bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Title

Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations

Authors

Srinivas Niranj Chandrasekaran[1], Beth A. Cimini[1], Amy Goodale[1], Lisa Miller[1], Maria Kost-Alimova[1], Nasim Jamali[1], John Doench[1], Briana Fritchman[1], Adam Skepner[1], Michelle Melanson[1], John Arevalo[1], Juan C. Caicedo[1], Daniel Kuhn[2], Desiree Hernandez[1], Jim Berstler[1], Hamdah Shafqat-Abbasi[1], David Root[1], Sussane Swalley[3], Shantanu Singh[1^#], Anne E. Carpenter[1^#]

^ co-senior author
# co-corresponding author

Affiliations

1. Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA, 02142
2. Merck Healthcare KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany
3. Biogen, Inc., 125 Broadway Street, Cambridge, MA 02139

Abstract

We present a new, carefully designed and well-annotated dataset of images and image-based profiles of cells that have been treated with chemical compounds and genetic perturbations. Each gene that is perturbed is a known target of at least two compounds in the dataset. The dataset can thus serve as a benchmark to evaluate methods for predicting similarities between compounds and between genes and compounds, measuring the effect size of a perturbation, and more generally, learning effective representations for measuring cellular state from microscopy images. Advancements in these applications can accelerate the development of new medicines.

Introduction

Computer vision has benefitted dramatically from the revolution in deep learning. Biomedical research is an exceptionally satisfying domain on which to apply advances in machine learning, and yet its application has been relatively limited to supervised learning for medical imaging from patients, including classification and segmentation of X rays and MRI, PET, and CT scans. Similarly, deep-learning based image analysis for cell biology has generally focused on supervised tasks, such as segmentation (Caicedo et al., 2019; Moen et al., 2019), while representation learning and other image modeling applications have lagged behind (Pratapa, Doron, & Caicedo, 2021).

One cell biology method – image-based profiling of cell samples – is proving increasingly useful for the discovery of disease underpinnings and useful drugs (Chandrasekaran, Ceulemans, Boyd, & Carpenter, 2020). In image-based profiling, human cells are cultured in samples of a few hundred cells, each sample treated with a different chemical or genetic perturbation. The resulting morphology (visual appearance) of each sample is observed by microscopy, then

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

compared to identify meaningful differences and similarities. More than a dozen applications in biology and drug discovery have been demonstrated, including: (a) identifying the mechanisms of a disease by comparing cells from patients with a disease to those without the disorder and (b) identifying gene functions or the impact of chemicals on cells by unsupervised clustering of large sets of samples to determine relationships among the genes or chemicals tested in the experiment. Thus, image-based profiling can reveal disease mechanisms and potential therapeutics.

However, image-based profiling has yet to fully benefit from the latest machine learning research. The vast majority of studies use classical segmentation and feature extraction; deep learning methods are beginning to be explored (Pratapa et al., 2021) and there is much room for advancement. Historically, the lack of ground truth has been a major limiting factor in the field, for instance, the “correct” relationships among perturbations (e.g. genes and compounds) are unknown. While this is exciting because the potential for biological discovery is high, the lack of ground truth presents a challenge for optimizing deep learning pipelines. In fact, image-based profiling applications typically can be described as representation learning tasks (in the absence of ground truth); if samples are represented optimally and ideal distance metrics are applied, then biologically meaningful differences between samples will be detectable and technical artifacts will be suppressed.

To push forward advancements in this field, we assembled a consortium of ten pharmaceutical companies, two non-profit institutions, and several supporting companies, known as the JUMP-Cell Painting Consortium (Joint Undertaking in Morphological Profiling). After extensive optimization of the main assay used in image-based profiling, called the Cell Painting assay (Bray et al., 2016), this Consortium created a ground truth dataset, presented here, to move methods in the field forward. We selected and curated a set of genes and compounds with (relatively) known relationships among each other, and designed an experimental layout to enable testing and comparing methods to quantify their relationships.

Here, we describe our design and creation of this dataset via a single large experiment comprising roughly three million images and seventy five million single cells, called CPJUMP1, which contains chemical and genetic perturbation pairs that target the same genes in cells and therefore ought to match. It allows exploring a number of technical and biological parameters that might affect matching ability and testing computational strategies to optimally represent the samples so that they can be compared and thus uncover valuable biological relationships.

Framing the dataset for the ML community

An essential aspect of what we present here is that a very limited amount of rather noisy ground truth exists; each known drug-gene interaction was painstakingly discovered after hundreds of thousands of dollars of effort over many years, and many pairings are uncertain. By contrast, many mainstream machine learning (ML) tasks are oriented to replicate specific human skills where ground truth can be collected at large scale (e.g. translation or image recognition), given sufficient resources.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Our hope is that novel ML methods developed using our dataset will be used to discover new gene-compound connections (Rohban et al., 2021). This can yield new therapeutics for particular diseases, or identify how a potential drug is working and thus add to ground truth for this problem in the future.

Further, our dataset provides a challenging, real world testbed for many kinds of ML algorithms. It is a large-scale perturbation experiment with complex multi-dimensional, hierarchical data (images displaying dozens of cells each), and we believe new ML strategies still need to be developed to realize its full potential. In addition to the prediction problems we present in this paper, it also opens up problems in high-level reasoning on experimental data, allowing the study of complex artificial intelligence strategies, such as causal inference (observations from interventional experiments), planning (optimizing the next intervention that maximizes discovery), and simulations (what would have happened if other interventions are applied).

