Title:
Morphology and gene expression profiling provide complementary information for mapping cell state

Authors:
Gregory P. Way[1,2,*], Ted Natoli[3,*], Adeniyi Adeboye[1,*], Lev Litichevskiy[3], Andrew Yang[3], Xiaodong Lu[3], Juan C. Caicedo[1], Beth A. Cimini[1], Kyle Karhohs[1], David J. Logan[1,+], Mohammad Rohban[1], Maria Kost-Alimova[4], Kate Hartland[4], Michael Bornholdt[1], Niranj Chandrasekaran[1], Marzieh Haghighi[1], Shantanu Singh[1,^], Aravind Subramanian[3,^,#], Anne E. Carpenter[1,^,#]

*co-first author
^co-senior author
[#]co-corresponding author

Affiliations:
1. Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
2. Center for Health AI and Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado, USA
3. Cancer Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
4. Center for the Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

Current Affiliations:
+ Pfizer Worldwide Research, Development and Medical, Internal Medicine Research Unit, Cambridge, Massachusetts, 02139, USA

Abstract:
Deep profiling of cell states can provide a broad picture of biological changes that occur in disease, mutation, or in response to drug or chemical treatments. Morphological and gene expression profiling, for example, can cost-effectively capture thousands of features in thousands of samples across perturbations, but it is unclear to what extent the two modalities capture overlapping versus complementary mechanistic information. Here, using both the L1000 and Cell Painting assays to profile gene expression and cell morphology, respectively, we perturb A549 lung cancer cells with 1,327 small molecules from the Drug Repurposing Hub across six doses. We determine that the two assays capture some shared and some complementary information in mapping cell state. We find that as compared to L1000, Cell Painting captures a higher proportion of reproducible compounds and has more diverse samples, but measures fewer distinct groups of features. In an unsupervised analysis, Cell Painting grouped more compound mechanisms of action (MOA) whereas in a supervised deep learning analysis, L1000 predicted more MOAs. In general, the two assays together provide a complementary view of drug mechanisms for follow up analyses. Our analyses answer fundamental biological questions comparing the two biological modalities and, given the numerous applications of profiling in biology, provide guidance for planning experiments that

profile cells for detecting distinct cell types, disease phenotypes, and response to chemical or genetic perturbations.

Introduction:

In a profiling experiment, biologists measure high-dimensional readouts from biological samples (e.g. single cells, organoids, tissue, whole organisms). The resulting profile contains measurements of hundreds to thousands of individual features that together form a systems biology representation of the sample of interest. Automation now allows biologists to probe thousands of chemical and genetic perturbations to assess their phenotypic impact [1–3]. Therefore, perturbational profiling results in a large number of samples measured across a common set of high-dimensional features. Biologists can then apply data mining and machine learning to these datasets to detect and quantify the similarities and differences among samples. These approaches have the potential to greatly advance drug discovery, functional genomics, and precision medicine, for example, by annotating uncharacterized small molecules, cataloging the mechanistic outcome of gene editing, and testing the impact of specific perturbations on disease-associated phenotypes [4–6].

Biologists can access different aspects of cell state through multiple profiling assays that capture different biological landscapes: DNA, RNA, epigenetic marks, metabolites, microbiota, proteins, kinases, and spatial information [7–13]. Some profiling approaches measure multiple data modalities in the same assay and are dubbed multi-omic or multi-modal assays [14,15]; others pool and de-multiplex perturbations to increase throughput [16]; still others test a single readout (e.g. viability) but across hundreds of cell types to yield a profile [17,18].

Gene expression and cell morphology are currently the two highest-throughput, lowest cost, high-dimensional profiling data types [3,19,20]. These readouts measure fundamentally different aspects of biology. In the L1000 bead-based assay, probes targeting 978 genes measure mRNA transcript levels (gene expression) in a cell population [3]. In the Cell Painting assay, after treating cells with six fluorescent dyes to mark eight cellular compartments, biologists use a microscope to image five channels and use software to analyze and extract several thousand morphology measurements from each cell [19]. Both the gene expression and morphology landscapes change as cells respond to perturbations.

Scientists have used individual profiling modalities to advance a variety of drug discovery applications, including improving screening library diversity, predicting cytotoxicity, prioritizing compounds for follow-up study, and inferring the mechanism of action of chemicals [21–28]. Integrating gene expression and morphology profiles with chemical structures revealed that each data type provides complementary information for predicting a drug's mechanism of action [29,30], for predicting the effects of perturbations [31], and for identifying nuisance compounds that can lead to false hits [32]. As well, to some degree, gene expression and morphology datasets contain sufficient information to predict changes in each other [29,30]. However, the field lacks a systematic study evaluating both assays' information content in terms of distinct versus overlapping signals, diversity of cell states, and performance in valuable tasks such as predicting compound mechanisms.

In this study, we collected L1000 and Cell Painting readouts from a common set of 1,327 different Drug Repurposing Hub compounds and controls across six doses representing 511 different mechanism-of-action (MOA) classes [33]. After data processing, we observed that while Cell Painting suffers from more batch and well position effects that must be carefully adjusted, the assay showed higher profile reproducibility than L1000. While L1000 includes more independent feature groups than Cell Painting, the latter provides a higher sample diversity. We test the practical implications of these properties by training top-performing deep learning models from a recent Kaggle competition to predict compound MOA [34]. While both assays predict a set of mechanisms consistently well, certain mechanisms are better captured in one assay or the other. In summary, we find that Cell Painting and L1000 each reproducibly measure a partly overlapping, partly distinct set of chemical compounds and MOAs. Based on this analysis we conclude that measuring both molecular and cellular phenotypes increases the ability to capture relevant biological mechanisms from unbiased compound screens.

Results

Measuring and processing morphology and gene expression data

One strategy to interrogate biological processes is to measure cell responses to various perturbations in high-throughput, high-dimensional profiling assays. Profiling assays vary in style and measurement, and the sensitivity and resolution with which they capture important biological signals depends on the assay chosen. In this experiment, we asked whether measuring the same perturbations using fundamentally different kinds of profiling assays provides advantages. We therefore created and analyzed two profiling datasets that capture different types of information: gene expression with the L1000 assay and morphology with the Cell Painting assay [3,19,20].

We perturbed A549 lung cancer cells with a common set of 1,327 compounds and controls from the Drug Repurposing Hub [33]. We chose the compounds to represent a diversity of mechanisms based on Drug Repurposing hub annotations and prior knowledge [33]. We measured a total of 1,258 compounds across six doses; the remainder had three to five doses.

We perturbed the cells under consistent experimental conditions, including the same 384-well plate layouts (**Supplementary Figure 1A**). The cells were exposed to compounds for 24 hours prior to L1000 profiling and for 48 hours prior to Cell Painting; we used standard assay time points based on past experience. We assayed data using 25 different plate maps (compound layouts), and, in most cases, we collected three replicate plates per plate map for L1000 and five replicate plates per plate map for Cell Painting, given its lower cost per plate. Each replicate plate contained 56 different compounds in six doses plus 24 Dimethyl sulfoxide (DMSO) negative controls and 24 proteasome inhibitor positive controls.

We applied standard data processing pipelines for each assay (see Methods) to normalize and transform the data prior to downstream analyses (**Supplementary Figure 1B**). Due to the limitations of the compound dispensing equipment, it was unfortunately infeasible to control for plate layout artifacts by scrambling perturbation locations within each plate across replicates.

Indeed, we observed plate position effects in the Cell Painting data, particularly in edge wells (**Supplementary Figure 2**). Therefore, we used a spherize transform to combine data across batches and adjust for these plate-position effects [35,36]. In all the downstream analyses, we use spherized Cell Painting profiles and the original, unspherized L1000 profiles unless indicated otherwise (L1000 did not benefit from spherizing, see **Supplementary Figure 2, bottom**).

Assessing profile reproducibility in L1000 and Cell Painting assays

To study a perturbation's function, a scientist must reliably and robustly measure its biological effect. Therefore, we introduced and calculated a reproducibility metric, which we call "percent replicating". Specifically, percent replicating captures the percentage of profile replicates (treatment of the same compound measured at the same dose) that are more similar to one another than to a randomly-permuted null distribution that adjusts for dose, sample size, and well position (see Methods for complete details).

As expected, percent replicating increased with dose in both assays, as higher concentrations of drug are more likely to impact cell systems (**Figure 1A**). However, we observed much higher percent replicating scores in Cell Painting (57% to 83%, from lowest to highest dose) compared to L1000 (16% to 35%, from lowest to highest dose) (**Figure 1A**). One possible explanation for higher Cell Painting reproducibility metrics is that plate layout effects artificially increased replicate correlations preferentially for that modality versus L1000. Indeed, we observed that percent replicating increased if we failed to adjust our null distribution for well position, and decreased if we failed to correct for plate-position effects by sphering profiles (**Supplementary Figure 3**). These results underscore the importance of proper construction of null distributions and proper profile normalization in Cell Painting. However, the same normalization did not improve scores for L1000. We observed similar results if we relaxed the null distribution constraints to only adjust for dose but not sample size nor well position (named "percent strong", see methods) (**Supplementary Figure 4**).

