OMNIMORPH: A CHANNEL-ADAPTIVE UNIFIED GENERATIVE MODEL FOR CELL PAINTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cell Painting uses multiplexed fluorescent dyes to label multiple subcellular structures of cells and is a standard assay for large-scale chemical and genetic screening. Due to the vast combinatorial space of chemical and genetic perturbations, using generative models to replace part of these experiments has become a promising avenue. However, the cell lines and perturbations covered by any single dataset are limited, so a model trained on that dataset alone struggles to generalize beyond them. Each dataset is also set up differently depending on its research purpose, such as whether it includes brightfield images or same-batch controls. Therefore, a unified generative model for Cell Painting that supports multiple datasets and multiple tasks is needed. Unlike natural images with a fixed RGB channel space, Cell Painting datasets vary in both the number of channels and the subcellular structures they label, which makes such a unified model difficult to train across datasets. To address this, we propose OmniMorph, to our knowledge the first unified generative model for Cell Painting. OmniMorph is a flow-matching model with a DiT backbone that encodes each fluorescence channel separately. We disentangle the semantics of a channel from its role with two embeddings. The channel embedding marks the subcellular structure the channel represents and is shared across datasets, and the role embedding marks the part it plays in the current task. The model is further conditioned on perturbation embeddings from pretrained encoders and on a natural-language instruction that specifies which task to perform. We train OmniMorph jointly on six datasets and on the four tasks of morphology generation, perturbation prediction, missing-channel completion and brightfield-to-fluorescence translation. It mostly outperforms task-specific methods on image-distribution and biological metrics. Ablations show that the channel embedding is critical and that joint training shares knowledge across datasets and yields gains across tasks, providing a scalable foundation for datasets of heterogeneous channel composition and further generative tasks. Code is available at `https://anonymous.4open.science/r/OmniMorph-2EBB/`.

1 INTRODUCTION

Cell Painting stains several subcellular structures with multiplexed fluorescent dyes and records the morphological response of cells to perturbations, and it is now widely used for large-scale chemical and genetic screening (Cimini et al., 2023) (Figure 1a). The combinatorial space of perturbations, cell lines and doses is vast, and with the growth of public datasets (Chandrasekaran et al., 2023), using generative models to replace part of these experiments has become a promising avenue. Generative models are currently applied to four main Cell Painting tasks (Figure 1c). The first two generate the morphology under a perturbation. Morphology generation synthesizes images from the perturbation and the cell line alone, anticipating the morphology a perturbation induces before the experiment (Yang et al., 2021; Navidi et al., 2025; Demirel et al., 2025), whereas perturbation prediction uses a control image as the reference and predicts the morphological change the perturbation induces (Palma et al., 2025; Zhang et al., 2025). The other two recover fluorescence that was not measured. Missing-channel completion predicts the staining of additional structures from a subset of channels (Zhang et al., 2026a; Sims et al., 2024; Nguyen et al., 2020), and brightfield-to-fluorescence translation infers the stained structures from unstained brightfield images (Ounkomol et al., 2018; Peysakhovich et al., 2025; Cross-Zamirski et al., 2023). Both make the readout of 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.

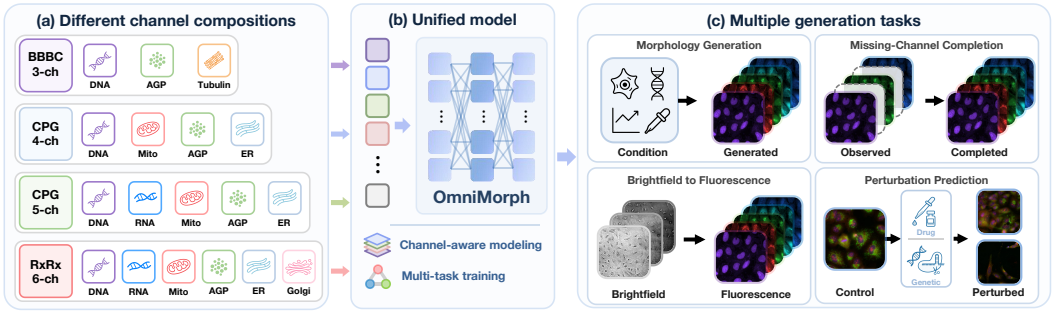

Figure 1: Overview of OmniMorph. (a) Public Cell Painting datasets differ in the subcellular structures they stain and in the number of channels. (c) Four generative tasks have been studied on these data, each with a different input. (b) OmniMorph tags each channel with the structure it shows and with its role in the current task, so that one model handles these datasets and tasks together.

structure available without staining it, which saves the cost of staining and lifts the limit on the number of fluorescence channels.

Around these tasks a large number of public datasets have accumulated, but each is designed for a specific research purpose and their channel composition differs (Figure 1a). For example, most follow the five-channel Cell Painting protocol (Cimini et al., 2023), RxRx3 (Fay et al., 2023) images actin and the Golgi apparatus separately in six channels, CPG0036 (Wolff et al., 2025) drops the RNA channel and has four, and BBBC021 (Caie et al., 2010) predates Cell Painting with only DNA, actin and tubulin in three. Datasets also differ in whether brightfield images are acquired and whether controls are included. Existing generative models are trained separately for a single dataset and a single task (Demirel et al., 2025; Navidi et al., 2025; Palma et al., 2025; Zhang et al., 2025), so each model is exposed only to the cell lines and perturbations of one dataset, cannot benefit from knowledge acquired on other tasks, and generalizes poorly beyond them. On natural images, unified generative models have shown that one model can perform several tasks and share knowledge across them (Xiao et al., 2025; Chen et al., 2025; Fu et al., 2025; Le et al., 2025; Han et al., 2025b). However, such a unified model is hard to build for Cell Painting because of the heterogeneity of its data. For example, the input and the target of every natural-image task are fixed RGB images, whereas the channels of Cell Painting images change with the dataset and the inputs change with the task. Therefore, a channel-adaptive unified model trained jointly across datasets and tasks is needed to overcome these limitations.

To this end, we propose OmniMorph, a channel-adaptive unified generative model for Cell Painting based on flow matching (Figure 1b). To adapt to any channel composition, OmniMorph processes each fluorescence channel separately. Every channel is replicated to three channels, compressed by a frozen SD-VAE and split into a set of tokens, and a DiT backbone models the tokens of all channels as one sequence. For one model to cover several datasets and tasks, we disentangle the semantics of a channel from its role in the task with two embeddings added to all tokens of the channel. A channel embedding marks the subcellular structure the channel represents, so that channels of the same structure share one representation across datasets. A role embedding marks whether the channel is a reference or a generation target in the current task, so that a change of task amounts to a reassignment of roles. Besides the image tokens, the sequence also carries the conditions as tokens. The identity of the perturbation is given as a representation from a pretrained encoder, and the dose and the cell line as structured vectors. The task is specified by a natural-language instruction that names the task, the target channels and a textual description of the cell line and the perturbation, encoded by a frozen language model (Lu et al., 2025). Heterogeneous datasets and multiple tasks can thus be trained within a single architecture. We train OmniMorph jointly on six public data sources that cover staining panels with three to six channels, five cell lines and two perturbation types, compounds and CRISPR, and compare it with task-specific methods on the four tasks. Our contributions are as follows.

- We propose OmniMorph, to our knowledge the first unified generative model for Cell Painting, trained jointly on datasets of different channel composition and on multiple tasks.

- We encode images channel by channel with a channel embedding and a role embedding and specify the task by a natural-language instruction, so that one architecture accommodates heterogeneous datasets and tasks.
- On six datasets and four tasks, OmniMorph mostly outperforms task-specific methods on image and biological metrics such as FID and mechanism-of-action (MoA) classification, and ablations show that the channel embedding is critical and that joint training shares knowledge across datasets and yields gains across tasks.

2 Related work

Cell Painting tasks. Public Cell Painting data come from academic benchmarks (Caie et al., 2010), the JUMP consortium (Chandrasekaran et al., 2024; Cimini et al., 2023; Way et al., 2022; Wolff et al., 2025) and corporate releases (Fay et al., 2023; Kraus et al., 2025). Existing generative work on Cell Painting addresses morphology generation (Yang et al., 2021; Navidi et al., 2025; Demirel et al., 2025; Wang et al., 2025; Hung et al., 2024; Jones et al., 2026), perturbation prediction (Palma et al., 2025; Zhang et al., 2025; 2026b; Lamiable et al., 2023; Bourou et al., 2024), channel completion (Nguyen et al., 2020; Sims et al., 2024; Zhang et al., 2026a; Bao et al., 2021; Wu et al., 2023; Sims et al., 2026; 2025) or brightfield-to-fluorescence translation (Ounkomol et al., 2018; Christiansen et al., 2018; Cross-Zamirski et al., 2022; 2023; Peysakhovich et al., 2025; Han et al., 2025a; Xing et al., 2024), usually with one model per source and per task, whereas OmniMorph puts the four tasks into one model trained jointly on six sources.

Unified generative models. On natural images, OmniGen (Xiao et al., 2025), UniVG (Fu et al., 2025), UniReal (Chen et al., 2025), OneDiffusion (Le et al., 2025) and ACE (Han et al., 2025b) unify multiple generative tasks through interleaved sequences, task tokens or per-image role embeddings, all targeting a fixed three-channel image space. On microscopy, FluoResFM and UniFMIR (Lu et al., 2026; Ma et al., 2024) unify multiple restoration tasks in one model, and CELL-E 2 and Vermeer (Khwaja et al., 2023; Kambhampati et al., 2026) are generative but serve the single task of producing protein-localization channels from protein sequences. Neither line handles variable panel composition together with multiple generative tasks.

3 Method

3.1 Problem formulation

We treat the set of single-channel fluorescence images acquired from one field of view as a sample $X = (x_1, \ldots, x_N)$, where the number of channels N depends on the data source. All four tasks in this paper generate fluorescence channels under a condition c,

$$\hat{X} \sim p_\theta(X \mid c)\,, \qquad c = (c_{\text{exp}}, R), \tag{1}$$

where the experimental condition c_{exp} always holds the perturbation type, the perturbation identity, the dose and the cell line, whereas the reference set R depends on the task. R is empty for morphology generation, the fluorescence images X^{ctrl} of a same-batch control for perturbation prediction, the observed channels $X^{\mathcal{O}}$ of the same field of view for missing-channel completion, which generates the remaining channels $X^{\mathcal{M}}$ so that X in Eq. (1) is $X^{\mathcal{M}}$, and the brightfield images X^{bf} of the same field of view, one or three focal planes in our data, for brightfield-to-fluorescence translation. The four tasks thus differ only in which images are given and which channels are generated, and Table 19 in the appendix summarizes this as an assignment of roles.

3.2 Architecture

Figure 2 gives the architecture. We use DiT (Peebles & Xie, 2023) as the backbone, which patchifies the latent representation into a token sequence and models it together with condition tokens by self-attention. To preserve the full subcellular structure information, each fluorescence channel is replicated to three channels and encoded separately by a frozen image autoencoder, $z_k = \text{Enc}(x_k)$, and Eq. (1) becomes

$$\hat{z} \sim p_\theta(z \mid c)\,, \tag{2}$$

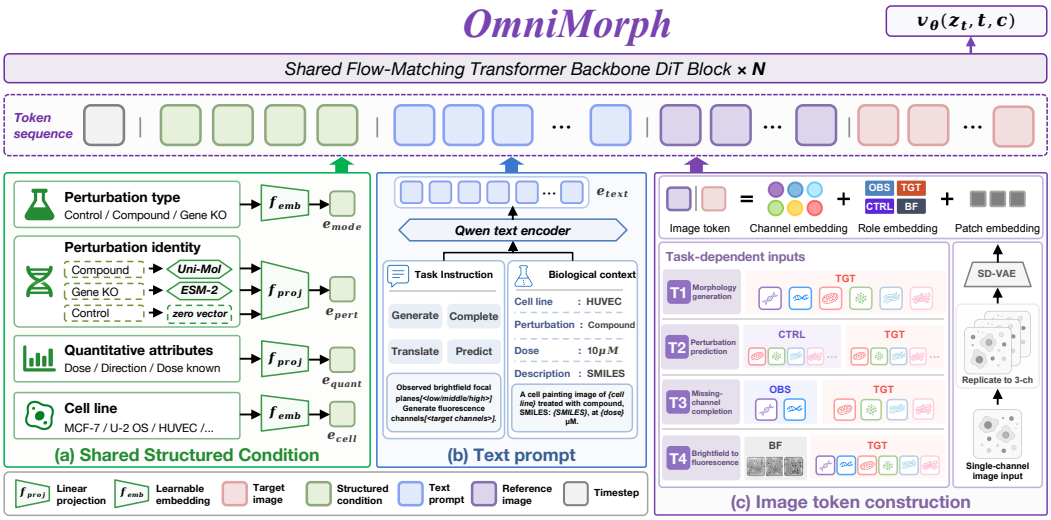

Figure 2: Architecture of OmniMorph. (a) The structured condition encodes the perturbation and the cell line, with compounds and genes passed through their own frozen encoders. (b) The text prompt is encoded by a frozen language model. (c) Each channel is encoded separately with a channel embedding and a role embedding, and the four tasks differ only in the roles of the image groups (Table 19). All tokens form one sequence for the flow-matching backbone.

where $z = \{z_k\}$ is the latent representation of the generated channels and the reference images in R are encoded channel by channel in the same way.

3.2.1 Channel semantics and task roles

Once the tokens of all channels are concatenated, the same sequence position corresponds to different subcellular structures in different datasets, and reserving a fixed position for every stain would not scale as new stains appear. We therefore attach a semantic embedding to every channel. Each of the seven subcellular structure classes in our data has a learnable vector in $\mathbb{R}^D$, and $e_{\mathrm{chan},\sigma(k)}$, where $\sigma(k)$ is the class of channel k, is added to all tokens of that channel. The embedding depends only on the structure, not on the dataset or the channel order, so channels showing the same structure share one vector across panels.

At the same time, the same channel plays different roles in different tasks, a reference in one task and a generation target in another, so we attach a role embedding to each channel (Chen et al., 2025). There are four roles, the four columns of Table 19, each with a learnable vector in $\mathbb{R}^D$.

The latent representation z_k of fluorescence channel k passes through $\mathrm{PatchEmbed}$ to give L tokens of width D, with all fluorescence channels sharing the same parameters. Each token then receives the two embeddings and the position encoding,

$$\tilde{z}_k^r = \mathrm{PatchEmbed}(z_k) + p + e_{\mathrm{chan},\sigma(k)} + e_{\mathrm{role},r}, \qquad r \in \{\mathrm{tgt}, \mathrm{obs}, \mathrm{ctrl}\}, \tag{3}$$

where r is the role of the channel in the current task, the spatial position encoding $p \in \mathbb{R}^{L\times D}$ is added token by token, and the two embeddings are vectors in $\mathbb{R}^D$ shared by the L tokens of the same channel. Brightfield images show no fluorescent structure, so their tokens are built in the same way with a separate $\mathrm{PatchEmbed}_{\mathrm{bf}}$, replacing e_{chan} with a focal-plane embedding $e_{\mathrm{pl},j}$ that marks the focus position of the j-th image, and taking role bf.

3.2.2 Condition encoding and input sequence

The experimental condition is encoded as four vectors. The perturbation type is one of control, compound and gene knockout and is mapped by a learnable embedding table f_{emb} to e_{mode}, and the cell line is mapped by a separate table of the same kind to e_{cell}. The perturbation identity and the

dose are computed from the input,

$$e_{\text{pert}} = f_{\text{proj}}\, f(a), \qquad e_{\text{quant}} = f_{\text{proj}}\,[\,\log_{10} d,\ s,\ m\,], \tag{4}$$

where a is the SMILES string of a compound or the protein sequence of a gene and f is the corresponding frozen encoder. Compounds are encoded by Uni-Mol (Zhou et al., 2023) into a single molecule-level vector, and genes are encoded by ESM-2 (Lin et al., 2023) with the per-residue representations averaged. Each f_{proj} is a separate two-layer MLP that maps to $\mathbb{R}^D$, and e_{pert} is the zero vector for controls. Because each perturbation type has its own encoder and projection and is identified by e_{mode}, one set of weights covers compounds, genes and controls. In e_{quant}, d is the compound concentration, $s \in \{-1, 0, +1\}$ marks gene knockout, control and compound, and $m \in \{0, 1\}$ indicates whether the dose is known, with the log dose set to zero when there is none.

The condition also contains a natural-language description, which a frozen Qwen2.5-14B (Qwen Team, 2024) encodes into a token sequence $e_{\text{text}} = f_{\text{proj}}\, g(y) \in \mathbb{R}^{L_{\text{text}} \times D}$, where y is the text, $g(y)$ is the sequence of per-token hidden states of the language model and f_{proj} projects it to width D. The text consists of two parts, a task instruction followed by a biological context. The templates of both parts and a complete example are given in Table 15 in the appendix.

Conditions enter the network in two ways, following DiT (Peebles & Xie, 2023) and SD3 (Esser et al., 2024). At the sequence level, the four vectors, the text tokens and the image tokens form the input sequence of the backbone,

$$v_\theta(z_t, t, c) \coloneqq \mathrm{DiT}\big(\,[\,\phi(t),\ e_{\text{mode}},\ e_{\text{pert}},\ e_{\text{quant}},\ e_{\text{cell}},\ e_{\text{text}},\ \tilde{z}^{\text{ref}},\ \tilde{z}^{\text{tgt}}\,]\,\big), \tag{5}$$

where $\phi(t)$ is the timestep embedding, $\tilde{z}^{\text{ref}}$ are the tokens of the reference channels from Eq. (3) with role obs, bf or ctrl, and $\tilde{z}^{\text{tgt}}$ are the tokens of the noised latent z_t of the generated channels with role tgt, with z_t defined in Section 3.3. The four tasks differ only in $\tilde{z}^{\text{ref}}$. At the output, the tokens at the positions of $\tilde{z}^{\text{tgt}}$ are mapped back to latent patches by a linear layer as the predicted velocity field. At the vector level, the four vectors are concatenated, passed through a zero-initialized projection W_0 and added to $\phi(t)$, $c_{\text{ada}} = \phi(t) + W_0[\,e_{\text{mode}}, e_{\text{pert}}, e_{\text{quant}}, e_{\text{cell}}\,]$, which modulates every DiT block through AdaLN-Zero. The text does not take part in modulation.

