ScaleFEx^{SM}: a lightweight and scalable method to extract fixed features from single cells in high-content imaging screens

Bianca Migliori, Neeloy Bose, Daniel Paull

The New York Stem Cell Foundation Research Institute, New York, NY, USA

Abstract

High-content imaging (HCI) is a popular technique that leverages high throughput datasets to uncover phenotypes of cell populations *in vitro*. When the differences between populations (such as a healthy and disease state) are completely unknown, it is crucial to build very large HCI screens to account for individual (donor) variation, as well as having enough replicates to create a reliable model. One approach to highlight phenotypic differences is to reduce images into a set of features using unbiased methods, such as embeddings or autoencoders. These methods are powerful at preserving the predictive power contained in each image while removing most of the unimportant image features and noise (e.g., background). However, they do not provide interpretable information about the features driving the decision process of the AI algorithm used. While tools have been developed to address this issue, such as CellProfiler, scaling this tool to large sample batches containing hundreds of thousands of images poses computational challenges. Additionally, the resulting feature vector, computationally expensive to have generated, is very large in size (containing over 3000 features) with many redundant features, making it challenging to perform further analysis and identify the truly relevant features. Ultimately, there is an increased risk of overfitting due to the presence of too many non-meaningful features that can ultimately skew downstream predictions.

To address this issue, we have developed ScaleFEx^{SM}, a Python pipeline that extracts multiple generic fixed features at the single cell level that can be deployed across large high-content imaging datasets with low computational requirements. This pipeline efficiently and reliably computes features related to shape, size, intensity, texture, granularity as well as correlations between channels. Additionally, it allows the measurement of additional features specifically related to mitochondria and RNA only, as they represent important channels with characteristics worth to be measured on their own. The measured features can be used to not only separate populations of cells using AI tools, but also highlight the specific interpretable features that differ between populations. We applied ScaleFEx^{SM} to identify the phenotypic shifts that multiple cell lines undergo when exposed to different compounds. We used a combination of recursive feature elimination, logistic regression, correlation analysis and dimensionality reduction representations to narrow down to the most meaningful features that described the drug shifts. Furthermore, we used the best scoring features to extract images of cells for each class closest to the average to visually highlight the phenotypic shifts caused by the drugs. Using this approach, we were able to identify features linked to the drug shifts in line with literature, and we could visually validate their involvement in the morphological changes of the cells.

ScaleFEx^{SM} can be used as a powerful tool to understand the underlying phenotypes of complex diseases and subtle drug shifts at the single cell level, bringing us a step closer to identifying disease-modifying compounds for the major diseases of our time.

1. Introduction

High-content imaging (HCI) is a popular technique that leverages large datasets generated through high-throughput processes to uncover phenotypes of cell populations in vitro (Yang et al., 2019). HCI can help identify phenotypic differences between populations, such as healthy and diseased states, as well as the effects of exogenous perturbations such as compounds and genetic manipulations (Mullard, 2019). However, when these differences are uncovered using unbiased analysis (no preconceived phenotype being measured), it is crucial to build very large datasets that allow the development of reliable models, with enough replicates to account for variables such as individual (donor) variation and other common confounding factors such as plates and position (Tegtmeyer et al., 2023; Venugopalan et al., 2020). However the scale-up of datasets poses great computational and analytical challenges making it harder to characterize the datasets and uncover differences between populations of interest (Ando et al., 2017; Ashdown et al., 2020; Schiff et al., 2022).
One approach to highlight phenotypic differences is to reduce images into a set of features using unbiased methods, such as embeddings (Ando et al., 2017;Ashdown et al., 2020; (Lippeveld et al., 2022; Schiff et al., 2022) or autoencoders (Kobayashi et al., 2022). These methods can retain the predictive power of each image while removing unimportant features and noise. However, it has proven very challenging to try and interpret the black box that is behind deep learning networks, which characterizes both embeddings and encoders (Zaritsky et al., 2021). One of the first approaches to morphological characterization was to compute and confront measurements that could define populations of cells based on hypothesized features of interest (Carpenter et al., 2006; Giedt et al., 2016). However, this method is biased to specific features of interest and thus not able to capture phenotypes not specifically measured. A way around this is Fixed Feature Extraction (FFE), in which measurements not linked to prior knowledge are computed on the entire dataset, resulting in a very large set of features that define the dataset beyond personal biases. Even though the features can become quite comprehensive, these approaches are not fully unbiased because the measurements are limited to what a human mind can think about measuring (Lippeveld et al., 2022; Schindelin et al., 2012). With the increase in the number of features though, it becomes much harder to compute those features in a timely manner or to scale up to bigger datasets leveraging automated ingestion and parallel processing. The most popular tool among the FFE softwares is CellProfiler (Carpenter et al., 2006). CellProfiler is an open-source software designed for high-throughput image analysis of biological samples, allowing the extraction of quantitative information from microscopy images of cells, tissue, and other biological samples. The software features an intuitive Graphical User Interface that allows the creation of custom pipelines with cascaded measurements to be performed on a set of multichannel images. The creators have provided the community with a premade pipeline (Bray et al., 2017) to be used for Cell Painting assays (Bray et al., 2016; Way et al., 2021), which is a widely spread immunohistochemistry assay that highlights generic parts of the cells in 5 different channels. Unfortunately, scaling this tool to large sample batches containing hundreds of thousands of images poses computational and cost challenges. The authors address some of those issues by proposing a cloud computing docker tool that performs distributed parallel computing over multiple virtual machines and optimizes the processes based on cost or time (depending on the user)(Weisbart and Cimini, 2022). The problem with this approach is that it requires a familiarity with cloud computing optimization and programming, else the setup and execution can be very cumbersome. Beyond the technical issues, the resulting output (feature vector) is very large, containing thousands of features (over 3000 features on a Cell Painting assay), many of which are redundant and highly correlated, making it challenging to perform downstream analysis. This can lead to computational limitations, feature selection difficulties, and the risk of overfitting.

To address these challenges, we have developed Scaled Feature Extraction (ScaleFEx^{SM}), a Python pipeline that extracts multiple generic fixed features at the single cell level that can be deployed across large high-content imaging datasets with low computation requirements and coding expertise. ScaleFEx^{SM} efficiently and reliably computes features related to shape, size, intensity, texture, and granularity, as well as correlations between channels. Furthermore, it allows the measurement of additional features, specific to mitochondria and RNA, which have been proven to play a role in

cellular dysfunctions and linked to many complex diseases (e.g. Parkinson's Disease ((Antony et al., 2020),(Hsieh et al., 2019),(Frye et al., 2021))). The disruption in the appearance of these cellular components makes an extra-measuring step worthwhile both to provide meaningful additional features and for ease of interpretability. In all, the features generated by ScaleFEx^{SM} can be used not only to separate and characterize populations of cells using AI tools, but also to highlight the specific interpretable features that differ between populations.

We leveraged ScaleFEx^{SM} to identify phenotypic shifts that cell lines undergo when exposed to different compounds and concentrations. Using Recursive Feature Selection (RFE) and Logistic Regression (LR) algorithms in combination with correlation analysis, pertinent features were selected, summarized in tables for ease of interpretability, and their distribution plotted and analyzed for statistical difference significance. Furthermore, the best scoring features were used to extract images of cells for each class closest to the average, to visually validate the phenotypic shifts caused by the drugs.

Altogether, ScaleFEx^{SM} provides a novel, computation-friendly approach to extract interpretable features from HCI screens. As screens grow ever larger in size and complexity, providing scientists with relevant interpretable images is key to understanding the biological changes induced in disease and drugged states.

2. Methods

2.1 Cell collection and expansion

Cell lines used in this study were obtained through the New York Stem Cell Foundation Repository. All fibroblast cell lines derived from biopsies had been previously collected under IRB approved protocols, as previously described (Paull et al., 2015). Cell lines were expanded and frozen into working aliquots using the NYSCF Global Stem Cell Array®, a fully automated cell culture platform, which consists of liquid handling systems (Hamilton STAR) that are connected to (amongst other devices) automated incubators, Celigo cell imagers (Nexcelom), VSpin centrifuges (Agilent), and Matrix tube decappers (Hamilton Storage Technologies) (Paull et al., 2015; Schiff et al., 2022). In brief, frozen master stock tubes were thawed in a 37°C water bath for 1 minute. Upon thawing, a manual count was taken, and cells were seeded into either one well of a 6 well plate, or one well of a 12 well plate in Fibroblast Expansion Media (FEM). Cells plated into 6 well plates immediately entered automated expansion workflows, while those remaining in 12 well plates were allowed to grow to 70-100% confluence before being passaged into 6 well plates and entering automated expansion workflows. Cells were fed every 2-3 days using automated scheduling and were passaged every 7 days. Automated methods dissociated cells using TrypLE Select (12604013, ThermoFisher) before being centrifuged into a cell pellet. The pellet was suspended in 2 mL media, with two 25 µL aliquots from each sample used in automated cell counting. Based on the total live cell count, cells were frozen, using the cryoprotectant Synthafreeze (A1254201, ThermoFisher), into cryovials at 50,000-100,000 cells per vial or replated into 6 well plates for further expansion. DNA from each cell line was obtained during the process and used to confirm cell line identity using the SNPTrace™ Assay (100-6280, Standard BioTools). Cells were additionally screened for their sterility and the absence of mycoplasma during this process (LT07-703, Lonza).

2.2 Automated screening

Novel automated procedures were developed to run a drug screen leveraging the automated platform of the NYSCF Global Stem Cell Array®. 20 unique donor lines were selected, and a single vial of each line was automatedly thawed into one well per line of a 12-well plate (Fisher Scientific, 07-200-91) for post-thaw recovery, using automated methods as previously described. Cells were

fed on Days 2 and 5, and by Day 7, when the cell lines had reached over 90% confluence, the plates were passaged. At this point, the liquid handler aspirated the remaining media from the seeded wells, washed them with Phosphate Buffered Saline (PBS) (Gibco™, 10010072), and rinsed them with TrypLE. Cells were incubated for 25 minutes at 37°C, and then brought back on deck where the TrypLE was neutralized with FEM. The cell suspensions were then pipetted into an intermediate block (Corning® 3958) and centrifuged on the Agilent VSpin. Post-centrifugation, the supernatant was aspirated, and the cell pellet was resuspended in FEM. An aliquot of cell suspension was incubated with Hoechst (H3570, ThermoFisher) and Propidium Iodide (P3566, ThermoFisher) in a Perkin Elmer ViewPlate (502105844, ThermoFisher) before being counted on a Celigo automated imager. The counts were fed back into the automated method, allowing the determination of the required volume of cell suspension needed to seed the wells of the destination plate; a targeted density of 500 cells per well was used in this study. The method was provided with a pre-designed layout, and the liquid handler was able to stamp the desired pattern, such that there were 12 sample replicates per donor per plate. Positionally, each replicate had a different location within the plate layout to minimize biasing the model with plate effects, and the outer two rows and columns on the plate were only seeded with media to avoid edge effects. In total, the Hamilton system was able to create five replicate 384-well imaging plates (CellVis, P384-1.5H-N). Cells were stored in an automated incubator (Cytomat, Thermofisher) before being treated with various compounds, stained, and imaged, as described below.

