A diffusion model conditioned on compound bioactivity profiles for predicting 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 compound mechanisms in 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 latent diffusion model conditioned on substructure-based molecular fingerprints or in silico bioactivity profiles to predict high-content images displaying the cellular outcomes induced by compound treatment. We trained and evaluated pDIFF models using high-content images from a Cell Painting assay profiling 3750 molecules with corresponding structural fingerprints and in silico bioactivity profiles. Using a realistic held-out set, we demonstrate that pDIFF models conditioned on in silico bioactivity profiles provide improved predictions of phenotypic responses of compounds with low chemical similarity to compounds in the training set compared to models trained on structural fingerprints only. In a virtual hit expansion scenario, pDIFF models conditioned on in silico bioactivity fingerprints yield significantly improved

expansion outcomes, thus showcasing the potential of generative imaging models 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 (Bray et al, 2016). While HCI assays can screen large compound collections, experimental assays are still restricted to a comparatively small sample of the synthetically accessible chemical space (Dobson, 2004; Drew et al, 2012; Grygorenko et al, 2020). 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 (Yang et al, 2021).

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 (Murphy, 2005; Zhao and Murphy, 2007; D. V. Sorokin et al, 2018). More recent approaches leverage advances in deep neural networks (LeCun et al, 2015) to learn image generation functions directly from training images (Osokin et al, 2017; Johnson et al, 2017; Goldsborough et al, 2017). 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 (Yang et al, 2021; Palma et al, 2023). Architecturally, these methods are built upon flow-based generative models (Yang et al, 2021) and generative adversarial networks (GANs) (Palma et al, 2023). Recent research points to the superior efficacy of diffusion−based generative architectures for conditional image generation (Dhariwal and Nichol, 2021), and more broadly,

for drug discovery tasks (Abramson et al, 2024; Corso et al, 2023). Chemically, these methods are conditioned on compound representations derived directly from molecular structures, limiting model generalizability to structurally similar chemical matter (Yang et al, 2021; Palma et al, 2023). Alternative representations based on imputed compound bioactivity (Martin et al, 2019) 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 chemical substructure or in silico bioactivity profiles to predict high-content images displaying the cellular outcomes induced by compound treatment. We build and validate pDIFF models with a collection of structural and bioactivity profiles as well as Cell Painting images corresponding to a diverse chemogenomic library of 3750 compounds. Using a "realistically novel" split for validation (Martin et al, 2019), we demonstrate that pDIFF models trained on in silico bioactivity profiles show improved extrapolation to novel chemical matter compared to models trained on structural profiles alone. We test the pDIFF models in a virtual hit expansion scenario and show that pDIFF models conditioned on bioactivity profiles yield significantly improved 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 (Rombach et al, 2022). At training time, pDIFF takes as input a compound profile (either a substructure-based molecular fingerprint or 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) (Kingma and Welling, 2022). A latent diffusion network conditioned on the compound's chemical or 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 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 chemistry" or the "language of compound bioactivity."

To generate an image for a specific compound with a trained pDIFF model, we take the structural or 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 structural or bioactivity profile together with a denoising diffusion probabilistic models (DDPM) schedule (Ho et al, 2020) (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) (Canham et al, 2020). These compounds were profiled in a Cell Painting assay (Bray et al, 2016), 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 for model building and evaluation.

We evaluated the ability of pDIFF to recapitulate compound-induced phenotypic outcomes with either molecular substructure or bioactivity profiles. As molecular profiles we use extended connectivity fingerprints (ECFPs), each folded into a 2048 count vector (Rogers and Hahn, 2010). As an alternative compound representation we used the in silico bioactivity profiles predicted by Profile-QSAR (pQSAR) (Martin et al, 2019), 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 each profile type (structural or bioactivity), we trained a pDIFF model. Once trained, we set each pDIFF model to generate 12 images per compound.

To validate the performance of the models, we used a "realistically novel" cluster-based split approach (Martin et al, 2019). 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** 1).

