Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment

Vebjorn Ljosa, Peter D. Caie, Rob ter Horst, Katherine L. Sokolnicki, Emma L. Jenkins, Sandeep Daya, Mark E. Roberts, Thouis R. Jones, Shantanu Singh, Auguste Genovesio, Paul A. Clemons, Neil O. Carragher and Anne E. Carpenter

J Biomol Screen 2013 18: 1321 originally published online 17 September 2013
DOI: 10.1177/1087057113503553

Journal of Biomolecular Screening
18(10) 1321–1329

DOI: 10.1177/1087057113503553
jbx.sagepub.com
SAGE

Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment

Vebjorn Ljosa[1], Peter D. Caie[2*#], Rob ter Horst[3#], Katherine L. Sokolnicki[1†], Emma L. Jenkins[4], Sandeep Daya[2], Mark E. Roberts[5], Thouis R. Jones[1‡], Shantanu Singh[1], Auguste Genovesio[1¶], Paul A. Clemons[1], Neil O. Carragher[2§], and Anne E. Carpenter[1]

Abstract

Quantitative microscopy has proven a versatile and powerful phenotypic screening technique. Recently, image-based profiling has shown promise as a means for broadly characterizing molecules' effects on cells in several drug-discovery applications, including target-agnostic screening and predicting a compound's mechanism of action (MOA). Several profiling methods have been proposed, but little is known about their comparative performance, impeding the wider adoption and further development of image-based profiling. We compared these methods by applying them to a widely applicable assay of cultured cells and measuring the ability of each method to predict the MOA of a compendium of drugs. A very simple method that is based on population means performed as well as methods designed to take advantage of the measurements of individual cells. This is surprising because many treatments induced a heterogeneous phenotypic response across the cell population in each sample. Another simple method, which performs factor analysis on the cellular measurements before averaging them, provided substantial improvement and was able to predict MOA correctly for 94% of the treatments in our ground-truth set. To facilitate the ready application and future development of image-based phenotypic profiling methods, we provide our complete ground-truth and test data sets, as well as open-source implementations of the various methods in a common software framework.

Keywords

phenotypic screening, high-content screening, image-based screening, drug profiling

Introduction

Image-based screens for particular cellular phenotypes are a proven technology contributing to the emergence of high-content screening as an effective drug- and target-discovery strategy.[1] Phenotypic screening has also been proposed as a strategy to assess the efficacy and safety of drug candidates in complex biological systems[2]; when applied at early stages in the drug-discovery process to relevant biological models, quantitative microscopy may help reduce the high levels of late-stage project attrition associated with target-directed drug-discovery strategies. Retrospective analysis of all drugs approved by the Food and Drug Administration (FDA) between 1999 and 2008 reveal that significantly more were discovered by phenotype-based screening approaches than by the more broadly adopted target-based screening model.[3] Screens for phenotypes that can be identified in a microscopy assay by a single measurement, such as cell size, DNA content, cytoplasm-nucleus translocation, or the intensity of a reporter stain, are widely used in pharmaceutical and academic labs, especially in standard cell

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]AstraZeneca Pharmaceuticals, UK
[3]Radboud University, Nijmegen, The Netherlands
[4]AstraZeneca Pharmaceuticals, Waltham, MA, USA
[5]Tessella plc, Abingdon, Oxfordshire, UK
*Present address: Breakthrough Research Unit, University of Edinburgh, Edinburgh, UK
†Present address: inviCRO, Boston, MA, USA
‡Present address: School of Engineering & Applied Sciences, Harvard University, Cambridge, MA, USA
§Present address: Edinburgh Cancer Research UK Centre, University of Edinburgh, Edinburgh, UK
¶Present address: École Normale Supérieure, Paris, France
#These authors contributed equally.

Received Mar 14, 2013, and in revised form Jul 4, 2013. Accepted for publication Jul 17, 2013.

Supplementary material for this article is available on the *Journal of Biomolecular Screening* Web site at http://jbx.sagepub.com/supplemental.

Corresponding Author:
Anne E. Carpenter, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA.
Email: anne@broadinstitute.org

lines and engineered reporter systems.[4] Even complex phenotypes, which require that machine learning be used to combine the measurements of many cellular properties, are now scored routinely in some laboratories.[5,6] Evidently, quantitative microscopy is a versatile and powerful readout for many cell states.

Profiling cell-based phenotypes is the next challenge for quantitative microscopy.[7] The principle of phenotypic profiling is to summarize multiparametric, feature-based analysis of cellular phenotypes of each sample so that similarities between profiles reflect similarities between samples.[8] Profiling is well established for biological readouts such as transcript expression and proteomics.[7,9] Comparatively, image-based profiling comes at a much lower cost, can be scaled to medium and high throughput with relative ease, and provides single-cell resolution. Although image-based screens aim to score samples with respect to one or a few known phenotypes, profiling experiments aim to capture phenotypes not known in advance, using label sets that can detect a variety of subtle cellular responses without focusing on particular pathways. Such unbiased, phenotypic profiling approaches provide an opportunity for more opportunistic, evidence-led drug discovery strategies that are agnostic to drug target or preconceived assumptions of mechanism of action (MOA). The potential applications of profiling are extensive:

- Predict the MOA of a new, unannotated compound by finding well-characterized compounds that have similar profiles
- Identify concentrations of compounds that have off-target effects
- Start with a large number of hit compounds yielding the same specific phenotype in a screen and select a subset for follow-up that represent their diversity in terms of overall cellular effects
- Identify compounds with a novel MOA, suggesting new targets
- Group a large collection of unannotated compounds into clusters that have the same MOA
- Discover synergistic effects of combinations of compounds
- Discover pathway targets possessing synergistic, additive, synthetically lethal, or chemosensitizing properties from combined genetic perturbation and small-molecule perturbation
- Provide iterative guidance to rational polypharmacology strategies
- Predict the protein target of a compound by finding the RNAi reagent that produces the most similar profile
- Identify compounds with cell line–specific effects by comparing the compounds' profiles across many cell lines, then relate to mutation status to further define MOA and develop patient-stratification hypotheses

