Scientific Reports

https://doi.org/10.1038/s41598-026-44976-6

Article in Press

A diffusion model conditioned on compound bioactivity profiles for generating high-content images

Received: 24 January 2025

Accepted: 15 March 2026

Published online: 03 April 2026

Cite this article as: Cook S., Chyba J., Gresoro L. *et al.* A diffusion model conditioned on compound bioactivity profiles for generating high-content images. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-44976-6

Steven Cook, Jason Chyba, Laura Gresoro, Doug Quackenbush, Minhua Qiu, Peter Kutchukian, Eric J. Martin, Peter Skewes-Cox & William J. Godinez

A diffusion model conditioned on compound bioactivity profiles for generating high-content images

Steven Cook[1*], Jason Chyba[1], Laura Gresoro[1], Doug Quackenbush[1], Minhua Qiu[1], Peter Kutchukian[2], Eric J. Martin[3], Peter Skewes-Cox[3], William J. Godinez[3*]

[1]Novartis Biomedical Research, San Diego, 92121, CA, USA.
[2]Novartis Biomedical Research, Cambridge, 02139, MA, USA.
[3]Novartis Biomedical Research, Emeryville, 94608, CA, USA.

*Corresponding author(s). E-mail(s): steven-1.cook@novartis.com; william_jose.godinez_navarro@novartis.com;

Abstract

High-content imaging (HCI) provides a rich snapshot of compound-induced phenotypic outcomes that augment our understanding of how compounds affect cellular systems. Generative imaging models for HCI provide a route towards anticipating the phenotypic outcomes of chemical perturbations in silico at unprecedented scale and speed. Here, we developed Profile-Diffusion (pDIFF), a generative method leveraging a profile-to-image latent diffusion model conditioned on in silico bioactivity profiles to generate high-content images displaying the cellular outcomes induced by compound treatment. We trained and evaluated a pDIFF model using high-content images from a Cell Painting assay profiling 3750 molecules (3375 training compounds and 375 held-out compounds) with corresponding in silico bioactivity profiles. Using the held-out set we demonstrate that pDIFF provides improved visual depictions of phenotypic responses of compounds that are structurally dissimilar to training compounds, compared to a baseline profile-to-image latent diffusion model trained on substructural molecular descriptors only. In a virtual hit expansion scenario, pDIFF yielded statistically significant improvement in expansion outcomes as measured by nearest-neighbor retrieval accuracy, compared to expansions based on compound structural representations, bioactivity profiles, and generative imaging models based only on substructural molecular descriptors, thus showcasing the potential of the methodology to speed up and improve the search for novel phenotypically active molecules.

Keywords: generative AI for drug discovery, in silico HCI, virtual screening

1 Introduction

Compound-induced phenotypes observed through high-content imaging (HCI) assays provide rich clues as to the activity and mechanisms of compounds in cellular systems [2]. While HCI assays can screen large compound collections, experimental assays are still restricted to a comparatively small sample of the synthetically accessible chemical space [9, 10, 14]. Generative models that predict images containing the specific phenotype induced by any compound can enable virtual screening and profiling of compound collections at unprecedented scale, speed, and cost [47].

Early work in generative models for cell microscopy images focused on mapping a symbolic representation of cellular biology (e.g., a state vector) onto an image using explicit appearance models for image generation [7, 33, 50]. More recent approaches leverage advances in deep neural networks [26] to learn image generation functions directly from training images [13, 22, 34]. Conditioning the generation functions with compound representations, such as compound fingerprints, allows these methods to generate high-content images showing the phenotypic outcomes induced by compound treatment [35, 47]. Architecturally, these methods are built upon flow-based generative models [47] and generative adversarial networks (GANs) [35]. Recent research points to the superior efficacy of diffusion−based generative architectures for conditional image generation [8], and more broadly, for drug discovery tasks [1, 6]. Chemically, these methods are conditioned on compound representations derived directly from molecular structures, limiting model generalizability to structurally similar chemical matter [35, 47]. Alternative representations based on imputed compound bioactivity [31] can potentially improve the ability of such generative methods to extrapolate to novel chemical matter.

In this paper, we introduce Profile Diffusion (pDIFF), an approach combining a stable diffusion-based generative model with in silico bioactivity profiles to generate high-content images displaying the cellular outcomes induced by compound treatment. Generating high-content images instead of predicting image features allows for flexible downstream application as any desired features can then be computed separately on the pDIFF output images. pDIFF model training only requires pairs of vectorial compound representations and associated high-content images, without requiring *a priori knowledge of the distribution of signal intensities or morphological features.* We build and validate a pDIFF model with a collection of bioactivity profiles and Cell Painting images corresponding to a diverse chemogenomic library of 3750 compounds [3] (see **Supplementary Figure 1** for calculated molecular properties). Using a "realistically novel" split for validation [31], we demonstrate that pDIFF improved extrapolation to novel chemical matter compared to a baseline diffusion model conditioned on chemical fingerprints. We also test the pDIFF model in a virtual hit expansion scenario and show that pDIFF yields statistically significant improvement in hit expansion outcomes.

