

Predicting cell health phenotypes using image-based morphology profiling

Gregory Way, Maria Kost-Alimova, Tsukasa Shibue, William Harrington, Stanley Gill, Federica Piccioni, Tim Becker, Hamdah Shafqat-Abbasi, William Hahn, Anne Carpenter, Francisca Vazquez, and Shantanu Singh

Corresponding author(s): Gregory Way, Broad Institute of MIT and Harvard

Review Timeline:

Submission Date:

2020-12-22

Accepted:

2021-01-11

Editor-in-Chief: Matthew Welch

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

RE: Manuscript #E20-12-0784

TITLE: "Predicting cell health phenotypes using image-based morphology profiling"

Dear Dr. Way:

I am pleased to accept your manuscript for publication in Molecular Biology of the Cell.

Sincerely,
Alexander Mogilner
Monitoring Editor
Molecular Biology of the Cell

Dear Dr. Way:

Congratulations on the acceptance of your manuscript.

A PDF of your manuscript will be published on MBoC in Press, an early release version of the journal, within 10 days. The date your manuscript appears at www.molbiolcell.org/toc/mboc/0/0 is the official publication date. Your manuscript will also be scheduled for publication in the next available issue of MBoC.

Within approximately four weeks you will receive a PDF page proof of your article.

Would you like to see an image related to your accepted manuscript on the cover of MBoC? Please contact the MBoC Editorial Office at mboc@ascb.org to learn how to submit an image.

Authors of Articles and Brief Communications are encouraged to create a short video abstract to accompany their article when it is published. These video abstracts, known as Science Sketches, are up to 2 minutes long and will be published on YouTube and then embedded in the article abstract. Science Sketch Editors on the MBoC Editorial Board will provide guidance as you prepare your video. Information about how to prepare and submit a video abstract is available at www.molbiolcell.org/science-sketches. Please contact mboc@ascb.org if you are interested in creating a Science Sketch.

We are pleased that you chose to publish your work in MBoC.

Sincerely,

Eric Baker
Journal Production Manager
MBoC Editorial Office
mbc@ascb.org

We were delighted by the Review Commons process and the sensible and constructive reviews, which were straightforward to address with new analyses and adjustments to our text. In our response below, we have improved the manuscript in three primary ways: 1) Improved rationale clarity and Cell Health assay reproducibility; 2) a greatly expanded section interrogating machine learning model robustness; and 3) a restructuring of the Drug Repurposing Hub cell health model application section, which includes a new validation analysis. We have also performed many additional minor improvements, as suggested.

Our manuscript introduced two ideas: 1) a new microscopy assay to measure 70 different cell health phenotypes and 2) a machine learning approach to predict these phenotypes using Cell Painting data. Cell Painting is a different microscopy assay that is unbiased (meaning that it does not target any specific phenotype), inexpensive, and high-throughput, and importantly Cell Painting data sets are becoming widely available. Overall, we show that, given Cell Painting data, many useful cell health annotations can be provided that are intuitive, informative, and useful for cell biologists. These annotations enable a richer understanding of perturbation mechanisms and provide a multivariate readout of cell health state, and come with no additional experimental costs.

In the manuscript, we presented two new assays to measure and predict various cell health indicators. These indicators included DNA damage, cell proliferation, cell cycle stage stalling, reactive oxygen species, and apoptosis. We collected data from these assays and trained machine learning models to predict cell health indicators using existing public Cell Painting data. Lastly, we applied these models to an external Cell Painting dataset of drug repurposing compounds, and validated predictions from four models. We made cell health predictions for all 1,571 of these compounds across 6 dose points and provided these predictions in an easy to navigate web application (<https://broad.io/cell-health-app>) as an example of what could be done for larger future Cell Painting datasets - for example, our laboratory is leading a Consortium creating Cell Painting data for 150,000 chemical and genetic perturbations.

Reviewer #1 (Evidence, reproducibility and clarity (Required)):

The submitted manuscript entitled 'Predicting cell health phenotypes using image-based morphology profiling' (RC-2020-00394) by Way et al. presents a set of seven dyes/staining (as two separate panels) to microscopically screen cell viability. For automatic classification a training/test set of 119 CRISPR (approximately 2 sgRNAs per gene) perturbations on 3 cancer cell lines were generated (lung A549, ovarian ES2, lung HCC44). After segmentation of cell nuclei a set of morphological cell measurements were extracted from each perturbation (total

952 features). The nature of these feature spanning cell cycle and viability phenotypes, enabled the authors to define 70 different phenotype classes, which are used to model a classifier by elastic linear regression. Specific definitions (cell cycle and ROS) were partly predicted/validated in an independent existing image data set (Drug Repurposing Hub project). The data is available as web-based application/visualization and the supplementary method is well described.

We thank the reviewer for their constructive comments and helpful feedback.

There is one subtle point that is worth raising given this description: The images we use to measure the cell cycle and viability phenotypes (two different staining panels in the Cell Health assays) are not the same images we use to extract morphology measurements (Cell Painting assay). This lack of connection, which is based on a light wavelength limitation present in all microscopes that limits the number of stains in a single assay, prevents us from developing a method that analyzes the same cells across the three assays. This distinction will become important later in the review, and we have made specific changes in the manuscript to increase clarity.

****Major concerns:****

(1)The only fundamental argument of this manuscript not to apply state-of-the-art deep learning (DL) machine-learning (mentioned in McCain et al. 2018), which does not require segmentation, feature extraction, abstraction, manual gating is the 'interpretability' of the predictions. However, performance, precision, scalability (by modern GPUs) with DL should clearly outperform 'manual' regression models. All recent machine vision benchmarks in microscopy confirm this, but also clearly shows 'real world' translational applications, e.g.

<https://www.nature.com/articles/s43018-020-0085-8>,

<https://www.biorxiv.org/content/10.1101/2020.07.02.183814v1.full.pdf>,

In other words, the presented methodology is not compared to DL, and is not convincing in terms of interpretability benefits.

(We've copied a similar critique from *Significance section* from Reviewer #1 in order to reduce redundancy) The author/co-authors have been instrumental/pioneered with their past work on cell-based image processing (CellProfiler software), but the presented methodology is simply outdated. Therefore, a revision towards a comparison and benchmarking with DL will also not help.

Ref (DL with MIL): <https://academic.oup.com/bioinformatics/article/32/12/i52/2288769>

We agree that deep learning approaches are exciting; much of our laboratory's work focuses on their application (see <https://doi.org/10.1073/pnas.2001227117>, <https://doi.org/10.1038/s41592-019-0612-7>; <https://doi.org/10.1002/cyto.a.23863>, <https://doi.org/10.1109/CVPR.2018.00970>), and we agree that they are likely to outperform simpler regression models trained using so-called hand-engineered features. We thank the reviewer for highlighting our failure to accurately and fully describe our rationale.

We intentionally did not use deep learning for this problem given (a) data limitations (b) the primary goal of the manuscript, which is to demonstrate feasibility.

Data limitations. There is no mechanism to link the cells of the assays (Cell Health and Cell Painting) together, which greatly reduces the available sample size. In the two referenced manuscripts, which each propose an exciting approach, the dataset is much larger (~17,000 and ~1,000 images respectively). Our dataset is only 357 perturbations that can only be linked between assays at the perturbation level rather than a single-cell level. Therefore, a deep learning approach is likely to produce models that don't generalize to other datasets. Furthermore, reviewer 3 commented in favor of the approach we presented: "Using elastic net regression models is well-suited to the problem due to the low number of observations."

Primary goal of the manuscript is to demonstrate feasibility. In addition, the primary goal of the manuscript is to add cell health annotations as functional readouts to perturbations. Our aim was to demonstrate feasibility of predicting cell health states, not to optimize performance. Optimizing performance would require collecting much more data, or developing new deep learning or data collection methods to account for the lack of matched single cell readouts.

To make this rationale more clear and concise, we have made the following changes in the manuscript:

In the first paragraph of page 3, we make some minor contextual updates ("To demonstrate proof of concept, we collected a small pilot dataset of 119 CRISPR knockout perturbations...") and replaced "We used simple machine learning methods, which are relatively easy to interpret compared to deep learning" with:

We used simple machine learning methods instead of a deep learning approach because of our limited sample size of 119 perturbations and the inability to increase the sample size by linking single cell measurements across assays.

We have also amended the Conclusions section to emphasize our primary goal and note possible deep learning extensions as future directions. The Conclusions now reads:

We have demonstrated feasibility that information in Cell Painting images can predict many different Cell Health indicators even when trained on a small dataset. The results motivate

collecting larger datasets for training, with more perturbations and multiple cell lines. These new datasets would enable the development of more expressive models, based on deep learning, that can be applied to single cells. Including orthogonal imaging markers of CRISPR infection would also enable us to isolate cells with expected morphologies. More data and better models would improve the performance and generalizability of Cell Health models and enable annotation of new and existing large-scale Cell Painting datasets with important mechanisms of cell health and toxicity.

(2)One aforementioned point of the methodology is cryptically/not described: Why it should be less expensive compared with other (which?) approaches (see introduction)?

We thank the reviewer for bringing up this point. We believe that part of this confusion stems from a slight misunderstanding about how images from the three assays (two Cell Health and one Cell Painting) are collected. The Cell Health assays are two distinct *panels* of targeted reagents that are separately prepared as two physically distinct assays. The Cell Painting assay is already an established assay used by many labs and companies around the world to mark cell morphology in an unbiased and relatively cheap way. We are comparing the expenses between the two Cell Health assays vs. the Cell Painting assay.

We believe that this misunderstanding likely results from our somewhat cryptic and inconsistent language when describing the Cell Health assays in the abstract and introduction. We've updated the third sentence of the abstract from "We developed two customized microscopy assays that use seven reagents to measure 70 specific cell health phenotypes..." to now read:

We developed two customized microscopy assays, one using four targeted reagents and the other three targeted reagents, to collectively measure 70 specific cell health phenotypes including proliferation, apoptosis, reactive oxygen species (ROS), DNA damage, and cell cycle stage.

For consistency, we have also updated the penultimate paragraph in the introduction to now read:

To do this, we first developed two customized microscopy assays, which collectively report on 70 different cell health indicators via a total of seven reagents applied in two reagent panels. Collectively, we call these assays "Cell Health".

With these clarifications in mind, we believe that the question of comparing monetary costs is more clear. We are comparing the costs of the targeted reagents in the two Cell Health assays to the unbiased reagents in the single Cell Painting assay. We've also modified the last two sentences in the first paragraph of the introduction to strengthen the connection between Cell Health assays, targeted reagents, and high cost:

Cell health is normally assessed by eye or measured by specifically targeted reagents, which are either focused on a single Cell Health parameter (ATP assays) or multiple, in combination, via FACS-based or image-based analyses, which involves a manual gating approach, complicated staining procedures, and significant reagent cost. These traditional approaches limit the ability to scale to large perturbation libraries such as candidate compounds in academic and pharmaceutical screening centers.

(3) Generalizability and/or training data size is essential for any model-based classification, but not evaluated or validated in the current manuscript. The independent validation on a A549 cell line only data might be not sufficient/convincing.

We separately address the two distinct points raised by the reviewer of 1) generalizability and 2) training data size:

1. Generalizability

We agree that any model-based classification must demonstrate generalizability. For this reason, we have taken careful consideration to assess the generalizability of all 70 models in two contexts. First, we assessed model performance in a single held out test set (15% of all data). All results we report in the main text (e.g. Figure 2) report performance on this test set. We see high performance in many (but not all) models, and we observe much better model performance compared to a negative control baseline (New Supplementary Figure S5). High performance in the test set indicates that, for some cell health indicators, the models generalize well.

Second, we also demonstrate that these models generalize to data from an entirely different experiment using a fundamentally different perturbation (CRISPR vs. drug compounds). We demonstrate generalizability to this external validation data in four different ways: 1) Validating a relatively simple model ("Number of Live Cells") with an orthogonal viability readout from the PRISM assay (barcoding-based cell viability; updated Figure 4); 2) Demonstrating that proteasome inhibitors, which are known to produce reactive oxygen species, are predicted to do so; 3) Demonstrating that PLK inhibitors, which are known to reduce entry to G1, show a robust dose response in the "G1 Cell Count" model; and 4) Demonstrating that aurora kinase and tubulin inhibitors are predicted to induce high DNA damage (gH2AX) in G1 cells. These two drug classes are known to cause "mitotic slippage" and double stranded DNA breaks. The fourth example was added in response to a comment by reviewer 3.

We've also added a series of enrichment tests, as described in the following new text:

We also chose to validate three additional models: *ROS*, *G1 cell count*, and *Number of gH2AX spots in G1 cells*. We observed that the two proteasome inhibitors (bortezomib and MG-132) in the Drug Repurposing Hub set yielded high *ROS* predictions (OR = 76.7; $p < 1 \times 10^{-15}$) (**Figure 4C**). Proteasome inhibitors are known to induce ROS (Han and Park, 2010; Ling et al., 2003).

As well, PLK inhibitors yielded low *G1 cell counts* (OR = 0.035; $p = 3.9 \times 10^{-8}$) (Figure 4C). The PLK inhibitor HM-214 showed an appropriate dose response (Figure 4D). PLK inhibitors block mitotic progression, thus reducing entry into the G1 cell cycle phase (Lee et al., 2014). Lastly, we observed that aurora kinase and tubulin inhibitors were enriched for high *Number of gH2AX spots in G1 cells* predictions (OR = 11.3; $p < 1 \times 10^{-15}$) (Figure 4E). In particular, we observed a strong dose response for the aurora kinase inhibitor barasertib (AZD1152) (Figure 4F). Aurora kinase and tubulin inhibitors cause prolonged mitotic arrest, which can lead to mitotic slippage, G1 arrest, DNA damage, and senescence (Orth et al. 2011; Cheng and Crasta 2017; Tsuda et al. 2017).