Most image-based profiling experiments thus far have been performed at the proof-of-principle scale, with a focus on developing computational methods for generating and comparing profiles. This article describes and compares five methods that have been proposed for profiling and shown to be effective in a particular experiment. The methods range from simple and fast to complicated and computationally intensive, and they differ greatly in how explicitly they take advantage of the individual-cell measurements to describe heterogeneous populations. Little is known about how the methods compare because each method was proposed as part of a more extensive methodology, often with different goals and with different types of data available (multiple concentrations, cell lines, or marker sets). We extracted the core profiling methods—namely, the algorithms for constructing per-sample profiles from per-cell measurements—from the larger methodologies, applied them to a typical experiment, and compared their ability to classify compounds into MOA. Our test experiment uses a physiologically relevant p53–wild-type breast cancer model system (MCF-7) and a mechanistically distinct set of targeted and cancer-relevant cytotoxic compounds that induces a broad range of gross and subtle phenotypes.[10] We provide our ground-truth and test data sets and open-source implementations of the methods to allow others to readily apply the methods and to extend the comparative analysis to additional methods and data sets.

Materials and Methods

Sample Preparation and Image Analysis

MCF-7 breast cancer cells were previously plated in 96-well plates; treated for 24 h with 113 compounds at eight concentrations in triplicate; labeled with fluorescent markers for DNA, actin filaments, and β-tubulin; and imaged as described.[10] Version 1.0.9405 of the image analysis software CellProfiler[11,12] measured 453 features (**Suppl. Table S1**) of each of the 2.2 million cells, using the pipelines provided (**Suppl. Data S1**).

Profiling

Before applying any of the profiling methods, the cell measurements were scaled linearly to remove interplate variation. For each feature, the first percentile of DMSO-treated cells was set to 0 and the 99th percentile was set to 1 for each plate separately. The same transformation functions were then applied to all compounds on the same plate, the assumption being that the DMSO distributions should be similar on each plate.

Per-sample profiles were computed from per-cell measurements by one of the profiling methods (see below). The treatment profile was constructed by taking the element-wise

median of the profiles of the three replicate samples. Using the cosine distance between the profiles as a measure of distance, each sample was predicted to have the MOA of the closest profile from a different compound ("nearest-neighbor classification"). The cosine distance is defined as

$$1 - \cos\theta = 1 - AB / (\|A\|\|B\|). \quad (1)$$

A cosine distance of 0 indicates that two vectors have identical directions, and a cosine distance of 2 indicates that two vectors have opposite directions. Two vectors are orthogonal if the cosine distance is equal to 1.

We chose simple, transparent methods for combining replicates, computing distances, and classifying profiles because our goal was to compare the core profiling methods rather than devise an optimal end-to-end analysis pipeline. In a real profiling application, other choices may be advantageous; for instance, the problem of classifying compounds into mechanisms is likely amenable to supervised classification approaches.

Profiling Methods

Means. The average is taken over all scaled features for each sample. Adams et al.[13] use this method but extend their profiles with means for different cell-cycle phases, some intensity proportions, and some standard deviations.

KS Statistic. The *i*-th element of the profile for a sample is the Kolmogorov-Smirnov (KS) statistic between the distribution of the *i*-th measurement of the cells in the sample with reference to mock-treated cells on the same microtiter plate. The KS statistic is calculated by taking the maximum distance between the empirical cumulative distribution functions (cdfs). Following Perlman et al.,[14] we used a nonstandard "signed" KS statistic that indicates whether the maximum distance is positive or negative.

Perlman et al.[14] describe this method in the context of a more extensive methodology that compares compounds over a range of concentrations, trying different alignments of the compounds' concentration ranges in order to produce a "titration-invariant similarity score." This procedure is independent of the underlying core profiling method and could therefore be used with any of the five methods tested here. We did not use it because the cosine distance was a stable measure of profile similarity in our experiment, even across concentrations (data not shown).

Normal Vector to Support-Vector Machine Hyperplanes. Support-vector machines (SVMs) were trained to distinguish the cells in each sample from mock-treated cells on the same microtiter plate.

SVM recursive feature elimination (SVM-RFE) starts by training an SVM model to distinguish a treatment from DMSO. The prediction accuracy is estimated using cross-validation. The *n* measurements with the lowest weight are then removed, and a new model is trained using the remaining measurements. This continues iteratively until one feature remains. Finally, the SVM model with the best prediction accuracy is selected. The best feature selection accuracy is theoretically obtained by removing one feature at a time (SVM-RFE1); however, this is computationally expensive. Therefore, following Loo et al.,[15] we used SVM-RFE2, which removes the 10% of the measurements with the lowest weight at each iteration. To eliminate more measurements, Loo et al.[15] eliminated measurements until the prediction accuracy fell below $0.9 \times ((C_{max} - C_{min}) + C_{min})$, where C_{max} is the maximum prediction accuracy and C_{min} the minimal prediction accuracy over the full range of a selected number of measurements.

