance instead amplifies the difference between the conditional

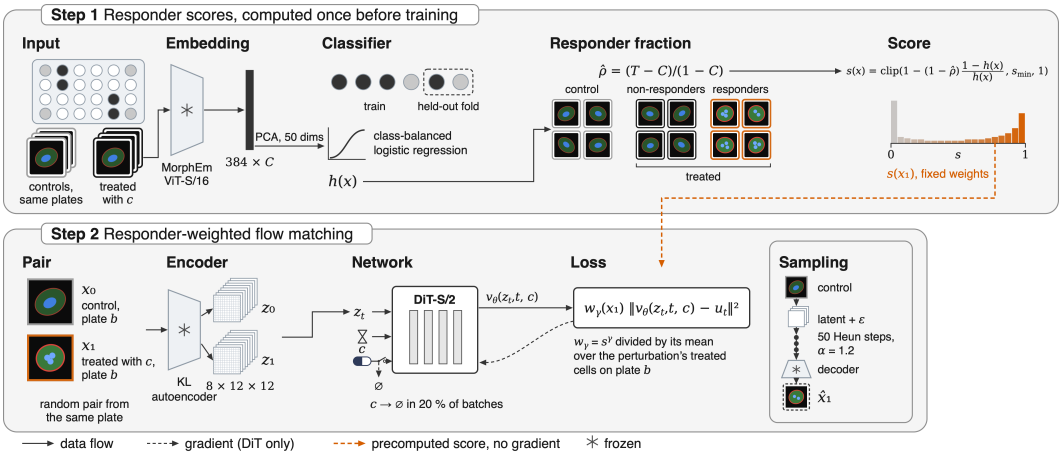

Figure 2: Method. **Top**, responder scoring, done once before training. A frozen MorphEm encoder embeds each image channel, and a class-balanced logistic regression, cross-fitted over wells, separates treated from same-plate control cells. Its out-of-fold scores give the responder fraction $\hat{\varrho}$ and the score s of each cell. **Bottom**, flow matching in the latent space of a frozen autoencoder, each pair of a control and a treated cell weighted by $s(x_1)^\gamma$, normalized within perturbation and plate.

and the unconditional field (Ho & Salimans, 2022; Zheng et al., 2023), and it does not sample from the intended distribution (Chidambaram et al., 2024).

3 Method

The method has two steps (Figure 2). A classifier first estimates which treated cells responded (Section 3.2), and flow matching then weights each cell by this estimate (Section 3.3). Section 3.1 sets up the responder model behind both steps, and Section 3.4 explains why the weighting is needed.

3.1 Setup and responder model

Consider one plate and one perturbation c, with control cells $x_0 \sim p_0$ and treated cells $x_1 \sim p_1$. A cell here is the latent code of a single-cell crop, on which the flow is trained. Each treated cell has an unobserved label $R \in \{0, 1\}$ that says whether it responded, and we assume the responder model

$$p_1 = \varrho\, p_R + (1 - \varrho)\, p_0, \tag{1}$$

with responder fraction $\varrho = \mathbb{P}(R = 1)$ and responder distribution p_R. It states that non-responding cells are distributed exactly like the controls of the same plate. This can fail, for example when responses are graded, and we test it in Section 5.1. The split into responders and non-responders is not unique, because control-like cells among the responder could just as well be counted as non-responders. We therefore take the smallest ϱ compatible with Eq. (1), so that it counts the cells whose morphology differs from controls (Appendix B). As in CellFlux, flow matching (Lipman et al., 2023) draws x_0 and x_1 independently from the same plate and trains a velocity field $v_\theta(x, t, c)$ to predict the direction $x_1 - x_0$ of the straight path $x_t = (1 - t)\, x_0 + t\, x_1$ (Liu et al., 2023), by minimizing $\mathbb{E}\|v_\theta(x_t, t, c) - (x_1 - x_0)\|^2$. New cells are generated by following v_θ from a control cell.

3.2 Responder scores

Scores are computed once per perturbation, before generative training. A perturbation here is a compound at one dose in BBBC021, an siRNA in RxRx1, and a compound, CRISPR guide or ORF in JUMP. Its treated cells can come from several wells and plates, and each plate contributes its own controls. Cells enter as 96×96 crops centered on the nucleus.

Embedding. Each image channel of a crop passes separately through MorphEm (Agrawal et al., 2026), a frozen ViT-S/16 trained with DINO (Caron et al., 2021) on single-channel microscopy images, and the channel embeddings are concatenated.

Cross-fitted classifier. A class-balanced logistic regression separates the treated cells of a perturbation from the same-plate controls. Its output $h(x)$ is the predicted probability that a cell is treated, and the odds $h/(1-h)$ estimate the density ratio $p_1(x)/p_0(x)$, which says how much more typical the cell is of treated than of control wells (Sugiyama et al., 2012). We cross-fit over wells, so each cell is scored by a classifier trained on other wells. Cells of one well also share their density, staining and focus, and a classifier that saw the well in training could mistake these for a response. We therefore split the wells into folds, train on all but one and score the held-out fold, so every cell is scored by a classifier that never saw its well (Chernozhukov et al., 2018).

Responder fraction. Take a threshold that a fraction C of the control scores exceed, say $C = 0.1$. Non-responders exceed it at the same rate, but responders exceed it more often. So, the fraction T of treated cells above it is larger. If all responders pass, $T = \varrho + (1-\varrho)\,C$, or, equivalently, $\varrho = (T-C)/(1-C)$. For example, 300 responders among 1,000 treated cells give $T = 0.3 + 0.7 \times 0.1 = 0.37$ and $\varrho = (0.37 - 0.1)/0.9 = 0.30$. We combine thresholds at the u-quantiles of the control scores, where $C(u) = 1-u$,

$$\hat{\varrho}_u = \underset{u \in [0.75,\, 0.95]}{\text{median}} \; \frac{T(u) - C(u)}{1 - C(u)}, \quad \text{clipped to } [0,1]. \tag{2}$$

Responders below the threshold lower the estimate, so $\hat{\varrho}$ does not overstate ϱ (Appendix B). What is important, is that the cells that respond clearly enough can be told apart from controls, which are the cells that weighting can emphasize.

Score of each cell. By Bayes' rule, a treated cell x responds with probability $\varrho\, p_R(x)/p_1(x)$. The responder models gives $\varrho\, p_R = p_1 - (1-\varrho\,)p_0$, so

$$\mathbb{P}(R = 1 \mid x) = \frac{\varrho\, p_R(x)}{p_1(x)} = \frac{p_1(x) - (1-\varrho)\, p_0(x)}{p_1(x)} = 1 - (1-\varrho)\,\frac{p_0(x)}{p_1(x)}.$$

The classifier odds estimate p_1/p_0, so replacing p_0/p_1 by $(1-h)/h$ and ϱ by $\hat{\varrho}$ results in the score

$$s(x) = \text{clip}\Big(1 - (1-\hat{\varrho})\,\frac{1-h(x)}{h(x)},\; s_{\min},\; 1\Big). \tag{3}$$

Cells that look like non-responders are scored with the lower bound $s_{\min}$, which keeps every weight positive. We cap $\hat{\varrho}$ at 0.95, because at $\hat{\varrho} = 1$ all scores would equal one, although the classifier still ranks some cells as more clearly changed than others. Calibrated scores average to ϱ over the treated cells of a perturbation, which gives us a check that needs no generative model.

3.3 Weighted flow matching

Latent space and network. As in CellFluxV2 (Zhang et al., 2026), a frozen KL autoencoder (Rombach et al., 2022) maps each crop to a $8 \times 12 \times 12$ latent code z. Each treated crop is paired with a random same-plate control crop, and a DiT-S/2 (Peebles & Xie, 2023) $v_\theta(z_t, t, c)$ is trained on the path

$$z_t = (1-t)\, z_0 + t\, z_1 + \sin^2(\pi t)\, \varepsilon, \qquad u_t = \mathrm{d}z_t/\mathrm{d}t = z_1 - z_0 + \pi \sin(2\pi t)\, \varepsilon, \tag{4}$$

with Gaussian noise ε, velocity target u_t and t drawn from a skewed distribution. The noise vanishes at $t = 0$ and $t = 1$, so Section 3.4 applies unchanged (Albergo et al., 2025). In BBBC021, c is the compound, while scores are computed per dose. Dropping c in 20% of the batches lets the model also train an unconditional field for guidance.

Weights and loss. With T_b the treated crops of a perturbation on plate b, each crop $x \in T_b$ gets a weight, and we minimize the weighted loss

$$w_\gamma(x) = \frac{s(x)^\gamma}{\frac{1}{|T_b|}\sum_{x' \in T_b} s(x')^\gamma}, \qquad \mathcal{L}_\gamma(\theta) = \mathbb{E}\big[w_\gamma(x_1)\, \|v_\theta(z_t, t, c) - u_t\|^2\big]. \tag{5}$$

The weights average to one within each perturbation and plate, so weighting changes which cells the model learns from, but not how much each perturbation and plate counts. With $\gamma = 0$ this is the CellFlux objective. We use $\gamma = 1$, which is exact for calibrated scores, and also report $\gamma \in \{0.25, 0.5, 2\}$. Weighting reduces the number of cells the model effectively learns from, which we measure with the effective sample size $\mathrm{ESS}_u = (\sum_{T_u} w)^2 / \sum_{T_u} w^2$ (Kish, 1965), the number of equally weighted cells that the weighted cells of perturbation u are worth. We train for 100 epochs from Gaussian noise and then for 1,000 epochs from controls.