3.3 Training and sampling

The model is trained with flow matching (Lipman et al., 2023; Liu et al., 2023) on the linear path $z_t = t\, z + (1 - t)\, \epsilon$ between noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the latent representation of the generated channels, where $t = 1$ is the data end and t is sampled from a logit-normal distribution (Esser et al., 2024). The network v_θ predicts the velocity field, and the loss

$$\mathcal{L} = \mathbb{E}_{t,\epsilon}\, \big\| v_\theta(z_t, t, c) - (z - \epsilon) \big\|_2^2 \tag{6}$$

is averaged over all elements of the generated channels. Reference images receive no noise, and their tokens are always built from the true latent representation. Training has two stages. The model is first pretrained on morphology generation for 200k steps and then trained jointly on the four tasks for 200k steps, with the tasks sampled in the ratio 1:3:3:3 and the data sources balanced by the square root of their sizes. The training data are subsets of six public sources, listed in Table 16 in the appendix. Sampling integrates the velocity field from pure noise with a second-order Heun solver for 50 steps, and in missing-channel completion the observed channels are reset to their true latent representation after every step, as in training. Resolution, latent shape, condition dropout and other implementation details are given in Section A.11.

4 Experiments

We evaluate OmniMorph on the four tasks and compare it with representative methods for each task. Except for the ablations in Section 4.4, **all OmniMorph results use a single model and checkpoint across all datasets and four tasks,** whereas baselines use separate models for each task and dataset. All tasks share one set of data splits, and every test image is excluded from the training pools of every task. All baselines are retrained by us at 512 pixels on the same splits. The number of models, steps, batch size and sampler of every method are listed in Table 20, and every iterative sampler uses 50 steps. In all tables, **bold** marks the best entry in a column within a block and underline the second best. A dash marks an entry that the conditioning of a method cannot express.

Table 1: FID↓ of morphology generation and perturbation prediction on the same frozen targets, in-distribution settings (ID) on the left and unseen perturbations (OOD) on the right. The perturbation-prediction block generates from a same-batch control image and marks only its best entry. CPG1, CPG4 and CPG36 abbreviate CPG0001, CPG0004 and CPG0036. KID is reported in Table 6.

Method	ID								OOD							
	BBBC021	CPG1	CPG4	CPG36	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G	BBBC021	CPG1	CPG4	CPG36	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G
Morphology generation																
Stable Diffusion	214.9	141.0	193.7	264.8	218.1	200.1	108.7	197.7	235.6	150.9	221.5	260.0	222.9	195.9	112.6	181.9
Mol2Image	139.0	216.9	60.8	138.0	119.9	–	190.4	–	128.1	211.0	52.7	142.0	111.6	–	186.3	–
MorphoDiff	142.2	136.3	128.2	277.8	202.3	192.3	116.5	209.2	180.1	117.3	150.0	272.0	208.0	190.3	120.7	192.6
MorphGen	60.5	**49.4**	47.7	**82.2**	110.1	106.2	40.0	75.8	–	–	–	–	–	–	–	–
MDTv2	46.9	74.8	**26.5**	104.7	71.5	**60.3**	53.1	46.1	–	–	–	–	–	–	–	–
OmniMorph	**43.9**	50.4	32.7	93.5	**52.9**	61.1	**29.5**	**25.7**	**32.3**	**90.0**	**27.0**	**93.1**	**51.2**	**50.9**	**29.1**	**26.3**
Perturbation prediction																
IMPA	141.7	110.6	60.2	151.3	84.4	198.8	101.2	66.7	95.4	125.3	56.5	151.7	76.6	224.0	98.5	68.5
OmniMorph	**58.6**	**41.3**	**35.2**	**86.7**	**41.6**	**59.6**	**28.1**	**24.8**	**27.5**	**52.2**	**30.8**	**84.6**	**42.5**	**47.7**	**28.4**	**27.2**

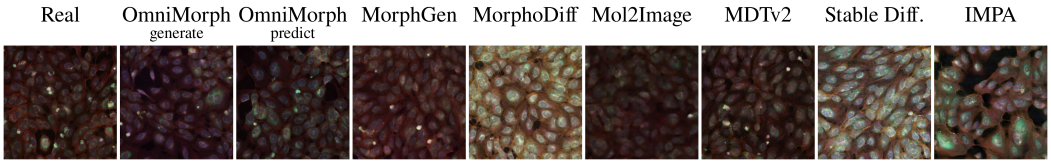

Figure 3: Morphology generation and perturbation prediction on one RxRx3 compound condition (ID). From left: the real image, OmniMorph from the condition alone (generate) and from the same-batch control (predict), and the baselines; only IMPA also starts from the control.

4.1 Morphology generation and perturbation prediction

Morphology generation synthesizes all fluorescence channels from the perturbation and the cell line alone, and perturbation prediction generates the same targets from a same-batch control image.

Data. Both tasks are evaluated on the same frozen targets of six datasets, BBBC021, CPG0001, CPG0004, CPG0036, CPG0000 and RxRx3, under in-distribution (ID) and unseen-perturbation (OOD) conditions. The compound and the gene experiment of CPG0000 and RxRx3 are trained and reported separately as CPG0-C, CPG0-G, RxRx3-C and RxRx3-G because some baselines support only one perturbation type. For perturbation prediction every target is additionally paired with a control image from the same plate and site.

Baselines. Morphology generation is compared with MorphoDiff (Navidi et al., 2025), MorphGen (Demirel et al., 2025), Mol2Image (Yang et al., 2021), MDTv2 (Gao et al., 2023) and an unconditional Stable Diffusion fine-tuned per dataset and perturbation type. MorphGen and MDTv2 condition on a closed class vocabulary and therefore appear only in the ID blocks. IMPA (Palma et al., 2025) is the only baseline that also starts from a control, so the other baselines enter perturbation prediction through the images they generate for the same targets.

Metrics. FID is computed per dataset against the real images of the same targets. Because several baselines output only RGB, all methods are compared under one fixed channel-to-RGB rendering per dataset, and methods that output fluorescence channels are also scored on their native channels. Biological identity is the balanced accuracy on five panels with mechanism-of-action or perturbation labels, from a classifier trained on real images only. The classifier uses OpenPhenom features (Kraus et al., 2024), with the panel construction and the classification protocol in Section A.1.1; Section A.5.1 repeats the five panels with CellProfiler features, under which OmniMorph is again the best generator on every panel. KID is given in Table 6, and the target lists and pairing rules in Section A.1.1.

In morphology generation under in-distribution conditions, OmniMorph ranks among the top two on all eight settings of Table 1 and first on four. The other four, CPG0001, CPG0004, CPG0036 and CPG0-G, are led by closed-vocabulary models, which learn a dedicated embedding for every perturbation seen in training and therefore hold an advantage on seen perturbations but cannot represent unseen ones. Under unseen perturbations, OmniMorph ranks first on all eight settings.

Table 2: Balanced accuracy↑ of a classifier trained on real images only, for morphology generation and perturbation prediction. All methods are scored on RGB renderings, and methods that output fluorescence channels are also scored on their native channels. Real images give the upper bound, and the two-row block marks only its best entry.

Method RGB / native	BBBC021 (MoA, 6)	CPG0001 (MoA, 10)	CPG0004 (cmpd, 10)	RxRx3-C (cmpd, 15)	RxRx3-G (gene, 15)
Real images (ref.)	0.954 / 0.975	0.815 / 0.818	0.571 / 0.642	0.620 / 0.684	0.760 / 0.781
Morphology generation					
Stable Diffusion	0.157	0.088	0.083	0.065	0.081
Mol2Image	0.262	0.095	0.195	0.099	–
MorphoDiff	0.253	0.092	0.108	0.087	0.085
MDTv2	0.250	0.299	0.175	0.237	0.242
MorphGen	0.670 / 0.661	0.532 / 0.586	0.236 / **0.477**	0.366 / 0.407	0.304 / 0.391
OmniMorph	**0.703 / 0.726**	**0.649 / 0.659**	**0.247** / 0.462	**0.476 / 0.479**	**0.526 / 0.708**
Perturbation prediction					
IMPA	0.164	0.135	0.211	0.126	0.074
OmniMorph	**0.730**	**0.565**	**0.231**	**0.488**	**0.467**

In perturbation prediction, OmniMorph attains a lower FID than IMPA on every setting of Table 1, in most settings by more than a factor of two. Relative to its own morphology generation, starting from the control improves the results on most datasets. On the unseen-perturbation setting of CPG0001 the FID decreases from 90.0 to 52.2, indicating that the control image supplies plate-level appearance that the perturbation identity alone does not. Figure 3 shows this on one RxRx3 condition: the prediction from the control matches the intensity and color balance of the real image better than the generation, whereas MorphoDiff, Stable Diffusion and IMPA are markedly brighter. Morphology generation remains better on CPG0004 and on BBBC021 under in-distribution conditions.

On the biological readout of Table 2, OmniMorph attains the highest balanced accuracy of morphology generation on all five panels under the RGB protocol, with MorphGen second on each. MDTv2 approaches OmniMorph in in-distribution FID, yet its balanced accuracy is below half that of OmniMorph on most panels, and MorphoDiff, Mol2Image and Stable Diffusion remain near chance level, so a close image distribution does not imply the correct perturbation identity. In perturbation prediction, OmniMorph exceeds every generation baseline on four of the five panels and is slightly below MorphGen on CPG0004, whereas IMPA remains near chance level and thus preserves the control morphology without introducing the perturbation. Under the native-channel protocol, both real images and OmniMorph attain higher balanced accuracy on every panel than under RGB, for OmniMorph on the RxRx3 gene panel 0.708 against 0.526. The RGB rendering thus discards channel information that distinguishes perturbations, and OmniMorph leads on four panels. OmniMorph therefore encodes the perturbation identity accurately enough both to generate the perturbed morphology from the condition alone and to move a control image toward it.

4.2 Channel completion

Channel completion infers the remaining fluorescence channels from the measured ones. It is evaluated on CPG0000, CPG0001, CPG0004 and RxRx3 with a frozen mask bank built on plates never seen in training. The mask bank fixes the observed and the hidden channels of every sample, so every method receives the same input and is scored only on the hidden channels. The baselines are a U-Net (Ronneberger et al., 2015), Pix2Pix (Isola et al., 2017) and MCD (Zhang et al., 2026a), each trained with random masking. The metrics are Pearson correlation, SSIM and MAE on the hidden channels, with MAE computed in the raw intensity range. The composition of the mask bank is given in Section A.2 and the configuration of each method in Table 20.

In channel completion, OmniMorph attains the highest Pearson correlation on three of the four datasets of Table 3 and the best value of all three metrics on CPG0004. U-Net leads on all three metrics on RxRx3 and is marginally better on some pixel metrics of the other datasets. Pix2Pix and MCD trail clearly on every dataset. The three baselines are trained directly with pixel-wise losses and therefore hold a natural advantage on these metrics. As a generative model, OmniMorph nev-

Table 3: Pearson correlation, SSIM and MAE of channel completion, computed on the hidden channels of the frozen mask bank, with MAE in the raw intensity range.

	CPG0000			CPG0001			CPG0004			RxRx3		
Method	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
MCD	0.762	0.597	0.0548	0.650	0.609	0.0910	0.750	0.461	0.0723	0.697	0.929	0.0108
Pix2Pix	0.814	0.940	0.0088	0.681	0.759	0.0558	0.835	0.958	0.0085	0.827	0.898	0.0123
U-Net	0.836	**0.949**	**0.0082**	0.750	0.820	**0.0464**	0.870	0.971	0.0070	**0.873**	**0.959**	**0.0071**
OmniMorph	**0.875**	0.942	0.0086	**0.824**	**0.823**	0.0468	**0.894**	**0.971**	**0.0065**	0.859	0.951	0.0077

Table 4: Pearson correlation, SSIM and MAE of brightfield-to-fluorescence translation, reported per dataset and overall, with the overall score weighted by the test-pool size of each dataset.

	CPG0000 (3-plane)			CPG0001 (1-plane)			Overall		
Method	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
CellResDM	**0.955**	0.844	0.0227	0.647	0.705	0.0618	0.823	0.784	0.0394
ISL	0.884	0.679	0.0327	0.700	0.773	0.0465	0.805	0.719	0.0386
cWGAN-GP	0.908	0.851	0.0271	0.756	**0.823**	**0.0439**	0.843	0.839	0.0343
OmniMorph	0.954	**0.905**	**0.0175**	**0.785**	0.811	0.0442	**0.882**	**0.865**	**0.0289**

ertheless matches them and leads in structural correlation, which lays the groundwork for extending to Cell Painting panels that label a wider range of subcellular structures.

4.3 Brightfield-to-fluorescence translation

Brightfield-to-fluorescence translation predicts the fluorescence channels from unstained transmitted-light images. It is evaluated on CPG0000 and CPG0001, whose brightfield inputs differ. CPG0000 provides 3 brightfield planes and CPG0001 a single one. Each method is therefore trained as one model that accepts both inputs and is evaluated on plates never seen in training. The single plane of CPG0001 carries fewer defocus cues and makes the mapping more underdetermined. The baselines are cWGAN-GP (Gulrajani et al., 2017), ISL (Christiansen et al., 2018) and the residual-diffusion model CellResDM (Xing et al., 2024). The metrics are those of channel completion, reported per dataset and overall, with the overall score weighted by the test-pool size of each dataset. The composition of the held-out plates is given in Section A.3 and the configuration of each method in Table 20.

In brightfield-to-fluorescence translation, OmniMorph attains the highest SSIM and the lowest MAE on CPG0000 in Table 4 and trails CellResDM in Pearson correlation by only 0.001. On CPG0001, OmniMorph attains the highest Pearson correlation, and its SSIM and MAE are slightly below those of cWGAN-GP. Figure 4 shows one CPG0001 field, on which the error map of OmniMorph is the darkest of the four methods and that of ISL the brightest. CellResDM matches OmniMorph on the three-plane CPG0000 but drops sharply on the single-plane CPG0001, whereas OmniMorph remains ahead under both inputs and is thus more robust to the amount of information in the input. cWGAN-GP and ISL are trained directly with pixel-wise losses, which favors them on these metrics. OmniMorph still attains the best overall value of all three metrics, which lays a foundation for extending the task to temporal sequences.

4.4 Ablations

The ablations test three design choices, namely the channel embedding, multi-dataset training and multi-task training. All variants share the architecture, optimizer and step budget of the OmniMorph model they are compared with. (a) Channel embedding compares the first-stage model, trained on morphology generation alone, with the same model without the channel embedding. (b) Multi-dataset training compares the first-stage model trained on each dataset separately with the one trained on all datasets. (c) Multi-task training compares, on each of the three other tasks, a single-task model trained from scratch, a model pretrained on morphology generation and then trained on that task alone, and OmniMorph. The FID version of (b) is given in Table 8.

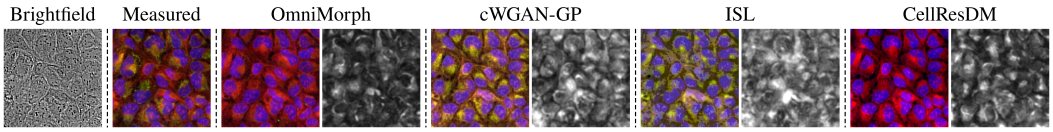

Figure 4: Brightfield-to-fluorescence translation on one CPG0001 field. Left: the brightfield input and the measured fluorescence; each method shows its prediction and, in gray, its absolute error to the measurement, averaged over the five channels on a shared scale.

Table 5: Ablations. (a) Mean FID↓ over eight datasets under the full-reference protocol of Section A.4; per-dataset values in Table 7. (b) Balanced accuracy↑ on the five panels of Table 2. (c) Mean in-distribution FID↓ and KID↓ ($\times 10^{-3}$) over seven perturbation-prediction settings, channel completion averaged over its four datasets, and the overall brightfield-to-fluorescence score of Table 4; per-dataset values in Table 9.

(b) Multi-dataset training

Model	BBBC021	CPG1	CPG4	RxRx3-C	RxRx3-G
Specialists	0.671	0.544	0.215	0.146	0.101
OmniMorph	**0.703**	**0.649**	**0.247**	**0.476**	**0.526**

(a) Channel embedding

Model	ID	OOD	Mean
Without channel embedding	150.7	155.1	152.9
OmniMorph	**49.3**	**51.4**	**50.4**

(c) Multi-task training

	Perturbation prediction		Channel completion			Brightfield-to-fluorescence		
Method	FID↓	KID↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
Single-task	81.2	54.8	0.762	0.882	0.0204	0.806	0.812	0.0354
Pretrained + single-task	54.7	20.9	0.803	0.897	0.0194	0.830	0.825	0.0334
OmniMorph	**41.3**	**8.4**	**0.863**	**0.922**	**0.0174**	**0.882**	**0.865**	**0.0289**

Channel embedding. Removing the channel embedding increases the FID on every dataset and split and roughly triples the mean FID in Table 5(a), from 50.4 to 152.9. In its absence, the model can infer the meaning of a channel only from its position, whereas the number of channels differs across datasets and the same structure appears different under different dyes and imaging protocols. The channel embedding labels the subcellular structure of each channel directly, so channels of the same structure share one representation across datasets, which is the basis of joint training. Removing the text condition instead (Table 10) more than doubles the morphology-generation FIDs of CPG0000 and CPG0036 and lowers the Pearson correlation of channel completion by 0.04 to 0.08.

Multi-dataset training. Joint training exposes the model to the perturbations of all six datasets. The shared model exceeds the per-dataset specialists on all five biological panels of Table 5(b), with the largest margins on the two RxRx3 panels.

Multi-task training. The single-task model trained from scratch is the weakest on every metric, so multi-dataset pretraining benefits every task. The pretrained single-task model and OmniMorph start from the same checkpoint and differ only in that OmniMorph divides its training budget among the four tasks, yet OmniMorph leads on every averaged metric of Table 5(c) and, per dataset, attains the lowest FID on all seven perturbation-prediction settings of Table 9. Joint multi-task training does not cause competition for capacity between tasks but yields gains across them.

5 Conclusion

We presented OmniMorph, a unified generative model for Cell Painting that handles datasets with different channel compositions and four generative tasks. It mostly outperforms task-specific methods, and ablations show that joint training shares knowledge across datasets and yields gains across tasks. The design remains limited in two respects. The task instruction is a template filled from structured fields rather than free-form language. Image metrics also do not tell whether the perturbation identity, the mechanism of action or the dose response is preserved, so biological tests in feature space remain necessary. Next, we plan to add datasets with more diverse channels and extend the channel vocabulary to open protein identities such as those in ProtiCelli (Sun et al., 2026).