Plate layout effects are a serious concern in profiling experiments where practical reasons require scientists to measure replicates of a sample at the same well position across physical plates; it is known that the location on the plate, especially distance from the edge of the plate, can impact many cell phenotypes. Therefore, to more closely study the impact of plate position on pairwise correlations, we performed a non-replicate diffusion analysis in which we systematically increased the well neighborhood size in calculating the null distribution of non-replicate correlations (see Methods). Briefly, we started with a diffusion size 0, which looks at the non-replicate correlations of different samples that are in the exact same well position, across different plate maps. As we increased diffusion (the well neighborhood size) to include adjacent and nearby wells, we observed a slight dampening of non-replicate correlations (**Supplementary Figure 5A**). While this analysis revealed increased plate-position effects in Cell Painting compared to L1000, this bias is relatively small compared to the overall replicate correlation signal (**Supplementary Figure 5B**). Taken together, plate layout effects are a critical component of profiling assays but are unlikely to have driven the strong performance we observed in this experiment. Nevertheless, when possible, we recommend scrambling replicates across wells to avoid this potential confounding effect.

We also suspected that Cell Painting might have a performance advantage over L1000 because we had, typically, five replicates of Cell Painting and only three replicates of L1000 per treatment. Indeed, a subsampling experiment that limited the number of Cell Painting replicates to match the number of L1000 replicates reduced percent replicating (from 57% to 37% and 83% to 67%, from lowest to highest dose), although still not to the level of L1000 (**Supplementary Figure 3**).

Comparing median pairwise replicate correlations across individual treatments, we observed that many compounds have higher correlations in Cell Painting compared to L1000, but many compounds are highly correlated in both assays (**Figure 1B**). However, we observed that certain compounds were only reproducible in one assay or the other (**Figure 1C**). We observed that 11% of compounds in the lowest dose (133 / 1,258) and 34% of compounds in the highest dose (422 / 1,258) produced a replicating profile in both assays. Considering both assays together, we found that 62% to 85% of compounds (from lowest to highest dose) produced a replicating profile (**Figure 1C**).

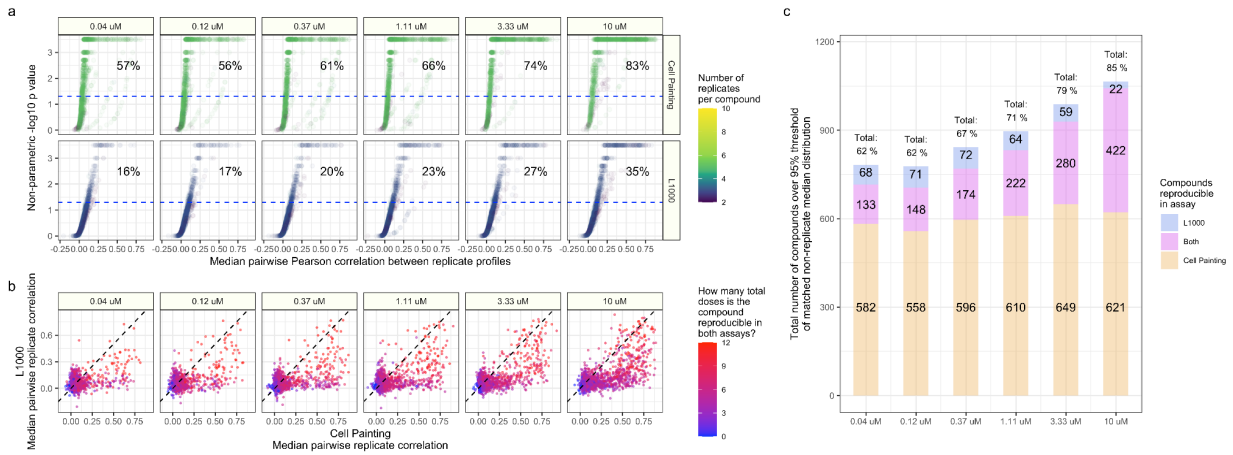

Figure 1. *Cell Painting and L1000 data provide complementary measurements of compound perturbations across doses.* **(a)** The percent replicating metric measures the percentage of profiles that correlate with each other to a level higher than a carefully-matched and randomly-sampled null distribution. See methods for full details about sampling and data processing. The dotted blue line indicates the 95th percentile of the matched non-replicate distribution. **(b)** Median pairwise replicate correlations between profiles measured by the L1000 assay (y axis) and Cell Painting assay (x axis). The dotted black line is the line y = x, so anything above is measured with a higher replicate correlation in L1000 and vice versa. **(c)** The L1000 and Cell Painting assays reproducibly measure a complementary set of compound perturbations. The three numbers represent (from top to bottom) the number of reproducible compounds unique to L1000, the number of compounds reproducible in both assays, and the number of reproducible compounds unique to Cell Painting.

Assessing the ability of Cell Painting and L1000 to capture compound mechanism of action

We next tested a more demanding, application-oriented metric based on a common use case when profiling compounds: determining a compound's MOA. We introduced "percent matching" as a metric to quantify how often a profiling assay can group together compound profiles that have the same annotated MOA (see Methods). Unlike percent replicating, this metric is not

influenced by plate layout effects, as compounds with the same annotated MOA are not located in the same well location across plate maps.

We observed higher percent matching scores for Cell Painting (ranging from 16 - 28%, depending on dose) than for L1000 (7 - 21%) (**Figure 2A**). Comparing percent matching scores between assays, we observed many overlapping, but also many assay-specific MOAs (**Figure 2B**). In general, we observed stronger signals in L1000 from a smaller number of MOAs, compared to weaker signals from a larger number of MOAs for Cell Painting, as indicated by more points above the dotted line for L1000 but higher percent matching scores for Cell Painting (**Figure 2B and 2C**).

The overlapping MOAs, captured by both L1000 and Cell Painting and including at least three different compounds, ranged from 3% of MOAs at the lowest dose (4 / 127) to 18% of MOAs at the highest dose (23 / 127). Moreover, when considering both assays together, they collectively captured 20% of MOAs at the lowest dose (25 / 127) and 31% of MOAs at the highest dose (39 / 127) (**Figure 2C**). At the highest dose, one captures 10% more MOAs by adding Cell Painting to L1000, and 3% more MOAs by adding L1000 to Cell Painting. These observations can guide researchers in selecting a particular profiling modality that provides more consistent measurements when studying specific compounds or MOAs (**Figure 2D**). We provide a full list of compounds and MOAs and their respective median pairwise correlations as **Supplementary Table S1**.

Certain MOA categories yielded differing patterns across L1000 and Cell Painting. For example, MEK inhibitors yielded strong consistency in both assays, particularly in L1000, while PLK inhibitors were also strong in both, but particularly strong in Cell Painting. Conversely, estrogen receptor agonists yielded uncorrelated profiles in both assays (**Supplementary Figure 6**). We note that in any MOA analysis, low matching scores may result from noise or technical limitations of the assays, but they may also reflect real biological signals resulting from either inaccurate annotations, which is a known challenge [37], or alternatively because the assay is capturing subtle mechanistic differences between compounds that are annotated with the same MOA.