Finally, unusual aspects of the data type that we present pose challenges to ML algorithms and will require they be pushed in different directions to adapt. This may spark creative solutions with broader impact in ML. For example:

- Multiplexed imaging (in our case, with five channels) will push the field of machine learning to adapt to domains outside of RGB natural images, where the number and relationship among the channels (e.g. the extent of correlation) is very different than for natural images. This may require more generic data modeling to include visual tasks other than natural images.
- Lack of manually-assigned ground truth, or inability to crowdsource specialized annotations: the traditional cross-validation experimental model needs rethinking how to ensure that observations are statistically consistent across experiments using quality metrics (e.g. replicate correlation) rather than hard ground truth comparisons. In addition to the ground truth we provide for the gene-compound benchmark, this dataset has additional weak labels in a real world domain (e.g. sample locations, batches, cell type, time point), which can guide experts on interpreting their findings, but cannot be used as traditional ground truth data to guide learning algorithms.
- Fairness in machine learning: batch effects in experimental biology can generate biases similar to those observed in other applications due to lack or unawareness of data distributions (e.g. underrepresented minorities). Training on one batch may fail to make accurate predictions in data from other batches. Other problems related to data distributions include unbalanced classes and rare phenotypes.

Related datasets

We are not aware of any other Cell Painting image-based datasets that include pairs of genetic and chemical perturbations with their relationships to each other annotated, and executed in parallel so as to minimize technical variations that may confound the signal; this latter point makes this dataset unique and highly valuable for developing methods to match chemical and genetic perturbations. Furthermore, this is the only image-based dataset with two different genetic perturbation types, CRISPR and ORF (described later), which allows exploring their relationship and efficacy. Although there are other datasets that can be used for the purpose of

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

representation learning (see next paragraph), this is a small, well-controlled dataset which in some contexts might be advantageous for training, over larger, disorderly datasets. Another advantage of this dataset is that it enables learning representations on one domain (a particular cell line or time point) and then testing on another domain. Finally, for the gene-compound matching problem, there are no other genetic perturbation Cell Painting datasets that are large enough for both training and testing.

Nevertheless, other Cell Painting datasets are public and may be useful to the community, for example as training data for self-supervised representation learning methods. These single-perturbation-type experiments include several datasets from the Carpenter-Singh laboratory (available through the Image Data Resource (Williams et al., 2017) at https://idr.openmicroscopy.org/search/?query=Publication%20Authors:Carpenter and the 2018 CytoData challenge https://github.com/cytodata/cytodata-hackathon-2018), one from the New York Stem Cell Foundation (Schiff et al., 2020) and several from Recursion, a clinical-stage biotechnology company (available at http://rxrx.ai). Almost all of these datasets involve chemical perturbations or different patient cell lines rather than genetic perturbations.

Data acquisition

Compound and gene selection

The CPJUMP1 dataset consists of images and profiles of cells that were perturbed separately by chemical and genetic perturbations, where both sets were chosen based on known "matching" relationships among them. Chemical perturbations are small molecules (i.e. chemical compounds) that affect the function of cells while the genetic perturbations are either *open reading frames* (ORFs) that overexpress genes (i.e. yield more of the gene's product in the cell) or *guide RNAs* that mediate CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats), which knockdown gene function (i.e. yield less of the gene's product in the cell).

We therefore designed CPJUMP1 such that for each gene, we have one ORF that produces a higher-than-normal amount of that gene's product, two CRISPR guides that yield a lower-than-normal amount of that gene's product, and one or two compounds that are thought to impact the cell by influencing the function of that gene's product.

Most compounds are thought to inhibit the function of their target gene's product (as opposed to making it overly active), so we expect image-based profiles from cells treated with CRISPR to generally correlate to (mimic) the corresponding compound's profile, whereas ORF profiles are generally expected to anti-correlate (oppose) the corresponding small molecule's profile, and ORFs and CRISPRs targeting the same gene should generally yield opposite (anti-correlated) effects on the cells' profiles. However, we strongly note that there will be numerous exceptions given the non-linear behavior of many biological systems and a number of distinct mechanisms by which these general principles may not hold (Rohban et al., 2021). In fact, one aim of generating this dataset is to quantify how often the expected relationships and directionalities occur.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We derived the list of compounds from Broad's Drug Repurposing Hub dataset (Corsello et al., 2017), a curated and annotated collection of FDA-approved drugs, clinical trial drugs, and pre-clinical tool compounds (Figure 1d). The genes perturbed by genetic perturbations were chosen because they are the annotated targets of the compounds. The specific criteria for compounds, genetic reagents (considering their on- and off-target effects), and controls, and their layout on the plates (Figure 1a-c), are described in the supplementary materials. After applying the filters and including controls, we selected a total of 306 compounds and 160 genes such that they could fit into three 384-well plates.