2.3 Drug dispensation

Based on prior work and previously published data (Way et al., 2021), three drugs were determined to be of interest for the high-throughput screen due to prior success in shifting the cellular morphology of diseased primary human fibroblasts: Y-39983 HCl (CAS: 173897-44-4), Pemigatinib (INCB054828) (CAS: 1513857-77-6), and CP21R7 (CP21) (CAS: 125314-13-8). Stocks of the compounds of interest (1mM) were thawed and an additional 0.2 mM dilution of each compound was made in 1mM DMSO (D2650, Sigma). An i.Dot 2.0 Dispenser (Cellink, SO20-7799) was used to dispense the compounds across all five plates. The final drug layout across the plates was determined algorithmically to ensure an unbiased approach with an equivalent number of replicates and well volume per experimental condition. The i.Dot additionally dispensed an equivalent volume of 1mM DMSO on top of the wells that did not receive any compound to ensure that all wells contained a total of 0.1% DMSO. Once the wells were treated with either compound or DMSO control, they were placed back into a Cytomat where they were imaged overnight on the Opera Phenix™ High Content Screening System. The plates were allowed to incubate for 72 hours, before they were removed for fixation and staining, as described below.

2.4 Experimental layout

The seeding system (described above) was designed to stamp a pattern 4 times from a 96 well plate to a 384. Two columns and rows were cut out from the outer rim edge to avoid edge effects leaving the plate with well 12 replicates per cell line. Each stamp was replicated 4 times, leaving a layout with 4 adjacent wells with the same cell line. The drugs were then evenly distributed among the cell lines, but for each of the 4 neighboring wells with the same cell line there was at least 1 DMSO, to balance local position effects. We generated a total of 5 plates, with 300 wells seeded with DMSO and 150 wells per condition.

2.5 Staining and imaging

We were then able to harness high-content image-based assays for unbiased morphological profiling with multiplexed fluorescent dyes. The protocol published in Bray et al. was adapted to

fluorescently label the cells using an automated liquid handling system, as previously described (Schiff et al., 2022). The plates were placed onto the deck of the Hamilton Star Liquid Handler. Spent media was aspirated and replaced with FEMcontaining MitoTracker (Invitrogen™ M22426). Plates were then incubated at 37 °C for 30 minutes, followed by fixation in 4% Paraformaldehyde (Electron Microscopy Sciences, 15710-S) before being washed with 1x HBSS (Thermo Fisher Scientific, 14025126). The plates were then permeabilized in 0.1% Triton X-100 (Sigma-Aldrich, T8787) diluted in 1x HBSS (Thermo Fisher Scientific, 14025126). Cells were stained at room temperature with the Cell Painting staining cocktail for 30 min after two additional washes in 1x HBSS. The stain cocktail included Hoechst 33342 trihydrochloride, trihydrate (Invitrogen™ H3570), Molecular Probes Wheat Germ Agglutinin, Concanavalin A, Alexa Fluor® 488 Conjugate (Invitrogen™ C11252), SYTO® 14 Green Fluorescent Nucleic Acid Stain (Invitrogen™ S7576), Alexa Fluor® 568 Phalloidin (Invitrogen™ A12380), Alexa Fluor 555 Conjugate (Invitrogen™ W32464). Plates were washed three times post-incubation with 1x HBSS, sealed with 1x HBSS in the wells, and refrigerated until removal for imaging.

Plates were imaged using the Opera Phenix™ High Content Screening System. The inner 240 wells of each plate were imaged using a non-confocal objective with 40x magnification (Water, NA 1.1), and 49 non-overlapping fields were acquired per well. To capture the entirety of the Cell Painting panel, five channels were created with differing combinations of excitation and emission spectra for each. Specifically, 375 nm and 435-480 for Hoechst 33342, 488 nm and 500-550 for Concanavalin A488, 488 nm and 570-630 for SYTO14, 561 nm and 570-630 for WGA and Phalloidin, and 640 nm and 650-760 for MitoTracker Deep Red.

2.6 Calculation of features

ScaleFEx^{SM} algorithm is publicly available here: DOI: 10.5281/zenodo.8121681 . Briefly, this code defines a class named "ScaleFEx_screens" which is the pipeline for extracting a vector of fixed features from HCI screens. The class has several attributes that can be specified when creating an instance of the class, such as the experiment folder location, the experiment name, the destination folder for saving results, the plates to be analyzed, the channels to be analyzed, the size of the images, the type of cells being analyzed and segmentation method. The "init" method initializes the instance variables of the class and runs the pipeline for each plate specified in the "Plates" attribute. If the "parallel" attribute is set to "True", the pipeline is run for each plate in parallel using multiprocessing. The "cascade_functions" method performs the actual vector extraction from the images. It first queries the data files for the specified plate, makes a list of wells and fields, and checks if a feature vector file already exists for this plate to load the file in case the calculation was already started and interrupted.

The code then loads each image file specified in the input files defined above and resizes them to a specified size depending on the desired downsampling ratio. It next applies a flat field correction and normalizes the image intensity to a range of 0 to 255. For each well and field of view, the code extracts single-cell features. It first identifies cell nuclei by detecting Hoechst-stained regions before extracting a region of interest of a specified size around each nucleus and applies a single-cell feature extraction method to extract quantitative features from the ROI (described in Supplementary File 1).The feature vectors are stored in a pandas dataframe. The dataframe has a row for each cell ROI and a column for each feature. The code can be run for a single well and FOV (Wells and fields parameters) or for multiple wells and FOVs. The code also allows the user to process only a subset of the input data (indSS parameter) and to skip processing of previously processed data (flag and flag2 parameters). Additional details on code execution and explanation can be found here “TBD”.

2.7 Analysis details

A jupyter notebook is provided in the git-hub repository together with the code. The output of ScaleFEx^{SM} is a csv file for each plate containing 1 cell per row and the information about location (well, field and coordinates). After using pandas libraries to load the csv, we normalized all the values between 0 and 1 and averaged at well level. Then we used a logistic regression algorithm (sklearn.linear_model.LinearRegression) to predict drug vs control and evaluated the result for each confounder removed. To remove the contribution of any given confounder, we removed the mean of each one and re-normalized at the end of the process to keep the values between 1 and 0. We then visually assessed the data using a umap (from umap library). Then we performed variance thresholding (sklearn.feature_selection.VarianceThresholding). To determine how many features were necessary to score the best performance, we used a Recursive Feature Elimination model (sklearn.feature_selection.RFE) with a logistic regression estimator and max iteration of 100000 adding 100 features at a time for features to select. Once the number of features was found for every drug, we used the same model to select the best necessary features. From a Logistic regression algorithm, we ranked the features and checked how many times the same feature occurred within the cross validation folds. The result is summarized in Supplementary File 2. We then used Seaborn and matplotlib libraries to visualize the heatmap and the data distributions. A two-sided Mann–Whitney U test, Bonferroni adjusted was used to assess the statistical difference between the 2 classes. The images of cells were extracted by selecting the well with the closest feature of interest to its mean. A 3x3 tiles portion of the well was visualized to maintain a decent resolution.

3. Results

3.1 Ingestion and pre-processing

The primary goals of ScaleFEx^{SM} are to efficiently extract as many measurable features as possible from an image of a single cell acquired in multiple channels, while also being easy to use and adaptable to different immunohistochemistry panels. ScaleFEx^{SM} is designed to reduce the high redundancy that is common in feature extraction techniques and to minimize user input by computing features on a per-channel basis instead of a compartment basis. The resulting feature vectors obtained by ScaleFEx^{SM} contain a vast amount of information in the form of numerical values, which can be leveraged in combination with machine learning-based feature selection techniques or statistical methods for specific feature comparison. To demonstrate the efficacy of this approach, we conducted an analysis using a dataset of fibroblasts derived from 20 donors that were treated with varying concentrations of three distinct compounds. These cells were stained using a CellPainting panel and imaged at high resolution, and ScaleFEx^{SM} was used to extract features from the acquired images (Figure 1). The cells were plated using a liquid handler (see Methods section for details) in equal numbers into 5 384 well plates. To avoid outliers due to the proximity of the edges (where there is known enhanced evaporation of media), only the inner 240 wells were seeded, before being treated with two different concentrations of three drugs. The resulting dataset consisted of 294000 images of 2160x2160 pixels, 5 channels each field, 49 fields for each well, 240 wells per plate and 5 wells in total.

To run the algorithm over such a large amount of images poses great challenges, especially as GPUs, which are typically used to speed up image-related analysis, cannot be used for complex measurements (Owens et al., 2007; Shi et al., 2012; Simbaña, 2020). For this reason, ScaleFEx^{SM} was designed to keep in mind computational power limitations while ensuring stability and resource optimization using parallel processing. We designed a pipeline (Figure 1) to loop over plates, wells, fields, single cells and channels in a hierarchical way, loading one field at a time, with plate parallelization.

This process starts from a file name search from the main experimental structure. Folder structure and naming conventions are to be kept the same for the algorithm to work, or an intermediate

Python coder would be able to easily modify the data-loader to adapt to they're custom naming structure.

The file names and paths are read into memory, with wells and tiles identified using scan-specific naming conventions. Using a specified number of random images, the background trend is computed and used for flat field correction (one matrix of values per channel). Next, the images of the channels belonging to the same tile are loaded, and if the acquisition was performed in multiple planes, the images for each channel are flattened into a max projection. The images to be processed are loaded, flattened, flat-field corrected, normalized, and converted into 8 bits. After these pre-processing steps, the segmentation step is performed to locate the cells and start the single-cell analysis. This step can be performed using multiple different approaches. For this paper, as the dataset we used consisted of fibroblasts, which are very easy to separate, we used an Otsu's threshold algorithm. For each of the detected cells, a crop of a specified size (in this case 598x598) was segmented from the centroid of the nuclei. In order to minimize computational time and ensure consistency in our measurements across cells, we adopted a cropping approach over using the entire cell irrespective of its size. By doing so, we were able to maintain a uniform reference for all the cells, resulting in more comparable and homogeneous measurements.

