nature portfolio

Peer Review File

SPACe: an open-source, single-cell analysis of Cell Painting data

Corresponding Author: Dr Michael Mancini

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors present a new open-source software package, SPACe, for phenotypic profiling analysis of high-content image data. The authors demonstrate several advantages of SPACe over the widely used CellProfiler software package in the context of high-throughput drug screening. These include speed of image segmentation, ability to run the software from a standard desktop, and performance enhancements gained by leveraging distributions of single-cell variation within and across treatments. Moreover, SPACe extracts fewer features from raw image data, and the authors confirm that this smaller feature set is equally informative to CellProfiler in terms of ability to assign MOA on the same set of images. The authors test several different cancer cell lines/models with a variety of reference compounds and draw interesting conclusions regarding cell-type specific responses to different classes of drugs. While the manuscript could be improved by clarifying some of the technical aspects of the analysis for the reader, overall SPACe seems to be a promising software for the field of high-throughput, high-content imaging and phenotypic profiling.

Specific comments/questions:

1) Fig S1: “Percent matching” and “Percent replicating”: Could the authors improve the legend for Figure S1 to indicate what these concepts mean? While this is explained in the text, the figure would be easier to interpret if these were briefly clarified in the legend.

2) Lines 145-146: “In percent matching, both SPACe and CellProfiler mean well values ranked significantly better than SPACe EMD values (Figure S1C-D).”
Could the authors elaborate on this point, perhaps in the Discussion? This is an interesting distinction that intuitively makes sense, and it highlights another potential advantage of using well distributions in addition to well central tendencies. This seems like a useful observation because molecules annotated with the "same MOA" can broadly affect the same biological pathway in a similar way, yet their precise molecular activity is rarely exactly the same – and some (many?) molecules may have off-target effects that have not been fully characterized.
This analysis would suggest that measures of central tendency can broadly classify molecules with similar MOAs (though with limited accuracy), while differences in distributions as measured by EMD can reveal more subtle phenotypic differences that may reflect differences in their molecular activities in the cell (which could provide a basis for further mechanistic studies).
In contrast, however, the RF analysis (Fog. S3) finds SPACe EMD to slightly outperform both CellProfiler mean and SPACe mean. So perhaps this distinction is not so clear? (Also see comments 7 and 9 below.)

3) Figure S2: It seems like the heatmaps in panels A-C should be symmetrical matrices, but the clustering seems different on the two axes. Could the authors explain more clearly what is going on here?

4) Figure S2: It looks like there are a lot more strongly correlated and strongly anti-correlated features based on SPACE analysis. Could the authors explain how they chose the 400 features?
The manuscript states that the feature set contains a (somewhat) higher proportion of uncorrelated features, but the authors also point out (lines 161-163) that “the feature set contains both redundancy and uniqueness sufficient to recapitulate the CellProfiler generated results from the reference dataset, similar to other published work that reduced the CellProfiler feature set to a little over 600 (28).”

However, usually the point of dimensional reduction is to reduce redundancy by culling uninformative features (or transforming the features using something like PCA and keeping only those features that explain most of the variation in the data). Did the authors consider comparing a reduced set of features with only low correlation vs. a set of similar size with more redundancy? Or choosing a subset of 400 features from CellProfiler with minimal redundancy and comparing that to the feature set used by SPACe?

5) Figure S2E: For panels A-C, the Spearman correlation ranges from -1 to +1. But in panel E, the scale ranges from 0-1. Do they mean the absolute value of the correlation here? How does this figure show “enrichment of CP features below 0.2 and SPACe above 0.8” (lines 154-156)? Also – the authors refer to this as a histogram, but it is not a histogram.

6) Figure S3: The main text (lines 167-168) indicate that "Each RF model was trained with half of the treatment replicates, randomly selected for each model replicate."
It was not clear whether this means (a) one replicate for all treatments, or (b) both replicates for half the treatments? If the reproducibility varies (which it does) depending on the analysis, then wouldn't using option (b) necessarily affect the results? The text, Methods, and Figure S3 legend state that 5 RF models were generated per dataset using bootstrapping, but the size of the training set is not specified, or how subsetting was performed. Could the authors please describe the RF analysis more thoroughly in the Methods section?

7) Figure S3: Panel B makes SPACe EMD look best, but what "Rank" means is not clear from reading either the text or the legend (it seems to be a comparison between the three analysis methods, but this should be clarified).

8) Figure S3: The methods don’t look that different from panel A; while C shows that SPACe mean or EMD most often slightly outperform CellProfiler mean, panels D-F look almost identical qualitatively across all MOAs. Is this simply a reflection of the information content available in the datasets themselves, or some congruence in feature extraction? That is, is there an intrinsic limit in how much information can be gleaned from the CellPainting assay, or will exploring further improvements in data analysis yield significant gains in classification performance, eventually?

9) Figure S3: The manuscript says that for MOA prediction using RFs (lines 493-494), “Model performance was evaluated by the percent of correct predictions (accuracy) and a confusion matrix generated.” This brings two questions to mind:
(a) What about mis-classification? Is there an “unclassified” category in addition to the “none” category (which seems to distinguish “active” vs. ‘“inactive”, which one assumes means “not distinguishable from control”)?
(b) Since a confusion matrix was generated for each model, presumably it would be possible to include some kind of ROC plots comparing accuracy-specificity or precision-recall? Would this add any value to the analysis?

10) Fig 2B: Is the clustering done for each “channel” separately, or were these separated after clustering on the combined data? What does the category "N/A" represent?

11) Figure 2D provides a concise summary of the EMD data. Did the authors compare performance of signed vs. unsigned EMD? Intuitively, one would think that including the sign (which is based on the direction of median change, correct?) would be advantageous, but it would be nice to see a comparison.

12) Could the authors list the stain and compartments affected in the assay somewhere in the manuscript? While CellPainting is a broadly available assay, it's not necessarily obvious where to go to find this information and it would be nice if the authors included a short description of the assay itself, and which sub-compartments or cellular components are surveyed (e.g. Both actin and tubulin cytoskeletons? Both peroxisomes and lysosomes? etc.).
Reminding the reader of which compartments are included in the CellPainting assay would be particularly helpful since the manuscript notes that SPACe segments the cell, cytoplasm, nucleus, nucleoli, and mitochondria. Are other compartments also segmented, or are the stain signals just analyzed for intensity and texture features without segmentation? Is this analysis tuned specifically to the CellPainting platform, or is it generalizable to markers of additional cellular structures that are not included in CellPainting (e.g. tubulin, biomolecular condensates such as stress granules)?

13) Is there any one cell line that shows greater responses across a larger number of bioactive chemicals? That is, if you had to choose one cell line to screen, or 2-3, say due to limited financial resources, which one(s) would they recommend? This could be a nice addition to the Discussion.

14) Do the authors discuss whether the software detects the presence of plate effects in Row/Columns effects, or batch effects? Did the authors consider plate design and how this would be implemented into the SPACe software?

15) Lines 454-455: "The reference distribution is here defined as the median of the DMSO distribution in each experiment. It is very confusing to label the reference (DMSO control) distribution as "the median of the DMSO distribution". This terminology obfuscates the narrative in many places. Why not just call it "the reference distribution"?

16) Methods section, "Statistical analysis" (beginning at line 497): “To compare fingerprints, Euclidean distance was measured between EMDs of all features in the treatment wells and the median DMSO control control wells”. Could the authors state how many DMSO wells were used? How were biological and technical replicates handled in the Euclidian distance calculation?

17) Minor editorial issues: (a) Figure 1D has an error in it — in panel D, top is percent replicating and bottom is percent matching, but the legend says both are percent replicating; (b) Fig. 5D legend is missing; (c) Figure legends not uniformly

formatted (capitalize each panel sentence or not?); (d) Line 134: “resources availability”, is this a typo?; (e) Line 469: “all datasets were processes” should be “processed”?

Reviewer #2

(Remarks to the Author)
"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

Reviewer #3

(Remarks to the Author)
Overall, this manuscript is well written and describes a new computational tool, SPACe, that should be helpful to those performing high-content image analyses, which is especially relevant with the ever growing use of cell painting and related morphological assays in early drug discovery. The manuscript is well written, reports a welcome scientific advance for cell image analysis, and is supported by several use cases including applicability to multiple cell lines (not just U-2 OS). However, in my best judgement, I would recommend that several major and minor comments be addressed before recommending publication.