Sampling. We add Gaussian noise to a same-plate control latent, integrate $v_\theta^\alpha = v_\theta(\cdot, \varnothing) + \alpha\,(v_\theta(\cdot, c) - v_\theta(\cdot, \varnothing))$ from $t = 0$ to $t = 1$ with 50 Heun steps spaced in time as in Karras et al. (2022), and decode, with the unconditional field $v_\theta(\cdot, \varnothing)$ and $\alpha = 1.2$ except in the guidance experiments.

3.4 Why training on all treated cells shrinks the effect

We now explain informally what a flow learns from the treated cells of Eq. (1) and why weighting changes it. Appendix B states and proves these results formally.

A perfect model reproduces the mixture. Flow matching regresses the path direction $x_1 - x_0$ on the current point x_t. The best regression function is the average direction of all paths through a point, $v^\star(x, t, c) = \mathbb{E}[x_1 - x_0 \mid x_t = x, c]$, and following $v^\star$ from a control cell reproduces p_1 exactly (Lipman et al., 2023; Liu et al., 2023). A perfectly trained model therefore reproduces the mixture, which shrinks the change of every feature.

Proposition 1 (Shrinkage)**.** *Under the responder model Eq. (1), the optimal flow started from p_0 generates $\hat{p} = p_1$, and every integrable feature ϕ, such as a CellProfiler measurement of nuclear area, satisfies*

$$\mathbb{E}_{\hat{p}}[\phi] - \mathbb{E}_{p_0}[\phi] = \varrho\,\big(\mathbb{E}_{p_R}[\phi] - \mathbb{E}_{p_0}[\phi]\big). \tag{6}$$

The predicted change of every feature is the responder effect times the responder fraction. The shrinkage comes from the training target, so more data or a larger network only converge to the shrunken answer.

Guidance cannot undo the shrinkage. Guidance (Ho & Salimans, 2022), which we use for sampling (Section 3.3), amplifies the difference between the conditional field and an unconditional field fitted to all perturbations pooled. At the start of sampling ($t = 0$), the optimal field points from the current cell to the mean of the target cells, so we can compute exactly what guidance does there.

Proposition 2 (Guidance)**.** *Let the perturbations share their controls, let δ_c be the mean change of the responders to perturbation c, and let $\bar{m}$ be the average of $\varrho_c \delta_c$ over the pooled training cells, to which the unconditional field is fitted. At $t = 0$, the guided optimal field $v_\alpha^\star = v^\star(\cdot, \varnothing) + \alpha\,(v^\star(\cdot, c) - v^\star(\cdot, \varnothing))$ moves control cells on average by $\bar{m} + \alpha\,(\varrho_c \delta_c - \bar{m})$, which for $\bar{m} = 0$ equals δ_c only if $\alpha = \alpha_c^\star = 1/\varrho_c$.*

When the changes of different perturbations roughly cancel ($\bar{m} \approx 0$), the required strength $1/\varrho_c$ differs between perturbations, so a single strength overshoots perturbations with many responders and undershoots those with few. This covers only the first step, and strong guidance also distorts the spread of the generated cells (Chidambaram et al., 2024), so the prediction is approximate. If a fraction $1 - q$ of the treated cells is replaced by controls, it predicts $\alpha^\star(q) = \alpha^\star(1)/q$.

Weighting targets the responders. Multiplying the loss of a cell by a weight acts like showing the model that cell more or less often, so a cell with weight near zero is almost ignored. Because the weight depends only on the treated cell, not on the control cell or on t, it changes only the distribution of treated cells that the flow is trained toward.

Proposition 3 (Weighting)**.** *If the loss of each pair is weighted by $s(x_1)^\gamma$, a score of the treated cell to the power $\gamma \geq 0$, the optimal flow transports p_0 to $p_1^{(\gamma)} = s^\gamma p_1 / Z_\gamma$, where $Z_\gamma = \mathbb{E}_{p_1}[s^\gamma]$. For the calibrated score $s(x) = \mathbb{P}(R = 1 \mid x)$ and $\gamma = 1$, $Z_1 = \varrho$ and $p_1^{(1)} = p_R$.*

For the calibrated score, Bayes' rule gives $s\,p_1 = \varrho\,p_R$, so the weights remove the non-responders exactly. Real scores are only estimates. If they leave weight on non-responders, $\gamma > 1$ pushes it

toward zero and concentrates the target on the most clearly changed cells, while $\gamma < 1$ moves the target toward all treated cells, so we also test other exponents.

4 EXPERIMENTAL SETUP

Datasets. BBBC021 (Caie et al., 2010) shows MCF-7 cells treated with compounds at up to eight doses, part of them annotated with 12 mechanism-of-action (MoA) classes (Ljosa et al., 2012). RxRx1 (Sypetkowski et al., 2023) shows U2OS cells with siRNA knockdowns. JUMP (Chandrasekaran et al., 2024), the CPJUMP1 pilot released as cpg0000 in the Cell Painting Gallery, shows Cell Painting images (Bray et al., 2016) of cells perturbed by compounds, CRISPR knockouts and ORF overexpression. We use the crops and splits of CellFlux (Zhang et al., 2025b). In the IID split, test cells come from perturbations seen in training, and in the OOD split of BBBC021 the test compounds are new. The responder analysis uses 84 BBBC021, 100 RxRx1 and 100 JUMP perturbations.

Recovered effect. We measure how much of an effect a model reproduces with CellProfiler features (Stirling et al., 2021), which describe size, shape and intensity and are independent of both the scorer's embedding and the generator's latent space. Features are computed on decoded crops and standardized by same-plate controls. With $\Delta(S)$ the mean feature shift of a set of crops S and H the held-out real treated crops of the same unit,

$$r(S) = \operatorname*{median}_{f} \frac{\Delta_f(S)}{\Delta_f(H)}, \qquad r_{\mathrm{proj}}(S) = \frac{\langle \Delta(S), \Delta(H) \rangle}{\|\Delta(H)\|^2}. \tag{7}$$

Each ratio $\Delta_f(S)/\Delta_f(H)$ is the fraction of the real change of the feature f that S reproduces, and their median is robust to features with a tiny real change, whose ratios are noisy. The projection r_{proj} measures how far the predicted effect goes along the real one. Both are one when every feature shifts by the real amount and zero without any change, and by the shrinkage result, an unweighted model trained with a fraction q of real treated crops should reach about q. Both can exceed one, because H contains the natural non-responders of the perturbation: a model of the responders alone should change more than H, and strong guidance can push cells further still. We report medians over perturbations. The evaluation was repeated in three runs with new samples (runs A, B and C), and we compare variants only within a run.

Dilution benchmark. Real data do not reveal which cells responded, so we build training sets where this is known. We take 31 BBBC021 perturbations from 11 compounds and hold out 2,261 real treated crops for evaluation. Training sets of 7,592 crops contain a fraction $q \in \{0.1, 0.25, 0.5, 0.75, 0.9, 1\}$ of real treated crops. The rest are same-plate controls, relabeled as treated and assigned to treated wells, so they act exactly like non-responders. The responder fraction is then $q\varrho_1$, where ϱ_1 is that of the undiluted units. We rescore every set and train with $\gamma = 0$ and $\gamma = 1$.

Real datasets and metrics. We train with $\gamma \in \{0, 0.25, 0.5, 1, 2\}$ on each dataset, with three seeds for the main variants, and report the mean and standard error (SE) over seeds. FID (Heusel et al., 2017) compares Inception features of all generated and real test crops. The energy distance (Székely & Rizzo, 2013) compares generated and real treated cells of each perturbation in the autoencoder's latent space, averaged over perturbations. It is zero only for equal distributions, and its unbiased estimator stays reliable for the few cells of one perturbation, unlike per-perturbation FID (Section 5.5). MoA accuracy (BBBC021) is the fraction of generated crops that a classifier trained on real crops assigns to the correct mechanism, and top-5 accuracy (RxRx1, JUMP) is the fraction whose perturbation is among the five most likely under such a classifier. On JUMP we also report the mean average precision (mAP).