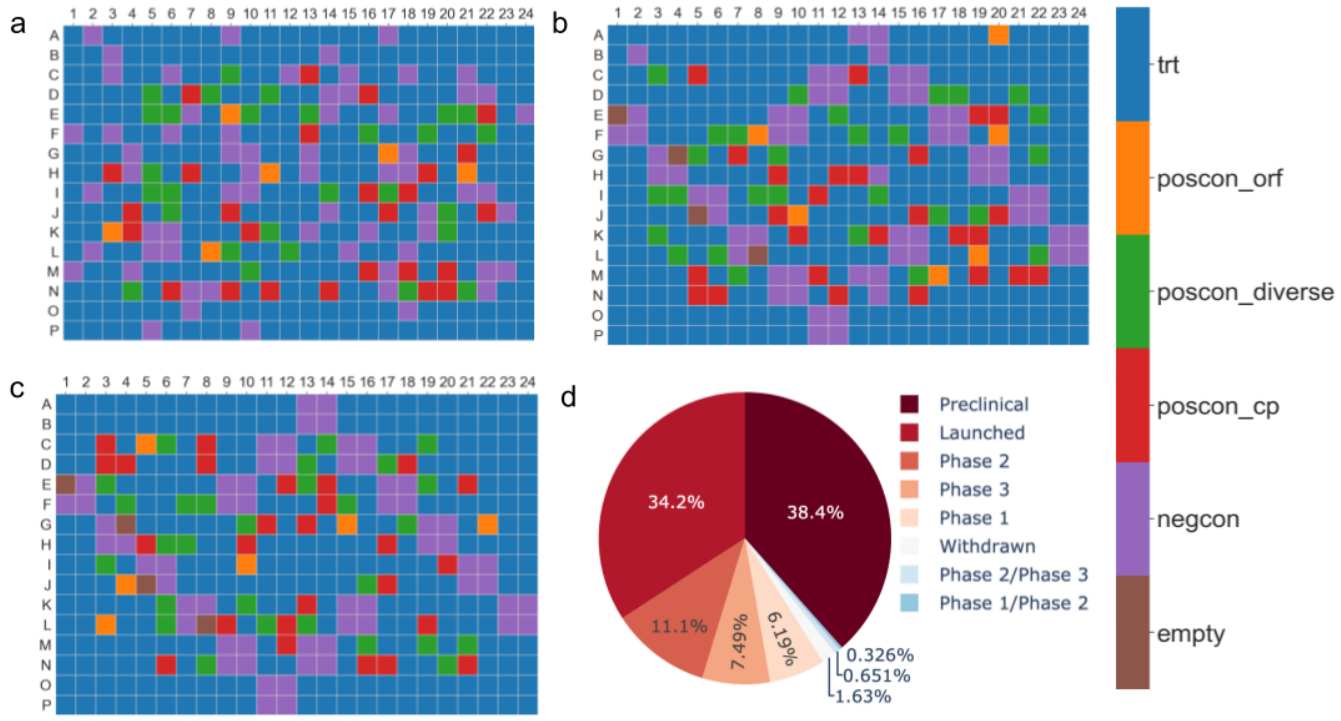

Figure 1: **Plate maps and clinical phase status**. Maps in a-c show a) Compound plate b) CRISPR plate and c) ORF plate. The control wells and the treatment (trt) wells are shown in different colors. Poscon are positive controls (additional details in the supplementary materials) and negcon is the negative control. d) Over a third of the compounds in the dataset have been launched whereas others have progressed to various stages of human clinical trials.

Experimental conditions

Although constrained by cost, we captured the compound, ORF, and CRISPR plates under various experimental conditions in order to identify those that improve gene-compound and gene-gene matching. We did more replicate plates for conditions that were less expensive or that were the most promising, and for the compound and CRISPR plates which had only

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

singlicates of most samples (as compared to ORFs which had duplicates within the plate; see supplementary materials for more details and to see the full list of treatments).

In addition to being used in optimizing the assay conditions, these plates can be used as multiple separate held-out test sets; tremendously valuable in any machine learning benchmark. As well, these conditions offer multiple “views” on cells treated with each chemical or genetic perturbation, and therefore can be used for many interesting machine learning explorations, such as style transfer e.g. to attempt prediction of one experimental condition from another.

Potential uses

Benchmarking representation learning methods

The holy grail of image-based profiling is a representation derived from the visual appearance of cell samples such that samples in biologically similar states have similar representations. Given such a representation, solutions for many of the applications discussed become immediately accessible, and can help us uncover mechanisms of disease and their potential cures.

Hand-engineered features have been carefully developed and optimized to capture cellular morphology variations in multi-dimensional representations. These features are the current standard in the field and still require extensive post-processing, including normalization, feature selection and dimensionality reduction. With advances in representation learning during the last decade, it is natural to ask what set of features could be automatically identified by machine learning algorithms directly from pixels.

Previous work in representation learning for image-based profiling has focused on using convolutional neural networks trained in four main ways: 1) Pre-trained in non-cellular datasets such as ImageNet (Ando, McLean, & Berndl, 2017; Pawlowski, Caicedo, Singh, Carpenter, & Storkey, 2016). Also known as transfer learning, this approach leverages generic feature extractors to measure cellular variations after proper feature normalization and rescaling. 2) Trained on cellular images using unsupervised learning (Lafarge et al., 2019; Lu, Kraus, Cooper, & Moses, 2019), including pretext tasks such as predicting certain image channels or all channels using autoencoders. 3) Trained on cellular images using weakly-supervised learning (Caicedo, McQuin, Goodman, Singh, & Carpenter, 2018; Cuccarese et al., 2020; Doan et al., 2020), where metadata about the experimental conditions has been used as weak labels to train models. 4) Self-supervised learning on cellular images (Perakis et al., 2021). More recently, approaches based on contrastive learning have been used to characterize cellular variations in image-based profiling as well.

All of these advancements have been explored in small-scale imaging screens, usually based on other imaging assays different from Cell Painting. The performance evaluation across approaches is not uniform, and it is unclear how these advancements could benefit the identification of phenotypic matches in large scale Cell Painting images with a diverse set of perturbations. Most works also make the assumption that it is ideal for convolutional nets to take the full stack of channels simultaneously (following standard practices in RGB image analysis), however, each image channel in Cell Painting has its own semantics. We envision the use of

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

other architectures, such as extensions of attention-based models (e.g. Vision Transformers (Dosovitskiy et al., 2020)) where the sequence of channels could be modeled in different and meaningful ways.