Major comments:

*A main theme of this manuscript is the comparison to Cell Profiler, and the authors make it a point that SPACe performs ~10X more efficiently than Cell Profiler. Notably, SPACe extracts 400 features, while the authors mention that CellProfiler extracts ~4000 features. Could the authors please comment if this computational efficiency is simply due to the reduced number of features that is being extracted? If not, perhaps make this more explicit. It would enhance the manuscript if the authors could further describe the details of the comparison, and whether the comparison was an "apples to apples" comparison. For example, would CellProfiler be just as efficient if it was asked to extract the same 400 features that SPACe used?

*The authors highlight berberine chloride as a case study. This compound has a promiscuous MOA and interestingly, it is also a highly colored compound. Did the authors consider if the apparent phenotypes observed were due to a true biological effect, or due to compound-mediated interference such as quenching or auto-fluorescence, or a combination? It is known that some compounds can produce apparent phenotypes in staining assays simply because the compounds themselves act as dyes, which could perhaps explain its high reproducibility and activity across many cell lines. If the activity is somehow related to compound interference, and SPACe classifies it as a MOA, that in itself is also a significant finding and a cautionary note to the CP community.

*"Compounds with discordant replicates or with obvious imaging artifacts were also excluded after manual inspection..." Could the authors please describe their criteria for discordance and determining "obvious artifacts"? Where images manually inspected? Whatever their approach, would it be feasible for a large screen? Related, I would actually recommend the authors include some examples of artifacts in the supplemental and how SPACe would have analyzed them. This would in fact be valuable information for users to know potential failure modes for this software, so that such "bad wells" can be identified.

*Several times in the manuscript, SPACe is compared to Cell Profiler. Could the authors please comment on whether there are any situations where Cell Profiler would actually be the more appropriate analysis tool? If so, these reasons should be mentioned to provide an overall balanced perspective.

*The authors note in the methods that cells were cultured in "ATCC suggested media". It is well known that the choice of cell culture media can impact compound activity, especially when the testing was performed on a cellular metabolism-based focused screening collection. Furthermore, the amount of serum can impact cell growth, for example, which may have an effect on the cell population distribution (cell state). To aid in interpretation of their results, and enhance reproducibility, it is recommended that the authors include a complete table (in the supplemental) with the specific media for each cell line (including any additives or serum).

*The authors describe one potential benefit of SPACe as the ability to analyze single-cell data. However, after reading the manuscript, it is still not clear how SPACe takes into account single cell data, and how analyzing data at the single-cell level would outperform well-level analyses. I would encourage the authors to better articulate these points, especially for less computational-savvy audiences, and perhaps articulate the potential benefits of single-cell analyses with a clear example (or if they believe one of their panels already does this, then better articulate to lay audiences).

*The authors describe a QC step that automatically removes wells with low object count. I presume this is to eliminate cytotoxic compounds or compounds that affect cell adherence. Are there any instances where this step could exclude potentially valuable data? If so, they should consider noting in the text.

*"It is unlikely to identify compounds that would act in a universal manner and can be used as controls across all

experimental models. This complicates the analysis, prediction, and interpretation of the MOA for compounds when based uniquely on phenotypic screening in a single cell model". This is hardly surprising. Do the authors, or others in the community, really expect that compounds would produce the same morphological changes (universal) across a variety of cell, and could they provide more context about why these analyses were done? In most practices that I am aware of, one would test a well-annotated MOA compound set in each cell line to define the morphological changes corresponding to the MOA according, and I would be skeptical of applying morphological fingerprints across cell lines. Related, this type of analysis has already been reported by the Harrill group at the US EPA, and I would recommend the authors consider citing their important work related to this topic.

Minor comments:

*"Ten compounds were the most toxic across all models (auranofin, SF1670, plumbagin, PR-619, CB-5083, PFK158, eeyarestatin 1, digitoxin, paclitaxel, and TG101348) and should be tested at lower concentrations to measure changes at non-toxic levels, as some of them show potentially very interesting phenotypes in the surviving cells." The significant phenotypes associated with cellular injury compounds has recently been described (PMID 36914634). The authors should consider citing this work. Have the authors considered analyzing some of this reference data using SPACe?

*I found the manuscript very well written, but at times it was hard to relate some of the technical terminology to how it impacts the end result (usually MOA calling). One recommendation that could enhance this manuscript's impact and appeal to a broader audience of end-users (e.g., bench scientists; non-computational biologists) would be to add less technical text that more clearly explain the practical implications of the various technical advances in SPACe.

*Throughout the manuscript, the authors attribute MOAs to reference compounds. The authors may want to state the source of these MOA annotations and provide reasonable caveats, as I would actually be hesitant to take some of these prescribed MOAs at face value. For example, while some of these compounds have well-defined MOAs, some of these compounds have well-known promiscuous MOAs (e.g., rottlerin). The authors should also consider whether the targets/MOAs, often determined at nM concentrations, are still relevant at uM concentrations.

*The authors should consider/add text that compounds that affect cellular adherence (but non-cytotoxic) may also lead to low object counts.

*"this implementation choice does not reduce in any way..." this is quite a strong statement. Is there really no possibility scenario that this implementation choice affects downstream performance?

*The authors use "small molecule inhibitors..." whereas "small-molecule inhibitors" is more generally preferred.

*Typo: "Perhaps interestingly, only seven compounds in the screed..."

*The correct name of the cell line is "U-2 OS" according to ATCC, not U2OS. The same for HepG2, which is actually Hep G2 according to ATCC. A minor point, but I recommend the authors check that these and other cell lines are correctly named.

*The authors use the term dose-response or dose several times, which should be reserved for actual administration of a drug dose in an in vivo setting; the more appropriate terminology for cell culture experiments should be "concentration-response" and its variations.

*The beginning of many sentences are not capitalized in many of the figure legends, most notably for figure panels. This should be corrected and consistent.

*Do the authors have a method they could cite for their mycoplasma testing?

*What was the source of the chemicals besides the Cayman metabolism library? This should be noted, especially for the key compounds in their study. What QC was performed on these compounds?

*"37C/5% CO2" should include the degree symbol, and chemical formulas should have subscripts. The authors should carefully review their manuscript and SI materials for similar issues.

*There should be spaces between the concentration number and the concentration units (e.g., "10 nM" instead of "10nM")

*Overall, the figures are beautiful, but boxes around several panels are clearly visible (e.g., Figure 2B, Figure 3A,B,D). Recommend these be touched up.

Reviewer #4

(Remarks to the Author)
The manuscript by Stossi et al. describes a Python-based software package, called SPACe, to perform single cell image-based morphology analysis for image data from CellPainting assays. The authors claim that their software is 10x faster at analysis than CellProfiling without loss in MOA results. They test their software on ~20 cell lines. The authors attribute their improvement over CellProfiling to two factors: 1) Utilization of a GPU and 2) Utilization of CellPose for segmentation.

The comparison between SPACe and CellProfiler is not a fair comparison. In line 152 the authors mention that SPACe extracts ~400 features while CellProfiler extracts ~4000 features. The difference between these two is 10X, which is oddly the same factor of improvement the authors are claiming in computing time. Is the computational saving a result of their implementation or from reduced feature selection? Since the features extracted by SPACe is a subset of CellProfiler's, it should be obvious for the authors to reduce the CellProfiler pipeline to the same feature set.

Furthermore, it's unclear if the impact of this work is as great as the authors claim. Computational power is improving all the time. While CellProfiler is a clunky piece of software, and improvements are always welcome, it's unclear that processing speed is limiting any biological findings. Importantly, the reported SPACe software is not enabling new findings.

The utility of CellProfiler is that it presents the image processing operations in an accessible way to biologists. The software presented by the authors method require significantly more expertise to operate.

Line 342 – I would say a major limitation to the wider adoption of CP-like phenotypic screening is the costs associated with the HCS hardware, cells, reagents, and expertise. Computing cost is less significant than these costs, and a full analysis is conducted infrequently.

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have made a good-faith and largely satisfactory effort to address most of the comments from the original round of reviews. We have only minor comments on the revised manuscript:

1. Figure 2B and Fig 3: Label (if room) and legend (at least) should specify “Signed EMD”.

2. Figure S2C: This third heatmap in the figure is still not symmetrical; different clustering is used on each axis (this is obvious by eye).