5 RESULTS

5.1 MANY PERTURBATIONS CHANGE ONLY PART OF THE CELLS

We first check that the scores reflect real effects rather than differences between wells. Even wells with identical conditions differ, so a classifier can partly separate two groups of wells without any perturbation effect. To measure this we shuffle the treated and control labels between whole wells five

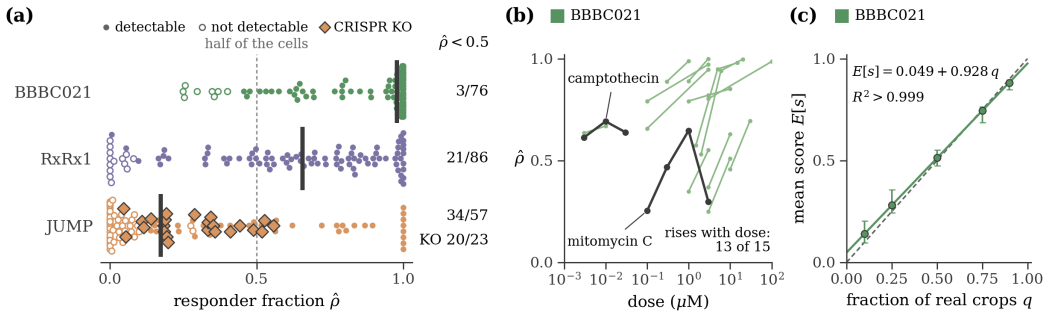

Figure 3: **(a)** Sorted $\hat{\varrho}$ for 84 BBBC021, 100 RxRx1 and 100 JUMP perturbations.**(b)** $\hat{\varrho}$ against log dose for the 15 non-saturated multi-dose BBBC021 compounds, the two that do not rise in black. **(c)** Mean score against the fraction q of real treated crops, fit $0.049 + 0.928\,q$, identity dashed.

times and score the perturbation again. The AUROC of these shuffled runs, in which labels reflect only difference between wells, form a null band. A perturbation is detectable when its real AUROC lies above its band, that is, when its effect exceeds what differences between wells explain. This holds for 76 of 84 BBBC021 treatments, 86 of 100 RxRx1 siRNAs and 57 of 100 JUMP perturbations (Figure 3a). Among the detectable ones, fewer than half of the cells respond in 3 of 76 BBBC021 treatments (4%), 21 of 86 RxRx1 siRNAs (24%) and 34 of 57 JUMP perturbations (60%), including 20 of 23 CRISPR knockouts, consistent with cells escaping the edit (Papalexi et al., 2021).

Three heuristic checks support these estimates. First, a higher dose should affect more cells, and $\hat{\varrho}$ rises with dose for 13 of the 15 multi-dose BBBC021 compounds whose $\hat{\varrho}$ is not already near one (Figure 3b). Second, $\hat{\varrho}$ is close to one for drugs that act on the cytoskeleton, on mitosis or on protein synthesis (class means 0.95 to 0.99), and much lower for kinase inhibitors, the epithelial class and drugs that damage DNA or block its replication (0.38 to 0.65, Figure 3d). Low values can mean that fewer cells respond or that many respond too weakly to be told apart from controls. An earlier analysis of BBBC021 found both: many perturbations change the cells of a well unevenly, and half of its mechanisms cannot be recognized by eye (Ljosa et al., 2013). Third, in the dilution benchmark, where the answer is known, calibrated scores should average to a value proportional to the share q of real crops. The mean score indeed grows along the line $0.049 + 0.928\,q$ ($R^2 > 0.999$, Figure 3c), whose intercept is the mean score of the injected controls.

The responder model also assumes that non-responders look exactly like controls. Treated and control cells that both score below a threshold should then be indistinguishable, with an AUROC near 0.5. The AUROC stays close to 0.5 on JUMP (0.48 to 0.58), but not on BBBC021 (0.63 to 0.80) or RxRx1 (0.70 to 0.77, Table 4), where even the least affected treated cells differ somewhat from controls, because cells respond to different degrees or wells differ technically. The weights then favor the most affected cells rather than select a clean set of responders. The next section tests how weighting copes with such non-responders.

5.2 Responder weighting recovers a diluted effect

Figure 1b shows the recovered effect r against the fraction q of real treated crops. A diluted training set changes by q times the full change, so its own recovered effect falls from 0.974 at $q = 1$ to 0.104 at $q = 0.1$, which a perfect unweighted model would reproduce. The trained unweighted model ($\gamma = 0$) lies even below it, at 0.847, 0.320 and -0.001 for $q = 0.9$, 0.5 and 0.1, against 0.879, 0.483 and 0.104 for its training data. At $q = 0.1$, half of its predicted feature changes point the wrong way. With responder weighting ($\gamma = 1$) the curve stays nearly flat, at 1.011, 0.975 and 0.858, and $\gamma = 1$ beats $\gamma = 0$ on at least 29 of 31 perturbations at every level. At $q = 0.1$, $\gamma = 1$ learns mostly from the few real crops, so its cells lie slightly closer to them than held-out cells do, a mild sign of memorization, but none is a copy (Appendix D).

So far, the injected controls come from the training plate and look exactly like controls. To mimic non-responders that differ from controls, we also inject controls from other plates, which differ from those of the training plate with an AUROC of 0.65 or 0.74 in two variants. At $q = 0.1$, $\gamma = 1$ still

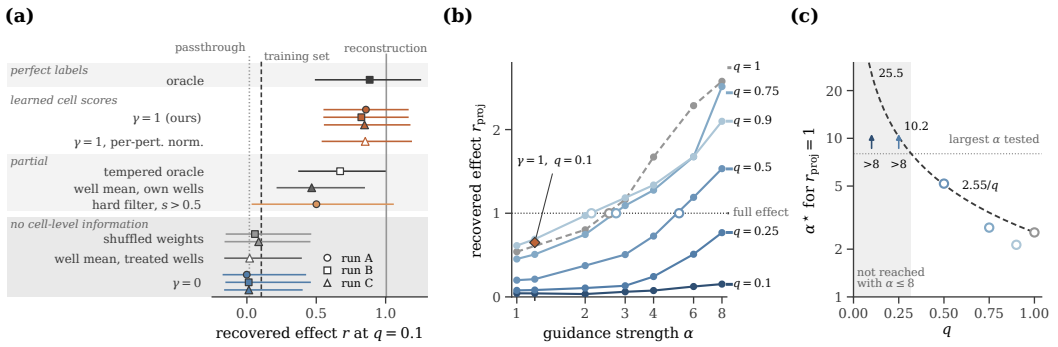

Figure 4: **(a)** Median and IQR of r over 31 perturbations per variant, marker shape by evaluation run. **(b)** r_{proj} of $\gamma = 0$ models against guidance strength. **(c)** $\alpha^\star$, measured (circles) and predicted (dashed). Numbers in Tables 7 and 8.

recovers 0.855 and 0.902, against 0.858 with same-plate controls, and an oracle that knows which crops are real treated cells recovers 0.889 and 0.894. The recovered effect of $\gamma = 0$ rises from -0.001 to 0.155 and 0.287, because part of the difference between plates points along the real effect. The second variant breaks the responder model about as strongly as the real BBBC021 data do (Table 9).

5.3 What drives the gain

To find out what produces the gain, we change one part of $\gamma = 1$ at a time, at $q = 0.1$ (Figure 4a). Samples differ between evaluation runs, so each variant is compared with $\gamma = 1$ from the same run. **Learned scores are close to perfect ones.** The oracle recovers 0.885, against 0.825 for $\gamma = 1$. **Fewer effective training cells do not explain the gain.** Weighting lets fewer cells dominate the loss, but two variants with the same effective number of training cells as $\gamma = 1$ do worse: shuffling the weights within each perturbation recovers only 0.058, and an oracle softened to the same spread of weights (scores raised to the power 0.5625) recovers 0.672 and loses on 29 of 31 perturbations. **Weights must be set per cell, not per well.** Giving each crop the mean score of its well works when the injected controls fill wells of their own (0.467), but not when they share wells with real treated crops (0.020), as real non-responders do, while $\gamma = 1$ recovers 0.848 in that run.

5.4 Guidance cannot replace the weights

Can stronger guidance restore the diluted effect instead? We sample the $\gamma = 0$ models of every dilution level, and one trained on the undiluted set, with α from 1 to 8 (Figure 4b), and call $\alpha^\star(q)$ the smallest strength at which r_{proj} reaches one. Guidance restores the full effect only for mild dilution ($q \geq 0.5$), and it needs more strength as dilution grows, 2.13 at $q = 0.9$ and 5.18 at $q = 0.5$ (Figure 4c). At $q = 0.25$ even $\alpha = 8$ reaches only 0.768, and at $q = 0.1$ it raises r_{proj} from 0.043 to just 0.153, while $\gamma = 1$ reaches 0.65 at the default strength.

The theory predicts $\alpha^\star(q) = \alpha^\star(1)/q$, with $\alpha^\star(1) = 2.55$ from the undiluted model. This value should be close to one, as nearly all cells of the undiluted perturbations respond, but even this model reproduces only 0.54 of the effect without guidance. The prediction fits $q = 0.5$ (5.10 against 5.18), explains why no tested strength suffices for $q \leq 0.25$, which would need about 10 and 25, and asks for somewhat more than needed at $q = 0.9$ and 0.75. No single strength suits all levels: even the most uniform one, $\alpha = 1$, leaves a spread of 0.57 across levels, against 0.18 for $\gamma = 1$. Strong guidance also costs image quality, raising the BBBC021 FID of $\gamma = 0$ from 9.73 at $\alpha = 1.2$ to 23.40 at $\alpha = 6$.