As a way to compare different representation methods, we created a benchmark based on detecting how many samples are measurably different from negative controls. Although other tasks might be defined, we chose perturbation detection as the task to evaluate representations because it is a task that often precedes other useful applications, and is equivalent to measuring statistical significance of the perturbation's signal. For example, a set of chemical or genetic perturbations might be filtered by this criterion before embarking on subsequent laboratory experiments, or prior to training a model, or other analysis that could be confounded by noisy signals. It can also be useful for determining what experimental protocol or computational analysis pipeline to use among several alternatives. It should be noted that even given perfect computational methods for feature extraction, batch correction, and profile comparison, not all samples will be detectably different from negative controls for several biological reasons. For example, a drug or genetic perturbation may only impact cell morphology in a particular cell type, under particular environmental conditions, at a particular time, or if particular stains were used, conditions which may not have been met in the experiment.

To detect the number of samples with a measurably distinct phenotype, we estimated the standard metric used in the field, *Percent Replicating* (Figure 2), which is the proportion of samples that are distinct from the null distribution built from samples that are non-replicates. A sample is considered to have a detectable signature if the median of the correlation between the replicates of the sample is greater than the 95th percentile of the null distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

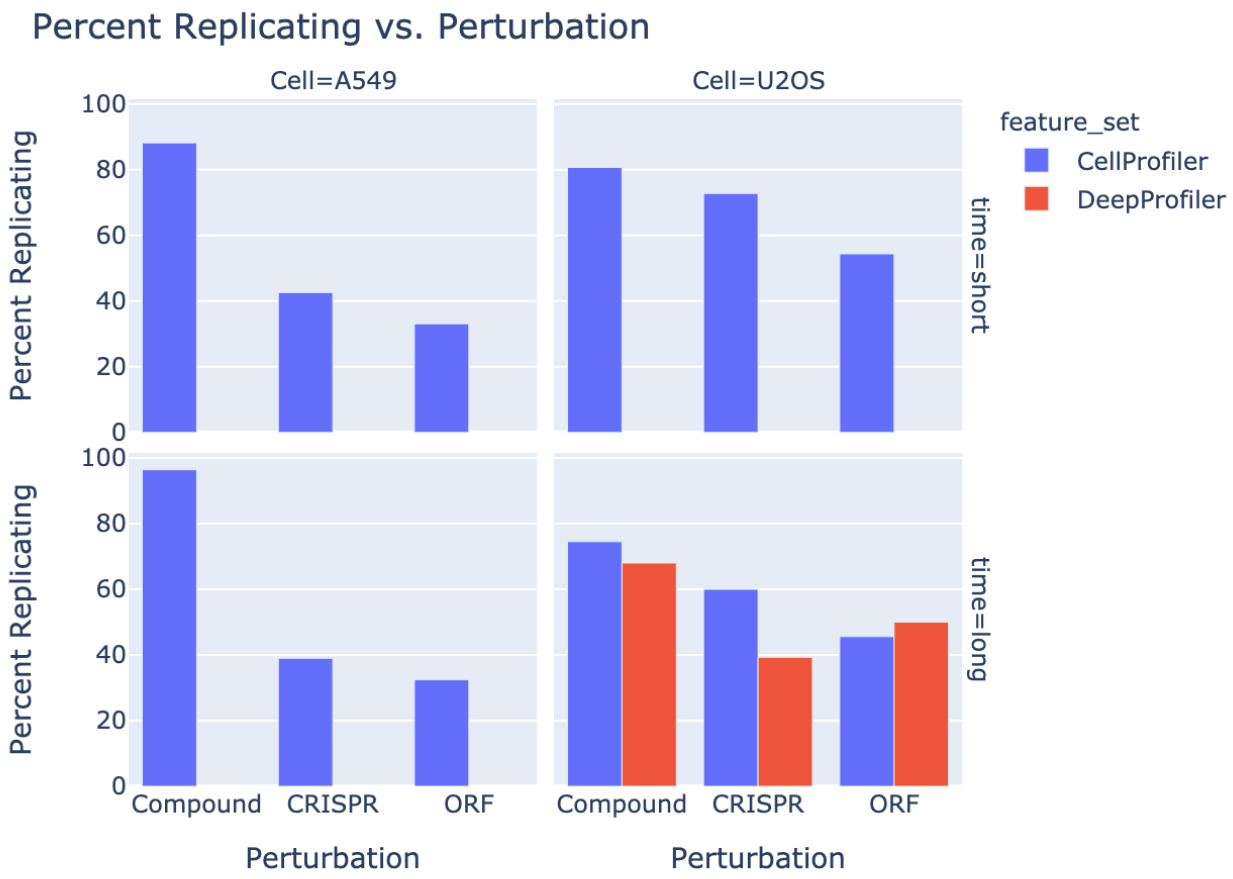

Figure 2: *Percent Replicating* for each perturbation modality and under several experimental conditions. Across all experimental conditions and using CellProfiler-derived features (blue), including cell type (columns) and time points (rows), compounds have a stronger within-replicate correlation compared to CRISPRs, which are themselves stronger than ORFs. The same is true for features extracted using an ImageNet pretrained network (EfficientNet; red).

We also created a deep-learning baseline by extracting features using an ImageNet pretrained network (EfficientNet (Tan & Le, 2019)) for a subset of samples. This is a useful baseline to evaluate transfer learning strategies, which have been reported to be competitive on this problem (Ando et al., 2017; Pawlowski et al., 2016). In this baseline, the *Percent Replicating* values (Figure 2, red) show a similar trend compared to that of CellProfiler-derived features, where the compounds have higher within-replicate correlation compared to the genetic perturbations. It seems likely that a more specialized network would perform better and we propose this as a ripe direction for future research.