For each of the channels, there is first a further segmentation step to identify the mask of the tagged portion of the cell within the crop, then channel-specific measurements are performed. In cases where multiple cells would be within the same crop, only the map overlapping with the nuclear map was considered part of the cell, and for the DNA channel, we used the mask that included the center pixel of the crop.

3.2 Features measurement

For each channel, a set of different measurements is extracted starting from the initial segmentation (Figure 1b, Supplementary Figure 1a) . The main categories of measurements are shape, texture, granularity, intensity, concentric measurements, overlap between channels, optional measurements specific to the mitochondria and RNA, and global annotations. All of the measurements are summarized and described in Supplementary Table 1. The total number of features is 1738 for a cell painting panel (5 channels), but it can vary depending on the number of dyes used and if the extra features for Mitochondria or RNA are selected. The distribution of summed features by category are shown in Figure 1b (Note that some features belong to more than 1 category, so they're counted twice).

3.2.1. Shape

Shape measurements are the features related to the shape and size of the cell. They are computed based on the per-channel segmentation mask, and measure features such as area, perimeter, radius and compactness. Shape measurements are important in understanding the high-level differences between populations and are the easiest to validate by eye.

3.2.2. Texture

Texture is a measure of a cell's definition, alongside the relative measure of dye taken up by any given cell. Texture is measured by computing a gray level co-occurrence matrix (GCLM), represented as a histogram describing the spatial relationship between pixels in an image and their gray levels. The GLCM is calculated by counting the number of times pairs of pixels with specific gray levels occur in a defined spatial relationship (calculated over multiple angles and distances). The resulting matrix is used to extract features that describe the texture, directionality, and other properties of the image. We computed texture for 5 distances and for 5 different angles. To avoid

background biases, we computed these measurements on the image multiplied by the mask (the background results in 0 values).

3.2.3. Granularity

Granularity in an image refers to the level of detail and resolution of its components. It is a measure of the size of the individual elements that make up an image, and how well these elements are defined and separated. A high-granularity image has a fine resolution with clearly defined elements, while a low-granularity image has a coarse resolution, and its elements appear blurred or blended. To measure granularity, a deconvolution process can be performed using round kernels of varying sizes on the image, and the average value can be calculated. Similarly, to the texture measurements, we assigned a 0 value to all the pixels outside the segmentation mask.

3.2.4. Intensity

Intensity measurements refer to the measurement of the brightness or darkness of individual pixels in an image. The intensity of a pixel is represented by a numerical value that corresponds to the amount of light that the pixel reflects or absorbs. The intensity measures help to identify if some dyes are particularly reactive or condensed on some cells in respect to others.

3.2.5. Concentric measurements

Concentric measurements are intensity and granularity measurements computed in a radial fashion. Concentric areas are used to define new masks to overlay to the image and compute over the reduced images the measurements described above. This set of features help to identify difference in the radial distribution within the cells.

3.2.6. Overlap between channels

This set of measurements calculates the relationships between different channels and how much they correlate to each other. They help highlight differences in the cell's compartments.

3.2.7. Mitochondria

This optional measurement is specific to the mitochondria channel. In computing this set of measurements, an additional segmentation step is performed within the mask, followed by a skeletonization to highlight the fine net-structures of mitochondria. From this second binary mask, volumetric information is extracted, and from its skeletonization (Supplementary Figure 1b), we can calculate the total extent of mitochondria, the number of branches and the number of end points.

3.2.8. RNA

Similarly, to the mitochondria measurements, the RNA measurements are also optional and calculate the number and size of the nuclear RNA organelles overlaid to the DNA mask. This measurement is performed through an extra step of segmentation skeletonization (Supplementary Figure 1b) and includes counts of organelles and total volume.

3.2.9. Global

Global measurements include the information retrieved at tile level and are the same for all the channels. It logs the density of the site (number of cells), and distance from the 2 closest cells. Having knowledge about the spatial surroundings of the cells is very important, as the cells can assume very different phenotypes depending on density and proximity to other cells.

3.3. Analysis

Following computation of features, we focused on extracting those most meaningful to our dataset using the steps shown in Figure 2a. We first evaluated the presence of confounding factors that might have masked or influenced the results as previously described (Schiff et al, Tegtmeyer et al.,Narayanaswamy et al., 2020). In brief, we evaluated the prediction score of a binary Linear Regression model built to identify each drug compared to control (DMSO) for each of the known confounders of the experiment (plates, wells, rows, columns, donor), on well level averages. The model was designed in 5 Cross Validation (CV) folds with a plate held out for testing for each fold, on data normalized between 0 and 1. We iteratively removed the contribution of each confounder both singularly and in combination and evaluated the prediction score and UMAPs. The result showed a UMAP more clustered around each drug, independently (Figure 2b), and an increase of the prediction score (Figure 2c) in comparison to the raw data (Supplementary Figure 2a-b) when correcting for plates, columns, rows, and donors.

The final data that we used for feature extraction was confounder-corrected and normalized, and it was averaged at well level to reduce noise and highlight the prevalent effects in most of the population of cells (Schiff et al., 2022). We next removed all features with low variance, which is particularly common for the concentric measures, as the circular outer parts of the cells might include empty pixels, leading to 0 measurements. We used a Recursive Feature Elimination (RFE) algorithm to eliminate the least important features for each drug based on a Logistic Regression (LR) model built to classify each drug vs control. To determine the minimum number of features necessary to reach the best accuracy using this method, we iteratively increased the features to keep and evaluated the prediction score for each drug, and picked the smallest number that yielded the best accuracy. After calculating the best features according to RFE, we used a LR algorithm to rank the features based on the importance score. We did it for each CV iteratively and then selected the features present in all the folds as the most meaningful. We then removed the most correlated features to reduce redundancy and to focus on the most important and distinct features that contribute to the drug shift, while avoiding over-emphasis on similar features. The resulting features are summarized in heatmaps (Figure 2d, Supplementary Figure 3-4) and in Supplementary File 2. To dig deeper into the phenotypes, we further looked at the highest ranking features for each drug and looked at the distribution of values (Figure 3, Supplementary Figure 3-4) to evaluate the differences in those specific features that were significantly enriched. The drug Y-39983 was found to significantly disrupt AGP solidity and mass displacement at varying doses, as well as affecting multiple mitochondrial phenotypes at both 0.2 and 1 nM. Significant changes in mitochondrial and endoplasmic reticulum eccentricity were seen in samples treated with 0.2 nM pemigatinib, with mitochondrial texture correlation significantly altered in the presence of both 0.2 and 1 nM of this drug. Similarly, CP21R7, amongst other features, significantly disrupted Mitochondrial mean volume at both 0.2 and 1.0 nM. We next selected examples close to the median of the distributions both for the drugs and DMSO controls as representative examples for the feature of interest, to visually understand and interpret the result. Overall, these results demonstrate the utility of ScaleFEx^{SM} in providing interpretable features disrupted through the exogenous addition of small molecules and the specificity to appropriate these features to specific drugs and doses.

4. Discussion

We developed and applied ScaleFEx^{SM} to identify the phenotypic shifts caused by drug treatments in skin fibroblasts taken from 20 different donors. This methodology allowed the characterization of cell populations beyond the confounding factors contributing to the variation, such as the donor'

themselves (based on their demographics) (Schiff et al, 2022), which can often mask/hide the drug-induced phenotypic shifts. From the UMAPs and linear regression prediction accuracy it is clear that the set of measurements consents to easily separate and cluster the cells treated with different compounds at different concentrations. Diving deeper into the specific features of interest that better characterize the drug shifts, we found that the compounds that caused a larger shift needed less features (Figure 3a) to correctly classify most of the wells, while for more subtle differences, a combination of features is needed (sup Figure 3,4). We evaluated the effects of 3 drugs (CP21R7, Pemigatinib and Y-39983-HCl) at 2 different concentrations (0.2mM and 1.0 mM). In general, all of the main features selected using our method showed a well separated distribution between control and treated wells with highly significant differences observed between treated and DMSO wells (Figure 3, Sup3, Sup4). The drug with the strongest effect was Y-39983-HCl, particularly at the highest concentration; only 2 features were necessary to confidently classify this population (Figure 3a): solidity of AGP and max radius of Mitochondria. Solidity is the ratio between the segmented area and its convex hull, which is the smallest convex polygon that completely encloses the object. Put simply, an image shape with low solidity would have a more irregular or fragmented contour. Figure 3a shows that cells treated with Y-39983-HCl at 1.0 nM possess a more irregular and edgy shape as compared to the controls. Similarly, the maximum radius of the mitochondria channel appears to be larger for the untreated cells. This is in line with literature where it has been widely described that Y-39983 is an inhibitor of Rho Kinase, thus capable of disrupting actin filament organization, changes in cell migration and stress fiber formation (Chrzanowska-Wodnicka and Burridge, 1996; Honjo et al., 2001; Tanna and Johnson, 2018).
These effects are also visible, though to a smaller extent, in the lower dose of treatment, highlighting additional features disrupted by Rho kinase inhibition. For instance, mass displacement is the measurement of the distance between the center of mass of the binary image and the center of mass of the grayscale image. A smaller value indicates that the density of pixels is evenly distributed in the shape, which is typical in rounded cells. For Y-39983-HCl at 0.2 mM this is true for both the AGP and mitochondria channels (Figure 3b). Similarly in the ER and DNA channel we can appreciate the rounder shape described by the Eccentricity measurement. Shape and intensity measures are much easier to validate by eye compared to others, but according to the algorithm there are more features that contribute to the shifts. A complete summary of scores and occurrences in each fold can be found in Supplementary File 2. In Y-39983-HCl together with shape and intensity related features, ER and mitochondria texture seem to play a role together with some features that measure the overlap and correlation between different channels. As Y-39983 is a ROCK inhibitor, the ability of the treated cells to deposit extracellular matrix is decreased, leading to less elastic and less motile cells, which can be easily interpreted from the highlighted features.

We saw similar effects with the drug Pemigatinib, with a shift towards a rounder shape (Supplementary Figure 3, Table 2), alongside the eccentricity and mass displacement importance in mitochondria and ER highlighted at both the lower and higher concentrations. This is in line with Pemigatinib's role as FGFR inhibitor, as the structure and function of the cytoskeleton is affected leading to a rounder appearance(Krook et al., 2020). Several mitochondrial texture features were highlighted suggesting an increase in mitochondrial fragmentation, a common feature encountered when proliferation is impaired(Greaves et al., 2012). The GSK-3β inhibitor, CP21R7, is known to disrupt a number of functions in normal cellular homeostasis as this enzyme has a diverse role in cellular physiology. Few descriptions of the effects of CP21R7 are found in the literature, but through our analysis pipeline we found significant differences in the mitochondria number and arrangement at both concentrations (Supplementary Figure 4, Supplementary File 2) , which is in line with previous reports that have shown CP21R7 can lead to the increase in expression of genes involved in metabolism(Lei et al., 2020) In addition, features related to the correlation of the AGP channel with ER, Mitochondria and DNA are relevant to characterize CP21R7. From the images it is visible that the overall size of the cells, highlighted in the AGP dye, is larger than the other inner compartments (either ER or Mitochondria) compared to the controls (Supplementary Figure 4).

