



# Image-based profiling for drug discovery: due for a machine-learning upgrade?

Srinivas Niranj Chandrasekaran<sup>1</sup>, Hugo Ceulemans<sup>2</sup>, Justin D. Boyd<sup>3</sup> and Anne E. Carpenter<sup>1</sup>✉

**Abstract** | Image-based profiling is a maturing strategy by which the rich information present in biological images is reduced to a multidimensional profile, a collection of extracted image-based features. These profiles can be mined for relevant patterns, revealing unexpected biological activity that is useful for many steps in the drug discovery process. Such applications include identifying disease-associated screenable phenotypes, understanding disease mechanisms and predicting a drug's activity, toxicity or mechanism of action. Several of these applications have been recently validated and have moved into production mode within academia and the pharmaceutical industry. Some of these have yielded disappointing results in practice but are now of renewed interest due to improved machine-learning strategies that better leverage image-based information. Although challenges remain, novel computational technologies such as deep learning and single-cell methods that better capture the biological information in images hold promise for accelerating drug discovery.

## Proteomic profiling

Measuring the levels of a large number of proteins in a sample, sometimes including their post-translationally modified forms.

## Metabolomic profiling

Measuring the levels of a large number of metabolites in a sample.

<sup>1</sup>Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>2</sup>Discovery Data Sciences, Janssen Pharmaceutica NV, Beerse, Belgium.

<sup>3</sup>High Content Imaging Technology Center, Internal Medicine Research Unit, Pfizer Inc., Cambridge, MA, USA.

✉e-mail: [anne@broadinstitute.org](mailto:anne@broadinstitute.org)

<https://doi.org/10.1038/s41573-020-00117-w>

In a drug development programme, evaluation of the efficacy and safety of all candidate compounds in humans, or even rodents, is ethically and practically unfeasible. Therefore, simpler model systems (cells, tissues and small model organisms) are used to map clinical efficacy and safety to a screening-amenable molecular target, pathway or phenotype in the process of screening<sup>1</sup> (BOX 1). The design of a screening assay balances on the one hand practicality and affordability, allowing broad exploration of chemical space, and on the other hand biological relevance to the disorder being studied or to possible safety concerns. Thus, the output, or readout, of a screening assay is typically chosen to be one or a few readily interpretable features that reflect biology already understood to be relevant for efficacy or safety<sup>2</sup>. Multiple such assays are used to test thousands to millions of small molecules to identify and triage hits (that is, attractive molecular starting points). The assays are also then used to advance hits to more drug-like leads and finally to optimize leads before preclinical development.

Profiling is an alternative strategy to screening. The word 'profiling' has two meanings: representing a sample by a profile (that is, a collection of features), and making predictions about a sample based on such a representation. Profiling aims to capture a wide variety of features, few or none of which may have previously

validated relevance to a disease or potential treatment. It may, therefore, reveal unanticipated biology at play. Profiling often relies on the same or similar model systems as screening assays (for example, fluorescently stained cells), but profiling represents these model systems with a more comprehensive set of features<sup>3–5</sup>. More features can be powerful — for example, to uncover a previously unexpected mechanism of the disease or to provide sensitive quality control for the stability of an assay system — but they can also add unhelpful noise or be more difficult to interpret than a carefully selected screening feature.

Feature profiles can be constructed in many ways using many assay types<sup>6</sup>. Readouts can be generated from panels of separate assays; examples widely used in industry<sup>7</sup> include cell viability across cell line panels<sup>8</sup>, enzymatic activity across a kinase panel<sup>9</sup> and binding to a panel of safety-relevant targets<sup>10</sup>. More cost-efficient profiles can be acquired in a single multiplexed assay<sup>11</sup> by combining one of several kinds of high-dimensional readout technologies with cell-based model systems. Proteomic profiling and metabolomic profiling would be very powerful but are too low throughput due to their cost<sup>12–15</sup>. Increasingly affordable sequencing approaches enable transcriptional profiling<sup>16</sup> to rival bead-based alternatives<sup>17,18</sup>; they also underlie highly multiplexed cell viability assays<sup>19</sup>.

## Box 1 | Approaches to screening

Screening is the workhorse of modern drug discovery; 'screening' describes testing many potential drugs in an assay that in some way detects an impact on a disease. A screening assay can detect engagement with a pre-identified disease-related protein, known as a target (in target-based assays), or a predefined molecular output event such as phosphorylation, translocation or gene activation for a disease-related pathway (in pathway-based assays), or a change in a molecularly agnostic disease-related phenotype (in phenotypic assays). Pathway-based assays, while target agnostic, are as reductive as target-based assays; they tend to be relatively simple, relying on a cultured cell line and one or a few prespecified molecular readouts. Molecularly agnostic phenotypic screens use a broader definition of phenotype; they aim to replicate the human disease state as closely as possible in an assay format that is nevertheless efficiently screenable. Some screens can be exceptionally complex, such as testing each compound's ability to alter the function of 3D organoids derived from patient cells. Interest in phenotypic screens has fluctuated, driven initially by the expectation of improved validation in the clinic, but dampened by disappointment with the lack of evidence thereof<sup>1,23,24,164–167</sup>.

Designing a screenable assay involves selecting a combination of a model system, stimulus and readouts that aim to maximize clinical trial success. A popular guiding strategy in the industry called the 'rule of three', for example, specifies (1) designing the model system to best mimic the disease condition, (2) selecting a stimulus to produce a disease-associated response and (3) choosing readouts as the most proximal quantifiable features that reflect the functional consequence of disease<sup>168</sup>.

To date, image-based profiling using automated microscopy is the least expensive of these high-dimensional profiling techniques, and it inherently offers single-cell resolution, which can capture important heterogeneous cell behaviours. Computer vision techniques have advanced dramatically in the past few years, enabling the extraction of a huge quantity of unbiased morphological information from images. Cell-based microscopy assays have also advanced (BOX 2), with assays such as Cell Painting inexpensively combining multiple stains in a robust assay yielding single-cell profiles composed of thousands of features<sup>20</sup>. The few reported comparative analyses indicate that image-based profiling may capture more biological information than high-throughput transcriptional profiling<sup>21,22</sup>.

Profiling is not a widely deployed strategy at the primary screening phase in the pharmaceutical industry, where well-validated, bespoke screening assays remain the preferred approach. However, profiling does play a role at other stages of the drug discovery process. In the discovery stages before screening — including target identification, target validation, phenotype discovery and assay development — up to thousands of compounds are profiled in exploratory assays that combine imaging, primary or induced pluripotent stem cell-derived cells and/or genetic editing<sup>1,23,24</sup>. As a consequence, profiling in the early stages of drug discovery can reveal key biological readouts that can be used for the subsequent phase of screening at high throughput.

Image-based profiling is also regularly applied downstream of screening to the hits that emerge from target-based, pathway-based or phenotypic screening and validation. It then functions as an unbiased secondary assay, applicable to hits from any screen and in any disorder. At minimum, profiling can organize hits into groups with biologically similar effects. At best, it can hint at a compound's mechanism of action (MOA) and previously unsuspected off-target activities. However,

it requires a considerable interpretation effort to derive these actionable insights from phenotypic features. With conventional analytical approaches, this interpretation step has been too onerous to scale up to screening full libraries.

Recently, machine-learning models were trained to predict the outcomes of hundreds of assays from a set of existing high-content images that were originally collected at high throughput for an image-based screen with a focused readout<sup>25</sup>. The study illustrated how machine learning can leverage side information — in this case, a large volume of assay activity labels. In another study, machine learning showed potential for image-based lead optimization<sup>26</sup>. Both of these studies were led by researchers at large pharmaceutical companies and reveal a renewed interest in profiling-inspired drug discovery. Not coincidentally, both studies were enabled by machine learning.

A new wave of biotechnology companies, including insitro and Recursion, are developing image-based profiling strategies with the explicit intent to feed a machine-learning effort<sup>27,28</sup>. They rely on moderately sized libraries and unbiased image-based profiling for phenotype discovery and primary screening, rather than on extensive decks of compounds in more time-consuming, customized assays. This enables more rapid exploration of hundreds of model systems that emulate disease states with genetic perturbations, even though some disease states may be less well represented in a generic cell system without customized assay readouts (BOX 2). Machine learning is deployed to learn to map chemical, genetic or pathological perturbations to their *in vitro*, *in vivo* or clinical effects and to transform raw profiles into a screening assay, reducing the need for expert-crafted screening assays.

In this Review, we focus on applications of image-based profiling that are immediately applicable to the drug discovery process. We begin with a brief introduction to image-based profiling assays and analysis approaches and then discuss the status, successes and limitations of various pharmaceutical applications of this technology.

### Image-based profiling in a nutshell

Image-based profiling does not require specialized equipment or reagents. All that is needed are images of biological samples that represent different cases (for example, categories of human patients) or treatment conditions (for example, chemical, genetic, time-point or other perturbations of the biological system) (FIG. 1).

First, the biological samples are prepared. This is typically done in arrayed multiwell plate format, although living cell microarrays and pooled imaging strategies are higher-throughput options<sup>29</sup>. Next, the samples are subjected to treatment conditions of interest and incubated. The samples are imaged, typically after fixation and staining, although one can instead conduct live, time-lapse imaging<sup>30,31</sup> and/or use label-free techniques such as those that predict a staining pattern from brightfield images using machine learning<sup>32,33</sup>.

The images are processed to extract features, which are aggregated into profiles. Methods for this step

#### Mechanism of action

(MOA). The description of how a compound interacts with a target and affects a biological system.

#### Side information

Further available measurements or metadata about samples that indirectly improve predictive performance.

#### Labels

Values for particular parameters in a given set of samples. For example, each compound in a dataset might have a mechanism of action label or a toxicity label.

#### Lead optimization

The process of narrowing down compounds after hit expansion to those with desired activity.

#### Brightfield images

Images captured from a sample without using any fluorescent illumination light.

## Supervision

In machine learning, supervised learning aims for the system to predict the correct answers for each input, on the basis of examples. By contrast, in unsupervised learning the goal is to learn useful representations of each sample such that the similarities and differences among them can be observed.

are developing rapidly, in particular moving from expert-defined feature extraction to data-driven deep learning<sup>34,35</sup>. Finally, the extracted profiles are analysed for biologically meaningful similarities and differences in a portion of the computational workflow that differs depending on which application has been chosen.

In essence, any set of images can be used for image-based profiling (BOX 2). The workflow described above requires adaptation depending on the samples, perturbations, stains and imaging modality used, but the overall strategy is the same. That said, profiling will be most powerful if each image contains a large number of instances (for example, cells or organisms) and a large amount of visible information about each instance (for example, high-resolution images of multiple corresponding stains).

## Analysis techniques evolve

**Challenges in profile analysis.** Thousands of features can be extracted from images, but not all are equally useful for a given task. Different disease states or different compound mechanisms of action (MOAs) are best captured by reading out changes in distinct combinations of features, such as the shape and size of cell structures or the intensity and texture of various stains. An extensive image-based profile encompasses many feature combinations, overlapping or not, that together contain the

information to document a broad spectrum of diseases and MOAs at once.