Benchmarking gene-compound matching methods

We next established a benchmark and evaluation setup for researchers to develop and test strategies for a real-world retrieval task, where we search for genes or compounds that have a similar impact on cell morphologies as the query gene or compound. This dataset presents a

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unique opportunity to match profiles of perturbations across modalities (chemical versus genetic), because genes in this dataset that are targeted by two types of genetic perturbations (ORF and CRISPR) are also targeted by two compounds. To establish a baseline approach to match profiles across modalities, we computed the Pearson correlation (used as a similarity metric) between all chemical and genetic perturbation pairs. We then evaluated the performance by estimating *Percent Matching* (Figure 3), which is the proportion of "true" connections (chemical-genetic perturbation pairs that target the same gene) that are distinct from a null distribution built from "false" connections (chemical-genetic perturbation pairs that are not known to target the same gene). A true connection is considered to be correctly detected if its correlation is greater than the 95th percentile of the null distribution.

The baseline results show that there is a signal in this dataset for matching chemical and genetic perturbations that target the same gene (~7-11%, against a false positive rate of 5%), but there is much room for improvement. It should be strongly noted, though, that significant time and resources are otherwise required to identify the target of a compound, and similarly to identify compounds that target a particular gene. Therefore, even the baseline's relatively low matching rates can accelerate drug development by yielding a list of possibilities for biologists to test directly in subsequent experiments; improving image representations and therefore the accuracy of predicted matches by a few percentage points could have a major impact on the pharmaceutical industry.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

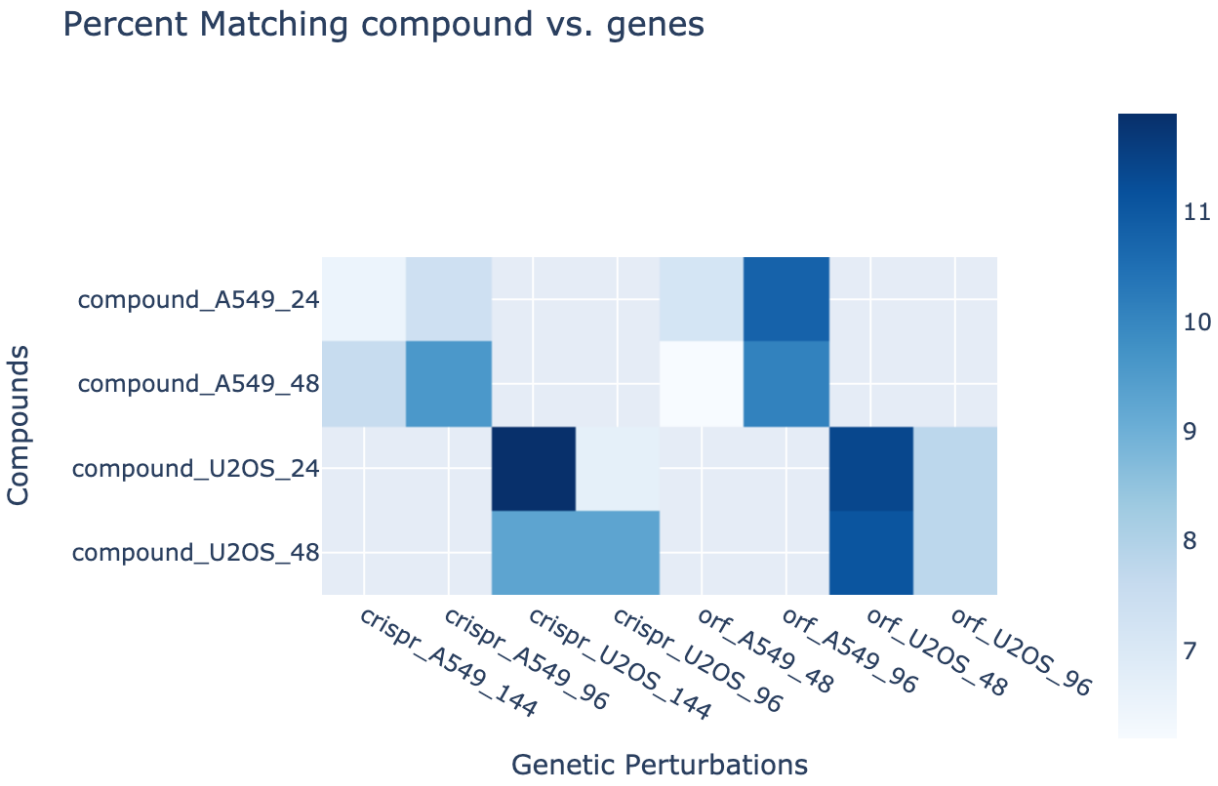

Figure 3: *Percent Matching* between compounds and genetic perturbations. Axis labels include the cell lines (A549 or U2OS) and the timepoints (48 hour, 96 hour, and 144 hour). Compound-compound matching could be evaluated similarly, given that there are pairs of compounds targeting the same gene; however, many larger datasets exist for this purpose.

Conclusion

Drug discovery would benefit greatly from the machine learning community turning its attention to rich, single-cell imaging data. The simple benchmarks we created aim to provide a foundation on which researchers can develop and test novel methods for representation learning, multi-view learning, information retrieval, and style transfer, among many others.

