Cell morphology-guided de novo hit design by conditioning generative adversarial networks on phenotypic image features

Oscar Méndez-Lucio, [1,2] Paula Marin-Zapata, [3] Joerg Wichard, [4] David Rouquié, [1] Djork-Arné Clevert, [3]

[1] Bayer SAS, Bayer CropScience, 355 rue Dostoïevski, CS 90153, 06906 Sophia Antipolis, France

[2] Bloomoon, 13 Avenue Albert Einstein, 69100 Villeurbanne, France

[3] Department of Machine Learning Research, Bayer AG, 13353 Berlin, Germany

[4] Department of Genetic Toxicology, Bayer AG, 13353 Berlin, Germany

E-mail: oscar.mendezlucio.ext@bayer.com, paula.marinzapata@bayer.com, joerg.wichard@bayer.com, david.rouquie@bayer.com, djork-arne.clevert@bayer.com,

Abstract

Developing new small molecules that are bioactive is time-consuming, costly and rarely successful. As a mitigation strategy, we apply, for the first time, generative adversarial networks to de novo design of small molecules using a phenotype-based drug discovery approach. We trained our model on a set of 30,000 compounds and their respective morphological profiles extracted from high content images; no target information was used to train the model. Using this approach, we were able to automatically design agonist-like compounds of different molecular targets.

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 more into the final desired biological response, such as in a systems-based approach, rather than the current 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 and 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 final conclusion due to the long time frame of drug discovery projects[6,7].

Despite the many advantages of the systems-based approach[3], only a few experimental settings have been used for large-scale screenings. One of these is transcriptomic analysis where the change in gene expression levels caused by perturbation (either chemical or biological) is used as readout to select active compounds. This approach has been used in the Connectivity Map project[8,9] to connect disease, genes and drugs and has been successfully applied to identify new active molecules[10–12], find new uses for known drugs[13,14], and even for mode of action identification[15,16]. An alternative is to observe the effect of the perturbation at morphological level rather than at transcriptomic level. In this regard, the Cell Painting approach[17], a new technique based on High-Content Imaging (HCI), has been extremely useful. Cell Painting uses different dyes to simultaneously stain several organelles and cellular components, thus capturing information of the complete cellular state[18]. In the context of profiling, images are usually processed by computational pipelines to extract feature representations (morphological profiles)[19–23], which serve as phenotype descriptors in further tasks. This technique has been used to cluster small molecules by similar phenotypic effect[18], for drug repurposing[24], to map cellular

morphology to gene function[25], and to predict results from biochemical assays[26,27] among other applications[28]. Interestingly, most cell painting applications to small molecules have focused on clustering or classification tasks and, to our best knowledge, no use-case has combined these information-rich assays with a molecular generative model[29,30] to directly guide molecular design.

In the current work we combine cell painting morphological profiles with generative adversarial networks (GANs)[31] in order to automatically design compounds that induce a specific morphological effect. GANs are generative models that have been extensively used for molecular design, specifically for the design of new active molecules and photovoltaic materials[32,33]. In this paper we show that by conditioning a GAN using morphological profiles we can automatically design compounds with high molecular similarity to known agonists of particular targets without having target or biological activity annotations in the training set.

Results

The framework used in this study is divided in two stages. The first stage consists in training a model capable of encoding molecular structures into a continuous latent space, but also able to transform any point of the latent space back into a valid molecular structure (Figure 1a). This encoder-decoder model was designed following a similar implementation of the one proposed by Winter et al[34], which is inspired on machine translation. The second stage is a conditional generative adversarial network (GAN) capable of designing compounds that can induce specific cell morphology changes (Figure 1b). More specifically, the GAN is composed of three neural networks, namely generator, discriminator and conditional. The task of the generator is to suggest points of the latent space that correspond to molecules able to induce a desired morphological profile. On the other hand, the discriminator evaluates if the generated point of the latent space indeed corresponds to a proper molecule whereas the conditional network calculates the probability of this molecule to induce the desired morphological profile. The GAN is trained in an adversarial

manner meaning that at every training step the generator comes with better suggestions to fulfil the criteria of the discriminator, but at the same time the evaluation by the discriminator becomes stricter each time. This conditional GAN was trained with 126,779 morphological profiles induced by 30,616 compounds in the U2OS cell line reported by Bray *et al*.[35] (see Methods for more details).