Notably, for a given goal, the presence of the relevant features is often not enough; we also need differential weighting: the amplification of features that are relevant to the goal and the suppression of features that are not. Importantly, the more extensive the profile, the more biological information it can potentially encode, but also the harder it is to extract the information for any given individual task from among irrelevant information. This phenomenon is referred to as the curse of dimensionality: the masking of the task-specific signal derived from comparatively few relevant features by the cumulative noise (related to technical variation) and confounding signal (related to unrelated biology) across the many other features. The problem can be addressed by combining features by weighted aggregation and/or more powerful representations produced by machine learning, which can have various levels of supervision (BOX 3). In a way, an assay developer applies feature weighting when exploring and optimizing biological models, reagents and readouts for a screening assay. Computationally extracting relevant information from feature weighting of a profile can be thought of as the virtual development of a profile-based screening assay.

**Early analysis techniques.** Since the first proposals to leverage image-based profiling for drug discovery<sup>36–38</sup>, the analysis of image-based profiles has largely relied on unsophisticated unsupervised clustering of the profiles, including some profiles for well-known controls. Profiles are used to place each sample into a single, shared high-dimensional representation space (panels a–c of the figure in BOX 3) — also referred to as a (weighted) feature or phenotypic space — where the closeness between pairs of compounds corresponds to their profile similarity. ‘Shared’ here refers to adjusting all features in the same way, regardless of which samples or classes are being studied.

The nearest-neighbour strategy is an approach to add an element of supervision. It predicts the biological activity or MOA of a query compound on the basis of that of a reference compound with a suitably similar profile. The biggest advantage of the nearest-neighbour approach is its information efficiency: in contrast to more advanced methods, it can be applied even if only a handful of reference compounds with existing annotation are available. Nevertheless, like those advanced methods, the approach improves with more annotated references. Importantly, reference compounds must be sufficiently active, fairly selective and, of course, accurately annotated — which is not assured even for the current marketed pharmacopoeia. As a result, the nearest-neighbour strategy has limited power.

Feature adjustment, transformation and normalization techniques are recommended to increase signal-to-noise ratios and thereby mitigate the curse of dimensionality<sup>39</sup> (panel b of the figure in BOX 3). These techniques, however, affect only the shared representation space in which samples are placed. Hence, they do not remedy the vulnerability of unsupervised approaches to confounding signals from sample biases

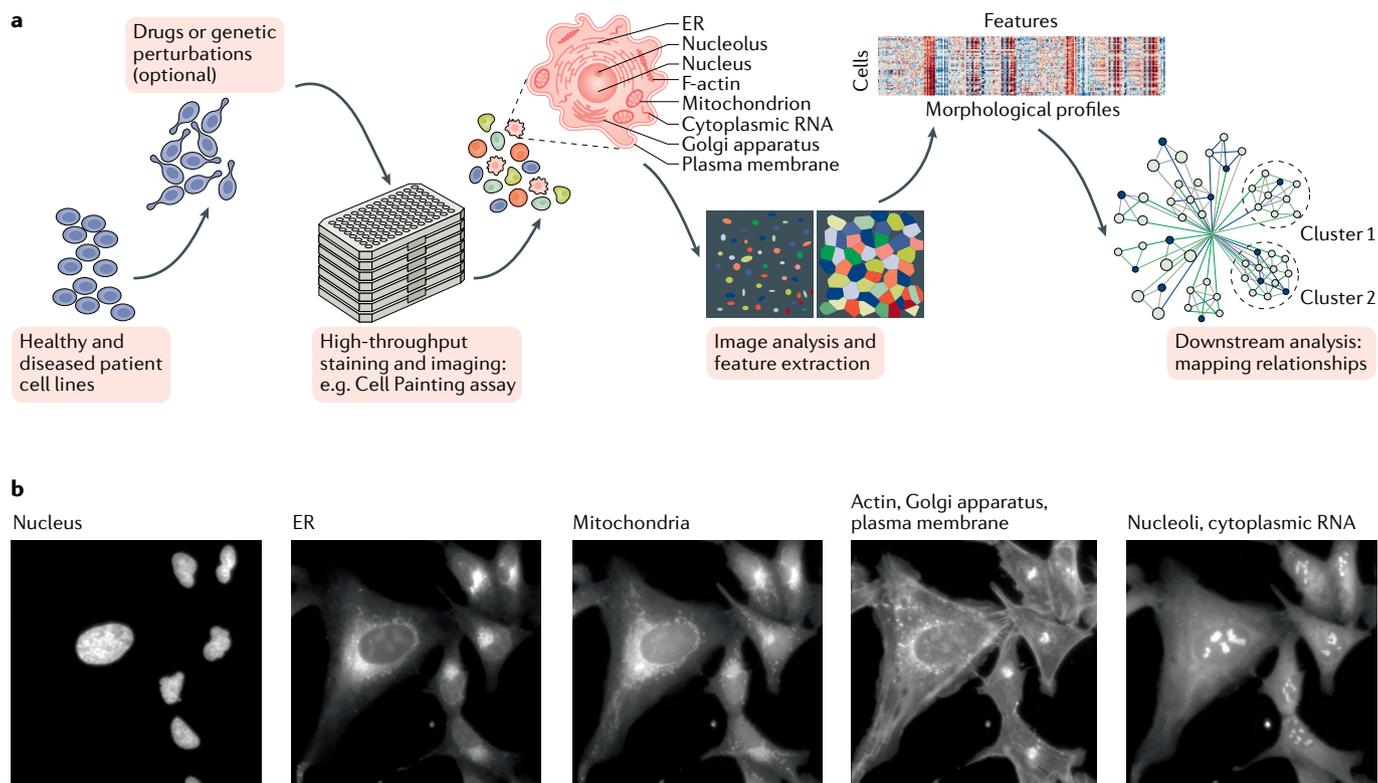
## Box 2 | Assays for image-based profiling

There are two approaches to choosing the staining conditions and biological model system for image-based profiling: customized versus unbiased. In the customized approach, one chooses a model system and fluorescent markers that are thought to be associated with specific disease properties. The unbiased approach uses a more generic model system (for example, a particular cultured cell line) and a more general set of stains, regardless of the disease under study. Although customization tends to more reliably provide information relevant to a disorder, there is growing evidence that unbiased marker sets can detect changes in a substantial proportion of biological pathways, even in a single assay using a single cultured cell line<sup>22,25,59,87</sup>.

The most commonly used unbiased assay for image-based profiling is Cell Painting<sup>20,104</sup>, whereby six inexpensive dyes are used to stain eight cell organelles and components, which are imaged in five channels that each capture fluorescent light of a particular wavelength. Developed by the laboratories of Stuart Schreiber and Anne Carpenter, the assay captures several thousand metrics for each imaged cell. Although alternatives exist<sup>169</sup>, most of the current publicly available image-based profiling data were obtained by Cell Painting<sup>82,170</sup>.

Customized image-based assays for profiling can range from a straightforward combination of stains for markers thought to be relevant to a disease to much higher complexity. For example, there are several techniques for performing multiple rounds of staining and destaining, which can yield dozens of channels/biomarkers detected per field of view<sup>171–174</sup>. Imaging mass cytometry also offers substantial multiplexing capacity, but requires expensive, slow instrumentation<sup>175,176</sup>. Multiplexed image profiles can also be created by combining readouts from multiple separate assays, each treated with a different fluorescent probe<sup>177</sup>, although preparing multiple wells per sample limits this approach to relatively small numbers of samples.

The vast majority of image-based profiling is cell based for practical reasons, but animal-based screening is possible and powerful, primarily in the worm *Caenorhabditis elegans*<sup>178</sup> and zebrafish<sup>179</sup>. Most image-based profiles use static images as input, but notable exceptions exist; successful studies include identifying antipsychotic-like compounds in zebrafish behavioural assays<sup>69</sup>, identifying dynamic mitotic phenotypes<sup>31</sup> and detecting drug and growth surface response in time-lapse imaging of neutrophil granulocytes from patients with asthma<sup>180</sup>. While standardizing on a single image-based profiling assay such as Cell Painting allows profiles across experiments and research groups to be usefully shared, these examples demonstrate compelling reasons to choose alternative staining and imaging protocols and biological systems.



**Fig. 1 | Image-based profiling.** **a** | Overview of the typical steps in the workflow for generating image-based profiles from biological samples. **b** | Example images from the Cell Painting assay often used for image-based profiling. It includes six stains labelling eight cellular components, which are imaged in five channels<sup>20</sup>. ER, endoplasmic reticulum.

or polypharmacology (BOX 4) in the shared representation space. Classical feature selection methods reduce thousands of raw features to hundreds according to feature correlation, but typically without a dramatic increase of discriminative power. The first step in the analysis of an image-based profiling study is often a further reduction of the high-dimensional phenotypic space to a 2D or 3D representation for visualization purposes using computational methods such as principal component analysis (PCA), *t*-distributed stochastic neighbour embedding (t-SNE), and uniform manifold approximation and projection (UMAP)<sup>40–42</sup>. Although such visualizations are a non-optimal representation of the rich information of high-dimensional profiles, they can provide a first sense of the similarity patterns in the experiment.

A substantial proportion of known bioactive compounds yield a significant change in image-based profiles: 68% in one study using the Cell Painting assay<sup>22</sup>. Yet, due to the curse of dimensionality, a limited number of MOAs can be readily distinguished by the nearest-neighbour strategy: typically only one or two dozen form well-delineated clusters, and rarely are new clusters identified. The commonly observed clusters tend to represent compounds with disruptive mechanisms, such as histone deacetylase inhibition and microtubule interference. The ability of image-based profiling to yield distinctive clusters of compounds for other, often subtler mechanisms is limited and has not increased much over the years<sup>43</sup>. In a recent study where 15 reporter cell lines were evaluated, even after feature selection only

20 of 83 MOA classes were readily distinguishable by image-based profiles in the best single cell line, and 41 were readily distinguishable, cumulatively across all 15 cell lines<sup>44</sup>.

**Machine learning: powering a new wave of image-based profiling.** Driven by the recent availability of ever-larger volumes of images, the field has begun to turn to machine learning, and specifically deep learning, to improve extraction of relevant signal from profiles<sup>45</sup>. In essence, a deep neural network is a concatenation of multiple layers, each trained to reweight and transform data to achieve a goal, such as detecting objects in images or predicting the compound to which a cell has been exposed. An input layer introduces data such as image-based profiles or raw images into the network. Each layer then reshapes the data as it flows through the network. Finally, the output layer exports the completely transformed data.

These techniques thus learn a more sophisticated representation than simple reweighting. Self-supervised methods focus on information that can be learnt from different documentations of the same object (perturbation, sample or cell type) (panel c of the figure in BOX 3). For example, if the same microscopy field is documented with two views with different stains, a method can be trained to predict one view from the other; in the process, it learns to filter out the noise in either view to leave only the robust signal<sup>46</sup>. Related methods train to differentiate multiple cells from a single view or multiple

#### Polypharmacology

The property of a compound whereby it interacts with more than a single target.

#### Neural network

A machine-learning architecture whereby features of a sample (for example, image pixels or image-derived metrics) are fed into a network of nodes, which collectively learn to produce the correct answer for that sample by each node adjusting its contribution (weight) to the final answer.

views (that is, replicates) that document a specific perturbation (for example, chemical or genetic) from those documenting any other perturbations<sup>47,48</sup>. These machine-learning methods require more substantial volumes of images for training but position all samples in a cleaner shared representation space than unsupervised methods. The more samples that are included, the more mechanisms and biological processes this shared representation space will try to simultaneously capture, challenging the representation's discriminatory power.