2 Results

2.1 Profile Diffusion: Stable diffusion model conditioned on compound profiles

We developed the Profile Diffusion (pDIFF) methodology based off the stable diffusion architecture [39]. At training time, pDIFF takes as input an in silico bioactivity profile as well as a high-content image displaying the cellular outcomes induced by that compound. The input image is compressed into a lower-resolution latent representation through a pre-trained variational autoencoder (VAE) [24]. A latent diffusion network conditioned on the compound's bioactivity profile is trained in the VAE's latent space to predict the noise added to the latent image at each step of a diffusion process (see **Figure 1** and **Methods**). Thus, in comparison with the original implementation of the Stable Diffusion model, where natural language embeddings are used to condition the denoising network, pDIFF instead induces the model to learn the "language of compound bioactivity".

To generate an image for a specific compound with a trained pDIFF model, we take the in silico bioactivity profile of that compound, sample a random noise tensor in the VAE's latent space, and iteratively denoise that tensor with the model conditioned on the compound's bioactivity profile together with a denoising diffusion probabilistic models (DDPM) schedule [19] (see **Methods**). Varying the number of denoising iterations trades off image quality against computation time, as the model must be invoked more often for more steps. Finally, the denoised tensor is passed through the VAE decoder to synthesize a high-content image. Different random noise tensors result in different images, thus allowing generation of a collection of images for each input compound.

2.2 Profile Diffusion generalizes to novel chemical matter

To validate pDIFF in a realistic drug discovery context, we used 3750 compounds of the Novartis Mechanism-of-Action (MoA) Box chemogenomic library (MoA Box) [3]. These compounds were profiled in a Cell Painting assay [2], where U-2 OS cells were compound-treated in triplicate at 12.5 μM (see **Methods**). Twenty-four high-content images were acquired per compound. We utilize the cytoplasm, mitochondria, and nuclei channels of the Cell Painting protocol to construct three-channel RGB images that, along with their corresponding compound profiles, form the dataset for model building and evaluation.

As conditioning profile for pDIFF, we used the in silico bioactivity profiles predicted by Profile-QSAR (pQSAR) [31], a massively multitask bioactivity machine learning model trained on over two million compounds across 14222 Novartis dose-response assays (see **Methods**). For each of the 3750 compounds in our dataset, pQSAR predicts activity in terms of pAC_{50} values (which is the negative logarithm of the half-maximal activity concentration) for each of the 14222 assays, thus resulting in a 14222-length profile per compound. For numerical tractability, we reduced predictions from the 14222 assays to 2018 assays through a target-focused assay selection scheme (see **Methods**).

For benchmarking purposes we also trained a stable diffusion model conditioned on chemical fingerprints. Specifically, we used extended-connectivity fingerprints (ECFPs) [38], each folded into a 2048 count vector. Briefly, ECFPs indicate the presence or absence of substructural compound motifs defined using hashing algorithms, permitting intersection-over-union similarity measures such as Tanimoto similarity. For this baseline model, we used the same training and inference regimes as used for pDIFF. At inference time, we set both pDIFF and the baseline model to generate 12 images per compound (see **Methods**).

To validate the performance of pDIFF and the baseline model, we used a "realistically novel" cluster-based split approach [31]. This approach aims to replicate a realistic screening scenario, where project teams are interested in testing compounds that differ significantly from those previously tested (see **Methods**). We chose 10% of the total dataset size for the held-out set, giving us 3375 training and 375 held-out compounds. This split results in a highly dissimilar median Tanimoto coefficient of 0.11 between train and test compounds (**Supplementary Figure 2**).

Example real and generated images for three held-out compounds are shown in **Figure** 2. These active compounds induce outcomes visually different from those observed in the neutral controls (see **Supplementary Figure 3**). We show images for Halofuginone, a glutamyl-prolyl-tRNA synthetase inhibitor [23] that is known to inhibit the viability of Osteosarcoma cell lines [25]. The baseline model conditioned on chemical fingerprints is unable to correctly predict the cell death outcome induced by this compound, which has a low chemical similarity to the training set (Tanimoto coefficient of 0.23 to the nearest neighbor compound in the training set). In contrast, the pDIFF model, which is conditioned on pQSAR bioactivity profiles, visually recapitulates this outcome. We also show images for CHEMBL2326002, a protein kinase C-theta inhibitor [21] which induces a slightly elongated phenotype with a punctuated pattern in the mitochondria channel. The baseline model does not predict the elongated punctuated pattern, whereas pDIFF generates images with cells closely resembling the phenotype induced by this compound. For Brusatol, which plays a role in DNA damage repair [28] by acting as an inhibitor of the NRF2 pathway [37], cells exhibit a toxic phenotype that is anticipated only by pDIFF.