Gaussian Mixture Modeling. To build Gaussian mixture (GM) profiles, 10% of the data were subsampled uniformly across all samples. This selection was mean-centered, after which the data were transformed using principal-component analysis (PCA), retaining enough principal components to explain 80% of the variance (~54 for our data set). Next, a GM model was fit to the data using the expectation-maximization (EM) algorithm. The algorithm was initialized with unit covariance and the centroid positions obtained using the k-means algorithm. The starting positions of the centroids in the k-means algorithm were initialized randomly, meaning the algorithm is nondeterministic. The Gaussians resulting from the EM algorithm were used as a model for the remaining 90% of the data. The rest of the data were centered using the mean of the data that was used to build GM models and projected into the same loading space. For each cell, the posterior probability of belonging to each of the Gaussians was computed. Profiles were constructed by averaging these posterior probabilities for each compound concentration. The number of values in a profile is thus equal to the number of Gaussians used to model the data. The best number of Gaussians was chosen empirically.

Factor Analysis. This method attempts to describe the covariance relationships between the image measurements $\mathbf{x}$ in terms of a few latent random variables $\mathbf{y}$ called factors. The factors are drawn from an isotropic Gaussian distribution. The observed image measurements $\mathbf{x}$ are modeled as an affine transformation $\mathbf{Ay} + \boldsymbol{\mu}$ of the factors and a measurement-specific noise term $\boldsymbol{\nu}$:

$$\mathbf{x} = \mathbf{Ay} + \boldsymbol{\mu} + \boldsymbol{\nu}. \quad (2)$$

The observed measurements are assumed to be conditionally independent given the factors; in other words, $\mathbf{v} \sim N(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$. We estimate $\mathbf{A}$, $\boldsymbol{\mu}$, and $\mathbf{\Sigma}$ by an EM algorithm[16] implemented in the MDP toolkit (http://mdp-toolkit.sourceforge.net/). Then, we can compute the profile of a sample as the maximum a posteriori estimate of $\mathbf{y}$:

$$\mathrm{E}\left[\mathbf{y} \mid \mathbf{x}_n\right] = \mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}} + \Sigma\right)^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}\right) \quad (3)$$

where $\mathbf{x}_n$ is the vector of per-cell measurements in sample n averaged over the cells in the sample.

Available Data

To facilitate the development and evaluation of additional profiling methods, we provide our ground-truth annotations (**Suppl. Table S2**) and the measurements of each of the ~450,000 cells whose treatments are annotated. The data are supplied as comma-delimited files together with scripts for loading them into a MySQL database (**Suppl. Data S2**). The data schemas are described (**Suppl. Text S1**).

The images and metadata have been deposited with the Broad Bioimage Benchmark Collection (http://www.broadinstitute.org/bbbc/),[17] accession number BBBC021.

Software Implementations

The profiling methods are implemented as part of the open-source image data-analysis software CellProfiler Analyst (http://cellprofiler.org/). The implementations do not make assumptions that are particular to our experiment and can be readily applied to measurement data from the widely used image-analysis software CellProfiler[11,12] or data from other sources that can be imported into CellProfiler Analyst or otherwise converted to CellProfiler's database schema. The implementations contain support for parallel processing on a cluster of computers. The profiling methods can be executed as scripts from the Unix command line or used in Python programs as a module (**Suppl. Text S2**).

Reproducibility

We provide complete source code to readily reproduce most figures, tables, and other results that involve computation (**Suppl. Text S3**; **Suppl. Data S3**). **Supplemental Table S6** was constructed manually/interactively and is not reproducible.

Results

We implemented five proposed methods[13–15,18,19] for constructing per-sample profiles from per-cell measurements in a common computational framework. We benchmarked the five methods on images we had previously collected of MCF-7 breast cancer cells treated for 24 h with a collection of 113 small molecules at eight concentrations (**Suppl. Table S3**). The cells were fixed; labeled for DNA, F-actin, and β-tubulin; and imaged by fluorescent microscopy. For this study, we measured 453 standard cytometric measurements (**Suppl. Table S1**) of each cell using CellProfiler[11,12] and applied each of the five profiling methods. To be able to evaluate the performance of the profiling methods, we limited our attention to a subset of the data (our "ground-truth" data set) for which we were confident that the primary MOA of compounds was achieved at the concentration tested during the course of the experiment. (The term *mechanism of action* is used rather loosely here and refers to a sharing of similar phenotypic outcomes among different compound treatments, rather than referring strictly to modulation of a particular target or target class.) The mechanistic classes were selected so as to represent a wide cross section of cellular morphological phenotypes. The differences between phenotypes were in some cases very subtle: We were able to identify only 6 of the 12 mechanisms visually; the remainder were defined based on the literature. This carefully collected ground-truth data set consisted of 38 compounds at active concentrations. Some compounds were active at only one concentration and some at up to seven concentrations, for a total of 103 treatments (active compound concentrations) spanning 12 mechanistic classes (**Suppl. Table S2**; **Suppl. Fig. S1**). The mock treatment DMSO was included as a negative control. Using the cosine distance as a measure of profile dissimilarity, we classified the 103 treatments into MOAs by assigning to each profile the MOA of the most similar profile (**Fig. 1**, top). When classifying a treatment, all concentrations of the same compound were held out from the training set in order to prevent overtraining. The samples were prepared and imaged in 10 batches, but classes and replicates were distributed across batches and plates, respectively, so as to avoid biasing the classification (**Suppl. Text S4**).[20] Using this experimental data set, we tested five profiling methods (**Fig. 1A–E**), as detailed below.