5.5 Responder weighting on real datasets

Figure 5 summarizes the results on the full datasets (numbers in Table 10). **Weighting improves BBBC021.** Over three seeds, the IID energy distance falls from 0.086 without weighting to 0.061 with $\gamma = 1$, a reduction by 29%, and to 0.058 with $\gamma = 2$. MoA accuracy rises from 0.618 to 0.633 and 0.639, and FID stays almost the same. Shuffled weights, which give the same weights to random

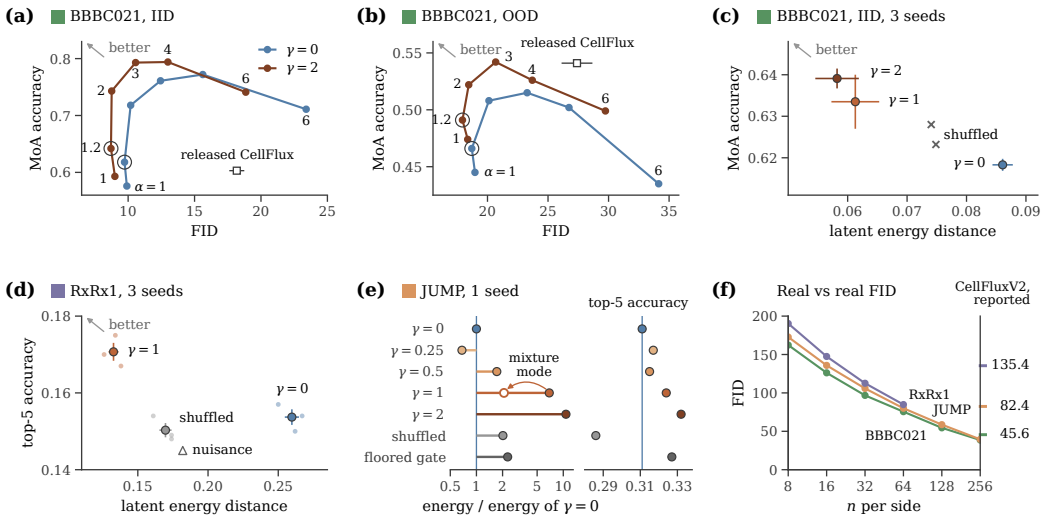

Figure 5: **(a, b)** FID against MoA accuracy for α from 1 to 6, one seed, rings at the default $\alpha = 1.2$. **(c, d)** Mean $\pm$ SE over three seeds, small dots in (d) for single seeds. **(e)** First seed. **(f)** FID between two disjoint sets of n real crops of one perturbation, ticks for the values reported for CellFluxV2. Numbers in Appendix E.

cells, lower the energy distance less (to 0.074), so part of the gain comes from reweighting as such and the rest from placing the weight on likely responders. On held-out compounds (OOD), only MoA accuracy improves clearly. At every guidance strength, $\gamma = 2$ has a lower FID and a higher MoA accuracy than $\gamma = 0$ (one seed, Figure 5a,b). In CellFlux's IID split, most test crops overlap a training crop from the same well, but the ranking also holds on those that overlap none. The appendix also compares our models with the released CellFlux checkpoint.

Weighting improves RxRx1. With $\gamma = 1$, the energy distance roughly halves (0.260 to 0.133), FID falls from 21.8 to 19.2 and top-5 accuracy rises from 0.154 to 0.171. Shuffled weights also lower the energy distance (to 0.170) but not top-5 accuracy (0.150), so reweighting as such improves the fit, while only responder weights make generated cells easier to assign to their perturbation. If $\gamma = 1$ has learned the responders, adding non-responders back should recover the treated population. In this mixture mode, each generated cell is kept with probability $\hat{\varrho}$, else replaced by a control, and the result matches the real treated cells as well as $\gamma = 0$ (1.02 times its energy distance, first seed at $\alpha = 1$).

Strong weighting hurts on JUMP, where by our estimate the median perturbation changes fewer than one cell in five ($\hat{\varrho} = 0.17$). Over three seeds, $\gamma = 1$ has 5.9 times the energy distance of $\gamma = 0$ and a higher FID (14.75 against 13.43), although it raises top-5 accuracy from 0.313 to 0.321 and mAP from 0.148 to 0.252 and beats shuffled weights on both (Table 17). Part of the gap is expected, as the reference contains all treated cells, most of which look like controls. In mixture mode, the ratio to $\gamma = 0$ falls from 6.9 to 2.1 for the first seed (Figure 5e), but stays above one. Mild weighting improves both the fit and top-5 accuracy, with $\gamma = 0.25$ reaching, over three seeds, 0.48 times the energy distance of $\gamma = 0$, FID 13.13 and top-5 accuracy 0.321.

6 Discussion and conclusion

When only some treated cells respond, even a perfect generative model trained on all of them shrinks every predicted change by the responder fraction. Weighting each cell's loss by its probability of response targets the responders, which recovered most of a known effect that an unweighted model missed, also with non-responders unlike controls, and improved the fit on BBBC021 and RxRx1.

Reproducibility statement

Section 3 defines every step of scoring and training, and Table 1 summarizes how each reported quantity is computed. Appendix B contains the proofs, and Appendices C to E list the settings and give the numbers behind every figure. We release our code: `https://anonymous.4open.science/r/responderFlowMatching-C5CB`

AI-Disclosure

We used generative AI tools (DeepL and Claude) only for translation and for polishing the writing. We reviewed all AI-assisted text and take full responsibility for the content of this paper.

References

Vidit Agrawal, John Peters, Tyler N. Thompson, Mohammad Vali Sanian, Chau Pham, Nikita Moshkov, Arshad Kazi, Aditya Pillai, Jack Freeman, Byunguk Kang, et al. CHAMMI-75: Pre-training multi-channel models with heterogeneous microscopy images. In *International Conference on Learning Representations*, 2026.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26, 2025.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11, 2010.

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 *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 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), 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, 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, 2024.

Daniel B. Burkhardt, Jay S. Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A. Gigante, Kevan C. Herold, Guy Wolf, Antonio J. Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39, 2021.

Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9), 2017.

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), 2010.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

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*, 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, 2024. doi: 10.1038/s41592-024-02241-6.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? A fine-grained analysis in a simple setting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

Aleksandrina Goeva, Michael-John Dolan, Judy Luu, Eric Garcia, Rebecca Boiarsky, Rajat M. Gupta, and Evan Macosko. HiDDEN: a machine learning method for detection of disease-relevant populations in case-control single-cell transcriptomics data. *Nature Communications*, 15:9468, 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. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Charles Jones, Emmanuel Noutahi, Jason Hartford, and Cian Eastwood. Elucidating the design space of flow matching for cellular microscopy. *arXiv preprint arXiv:2603.26790*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Leslie Kish. *Survey Sampling*. John Wiley and Sons, 1965.

D. Klein, J. S. Fleck, D. Bobrovskiy, L. Zimmermann, S. Becker, A. Palma, et al. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7), 2012.

Vebjorn Ljosa, Peter D. Caie, Rob ter Horst, Katherine L. Sokolnicki, Emma L. Jenkins, Sandeep Daya, Mark E. Roberts, Thouis R. Jones, Shantanu Singh, Auguste Genovesio, Paul A. Clemons, Neil O. Carragher, and Anne E. Carpenter. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *Journal of Biomolecular Screening*, 18(10):1321–1329, 2013. doi: 10.1177/1087057113503553.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP: Learning perturbation effects in cell painting via text-guided contrastive learning. In *Advances in Neural Information Processing Systems*, 2025.

Zeinab Navidi, Jun Ma, Esteban A. 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.

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

Efthymia Papalexi, Eleni P. Mimitou, Andrew W. Butler, Samantha Foster, Bernadette Bracken, William M. Mauck, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z. Yeung, Peter Smibert, and Rahul Satija. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Andrea Rubbi, Amir Akbarnejad, Mohammad Vali Sanian, Aryan Yazdan Parast, Hesam Asadollahzadeh, Arian Amani, Naveed Akhtar, Sarah Cooper, Andrew Bassett, Pietro Liò, Lassi Paavolainen, Sattar Vakili, and Mo Lotfollahi. Shortest-path flow matching with mixture-conditioned bases for OOD generalization to unseen conditions. *arXiv preprint arXiv:2601.11827*, 2026.

Michael D. Slack, Elisabeth D. Martinez, Lani F. Wu, and Steven J. Altschuler. Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences*, 105(49), 2008.

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*, 22:433, 2021.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. RxRx1: A dataset for evaluating experimental batch correction methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. doi: 10.1109/CVPRW59228.2023.00451.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8), 2013.

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. *arXiv preprint arXiv:2603.21743*, 2026a.

Dongxia Wu, Yuhui Zhang, Serena Yeung-Levy, Emma Lundberg, and Emily B. Fox. Uncertainty-aware distribution-to-distribution flow matching for scientific imaging. *arXiv preprint arXiv:2603.21717*, 2026b.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Shiyuan Zhang, Weitong Zhang, and Quanquan Gu. Energy-weighted flow matching for offline reinforcement learning. In *International Conference on Learning Representations*, 2025a.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating cellular morphology changes via flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025b.

Yuhui Zhang, Yuchang Su, Zoe Wefers, Shiye Su, He Li, Tianhong Li, Chenyu Wang, James Burgess, Alejandro Lozano, Linqi Zhou, Daisy Ding, Jeffrey Nirschl, Emma Lundberg, and Serena Yeung-Levy. CellFluxV2: An image generative foundation model for virtual cell modeling. *bioRxiv*, 2026. doi: 10.64898/2026.01.19.696785.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

7 Appendix

A Notation and computed quantities

Table 1 lists every quantity that appears in the paper, how it is computed and where it is used.

Table 1: Computed quantities. Scores are computed within one perturbation unit u, and control statistics come from the controls of the same plates.