Additional images for molecules are shown in **Supplementary Figure 4**. Across a variety of molecules, mechanisms of action (viz. actin polymerization, LATS inhibition, antifungal activity, v-ATPase inhibition, viral fusion inhibition, HSP90 inhibition, BRD4 inhibition), and phenotypic responses (e.g., punctae, elongation, filamentation, clumping), pDIFF generated images displaying phenotypes that closely resembled those observed in the real images. For Thailandepsin A [46], an HDAC inhibitor, pDIFF generated images displaying elongated cells while the real images display a phenotype more similar to the neutral controls. As cell lines vary in their expression of cellular targets and pathways [17], we hypothesize that pDIFF, learning from other HDAC inhibitors in the training set, predicted an outcome not induced in the U2-OS cell line used in this study. We also show images for pterocarpanquinone, which modulates FoxO3a activity [32]. For this molecule both the baseline diffusion model and the pDIFF model did not recapitulate the apoptotic phenotype.

To quantitatively assess the ability of pDIFF to recapitulate phenotypic outcomes, we follow a previous scheme [47] by computing biologically relevant features (viz. coverage, cell count, cell size) for real and generated image sets. To calculate these segmentation-dependent features, we merge the channels to create grayscale images, then utilize the Cellpose [43] segmentation model to create cell masks from which image-level cell coverage, count, and average cell size metrics can be directly computed. Features are averaged over all images for each compound, then Spearman correlation coefficients between these compound-aggregated features of real and generated image sets are calculated (see **Methods**).

Table 1 shows the real vs. generated correlation coefficients across the 375 held-out (test) compounds for three features. As upper bound, we compute the correlation among real images of the held-out compounds by splitting the 24 real images into two halves, calculating features for each half, and computing the correlation coefficients between the features of both halves. Correlation values for the image features between the two sets of real images range from 0.48 to 0.67. For this challenging held-out set of compounds with low chemical similarity to the training set (median Tanimoto coefficient of 0.11 between train and test compounds, cf. **Supplementary Figure 2.)** , the baseline diffusion model conditioned on chemical fingerprints yielded correlation values ranging from 0.04 to 0.13. The pDIFF model yielded correlation coefficients between 0.12 to 0.48, thus exhibiting substantially improved performance on novel chemical matter compared to the baseline model.

2.3 Profile Diffusion enables improved retrieval of phenotypically similar molecules for hit expansion

Having ascertained the ability of pDIFF to generalize to novel chemical matter, we proceeded to test pDIFF in a virtual hit expansion scenario, where the goal is to find compounds inducing phenotypic outcomes similar to those induced by known active compounds (i.e., phenotypic screening hits). To set the ground truth for this scenario, we take the real images of 101 active query compounds from the training set and compare these to the real images of the 375 compounds in the held-out (test) set. To compare the similarity of two compounds via images, we use the Sinkhorn divergence [11] on the sets of Cellpose [43] feature vectors derived from the compounds' corresponding images. Using these Sinkhorn divergences, we retrieve the 50 nearest-neighbor compounds in the test set for each query compound in the training set. We repeat the nearest neighbor retrieval but with pDIFF-generated images instead of real images for the test compounds, thus comparing real images for query compounds with pDIFF synthetic images for test compounds. As performance measure for this task, we calculate the percentage overlap between the ground truth nearest neighbors and those retrieved using pDIFF images.

We report the distribution of percentage overlap between the real and pDIFF test-set nearest-neighbors for the 101 query compounds in **Figure** 3. As baseline, we show the percentage overlap values for 101 random selections of 50 test compounds, which yields a median percentage overlap of 14%. As additional baselines, we show the percentage overlap values for nearest neighbors retrieved with chemical fingerprints as well as bioactivity profiles (ECFP and pQSAR profiles, respectively) using a Tanimoto

or cosine distance. These approaches lead to median percentage overlaps of 16% and 38%, respectively. A baseline diffusion model conditioned on ECFPs leads to a median percentage overlap of 16%. Finally, retrieval with images generated by pDIFF leads to a median percentage overlap of 50%. A two-sample Kolmogorov–Smirnov test [5] as implemented in the SciPy library [45] was used to determine the statistical significance of differences among the percentage overlap values of the different approaches. The resulting p-values were corrected for false discovery rate with the Benjamini–Hochberg method (see **Supplementary Table 1**). The differences between the distribution of percentage overlap values of the pDIFF approach and those of all other approaches were significant. This outcome suggests that pDIFF enables improved retrieval of phenotypically similar molecules, thus enhancing virtual expansion outcomes.

3 Discussion

High-content screening assays provide rich mechanistic and holistic information about cellular responses to chemical perturbations. Overcoming fundamental limitations in screening capacity and compound synthesis via in silico generated high-content images holds potential to vastly expand the compound search space. To this end, we developed a purely virtual generative imaging approach, pDIFF, necessitating only in silico bioactivity profiles as input for image generation. Using a realistic validation scheme, we show that pDIFF provides remarkably improved image predictions for novel chemical matter.

Generating HCI images instead of predicting image measurements directly from compound representation benefits from state-of-the-art diffusion models that both excel at generating high-quality, high-dimensional images as well as exhibiting more stable convergence [20]. Relevant image-derived measurements reflective of phenotype can subsequently be extracted from the generated images with existing image analysis pipelines. In our study, for example, we use the same analysis pipeline to segment and extract image features for both real and pDIFF-generated images. While the image features could be predicted directly from in silico compound profiles, the ability to generate images is key to gaining the confidence of experimentalists as the generated images readily provide visual insight informing the image-derived measurements. pDIFF allows for visual review of HCI-based virtual screens just as chemists visually review virtual screens from chemical similarity or docking.