Like unsupervised or self-supervised approaches, supervised methods require a large volume of images for

training, but in addition they need training data, including training data points with activity labels for predefined tasks, and they benefit from side information. If large label sets are available that annotate compounds with their activities in validated assays, supervised methods can be trained, for example, to deconvolute the relevant signal from raw images or image-based profiles, including the output of self-supervised methods<sup>25</sup>. In contrast to unsupervised or self-supervised approaches, which learn a single representation space to position samples irrespectively of tasks of interest, the output layers of advanced supervised approaches produce

### Box 3 | Increasing discriminative ability in phenotypic space

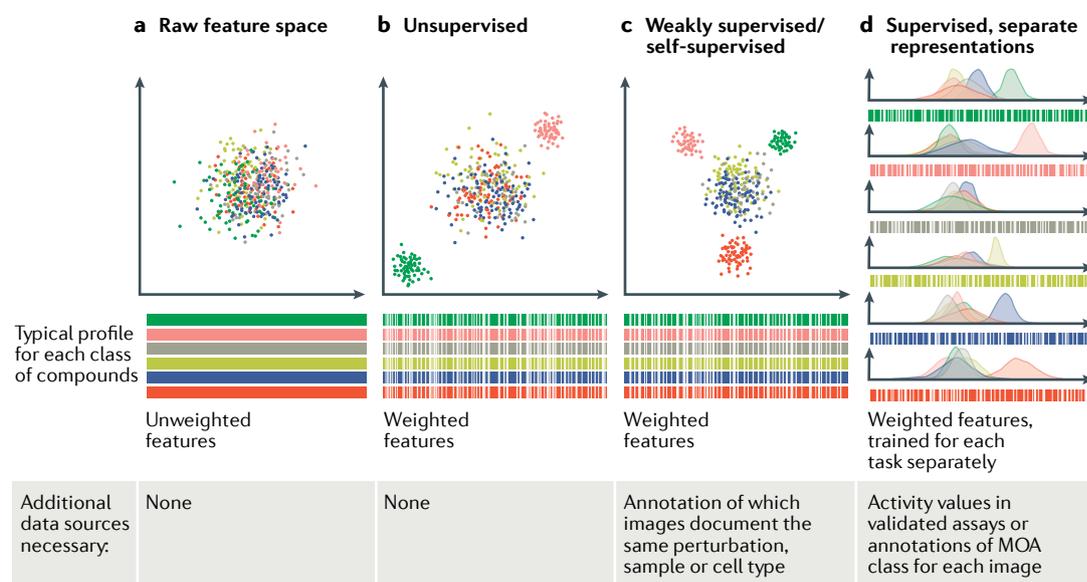
Various computational strategies can improve results in image-based profiling, shown here as a theoretical example for identifying the mechanisms of action (MOAs) for a set of compounds. Each tested sample (in this case, each compound) is represented as a dot in the phenotypic space, where distances between dots reflect the similarity of images of cells treated with the compounds. Even with the best strategies, many MOA classes will not be readily detectable or distinguishable. The strategies shown in panels a–c in the figure are useful to get a quick view of clusters of samples in a given dataset even if no sample annotations (for example, MOA labels) are available to indicate what each cluster represents. If these strategies are to be used to assign an MOA class to each sample, the approach would be called 'semisupervised', because after creation of this shared space, close proximity to compounds with known MOAs (if any) would be used to assign MOAs.

If raw features are extracted from images and placed into phenotypic space with no adjustment, samples do not typically form noticeable clusters (panel a). All features are equally weighted, such that those most relevant to the task at hand are typically drowned out by irrelevant, noisy or redundant features.

An unsupervised machine-learning method can select the appropriate weights for each feature in order to emphasize important ones and suppress noisy or redundant ones (panel b). This process can be performed by techniques ranging from principal component analysis to autoencoders and is depicted as individual features becoming darker or lighter in the profiles shown below the scatterplot. This weighting typically allows classes of compounds with strong morphological phenotypes (pink and green, here) to be distinguished.

Self-supervision strategies leverage redundant image information (for example, multiple replicates per perturbation, different channels for each microscopy view or multiple instances per cell type) to distil a robust signal (panel c). The phenotypic space of reweighted features learned by the network often yields a better discrimination of MOA classes.

When MOA annotations or assay activity values are available, the problem becomes supervised (panel d). Although one can learn a common (shared) representation using supervision (not shown), in practice this does not typically work well given the scale of MOA annotations available. Instead, here, for each MOA class, features in the profile are specifically weighted by training on samples with known activity in a corresponding assay. This can be especially helpful to handle polypharmacology; in our example, the blue and yellow-green compounds are distinguishable from other compounds via individual models in panel d but not by any shared model (panel a, b or c). This could result from all compounds in the blue and yellow-green classes having the MOA 'X', but half of these compounds may have an additional MOA, 'Y', whereas the other half have a different additional MOA, 'Z'. Training separate models for MOAs 'X', 'Y' and 'Z' allows the features to be weighted so that each model ignores features associated with the other MOAs and more readily focuses on the targeted MOA.



## Box 4 | Inherent challenges in profiling

Apart from the curse of dimensionality (the masking of a signal of interest by the noise and confounding signal of other features in a profile), all high-dimensional profiling methods, including imaging, face additional challenges.

Technical artefacts such as batch effects and plate layout effects typically make images from particular locations on a plate or from particular batches look more similar to each other than to images of the same sample in different plate locations or different batches<sup>181</sup>. Even different lineages of a given cultured cell line can be distinguished on the basis of image-based profiles<sup>61</sup>. Although arguably the ability to detect technical artefacts speaks to the sensitivity of image-based assays, how to best mitigate the effects remains unresolved.

When one is identifying disease phenotypes from patient samples, there is a risk of confounding by various genetic or sample biases that may not be relevant for the disease. For example, samples from diseased patients might have a different demographic balance or might be more affected by particular technical artefacts than controls, such that features associated with the disease samples lack true predictive power. It is difficult and expensive to carefully identify and control for all possible confounding factors, increase sample sizes and/or test separately obtained cohorts, so these present real risks.

For compounds, a unique challenge is polypharmacology (that is, the simultaneous engagement of a small molecule with multiple targets or processes). Historically, successful drugs have been thought to be highly specific in their activity, but it has become clear that polypharmacology is the rule rather than the exception, even for marketed drugs<sup>182,183</sup>. Polypharmacology results in profiles that convolve multiple signals, only one or two of which may be relevant for the intended activity, while the other signals may indicate favourable, unfavourable or undefined effects of the compound. Attempting to identify the mechanisms of a polypharmacological compound can be a challenge: at best, its profile shows strong similarity to the single most prominent mechanism it features; at worst, the profile becomes suspended in a poorly interpretable 'no-man's land'.

a representation for each predefined task (panel d of the figure in BOX 3).

Drug development companies are increasingly treating their existing data as a resource to be leveraged for machine learning. The promise of improved profile resolution through rich data and machine learning is currently rekindling industrial interest in phenotypic profiling. Although the pay-off may be great, serious machine-learning effort comes at a high cost, largely in terms of recruiting and retaining highly sought-after experts and giving them the time and computational resources to build and maintain suitable software. Machine-learning strategies also risk problems such as overfitting, technical artefacts or confounders in the data, and bias<sup>49,50</sup> (BOX 4). The ideal experimental set-up is often cost-prohibitive, and avoiding all possible confounding factors is impossible. Thus, practically speaking, practitioners must carefully design a combined experimental/data strategy that diminishes the risks; this requires deep understanding of the problem. One mitigant of machine-learning confounders in drug discovery is that for many of the applications we will discuss, the goal is simply to triage compounds that will subsequently be tested thoroughly, such that false positives do not proceed. Nevertheless, in biomedicine, machine-learning methods that are interpretable — that is, where one can learn what features of samples are being used in decision-making — can sometimes ensure that technical artefacts are not the source of the signal and can provide insight into disease and drug mechanisms.

**Profile-based phenotype discovery and screening**

It takes many months to years to develop a conventional image-based assay for screening, even after years of basic research into the mechanisms of a disease. One time-consuming step is hypothesizing and engineering one or a few relevant assay readouts, most often the staining of a molecule or other cell component. The time spent on this step can be dramatically reduced by using a generic staining approach such as Cell Painting (BOX 2); applying image-based profiling to such images typically yields more extensive feature sets than customized staining. In conventional image-based assay development, researchers also spend considerable effort selecting suitable conditions for testing a drug's impact on putative disease-associated cell phenotypes, which might include the cell type and culture conditions, appropriate stimuli and the duration of drug exposure. Here too, timelines can be cut by taking a more generic approach; for instance, one that relies on a simple cultured cell system and standardized assay conditions that are not customized to the diseases at hand. However, developing customized assay readouts with a careful selection of sample material, stimuli and time points may yield a higher probability of finding a disease-relevant phenotype, perhaps justifying much longer assay development timelines. Furthermore, many disease programmes will involve highly customized assays in the long run as secondary assays, so one could argue these may as well be developed up front for primary screening unless their throughput is limiting.

Regardless of where an image-based assay is on the spectrum from unbiased to customized, the typical steps to profile-based phenotype discovery and screening (also known as signature discovery and signature-based screening) are as follows.

- Prepare sets of biological samples that represent the disease state and the healthy state via strategies described in detail later in this section (TABLE 1).
- Capture image-based profiles and attempt to identify any reproducible phenotypic difference between the diseased and healthy samples. This phenotypic difference will become the screening objective — that is, the phenotypic assay readout. This readout might be a single feature extracted from a single image channel (in essence, a conventional high-content assay), or it might be a multifeature profile that discriminates between the diseased and healthy states. Machine learning and side information may be required to filter out confounding signals and noise. The discovery of novel phenotypes associated with a disease may itself yield new mechanistic insights into the disorder.
- Optionally, simplify the assay (for example, remove unnecessary fluorescent markers) to reduce its cost, or add markers that serve a useful triaging function for hits.
- Use the identified processed phenotype or profile to (a) test thousands of chemicals for their ability to reverse the disease morphology to resemble the healthy state or (b) virtually query an existing dataset of image-based profiles from chemical perturbations of healthy cells to identify those whose perturbation yields the 'opposite' (anticorrelated)

phenotype, indicating a favourable impact on the same pathways as are impacted by the disease. In addition, compounds that produce the same (correlated) profile as the disease can potentially provide useful mechanistic information.

- Optionally, identify or validate novel targets for the disorder by (a) testing a genome-scale set of genetic perturbations for their ability to modify the disease-related phenotype or (b) virtually querying an existing genome-scale dataset of image-based profiles from genetic perturbations of healthy cells to identify or validate genes whose perturbation yields the same (correlated) or opposite (anticorrelated) phenotype. Novel, validated targets could then be fed into conventional target-based drug discovery pipelines.