3. Revised manuscript, lines 155-167 and Figure S2E: Referring to Reviewer 1, Question #5 in the first review cycle, pertaining to the "histogram" or "frequency distribution" of Spearman correlation between features:

a. Thanks for clarifying that the diagram represents a frequency distribution, and that overlaid histograms were converted to lines for better visual comparison. Perhaps the usage of "histogram" vs. "frequency distribution" seems semantic to the authors, and it is a minor point, but we would expect a histogram to show a "grouped frequency distribution", in which the area of adjacent rectangles is proportional to the frequency (or relative frequency) of observations in each interval. Especially considering that the feature sets are of different magnitudes, and apparently not binned to the same interval sizes, calling this a histogram seemed visually incongruous. Also note that, technically, this is a "relative frequency distribution".

b. Line 158: When the authors say "A large fraction of features is highly correlated (>0.8), regardless of the analysis method being used", how large is this number really? Perhaps not a "large" fraction, but a "substantial" or even "minor" fraction? From Fig. S2E, it looks like maybe around 20% of SPACe features, but a much smaller proportion of CP features (even though in absolute terms still large given the size of the full feature set)? This may be why the authors say that there is an "enrichment" of features for CP below 0.2 and SPACe above 0.8, but this phrasing still trips us up. The reason is that Fig. S2E shows only a slightly higher proportion of CP features with very low pairwise correlations (Spearman correlation < 0.2) relative to the SPACe feature set; and this difference is probably negligible taking all other factors into account (?). In contrast, SPACe gives rise to a pretty flat feature distribution from a correlation of 0 to ~0.9, so any "enrichment" of SPACe features with correlation > 0.8 is specifically relative to the CP distribution. It would seem simpler just to point out that while the SPACe feature set contains a greater proportion of highly correlated features (Spearman correlation > 0.8) than the CP feature set (Fig. S2E), SPACe contains a significantly higher proportion of "unique" features (defined as having no correlation with other features > 0.95) (Fig. S2D). Again, relatively minor point, but perhaps this description could be improved here.

4. Revised manuscript, lines 505-510: Referring to Reviewer 1, Question #15 in the first review cycle: This pertains to the definition of the "reference distribution" for DMSO controls. The manuscript says,

"The QC routine is designed to establish a reliable ground truth for single cell distributions in control samples (e.g., DMSO). The idea stems from our prior publication (19) that demonstrated the value of distribution analysis as a quality control step for high throughput microscopy assays and subsequent single cell analyses. The QC step establishes a reference distribution for the DMSO negative control wells (eliminating outliers because of low object count or aberrant phenotypic profile). The reference distribution is defined as the median of the DMSO distribution in each experiment."

What was not clear in the first round of review is exactly how the "median of the DMSO distribution in each experiment" is computed to establish the reference distribution (though this must seem obvious to the authors). The text does not clarify the procedure used to obtain this, but perhaps it is described in the mentioned prior publication? We infer that the authors compute a "per-well" distribution for each feature across all of the DMSO wells on a plate, and that these per-well distributions become the basis for the "median" reference distribution. Is that correct? It would be more clear if the authors explained this more clearly in the Methods, which would alleviate the original source of the confusion.

5. Reviewer 1, Question #14, regarding controlling for row/column or batch effects:

The authors state that control wells can be anywhere on a plate, which is true, but we and others have noticed that there can be major differences across plates due to issues with reagent dispensing or differential humidity. There can also be differences in the DMSO "median" reference distribution for the controls across different plates. Without normalization of some kind to account for such differences, it is not clear how the QC step would address this. The authors state in their rebuttal that QC will flag bad wells, fair enough, but while "plotting the EMD data" could indeed detect edge/column/row effects and probably also plate effects (?), this does not really address the issue. Are we missing something?

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
I have reviewed the revised manuscript and materials, as well as the response letter. I believe the revisions and author comments adequately address my comments. Recommend for publication.

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We would like to thank the reviewers for the comments and suggestions that have greatly improved the manuscript. Please see our detailed responses to each request in blue.

Reviewer #1:

1) Fig S1: “Percent matching” and “Percent replicating”: Could the authors improve the legend for Figure S1 to indicate what these concepts mean? While this is explained in the text, the figure would be easier to interpret if these were briefly clarified in the legend.

We have added a basic explanation in the legend, as suggested.

2) Lines 145-146: “In percent matching, both SPACe and CellProfiler mean well values ranked significantly better than SPACe EMD values (Figure S1C-D).”
Could the authors elaborate on this point, perhaps in the Discussion? This is an interesting distinction that intuitively makes sense, and it highlights another potential advantage of using well distributions in addition to well central tendencies. This seems like a useful observation because molecules annotated with the "same MOA" can broadly affect the same biological pathway in a similar way, yet their precise molecular activity is rarely exactly the same – and some (many?) molecules may have off-target effects that have not been fully characterized.
This analysis would suggest that measures of central tendency can broadly classify molecules with similar MOAs (though with limited accuracy), while differences in distributions as measured by EMD can reveal more subtle phenotypic differences that may reflect differences in their molecular activities in the cell (which could provide a basis for further mechanistic studies).
In contrast, however, the RF analysis (Fog. S3) finds SPACe EMD to slightly outperform both CellProfiler mean and SPACe mean. So perhaps this distinction is not so clear? (Also see comments 7 and 9 below.)

We thank the reviewer for this relevant comment. The comparison between mean and EMD suggests that central tendency metrics (mean values) are more effective at capturing the relevant features for MOA classification than distribution-based metrics (EMD values) when all features contribute equally, as is the case for ‘Percent Matching’. One possible explanation is that central tendency metrics can more robustly summarize the overall characteristics of a well, making them less sensitive to variations and noise present at the single cell level/measurements. In the predictive RF analysis, the ability to predict MOAs accurately might benefit more from the detailed distribution information captured by SPACe EMD feature sets. EMD indeed provides a more sensitive measure of the differences between feature sets, which allows the RF model to capture phenotypic variations more efficiently resulting in more accurate MOA prediction. In addition, in contrast to ‘Percent Matching,’ the weight/contribution of each feature to the prediction can differ in RF models. Therefore, the difference in the relative performance of EMD-based features in ‘Percent Matching’ and RF model outcomes suggest there is likely a subset of EMD-based features that better capture the MOA phenotype than any subset of mean-based features. We have added this comment to the discussion in the revised manuscript.

3) Figure S2: It seems like the heatmaps in panels A-C should be symmetrical matrices, but the clustering seems different on the two axes. Could the authors explain more clearly what is going on here?

The heatmaps are indeed symmetrical. We have updated the figures with higher resolution versions that better show that the clustering dendrograms are identical.

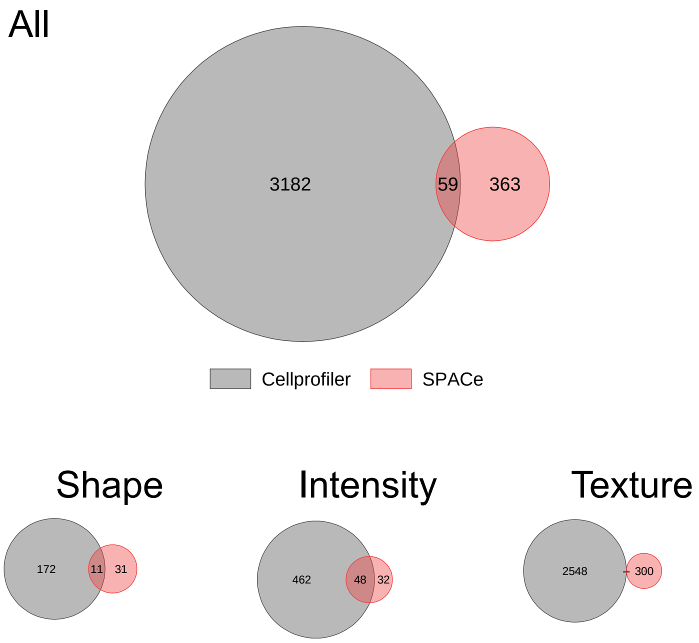
4) Figure S2: It looks like there are a lot more strongly correlated and strongly anti-correlated features based on SPACE analysis. Could the authors explain how they chose the 400 features?
The manuscript states that the feature set contains a (somewhat) higher proportion of uncorrelated features, but the authors also point out (lines 161-163) that “the feature set contains both redundancy and uniqueness

sufficient to recapitulate the CellProfiler generated results from the reference dataset, similar to other published work that reduced the CellProfiler feature set to a little over 600 (28)."
However, usually the point of dimensional reduction is to reduce redundancy by culling uninformative features (or transforming the features using something like PCA and keeping only those features that explain most of the variation in the data). Did the authors consider comparing a reduced set of features with only low correlation vs. a set of similar size with more redundancy? Or choosing a subset of 400 features from CellProfiler with minimal redundancy and comparing that to the feature set used by SPACe?