Quantity	How it is computed	Used in
T_b	A perturbation is a compound at one dose (BBBC021), an siRNA (RxRx1), or a compound, CRISPR guide or ORF (JUMP). T_b are its treated crops on plate b. In BBBC021 the flow is conditioned on the compound, so all doses of a compound share one condition.	Everywhere
$h(x)$	Out-of-fold probability of a class-balanced ℓ_2 logistic regression ($C = 1$) on 50 principal components of MorphEm embeddings, scaled robustly to the training controls of the fold. A fold holds one well when a class has at most three wells, otherwise there are $\min(5, \text{wells})$ folds.	Section 3.2
$\hat{\varrho}_u$	Median over $u \in [0.75, 0.95]$ of $(T(u) - C(u))/(1 - C(u))$, where T and C are the fractions of treated and control crops above the u-quantile of the control scores, clipped to $[0, 1]$. $2\ \mathrm{AUROC} - 1$ is reported alongside.	Eq. (2), Figure 3
$s(x)$	$\mathrm{clip}(1 - (1 - \hat{\varrho}_u)(1 - h)/h,\ s_{\min},\ 1)$, with $\hat{\varrho}_u$ capped at 0.95.	Eq. (3)
$w_\gamma(x)$	$s(x)^\gamma$ divided by its mean over $T_{u,b}$.	Eq. (5)
ESS	$(\sum w)^2 / \sum w^2$ over T.	Section 5.2
Null band, floor	Band: AUROC under five well-level permutations of the labels per perturbation, pooled within strata of the real AUROC (below 0.7, 0.7 to 0.9, at least 0.9). A perturbation is detectable when its AUROC lies above its band. Floor: permutations of control wells against control wells, pooled per dataset.	Section 5.1, floored gate
q	Fraction of real treated crops in a dilution set, the rest being same-plate controls placed in treated wells.	Sections 4, 5.2
$r(S)$, $r_{\mathrm{proj}}(S)$	Median over CellProfiler features of $\Delta_f(S)/\Delta_f(H)$, and the projection $\langle \Delta(S), \Delta(H) \rangle / \|\Delta(H)\|^2$, where Δ is the mean shift standardized by same-plate controls and H the held-out real treated crops. Medians over perturbations.	Eq. (7), Figures 1 and 4
α, $\alpha^\star$	Guidance strength in $v(\cdot, \varnothing) + \alpha\,(v(\cdot, c) - v(\cdot, \varnothing))$. $\alpha^\star$ is the smallest α at which r_{proj}, interpolated over the grid, reaches one.	Section 5.4
Energy distance	$2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|$ between generated and real treated latents of a perturbation, with the $i = j$ terms left out, averaged over perturbations.	Table 10
FID	Fréchet distance between Inception features of generated and real crops, over all evaluation crops, with a well-bootstrap SE.	Figure 5
MoA, top-k, mAP	MoA accuracy from CellFlux's own IID and OOD classifiers on BBBC021, top-5 accuracy of identifying the perturbation on RxRx1 and JUMP, and top-1 accuracy and mAP on JUMP.	Tables 10 and 17
Mixture mode	Each evaluation crop is replaced by a generated crop with probability $\hat{\varrho}_u$ and by a same-plate control otherwise.	Section 5.5

B PROOFS

This appendix states the results of Section 3.4 precisely and proves them step by step. Throughout, we fix one plate and one perturbation c, drop c from the notation where it plays no role, and write p_0, p_1 and p_R for the densities of control, treated and responding cells (Eq. (1)). The control cell $x_0 \sim p_0$ and the treated cell $x_1 \sim p_1$ are independent, and $x_t = (1 - t)\, x_0 + t\, x_1$.

Proposition 1. With independent coupling and the straight path, the flow-matching loss is minimized by $v^\star(x, t) = \mathbb{E}[x_1 - x_0 \mid x_t = x]$, and the solution of $\dot{x} = v^\star(x, t)$ started from $x_0 \sim p_0$ has at every time t the form of $(1 - t)x_0 + tx_1$ (Lipman et al., 2023). Liu et al. (2023) prove the same marginal-preserving property for rectified flow. At $t = 1$ the generated law is therefore p_1. Taking expectations of ϕ under Eq. (1) gives $\mathbb{E}_{p_1}[\phi] = \varrho\, \mathbb{E}_{p_R}[\phi] + (1 - \varrho)\, \mathbb{E}_{p_0}[\phi]$, and subtracting $\mathbb{E}_{p_0}[\phi]$ gives Eq. (6). Nothing in the argument depends on the form of ϕ, so it applies to $\phi \circ \mathrm{Dec}$. For the noisy interpolant of Eq. (4), the field $\mathbb{E}[u_t \mid z_t = z]$ transports the same endpoint laws, because the noise vanishes at $t = 0$ and $t = 1$ (**?**).

The velocity field is not rescaled. Conditioning on R inside the conditional expectation gives $v^\star = w\, v_R + (1-w)\, v_0$, with $v_r(x,t) = \mathbb{E}[x_1 - x_0 \mid x_t = x, R = r]$ and $w(x,t) = \Pr(R = 1 \mid x_t = x)$. At $t = 0$ the state equals x_0, which is independent of R, so $w = \varrho$ everywhere. For $t > 0$ Bayes' rule gives

$$w(x,t) = \frac{\varrho\, p_t^{(1)}(x)}{\varrho\, p_t^{(1)}(x) + (1-\varrho)\, p_t^{(0)}(x)},$$

where $p_t^{(r)}$ is the law of x_t given $R = r$. Both laws are convolutions of a rescaled p_0 with a rescaled target law. If the characteristic function of p_0 has no zeros, as holds when the source is convolved with Gaussian noise, $w(\cdot, t)$ is constant only when $p_R = p_0$. Writing the optimal field as ϱ times a responder field plus a null field is therefore wrong at every interior time for any perturbation that has an effect, which is why we make only distributional claims.

Effect sizes under mixing. Let a feature have mean μ_0 and variance σ^2 under p_0, and mean $\mu_0 + \delta$ and the same variance under p_R, and let $d_1 = \delta/\sigma$. Under the mixture its mean is $\mu_0 + \varrho\delta$ and its variance is $\sigma^2 + \varrho(1-\varrho)\delta^2$. The effect standardized by the control spread is ϱd_1, which is exactly linear. Cohen's d with the pooled standard deviation is

$$d(\varrho) = \frac{\varrho\, d_1}{\sqrt{1 + \frac{1}{2}\varrho(1-\varrho)\, d_1^2}},$$

which falls faster than linearly. For $d_1 = 2$ and $\varrho = 0.5$ it gives 0.82 instead of 1.00. We standardize by controls so that the linearity of Proposition 1 carries over to the metric.

Proposition 2 Let μ_0 be the mean of p_0. At $t = 0$ the state equals x_0, which is independent of x_1, so $v^\star(x, 0, c) = \mathbb{E}[x_1 \mid c] - x = \mu_0 + \varrho_c\delta_c - x$ by Eq. (1). The field trained with the null token sees targets from the pooled distribution and equals $\mathbb{E}[x_1] - x = \mu_0 + \bar{m} - x$. Substituting both into $v_\alpha^\star = v^\star(\cdot, \varnothing) + \alpha\,(v^\star(\cdot, c) - v^\star(\cdot, \varnothing))$ gives $v_\alpha^\star(x, 0, c) = (\mu_0 - x) + \bar{m} + \alpha\,(\varrho_c\delta_c - \bar{m})$. Averaging over $x \sim p_0$ removes $\mu_0 - x$ and leaves the average move $\bar{m} + \alpha\,(\varrho_c\delta_c - \bar{m})$ of Proposition 2. With $\bar{m} = 0$, this move equals δ_c only when $\alpha = 1/\varrho_c$. In the dilution benchmark $\varrho = q\varrho_1$, so $\alpha^\star(q) = 1/(q\varrho_1) = \alpha^\star(1)/q$. The argument concerns the initial velocity only. For $t > 0$ the responder weight $w(x,t)$ of the optimal field varies with x, so the needed correction $1/w(x,t)$ is not a constant, and guidance with $\alpha > 1$ also changes higher moments (Chidambaram et al., 2024). The prediction therefore holds to first order.

Proposition 3. Write the weighted loss as an integral over x_0, x_1 and t with density $p_0(x_0)\, p_1(x_1)$. Because the weight depends on x_1 alone, $s(x_1)^\gamma p_1(x_1) = Z_\gamma\, p_1^{(\gamma)}(x_1)$, and

$$\mathcal{L}_\gamma(\theta) = Z_\gamma\, \mathbb{E}_{x_0 \sim p_0,\, x_1 \sim p_1^{(\gamma)},\, t} \big\| v_\theta(x_t, t, c) - (x_1 - x_0) \big\|^2.$$

The factor $Z_\gamma > 0$ does not depend on θ, so the minimizers are those of unweighted flow matching toward $p_1^{(\gamma)}$, and the argument for Proposition 1 shows that the minimizer transports p_0 to $p_1^{(\gamma)}$. For the calibrated posterior, $s\, p_1 = \varrho\, p_R$, which integrates to ϱ, so $Z_1 = \varrho$ and $p_1^{(1)} = p_R$. Independent noise in the path, as in Eq. (4), does not change the argument, and the normalization of Eq. (5) divides each perturbation and plate by an estimate of its own Z_γ.