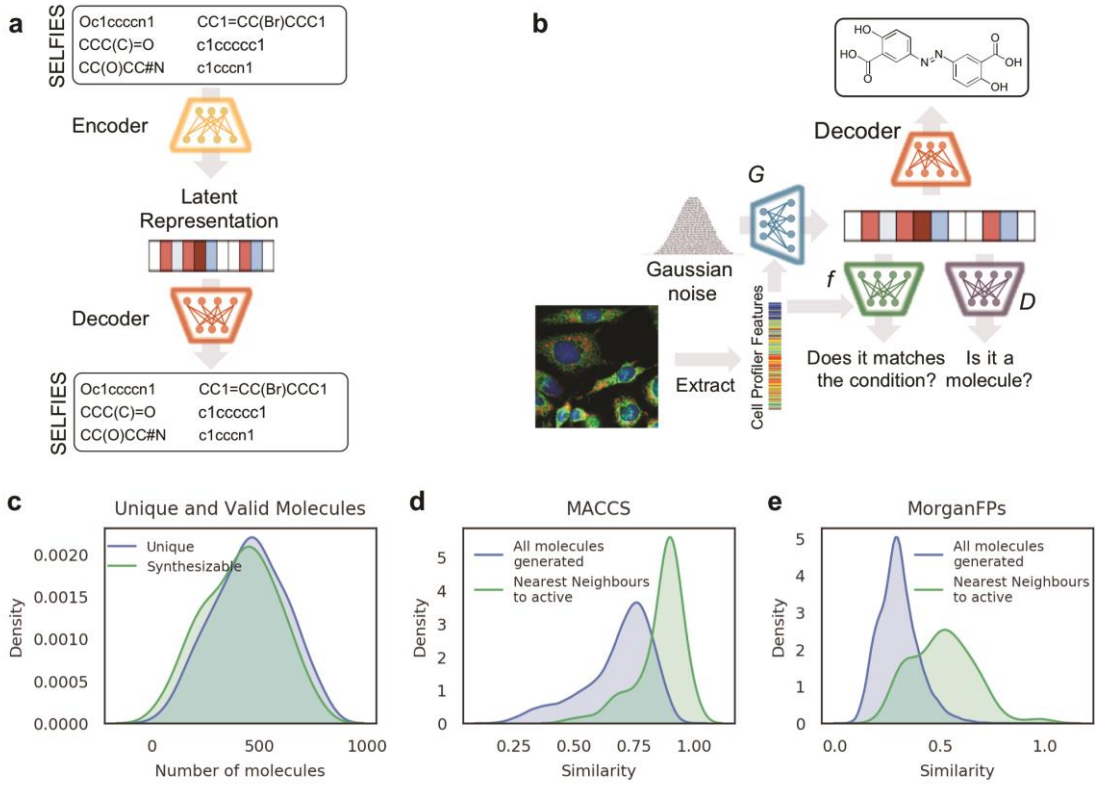

Figure 1. Graphical representation of the models and distribution of similarity. Molecules were encoded using a model that transforms SELFIES[36] of a molecule into a latent representation that can be later decoded back into SELFIES (a). The generative adversarial network in (b) has a generator (G in blue) that takes the desired morphological profile together with a vector of random noise to produce a molecular representation that can be decoded into SELFIES using the decoder (in red). The discriminator (D in purple) calculates the probability of the molecular representation to be a real molecule and the conditional network (f in green) calculates the probability of the molecular representation to match the morphological profile. Distributions shown in (c) represent the number of unique (blue) and synthesizable (green) molecules generated across the 114 morphological features from overexpressed genes.

Distributions of similarity measurements between all generated molecules and their most similar known agonist were calculated using MACCS (d) and Morgan fingerprints (e).

