Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2023, **2**, 91

Cell morphology-guided *de novo* hit design by conditioning GANs on phenotypic image features†

Paula A. Marin Zapata, ‡*[a] Oscar Méndez-Lucio, ‡[bc] Tuan Le, [a] Carsten Jörn Beese, [d] Jörg Wichard,[e] David Rouquié[b] and Djork-Arné Clevert [a]

Received 3rd August 2022
Accepted 18th November 2022

DOI: 10.1039/d2dd00081d

rsc.li/digitaldiscovery

Developing novel bioactive molecules is time-consuming, costly and rarely successful. As a mitigation strategy, we utilize, for the first time, cellular morphology to directly guide the *de novo* design of small molecules. We trained a conditional generative adversarial network on a set of 30 000 compounds using their cell painting morphological profiles as conditioning. Our model was able to learn chemistry-morphology relationships and influence the generated chemical space according to the morphological profile. We provide evidence for the targeted generation of known agonists when conditioning on gene overexpression profiles, even though no information on biological targets was used during training. Based on a target-agnostic readout, our approach facilitates knowledge transfer between biological pathways and can be used to design bioactives for many targets under one unified framework. Prospective application of this proof-of-concept to larger chemical spaces promises great potential for hit generation in drug and phytopharmaceutical discovery and chemical safety.

Introduction

Recent studies have categorized pharmaceutical research and development (R&D) as a very inefficient process in terms of the high cost and small number of approved molecules.[1,2] Given that the cost of bringing a new drug to market is doubling approximately every 9 years, the current R&D model might not be sustainable in a few more decades.[1] This concern raises the question of how to improve the current pharmaceutical R&D efficiency. Some authors propose to focus on desired biological responses, such as in systems-based approaches, moving away from the currently popular target-based approach.[3–5] The latter is characterized for focusing only on a well-characterized target or mode of action. On the other hand, systems-based drug discovery aims to identify or optimize a molecule with little knowledge of the biological target or mode of action, relying more on phenotypic changes. Recent studies have compared the efficiency of these approaches based on the number of first-in-class drugs approved by the FDA but it is too soon to draw a conclusion due to the long time frame of drug discovery projects.[6,7] The same principles apply to the phytopharmaceutical sector where innovative active substances are needed to find new mode of action in target species (plant, insect, fungi) while having an optimized safety profiles for humans and the environment.[8]

Despite the many advantages of the systems-based approach,[3] only a few experimental settings have been used in large-scale screens. One of these is transcriptomic analysis where the change in gene expression levels caused by perturbation (either chemical or biological) is used as a readout to select active compounds. This approach has been used in the Connectivity Map project[9,10] to connect disease, genes and drugs and has been successfully applied to identify new active molecules,[11–13] drug repurposing,[14,15] and mode of action identification.[16,17] An alternative is to observe the effect of the perturbation at the morphological level rather than at the transcriptomic level. In this regard, the cell painting assay, a new technique based on High-Content Imaging (HCI), has been extremely useful. Cell painting uses six stains to label eight cellular components, thus capturing a good overview of the cellular state.[18] In the context of profiling, images are usually processed by computational pipelines to extract feature representations called morphological profiles[19–24] (reviewed in ref. 25), which serve as phenotype descriptors in further tasks. This technique has been used for drug repurposing,[26] to cluster small molecules by similar phenotypic effect,[19] map cellular morphology to gene function,[27] predict biochemical assays,[28,29] infer modes of action,[30] and characterize toxicity[31] among other applications (reviewed in ref. 32 and 33).

[a]*Bayer AG, Machine Learning Research, Pharmaceuticals, Berlin, Germany. E-mail: paula.marinzapata@bayer.com*

[b]*Bayer SAS, Early Toxicology, Crop Science, Sophia Antipolis, France*

[c]*Bloomoom, 13 Avenue Albert Einstein 69100, Villeurbane, France*

[d]*Leibniz-Forschungsinstitut für Molekulare Pharmakologie, Berlin, Germany*

[e]*Bayer AG, Genetic Toxicology, Pharmaceuticals, Berlin, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2dd00081d

‡ Equal contribution: these authors contributed equally and the order was decided alphabetically.

Interestingly, most cell painting applications in small molecule research have focused on clustering or classification tasks. To our best knowledge, no work has used this information-rich assay to guide *de novo* molecular design in drug discovery.[34,35] *De novo* molecular design aims to propose novel chemical structures with desired profiles, for example, bioactivity on a target protein[36] or optimized molecular properties.[37,38] A popular method in the field is the use of generative adversarial neural networks (GANs),[36] having numerous applications to drug discovery[39–43] and materials sciences.[44] These techniques promise to play a crucial role in pharmaceutical and phytopharmaceutical research by allowing a cost-effective exploration of the vast drug-like chemical space (estimated to be in the range of 10^{23} to 10^{60}).[45]

In our previous work,[46] we proposed to guide the *de novo* generation of active compounds using gene expression signatures. While transcriptional profiling is costly, high content imaging provides a less expensive profiling technology, which is broadly accessible and offers higher throughput. Therefore, in this work, we combine cell painting with generative adversarial networks to design compounds inducing a specific morphological effect. We show that by conditioning a GAN on morphological profiles we can influence the generated chemical space and suggest compounds with high molecular similarity to known agonists without using biological activity annotations in the training set.

Results

Model architecture

To design phenotype tailored compounds, we propose a modular architecture with two main components: a molecular autoencoder and a conditional Wasserstein generative adversarial network (cGAN)[47] (Fig. 1). The goal of the molecular autoencoder is to provide a representation of compounds (molecular embedding) and decode those representations back to compounds. Inspired by recent work encoding molecules into continuous spaces,[38,48,49] we trained a variational autoencoder[50] to reconstruct discrete representations of compounds and used the bottleneck layer as continuous molecular embedding. As input, we chose SELFIES[51] (self-referencing embedded strings) since they offer 100% validity to describe molecular graphs. The autoencoder architecture follows the implementation proposed by Winter *et al.*[48] based on stacked Gated Recurrent Unit (GRU) cells, and it was trained on ChEMBL22 (ref. 52) to cover a large chemical space. The goal of the cGAN is to produce phenotype-tailored compounds using morphological profiles from cell painting images as conditioning input. The model is composed of three neural networks: a generator, a discriminator, and a condition network. The generator takes a random noise vector and a morphological profile to generate a vector in the molecular embedding space targeted to the input profile, the discriminator evaluates if the generated molecular embedding corresponds to a real or generated molecule, and the condition network evaluates whether the molecular embedding matches the morphological profile. The cGAN was trained in an adversarial manner on the compound profiling dataset BBBC036v1,[53] which contains 126 779 morphological profiles induced by 30 616 compounds in U2OS cells (see Methods for more details).

Generating molecules from compound-induced morphological profiles

First, we examined the overall generative performance of our model. To this end, we generated 30 000 molecules conditioned on 30 000 morphological profiles randomly selected from the training set (*i.e.*, 1 molecule per profile) and computed several statistics on generated compounds (Table 1, column Cpds). Although only ~50% of the molecules were valid, most of them were unique (99%), displayed low similarity to the training set (90%), and showed relatively high scaffold diversity (83% Murcko scaffold uniqueness *vs.* 70% in the training set). Furthermore, a Fréchet ChemNet Distance (FCD)[54] in the range of 12–15 shows a modest difference between the distribution of generated molecules and the training set. These results indicate that the generative adversarial network was able to generate novel and diverse sets and did not suffer from the mode-collapse problem commonly observed in this type of model.