Means

We first constructed profiles in the simplest way we could envision: average each measurement over the cells in the sample (**Fig. 1A**). A profile thus consists of a single value for each of the 453 features. This was the main approach used by Tanaka et al.[21] to discover an inhibitor of carbonyl reductase 1, although their profiles also included some statistics other than the mean.[13] With this profiling method, 83% of the compound-concentration profiles could be classified correctly (**Table 1**). The cosine distance remained effective despite the high dimensionality of the measurements, so there is no

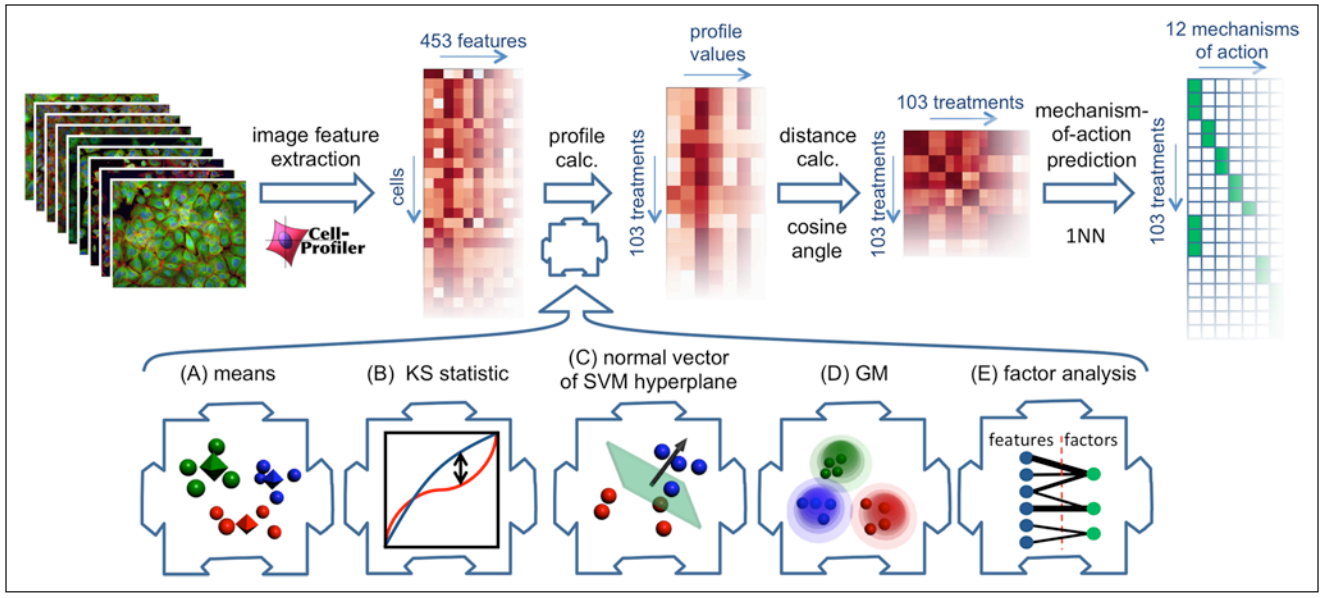


Figure 1. Overview of approach. (Top) Experimental design. Cultured cells in microtiter plates were compound treated, labeled, fixed, and imaged. The image analysis software CellProfiler measured 453 properties of each cell. One of the profiling methods under investigation condensed these measurements into a profile (vector of numbers) that describes each sample. A sample with unknown mechanism of action (MOA) was predicted to have the same MOA as the sample whose profile is most similar to that of the unknown sample, using the cosine of the angle between the profiles as measure of similarity. (Bottom) Illustrations of the five profiling methods tested. (**A**) Means of raw per-cell features. (**B**) Kolmogorov-Smirnov (KS) statistic relative to negative control. (**C**) Normal vector of decision plane of linear support-vector machine (SVM) versus negative control. (**D**) Proportion of cells in each component of a Gaussian mixture (GM). (**E**) Latent feature extraction using factor analysis.

Table 1. Accuracies for classifying compound treatments into mechanisms of action.

Method	Accuracy, %
Means	83
KS statistic	83
Normal vector to support-vector machine hyperplane	81
With recursive feature elimination	64
Distribution over Gaussian mixture components	83
Factor analysis + means	94

significant compression of distances, a common problem in high-dimensional data analysis in which the distance to the nearest point approaches the distance to the farthest point (**Suppl. Fig S2**). This indicates that most of the measurements contribute information about MOA and are not simply redundant measurements that add noise.[22]

That small-molecule effects could be characterized so well by the shift in means was unexpected because many treatments induce a heterogeneous phenotypic response across the cell population in each sample. For instance, treatment with microtubule destabilizers produced a mixture of ~44% mitotic cells, ~27% cells with fragmented nuclei, ~16% cells with diffuse and faint tubulin staining, and ~12% cells with an appearance similar to mock-treated cells. Even though the "means" method made no attempt to model the subpopulations of cells, it was mostly able to distinguish microtubule destabilizers from microtubule stabilizers, which also block in M-phase and therefore also caused a high proportion of mitotic cells (**Suppl. Table S4**). There was room for improvement, however; in particular, many microtubule stabilizers and actin disruptors were misclassified as other MOAs. DNA damage agents and DNA replication inhibitors were consistently confused.

Although the image features that are most influential in distinguishing each mechanism of action from the rest (**Suppl. Table S5**) are largely expected (e.g., the texture of actin staining in the cytoplasm is important for distinguishing actin disruptors), it is notable that the profiles generally obtain their discriminatory power from a combination of image features.

Some other population statistics (medians, modes, and means combined with standard deviations) gave similar results. Medians combined with median absolute deviations achieved higher accuracy (88%), mainly by being better able to distinguish DNA damage agents and DNA replication inhibitors (**Suppl. Fig. S3**).

KS Statistic

Perlman et al.[14] used the KS statistic as part of their titration-invariant similarity score profiling method. The KS

statistic is calculated separately for each treatment and measurement. It is the maximal difference between the cumulative distribution function (cdf) of the measurements of the treated cells and the corresponding cdf of mock-treated cells (**Fig. 1B**). This method is more computationally expensive than simply computing the mean but can be more sensitive: For example, a hypothetical treatment that causes some of the cells to shrink and the rest to grow could leave the mean cell size unchanged but would increase the KS statistic.