We used this framework to generate compounds that mimic the morphological changes observed when overexpressing a specific target gene in the U2OS cell line using cDNA. We focused on 12 genes that where individually overexpressed by Rohban *et al*.[25] in U2OS using different cDNA open reading frames (ORFs), namely *BRCA1*, *CREBBP*, *HIF1A*, *HSPA5*, *JUN*, *NFKB1, NFKBIA*, *PDPK1*, *PRKAA1, STAT1*, *STAT3*, and *TP53*. Considering all ORFs per gene, we computed a total of 114 morphological profiles and generated 1,000 molecules for each of them. On average, each morphological profile produced 46.7% of valid molecules from which most of them (45.6%) correspond to unique compounds and 41.4% presented a synthetic accessibility score[37] < 4.5, meaning they have high chances to be synthesized (Figure 1c). To test the ability of our model to generate phenotype-tailored molecules, we evaluated the similarity between the generated compounds and known agonists of the 12 selected targets. Agonists were retrieved from ExCAPE database[38] making sure to exclude all of them that were present in the training set. Overall, generated molecules share similar chemical groups with the known agonist compounds (median MACCS[39] similarity = 0.72, IQR = 0.18) but also share a good proportion of molecular substructures (median Morgan Fingerprints similarity = 0.30, IQR = 0.12) as shown in Figures 1d and 1e. Figure 2 shows examples of the generated molecules presenting the highest similarity to a known active molecule for each of the overexpressed targets.