Our strategy for building the SPACe features library reflects the emergence of well-organized feature libraries in specialized repositories for image analysis. In particular, the shape features in our library contain all features that are part of the current Python sci-kit image package, where features were selected according to their expected representation and discriminatory power in the class of natural images. Similarly, our texture features are based on a current, widely used database of texture feature (Pyradiomics) which was constructed and extensively demonstrated in applications on biomedical images. In addition, the numerical implementation of our features was optimized to achieve data parallelization, *e.g.,* using multiprocessing. This was very important to reduce computing time.

By contrast, the feature library of CellProfiler was built over the course of several years and software versions incorporating increasingly more features over time and from different sources. This has caused the CellProfiler feature library to become large and somewhat redundant.

The standard CellProfiler Cell Painting pipeline tends to collect all features from all regions from all images (where applicable), resulting in a very large number of total features. We agree that PCA is an effective tool to reduce the dimensionally of the generated datasets, however, one of our goals in developing SPACe was to maintain feature interpretability, which is mostly lost using PCA: reduction to Principal Components makes it very difficult to directly interpret the observed response since all features contribute (with different weights) to each component. Considering that SPACe utilizes 2 additional ROIs per cell (nucleoli, mitochondria), we decided it was important to select a more focused initial feature set. Further, due to the increased number of ROIs per cell, a decision was made to move away from the 'all features, all ROIs, all images' approach towards a more selective approach where ROI and channel combinations predicted to be more biologically relevant (*i.e.*, Nucleus ROI + DNA stain) are prioritized over those predicted to be less useful (*i.e.,* Nucleus ROI + Mitotracker).



Response Figure 1. **Feature Set Comparisons**. Venn diagrams of single-cell feature overlap in datasets generated using CellProfiler and SPACe CPA pipelines.

To clarify the similarities and differences between CellProfiler and SPACe CPA feature sets, we annotated the ROI, images, feature, and feature type (shape, intensity, texture) for every single-cell feature from each analysis pipeline (Response Figure 1). For the purposes here, CellProfiler 'granularity' features are considered as texture features. As expected for shape and intensity features, there are subsets that overlap between CellProfiler and SPACe. In addition to features related to the 2 additional ROIs, the features unique in SPACe include 6 shape-properties (convex area, Crofton perimeter, circularity, EFC ratio, equivalent diameter area, maximum Feret diameter) and 2 additional intensity percentiles.

For texture features, by far the largest proportion of features for both analysis pipelines, we don't consider there to be any overlap. In addition to the granularity features, texture features within CellProfiler are calculated using a 2-D grayscale correlation matrix (GLCM) for each combination of ROI, image, direction, and scale. A panel of texture features are then extracted from this collection of GLCMs. In SPACe, the texture features are based on the well-established Pyrodiomics library (van Griethuysen et al., Cancer Research, 77(21), e104–e107, 2017).

To improve computational efficiency, SPACe calculates texture features across all distances and angles in the GLCM based on objects that are rescaled to 20 x 20 pixels and image intensities rescaled to 8 grayscale levels based on minimum and maximum object intensity. For each texture feature (contrast, dissimilarity, homogeneity, energy, correlation), SPACe generates a set of values corresponding to the various distance-angle combinations. SPACe then computes statistical descriptors—percentiles, mean, standard deviation (SD), and mean absolute deviation (MAD)—from these sets of values for each texture feature. The output is a vector that includes these statistical descriptors for each texture feature, essentially forming a 4D array where each dimension represents a texture feature and its corresponding statistical descriptors. We have added additional text to the methods section that specifies how texture features were extracted.

While we considered testing reduced sets of features, due to the difference in texture calculations, we felt any comparison would still be an 'apples to oranges' since texture features make up the bulk of the features extracted. In addition, others have already applied methods to reduce the scope of the feature set produced by the CellProfiler pipeline and our goal was to merely show equivalent performance using SPACe versus the full CellProfiler analysis. Finally, the training process for the Random Forest predictive models used would automatically select those features that best capture the variations between MOAs, negating the need to manually reduce the feature set prior to training.

For clarification, we have updated lines 161-163 "the feature set contains both redundancy and uniqueness sufficient to recapitulate the CellProfiler generated results from the reference dataset, similar to other published work that reduced the CellProfiler feature set to a little over 600 (28)," to ""the feature set contains sufficient diversity to recapitulate the CellProfiler generated results from the reference datasets, similar to other published work that reduced the CellProfiler feature set to a little over 600 (28)."

5) Figure S2E: For panels A-C, the Spearman correlation ranges from -1 to +1. But in panel E, the scale ranges from 0-1. Do they mean the absolute value of the correlation here? How does this figure show "enrichment of CP features below 0.2 and SPACe above 0.8" (lines 154-156)? Also – the authors refer to this as a histogram, but it is not a histogram.

Yes, Figure S2E represents the absolute value correlation. We have updated the figure for accuracy. However, we argue that this figure is a histogram, showing the distribution of feature correlation values in each feature set. We chose to use lines as opposed to bars to allow the overlay of histograms. To avoid confusion, we have updated the figure legend.

6) Figure S3: The main text (lines 167-168) indicate that "Each RF model was trained with half of the treatment replicates, randomly selected for each model replicate."
It was not clear whether this means (a) one replicate for all treatments, or (b) both replicates for half the treatments? If the reproducibility varies (which it does) depending on the analysis, then wouldn't using option (b) necessarily affect the results?
The text, Methods, and Figure S3 legend state that 5 RF models were generated per dataset using bootstrapping, but the size of the training set is not specified, or how subsetting was performed. Could the authors please describe the RF analysis more thoroughly in the Methods section?

We have updated the Methods section to clarify how the RF models were trained and then tested.

7) Figure S3: Panel B makes SPACe EMD look best, but what "Rank" means is not clear from reading either the text or the legend (it seems to be a comparison between the three analysis methods, but this should be

clarified).

There were no significant differences in the accuracy in predicting MOA using the library of RF models generated using CellProfiler mean, SPACe mean, or SPACe EMD (Figure S3A), however, as mentioned in the comment below, we did observe that there might a potential pattern in terms of how RF models were generated using each type of data performed. Therefore, we used rank analysis to quantify a performance pattern when considering CellProfiler and SPACe derived data. In rank analysis, for each MOA the average accuracy of the 5 RF models generated using each data type (CellProfiler-Mean, SPACe-Mean, SPACe-EMD) is ordered from highest to lowest. The highest is given the rank of 1, the lowest the rank of 3. To generate the data in Figure S3B, the rank value for each date type is averaged over all MOAs. To clarify for the reader, we have modified the figure legend.

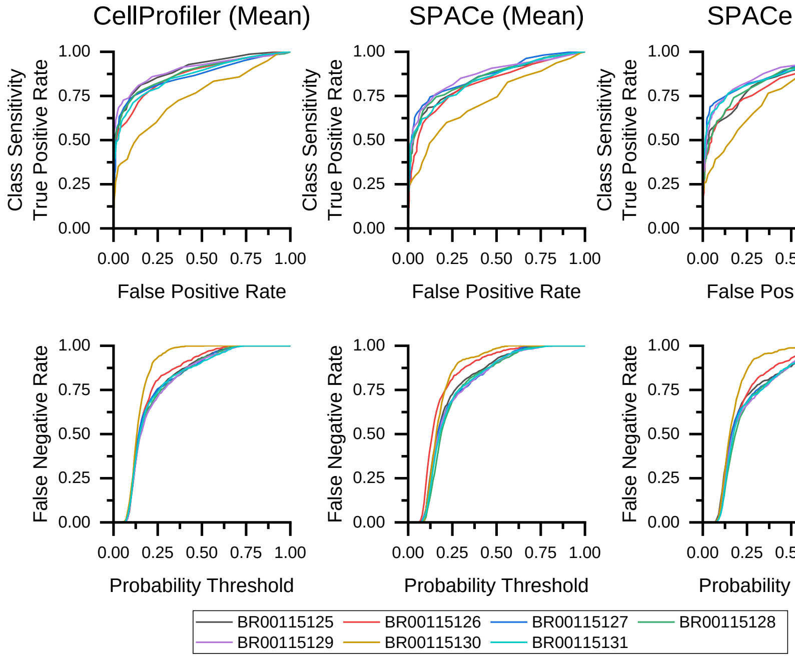
8) Figure S3: The methods don't look that different from panel A; while C shows that SPACe mean or EMD most often slightly outperform CellProfiler mean, panels D-F look almost identical qualitatively across all MOAs. Is this simply a reflection of the information content available in the datasets themselves, or some congruence in feature extraction? That is, is there an intrinsic limit in how much information can be gleaned from the CellPainting assay, or will exploring further improvements in data analysis yield significant gains in classification performance, eventually?