The method based on the KS statistic reaches a prediction accuracy of 83% (**Table 1**). As with the means method, DNA damage agents and DNA replication inhibitors were confused (**Suppl. Fig. S4B**). Many DNA damage agents were additionally misclassified as Aurora kinase inhibitors, and there was some confusion between microtubule destabilizers and Eg5 kinesin inhibitors.

Normal Vectors to SVM Hyperplanes

Loo et al.[15] describe a multivariate method that trains a linear SVM[23] to distinguish compound-treated cells from mock-treated cells. The SVM constructs the maximal-margin hyperplane that separates the compound-treated and mock-treated cells in the feature space. The normal vector of this hyperplane is adopted as a profile of the sample (**Fig. 1C**). The method classified 81% of the treatments correctly (**Table 1**).

The methodology of Loo et al.[15] additionally uses SVM-RFE to remove redundant and noninformative measurements from profiles and replace them with zeros in order to increase the sensitivity of analysis and make profiles more interpretable. This feature elimination is done independently for each treatment. Adding this step reduced the classification accuracy to 64% (**Table 1**). Inspecting the lists of features chosen gives a clue to why: The SVM is being trained to distinguish a compound from DMSO, so the features most useful for this purpose are selected. These features are not generally the same features that are useful for distinguishing compounds with different MOA. Indeed, features preferentially retained by the feature-elimination step are often correlated with reduced cell count, as almost every active compound has some cytotoxic effects: Three of the five most frequently selected features are clearly influenced by cell count, having to do with number of neighbors and number of cells touching (**Suppl. Table S6**). This behavior is not a flaw in SVM-RFE: It simply magnifies the tendency of the normal-vector method to emphasize the features that most clearly separate the treated cells from mock-treated cells.

Distribution over GM Components

To better characterize heterogeneous cell populations, Slack et al.[18] proposed modeling the data as a mixture of a small number of Gaussian distributions and profiling each sample by the mean probabilities of its cells belonging to each of the Gaussians. This GM method assumes that compound treatment causes cells to shift between a limited number of general states. It is indeed generally true that cellular phenotypes induced by perturbations can usually be found, albeit at low levels, in wild-type cell populations.[5] GM models have been used in other phenotype-detection applications as well.[24]

We fitted different mixtures of Gaussians to a subsample of ~45,000 cells (10% of the cells), with the number of components ranging from 2 to 30. A nondeterministic EM algorithm was used to fit Gaussians to the data; therefore, the model construction and cross-validation was performed 20 times to assess model variability. Twenty-five Gaussians resulted in a prediction accuracy of ~83% (**Table 1**) but with large variation depending on the initial conditions (**Suppl. Fig. S5**). Increasing the number of Gaussians beyond 25 does not improve the accuracy (**Suppl. Fig. S5**). Some classification mistakes occurred in only some models, whereas others were consistent across models (**Suppl. Fig. S4E**).

The GM method performs equally well whether created from control cells or treated cells (**Fig. S6**), so the mixture components may be mainly modeling cellular phenotypes that are widely represented rather than phenotypes induced by only particular treatments.

Factor Analysis

Although we measured 453 morphological features of each cell, it is the underlying biological effects that are of interest. Young et al.[19] used factor analysis to discover such underlying effects under the assumption that an underlying process (factor) affects several measurements and that variations restricted to a single measurement are noise.

We trained a factor model on a random subsample of ~45,000 control cells (15% of the control cells in the experiment). We computed the maximum a posteriori estimate of the factors given each cell and averaged these values over all cells treated with the same compound and concentration to obtain a profile of the treatment. Varying the number of factors, we found that the performance was similar to the other methods when using ~25 factors but that performance increased gradually with the number of factors, reaching a plateau at ~50 factors (**Fig. 2**). Although the procedure is nondeterministic, the accuracy generally does not change more than 3 percentage points in either direction with a given number of factors. With 50 factors, the prediction accuracy was 94%, which is substantially better than any of the other methods that were tested (**Table 1**). There was still some confusion between DNA damage agents and DNA replication inhibitors (**Fig. 3**).

The improvement in accuracy was not simply due to the method's implicit dimensionality reduction: Reducing the dimensionality to 50 by PCA did not lead to

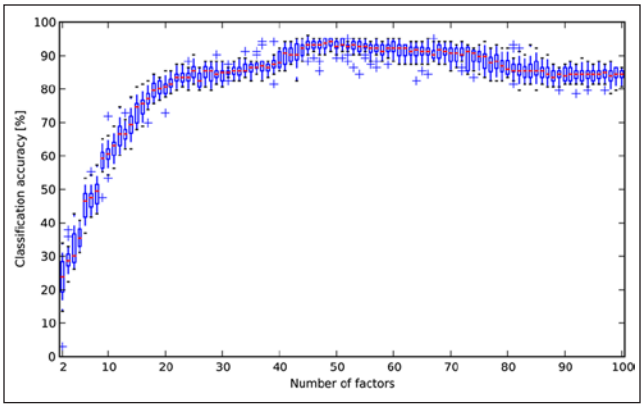


Figure 2. Distributions of classification accuracies for 20 runs of the factor analysis method for each possible choice of the number of factors from 2 to 100. The performance was similar to the other methods when using ~25 factors, but the accuracy increased gradually with the number of factors, reaching a plateau at ~50 factors.