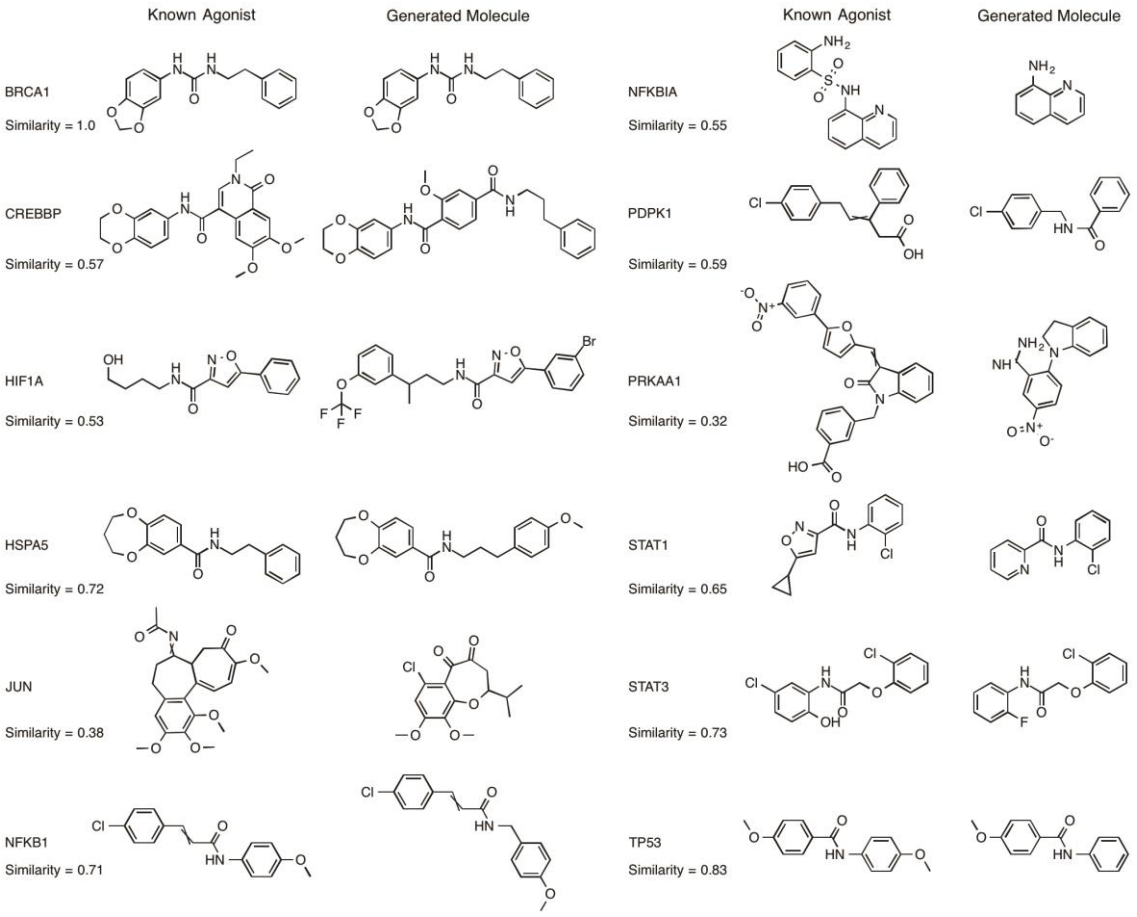

Figure 2. Chemical structures of molecules generated by the model. These molecules were obtained by conditioning the model using the microscopy images of 12 different overexpressed genes and are depicted next to their closest active nearest neighbour.

A scaffold analysis of the generated molecules revealed that the model can generate molecules containing the same scaffolds as the known active molecules. Tables 1 and 2 show the number of shared scaffolds per target. Interestingly, more than 20% of molecules generated for *BRCA1* and *NFKB1* contain a Murcko scaffold that is known to be active. This rate is considerably increased when using Murcko generic scaffolds, which is an abstraction form of scaffold where all atoms are converted into carbon atoms and all bonds transform into single bonds. In this case, more than 50% of molecules generated for *BRCA1* and *NFKB1* and more than 25% for *TP53* and *HSPA5* contain a known active generic Murcko scaffold. Interestingly, none of the molecules generated for *JUN* and *PRKAA1* contain a known active Murcko scaffold or generic scaffold. Nonetheless, it is worth noting that only less than

five Murcko scaffolds have been reported in the ExCAPE database to have an agonist effect on these targets.

Table 1. Comparison of Murcko scaffolds in generated molecules and known agonists from ExCAPE database.

	Valid and synthetic accessible molecules	Unique scaffolds in known agonists	Generated molecules containing an active scaffold	Unique active scaffolds present in generated molecules
PDPK1	4624	1	0	0
BRCA1	3444	4632	861	97
NFKB1	3343	256	703	39
STAT1	4304	78	61	10
NFKBIA	2435	4	93	1
STAT3	6824	56	359	9
PRKAA1	2410	3	0	0
CREBBP	2591	80	0	0
HIF1A	4665	38	1	1
JUN	5787	4	0	0
TP53	4250	7522	718	126
HSPA5	2592	551	133	24

Table 2. Comparison of Murcko generic scaffolds in generated molecules and known agonists from ExCAPE database.

	Valid and synthetic accessible molecules	Unique generic scaffolds in known agonists	Generated molecules containing an active generic scaffold	Unique active generic scaffolds present in generated molecules
PDPK1	4624	1	19	1
BRCA1	3444	2146	2248	237
NFKB1	3343	201	1783	66
STAT1	4304	62	562	36
NFKBIA	2435	4	109	1
STAT3	6824	51	835	21
PRKAA1	2410	2	0	0
CREBBP	2591	76	100	14
HIF1A	4665	32	141	7
JUN	5787	4	0	0
TP53	4250	4627	1851	374
HSPA5	2592	406	677	90

In order to choose the most promising compounds for experimental evaluation we filtered all generated compounds based on physicochemical properties, namely LogP between 0 and 7, molecular weight less than 750 daltons, number of hydrogen bond donors and acceptors below 10 and less than 10 rotatable bonds. This filtering step reduced the total number of molecules to 39,801. Additionally, we removed those compounds that contained non medicinal chemistry friendly SMART and toxalert[40,41] leaving a total number of 33,341 compounds. Since we are interested in novel molecules, we filtered out those compounds with Morgan Fingerprints similarity higher than 0.8 compared to compounds in the training set, which reduced the number of compounds to a total of 33,309. In a further step, we used in-house models to rank generated molecules based on their probability to cross the cellular membrane in a Caco-2 assay. These helped us to choose at least 100 molecules for each target that were examined by an experienced chemist. Some of these molecules are currently being synthesized and tested on a cell painting assay.