This tool is designed specifically to address the scalability issues, both in terms of ease of deployment and minimization of computational power. As highly-automated laboratories continue to grow(Allen and Nilsson, 2021; Chandrasekaran et al., 2021; Gu et al., 2018; Shah et al., 2020), often culturing multiple different cell types across many different assays, it is crucial to have a tool that can easily be adapted and deployed on multiple datasets without laborious parameter tweaking. Thus, ScaleFEx^{SM} is generalizable to any type of cell and acquisition method natively. For ease of use, the number of user inputs needed are very limited, while preserving the possibility of tweaking important parameters as desired. This is in contrast to other tools that offer a high range of customization and different settings that on the one end offer more options (Lippeveld et al., 2022), but on the other make it much harder to deploy without error across large, diverse datasets.

In addition to its ease of use, ScaleFEx^{SM} can be deployed on any computer, without any specific requirements (other than a memory sufficient to load 5 images simultaneously). While it is a given that increased performance will be seen with the use of multiple fast CPUs, thanks to plate parallelization and ScaleFEx^{SM}'s slimmer nature, the run time is significantly reduced compared to typical tools such as Cell Profiler when run on similar resources (e.g. smaller desktop devices). As a future step, we are working towards implementing distributed computing options to make ScaleFEx^{SM} faster and cheaper to compute at a very large scale.
The dataset created for this study was intended to validate the method to confirm that features were accurately computed and that the analysis pipeline correctly reflected the characterization that could be visually confirmed by extracting the most meaningful images representing those features. Moving forward, we intend to use ScaleFEx^{SM} and the analysis pipeline described above to highlight subtle phenotypes that are not easily identifiable by eye such as complex disease phenotypes or shifts from drugs with mild effects. The ease of use and scalability of this tool makes it ideal for large drug, genetic and disease screens.

5. Conclusion

In this study, we describe the development of ScaleFEx^{SM}, a Python pipeline that extracts multiple fixed features in single cells and can be deployed across large high-content imaging datasets with low computational requirements. We demonstrate the utility of ScaleFEx^{SM} for separating different drug treatments with high accuracy, as well as identifying specific interpretable features underlying the decision process of the classifier. We were able to visually confirm the effects of the drugs on cells by extracting wells with the closest feature to the median of the population and comparing it with controls closest to their medians. This approach facilitated the direct link between drugs and their effects, which could be visually confirmed, thereby providing a reliable and credible tool compared to most AI-based dimensionality reduction tools that are more black-box style. Overall, this study highlights the potential of ScaleFEx^{SM} as a tool for identifying and characterizing drug effects on cellular populations, paving the way for the discovery of new therapeutics and disease treatments.

6. Acknowledgements

We thank members of the NYSCF Global Stem Cell Array® Team for cell expansion handling and method development that led to the final experiment. We are grateful to Sandra Capellera-Garcia, Corvis Richardson and Raeka Aiyar for manuscript input and review. We thank Bjarki Johannesson for the discussions and contribution at the start of the project. We are grateful to all the study participants who donated samples for this research.

7. Code and Data Availability

All code will be made available under Business Source License 1.1 and shared through a GitHub repository. Raw data will be downloadable upon publication of the manuscript from an Azure hosted bucket.

8. References

Allen, A., Nilsson, L., 2021. The Drug Factory: Industrializing How New Drugs Are Found. SLAS Discov. Adv. Life Sci. R D 26, 1076–1078. https://doi.org/10.1177/24725552211028124

Ando, D.M., McLean, C.Y., Berndl, M., 2017. Improving Phenotypic Measurements in High-Content Imaging Screens. https://doi.org/10.1101/161422

Antony, P.M.A., Kondratyeva, O., Mommaerts, K., Ostaszewski, M., Sokolowska, K., Baumuratov, A.S., Longhino, L., Poulain, J.F., Grossmann, D., Balling, R., Krüger, R., Diederich, N.J., 2020. Fibroblast mitochondria in idiopathic Parkinson's disease display morphological changes and enhanced resistance to depolarization. Sci. Rep. 10, 1569. https://doi.org/10.1038/s41598-020-58505-6

Ashdown, G.W., Dimon, M., Fan, M., Sánchez-Román Terán, F., Witmer, K., Gaboriau, D.C.A., Armstrong, Z., Ando, D.M., Baum, J., 2020. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. Sci. Adv. 6, eaba9338. https://doi.org/10.1126/sciadv.aba9338

Bray, M.-A., Gustafsdottir, S.M., Rohban, M.H., Singh, S., Ljosa, V., Sokolnicki, K.L., Bittker, J.A., Bodycombe, N.E., Dančík, V., Hasaka, T.P., Hon, C.S., Kemp, M.M., Li, K., Walpita, D., Wawer, M.J., Golub, T.R., Schreiber, S.L., Clemons, P.A., Shamji, A.F., Carpenter, A.E., 2017. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. GigaScience 6, giw014. https://doi.org/10.1093/gigascience/giw014

Bray, M.-A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., Carpenter, A.E., 2016. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat. Protoc. 11, 1757–1774. https://doi.org/10.1038/nprot.2016.105

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., Sabatini, D.M., 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 7, R100. https://doi.org/10.1186/gb-2006-7-10-r100

Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D., Carpenter, A.E., 2021. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat. Rev. Drug Discov. 20, 145–159. https://doi.org/10.1038/s41573-020-00117-w

Chrzanowska-Wodnicka, M., Burridge, K., 1996. Rho-stimulated contractility drives the formation of stress fibers and focal adhesions. J. Cell Biol. 133, 1403–1415. https://doi.org/10.1083/jcb.133.6.1403

Frye, R.E., Lionnard, L., Singh, I., Karim, M.A., Chajra, H., Frechet, M., Kissa, K., Racine, V., Ammanamanchi, A., McCarty, P.J., Delhey, L., Tippett, M., Rose, S., Aouacheria, A., 2021. Mitochondrial morphology is associated with respiratory chain uncoupling in autism spectrum disorder. Transl. Psychiatry 11, 1–17. https://doi.org/10.1038/s41398-021-01647-6

Giedt, R.J., Fumene Feruglio, P., Pathania, D., Yang, K.S., Kilcoyne, A., Vinegoni, C., Mitchison, T.J., Weissleder, R., 2016. Computational imaging reveals mitochondrial morphology as a biomarker of cancer phenotype and drug response. Sci. Rep. 6, 32985. https://doi.org/10.1038/srep32985

Greaves, L.C., Reeve, A.K., Taylor, R.W., Turnbull, D.M., 2012. Mitochondrial DNA and disease. J. Pathol. 226, 274–286. https://doi.org/10.1002/path.3028

Gu, H., Huang, X., Xu, J., Song, L., Liu, S., Zhang, X.-B., Yuan, W., Li, Y., 2018. Optimizing the method for generation of integration-free induced pluripotent stem cells from human peripheral blood. Stem Cell Res. Ther. 9, 163. https://doi.org/10.1186/s13287-018-0908-z

Honjo, M., Tanihara, H., Inatani, M., Kido, N., Sawamura, T., Yue, B.Y.J.T., Narumiya, S., Honda, Y., 2001. Effects of Rho-Associated Protein Kinase Inhibitor Y-27632 on Intraocular Pressure and Outflow Facility. Invest. Ophthalmol. Vis. Sci. 42, 137–144.

Hsieh, C.-H., Li, L., Vanhauwaert, R., Nguyen, K.T., Davis, M.D., Bu, G., Wszolek, Z.K., Wang, X., 2019. Miro1 Marks Parkinson's Disease Subset and Miro1 Reducer Rescues Neuron Loss in Parkinson's Models. Cell Metab. 30, 1131-1140.e7. https://doi.org/10.1016/j.cmet.2019.08.023

Kobayashi, H., Cheveralls, K.C., Leonetti, M.D., Royer, L.A., 2022. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Nat. Methods 19, 995–1003. https://doi.org/10.1038/s41592-022-01541-z

Krook, M.A., Lenyo, A., Wilberding, M., Barker, H., Dantuono, M., Bailey, K.M., Chen, H.-Z., Reeser, J.W., Wing, M.R., Miya, J., Samorodnitsky, E., Smith, A.M., Dao, T., Martin, D.M., Ciombor, K.K., Hays, J., Freud, A.G., Roychowdhury, S., 2020. Efficacy of FGFR inhibitors and combination therapies for acquired resistance in FGFR2-fusion cholangiocarcinoma. Mol. Cancer Ther. 19, 847–857. https://doi.org/10.1158/1535-7163.MCT-19-0631

Lei, Z., Yang, L., Yang, Y., Yang, J., Niu, Z., Zhang, X., Song, Q., Lei, Y., Wu, H., Guo, J., 2020. Activation of Wnt/β-catenin pathway causes insulin resistance and increases lipogenesis in HepG2 cells via regulation of endoplasmic reticulum stress. Biochem. Biophys. Res. Commun. 526, 764–771. https://doi.org/10.1016/j.bbrc.2020.03.147

Lippeveld, M., Peralta, D., Filby, A., Saeys, Y., 2022. A scalable, reproducible and open-source pipeline for morphologically profiling image cytometry data. https://doi.org/10.1101/2022.10.24.512549

Mullard, A., 2019. Machine learning brings cell imaging promises into focus. Nat. Rev. Drug Discov. 18, 653–655. https://doi.org/10.1038/d41573-019-00144-2

Narayanaswamy, A., Xia, C., Ando, M., Venugopalan, S., Qian, W., 2020. GAN-Mediated Batch Equalization. bioRxiv.