AI USE STATEMENT

Generative AI tools were used for code organization, language polishing, and assistance with literature searches. All resulting content was independently reviewed and confirmed by the authors.

ETHICS STATEMENT

This work uses only publicly released Cell Painting image datasets of cultured cell lines; no human subjects, patient data or animal experiments are involved, and every dataset is used under its license (Section A.13). OmniMorph generates fluorescence images of cells under chemical or genetic perturbations in order to reduce the cost of screening experiments in research. It does not replace experimental validation, and generated images should not be taken as evidence of a biological effect without such validation. Beyond the risks common to generative image models, we see no direct pathway to harmful applications.

REPRODUCIBILITY STATEMENT

All datasets are public: BBBC021, the Cell Painting Gallery sets CPG0000, CPG0001, CPG0004 and CPG0036, and RxRx3-core. Section 3 and Section A.11 give the architecture, the training schedule, the sampler and the implementation details, Section A.12 the configuration of every baseline, and Sections A.1.1, A.2, A.3 and A.10 the frozen target lists, control pairings, mask bank, held-out plates and data splits behind every table; Section A.13 lists licenses and compute. The code, the frozen target lists, the mask bank, the control pairings, the panel definitions and the trained weights will be released, so that every number in the paper can be recomputed from the same frozen inputs.

REFERENCES

Shunxing Bao et al. Random multi-channel image synthesis for multiplexed immunofluorescence imaging. In *Proceedings of Machine Learning Research (MICCAI Workshop on Computational Pathology)*, volume 156, pp. 36–46, 2021.

Anis Bourou et al. PhenDiff: Revealing subtle phenotypes with diffusion models in real images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024. doi: 10.1007/978-3-031-72384-1_34.

Peter D. Caie, Rebecca E. Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E. Roberts, and Neil O. Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, 2010. doi: 10.1158/1535-7163.MCT-09-1148.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, Laurent Brino, Patrick J. Byrne, Hugo Ceulemans, Carolyn Ch'ng, Beth A. Cimini, Djork-Arne Clevert, Nicole Deflaux, John G. Doench, Thierry Dorval, Regis Doyonnas, Vincenza Dragone, Ola Engkvist, Patrick W. Faloon, Briana Fritchman, Florian Fuchs, Sakshi Garg, Tamara J. Gilbert, David Glazer, David Gnutt, Amy Goodale, Jeremy Grignard, Judith Guenther, Yu Han, Zahra Hanifehlou, Santosh Hariharan, Desiree Hernandez, Shane R. Horman, Gisela Hormel, Michael Huntley, Ilknur Icke, Makiyo Iida, Christina B. Jacob, Steffen Jaensch, Jawahar Khetan, Maria Kost-Alimova, Tomasz Krawiec, Daniel Kuhn, Charles-Hugues Lardeau, Amanda Lembke, Francis Lin, Kevin D. Little, Kenneth R. Lofstrom, Sofia Lotfi, David J. Logan, Yi Luo, Franck Madoux, Paula A. Marin Zapata, Brittany A. Marion, Glynn Martin, Nicola Jane McCarthy, Lewis Mervin, Lisa Miller, Haseeb Mohamed, Tiziana Monteverde, Elizabeth Mouchet, Barbara Nicke, Arnaud Ogier, Anne-Laure Ong, Marc Osterland, Magdalena Otrocka, Pieter J. Peeters, James Pilling, Stefan Prechtl, Chen Qian, Krzysztof Rataj, David E. Root, Sylvie K. Sakata, Simon Scrace, Hajime Shimizu, David Simon, Peter Sommer, Craig Spruiell, Iffat Sumia, Susanne E. Swalley, Hiroki Terauchi, Amandine Thibaudeau, Amy Unruh, Jelle Van de Waeter, Michiel Van Dyck, Carlo van Staden, Michał Warchoł, Erin Weisbart, Amélie Weiss, Nicolas Wiest-Daessle, Guy Williams, Shan Yu, Bolek Zapiec, Marek Żyła, Shantanu Singh, and Anne E. Carpenter. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. bioRxiv preprint, 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, Alexandr A. Kalinin, John Arevalo, Marzieh Haghighi, Juan C. Caicedo, Daniel Kuhn, Desiree Hernandez, James Berstler, Hamdah Shafqat-Abbasi, David E. Root, Susanne E. Swalley, Sakshi Garg, Shantanu Singh, and Anne E. Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6): 1114–1121, 2024. doi: 10.1038/s41592-024-02241-6.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. UniReal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12501–12511, 2025. doi: 10.1109/CVPR52734.2025.01166.

Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, Piyush Goyal, William Fedus, Ryan Poplin, Andre Esteva, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner. In Silico Labeling: Predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803.e19, 2018. doi: 10.1016/j.cell.2018.03.040.

Beth A. Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, et al. Optimizing the Cell Painting assay for image-based profiling. *Nature Protocols*, 18:1981–2013, 2023. doi: 10.1038/s41596-023-00840-9.

Jan Oscar Cross-Zamirski et al. Label-free prediction of Cell Painting from brightfield images. *Scientific Reports*, 12:10001, 2022. doi: 10.1038/s41598-022-12914-x.

Jan Oscar Cross-Zamirski et al. Class-guided image-to-image diffusion: Cell Painting from brightfield images with class labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3800–3809, 2023.

Berker Demirel, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. MorphGen: Controllable and morphologically plausible generative cell-imaging. arXiv:2510.01298, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Marta M. Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, Jordan Christensen, Berton A. Earnshaw, Imran S. Haque, and Ben Mabey. RxRx3: Phenomics map of biology. bioRxiv preprint, 2023.

Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. UniVG: A generalist diffusion model for unified image generation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17160–17170, 2025. doi: 10.1109/ICCV51701.2025.01594.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23164–23173, 2023.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yikai Han et al. DiffStain: Conditioned diffusion-based semantic virtual staining with mask guidance. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025a. doi: 10.1007/978-3-032-05169-1_14.

Zhen Han et al. ACE: All-round creator and editor following instructions via diffusion transformer. In *International Conference on Learning Representations (ICLR)*, 2025b.

Albert Hung, Charles J. Zhang, Jonathan Z. Sexton, Matthew J. O'Meara, and Joshua D. Welch. LUMIC: Latent diffUsion for multiplexed images of cells. bioRxiv, 2024.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

Sandeep Kambhampati, Eric Zimmermann, Emre Hayir, Kevin K. Yang, Fei Chen, and Alex X. Lu. Vermeer: Autoregressive generative modeling of microscopy predicts protein localization. bioRxiv, 2026.

Emaad Khwaja, Yun S. Song, Aaron Agarunov, and Bo Huang. CELL-E 2: Translating proteins to pictures and back with a bidirectional text-to-image transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Oren Kraus, Federico Comitani, John Urbanik, Kian Kenyon-Dean, Lakshmanan Arumugam, Saber Saberian, Cas Wognum, Safiye Celik, and Imran S. Haque. RxRx3-core: Benchmarking drug-target interactions in high-content microscopy. arXiv preprint arXiv:2503.20158, 2025.

Alexis Lamiable et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 14:6386, 2023. doi: 10.1038/s41467-023-42124-6.

Duong H. Le et al. One diffusion to generate them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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

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 (NeurIPS)*, 2025.

Qiang Lu, Xiao Liu, Qi Feng, et al. A foundation model for multi-task cross-distribution restoration of fluorescence microscopy images. *Nature Communications*, 17:3729, 2026. doi: 10.1038/s41467-026-70307-4.

Chenxi Ma, Weimin Tan, Ruian He, and Bo Yan. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nature Methods*, 21:1558–1567, 2024. doi: 10.1038/s41592-024-02244-3.

Zeinab Navidi et al. MorphoDiff: Cellular morphology painting with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Thanh C. Nguyen, Vy Bui, Anh Thai, Van K. Lam, Christopher B. Raub, Lin-Ching Chang, and Georges Nehmetallah. Virtual organelle self-coding for fluorescence imaging via adversarial learning. *Journal of Biomedical Optics*, 25(9):096009, 2020. doi: 10.1117/1.JBO.25.9.096009.

Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11):917–920, 2018. doi: 10.1038/s41592-018-0111-2.

Alessandro Palma, Fabian J. Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025. doi: 10.1038/s41467-024-55707-8.

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

Alexander Peysakhovich, William Berman, Joseph Rufo, Felix Wong, and Maxwell Z. Wilson. MONET: Virtual Cell Painting of brightfield images and time lapses using reference consistent diffusion. arXiv:2512.11928, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Zachary Sims, Gordon B. Mills, and Young Hwan Chang. MIM-CyCIF: masked imaging modeling for enhancing cyclic immunofluorescence with panel reduction and imputation. *Communications Biology*, 7:409, 2024. doi: 10.1038/s42003-024-06110-y.

Zachary Sims, Sandhya Govindarajan, Gordon B. Mills, Sebnem Ece Eksi, and Young Hwan Chang. Language of stains: Tokenization enhances multiplex immunofluorescence and histology image synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15963 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-3-032-04965-0_26.

Zachary Sims, Sandhya Govindarajan, Kaoutar Ait-Ahmad, Cigdem Ak, Marigold Kuykendall, Gordon B. Mills, Sebnem Ece Eksi, and Young Hwan Chang. miniMTI: minimal multiplex tissue imaging enhances biomarker expression prediction from histology. Research Square preprint, 2026.

Huangqingbo Sun, Konstantin Kahnert, Jan N. Hansen, William Leineweber, Mingyang Li, Wanyue Feng, Frederic Ballllosera, Ulrika Axelsson, Wei Ouyang, and Emma Lundberg. Generative machine learning unlocks the first proteome-wide image of human cells. bioRxiv, 2026.

Xuesong Wang, Yimin Fan, Yucheng Guo, Chenghao Fu, Kinhei Lee, Khachatur Dallakyan, Yaxuan Li, Qijin Yin, Yu Li, and Le Song. Prediction of cellular morphology changes under perturbations with a transcriptome-guided diffusion model. *Nature Communications*, 16:8210, 2025. doi: 10.1038/s41467-025-63478-z.

Gregory P. Way, Ted Natoli, Adeniyi Adeboye, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Systems*, 13(11):911–923, 2022. doi: 10.1016/j.cels.2022.10.001.

Christopher Wolff, Martin Neuenschwander, Carsten Jörn Beese, Divya Sitani, Maria C. Ramos, Alzbeta Srovnalova, María José Varela, Pavel Polishchuk, Katholiki E. Skopelitou, Ctibor Škuta, Bahne Stechmann, José Brea, Mads Hartvig Clausen, Petr Dzubak, Rosario Fernández-Godino, Olga Genilloud, Marian Hajduch, María Isabel Loza, Martin Lehmann, Jens Peter von Kries, Han Sun, and Christopher Schmied. Morphological profiling data resource enables prediction of chemical compound properties. *iScience*, 28(5):112445, 2025. doi: 10.1016/j.isci.2025.112445.

Eric Wu, Alexandro E. Trevino, Zhenqin Wu, et al. 7-UP: Generating in silico CODEX from a small set of immunofluorescence markers. *PNAS Nexus*, 2(6):pgad171, 2023. doi: 10.1093/pnasnexus/pgad171.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304, 2025. doi: 10.1109/CVPR52734.2025.01241.

Xiaodan Xing, Chunling Tang, Siofra Murdoch, Giorgos Papanastasiou, Yunzhe Guo, Xianglu Xiao, Jan Cross-Zamirski, Carola-Bibiane Schönlieb, Kristina Xiao Liang, Zhangming Niu, Evandro Fei Fang, Yinhai Wang, and Guang Yang. Artificial immunofluorescence in a flash: Rapid synthetic imaging from brightfield through residual diffusion. arXiv:2407.17882, 2024.

Karren Yang et al. Mol2Image: Improved conditional flow models for molecule to image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Haoran Zhang, Mingyuan Zhou, and Wesley Tansey. Controllable diffusion-based generation for multi-channel biological data. In *International Conference on Learning Representations (ICLR)*, 2026a.

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 *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 76111–76138, 2025.

Yuhui Zhang et al. CellFluxV2: An image generative foundation model for virtual cell modeling. bioRxiv, 2026b.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

A APPENDIX

The appendix follows the order of Section 4. For each task it gives the setup details that the main text refers to, the qualitative results and the tables that the main text summarises: the KID companion of Table 1, the FID version of the multi-dataset ablation and the per-dataset values of the multi-task ablation. The configuration under which every baseline was retrained is given with the other reproduction details. Analyses of the biological feature space come next, and the reproduction details of the data, of the model and of the baselines close the appendix.

A.1 MORPHOLOGY GENERATION AND PERTURBATION PREDICTION

A.1.1 EVALUATION PROTOCOL: TARGETS, CONTROLS AND PANELS

Every dataset and split of Table 1 fixes one frozen list of targets, drawn evenly spaced from the frozen test manifest of that dataset, perturbation type and split. An ID split holds perturbations whose identity occurred in training, and an OOD split holds compound or gene identities that did not. Every method generates one image per target (50 sampler steps for the diffusion and flow-matching models, one pass for Mol2Image and IMPA), and the reference set is the real images of exactly the same targets, so within a setting every method is compared against the same images at the same count. FID and KID use the 2048-dimensional pool features of Inception-V3 on the frozen per-dataset channel-to-RGB rendering, which is shared by every method and by both sides of the comparison. With (μ_g, Σ_g) and (μ_r, Σ_r) the mean and covariance of the generated and the real features of a setting,

$$\mathrm{FID} = \|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}\Big(\Sigma_g + \Sigma_r - 2\,(\Sigma_g\Sigma_r)^{1/2}\Big)\,, \tag{7}$$

and KID is the unbiased estimate of the squared maximum mean discrepancy under the cubic polynomial kernel $k(x, y) = (x^\top y/d + 1)^3$, $d = 2048$,

$$\mathrm{KID} = \frac{1}{m(m-1)}\sum_{i\neq j} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(y_i, y_j) - \frac{2}{mn}\sum_{i,j} k(x_i, y_j), \tag{8}$$

with $x_1, \ldots, x_m$ the generated and $y_1, \ldots, y_n$ the real features, averaged over 100 random subsets of 512 and reported in units of 10^{-3}.

Perturbation prediction is scored on the same targets. Each target is paired with a negative-control image of the same plate, taken at the same site where one exists, and these control images never enter a training pool. Targets without such a protected control are left out of the perturbation-prediction rows. Only OmniMorph and IMPA start from a control image; the remaining baselines are compared through their generations of the same targets and are not scored a second time.

Balanced accuracy (Table 2) is measured on five frozen panels: BBBC021 with its six mechanism-of-action classes (235 query images), CPG0001 with ten mechanism-of-action classes (1,110), CPG0004 with ten compounds (112), and RxRx3 with fifteen compounds (149) and fifteen gene knockouts (1,087). Each panel was frozen before any generated image was scored. BBBC021 keeps the complete label set of its library; the other panels keep the classes that a real-image screen ranks strongest, by cross-plate retrieval of real images with the query's plate excluded from the gallery, requiring at least four held-out query images per kept class. The CPG0004 panel pins each compound to a single dose and drops compounds whose held-out images span several doses.

Features come from the OpenPhenom microscopy foundation model: each channel enters as one 512×512 grayscale image, its patch tokens are mean-pooled to a 384-dimensional vector, and the per-channel vectors occupy six fixed slots, giving a 2304-dimensional image feature. Under the RGB protocol the three channels of the frozen per-dataset rendering fill the first three slots, and every method is scored this way; under the native-channel protocol each measured fluorescence channel fills its own slot, which only the methods that output fluorescence channels support.

The classifier of a panel is a one-vs-rest logistic regression with balanced class weights on z-scored features. The z-scoring statistics and the classifier are fit on real images of the training split only; generated images never train the classifier and enter only as queries, one image per held-out query row. Balanced accuracy is the macro recall over the K classes present among the queries, $\mathrm{BAcc} =$

$\frac{1}{K}\sum_{k=1}^{K}\mathrm{TP}_k/(\mathrm{TP}_k+\mathrm{FN}_k)$, so that a classifier predicting one class for every query scores $1/K$. In Table 2, the queries of a morphology-generation row are the generations of that method for the panel's query conditions, and the queries of a perturbation-prediction row are generated from the control image paired with each query row, so rows without a control drop out: 1,103 of 1,110 on CPG0001, 111 of 112 on CPG0004, and 860 of 1,087 on the RxRx3 gene panel, while on BBBC021 the control is taken from the same plate and site. The real-image row applies the identical classifier to the real query images of the full panel.

A.1.2 KID AND PER-IMAGE DISTANCES

Table 6 gives the KID of every entry of Table 1, and Figure 5 looks behind these single numbers at the individual images: for every generated image of a method, the distance in the Inception feature space in which FID and KID are computed to the nearest real target of the setting (fidelity), and for every real target the distance to the nearest generation of the method (coverage). A box collects the values of one method on one setting, so the 512 targets of a setting enter one by one instead of as one number, and the grey box is the reference of the real targets against each other, each matched to its nearest other real target. Under in-distribution conditions OmniMorph is closest to the real targets on BBBC021 and the two RxRx3 settings, and on the CPG settings it is within 0.1 (CPG0036, behind MorphGen) to 0.8 (CPG0001, behind MDTv2) of the closest method, with the same order in coverage, where OmniMorph and MDTv2 are within 0.4 of each other. Under unseen perturbations OmniMorph has the lowest fidelity distance on every setting except CPG0001, where IMPA, Stable Diffusion and Mol2Image are lower, and the lowest coverage distance on every setting. On RxRx3 its generations are as close to the real targets as the real targets are to each other. Methods appear where their generations for the frozen targets are available, as in Table 1.

Table 6: KID↓ in units of 10^{-3} for every entry of Table 1, on the same targets and the same reference sets, so the two tables describe the same generated images entry for entry; in-distribution settings in the upper half, unseen perturbations in the lower half. Blocks, marks, dashes and dataset abbreviations follow Table 1.