We agree, the ability to accurately predict MOA is equivalent when using either CellProfiler or SPACe, or when using mean- or EMD-aggregated data. Although not shown, we did generate EMD-aggregated CellProfiler data and found no significant gains in prediction accuracy. Given the comprehensive nature of the CellProfiler feature set, we feel this is likely due to an intrinsic limitation of the information content available using the standard CPA dye set. While these dyes are sufficient to capture a phenotype that predicts several MOAs with high accuracy (> 85%), they do not label cells in a manner to generate an observable phenotype for all MOAs. Substantial gains may be realized using additional dyes that label different subcellular compartments. Further, one must keep in mind that it is not known if all treatments were active in the U2OS cell line at concentrations used to generate the reference MOA datasets. The CPA dyes may fail to capture a phenotype simply because there is no phenotype to begin with; therefore, in this case, no degree of optimization of data analysis will yield gains in prediction performance.

9) Figure S3: The manuscript says that for MOA prediction using RFs (lines 493-494), "Model performance was evaluated by the percent of correct predictions (accuracy) and a confusion matrix generated." This brings two questions to mind:
(a) What about mis-classification? Is there an "unclassified" category in addition to the "none" category (which seems to distinguish "active" vs. '"inactive", which one assumes means "not distinguishable from control")?
(b) Since a confusion matrix was generated for each model, presumably it would be possible to include some kind of ROC plots comparing accuracy-specificity or precision-recall? Would this add any value to the analysis?

For the purposes of the analysis presented in the manuscript, we did not include an 'unclassified' category, rather we had the RF models predict the MOA for all samples based upon the highest probability. We adopted this approach because we were interested in observing differences in classification errors observed using data generated from the different pipelines. Indeed, although the confusion matrixes are similar between RF models trained with data generated from CellProfiler and SPACe, there are subtle differences as described in the manuscript that help understand the impact of different feature sets and data aggregation methods.

The primary focus of the presented work was to understand how the differing feature sets produced by the reference standard CellProfiler and SPACe affected model behavior, not seeking the best model performance since we do not attempt to classify the MOA in the subsequent work presented in the manuscript. However, it is more than likely accuracy could be increased, at the cost of sensitivity, by introducing a probability threshold and an unclassified category. For reference, we show the average False Positive Rate vs. True Positive Rate (ROC) and Probability Threshold vs. False Negative Rate curves for models trained using each reference dataset (Response Figure 2). Due to the limited difference, we feel this does not add sufficient value to the manuscript as it reaffirms that data generated by CellProfiler and SPACe equally captures the phenotypes associated with the various MOAs.



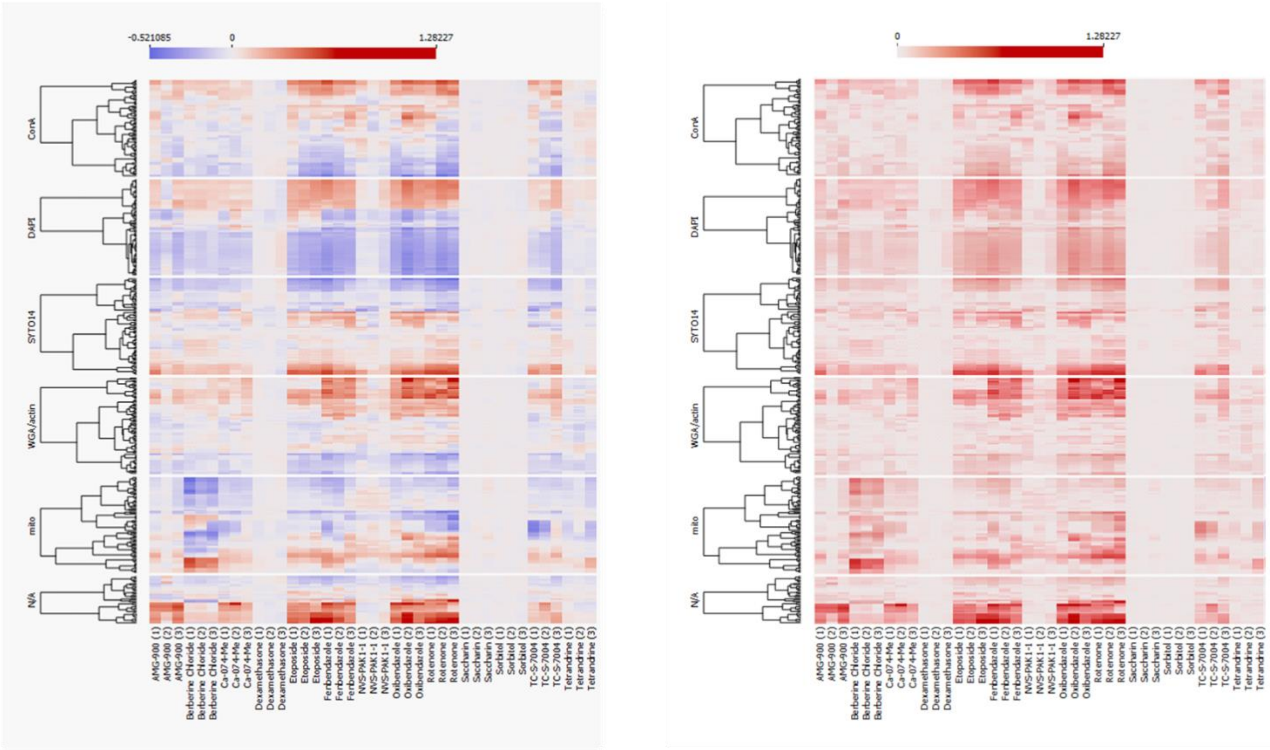
Response Figure 2. RF model performance curves.

10) Fig 2B: Is the clustering done for each “channel” separately, or were these separated after clustering on the combined data? What does the category "N/A" represent?

We apologize for the confusion. Separation was carried out after clustering. The N/A category represents the features not linked to fluorescence-based measurements, for example sizes and ratios (nuclear size/cell size or mitochondrial area/cytoplasm area). We added this definition in the Figure Legends

11) Figure 2D provides a concise summary of the EMD data. Did the authors compare performance of signed vs. unsigned EMD? Intuitively, one would think that including the sign (which is based on the direction of median change, correct?) would be advantageous, but it would be nice to see a comparison.

We show below a plot with unsigned EMD in Response Figure 3 below. In our opinion, the plot using sign is more informative, as both the strength of the alteration and its direction are visible. For example, if the cell area is altered in response to a compound, the plot using sign also shows whether the area became larger or smaller.



Response Figure 3. Comparison between signed (left) and unsigned (right) heatmaps and clustering.

12) Could the authors list the stain and compartments affected in the assay somewhere in the manuscript? While Cell Painting is a broadly available assay, it's not necessarily obvious where to go to find this information and it would be nice if the authors included a short description of the assay itself, and which sub-compartments or cellular components are surveyed (e.g. Both actin and tubulin cytoskeletons? Both peroxisomes and lysosomes? etc.).
Reminding the reader of which compartments are included in the CellPainting assay would be particularly helpful since the manuscript notes that SPACe segments the cell, cytoplasm, nucleus, nucleoli, and mitochondria. Are other compartments also segmented, or are the stain signals just analyzed for intensity and texture features without segmentation? Is this analysis tuned specifically to the CellPainting platform, or is it generalizable to markers of additional cellular structures that are not included in CellPainting (e.g. tubulin, biomolecular condensates such as stress granules)?

We have added a description of the dyes used and of their respective compartments, essentially following the standard reference: Cimini et al., Nature Protocols 2023. We remark that the SPACe pipeline is modular and can be generalized to other structures by adding segmentation steps of the compartments of interest.

13) Is there any one cell line that shows greater responses across a larger number of bioactive chemicals? That is, if you had to choose one cell line to screen, or 2-3, say due to limited financial resources, which one(s) would they recommend? This could be a nice addition to the Discussion.

Thank you for the question. If we consider all 18 cell lines we tested with known active compounds, U2OS is the most responsive, followed by A549, Hela, MCF10A, 5637, and MDA-MB-231. If we look at the cell metabolism library data, 5637 had the highest hit rate, followed by MDA-MB-231, U2OS, and HepG2. From these somewhat limited screens we would suggest U2OS (as they appear to be a CP gold standard), plus a couple of cell lines that are relevant to the screening campaign needing to be performed. If a lab has a few cell lines available, we suggest testing them with a set of control chemicals/chemicals of interest ahead of screening campaigns to pinpoint the best models. We added a comment to this effect in the Discussion.