Owens, J.D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A.E., Purcell, T.J., 2007. A Survey of General-Purpose Computation on Graphics Hardware. Comput. Graph. Forum 26, 80–113. https://doi.org/10.1111/j.1467-8659.2007.01012.x

Paull, D., Sevilla, A., Zhou, H., Hahn, A.K., Kim, H., Napolitano, C., Tsankov, A., Shang, L., Krumholz, K., Jagadeesan, P., Woodard, C.M., Sun, B., Vilboux, T., Zimmer, M., Forero, E., Moroziewicz, D.N., Martinez, H., Malicdan, M.C.V., Weiss, K.A., Vensand, L.B., Dusenberry, C.R., Polus, H., Sy, K.T.L., Kahler, D.J., Gahl, W.A., Solomon, S.L., Chang, S., Meissner, A., Eggan, K., Noggle, S.A., 2015. Automated, high-throughput derivation, characterization and differentiation of induced pluripotent stem cells. Nat. Methods 12, 885–892. https://doi.org/10.1038/nmeth.3507

Schiff, L., Migliori, B., Chen, Y., Carter, D., Bonilla, C., Hall, J., Fan, M., Tam, E., Ahadi, S., Fischbacher, B., Geraschenko, A., Hunter, C.J., 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, E.A., Andres-Martin, L., Pritchard, O., Duren-Lubanski, A., Daigavane, A., Reggio, K., NYSCF Global Stem Cell Array® Team, Nelson, P.C., Frumkin, M., Solomon, S.L., Bauer, L., Aiyar, R.S., Schwarzbach, E., Noggle, S.A., Monsma, F.J., Paull, D., Berndl, M., Yang, S.J., Johannesson, B., 2022. Integrating deep learning and unbiased automated high-content screening to identify complex disease signatures in human fibroblasts. Nat. Commun. 13, 1590. https://doi.org/10.1038/s41467-022-28423-4

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682. https://doi.org/10.1038/nmeth.2019

Shah, K., Bedi, R., Rogozhnikov, A., Ramkumar, P., Tong, Z., Rash, B., Stanton, M., Sorokin, J., Apaydin, C., Batarse, A., Bergamaschi, J., Blattner, R., Brown, S., Bosshardt, A., Castrillo, C., Dang, B., Drusinsky, S., Enriquez, L., Grayson, D., Hilliard, J., Hsu, P.-K., Johnson, C., Jones, R., Lash, A., Lee, C.-Y., Li, K., McKay, A., Mount, E., Nicola, J., Oumzil, I., Paek, J., Pascoe, D., Piepho, A., Poust, S., Quang, D., Schultz, M., Sims, J., Taylor, P., Treiman, G., Wueseke, O., Young, N., Pollen, A., Flanzer, D., Chao, D., Skibinski, G., Kato, S., Escola, G.S., 2020. Optimization and scaling of patient-derived brain organoids uncovers deep phenotypes of disease. https://doi.org/10.1101/2020.08.26.251611

Shi, L., Liu, W., Zhang, H., Xie, Y., Wang, D., 2012. A survey of GPU-based medical image computing techniques. Quant. Imaging Med. Surg. 2, 188–206. https://doi.org/10.3978/j.issn.2223-4292.2012.08.02

Simbaña, M., n.d. Computer Architecture A Quantitative Approach (5th edition).

Tanna, A.P., Johnson, M., 2018. Rho Kinase Inhibitors as a Novel Treatment for Glaucoma and Ocular Hypertension. Ophthalmology 125, 1741–1756. https://doi.org/10.1016/j.ophtha.2018.04.040

Tegtmeyer, M., Arora, J., Asgari, S., Cimini, B.A., Peirent, E., Liyanage, D., Way, G., Weisbart, E., Nathan, A., Amariuta, T., Eggan, K., Haghighi, M., McCarroll, S.A., Carpenter, A.E., Singh, S., Nehme, R., Raychaudhuri, S., 2023. High-dimensional phenotyping to define the genetic basis of cellular morphology. https://doi.org/10.1101/2023.01.09.522731

Venugopalan, S., Narayanaswamy, A., Yang, S., Geraschenko, A., Lipnick, S., Makhortova, N., Hawrot, J., Marques, C., Pereira, J., Brenner, M., Rubin, L., Wainger, B., Berndl, M., 2020. It's easy to fool yourself: Case studies on identifying bias and confounding in bio-medical datasets. https://doi.org/10.48550/arXiv.1912.07661

Way, G.P., Kost-Alimova, M., Shibue, T., Harrington, W.F., Gill, S., Piccioni, F., Becker, T., Shafqat-Abbasi, H., Hahn, W.C., Carpenter, A.E., Vazquez, F., Singh, S., 2021. Predicting cell health phenotypes using image-based morphology profiling. Mol. Biol. Cell 32, 995–1005. https://doi.org/10.1091/mbc.E20-12-0784

Weisbart, E., Cimini, B.A., 2022. Distributed-Something: scripts to leverage AWS storage and computing for distributed workflows at scale. https://doi.org/10.48550/arXiv.2210.01073

Yang, S.J., Lipnick, S.L., Makhortova, N.R., Venugopalan, S., Fan, M., Armstrong, Z., Schlaeger, T.M., Deng, L., Chung, W.K., O'Callaghan, L., Geraschenko, A., Whye, D., Berndl, M., Hazard, J., Williams, B., Narayanaswamy, A., Ando, D.M., Nelson, P., Rubin, L.L., 2019. Applying Deep Neural Network Analysis to High-Content Image-Based Assays. SLAS Discov. Adv. Life Sci. R D 24, 829–841. https://doi.org/10.1177/2472555219857715

Zaritsky, A., Jamieson, A.R., Welf, E.S., Nevarez, A., Cillay, J., Eskiocak, U., Cantarel, B.L., Danuser, G., 2021. Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma. Cell Syst. 12, 733-747.e6. https://doi.org/10.1016/j.cels.2021.05.003

9. Figure captions

Figure 1. An overview of the complete experimental pipeline.
a) Experimental pipeline: Using a collection of skin fibroblasts from a cohort of 20 donors, we generated 5, reproducibly seeded, 384 well plates for compound profiling. We applied 3 drugs at 2 different concentrations alongside DMSO controls via automated dispensing. After 72 hours of growth in the presence of the drugs, the cells were fixed, Cell-Painted and imaged using an automated microscope.

b) Pre-processing and feature extraction: Following imaging, initial preprocessing was performed before each cell was segmented and measurements related to Shape, Texture, Granularity, Intensity, Concentric rings, Colocalization, Mitochondria and RNA specific were computed for each channel (here the total count for a Cell Painting assay is shown).
c) Analysis and shift characterization: The resulting vector was then used to evaluate the effects of the compounds as compared to the DMSO controls using machine learning algorithms.

Figure 2. Analysis and Feature Extraction Overview.
a) Following normalization of the measurements to values between 0 and 1 and the removal of confounders by subtraction of the mean of the confounder, using well-leveral averages, results are visualized using UMAPs. First, features with low variance are removed, then using recursive feature elimination (RFE) unimportant features were iteratively removed, leaving behind the most impactful features contributing to prediction scores. Highly correlated features are removed, before the remaining features are ranked in order to identify the most promising ones. Finally, we extract representative images of the features of interest and compare them to their ranking.
b) UMAP of the feature vector color-coded by drug and condition. The values were corrected for plate, donor, column and row effects, and normalized between 0 and 1.
c) Distribution of Binary predictions of drug VS control using a Linear Regression model and 5 CV folds, where in each fold the training is performed on all the plates but one held out for testing. The model is trained to classify wells treated with each drug iteratively and DMSO controls. The values were corrected for plate, donor, column and row effects, and normalized between 0 and 1. Boxes represent median, upper and lower quartiles, the whiskers represent the standard deviation, and the diamonds the outliers.
d) Example of feature summarization, by channel and category for cells treated with Y-3998-HCl at 0.2 mM. In the heatmap, the darker the color, the higher the number of features. The number states the actual number of feature per class and channel. The half numbers are for those features with double class or double channel.

Figure 3 - Value distribution and image visualization of the most important features to correctly classify Y-39983 HCl.
a) Features highlighted at 1.0 mM and (b) 0.2 mM. The pictures depict the wells with the closest value to the mean of the feature. In interest of space, we only visualize 9 out of the 49 total tiles that make a well. The feature distribution values are all normalized between 0 and 1. Boxes represent median, upper and lower quartiles, the whiskers represent the standard deviation, and the diamonds the outliers. **** = $p < 0.0001$.

Supplementary Figure 1
a) Example of segmentation masks for each channel in a single cell and additional segmentation masks used for RNA and Mitochondria (skeleton only)
b) Illustration to describe the additional features related to Mitochondria and RNA that are computed. An additional mask was first computed to segment the sub-structures of the channel. For RNA, we then calculated the values related to number of organelles and volume, shown in the mask (bottom right, panel b - RNA Mask).
For mitochondria, we first compute measurements related to volume on the mask (purple portion in the schematics), before computing the mask's skeleton and subsequently branches, endpoints, total length and number.

Supplementary Figure 2
a) UMAP of the feature vector color-coded by drug and condition. The values were normalized between 0 and 1.
b) Distribution of Binary predictions of drug VS control using a Linear Regression model and 5 CV folds, where in each fold the training is performed on all the plates, but one held out for testing. The

model is trained to classify wells treated with each drug iteratively and DMSO controls. The values were normalized between 0 and 1. Boxes represent median, upper and lower quartiles, the whiskers represent the standard deviation, and the diamonds the outliers.

Supplementary Figure 3

Value distribution and image visualization of the most important features to classify Pemigatinib at a) 0.2 mM and b) 1.0 mM. The pictures depict the wells with the closest value to the mean of the feature. In interest of space, we only visualize 9 out of the 49 total tiles that make a well.
In the heatmaps, the darker the color, the higher the number of features. The number states the actual number of features per class and channel. The half numbers are for those features with double class or double channel. The feature distribution values are all normalized between 0 and 1. Boxes represent median, upper and lower quartiles, the whiskers represent the standard deviation, and the diamonds the outliers. **** = $p < 0.0001$.

Supplementary Figure 4

Value distribution and image visualization of the most important features to correctly classify CP21R7 at a) 0.2 mM and b) 1.0 mM. The features of interest are shown on top of the images. The pictures depict the wells with the closest value to the mean of the feature. In interest of space, we only visualize 9 out of the 49 total tiles that make a well.
In the heatmaps, the darker the color, the higher the number of features. The number states the actual number of features per class and channel. The half numbers are for those features with double class or double channel. The feature distribution values are all normalized between 0 and 1. Boxes represent median, upper and lower quartiles, the whiskers represent the standard deviation, and the diamonds the outliers. **** = $p < 0.0001$.

10. Table captions

Table 1

Summary of all the features computed by ScaleFEx. Each column specifies feature name, category, the formula used to compute the feature, the package (if any) used for the computation, an explanation of their meaning and if the feature is computed by channel.
In this table we removed the channel name and the rounds of features (but one) that went through multiple iterations to avoid redundant information in the table.