In our study, we used the bioactivity profiles calculated by the in-house Profile-QSAR (pQSAR) model [31] trained on over 2M compounds and 14k Novartis internal dose-response assays. The pQSAR methodology is publicly available, and useful models have been built on public bioactivity databases, such as ChEMBL [49]. pDIFF could also be trained on other experimental profiles such as gene expression signatures [27, 44, 48]. Furthermore, augmenting the bioactivity profiles with cell type and concentration information with corresponding image data could expand the predictive scope of a single pDIFF model.

Given its ability to predict phenotypic outcomes for novel chemical matter, we envision pDIFF having impact in a number of drug discovery scenarios. As shown by our findings, hit expansion campaigns searching for molecules inducing phenotypic

outcomes similar to those induced by hit molecules can benefit from the improved image-based similarity calculations enabled by pDIFF. Profiling activities [4] aiming to reveal the mechanism-of-action of molecules can likewise leverage pDIFF generated images for similarity calculations relative to images of a reference set of annotated molecules. Image-based readouts have been shown to be predictive of cardiotoxocity [29, 41] and we anticipate pDIFF-generated images to help flag adverse cardiac effects of compounds. Historical imaging assays could be revisited virtually through pDIFF, so past image screening investments can be leveraged to enable future follow up work.

Training pDIFF is computationally demanding due to the need to re-learn the space of high-content images and their corresponding profiles instead of the CLIP text embeddings Stable Diffusion was trained with. Our training set of 12 fields of view (FOVs) per each of 3,375 compounds, took 360 GPU-hours to train for 30,000 steps. A trained pDIFF model can predict one 512×512 pixel output patch per GPU-second using a conservative, high-quality inferencing schedule (DDPM [19], 100 network inference steps). Alternative schedulers such as the Diffusion Probabilistic Models (DPM) solver [30] purport to give a similar quality of results using only tens of steps, making this an attractive area for further optimization. Diffusion model inference is more expensive than other generative models like GANs due to the need to repeatedly invoke the network. Even with our conservative settings, the entire MoA box imageset of 90,000 images (3,750 compounds × 24 FOVs) could be re-generated via inference in 25 GPU-hours, without need for expensive and time-consuming compound synthesis, plating, image acquisition, etc.

In conclusion, our work shows the potential of in silico image prediction to anticipate induced phenotypic outcomes of compounds, including those generated through other machine learning methods [12, 40, 42], much earlier in the drug discovery pipeline. The application of pDIFF allows us to explore broader and more diverse compound collections at unprecedented speed, scale, and efficiency.

4 Acknowledgements

We thank Frederick Lo for help with data logistics and pre-processing. We thank Mark A. Bray for fruitful discussions.

Conflict of interest

All authors are (or were at the time of their involvement with the studies) employees of Novartis.

Data availability

The data used in this study are proprietary to Novartis. The data are not publicly available due to intellectual property restrictions. An example dataset is available in the pDIFF code respository.

Code availability

The code and an example dataset for pDIFF is available in **Supplementary Code** and at https://github.com/Novartis/pDIFF

Author contribution

S.C. and W.J.G. designed and led the study. S.C. developed, implemented, and evaluated pDIFF. J.C., L.G., and D.Q. developed and ran the imaging assay. E.J.M. developed the algorithm to compute the in silico bioactivity profiles and provided feedback. M.Q., P.K., and P.S.-C. provided feedback. S.C., M.Q., and W.J.G. analyzed and interpreted the results. S.C. and W.J.G. wrote the article. All authors reviewed the manuscript.

5 Figures

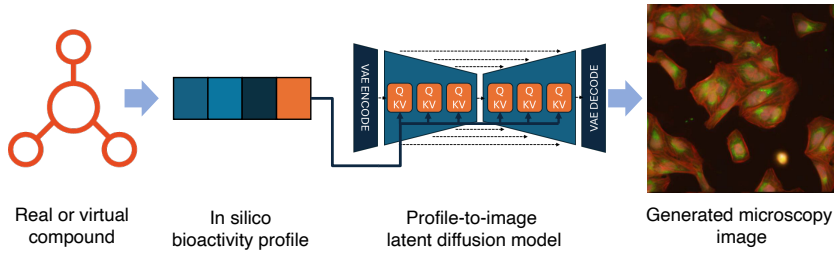

Fig. 1 Profile-Diffusion (pDIFF) workflow. A virtual or real compound is represented computationally through an in silico bioactivity profile computed by pQSAR. The representation is used to condition the generative process of the stable diffusion model underlying pDIFF to generate a high-content image showing the phenotypic outcome induced by compound treatment.