	In-distribution (ID)							
Method	BBBC021	CPG1	CPG4	CPG36	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G
Morphology generation								
Stable Diffusion	216.3	117.5	208.6	372.5	192.7	162.5	117.2	258.3
Mol2Image	70.6	216.4	43.2	171.4	75.2	–	252.0	–
MorphoDiff	119.7	111.5	98.9	392.9	182.0	154.5	129.5	276.7
MorphGen	19.7	18.4	17.5	**87.7**	80.7	59.9	27.1	34.9
MDTv2	4.3	48.1	**7.0**	109.0	46.4	28.9	43.8	33.1
OmniMorph	**4.1**	**11.0**	9.2	109.0	**18.3**	**13.6**	**13.6**	**7.0**
Perturbation prediction								
IMPA	78.7	76.6	32.6	197.7	43.8	183.0	106.5	64.2
OmniMorph	**8.8**	**6.3**	**9.2**	**100.3**	**7.8**	**7.3**	**11.8**	**7.8**
	Unseen perturbations (OOD)							
Method	BBBC021	CPG1	CPG4	CPG36	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G
Morphology generation								
Stable Diffusion	266.1	143.8	241.4	361.5	201.1	169.5	121.0	231.5
Mol2Image	74.0	269.8	40.1	177.6	77.5	–	243.1	–
MorphoDiff	197.3	89.4	115.1	380.2	183.5	162.1	135.5	248.9
MorphGen	–	–	–	–	–	–	–	–
MDTv2	–	–	–	–	–	–	–	–
OmniMorph	**7.5**	**59.2**	**9.4**	**109.6**	**16.6**	**13.2**	**12.3**	**9.4**
Perturbation prediction								
IMPA	53.3	102.3	39.2	200.6	38.9	200.0	101.1	65.5
OmniMorph	**3.9**	**25.5**	**8.0**	**96.2**	**9.1**	**6.6**	**10.7**	**11.8**

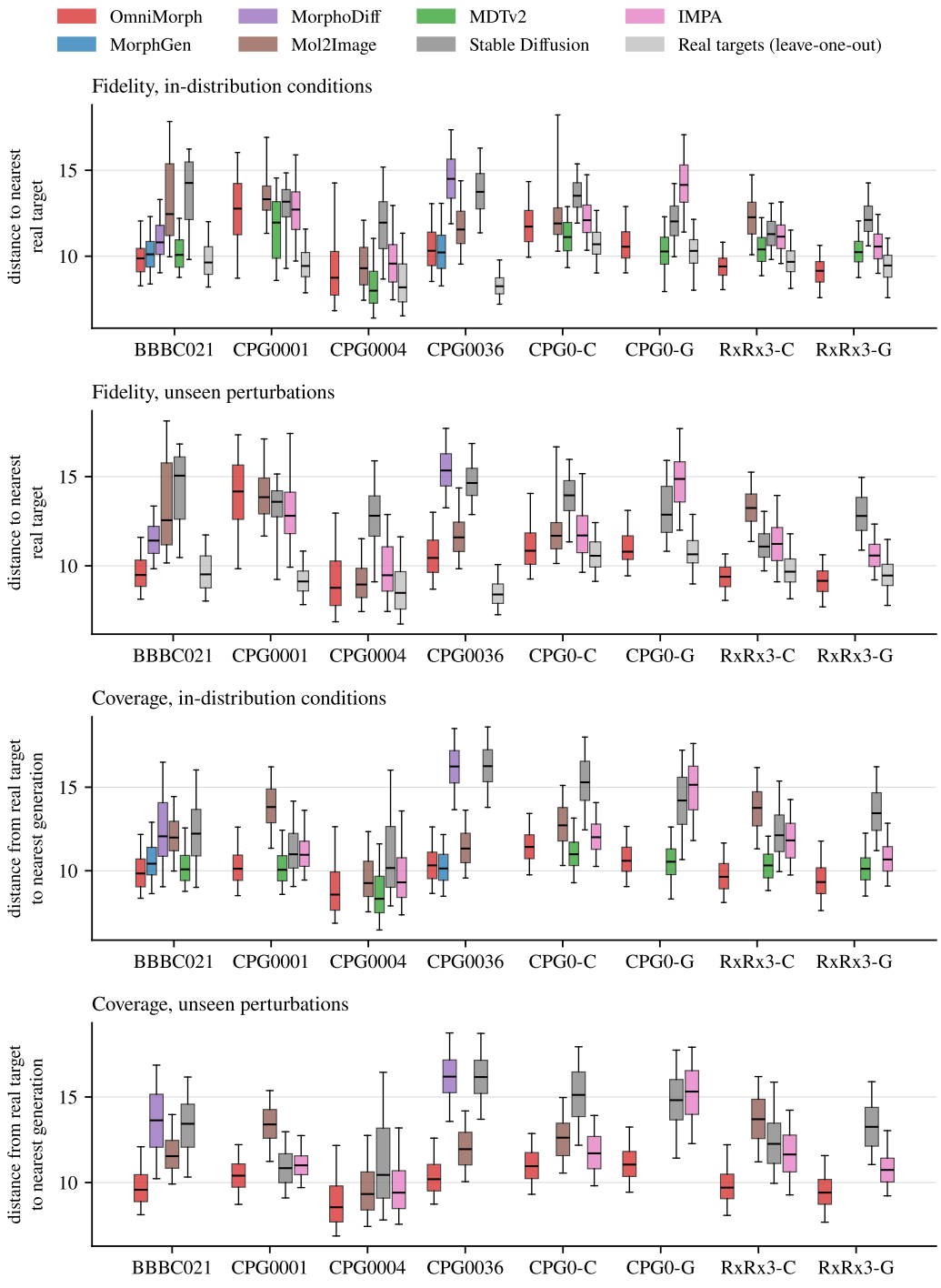

Figure 5: Per-image distances in Inception feature space on the frozen targets of Table 1. Fidelity (top two panels): for every generated image of a method, the distance to the nearest real target of the setting. Coverage (bottom two): for every real target, the distance to the nearest generation of the method. A box collects the values of one method on one setting (box: quartiles, line: median, whiskers: 5th to 95th percentile); within each measure, in-distribution conditions above and unseen perturbations below. Grey: the real targets against each other, each matched to its nearest other real target. Methods appear where their generations for the frozen targets are available, as in Table 1.

A.1.3 QUALITATIVE RESULTS

Figures 6 and 7 show morphology generation and Figures 8 and 9 perturbation prediction, with one dataset setting per column and one method per row. Within a column every method is shown on the same target, and every tile is rendered with the channel-to-RGB rule of the FID protocol. On CPG0036, where MorphoDiff and Stable Diffusion return near-uniform colour fields for most targets, their tiles and the MDTv2 tile are representative samples from the generation sets. A cell is marked N/A when a method cannot serve the setting: MorphGen and MDTv2 condition on a closed vocabulary and have no OOD generations, and Mol2Image conditions on SMILES strings and has no gene settings. The perturbation-prediction figures add the control image from which OmniMorph and IMPA generate. Figure 10 adds four further examples per ID setting: the presentable generations of OmniMorph closest to their real image, on distinct perturbations, next to those real images.

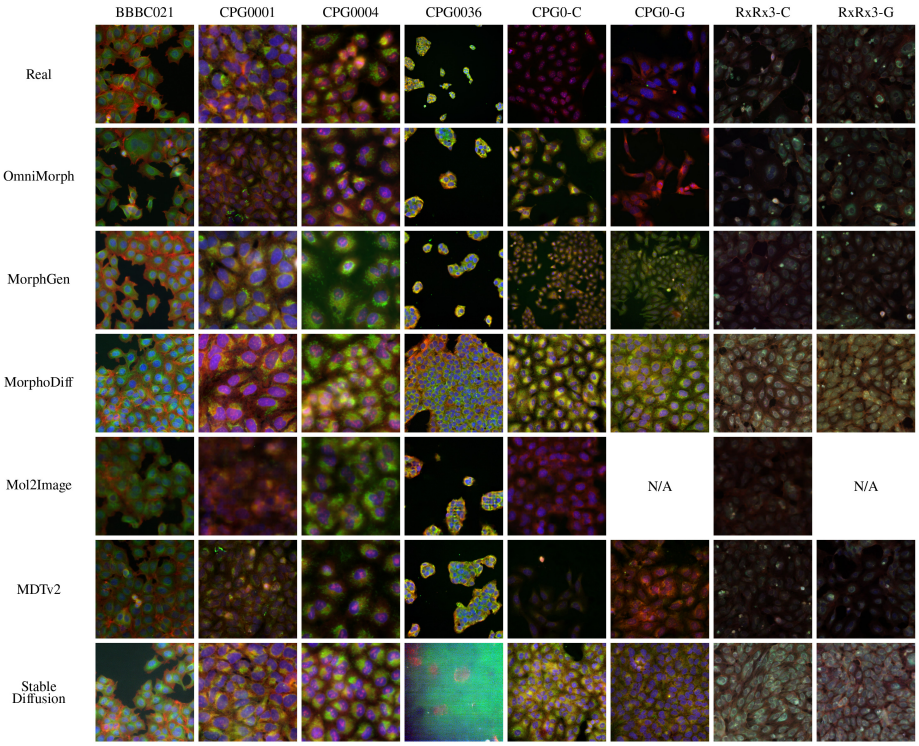

Figure 6: Morphology generation under in-distribution conditions. One perturbation per dataset setting (columns) and one method per row; the top row is the real image of the target. All tiles use the frozen channel-to-RGB rendering of the dataset.

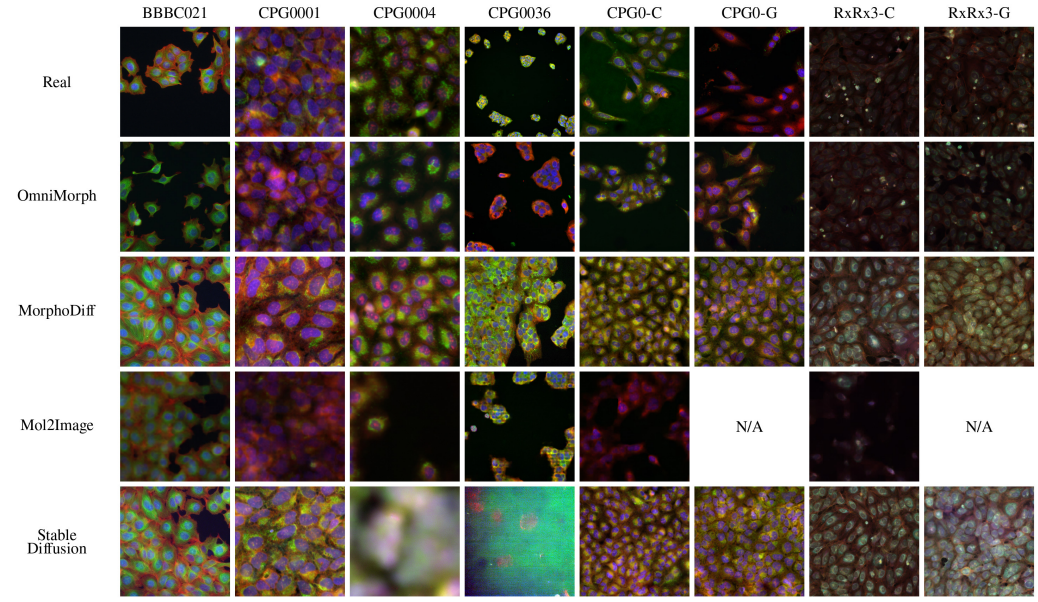

Figure 7: Morphology generation under unseen perturbations, laid out as Figure 6. MorphGen and MDTv2 have no generations for unseen perturbations, and Mol2Image has no gene settings.

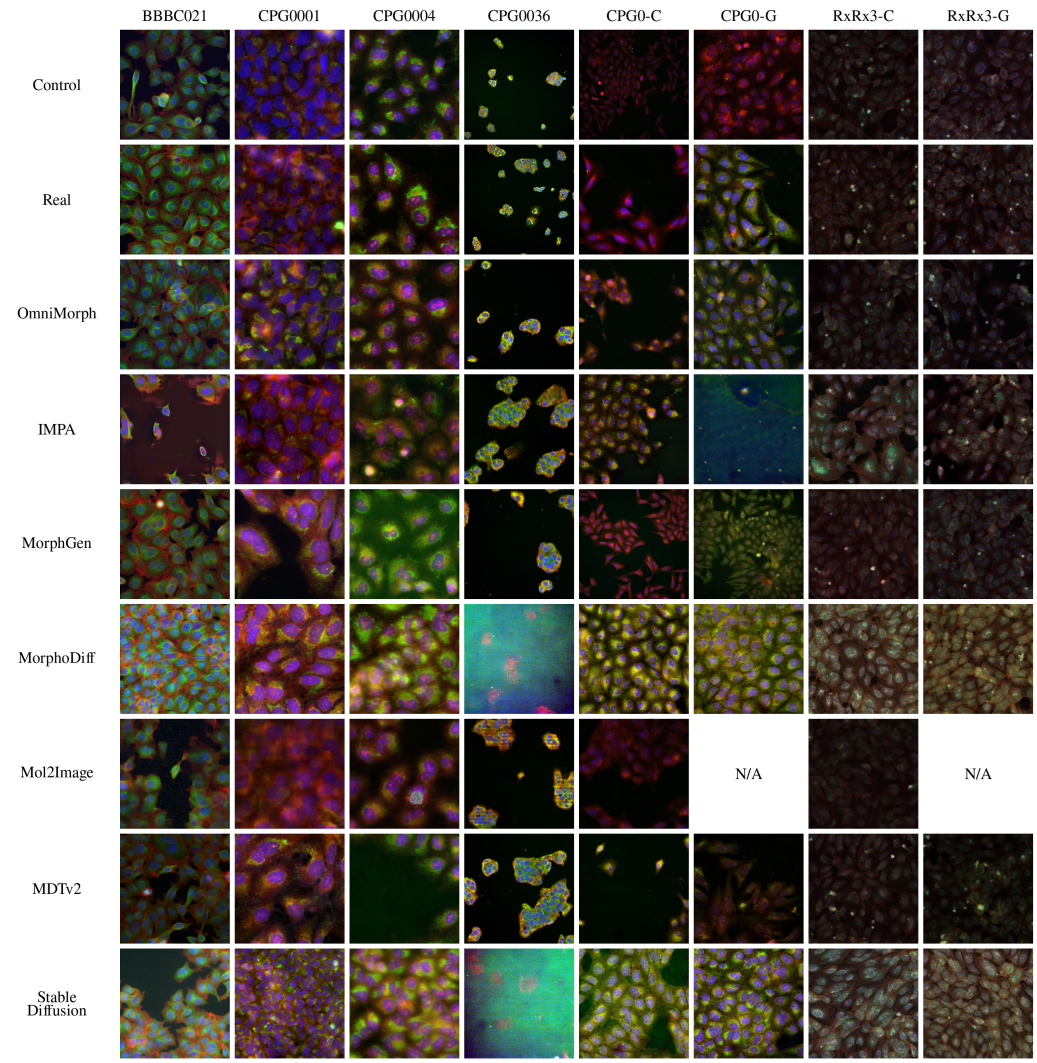

Figure 8: Perturbation prediction under in-distribution conditions. The first row is the control image from which OmniMorph and IMPA generate, the second the real image of the target; the remaining rows are the generations of the other methods for the same target.

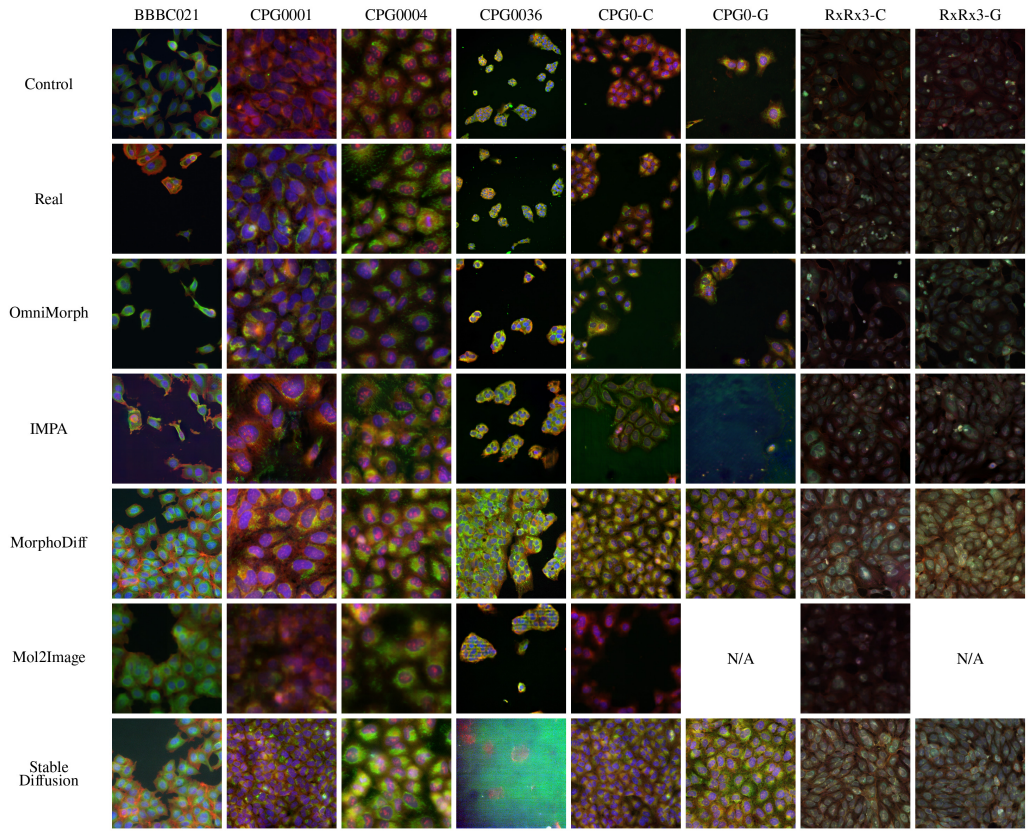

Figure 9: Perturbation prediction under unseen perturbations, laid out as Figure 8.