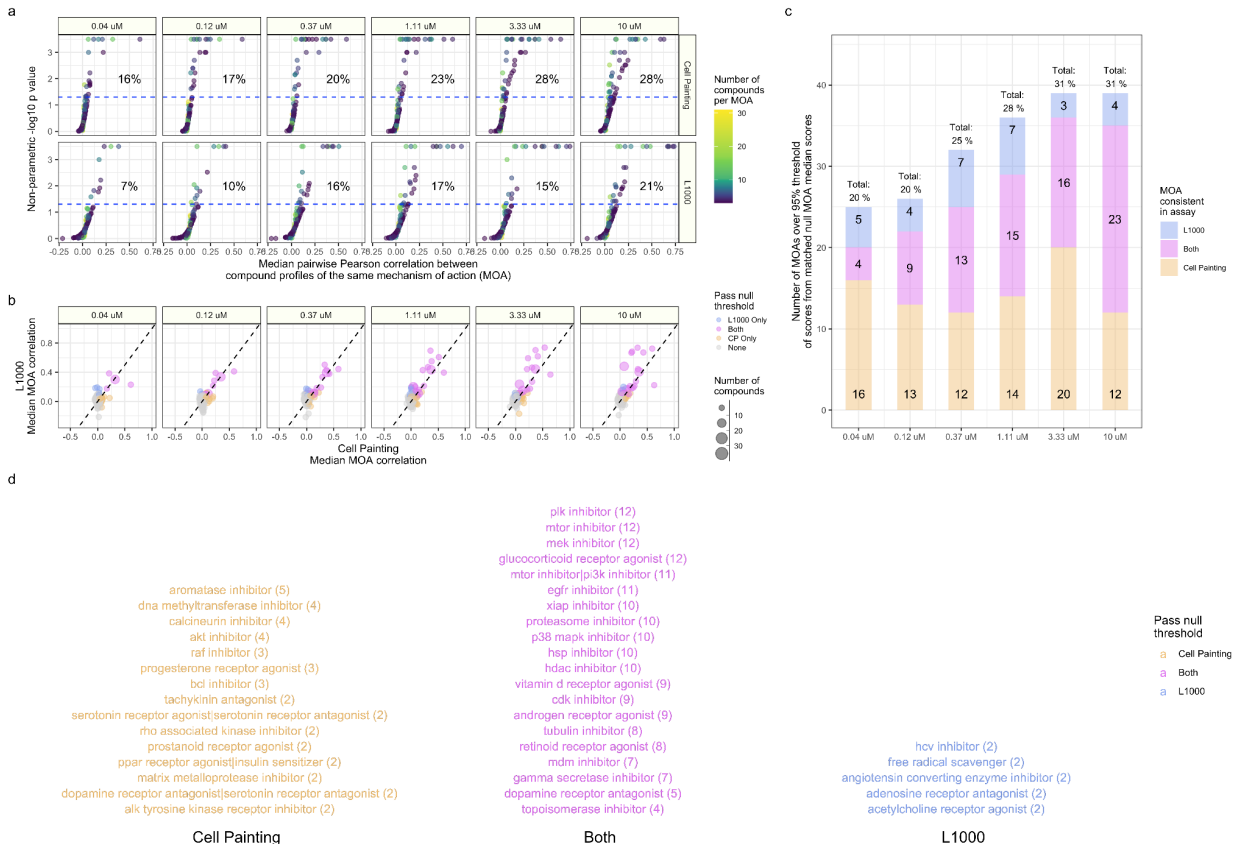

Figure 2. *Cell Painting and L1000 differentially measure compound perturbations by mechanism of action (MOA)*. **(a)** Median pairwise replicate correlations of groups of compounds with a given MOA annotation, measured in both assays and across doses. The color of the point represents how many compounds were annotated to a given MOA class. **(b)** Median correlation between compounds annotated with the same MOA. We derived the null threshold through a nonparametric permutation test of randomly sampled compounds (see Methods). The size of the points represent how many compounds belong to the MOA class. **(c)** The L1000 and Cell Painting assays reproducibly measure a complementary set of MOAs. The three numbers represent (from top to bottom) the number of reproducible MOAs unique to L1000, the number of MOAs reproducible in both assays, and the number of reproducible MOAs unique to Cell Painting. **(d)** All 40 MOAs that are consistently measured by Cell Painting, L1000, or both in at least two doses. Brackets indicate the number of times the MOA was reproducible (per dose and assay). For example, PLK inhibitors were reproducible in every dose and in each assay (6 doses x 2 assays).

Analyzing the diversity of perturbed cell states manifesting in gene expression and morphology
While percent replicating captures the proportion of compounds that significantly change cell state, it does not quantify the diversity of those cell states when considering different compounds. Diversity of cell states is critical for many applications, such as MOA matching as described above, because more diversity indicates more potential for interesting biological findings. For example, quantifying cell state diversity is critical when selecting compounds for inclusion in a screening library, as the goal is typically to maximize phenotypic diversity among the compounds; strategies that reduce redundancy allow inclusion of more diverse phenotypes and are therefore more likely to result in drug discovery pipeline “hits” [21].

To qualitatively assess the diversity of profiles produced by each profiling assay, we applied a unified manifold approximation (UMAP) transform [38]. We observed that, in both assays, many

compounds form distinct islands that consistently group specific MOAs, while a sizable set of compounds are relatively similar to negative controls (**Figure 3A**). The islands separated more with increasing dose, and we identified similarly distributed clusters using t-distributed stochastic neighbor embedding (t-SNE) [39] (**Supplementary Figure 7)**. The primary data collection for this project used a single cell line, A549; a small dataset we gathered using three cell lines (A549, MCF7, and U2OS) showed more phenotypic separation according to cell line and incubation period (**Supplementary Figure 8**). This separation demonstrated higher biological diversity induced by inherent cell line differences as compared to diversity induced by different perturbations, which is consistent with observations for L1000 data[40].

For a quantitative analysis, we fit different clustering algorithms to approximate the number of unique groups of compounds that manifest in each data modality. We observed more distinct clusters in Cell Painting compared to L1000 readouts. This observation was consistent across different clustering solutions (from k=2 to k=40), with different clustering algorithms (k-means clustering and Gaussian mixture models) and using three different metrics (Silhouette scores, Davies Bouldin scores, and Bayesian information criterion) (**Supplementary Figure 9**). Observing global patterns of pairwise sample correlations in a heatmap provides further evidence of increased diversity in Cell Painting measurements as indicated by lower pairwise correlations across different compounds (**Figure 3B**). Taken together, this analysis suggests that morphology profiles measured by Cell Painting capture more diverse cell states than the gene expression profiles measured by L1000, under the experimental conditions tested.

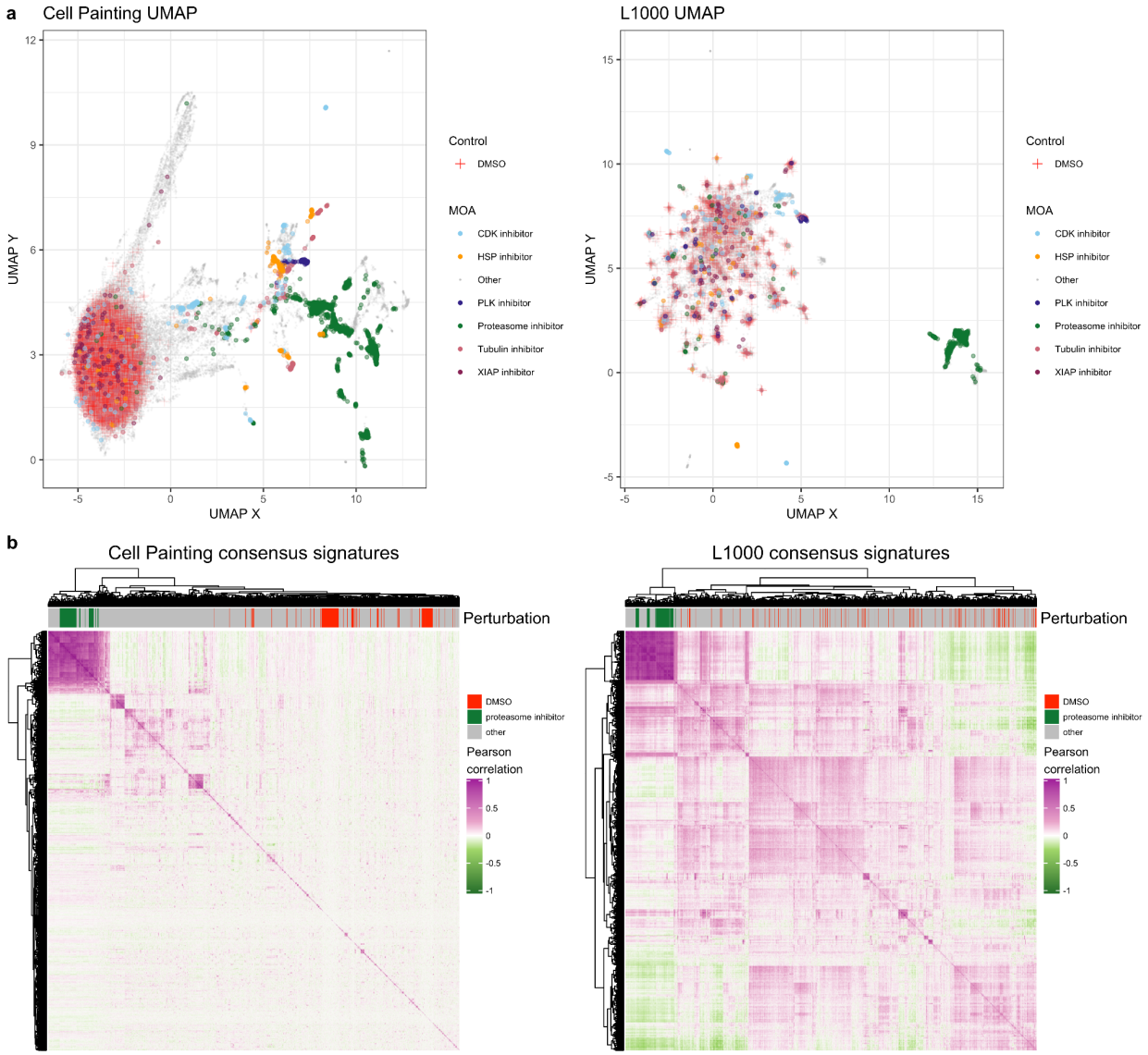

Figure 3. *Cell Painting captures a more diverse sample space than L1000.* **(a)** Uniform manifold approximation (UMAP) coordinates of all perturbations (level 4 replicates) across all doses by Cell Painting (left) and L1000 profiles (right). We highlight select MOAs that are consistently different from DMSO controls in either modality. Note that Cell Painting data is spherized and L1000 data is not, as explained in the main text; here this manifests in quite different patterns for the negative control DMSO samples. In particular, many of the otherwise-distinct islands of compounds for L1000 are populated by negative control DMSO. **(b)** Heatmaps of pairwise Pearson correlations of all measured compounds' consensus signatures (see Methods) in either assay at the highest dose (10 μM) plus positive-control proteasome inhibitors at 20 μM and DMSO negative controls.

Assessing the complementarity of profiling morphology and gene expression features

By design, different profiling technologies measure different biological features. L1000 is a gene expression assay and therefore measures molecular features; specifically, mRNA transcript levels in a biological sample. Cell Painting is an image-based assay that instead measures cell features; both morphological and spatial. Nevertheless, biological signals are often related or even tightly coupled. We therefore sought to approximate how many independent groups of

features exist in both modalities. This is distinct from our analysis of the number of groups of *samples* described in the prior section.