**Identifying a disease-associated phenotype.** The first step, identifying a disease-associated phenotype in images, is crucial<sup>51</sup>. Several strategies exist for identifying a cellular disease state with a profile that differs from that of the healthy state (TABLE 1). First, patient-derived cells are a physiologically relevant choice, assuming a sufficient number of independent patients are available to yield confidence that phenotypic differences are associated with the disease rather than due to the inherent morphological variability of cell lines across patients. Caution must be exercised, as high-dimensional profiles are prone to confounding factors (BOX 4), whereby features that seemingly distinguish between healthy and diseased states may in fact reflect age, genetic, exposure or sample biases that are not relevant to the disease. Nevertheless, many reproducible image-based phenotypes have been discovered, often inadvertently, as scientists stained and visually examined cells, typically using common markers such as organelle dyes. For example, unusual mitochondrial structure was identified in fibroblasts and lymphocytes from patients with bipolar disorder<sup>52</sup> and in fibroblasts from patients with Leigh

syndrome<sup>53</sup>, and normal human fibroblasts can be differentiated from Huntington disease fibroblasts using only tubulin staining<sup>54</sup>. Image-based profiling offers a way to scale-up and systematize this kind of serendipitous discovery.

A second approach to identifying a disease-associated phenotype is especially suited to disorders caused by loss-of-function mutations in single genes. A gene's expression is decreased using RNAi or CRISPR and then the morphological impact on cells is examined, taking care to identify off-target effects<sup>55</sup>. This approach was used to detect an impact on cell structure (as detected by stains for DNA, actin and VE-cadherin), which was obvious by eye, of RNAi knockdown of *CCM2*, the gene associated with the loss-of-function disorder cerebral cavernous malformation<sup>56</sup>. Through screening, the researchers identified small molecules that reverse the phenotype. Remarkably, compounds chosen by a computational analysis outperformed those chosen by eye in secondary physiological experiments. This study, conducted at the University of Utah, led to the launch of the biotechnology company Recursion. The company has since identified hundreds of disease-associated image-based phenotypes available for parallel screening and has placed four drugs into clinical trials. In a similar strategy, an academic research team took a comprehensive approach to investigate alleles from genome sequencing studies: in adipocytes differentiated in vitro, they ablated 125 genes associated with type 2 diabetes and clustered the resulting image-based profiles, identifying novel lipodystrophy genes<sup>57</sup>. Another team mutated zebrafish orthologues near 132 schizophrenia-associated alleles and created behavioural and brain structural image-based profiles, prioritizing candidates for further study<sup>58</sup>.

A third approach to phenotype identification for screening, gene overexpression, is especially suited for testing alleles in protein-coding regions of genes that are known or hypothesized to cause disorders. The mutant form of a given protein is exogenously expressed in cells, and its image-based profile is compared with that of cells overexpressing the wild-type form<sup>59</sup>. Although overexpression may itself impact the cell's structure and function, observing a differential signature between wild-type and mutant overexpression yields a strong hypothesis for disease-related impact. The Taipale laboratory at the University of Toronto is using this strategy for a set of monogenic disorders, with the twist that each disease-associated protein is flag-tagged so that disease-associated changes in protein localization can be detected in addition to changes in cell morphology (M. Taipale and J. Lacoste, personal communication). One could envision this strategy becoming the routine next step after every sequencing study that yields long lists of hypothesized disease-associated variants: generate appropriate genetic reagents for all alleles and test their image-based impact in a suitable cell line. This approach might allow clustering of alleles into different functional groups, such that a subset of each group could be studied in more disease-specific assays. Already, image-based profiling of gene overexpression is showing promise in determining the impact

Table 1 | Strategies for identifying a 'disease state in a dish'

Strategy to create disease state	Disease state (example)	Healthy state (example)
Patient-derived cell lines	Cells taken from patients with asthma	Cells from healthy volunteers
Gene knockdown or knockout	Cells with loss-of-function disease-associated gene <i>CCM2</i> knocked down by RNAi or CRISPR	Mock-treated control cells
Allele overexpression (optional: tag the protein of interest to examine its localization in addition to the cell's overall morphology)	Cells overexpressing a variant associated with lung cancer	Cells overexpressing the wild-type form
Cell lines engineered by gene-editing techniques	Cells containing a non-coding variant associated with schizophrenia, in its endogenous location	Mock-treated control cells lacking the variant
Existing small molecules with known beneficial effects	Any cell-based or organism-based model system	Treatment with small molecules of known benefit for the disorder

of so-called variants of unknown significance in patient tumours (J. Caicedo and J. Boehm, personal communication), analogous to a prior successful mRNA profiling approach<sup>60</sup>.

A fourth strategy is to genetically modify cells, which, unlike overexpression, allows interrogation of both coding and non-coding variants. For now, the single-cell cloning procedures needed for gene editing techniques are too slow and inconsistent for large-scale use, and lineage artefacts may confound accurate phenotype detection<sup>61</sup>. Nevertheless, the approach can work well for testing individual alleles. Recently a change in image-based profile was detected for cancer cells genetically modified to express mutant focal adhesion kinase (FAK) versus its wild-type form<sup>62</sup>. This genetic perturbation was intended to mimic pharmacological treatment targeted to FAK and simplify the pathway engagement while avoiding off-target concerns with FAK small-molecule inhibitors. The study authors then screened small molecules to identify putative synergistic combinations with FAK inhibitors, identifying histone deacetylase inhibitors as potential novel kinase inhibitor drug combinations for cancer. It may soon be feasible to rapidly and systematically genetically engineer cell lines containing disease-associated alleles in their endogenous locations, which opens the door for this approach to be conducted more systematically. One can even engineer model organisms to carry human disease-associated alleles to create 'avatars' for identifying image-based phenotypes and subsequent testing of compounds<sup>63–67</sup>.

A fifth approach to identify a screenable phenotype is to identify changes in image-based profiles associated with existing drugs for a particular disease of interest, then use those phenotypes to identify new candidate drugs and often their MOA as well. This approach has identified new behavioural phenotypes and potential therapeutics in zebrafish time-lapse imaging using existing psychotropics<sup>68</sup>, antipsychotics<sup>69</sup>, appetite modulators<sup>70</sup> and anaesthetics<sup>71</sup> as the query compounds. Image-based profiling also revealed morphological changes associated with 61 structurally diverse free fatty acids, identifying those that are associated with lipotoxicity in insulin-secreting pancreatic  $\beta$ -cells (N. Wieder and A. Greka, personal communication).

These five strategies for identifying screenable phenotypes differ in terms of the ease of assay development, the degree of customization required and how closely they reflect the human disease state. At one end of the spectrum, one can choose a single cell type and set of staining conditions to search for many disease phenotypes in parallel. In this case, the assay development time for each disorder is zero, but many phenotypes will be missed due to a lack of a suitable stain or appropriate biological conditions. Still, the simplicity, cost-efficiency and scalability of the strategy make it an attractive approach: although customized follow-up validation assays must still be created, they can be more expensive and of lower throughput, and they need to be made only for the subset of disorders showing promising hits. By contrast, there will be more chance of identifying a phenotype for any particular disease if the assay is more specifically tailored to the disorder, although this

increases the assay development time. For example, one could explore several healthy and diseased systems to be imaged, such as several cell lines, primary cell types, mixtures of co-cultured cells or even differentiated cell cultures. In addition to more physiologically relevant cell/organism systems, stains can be customized to reflect several biological processes associated with the disorder. At the limit, these customizations end up as equivalent to conventional phenotypic assay development.

It should be noted that reversing disease-associated phenotypes identified via the strategies described above will not always yield effective drugs. First, cultured cells do not reflect the full intricacies seen in a human organism; many disease mechanisms are not cell autonomous, and many drugs will have different effects in patients' whole bodies, especially considering their different genetic backgrounds and environmental exposure. These limitations are shared with all biochemical or cell-based drug screening methods and are fairly well appreciated. However, profile-based phenotype discovery introduces additional concerns. For example, phenotypes detected as described above might be completely incidental to the phenotype that causes symptoms for patients, such that drugs reversing the phenotype are ineffective. Even worse, the detected phenotype might reflect the cells' attempt to mitigate the impact of the disease perturbation, such that drugs reversing the phenotype would aggravate the condition in patients. Nevertheless, the speed that profiling offers allows extra time to rule out these kinds of problems.

The drug industry as a whole has begun to adopt image-based profiling to inform target identification and validation, phenotype discovery and assay development before screening. However, for screening itself, wherever possible, the industry tends to prefer a customized assay focused on a molecularly defined target or pathway that adequately reproduces the profile-based findings. Focused assays facilitate determining structure–activity relationships (SARs), selecting mechanistically inspired biomarkers of efficacy and identifying on-target liabilities that correlate with desired on-target activity. Moreover, contrary to expectations, compounds identified in early mRNA and image profile-based phenotypic screening efforts do not seem to have been validated in animal models more often than compounds identified in more straightforward target or pathway-based screening approaches — although, admittedly, no systematic analysis has been reported to date. Large pharmaceutical companies also tend to focus on selected disease areas, and in those areas they prefer to maximize the chances of success by screening large compound libraries, which can be more efficiently accommodated in streamlined customized assays, especially if profiling will nevertheless eventually be used in follow-up assays. However, thousands of diseases fall out of industry scope. As a result, biotechnology start-ups (Recursion and insitro, most notably<sup>27</sup>) are exploring this space. As mentioned in the introduction, they rely on profile-based phenotypic screening of smaller libraries in generic assays to more efficiently evaluate the potential of compound intervention across that less frequently studied disease spectrum.

### Lead generation

In the drug discovery process, screening is followed by lead generation, in which hundreds of screening hits are narrowed down to just a few lead candidates. The various applications of image-based profiling for lead generation can use either unbiased assays (which could be systematically applied across disease areas) or customized assays with relevant biomarkers (which can be tailored to particular disease areas) as described in the previous section and BOX 2.

### Hit expansion, lead optimization and SAR studies.

During lead generation, two complementary activities — hit expansion and lead optimization — are pursued to triage and modify compounds for the most favourable attributes for further drug development. Conventionally, lead generation is guided by several individual assays, including the primary screening assay. These typically have simple and rapid readouts, reflecting single effects of molecules, such as enzyme activity, reporter expression, aggregation of a target protein or cell viability. Image-based profiling can be a tractable path for lead optimization, because it is quick, is sensitive and covers a broad, though not comprehensive, range of biology. Rather than relying on multiple assay readouts to select compounds with desired qualities, image-based profiling presents the possibility of selecting compounds with several desired qualities using a single assay<sup>72</sup>. Moreover, image-based profiling can quickly and comprehensively establish relationships among hits into biological clusters, independently of compound structure, providing direction for prioritization<sup>37,73,74</sup>. In addition, hits can be triaged quickly across disease-specific cell types with different genetic backgrounds using the same image-based profiling assay to assess the efficacy of compounds<sup>75,76</sup>.

Recently, several SAR studies applied image-based profiling using Cell Painting to assess the biological activity of newly synthesized compounds and build diversity sets for focused libraries<sup>77–80</sup>. The quick deployment and richness of the readouts enabled researchers to efficiently examine the effects of compound modifications across a wide range of biological activity. The unbiased assay proved to be sensitive enough to distinguish the effects of stereochemistry on activity. Furthermore, because an unbiased image-based profiling assay covers such a broad spectrum of activity, modifications that act on different mechanisms can be identified<sup>78,79</sup> — an advantage over conventional SAR studies.