14) Do the authors discuss whether the software detects the presence of plate effects in Row/Columns effects, or batch effects? Did the authors consider plate design and how this would be implemented into the SPACe software?

Thank you for the question. While SPACe does not formally address plate effects, its quality control step will flag bad wells and ultimately the EMD data can be plotted to detect edge/column/row effects in the data. The plate design is not essential for SPACe as the anchor wells, labeled with DMSO, can be in any position on the plate.

15) Lines 454-455: "The reference distribution is here defined as the median of the DMSO distribution in each experiment. It is very confusing to label the reference (DMSO control) distribution as "the median of the DMSO distribution". This terminology obfuscates the narrative in many places. Why not just call it "the reference distribution"?

We apologize for the convoluted narrative, we sought to emphasize how the reference distribution was being calculated. We simplified the text according to the reviewer's suggestion.

16) Methods section, "Statistical analysis" (beginning at line 497): "To compare fingerprints, Euclidean distance was measured between EMDs of all features in the treatment wells and the median DMSO control control wells". Could the authors state how many DMSO wells were used? How were biological and technical replicates handled in the Euclidian distance calculation?

We apologize for the lack of details – in the screening campaigns we had 32 DMSO wells/plate; in any other experiment we had at least 8 DMSO wells. For EMD calculations, every plate is considered a separate unit (biological replicate when run on different days or technical replicate if run the same day) and we measured EMD for each well in each plate (including each DMSO well that passed QC). In this way, we would have, for example, 32 EMD values for DMSO plus 352 EMD values for all other treatments in the plate. We added a note to the "statistical analysis" section to emphasize this effect.

17) Minor editorial issues: (a) Figure 1D has an error in it — in panel D, top is percent replicating and bottom is percent matching, but the legend says both are percent replicating; (b) Fig. 5D legend is missing; (c) Figure legends not uniformly formatted (capitalize each panel sentence or not?); (d) Line 134: "resources availability", is this a typo?; (e) Line 469: "all datasets were processes" should be "processed"?

We apologize for these errors and have corrected them in the new draft.

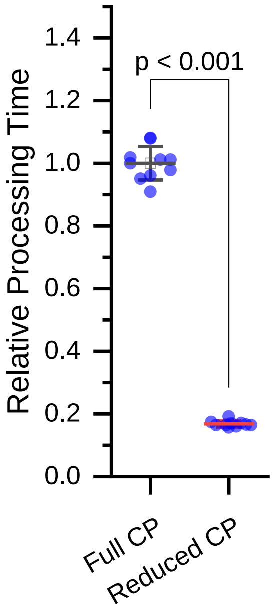
Reviewer #3:

Major comments:

*A main theme of this manuscript is the comparison to Cell Profiler, and the authors make it a point that SPACe performs ~10X more efficiently than Cell Profiler. Notably, SPACe extracts 400 features, while the authors mention that CellProfiler extracts ~4000 features. Could the authors please comment if this computational efficiency is simply due to the reduced number of features that is being extracted? If not, perhaps make this more explicit. It would enhance the manuscript if the authors could further describe the details of the comparison, and whether the comparison was an "apples to apples" comparison. For example, would CellProfiler be just as efficient if it was asked to extract the same 400 features that SPACe used?

The reviewer raises an interesting question that complements the point by Review #1, point #4 above. As we described above, the features extracted by SPACe are a curated subset of features related to those collected by CellProfiler and expanded through additional ROIs.

To address the issue of computational efficiency, it is important to note some limitations that exist. First, it is not possible to make an apple-to-apples comparison because CellProfiler does not allow users to select a subset of shape, intensity, or texture features to extract. In addition, as described above in response to Reviewer #1, the way texture features are calculated differs significantly between CellProfiler and SPACe. This is important because the extraction of single-cell texture features is a major consumer of computational time in both CellProfiler and SPACe pipelines.



Response Figure 4. Relative processing time using a simplified CellProfiler pipeline.

However, to understand the impact of reducing the number of features extracted on CellProfiler computing time, we generated a SPACe-like variant of the CellProfiler pipeline. To do this, we first removed feature modules representing metrics not collected by SPACe (MeasureImageQuality, MeasureColocalization, MeasureGranularity, MeasureObjectNeighbors, MeasureObjectIntensityDistribution). Next, we reduced the collection of texture features to ROI and image combinations collected by SPACe. Finally, we reduced the number of texture pixel scales from 3 different scales to 1 even though this does not reflect how SPACe extracts texture using the Pyrodiomics library. Using this reduced CellProfiler pipeline, the total number of singe-cell features extracted is reduced from to 916 and the difference in the number of texture features extracted between CellProfiler and SPACE is reduced to 116. With this reduced pipeline, the processing time per dataset for CellProfiler is reduced by approximately 6-fold relative to the standard pipeline (Response Figure 4). Based on these highlighted limitations, we believe it is fair to say that SPACe gains efficiency due to the curated set of features and the use of the Pyrodiomics library. However, this remains an apples-to-oranges comparison as the extraction of texture features is the primary contribution to processing time for CellProfiler. In contrast, the primary contribution of processing time for SPACe is the segmentation of nuclei and cells using the pretrained Cellpose cNN models. To make the comparison clearer, we have expanded the description of the comparison between CellProfiler and SPACe starting at line 137 in the original submission.

*The authors highlight berberine chloride as a case study. This compound has a promiscuous MOA and interestingly, it is also a highly colored compound. Did the authors consider if the apparent phenotypes observed were due to a true biological effect, or due to compound-mediated interference such as quenching or auto-fluorescence, or a combination? It is known that some compounds can produce apparent phenotypes in staining assays simply because the compounds themselves act as dyes, which could perhaps explain its high reproducibility and activity across many cell lines. If the activity is somehow related to compound interference, and SPACe classifies it as a MOA, that in itself is also a significant finding and a cautionary note to the CP community.

We thank the reviewer for the insight and indeed it would be interesting if that was the case. However, berberine chloride excites at 350nm (closest laser would be 405nm), but its effects are only apparent in the far-red channel showing a clear change in mitochondrial morphology and distribution indicating that there is no interference from the known fluorescent properties of the compound. Of course, the phenotype could perhaps be due to berberine chloride (or a metabolite) accumulating within mitochondria, which we are not aware of. Also, a similar phenotype was shown by Willis et al., SLAS Discovery 2020, and they have not identified it as an artifact in their analysis.

*"Compounds with discordant replicates or with obvious imaging artifacts were also excluded after manual inspection..." Could the authors please describe their criteria for discordance and determining "obvious artifacts"? Where images manually inspected? Whatever their approach, would it be feasible for a large screen? Related, I would actually recommend the authors include some examples of artifacts in the supplemental and how SPACe would have analyzed them. This would in fact be valuable information for users to know potential failure modes for this software, so that such "bad wells" can be identified.

We thank the reviewer for this question. In general, bad images/discordant replicates will be flagged in the QC step especially when object count is very low (where the threshold for low can be manually adjusted by the user) or in post-analysis when one of the replicate wells has a widely different EMD value (either high or low) as compared to the other technical replicates. Manual inspection always follows for the wells with a discordant EMD as these can often derive from technical issues (for example, missed treatment, under or over labeling, dry well etc). We understand that for large screens this could be time consuming and is an issue of user choice and dependent upon the number of hits to be followed on.

*Several times in the manuscript, SPACe is compared to Cell Profiler. Could the authors please comment on whether there are any situations where Cell Profiler would actually be the more appropriate analysis tool? If so, these reasons should be mentioned to provide an overall balanced perspective.

As described above, there are several measurement modules implemented in CellProfiler that are not currently replicated by SPACe. Among these is the MeasureObjectNeighbors module, meaning SPACe does not directly measure the special relationship between cells in samples. For samples in which this might be an important endpoint, such as more complex 3D or tissue samples, CellProfiler might yield a more complete analysis. However, the open-source nature of SPACe provides a foundation that users can use to modify or add additional features as desired. It is important to note, CellProfiler has a large user base, so it is relatively easy to find discussion forums online. In addition, there is a relatively large team currently maintaining the software and helping users with more complex image analysis problems.

In response to this comment, we have modified the discussion starting at line 360 in the original submission.