Table 2

Summary of the best scoring features for each drug. Each sheet contains the features relative to the drug and concentration specified in the sheet-name. The features are sorted by the total counts of features making the cut for each plate and the sum of the coefficients. In the remaining columns there are the coefficients for each of the plates

Figure 1

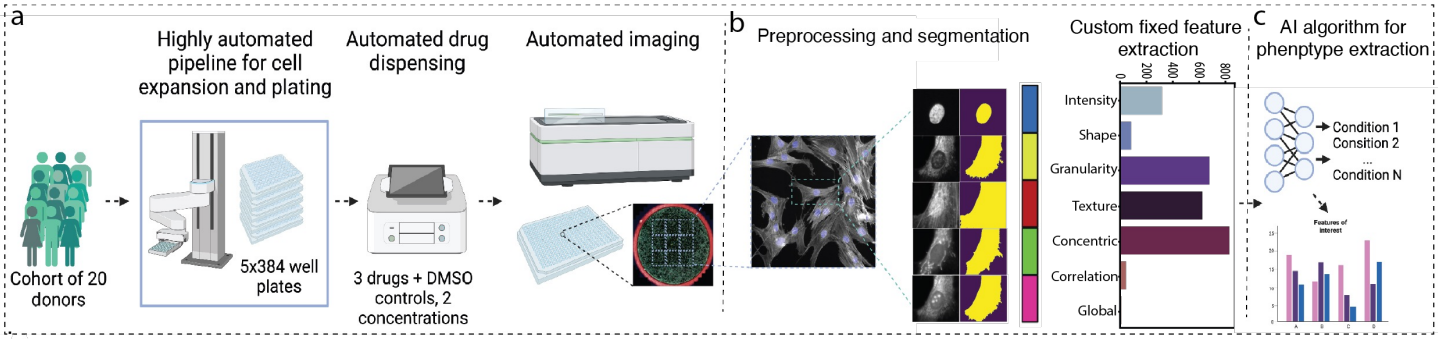

a
Highly automated pipeline for cell expansion and plating
Automated drug dispensing
Automated imaging
Cohort of 20 donors
5x384 well plates
3 drugs + DMSO controls, 2 concentrations
b
Preprocessing and segmentation
Custom fixed feature extraction
0
200
400
600
800
Intensity
Shape
Granularity
Texture
Concentric
Correlation
Global
c
AI algorithm for phenptype extraction
Condition 1
Consition 2
...
Condition N
Features of interest

Figure 2

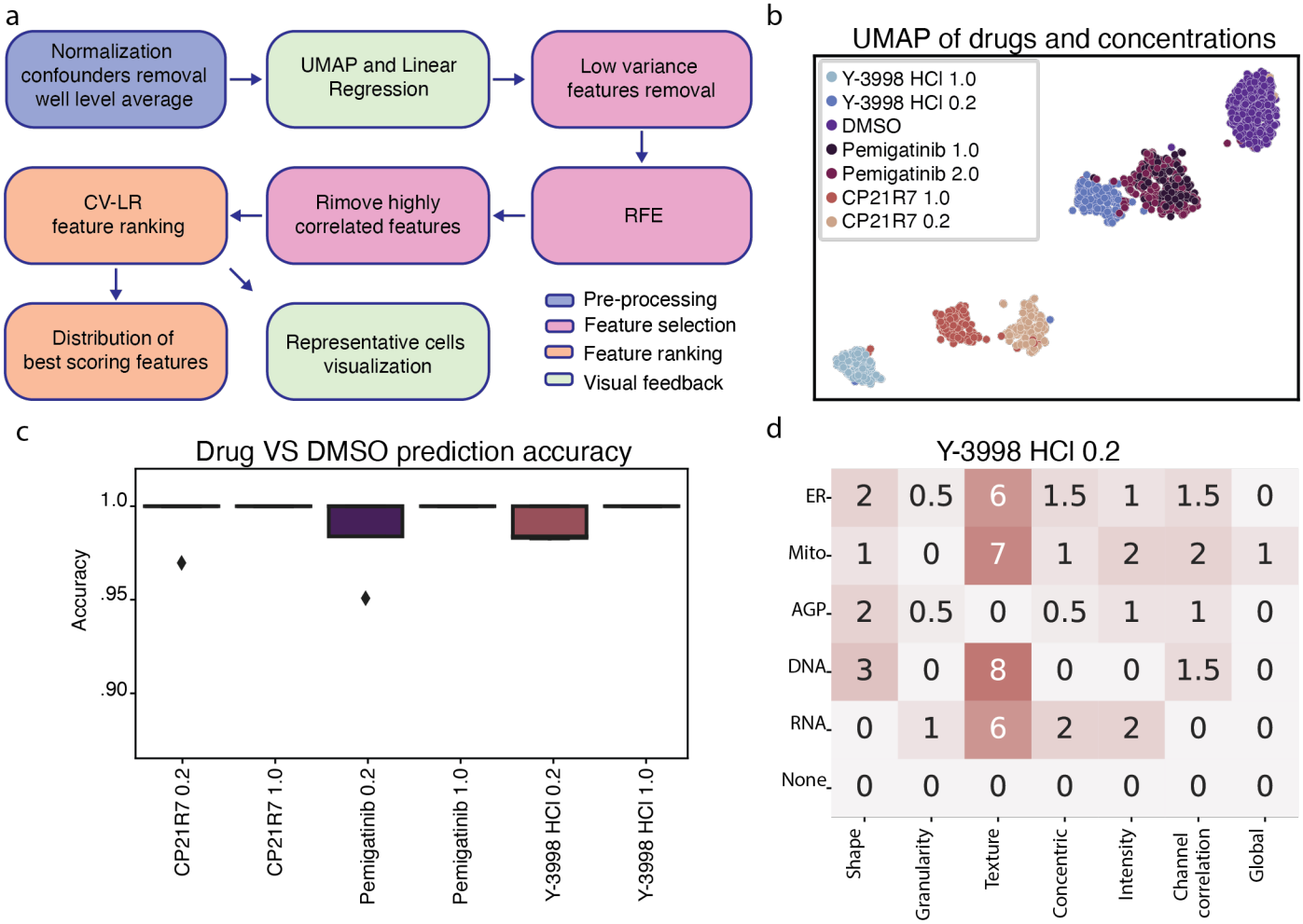
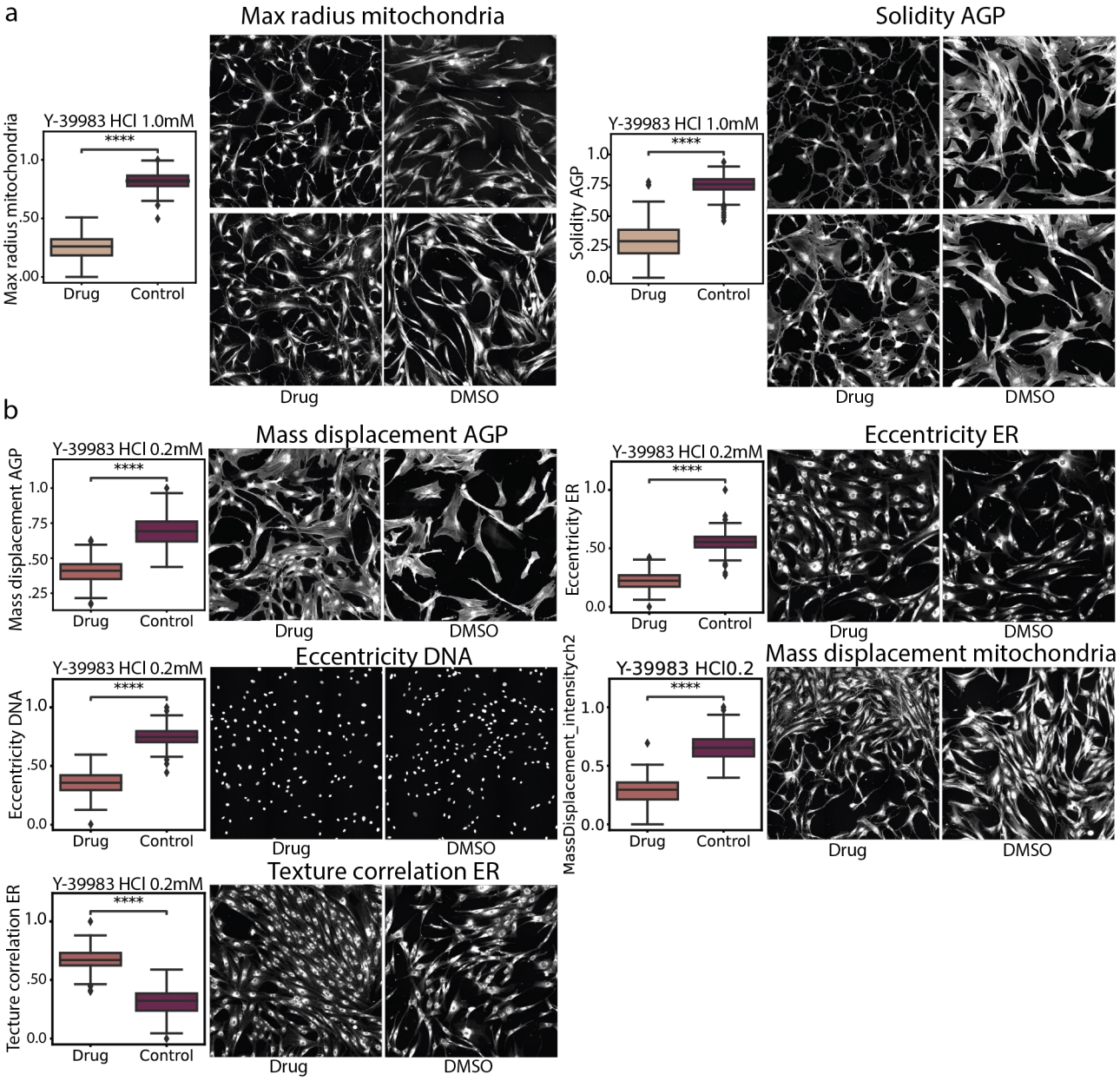
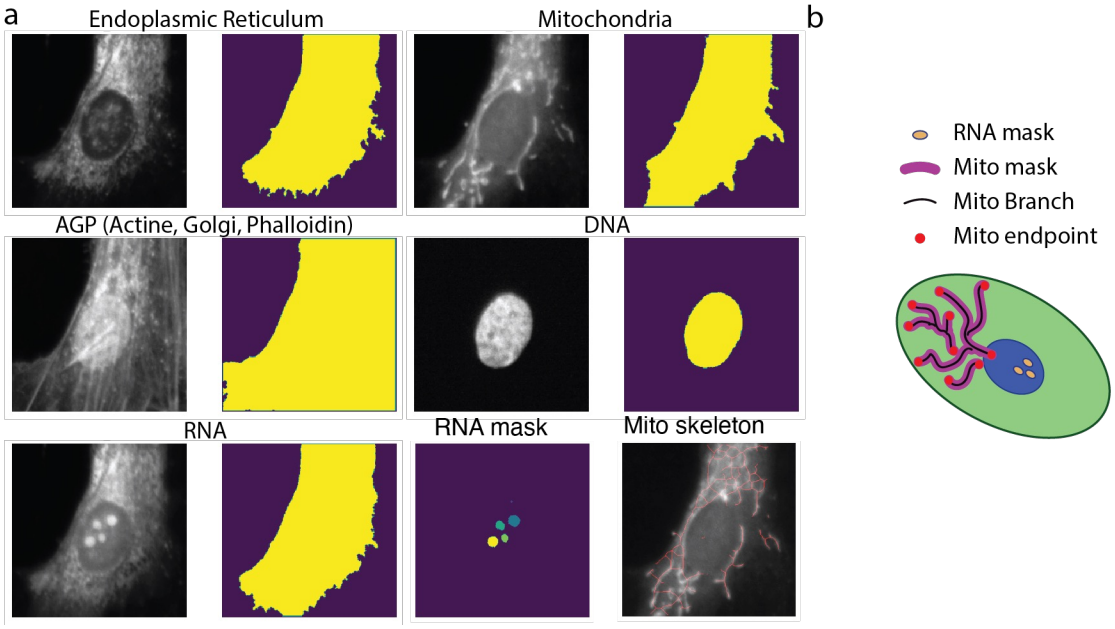