Because these readouts reflect the consequences of molecular modifications for biological activity, this approach can be used to examine biological activity of hits from targeted screening, provided the cellular model system contains the molecular target in a biologically relevant context. Importantly, one must confirm that observed effects are indeed the result of target modification by the test compound and not the result of cellular adaptation via an alternative mechanism or polypharmacology. Furthermore, and more challenging, is the possibility of no observed response. The absence of a response may not be due to the absence of a target or target binding, but could be due to complex cellular pharmacology — a combination of compound concentration, duration of

treatment and cellular metabolism — which yields an observable morphological profile similar to no-treatment controls.

In this context, it is helpful to understand the relative association of image-based features with the desired phenotype or with off-target activities. However, this interpretation step is still a challenge beyond simplistic cases in which a small number of image-based features are strongly affected and easily interpretable, such as a change in protein localization. Although it has yet to be demonstrated for lead optimization, machine learning is emerging as a powerful tool to deconvolute multiplexed profiles of studied hits into virtual assays for distinct biological activities (as described in the next section), which can be favourable, neutral or unfavourable. Whether translated into interpretable activities by nearest-reference or by more advanced machine-learning approaches, the rich information in image-based profiles for newly synthesized compounds can be used to track the incremental effects of subsequent compound modifications during lead optimization.

**Predicting assay activity.** Compared with traditional hit expansion approaches, machine learning offers some attractive alternatives that make use of unbiased image data and provide an advantage over conventional SAR studies. By nature, image-based prediction can generate hits that are structurally diverse because predictions are based on activity in a biological system rather than chemical structure. Also, being inherently multiplexed, image-based methods are potentially capable of predicting the activity of hit compounds in assays unrelated to the assay that generated the image-based profiles.

This hypothesis was verified by a multi-institution team using image-based profiles from Janssen to successfully predict the activity of structurally diverse hit compounds in screening assays. In two validation studies, the strategy yielded a 60-fold to 250-fold increase in hit rates compared with the original screening assays<sup>25</sup>. Using the program CellProfiler to perform classical image segmentation and feature extraction, they leveraged supervised machine learning to predict the activity of a large set of compounds in given assays, based on archived imaging data on those compounds in an unrelated assay. Building on this research, a later team constructed a novel network architecture, GapNet, to predict ChEMBL-derived<sup>81</sup> compound annotations for a 30,000-compound Cell Painting dataset<sup>82</sup>: 32% of the assays could be well predicted<sup>83</sup>. In addition to the practical implication — that many expensive screens might be replaced effectively with a computational prediction step — this result also indicates that a substantial proportion of biological pathways of interest are captured in a single imaging assay. Furthermore, by combining image profiles with compound structural information, machine learning was able to predict active chemical structures *de novo*<sup>26</sup>.

If these machine-learning strategies work well in practice, one could envision the sizes of primary screens being so substantially reduced in the future as to fundamentally change the typical drug discovery process, with

#### Hit expansion

The selection of compounds that were not tested in the primary screen, to broaden the diversity of the chemical space for hit selection. Compounds are selected on the basis of similarities in structure or biological activity to candidate hits.

#### SAR studies

An iterative process for lead optimization in which assays are applied to determine the effect of successive structural modifications to a compound on activity.

heavier reliance on computational predictions rather than experimentation in early stages.

**Predicting toxicity.** Identifying a compound's liabilities during lead generation is as important as assessing its efficacy; roughly 17% of phase III trial failures are due to safety concerns<sup>84</sup>. A simple cell-based assay such as Cell Painting, which could readily be scaled to an entire million-compound collection, is appealing for toxicity prediction given widespread enthusiasm for reducing the use of intact animals for testing both pharmaceuticals and environmental chemicals<sup>85,86</sup>.

Recently, a team at the US Environmental Protection Agency used image-based profiles from the Cell Painting assay to characterize selected chemicals' bioactivity and toxicity<sup>87</sup>. One nuance of toxicity testing is that the dose at which effects are seen is paramount to discerning whether the compound will be toxic in humans. By comparing the toxic concentration thresholds for each chemical generated with other methods with thresholds generated with image-based profiling, the team concluded that image-based profiling is a viable, cost-effective alternative for chemical safety assessments. Image-based profiling of full dose–response curves for large-scale chemical libraries (for example, six dose points for millions of compounds) would require significant investment, but subsets of doses or subsets of compounds are readily feasible. Cell Painting's ability to predict various specific cell health assay readouts that provide mechanistic information, such as stalling in various cell cycle states<sup>88</sup>, provides further motivation for the collection of large-scale Cell Painting data on compounds for pharmaceutical, agricultural and environmental use.

Other methods combine image-based profiling with machine learning to identify the MOAs of toxicants in different types of cells. Image-based features can predict nephrotoxicity of drugs, chemicals and toxicants targeting human renal proximal tubular cells<sup>89</sup>. A high-throughput *in vitro* phenotypic profiling for toxicity prediction (HIPPTox) system predicted pulmonary toxicity more accurately than cell viability assays in human lung cell lines<sup>90</sup>. Image-based profiling identified rosmarinic acid as a candidate with cardioprotective effects against the toxicity of doxorubicin<sup>91</sup>. This approach also revealed the toxic effects of bisphenol A and its analogues on a testicular cell co-culture model by simultaneously measuring multiple adverse end points such as changes to nuclear morphology and cytoskeletal structure<sup>92</sup>.

### Identifying the MOA

Elucidating the MOA of a drug provides a deeper understanding of its biological activity<sup>68,93–96</sup>, increases its chances of clinical approval and enables the design of novel drugs<sup>97</sup>, and provides insights into the potential to separate unfavourable effects from favourable effects if they are driven by different targets<sup>98</sup>.

Identifying the MOA and/or targets of a hit or lead presents a major challenge, particularly for candidates arising from phenotypic rather than target-based screens. No single experimental technology can definitively

identify a compound's MOA. Instead, the most likely MOA is often inferred from a combination of several complementary methods<sup>99</sup>. Proteomics approaches can be used to identify unknown protein targets after pull-down with a tagged compound or if an untagged compound shifts the target's stability<sup>100</sup>. Assay panels can readily identify the engagement of a compound with a set of predefined targets such as kinases, but only a fraction of the druggable proteome can be queried in this way, and every assay has its limitations. Finally, phenotypic profiling, including image-based profiling, can also contribute to MOA identification, as we describe in this section.

There are three broad categories of image-based profiling approaches to determining the MOA. The first approach, known as guilt-by-association with annotated compounds, involves comparing image-based profiles of compounds with unknown MOAs with those of well-annotated compounds to identify neighbours<sup>101</sup>. In this approach, image-based profiles are clustered according to the assumption that drugs with similar MOAs generate similar phenotypic signatures<sup>102</sup>. This strategy has been successfully implemented across studies spanning more than a decade<sup>73,75,102–106</sup>. However, as described in the section Analysis techniques evolve, image-based profiles and classical methods have effectively grouped compounds for many classes of MOAs but are ineffective for other classes, and polypharmacology is confounding. Nevertheless, there are many examples of novel MOA discoveries being made by this route, including identifying novel inhibitory activity of siltitasertib<sup>107</sup>, the MOA of autoquin (an autophagy inhibitor with a non-protein target)<sup>108</sup>, the translation inhibition of phenomycin<sup>109</sup>, the activity of synthesized pseudo-natural products, pyranofuropryridones, which remained inactive in other common assays<sup>110</sup>, and useful components of natural products<sup>111–114</sup>.

We expect that in coming years this guilt-by-association strategy will become increasingly sophisticated by leveraging advancements in machine learning, availability of more extensive reference annotation, integration with other data sources and optimization of cell lines. As the choice of cell line informs MOA prediction accuracy<sup>21</sup>, effort has been made towards engineering reporter cell lines<sup>43</sup> and identifying cell line-invariant features<sup>115</sup>. Others have interrogated genetic heterogeneity, finding distinct morphological responses to serotonin modulators across breast cancer cell lines<sup>116</sup>. The integration of image and transcriptome data into ensemble approaches has shown promise to improve MOA determination for synthetic small molecules, natural products and identified bioactive metabolites<sup>21,117</sup>. Image-based profiles of cells treated with natural products have also been combined with mass spectral features of the same natural products to discover a novel family that causes endoplasmic reticulum stress<sup>118</sup>. If image-based and other profile types carry complementary information, then creating and integrating both kinds of profiles will be useful; if they are instead largely redundant, then researchers could acquire only one modality and build translators to convert one type of profile to the other<sup>119</sup>.

In the second approach, known as guilt-by-association with perturbed genes, the image-based profile of the drug is matched to that of cells perturbed with a specific genetic reagent, thereby yielding a hypothesis for the drug's target and MOA<sup>120</sup>. As pioneered with yeast mutants<sup>121–123</sup>, this approach has shown success in cell-based, image-based profiling experiments, although it is not yet widely used. Genetic perturbation can be achieved by a variety of techniques. Historically, perturbation by small interfering RNA (siRNA) has been most common<sup>124–126</sup>, although seed effects can be confounding<sup>127</sup>. Another strategy is to use CRISPR–Cas9 (REF.<sup>128</sup>) to knock out genes of interest as it is less susceptible to undesirable and confounding off-target effects of siRNA<sup>55,129</sup>. Overexpressing genes is another option that can yield distinctive image-based profiles<sup>59</sup>, which in theory could be matched to a chemical's image-based profile. Most chemical compounds inhibit a protein's function, but it is currently unclear whether matching a suppressed gene more effectively determines the MOA as opposed to looking for a profile that is anticorrelated to that of an overexpressed gene.

The third approach, which is based on rescue experiments, might be more accurate than the similarity-matching strategies described above, but would require new, large-scale experiments for each query molecule rather than a simple computational matching exercise. In a rescue experiment to determine the MOA, cells treated with a given drug that induces a particular image-based profile would be treated with genetic perturbations to identify any that can reverse the profile, such that the cells resemble untreated cells. There are currently no published examples of this approach, although pooled optical profiling puts this type of experiment closer to reality by offering the capability to test hundreds to thousands of genetic reagents in parallel<sup>29</sup>.

As mentioned in the section Analysis techniques evolve, most image-based MOA prediction studies use a nearest-neighbour approach. This procedure assumes that each drug has a single MOA, but it is now well appreciated that few chemicals impact only a single protein within a cell. Instead, polypharmacology, including off-target effects, is commonly seen (BOX 4). One or two of these targets may be relevant in the context of the disease of interest, but others may also be reflected in the unweighted profile and obscure the profiles resulting from the targets of interest. For certain MOAs, particularly those with a strong and broad impact on cellular morphology, clusters of compounds are easily identified with straightforward methods<sup>130</sup> but such methods fail for many MOA classes. Hence there is a need for MOA determination methods that account for polypharmacology. Deep learning-based approaches have been developed<sup>25,83</sup> that are inherently well suited for tackling this problem, as they can produce MOA-specific, reweighted image-based profiles, thereby deconvoluting the complex phenotypes arising from polypharmacology<sup>44</sup>. We are only beginning to scratch the surface of image-based profiling of single perturbations, but testing pairs of reagents has begun for limited sets of small molecules<sup>131</sup> and genetic perturbations<sup>132</sup>. Larger such datasets should lead to a

better understanding of how to identify and deconvolute polypharmacology in profiles. Regardless of how MOA hypotheses arise, ideally more than one structurally different and accurately annotated compound targeting the same target/pathway will be available to lend evidence to a hypothesis.