To provide a comprehensive characterization of generated molecules, Table 1 also lists physicochemical and drug-like properties, although they were not part of our optimization objective. On average, generated molecules present drug-like physicochemical properties, most of them fulfilling the 'Rule of 5' proposed by Lipinski,[55] which is considered a standard of drug-likeness (molecular weight < 500D, Log P < 5, H-bond donors < 5, H-bond acceptors < 10, and number of rotatable bonds < 10). Accordingly, more than 40% of the generated molecules have a high Quantitative Estimate of Drug-likeness[56] (QED > 0.5), and the similarity between ChEMBL and generated compounds is comparable to that between ChEMBL and the training set in terms of FCD score. In addition, more than 50% of the generated molecules are expected to be synthesizable

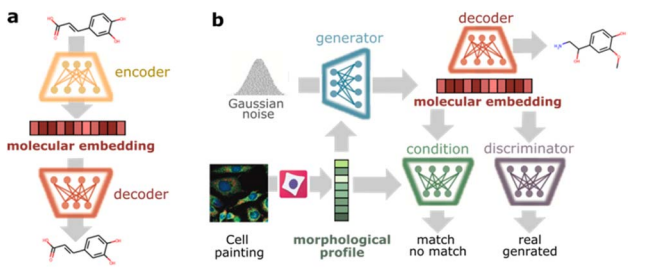

Fig. 1 Model architecture. (a) A variational autoencoder is trained on reconstructing SELFIES representations of small molecules. Encoder: provides a continuous representation of molecules (molecular embedding). Decoder: decodes molecular embeddings into SELFIES. (b) A conditional generative adversarial network is trained to generate molecular embeddings conditioned on morphological profiles. Generator: takes a morphological profile and random noise vector to produce molecular embeddings targeted to the input profile. Discriminator: calculates the probability of a molecular embedding to come from a real or generated molecule. Condition: calculates the probability of a molecular embedding to match the morphological profile.

Table 1 Properties of generated molecules compared to the training set. Molecules were generated in 3 repetitions conditioned on 10 000 randomly selected training set profiles per condition group and repetition. Cpds: compounds, DMSO: neutral controls, %: percentage of valid unique molecules satisfying specified constraints. Values report mean ± standard deviation among 3 repetitions, except for Physchem properties and ChEMBL FCD, which report mean and std among all generated molecules. NN: nearest neighbor, scaff: scaffold, dice sim: dice similarity of Morgan fingerprints, CFD: Fréchet ChemNet distance

		Generated		
Property	Details	Cpds	DMSO	Train
Validity	% Valid	49.8 ± 0.4	56.6 ± 0.3	—
Uniqueness	% Unique	99.9 ± 0.1	99.7 ± 0.1	—
Novelty	% Dice sim to train NN < 0.5	90.8 ± 0.5	81.5 ± 0.4	—
	Train set FCD	15.2 ± 0.025	12.3 ± 0.028	
Diversity	% Unique Murcko scaff	82.8 ± 0.4	83.0 ± 0.4	69.6
Condition match	% Class prob > 0.75	99.3 ± 0.0	97.6 ± 0.2	73.9
	Shuffled samples	9.4 ± 0.6	10.9 ± 0.3	13.5
	Shuffled features	10.3 ± 0.1	8.1 ± 0.3	5.3
Synthesizability	% SA score < 4.5	81.6 ± 0.6	92.3 ± 0.1	96.2
	% RSA score > 0.5	54.0 ± 0.5	74.6 ± 0.4	76.1
	% (RSA > 0.5) & (SA < 4.5)	50.9 ± 0.7	72.6 ± 0.6	75.2
Drug-likeness	ChEMBL FCD	8.5 ± 0.07	4.9 ± 0.05	8.4
	% QED > 0.5	41.5 ± 1.1	63.9 ± 0.2	75.8
Toxicity	% cpds with toxicophores	24.8 ± 0.2	21.9 ± 0.4	12.5
Physchem	Molecular weight	443.6 ± 124.5	391.3 ± 101.8	381.0 ± 96.4
	Log *P*	2.5 ± 1.9	3.0 ± 1.6	3.2 ± 1.3
	HBA + HBD	9.0 ± 3.2	7.0 ± 2.7	6.5 ± 2.3
	NRB	7.6 ± 4.4	5.9 ± 3.3	4.9 ± 2.4

since they display an acceptable Retrosynthetic Accessibility Score[57] (RSA > 0.5) and a low Synthetic Accessibility score[58] (SA < 4.5). Finally, only less than 25% of the generated molecules contained one or more toxicophores, *i.e.*, substructures that are highly correlated with adverse effects.[59] Together, these findings demonstrate that our model can generate molecules with drug-like characteristics.

As a preliminary assessment of the effect of morphological conditioning, we compared conditioning on training set compounds *vs.* conditioning on Dimethyl Sulfoxide (DMSO) negative controls, finding subtle differences between both groups. DMSO-conditioned molecules (Table 1, column DMSO) display higher validity, synthesizability and slightly closer physicochemical properties to the training set than compound-conditioned molecules (Table 1, column Cpds). However, both groups display similar diversity, uniqueness, and toxicophore content. The high similarity of physicochemical properties between DMSO-conditioned molecules and the training set might be due to the large proportion of compounds without distinguishable morphological signature in the training set. This is evident from the overlap between morphological profiles from compounds and DMSO (ESI Fig. 1†), leading to an increased density of training samples near the DMSO phenotypic neighborhood. Such overlap also advocates for a more fine-grained stratification of morphological profiles to judge the effect of morphological conditioning.

Translating morphological similarity into chemical similarity

Next, we asked whether morphological conditioning could influence the chemical space of the generated compounds in a sensible manner. To answer this, we clustered morphological features and compared the chemical similarity of the generated molecules conditioned on profiles from different clusters (more details in Methods). To get an impression of the cluster phenotypes and their proximity in the feature space, we created a minimum spanning tree (MST) based on the cluster centroids (Fig. 2a) and displayed representative images of the closest neighbors to selected centroids (Fig. 2b). For easy reference, clusters were enumerated by increasing distance to the cluster with most DMSO control samples (cluster 0).

Comparison of the generated chemical space between cluster pairs showed that our model effectively translates morphological similarity into chemical similarity and even preserves distance relationships between both levels. Taking cluster 0 as a reference and following the hierarchical paths proposed by the spanning tree, we observe a reduced chemical similarity to cluster 0 for increasing distances in the morphological space (Fig. 2c). This trend holds for most pairwise comparisons (except clusters 11 and 15) and is observable based on molecular fingerprints, as well as molecular embeddings. The latter indicates that distances in the chemical embedding space learned by the model are in accordance with established metrics of chemical similarity. Additionally, we used t-map[60] projections of Morgan fingerprints to visualize differences in the generated chemical space between pairs of nearby (Fig. 2d) and distant (Fig. 2e) clusters judging from the MST. While proximal clusters do not segregate well, pairs of distant clusters display a clear differentiation of their chemical space. Moreover, visualization of 100 000 generated molecules from all clusters depicts a good distinction between the 10 closest and 10 farthest clusters to the DMSO control reference (Fig. 2f). Similar segregation patterns were observed among the training