C Scorer details and responder statistics

MorphEm (Agrawal et al., 2026) is the ViT-S/16 released by its authors, used without fine-tuning. Before scoring, a crop is removed when its fraction of saturated pixels, or when the share of a channel's intensity that an illumination ramp explains (R^2), exceeds a per-channel threshold. The thresholds remove about 0.1% of the control crops per criterion, and on BBBC021 the saturation threshold is 0.1% of the pixels. The control floor comes from permuting control wells against control wells, pooled per dataset. It has a mean AUROC of 0.50, 0.53 and 0.47 on BBBC021, JUMP and RxRx1, with maxima of 0.54, 0.59 and 0.89. RxRx1 estimates it from only 12 control wells, so its maximum is the relevant ceiling for a null draw. A perturbation's estimate of ϱ is flagged as unstable when the terms of Eq. (2) spread by more than 0.10. Table 2 summarizes the responder statistics under both estimators of ϱ, Table 3 gives the class means of Figure 3d, and Table 4 tests the second half of the responder model.

Folds and features. These are the details we omitted in Section 3.2. The wells of a perturbation, treated and control, are split into folds. A class with at most three wells gets one well per fold, and otherwise we use $\min(5, \text{number of wells})$ folds, stratified by class. Within each fold, the embeddings are scaled robustly (median and median absolute deviation) to the control crops of the training wells and projected onto 50 principal components fitted on the training rows. The classifier is an ℓ_2-regularized logistic regression with $C = 1$ and class weights $N/(2N_1)$ and $N/(2N_0)$, where N_1 and N_0 count treated and control crops and $N = N_0 + N_1$.

Null band. For each perturbation, we permute the treated and control labels between whole wells five times and recompute the AUROC. Five permutations give a noisy band for a single perturbation, so we pool the permutation AUROCs of all perturbations whose real AUROC falls into the same stratum, below 0.7, 0.7 to 0.9, or at least 0.9, and use 2.5th to 97.5th percentiles of the pooled distribution as the band of each of them. A perturbation is detectable when its AUROC lies above its band.

Table 2: Responder statistics per dataset. Counts of low responder fractions refer to detectable perturbations, first with the plateau estimate of Eq. (2), which every number in the paper uses, then with $2\ \mathrm{AUROC} - 1$.

Dataset	Pert.	Detectable	Inside band	Below band	$\hat{\varrho} < 0.5$	$2\ \mathrm{AUROC} - 1 < 0.5$
BBBC021	84	76	8	0	3 (4%)	0
RxRx1	100	86	14	0	21 (24%)	18
JUMP	100	57	42	1	34 (60%)	29
CRISPR knockouts	44	23	21	0	20 (87%)	18
ORF overexpression	16	7	9	0	5 (71%)	5
Compounds	40	27	12	1	9 (33%)	6

Table 3: Mean $\hat{\varrho}$ per mechanism of action in BBBC021, averaged over compounds and doses (Figure 3d).

MoA	Mean $\hat{\varrho}$	MoA	Mean $\hat{\varrho}$
Kinase inhibitors	0.38	Microtubule destabilizers	0.95
Epithelial	0.50	Aurora kinase inhibitors	0.97
DNA damage	0.60	Actin disruptors	0.98
DNA replication	0.65	Protein synthesis	0.98
Protein degradation	0.79	Eg5 inhibitors	0.98
Cholesterol-lowering	0.83	Microtubule stabilizers	0.99

Table 4: Test of the second half of the responder model. For each level τ, treated and control crops that score below the same threshold are compared, and the table gives their mean AUROC over perturbations. If non-responders look like controls the AUROC is near 0.5. The last two columns give the range of the same statistic in simulations with the design of each dataset, with responder fractions of 0.3, 0.6 and 0.9 when the assumption holds and 0.5 when non-responders differ from controls.

	Real data, τ					Simulation	
Dataset	0.1	0.25	0.5	0.75	0.9	Assumption holds	Assumption fails
BBBC021	0.634	0.636	0.715	0.749	0.795	0.48 to 0.53	0.55 to 0.59
JUMP	0.481	0.516	0.531	0.560	0.580	0.48 to 0.59	0.48 to 0.62
RxRx1	0.700	0.745	0.749	0.768	0.751	0.45 to 0.54	0.55 to 0.58

D DILUTION BENCHMARK

The 31 perturbations come from 11 compounds, namely ALLN, AZ138, AZ258, anisomycin, cytochalasin B, demecolcine, emetine, mevinolin, nocodazole, taxol and vincristine. Every training set holds 7,592 crops, the held-out reference 2,261 crops, and the CellProfiler evaluation samples 150 crops per perturbation from each diluted set. The autoencoder reconstructs BBBC021 crops with a mean absolute error of 0.04 and an LPIPS (Zhang et al., 2018) of 0.13, and JUMP crops with 0.027 and 0.09. Every CellProfiler evaluation run passed four gates before use, namely reconstruction near one, passthrough near zero, and two splits of one population, controls against controls near zero and real crops against real crops near one. Measuring segmented cells instead of whole crops gives the same picture at $q = 0.1$, 0.91 for $\gamma = 1$ against 0.21 for $\gamma = 0$. Because the flow conditions on the compound, a compound's generated crops can also be compared with its real reference pooled over its doses. This moves r at $q = 0.1$ from 0.858 to 0.803 for $\gamma = 1$ and from -0.001 to 0.032 for $\gamma = 0$, and every table here uses the per-perturbation reference. Table 6 gives the composition of the weights and the latent energy distance to the held-out crops, Table 7 the attribution with paired tests, Table 8 the guidance sweep and Table 9 the variant with injected controls from other plates. Held-out reference crops lie within one crop width of a real training crop at $q = 0.1$ in 7% of cases.

Table 5: Dilution benchmark on 31 BBBC021 perturbations, first evaluation run (Figures 1b and 3c). Diluted real is the training set itself. Model columns give median r with the IQR in brackets and median r_{proj}. $\mathbb{E}[s]$ is the mean score of the diluted set with the range over perturbations. No models were trained at $q = 1$.

	Diluted real		$\gamma = 0$		$\gamma = 1$		
q	r	r_{proj}	r [IQR]	r_{proj}	r [IQR]	r_{proj}	$\mathbb{E}[s]$ [range]
0.10	0.104	0.093	-0.001 [-0.175, 0.427]	0.042	0.858 [0.552, 1.166]	0.650	0.140 [0.097, 0.204]
0.25	0.247	0.233	0.085 [-0.124, 0.463]	0.082	0.934 [0.594, 1.227]	0.774	0.281 [0.244, 0.357]
0.50	0.483	0.465	0.320 [0.119, 0.689]	0.212	0.975 [0.681, 1.232]	0.772	0.514 [0.475, 0.552]
0.75	0.730	0.722	0.671 [0.476, 0.964]	0.508	1.024 [0.778, 1.235]	0.828	0.745 [0.687, 0.764]
0.90	0.879	0.863	0.847 [0.655, 1.099]	0.690	1.011 [0.770, 1.227]	0.822	0.882 [0.847, 0.897]
1.00	0.974	0.957					

Table 6: Composition of the $\gamma = 1$ weights on the dilution sets, and the unbiased latent energy distance of generated to held-out crops.

q	Injected share of crops	Injected share of weight	ESS$/N$	Energy, $\gamma = 0$	Energy, $\gamma = 1$
0.10	0.906	0.304	0.136	1.356	0.715
0.25	0.755	0.135	0.283	1.222	0.647
0.50	0.503	0.049	0.518	0.755	0.344
0.75	0.255	0.019	0.749	0.306	0.194
0.90	0.106	0.006	0.886	0.172	0.186

Table 7: Attribution at $q = 0.1$ (Figure 4a). Median r over 31 perturbations with the IQR. Runs A, B and C are separate evaluation runs with new generated samples and the same real crops, and each paired Wilcoxon test compares a variant with $\gamma = 1$ in the same run. For $\gamma = 0$ the range covers the three runs and, in run A, every q. The hard filter was evaluated with the samples of run A and trained for only 4,400 of the 31,900 steps of the other variants (Appendix F).

Run	Variant	Median r [IQR]	Perturbations won by $\gamma = 1$	p
A	Reconstruction	1.005 [0.991, 1.117]		
	Diluted real	0.104 [0.056, 0.153]		
	$\gamma = 1$	0.858 [0.552, 1.166]		
	$\gamma = 0$	-0.001 [-0.175, 0.427]	29 to 31 of 31	$< 10^{-5}$
	Passthrough	0.018 [-0.152, 0.513]		
	Hard filter, $s > 0.5$	0.501 [0.034, 1.059]		
B	Oracle	0.885 [0.490, 1.256]	7 of 31	0.001
	$\gamma = 1$	0.825 [0.551, 1.166]		
	Tempered oracle	0.672 [0.370, 1.001]	29 of 31	$< 10^{-5}$
	Shuffled weights	0.058 [-0.153, 0.464]	30 of 31	$< 10^{-5}$
	$\gamma = 0$	0.013 [-0.155, 0.461]	29 to 31 of 31	$< 10^{-5}$
C	$\gamma = 1$	0.848 [0.555, 1.179]		
	$\gamma = 1$, per-pert. normalization	0.853 [0.539, 1.190]		0.43
	Well mean, own wells	0.467 [0.214, 0.851]		
	Well mean, treated wells	0.020 [-0.163, 0.396]		
	Shuffled weights	0.085 [-0.161, 0.458]	30 of 31	$< 10^{-5}$
	$\gamma = 0$	0.014 [-0.164, 0.402]	29 to 31 of 31	$< 10^{-5}$