### Beyond drug discovery

Outside the scope of this Review are many other successes of image-based profiling applied to a wide variety of important biological phenomena and cellular structures, including identifying relevant genes via functional genomics and studying cell responses to growth topologies and differentiation factors<sup>133–138</sup>. Although identifying biomarkers of disease or drug response can be a first step towards drug discovery and a helpful aid in clinical trials, we do not cover in this Review the tremendous strides that have been made in image-based diagnostics, in some cases by using deep learning on unlabelled samples to create label-free diagnostics<sup>139</sup>. Examples range from suggesting a diagnosis<sup>140</sup> to predicting patient outcomes<sup>141–143</sup> or even molecular phenotypes<sup>144,145</sup>. Lastly, we do not cover interventional, personalized medicine applications here, where patient samples are treated with various potential therapies and imaging is used to measure responses. Examples include the image-based profiling of the response of a bacterial strain isolated from a patient to various antibiotics<sup>146</sup>, of organoids derived from a patient with cystic fibrosis to drugs<sup>147</sup> and of tumour cells or organoids to various chemotherapies<sup>148–150</sup>.

### Future directions

We fully expect advancements in the field of image-based profiling — both computational and biological — will progress rapidly in the next 5 years as the approach gains attention.

On the computational side, deep learning is already beginning to accelerate drug discovery by tackling diverse problems in the process<sup>151</sup>, and image-based profiling will be among the major beneficiaries of advancements in computer vision and predictive algorithms<sup>152–154</sup>. Deep learning can process raw microscopy images to produce representations that are better suited for downstream analysis and interpretation; cells or cellular substructures can be identified more accurately<sup>155,156</sup>, and improved image-based descriptors can be derived during feature extraction<sup>157</sup>. Deep learning may therefore eventually replace classical image processing and feature extraction algorithms, such as those in the currently most commonly used software program, CellProfiler<sup>158</sup>, or one of its commercial counterparts. Deep networks have been trained for the interpretation of image-based profiles: they can recognize the biological states of imaged cells<sup>159</sup> or predict the biological activities in validated assays for imaged compounds<sup>25</sup>. Convolutional neural networks can also integrate bespoke feature extraction and interpretive tasks in a single process<sup>160</sup>. Emerging anecdotal evidence suggests that provided with enough activity labels, these single-step, end-to-end networks have a predictive performance superior to that of conventional feature extraction

and profiling. This may stem from deep learning's ability to learn specialized composite image features beyond the ones predefined by feature extraction software. Alternatively, it may reflect an implicit encoding of the heterogeneity of single cells within a microscopy image<sup>134</sup>. Capturing single-cell heterogeneity after feature extraction has indeed been found to improve image-based phenotypic clustering<sup>161</sup>. Finally, the flexible architecture of neural networks enables information to flow in from alternative data sources and formats, as input, or as side information.

On the biology side, drug discovery scientists are adopting increasingly complex model systems for their image-based profiling, such as differentiated cell types, tissues, organoids and whole model organisms. These assay systems may more routinely become higher-resolution and multidimensional systems, including 3D and time-lapse image capture. Sequencing-based barcoding methods are enabling larger-scale genetic perturbation libraries to be profiled by imaging<sup>29</sup>. We also expect that data generation and machine-learning approaches will become increasingly intertwined. For example, networks can learn to predict fluorescence patterns from transmitted light images<sup>33</sup> when provided with enough training pairs. This may in the future enable cost-efficient label-free image capture.

The ability to make broad activity predictions on the basis of microscopy screen images that were originally collected for a single mechanism of interest<sup>25</sup> has created an appetite to combine similar approaches with the richer profiles from multiplexed generic staining protocols such as Cell Painting<sup>83</sup>. Biotechnology and pharmaceutical companies are already investing in generating purpose-built image sets that document genetic and compound-induced perturbations. This investment occurs in proprietary settings, but also in the context of public-private partnerships involving multiple pharmaceutical companies, which will ultimately boost the availability of such datasets in the public domain<sup>27</sup>. It remains to be seen whether image-based profiles have reproducibility problems as recently suggested for mRNA profiles in the largest single-site public dataset<sup>162</sup>, and whether machine-learning approaches can reconcile image-based profiles across experimental batches and data generation sites. Nevertheless, we hope that increased availability of high-quality queryable image datasets<sup>163</sup> paired with side information on imaged compounds, genetic perturbations or disease models will in turn inspire the design of yet more powerful machine-learning methods, driving a virtuous circle of discovery.

Published online 22 December 2020

- Haasen, D. et al. How phenotypic screening influenced drug discovery: lessons from five years of practice. *Assay Drug Dev. Technol.* **15**, 239–246 (2017). **This primer describes the hype and eventual more nuanced positioning of phenotypic screening, providing a pharmaceutical industry perspective as machine learning is ready to spark a new wave of opportunities in this space.**
- Singh, S., Carpenter, A. E. & Genovesio, A. Increasing the content of high-content screening: an overview. *J. Biomol. Screen.* **19**, 640–650 (2014).
- Dorval, T., Chanrion, B., Cattin, M.-E. & Stephan, J. P. Filling the drug discovery gap: is high-content screening the missing link? *Curr. Opin. Pharmacol.* **42**, 40–45 (2018).
- Boutros, M., Heigwer, F. & Laufer, C. Microscopy-based high-content screening. *Cell* **163**, 1314–1325 (2015).
- Caicedo, J. C., Singh, S. & Carpenter, A. E. Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.* **39**, 134–142 (2016).
- Herholt, A., Galinski, S., Geyer, P. E., Rossner, M. J. & Wehr, M. C. Multiparametric assays for accelerating early drug discovery. *Trends Pharmacol. Sci.* <https://doi.org/10.1016/j.tips.2020.02.005> (2020).
- David, L. et al. Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Front. Pharmacol.* **10**, 1303 (2019).
- Stinson, S. F. et al. Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
- Merget, B., Turk, S., Eid, S., Rippmann, F. & Fulle, S. Profiling prediction of kinase inhibitors: toward the virtual assay. *J. Med. Chem.* **60**, 474–485 (2017).
- Bowes, J. et al. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discov.* **11**, 909–922 (2012).
- Kurita, K. L. & Lington, R. G. Connecting phenotype and chemotype: high-content discovery strategies for natural products research. *J. Nat. Prod.* **78**, 587–596 (2015).
- Litichevskiy, L. et al. A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. *Cell Syst.* **6**, 424–443.e7 (2018).
- Bouzekri, A., Esch, A. & Ornatsky, O. Multidimensional profiling of drug-treated cells by imaging mass cytometry. *FEBS Open Bio* **9**, 1652–1669 (2019).
- Zampieri, M., Sekar, K., Zamboni, N. & Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **36**, 15–23 (2017).
- Dubuis, S., Ortmayr, K. & Zampieri, M. A framework for large-scale metabolome drug profiling links coenzyme A metabolism to the toxicity of anti-cancer drug dichloroacetate. *Commun. Biol.* **1**, 101 (2018).
- Ye, C. et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.* **9**, 4307 (2018).
- Keenan, A. B. et al. Connectivity mapping: methods and applications. *Annu. Rev. Biomed. Data Sci.* **2**, 69–92 (2019).
- Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
- Corsello, S. M. et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
- Bray, M.-A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016). **This step-by-step protocol article details Cell Painting, the most popular image-based profiling assay; it combines six stains to label eight cell components in a straightforward microscopy assay.**
- Lapins, M. & Spjuth, O. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. Preprint at *bioRxiv* <https://doi.org/10.1101/580654> (2019).
- Wawer, M. J. et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl Acad. Sci. USA* <https://doi.org/10.1073/pnas.1410933111> (2014).
- Weigle, S., Martin, E., Voegtli, A., Wahl, B. & Schuler, M. Primary cell-based phenotypic assays for pharmacologically and genetically study fibrotic diseases in vitro. *J. Biol. Methods* **6**, e115 (2019).
- Johnson, K. et al. A stem cell-based approach to cartilage repair. *Science* **336**, 717–721 (2012).
- Simm, J. et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.* **25**, 611–618.e3 (2018). **This study applies machine learning to microscopy images designed in the context of a specific screening campaign to inform on the activity of compounds in a broad spectrum of biological processes; this has applications in early drug discovery.**
- Méndez-Lucio, O., Zapata, P. A. M., Wichard, J., Rouquié, D. & Clevert, D.-A. Cell morphology-guided de novo hit design by conditioning generative adversarial networks on phenotypic image features. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv.11594067> (2020). **This study proposes an approach to image-based lead optimization: a generative adversarial network is trained on small-molecule structures with and without matching image-based profiles to generate new small-molecule structures that mimic the morphological profiles induced by overexpression of 12 individual genes.**
- Mullard, A. Machine learning brings cell imaging promises into focus. *Nat. Rev. Drug Discov.* **18**, 653–655 (2019).
- Mullard, A., Daphne Koller. *Nat. Rev. Drug Discov.* **18**, 576–577 (2019).
- Feldman, D. et al. Optical pooled screens in human. *Cells* **179**, 787–799.e17 (2019).
- Cooper, S., Sadok, A., Bousgouni, V. & Bakal, C. Apolar and polar transitions drive the conversion between amoeboid and mesenchymal shapes in melanoma cells. *Mol. Biol. Cell* **26**, 4163–4170 (2015).
- Eismann, B. et al. Automated 3D light-sheet screening with high spatiotemporal resolution reveals mitotic phenotypes. *J. Cell Sci.* **133**, jcs245043 (2020).
- Christiansen, E. M. et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* **173**, 792–803.e19 (2018).
- Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).
- Grys, B. T. et al. Machine learning and computer vision approaches for phenotypic profiling. *J. Cell Biol.* **216**, 65–71 (2017).
- Caicedo, J. C. et al. Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017). **Members of the CytoData Society across most laboratories in the field of image-based profiling come together to define best practices for informatics; remarkably, these are still relatively up to date, with the exception of deep learning advancements.**
- Price, J. H. et al. Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J. Cell. Biochem. Suppl.* **39**, 194–210 (2002).