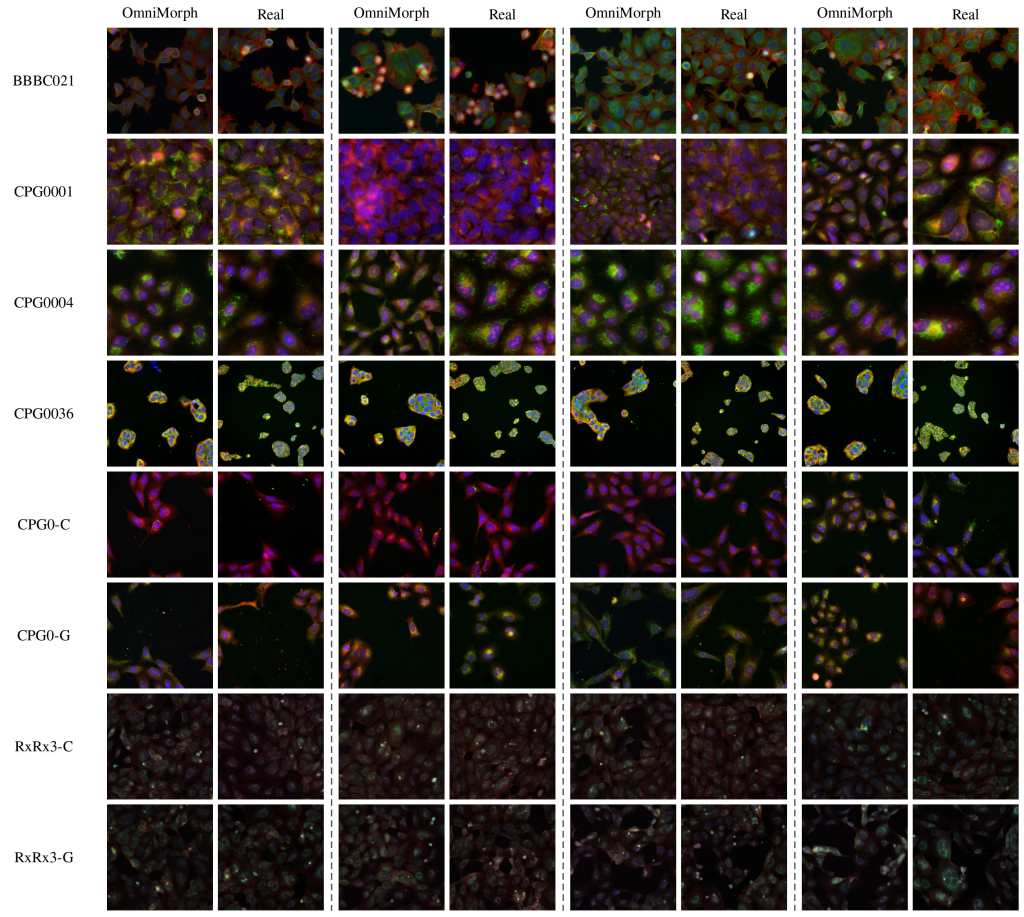

Figure 10: Further generations under in-distribution conditions. For every setting, four targets of distinct perturbations, each as a pair with the OmniMorph generation on the left and the real image of the same target on the right; dashed lines separate the pairs. All tiles use the frozen channel-to-RGB rule of the dataset.

A.2 CHANNEL COMPLETION

Channel completion (Table 3) is scored on a frozen mask bank of 2,048 masked inputs built on plates never seen in training, 512 per dataset and of three kinds: 128 observe only the DNA channel, 256 hide exactly one non-DNA channel, balanced over channels, and 128 hide a per-sample random subset of the non-DNA channels; the DNA channel is observed in every mask. The bank defines 4,523 hidden channel images in total. Every method receives the same observed channels and is scored only on the hidden, genuinely measured channels in the original $[0, 1]$ intensity range; Pearson correlation, SSIM and MAE are computed per hidden channel image and averaged. The configuration of every method is given in Table 20. Figures 11 and 12 show completions from the DNA channel alone and Figures 13 and 14 completions of one hidden channel, three held-out fields per dataset. Each field takes two rows: the observed channels, the panel assembled by each method and the measured channels on the first, and under every method the absolute difference between its composite and the measured composite on the second, as a grey map with one scale per field, so that darker is closer.

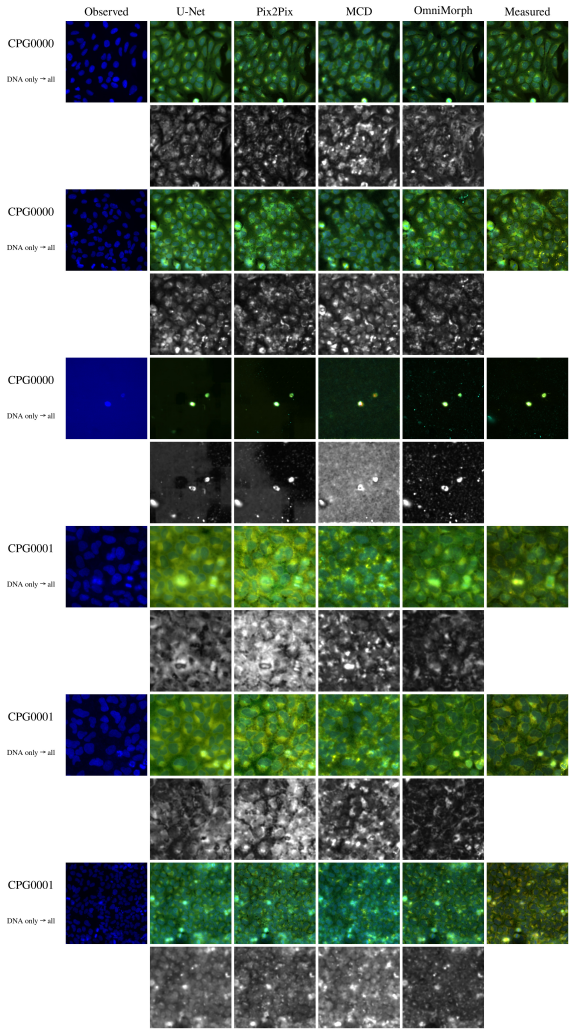

Figure 11: Channel completion from the DNA channel alone on CPG0000 and CPG0001, three held-out fields each. First row of a field: the observed DNA channel, the panel completed by each method and the measured channels, rendered with the channel-to-RGB rule of the dataset. Second row: absolute difference between each completed composite and the measured one, one grey scale per field (dark is close).

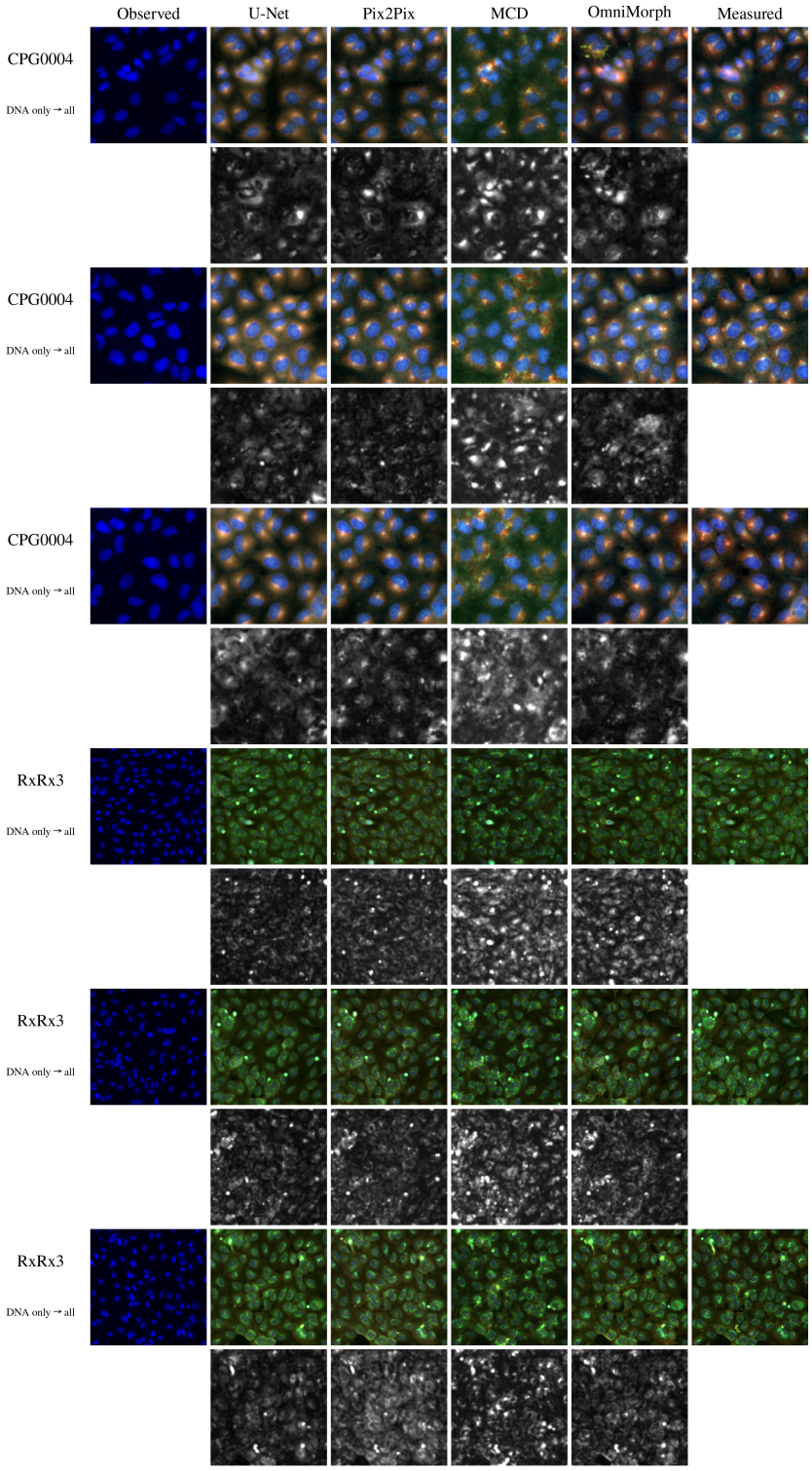

Figure 12: Channel completion from the DNA channel alone on CPG0004 and RxRx3, laid out as Figure 11.

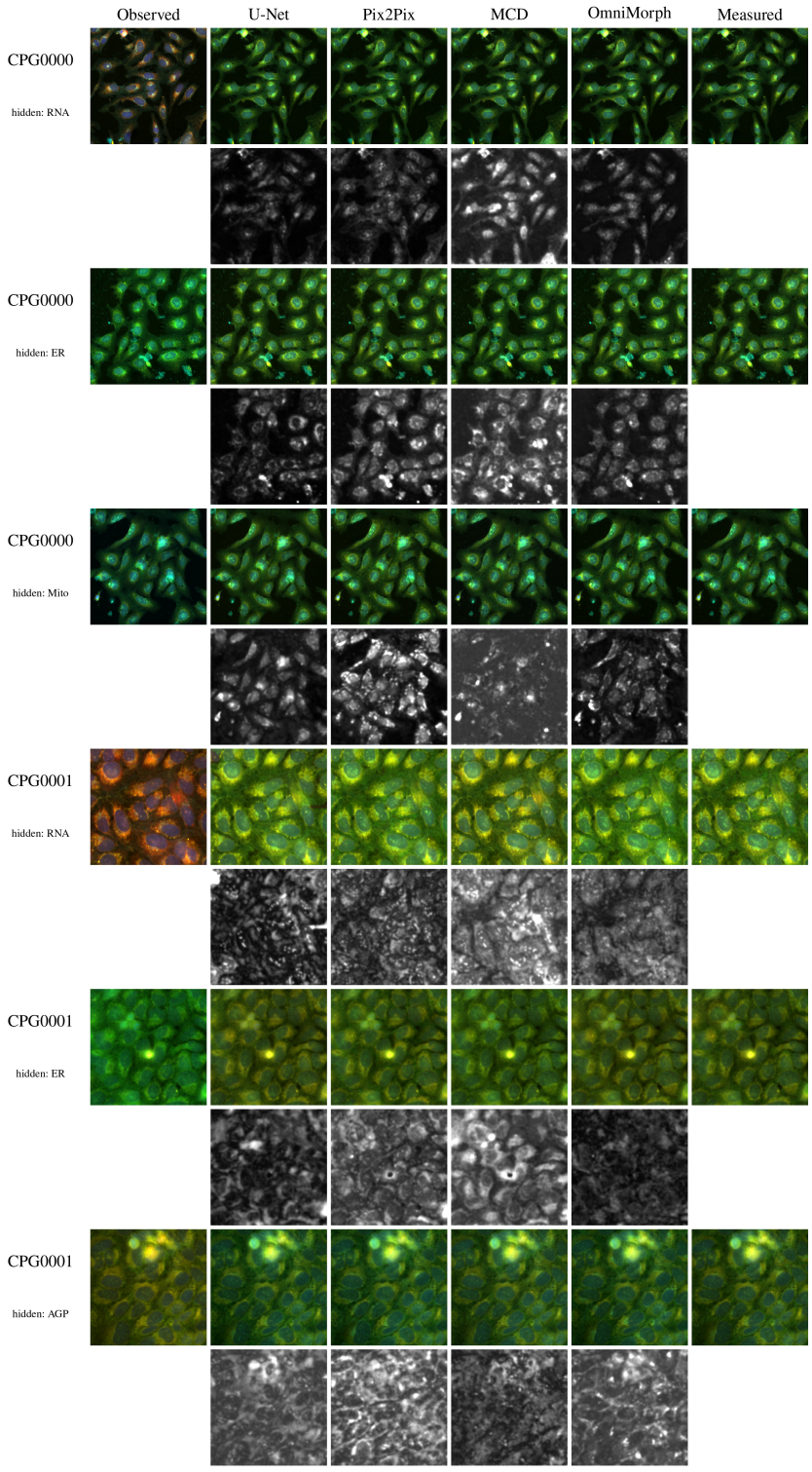

Figure 13: Channel completion with one hidden channel, named in the row label, on CPG0000 and CPG0001. The observed panel contains every other channel and each method fills in the hidden one; the composite shows the assembled panel, and the second row of each field the absolute difference to the measured composite.

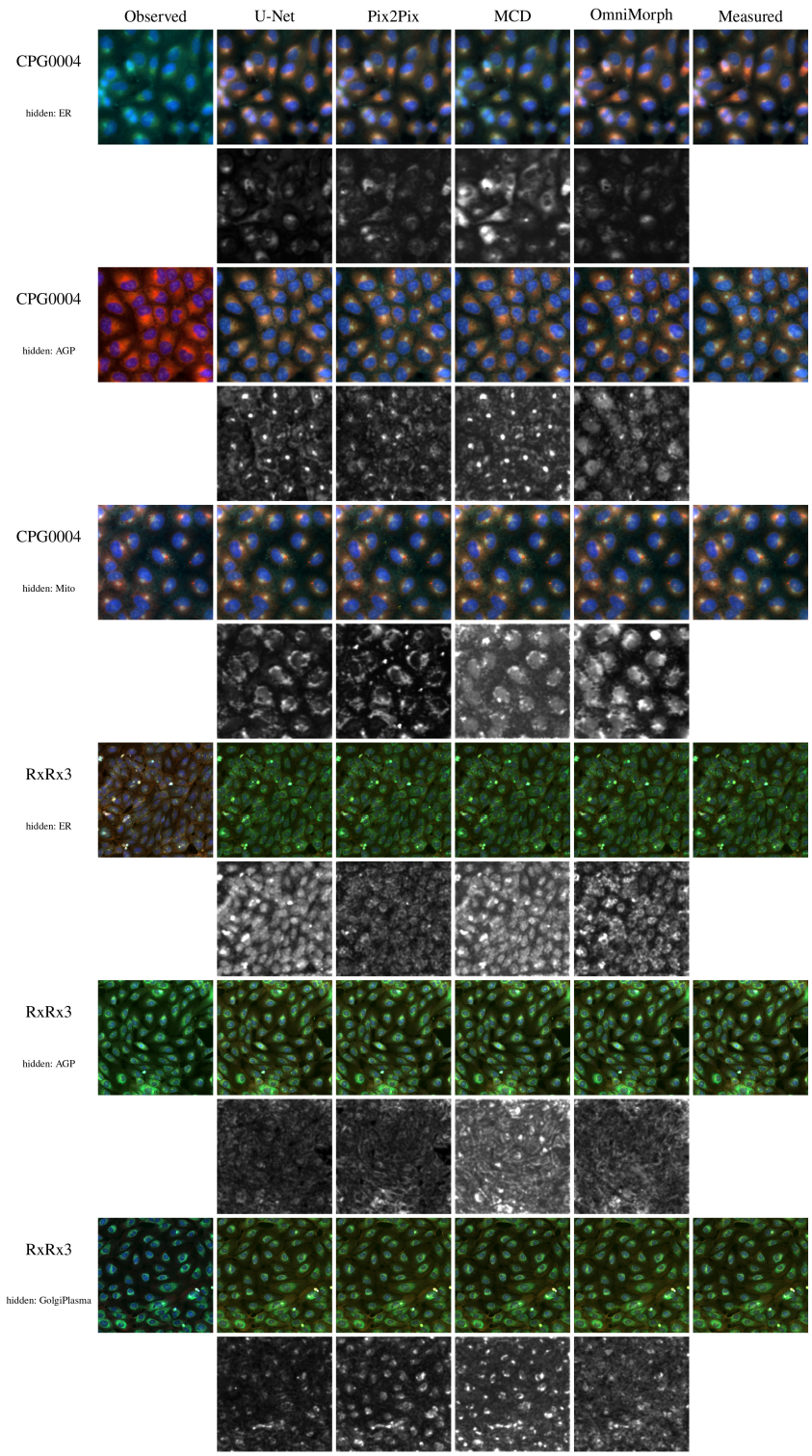

Figure 14: Channel completion with one hidden channel on CPG0004 and RxRx3, laid out as Figure 13.

A.3 BRIGHTFIELD-TO-FLUORESCENCE TRANSLATION

Brightfield-to-fluorescence translation (Table 4) is evaluated on five held-out plates, three of CPG0000 and two of CPG0001, with 512 evenly spaced test fields per dataset, identical for every method; one CPG0000 plate is excluded after an image-quality audit of its fluorescence targets, and the overall column weights the two datasets by the sizes of their remaining test pools. Brightfield inputs and fluorescence targets share one center crop without resizing, and the metrics follow the channel-completion definitions. The configuration of every method is given in Table 20. Figures 15 and 16 show four fields of each dataset with the brightfield input, the prediction of every method and the measured fluorescence, and under every prediction its absolute difference to the measured composite.

