

# Unbiased Phenotype Detection Using Negative Controls

SLAS Discovery  
2019, Vol. 24(3) 234–241  
© 2019 Society for Laboratory  
Automation and Screening



DOI: 10.1177/2472555218818053  
slasdisc.sagepub.com



Antje Janosch<sup>1</sup>, Carolin Kaffka<sup>2</sup>, and Marc Bickle<sup>1</sup> 

## Abstract

Phenotypic screens using automated microscopy allow comprehensive measurement of the effects of compounds on cells due to the number of markers that can be scored and the richness of the parameters that can be extracted. The high dimensionality of the data is both a rich source of information and a source of noise that might hide information. Many methods have been proposed to deal with this complex data in order to reduce the complexity and identify interesting phenotypes. Nevertheless, the majority of laboratories still only use one or two parameters in their analysis, likely due to the computational challenges of carrying out a more sophisticated analysis. Here, we present a novel method that allows discovering new, previously unknown phenotypes based on negative controls only. The method is compared with L1-norm regularization, a standard method to obtain a sparse matrix. The analytical pipeline is implemented in the open-source software KNIME, allowing the implementation of the method in many laboratories, even ones without advanced computing knowledge.

## Keywords

high-content screening, multiparametric analysis, fingerprinting

## Introduction

Of the many ways to unlock the mechanisms by which cells respond to their environment, high-content screening has established itself as a powerful technology. Several reasons can be invoked for this popularity, the major one being the richness of information that can be obtained from images. The richness comes in many layers. First, several organelles or proteins can be marked simultaneously, allowing the gathering of information about several concurrent biological effects. Second, for each labeled cellular component many parameters can be extracted that not only measure its amount but also its morphology and its distribution in the cell with respect to itself and other structures in the cell. Lastly, this analysis can be carried out on a cell-by-cell basis, giving insight into population distribution instead of simply its average. Few other methods yield such diverse information about cells.

The richness of the information extracted is also a challenge, both mathematically and computationally. Depending on the resolution of the images, between 200 and 2000 cells are commonly imaged per condition in a screen. A screen with 100,000 data points and 1000 parameters yields a matrix of  $2 \times 10^8$  rows and  $1 \times 10^3$  columns whose size can be nearly as large as the original image data (no compression). The data are also complex in that they represent different measurements with different units, some of which are

correlated with others, some of which are not informative and add noise without helping to discriminate the signal, and lastly, some that are noisy and tend to mask the signal. A further complication is that some parameters do not help the biologist understand what the phenotype actually looks like and its biological significance. This is particularly true for texture features that are hard to interpret in a biologically meaningful manner. Since data mining is the task of transforming data into knowledge or insights, some parameters need to be ignored when interpreting phenotypes.

The challenging nature of data mining is reflected in the fact that most high-content screening publications to date make use of only a few parameters instead of delving into the richness of the data they have obtained.<sup>1</sup> Obviously, to discover hits in a very focused assay, only a few parameters

<sup>1</sup>Technology Development Studio, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

<sup>2</sup>Fraunhofer-Institut für Verkehrs- und Infrastruktursysteme, Dresden, Germany

Received Sept 20, 2018, and in revised form Nov 13, 2018. Accepted for publication Nov 19, 2018.

### Corresponding Author:

Marc Bickle, Technology Development Studio, Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany.

Email: bickle@mpi-cbg.de

are required. Nevertheless, when more parameters are exploited, the hits can be profiled according to their phenotypes, allowing a stratification strategy for selecting compounds for further development.<sup>2</sup> Furthermore, if the assay has been well characterized with a set of compounds with a known mode of action, the hits can potentially already be classified with respect to their mode of action, allowing quick identification of compounds acting via novel mechanisms.

Several strategies have been developed to use the full range of parameters extracted from the images, each with their own advantages and disadvantages. These can be broadly split into supervised methods and unsupervised methods. Supervised methods use various machine learning algorithms (support vector machines,<sup>3</sup> random forests,<sup>4</sup> deep neural networks,<sup>5</sup> bagging,<sup>6</sup> etc.) to classify cells into phenotype categories defined with a training set. These methods use many parameters and are very powerful and relatively easy to use. Their main disadvantage is that they have to classify objects into one of the user-defined categories, so that novel, previously unseen phenotypes will be wrongly classified (open world vs closed world).<sup>7</sup> For example, training a machine learning algorithm on cat images will wrongly classify a dog as a breed of cat. This problem is not addressed in the field, with a few exceptions.<sup>8</sup> A further problem of supervised methods is the danger of overfitting or training the algorithm on some latent feature.<sup>9</sup> Overfitting leads to overestimation of the efficacy of the algorithm and poor performance on real, unknown data. Training on latent features can lead the algorithm to recognize features outside of the objects of interest. For instance, if the controls used for training are always in the same position, the algorithm might recognize the position due to peculiarities in the way the microscope images those positions.