37. Perlman, Z. E. et al. Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).  
**This article arguably launched the field of image-based profiling, as it presents the high-dimensional analysis of images of cells in the presence of chemical perturbations.**
38. Abraham, V. C., Taylor, D. L. & Haskins, J. R. High content screening applied to large-scale cell biology. *Trends Biotechnol.* **22**, 15–22 (2004).
39. Michael Ando, D., McLean, C. Y. & Berndt, M. Improving phenotypic measurements in high-content imaging screens. Preprint at *bioRxiv* <https://doi.org/10.1101/161422> (2017).
40. van der Maaten, L. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
41. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* <https://arxiv.org/abs/1802.03426> (2018).
42. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
43. Kang, J. et al. Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. *Nat. Biotechnol.* **34**, 70–77 (2016).
44. Cox, M. J. et al. Tales of 1,008 small molecules: phenomic profiling through live-cell imaging in a panel of reporter cell lines. *Sci. Rep.* **10**, 13262 (2020).  
**This study quantifies the improvement in the ability of image-based profiling using Cell Painting to distinguish compound MOAs in 15 cell lines versus the best single cell line (approximately twofold); use of data from all 15 cell lines distinguished only about half of the MOAs.**
45. Scheeder, C., Heigwer, F. & Boutros, M. Machine learning and image-based profiling in drug discovery. *Curr. Opin. Syst. Biol.* **10**, 43–52 (2018).
46. Lu, A. X., Kraus, O. Z., Cooper, S. & Moses, A. M. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLoS Comput. Biol.* **15**, e1007348 (2019).
47. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**, 44–53 (2018).
48. Caicedo, J. C., McQuin, C., Goodman, A., Singh, S. & Carpenter, A. E. Weakly supervised learning of single-cell feature embeddings. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* <https://doi.org/10.1109/CVPR.2018.00970> (2018).
49. Teschendorff, A. E. Avoiding common pitfalls in machine learning omic data science. *Nat. Mater.* **18**, 422–427 (2019).
50. Riley, P. Three pitfalls to avoid in machine learning. *Nature* **572**, 27–29 (2019).
51. Pegoraro, G. & Misteli, T. High-throughput imaging for the discovery of cellular mechanisms of disease. *Trends Genet.* **33**, 604–615 (2017).
52. Cataldo, A. M. et al. Abnormalities in mitochondrial structure in cells from patients with bipolar disorder. *Am. J. Pathol.* **177**, 575–585 (2010).
53. Blanchet, L. et al. Quantifying small molecule phenotypic effects using mitochondrial morpho-functional fingerprinting and machine learning. *Sci. Rep.* **5**, 8035 (2015).
54. Hung, C. L.-K. et al. A patient-derived cellular model for Huntington's disease reveals phenotypes at clinically relevant CAG lengths. *Mol. Biol. Cell* **29**, 2809–2820 (2018).
55. Smith, I. et al. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol.* **15**, e2003213 (2017).
56. Gibson, C. C. et al. Strategy for identifying repurposed drugs for the treatment of cerebral cavernous malformation. *Circulation* **131**, 289–299 (2015).
57. Jiao, Y. et al. Discovering metabolic disease gene interactions by correlated effects on cellular morphology. *Mol. Metab.* **24**, 108–119 (2019).
58. Thyme, S. B. et al. Phenotypic landscape of schizophrenia-associated genes defines candidates and their shared functions. *Cell* **177**, 478–491.e20 (2019).
59. Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via cell painting. *eLife* **6**, e24060 (2017).
60. Berger, A. H. et al. High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* **32**, 884 (2017).
61. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
62. Dawson, J. C. et al. A synergistic anti-cancer FAK and HDAC inhibitor combination discovered by a novel chemical-genetic high-content phenotypic screen. *Mol. Cancer Ther.* <https://doi.org/10.1101/590802> (2020).
63. Papanikolopoulou, K., Mudher, A. & Skoulakis, E. An assessment of the translational relevance of *Drosophila* in drug discovery. *Expert Opin. Drug Discov.* **14**, 303–313 (2019).
64. Costa, B., Estrada, M. F., Mendes, R. V. & Fior, R. Zebrafish avatars towards personalized medicine—a comparative review between avatar models. *Cells* **9**, 293 (2020).
65. Astone, M., Dankert, E. N., Alam, S. K. & Hoepfner, L. H. Fishing for cures: the allLURE of using zebrafish to develop precision oncology therapies. *NPJ Precis. Oncol.* **1**, 39 (2017).
66. Vashi, N. & Justice, M. J. Treating Rett syndrome: from mouse models to human therapies. *Mamm. Genome* **30**, 90–110 (2019).
67. Dar, A. C., Das, T. K., Shokat, K. M. & Cagan, R. L. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486**, 80–84 (2012).
68. Rihel, J. et al. Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science* **327**, 348–351 (2010).
69. Bruni, G. et al. Zebrafish behavioral profiling identifies multitarget antipsychotic-like compounds. *Nat. Chem. Biol.* **12**, 559–566 (2016).
70. Jordi, J. et al. High-throughput screening for selective appetite modulators: a multibehavioral and translational drug discovery strategy. *Sci. Adv.* **4**, eaav1966 (2018).
71. McCarroll, M. N. et al. Zebrafish behavioural profiling identifies GABA and serotonin receptor ligands related to sedation and paradoxical excitation. *Nat. Commun.* **10**, 4078 (2019).
72. Kummel, A. et al. Differentiation and visualization of diverse cellular phenotypic responses in primary high-content screening. *J. Biomol. Screen.* **17**, 843–849 (2012).
73. Loo, L.-H., Wu, L. F. & Altschuler, S. J. Image-based multivariate profiling of drug responses from single cells. *Nat. Methods* **4**, 445–453 (2007).
74. Vial, M.-L. et al. A grand challenge. 2. Phenotypic profiling of a natural product library on Parkinson's patient-derived cells. *J. Nat. Prod.* **79**, 1982–1989 (2016).
75. Caie, P. D. et al. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol. Cancer Ther.* **9**, 1913–1926 (2010).
76. Hughes, R. E. et al. High-content phenotypic profiling in esophageal adenocarcinoma identifies selectively active pharmacological classes of drugs for repurposing and chemical starting points for novel drug discovery. *SLAS Discov.* <https://doi.org/10.1177/247255520917115> (2020).
77. Gerry, C. J. et al. Real-time biological annotation of synthetic compounds. *J. Am. Chem. Soc.* **138**, 8920–8927 (2016).  
**This study launched many further examples using Cell Painting interactively with synthetic chemistry for medicinal chemistry and SAR determination.**
78. Nelson, S. D. Jr, Wawer, M. J. & Schreiber, S. L. Divergent synthetic and real-time biological annotation of optically active tetrahydrocyclopenta[*c*]pyranone derivatives. *Org. Lett.* **18**, 6280–6283 (2016).
79. Melillo, B. et al. Synergistic effects of stereochemistry and appendages on the performance diversity of a collection of synthetic compounds. *J. Am. Chem. Soc.* **140**, 11784–11790 (2018).
80. Zimmermann, S. et al. A scaffold-diversity synthesis of biologically intriguing cyclic sulfonamides. *Chemistry* **25**, 15498–15503 (2019).
81. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
82. Bray, M.-A. et al. A dataset of images and morphological profiles of 30,000 small-molecule treatments using the Cell Painting assay. *Gigascience* <https://doi.org/10.1093/gigascience/giw014> (2017).
83. Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S. & Klambauer, G. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *J. Chem. Inf. Model.* **59**, 1163–1171 (2019).  
**This article describes how combining convolutional neural networks with images of broad, generic stains boosts biological information extraction from microscopy images.**
84. Hwang, T. J. et al. Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern. Med.* **176**, 1826–1833 (2016).
85. Thomas, R. S. et al. The next generation blueprint of computational toxicology at the US Environmental Protection Agency. *Toxicol. Sci.* **169**, 317–332 (2019).
86. Paul Friedman, K. et al. Utility of in vitro bioactivity as a lower bound estimate of in vivo adverse effect levels and in risk-based prioritization. *Toxicol. Sci.* **173**, 202–225 (2020).
87. Nyffeler, J. et al. Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling. *Toxicol. Appl. Pharmacol.* **389**, 114876 (2020).  
**This careful analysis describes the use of Cell Painting to report on various aspects of toxicity.**
88. Way, G. P. et al. Predicting cell health phenotypes using image-based morphology profiling. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.07.08.193938> (2020).
89. Su, R., Xiong, S., Zink, D. & Loo, L.-H. High-throughput imaging-based nephrotoxicity prediction for xenobiotics with diverse chemical structures. *Arch. Toxicol.* **90**, 2793–2808 (2016).
90. Lee, J.-Y. J., Miller, J. A., Basu, S., Kee, T.-Z. V. & Loo, L.-H. Building predictive in vitro pulmonary toxicity assays using high-throughput imaging and artificial intelligence. *Arch. Toxicol.* **92**, 2055–2075 (2018).
91. Zhang, O., Li, J., Peng, S., Zhang, Y. & Qiao, Y. Rosmarinic acid as a candidate in a phenotypic profiling cardio/cytotoxicity cell model induced by doxorubicin. *Molecules* **25**, 836 (2020).
92. Yin, L. et al. High-content image-based single-cell phenotypic analysis for the testicular toxicity prediction induced by bisphenol A and its analogs bisphenol S, bisphenol AF, and tetrabromobisphenol A in a three-dimensional testicular cell co-culture model. *Toxicol. Sci.* **173**, 313–335 (2020).
93. Godinez, W. J. et al. Morphological deconvolution of beta-lactam polyspecificity in *E. coli*. *ACS Chem. Biol.* **14**, 1217–1226 (2019).
94. Tanaka, M. et al. An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol.* **3**, e128 (2005).
95. Lin, S. et al. Diversity focused semisyntheses of tetronate polyether ionophores. Preprint at *ChemRxiv* <https://doi.org/10.26434/chemrxiv.8299715> (2019).
96. Mayer, T. U. et al. Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* **286**, 971–974 (1999).
97. Mechanism matters. *Nat. Med.* **16**, 347 (2010).
98. MacDonald, M. L. et al. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.* **2**, 329–337 (2006).
99. Schenone, M., Dančik, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**, 232–240 (2013).
100. Kubota, K., Funabashi, M. & Ogura, Y. Target deconvolution from phenotype-based drug discovery by using chemical proteomics approaches. *Biochim. Biophys. Acta Proteins Proteom.* **1867**, 22–27 (2019).
101. Woehrmann, M. H. et al. Large-scale cytological profiling for functional analysis of bioactive compounds. *Mol. Biosyst.* **9**, 2604–2617 (2013).
102. Slack, M. D., Martinez, E. D., Wu, L. F. & Altschuler, S. J. Characterizing heterogeneous cellular responses to perturbations. *Proc. Natl. Acad. Sci. USA* **105**, 19306–19311 (2008).
103. Ljosa, V. et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
104. Gustafsdottir, S. M. et al. Multiplex cytological profiling assay to measure diverse cellular States. *PLoS ONE* **8**, e80999 (2013).
105. Gebre, A. A. et al. Profiling of the effects of antifungal agents on yeast cells based on morphometric analysis. *FEMS Yeast Res.* **15**, fov040 (2015).
106. Futamura, Y. et al. Morphbase, an encyclopedic cell morphology database, and its use for drug target identification. *Chem. Biol.* **19**, 1620–1630 (2012).
107. Reisen, F. et al. Linking phenotypes and modes of action through high-content screen fingerprints. *ASSAY. Drug Dev. Technol.* **13**, 415–427 (2015).
108. Lariaia, L. et al. Image-based morphological profiling identifies a lysosomotropic, iron-sequestering autophagy inhibitor. *Angew. Chem. Int. Ed.* <https://doi.org/10.1002/ange.201913712> (2020).
109. Hansen, B. K. et al. Structure and function of the bacterial protein toxin phenomycin. *Structure* **28**, 528–539.e9 (2020).