Code and Data availability

Morphological profiles (both CellProfiler and DeepProfiler-derived), image analysis pipelines, profile generation pipelines, plate maps and plate and compound metadata, and instructions for retrieving the cell images are publicly available online at https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted. The code for reproducing the benchmark results is available at https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/tree/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/benchmark

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Code for generating Figure 1 is available at https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/tree/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/visualization

Code for generating Figure 2 is available at https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/blob/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/benchmark/1.percent_replicating.ipynb, https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/blob/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/benchmark/4.percent_replicating_dl.ipynb and https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/blob/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/benchmark/7.generate_figure.ipynb

Code for generating Figure 3 is available at https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/blob/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/benchmark/2.percent_matching_across_modalities.ipynb

Methods

Sample preparation and image acquisition

The Cell Painting assay involves staining eight components of cells with six fluorescent dyes: nucleus (Hoechst), nucleoli and cytoplasmic RNA (SYTO 14), endoplasmic reticulum (concanavalin A), Golgi and plasma membrane (wheat germ agglutinin; WGA), mitochondria (MitoTracker), and the actin cytoskeleton (phalloidin) (Figure 4). We optimized the Cell Painting assay described in (Bray et al., 2016) by changing the concentrations of Hoechst, phalloidin, concanavalin A and SYTO14 and combining dye addition and dye permeabilization steps. These changes will be described in more detail in (Cimini et al., in preparation) and are currently publicly available at https://github.com/carpenterlab/2016_bray_natprot/wiki#updates-to-the-cell-painting-protocol. The images were acquired across five fluorescent channels plus three brightfield planes using a Perkin Elmer Opera Phenix HCI microscope at 20x magnification.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

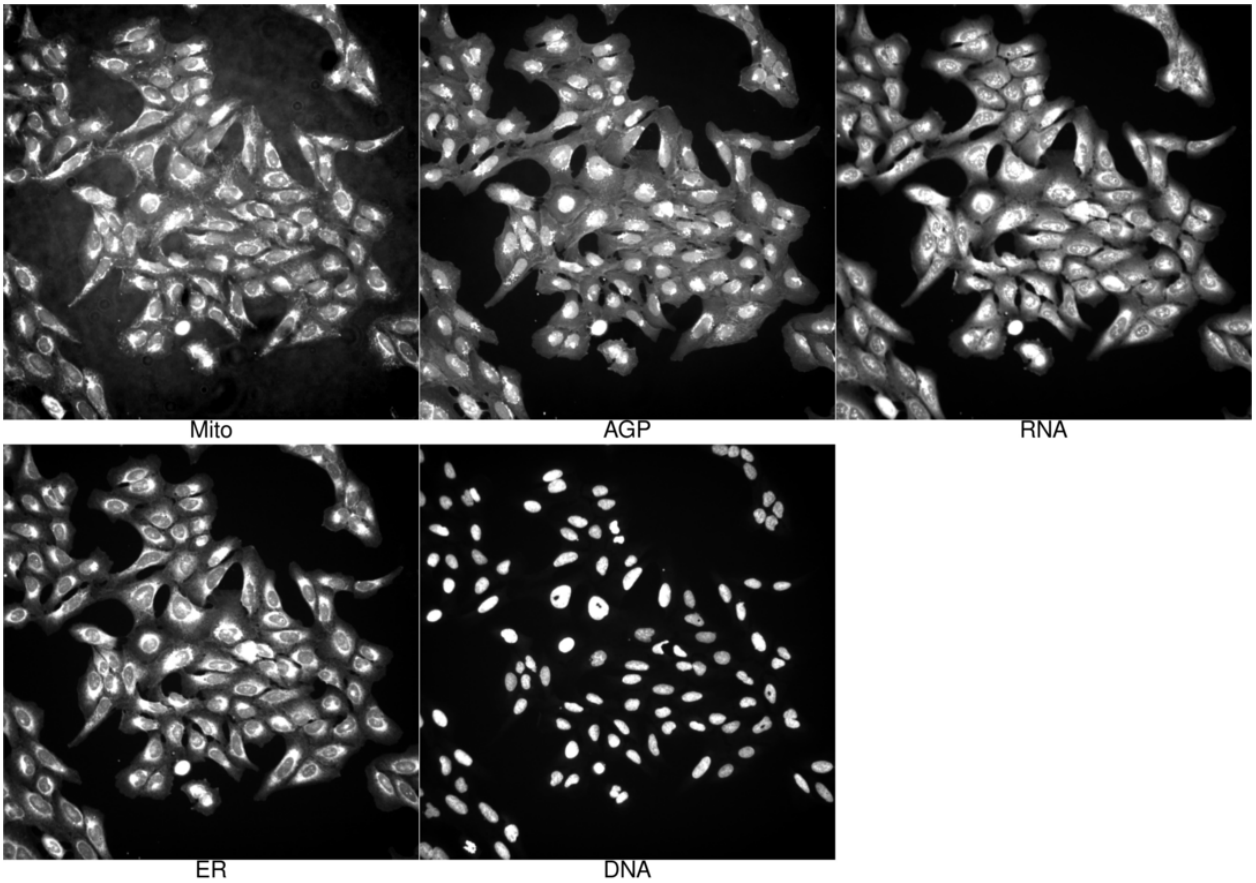

Figure 4: **Sample images from the dataset.** A single 5-channel image of U2OS cells treated with the compound PFI-1. The channel names indicate the cellular structures identified in each image (see main text for details). The width of each image is 645 um (and are 1080 pixels height and width). Other example images are available on https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/tree/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/example_images.

Image processing

We used the CellProfiler (McQuin et al., 2018) bioimage analysis software to process the images. We corrected for variations in background intensity, and then segmented cells, distinguishing between nuclei and cytoplasm. Then, across the various channels captured, we measure various features of cells across several categories including fluorescence intensity, texture, granularity, density, location (see http://cellprofiler-manual.s3.amazonaws.com/CellProfiler-3.0.0/index.html for more details). Following the image analysis pipeline (see (Bray et al., 2016) for the pipeline), we obtain more than 75 million cells and 5792 feature measurements.