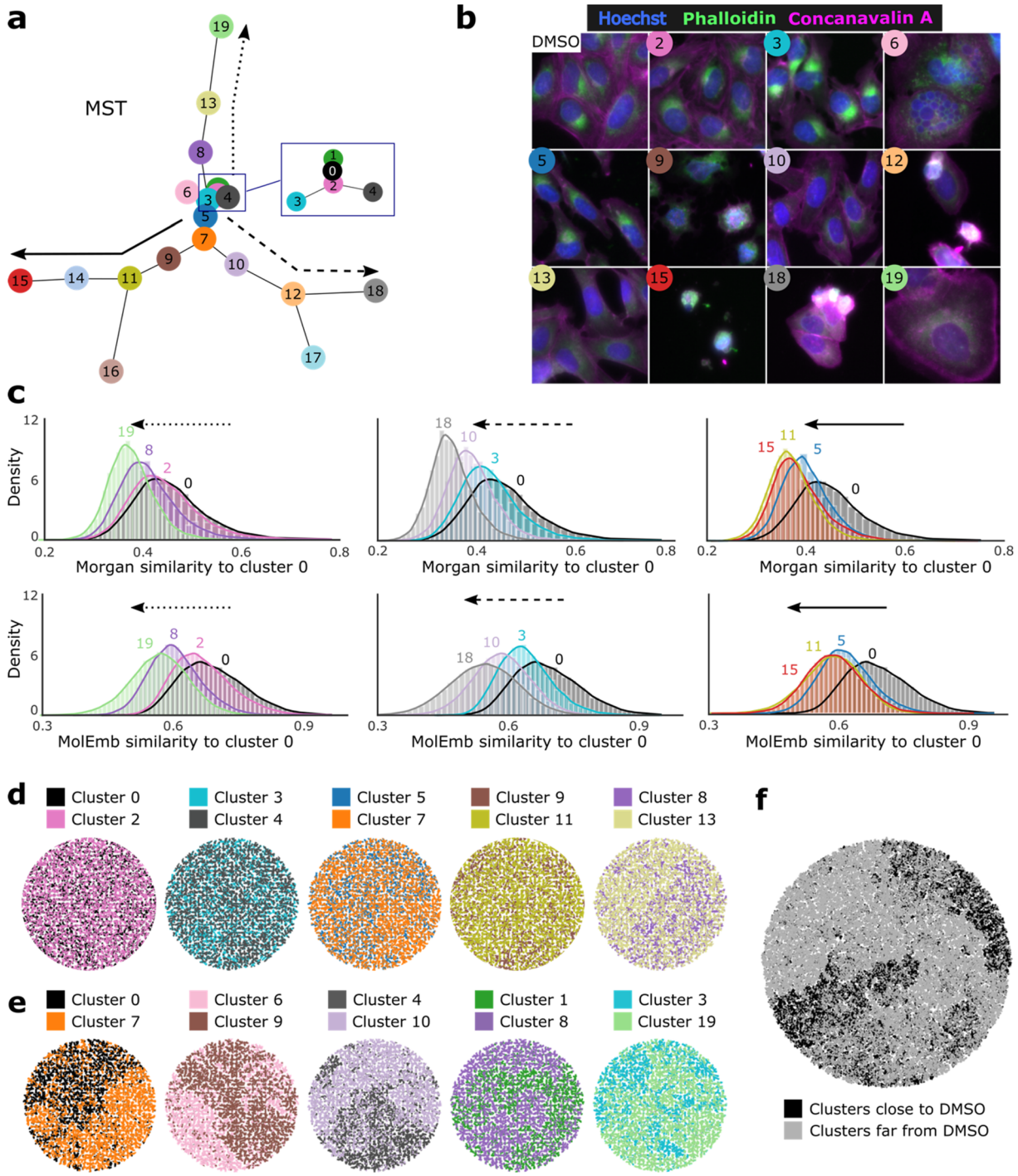

Fig. 2 Translation of phenotypic similarity to generated chemical similarity observed from *k*-means clustering of training set profiles. (a) MST: minimum spanning tree calculated on cluster centroids. Cluster 0 has the largest proportion of DMSO samples, and all other clusters are enumerated by increasing distance to cluster 0. (b) Representative cell painting images of the closest samples to selected cluster centroids. Only 3 out of the 5 fluorescent channels from the original dataset (BBBC036v1) are displayed. Hoechst: nucleus (blue), phalloidin: actin, golgi and plasma membrane (magenta), concanavalin A: ER (green). (c) Distributions of the chemical similarity between generated molecules conditioned on cluster 0 and selected clusters along the arrow paths marked in (a). Values report pairwise Morgan dice similarities and molecular embedding cosine similarities between all generated molecules of a given cluster and their closest neighbor in cluster 0. (d–f) t-map projections of Morgan fingerprints of generated molecules, color-coded by the cluster of their conditioning profile (5000 randomly selected samples per cluster). (d) Pairs of clusters that are direct neighbors in the MST. (e) Pairs of clusters distant in the MST. (f) All clusters (100 000 molecules) binary categorized as the 10 closest and 10 farthest centroids from cluster 0.

set molecules (ESI Fig. 2a†), although no obvious morphology-chemistry associations could be observed from bulk, unclustered profiles (ESI Fig. 2b and c†). Together, these results indicate that our model learned morphology-chemistry relationships from the training set and can manipulate the generated chemical space through its morphological conditioning in a specific and logical manner.

Generating molecules from genetic perturbations

We sought to find evidence for the generalizability of the model and biological relevance of generated molecules by conditioning on profiles outside the training set and comparing them to known bioactive molecules. We used profiles from the gene overexpression profiling dataset BBBC037v1 (ref. 27) and compared generated molecules to known agonists retrieved from the ExCAPE database,[61] following the reasoning that overexpression should resemble agonism. We could retrieve known agonists for 9 genes out of 220 overexpressed genes (summarized in ESI Table 1†), making sure to exclude those compounds present in the training set.

First, we visualized the overexpression phenotypes to inspect the strength of the morphological signal (Fig. 3 and ESI Fig. 3†). Profile projections indicate that only NFKB1, BRCA1 and HSPA5 strongly to mildly differ from DMSO controls, while other genes show low differentiation (TP53, CREBBP, STAT1, STAT3, HIF1A, NFKBIA). Nevertheless, generated molecules conditioned on most genes showed some differentiation compared to conditioning on DMSO (ESI Fig. 4†). We include all genes in subsequent analyses, later confirming that only differentiable genes lead to conclusive results. Good overlap between DMSO controls from the overexpression (test set) and training set is also evident, corroborating the comparability of both datasets after applying quantile transform (see Methods).

Scaffold enrichment analysis identifies morphology-tailored molecular substructures

Next, we generated molecules conditioned on profiles from overexpressed genes (see methods) and examined if the generated molecules were similar to known agonists. A key question is whether resemblance to a known agonist happens due to the conditioning or by chance, since the larger the generated sample size, the higher the probability to find an active-like molecule. To address this, we first performed an enrichment analysis to select scaffolds that are significantly overrepresented for each gene compared to DMSO, thus ensuring, to a certain extent, the causality of the morphological conditioning (see Methods). Subsequently, we checked whether known agonists contain enriched scaffolds from the corresponding gene. Highlighting the fairness of our scaffold selection process is the fact that it does not guarantee a given number of scaffolds per target (if any).

On average, generated molecules conditioned on overexpressed genes displayed 96% and 79% compound and scaffold uniqueness, respectively (ESI Table 1†), closely following the results from the training set. Scaffold enrichment analysis relative to DMSO showed that overrepresented scaffolds were consistent and slightly discriminative. Within a given gene conditioning, we observed groups of scaffolds sharing a good proportion of molecular substructures (ESI Fig. 5†), reassuring that the model learned chemical patterns. Between different genes, we observed some diversity in enriched scaffolds, although there was a frequent pattern of two aromatic rings linked by an aliphatic chain of variable length.

Among the 9 tested genes, we could identify ExCAPE agonists with enriched scaffolds for only two genes (BRCA1 and NFKB1) which interestingly, displayed the strongest phenotypic signal (Fig. 3). For BRAC1, we proposed 8 active scaffolds, among which, 1 was found in a potent BRCA1 agonist (hit rate of 12.5%). For NFKB1, 2 out of 25 proposed scaffolds were found

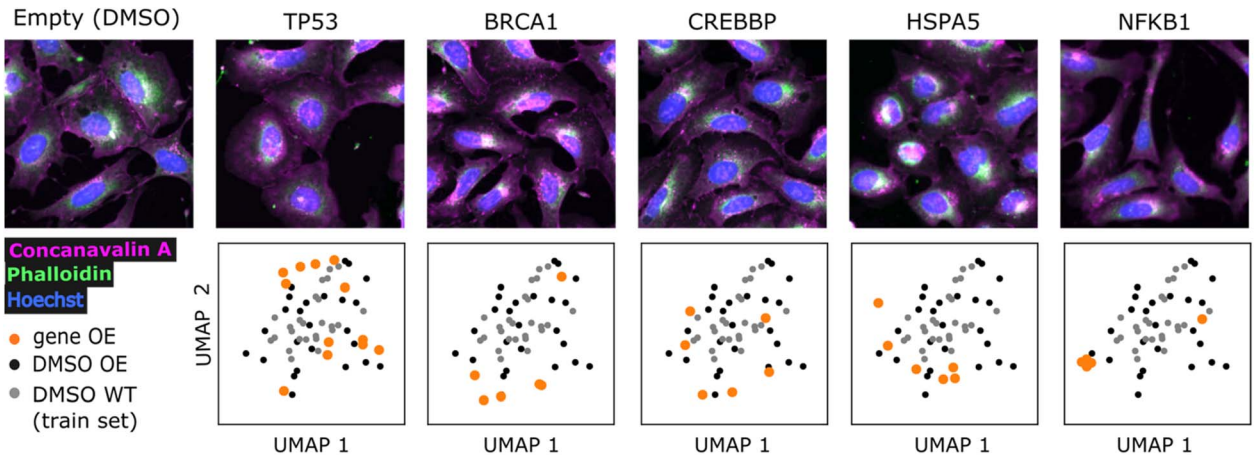

Fig. 3 Overexpression phenotypes. Representative images and UMAP projections of phenotypic profiles for overexpressed genes (gene OE) and 50 randomly selected DMSO controls. DMSO WT: neutral controls from the training set. DMSO OE: empty vector controls from the overexpression dataset (test set). The 5 genes with most ExCAPE agonists are presented. Only 3 out of the 5 fluorescent channels from the original dataset (BBBC037v1) are displayed. Hoechst: nucleus (blue), phalloidin: actin, golgi and plasma membrane (magenta), concanavalin A: ER (green).