Several unsupervised methods have been proposed; in particular, two publications did a systematic review and tested many of these.<sup>10,11</sup> In both papers, the task was to evaluate the ability of various dimension reduction methods and, for Reisen et al., evaluate various distance measures to classify or score a high-dimensional phenotypic dataset. Reisen et al. demonstrated that the full-length phenotypic readout distinguishes less well between a positive and a negative control than the best single parameter, demonstrating the need to perform dimension reduction. They then demonstrated that single parameters are very poor at profiling compounds and finding compounds of a similar mode of action. Instead, the full-length fingerprint appears to recognize similar phenotypes slightly better than reduced dimensionality data by principal component analysis (PCA) or random forest scales (RFSs). Ljosa et al. demonstrated that for profiling compounds, their full fingerprint performs very well and is not outperformed by methods meant to capture the

population response. In their analysis, dimension reduction by factor analysis does significantly improve the classification of the compounds. This demonstrates that some parameters in their fingerprint are noisy and affect the classification.

We sought to develop a profiling method that is able to discover previously unseen phenotypes, exploit the population response, and avoid noisy or uninformative parameters. We reasoned that if we obtain the set of parameters that describes the negative population precisely and reproducibly, we would reduce the dimensionality without sacrificing the ability to discover new phenotypes, while eliminating noisy, unreliable parameters. For instance, if the area of nuclei is always to be found within a given range in all the wells of the profiling experiment, then any condition where the area diverges significantly is a phenotype. Conversely, if another parameter exhibits diverging ranges in different wells, this would not be reliable and the parameter should be eliminated. To achieve this, our method analyzes the distribution of populations in all wells for each parameter available and selects parameters that are neither too noisy nor too uninformative. We used the publicly available data of BBBC021 and BBBC022 from the Broad Institute to test our novel method and demonstrate its performance. We compared our method for reducing parameter space using negative controls only with the classic method of L1-norm regularization.<sup>12</sup> We found that our method improves the classification performance compared with the full matrix, but it does not outperform the sparse matrix obtained by L1-norm. We then showed that our method generalizes better when an unseen class is added to the classification problem compared with the sparse matrix obtained by L1-norm.

## Materials and Methods

### Data

We used image set BBBC021v1,<sup>13</sup> available from the Broad Bioimage Benchmark Collection.<sup>14</sup> The BBBC021 dataset is of MCF7 breast cancer cells that were plated and treated with compounds of a known mode of action.<sup>13</sup> The images were reanalyzed with CellProfiler by Ljosa et al.<sup>11</sup> The resulting data were made publicly available and are available on the Broad Institute's website. The data were imported into a mysql database as described by Ljosa et al. and imported into KNIME with corresponding database reader nodes. The data were then normalized per plate as described in by Ljosa et al. by linearly scaling the data, setting the 1st percentile of DMSO to 0 and the 99th percentile to 1.

We used image set BBBC022v1,<sup>15</sup> available from the Broad Bioimage Benchmark Collection.<sup>14</sup> The BBBC022

dataset comprises U2OS cells treated with 1600 known bioactive compounds. The cells were labeled with six dyes to characterize seven organelles (Cell Painting Assay).<sup>15</sup> We downloaded the images from the Broad website (<https://data.broadinstitute.org/bbbc/BBBC022/>) and analyzed them with the CellProfiler pipeline provided by Gustafsdottir et al. without image correction. The fingerprints were annotated with the mode of action provided in Rohban et al.<sup>16</sup> The BBBC022 data were not scaled, but the well averages were z-normalized per plate with respect to the DMSO controls. The BBBC021 dataset consists of three replicates, whereas the BBBC022 dataset consists of replicates. Either the replicates were analyzed separately or the median of the replicates was calculated.

### Binning Analysis

For subpopulation analysis, we applied the binning analysis that was previously described.<sup>17</sup> Briefly, the DMSO population for each parameter was sorted from the smallest to the biggest value and the cutoff values at the 20th, 40th, 60th, 80th, and 100th percentiles were determined, creating five bins. Each DMSO well then had five bins per parameter, all containing approximately 20% of the population. The distribution of the counts was approximately normal, allowing us to compute a z score for the counts in each bin and each parameter of each experimental well. The binning analysis was carried out per plate with the DMSO population of the entire plate.