a
Normalization confounders removal well level average
UMAP and Linear Regression
Low variance features removal
RFE
Rimove highly correlated features
CV-LR feature ranking
Distribution of best scoring features
Representative cells visualization
Pre-processing
Feature selection
Feature ranking
Visual feedback
b
UMAP of drugs and concentrations
Y-3998 HCl 1.0
Y-3998 HCl 0.2
DMSO
Pemigatinib 1.0
Pemigatinib 2.0
CP21R7 1.0
CP21R7 0.2
c
Drug VS DMSO prediction accuracy
Accuracy
1.0
.95
.90
CP21R7 0.2
CP21R7 1.0
Pemigatinib 0.2
Pemigatinib 1.0
Y-3998 HCl 0.2
Y-3998 HCl 1.0
d
Y-3998 HCl 0.2
ER- 2 0.5 6 1.5 1 1.5 0
Mito- 1 0 7 1 2 2 1
AGP- 2 0.5 0 0.5 1 1 0
DNA- 3 0 8 0 0 1.5 0
RNA- 0 1 6 2 2 0 0
None- 0 0 0 0 0 0 0
Shape
Granularity
Texture
Concentric
Intensity
Channel correlation
Global

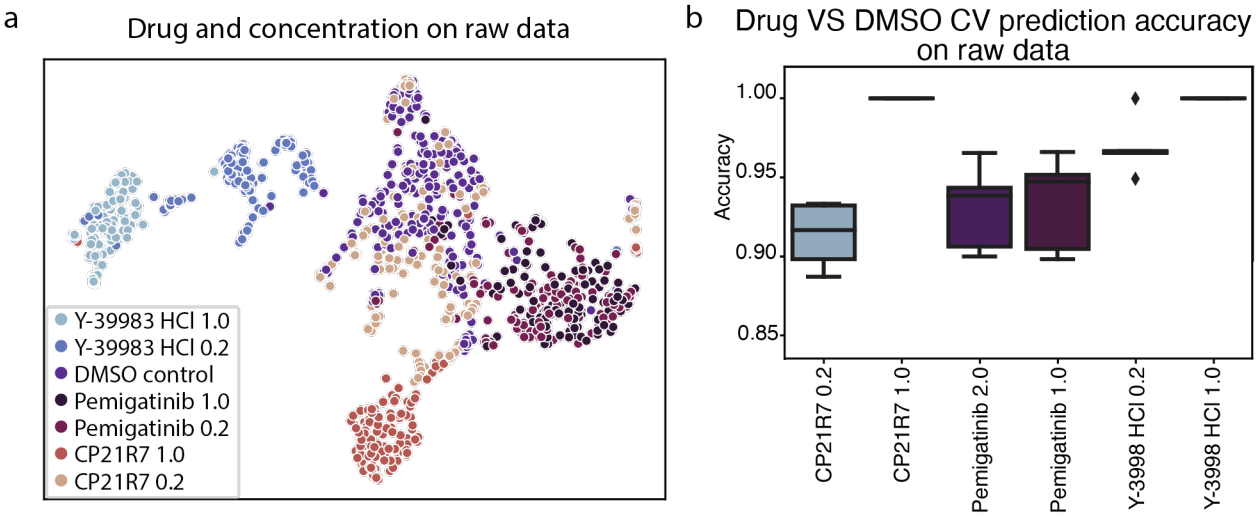
Figure 3


a
Max radius mitochondria
Y-39983 HCl 1.0mM
****
Max radius mitochondria
1.0
.50
.00
Drug
Control
Drug
DMSO
Solidity AGP
Y-39983 HCl 1.0mM
****
Solidity AGP
1.0
.75
.50
.25
0.0
Drug
Control
Drug
DMSO
b
Mass displacement AGP
Y-39983 HCl 0.2mM
****
Mass displacement AGP
1.0
.75
.50
.25
Drug
Control
Drug
DMSO
Eccentricity ER
Y-39983 HCl 0.2mM
****
Eccentricity ER
1.0
.50
0.0
Drug
Control
Drug
DMSO
Eccentricity DNA
Y-39983 HCl 0.2mM
****
Eccentricity DNA
1.0
.50
0.0
Drug
Control
Drug
DMSO
Mass displacement mitochondria
Y-39983 HCl0.2
****
MassDisplacement_intensitych2
1.0
0.5
0.0
Drug
Control
Drug
DMSO
Texture correlation ER
Y-39983 HCl 0.2mM
****
Tecture correlation ER
1.0
.50
0.0
Drug
Control
Drug
DMSO