in an agonist (hit rate of 8%). The training set provides a comparison baseline, where 7.6% and 0.23% of 9750 unique scaffolds were found in BRCA1 and NFKB1 agonists, respectively. Thus, we provide a respective increase in hit rate of 1.6 and 34-fold, but most importantly, we dramatically reduce the search possibilities (from 9750 to just 8 or 25). The small number of proposed active scaffolds would make it feasible to prospectively conduct experimental validation. However, since chemical synthesis of *de novo* molecules is time and resource-consuming, we consider it is out of scope for this proof-of-concept work.

Comparative analyses reassure attribution to the morphological conditioning

Having identified 3 active scaffolds suggested by the model, we further scrutinize the contribution of the morphological conditioning by checking the specificity with which active-like molecules are generated. For a given agonist, we compared the number of similar generated molecules conditioned on the respective gene *vs.* the remaining 8 overexpressed genes and DMSO. To provide a robust comparison, we also included a set of 10 highly differentiable genes (see Methods) and report counts for 3 Morgan fingerprint similarity thresholds (0.6, 0.7 and 0.8) to designate active-like compounds (Fig. 4). Results are only reported for 2 of the ExCAPE agonists since no similar molecules were found for one of the NFKB1 compounds (ESI Fig. 6†).

Significant differences in the frequency of active-like molecules among the 20 tested genes reveal good specificity of the generation process towards the corresponding gene targets (Fig. 4a). Both BRCA1 and NFKB1 agonists displayed significantly increased active-like counts when conditioning on their respective gene compared to any other gene, with only NFKB1 lacking significance *vs.* NFKBIA for the lowest similarity threshold (0.6). Besides this, active-like generated compounds showed different substituents and side chains at several locations, suggesting that the model can mimic diverse medicinal chemistry decorations (Fig. 4b). Interestingly, one of the generated compounds for BRCA1 was a perfect match to the known agonist (highlighted in Fig. 4b).

As a second check, we used the condition network to quantify changes in the matching probability towards both ExCAPE agonists as a function of the phenotypic distance to their respective gene overexpression profile (ESI Fig. 7†). Following our expectations, the classification probability towards the NFKB1 agonist strongly decreased for training set and overexpression profiles with increasing cosine distance to the NFKB1 overexpression profile. A similar but milder trend was observed for BRCA1, which as formerly pointed out, has a less distinctive morphological signature. Overall, these results

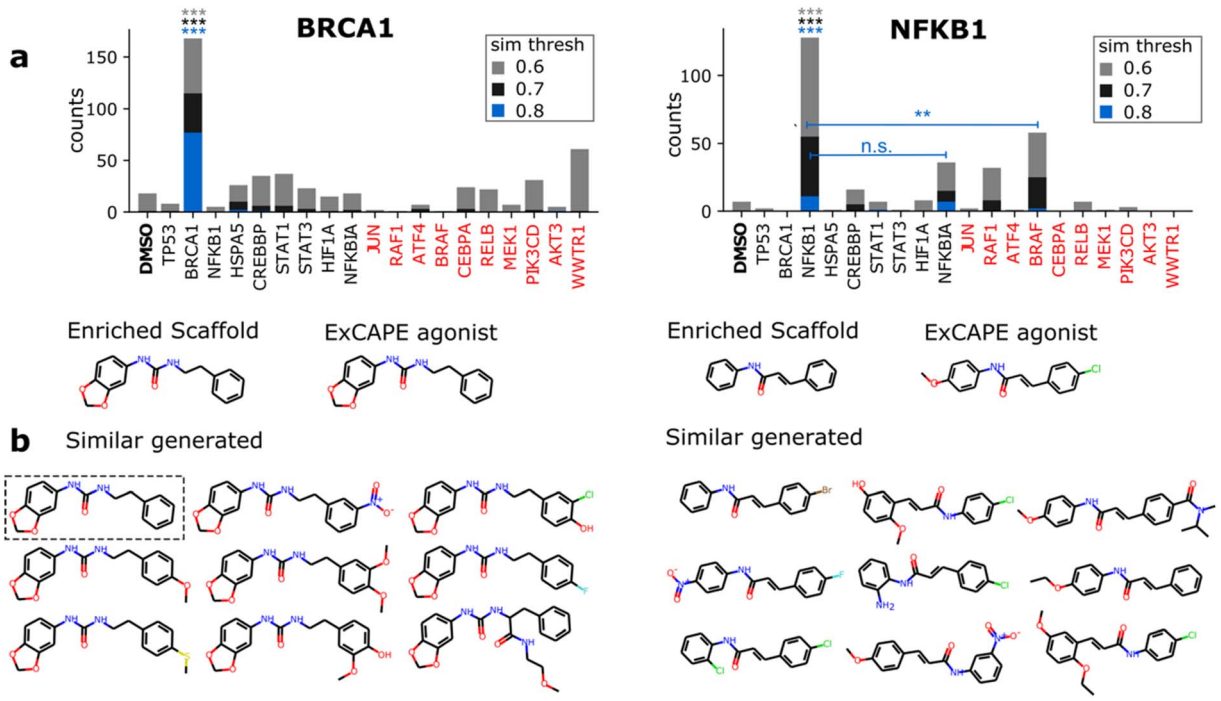

Fig. 4 Specificity of generated active-like molecules for ExCAPE agonists containing significantly enriched scaffolds. (a) Number of generated molecules with Morgan dice similarity to the specified agonist above several thresholds (color-coded) conditioned on profiles from overexpressed genes and DMSO (empty control). The set of genes contains those with known ExCAPE agonists (black) plus the 10 most differentiable genes (red). Statistical significance was determined with Fisher's exact test on the subset of 200 most similar-to-active molecules for each gene. Reported significance follows the color code of similarity thresholds. ***: $p < 0.001$, **: $p < 0.01$, n.s: non-significant. Stars above BRCA1 and NFKB1 summarize significance against all genes, except those genes indicated otherwise by horizontal lines. (b) Representative generated molecules with Morgan dice similarity above 0.6 to the indicated agonist. Molecules were generated conditioned on the specified gene. Gray dotted box highlights a perfect match to the agonist.

encourage the idea that the generation of active-like molecules was attributed to the morphological conditioning.

Interpolation of chemical and morphological spaces

A benefit of using continuous chemical descriptors as well as continuous biological readouts is that it allows us to smoothly navigate both spaces. We leverage this to further assess the behavior of our model with respect to the two selected ExCAPE agonists, performing linear interpolation and monitoring the predictions of the condition network. Concretely, we study how the matching probability towards a given gene changes as we navigate the chemical space (Fig. 5a), and conversely, how the matching probability towards a given compound changes as we navigate the phenotypic space (Fig. 5b).