Table 3. Number of generated molecules per target after each of the filtering steps.

	Valid and synthetic accessible molecules	After physicochemical filters	After removing compounds with ToxAlerts	Compounds with low similarity to the training set*
BRCA1	3444	2843	2347	2342
CREBBP	2591	2185	1905	1902
HIF1A	4665	4107	3564	3562
HSPA5	2592	2373	2104	2102
JUN	5787	4071	3405	3403
NFKB1	3343	3184	2634	2626
NFKBIA	2435	2136	1699	1699
PDPK1	4624	4059	3447	3443
PRKAA1	2410	1883	1489	1486
STAT1	4304	3849	3337	3335
STAT3	6824	5589	4664	4663
TP53	4250	3522	2746	2746

*Similarity threshold set to 0.8 using Morgan Fingerprints and Dice distance

Conclusions

In this work, we showed that it is possible to combine artificial intelligence with morphological cellular images in order to design new compounds. In this specific proof of concept, the model was able to use the information contained in a genetically induced phenotypic profile to produce molecules that present high structural similarity to known agonists of a particular target. We expect that this work will help to attract more attention to systems-based approach, particularly to high throughput imaging screening. In addition, we expect that more computational tools, like the one presented in this paper, will help increase the efficiency of the systems-based approaches in order to accelerate drug discovery R&D.

Methods

Data set: For this study we used a data set coming from a high-content screening of 30,616 compounds that followed the cell painting protocol.[35] More specifically, Human Bone Osteosarcoma Epithelial cells (U2OS cells) were plated and treated with each of these compounds (details can be found elsewhere[17]) in quadruplicate. Each well was imaged using 5 fluorescent channels and 6 fields of view. The different dyes used in cell painting were chosen to stain different organelles or cellular components including: nucleus, endoplasmic reticulum, nucleoli, cytoplasmic RNA, cytoskeleton, Golgi, plasma membrane and mitochondria.[17,18] Raw images from different fields and channels were processed using CellProfiler[23] to extract morphological profiles composed by 1,783 quantitative features (e.g. cellular shape, intensity, texture, etc.) able to capture slight morphological changes induced by the treatment. More details on the experimental and processing protocols can be found in the original publication.[35] In the end, this data set is composed by 126,779 morphological profiles corresponding to 30,616 compounds. Notice that each compound is associated with at least 4 population-averaged morphological profiles coming from replicates, which were treaded independently.

As a test set we used morphological profiles calculated from images taken during a high-content profiling campaign of gene gain-of-function recently published.[25] In that work, Rohban *et al.* profiled the overexpression of 323 open reading frames (ORF) constructs also following the cell painting protocol. Illumination corrected images for the test set were obtained from The Image Data Resource (IDR) web API. To assure comparability of profiles between test and training sets, we closely followed the feature extraction protocol from Bray et al.[35]. Single-cell profiles were calculated with CellProfiler 2.2.0 using the analysis.cppipe pipeline, which was slightly modified to take input .csv files. Population-averaged profiles were computed as the median among all cells from all fields of view per well. As 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.[35] for a randomly chosen plate (plate number 25690).

All molecular structures were pre-processed using the MolStandardize module from RDKIT[42]. First, molecules as SMILES codes were standardized by removing hydrogens, sanitizing the molecules, disconnecting metals and normalizing the structure. Then, only the largest fragment of each molecule was kept, and their charges were neutralized. Finally, the uncharged molecules were transformed back into a non-isomeric form of canonical SMILES. Despite SMILES can capture information regarding the molecular structure, they do not capture the molecular graph connectivity. In order to overcome this issue, we further transform SMILES into self-referencing embedded strings (SELFIES)[36] which are able to encode molecular graphs in a simple and deterministic way. In order to use SELFIES as input for our model, they were transformed into a one-hot encoding format just keeping those tokens that appear in at least 100 molecules from the 1.5 million contained in ChEMBL22[43] (Figure 3). Molecules containing different SELFIES tokens to the ones selected previously or SMILES strings larger than 120 character were removed from the dataset.