### Stability of Parameter Analysis

The range of each 20% bin of each parameter was determined over the entire DMSO population of a plate. For each well of the plate, it was then expected that 20% of the population should be found in each bin for each parameter. The actual counts of cells per well and parameter followed a multinomial distribution since the procedure is similar to drawing a ball from an urn with replacement. Therefore, to test whether the obtained distribution of counts for each parameter and each well corresponded to a 20%, 20%, 20%, 20%, 20% distribution, a chi-square test was applied. The resulting *p* value was adjusted for multiple testing with the Bonferroni method.

### L1 Regularization

L1 regularization was implemented in the KNIME workflow in a Python scripting node. LogisticRegression from the sklearn library was used, and the tuning parameter alpha was set at 0.01, 0.05, 0.1, 0.5, 1, 2, 4, 6, 8, and 10. At each alpha, some parameters had 0 coefficients and fingerprints were constructed without these parameters and used for classification.

### Classification

The distance matrix of the profiles was calculated with cosine as the distance measure. The nearest-neighbor algorithm with leave-one-out cross-validation was applied to determine the most similar compound and its mode of action. The number of correctly classified compounds was reported.

### Software

The workflows were implemented in KNIME. The normalization script was written in Python and a graphical user interface (GUI) was generated for KNIME with the RGG interface. The chi-square test was carried out with R from within the KNIME workflow using a R snippet node. All workflows and instructions for installing KNIME with extensions are available for download at <https://publications.mpi-cbg.de/publications-sites/7217/>.

## Results

### Fingerprint Profiling with Well Averages

We wished to test whether analyzing subpopulations and reducing the parameter space by selecting parameters that describe a negative control population robustly could improve phenotypic fingerprint classification tasks. To this end, we built an analysis pipeline based on the benchmarking datasets BBBC021 and BBBC022 of the Broad Institute. For the BBBC021 dataset, we downloaded the fingerprints, and for the BBBC022 dataset, we downloaded the images and performed image analysis with the CellProfiler pipeline provided with the dataset. We thereby obtained two annotated datasets of individual cells treated with compounds and described by many quantitative parameters. The BBBC021 dataset has 454,695 cells treated with 96 compounds in triplicate and 7 compounds in duplicate described by 460 parameters. Only a subset of the BBBC022 dataset was used in this analysis, representing 2,894,683 cells treated with 591 compounds classified in 59 modes of action.

First, we tested the simplest analysis possible for classifying fingerprints by calculating the arithmetic mean of the single-cell data of each well. We classified both the individual profiles and the median of the replicates of both datasets. For classification, we found the mode of action of the most similar compound based on its cosine distance. We found that for the BBBC021 dataset, 289 out of 302 (94.7%) individual fingerprints were correctly classified (**Table 1**). The BBBC022 dataset showed poorer classification, with 419 out of 2372 individual fingerprints (17.7%) correctly classified. In summary, the BBBC021 dataset can be classified with very high accuracy using the average of the cell data, whereas the BBBC022 dataset shows much more modest accuracy.

**Table 1.** Classification Accuracy of the Various Methods with the BBBC021 and BBBC022 Datasets.

Method	BBBC021			BBBC022		
	% Correct MoA	Correct MOA	Wrong MOA	% Correct MoA	Correct MOA	Wrong MOA
Well averages	95.70%	289	13	17.66%	419	1953
Binned parameters	94.70%	286	16	19.39%	460	1912
Binned parameters, 20% and 100% bins	94.04%	284	18	19.65%	466	1906
Well averages, L1-norm regularization	92.05%	278	24	20.53%	487	1885
Binned parameters, L1-norm regularization	95.03%	287	15	19.56%	464	1908
Binned stable parameters	95.03%	287	15	19.22%	456	1916
Binned stable parameters, 20% and 100% bins	93.38%	282	20	20.19%	479	1893

MoA = mode of action.