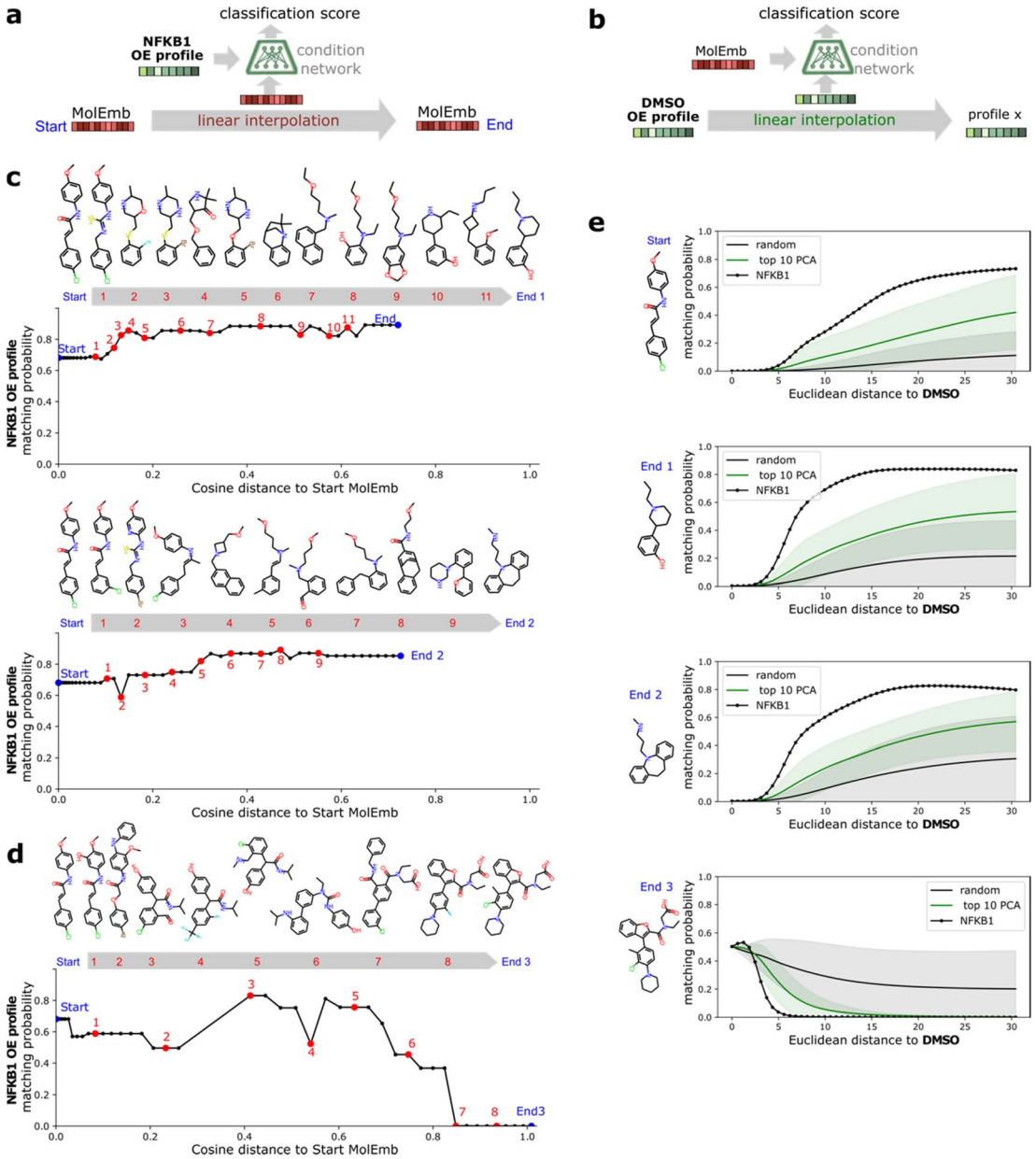

Fig. 5 Interpolation of chemical and morphological spaces and its effect on the condition match for NFKB1. (a) Schematic of the chemical interpolation experiment. The condition network is used to compute the matching probability between the median NFKB1 overexpression profile and selected molecular embeddings along a linear interpolation trajectory between a start and end molecule. Larger gaps between interpolation points correspond to regions in the molecular embedding space without valid decoded molecules. (b) Positive direction: end molecules are other NFKB1 ExCAPE agonists with higher matching probability than the start molecule. (c) Negative direction: end molecule is a random generated molecule with lower matching probability than the start molecule. (d) Schematic of the morphological interpolation experiment. The condition network is used to compute the matching probability between the molecular embedding from a selected compound and morphological profiles along linear interpolation trajectories. (e) Interpolation curves between DMSO and NFKB1, 100 random directions or the top 10 axes of variation determined by PCA. Continuous lines and shadows report mean and std, respectively. Selected compounds correspond to those used in (b and c) for chemical interpolation.

For interpolation of the chemical space, we pick the NFKB1 and BRCA1 ExCAPE agonists with enriched scaffolds as starting points and evaluate the matching probability against NFKB1 and BRCA1 overexpression profiles, respectively (see Methods). To cover increased and decreased activity directions, we choose interpolation endpoints from other ExCAPE agonists outside the training set with higher matching probability towards the corresponding gene (Fig. 5c) as well as random generated molecules with lower matching probability (Fig. 5d). Further examples are provided in Fig. 8.† Chemical interpolation experiments show that incremental shifts in the molecular embedding correspond to smooth transitions in the discrete chemical space and induce sensible changes in the output from the condition network. The matching probability towards NFKB1 and BRCA1 overexpression profiles increased progressively along interpolation trajectories towards increased activity directions, reinforcing the belief that our model learned meaningful trends instead of isolated phenotype-compound matching occurrences. Interpolation towards decreased activity directions showed more irregular responses, possibly due to the randomness of the selected direction and increased distance to the start molecule.

For interpolation of the morphological space, we follow the matching probability against the compounds used in Fig. 5c and d and pick the DMSO profile from the overexpression dataset as starting point (Fig. 5e) and NFKB1 and BRCA1 profiles as endpoints. As comparison baselines, we also include trajectories towards random directions or along the axes of highest variation in the training set (see Methods). Phenotypic interpolation revealed a dose–response towards the distance to DMSO and confirmed boosted matching in the direction of overexpression profiles. The matching probability against the 3 NFKB1 agonists (start, end1, end2 in Fig. 5e) increased from DMSO towards the NFKB1 phenotype, while decreasing for the NFKB1 inactive molecule (end3). Interestingly, the DMSO probability was zero for the 3 agonists, indicating that they were recognized as bioactives even though neither bioactivity labels nor DMSO profiles were used during training. Notably, there was a general trend towards higher matching probabilities for larger distances to DMSO (random curve). However, the trend was more pronounced along axes of highest variation (top 10 PCA) and even more towards the NFKB1 phenotype, with similar results for BRCA1 (Fig. 9†). This suggests that the model does not simply rely on overall phenotypic strength but displays some specificity towards different regions in the morphological space.

Together, interpolation experiments highlight the quality of our molecular descriptors and the soundness of the morphology-chemistry associations learned by the model. Moreover, they demonstrate the potential of our framework for molecular optimization.

Discussion

We propose an approach to generate phenotype-customized molecules by conditioning a generative adversarial network on cell painting morphological profiles and demonstrate how our model can influence the chemical space and propose tailored bioactives without prior information of biological targets. By using a target-agnostic readout, we learn general relationships between cell morphology and molecular structure and intrinsically provide a framework for knowledge transfer between diverse biological pathways. Our work provides a first step towards the systematic use of high content imaging for molecular design.