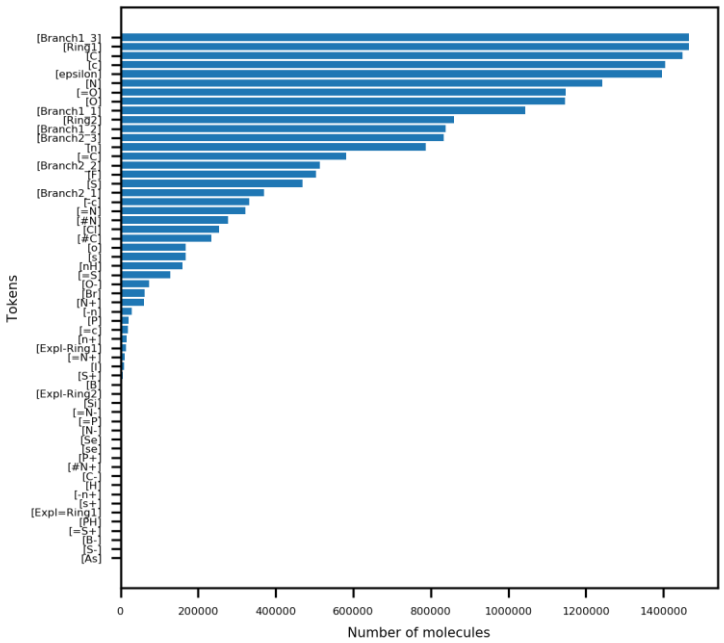

Figure 3. Tokens that appear in a least 100 molecule from the 1.5 million contained in ChEMBL22[43]

Molecular Embedding: Recent studies have shown that is possible to encode molecules (as SMILES[34,44,45], graphs or SELFIES[36]) into a continuous representations. As before[34,45,46], we used a molecular autoencoder but in this

case it learns to encode SELFIES of a molecule (as one-hot encoding) into a continuous latent vector, which then can be decoded back into SELFIES. The architecture of this model is based on the previous work of Winter et al.[34,45], where the encoder is comprised by three stacked Gate Recurrent Unit (GRU) cells where their resulting cell states are concatenated and fed into two independent fully connected layers which predict the mean and standard deviation of a distribution from which the latent vector will be sampled for the variational approach. After sampling, the latent vectors are compressed by a tanh activation function to produce the final latent vectors 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 two expand them from 256 to 768 dimensions. This vector is split into three vectors of 256 dimensions and used as initial states of another three stacked GRU cells. The output of the stacked GRU 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[47] scheme during 10 epochs on 1.25 million molecules extracted from ChEMBL 22[43] complemented by the molecules in the training set.

Generator architecture (G(z,c)): The generator receives as input the condition, in this case a morphological profile (of 1,449 features), and a 1,000-dimmensional 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 activation function. At the same time, 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 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 architecture: 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. In order to reduce overfitting, dropout with 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 architecture: This network estimates the probability of a molecular representation to match a morphological profile. Here 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 rate of 0.4. In a similar way, 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.

Training: The conditional generative adversarial network was trained during 500 epochs using a batch size of 256. Here, an epoch was composed by 496 steps where the weights of the discriminator were updated after each step, whereas those of the generator every 10 steps. The network was trained using the RMSprop optimizer with learning rate of 1×10^{-4} for both the generator and discriminator. The weights of the conditional network were pretrained and frozen during the GAN training process. All neural networks were built and trained using Tensorflow 1.14[48].

Model evaluation: During training, we generated a molecular representation for each morphological profile in the training set at the end of each epoch. These were used to evaluate the similarity between the generated and real molecular representations using Fréchet distance. The Fréchet distance measures the similarity between two distributions (in this case the real and generated) and was recently proposed as an efficient way to evaluate the efficacy of generative models.[49] This metric takes the mean and covariance of the real distribution (μ_r and C_r, respectively), together with the mean and covariance of the generated distribution (μ_g and C_g) in the following formulation:

$$d^2\left((\mu_r, C_r), (\mu_g, C_g)\right) = \|\mu_g - \mu_r\|_2^2 + Tr(C_g + C_r - 2(C_g C_r)^{1/2})$$