The updated methods section describing our approach to assess generalizability perform the enrichment tests now states:

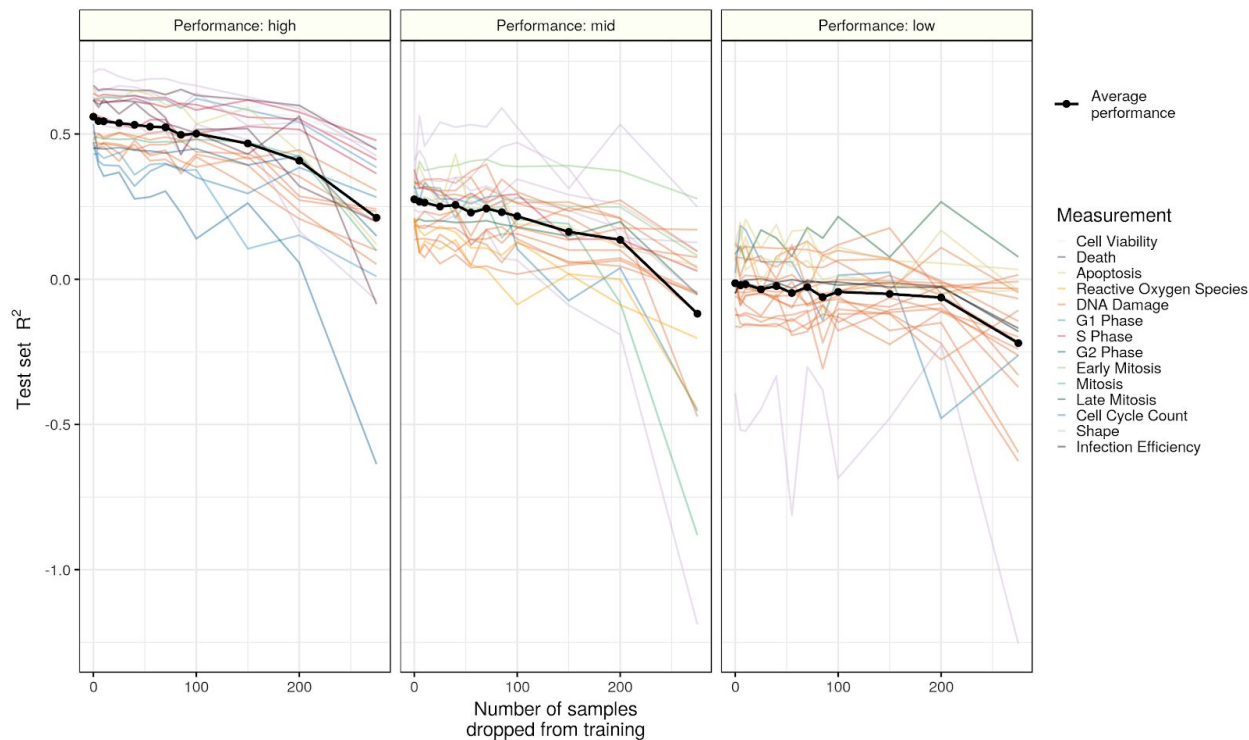
Assessing generalizability of cell health models applied to Drug Repurposing Hub data

We used our cell health webapp (<https://broad.io/cell-health-app>) to identify compounds with high predictions for three models with high or intermediate performance: *ROS*, *Number of G1 cells*, and *Number of gH2AX spots in G1 cells*. For each model, we identified classes of compounds with consistently high scores, then tested for statistical enrichment: for proteasome inhibitors in the *ROS* model, PLK inhibitors in the *Number of G1 cells* model, and aurora kinase and tubulin inhibitors in the *Number of gH2AX spots in G1 cells* model. We used one-sided Fisher's exact tests to quantify differences in expected proportions between high and low model predictions. For each case, we determined high and low predictions based on the 50% quantile threshold for each model independently.

We acknowledge that prospectively making predictions and measuring Cell Health readouts directly in a new experiment would be more convincing, but we note that our existing assessment of generalizability in an external experiment is already unusual in machine learning publications. Additionally and unfortunately, collecting a second validation dataset for this manuscript is not currently feasible given experiments backlogged from COVID.

2. Training data size

We also agree that a more comprehensive analysis on training data size would be an important indicator of model limitations. Therefore, we performed a sample titration analysis in which we randomly dropped samples from the training procedure, and tracked performance of the held out test set. We add the following figure, figure legend, and results text to describe and interpret the results.



Supplementary Figure S13: Dropping samples from training reduces test set model performance in high, mid, and low performing models. We determined model performance stratification by taking the top third, mid third, and bottom third of test set performance when using all data. We performed the sample titration analysis with 10 different random seeds and visualized the median test set performance for each model.

We updated the results section to introduce and discuss this result:

Lastly, we performed a sample size titration analysis in which we randomly removed a decreasing amount of samples from training. For the high and mid performing models, we observed a consistent performance drop, suggesting that increasing sample size would result in better overall performance (**Supplementary Figure 13**).

Finally, the updated methods section describing our sample titration analysis now reads:

Machine learning robustness: Investigating the impact of sample size

We performed an analysis in which we randomly dropped an increasing amount of samples from the training set before model training. After dropping the predefined number of samples, we retrained all 70 cell health models and assessed performance on the original holdout test set. We performed this procedure ten times with ten unique random seeds to mirror a more realistic scenario of new data collection and to reduce the impact of outlier samples on model training.

All software updates introducing this analysis can be viewed at <https://github.com/broadinstitute/cell-health/pull/143>

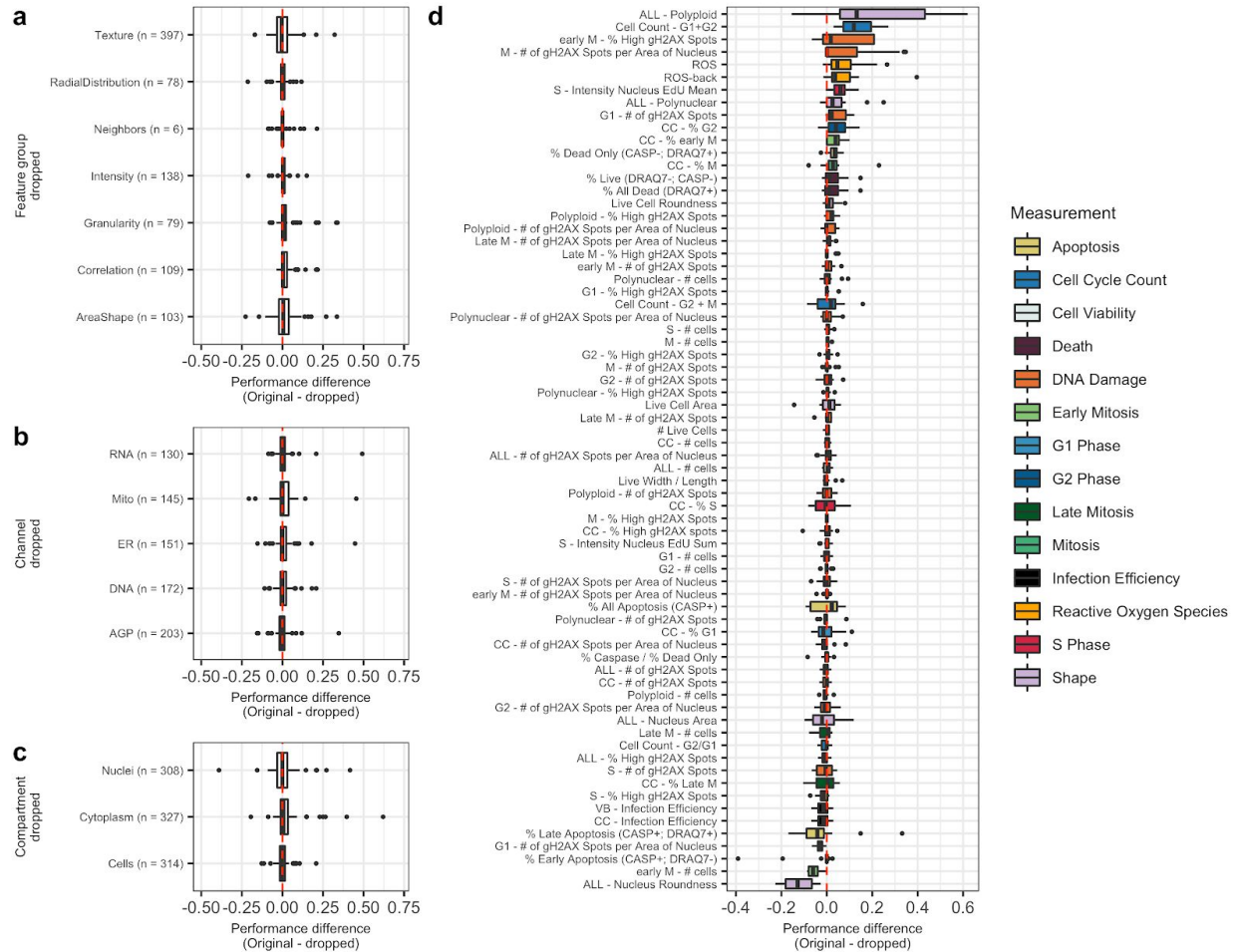
****Minor concerns:****

(1) Highest test performance comprises that precision is mainly driven by cell cycle/count and live status and could be probably derived from DRAQ7 (Fig. 2) and DNA granularity (Fig. 3, bottom right) and would argue for rigid feature selection across channels and features.

We believe that clarifying the confusion between the two Cell Health assays we developed and the well-established Cell Painting assay addresses part of this concern. The DRAQ7 dye marks dead cells, and is measured in Cell Health. In other words, readouts from this reagent are what we aim to predict, not what we use for training. Indeed, DRAQ7-based phenotypes are among the top predicted models, which is a result we present in Supplementary Figure S7 - this figure uncovers which Cell Health phenotypes are more easily predicted by Cell Painting.

The DNA granularity morphology measurements are collected from the Cell Painting assay and thus *are* available for training, and, as noted by the reviewer, encode a high proportion of signal in predicting the various cell health phenotypes. In our most common processing workflows for other projects, we do apply a rigid feature selection pipeline to all Cell Painting profiles before analysis, but we do not do this in this analysis since we were using a model with a sparsity-inducing penalty (elastic net).

To directly answer the question of how channels and feature groups influence model performance, we've performed a systematic experiment removing different channel, compartment, and feature groups and retraining all models with the specific group dropped. We now include the following supplementary figure:



Supplementary Figure S12: Systematically removing classes of features has little impact on most models' performance. We retrained all 70 cell health models after dropping features associated with specific **(a)** feature groups, **(b)** channels, and **(c)** compartments. Each dot is one model (predictor), and the performance difference between the original model and the retrained model after dropping features is shown on the x axis. Any positive change indicates that the models got worse after dropping the feature group. **(d)** Individual model differences in performance after dropping features. Each dot is one class of features removed (as in a-c).

Additionally, we updated the results section to introduce and discuss this result:

We also performed a systematic feature removal analysis, in which we retrained cell health models after dropping features that are measured from specific groups, compartments, and channels. We observed that most models were robust to dropping entire feature classes during training (**Supplementary Figure 12**). This result demonstrates that many Cell Painting features are highly correlated, which might permit prediction “rescue” even if the directly implicated morphology features are not measured. Because of this, we urge caution when generating hypotheses regarding causal relationships between readouts and individual Cell Painting features.

And we add the following to the methods section:

Machine learning robustness: Systematically removing feature classes

We performed an analysis in which we systematically dropped features measured in specific compartments (Nuclei, Cells, and Cytoplasm), specific channels (RNA, Mito, ER, DNA, AGP), and specific feature groups (Texture, Radial Distribution, Neighbors, Intensity, Granularity, Correlation, Area Shape) and retrained all models. We omitted one feature class and then independently optimized all 70 cell health models as described in the Machine learning framework results section above. We repeated this procedure once per feature class.

All software updates introducing this analysis can be viewed at <https://github.com/broadinstitute/cell-health/pull/143>

(2)Any H2AX and 'polynuclear' would probably fail in any cell line with this size of training data.

Indeed we would expect certain cell health phenotype models to fail if they had few hits and a relatively low variance of output values. This hit rate is directly associated with the phenotypes that the CRISPR perturbations induce, which is why we intentionally selected them to span multiple gene pathways in an attempt to maximize morphology diversity (see Supplementary Table S1).

We did indeed observe that the polynuclear model had few hits in the training data and relatively poor performance. We did not expect this result, given that DNA stains are captured in the Cell Health and Cell Painting assays. We suspect the poor performance in this model is likely because so few cells were classified as polynuclear in our gating strategy, making it perhaps an inconsistently measured readout.

By contrast, some gH2AX models did have relatively good performance. In the conclusion, we note that increased training data size using more perturbations is likely to improve model performance:

The results motivate collecting larger datasets for training, with more perturbations and multiple cell lines. These new datasets would enable the development of more expressive models, based on deep learning, that can be applied to single cells. Including orthogonal imaging markers of CRISPR infection would also enable us to isolate cells with expected morphologies. More data and better models would improve the performance and generalizability of Cell Health models and enable annotation of new and existing large-scale Cell Painting datasets with important mechanisms of cell health and toxicity.

(3)To what refers the 'weights' of the model in Fig. 1c?

We thank the reviewer for pointing out that we never defined this term in the Figure 1 legend. We use “weights” to refer to the coefficients from the regression model. To make this more

clear, we have updated the legend to now read: “Model coefficient weights” and the text in Figure 1C to now read “model weights”.

Reviewer #1 (Significance (Required)):

This manuscript is not advanced in the context of latest improvements/developments of cell-based microscopic classification. Rationale in the introduction and the conclusion are not linked (interpretability, generalizability, costs). It seems to be unfinished or unformatted to this end?

Since responding to these reviews, we believe that our primary motivation - to demonstrate proof-of-concept of predicting cell health phenotypes directly from Cell Painting data - is now much clearer, holistically. We provide below an updated introduction, which improves rationale.

Perturbing cells with specific genetic and chemical reagents in different environmental contexts impacts cells in various ways (Kitano, 2002). For example, certain perturbations impact cell health by stalling cells in specific cell cycle stages, increasing or decreasing proliferation rate, or inducing cell death via specific pathways (Markowitz, 2010; Szalai et al., 2019). Cell health is normally assessed by eye or measured by specifically targeted reagents, which are either focused on a single Cell Health parameter (ATP assays) or multiple, in combination, via FACS-based or image-based analyses, which involves a manual gating approach, complicated staining procedures, and significant reagent cost. These traditional approaches limit the ability to scale to large perturbation libraries such as candidate compounds in academic and pharmaceutical screening centers.

Image-based profiling assays are increasingly being used to quantitatively study the morphological impact of chemical and genetic perturbations in various cell contexts (Caicedo et al., 2016; Scheeder et al., 2018). One unbiased assay, called Cell Painting, stains for various cellular compartments and organelles using non-specific and inexpensive reagents (Gustafsdottir et al., 2013). Cell Painting has been used to identify small-molecule mechanisms of action (MOA), study the impact of overexpressing cancer mutations, and discover new bioactive mechanisms, among many other applications (Caicedo et al., 2018; Christoforow et al., 2019; Hughes et al., 2020; Pahl and Sievers, 2019; Rohban et al., 2017; Simm et al., 2018; Wawer et al., 2014). Additionally, Cell Painting can predict mammalian toxicity levels for environmental chemicals (Nyffeler et al., 2020) and some of its derived morphology measurements are readily interpreted by cell biologists and relate to cell health (Bray et al., 2016). However, no single assay enables discovery of fine-grained cell health readouts.