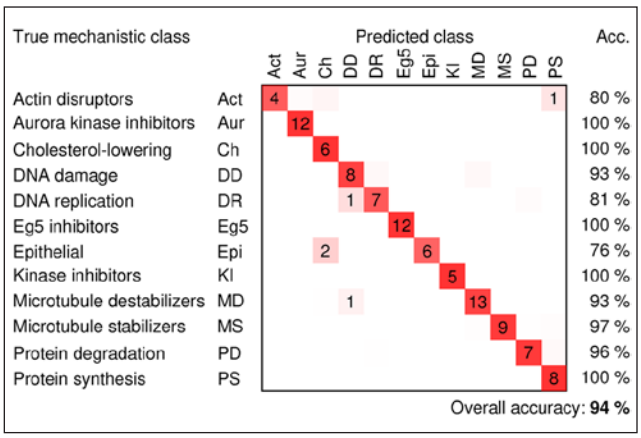


Figure 3. Confusion matrix for the factor-analysis method, showing the number of compound concentrations that were classified correctly (on the diagonals) and incorrectly (off the diagonals), the classification accuracies for each mechanism of action (right columns), and overall classification accuracy (number of correctly classified compound concentrations divided by the total number of compound concentrations). Average outcomes over 20 models are shown; dimly colored squares without numbers indicate classification outcomes that occurred fewer than 0.5 times on average.

an improvement over the means method, and selecting the feature most heavily loaded on each of the 50 factors decreased the accuracy to 63% (**Suppl. Table S7**).

The factor-analysis method can be viewed as the means method with a preprocessing step that transforms the measurements of each cell into the latent factor space. Although factor analysis greatly improves the means method, it does not improve the KS statistic method as much. Using it as a preprocessing step before any of the other profiling methods is not helpful (**Suppl. Fig. S7**).

Most of the factors cannot be readily interpreted by their feature loadings (**Suppl. Table S8**). This is an Occam dilemma[25]: When the number of factors is high enough to yield good predictive accuracy, the factors are difficult to interpret because they combine numerous features in order to pick up on subtle phenotypic differences. Although we cannot use direct interpretation to verify that the factors are biologically relevant, careful cross-validation and experimental design can guard against bias by batch effects and other artifacts[20] (**Suppl. Text S4**).

The factor model performs equally well whether created from control cells or treated cells (**Suppl. Fig. S9**). Because the wild-type variation is sufficient to elucidate the relationships between image features and latent factors, the factors may be capturing stable, fundamental modes of variation for the cell line (viewed through a particular assay and feature set) and not the extreme changes induced by particular treatments.

Discussion

We compared five methods[13–15,18,19] for generating per-sample profiles from image-based cell data in the context of classifying small molecules into 12 MOAs based on cellular morphology. All methods had previously been demonstrated in distinct experiments, mostly proof-of-principle studies, with some yielding biological discovery. However, these methods had never before been directly compared on a common data set. Each method was previously proposed as part of a larger methodology, sometimes including strategies for particular contexts, such as making use of information from multiple cell lines or multiple concentrations. These strategies can be applied independently of the core profiling method; here, we compared only the computational cores of the profiling methods. We did not evaluate the underlying statistical methods (KS, SVM, GM, factor analysis), which have solid theoretical foundations and an excellent record of solving analysis problems of many kinds.

On our data set, the simplest method, which profiles compounds by the population means of the measurements of the treated cells, performed better than expected, achieving 83% accuracy in predicting MOA. Because many of the measurements are non-Gaussian, we expected nonparametric KS statistics to be superior, but that was not the case. Describing a compound by the decision boundary of a linear SVM trained to distinguish compound-treated cells from mock-treated cells did not offer improvement either (83%), and adding a feature-reduction step reduced performance (64%). A GM method that tries to model subpopulations of cells with a mixture model might be expected to have an advantage in experiments in which the perturbations lead to shifts between a small number of discernible cell states (e.g., cell-cycle states), but we did not observe

this: Although the treated samples were heterogeneous with respect to cellular phenotype, and some phenotypes were not specific to any mechanistic class, the GM model performed no better than other methods (83%). The profiles that best represented the phenotypes were obtained using factor analysis (94% accuracy in predicting MOA). This method's potential limitation of excluding important nonredundant image-based features as noise has been demonstrated in a screening context in which only 29 measurements were made of each cell,[26] but with our higher-dimensional features, the method proved effective at extracting the underlying sources of variation.

Because all of the profiling methods we tested operate on measurements at the resolution of single cells, there was the potential that some of them might detect effects that are present in only a small subpopulation of the cells in the sample. However, only the GM method makes explicit attempts to model cell subpopulations across samples. It was therefore surprising that even the means method was sufficient to characterize treatments producing heterogeneous phenotypic response. Because compound treatments typically affect most cells in a sample (although frequently in different ways), our experimental results are insufficient to predict the methods' relative performance in RNAi screens in which the interference is effective in only a small percentage of the cells. It is possible that the KS statistic may work better than the mean in such experiments or that the GM method may be able to detect a globally popular phenotype even though it occurs at a low proportion in a particular sample. It is also possible that new profiling methods will be required to fully realize the potential of using single-cell measurements to profile samples that are distinguished only by small, subtle subpopulations of cells or to be robust to off-target effects.

The assay and compound collection chosen for this study are typical of a profiling experiment: Morphology assays are attractive for profiling because they can capture a wide variety of subtle cellular responses without focusing on particular pathways. However, there may be particular MOAs that are not displayed within the assay parameters described in this study. One important parameter is time following compound exposure. In this study, we chose 24 h following compound treatment of cells as this produced an optimal mitotic arrest phenotype in the MCF-7 cell line studied. For other cell lines or other compound classes, there may be added value gained from increasing the biological space of profiling studies by combining features quantified from multiple assays and applying the profiling methods across multiple time points following compound treatment. The choice of assay and optimal time point for profiling will likely depend on the scientific questions being asked. The chemical compounds we tested are commonly studied bioactive compounds. Therefore, the present study is valuable in providing a comparative analysis of methods in the context of one particular (but representative) profiling experiment. Creating and annotating a ground-truth set of compounds with known MOA is not trivial; we hope this work provides a template for future creation of ground-truth data sets.