In general, we observed higher diversity of feature signals in L1000 compared to Cell Painting (**Figure 4A**). Much higher absolute value pairwise correlations among Cell Painting features, even after feature selection (see Methods) indicate more redundant measurements compared to L1000 (**Figure 4B)**. Indeed, the top Principal Components (PCs) explain a higher proportion of variance in Cell Painting compared to L1000 data, providing further evidence of increased feature redundancy in Cell Painting (**Figure 4C**). Both assays attempt to reduce redundancy in some way. For L1000, scientists deliberately chose L1000's features (978 distinct molecular entities) to minimize redundancy in measurements while maximizing the ability to infer transcriptome-wide gene expression [3]. Following the standard image-based profiling pipeline [41], we also removed highly correlated features from the Cell Painting assay. Taken together, this analysis suggests that there is a higher diversity of gene expression features than morphology features, as measured by these two assays.

Because we collected thousands of perturbations with replicates, we can study the specific features, in either assay, that were highly impacted by individual compound treatments. Calculating a metric called "activity score" [3], which combines both replicate reproducibility and number of impacted features (see Methods), we observed that certain compounds disrupt L1000 and Cell Painting features with different strengths in a dose-dependent manner (**Figure 4D**). Nearly all of the compound perturbations disrupted more morphology readouts than expression readouts. This observation was even more pronounced when we focused on compounds that were reproducible in both assays in at least three different doses (**Figure 4E**). In particular, Dasatinib, Alisertib, Brequinar, Aphidicolin, and STA-5326 consistently induced many morphological changes while changing relatively few expression values. By performing separate pathway analyses using the few genes most disrupted by each of the aforementioned compounds, we observed compound-specific associations with cholesterol biosynthetic process (GO:0006695), sulfur compound biosynthetic process (GO:0044272), cellular response to starvation (GO:0009267), and cellular response to UV (GO:0034644) (**Figure 4F**). Conversely, l-Ergothioneine induced many more transcriptional changes than morphological changes, which influenced genes associated with RNA splicing (GO:0008380) (**Figure 4F**). This type of analysis opens the door to exploring relationships between particular mRNA levels and specific morphologies when perturbing cells [29,30].

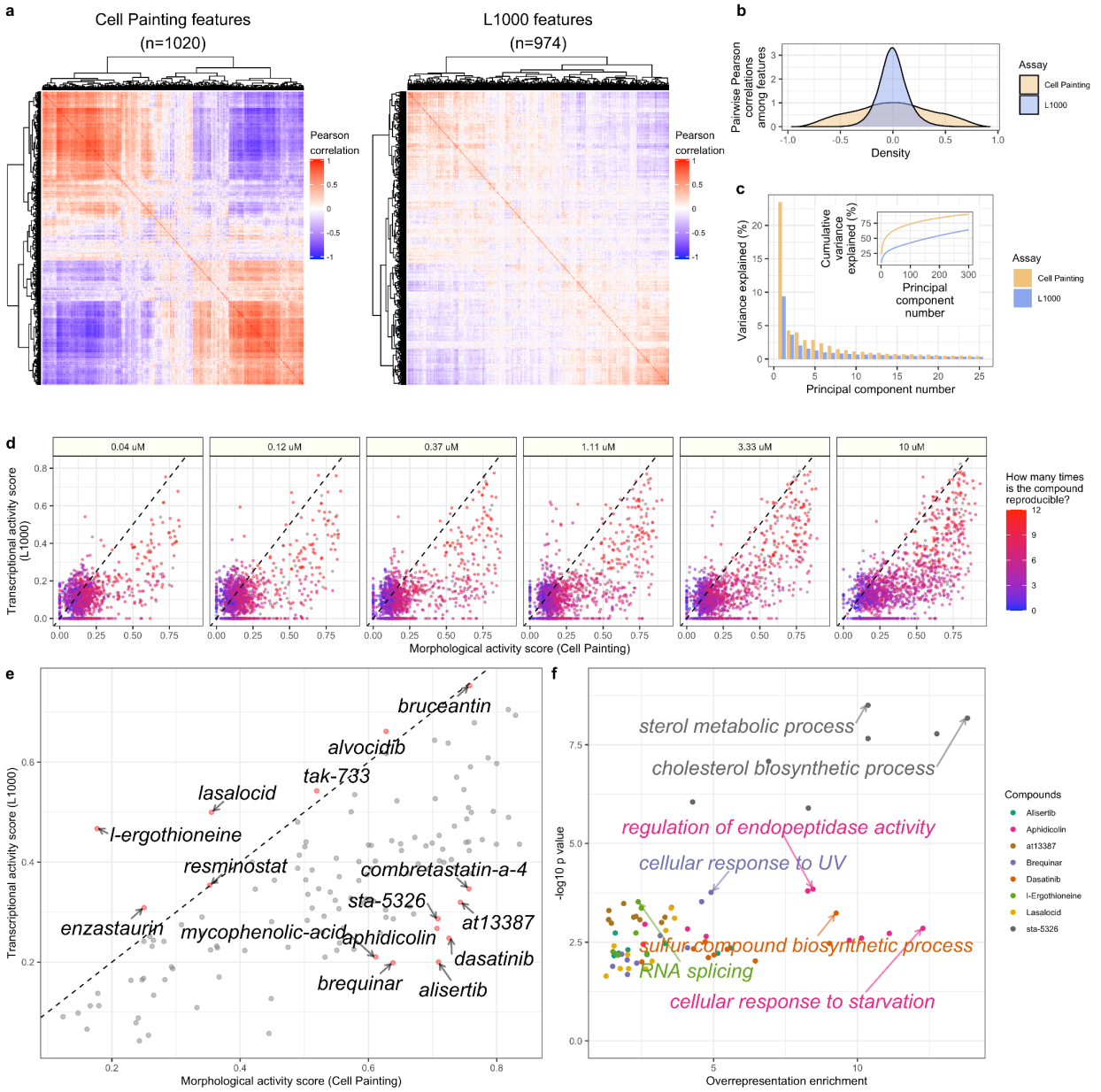

Figure 4. *Cell Painting morphology features are more redundant than L1000 gene expression features.* **(a)** Heatmaps of pairwise Pearson correlations of 1,020 Cell Painting features and 974 L1000 features, in each case derived from feature-selected consensus signatures of the same compound treatments at 10μM. **(b)** The same data plotted as a density plot shows the distribution of correlations between pairs of L1000 or Cell Painting features. **(c)** The percentage of variance explained for the top 30 principal components derived from a Principal Component Analysis (PCA) in Cell Painting or L1000 readouts. **(d)** Comparing activity scores for highly reproducible compound perturbations (as defined by having 3 or more doses passing the percent strong threshold) reveals that most compounds induce a higher number of morphological changes than gene expression changes. **(e)** The mean MAS and TAS for compounds that are reproducible in at least three doses, with labels for compounds with the largest difference between MAS and TAS. **(f)** Overrepresentation analysis (ORA) of the genes associated with each individual compound reveals certain gene ontology (GO) gene sets associated with compounds that induce strong morphological changes. Each point is a gene set, comprising L1000 landmark genes that were consistently modulated by that compound.

Complementary predictions of compound mechanism of action by different profiling modalities
A large range of perturbation experiments have mechanistic prediction as a central goal [42]. Therefore, we assessed the ability of Cell Painting and L1000 readouts to predict compound MOA. Unlike the unsupervised, correlation-based percent matching analysis described above, here we took a supervised approach.

Repurposing the top models from a related Kaggle competition to predict MOA from L1000 readouts and cell viability data [34], we observed that in our dataset, L1000 performed slightly better at MOA prediction than Cell Painting across a wide range of different deep learning architectures (**Figure 5A**). We observed that performance increased in step with treatment dose (**Supplementary Figure 10**). However, the Kaggle competition selected for models especially suited to L1000 and cell viability data and that it is possible that alternate architectures might favor Cell Painting data.

The ability of an MOA to be predicted was correlated between the two modalities (Spearman correlation = 0.7, $p < 2.2e-16$), but some MOAs were predicted better in one assay compared to the other (**Figure 5B**). For example, some MOAs, like proteasome inhibitors, XIAP inhibitors, and MEK inhibitors were predicted nearly perfectly using data from either assay. However, other MOAs, like HSP inhibitors and NFkB inhibitors were predicted more robustly using L1000 readouts while p38 MAPK inhibitors, PLK inhibitors, and dehydrogenase inhibitors were predicted more robustly using Cell Painting readouts (**Figure 5B**). Poor predictions might be a result of noisy readouts or the ability of the data type to reveal more subtle compound-specific signals such as off-target effects.