We hypothesized that we could predict many cell health readouts directly from the Cell Painting data, which is already available for hundreds of thousands of perturbations. This would enable the rapid and interpretable annotation of small molecules or genetic perturbations. To do this, we first developed a customized microscopy assays, which collectively report on 70 different cell

health indicators via a total of seven reagents applied in two reagent panels. Collectively, we call these assay panels “Cell Health”.

To demonstrate proof of concept, we collected a small pilot dataset of 119 CRISPR knockout perturbations in three different cell lines using Cell Painting and Cell Health. We used the Cell Painting morphology readouts to train 70 different regression models to predict each Cell Health indicator independently. We used simple machine learning methods instead of a deep learning approach because of our limited sample size and the inability to increase it by linking single cell measurements from both assays. We predicted certain readouts, such as the number of S phase cells, with high performance, while performance on other readouts, such as DNA damage in G2 phase cells, was low. We applied and validated these models on a separate set of existing Cell Painting images acquired from 1,571 compound perturbations measured across six different doses from the Drug Repurposing Hub project (Corsello et al., 2017). We provide all predictions in an intuitive web-based application at <http://broad.io/cell-health-app>, so that others can extend our work and explore cell health impacts of specific compounds.

Reviewer #2 (Evidence, reproducibility and clarity (Required)):

This report from Way et al describes a method of extending a very popular screening technology called Cell Painting developed by the Carpenter Lab. The authors are contending with an important issue and as such this paper potentially will be of great interest to the community. Cell Painting provides quantitative fingerprints of cell phenotypes in response to changes in the molecular or physiological status of cells. However the molecular basis or even the candidate pathways for those changes is not always clear. Here, the authors take specific markers of cell physiology, e.g., DNA damage, ROS production, cell cycle progression etc. and relate them to Cell Painting features. The authors are trying to address the issue that running many probes of cell physiology is expensive and time consuming and that identifying proxies for these assays using much simpler Cell Painting technologies would be a useful and potentially powerful approach. The overall goal is to develop some type of regression model that can link the state of cells (the "health") to Cell Painting fingerprints.

The authors use three separate cell lines and CRISPR knockouts delivered through lentivirus that target 59 genes to establish a range of cell physiologies that they directly measure (the "Cell Health") and then relate to similar assays performed by Cell Painting. Ultimately they aim to use Cell Painting models to predict Cell Health.

We thank the reviewer for their succinct summary of our goals and rationale for this manuscript, and for the constructive and valuable comments herein.

****Major Issues:****

It appears that the phenotypes that are detected at a high enough level of significance (see Fig. 2), e.g DNA damage (gH2Ax), apoptosis (Caspase 3/7), dead cells, ROS (CellROX), etc. are probably most easily detected by simply monitoring DAPI signal in these screens. To detect many of the phenotypes, the authors have presented a fairly complex method of doing much simpler assays. The authors correctly highlight in Fig. 3 that the phenotypes they are detecting go beyond pure signals from DAPI. They report power in their models from Radial Distribution across many different components of the Cell Painting feature set.

We agree that the two assays we're collectively calling "Cell Health" are indeed fairly complex - we use two different panels of multiplexed stains and a series of gating strategies to measure phenotypes in various cell subpopulations. However, the fundamental message in the manuscript is that we may no longer need to perform these complex assays if we get this information from the simpler Cell Painting assay.

We agree that our machine learning approach to predict the various cell health phenotypes uses signals beyond nucleus-based stains. However, even if we are predicting just DAPI signals, this reinforces our argument that the specific stains in the Cell Health assays (which are commonly used in targeted experiments) are not necessary to measure specifically. Instead, in certain circumstances, a scientist should just use unbiased stains to capture their biology of interest, since the stains are cheaper at scale and one has access to much more information.

It is also worth noting that the DNA damage phenotypes in specific cell subpopulations (e.g. DNA Damage in G1 cells) would not be possible to measure with high precision without EdU co-staining.

However these appear to give outputs that won't be that useful. It is hard to tell whether this is simply because they don't have enough images or whether their signal is confounded by using cell lines where the lentivirus CRISPR knockouts are working less efficiently.

(Reviewer 2 introduced a similar critique below, which we now move here) A fundamental issue that the authors mention but do not address is the efficiency of the CRISPR KOs. The authors should measure the efficiency of representative guides and present these data to help support the interpretation of their models.

We definitely agree that sample size is a limitation in this manuscript. Our primary goal with this paper was to demonstrate feasibility of the approach to predict the targeted Cell Health readouts using a simpler (and more affordable/scalable) assay in Cell Painting. The promising results we observed, especially given this sample size limitation, motivates collecting a larger dataset using more perturbations.

Potentially confounded signal by low efficiency CRISPR knockouts is also an interesting topic. We do provide Supplementary Figure S8 to describe a subtle relationship that we observed regarding CRISPR infection efficiency. We also discuss this in the results as: "We observed

overall better predictivity in ES2 cells, which had the highest CRISPR infection efficiency (**Supplementary Figure 8**), suggesting that stronger perturbations provide better information for training and that training on additional data should provide further benefit.”

Additionally, we made a substantial effort to maximize CRISPR efficiency by independently optimizing lentivirus volumes for each sgRNA. In general, we observed that some cell lines are easier to CRISPR, probably based on more factors beyond Cas9 expression. However, we note that CRISPR is being used simply as a perturbation to elicit a variable morphology response. In other words, the type, efficacy, and even accuracy of perturbation does not matter as long as it satisfies two constraints: 1) induces a morphology response for a sufficient number of perturbations, and 2) is consistent between the two assays (Cell Health and Cell Painting). Our setup satisfies both constraints.

However, this experiment (and data from the experiment) can be used in other contexts in which the CRISPR efficiency *is* extremely important. Therefore, we added three columns to Supplementary Table 1 providing the efficiency readouts for the three cell lines. (This information was already present in GitHub, but we moved it to a more obvious location in Supplementary Table 1). Code describing this change can be viewed here: <https://github.com/broadinstitute/cell-health/pull/142>

In regards to the first sentence of this concern: “However these appear to give outputs that won’t be that useful” - indeed, we fully expected that many cell health readouts would be difficult to predict. In the original submission, we included the following explanation for potential sources of low performing models: “Performance differences might result from random technical variation, small sample sizes for training models, different number of cells in certain Cell Health subpopulations (e.g. mitosis or polynuclear cells), fewer cells collected in the viability panel (see methods), or the inability of Cell Painting reagents to capture certain phenotypes.”

It seems misleading (or perhaps the explanation lacks clarity) to describe in the same paragraph the need to validate the model by applying it to new datasets, namely the Drug Repurposing Hub project, then describe gradients in cell health features across UMAP coordinates.

We thank the reviewer for pointing out this source of confusion and for providing an opportunity to improve the clarity of this section. Our major revisions here are as follows: 1) Introduce the Drug Repurposing Hub as an external dataset for validation; 2) Validate a high performing and simple model (*number of live cells*) by comparing model readout predictions from the Drug Repurposing Hub Cell Painting profiles against orthogonal PRISM viability readouts (in compounds with slightly different doses); 3) Validate three additional models: enrichment of proteasome inhibitors in the *ROS* model, enrichment of PLK inhibitors in the *G1 cell count* model, and enrichment of tubulin-destabilizing compounds in the *Number of gH2Ax spots in G1 cells* model; 4) Display a global structure of Cell Health predictions in UMAP space for select models. Note that for the fourth point, we are using the UMAP gradients to observe patterns, and not to validate models.

In order to encapsulate the updated flow, we've pasted below the entire Drug Repurposing Hub results/discussion section, which introduces two additional analyses and new text in response to various other reviewer comments. We feel that the updated section improves clarity and purpose.

The updated section now reads:

“Predictive models of cell health would be most useful if they could be trained once and successfully applied to data sets collected separately from the experiment used for training. Otherwise one could not annotate existing datasets that lack parallel Cell Health results, and Cell Health assays would have to be run alongside each new dataset. We therefore applied our trained models to a large, publicly-available Cell Painting dataset collected as part of the Drug Repurposing Hub project (Corsello et al., 2017). The data derive from A549 lung cancer cells treated with 1,571 compound perturbations measured in six doses.

We first chose a simple, high-performing model to validate. The *number of live cells* model captures the number of cells that are unstained by DRAQ7. We compared model predictions to orthogonal viability readouts from a third dataset: Publicly available PRISM assay readouts, which count barcoded cells after an incubation period (Yu et al., 2016). Despite measuring perturbations with slightly different doses and being fundamentally different ways to count live cells (**Figure 4A**), the predictions correlated with the assay readout (Spearman's Rho = 0.35, $p < 1 \times 10^{-3}$; **Figure 4B**).

We also chose to validate three additional models: *ROS*, *G1 cell count*, and *Number of gH2AX spots in G1 cells*. We observed that the two proteasome inhibitors (bortezomib and MG-132) in the Drug Repurposing Hub set yielded high *ROS* predictions (OR = 76.7; $p < 1 \times 10^{-15}$) (**Figure 4C**). Proteasome inhibitors are known to induce ROS (Han and Park, 2010; Ling et al., 2003). As well, PLK inhibitors yielded low *G1 cell counts* (OR = 0.035; $p = 3.9 \times 10^{-8}$) (**Figure 4C**). The PLK inhibitor HM-214 showed an appropriate dose response (**Figure 4D**). PLK inhibitors block mitotic progression, thus reducing entry into the G1 cell cycle phase (Lee et al., 2014). Lastly, we observed that aurora kinase and tubulin inhibitors yielded high *Number of gH2AX spots in G1 cells* predictions (OR = 11.3; $p < 1e-15$) (**Figure 4E**). In particular, we observed a strong dose response for the aurora kinase inhibitor barasertib (AZD1152) (**Figure 4F**). Aurora kinase and tubulin inhibitors cause prolonged mitotic arrest, which can lead to mitotic slippage, G1 arrest, DNA damage, and senescence (Orth et al. 2011; Cheng and Crasta 2017; Tsuda et al. 2017).

We applied uniform manifold approximation (UMAP) to observe the underlying structure of the samples as captured by morphology data (McInnes et al., 2018). We observed that the UMAP space captures gradients in predicted *G1 cell count* (**Supplementary Figure S14A**) and in predicted *ROS* (**Supplementary Figure S14B**). We also observed similar gradients in the ground truth cell health readouts in the CRISPR Cell Painting profiles used for training cell

health models (**Supplementary Figure S15**). Gradients in our data suggest that cell health phenotypes manifest in a continuum rather than in discrete states.

Lastly, we observed moderate technical artifacts in the Drug Repurposing Hub profiles, indicated by high DMSO profile dispersion in the Cell Painting UMAP space (**Supplementary Figure 14C**). This represents an opportunity to improve model predictions with new batch effect correction tools. Additionally, it is important to note that the expected performance of each Cell Health model can only be as good as the performance observed in the original test set (see **Figure 2**), and that all predictions require further experimental validation.“

Updated Figure 4:

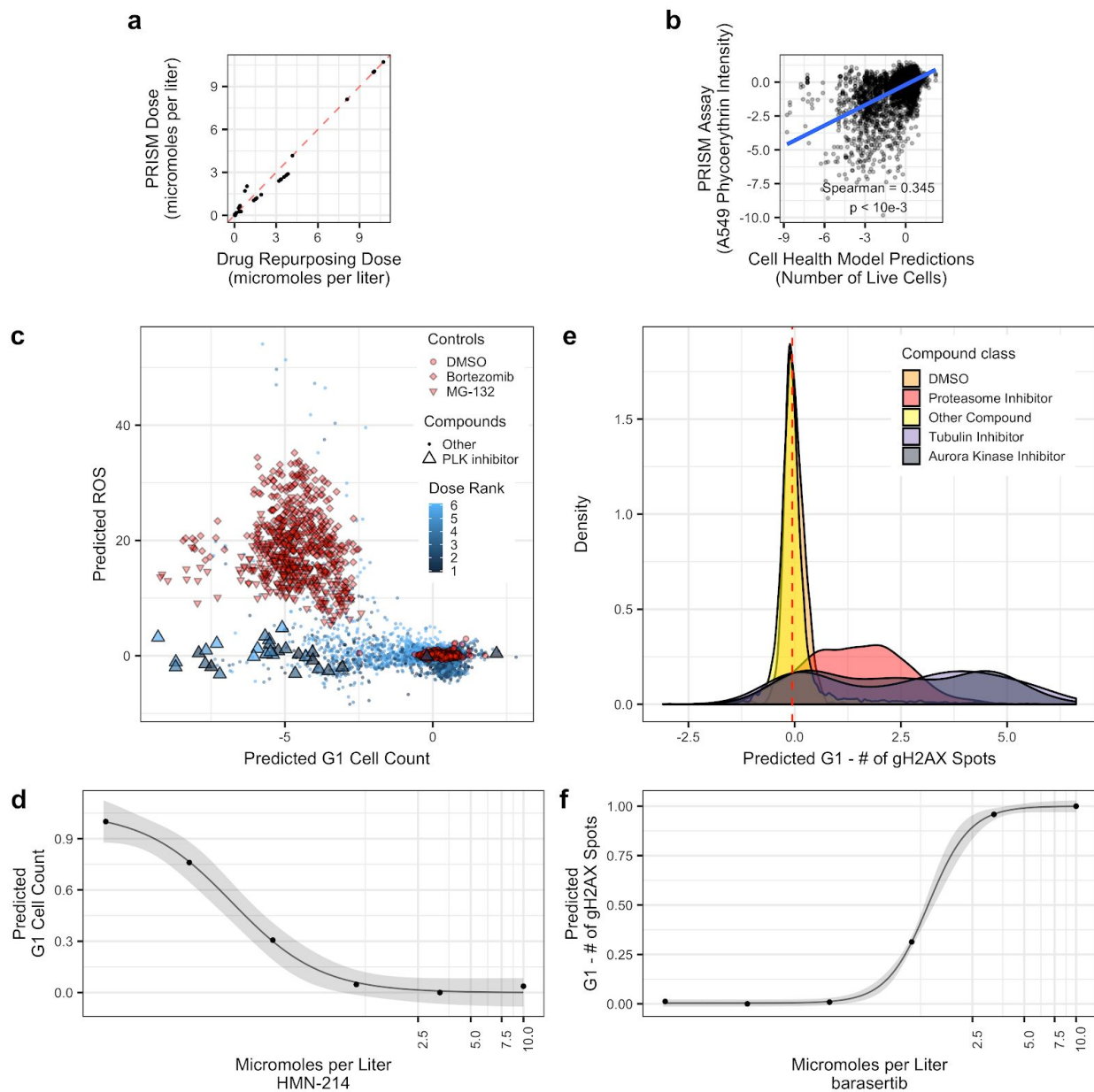
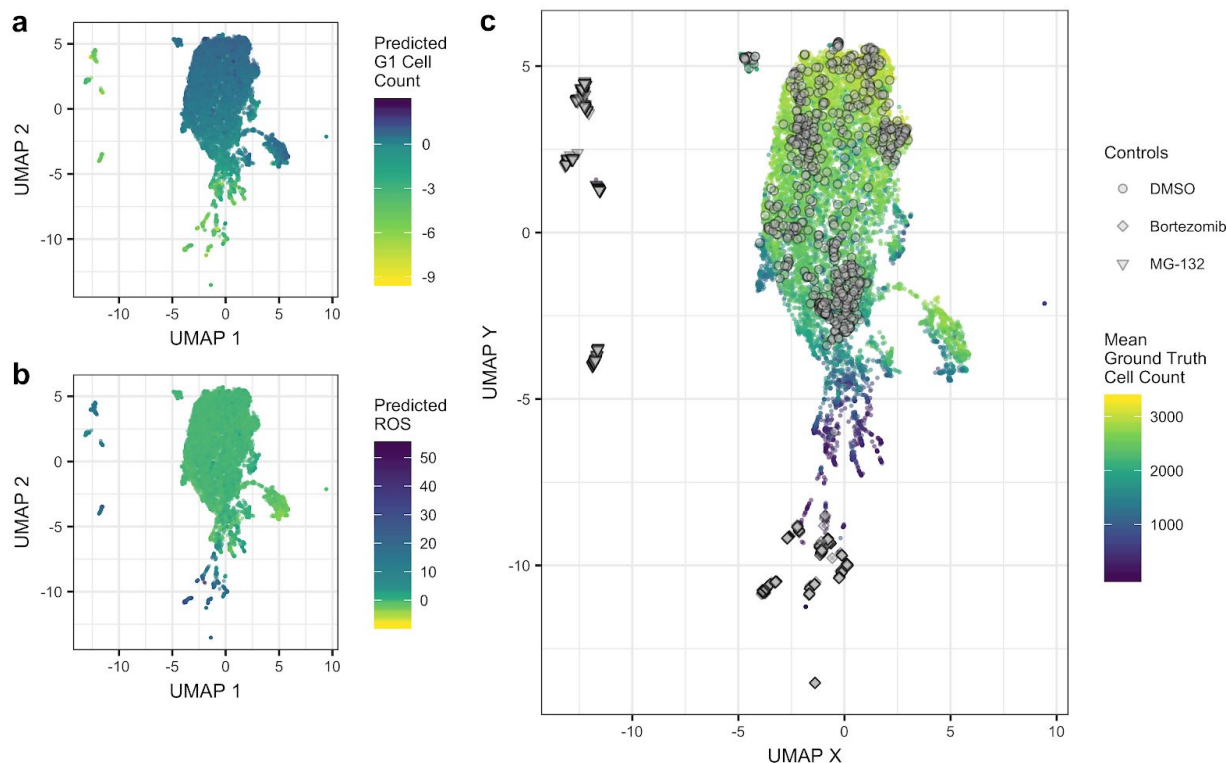