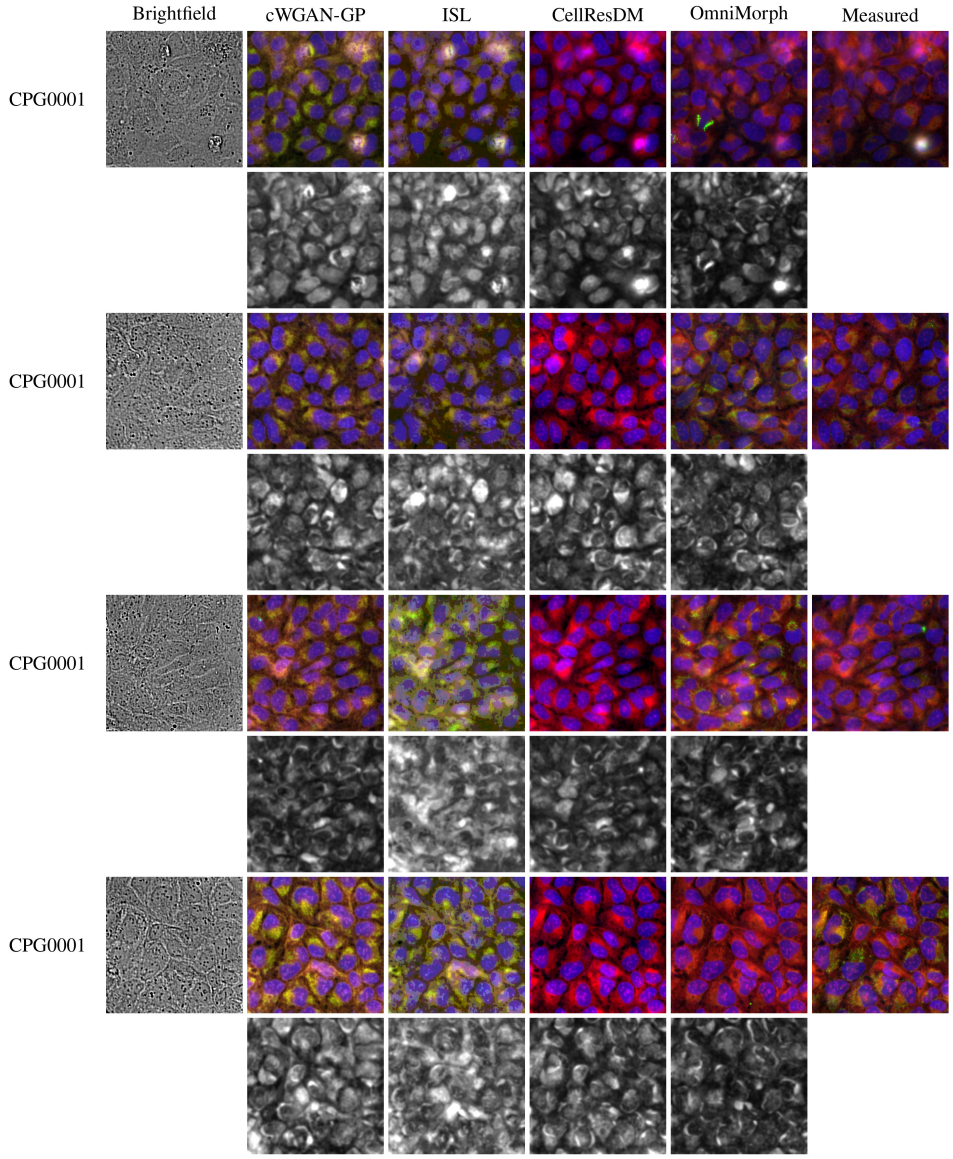

Figure 15: Brightfield-to-fluorescence translation on four held-out fields of CPG0001. First row of a field, from left: the brightfield input (CPG0001 provides a single focal plane), the prediction of each method and the measured fluorescence, all rendered with the same channel-to-RGB rule. Second row: absolute difference between each prediction and the measured composite, one grey scale per field (dark is close).

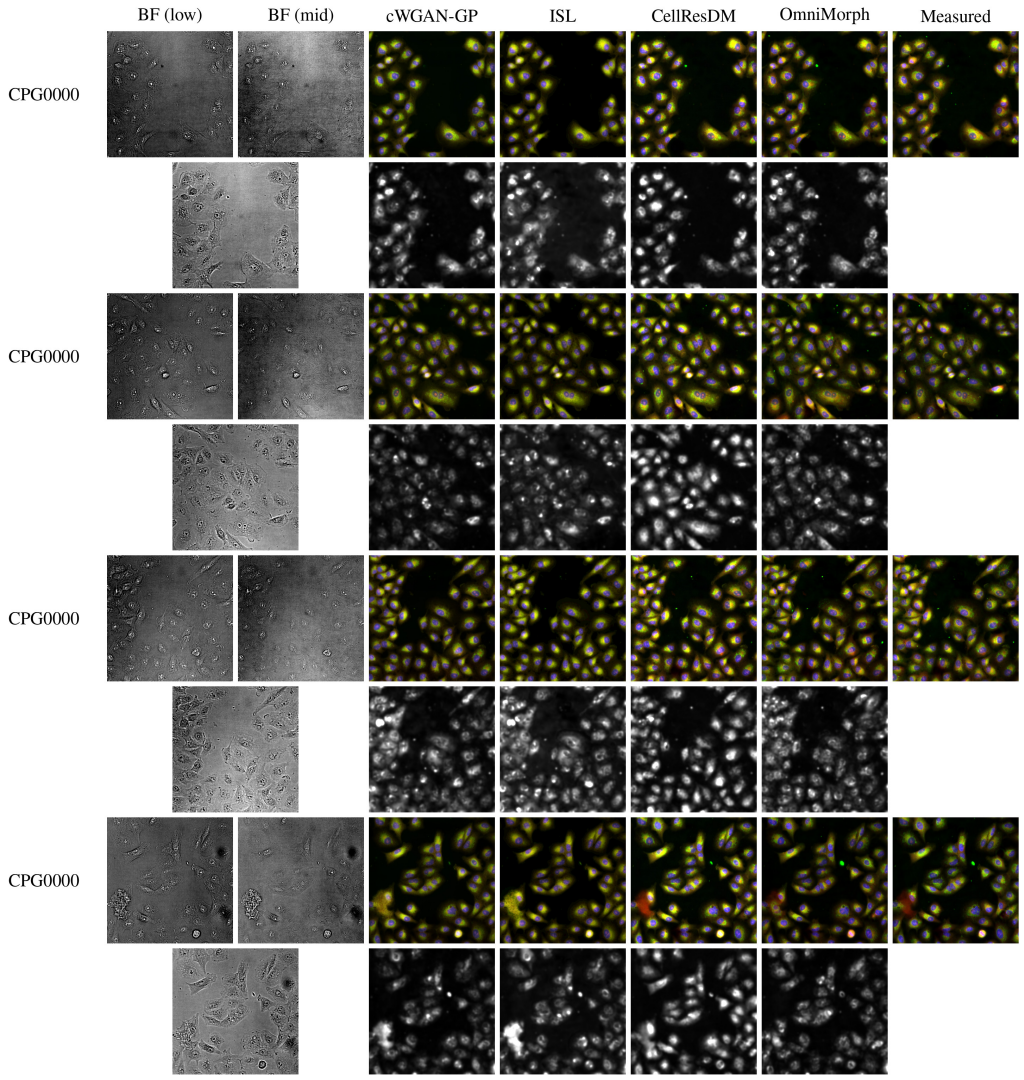

Figure 16: Brightfield-to-fluorescence translation on four held-out fields of CPG0000, laid out as Figure 15. The brightfield input has three focal planes: the low and middle planes in the first row of a field and the high plane beneath them.

A.4 Ablations

The channel-embedding ablation of Table 5(a) is scored under the full-reference protocol: each model generates 1024 images per setting from the frozen condition manifest, and the reference set is 1024 evenly spaced real images of the frozen test pool of that dataset and split. Both rows are trained with the first-stage recipe for the same number of steps; the ablated row removes the semantic channel embedding, so a channel is identified only by its position in one fixed slot layout. Table 8 gives the multi-dataset ablation of Table 5(b) as FID under the protocol of Table 1, and Table 9 the per-dataset values behind Table 5(c).

Table 10 removes the text condition. The model is retrained under the same two-stage schedule with the caption replaced by the embedding of an empty string at every step, so it receives the structured conditions and the task instruction but no natural-language description of the sample, and the result is compared with OmniMorph. Without captions the Pearson correlation of channel completion is lower by 0.04 to 0.08 on the four datasets. In morphology generation the RxRx3 settings move by at most two FID points, BBBC021 worsens by 5 to 11, CPG0001 improves and CPG0004 worsens by about ten, and the CPG0000 and CPG0036 settings collapse, with FID rising from 51–61 to 116–168 on CPG0000 and from 93 to over 200 on CPG0036: the generations for these two datasets resemble CPG0001 images, so the caption is what tells the model which dataset it is drawing from. Perturbation prediction follows the same pattern: CPG0001 and the RxRx3 compound setting improve slightly, CPG0004 worsens by four FID points, and the two CPG0000 settings rise from 42 and 60 to 114 and 94.

Table 7: Per-dataset FID↓ of the channel-embedding ablation summarised in Table 5(a): the first-stage model with and without the channel embedding, with an ID and an OOD column per dataset, under the full-reference protocol of Section A.4. **Bold** marks the better row.

	BBBC021		CPG0001		CPG0004		CPG0036		CPG0-C		CPG0-G		RxRx3-C		RxRx3-G	
Method	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Without channel embedding	137.5	122.2	191.9	210.0	125.7	132.2	198.7	214.3	176.8	187.5	122.2	115.1	132.5	132.5	120.2	126.6
OmniMorph	**30.7**	**31.0**	**46.4**	**67.3**	**33.7**	**25.4**	**83.3**	**93.5**	**101.3**	**99.2**	**59.3**	**52.4**	**19.9**	**21.8**	**19.9**	**21.0**

Table 8: FID↓ of the multi-dataset training ablation under the protocol of Table 1, with an ID and an OOD column per dataset. Both rows use the first-stage recipe, each column of the per-dataset row comes from the model trained on that dataset only, and the multi-dataset row is one model trained on all datasets. **Bold** marks the better of the two rows.

	BBBC021		CPG0001		CPG0004		CPG0036		CPG0-C		CPG0-G		RxRx3-C		RxRx3-G	
Method	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Per-dataset training	**43.8**	**27.3**	**38.1**	**25.5**	43.0	**29.7**	99.0	100.2	**36.5**	**32.6**	**56.4**	**45.7**	29.9	31.1	29.2	28.1
OmniMorph	45.2	32.3	79.2	164.4	**37.7**	36.7	**97.8**	**93.1**	52.9	51.2	61.1	50.9	**29.4**	**29.1**	**25.7**	**26.2**

A.5 Biological feature analyses

The analyses below ask whether the biological identity that Table 2 measures with OpenPhenom features survives a change of descriptor, and how real and generated images relate to each other in feature space. Every z-scoring statistic and every PCA basis is fit on real features only, and generated features are projected onto them.

A.5.1 Classification with CellProfiler features

The five panels of Table 2 are scored a second time with CellProfiler features in place of OpenPhenom. The same query images, the same real training images and the same classifier are used; only the descriptor changes. Every image is segmented and measured with one CellProfiler pipeline, and its profile is the mean, median and standard deviation of the per-object measurements of nuclei, cells and cytoplasm together with the image-level measurements. Images in which no object is found keep their row and are imputed with the medians of the real training images, so no method is scored on

Table 9: Per-dataset values of the multi-task ablation summarised in Table 5, with one block each for missing-channel completion, brightfield-to-fluorescence translation and perturbation prediction under the protocols of Table 3, Table 4 and Table 1, the last reporting FID↓ and KID↓ on in-distribution conditions, with KID in units of 10^{-3}. **Bold** marks the best entry in a column and underline the second best.

Channel completion

	CPG0000			CPG0001			CPG0004			RxRx3		
Method	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
Single-task	0.790	0.893	0.0120	0.688	0.778	0.0489	0.814	0.942	0.0097	0.755	0.914	0.0111
Pretrained + single-task	0.821	0.917	0.0103	0.744	0.787	0.0496	0.850	0.957	0.0080	0.796	0.928	0.0096
OmniMorph	**0.875**	**0.942**	**0.0086**	**0.824**	**0.823**	**0.0468**	**0.894**	**0.971**	**0.0065**	**0.859**	**0.951**	**0.0077**

Brightfield-to-fluorescence translation

	CPG0000 (3-plane)			CPG0001 (1-plane)			Overall		
Method	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
Single-task	0.918	0.853	0.0242	0.657	0.756	0.0505	0.806	0.812	0.0354
Pretrained + single-task	0.931	0.869	0.0219	0.695	0.766	0.0487	0.830	0.825	0.0334
OmniMorph	**0.954**	**0.905**	**0.0175**	**0.785**	**0.811**	**0.0442**	**0.882**	**0.865**	**0.0289**

Perturbation prediction (in-distribution conditions)

	BBBC021		CPG0001		CPG0004		CPG0-C		CPG0-G		RxRx3-C		RxRx3-G	
Method	FID	KID	FID	KID	FID	KID	FID	KID	FID	KID	FID	KID	FID	KID
Single-task	82.4	31.8	82.5	57.3	72.0	47.7	90.3	55.8	129.3	89.9	63.7	61.3	48.0	39.7
Pretrained + single-task	60.0	**5.2**	44.0	13.8	38.4	**8.6**	88.5	49.4	88.6	33.2	33.4	20.5	30.3	15.4
OmniMorph	**58.6**	8.8	**41.3**	**6.3**	**35.2**	9.2	**41.6**	**7.8**	**59.6**	**7.3**	**28.1**	**11.8**	**24.8**	**7.8**

Table 10: Text-condition ablation. The model is retrained under the same schedule with the caption replaced by the embedding of an empty string, the task instruction kept, and compared with OmniMorph: (a) channel completion under the protocol of Table 3; (b) morphology-generation FID↓ on the settings of Table 1; (c) perturbation-prediction FID↓ on the in-distribution settings of Table 1. **Bold** marks the better row.

(a) Channel completion

	CPG0000			CPG0001			CPG0004			RxRx3		
Method	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓	Pear.↑	SSIM↑	MAE↓
Without captions	0.819	0.908	0.0112	0.747	0.786	0.0500	0.852	0.958	0.0078	0.797	0.927	0.0096
OmniMorph	**0.875**	**0.942**	**0.0086**	**0.824**	**0.823**	**0.0468**	**0.894**	**0.971**	**0.0065**	**0.859**	**0.951**	**0.0077**

(b) Morphology generation, FID

Method	BBBC021	CPG1	CPG4	CPG36	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G
In-distribution conditions								
Without captions	48.7	**43.2**	42.4	209.3	168.0	127.3	29.9	27.7
OmniMorph	**43.9**	50.4	**32.7**	**93.5**	**52.9**	**61.1**	**29.5**	**25.7**
Unseen perturbations								
Without captions	43.1	**54.6**	42.4	200.6	160.8	116.4	**28.9**	28.5
OmniMorph	**32.3**	90.0	**27.0**	**93.1**	**51.2**	**50.9**	29.1	**26.3**

(c) Perturbation prediction, FID, in-distribution settings

Method	CPG1	CPG4	CPG0-C	CPG0-G	RxRx3-C	RxRx3-G
Without captions	**38.0**	39.2	113.5	94.4	**26.4**	–
OmniMorph	41.3	**35.2**	**41.6**	**59.6**	28.1	24.8

a subset it selected for itself. CellProfiler measurements are redundant, so the 6,322 raw measurements are reduced on the real training images of each panel before anything else, keeping those with less than 5% missing values, non-constant, and with a pairwise correlation below 0.9 (1,348 to 2,136 features per panel). On the RxRx3 gene panel the training library is capped at 4,021 images, drawn evenly within every gene. Table 11 puts the two descriptors side by side. With CellProfiler features OmniMorph has the highest balanced accuracy of the generators on all five panels, 0.74, 0.67, 0.30,

0.41 and 0.26, against 0.73, 0.54, 0.25, 0.41 and 0.13 for MorphGen, and MorphGen is second on every panel except CPG0004, where MDTv2 takes second place. The two descriptors differ in what they see: CellProfiler raises the scores of the three leading generators on BBBC021, CPG0001 and CPG0004 while lowering the real-image reference there (0.93, 0.71 and 0.48 against 0.95, 0.82 and 0.57), and it is much weaker on the RxRx3 gene panel, where it recovers 0.37 for the real images against 0.76 with OpenPhenom, so the gap between generated and real images is a property of the descriptor there rather than of the generations. MorphoDiff, Stable Diffusion and IMPA stay near chance under both descriptors.

Table 11: Balanced accuracy↑ of the five biology panels with CellProfiler features (CP) next to the OpenPhenom values of Table 2 (OP), RGB protocol, morphology generation. Same query images, same real training images and same classifier; only the descriptor differs. IMPA is scored on its perturbation predictions. MDTv2 on BBBC021 (72 images), MorphoDiff on CPG0001 (558) and IMPA on the RxRx3 gene panel (857) are scored on the subsets of the panel they generated, as in Table 2. **Bold** marks the best method per column and underline the second best; the real-image row is a reference and is not ranked.

	BBBC021 (MoA, 6)		CPG0001 (MoA, 10)		CPG0004 (cpd., 10)		RxRx3-C (cpd., 15)		RxRx3-G (gene, 15)	
Method	OP	CP	OP	CP	OP	CP	OP	CP	OP	CP
Real images (ref.)	0.954	0.934	0.815	0.705	0.571	0.483	0.620	0.517	0.760	0.370
IMPA	0.164	0.209	0.135	0.202	0.211	0.089	0.126	0.051	0.074	0.101
Stable Diffusion	0.157	0.189	0.088	0.088	0.083	0.073	0.065	0.059	0.081	0.046
MorphoDiff	0.253	0.119	0.092	0.153	0.108	0.163	0.087	0.051	0.085	0.040
MDTv2	0.250	0.296	0.299	0.368	0.175	0.256	0.237	0.278	0.242	0.189
MorphGen	0.670	0.731	0.532	0.537	0.236	0.245	0.366	0.412	0.304	0.125
OmniMorph	**0.703**	**0.740**	**0.649**	**0.671**	**0.247**	**0.299**	**0.476**	**0.414**	**0.526**	**0.260**

A.5.2 Generated and real images of single conditions

Figure 17 looks at one condition at a time, one point per image, for twelve conditions whose real images form a clear cluster in this space: four compounds of BBBC021, four of CPG0001, and two compounds and two gene knockouts of RxRx3. Every real image of the condition in the panel gallery is drawn together with every generation of OmniMorph for it: the generations for the panel's query conditions, further seeds for the same conditions, and generations for the conditions of the training images, so that the generated set is comparable in size to the real one. Where the dataset has negative-control images with features, an equal number of controls is drawn as well. The axes are the two leading components of the real images of the condition and the controls, fit on real data only. For each condition we count the fraction of generations whose nearest real image is an image of the condition rather than a control, and compare it with the same fraction for the real test images of the condition, which is what a perfect generator would reach. Real sets have 87 to 700 images per condition and generated sets 87 to 503. The generations reach the real-test fraction for nine of the twelve conditions, cytochalasin D and RPS11 stay within 0.1 of it, and MLN9708 falls short (0.82 against 1.00), with part of its generations nearer to the other conditions of the panel than to the compound.