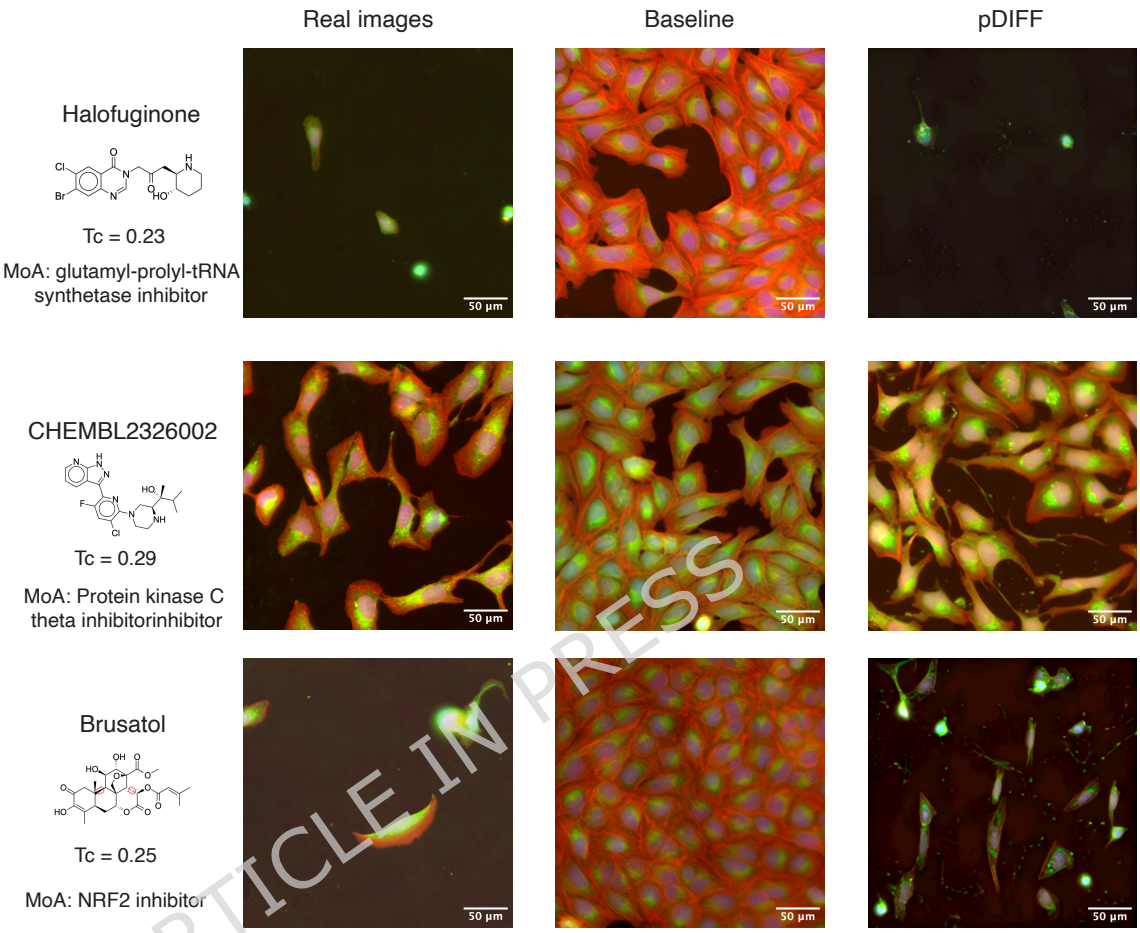

Fig. 2 Molecules in the realistic held-out set along with their corresponding real images (Nuclei, blue; mitochondria, green; red, cytoplasm) as well as images generated with diffusion models. We show images generated by a baseline diffusion model conditioned on chemical fingerprints (Baseline) as well as by pDIFF. For each molecule, we list the Tanimoto coefficient (Tc) to the nearest neighbor molecule in the training set as well as the associated mechanism of action (MoA).

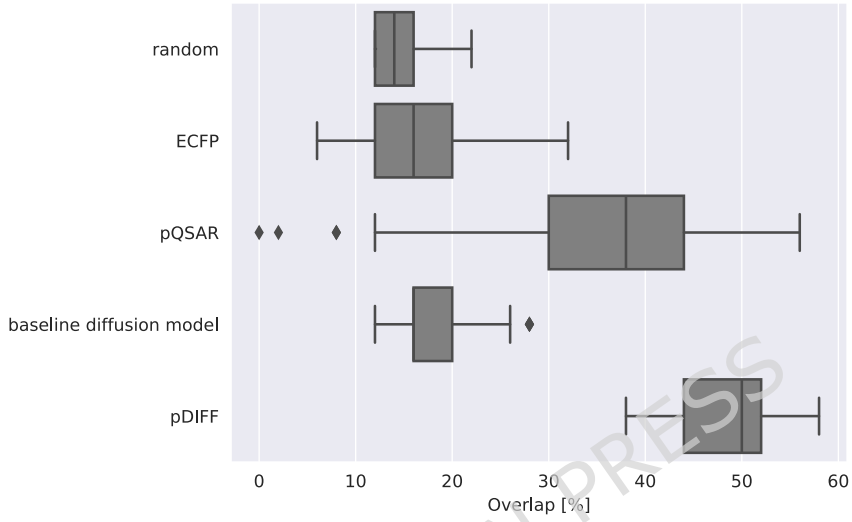

Fig. 3 Distribution of percentage overlap values between 50 ground truth nearest neighbors and those retrieved with virtual expansion approaches for 101 query compounds. The query compounds are actives from the training set, and we query into the 375 compounds of the held-out set. We show the performance for an approach selecting molecules randomly ('random') as well as for similarities computed using chemical or bioactivity profiles (ECFP and pQSAR, respectively). The results for approaches computing similarities using image profiles derived from images generated through either a baseline diffusion model or pDIFF are also shown.