Figure 4: Validating Cell Health models applied to Cell Painting data from The Drug Repurposing Hub. The models were not trained using the Drug Repurposing Hub data. **(a)** The results of the dose alignment between the PRISM assay and the Drug Repurposing Hub data. This view indicates that there was not a one-to-one matching between perturbation doses. **(b)** Comparing viability estimates from the PRISM assay to the predicted *number of live cells* in the Drug Repurposing Hub. The PRISM assay estimates viability by measuring barcoded A549 cells after an incubation period. **(c)** Drug Repurposing Hub profiles stratified by *G1 cell count* and *ROS* predictions. Bortezomib and MG-132 are proteasome inhibitors and are used as positive controls in the Drug Repurposing Hub set; DMSO is a negative control. We also highlight all PLK inhibitors in the dataset. **(d)** HMN-214 is an example of a PLK inhibitor that shows strong dose response for *G1 cell count* predictions. **(e)** Tubulin and aurora kinase inhibitors are

predicted to have high *Number of gH2AX spots in G1 cells* compared to other compounds and controls. **(f)** Barasertib (AZD1152) is an aurora kinase inhibitor that is predicted to have a strong dose response for *Number of gH2AX spots in G1 cells* predictions.

Updated Supplementary Figure:



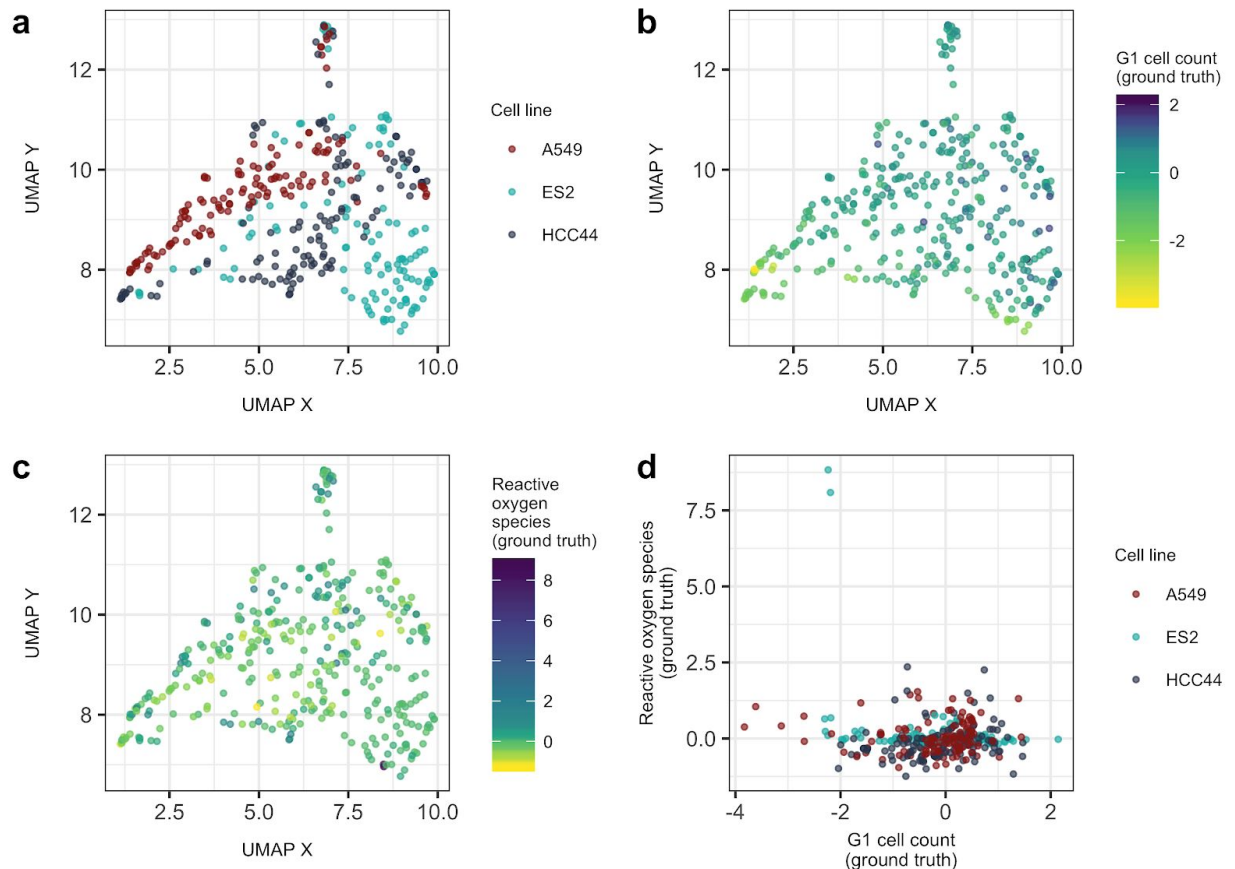
Supplementary Figure S14: Applying a Uniform Manifold Approximation (UMAP) to Drug Repurposing Hub consensus profiles of 1,571 compounds across six doses. The models were not trained using the Drug Repurposing Hub data. **(a)** The point color represents the output of the Cell Health model trained to predict the number of cells in G1 phase (G1 cell count). **(b)** The same UMAP dimensions, but colored by the output of the Cell Health model trained to predict reactive oxygen species (ROS). **(c)** In the UMAP space, we highlight DMSO as a negative control, and Bortezomib and MG-132 as two positive controls (proteasome inhibitors) in the Drug Repurposing Hub set. We observe moderate batch effects in the negative control DMSO profiles, based on their spread in this visualization. The color represents the predicted number of live cells. The positive controls were acquired with a very high dose and are expected to result in a very low number of predicted live cells.

All software updates required to update these figures can be viewed at <https://github.com/broadinstitute/cell-health/pull/145>

Is it surprising that cell health phenotypes and gradients therein are present in a dataset describing cell health perturbations?

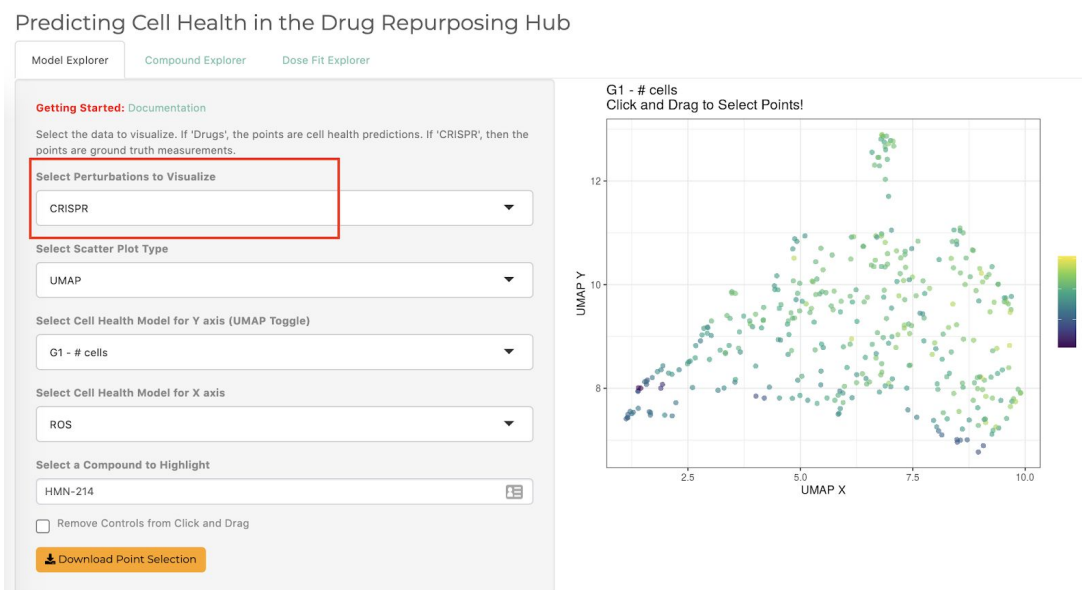
This was not surprising to us, and we thank the reviewer for asking the question. We have now added a new Supplementary Figure to present a UMAP with ground truth cell health measurements in the CRISPR dataset (pasted below). By adding the figure, we show how Cell Health predictions are expected to show gradients in UMAP space. In fact, for any lower-dimensional embedding that is able to preserve local neighborhoods of the high-dimensional space, we should expect all linear transformations of the input data (in the high-dimensional space) to vary smoothly across the lower-dimensional embedding. However, it is still informative to observe where the specific Cell Health phenotype predictions manifest in relation to global morphology structure. We add the following sentence in the Drug Repurposing Hub paragraph juxtaposed to the other UMAP gradient observations:

We applied uniform manifold approximation (UMAP) to observe the underlying structure of the samples as captured by morphology data (McInnes et al., 2018). We observed that the UMAP space captures gradients in predicted *G1 cell count* (Supplementary Figure S14A) and in predicted *ROS* (Supplementary Figure S14B). We also observed similar gradients in the ground truth cell health readouts in the CRISPR Cell Painting profiles used for training cell health models (Supplementary Figure S15). Gradients in our data suggest that cell health phenotypes manifest in a continuum rather than in discrete states.



Supplementary Figure S15: Applying a Uniform Manifold Approximation (UMAP) to the Cell Painting consensus profile data of CRISPR perturbations. UMAP coordinates visualized by (a) cell line, (b) ground truth G1 cell counts, and (c) ground truth ROS counts. (d) Visualizing the distribution of ground truth ROS compared against G1 cell count. The two outlier ES2 profiles are CRISPR knockdowns of *GPX4*, which is known to cause high ROS.

We have also added the option to explore the CRISPR profile Cell Health ground truth in our shiny app <https://broad.io/cell-health> (screenshot pasted below)



Modifications to the software introducing these changes can be viewed at <https://github.com/broadinstitute/cell-health/pull/141>.

The actual test of the model's performance is in the paragraph below, but the data associated with the Spearman correlation is hidden in Fig. S10b. The data is not convincing by eye, and the artifactually low p value suggests that proper statistical corrections were not applied.

We have moved the Spearman correlation figure (previously Supplementary Figure S10B) into a main figure, along with a complete restructuring of the results and discussion in the Drug Repurposing Hub section.

We appreciate the careful observations and interpretations, and confirm the statistical test performed here is sound and the p value is correct (there is no need to account for multiple testing since there is only one test being applied, a test of correlation between two variables).

We add this rationale to the "Comparing viability predictions to an orthogonal readout" methods section:

We performed the non-parametric Spearman correlation test because 1) the doses were not aligned between the datasets we compared, and 2) it is possible that a strong nonlinear correlation exists between readouts from two fundamentally different ways to measure viability.

It is definitely valid to critique the scatter plot relationship to understand that the mean squared error is quite high (i.e. if two datasets had viability measurements using the two approaches, it would be wrong to assume that lower measurements in one assay automatically could be compared to lower measurements of the other assay). This level of variability would be lost if all we did was report the test statistic, which is the reason why we included the scatter plot as a figure.

It may also be important to mention that the authors of the PRISM paper also noted high variation in their estimates (from Corsello et al <https://doi.org/10.1038/s43018-019-0018-6>): "At the level of individual compound dose–responses, we note that the PRISM Repurposing dataset tends to be somewhat noisier, with a higher standard error estimated from vehicle control measurements (Extended Data Fig. 5c and Extended Data Fig. 6a–c)."

Nevertheless, we agree that the current way we report this p value is distracting and potentially misleading, depending on how the p value is interpreted. Therefore, we have updated the reporting of all p values to say that they are less than a predefined cutoff. The figure now states that $p < 10e-3$. This decision was inspired by the suggestion provided in <https://twitter.com/rafalab/status/1310610623898808320>

Fig 1A and associated methods are not sufficient information to describe the manual gating strategy and any variability found across iterations in these gates. Effort should be taken to quantify where these manual boundaries were set and why.