110. Christoforow, A. et al. Design, synthesis, and phenotypic profiling of pyrano-furo-pyridone pseudo natural products. *Angew. Chem. Int. Ed.* **58**, 14715–14723 (2019).
111. Peters, C. E. et al. Rapid inhibition profiling identifies a keystone target in the nucleotide biosynthesis pathway. *ACS Chem. Biol.* **13**, 3251–3258 (2018).
112. Schulze, C. J. et al. 'Function-first' lead discovery: mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20**, 285–295 (2013).
113. Ochoa, J. L., Bray, W. M., Lokey, R. S. & Linington, R. G. Phenotype-guided natural products discovery using cytological profiling. *J. Nat. Prod.* **78**, 2242–2248 (2015).
114. Kremb, S., Müller, C., Schmitt-Kopplin, P. & Voelstra, C. R. Bioactive potential of marine macroalgae from the central red sea (Saudi Arabia) assessed by high-throughput imaging-based phenotypic profiling. *Mar. Drugs* **15**, 80 (2017).
115. Boyd, J. C., Pinheiro, A., Nery, E. D., Reygal, F. & Walter, T. Domain-invariant features for mechanism of action prediction in a multi-cell-line drug screen. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btz774> (2019).
116. Warchal, S. J. et al. High content phenotypic screening identifies serotonin receptor modulators with selective activity upon breast cancer cell cycle and cytokine signaling pathways. *Bioorganic Medicinal Chem.* **28**, 115209 (2020).
117. Hight, S. K. et al. High-throughput functional annotation of natural products by integrated activity profiling. *Pharmacol. Toxicol.* <https://doi.org/10.1101/748129> (2019).
118. Kurita, K. L., Glassey, E. & Linington, R. G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl Acad. Sci. USA* **112**, 11999–12004 (2015).
119. Nassiri, I. & McCall, M. N. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic Acids Res.* **46**, e116 (2018).
120. Breinig, M., Klein, F. A., Huber, W. & Boutros, M. A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells. *Mol. Syst. Biol.* **11**, 846 (2015). **This is a landmark study on the interaction of genes and compounds by use of isogenic cell lines, all mapped using image-based phenotypes.**
121. Ohnuki, S., Oka, S., Nogami, S. & Ohya, Y. High-content, image-based screening for drug targets in yeast. *PLoS ONE* **5**, e10177 (2010). **This is a pioneering study in yeast demonstrating successful identification of small-molecule mimics of gene signatures (and vice versa for target identification) using image-based data.**
122. Piotrowski, J. S. et al. Plant-derived antifungal agent poaic acid targets  $\beta$ -1,3-glucan. *Proc. Natl Acad. Sci. USA* **112**, E1490–E1497 (2015).
123. Iwaki, A., Ohnuki, S., Suga, Y., Izawa, S. & Ohya, Y. Vanillin inhibits translation and induces messenger ribonucleoprotein (mRNP) granule formation in *Saccharomyces cerevisiae*: application and validation of high-content, image-based profiling. *PLoS ONE* **8**, e61748 (2013).
124. Sundaramurthy, V. et al. Integration of chemical and RNAi multiparametric profiles identifies triggers of intracellular mycobacterial killing. *Cell Host Microbe* **13**, 129–142 (2013).
125. Sundaramurthy, V. et al. Deducing the mechanism of action of compounds identified in phenotypic screens by integrating their multiparametric profiles with a reference genetic screen. *Nat. Protoc.* **9**, 474–490 (2014).
126. Eggert, U. S. et al. Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.* **2**, e379 (2004).
127. Singh, S. et al. Morphological profiles of RNAi-induced gene knockdown are highly reproducible but dominated by seed effects. *PLoS ONE* **10**, e0131370 (2015).
128. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
129. Jackson, A. L. et al. Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**, 635–637 (2003).
130. Young, D. W. et al. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **4**, 59–68 (2008).
131. Caldera, M. et al. Mapping the perturbome network of cellular perturbations. *Nat. Commun.* **10**, 5140 (2019).
132. Fischer, B. et al. A map of directional genetic interactions in a metazoan cell. *eLife* **4**, e05464 (2015).
133. Yin, Z. et al. A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. *Nat. Cell Biol.* **15**, 860–871 (2013).
134. Usaj, M. M. et al. Systematic genetics and single-cell imaging reveal widespread morphological pleiotropy and cell-to-cell variability. *Mol. Syst. Biol.* **16**, e9243 (2020).
135. Chong, Y. T. et al. Yeast proteome dynamics from single cell imaging and automated analysis. *Cell* **161**, 1413–1424 (2015).
136. Neumann, B. et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
137. Unadkat, H. V. et al. An algorithm-based topographical biomaterials library to instruct cell fate. *Proc. Natl Acad. Sci. USA* **108**, 16565–16570 (2011).
138. Ruan, X. et al. Image-derived models of cell organization changes during differentiation of PC12 cells. Preprint at *bioRxiv* <https://doi.org/10.1101/522763> (2019).
139. Doan, M. & Carpenter, A. E. Leveraging machine vision in cell-based diagnostics to do more with less. *Nat. Mater.* **18**, 414–418 (2019).
140. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
141. Zhu, Y. et al. An image informatics pipeline for imaging mass cytometry to characterize the immune landscape in pre- and on-treatment immune therapy and its application in recurrent platinum-resistant epithelial ovarian cancer. in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)* 1–4 (IEEE, 2019).
142. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
143. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
144. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
145. Ash, J. T., Darnell, G., Munro, D. & Engelhardt, B. E. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. Preprint at *bioRxiv* <https://doi.org/10.1101/458711> (2018).
146. Quach, D. T., Sakoulas, G., Nizet, V., Pogliano, J. & Pogliano, K. Bacterial cytological profiling (BCP) as a rapid and accurate antimicrobial susceptibility testing method for *Staphylococcus aureus*. *EBioMedicine* **4**, 95–103 (2016).
147. Dekkers, J. F. et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* **19**, 939–945 (2013).
148. Snijder, B. et al. Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *Lancet Haematol.* **4**, e595–e606 (2017).
149. Irmisch, A. et al. The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support. *Oncology* <https://doi.org/10.1101/2020.02.13.20017921> (2020).
150. Betge, J. et al. Multiparametric phenotyping of compound effects on patient derived organoids. Preprint at *bioRxiv* <https://doi.org/10.1101/660993> (2019).
151. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
152. Kraus, O. Z. & Frey, B. J. Computer vision for high content screening. *Crit. Rev. Biochem. Mol. Biol.* **51**, 102–109 (2016).
153. Moen, E. et al. Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
154. Chessel, A. & Carazo Salas, R. E. From observing to predicting single-cell structure and function with high-throughput/high-content microscopy. *Essays Biochem.* **63**, 197–208 (2019).
155. Wollmann, T. et al. GRUU-Net: integrated convolutional and gated recurrent neural network for cell segmentation. *Med. Image Anal.* **56**, 68–79 (2019).
156. Caicedo, J. C. et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
157. Jackson, P. T. et al. Phenotypic profiling of high throughput imaging screens with generic deep convolutional features. in *2019 16th International Conference on Machine Vision Applications (MVA)* 1–4 (IEEE, 2019).
158. McQuin, C. et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
159. Eulenberg, P. et al. Reconstructing cell cycle and disease progression using deep learning. *Nat. Commun.* **8**, 463 (2017).
160. Kensert, A., Harrison, P. J. & Spjuth, O. Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS Discov.* **24**, 466–475 (2019).
161. Rohban, M. H., Abbasi, H. S., Singh, S. & Carpenter, A. E. Capturing single-cell heterogeneity via data fusion improves image-based profiling. *Nat. Commun.* **10**, 2082 (2019).
162. Lim, N. & Pavlidis, P. Evaluation of connectivity map shows limited reproducibility in drug repositioning. Preprint at *bioRxiv* <https://doi.org/10.1101/845693> (2019).
163. Ellenberg, J. et al. A call for public archives for biological image data. *Nat. Methods* **15**, 849–854 (2018).
164. Swinney, D. C. The contribution of mechanistic understanding to phenotypic screening for first-in-class medicines. *J. Biomol. Screen.* **18**, 1186–1192 (2013).
165. Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: origins and evolution. *Nat. Rev. Drug Discov.* **13**, 577–587 (2014).
166. Warchal, S. J., Unciti-Broceta, A. & Carragher, N. O. Next-generation phenotypic screening. *Future Med. Chem.* **8**, 1331–1347 (2016).
167. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **10**, 507–519 (2011).
168. Vincent, F., Loria, P., Pregel, M. & Stanton, R. Developing predictive assays: the phenotypic screening 'rule of 3'. *Sci. Transl. Med.* <https://doi.org/10.1126/scitranslmed.aab1201> (2015).
169. Lau, T. A., Bray, W. M. & Lokey, R. S. macrophage cytological profiling and anti-inflammatory drug discovery. *Assay. Drug Dev. Technol.* **17**, 14–16 (2019).
170. RxRx. <https://www.rxrx.ai/>.
171. Lin, J.-R. et al. Highly multiplexed immunofluorescence imaging of human tissues and tumors using T-CyCIF and conventional optical microscopes. *eLife* **7**, e31657 (2018).
172. Gerdes, M. J. et al. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl Acad. Sci. USA* **110**, 11982–11987 (2013).
173. Bolognesi, M. M. et al. Multiplex staining by sequential immunostaining and antibody removal on routine tissue sections. *J. Histochem. Cytochem.* **65**, 431–444 (2017).
174. Glass, G., Papin, J. A. & Mandell, J. W. SIMPLE: a sequential immunoperoxidase labeling and erasing method. *J. Histochem. Cytochem.* **57**, 899–905 (2009).
175. Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass cytometry imaging for the study of human diseases-applications and data analysis strategies. *Front. Immunol.* **10**, 2657 (2019).
176. Rappez, L. et al. Spatial single-cell profiling of intracellular metabolomes in situ. Preprint at *bioRxiv* <https://doi.org/10.1101/510222> (2019).
177. Kang, Z. B. et al. Fluopack screening platform for unbiased cellular phenotype profiling. *Sci. Rep.* **10**, 2097 (2020).
178. Perez-Gomez, A. et al. A phenotypic caenorhabditis elegans screen identifies a selective suppressor of antipsychotic-induced hyperphagia. *Nat. Commun.* **9**, 5272 (2018).
179. Cassar, S. et al. Measuring drug absorption improves interpretation of behavioral responses in a larval zebrafish locomotor assay for predicting seizure liability. *J. Pharmacol. Toxicol. Methods* **88**, 56–63 (2017).
180. Becker, T., Caicedo, J. C., Singh, S., Weckmann, M. & Carpenter, A. E. Combining morphological and migration profiles of in vitro time-lapse data. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 965–968 (IEEE, 2018).
181. Qian, W. W. et al. Batch equalization with a generative adversarial network. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.02.07.939215> (2020).
182. Rastelli, G. & Pinzi, L. Computational polypharmacology comes of age. *Front. Pharmacol.* **6**, 157 (2015).

183. Proschak, E., Stark, H. & Merk, D. Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. *J. Med. Chem.* **62**, 420–444 (2019).

#### Acknowledgements

The authors appreciate helpful comments from S. Jaensch, S. Singh, J. Caicedo, N. Rindtorff and all members of the Carpenter laboratory. The authors acknowledge funding

support for S.N.C. and A.E.C. from the US National Institutes of Health (R35 GM122547 to A.E.C.).

#### Author contributions

The authors contributed equally to all aspects of the article.

#### Competing interests

H.C. and J.D.B. are employed by Janssen and Pfizer, respectively. A.E.C. is on the Scientific and Technical Advisory

Board of, has optional ownership interest in and receives income from Recursion. S.N.C. declares no conflict of interest.

#### Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2020