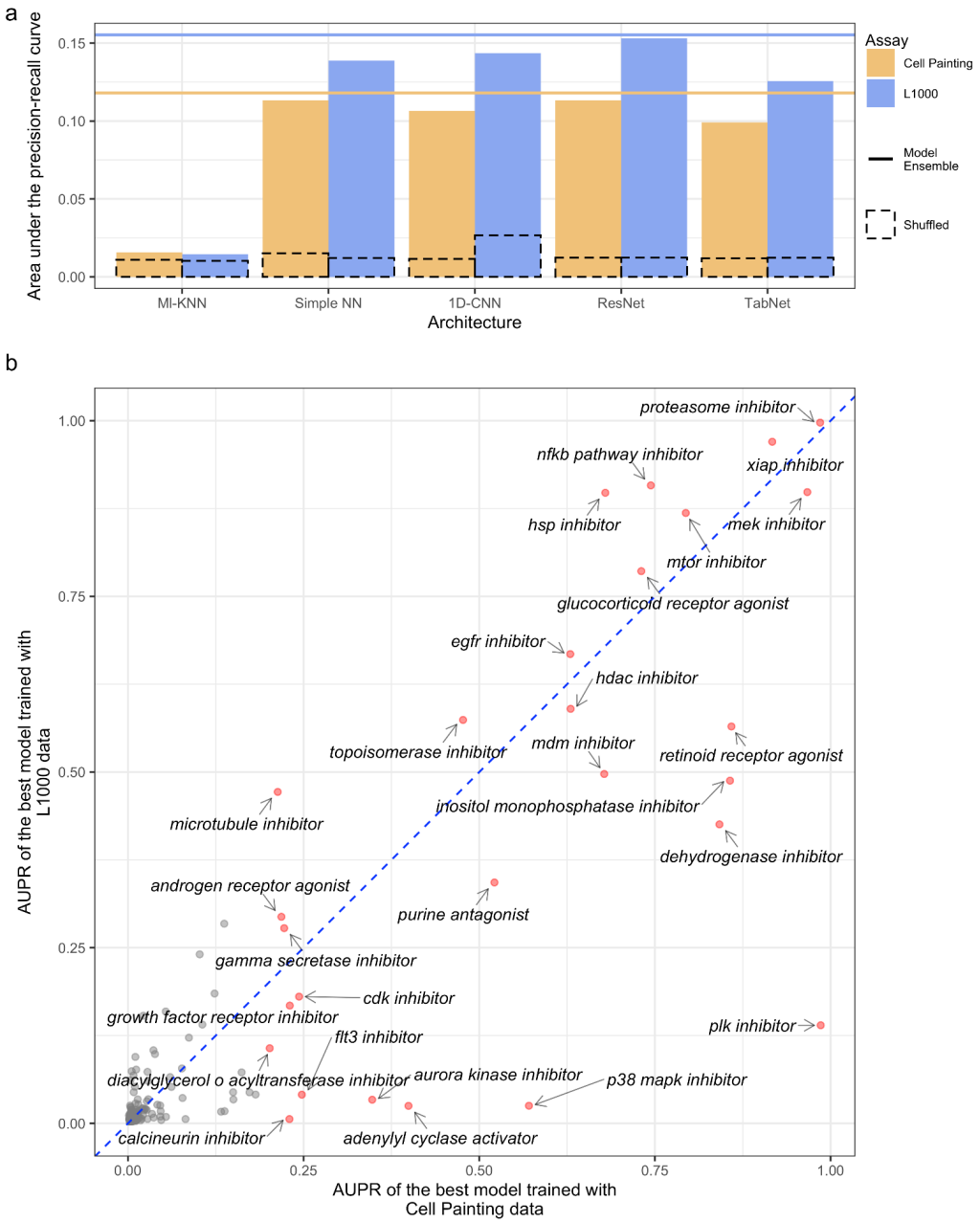

Figure 5. *Predicting compound mechanisms of action (MOA) in Cell Painting and L1000 reveals overlapping and complementary performance for different mechanisms.* **(a)** Test set model performance measured by area under the precision-recall curve (AUPR) for MOA prediction for our multi-label, multi-class prediction framework. We trained models from a recent Kaggle competition plus a K nearest neighbors baseline model. The dotted bar chart represents a negative control in which we trained models with shuffled labels. The solid lines indicate ensemble model performance by blending model predictions (see Methods). **(b)** Individual MOA performance by AUPR in the top performing model using Cell Painting and L1000 data. We report all performance metrics using test set profiles, which are held out as described in Methods.

Discussion

Large-scale perturbational profiling experiments are time- and cost-intensive; comparing their relative abilities is important information for experimental design and planning. We found that mRNA profiling (via L1000) and morphological profiling (via Cell Painting) were generally complementary. Cell Painting had a higher diversity of samples and matched MOAs more consistently in an unsupervised setting, while L1000 had a higher diversity of features and better performance predicting MOAs in a supervised setting. Cell Painting is less expensive, enabling larger experiments for a given budget, but L1000 offers a much larger pool of publicly available data to query [3]. A wide variety of biological pathways are readily captured by both data types, but some are better observed in one modality versus the other.

We had anticipated that morphological changes would generally not occur without concomitant changes in mRNA levels but found examples of compound treatments where this happens, and vice versa. It may be that mRNA changes do occur, but not at the short timescale of the mRNA detection (24 hours) as compared to the image capture (48 hours), or that the L1000 assay does not capture the mRNAs in the cell that change, due to technical noise or because it measures only ~5% of the transcripts in the cell. It is also possible that the different incubation times for compound treatments increased our ability to detect changes, but it is unlikely any single time point is optimal for all treatments, given potential differences in cell and molecular response times [43].

MOA prediction is a major challenge, and no assay exists that can reliably succeed across a majority of known MOA classes [42]; in light of past studies [28,44], we therefore did not expect to reliably predict most MOAs in our dataset but rather to compare the assays' performance in relative terms. Note that in our "percent matching" experiments, which quantifies MOA pairwise correlations in an unsupervised way, we observed lower correlation between modalities compared to the deep learning supervised approach. The deep learning models may exploit non-linear structures in the profile feature spaces that our percent-matching experiments might miss. However, the deep learning models cloud prediction interpretation, while also requiring accurate annotations for training.

One can imagine future developments in both assays that could improve performance. For perturbational profiling of mRNA expression, L1000 is the only assay at a cost reasonable for large-scale compound experiments; fortunately, expression profiles from L1000 are highly similar to RNA-seq equivalents: out of 3,176 patient-derived RNA samples profiled on both platforms, 3,103 (98%) had high quality cross-platform correlations [45]. For imaging assays, deep learning-based segmentation and feature extraction offers promise (roughly two-fold improvement in MOA prediction; Juan Caicedo, personal communication) but deep learning is not yet routine for image-based profiling [4,46]. Using additional stains is another sensible route, although initial testing indicates it does not seem to dramatically improve MOA prediction performance [47]. For both, screening additional cell types [44,47,48] and timepoints might increase the ability to detect and characterize perturbations in cell state. If experiments capture both profiling types, the profiles can be integrated and increase their power and resolution [28,29,49,50].

Overall, this paper will help people to better understand the pros and cons of the two currently largest and cheapest methods of large-scale drug profiling.

Methods

Sample preparation: Cell Painting

We generated Cell Painting data according to Bray et al, 2016 [19]. Briefly, we cultured cells on 384 well plates and exposed them to compound treatment at various doses for 24, 48, or 72 hours. After exposure, we fixed, stained, and then imaged all cells. We performed all imaging using a Phenix Opera with a 20X/1.0NA water objective, 1x1 binning, and filter sets described in Bray et al 2016 Supplementary Note 1.

Sample preparation: L1000

We generated the L1000 data according to the protocol outlined in Subramanian et al, 2017 [3]. Briefly, we cultured cells on 384 well plates and exposed them to compound treatment at various doses for 24, 48, or 72 hours. After the incubation time, we lysed cells and subjected them to ligation-mediated amplification (LMA) and detection. We captured mRNA using oligo-dT coated beads and reverse transcribed the sequences into cDNA. We PCR amplified the cDNA using biotinylated, barcoded primers and gene-specific juxtaposed probe pairs resulting in gene-specific, barcoded, and biotinylated PCR amplicons. We then hybridized these amplicons to Luminex beads, stained them with streptavidin R-phycoerythrin (SAPE), and detected them using a Luminex FlexMAP 3D scanner. Therefore, each bead reports the barcode, which determines gene identity, and we measure the SAPE fluorescent intensity, which indicates transcript abundance.

L1000 data processing

We processed L1000 data into perturbagen-specific differential expression signatures as described in Subramanian et al 2017 [3]. Briefly, we captured raw fluorescent intensities (FI) from the Luminex FlexMAP 3D scanner for each of the 978 L1000 landmark genes (Level 1 data). We then deconvoluted FI data to extract the median FI (MFI) for the two genes being measured by each Luminex bead barcode (Level 2 data). We loess-normalized the MFI values to the ten L1000 invariant gene sets within each well, and then quantile normalized all wells on the same detection plate, which resulted in each sample having the same empirical distribution (Level 3 data). We then computed gene-wise robust z-scores per sample, using all other samples on the same plate as the reference distribution (Level 4 data). Lastly, we collapsed biological replicates into consensus signatures using a weighted average, where each replicate was weighted by its average correlation with the others (Level 5 data). We made all data and metadata publicly available on figshare [51].