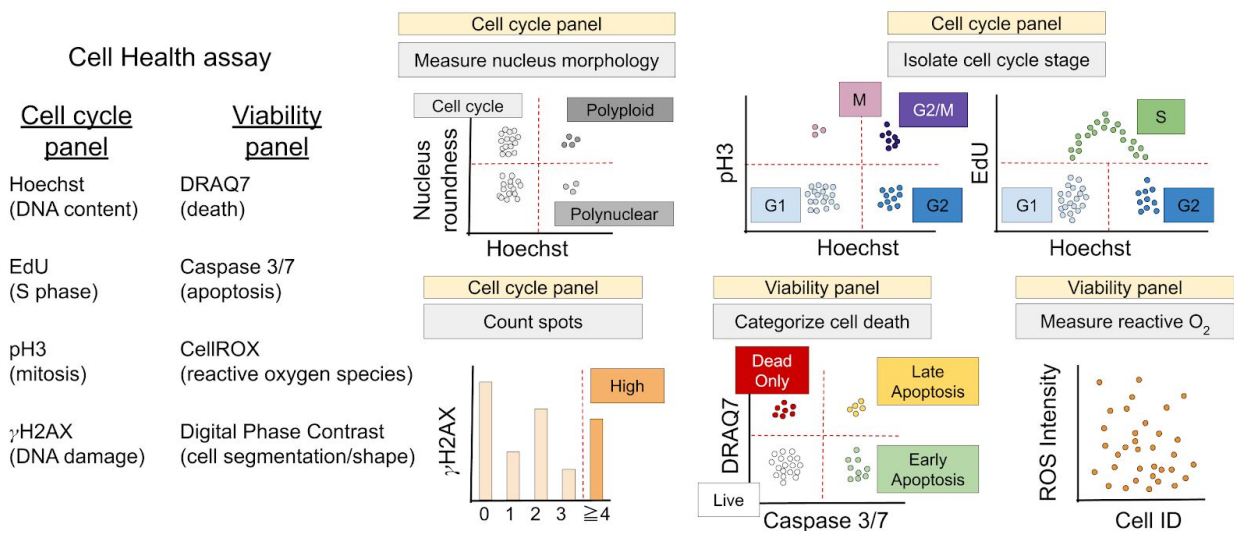
We describe the manual gating strategies in much detail in the methods section "*Cell Health assay: Image analysis*". However, we agree that a description of measurement variability and experimental approach requires more detail, and we agree that the manuscript would benefit from a visual example of these gates. These improvements required us to rearrange Figure 1.

With a goal of increasing reproducibility in the cell health assay, we've (1) moved example images of the Cell Health assay to Figure 1A; (2) Moved the existing gating strategies drawing to Supplementary Figure 1; (3) Added real data examples of the manual gating strategy as a new Supplementary Figure 2. We show all updates below:

Updated Figure 1:

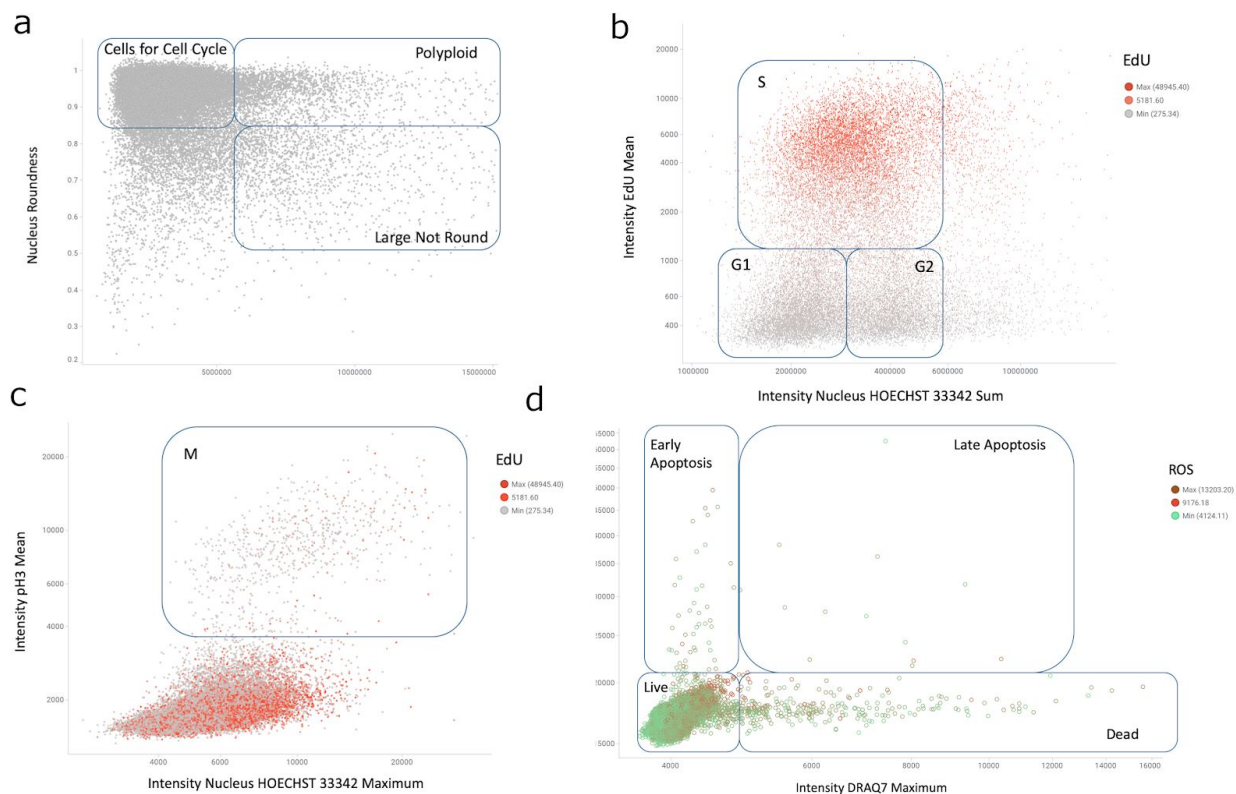
subpopulations and to generate cell health readouts for each perturbation. (top) In the “Cell Cycle” panel, in each nucleus we measure Hoechst, EdU, PH3, and γ H2AX. (bottom) In the “Cell Viability” panel, we capture digital phase contrast images, measure Caspase 3/7, DRAQ7, and CellIROX. (b) Example Cell Painting image across five channels, plus a merged representation across channels. The image is cropped from a larger image and shows ES2 cells. Below are the steps applied in an image-based profiling pipeline, after features have been extracted from each cell’s image. (c) Modeling approach where we fit 70 different regression models using CellProfiler features derived from Cell Painting images to predict Cell Health readouts.

Updated Supplementary Figure S1:



Supplementary Figure S1: Illustration of the gating strategy in the Cell Health assays. We extract 70 different readouts from the Cell Health imaging assay. The assay consists of two customized reagent panels, which use measurements from seven different targeted reagents and one channel based on digital phase contrast (DPC) imaging; shown are five toy examples to demonstrate that individual cells are isolated into subpopulations by various gating strategies to define the Cell Health readouts.

Updated Supplementary Figure S2 (Example gating strategies):



Supplementary Figure S2: Real data of manual gating in the Cell Health assays.

For each cell line, we apply a series of manual gating strategies defined by various stain measurements in single cells to define cell subpopulations. **(a)** In the cell cycle panel, we first select cells that are useful for cell cycle analysis based on nucleus roundness and Hoechst intensity measurements. We also identify polyploid and “large not round” (polynuclear) cells. **(b)** We then subdivide the cells used for cell cycle to G1, G2, and S cells based on total Hoechst intensity (DNA content) and EdU incorporation signal intensity. **(c)** We use Hoechst and PH3 nucleus intensity to define mitotic cells. The points are colored by EdU intensity in the nucleus in both (b) and (c). **(d)** Example gating in the viability panel. We use DRAQ7 and CellEvent (Caspase 3/7) to distinguish alive and dead cells, and categorize early or late apoptosis. See Methods for more details about how the Cell Health measurements are made.

We’ve also added the following to the methods section:

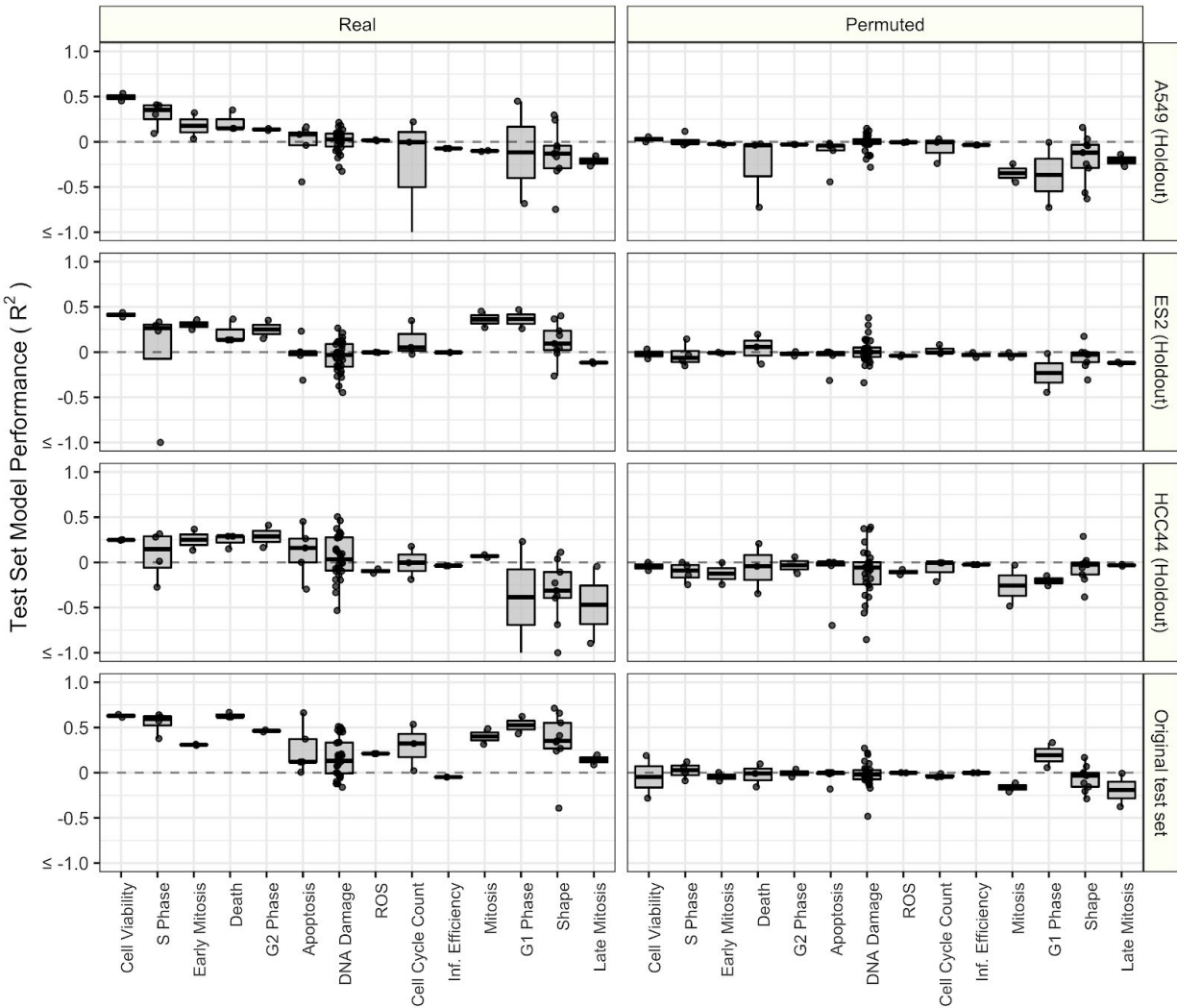
Additionally, we set these gates for each cell subpopulation using a set of random wells from each cell line and experiment independently. We observed that the intensity measurements used to form the gates were consistent across wells and plates, and generally formed distinct cell subpopulation clusters. After using the random wells to set the gates, we used the Harmony microscope software to apply the gates to the remaining wells and plates.

In general however, the need to clearly define this process further emphasizes a strength in our approach: There is great potential for inconsistencies when different humans draw gates. We aim to reduce these inconsistencies by predicting these readouts from Cell Painting images directly.

The authors conclude that their results motivate further data acquisition and model training, and that this will improve model performance. This is only true if their lack of predictive power comes from the data volume itself, and not in larger problems of data quality, variability and the core assumptions of their method. The authors note the better predictability in ES2 cells, likely due to higher CRISPR efficiency and therefore stronger phenotypes. It is possible, as I believe the authors suggest, that the ES2 cells provide information that improves the predictive power of cells with poor infection efficiency. It is instead possible that only the ES2 cells with strong phenotypes yield predictive power, pulling the average of the dataset up. Authors could train the cell line specific datasets independently and compare relative changes in predictive performance. Otherwise, is it possible that subtle or highly complex phenotypes simply cannot be detected by this method and more data will be unlikely to improve predictability in modest perturbations.

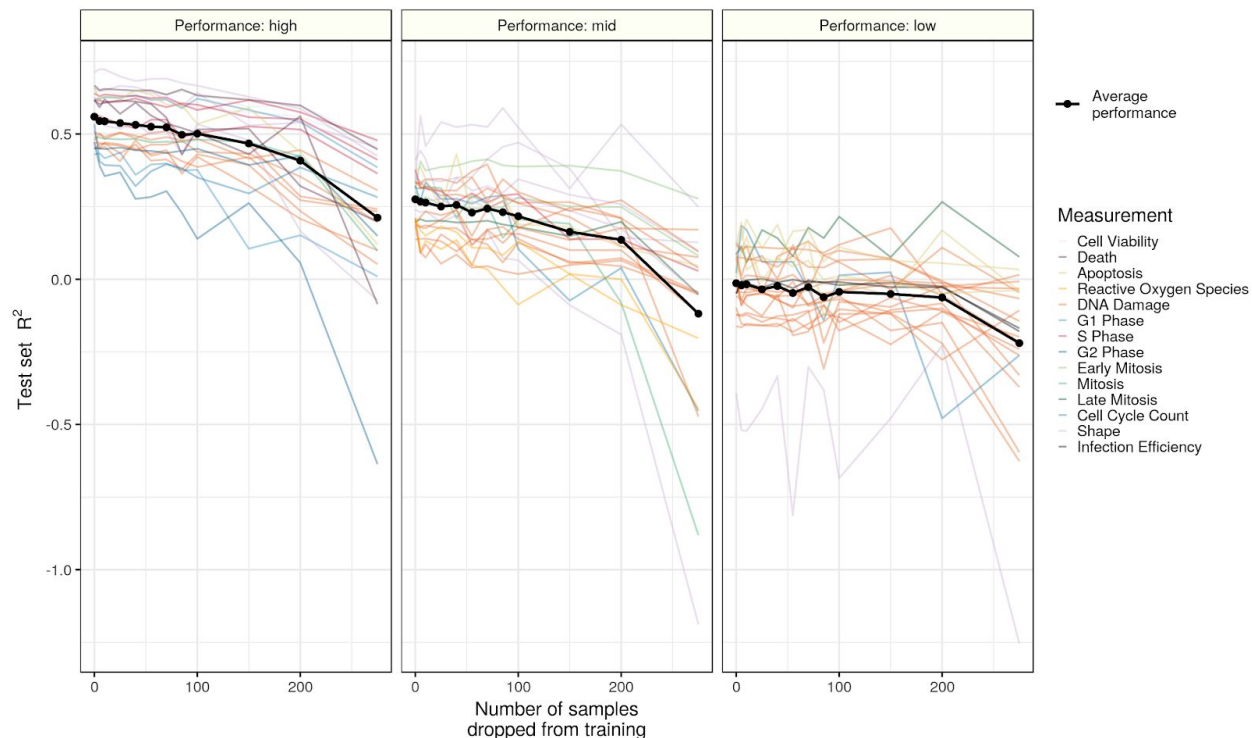
We thank the reviewers for raising this possibility. To explore this, we performed a cell-line holdout analysis in which we retrained (and individually reoptimized) all 70 cell health models on every combination of two cell lines and predicted readouts from the held out third cell line.