Generating molecules from features of morphological images: We used 114 morphological profiles from the ones reported by Rohban et al. after overexpressing open reading frames (ORFs) corresponding to 12 different genes: *BRCA1*, *CREBBP*, *HIF1A*, *HSPA5*, *JUN*, *NFKB1, NFKBIA*, *PDPK1*, *PRKAA1, STAT1*, *STAT3*, and *TP53*. Known agonists of the corresponding protein expressed by these genes were extracted from the ExCAPE database[38]. For these 12 targets we generated 1000 molecular representation for each morphological profile (114 in total) and evaluated the model by comparing the generated molecules to the known agonists of these targets. Dice similarity between the generated compounds and the molecules with known agonistic effect was computed using MACCS keys and Morgan Fingerprints (radius = 3, 1024 bits) with RDKIT.[42] Despite that the similarity between each generated molecule and each known agonist was calculated, only the one corresponding to the closest known agonist was recorded for each generated molecule and used for further analysis. It is worth mentioning that during these validation tasks both the number of valid molecules (validity measure) and the number of unique molecules (uniqueness measure) were recorded as sanity check for the generative model.

Author contributions

O.M.L. performed the computational experiments and designed the deep learning model. P.M.Z pre-processed all images in the data set, calculated features from morphological images and helped with the manuscript 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.M.Z, J.W., D.R. and D.A.C. read and approved the manuscript.

Acknowledgments

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.

Competing interests

D.A.C., P.M.Z and J.W. are employees of Bayer AG. O.M.L and D.R. work

directly or indirectly for Bayer SAS**.**

Data availability

The data that support the findings of this study will be available on request from the corresponding author after the final publication of this manuscript.

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.

References

1. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* vol. 11 191–200 (2012).

2. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **10**, 428–438 (2011).

3. Butcher, E. C. *Can cell systems biology rescue drug discovery? Nature Reviews Drug Discovery* vol. 4 www.nature.com/reviews/drugdisc (2005).

4. Zheng, W., Thorne, N. & McKew, J. C. Phenotypic screens as a renewed approach for drug discovery. *Drug Discov. Today* **18**, 1067–1073 (2013).

5. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: An industry perspective. *Nat. Rev. Drug Discov.* **16**, 531–543 (2017).

6. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **10**, 507–519 (2011).

7. Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: Origins and evolution. *Nat. Rev. Drug Discov.* **13**, 577–587 (2014).

8. Lamb, J. *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science (80-.).* **313**, 1929–1935 (2006).

9. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

10. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).

11. Wei, G. *et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer*

376 *Cell* **10**, 331–342 (2006).

377 12. De Wolf, H. *et al.* High-Throughput Gene Expression Profiles to Define
378 Drug Similarity and Predict Compound Activity. *Assay Drug Dev. Technol.*
379 **16**, 162–176 (2018).

380 13. Aliper, A. *et al.* Deep learning applications for predicting pharmacological
381 properties of drugs and drug repurposing using transcriptomic data. *Mol.*
382 *Pharm.* **13**, 2524–2530 (2016).

383 14. Iorio, F., Rittman, T., Ge, H., Menden, M. & Saez-Rodriguez, J.
384 Transcriptional data: A new gateway to drug repositioning? *Drug Discov.*
385 *Today* **18**, 350–357 (2013).

386 15. Iwata, M., Sawada, R., Iwata, H., Kotera, M. & Yamanishi, Y. Elucidating
387 the modes of action for bioactive compounds in a cell-specific manner by
388 large-scale chemically-induced transcriptomics. *Sci. Rep.* **7**, 40164
389 (2017).

390 16. Wacker, S. A., Houghtaling, B. R., Elemento, O. & Kapoor, T. M. Using
391 transcriptome sequencing to identify mechanisms of drug action and
392 resistance. *Nat. Chem. Biol.* **8**, 235–237 (2012).

393 17. Bray, M. A. *et al.* Cell Painting, a high-content image-based assay for
394 morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.*
395 **11**, 1757–1774 (2016).