Deep learning feature extraction

Segmentation bounding boxes around each cell are produced by CellProfiler and centered and cropped to 224x224 pixels keeping the aspect ratio. Each of the five channels is copied three

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

times to mimic RGB images, so that it can be given as input to the EfficientNet (Tan & Le, 2019) pretrained in the ImageNet classification task. The five vectors extracted from the second-to-last layer are concatenated, generating 6400 features that represent the cell's profile. Finally, all the cell feature vectors extracted from the same well are averaged to get a single representation for the well. This process was implemented using the DeepProfiler open source library (https://github.com/cytomining/DeepProfiler). The well-level aggregated features are spherized (Ando et al., 2017) for each set of replicate plates.

Image-based profiling

We used cytominer (Singh et al., 2020) (https://cytomining.github.io/profiling-handbook/) and pycytominer workflows (https://github.com/jump-cellpainting/profiling-recipe) to process the single cell features. We aggregated the single cell profiles by computing the mean. We then normalized the averaged profiles by subtracting the median and dividing by the median absolute deviation (m.a.d.) of each feature. This was done in two ways: using the median and m.a.d. of (i) the negative control wells on the plate (used in the analysis shown here), and (ii) all the wells on the plate. Finally, we filtered out redundant features as well as features with low variance. All the steps in the profiling workflow were performed for each individual plate separately.

Recommended dataset splits

The methods presented in the benchmarks do not involve any training (we simply use a predetermined similarity metric and hand-engineered features or a pre-trained model) and thus did not require creating the typical train-validate-test data splits. For the two benchmarks, representation learning and gene-compound matching, we offer the following guidelines for creating data splits when training is involved:

Representation learning: Depending on the use case, we suggest using different splits. For a general representation learning task, the compound, CRISPR or ORF dataset could be used with a 60-20-20 split. For a domain adaptation task, one could train on the dataset from one cell line or time point and test it on the other cell line or time point. All the replicates of a perturbation should be in the same split.

Gene-compound matching:

1. All replicates of a perturbation should be in the same split.
2. For the CRISPR dataset, both guides should be in the same split.
3. Three of the compounds (BVT-948, dexamethasone, and thiostrepton) have two different identifiers each in the dataset (because of small differences in structures) but the same compound name. Each pair should be in the same split.
4. If analyzing data at the single cell level, all cells from a well should be in the same split.

We provide recommended data splits in https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted/tree/ed0f918e5920021a3c6f9a8c2d63cf22c2067039/datasplits

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Limitations and Ethical concerns

This data can accelerate drug discovery and therefore improve human health and reduce drug development costs. Nevertheless, we note an ethical concern: the cell types are commonly used historical lines derived from two white patients, one male (A549) and one female (U2OS). Therefore, conclusions from this data may only hold true for the demographics or genomics of those persons and not broader groups. They were chosen because the lines are both well-suited for microscopy and they offer the advantage of enabling direct comparison to extensive prior studies using them.

There are additional limitations for the presented datasets, aside from their data quality as already noted. The number of gene perturbations captured in these datasets are in the few hundreds whereas there are roughly 21,000 genes in the genome and numerous variations within each. Likewise, a few hundred compounds are tested here but pharmaceutical companies often have collections of compounds numbering in the millions. In terms of the assay itself, the Cell Painting assay includes only six stains, which is insufficient to capture the localization and morphological variation of all cellular components. The only limitation for expanding these datasets are the financial resources to carry out the experiments (more samples, different microscopes, etc.).

We created this dataset using a single cell line and single facility; this choice limits the potential for generalizability of any models using it as training data. We note that generalizability of models across datasets is often unnecessary in biology experiments where controls can be included within each experiment; in fact, we recommend those creating large datasets to include these sets of controls in the experiment in order to have internal controls/landmarks for the assay. We also made this choice in order to minimize technical variability and therefore maximize the biological signal in the data. Given the relatively low percent matching of genes to compounds, the primary aim in the field is to develop methods that work well within a single dataset and only later is it worthwhile to aim for generalizability. Nevertheless, our consortium is currently collecting new data across ten sites as a resource for addressing the generalizability problem.

Acknowledgements and Disclosure of Funding

The authors appreciate the more than 100 scientists who have contributed to the organization and scientific direction of the JUMP Cell Painting Consortium. We thank Max Macaluso (operations) and Tanaz Abid (technical) at the Broad Institute for their assistance as well.

The authors gratefully acknowledge a grant from the Massachusetts Life Sciences Center Bits to Bytes Capital Call program for funding the data production. We appreciate funding to support data analysis and interpretation from members of the JUMP Cell Painting Consortium and from the National Institutes of Health (NIH MIRA R35 GM122547 to AEC). The authors also gratefully acknowledge the use of the PerkinElmer Opera Phenix High-Content/High-Throughput imaging system at the Broad Institute, funded by the S10 Grant NIH OD-026839-01.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

AEC has optional ownership interest in Recursion, a public biotechnology company using image-based profiling for drug discovery. SES is an employee of Dewpoint Therapeutics. Daniel Kuhn is an employee of Merck Healthcare KGaA, Darmstadt, Germany.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix

The landing page of the GitHub repository for this dataset has all the relevant additional information: https://github.com/jump-cellpainting/2021_Chandrasekaran_submitted.

We have released the data with a CC0 license and the code with a BSD 3-Clause license.

We have chosen GitHub as the hosting platform, and use GitLFS to store large files.