Despite there being fewer samples in the training set in the cell line holdout test compared to the original test set (66% vs. 85%) and the fact that each model had never seen the held out cell line before, many cell health phenotypes could still be predicted. We add the following results in a new Supplementary Figure:



Supplementary Figure S11: Results from a cell line holdout analysis. We trained and evaluated all 70 cell health models in three different scenarios using each combination of two cell lines to train, and the remaining cell line to evaluate. For example, we trained all 70 models using data from A549 and ES2 and evaluated performance in HCC44. We bin all cell health models into 14 different categories (see Supplementary Table S3 and <https://github.com/broadinstitute/cell-health/6.ml-robustness> for details about the categories and scores). We also provide the original test set (15% of the data, distributed evenly across all cell types) performance in the last row, as well as results after training with randomly permuted data. This cross-cell-type analysis yields worse performance overall. Nevertheless, despite the models never encountering certain cell lines, and having fewer training data points, many models still have predictive power across cell line contexts. Note that we truncated the y axis to remove extreme outliers far below -1. The raw scores are available on <https://github.com/broadinstitute/cell-health>.

We've also performed a sample size titration analysis, which suggests that more data would indeed improve model performance. More data would also enable a deep learning approach, which is also likely to improve performance.



Supplementary Figure S13: Dropping samples from training reduces test set model performance in high, mid, and low performing models. We determined model performance stratification by taking the top third, mid third, and bottom third of test set performance when using all data. We performed the sample titration analysis with 10 different random seeds and visualized the median test set performance for each model.

We also update the results section to introduce and discuss this result:

Lastly, we performed a sample size titration analysis in which we randomly removed a decreasing amount of samples from training. For the high and mid performing models, we observed a consistent performance drop, suggesting that increasing sample size would result in better overall performance (**Supplementary Figure 13**).

And an updated methods describing this analysis now reads:

Machine learning robustness: Investigating the impact of sample size

We performed an analysis in which we randomly dropped an increasing amount of samples from the training set before model training. After dropping the predefined number of samples, we retrained all 70 cell health models and assessed performance on the original holdout test set. We performed this procedure ten times with ten unique random seeds to mirror a more

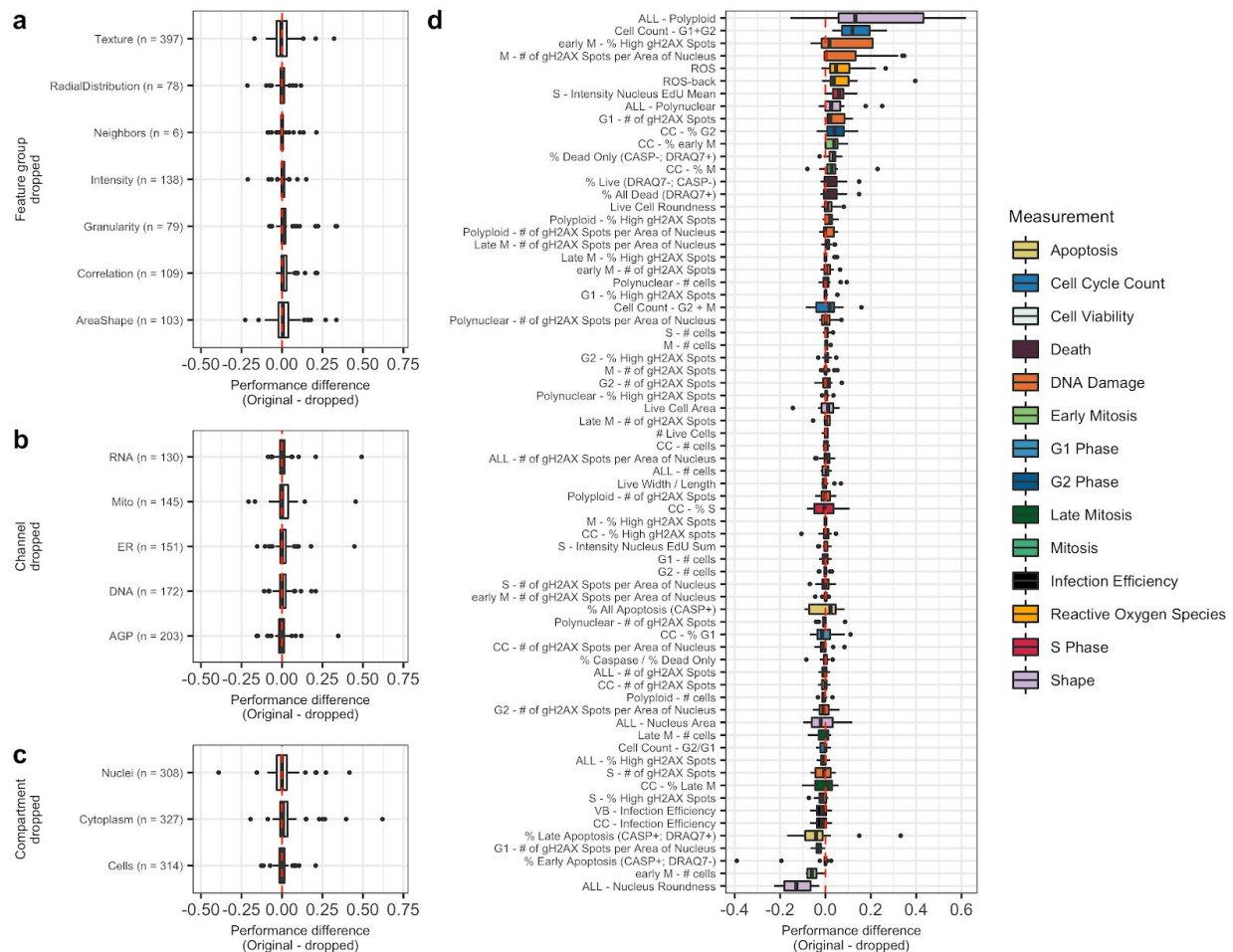
realistic scenario of new data collection and to reduce the impact of outlier samples on model training.

All software updates introducing this analysis can be viewed at <https://github.com/broadinstitute/cell-health/pull/143>

Although the authors argue that the Cell Painting assay is capturing complex health phenotypes using a variety of morphological features, there is a clear overweighting of a particular few (in fact two...). It would be interesting to systematically retrain with exclusion of particular features to determine if equalizing the weight across features changes performance. These are also notably the feature groups with the fewest features-- how many individual features within these feature groups are pulling all the weight?

We agree that an additional computational analysis including a systematic feature removal would be interesting and valuable. We've included this analysis as part of a new results subsection in which we assess where classification improvements are likely to come from by testing robustness of the ML models.

Specifically, we've systematically removed individual features that belong to specific feature groups, channels, and compartments to determine how much their absence negatively affects model performance. The added supplementary figure is pasted below.



Supplementary Figure S12: Systematically removing classes of features has little impact on most models' performance. We retrained all 70 cell health models after dropping features associated with specific **(a)** feature groups, **(b)** channels, and **(c)** compartments. Each dot is one model (predictor), and the performance difference between the original model and the retrained model after dropping features is shown on the x axis. Any positive change indicates that the models got worse after dropping the feature group. **(d)** Individual model differences in performance after dropping features. Each dot is one class of features removed (as in a-c).

We conclude that the majority of cell health models are robust to missing feature groups. Some models actually improve with a reduction in the feature space. Combined with the feature heatmap presented in Figure 3, these results tell us that a lot of the morphology signal is redundant across Cell Painting features.

We add the following text to the results:

We also performed a systematic feature removal analysis, in which we retrained cell health models after dropping features that are measured from specific groups, compartments, and

channels. We observed that most models were robust to dropping entire feature classes during training (**Supplementary Figure 12**). This result demonstrates that many Cell Painting features are highly correlated, which might permit prediction “rescue” even if the directly implicated morphology features are not measured. Because of this, we urge caution when generating hypotheses regarding causal relationships between readouts and individual Cell Painting features.

And the following to the methods:

Machine learning robustness: Systematically removing feature classes

We performed an analysis in which we systematically dropped features measured in specific compartments (Nuclei, Cells, and Cytoplasm), specific channels (RNA, Mito, ER, DNA, AGP), and specific feature groups (Texture, Radial Distribution, Neighbors, Intensity, Granularity, Correlation, Area Shape) and retrained all models. We omitted one feature class and then independently optimized all 70 cell health models as described in the Machine learning framework results section above. We repeated this procedure once per feature class.

All software updates introducing this analysis can be viewed at <https://github.com/broadinstitute/cell-health/pull/143>

In summary there is a very interesting concept here, but for several possible, currently undefined reasons, the authors are reporting a very weak measurement. The authors allude to these limitations, but it would be great if the authors could address these issues and provide a stronger dataset.

We thank the reviewers for their encouraging remarks. We believe that with the added robustness analyses and with increased clarity about the motivation behind the paper, we’ve successfully demonstrated a proof of concept for the approach to predict cell health phenotypes from Cell Painting images. We believe that we’ve provided sufficient evidence to a reader to demonstrate the benefits of the prediction approach. As well, given the additional details describing the Cell Health assay reproducibility, that the paper also successfully introduces a new assay paradigm.

Furthermore, while many of the cell health measurements are definitely weak (and unreliable), it is not fair to generalize all predictions as weak (especially given the sample size limitations).

It is also worth noting that, under the current circumstances, separating the one dataset we have into a train/test set and validating the model in an external set is the best we could do; we do not have additional budget to run further wet lab experiments (which would also face a COVID backlog in our chemical screening group). We agree that additional datasets would benefit the field; our current data is now public, all of our future data will be public (to the extent possible), and we hope that others building on our work will make their data public too to address these questions.

Lastly, in response to the “currently undefined reasons” comment, as well as other comments throughout, we’ve now included a new subsection in the Results/Discussion subsection to more directly answer some of the reasons why many models may have underperformed. Specifically, and as mentioned previously in this response, we perform three distinct robustness analyses: 1) Cell line holdout; 2) feature holdout; 3) sample size titration.

Authors should include representative images of their Cell Health assay in the main figures. A full figure of all labels and examples of manual gating should be included (S1 is too limited) Scale bars need to be included in all images, some are missing in S1

We thank the reviewers for this suggestion. We have since substantially updated figure 1 and supplementary figure S1. We have also added a new supplementary figure S2 as an example of the manual gating strategies, and we have updated all scale bars appropriately. We’ve attached the specific figure updates in an earlier response.

"20x water objective in confocal mode" is not a sufficient level of detail on image acquisition parameters especially considering the lack of representative images. At the very least, NA and if appropriate pinhole size should be reported. Similarly, "9 FOV per well" is not sufficient. Pixel size and FOV area/dimensions are necessary.

We have added these necessary details in their representative methods sections:

We acquired all cell images using an Opera Phenix High Content Imaging Instrument (PerkinElmer) with a 20X water objective (a numerical aperture (NA) of 1.0), in confocal mode (a pinhole size of 50µm). The effective pixel size was 0.65µm/pixel. We acquired images in four channels using default excitation / emission combinations: for the blue channel (Hoechst) 405/435-480; for the green channel (Alexa 488 and CellEvent) 488/500-550; for the orange channel (Alexa 568 and CellRox Orange) 561/570-630 and for the far-red channel (Alexa 647 and DRAQ7) 640/650-760. We applied the Cell Health reagents for cell viability and for cell cycle in two separate plates.

The legends for the different parts of Fig S10 are transposed which makes the figure quite confusing. The authors should amend or clarify the language of "guide perturbation" and "guide profile".

Wow! We thank the reviewers for pointing out this oversight, and for their careful attention to detail. This figure is now completely different after the restructuring of the Drug Repurposing Hub results/discussion section. The legends for all figures are now correct.

EdU is defined after it is abbreviated in methods

We thank the reviewers for noting this. We've now fixed where these acronyms are abbreviated in the methods section and removed their definition in later sections where redundant:

The authors should address the following image processing reproducibility concerns:

Segmentation and feature extraction parameters are not included in the Supplementary Information. Either attach the CellProfiler pipeline or add a table with parameters and settings used for each module.

CellProfiler and Harmony versions are missing.

We thank the reviewers for pointing out these very important omissions. We have since rectified in the methods section:

We built a CellProfiler image analysis and illumination correction pipeline (version 2.2.0) to extract these image-based features (McQuin et al., 2018). We include the CellProfiler pipelines in our github repository.

We developed and ran two distinct image analysis pipelines in Harmony software (version 4.1; PerkinElmer) for each of the Cell Health plates.

We also add the CellProfiler pipelines to our GitHub repository. A pull request introducing this change can be viewed here: <https://github.com/broadinstitute/cell-health/pull/149>

Subpopulation definition (page 14) should be defined in a way that the algorithms (pipelines) could be reproduced, e.g.: "unusually high intensity of Hoechst max" requires a stricter definition.

These definitions are subjective by nature. Gating decisions will be different depending on the scientist performing the image analysis. We feel that the sentence: "We excluded outlier nuclei with unusually high intensity of Hoechst max" conveys this subjectivity well. One of the strengths of the proposed approach to predict cell health phenotypes directly from the Cell Painting images is the removal of gating subjectivity.

Why is the nucleus roundness calculated in PE Harmony and not in the CellProfiler pipeline itself?

We used the nucleus roundness measurements as calculated in PE Harmony to define the "cells selected for cell cycle" subpopulation in the first panel of the Cell Health assay. I.e. this measurement was integral to the Cell Health assay itself. We believe that the addition of example gates (in supplementary figure 2) clears up this confusion.