Image feature extraction

To extract image features, we built a CellProfiler (version 2.2.0) [52] image analysis pipeline and ran it on Amazon Web Services using Distributed-CellProfiler [53]. We also performed illumination correction to standardize readouts and account for confounding factors by homogenizing light across all fields of view [54]. The image analysis pipeline segments cells by distinguishing nuclei from cytoplasm and then extracts measurements for specific features related to the various

channels captured (see Sample preparation: Cell Painting). Specifically, we measured fluorescence intensity, texture, granularity, density, location, and various other measurements for each single cell (see https://cellprofiler-manual.s3.amazonaws.com/CPmanual/index.html for more details). Following the image-analysis pipeline, we obtain 110,012,425 cells and 1,790 feature measurements across 136 different plates. The pipelines are available online here https://github.com/broadinstitute/imaging-platform-pipelines/tree/3eb4ff5676aa7889666f09b606cd915c8b9ea839/cellpainting_a549_20x_phenix_bin1.

Cell Painting image-based profiling

After extracting CellProfiler readouts from all Cell Painting images of segmented single cells, we applied an image-based profiling pipeline to process morphology readouts (**Supplementary Figure 1B**). In the first step of this pipeline, we used cytominer-database (https://github.com/cytomining/cytominer-database) to collect and validate all CellProfiler output measurements from Cells, Cytoplasm, and Nuclei compartments for every site (field of view). The output of this first step is a set of SQLite files that contain raw single cell profiles per plate (level 2 data).

Next, we used pycytominer to process the single cell readouts [55]. We performed a standard image-based profiling pipeline [41] consisting of profile aggregation, annotation (level 3 profiles), normalization (level 4a), feature selection (level 4b), and forming consensus signatures (level 5). We performed median aggregation and normalized aggregated profiles using the "mad_robustize" method, which scales features independently by subtracting each value by the median and dividing by the median absolute deviation. We normalized each plate using the DMSO controls only, which allows us to more easily compare profiles across plates. We also performed several standard feature selection operations to remove features with missing data ("drop_na_columns"), remove features with low variance ("variance_threshold"), remove features that are highly correlated with other features ("correlation_threshold"), and remove blocklist features ("blocklist"). These blocklist features include CellProfiler features that we've previously observed to be unstable and noisy [56].

Because the negative control DMSO profiles were noisy due to technical artifacts, we applied a spherize transform (also known as whitening) to mitigate the impact of well positioning [35,36]. We also formed consensus signatures (level 5) by moderated z-score (MODZ) aggregating all replicate wells across plate maps into a single signature. We applied feature selection to the consensus signatures and batch effect corrected profiles separately using the same operations as described above. We applied the same pipeline to batch 1 (A549) and batch 2 (A549, MCF7, and U2OS) Cell Painting datasets.

We provide all the image-based profiles (level 3 and up) and the data processing pipelines in a versioned and publicly available github repository at https://github.com/broadinstitute/lincs-cell-painting/ [57].

Constructing an appropriate null distribution to calculate reproducibility metrics
In order to calculate percent replicating, we designed null distributions to control for three things: 1) different replicate cardinalities between different compound treatments, 2) well position on the 384 well plate, and 3) treatment dose.

Specifically, for percent replicating, for a given perturbation *x* located on well *w* measured across *n* replicates and treated with dose *p*, we randomly sampled *n* non-replicate profiles assayed in well *w* (but from different plate maps) from all perturbations that were treated with dose *p*. We performed this sampling procedure 1,000 times per replicate cardinality (e.g. compounds with 3 replicates, 4 replicates, 5 replicates, etc.) with two additional restrictions: (1) the random sample did not include replicates for perturbation *x*, and (2) no two compounds of the same non-*x* perturbation were included in the same null group. For example, in cases where a compound treatment at a specific dose had five replicates, we sampled 1,000 groups of five randomly sampled non-replicate profiles of the same dose. We used level 4 profiles considering compound and dose information as replicates, and we considered a replicating profile one in which the ground truth median pairwise replicate correlation was higher than 95% of the null distribution. We therefore calculate the percent replicating metric as the proportion of all replicating profiles over all common perturbations. This 95% thresholding procedure is equivalent to calculating per-treatment non-parametric p values (by counting how many times the replicate pairwise correlation was greater than the non-replicate null distributions) and reporting how many compounds were above an alpha p-value threshold of 0.05. We report this percent replicating implementation in **Figure 1**.

We also calculated percent replicating by relaxing the null distribution constraints. We performed the procedure as described above except we did not require the non-replicates be drawn from the same well position. We performed this analysis to provide evidence of the impact of well position on percent replicating interpretation. We compare same-well and any-well percent replicating metrics in **Supplementary Figure 3**.

We also introduced and calculated a second metric, which we called “percent strong” (see **Supplementary Figure 4**). In percent strong, we construct the non-replicate null distribution without adjusting for well position or replicate cardinality. We did still, however, calculate dose-specific null distributions. Specifically, for each modality and normalization strategy independently, we calculate a single null distribution for each dose by randomly sampling 1,000 groups of non-replicate profiles per replicate cardinality (the same null distributions for percent replicating, but we ignore replicate cardinality) and compute median non-replicate pairwise correlations. We subsequently calculate percent strong as the percentage of replicate median pairwise Pearson correlations greater than 95% of the full non-replicate null distribution. Percent strong provides more possible combinations of non-replicate sampling and therefore is not as susceptible to sampling biases as percent replicating. In other words, because percent replicating strictly samples non-replicates from the same well, if a specific well, by chance, housed very similar perturbations, the non-replicate distribution might be unduly skewed and deflate percent replicating scores. Percent strong is the least constrained null distribution and is robust to normalization strategy and subsampling (see Subsampling subsection).

We calculated percent replicating and percent strong using Cell Painting and L1000 input data with five different normalization strategies: 1) Cell Painting level 4 spherized profiles; 2) Cell Painting level 4 non-spherized profiles (median aggregated features with z-score normalization); 3) Cell Painting level 4 spherized subsampled profiles (see below); 4) L1000 level 4 spherized profiles; and 5) L1000 level 4 non-spherized profiles.

For percent matching, we performed a similar procedure as percent replicating. The only differences were that we (1) used level 5 consensus signatures from both data sets and (2) considered MOA and dose information as replicates. We used level 5 consensus signatures instead of level 4 replicate signatures, because consensus signatures are less noisy and correct for potential different replicate cardinalities per compound within an MOA.

We subsequently constructed dose and MOA compound cardinality-specific null distributions to compare against. Specifically, for each MOA, we calculated its median pairwise replicate correlation. We next randomly sampled 1,000 groups of level 5 consensus profiles of the same cardinality of the MOA compound count. For example, if an MOA contained 10 compounds, we formed one group by randomly sampling 10 compounds from different MOAs. We only sampled compounds measured at the same dose, and we did not include any two compounds of the same MOA in each random sample. For each of the 1,000 randomly groups, we calculated median pairwise correlations, which formed our percent matching null distribution. Lastly, we calculated a compound specific p value by dividing how many times the real median pairwise correlation of replicates was higher than all 1,000 randomly sampled groups of median pairwise correlations. We considered a matched MOA one in which the ground truth MOA median pairwise correlation was higher than 95% of the null distribution. We therefore calculate the percent matching metric as the total number of matched MOAs over all common MOAs.

Subsampling Cell Painting level 4 profiles to match L1000 replicate count
We collected fewer L1000 profiles than Cell Painting profiles. In most cases, with some exceptions, we collected three L1000 replicates and five Cell Painting replicates. We collected samples according to standard operating procedures for both assays, which pertain to sample handling and costs.

To determine the extent to which our percent replicating metrics were biased by replicate count, we performed a subsampling experiment using the spherized Cell Painting profiles. Specifically, we randomly sampled Cell Painting profiles without replacement to match exactly the same number of L1000 replicates for the individual compound of interest. Using this subsampled dataset, we calculated percent replicating. We also recalculated the null distribution using subsampled profiles.

Plate diffusion analysis to test the impact of plate position effects
We performed a plate diffusion analysis to assess plate position biases in Cell Painting and L1000 data. Specifically, for a given well w with treatment x collected on plate map P, we collected all non-replicate samples across all plate maps except P in a specific well

neighborhood as defined by diffusion parameter *d*. In other words, we selected all non-replicate wells in a predefined local neighborhood around well *w*. We used five different diffusion parameters (0, 1, 2, 3, and 4) to define this neighborhood. For d=0, we only included non-replicate samples from the same well, for d=1, we included all adjacent neighbors of well w on different plate maps, for d=2, we included all adjacent neighbors plus all neighbors' neighbors on different plate maps, and so on. After defining these non-replicate samples, we calculated all combinations of pairwise replicate correlations between treatment *x* and all non-replicate samples and calculated the mean of the distribution of well-neighborhood pairwise correlations.

Furthermore, we not only considered the local neighborhood around well *w*, but also the local neighborhood around the hypothetical plate-flipped version of well *w* (e.g. well P24 is the flipped version of well A01) in collecting non-replicates to analyze. In practice, the scientists collecting the data put the 384-well plate in the data collection machine in one of two orientations. Including this mirror parameter ensures that our diffusion analysis captures any technical plate effects introduced by different plate orientations.

We used the same five level 4 input data sets with different normalization strategies as we defined in the percent replicating and percent strong methods subsection. We report the mean of the total well-neighborhood pairwise correlations to determine consistent plate position technical artifacts per well position. If a strong plate position effect were present, then we would expect to see neighborhood correlations substantially drop with increasing diffusion.