With the emergence of image-based high-content screening across more complex and diverse assay formats incorporating co-cultures, stem cells, and model organisms, future studies may demonstrate that particular profiling methods perform better on specific assays, cell types, or even focused compound or siRNA libraries. Thus, we foresee additional value in providing an analysis framework and a ground-truth data set to facilitate further comparisons in the field using alternate data sets or methods. We have implemented all five methods and offer the source code (**Suppl. Text S2**), along with our entire set of cellular measurements for our ground-truth data set (**Suppl. Data S2**) so that they can aid in the future application, development, and comparison of image-based phenotypic profiling approaches.

Acknowledgments

We would like to thank Christopher Denz, Lisa Drew, Tom Houslay, Zhongwu Lai, David Logan, Melissa Passino, Tejas Shah, and J. Anthony Wilson for helpful input.

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 disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Institutes of Health (grant U54-HG005032), the National Science Foundation (grant CAREER DBI 1148823), and AstraZeneca Pharmaceuticals.

References

1. Bickle, M. High-Content Screening: A New Primary Screening Tool? *IDrugs* **2008**, *11*, 822–826.
2. Lee, J. A.; Uhlik, M. T.; Moxham, C. M.; et al. Modern Phenotypic Drug Discovery Is a Viable, Neoclassic Pharma Strategy. *J. Med. Chem*. **2012**, *55*, 4527–4538.
3. Swinney, D. C.; Anthony, J. How Were New Medicines Discovered? *Nat. Rev. Drug Discov*. **2011**, *10*, 507–519.
4. Carpenter, A. E. Image-Based Chemical Screening. *Nat. Chem. Biol*. **2007**, *3*, 461–465.
5. Jones, T. R.; Carpenter, A. E.; Lamprecht, M. R.; et al. Scoring Diverse Cellular Morphologies in Image-Based Screens with Iterative Feedback and Machine Learning. *Proc. Natl. Acad. Sci. U.S.A*. **2009**, *106*, 1826–1831.
6. Neumann, B.; Held, M.; Liebel, U.; et al. High-Throughput RNAi Screening by Time-Lapse Imaging of Live Human Cells. *Nat. Methods* **2006**, *3*, 385–390.

7. Feng, Y.; Mitchison, T. J.; Bender, A.; et al. Multi-parameter Phenotypic Profiling: Using Cellular Effects to Characterize Small-Molecule Compounds. *Nat. Rev. Drug Discov*. **2009**, *8*, 567–578.
8. Wagner, B. K.; Clemons, P. A. Connecting Synthetic Chemistry Decisions to Cell and Genome Biology Using Small-Molecule Phenotypic Profiling. *Curr. Opin. Chem. Biol*. **2009**, *13*, 539–548.
9. Lamb, J.; Crawford, E. D.; Peck, D.; et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313*, 1929–1935.
10. 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.
11. Carpenter, A. E.; Jones, T. R.; Lamprecht, M. R.; et al. CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. *Genome Biol*. **2006**, *7*, R100.
12. Kamentsky, L.; Jones, T. R.; Fraser, A.; et al. Improved Structure, Function and Compatibility for CellProfiler: Modular High-Throughput Image Analysis Software. *Bioinformatics* **2011**, *27*, 1179–1180.
13. Adams, C. L.; Kutsyy, V.; Coleman, D. A.; et al. Compound Classification Using Image-Based Cellular Phenotypes. *Methods Enzymol*. **2006**, *414*, 440–468.
14. Perlman, Z. E.; Slack, M. D.; Feng, Y.; et al. Multidimensional Drug Profiling by Automated Microscopy. *Science* **2004**, *306*, 1194–1198.
15. Loo, L.-H.; Wu, L. F.; Altschuler, S. J. Image-Based Multivariate Profiling of Drug Responses from Single Cells. *Nat. Methods* **2007**, *4*, 445–453.
16. McLachlan, G. J. *The EM Algorithm and Extensions*; Wiley: New York, 2008.
17. Ljosa, V.; Sokolnicki, K. L.; Carpenter, A. E. Annotated High-Throughput Microscopy Image Sets for Validation. *Nat. Methods* **2012**, *9*, 637.
18. Slack, M. D.; Martinez, E. D.; Wu, L. F.; et al. Characterizing Heterogeneous Cellular Responses to Perturbations. *Proc. Natl. Acad. Sci. U.S.A*. **2008**, *105*, 19306–19311.
19. Young, D. W.; Bender, A.; Hoyt, J.; et al. Integrating High-Content Screening and Ligand-Target Prediction to Identify Mechanism of Action. *Nat. Chem. Biol*. **2008**, *4*, 59–68.
20. Shamir, L. Assessing the Efficacy of Low-Level Image Content Descriptors for Computer-Based Fluorescence Microscopy Image Analysis. *J. Microsc*. **2011**, *243*, 284–292.
21. Tanaka, M.; Bateman, R.; Rauh, D.; et al. An Unbiased Cell Morphology-Based Screen for New, Biologically Active Small Molecules. *PLoS Biol*. **2005**, *3*, e128.
22. Durrant, R. J.; Kabán, A. When Is "Nearest Neighbour" Meaningful: A Converse Theorem and Implications. *J. Complex*. **2009**, *25*, 385–397.
23. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.
24. Yin, Z.; Zhou, X.; Bakal, C.; et al. Using Iterative Cluster Merging with Improved Gap Statistics to Perform Online Phenotype Discovery in the Context of High-Throughput RNAi Screens. *BMC Bioinform*. **2008**, *9*, 264.
25. Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci*. **2001**, *16*, 199–231.
26. Kümmel, A.; Selzer, P.; Beibel, M.; et al. Comparison of Multivariate Data Analysis Strategies for High-Content Screening. *J. Biomol. Screen*. **2011**, *16*, 338–347.