Two questions are asked of the feature space, on the four or five classes of each of four panels whose real images separate best (chosen on real images only by the silhouette of each class): do the generations of a condition fall on the real images of that condition, and do the generations of different conditions separate from each other. Figure 18 answers both at the level of single images. Two linear discriminant axes are fit on the real training images of the chosen classes and every image is projected on them; generations are only projected; the generations are those for the panel's query conditions, two further seeds for the same conditions and generations for the training conditions of the drawn classes, so that each class carries several times as many generated as real test images. The top row overlays the generations on the real images, training images faded and test images filled: the fraction of generations whose nearest real image belongs to the same class is 0.68, 0.59, 0.74, 0.98 on BBBC021, CPG0001, the RxRx3 compound panel and the RxRx3 gene panel. The bottom row shows the generations alone, with a silhouette by class of 0.33, 0.27, 0.30, 0.63, against 0.56, 0.37, 0.44, 0.48 for the real test images on the same axes: the generated conditions separate in

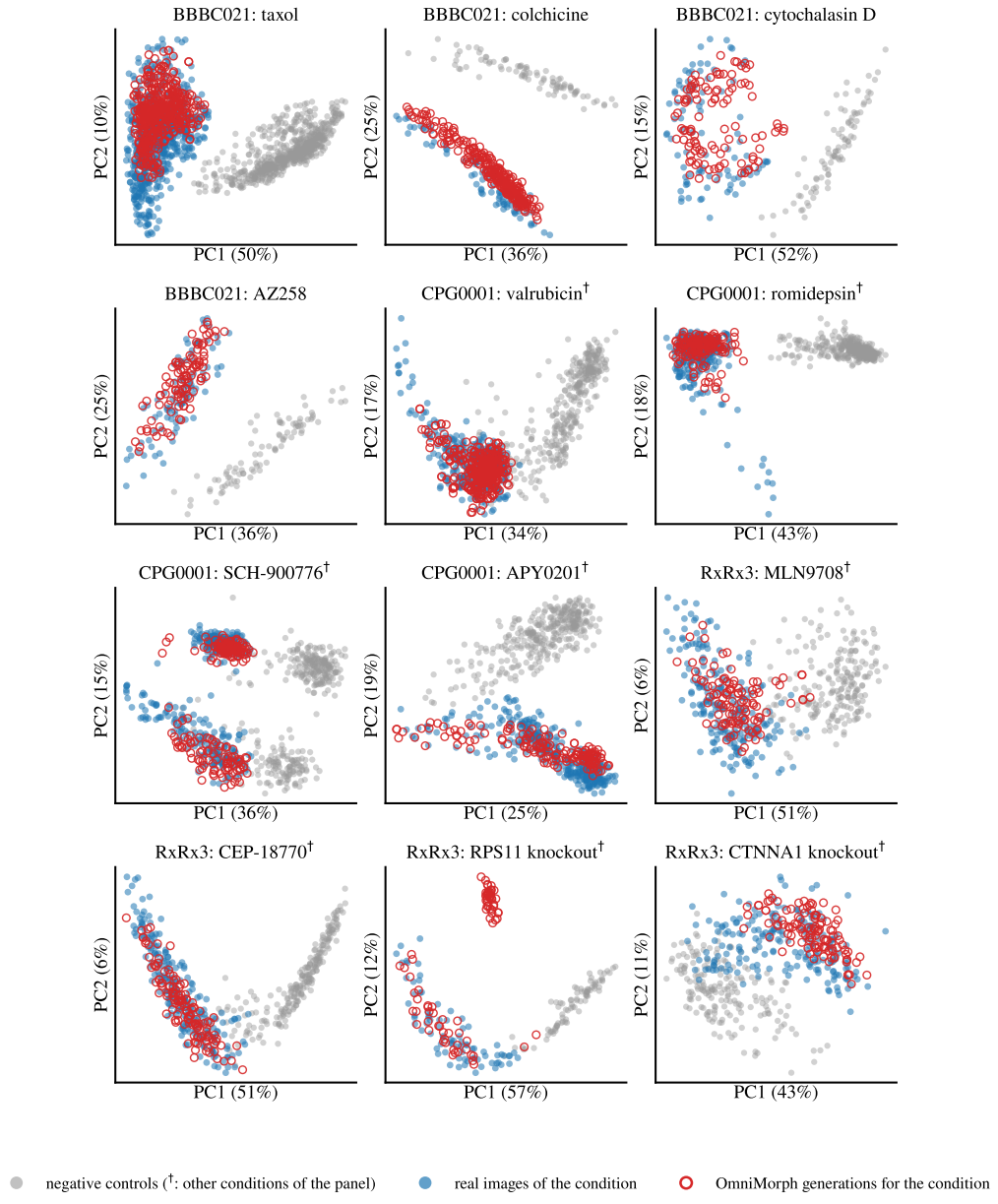

Figure 17: Single conditions in OpenPhenom space, one point per image. Real images of the condition (blue), OmniMorph generations for the condition (red, hollow), the RxRx3 knockouts marked as such in the panel titles and, where available, an equal number of negative-control images of the dataset (grey), on the two leading components of the real images and controls. † marks datasets without control images, where the other conditions of the panel stand in as the background. Points beyond 2.5σ of their group centre (at most a tenth of a group) are omitted.

the same arrangement as the real ones, with wider clouds. CPG0004 is not shown: its single-dose compounds are the closest to each other and neither real nor generated images form separate groups there. Points farther than 2σ from their class centre, at most fifteen percent of a class, are not drawn on either side. The same analysis with CellProfiler features gives the same ranking of the panels with weaker separation, and is not shown.

Figure 19 quantifies the agreement condition by condition. The mean z-scored profile of the real images of a condition is correlated with the mean profile of a method's generations for the same condition, and the mean Spearman correlation over the conditions of a panel is compared with two references: the correlation between two random halves of the real images, which bounds what any generator can reach, and the correlation with the generations of another condition, which is what a generator earns without encoding the condition. OmniMorph reaches 0.50, 0.60, 0.56, 0.66 and 0.59 on the five panels against upper bounds of 0.76, 0.86, 0.69, 0.70 and 0.70. It leads on four panels and is within 0.03 of MDTv2 and MorphGen on CPG0004, where the single-dose compounds resemble each other and even a mismatched condition correlates at 0.34. MorphoDiff, Stable Diffusion and IMPA stay near the mismatched reference on every panel, so their generations carry little condition-specific profile. We do not report the agreement between k-means clusters and labels that is sometimes used for this purpose: it rewards a generator that collapses every condition onto one mode, and on our panels it ranked Mol2Image and IMPA highest for that reason.

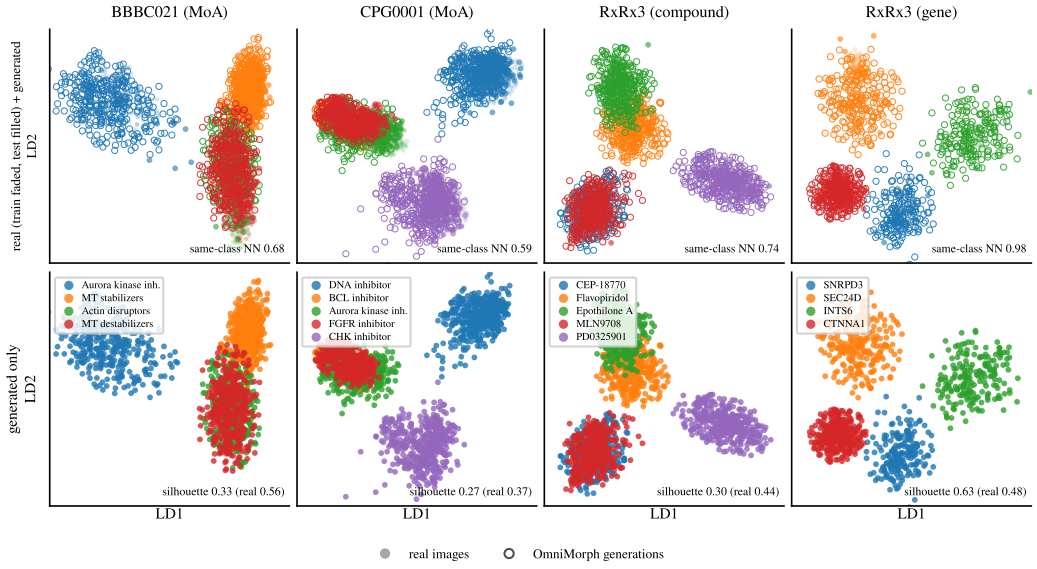

Figure 18: Coverage and separation in feature space (OpenPhenom, RGB protocol), one point per image. For each panel, the four or five classes whose real images separate best, on two linear discriminant axes fit on the real training images of these classes; generations are only projected. Top: real training images (faded), real test images (filled) and OmniMorph generations for the same conditions (hollow), with the fraction of generations whose nearest real image has the same class. Bottom: the generations alone, coloured by class, with their silhouette by class and that of the real test images on the same axes. Points beyond 2σ of their class centre (at most fifteen percent of a class) are omitted.

A.5.3 Dose response

Dose is part of the condition of OmniMorph, and none of the compared generators takes it as input, so this analysis has no baseline. For the fourteen compounds of CPG0004 and the sixteen of RxRx3 whose real images show the clearest dose response, ranked on real images alone by how monotonically the distance from the lowest dose grows with dose, OmniMorph generates images at every dose of the compound from the conditions of the real images, 7 to 42 per dose. In the z-scored OpenPhenom space, the centroids of the real images of the doses are joined in dose order, and so are the centroids of the generated images; the cosine between the real and the generated tra-

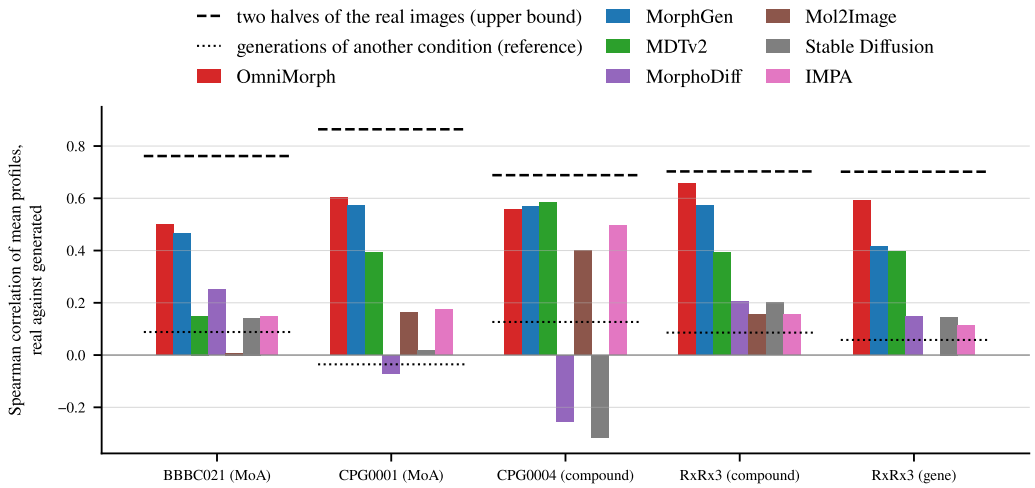

Figure 19: Condition-level profile agreement. Spearman correlation between the mean z-scored OpenPhenom profile of the real images of a condition and that of a method's generations for the same condition, averaged over the conditions of each panel. Dashed: correlation between two random halves of the real images (upper bound). Dotted: correlation with the generations of a different condition (specificity reference).

jectory of centroid displacements measures whether the two move the same way. Figure 20 shows the generated images of the three compounds with the clearest generated dose gradient, MLN9708 and Flavopiridol on RxRx3 and oprozomib on CPG0004 (trajectory cosines 0.80, 0.73 and 0.42), coloured by dose on the axes of the real images: the doses order along the leading component. Vincristine also follows the real direction (cosine 0.68), but its generated centroids move too little for the doses to separate. The generated displacement is smaller than the real one, a third of the real span at the highest dose for the median compound and between a third and nine tenths for the four shown, and for the compounds with weaker real responses the generations stay close to the low-dose centroid, so the dose direction is recovered for the strongest responders only: three of sixteen compounds on RxRx3 and two of fourteen on CPG0004 reach a trajectory cosine above 0.3, and the second of these, CYT-997, has no monotonic real response to follow.

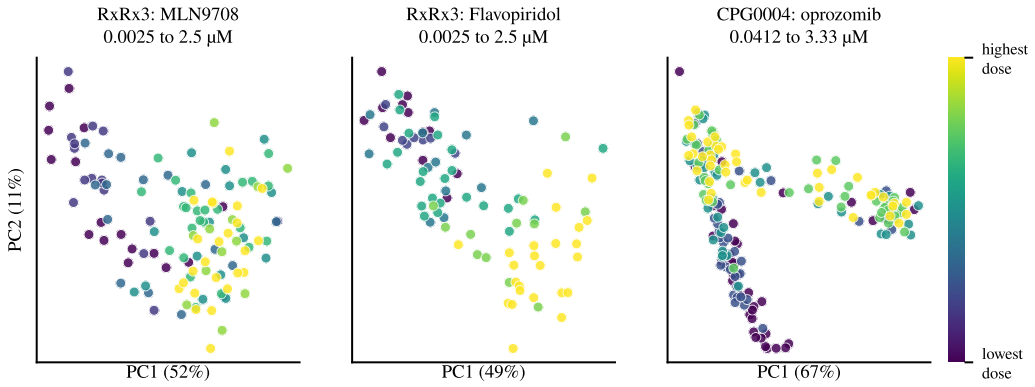

Figure 20: Dose response of OmniMorph: generated images of the three compounds with the clearest generated dose gradient, coloured by dose from the lowest (purple) to the highest (yellow), on the two leading components of the real images of the compound (fit on real images only, which are not drawn). The dose range of each compound is given in the panel title.

A.5.4 EFFECT DIRECTION AND MAGNITUDE AGAINST NEGATIVE CONTROLS

Image-distribution metrics and the classification panels do not test whether a generated image moves away from the untreated state in the right direction and by the right amount. Table 12 tests this on the five biology panels against the negative controls of the panel plates: the DMSO wells of BBBC021, the negative-control wells of CPG0001 and CPG0004, and the empty and CRISPR-control wells of the RxRx3 plates. For every condition, the real effect is the difference between the mean feature of its real images and the mean feature of the controls, the generated effect is the same difference with a method's generations, and the two are compared by their cosine and by the ratio of their magnitudes. Three references bound the numbers. Two halves of the real images give the ceiling. Pairing the real effect of a condition with the generated effect of another condition gives what a method earns without using the perturbation, because every treatment shares a common displacement from the controls; this shared part alone yields a cosine of 0.78 on BBBC021 and 0.35 to 0.51 on CPG0001 and the RxRx3 panels, so the specific part of the agreement is the margin above it. A method that returns a control image scores the cosine between two halves of the controls, between -0.17 and 0.25.

OmniMorph produces effects of the real magnitude on every panel (ratios 0.96 to 1.07) and reaches the ceiling on BBBC021 (0.93 against 0.94). Its margin over the shuffled reference is the largest of all methods on BBBC021 (0.15), the RxRx3 compound panel (0.15 from the condition, 0.21 from the control image) and the RxRx3 gene panel (0.34), and on CPG0004 it is 0.40 against a shuffled reference of zero. On CPG0001 and CPG0004 MorphGen is slightly higher (0.75 against 0.72 and 0.48 against 0.40), with a smaller margin on CPG0001 because its shuffled reference is higher too. Starting from the control image (perturbation prediction) gives a higher cosine than generating from the condition alone on BBBC021, CPG0004 and the RxRx3 compound panel, the same on CPG0001 (0.71 against 0.72) and a lower one on the RxRx3 gene panel. MorphoDiff and Stable Diffusion do not separate matched from shuffled conditions and on CPG0001, CPG0004 and the RxRx3 panels their effects are two to nine times larger than the real one and, except on CPG0004, point away from it, the signature of images that are far from the plate rather than perturbed. Mol2Image and MDTv2 follow the real direction with a smaller margin and, for Mol2Image, a doubled magnitude. IMPA, which edits the control image, keeps the real magnitude except on CPG0004, and its direction is within 0.06 of the shuffled reference on every panel but CPG0004 (0.17 against 0.01). A fourth control, replacing the reference control image of perturbation prediction by one from another plate, requires new generations and is left for a revision.

Table 12: Perturbation effect against negative controls on the biology panels of Table 2. For every condition of a panel, the real effect is the mean OpenPhenom feature of its real images minus the mean feature of the negative-control images of the panel plates (DMSO wells on BBBC021, negative-control wells on CPG0001 and CPG0004, empty and CRISPR-control wells on RxRx3), and the generated effect is the same difference with a method's generations for that condition. *cos*: cosine between the two effects, mean over conditions. *shuf.*: the same cosine when the real effect of a condition is paired with the generated effect of another condition (mean over the other conditions), which is what a method earns without using the perturbation. $|\Delta_{\text{gen}}|/|\Delta_{\text{real}}|$: median ratio of the effect magnitudes. *Real, two halves* compares the effect of one half of the real images with that of the other half, and *Control returned* is a method that outputs a control image, computed as the effect between two halves of the controls. **Bold** marks the best method per column within a block and underline the second best (for the magnitude ratio, the value closest to 1); the shuffled cosine is a chance level and is not ranked.