*The authors note in the methods that cells were cultured in "ATCC suggested media". It is well known that the choice of cell culture media can impact compound activity, especially when the testing was performed on a cellular metabolism-based focused screening collection. Furthermore, the amount of serum can impact cell growth, for example, which may have an effect on the cell population distribution (cell state). To aid in interpretation of their results, and enhance reproducibility, it is recommended that the authors include a complete table (in the supplemental) with the specific media for each cell line (including any additives or serum).

We added the media information in the Materials and Methods section.

*The authors describe one potential benefit of SPACe as the ability to analyze single-cell data. However, after reading the manuscript, it is still not clear how SPACe takes into account single cell data, and how analyzing data at the single-cell level would outperform well-level analyses. I would encourage the authors to better articulate these points, especially for less computational-savvy audiences, and perhaps articulate the potential benefits of single-cell analyses with a clear example (or if they believe one of their panels already does this, then better articulate to lay audiences).

To avoid a possible misunderstanding, we want to clarify that by “single-cell level” analysis we mean that our analysis does not aggregate measurements from an entire well into a single central-tendency measure, such as the mean or the median. Instead, we use single-cell measurements to compute a distribution of responses so that we can fully account for cell heterogeneity.

We believe that single-cell analysis is critical to faithfully reproduce and understand cell heterogeneity. As an extreme “idealized” example, one could consider the case where, in a well, half of the cells exhibit positive response (with respect to DMSO) and the other half of the cells have negative response to a perturbation, so that the average response is zero (with respect to DMSO). Clearly, the average response per well (which aggregates all cells responses into a single measure of central tendency) will not detect a “hit” as it misses the heterogeneous behavior of the cell population. By contrast, a single-cell approach in the sense that we have described collects single-cell information and generates a distribution of responses much more robust to detect a biologically meaningful change in the distribution of the cell population with respect to DMSO.

*The authors describe a QC step that automatically removes wells with low object count. I presume this is to eliminate cytotoxic compounds or compounds that affect cell adherence. Are there any instances where this step could exclude potentially valuable data? If so, they should consider noting in the text.

Yes, the reviewer is correct. We added a comment in the Methods section as the threshold for QC cut-off can be manually adjusted by the user and could be useful for studies employing a low number of plated cells and cytostatic compounds.

*"It is unlikely to identify compounds that would act in a universal manner and can be used as controls across all experimental models. This complicates the analysis, prediction, and interpretation of the MOA for compounds when based uniquely on phenotypic screening in a single cell model". This is hardly surprising. Do the authors, or others in the community, really expect that compounds would produce the same morphological changes (universal) across a variety of cell, and could they provide more context about why these analyses were done? In most practices that I am aware of, one would test a well-annotated MOA compound set in each cell line to define the morphological changes corresponding to the MOA according, and I would be skeptical of applying morphological fingerprints across cell lines. Related, this type of analysis has already been reported by the Harrill group at the US EPA, and I would recommend the authors consider citing their important work related to this topic.

We appreciate the comment and have modified the Discussion and have the suggested reference.

Minor comments:

*"Ten compounds were the most toxic across all models (auranofin, SF1670, plumbagin, PR-619, CB-5083, PFK158, eeyarestatin 1, digitoxin, paclitaxel, and TG101348) and should be tested at lower concentrations to measure changes at non-toxic levels, as some of them show potentially very interesting phenotypes in the surviving cells." The significant phenotypes associated with cellular injury compounds has recently been described (PMID 36914634). The authors should consider citing this work. Have the authors considered analyzing some of this reference data using SPACe?

We thank the reviewer for the very valuable comment, and we apologize for omitting discussion of the recent study, which is indeed highly informative. We have added the reference and commented on it both in the Results and in the Discussion sections. We have not undertaken the analysis of this dataset with SPACe preferring to use the more utilized Broad library. We are in the process of exploring this dataset and identifying toxic/interfering signatures, but this is beyond the scope of this manuscript and will be a topic for a follow-up study.

*I found the manuscript very well written, but at times it was hard to relate some of the technical terminology to how it impacts the end result (usually MOA calling). One recommendation that could enhance this manuscript's impact and appeal to a broader audience of end-users (e.g., bench scientists; non-computational biologists) would be to add less technical text that more clearly explain the practical implications of the various technical advances in SPACe.

We appreciate the comment regarding the complexity in writing, and we have simplified the text toward that goal.

i

*Throughout the manuscript, the authors attribute MOAs to reference compounds. The authors may want to state the source of these MOA annotations and provide reasonable caveats, as I would actually be hesitant to take some of these prescribed MOAs at face value. For example, while some of these compounds have well-defined MOAs, some of these compounds have well-known promiscuous MOAs (e.g., rottlerin). The authors should also consider whether the targets/MOAs, often determined at nM concentrations, are still relevant at uM concentrations.

We acknowledge and thank the reviewer for this comment and have rewritten some parts of the manuscript to this effect. In most cases the MoAs were annotated by the JUMP consortium and were used as such. At the uM concentrations used, in some cases it is likely that the MoA could be the result of interactions with multiple targets and secondary effects. We added comments in the manuscript to this effect.

*The authors should consider/add text that compounds that affect cellular adherence (but non-cytotoxic) may also lead to low object counts.

Thank you for the comment, we added a sentence to that effect.

*"this implementation choice does not reduce in any way..." this is quite a strong statement. Is there really no possibility scenario that this implementation choice affects downstream performance?

We changed the text to reduce the strength of the statement

*The authors use "small molecule inhibitors..." whereas "small-molecule inhibitors" is more generally preferred.

Corrected.

*Typo: "Perhaps interestingly, only seven compounds in the screed..."

Thank you for catching the typo, it has been corrected.

*The correct name of the cell line is "U-2 OS" according to ATCC, not U2OS. The same for HepG2, which is actually Hep G2 according to ATCC. A minor point, but I recommend the authors check that these and other cell lines are correctly named.

Corrected.

*The authors use the term dose-response or dose several times, which should be reserved for actual administration of a drug dose in an in vivo setting; the more appropriate terminology for cell culture experiments should be "concentration-response" and its variations.

Corrected.

*The beginning of many sentences are not capitalized in many of the figure legends, most notably for figure panels. This should be corrected and consistent.

We apologize, it has been corrected.

*Do the authors have a method they could cite for their mycoplasma testing?

Here are some relevant publications that we cite in the revised manuscript and are included in the revised reference section:

Jung H, Wang SY, Yang IW, Hsueh DW, Yang WJ, Wang TH, Wang HS. Detection and treatment of mycoplasma contamination in cultured cells. Chang Gung Med J. 2003 Apr;26(4):250-8. PMID: 12846524.

And here is one of many links to a company website:

https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/microbiological-testing/mycoplasma-testing/testing-for-mycoplasma

*What was the source of the chemicals besides the Cayman metabolism library? This should be noted, especially for the key compounds in their study. What QC was performed on these compounds?

We added the requested information on the origins of the compounds. The compounds were used as provided. (No QC was performed on the compounds tested).

*"37C/5% CO2" should include the degree symbol, and chemical formulas should have subscripts. The authors should carefully review their manuscript and SI materials for similar issues.

Corrected.

*There should be spaces between the concentration number and the concentration units (e.g., "10 nM" instead of "10nM")

Corrected.

*Overall, the figures are beautiful, but boxes around several panels are clearly visible (e.g., Figure 2B, Figure 3A,B,D). Recommend these be touched up.

Figures were touched up

Reviewer #4:

The comparison between SPACe and CellProfiler is not a fair comparison. In line 152 the authors mention that SPACe extracts ~400 features while CellProfiler extracts ~4000 features. The difference between these two is 10X, which is oddly the same factor of improvement the authors are claiming in computing time. Is the computational saving a result of their implementation or from reduced feature selection? Since the features extracted by SPACe is a subset of CellProfiler's, it should be obvious for the authors to reduce the CellProfiler pipeline to the same feature set.

Reviewers #1 and #2 shared similar and related comments. Please see the responses above that address this question

Furthermore, it's unclear if the impact of this work is as great as the authors claim. Computational power is improving all the time. While CellProfiler is a clunky piece of software, and improvements are always welcome, it's unclear that processing speed is limiting any biological findings. Importantly, the reported SPACe software is not enabling new findings.

The main goal of this manuscript was to introduce a more practical and portable computational platform for cell profiling that would make the Cell Painting approach more accessible. Our direct experience is that memory requirements and computing requirements of CellProfiler are such that it is sometimes impossible to process data. In such situations, the problem is not about waiting extra hours of processing time but rather about being able to analyze the data. This experience was part of our motivation for developing SPACe. We believe that, by making Cell Painting more accessible, we will increase the opportunities for investigators to generate new findings.