References

Ando MD, McLean C, & Berndl M. (2017). Improving Phenotypic Measurements in High-Content Imaging Screens (p. 161422). doi: 10.1101/161422

Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, & Carpenter AE. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.*, *11*(9), 1757–1774. doi: 10.1038/nprot.2016.105

Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghighi M, Heng C, Becker T, Doan M, McQuin C, Rohban M, Singh S, & Carpenter AE. (2019). Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods*, *16*(12), 1247–1253. doi: 10.1038/s41592-019-0612-7

Caicedo JC, McQuin C, Goodman A, Singh S, & Carpenter AE. (2018). Weakly supervised learning of single-cell feature embeddings (p. 293431). doi: 10.1101/293431

Chandrasekaran SN, Ceulemans H, Boyd JD, & Carpenter AE. (2020). Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.*, 1–15. Retrieved from https://www.nature.com/articles/s41573-020-00117-w

Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, Johnston SE, Vrcic A, Wong B, Khan M, Asiedu J, Narayan R, Mader CC, Subramanian A, & Golub TR. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.*, *23*(4), 405–408. doi: 10.1038/nm.4306

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cuccarese MF, Earnshaw BA, Heiser K, Fogelson B, Davis CT, McLean PF, Gordon HB, Skelly K-R, Weathersby FL, Rodic V, Quigley IK, Pastuzyn ED, Mendivil BM, Lazar NH, Brooks CA, Carpenter J, Jacobson P, Glazier SW, Ford J, Jensen JD, Campbell ND, Statnick MA, Low AS, Thomas KR, Carpenter AE, Hegde SS, Alfa RW, Victors ML, Haque IS, Chong YT, & Gibson CC. (2020). *Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery* (p. 2020.08.02.233064). doi: 10.1101/2020.08.02.233064

Doan M, Sebastian JA, Caicedo JC, Siegert S, Roch A, Turner TR, Mykhailova O, Pinto RN, McQuin C, Goodman A, Parsons MJ, Wolkenhauer O, Hennig H, Singh S, Wilson A, Acker JP, Rees P, Kolios MC, & Carpenter AE. (2020). Objective assessment of stored blood quality by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* doi: 10.1073/pnas.2001227117

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, & Houlsby N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Retrieved from http://arxiv.org/abs/2010.11929

Lafarge MW, Caicedo JC, Carpenter AE, Pluim JPW, Singh S, & Veta M. (2019). *Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning*. *102*, 315–325. Retrieved from http://proceedings.mlr.press/v102/lafarge19a.html

Lu AX, Kraus OZ, Cooper S, & Moses AM. (2019). Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLoS Comput. Biol.*, *15*(9), e1007348. doi: 10.1371/journal.pcbi.1007348

McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, Doan M, Ding L, Rafelski SM, Thirstrup D, Wiegraebe W, Singh S, Becker T, Caicedo JC, & Carpenter AE. (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.*, *16*(7), e2005970. doi: 10.1371/journal.pbio.2005970

Moen E, Bannon D, Kudo T, Graf W, Covert M, & Van Valen D. (2019). Deep learning for cellular

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

image analysis. *Nat. Methods*, *16*(12), 1233–1246. doi: 10.1038/s41592-019-0403-1

Pawlowski N, Caicedo JC, Singh S, Carpenter AE, & Storkey A. (2016). Automating Morphological Profiling with Generic Deep Convolutional Networks (p. 085118). doi: 10.1101/085118

Perakis A, Gorji A, Jain S, Chaitanya K, Rizza S, & Konukoglu E. (2021). Contrastive Learning of Single-Cell Phenotypic Representations for Treatment Classification. Retrieved from http://arxiv.org/abs/2103.16670

Pratapa A, Doron M, & Caicedo JC. (2021). Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.*, *65*, 9–17. doi: 10.1016/j.cbpa.2021.04.001

Rohban MH, Fuller AM, Tan C, Goldstein JT, Syangtan D, Gutnick A, Nijsure MP, Rigby M, Sacher JR, Corsello SM, Peppler GB, Bogaczynska M, Ciotti GE, DeVine A, Doan M, Gale JP, Derynck R, Turbyville T, Boerckel JD, Singh S, Kiessling LL, Schwarz TL, Varelas X, Wagner FF, Kafri R, Karin Eisinger-Mathason TS, & Carpenter AE. (2021). Discovery of small molecule pathway regulators by image profile matching (p. 2021.07.29.454377). doi: 10.1101/2021.07.29.454377

Schiff L, Migliori B, Chen Y, Carter D, Bonilla C, Hall J, Fan M, Tam E, Ahadi S, Fischbacher B, Geraschenko A, Hunter CJ, Venugopalan S, DesMarteau S, Narayanaswamy A, Jacob S, Armstrong Z, Ferrarotto P, Williams B, Buckley-Herd G, Hazard J, Goldberg J, Coram M, Otto R, Baltz EA, Andres-Martin L, Pritchard O, Duren-Lubanski A, Reggio K, NYSCF Global Stem Cell Array Team, Bauer L, Aiyar RS, Schwarzbach E, Paull D, Noggle SA, Monsma FJ, Berndl M, Yang SJ, & Johannesson B. (2020). Deep learning and automated Cell Painting reveal Parkinson's disease-specific signatures in primary patient fibroblasts (p. 2020.11.13.380576). doi: 10.1101/2020.11.13.380576

Singh S, Goodman A, Becker T, McQuin C, Rohban MH, Way GP, Cimini BA, & Carpenter AE. (2020). Cytominer: Methods for Image-Based Cell Profiling (Version 0.2.2). Retrieved from https://cran.r-project.org/package=cytominer

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.05.475090; this version posted January 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tan M, & Le Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 6105–6114. Retrieved from http://proceedings.mlr.press/v97/tan19a.html

Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U, Brazma A, Salas REC, & Swedlow JR. (2017). The Image Data Resource: A Bioimage Data Integration and Publication Platform. *Nat. Methods*, *14*(8), 775–781. doi: 10.1038/nmeth.4326