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** 2). We show images for Halofuginone, which is known to inhibit the viability of Osteosarcoma cell lines (Lamora et al, 2015). The pDIFF model conditioned on structural 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 conditioned on pQSAR bioactivity profiles visually recapitulates this outcome. We also show images for CHEMBL2326002, a protein kinase C-theta inhibitor which induces a slightly elongated phenotype with a punctuated pattern in the mitochondria channel. The model conditioned on substructural fingerprints does not predict the elongated punctuated pattern, whereas the approach conditioned on pQSAR bioactivity profiles generates images with cells closely resembling the phenotype induced by this compound. For Brusatol, which plays a role in DNA damage repair (Li et al, 2023) by acting as an inhibitor of the NRF2 pathway (Ren et al, 2011), cells exhibit a toxic phenotype that is anticipated by the pDIFF model conditioned on bioactivity only.

To quantitatively assess the ability of pDIFF to recapitulate phenotypic outcomes, we follow a previous scheme (Yang et al, 2021) by collecting two image-level features (coverage, cell count) and four cell-level features (cell size and average intensity for three channels within cell boundaries) for the real and generated image sets. 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 each of the six 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 0.46 to 0.76. For this challenging held-out set of compounds with low chemical similarity to the training set, the pDIFF model conditioned on substructural fingerprints yielded correlation values ranging from -0.07 to 0.13. The model conditioned on bioactivity profiles yielded

correlation coefficients between 0.06 to 0.46, thus exhibiting substantially improved performance on novel chemical matter.

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 (e.g. 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 (Feydy et al, 2019) on the sets of Cellpose (Stringer et al, 2021) 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 the 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 structural or bioactivity profiles only (ECFP and pQSAR, respectively) using a Tanimoto or

cosine distance. These approaches lead to median percentage overlaps of 16% and 38%, respectively. Using these profiles as conditioning input to pDIFF (ECFP pDIFF and pQSAR pDIFF, respectively) results in median percentage overlaps of 16% and 50%. A two-sample Kolmogorov–Smirnov test (Conover, 1971) 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** 2). The differences between the distribution of percentage overlap values of the pQSAR pDIFF approach and those of all other approaches were significant. The approach based on pQSAR 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 compound representations, e.g., substructural fingerprints or predicted bioactivity profiles, as input for image generation. Using a realistic validation scheme, we show that pDIFF conditioned on bioactivity profiles 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 (Huang et al, 2024). 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.

Our best results require in silico bioactivity profiles to be available for all desired compounds. In our study, we used the bioactivity profiles calculated by the in-house Profile-QSAR (pQSAR) model (Martin et al, 2019) 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 (Zdrazil et al, 2023). Using pretrained pQSAR models eliminates the need to train an embedding model for the profiles themselves as a precondition for pDIFF training. pDIFF could also be trained on other profiles such as gene expression signatures (Ye et al, 2018; Li et al, 2012; Subramanian et al, 2017) or alternative cheminformatic fingerprints (Hert et al, 2004; Riniker and Landrum, 2013). Furthermore, training on multi-modal profiles combining cell type and concentration information with the compound representations together with corresponding cell-specific and concentration-resolved experimental 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 (Chandrasekaran et al, 2021) 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 (Garcia de Lomana et al, 2023; Seal et al, 2024) 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 (Ho et al, 2020), 100 network inference steps). Alternative schedulers such as the Diffusion Probabilistic Models (DPM) solver (Lu et al, 2022) 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 (Sanchez-Lengeling and Aspuru-Guzik, 2018; Godinez et al, 2022; Shen et al, 2024), 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 is proprietary to Novartis. The data is 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 [1].

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.

[1] GitHub repository to be published upon manuscript acceptance

5 Figures

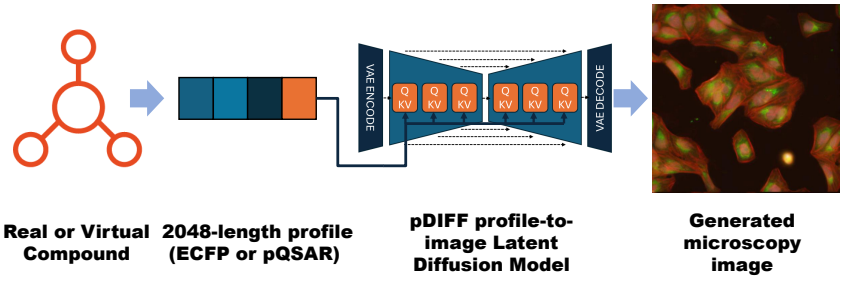


Fig. 1 pDIFF workflow. Virtual or real compounds are represented computationally through a substructural molecular fingerprint (ECFP) or an in silico bioactivity profile (pQSAR). The resulting representations are used to condition the generative process of the stable diffusion model underlying our Profile-Diffusion (pDIFF) approach to generate a high-content image showing the phenotypic outcome induced by compound treatment.