To our knowledge, we are the first to use cell painting data to guide the *de novo* design of small molecules. A recent work tackled the inverse problem *i.e.*, synthesizing cell painting images conditioned on molecular structures.[62] We believe that our approach is more applicable to small molecule discovery since it directly proposes compounds given a desired biological response and allows straightforward integration of genetic perturbations. Other related works have used gene expression profiles to guide *de novo* molecular design.[46,63] However, given the low cost of high content imaging, the large-scale data generation needed to bring these methods to their full potential will favor the morphological over the transcriptomic domain. For instance, joint initiatives, such as the JUMP-CP consortium, promise to release morphological profiling data for over 140 000 chemical and genetic perturbations.[64]

Evaluating the significance of model outputs in generative chemistry is not straightforward.[65] In this proof-of-concept, we focused on demonstrating a sensible influence of morphological conditioning on the generated chemical space and evaluating the relative merit of the conditioning. Using gene overexpression profiles, we provide evidence of the targeted generation of agonist compounds, where we could attribute the overrepresentation of active-like substructures to morphological conditioning. However, assessment of the potentiality of our approach to generate truly novel compounds inducing a desired phenotype will require larger datasets or experimental testing. Our evaluation strategy relied on public datasets and was based on the premise that gene overexpression should resemble agonism. We were therefore limited by the low number of agonists we could retrieve from public resources, which reduced our evaluation set to only 9 genes, many of them with rather weak phenotypes. Although gene knock-down perturbations would have certainly provided more reference compounds (in this case antagonists), we avoided using public cell painting RNAi data due to the reported seeding effects reflected in poor specificity.[66]

Agreeing with our expectations, we only found enriched scaffolds present in ExCAPE agonists for a subset of two genes displaying good morphological differentiation from DMSO controls (NFKB1 and BRCA1). A deeper look into the ExCAPE compounds shows that both agonists have strong activity towards their target. The NFKB1 agonist was identified as a potent activator of the NFKB1 pathway in NFkB-bla cells (PubChem AID: 928, PubChem CID: 1628214), with half-maximal efficacy (EC50) of 1.4 μM. The BRCA1 agonist showed high efficacy on a BRCA1 expression activation assay in MCF7 cells (EC50 = 3.4 μM), although the activity value was calculated from only a partial Hill curve reaching ~80% efficacy at the highest concentration (PubChem AID: 624 202, PubChem

CID: 6465485). Interestingly, this assay directly measured BRCA1 expression levels, matching the overexpression perturbation on which we based our analysis. The fact that both agonists showed micromolar potencies in cell-based assays reassures their functionality activating their respective target in a relevant biological context.

A notable aspect of our approach is the continuous nature of our chemical and biological descriptors, which enables efficient exploration and interpolation of both spaces. We exploited this to confirm the consistency of our model predictions regarding the two selected ExCAPE agonists, revealing logical dose-like effects and smooth, meaningful changes in chemical structure. These observations reinforce the notion that the observed phenotype-chemistry matches resulted from relevant associations learned by the model. Furthermore, they underline the quality of our molecular representations and provide useful evidence about the remarkable performance of SELFIES, as it was recently reported in neural translation tasks for molecular representation.[67]

Conditioning on morphological profiles opens many possibilities inherent to profiling itself. The multiparametric, target-agnostic nature of the readout allows for capturing complex biological responses, being particularly helpful when experimental endpoints are not easily defined. Moreover, morphological profiles can implicitly capture toxicity or off-target effects, thereby providing an alternative to multi-objective optimization for *de novo* molecular design and chemical de-risking. Finally, having a standardized cell painting protocol will potentially ameliorate the lack of comparability between experimental set-ups and dependence of labels on external context, two common difficulties in modeling biological systems.[68] It is however worth noting that we observed differences between the distributions of morphological profiles from chemical and genetic perturbations, even though both datasets were produced in the same research group. Although distributions became more comparable after applying a quantile transformation, this emphasizes the need for profile harmonization methods to enable the collaborative assembly of large morphological datasets.

Conclusions

We were able to combine generative models with cellular morphology information to design compounds that potentially induce a desired biological response. Our work bridges the fields of *de novo* molecular design and morphological profiling, taking the first steps into an exciting wave of system-based approaches which promise to overcome reductionist, single-target views of biological systems. By using larger datasets, we expect that we will be able to push our approach beyond proof-of-concept and ultimately, accelerate small molecule discovery by providing fast access to innovative chemistry influencing diverse biological pathways.

Methods

Model architecture and training

Molecular autoencoder. A variational molecular autoencoder was trained to reconstruct SELFIES (as one-hot encoding). To compute SELFIES, all molecular structures were pre-processed using the MolStandardize module from RDKIT.[69] First, SMILES were standardized by removing hydrogens, sanitizing the molecules, disconnecting metals and normalizing the structure. Then, only the largest fragment was kept, charges were neutralized and the uncharged molecules were transformed back into a non-isomeric form of canonical SMILES. SMILES were finally transformed to SELFIES in a one-hot encoding format just keeping those tokens that appear in at least 100 molecules from the 1.5 million contained in ChEMBL22 (ref. 52) (ESI Fig. 10†). Molecules containing different SELFIES tokens to the ones selected previously or SMILES strings larger than 120 characters were removed from the dataset. The encoder is composed of three stacked Gate Recurrent Unit (GRU). Their resulting cell states are concatenated and fed into two independent fully connected layers which, following the variational approach, predict the mean and standard deviation of the distribution from which the latent vectors will be sampled. After sampling, a tanh activation function produces the final molecular embedding of 256 dimensions with values between −1 and 1. The decoder takes as input the latent vectors which are fed into a fully connected layer to expand them from 256 to 768 dimensions. This vector is split into three vectors of 256 dimensions and used as the initial state of another set of three stacked GRU cells. The output of the stacked GRU cells was followed by a dropout layer (rate of 0.2) and a dense layer (with a softmax activation function), which generates a probability distribution over all possible SELFIES tokens for each time step. This variational molecular autoencoder was trained following a teacher-forcing[70] scheme during 10 epochs on 1.25 million molecules extracted from ChEMBL22,[52] complemented by the molecules in the training set.

Generator. The generator receives as input the conditioning (morphological profile with 1449 features) and a 1000-dimensional noise vector sampled from a normal distribution. The input noise is processed by a 2-layer multilayer-perceptron (MLP) with 512 and 256 nodes, respectively, where each layer uses LeakyRelu as an activation function. The morphological features are processed by an MLP of size [1024, 512, 256] also using LeakyRelu after each layer. The two resulting tensors are concatenated and used as input for another 2-layer MLP, where the first layer has 256 nodes and a LeakyRelu activation function. The second layer acts as an output layer (*i.e.* the number of nodes is equal to the dimensionality of the latent space) and is followed by a tanh activation function.

Discriminator. The discriminator is composed of a 4-layer MLP of [256, 256, 256, 1] hidden units with LeakyRelu activation function in the first three layers. To reduce overfitting, dropout with a rate of 0.4 was used between the second and third hidden layers and between the third and the last layer of the MLP.

Conditional network. The morphological profile is processed by a MLP of 3 layers with 1024, 512 and 256 units, respectively, and then regularized by a dropout layer with a rate of 0.4. The latent space coordinates of the compound are also fed into a 2-layer MLP with dimension [256, 256] and finalized with a dropout layer. The outputs of these two MLP, corresponding

to the processed morphological profile and compound information, were concatenated and used as input of an MLP of size [256, 1] using LeakyRelu and sigmoid activation functions, which estimates the probability of a molecule to produce a certain morphological profile.

Model training. The conditional generative adversarial network was trained using a Wasserstein loss.[71] The loss functions for the generator ($G_0(z, c)$) and discriminator ($D_0(x)$) are:

$$L_{D_0} = E_{\tilde{x}p_{\text{real}}}[-D_0(x)] + E_{\tilde{z}p_z,\, \tilde{c}p_{\text{real}}}[D_0(G_0(z,\ c))]$$

$$+\lambda E_{\hat{x}p_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D_0(\hat{x})\|_2 - 1\right)^2\right], L_{G_0}$$
$$= E_{\tilde{z}p_z,\, \tilde{c}p_{\text{real}}}[\, -D_0(G_0(z,\ c))$$
$$- \alpha \log(f_0(G_0(z,\ c),\ c))]$$

where x and c are a molecule representation and a morphology profile, respectively, sampled from the real data distribution p_{real}, z is a vector with random noise sampled from a Gaussian distribution (p_z) and f_0 is a function (in this case, a neural network) that measures the probability of a morphology profile to correspond to a molecular representation. The λ and α terms are regularization parameters, both empirically set to 10. λ weights the influence of the gradient penalty in the discriminator loss. α term weights the influence of f_0 in the generator loss.