### Fingerprint Profiling with Subpopulation

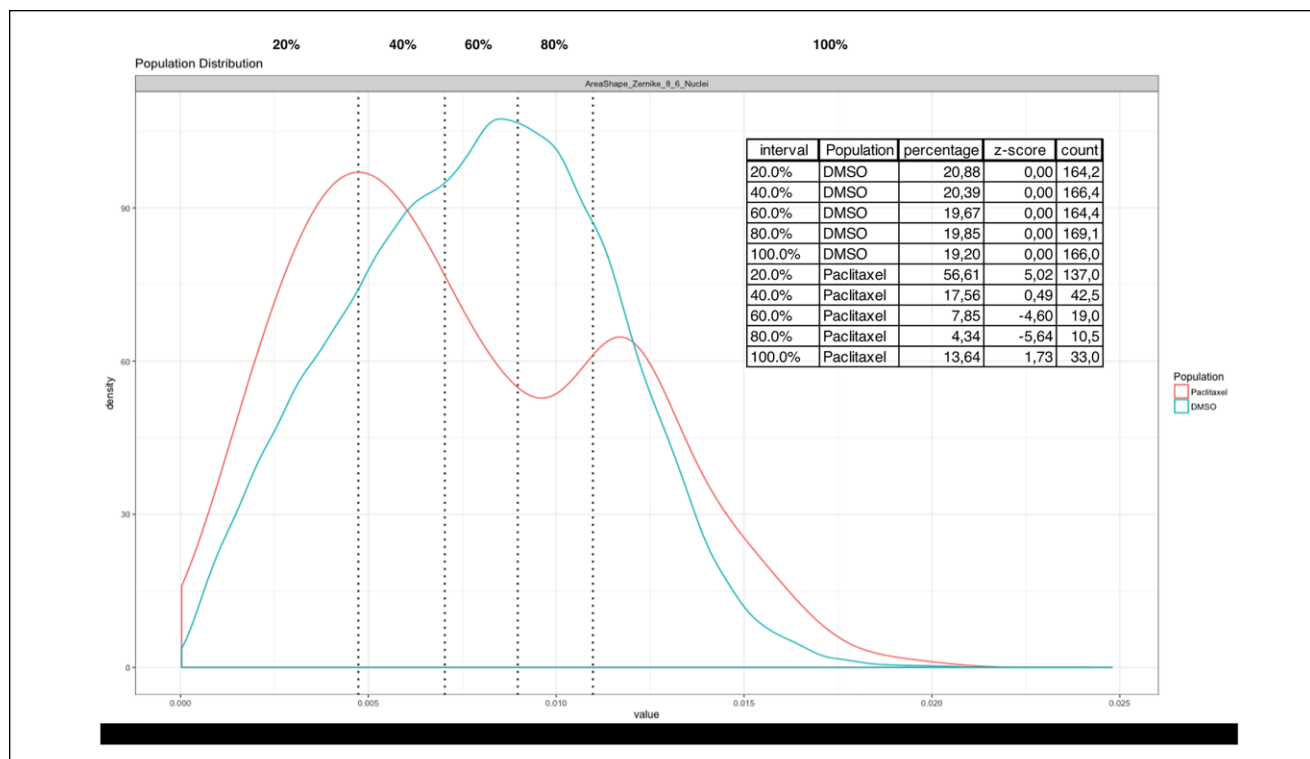
After establishing a baseline performance for profiling compounds using population means, we wanted to investigate whether incorporating subpopulation information could be beneficial for the accuracy of the classification. We have previously developed a method for subpopulation analysis called binning analysis.<sup>17</sup> Briefly, users define how many equally sized bins into which to split the populations in the wells. For five bins, the 20th, 40th, 60th, and 80th percentiles are determined for the control population for each parameter. Control populations can be all the negative wells in a plate, or in batches of plates or an entire screen. Each bin therefore determines a parameter range for each percentile. The limits of the ranges are then applied to experimental wells to determine the number of cells in each bin. For each well, each parameter, and each bin, the output of our method yields the population percentage, the counts, and the z scores based on the mean and standard deviation of the percentages of the negative control population. Phenotypic changes result in a shift of the population distribution among the bins compared with the negative control. To illustrate the behavior, we show for the BBBC022 dataset the eighth order of the Zernike polynomial at a scale of 2 pixels for DMSO and paclitaxel with five bins. The population distribution of paclitaxel is shifted toward the higher (100%) and lower (20%) bins (**Fig. 1**). Averaging such a parameter for the well population yields a z score of  $-3$ , whereas the binning analysis yields a z score of 6.7 for the 20% bin, 2 for the 40% bin, 6.6 for the 60% bin,  $-9.1$  for the 80% bin, and 4.7 for the 100% bin. This method therefore allows us to detect weak phenotypes and explains the population distribution inherently.