	BBBC021 (14 compounds)			CPG0001 (16 compounds)			CPG0004 (10 compounds)			RxRx3-C (15 compounds)			RxRx3-G (15 genes)		
Method	cos↑	shuf.	$\lvert\Delta_{\text{gen}}\rvert/\lvert\Delta_{\text{real}}\rvert$	cos↑	shuf.	$\lvert\Delta_{\text{gen}}\rvert/\lvert\Delta_{\text{real}}\rvert$	cos↑	shuf.	$\lvert\Delta_{\text{gen}}\rvert/\lvert\Delta_{\text{real}}\rvert$	cos↑	shuf.	$\lvert\Delta_{\text{gen}}\rvert/\lvert\Delta_{\text{real}}\rvert$	cos↑	shuf.	$\lvert\Delta_{\text{gen}}\rvert/\lvert\Delta_{\text{real}}\rvert$
Morphology generation															
Stable Diffusion	0.18	0.18	1.91	-0.14	-0.14	3.01	0.29	0.29	9.12	-0.59	-0.59	2.38	-0.19	-0.19	3.12
Mol2Image	0.46	0.42	2.56	0.40	0.31	2.28	0.45	0.08	2.19	0.47	0.37	2.31	–	–	–
MorphoDiff	0.04	0.02	1.59	-0.18	-0.19	5.81	0.31	0.29	9.20	-0.57	-0.58	2.49	-0.19	-0.19	3.32
MDTv2	0.73	0.66	1.12	0.49	0.34	**1.00**	0.42	0.11	1.69	0.38	0.27	1.32	0.60	0.26	1.32
MorphGen	0.90	0.76	1.08	**0.75**	0.54	1.06	**0.48**	-0.02	1.11	0.59	0.41	0.90	0.51	0.27	1.10
OmniMorph	**0.93**	0.78	**1.07**	0.72	0.47	1.02	0.40	0.00	**0.96**	**0.66**	0.51	**1.02**	**0.69**	0.34	**1.01**
Perturbation prediction															
IMPA	0.68	0.66	**0.98**	0.49	0.43	1.21	0.17	0.01	2.14	0.41	0.41	1.10	0.42	0.40	**1.05**
OmniMorph	**0.94**	0.81	1.03	**0.71**	0.53	**1.17**	**0.44**	0.00	**1.00**	**0.73**	0.52	**1.04**	**0.52**	0.33	1.08
References															
Control returned	0.25	–	0.00	0.05	–	0.00	0.15	–	0.00	-0.15	–	0.00	-0.17	–	0.00
Real, two halves	0.94	–	1.05	0.90	–	1.00	0.64	–	1.27	0.75	–	0.91	0.84	–	0.98

A.5.5 PER-CHANNEL CORRELATION AND FID

The native-channel protocol of Table 2 shows that the generated channels carry more identity than their RGB rendering, and Table 13 gives the per-channel version of Table 3 and Table 4: in channel completion OmniMorph has the highest Pearson correlation on every hidden channel of the three CPG datasets and trails U-Net by 0.01 to 0.03 on the five channels of RxRx3, and in brightfield-to-fluorescence translation it leads on every channel of both datasets, with the largest margin on the DNA channel of CPG0000. Table 14 scores each native channel on its own, as an FID between the real query images of a biology panel and the generations for the same conditions after the per-image scaling of the RGB protocol. Like the other analyses of this section it uses the image sets of Table 2, the same images as the biological readout, and not the frozen target lists of Table 1, so its values are not comparable with that table. OmniMorph has the lower FID on every channel of BBBC021, CPG0001 and the two RxRx3 panels, by a wide margin on the RxRx3 gene panel (15 to 23 against 55 to 72), and trails MorphGen on four of the five channels of CPG0004, whose 112 images make its estimate the noisiest of the five.

Table 13: Per-channel Pearson correlation of channel completion (on the hidden channels of the frozen mask bank; the DNA channel is always observed) and of brightfield-to-fluorescence translation (on the held-out fields), for the runs of Table 3 and Table 4; the CellResDM evaluator reports no per-channel values. **Bold** marks the best method per channel and underline the second best.

Channel completion

	CPG0000				CPG0001			
Method	RNA	ER	AGP	Mito	RNA	ER	AGP	Mito
MCD	0.882	0.740	0.672	0.752	0.759	0.669	0.585	0.589
Pix2Pix	0.918	0.857	0.770	0.704	0.805	0.686	0.677	0.645
U-Net	0.934	0.894	0.794	0.763	0.848	0.777	0.772	0.692
OmniMorph	**0.941**	**0.907**	**0.826**	**0.826**	**0.900**	**0.849**	**0.808**	**0.742**

	CPG0004				RxRx3				
Method	RNA	ER	AGP	Mito	RNA	ER	AGP	Mito	Golgi
MCD	0.856	0.808	0.662	0.671	0.792	0.774	0.512	0.697	0.707
Pix2Pix	0.886	0.882	0.860	0.726	0.911	0.888	0.634	0.800	0.844
U-Net	0.900	0.908	0.892	0.805	**0.941**	**0.925**	**0.751**	**0.860**	**0.892**
OmniMorph	**0.920**	**0.926**	**0.908**	**0.821**	0.928	0.909	0.725	0.850	0.880

Brightfield-to-fluorescence translation

	CPG0000					CPG0001				
Method	DNA	RNA	ER	AGP	Mito	DNA	RNA	ER	AGP	Mito
ISL	0.848	0.922	0.903	0.871	0.876	0.726	0.767	0.692	0.669	0.645
cWGAN-GP	0.890	0.937	0.923	0.889	0.900	0.778	0.810	0.747	0.731	0.716
OmniMorph	**0.949**	**0.974**	**0.961**	**0.933**	**0.952**	**0.805**	**0.849**	**0.780**	**0.749**	**0.743**

Table 14: FID↓ per native channel. For every biology panel of Table 2, its real query images and the native-channel generations of OmniMorph and of MorphGen for the same conditions are compared one fluorescence channel at a time (these are the panel image sets, not the 512-target lists of Table 1); each plane is scaled per image to $[0, 255]$ as in the RGB protocol, replicated to three channels and passed through the same Inception network as Table 1. **Bold** marks the lower FID.

	BBBC021			CPG0001					CPG0004				
Method	DNA	Tubulin	AGP	DNA	RNA	ER	AGP	Mito	DNA	RNA	ER	AGP	Mito
MorphGen	42.4	57.4	52.7	36.9	47.4	42.7	55.1	30.6	**103.7**	**118.9**	**73.0**	79.5	**86.5**
OmniMorph	**34.8**	**50.0**	**43.3**	**32.7**	**44.1**	**39.0**	**49.1**	**26.4**	112.0	125.6	79.0	**77.9**	104.7

	RxRx3-C						RxRx3-G					
Method	DNA	ER	AGP	RNA	Mito	Golgi	DNA	ER	AGP	RNA	Mito	Golgi
MorphGen	46.1	38.3	82.1	46.6	66.8	67.4	54.9	63.8	69.0	68.7	71.5	67.9
OmniMorph	**27.4**	**30.4**	**68.9**	**40.7**	**54.5**	**56.2**	**15.3**	**14.9**	**21.6**	**20.2**	**22.8**	**21.0**

A.6 TEXT TEMPLATES

The text condition is the task instruction followed by the biological context, and the templates of both together with a complete example are given in Table 15. The cell line is written as MCF-7 breast cancer cells, HUVEC endothelial cells, U-2 OS osteosarcoma cells, A549 lung cancer cells or HepG2 hepatocellular carcinoma cells. SMILES strings longer than 200 characters and protein sequences longer than 256 are truncated and end with `...`, doses are printed with `%.6g`, and the unconditional branch uses the text `A cell painting image.`

Table 15: Templates and an example of the text condition. (a) The task instruction depends on the task, and each $[\cdot]$ is filled with channel names, except the first one in the brightfield instruction, which lists the focal planes as low, middle or high. (b) The biological context depends on the perturbation type and is shared by the four tasks. (c) The complete text of one missing-channel completion input with a perturbation taken from CPG0004, formed by (a) followed by (b).

(a) Task instruction

Morphology generation	`Generate fluorescence channels [·] from noise.`
Channel completion	`Observed fluorescence channels [·]. Generate missing fluorescence channels [·].`
Brightfield-to-fluorescence	`Observed brightfield focal planes [·]. Generate fluorescence channels [·].`
Perturbation prediction	`Observed source-state fluorescence channels [·]. Generate target-state fluorescence channels [·] under the stated perturbation.`

(b) Biological context

Compound	`A cell painting image of {cell line} treated with compound, SMILES: {SMILES}, at {dose} µM.`
Gene knockout	`A cell painting image of {cell line} treated with CRISPR knockout, targeting gene: {gene}, protein sequence: {sequence}.`
Control	`A cell painting image of {cell line}, no perturbation applied, negative control.`

(c) Complete example

```
Observed fluorescence channels [DNA/nucleus, mitochondria]. Generate
missing fluorescence channels [RNA/nucleoli, endoplasmic reticulum,
actin/AGP]. A cell painting image of A549 lung cancer cells treated with
compound, SMILES: O=c1[nH]cc(F)c(=O)[nH]1, at 3.33 µM.
```

A.7 DATA SOURCES

This study uses six publicly available high-content microscopy datasets whose fluorescence panels range from three to six channels, two of which provide additional brightfield focal planes. Table 16 summarizes the imaging panel, the cell context, the perturbation types, and the number of training images of each dataset. The six datasets originate from different laboratories and acquisition protocols and differ in channel composition, and this heterogeneity is the central difficulty that a unified model must address. The subset of each source is defined by a frozen plate list that is released with the splits. Plates were selected before training for complete perturbation, dose and plate metadata and to cover the perturbation panels used in evaluation, and the held-out plates of Section A.10 were removed from every training pool.

Table 16: Overview of the datasets. The fluorescence count excludes brightfield images, BF denotes the number of brightfield focal planes, and the last column gives the number of 512×512 fields of view in the training set.

Dataset	Fluor.	BF	Cell line	Perturbations	Train images
BBBC021	3	0	MCF-7	Compounds	7,457
CPG0000	5	3	U-2 OS/A549	Compounds and genes	26,757
CPG0001	5	1	U-2 OS	Compounds	18,163
CPG0004	5	0	A549	Compounds with a dose series	19,630
CPG0036	4	0	HepG2	Compounds	9,405
RxRx3	6	0	HUVEC	Compounds and genes	173,409
Total					254,821

A.8 CONTROLLED CHANNEL VOCABULARY

The datasets stain the same subcellular structures with different dyes and label the resulting channels differently. Dataset-specific channel labels are therefore mapped to a controlled biological vocabulary before they enter the shared model. The vocabulary contains seven fluorescence semantics: nucleus/DNA, nucleoli/RNA, endoplasmic reticulum, actin/AGP-like cytoskeleton, Golgi/plasma-membrane-like structures, mitochondria, and microtubules/tubulin. The mapping aligns the biological content that a channel marks while retaining the source identity, and a shared label does not imply that stains or acquisition protocols from different studies are experimentally identical.

The vocabulary makes differences in panel composition explicit. The three channels of BBBC021 mark the nucleus, microtubules, and actin, and must not be interpreted as the first three channels of a standard five-channel panel, and CPG0036 is a genuine four-channel panel that lacks an RNA channel rather than a five-channel sample with an unobserved RNA target. Table 17 lists the semantics measured in each dataset.

Table 17: Biological channel semantics in each dataset. A check mark denotes a measured fluorescence channel; a dash denotes that the semantic type is not part of that dataset's panel.

Dataset	Nucleus	RNA	ER	Actin/AGP	Golgi/PM	Mito	Tubulin
BBBC021	✓	–	–	✓	–	–	✓
CPG0000	✓	✓	✓	✓	–	✓	–
CPG0001	✓	✓	✓	✓	–	✓	–
CPG0004	✓	✓	✓	✓	–	✓	–
CPG0036	✓	–	✓	✓	–	✓	–
RxRx3	✓	✓	✓	✓	✓	✓	–

Each channel position records a semantic type identifier and a presence mask, so channel meaning is determined by metadata rather than by tensor position. The model orders the channels that actually exist, masks absent positions from attention and loss, and restores its outputs to the channel order of the dataset or of the user request.

A.9 PREPROCESSING AND QUALITY CONTROL

Raw images are converted into 512×512 single-channel fields of view, and each fluorescence channel is linearly rescaled to the $[0, 1]$ intensity range and is stored and encoded independently rather than merged into an RGB image during training. Brightfield-to-fluorescence translation requires the brightfield and fluorescence images to cover the same physical field, so both modalities undergo the same geometric transform, a center crop without resizing. Every field of view passes a quality screening before entering training or evaluation, which discards out-of-focus, saturated, or nearly empty fields and requires complete perturbation, dose, and plate metadata. Each control–perturbed image pair of perturbation prediction is drawn from the same experimental batch, and no pixel-level alignment between the two images is assumed.

A.10 DATA SPLITS AND LEAKAGE CONTROL

All tasks share one set of data splits, and the test sets are frozen at the level of images, conditions, and plates. Morphology generation and perturbation prediction test condition generalization and are evaluated both on conditions seen during training (ID) and on held-out compound or gene identities (OOD). Channel completion and brightfield-to-fluorescence translation test plate generalization, so their test images come entirely from plates unseen in training; brightfield-to-fluorescence translation holds out five plates, three of CPG0000 and two of CPG0001, and its test set contains 14,285 spatially aligned examples with no plate overlap with the 37,248 training examples, and one of the three CPG0000 plates is excluded from the evaluation after the image-quality audit of Section A.3. The splits enforce one global rule: an image included in any test set is removed from the training data of every task, so that a held-out target of one task cannot re-enter the model through the training pool of another, and a zero-overlap audit of all protected samples is performed before training starts. Table 18 lists which datasets each task trains on and which it is evaluated on. The frozen plate lists behind every training pool and every evaluation set, together with the perturbation types of each dataset, are tabulated in the supplementary spreadsheet (Table S1).

Table 18: Dataset–task coverage. "Train/eval" means the dataset is in the task's training pool and in its evaluation, and "Train" means it contributes to training only.

Dataset	Morphology generation	Channel completion	Brightfield-to-fluorescence	Perturbation prediction
BBBC021	Train/eval	Train	–	Train/eval
CPG0000	Train/eval	Train/eval	Train/eval	Train/eval
CPG0001	Train/eval	Train/eval	Train/eval	Train/eval
CPG0004	Train/eval	Train/eval	–	Train/eval
CPG0036	Train/eval	Train	–	Train/eval
RxRx3	Train/eval	Train/eval	–	Train/eval

A.11 IMPLEMENTATION DETAILS

Table 19 lists the role that each task assigns to its image groups.

Table 19: Role assignment of the four tasks. Each column is one role embedding, each row lists the images a task assigns to that role, and a dash marks a role the task does not use.

	Reference			Generated
Task	Observed ($e_{\mathrm{role,obs}}$)	Brightfield ($e_{\mathrm{role,bf}}$)	Control ($e_{\mathrm{role,ctrl}}$)	Target ($e_{\mathrm{role,tgt}}$)
Morphology generation	—	—	—	X
Missing-channel completion	$X^{\mathcal{O}}$	—	—	$X^{\mathcal{M}}$
Brightfield-to-fluorescence	—	X^{bf}	—	X
Perturbation prediction	—	—	X^{ctrl}	X

Images have a resolution of 512×512, and Enc produces a $4 \times 64 \times 64$ latent representation per channel, which is divided into 2×2 patches to give $L = 1024$ tokens. The backbone is the DiT-XL

configuration with 28 blocks, a width of 1152 and 16 attention heads. The text encoder is Qwen2.5-14B-Instruct, whose penultimate-layer hidden states are precomputed and cached. Conditions are dropped with probability 0.1 during training to learn the unconditional branch. During sampling, brightfield and control images serve only as conditions and are not integrated, and after integration the generated channels are decoded into images and restored to the original channel order.

Both training stages use AdamW with a constant learning rate of 5×10^{-5} and weight decay 0.05, an effective batch of eight images, and an exponential moving average of the weights with decay 0.9999. The first stage trains on morphology generation for 200k steps, and the second continues from it on the four tasks for 200k steps with the task ratio and the data balancing of Section 3. Sampling uses the second-order Heun solver with 50 steps and no guidance.

A.12 Baseline configuration

Table 20 lists, for every compared method, what one set of weights covers, the number of models trained, the training exposure and the sampler. All baselines are retrained by us at 512 pixels on the splits of Section A.10.

Table 20: Baseline configuration. Every baseline is retrained by us at 512 pixels on our splits; none is taken from published numbers and none is evaluated at a lower resolution and upsampled. *Scope* is what one set of weights covers: a *view* is one source and one perturbation type, and methods that condition through a per-dataset embedding table or a molecular encoder cannot span both types. *Exposure* is steps $\times$ effective batch, the length of the training run; **for every method the checkpoint reported is the best-scoring one saved during its run, selected on the validation split.** Sampling is fixed per method and never tuned per cell.

Method	Scope	Models	Steps $\times$ batch	Channels	Sampler
Morphology generation					
MorphGen	view	7	200k $\times$ 8	native, per-channel latents	Euler–Maruyama 50, cfg 1
MorphoDiff	view	8	200k $\times$ 8	RGB composite	DDIM 50, no CFG
Mol2Image	compound view	6	200k $\times$ 8	RGB composite	progressive flow
MDTv2-XL/2	source	4	200k $\times$ 8	RGB composite	50 steps, cfg 1
Stable Diffusion	view	8	200k $\times$ 8	RGB composite	DDIM 50, no CFG
Channel completion					
U-Net	channel count	2	200k $\times$ 8	native, 5 or 6	deterministic
Pix2Pix	channel count	2	200k $\times$ 8	native, 5 or 6	deterministic
MCD	channel count	2	200k $\times$ 8	native, 5 or 6	EDM sampler
Brightfield-to-fluorescence					
cWGAN-GP	all	1	200k $\times$ 8	native, 6 in / 5 out	deterministic
ISL	all	1	200k $\times$ 8	native, 6 in / 5 out	deterministic
CellResDM	all	1	200k $\times$ 8	native, 5 in / 5 out	official ResShift
Perturbation prediction					
IMPA	source	4	200k $\times$ 8	native, 5 or 6	single forward pass
OmniMorph	all sources, all tasks	1	(200k+200k) $\times$ 8	native, up to 6	Heun 50, cfg 1

A.13 Licenses and reproducibility

BBBC021 is distributed by the Broad Bioimage Benchmark Collection under the Creative Commons Attribution 3.0 license, the Cell Painting Gallery datasets CPG0000, CPG0001, CPG0004 and CPG0036 under CC0 1.0, and RxRx3-core by Recursion under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license. The shared model is trained for 200k steps per stage and every baseline for 200k steps, all at an effective batch of eight (Table 20). These are the lengths of the training runs: **for every method, ours included, the checkpoint reported is the best-scoring checkpoint saved during its run, selected on the validation split, rather than the last one.** The code, the frozen target lists, the mask bank, the control pairings, the panel definitions and the trained weights will be released, so that every number of the paper can be recomputed from the same frozen inputs.