The model was trained for 500 epochs with a batch size of 256 (496 steps per epoch). The discriminator was updated after each step, whereas the generator every 10 steps. The network was trained using the RMSprop optimizer with a learning rate of 1×10^{-4} for both the generator and discriminator. During training, we monitored the similarity between real and generated molecular representations using Fréchet distance[49] (ESI Fig. 11†). The weights of the condition network were pre-trained on a binary cross-entropy loss and frozen during the GAN training process. All neural networks were built and trained using Tensorflow 1.14.[72]

Morphological profiles

Training set. We used the BBC036v1 (ref. 53) dataset available from the Broad Bioimage Benchmark Collection, consisting of U2OS cells treated with 30 616 compounds at 10 μM, 24 h and 4 replicates, and imaged with the cell painting protocol. Per-well profiles were downloaded from Bray *et al.*[53] and normalized relative to negative controls per plate using robust *z*-scoring. *i.e.*, subtraction of DMSO median and division by DMSO median absolute deviation (mad). 334 Features with zero mad were removed, reducing the number of features to 1449. Profiles were preprocessed with the scikit-learn[73] implementation of a quantile transformer fitted on the training set to improve the comparability between training and overexpression datasets. In total, the training set consisted of 126 779 morphological profiles which were treated independently.

Overexpression set. We used the BBBC037v1 (ref. 27) dataset available from the Broad Bioimage Benchmark Collection, covering the overexpression of 220 genes in U2OS cells imaged with the cell painting protocol and 5 replicates (all ORFs per gene were considered). Illumination corrected images were obtained from The Image Data Resource (IDR) web API. We closely followed the feature extraction protocol from Bray *et al.*[53] to produce comparable features to the training set: single-cell profiles were calculated with CellProfiler 2.2.0 using the analysis.cppipe pipeline, which was slightly modified to take input csv files. Per-well profiles were computed as the median among all cells from all fields-of-view per well. As a sanity check, we confirmed that our feature extraction methodology was able to reproduce the single-cell and population-averaged profiles reported by Bray *et al.*[53] for a randomly chosen plate (plate number 25690). Profile normalization followed as described for the training set.

Molecule generation and post inference analyses

Molecule generation from morphological clusters. *k*-means clustering was performed on the median profiles per standardized SMILES plus a random selection of 1% DMSO profiles using scikit-learn[73] with 20 clusters and Euclidean distance. The MST was computed with the python library Networkx using a kamada_kawai layout. 15 000 valid molecules were generated based on a random sample of 500 profiles per cluster (or less for clusters with less than 500 samples).

Molecule generation from overexpression profiles. 20 000, valid molecules passing custom physicochemical filters were generated per gene using per-well profiles. Physicochemical filters were applied to enhance similarity to drug-like molecules and included: $-2 < \log P < 7$, $-2 <$ Mol weight < 7, H acceptors + H donors < 10, TPSA < 150, rotatable bonds < 150, and Sure ChEMBL alerts $= 0$. The generation experiment was repeated 3 times using different random seeds.

Scaffold enrichment analysis. For each overexpressed gene and DMSO, Murcko scaffolds were computed with RDKIT[69] for all generated molecules, and the number of molecules containing each scaffold was counted. For each scaffold, count data was used to test whether the scaffold was significantly overrepresented in molecules coming from the conditioned gene *vs.* DMSO, using Fisher's exact test ($p = 0.01$). Only scaffolds that were consistently enriched in 3 generation repetitions and larger than 15 atoms were kept. Enrichment analysis was also performed on a set of 10 highly differentiable genes which were selected as follows: for each gene, pairwise distances between profiles from the same gene (intra distances) and each other gene (inter distances) were computed. For each gene pair, a *t*-test ($p = 0.05$) was used to test for significant differences between intra and inter distances. The 10 genes with the most statistically significant comparisons were selected. All statistical tests were performed with the scipy.stats Python library.

Molecular properties. All molecular properties, SA and QED scores were calculated with RDKIT. Toxicity alerts were estimated with a substructure match to Sure ChEMBL Alerts.[59] The Chembl Fréchet Inception scores[54] were calculated with respect to the subset of ChEMBL provided by the GuacaMol Benchmark.[74] The retrosynthetic accessibility score (RSA) was estimated using the recently developed model by Thakkar *et al.*,[57]

which predicts whether a compound would pass a synthetic planning route, based on AiSynthFinder.[75]

Linear interpolation experiments. Interpolation trajectories ($\boldsymbol{y}$) were computed using 50 equally spaced steps between the start ($\boldsymbol{y}_0$) and end vector ($\boldsymbol{y}_e$), *i.e.*, $\boldsymbol{y}_i = \boldsymbol{y}_0 + i(\boldsymbol{y}_e - \boldsymbol{y}_0)/49$, with i in {0,1,…,49}. For chemical interpolation, whenever $\boldsymbol{y}_i$ did not decode to a valid molecule, we added a random noise vector ($\boldsymbol{y}_i'$) and reran the embedding to SELFIES decoding step. This was repeated until a valid molecule was found or up to maximum of 20 iterations. $\boldsymbol{y}_i'$ was sampled from a uniform distribution between 0, and 2% of the per-dimension values. Morphological interpolation was performed on quantile transformed profiles. To increase comparability between all trajectories starting from DMSO towards different directions, we subtract the median DMSO profile from the overexpression dataset, normalize the end profiles to their unit vector and scale this vector with the magnitude of the overexpression profile (BRCA1 or NFKB1) being compared against. Thus, morphological endpoints point to different directions but preserve the same distance to DMSO as the compared overexpression profile. In this setting, all trajectories have a constant cosine distance to DMSO, and therefore, we report the Euclidean distance. The random baseline directions were chosen from a uniform distribution. The highest variation directions were obtained from the coefficients of a PCA model with 10 components fitted on the per-compound median profiles of the training set (after subtracting the DMSO profile).

Data availability

The data that support the findings of this study are available on **https://github.com/Bayer-Group/CPMolGAN**.

Code availability

The code used to generate results shown in this study will be available from the corresponding author upon request after the final publication of this manuscript.

Author contributions

P. A. M. Z. processed images and profiles, ran inference experiments, prepared figures and wrote the manuscript. O. M. L. trained and designed the deep learning architecture and helped writing the manuscript. T. L. helped with inference experiments and manuscript preparation. C. J. B. made image compositions and helped with manuscript and figures preparation. D. R. and J. W. provided guidance and helped with the manuscript preparation. D. A. C. conceived the study and supervised the work. O. M. L., P. A. M. Z., T. L., C. J. B., J. W., D. R. and D. A. C. read and approved the manuscript.

Conflicts of interest

During the development of this work, D. A. C., P. A. M. Z. and J. W. were employees of Bayer AG, D. R. and O. M. L. were employees of Bayer SAS, and C. B. was an employee of the Leibniz-Forschungsinstitut für Molekulare Pharmakologie.

Acknowledgements

The authors thank Joerg Tiebes for providing chemistry expert feedback and for his useful comments. We are also grateful to Arwa Al-Dilaimi and Angela Becker for supporting the project and for insightful discussions.

References

1 W. J. Scannell, A. Blanckley, H. Boldon and B. Warrington, *Nat. Rev. Drug Discovery*, 2012, **11**, 191–200.
2 F. Pammolli, L. Magazzini and M. Riccaboni, *Nat. Rev. Drug Discovery*, 2011, **10**, 428–438.
3 E. C. Butcher, *Nat. Rev. Drug Discovery*, 2005, **5**, 7.
4 W. Zheng, N. Thorne and J. C McKew, *Drug Discovery Today*, 2013, **18**, 1067–1073.
5 J. G. Moffat, F. Vincent, J. A. Lee, J. Eder and M. Prunotto, *Nat. Rev. Drug Discovery*, 2017, **16**, 531–543.
6 D. C. Swinney and J. Anthony, *Nat. Rev. Drug Discovery*, 2011, **10**, 507–519.
7 J. Eder, R. Sedrani and C. Wiesmann, *Nat. Rev. Drug Discovery*, 2014, **13**, 577–587.
8 C. Lamberth, S. Jeanmart, T. Luksch and A. Plant, *Science*, 2013, **341**, 742–746.
9 J. Lamb, E. D. Crawford, D. Peck, *et al.*, *Science*, 2006, **313**, 1929–1935.
10 A. Subramanian, R. Narayan, S. M. Corsello, *et al.*, *Cell*, 2017, **171**, 1437–1452.
11 H. Hieronymus, J. Lamb, K. N. Ross, *et al.*, *Cancer Cell*, 2006, **10**, 321–330.
12 G. Wei, D. Twomey, J. Lamb, *et al.*, *Cancer Cell*, 2006, **10**, 331–342.
13 H. De Wolf, L. Cougnaud, K. Van Hoorde, *et al.*, *Assay Drug Dev. Technol.*, 2018, **16**, 162–176.
14 A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina and A. Zhavoronkov, *Mol. Pharm.*, 2016, **13**, 2524–2530.
15 F. Iorio, T. Rittman, H. Ge, M. Menden and J. Saez-Rodriguez, *Drug Discovery Today*, 2013, **18**, 350–357.
16 M. Iwata, R. Sawada, H. Iwata, M. Kotera and Y. Yamanishi, *Sci. Rep.*, 2017, **7**, 40164.
17 S. A. Wacker, B. R. Houghtaling, O. Elemento and T. M. Kapoor, *Nat. Chem. Biol.*, 2012, **8**, 235–237.
18 M. A. Bray, S. Singh, H. Han, *et al.*, *Nat. Protoc.*, 2016, **11**, 1757–1774.
19 S. M. Gustafsdottir, V. Ljosa, K. L. Sokolnicki, *et al.*, *PLoS One*, 2013, **8**, e80999.
20 C. Scheeder, F. Heigwer and M. Boutros, *Curr. Opin. Syst. Biol.*, 2018, **10**, 43–52.
21 E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert and D. Van Valen, *Nat. Methods*, 2019, **16**, 1233–1246.
22 P. T. Jackson, Y. Wang, S. Knight, *et al.*, *16th International Conference on Machine Vision Applications (MVA) 1–4*, IEEE, 2019, DOI: **10.23919/MVA.2019.8757871**.