Here, we carried out the binning with five bins using the DMSO population distribution per plate. To estimate the effect of binning on the ability to distinguish the various

modes of action present in the dataset, we again performed nearest-neighbor classification. Since for some parameters not all bins are instantiated, we first remove parameters with missing values in order to calculate the cosine distance matrix used in nearest-neighbor classification. Binning the BBBC021 dataset resulted in a vector with 2247 parameters for the individual replicates. With the binned data, we correctly classified 286 out of 302 compounds (94.7%) (**Table 1**). The subpopulation analysis of the binned profiles correctly classified one compound less than the mean profiles. We reasoned that the accuracy could be improved just by carrying out the classification using either the lowest-range bin (first 0%–20% bin) or the largest-range bin (80%–100% bin), since changes in phenotypes will be reflected by parameters becoming either smaller or greater, and thus only the extreme bins will be affected. Carrying out nearest-neighbor classification with the 100% and 20% bins decreased the classification performance to 284 out of 302 replicates (94.0%). The same analysis for the BBBC022 dataset yielded a classification with all bins of 460 correctly classified profiles (19.4%), compared with 419 with well population average. Using only the 100% and 20% bins, we were able to further improve the classification to 466 correctly classified profiles (19.7%). In summary, classifying compounds with the 100% and 20% bin profiles significantly improved the classification accuracy for the BBBC022 dataset, but slightly decreased the classification accuracy for the BBBC021 dataset.

### L1-Norm Regularization for Parameter Reduction

A disadvantage of our binning analysis is that the number of parameters increases, which worsens the dimensionality problem for distinguishing phenotypes. We therefore wanted to reduce the number of parameters and obtain a



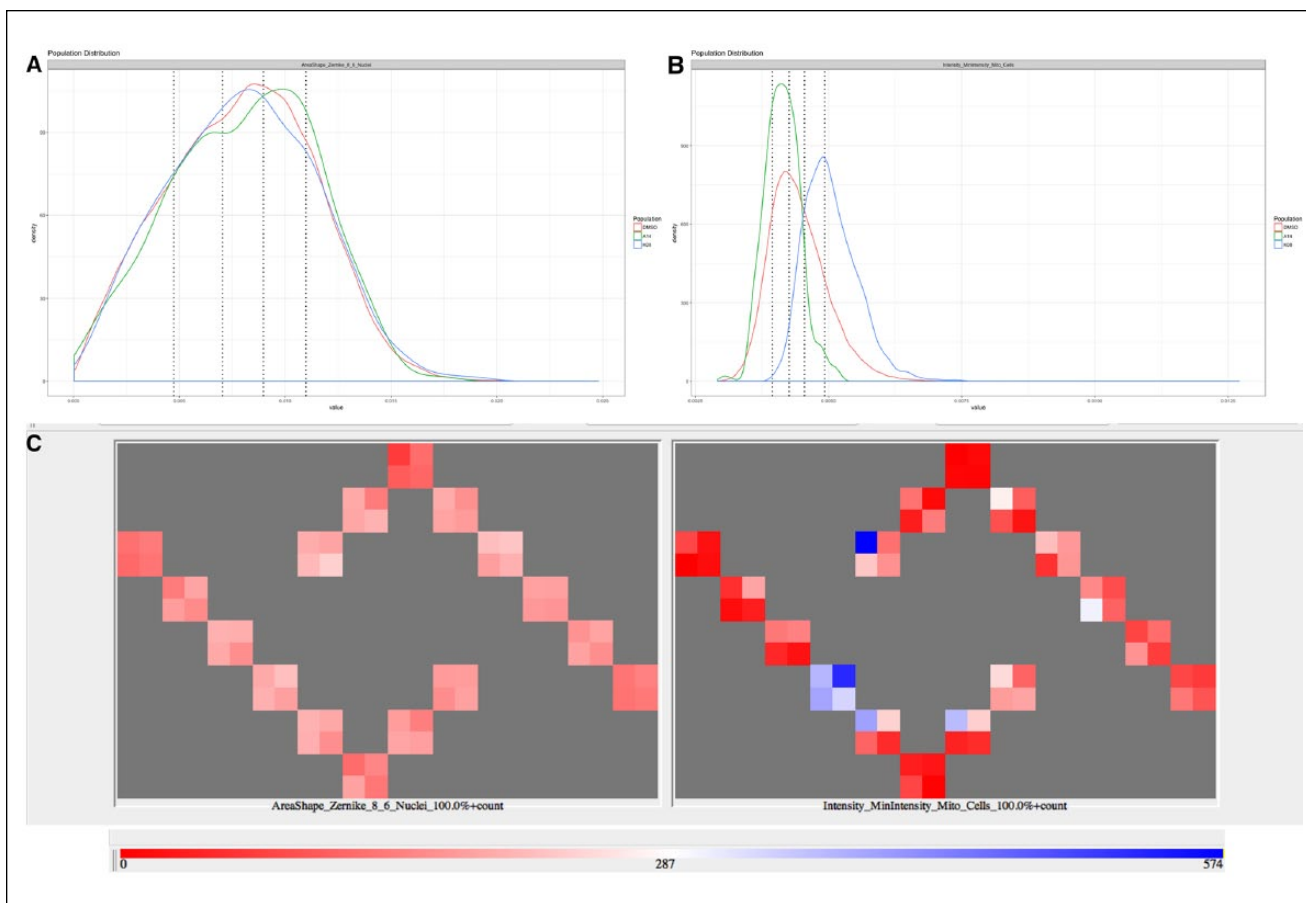
**Figure 1.** Distribution of the cell populations of DMSO and paclitaxel in plate 20585 of the BBBC022 dataset for the eighth order of the Zernike polynomial at a scale of 6 pixels of the nuclei. The dotted lines represent the separation between the 20th, 40th, 60th, and 80th percentiles. Paclitaxel transforms a unimodal distribution to a bimodal distribution, resulting in an increased number of objects in the low 20% and 100% bins. The inset shows the percentage, z score, and count for each bin for the DMSO population of the plate and the paclitaxel well.