6 Tables

Table 1 Results for held-out validation compounds. Values given are Spearman correlation coefficients between compound-aggregated average values for each image feature. The first row shows the correlation values for each half of the real images compared against each other. pDIFF model performs better than the baseline diffusion model conditioned on chemical fingerprints.

	Coverage	Cell Count	Cell Size
Real images upper bound	.67	.56	.48
Baseline diffusion model	.13	.10	.04
pDIFF	.48	.28	.12

7 Methods

7.1 Image Acquisition and Preparation

U2-OS cells were treated with 3750 compounds from the Novartis MoA Box [3] at a fixed concentration of 12.5µM and incubated for 24 hours. The Cell Painting staining protocol [2] was applied. Image acquisition was performed using a 4 HP laser, 4 camera Phenix imaging system and a 20x NA1.0 water immersion objective. Acquired images were background corrected using the BaSIC algorithm [36], rescaled by a factor of $\frac{1}{2}$ from 2160x2160 pixels to 1080x1080, then random crops of 512x512 are extracted from each image for training. RGB images were assembled from the F-actin Cytoskeleton, Mitochondria, and Nucleus channels (Phalloidin/Alexa, MitoTracker Deep Red, and Hoescht 33342 stainings, respectively).

7.2 Profile calculations

To compute the in silico bioactivity profiles for each compound, we used 14222 models from the massively-multitask pQSAR algorithm for predicting pAC_{50} values for 14222 Novartis-internal biochemical and cellular dose-response assays. We narrowed the resulting 14222-D bioactivity profile by first selecting only the assays where pQSAR models provided useful predictions on the challenging "realistically novel" held-out set per assay, as measured by the squared Pearson correlation coefficient $r^2 > 0.3$ between predicted and experimental pAC_{50} values. We then selected the target-based biochemical and cellular assays in the narrowed assay collection, grouped them by target protein, and selected the best-performing pQSAR model based on the models' r^2 values per target. For the remaining purely-phenotypic assays, we grouped them by drug discovery project, and likewise selected the best-performing pQSAR model per project. For the remaining set of assays with neither target nor project annotations, we selected only very high-quality pQSAR models ($r^2 > 0.9$). This selection process resulted 2018 assays, amounting to a 2018-D in silico bioactivity profile per compound. A block diagram describing this selection process is shown in **Supplementary Figure 5**. The profile was zero-padded i.e., 30 zeros appended to the vector to reach the 2048-D input length required for the pDIFF model.

7.3 Model Architecture

The Stable Diffusion 2.1 model [39] was used as the backbone for pDIFF. The lengths of the cross-attention weights are changed to 2048, and all existing weights are re-initialized. As no natural language inputs are needed, the CLIP module is removed. The pre-trained VAE is retained and its weights are frozen. Following [39], we train the model to minimize the mean squared error between actual noise ϵ and predicted noise $\epsilon_\theta(z_t, t, y)$ (Eq. 1), where t is a current timestep of the noise schedule, θ are the network parameters, z_t is the latent image representation at step t of the noise schedule, $\mathcal{E}$ is the VAE used to encode the image x, and c is the conditioning profile.

$$L = \mathbb{E}_{\mathcal{E}(x),c,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2\right] \quad (1)$$

7.4 Model Training

pDIFF was trained for 30k steps on 4x Nvidia A100 GPUs using huggingface Accelerate, no mixed precision, in distributed data parallel mode. Starting from 24 images per compound in the training set, pDIFF is trained using 12 of those images, while the 12 additional images per compound in the training set are held out from training and reserved for calculating an upper bound for expected image similarity. Training time was approximately 90 wall-clock hours (360 GPU hours) for the 40k images in the training set (12 images per compound x 3375 compounds). The Min-SNR weighting strategy [16] was applied to accelerate convergence, and offset noise [15] was applied to help the model learn to generate empty and nearly-empty FOV images with very low average intensity. Training hyperparameters are presented in **Supplementary Table 2**.

7.5 Inference

The DDPM scheduler [19] with 100 steps was used for inference with the model. 12 images were generated per compound profile, at a time cost of approximately 1 second per 512x512 output image per GPU. We use the guidance algorithm [18] to drive the generation process. In Eq. 2 w is the guidance scale, c is the conditioning profile, z_t is the output result of the previous denoising step, $\emptyset$ is a null conditioning profile, $\epsilon_t(z_t, \emptyset)$ is the model's unconditional noise prediction, $\epsilon_t(z_t, c)$ is the conditional noise prediction, and $\tilde{\epsilon_t}$ is the resulting guided noise [18].

$$\tilde{\epsilon_t} = (1 + w)\epsilon_t(z_t, c) - w\epsilon_t(z_t, \emptyset) \quad (2)$$

Inference-time classifier-free guidance scale value $w = 4$ was chosen empirically.