23 A. X. Lu, O. Z. Kraus, S. Cooper and A. M. Moses, *PLoS Comput. Biol.*, 2019, **15**, e1007348.
24 A. E. Carpenter, T. R. Jones, M. R. Lamprecht, *et al.*, *Genome Biol.*, 2006, **7**, R100.
25 J. C. Caicedo, S. Cooper, F. Heigwer, *et al.*, *Nat. Methods*, 2017, **14**, 849–863.
26 C. C. Gibso, W. Zhu, C. T. Davis, *et al.*, *Circulation*, 2015, **131**, 289–299.
27 M. H. Rohban, S. Singh, X. Wu, *et al.*, *eLife*, 2017, **6**, e24060.
28 J. Simm, G. Klambauer, A. Arany, *et al.*, *Cell Chem. Biol.*, 2018, **25**, 611–618.e3.
29 M. Hofmarcher, E. Rumetshofer, D.-A. Clevert, S. Hochreiter and G. Klambauer, *J. Chem. Inf. Model.*, 2019, **59**, 1163–1171.
30 M. J. Cox, S. Jaensch, J. Van de Waeter, *et al.*, *Sci. Rep.*, 2020, **10**, 13262.
31 J. Nyffeler, C. Willis, R. Lougee, A. Richard, K. Paul-Friedman and J. A. Harrill, *Toxicol. Appl. Pharmacol.*, 2020, **389**, 114876.
32 J. C. Caicedo, S. Singh and A. E. Carpenter, *Curr. Opin. Biotechnol.*, 2016, **39**, 134–142.
33 S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd and A. E. Carpenter, *Nat. Rev. Drug Discovery*, 2021, **20**, 145–159.
34 J. Meyers, B. Fabian and N. Brown, *Drug Discovery Today*, 2021, **26**, 2707–2715.
35 D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Mol. Syst. Des. Eng.*, 2019, **4**, 828–849.
36 M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, *ACS Cent. Sci.*, 2018, **4**, 120–131.
37 P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan and E. J. Bjerrum, *Nat. Mach. Intell.*, 2020, **2**, 254–265.
38 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, *et al.*, *ACS Cent. Sci.*, 2018, **4**, 268–276.
39 G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias and A. Aspuru-Guzik, *arXiv*, 2018, preprint, arXiv:1705.10843, DOI: **10.48550/arXiv.1705.10843**.
40 A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, *Mol. Pharm.*, 2017, **14**, 3098–3104.
41 B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, *ChemRxiv Camb. Camb. Open Engage*, 2017, preprint, DOI: **10.26434/chemrxiv.5309668.v3**.
42 N. De Cao and T. Kipf, *arXiv*, 2018, preprint, arXiv:1805.11973, DOI: **10.48550/arXiv.1805.11973**.
43 E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, *et al.*, *J. Chem. Inf. Model.*, 2018, **58**, 1194–1204.
44 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
45 W. P. Walters, *J. Med. Chem.*, 2019, **62**, 1116–1124.
46 O. Méndez-Lucio, B. Baillif, D.-A. Clevert, D. Rouquié and J. Wichard, *Nat. Commun.*, 2020, **11**, 10.
47 M. Mirza and S. Osindero, *arXiv*, 2014, preprint, arXiv1411.1784v1, DOI: **10.48550/arXiv.1411.1784**.
48 R. Winter, F. Montanari, F. Noé and D.-A. Clevert, *Chem. Sci.*, 2019, **10**, 1692–1701.
49 R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé and D.-J. Clevert, *Chem. Sci.*, 2019, **10**, 8016–8024.
50 D. P. Kingma and M. Welling, *arXiv*, 2014, preprint, arXiv:1312.6114v10, DOI: **10.48550/arXiv.1312.6114**.
51 M. Krenn, F. Häse, A. Nigam, P. Friederichand and A. Aspuru-Guzik, *33rd Conference on Neural Information Processing Systems*, NeurIPS 2019, 2019.
52 A. Gaulton, A. Hersey, M. Nowotka, *et al.*, *Nucleic Acids Res.*, 2017, **45**, D945–D954.
53 M.-A. Bray 1, S. M. Gustafsdottir, M. H. Rohban, *et al.*, *GigaScience*, 2017, **6**, 1–5.
54 K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter and G. Klambauer, *J. Chem. Inf. Model.*, 2018, **58**, 1736–1741.
55 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
56 G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nat. Chem.*, 2012, **4**, 90–98.
57 A. Thakkar, V. Chadimova, E. J. Bjerrum, O. Engkvist and J.-L. Reymond, *Chem. Sci.*, 2021, **12**, 3339–3349.
58 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 8.
59 I. Sushko, E. Salmina, V. A. Potemkin, G. Poda and I. V. Tetko, *J. Chem. Inf. Model.*, 2012, **52**, 2310–2316.
60 D. Probs and J. L. Reymond, *J. Cheminf.*, 2020, **12**, 12.
61 J. Sun, N. Jeliazkova, V. Chupakhin, *et al.*, *J. Cheminf.*, 2017, **9**, 17.
62 K. Yang, S. Goldman, W. Jin, *et al.*, *arXiv*, 2020, preprint, arXiv:200608532, DOI: **10.48550/arXiv.2006.08532**.
63 J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert and M. Rodríguez Martínez, *iScience*, 2021, **24**, 102269.
64 JUMP-Cell Painting Consortium. **https://jump-cellpainting.broadinstitute.org/**.
65 W. P. Walters and M. Murcko, *Nat. Biotechnol.*, 2020, **38**, 143–145.
66 S. Singh, X. Wu, V. Ljosa, *et al.*, *PLoS One*, 2015, **10**, e0131370.
67 K. Rajan, A. Zielesny and C. Steinbeck, *J. Cheminf.*, 2021, **13**, 34.
68 A. Bender and I. Cortes-Ciriano, *Drug Discovery Today*, 2021, **26**, 1040–1052.
69 G. A. Landrum, *RDKit-Open-source cheminformatics*, **https://www.rdkit.org**.
70 R. J. Williams and D. Zipser, *Neural Comput.*, 1989, **1**, 270–280.
71 I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville, *arXiv*, 2017, preprint, arXiv:1704.00028, DOI: **10.48550/arXiv.1704.00028**.
72 M. Abadi, A. Agarwal, P. Barham, *et al.*, *arXiv*, 2016, preprint, arXiv:160304467, DOI: **10.48550/arXiv.1603.04467**.
73 F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
74 N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, *J. Chem. Inf. Model.*, 2019, **59**, 1096–1108.
75 S. Genheden, A. Thakkar, V. Chadimová, *et al.*, *J. Cheminf.*, 2020, **12**, 70.