sparse matrix capable of distinguishing better phenotypes. A popular approach to reduce the parameter space when a ground truth is known is L1-norm regularization.<sup>12</sup> The ground truth is used to carry multiple logistic regression, and L1-norm regularization sets parameters to 0. A tuning factor alpha is scanned with values between 0.01 and 10, with 10 different values yielding 10 fingerprints with different selections of parameters. The classification task was carried out with each fingerprint and the accuracy assessed. We first applied L1-norm regularization to both the well-averaged populations and the binned population. The accuracy of the classification with L1-norm regularization for the well-averaged populations of the BBBC021 dataset was decreased, with 278 compounds correctly classified, compared with 289 for the entire parameter set. L1-norm regularization modestly improved the classification of the binned individual replicates to 287 correct classifications, compared with 286 for all the bins. For the well-averaged BBBC022 dataset, L1-norm regularization increased the classification accuracy to 487 (20.5%) profiles compared with 419 for the well averages. L1-norm regularization did not improve the classification of the binned data, with 464 (19.7%) of correctly classified

profiles compared with 466 profiles with the high and low bin profiles only.

### Removal of Noisy Parameters

Obtaining a sparse matrix with L1-norm regularization biases the model toward detecting phenotypes that are present in the training set. The optimized parameter set might not generalize well and might be unable to correctly classify unseen phenotypes. To avoid the problem of overfitting, we focused on selecting parameters only based on negative controls. We reasoned that parameters that describe the negative population faithfully and reproducibly should be sensitive to detecting deviations from the unperturbed state. We therefore analyzed the distribution of counts for each parameter in each well to assess whether it followed the expected distribution of the 20th percentile of the population in each bin. The distribution of the counts per bin is a multinomial distribution, and a chi-square test yields a  $p$  value for the likelihood that the observed counts belong to the expected multinomial distribution. The  $p$  value was then adjusted for multiple testing and a cutoff value of 0.05 was chosen to classify whether the counts obtained for a parameter in a



**Figure 2.** (A) Population distribution of a stable parameter (eighth order of the Zernike polynomial at a scale of 6 pixels) for the entire DMSO population of a plate (red) and two DMSO wells from the same plate (green, blue). (B) Population distribution of a noisy parameter (minimal intensity of Mitotracker in cells) for the entire DMSO population of a plate (red) and two DMSO wells from the same plate (green, blue). (C) Plate heatmap of the number of cells in the 100% bin for the DMSO wells of a single plate for a stable parameter (eighth order of the Zernike polynomial at a scale of 6 pixels; left) and a noisy parameter (minimal intensity of Mitotracker in cells; right).

well follow the expected multinomial distribution. For some parameters, only few wells showed the expected distribution, which we call noisy parameters, whereas for other parameters all wells were within the expected distribution, and we call these parameters stable (Fig. 2). We then built phenotypic fingerprints removing stepwise noisy parameters, obtaining fingerprints with more and more stable parameters (above the 0.05 cutoff). To build the fingerprints, we first removed parameters where only 1% of DMSO wells were above the 0.05 cutoff, then 6%, and so forth, until 96% of the DMSO wells were above the 0.05 cutoff in steps of 5%. We then performed the classification task with the 20 resulting fingerprints. For the BBBC021 dataset, we found that 287 compounds were correctly classified (95.0%) with the all bins, and using only the 100% and 20% bins, 282 compounds were correctly assigned their mode of action (93.4%). For the BBBC022 datasets, 456 (19.2%) and 479 (20.2%) compounds were correctly classified with all the

bins or the 100% and 20% bins only, respectively. In summary, selecting stable parameters increased the classification accuracy compared with all binned parameters. Selecting parameters by L1-norm regularization was in most cases more powerful, though.