Reviewers:

Jason Swedlow
Melpi Platani
Erin Diel
Emil Rozbicki

Reviewer #2 (Significance (Required)):

Nature and Significance: This study aims to demonstrate how phenotypic studies using different markers can be combined and linked to deliver wider application and value.

Relationship to Published Work: This study extends previous work from the same group and attempts a novel extension. The approach is a useful concept and potentially important.

Audience: The method this paper proposes will be of interests to scientists involved with drug discovery and/or computational biology.

Reviewer's Expertise: Cell Biology, Imaging, Imaging Informatics, Machine Learning, Computer Vision

We would like to again express thanks to these reviewers for their careful read, very helpful comments, and encouraging remarks.

Reviewer #3 (Evidence, reproducibility and clarity (Required)):

The authors present a novel idea on predicting various cell health readouts based on a general set of markers and cell painting assay. The cell health readouts are based on more specific markers performed in different assays measuring cell proliferation and death. The authors suggest that such an approach can reduce the number of experiments needed. The paper is well written, and the figures are clear and comprehensive.

We thank the reviewer for their helpful comments and encouragement!

****Major comments:****

Some of the health readouts are based on general morphology (cell and nucleus) which can be obtained based on cell painting assay. Although some of these models perform well, it is surprising that the model of nuclear roundness did not perform very well especially for HCC4 (R-square reaching zero). This is surprising as these data can be extracted from cell painting assays. Can the author elaborate on why this is the case?

We agree that the performance of the *live cell roundness* and *nucleus roundness* models were unexpectedly low. One would expect that these shape features as measured by PerkinElmer

Harmony software, would be easily predicted from CellProfiler readouts from the Cell Painting assay.

The roundness property was used in Harmony versions, <=4.9 and calculated with an empirical formula:

$$2*\sqrt{\pi}*\sqrt{(\text{Area}-\text{BorderArea}/2.0)/\text{BorderArea}-0.1}$$

where Area is object area in pixels and BorderArea is border area in pixels (we thank Joe Trask, Olavi Ollikainen, Hartwig Preckel, and Kaupo Palo at PerkinElmer for this information.)

No single feature in the CellProfiler readouts measures roundness directly; instead, CellProfiler will measure a combination of shape features that together could synthesize the idea of “roundness”. However, given that the elastic net approach is well-suited for this type of synthesis, it remains unclear why roundness is not predicted well.

One possible explanation is that shape features are the most different measurements across cell lines and they are measured precisely in both assays. Precise measurements coupled with our training strategy of using all three lines together, might lead to poor performance in predicting certain cell-line intrinsic features.

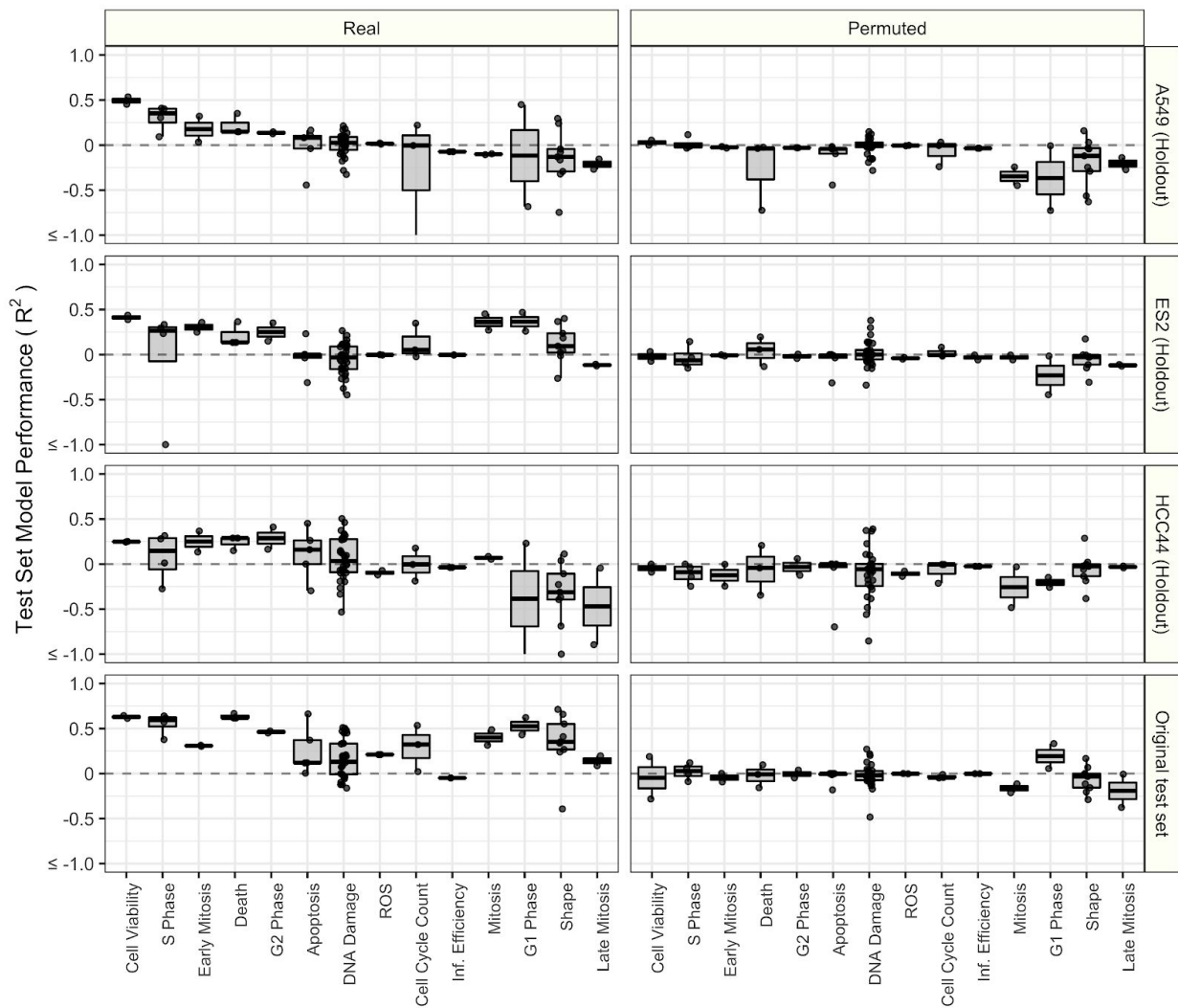
We tested this shape result directly (and also generally to the other cell health features) in a “cell line holdout” analysis, which we describe in more detail in response to the next comment. In this analysis, we tested how well models generalized to cell lines not encountered in the training process. In this analysis, we trained on every combination of two cell lines and applied the trained models to the third. We observed that cell line intrinsic features, like shape, are predicted poorly if a model was not trained using the cell line.

Using elastic net regression models is well-suited to the problem due to the low number of observations. However, there is a significant difference between the performance of different cell lines. Does the performance of the models improve if different models were trained for every cell line? Leave one out approach can be used to accommodate the scarcity of samples.

We thank the reviewer for this important question. We also appreciate how different certain models behaved with certain cell lines. We would like to stress that the results presented here represent a small pilot study that is not meant to optimize model performance. Instead, the motivation of the manuscript is to demonstrate proof-of-concept of the approach to predict specific cell health phenotypes directly from Cell Painting images. We believe that the current results demonstrate positive proof, which warrants an expansion of data collection and an improvement of the classification methodology.

Nevertheless, with our current data, we can answer an important question about the feasibility of signal transfer between cell lines. Therefore, we performed an additional “cell line holdout”

analysis. We believe that the cell line holdout analysis tells us that signals can be transferred across contexts, but that any leading observations must be followed up with experiments performed directly in the cell line of interest. This signal transfer is diluted compared to the original test set performance, but it is also worth noting that the models presented in Supplementary Figure 11 (pasted below) were trained on only 66% of the data in the holdout cell line analysis and 85% of the data in the original analysis.



Supplementary Figure S11: Results from a cell line holdout analysis. We trained and evaluated all 70 cell health models in three different scenarios using each combination of two cell lines to train, and the remaining cell line to evaluate. For example, we trained all 70 models using data from A549 and ES2 and evaluated performance in HCC44. We bin all cell health models into 14 different categories (see Supplementary Table S3 and <https://github.com/broadinstitute/cell-health/6.ml-robustness> for details about the categories and scores). We also provide the original test set (15% of the data, distributed evenly across all cell types) performance in the last row, as well as results after training with randomly permuted data.

This cross-cell-type analysis yields worse performance overall. Nevertheless, despite the models never encountering certain cell lines, and having fewer training data points, many models still have predictive power across cell line contexts. Note that we truncated the y axis to remove extreme outliers far below -1. The raw scores are available on <https://github.com/broadinstitute/cell-health>.

And we add the following text to the results section:

We performed a series of analyses to determine certain parameters and options that are likely to improve models in the future. First, we performed a “cell line holdout” analysis, in which we trained models on two of three cell lines and predicted cell health readouts on the held out cell line. We observed that certain models including those based on viability, S phase, early mitotic and death phenotypes could be moderately predicted in cell lines agnostic to training (**Supplementary Figure 11**). Not surprisingly, shape-based phenotypes could not be predicted in holdout cell lines, which emphasizes the limitations of transferring certain cell-line specific measurements across cell lines.

All software updates introducing this analysis can be viewed at <https://github.com/broadinstitute/cell-health/pull/143>

The authors chose to validate based on the number of live cells as it is one of the best models. However, this readout can be obtained using simple viability assays. It would be more convincing to validate on a more complex phenotype that can only be attained using imaging such as #gH2AX spots.

It is worth noting that we do also show generalizability in the Drug Repurposing Hub for two other models: *ROS* and *G1 cell count*. We show that proteasome inhibitors significantly induce high ROS and PLK inhibitors restrict entry to G1. We have also added enrichment tests demonstrating high statistical significance for these compound mechanisms.

While we recognize that these two examples provide anecdotal evidence, they suggest the ability and power of the approach to assign phenotypes to Cell Painting images.

Nevertheless, we thank the reviewer for bringing up this critical point and certainly appreciate the benefit of validating a gH2AX model. Therefore, we’ve added a similar analysis in which we demonstrate generalizability of the top performing gH2Ax model: *Number of gH2AX spots in G1 cells*. We discuss these changes in an updated section:

We also chose to validate three additional models: *ROS*, *G1 cell count*, and *Number of gH2AX spots in G1 cells*. We observed that the two proteasome inhibitors (bortezomib and MG-132) in the Drug Repurposing Hub set yielded high *ROS* predictions (OR = 76.7; $p < 1 \times 10^{-15}$) (**Figure 4C**). Proteasome inhibitors are known to induce ROS (Han and Park, 2010; Ling et al., 2003). As well, PLK inhibitors yielded low *G1 cell counts* (OR = 0.035; $p = 3.9 \times 10^{-8}$) (**Figure 4C**). The

PLK inhibitor HM-214 showed an appropriate dose response (**Figure 4D**). PLK inhibitors block mitotic progression, thus reducing entry into the G1 cell cycle phase (Lee et al., 2014). Lastly, we observed that aurora kinase and tubulin inhibitors yielded high *Number of gH2AX spots in G1 cells* predictions (OR = 11.3; $p < 1e-15$) (**Figure 4E**). In particular, we observed a strong dose response for the aurora kinase inhibitor barasertib (AZD1152) (**Figure 4F**). Aurora kinase and tubulin inhibitors cause prolonged mitotic arrest, which can lead to mitotic slippage, G1 arrest, DNA damage, and senescence (Orth et al. 2011; Cheng and Crasta 2017; Tsuda et al. 2017).

We also modify the abstract summarizing this result:

For Cell Painting images from a set of 1,500+ compound perturbations across multiple doses, we validated predictions by orthogonal assay readouts, and by confirming mitotic arrest, ROS, and DNA damage phenotypes via PLK, proteasome, and aurora kinase/tubulin inhibition, respectively.

And we add this analysis to an updated Figure 4:

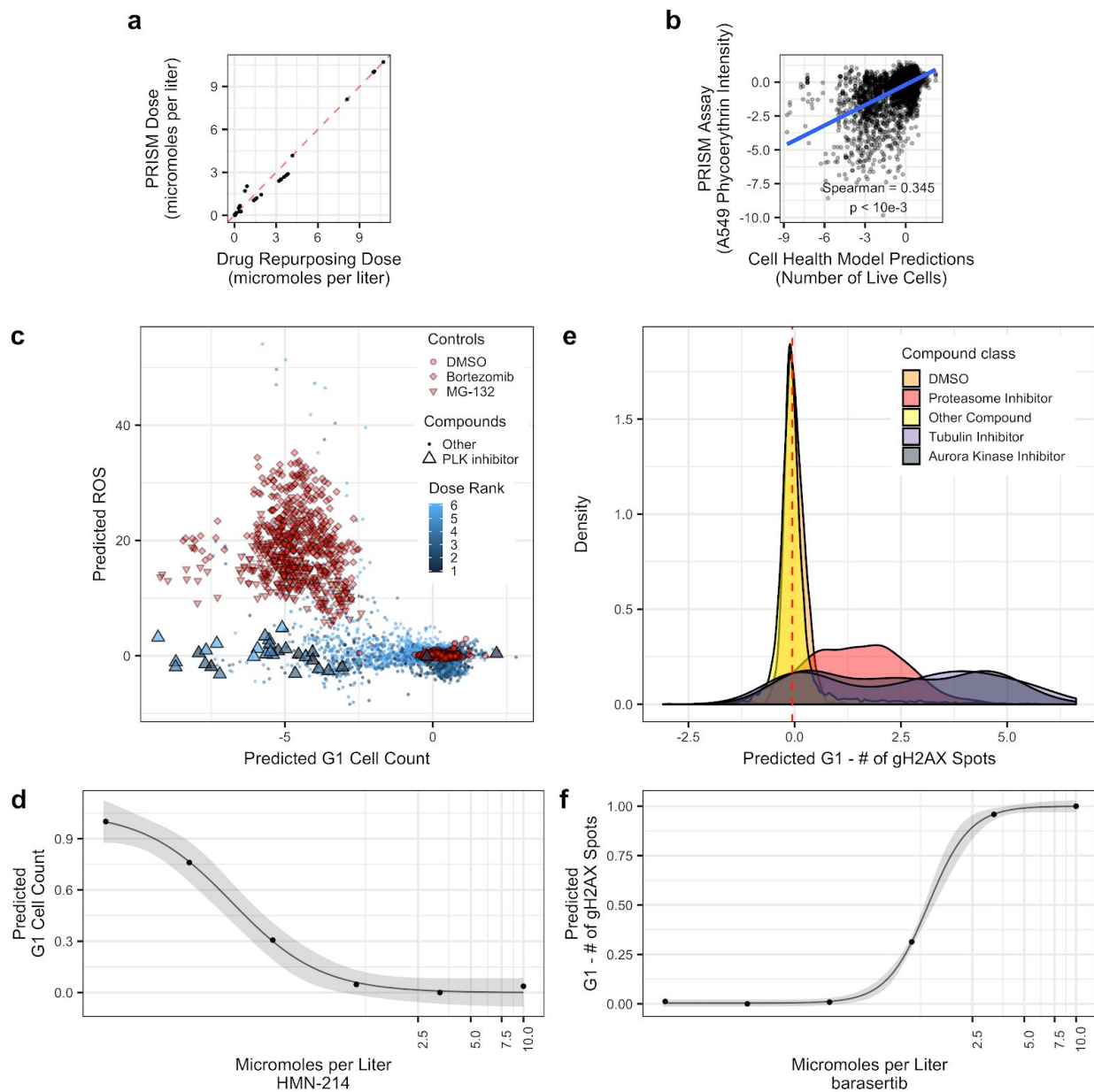


Figure 4: Validating Cell Health models applied to Cell Painting data from The Drug Repurposing Hub. The models were not trained using the Drug Repurposing Hub data. **(a)** The results of the dose alignment between the PRISM assay and the Drug Repurposing Hub data. This view indicates that there was not a one-to-one matching between perturbation doses. **(b)** Comparing viability estimates from the PRISM assay to the predicted *number of live cells* in the Drug Repurposing Hub. The PRISM assay estimates viability by measuring barcoded A549 cells after an incubation period. **(c)** Drug Repurposing Hub profiles stratified by *G1 cell count* and *ROS* predictions. Bortezomib and MG-132 are proteasome inhibitors and are used as positive controls in the Drug Repurposing Hub set; DMSO is a negative control. We also highlight all PLK inhibitors in the dataset. **(d)** HMN-214 is an example of a PLK inhibitor that shows strong dose response for *G1 cell count* predictions. **(e)** Tubulin and aurora kinase inhibitors are