396 18. Gustafsdottir, S. M. *et al.* Multiplex cytological profiling assay to measure
397 diverse cellular states. *PLoS One* **8**, 1–7 (2013).

398 19. Scheeder, C., Heigwer, F. & Boutros, M. Machine learning and image-
399 based profiling in drug discovery. *Current Opinion in Systems Biology* vol.
400 10 43–52 (2018).

401 20. Moen, E. *et al.* Deep learning for cellular image analysis. *Nat. Methods*
402 doi:10.1038/s41592-019-0403-1 (2019) doi:10.1038/s41592-019-0403-1.

403 21. Jackson, P. T. *et al.* Phenotypic profiling of high throughput imaging
404 screens with generic deep convolutional features. *Proc. 16th Int. Conf.*

405 *Mach. Vis. Appl. MVA 2019* 1–4 (2019)
406 doi:10.23919/MVA.2019.8757871.

407 22. Lu, A. X., Kraus, O. Z., Cooper, S. & Moses, A. M. Learning unsupervised
408 feature representations for single cell microscopy images with paired cell
409 inpainting. *PLoS Comput. Biol.* **15**, 1–27 (2019).

410 23. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying
411 and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).

412 24. Gibson, C. C. *et al.* Strategy for identifying repurposed drugs for the
413 treatment of cerebral cavernous malformation. *Circulation* **131**, 289–299
414 (2015).

415 25. Rohban, M. H. *et al.* Systematic morphological profiling of human gene
416 and allele function via Cell Painting. *Elife* **6**, 1–23 (2017).

417 26. Simm, J. *et al.* Repurposing High-Throughput Image Assays Enables
418 Biological Activity Prediction for Drug Discovery. *Cell Chem. Biol.* **25**, 611-
419 618.e3 (2018).

420 27. Hofmarcher, M., Rumetshofer, E., Clevert, D. A., Hochreiter, S. &
421 Klambauer, G. Accurate Prediction of Biological Assays with High-
422 Throughput Microscopy Images and Convolutional Networks. *J. Chem.*
423 *Inf. Model.* **59**, 1163–1171 (2019).

424 28. Caicedo, J. C., Singh, S. & Carpenter, A. E. Applications in image-based
425 profiling of perturbations. *Current Opinion in Biotechnology* vol. 39 134–
426 142 (2016).

427 29. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A.
428 Machine learning for molecular and materials science. *Nature* vol. 559
429 547–555 (2018).

430 30. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design
431 using machine learning:Generative models for matter engineering.
432 *Science (80-.).* **361**, 360–365 (2018).

433 31. Goodfellow, I. J. *et al.* Generative adversarial nets. in *Advances in Neural*

Information Processing Systems vol. 3 2672–2680 (Curran Associates, Inc., 2014).

32. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. Preprint at http://arxiv.org/abs/1705.10843 (2017).

33. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. DruGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* **14**, 3098–3104 (2017).

34. Winter, R., Montanari, F., Noé, F. & Clevert, D. A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).

35. Bray, M. A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).

36. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. (2019).

37. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 1–11 (2009).

38. Sun, J. *et al.* ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminform.* **9**, 1–9 (2017).

39. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).

40. Non MedChem-Friendly SMARTS. https://www.surechembl.org/knowledgebase/169485-non-medchem-

friendly-smarts.

41. Sushko, I., Salmina, E., Potemkin, V. A., Poda, G. & Tetko, I. V. ToxAlerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* **52**, 2310–2316 (2012).

42. Landrum, G. A. RDKit: Open-source cheminformatics. http://www.rdkit.org.

43. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).

44. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).

45. Winter, R. *et al.* Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **10**, 8016–8024 (2019).

46. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De Novo Generation of Hit-like Molecules from Gene Expression Signatures Using Artificial Intelligence. doi:10.26434/chemrxiv.7294388.v1.

47. Williams, R. J. & Zipser, D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Comput.* **1**, 270–280 (1989).

48. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Preprint at http://arxiv.org/abs/1603.04467 (2016).

49. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **58**, 1736–1741 (2018).