The utility of CellProfiler is that it presents the image processing operations in an accessible way to biologists. The software presented by the authors method require significantly more expertise to operate.

We agree with the reviewer that Cell Profiler's GUI makes it user-friendly. However, SPACe has a very simple menu of operations that we designed specifically to analyze cell painting data and requires *significantly less parameter tuning than Cell Profiler*. For instance, Cell Profiler uses an intensity thresholding approach for cell segmentation that requires significantly more manual parameter tuning as compared to our approach based on Cellpose (using a very robust generalist data-driven approach). Our method is implemented in PyTorch which is a well-established and widely used programming language, with a very large user base. The user does not need to be knowledgeable about PyTorch to use SPACe, but a knowledgeable user can take advantage of the Pytorch implementation to potentially add additional image data processing modules.

Line 342 – I would say a major limitation to the wider adoption of CP-like phenotypic screening is the costs associated with the HCS hardware, cells, reagents, and expertise. Computing cost is less significant than these costs, and a full analysis is conducted infrequently.

We think this is a rather subjective comment that was not shared by other reviewers or our beta users. We feel that the cost, access and availability of cloud computing is still a significant barrier to HTS, and that SPACe would somewhat help to mitigate it. We understand the steep impact of the cost of instruments, however, which can be (and are) available to users at low cost/hour in an increasingly larger number of academic core facilities, including our own.

We thank the reviewer for all the additional comments that were worthwhile in making the manuscript better. Please see our responses (in blue) to the questions raised.

Reviewer #1 (Remarks to the Author):

The authors have made a good-faith and largely satisfactory effort to address most of the comments from the original round of reviews. We have only minor comments on the revised manuscript:

1. Figure 2B and Fig 3: Label (if room) and legend (at least) should specify “Signed EMD”.

We added “Signed EMD to figures, legends and in the manuscript, wherever applicable.

2. Figure S2C: This third heatmap in the figure is still not symmetrical; different clustering is used on each axis (this is obvious by eye).

We have identified the underlying issue that was causing the asymmetry in how the heat map was displayed and corrected the figure.

3. Revised manuscript, lines 155-167 and Figure S2E: Referring to Reviewer 1, Question #5 in the first review cycle, pertaining to the "histogram" or "frequency distribution" of Spearman correlation between features:

a. Thanks for clarifying that the diagram represents a frequency distribution, and that overlaid histograms were converted to lines for better visual comparison. Perhaps the usage of "histogram" vs. "frequency distribution" seems semantic to the authors, and it is a minor point, but we would expect a histogram to show a "grouped frequency distribution", in which the area of adjacent rectangles is proportional to the frequency (or relative frequency) of observations in each interval. Especially considering that the feature sets are of different magnitudes, and apparently not binned to the same interval sizes, calling this a histogram seemed visually incongruous. Also note that, technically, this is a "relative frequency distribution".

We have ensured that the manuscript refers to this figure as a ‘relative frequency plot’.

b. Line 158: When the authors say "A large fraction of features is highly correlated (>0.8), regardless of the analysis method being used", how large is this number really? Perhaps not a "large" fraction, but a "substantial" or even "minor" fraction? From Fig. S2E, it looks like maybe around 20% of SPACe features, but a much smaller proportion of CP features (even though in absolute terms still large given the size of the full feature set)? This may be why the authors say that there is an "enrichment" of features for CP below 0.2 and SPACe above 0.8, but this phrasing still trips us up. The reason is that Fig. S2E shows only a slightly higher proportion of CP features with very low pairwise correlations (Spearman correlation < 0.2) relative to the SPACe feature set; and this difference is probably negligible taking all other factors into account (?). In contrast, SPACe gives rise to a pretty flat feature distribution from a correlation of 0 to ~0.9, so any "enrichment" of SPACe features with correlation > 0.8 is specifically relative to the CP distribution. It would seem simpler just to point out that while the SPACe feature set contains a greater proportion of highly correlated features (Spearman correlation > 0.8) than the CP feature set (Fig. S2E), SPACe contains a significantly higher proportion of "unique" features (defined as having no correlation with other features > 0.95) (Fig. S2D). Again, relatively minor point, but perhaps this description could be improved here.

This is an excellent suggestion, and we have revised the paragraph to:

“To understand the uniqueness of the features collected by each method, Spearman correlation between features across all samples in the JUMP reference datasets were examined (Figure S2A-C). The relative frequency of features with correlation values between 0.2 and 0.8 were similar for each method, however, SPACe extracted feature sets contain a greater proportion of highly correlated features (Spearman correlation > 0.8) than the CP feature set (Fig. S2E). SPACe mean and EMD feature sets contain a higher proportion (24% and 32%, respectively) of ‘unique’ features (defined as a feature with no correlation compared to the CellProfiler feature set (16%), despite the absolute number of unique features being lower (Figure S2D). This suggests that, although SPACe collects a smaller feature set, the feature set contains sufficient diversity to recapitulate the CellProfiler generated results from the reference datasets, similar to other published work that reduced the CellProfiler feature set to a little over 600 (28).”

4. Revised manuscript, lines 505-510: Referring to Reviewer 1, Question #15 in the first review cycle: This pertains to the definition of the "reference distribution" for DMSO controls. The manuscript says,

"The QC routine is designed to establish a reliable ground truth for single cell distributions in control samples (e.g., DMSO). The idea stems from our prior publication (19) that demonstrated the value of distribution analysis as a quality control step for high throughput microscopy assays and subsequent single cell analyses. The QC step establishes a reference distribution for the DMSO negative control wells (eliminating outliers because of low object count or aberrant phenotypic profile). The reference distribution is defined as the median of the DMSO distribution in each experiment."

What was not clear in the first round of review is exactly how the "median of the DMSO distribution in each experiment" is computed to establish the reference distribution (though this must seem obvious to the authors). The text does not clarify the procedure used to obtain this, but perhaps it is described in the mentioned prior publication? We infer that the authors compute a "per-well" distribution for each feature across all of the DMSO wells on a plate, and that these per-well distributions become the basis for the "median" reference distribution. Is that correct? It would be more clear if the authors explained this more clearly in the Methods, which would alleviate the original source of the confusion.

We apologize to the reviewer for the lack of clarity. The definition of the reference distribution is indeed described in our prior publication. The reviewer is correct, a distribution is calculated for each DMSO well. For each fixed feature, its distribution is calculated for each DMSO well on the plate. Next, from these distributions, a reference distribution for DMSO is calculated by taking the median of these distributions. We added a comment in the Methods and in the Result sections.

5. Reviewer 1, Question #14, regarding controlling for row/column or batch effects:

The authors state that control wells can be anywhere on a plate, which is true, but we and others have noticed that there can be major differences across plates due to issues with reagent dispensing or differential humidity. There can also be differences in the DMSO "median" reference distribution for the controls across different plates. Without normalization of some kind to account for such differences, it is not clear how the QC step would address this. The authors state in their rebuttal that QC will flag bad wells, fair enough, but while "plotting the EMD data" could indeed detect edge/column/row effects and probably also plate effects (?), this does not really address the issue. Are we missing something?

We agree with the reviewer that, depending on the setting utilized for HTS, there could be major differences inside the same plate and across plates. In our experience, utilizing the information embedded in the distribution is more forgiving and stable as we demonstrated in our prior publication. As the QC step is performed on a plate-by-plate basis, we found that there is no need for normalization across plates. In our previous work (albeit not on cell painting), we found that the reference distribution was very stable across time (over 3 years) and random experimental variation, which of course should be validated for cell painting in terms of which features are indeed more stable and reproducible over time. In a simplistic way, the key is to have enough control wells in each plate so that the reference distribution is as close as possible to the true distribution.

In general, users can examine the plate distribution of QC flagged wells. If there is a consistent pattern to these wells, that would suggest a plate effect that may need to be addressed. Due to the modular nature of SPACe, those needing to address plate effects can insert that function using one of the established methods after single cell feature extraction, before QC analysis, and before final EMD calculations.

We remark that our screening environment is one in which screening campaigns are relatively focused, incubation times are short, and precise liquid handling systems are used to process the multi-well plates. Likely due to these factors, we have not observed significant plate effects in our data. We have also not yet adopted randomized/semi-randomized plate layouts that would facilitate the calculation of correction factors for row/column effects. Due to this and the degree of data manipulation involved in determining a correction factor for each feature in each well and propagating that factor to all single cell values, we have elected not to include that functionality in SPACe.