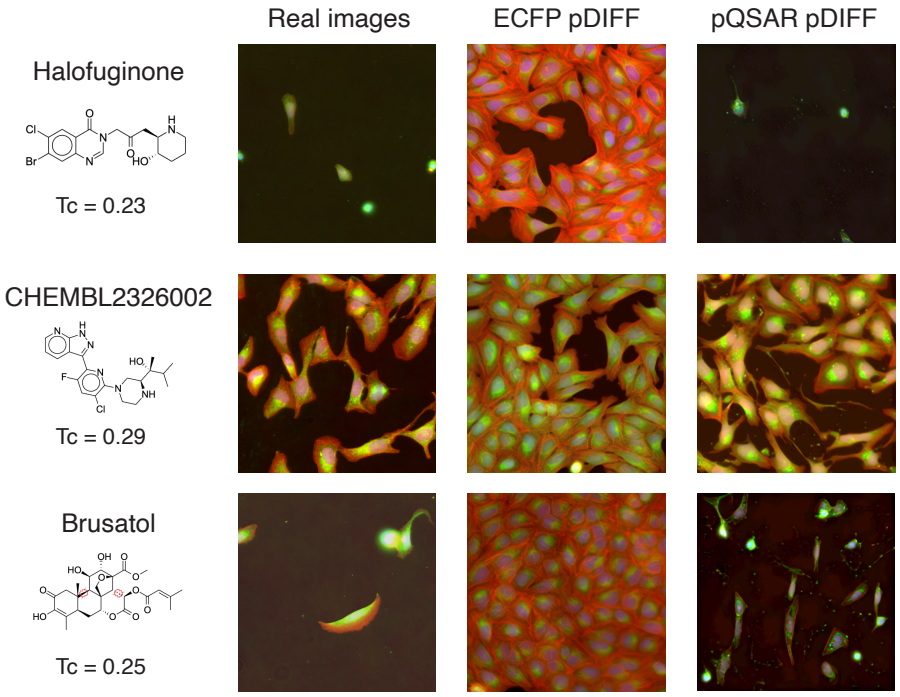


Fig. 2 Molecules in the realistic held-out set along with their corresponding real images (Nuclei, blue; mitochondria, green; red, cytoplasm) as well as images generated with pDIFF models. We show the generated images for models conditioned on structural profiles (ECFP pDIFF) as well as bioactivity profiles (pQSAR pDIFF). For each molecule, we list the Tanimoto coefficient (Tc) to the nearest neighbor molecule in the training set.

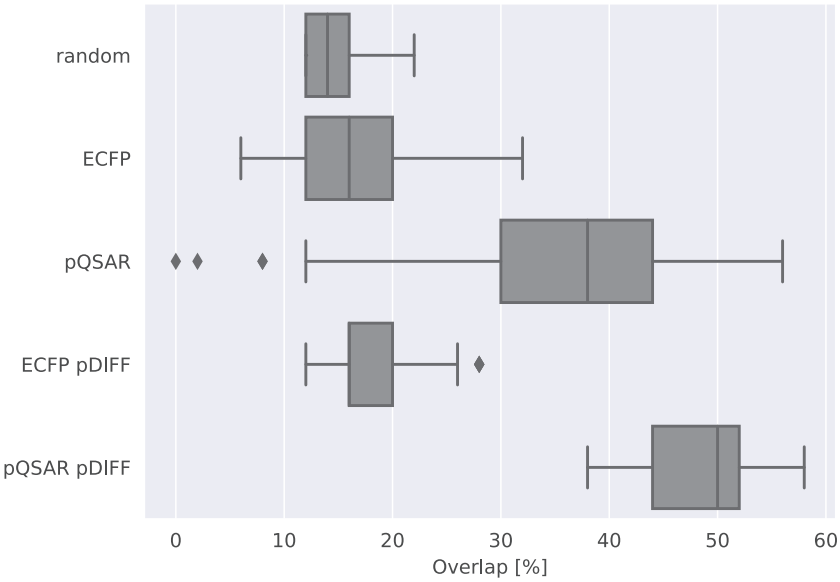


Fig. 3 Distribution of percentage overlap values between 50 ground truth nearest neighbors and those retrieved with virtual expansion approaches for 101 query compounds. We show the performance for an approach selecting molecules randomly ('random') as well as for similarities computed using molecular or bioactivity profiles (ECFP and pQSAR, respectively). The results for approaches computing similarities using image profiles derived from pDIFF generated images with either molecular or bioactivity profiles (ECFP pDIFF or pQSAR 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. The pQSAR-based pDIFF model performs better than the pDIFF model conditioned on structural fingerprints (ECFP).