### Generalization of the Selected Parameters

Our rationale to select parameters based only on the description of the negative control was that we were not biasing the model toward known modes of action, whereas L1-norm regularization optimizes the model on the known modes of action. To directly test whether our method generalizes better, we retrained the models omitting a mode of action with many compounds. We chose dopamine receptor antagonist with 180 compounds and glucocorticoid receptor agonist with 140 compounds. We trained the L1-norm regularization model and selected the best threshold for stable parameters without

one of these modes of action and verified the classification accuracy after adding the compounds and their mode of action back in the test set. Thus, the models were trained without seeing a class, which was then later added. We found that for the dopamine receptor antagonist mode of action, 118 compounds were not correctly classified with L1-norm regularization, whereas 115 compounds were not correctly classified with the stable parameter selection for either all bins or just the 20% and 100% bins. For the glucocorticoid receptor agonist mode of action, we found that 45 compounds were misclassified by L1-norm regularization compared with 49 and 37 for the stable parameter selection model. These results suggest that our stable parameter selection method with the 20% and 100% bins only allows building models that generalize better than L1-norm regularization.

## Discussion

We have taken advantage of two publicly available annotated high-content screening datasets to explore fingerprint analysis methods. We have compared a simple averaging of the well populations to subpopulation analysis and have compared the effect of dimensionality reduction on the classification task. The subpopulation analysis is based on the user-defined percentiles of the negative control population determining their parameter range. These parameter ranges are then used to determine the count of cells in each well. The number of bins in the analysis is important. At one extreme, having a single bin reduces the analysis to a simple cell count. At the other extreme, using too many bins, some bins will be empty or have only very few objects. It is therefore important to strike a balance between a high-resolution analysis of the data and no resolution at all. Additionally, some parameters cannot be binned, as they have too few values or a vast majority of the population have only one of the values. For instance, the number of nuclei per cell is often one. A binning of five for the number of nuclei per cell will not succeed, since a vast majority of cells will have a value of one and only a very few multinucleate cells will not be sufficient to create several 20% bins. In such a case, only the low-interval bin (20%) will be instantiated.

In our binning analysis, phenotypes causing small changes in parameters shifted the distribution of the population away from the user-defined percentiles. The significance of the distribution changes can be tested with a chi-square test. This method has the advantage that it handles various distribution types, such as bimodal or normal, equally well. One disadvantage is that statistical tests are sensitive to the number of objects, as the power of the test increases with larger populations. With small populations of a few hundred objects, large deviations from the expected quantiles are tolerated by the chi-square test, whereas only small deviations are tolerated with populations of several thousand cells. Due to this caveat, our method to detect

noisy parameters for dimensionality reduction needs to be optimized for each dataset. Noisy parameters are determined by analyzing the negative controls only, thereby avoiding biasing the selection of parameters toward known compounds. The estimation of noisiness of a parameter is carried out by testing how many wells of the negative control population do not show the expected quantiles. We show that removal of the noisiest parameters does improve the classification accuracy substantially, although the best accuracy in the BBBC021 and BBBC022 is not achieved by this method. For the BBBC021, the best accuracy is achieved by well average, which is the simplest method. It must be noted that the BBBC021 dataset is very biased, containing very well-behaved experimental wells hand-picked out of a larger dataset, and therefore achieves very high accuracy. Such accuracy is unlikely to be found in other datasets. For the BBBC022, the best accuracy is achieved with L1-norm regularization of the averaged population. For both the BBBC021 and the BBBC022 datasets, our method with binned, stable parameters is the second best. Given the good performance of our method in these two examples, it is likely that it will also perform well with other datasets. Generalization of the model obtained is one of the interesting features of our method. We show that selecting stable parameters with an unseen mode of action and then classifying the data with a new class leads to better accuracy than L1-norm regularization. This suggests that it is possible to train a model and reuse it even though the number and quality of classes have changed. Thus, when a researcher finds that the annotation of a compound has changed or been determined, all the compounds can be reclassified without first selecting stable parameters again.