Table 8: Guidance sweep for $\gamma = 0$ models (Figures 4b and c). Entries are r_{proj}. $\alpha^\star$ is derived by linear interpolation to a recovered effect of one, and the next column is the prediction $2.55/q$. The last row gives the spread of r_{proj} across q at each α, against 0.18 for $\gamma = 1$ at $\alpha = 1.2$. The undiluted model was not sampled at $\alpha = 1.2$.

q	$\alpha = 1$	1.2	2	3	4	6	8	$\alpha^\star$	$2.55/q$
0.10	0.043	0.042	0.034	0.061	0.076	0.124	0.153	not reached	25.5
0.25	0.078	0.082	0.106	0.133	0.243	0.510	0.768	not reached	10.2
0.50	0.200	0.212	0.374	0.506	0.727	1.190	1.533	5.18	5.10
0.75	0.453	0.508	0.746	1.092	1.276	1.675	2.514	2.74	3.40
0.90	0.613	0.690	0.972	1.184	1.337	1.676	2.099	2.13	2.83
1.00	0.540		0.802	1.163	1.672	2.286	2.577	2.55	2.55
Spread	0.57	0.65	0.94	1.12	1.60	2.16	2.43		

Table 9: Dilution with injected controls from other plates, which violates the responder model. Each injected crop is replaced by a control crop from a donor well on another plate, chosen by the AUROC between the controls of the two plates, with targets of 0.65 (near) and 0.78 (far). The first columns describe the sets at $q = 0.1$: the AUROC of the injected crops against the controls of the training plate, and the threshold check of Table 4 at $\tau = 0.1$ and 0.25, which gives 0.63 and 0.64 on real BBBC021. The other columns give the median recovered effect r over 31 perturbations. The same-plate row is the original benchmark, evaluated in run A.

		Threshold check		r at $q = 0.1$			r at $q = 0.5$	
Injected controls	AUROC	$\tau = 0.1$	$\tau = 0.25$	$\gamma = 0$	$\gamma = 1$	Oracle	$\gamma = 0$	$\gamma = 1$
Same plate	0.43	0.43	0.44	-0.001	0.858		0.320	0.975
Other plate, near	0.65	0.45	0.50	0.155	0.855	0.889	0.494	0.951
Other plate, far	0.74	0.62	0.73	0.287	0.902	0.894	0.606	0.973

E REAL SCREENS

Table 10: Real screens at the default guidance strength $\alpha = 1.2$. Energy is the energy distance between generated and real treated cells in latent space, averaged over perturbations. Values with $\pm$ give the mean and standard error over three seeds, the others come from one seed. BBBC021 values are for the IID split (OOD in Table 12).

Dataset	Variant	FID ↓	Energy ↓	MoA acc. ↑	Top-5 ↑
BBBC021	$\gamma = 0$	9.76 ± 0.07	0.086 ± 0.002	0.618 ± 0.001	
	$\gamma = 1$	9.49 ± 0.16	0.061 ± 0.004	0.633 ± 0.007	
	$\gamma = 2$	9.39 ± 0.36	0.058 ± 0.004	0.639 ± 0.002	
	$\gamma = 1$ shuffled	9.99	0.074	0.628	
RxRx1	$\gamma = 0$	21.84 ± 0.36	0.260 ± 0.005		0.154 ± 0.002
	$\gamma = 1$	19.15 ± 0.07	0.133 ± 0.004		0.171 ± 0.002
	Shuffled	20.99 ± 0.15	0.170 ± 0.004		0.150 ± 0.002
JUMP	$\gamma = 0$	13.43 ± 0.08	0.0061 ± 0.0002		0.313 ± 0.001
	$\gamma = 0.25$	13.13 ± 0.06	0.0029 ± 0.0006		0.321 ± 0.002
	$\gamma = 1$	14.75 ± 0.08	0.0356 ± 0.0028		0.321 ± 0.002
	Shuffled	15.80 ± 0.03	0.0104 ± 0.0006		0.292 ± 0.003
	Floored gate	13.08 ± 0.13	0.0127 ± 0.0009		0.334 ± 0.004

The autoencoder is a frozen KL autoencoder with EMA weights, channel width 128 and multipliers (1, 2, 4, 4), taken at epoch 227 for BBBC021 and after 500k steps for JUMP and RxRx1, and its latents are divided by their standard deviation (0.951, 0.877 and 0.967). Training uses AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4}, $\beta = (0.9, 0.95)$, batch size 256 and an EMA of 0.9999. On BBBC021 an epoch has 246 steps, so the two stages end at steps 24,600 and 270,600, and 100 epochs of the second stage, the length of the released CellFlux run, end at step 49,200. Half of the source latents receive Gaussian noise during training, and every source latent does at sampling. The evaluation sets hold 7,027 IID and 1,911 OOD crops on BBBC021, 9,000 on JUMP and 2,808 on RxRx1. All numbers use the last checkpoint of each run, and none is selected on test metrics. Four evaluation runs of one $\gamma = 0$ model spread the energy estimate of the evaluation code, which keeps the $i = j$ terms, by 0.009, about four times its spread over training seeds. With the unbiased estimator, we resampled all three-seed BBBC021 models three more times (Table 13). The evaluation run moves $\gamma = 0$ by an SD of 0.0046, more than its SD of 0.0027 over training seeds, but it moves all models together, so the gain of $\gamma = 1$ varies by an SD of only 0.0017. Averaged over the four runs, the gains of $\gamma = 1$ and $\gamma = 2$ are 0.024 and 0.025 (5.8 and 6.4 SE), and shuffled weights keep 40% and 39% of them. On RxRx1, in mixture mode with the first seed at $\alpha = 1$, $\gamma = 1$ has 1.02 times the energy distance of $\gamma = 0$, and shuffled weights 1.25 times.

Table 11: BBBC021 guidance sweep with one seed (Figures 5a and b). The released CellFlux checkpoint, evaluated with the same pipeline at its own strength, reaches FID 18.18 and MoA accuracy 0.603 on IID and FID 27.39 and MoA accuracy 0.541 on OOD.

	IID				OOD			
	$\gamma = 0$		$\gamma = 2$		$\gamma = 0$		$\gamma = 2$	
α	FID	MoA	FID	MoA	FID	MoA	FID	MoA
1	9.904	0.576	9.003	0.593	18.966	0.445	18.366	0.474
1.2	9.729	0.618	8.708	0.642	18.704	0.466	17.932	0.491
2	10.191	0.718	8.762	0.743	20.121	0.508	18.451	0.522
3	12.438	0.761	10.563	0.793	23.276	0.515	20.676	0.542
4	15.617	0.772	12.992	0.794	26.732	0.502	23.700	0.526
6	23.401	0.711	18.855	0.741	34.139	0.435	29.741	0.499

Table 12: BBBC021 at $\alpha = 1.2$. Rows with $\pm$ give mean and SE over three training seeds, shuffled variants one seed. On IID the energy gains of $\gamma = 1$ and $\gamma = 2$ are 5.7 and 6.9 SE (5.8 and 6.4 SE over four evaluation runs, Table 13) and the MoA gains 2.3 and 7.4 SE. On OOD only the MoA gains exceed two SE (2.1 and 2.5). FID differences stay within seed noise on both splits.

Split	Variant	FID ↓	Latent energy ↓	MoA acc. ↑
IID	$\gamma = 0$	9.76 ± 0.07	0.0861 ± 0.0017	0.618 ± 0.001
	$\gamma = 1$	9.49 ± 0.16	0.0613 ± 0.0040	0.633 ± 0.007
	$\gamma = 2$	9.39 ± 0.36	0.0582 ± 0.0037	0.639 ± 0.002
	$\gamma = 1$ shuffled	9.99	0.0741	0.628
	$\gamma = 2$ shuffled	9.43	0.0749	0.623
	Released CellFlux	18.18 ± 0.57		0.603
OOD	$\gamma = 0$	18.39 ± 0.25	0.3317 ± 0.0111	0.472 ± 0.003
	$\gamma = 1$	18.48 ± 0.43	0.3043 ± 0.0126	0.493 ± 0.010
	$\gamma = 2$	18.85 ± 0.51	0.3273 ± 0.0020	0.490 ± 0.007
	$\gamma = 1$ shuffled	19.17	0.3470	0.451
	$\gamma = 2$ shuffled	19.57	0.3890	0.480
	Released CellFlux	27.39 ± 1.27		0.541

Table 13: BBBC021 IID, unbiased latent energy distance in four evaluation runs with new samples. The first row gives $\gamma = 0$ averaged over seeds, the others the gain over $\gamma = 0$, where positive is better. The SE is over training seeds, averaged over runs. Shuffled variants have one seed, and the last column gives their share of the gain of the unshuffled variant.