7.6 Model Validation

We used the "realistically novel" split [31] to evaluate pDIFF. In this approach, the compounds are clustered by Tanimoto similarity, and the clusters ranked in order of size. We allocated compounds from the larger clusters to the training set until 3375 (90%) of the compounds were included. The remaining 375 (10%) of compounds coming from the singletons and smaller clusters were allocated to the realistic held-out set.

To quantify model performance, we first segmented real and generated images using Cellpose with the default cyto2 model in grayscale mode [43]. Following a previous approach [47], hand-engineered features were calculated for each image. Specifically, we computed the total image area covered by segmented cells (coverage), the number of segmented cells in the image (cell count), and the size of the segmented cells. We averaged the values of these features over all images corresponding to the same

compound and calculated Spearman correlations coefficients between features derived from real and generated images across the 375 held-out, or test, compounds.

7.7 Reproducing the Training Set

We check that models learned to reproduce phenotypic outcomes by evaluating the performance on the training set. **Supplementary Table 3** shows the Spearman correlation coefficients comparing pDIFF generated images for the 3375 molecules in the training set to the real images. As upper bound, we report the correlations of coverage, count, and size metrics between the 12 real images per . Resulting correlation values for the coverage, count, and size metrics in this real-vs-real comparison for the training set compounds range from 0.44 to 0.62. A baseline diffusion model conditioned on chemical fingerprints yields correlation values between 0.15 to 0.41. The pDIFF model shows very realistic performance with correlation values ranging from 0.40 to 0.59.

References

[1] Abramson J, Adler J, Dunger J, et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630(8016):493–500

[2] Bray MA, Singh S, Han H, et al (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature Protocols 11(9):1757–1774

[3] Canham SM, Wang Y, Cornett A, et al (2020) Systematic Chemogenetic Library Assembly. Cell Chemical Biology 27(9):1124–1129

[4] Chandrasekaran SN, Ceulemans H, Boyd JD, et al (2021) Image-based profiling for drug discovery: due for a machine-learning upgrade? Nature Reviews Drug Discovery 20(2):145–159

[5] Conover W (1971) One-sample Kolmogorov test/two-sample Smirnov test. In: B W (ed) Practical Nonparametric Statistics. Wiley, New York, p 295–314

[6] Corso G, Stärk H, Jing B, et al (2023) Diffdock: Diffusion steps, twists, and turns for molecular docking. URL https://arxiv.org/abs/2210.01776, 2210.01776

[7] D. V. Sorokin, I. Peterlík, V. Ulman, et al (2018) FiloGen: A Model-Based Generator of Synthetic 3-D Time-Lapse Sequences of Single Motile Cells With Growing and Branching Filopodia. IEEE Transactions on Medical Imaging 37(12):2630–2641. https://doi.org/10.1109/TMI.2018.2845884

[8] Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. URL https://arxiv.org/abs/2105.05233, 2105.05233

[9] Dobson CM (2004) Chemical space and biology. Nature 432(7019):824–828

[10] Drew KLM, Baiman H, Khwaounjoo P, et al (2012) Size estimation of chemical space: how big is it? Journal of Pharmacy and Pharmacology 64(4):490–495

[11] Feydy J, Séjourné T, Vialard FX, et al (2019) Interpolating between optimal transport and mmd using sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp 2681–2690

[12] Godinez WJ, Ma EJ, Chao AT, et al (2022) Design of potent antimalarials with generative chemistry. Nature Machine Intelligence 4(2):180–186

[13] Goldsborough P, Pawlowski N, Caicedo JC, et al (2017) CytoGAN: Generative Modeling of Cell Images. bioRxiv p 227645

[14] Grygorenko OO, Radchenko DS, Dziuba I, et al (2020) Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. iScience 23(11):101681

[15] Guttenberg N (2023) Diffusion with Offset Noise. URL https://www.crosslabs.org//blog/diffusion-with-offset-noise

[16] Hang T, Gu S, Li C, et al (2023) Efficient diffusion training via min-snr weighting strategy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7441–7451

[17] Heinrich L, Kumbier K, Li L, et al (2023) Selection of Optimal Cell Lines for High-Content Phenotypic Screening. ACS Chemical Biology 18(4):679–685. Publisher: American Chemical Society

[18] Ho J, Salimans T (2022) Classifier-free diffusion guidance. URL https://arxiv.org/abs/2207.12598, 2207.12598

[19] Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. URL https://arxiv.org/abs/2006.11239, 2006.11239

[20] Huang Z, Zhou P, Yan S, et al (2024) Scalelong: towards more stable training of diffusion model via scaling network long skip connection. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA

[21] Jimenez JM, Boyall D, Brenchley G, et al (2013) Design and Optimization of Selective Protein Kinase C theta (PKCTheta) Inhibitors for the Treatment of Autoimmune Diseases. Journal of Medicinal Chemistry 56(5):1799–1810. Publisher: American Chemical Society

[22] Johnson GR, Donovan-Maiye RM, Maleckar MM (2017) Generative modeling with conditional autoencoders: Building an integrated cell. URL https://arxiv.org/abs/1705.00092, 1705.00092

[23] Keller TL, Zocco D, Sundrud MS, et al (2012) Halofuginone and other febrifugine derivatives inhibit prolyl-tRNA synthetase. Nature Chemical Biology 8(3):311–317

[24] Kingma DP, Welling M (2022) Auto-encoding variational bayes. URL https://arxiv.org/abs/1312.6114, 1312.6114

[25] Lamora A, Mullard M, Amiaud J, et al (2015) Anticancer activity of halofuginone in a preclinical model of osteosarcoma: inhibition of tumor growth and lung metastases. Oncotarget 6(16):14413–14427

[26] LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539, URL https://doi.org/10.1038/nature14539

[27] Li H, Qiu J, Fu X (2012) RASL-seq for Massively Parallel and Quantitative Analysis of Gene Expression. Current Protocols in Molecular Biology 98(1)

[28] Li J, Xu C, Liu Q (2023) Roles of NRF2 in DNA damage repair. Cellular Oncology 46(6):1577–1593

[29] Garcia de Lomana M, Marin Zapata PA, Montanari F (2023) Predicting the Mitochondrial Toxicity of Small Molecules: Insights from Mechanistic Assays and Cell Painting Data. Chemical Research in Toxicology 36(7):1107–1120

[30] Lu C, Zhou Y, Bao F, et al (2022) Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. URL https://arxiv.org/abs/2206.00927, 2206.00927

[31] Martin EJ, Polyakov VR, Zhu XW, et al (2019) All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. Journal of Chemical Information and Modeling 59(10):4450–4459

[32] Nestal de Moraes G, Castro CPereira, Salustiano EJesus, et al (2014) The pterocarpanquinone LQB-118 induces apoptosis in acute myeloid leukemia cells of distinct molecular subtypes and targets FoxO3a and FoxM1 transcription factors Corrigendum in /10.3892/ijo.2019.4874. International Journal of Oncology 45(5):1949–1958

[33] Murphy R (2005) Location proteomics: a systems approach to subcellular location. Biochemical Society Transactions 33(3):535–538

[34] Osokin A, Chessel A, Salas REC, et al (2017) Gans for biological image synthesis. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 2252–2261, https://doi.org/10.1109/ICCV.2017.245

[35] Palma A, Theis FJ, Lotfollahi M (2023) Predicting cell morphological responses to perturbations using generative modeling. bioRxiv p 2023.07.17.549216

[36] Peng T, Thorn K, Schroeder T, et al (2017) A BaSiC tool for background and shading correction of optical microscopy images. Nature Communications 8(1):14836

[37] Ren D, Villeneuve NF, Jiang T, et al (2011) Brusatol enhances the efficacy of chemotherapy by inhibiting the nrf2-mediated defense mechanism. Proceedings of the National Academy of Sciences 108(4):1433–1438

[38] Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling 50(5):742–754

[39] Rombach R, Blattmann A, Lorenz D, et al (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10684–10695

[40] Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: Generative models for matter engineering. Science 361(6400):360–365

[41] Seal S, Spjuth O, Hosseini-Gerami L, et al (2024) Insights into Drug Cardiotoxicity from Biological and Chemical Data: The First Public Classifiers for FDA Drug-Induced Cardiotoxicity Rank. Journal of Chemical Information and Modeling 64(4):1172–1186

[42] Shen L, Fang J, Liu L, et al (2024) Pocket Crafter: a 3D generative modeling based workflow for the rapid generation of hit molecules in drug discovery. Journal of Cheminformatics 16(1):33

[43] Stringer C, Wang T, Michaelos M, et al (2021) Cellpose: a generalist algorithm for cellular segmentation. Nature Methods 18(1):100–106

[44] Subramanian A, Narayan R, Corsello SM, et al (2017) A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171(6):1437–1452.e17

[45] Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17:261–272

[46] Wang C, Henkes LM, Doughty LB, et al (2011) Thailandepsins: Bacterial Products with Potent Histone Deacetylase Inhibitory Activities and Broad-Spectrum Antiproliferative Activities. Journal of Natural Products 74(10):2031–2038. Publisher: American Chemical Society

[47] Yang K, Goldman S, Jin W, et al (2021) Mol2Image: Improved Conditional Flow Models for Molecule to Image Synthesis. pp 6688–6698

[48] Ye C, Ho DJ, Neri M, et al (2018) DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. Nature Communications 9(1):4307

[49] Zdrazil B, Felix E, Hunter F, et al (2023) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research 52(D1):D1180–D1192

[50] Zhao T, Murphy RF (2007) Automated learning of generative models for subcellular location: Building blocks for systems biology. Cytometry Part A 71A(12):978–990. Publisher: John Wiley & Sons, Ltd