Quantitative assessment of profile clustering

Using spherized Cell Painting level 4 profiles and non-spherized L1000 level 4 profiles, we performed three iterative clustering analyses in which we fit algorithms across a range of cluster numbers between k = 2 and k = 40 and acquired three goodness-of-fit heuristics (Silhouette scores, Davies Bouldin scores, and Bayesian Information Criterion (BIC) scores) for both datasets. Briefly, the Silhouette score is a metric indicating how separable clustering solutions are, with a score of 1 indicating that the identified clusters are clearly separable [58]. The Davies Bouldin score quantifies the ratio of within-cluster distances to between-cluster distances when comparing each cluster to their most similar neighboring cluster, and a lower value indicates more separable clusters [59]. BIC is a measurement of cluster likelihood and cluster predictability with an added penalty for increased cluster number, and a lower value indicates better clustering [60]. We visualize the tradeoff of these heuristics as we fit clustering algorithms with increasing cluster numbers.

For each model fitting, we used all 1,327 common compounds transformed into PCA space using 350 components. Therefore, we fit all clustering algorithms and calculated goodness-of-fit metrics using data of the same feature dimension, which, if not identical, can skew metrics and make comparison difficult.

Specifically, we applied k-means clustering with a maximum of 1,000 random iterations, across the k=2 to k=40 cluster number range, and calculated Silhouette and Davies Bouldin scores

from the resulting cluster solutions. We also fit full covariance Gaussian Mixture Models (GMM) with a k-means initialization and 1,000 maximum iterations, and we calculated BIC scores from the resulting clustering solutions. We performed this procedure using profiles resulting from each of the six different treatment doses independently, as well as using all profiles combined.

Calculating signature strength and activity score

To compare how different compound perturbations impacted individual feature measurements for both L1000 gene expression and Cell Painting morphology assays, we calculated signature strengths and activity scores as previously described [3]. Specifically, signature strength counts the number of features that substantially change when a sample is perturbed with a specific compound. We determined a substantially changed feature as one with a value greater than 2 after multiplying its z-score (transformed with respect to all compounds) by the square root of the number of replicates. We multiply by the square root of the number of replicates to enable more direct comparison of scores across compounds with different replicate counts.

Counting features in this fashion is equivalent to computing the absolute magnitude of change – we are implicitly transforming each feature so that values above 2 (or below -2) are mapped to 1 (or -1) and the rest are mapped to 0 (a “hard” sigmoid), and are then measuring the L1 norm (or L1 magnitude) of the resulting transformed vector. Intuitively, compounds that induce many features to high absolute value z scores are very disruptive of steady state, and compounds that don’t change very many features are not broadly strong perturbations. Instead, these compounds may either have little impact or be highly specific, meaning they only target one, or a few features strongly.

Activity score, either Morphological Activity Score (MAS) or Transcriptional Activity Score (TAS) for the Cell Painting or L1000 assays respectively, is the geometric mean of signature strength and median replicate correlation, normalized by the square root of number of features in the assay such that the resulting metric ranges between 0 and 1. A high activity score indicates compounds that reproducibly induce large changes in many features for a particular assay readout.

Identifying independent groups of features in assay measurements

To analyze feature redundancy and estimate the number of feature modules per assay, we calculated pairwise Pearson correlations of level 5 consensus profiles of Cell Painting (spherized) and L1000 (non-spherized) assays. We applied the same feature selection procedure in both assays, using pycytominer [55]. Specifically, we removed redundant features (as defined as having pairwise Pearson correlations < 0.9), features with low variance, and blocklist features [56]. This resulted in 1,020 Cell Painting features and 974 L1000 features. We calculated pairwise Pearson correlations of these features for all common compounds perturbed with 10 μM of compound. We visualized feature-level correlations using ComplexHeatmap [61].

Using sci-kit learn [62], we applied principal component analysis (PCA) with n_components = 150 using feature-selected level 4 profiles for each assay independently. PCA provides the

percentage of variance explained for each orthogonal component, and we use this information to determine the variety of signals in each feature space [63].

Supervised mechanism of action prediction: Multilabel-classification framework
We structured the classification task to predict compound MOAs from different input profiling modalities. Specifically, we created a binary label matrix for each individual MOA with corresponding labels for each compound. This formulation created a multi-label framework because many compounds have previously been annotated with two or more specific mechanisms [33]. For example, if a compound is annotated to mechanism “A” and mechanism “B”, the binary matrix would include positive labels for two different columns.

We used Cell Painting and L1000 profiles to predict the same MOA binary matrix. In all cases, we used level 4 replicate profiles as input for model classification. For Cell Painting, we used feature-selected spherized profiles (level 4bs) and for L1000 we used non-spherized profiles. We treated each input datasets in exactly the same fashion as we describe in the subsections below.

Supervised mechanism of action prediction: Training and test splits
In order to prevent signal leakage from the training set into the test set, we carefully split the compounds as input into the training and test sets. Specifically, we first split compounds based on MOA count. This means, for example, that if an MOA was represented by just one compound, we placed that compound in the training set. However, if an MOA had more than one compound, we split the compounds for that individual MOA between training and test set based on the 80/20 train/test ratio. Because some compounds are annotated to more than one MOA (hence “multi-label”), we needed to iterate, repeating the random splits, until these conditions were satisfied for all MOAs. Ultimately, this results in zero overlap of compounds in the training set compared to the test set. We used the same exact training and test set compounds for each assay.

To ascertain and verify that the classification models are learning from the training set and that they could generalize well on test set data, we created a shuffle data set using data in the training set. The shuffle data set consists of the same features and data as the normal training set, but we randomly shuffled target labels. We provided incorrect MOAs for all replicate profiles, and retrained and reevaluated all models on the same tasks.

Supervised mechanism of action prediction: Cross validation and model selection
To account for class imbalance in compound replicates in each multi-label MOA in the training set, we divided the compounds into two major groups based on treatment replicate count: less frequent and highly frequent. The same compound may be annotated to multiple MOAs, but we considered each MOA label independently when splitting data for cross validation.

We then applied a 5-fold double-stratified cross-validation strategy to the training set; we stratified the folds using both multi-label and treatment frequency categories. We evenly distributed compound replicates of different MOAs across folds while handling the rare and

common replicates differently. Specifically, we assigned less frequent compounds to the same fold, but we distributed highly frequent compounds evenly across folds. The threshold for dividing the compounds into rare/less frequent and highly frequent categories is 24. This threshold number means if the compound is found in less than 24 compound replicates in the training set it is considered less frequent, otherwise is considered highly frequent. In practice, most compounds in our training set belonged to the “less frequent” category. This procedure, termed “drug stratification” in the MOA Kaggle competition [34], caused our training folds to be evenly distributed by MOA category and to mostly contain unique compound perturbations.

We performed drug stratification using the training set to estimate each classification model’s performance, and to tune the hyperparameters needed to enhance model generalizability. Specifically, for every k-fold iteration, we trained our models using four random training subsets and tested on the 5th hold-out validation subset. In addition to testing cross validation performance on the held-out fold subset, we also made predictions on the entire testing dataset (using 4/5 folds) and averaged the results.

Supervised mechanism of action prediction: Model architecture

We chose models for the multi-label MOA predictions from the top-2 winners from the MOA Kaggle competition [34]. The models included 1D-Convolutional Neural Network (1D-CNN), TabNet (Attentive Interpretable Tabular Learning), Residual Neural Network (ResNet) and Simple Neural Network (Simple-NN) [64–67]. We modified the winning architectures to handle different assay input dimensions.

Specifically, the 1D-CNN architecture consisted of four convolutional layers with kernel sizes of 3 and 5, stride of 1 and padding sizes of 2 and 1. We added adaptive and max pooling layers, as well as batch normalization [68] and drop-out layers within the convolutional architecture to encourage better model generalization. The TabNet architecture consisted of a width of 64 for the decision prediction layer, a width of 128 for the attention embedding for each mask, 1 step in the architecture and gamma value of 1.3. The ResNet architecture consisted of six fully-connected layers with batch normalization and drop-out layers included within the architecture. We used rectified linear units (RELU) and exponential linear units (ELU) as activation functions between layers [69,70]. The Simple-NN architecture consisted of three fully-connected layers accompanied with batch normalization layers, drop-out and linear activation function layers. The optimization phase for all the models was done using Adam Optimizer with varying learning rates [71]. We independently optimized each architecture using data from each assay using the cross validation strategy as described above.

We also used an ensemble of the above-mentioned models in the MOA predictions by combining individual model predictions weighted by test set performance, then averaging the predictions to get an ensemble/blended version of all the models. We used multi-label k-nearest neighbors (K-NN) as a baseline model to compare performance [72,73].

For complete details of all architectures and implementation instructions, refer to https://github.com/broadinstitute/lincs-profiling-complementarity [74].

Supervised mechanism of action prediction: Feature engineering and data normalization
Prior to model training, we added features to the training and test sets. These features included principal components, UMAP features, factor analysis components, and statistical features such as sum, mean, kurtosis and standard deviation of all the features, for all four input datasets. Specifically, we added 25 UMAP features and 50 factor analysis components from the existing data prior to the Simple-NN, and we added 25 principal components to the 1D-Convolutional Neural Network, TabNet and ResNet models. Lastly, we added statistical features to the TabNet model. We normalized all features using z-score normalization prior to model training.