	Coverage	Cell Count	Cell Size	Cytoplasm	Mitochondria	Nuclei
Real images upper bound	.67	.56	.48	.46	.76	.51
ECFP pDIFF	.13	.10	.04	-.07	.10	-.01
pQSAR pDIFF	.48	.28	.12	.06	.12	.06

7 Methods

7.1 Image Acquisition and Preparation

U2-OS cells were treated with 3750 compounds from the Novartis MoA Box (Canham et al, 2020) at a fixed concentration of 12.5µM and incubated for 24 hours. The Cell Painting staining protocol (Bray et al, 2016) 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 (Peng et al, 2017), 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 Compound fingerprint and profile calculations

To compute the substructure-based molecular fingerprints, we used the extended-connectivity fingerprints (ECFP) algorithm (Rogers and Hahn, 2010) implemented in RDKit resulting in 2048-D count vectors.

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. The profile was zero-padded to reach the 2048-D input required into the pDIFF model.

7.3 Model Architecture

The Stable Diffusion 2.1 model (Rombach et al, 2022) was used as the backbone for pDIFF. The lengths of the cross-attention weights are changed to 2048 (for ECFP and Morgan fingerprints), 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 (Rombach et al, 2022), 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.

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

7.4 Model Training

All models were trained for 30k steps on 4x Nvidia A100 GPUs using huggingface Accelerate, no mixed precision, in distributed data parallel mode. All models are trained on 12 images per training set compound, as the remaining 12 are 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

(Hang et al, 2023) was applied to accelerate convergence, and offset noise (Guttenberg, 2023) 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 3. Code is available at https://github.com/Novartis/pDIFF.

7.5 Inference

The DDPM scheduler (Ho et al, 2020) 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 (Ho and Salimans, 2022) 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 (Ho and Salimans, 2022).

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

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

7.6 Model Validation

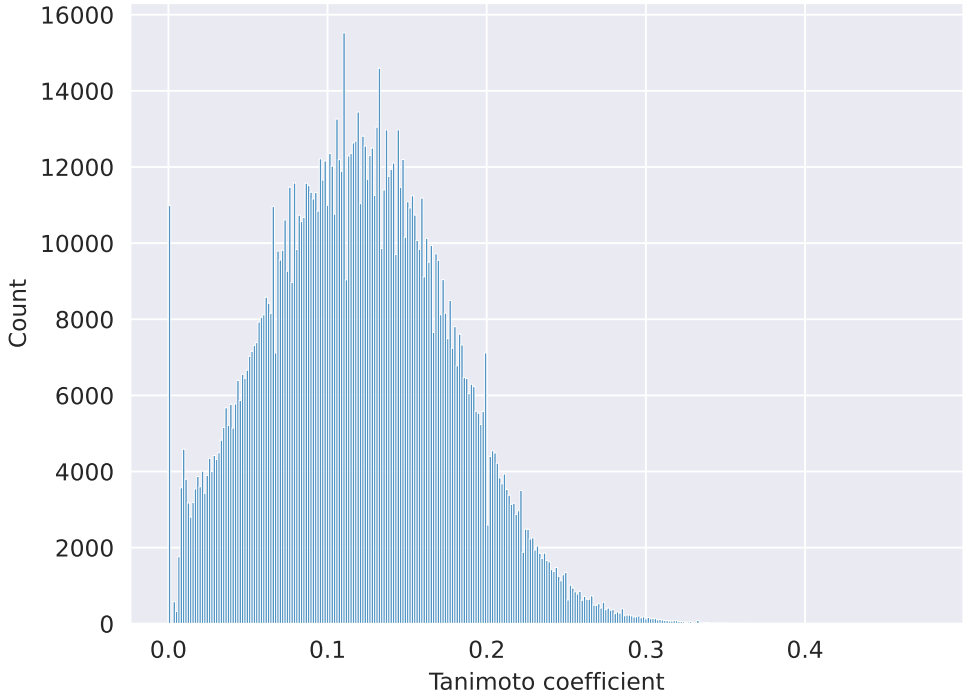
We used the "realistically novel" split (Martin et al, 2019) to evaluate pDIFF. In this approach, the compounds are clustered by similarity, and the clusters ranked in order of size. We allocated compounds from the larger clusters to the training set until 3375 (90%) of the compounds were included. The 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 (Stringer et al, 2021). Following a previous approach (Yang et al, 2021), six 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 also measured the intensity levels in each of three channels for 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.