Our analysis method can also be applied for tasks unrelated to fingerprint analysis. In our experience, the subpopulation analysis allows the detection of phenotypes where only a minority of cells show a phenotype. Our KNIME node calculates a z score for each parameter and each bin, based on the mean and standard deviation of the percentages in the DMSO wells. These z scores are used to detect phenotypes in screens, and we have detected weak, reproducible hits that were not detectable by population averages.

Our methodology could be further improved with more informative fingerprints. Here, we used only low-level features extracted from segmented images. Another approach could consist of applying machine learning such as deep learning to combine low-level parameters, either directly from the pixel intensities or from segmentation, and learn the relevant features.<sup>18</sup> Such fingerprints containing high-level features might perform much better than the low-level features we used here.

We created our analysis in KNIME and made it publicly available with the analyzed data. Thanks to the intuitive user interface, non-computer scientists will be able to utilize the method for their own work. We hope that the



community will take advantage of the method and provide feedback to further improve its performance.

### Acknowledgments

We are thankful to Prof. Yannis Kalaidzidis and Dr. Christophe Zechner for discussions about the methodology.

### Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD

Marc Bickle  <https://orcid.org/0000-0002-8918-1869>

### References

1. Singh, S.; Carpenter, A. E.; Genovesio, A. Increasing the Content of High-Content Screening: An Overview. *J. Biomol. Screen.* **2014**, *19*, 640–650.
2. Zhang, X.; Götte, M.; Ibig-Rehm, Y.; et al. Identification of SPPL2a Inhibitors by Multiparametric Analysis of a High-Content Ultra-High-Throughput Screen. *SLAS Discov.* **2017**, *22*, 1106–1119.
3. Tao, C. Y.; Hoyt, J.; Yan, F. A Support Vector Machine Classifier for Recognizing Mitotic Subphases Using High-Content Screening Data. *J. Biomol. Screen.* **2007**, *12*, 490–496.
4. Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; et al. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3*, 4713–4723.
5. Kraus, O. Z.; Grys, B. T.; Ba, J.; et al. Automated Analysis of High-Content Microscopy Data with Deep Learning. *Mol. Syst. Biol.* **2017**, *13*, 924–924.
6. Dao, D.; Fraser, A. N.; Hung, J.; et al. CellProfiler Analyst: Interactive Data Exploration, Analysis and Classification of Large Biological Image Sets. *Bioinformatics* **2016**, *32*, 3210–3212.
7. Bendale, A.; Boulton, T. E. Towards Open World Recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 7–12, 2015; IEEE: Piscataway, NJ, 2015; pp 1893–1902.
8. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Arxiv* **2015**, arXiv:1506.02142.
9. Shamir, L. Assessing the Efficacy of Low-Level Image Content Descriptors for Computer-Based Fluorescence Microscopy Image Analysis. *J. Microsc.* **2011**, *243*, 284–292.
10. Reisen, F.; Zhang, X.; Gabriel, D.; et al. Benchmarking of Multivariate Similarity Measures for High-Content Screening Fingerprints in Phenotypic Drug Discovery. *J. Biomol. Screen.* **2013**, *18*, 1284–1297.
11. Ljosa, V.; Caie, P. D.; ter Horst, R.; et al. Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment. *J. Biomol. Screen.* **2013**, *18*, 1321–1329.
12. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, *58*, 267–288.
13. Caie, P. D.; Walls, R. E.; Ingleston-Orme, A.; et al. High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells. *Mol. Cancer Ther.* **2010**, *9*, 1913–1926.
14. Ljosa, V.; Sokolnicki, K. L.; Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat Methods* **2012**, *9*, 637.
15. Gustafsdottir, S. M.; Ljosa, V.; Sokolnicki, K. L.; et al. Multiplex Cytological Profiling Assay to Measure Diverse Cellular States. *PLoS One* **2013**, *8*, e80999.
16. Rohban, M. H.; Singh, S.; Carpenter, A. E. Capturing Single-Cell Heterogeneity via Data Fusion Improves Image-Based Profiling. *bioRxiv* **2018**. DOI: 10.1101/328542.
17. Stöter, M.; Niederlein, A.; Barsacchi, R.; et al. CellProfiler and KNIME: Open Source Tools for High Content Screening. In *Target Identification and Validation in Drug Discovery: Methods and Protocols*; Moll, J., Colombo, R., Eds.; Humana Press: New York, **2013**; pp 105–122.
18. Caicedo, J. C.; McQuin, C.; Goodman, A.; et al. Weakly Supervised Learning of Single-Cell Feature Embeddings. *bioRxiv* **2018**. DOI: 10.1101/293431.