Supplementary Figure 1

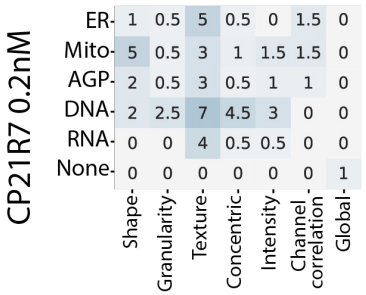


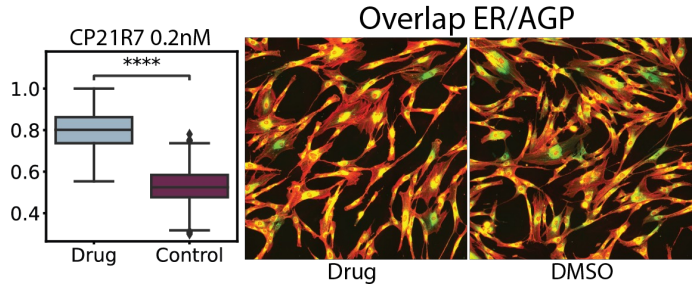
Supplementary Figure 2

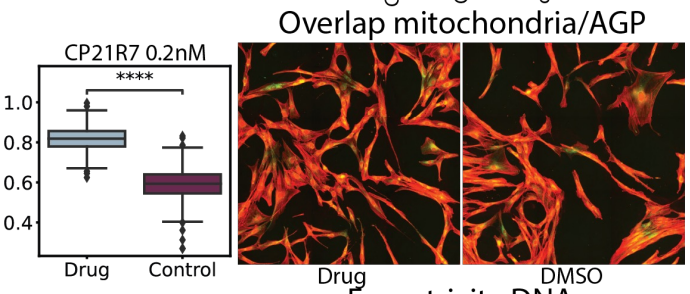


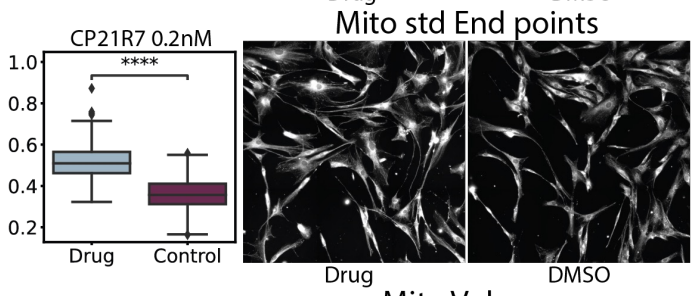
Supplementary Figure 3

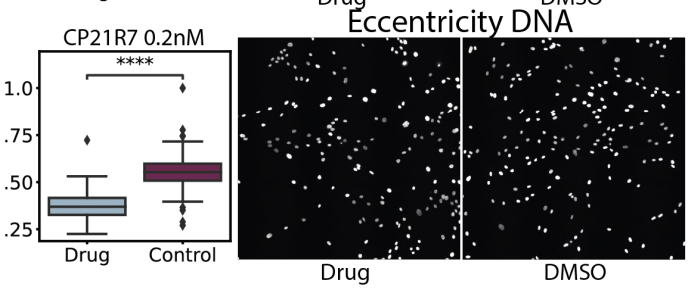
a

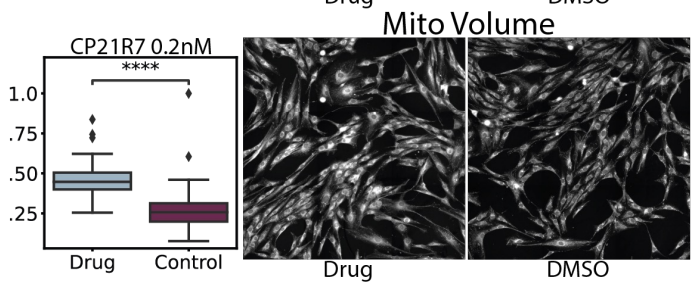




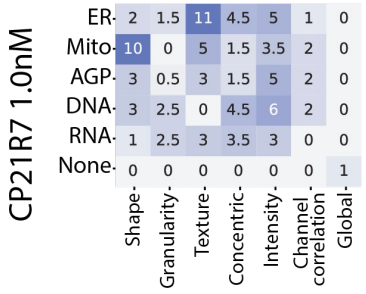


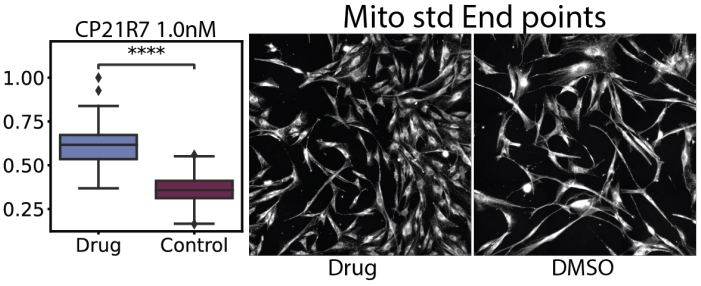


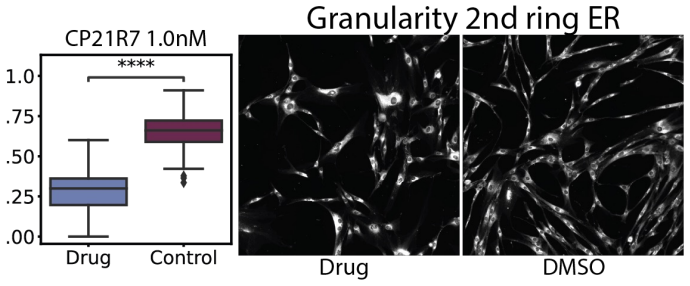


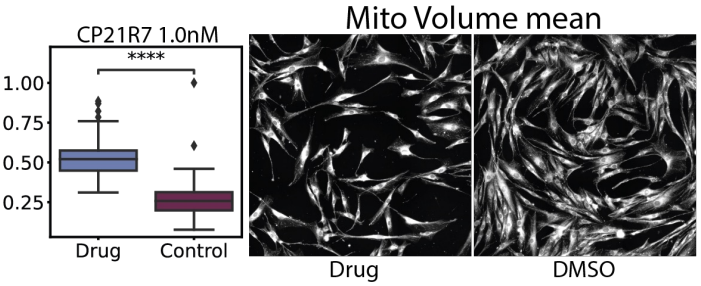


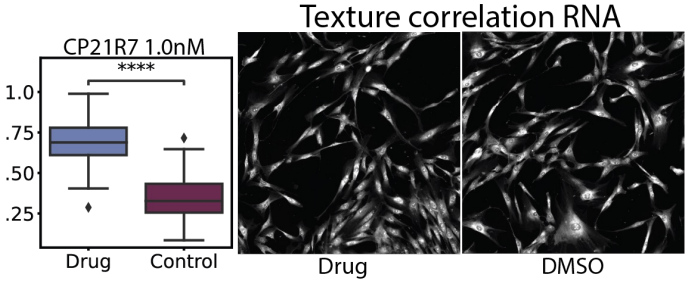
b

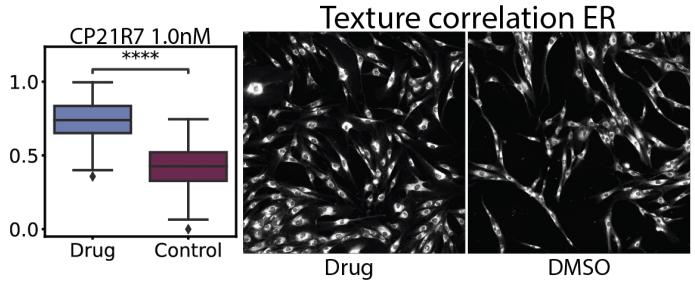






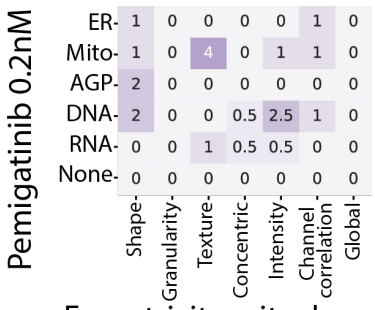


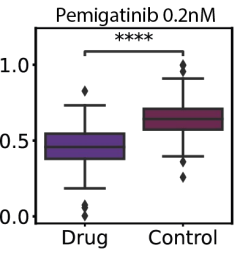




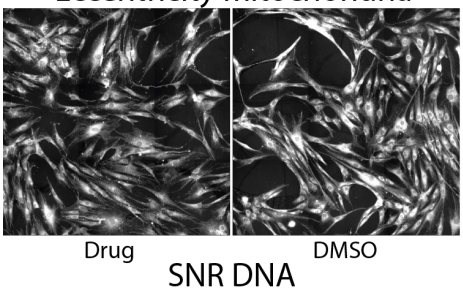
Supplementary Figure 4

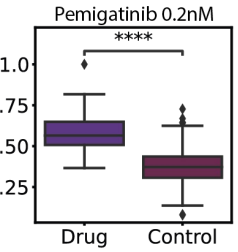
a



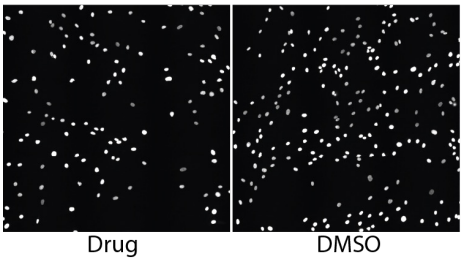


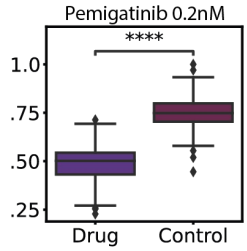
Eccentricity mitochondria



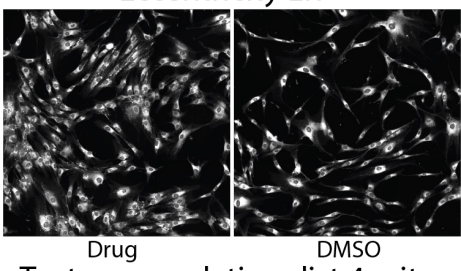


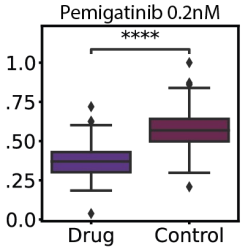
SNR DNA



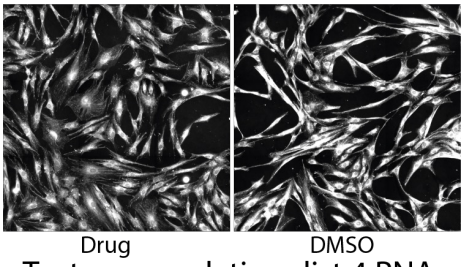


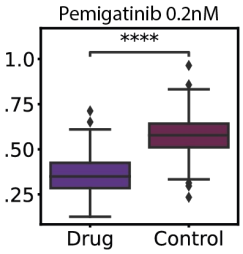
Eccentricity ER



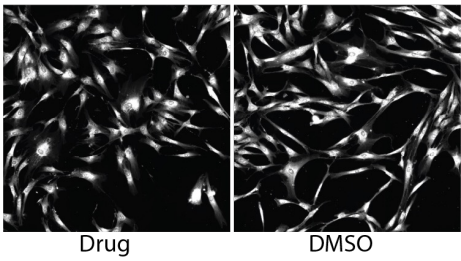


Texture correlation dist 4 mito

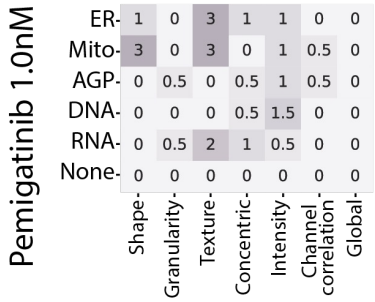


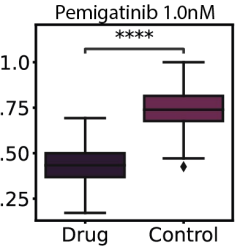


Texture correlation dist 4 RNA

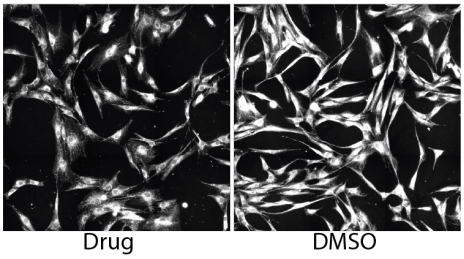


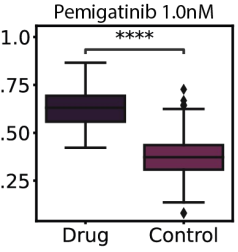
b



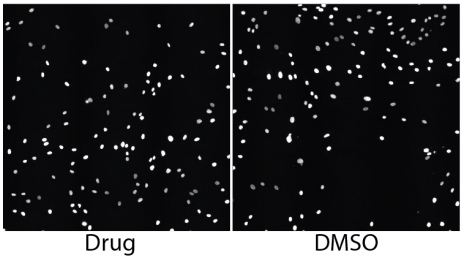


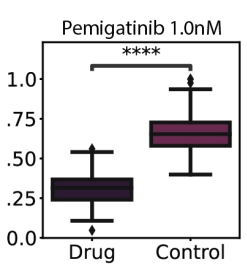
Texture correlation dist 5 mito



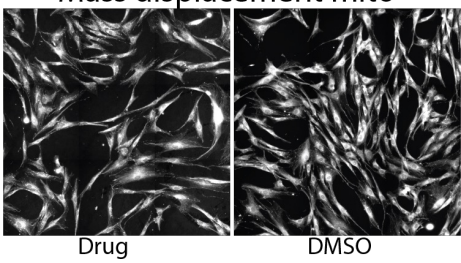


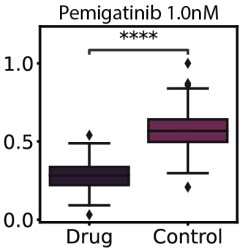
SNR DNA



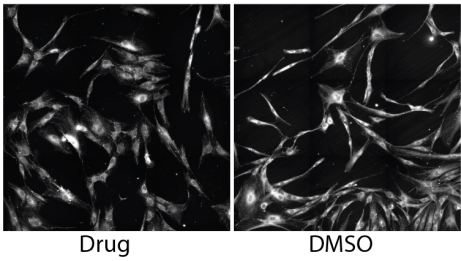


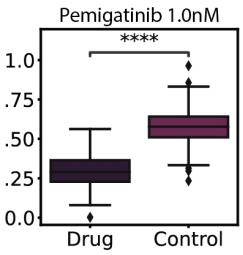
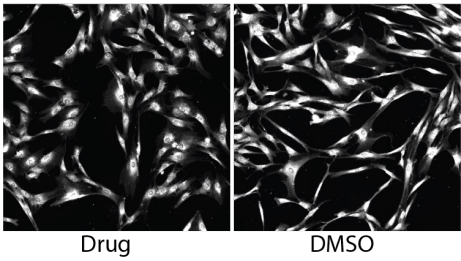
Mass displacement mito





Texture correlation dist 4 mito





Texture correlation dist 4 DNA