Supervised mechanism of action prediction: Model evaluation
The output of all the models is a probabilistic value between 0 and 1 corresponding to the probability of the model predicting a given MOA class label. We evaluated models calculating area under the Precision-Recall curve (AUPR). AUPR is a threshold-invariant metric that takes into account recall and precision, of which precision is particularly important because it measures the fraction of correct predictions among the positive predictions. AUPR accounts for imbalanced datasets, which is useful for evaluating classification tasks in highly imbalanced datasets [75]. We also calculated AUPR in randomly shuffled MOA class labels. To create this randomly shuffled matrix, we kept the MOA label count the same per MOA.

We used micro-averaging in our AUPR calculation for both "global" performance and per-MOA metrics. For the "global" AUPR (total performance) we aggregate the contributions of all compounds to compute the average metric. For the per-MOA AUPR we aggregate the contributions of all compounds annotated to the specific MOA.

Data availability
All code to reproduce this analysis is located at https://github.com/broadinstitute/lincs-profiling-complementarity [74]. All code to reproduce the Cell Painting image-based profiling pipeline is available at https://github.com/broadinstitute/lincs-cell-painting [57]. The L1000 data are available at figshare [51]. We are currently in the process of uploading all 45.8 TB of images to IDR, and we will update the manuscript once we have an accession number and direct object identifier to reference.

Acknowledgements
The authors thank Joshua Sacher for his help in curating Drug Repurposing Hub compound metadata. The authors gratefully acknowledge funding from the National Institutes of Health (NIH R35 GM122547 to AEC) and funding from the National Institutes of Health Common Fund LINCS Grant (NIH U54 HG008699 to AS). We would also like to acknowledge the use of the Opera Phenix High-Content/High-Throughput imaging system at the Broad Institute, funded by the S10 Grant NIH OD-026839-01.

References

1. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA

Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).

2. Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Syst* **6**, 13–24 (2018).
3. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017).
4. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
5. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. *Genome Med.* **12**, 8 (2020).
6. Musa, A. *et al.* A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinform.* **19**, 506 (2018).
7. Cazaly, E. *et al.* Making Sense of the Epigenome Using Data Integration Approaches. *Front. Pharmacol.* **10**, 126 (2019).
8. Di Minno, A., Gelzo, M., Stornaiuolo, M., Ruoppolo, M. & Castaldo, G. The evolving landscape of untargeted metabolomics. *Nutr. Metab. Cardiovasc. Dis.* (2021) doi:10.1016/j.numecd.2021.01.008.
9. 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).
10. Ottestad, A. L., Wahl, S. G. F., Grønberg, B. H., Skorpen, F. & Dai, H. Y. The relevance of tumor mutation profiling in interpretation of NGS data from cell-free DNA in non-small cell lung cancer patients. *Exp. Mol. Pathol.* **112**, 104347 (2020).
11. Wang, Y. & Ma, H. Protein kinase profiling assays: a technology review. *Drug Discov. Today Technol.* **18**, 1–8 (2015).
12. Waylen, L. N., Nim, H. T., Martelotto, L. G. & Ramialison, M. From whole-mount to

single-cell spatial assessment of gene expression in 3D. *Commun Biol* **3**, 602 (2020).

13. Yang, X. *et al.* High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. *Front. Genet.* **11**, 19 (2020).
14. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
15. Hu, Y. *et al.* Single Cell Multi-Omics Technology: Methodology and Application. *Front Cell Dev Biol* **6**, 28 (2018).
16. McFarland, J. M. *et al.* Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
17. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
18. Yu, C. *et al.* High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat. Biotechnol.* **34**, 419–423 (2016).
19. 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).
20. Gustafsdottir, S. M. *et al.* Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **8**, e80999 (2013).
21. 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. U. S. A.* **111**, 10911–10916 (2014).
22. Way, G. P. *et al.* Predicting cell health phenotypes using image-based morphology profiling. *Mol. Biol. Cell* **32**, 995–1005 (2021).
23. Filzen, T. M., Kutchukian, P. S., Hermes, J. D., Li, J. & Tudor, M. Representing high throughput expression profiles via perturbation barcodes reveals compound targets. *PLoS Comput. Biol.* **13**, e1005335 (2017).
24. Feng, Y., Mitchison, T. J., Bender, A., Young, D. W. & Tallarico, J. A. Multi-parameter

phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discov.* **8**, 567–578 (2009).

25. 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).
26. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
27. Nyffeler, J. *et al.* Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling. *Toxicol. Appl. Pharmacol.* **389**, 114876 (2020).
28. Lapins, M. & Spjuth, O. Evaluation of Gene Expression and Phenotypic Profiling Data as Quantitative Descriptors for Predicting Drug Targets and Mechanisms of Action. *bioRxiv* 580654 (2019) doi:10.1101/580654.
29. Haghighi, M., Singh, S., Caicedo, J. & Carpenter, A. High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations. *bioRxiv* 2021.09.08.459417 (2021) doi:10.1101/2021.09.08.459417.
30. Nassiri, I. & McCall, M. N. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic Acids Res.* **46**, e116–e116 (2018).
31. Caicedo, J. C. *et al.* Predicting compound activity from phenotypic profiles and chemical structures. *bioRxiv* 2020.12.15.422887 (2021) doi:10.1101/2020.12.15.422887.
32. Dahlin, J. L. *et al.* Nuisance compounds in cellular assays. *Cell Chem Biol* **28**, 356–370 (2021).
33. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
34. Kaggle.com, Laboratory for Innovation Science at Harvard, Connectivity Map & NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS). Mechanisms of action (MoA) prediction. *Kaggle* https://www.kaggle.com/c/lish-moa (2020).
35. Kessy, A., Lewin, A. & Strimmer, K. Optimal Whitening and Decorrelation. *Am. Stat.* **72**,

309–314 (2018).

36. Ando, M. D., McLean, C. & Berndl, M. Improving Phenotypic Measurements in High-Content Imaging Screens. *bioRxiv* 161422 (2017) doi:10.1101/161422.
37. Lin, A. *et al.* Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Sci. Transl. Med.* **11**, (2019).
38. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
39. van der Maaten, L. Visualizing Data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwAR0Bgg1eA5TFmqOZeCQXsIoL6PKrVXUFaskUKtg6yBhVXAFFvZA6yQiYx-M (2008).
40. Squires, C., Shen, D., Agarwal, A., Shah, D. & Uhler, C. Causal Imputation via Synthetic Interventions. (2020).
41. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
42. Schenone, M., Dančík, 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).
43. Niepel, M. *et al.* Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat. Commun.* **8**, 1–11 (2017).
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**, 1–14 (2020).
45. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *bioRxiv* 136168 (2017) doi:10.1101/136168.
46. Pratapa, A., Doron, M. & Caicedo, J. C. Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.* **65**, 9–17 (2021).
47. Rose, F. *et al.* Compound Functional Prediction Using Multiple Unrelated Morphological

Profiling Assays. *SLAS Technol* **23**, 243–251 (2018).

48. Boyd, J. C., Pinheiro, A., Del Nery, E., Reyal, F. & Walter, T. Domain-invariant features for mechanism of action prediction in a multi-cell-line drug screen. *Bioinformatics* **36**, 1607–1613 (2019).
49. Caicedo, J. C. *et al.* Predicting compound activity from phenotypic profiles and chemical structures. *bioRxiv* 2020.12.15.422887 (2021) doi:10.1101/2020.12.15.422887.
50. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* **0**, (2017).
51. Natoli, T. *et al.* L1000 data for LINCS profiling complementarity analysis. (2021) doi:10.6084/m9.figshare.13181966.
52. Kamentsky, L. *et al.* Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179–1180 (2011).
53. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
54. Singh, S., Bray, M.-A., Jones, T. R. & Carpenter, A. E. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.* **256**, 231–236 (2014).
55. Way, G.P., Chandrasekaran, S.N., Bornholdt, M., Fleming, S.J., Tsang, H., Adeboye, A., Cimini, B., Weisbart, E., Ryder, P., Stirling, D., Jamali, N., Carpenter, A.E., Singh, S. *Pycytominer: Data processing functions for profiling perturbations*. (2021).
56. Way, G. Blocklist Features - Cell Profiler. (2020) doi:10.6084/m9.figshare.10255811.v3.
57. Natoli, T. *et al. broadinstitute/lincs-cell-painting: Full release of LINCS Cell Painting dataset*. (2021). doi:10.5281/zenodo.5008187.
58. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
59. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 224–227 (1979).

60. Schwarz, G. Estimating the Dimension of a Model. *aos* **6**, 461–464 (1978).

61. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

62. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

63. Jolliffe, I. T. Principal Component Analysis and Factor Analysis. *Principal Component Analysis* 115–128 (1986) doi:10.1007/978-1-4757-1904-8_7.

64. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).

65. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) doi:10.1109/cvpr.2016.90.

66. Arik, S. O. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. (2019).

67. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

68. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015).

69. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). (2015).

70. Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). (2018).

71. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014).

72. Fix, E. & Hodges, J. L. Discriminatory analysis: Nonparametric discrimination: Consistency properties. *PsycEXTRA Dataset* (1951) doi:10.1037/e471672008-001.

73. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* vol. 46 175–185 (1992).

74. Way, G., Adeboye, A. & Singh, S. broadinstitute/lincs-profiling-complementarity: Paper submission DOI. (2021) doi:10.5281/zenodo.5580032.

75. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).