predicted to have high *Number of gH2AX spots in G1 cells* compared to other compounds and controls. (f) Barasertib (AZD1152) is an aurora kinase inhibitor that is predicted to have a strong dose response for *Number of gH2AX spots in G1 cells* predictions.

All software updates required to update these figures can be viewed at <https://github.com/broadinstitute/cell-health/pull/145>

It is also worth noting that collecting more data for this manuscript is not currently feasible given the amount of projects backlogged from COVID. We feel that given that the motivation of the project is to demonstrate feasibility of the approach, with our current training/testing machine learning framework and the application to Drug Repurposing Hub data is sufficient.

The text would benefit from expanding the discussion to include the advantages and limitations of their approach.

We thank the reviewer for bringing up this concern, and we agree that it is worth an increased discussion about advantages and limitations of the approach. Indeed, we've added a full new results/discussion subsection directly testing many of the assumptions for why some models performed well and others didn't. The new section introduces many model limitations:

We performed a series of analyses to determine certain parameters and options that are likely to improve models in the future. First, we performed a "cell line holdout" analysis, in which we trained models on two of three cell lines and predicted cell health readouts on the held out cell line. We observed that certain models including those based on viability, S phase, early mitotic and death phenotypes could be moderately predicted in cell lines agnostic to training (**Supplementary Figure 11**). Not surprisingly, shape-based phenotypes could not be predicted in holdout cell lines, which emphasizes the limitations of transferring certain cell-line specific measurements across cell lines. We also performed a systematic feature removal analysis, in which we retrained cell health models after dropping features that are measured from specific groups, compartments, and channels. We observed that many models were robust to dropping entire feature classes during training (**Supplementary Figure 12**). This result demonstrates that many Cell Painting features are highly correlated, which might permit prediction "rescue" even if the directly implicated morphology features are not measured. Because of this, we urge caution when generating hypotheses regarding causal relationships between phenotypes and individual Cell Painting features. Lastly, we performed a sample size titration analysis in which we randomly removed a decreasing amount of samples from training. For the high and mid performing models we observed a consistent performance drop, suggesting that increasing sample size would result in better overall performance (**Supplementary Figure 13**).

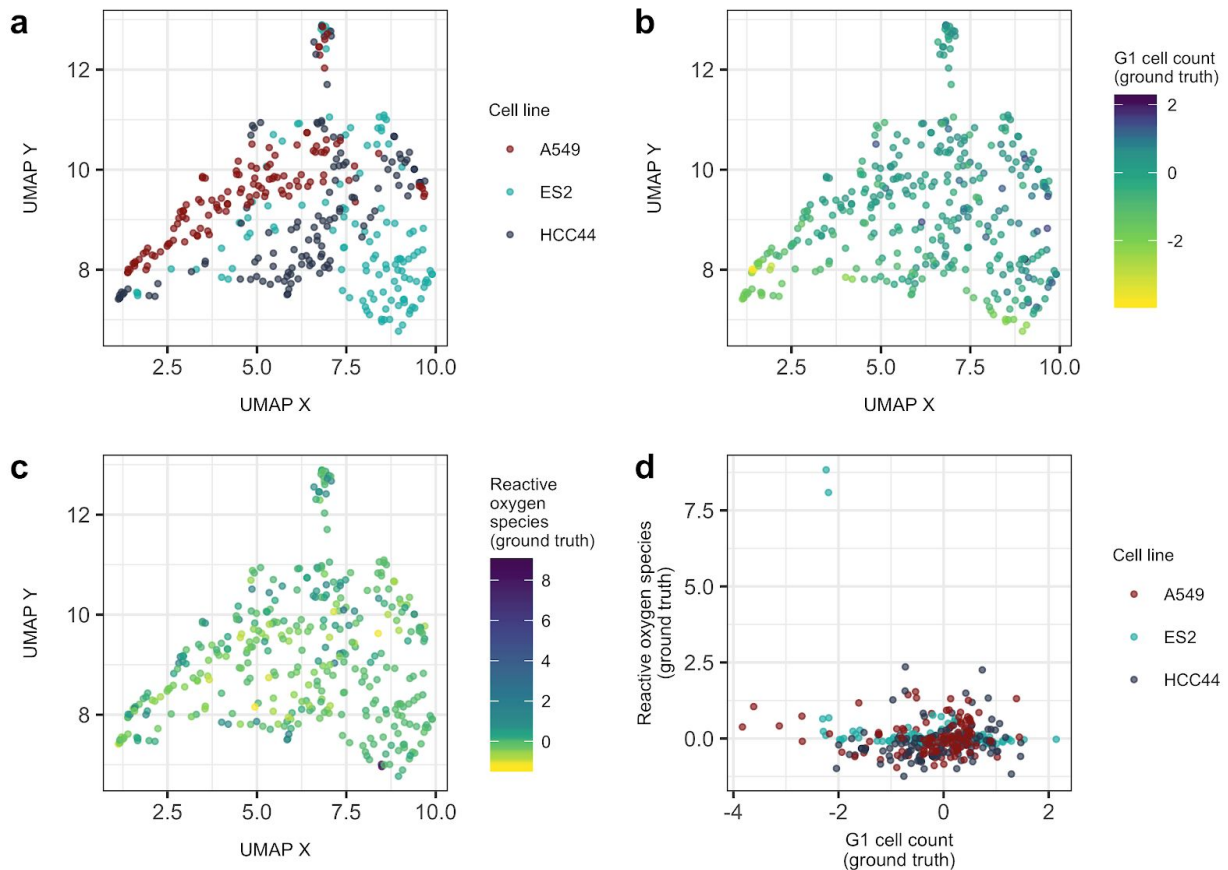
****Minor comments****

Page 8: The authors visualize the predicted G1 cell count and ROS when overlaid on a UMAP based on cell painting data from Drug Repurposing Hub. How these visualisations look like if applied to the original CRISPR training data.

We address this comment by adding a supplementary figure showing ground truth G1 cell count and ROS readouts.

We applied uniform manifold approximation (UMAP) to observe the underlying structure of the samples as captured by morphology data (McInnes et al., 2018). We observed that the UMAP space captures gradients in predicted G1 cell count (Supplementary Figure S14A) and in predicted ROS (Supplementary Figure S14B). We also observed similar gradients in the ground truth cell health readouts in the CRISPR Cell Painting profiles used for training cell health models (Supplementary Figure S15). Gradients in our data suggest that cell health phenotypes manifest in a continuum rather than in discrete states.

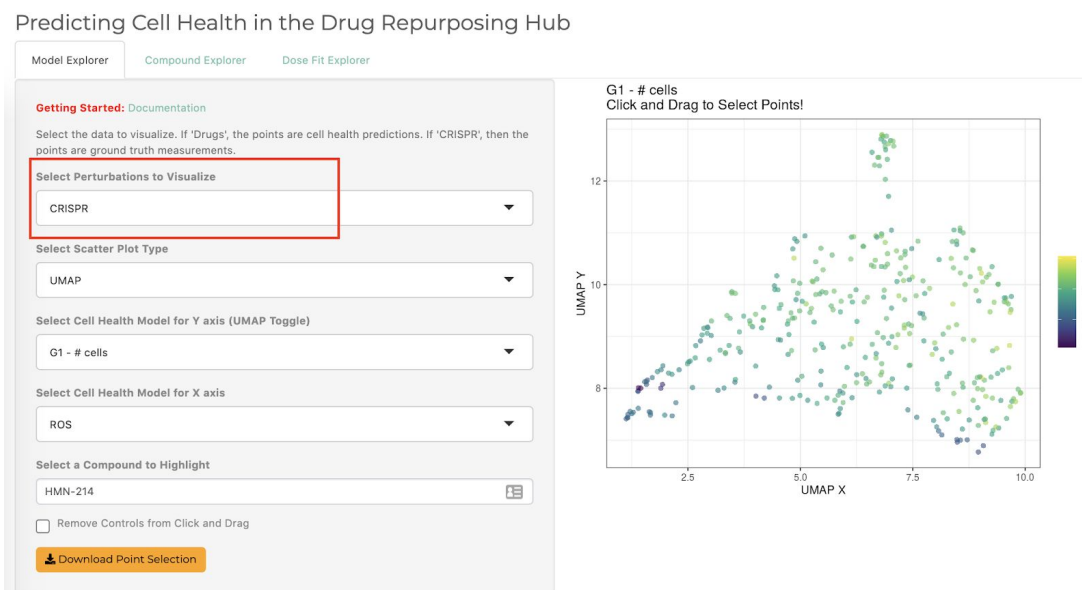
Where Supplementary Figure 15 is pasted below:



Supplementary Figure S15: Applying a Uniform Manifold Approximation (UMAP) to the Cell Painting consensus profile data of CRISPR perturbations. UMAP coordinates visualized by (a) cell line, (b) ground truth G1 cell counts, and (c) ground truth ROS counts. (d) Visualizing the

distribution of ground truth ROS compared against G1 cell count. The two outlier ES2 profiles are CRISPR knockdowns of *GPX4*, which is known to cause high ROS.

We have also added the option to explore the CRISPR profile Cell Health ground truth in our shiny app <https://broad.io/cell-health> (screenshot pasted below)



Modifications to the software introducing these changes can be viewed at <https://github.com/broadinstitute/cell-health/pull/141>.

The second part of the last paragraph on page 8 is confusing as it is not related to the first part using the PRISM data.

We thank the reviewer for noting this. We agree that the clarity of this section could be improved. We have now completely reworked the final section of applying the cell health models to the Drug Repurposing Hub data.

In particular, we've moved the PRISM data section as the first, most simple model to validate, and moved these results to Figure 4. We then describe validation for three other models: *ROS*, *G1 cell count* and *Number of gH2Ax spots in G1 cells*. And we end with the UMAP discussion, which is the original second part of the last paragraph on page 8.

The PRISM section now reads:

We first chose a simple, high-performing model to validate. The *number of live cells* model captures the number of cells that are unstained by DRAQ7. We compared model predictions to orthogonal viability readouts from a third dataset: Publicly available PRISM assay readouts, which count barcoded cells after an incubation period (Yu et al., 2016). Despite measuring

perturbations with slightly different doses and being fundamentally different ways to count live cells (Figure 4A), the predictions correlated with the assay readout (Spearman's Rho = 0.35, $p < 1 \times 10^{-3}$; Figure 4B).

Reviewer #3 (Significance (Required)):

This approach can be of wide interest as it is easy to implement, cost-effective and lead to interpretable models. It would be interesting to see if the results improve when increasing the sample size. Another aspect that can be useful to investigate in the future is whether including a separate marker that indicates infected cells only in the more detailed assays would result in better accuracies.

We thank the reviewer for their enthusiasm and for this concluding idea. Indeed, we also feel that including a separate marker to indicate infected cells could lead to improved accuracy. We add this thought to the concluding section as a future direction. The full updated conclusion reads as follows:

We have demonstrated feasibility that information in Cell Painting images can predict many different Cell Health indicators even when trained on a small dataset. The results motivate collecting larger datasets for training, with more perturbations and multiple cell lines. These new datasets would enable the development of more expressive models, based on deep learning, that can be applied to single cells. Including orthogonal imaging markers of CRISPR infection would also enable us to isolate cells with expected morphologies. More data and better models would improve the performance and generalizability of Cell Health models and enable annotation of new and existing large-scale Cell Painting datasets with important mechanisms of cell health and toxicity.

RE: Manuscript #E20-12-0784

TITLE: "Predicting cell health phenotypes using image-based morphology profiling"

Dear Dr. Way:

I am pleased to accept your manuscript for publication in Molecular Biology of the Cell.

Sincerely,
Alexander Mogilner
Monitoring Editor
Molecular Biology of the Cell

Dear Dr. Way:

Congratulations on the acceptance of your manuscript.

A PDF of your manuscript will be published on MBoC in Press, an early release version of the journal, within 10 days. The date your manuscript appears at www.molbiolcell.org/toc/mboc/0/0 is the official publication date. Your manuscript will also be scheduled for publication in the next available issue of MBoC.

Within approximately four weeks you will receive a PDF page proof of your article.

Would you like to see an image related to your accepted manuscript on the cover of MBoC? Please contact the MBoC Editorial Office at mboc@ascb.org to learn how to submit an image.

Authors of Articles and Brief Communications are encouraged to create a short video abstract to accompany their article when it is published. These video abstracts, known as Science Sketches, are up to 2 minutes long and will be published on YouTube and then embedded in the article abstract. Science Sketch Editors on the MBoC Editorial Board will provide guidance as you prepare your video. Information about how to prepare and submit a video abstract is available at www.molbiolcell.org/science-sketches. Please contact mboc@ascb.org if you are interested in creating a Science Sketch.

We are pleased that you chose to publish your work in MBoC.

Sincerely,

Eric Baker
Journal Production Manager
MBoC Editorial Office
mbc@ascb.org