	Evaluation run					
	1	2	3	4	Mean	SE or share
Energy of $\gamma = 0$	0.0861	0.0841	0.0787	0.0762	0.0813	
Gain of $\gamma = 1$	0.0248	0.0256	0.0225	0.0220	0.0237	0.0041
Gain of $\gamma = 2$	0.0279	0.0260	0.0235	0.0235	0.0252	0.0039
Gain of $\gamma = 1$, shuffled	0.0120	0.0112	0.0067	0.0084	0.0096	40%
Gain of $\gamma = 2$, shuffled	0.0112	0.0100	0.0095	0.0089	0.0099	39%

F FLOORED GATE, NUISANCE WEIGHTS, SAMPLER AND HARD FILTER

The floored gate gives every crop of a perturbation a flat weight, the perturbation's mean score and so weight one after normalization, when the perturbation's AUROC does not clear the control floor of Appendix C. Nuisance weights repeat the scorer with crop covariates in place of the embedding, namely the number of cells in the field and the foreground fraction, plus the distance to the image edge on RxRx1, in the scorer's own folds and with ϱ estimated as $2\ \mathrm{AUROC} - 1$. The weighted sampler draws crops with probability proportional to w, with replacement, and trains with an unweighted loss. Hard filtering drops the crops with $s \leq 0.5$ and trains $\gamma = 0$ on the rest. At $q = 0.1$ it keeps 1,083 crops, 97% of the real ones (a median of 22 per perturbation) and 387 injected controls. Its samples do not reach the real data (FID 132, precision 0), so its recovered effect of 0.501 is not a like-for-like comparison, which needs matched steps. On the undiluted BBBC021 set the filter keeps 84.5% of the crops and trains for 222,200 instead of 270,600 steps. With one seed it reaches an IID FID of 8.95 and MoA accuracy of 0.640, within the range of the $\gamma = 1$ and $\gamma = 2$ seeds and outside that of $\gamma = 0$, and an OOD FID of 17.87 and MoA accuracy of 0.491.

Table 14: IID crops that overlap no training crop. CellFlux's IID split separates crops within wells, so 6,194 of the 7,027 IID evaluation crops lie within one crop width of a training crop of the same perturbation, none of them a copy. On the 834 crops that share no pixels with any training crop the ranking holds, with $\gamma = 1$ and $\gamma = 2$ ahead of $\gamma = 0$ by 4.1 and 5.4 SE in energy and $\gamma = 2$ by 2.7 SE in FID. The higher absolute values reflect both sample size, since 834 overlapping crops give an FID of about 26, and composition, since these crops come from sparser fields and are enriched for cytotoxic compounds.

Metric	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	Released CellFlux
FID	47.0	46.1	44.7	60.8
Latent energy	0.84	0.76	0.72	

Table 15: RxRx1 at $\alpha = 1.2$ per seed (Figure 5d).

Variant	Metric	Seed 1	Seed 2	Seed 3	Mean $\pm$ SE
$\gamma = 0$	FID	22.38	21.97	21.16	21.84 ± 0.36
	Latent energy	0.267	0.250	0.262	0.260 ± 0.005
	Top-5	0.154	0.157	0.150	0.154 ± 0.002
$\gamma = 1$	FID	19.09	19.29	19.08	19.15 ± 0.07
	Latent energy	0.126	0.134	0.138	0.133 ± 0.004
	Top-5	0.170	0.175	0.167	0.171 ± 0.002
Shuffled	FID	20.85	20.82	21.29	20.99 ± 0.15
	Latent energy	0.174	0.161	0.174	0.170 ± 0.004
	Top-5	0.148	0.154	0.149	0.150 ± 0.002

Table 16: RxRx1 variants with one seed at $\alpha = 1.2$. The ratio compares the latent energy with the three-seed mean of $\gamma = 0$.

Variant	FID	Latent energy	Ratio to $\gamma = 0$	Top-5
$\gamma = 0.25$	18.66	0.186	0.72	0.162
$\gamma = 0.5$	19.60	0.156	0.60	0.163
$\gamma = 2$	19.55	0.139	0.54	0.161
Floored gate	18.51	0.129	0.50	0.168
Nuisance weights	18.29	0.182	0.70	0.145
Weighted sampler	21.26	0.140	0.54	0.162
Shuffled sampler	22.73	0.192	0.74	0.142

Table 17: JUMP at $\alpha = 1.2$, mean $\pm$ SE over three seeds (Figure 5e). Latent energy is the unbiased energy distance to all real treated crops, and the ratio divides its mean by that of $\gamma = 0$. Top-1 accuracy has an SE of at most 0.001.

Variant	FID	Latent energy	Ratio	Top-1	Top-5	mAP
$\gamma = 0$	13.43 ± 0.08	0.0061 ± 0.0002	1.00	0.093	0.313 ± 0.001	0.148 ± 0.002
$\gamma = 0.25$	13.13 ± 0.06	0.0029 ± 0.0006	0.48	0.096	0.321 ± 0.002	0.159 ± 0.002
$\gamma = 1$	14.75 ± 0.08	0.0356 ± 0.0028	5.88	0.094	0.321 ± 0.002	0.252 ± 0.002
Shuffled	15.80 ± 0.03	0.0104 ± 0.0006	1.71	0.082	0.292 ± 0.003	0.190 ± 0.003
Floored gate	13.08 ± 0.13	0.0127 ± 0.0009	2.09	0.102	0.334 ± 0.004	0.197 ± 0.002

Table 18: JUMP at $\alpha = 1.2$, first seed, all variants. Latent energy is also given relative to $\gamma = 0$ of the same seed, against the full real reference, against the treated crops the scorer ranks in the top half of their perturbation, and for the mixture mode, in which each evaluation crop is replaced by a generated crop with probability $\hat{\varrho}_u$ and by a same-plate control otherwise (mean of five draws).

		Latent energy				
Variant	FID	Value	Ratio	Ratio, top-half reference	Ratio, mixture mode	Top-5
$\gamma = 0$	13.36	0.0057	1.00	1.00		0.311
$\gamma = 0.25$	13.17	0.0039	0.68	0.45		0.317
$\gamma = 0.5$	14.71	0.0098	1.72	1.21		0.315
$\gamma = 1$	14.87	0.0396	6.90	5.48	2.08	0.324
$\gamma = 2$	15.41	0.0616	10.74	8.77	2.01	0.332
Shuffled	15.79	0.0116	2.02	1.66	2.58	0.286
Floored gate	12.86	0.0131	2.29	1.77		0.327
Nuisance weights	14.29	0.0510	8.90	7.22		0.298
Weighted sampler	19.03	0.0541	9.45	7.71		0.317

Table 19: FID between two disjoint samples of n real crops of the same perturbation (Figure 5f). No RxRx1 perturbation has enough crops for two samples of 128. The lower block gives per-perturbation FID at the sample sizes of our evaluation sets, for two real samples, for our $\gamma = 0$ model ($\alpha = 1.0$ on RxRx1) and for the released CellFlux checkpoint, and the values reported for CellFluxV2 (Zhang et al., 2026).

n per side	BBBC021	JUMP	RxRx1
8	162.3	172.8	190.3
16	126.1	135.8	147.5
32	96.8	105.8	112.6
64	75.6	80.0	84.8
128	54.5	58.7	
256	38.6	39.5	
Real against real, evaluation n	42.4	73.7	118.0
$\gamma = 0$	53.5	84.7	138.8
Released CellFlux	59.8	84.6	166.8
CellFluxV2, reported	45.6	82.4	135.4

G QUALITATIVE EXAMPLES

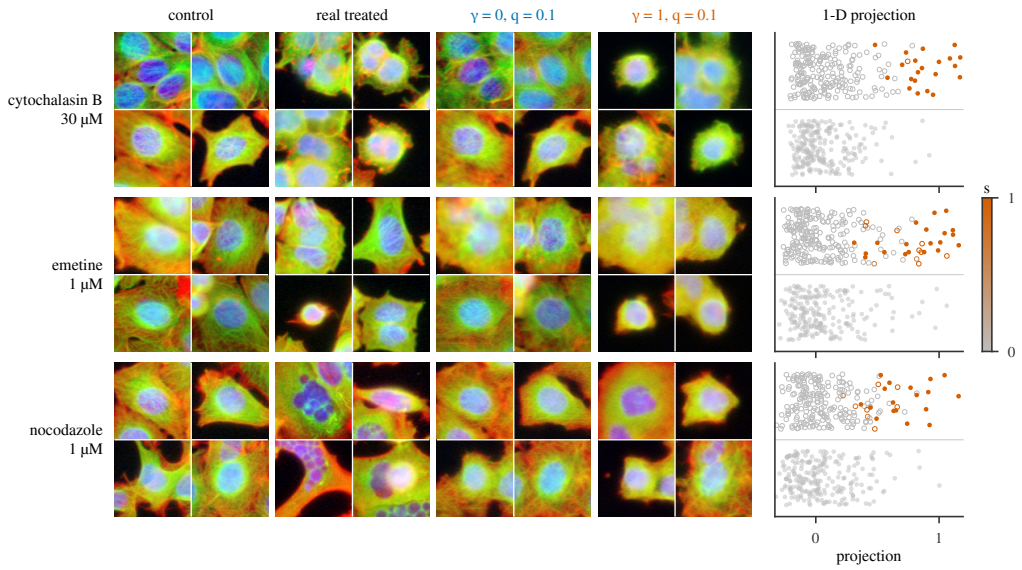

Figure 6: Qualitative examples from the dilution benchmark at $q = 0.1$ for cytochalasin B at 30 μM, emetine at 1 μM and nocodazole at 1 μM. Columns show a control crop, a real treated crop, generations of the $\gamma = 0$ and $\gamma = 1$ models, and a one-dimensional projection in which controls are grey and treated crops are colored by their score s, from grey to vermillion. Every crop is a real image or a model output.