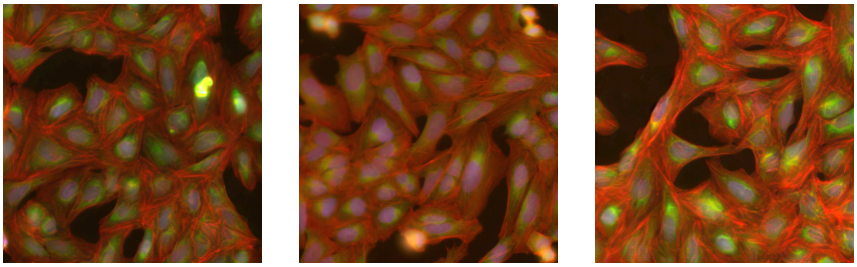
7.7 Reproducing the Training Set

We check that models learned to reproduce phenotypic outcomes by evaluating the performance on the training set. **Supplementary Table** 1 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 correlation values computed between the 12 real images used for model building and the remaining 12 held-out real images of the training molecules. Correlation values among real images across the calculated features range from 0.43 to 0.77. The ECFP pDIFF model yields correlation values between -0.08 to 0.41. The pQSAR pDIFF model shows very realistic performance with correlation values ranging from 0.40 to 0.68.

Supplementary Figures



Supplementary Figure 1 Histogram of Tanimoto coefficients between compounds in the training and held-out set for a total of 1265625 comparisons. The median Tanimoto coefficient is 0.11.



Supplementary Figure 2 Examples images of the neutral control treatments (dimethyl sulfoxide, DMSO).

Supplementary Tables

Supplementary Table 1 Results for held-out images from training compounds. Values given are Spearman correlation coefficients between compound-aggregated average values for each image feature. The first row shows the correlation values between training images and held-out images.

Configuration	Coverage	Cell Count	Cell Size	Cytoplasm	Mitochondria	Nuclei
Real images upper bound	.62	.53	.44	.43	.77	.49
ECFP pDIFF	.41	.36	.15	.04	.34	-.08
pQSAR pDIFF	.59	.46	.40	.40	.68	.42

Supplementary Table 2 Benjamini-Hochberg-adjusted p-values of two-sample Kolmogorov-Smirnov tests. Each test is carried out by comparing the distributions of percentage overlap values of two approaches.

	random	ECFP	pQSAR	ECFP pDIFF	pQSAR pDIFF
random		1.55e-03	3.45e-43	5.24e-07	1.85e-59
ECFP	1.55e-03		5.45e-28	3.97e-03	1.85e-59
pQSAR	3.45e-43	5.45e-28		9.19e-29	2.05e-13
ECFP pDIFF	5.24e-07	3.97e-03	9.19e-29		1.85e-59
pQSAR pDIFF	1.85e-59	1.85e-59	2.05e-13	1.85e-59	

Supplementary Table 3 Training parameters for both ECFP and pQSAR pDIFF models.

Parameter	Value
batch size	6
gradient accumulation steps	10
GPUs	4x Nvidia A100 40GB
effective batch size	240
lr	1e-4
max grad norm	.1
training steps	30,000
training schedule	constant, 1000 step warmup
noise offset	.2
SNR gamma	5.0
optimizer	AdamW
adam beta1, beta2	.9, .999
adam weight decay	1e-2
adam epsilon	1e-8

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hert J, Willett P, Wilton DJ, et al (2004) Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. Journal of Chemical Information and Computer Sciences 44(3):1177–1185

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

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

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

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

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

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

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

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)

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

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

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

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

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

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

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

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

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

Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of Cheminformatics 5(1):26

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

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

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

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

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

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

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

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

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

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

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