SUPPLEMENTARY TEXTS

Text S1

Data schemas

This section and **Text S3** describe the provided data and computer programs in sufficient detail to reproduce the results of the paper and to make use of the data and programs in the development and evaluation of new methods. The data are stored in four relational tables:

- supplement_Object: one row for each of the 454,793 cells. The primary key (TableNumber, ImageNumber, ObjectNumber) identifies a cell. The pair of columns (TableNumber, ImageNumber) is a foreign key into the supplement_Image table. The remaining 476 columns are the 453 measurements as well as 23 measurements that were not used for profiling (e.g., horizontal and vertical position of the cell in the field of view).
- supplement_Image: one row for each of the 2,528 fields of view. The primary key (TableNumber, ImageNumber) identifies an image. The pair of columns (Image_Metadata_Plate_DAPI, Image_Metadata_Well_DAPI) indicates the well the image came from. The column Image_Metadata_Compound contains the name of the compound the sample was treated with; it is a foreign key into the supplement_Compound table. The column Image_Metadata_Concentration is the concentration of the compound treatment in M. Together, the pair of columns (Image_Metadata_Compound, Image_Metadata_Concentration) is a foreign key into the supplement_GroundTruth table.
- supplement_GroundTruth: one row for each of the 104 compound-concentrations in our ground-truth set. In addition to the primary key (compound, concentration), there is a column moa that contains the mechanism of action of the compound. For the mock treatment DMSO, the moa column contains "DMSO".
- supplement_Compound: one row for each of the 39 compounds in our ground-truth set. In addition to the primary key compound, there is a column smiles that contains the chemical structure of the compound. For four of the compounds proprietary to AstraZeneca, structures are not available. No structure is provided for the mock treatment DMSO.

We provide the data as tab-delimited text files, together with SQL statements and scripts for importing them into a MySQL database. Although our scripts expect to read the data from a MySQL database, the text files can also be used directly. The ZIP file that contains the text files is 754 MB. Unpacked, the text files are 2.1 GB. Imported into the database, the tables are 835 MB.

The script load_supplement_tables.sh produces a sequence of SQL statements to create the necessary tables and import the data from the text files. It takes as argument the directory that contains the text files. Run the script, then send its output to MySQL. If you prefer, you can create the tables using the .sql files in the zip file, then import the data by issuing LOAD DATA INFILE statements manually to MySQL.

Text S2

Implementations

The profiling scripts are written in Python as part of CellProfiler Analyst (CPA), and depend on the following freely-available python modules:

- numpy 1.5.1
- CellProfiler Analyst (http://github.com/CellProfiler/CellProfiler-Analyst/) rev. e33def16b2
- scipy 0.7.1
- progressbar 2.3
- sklearn (scikits.learn) 0.12
- MySQL-python 1.2.3
- MDP toolkit 3.3

The visualization scripts additionally require matplotlib 1.2. Unless otherwise noted, all dependencies are available through the Python Package Index (PyPi, http://pypi.python.org/pypi/).

Data access is directed by the CellProfiler Analyst properties file supplement.properties. You should modify this file to contain the hostname, etc., for your database server.

All script can be run with the -h option to see a comprehensive list of options and arguments.

Cache: Before you can run any profiling scripts, you must make a binary cache of the per-cell data using cpa.profiling.cache, as follows:

```
python -m cpa.profiling.cache \
  supplement.properties /path/to/cache \
  "Image_Metadata_Compound = 'DMSO'"
```

The arguments are the CPA properties file, the directory in which to store the cache, and a SQL predicate that identifies the images of mock-treated samples, which will be used as reference for normalization (see Online Methods).

Subsampling: The GM profiling method and the preprocessors construct their models using a subsample of the cells. The script subsample creates such a subsample. It takes four positional arguments (CPA properties file, cache directory, subsample output file, and the number of cells to include in the subsample). The -f option can be used to specify a CPA filter to select the cells to sample from.

Preprocessors: The scripts cpa.profiling.factor_analysis, cpa.profiling.pca, and cpa.profiling.fasel construct models that the profiling scripts can use to preprocess the image features. All three scripts take the same positional arguments:

- the filename of a subsample,
- model size (the number of factors or principal components), and
- an output file in which to store the model.

Profiling scripts: The profiling scripts are:

- cpa.profiling.profile_mean
- cpa.profiling.profile_ksstatistic
- cpa.profiling.profile_svmnormalvector
- cpa.profiling.profile_gmm
- cpa.profiling.profile_factoranalysis_mean

All take the following positional arguments:

– CPA properties file
– Cache directory
– CPA group defining sets of images to be profiled together

In addition, all take the following options:

– `--multiprocessing` (run on multiple CPUs or CPU cores)
– `--lsf-directory` *TEMPORARY-DIRECTORY* (run in parallel on a computing cluster using LSF)
– `-o` *OUTPUT-FILENAME* (file to store the profiles in)
– `-f` *FILTER-NAME* (CPA filter indicating which images to profile)

As an example, the following command profiles each well (sample) of the experiment using the "means" method:

```
python -m cpa.profiling.profile_mean \
  -o my_mean_profiles.txt -f noncontrols \
  supplement.properties /path/to/cache Well
```

The profile_ksstatistic and profile_svmnormalvector scripts expect an additional positional argument: the name of a CPA filter that defines the images that are of mock-treated samples, to serve as reference distribution for the KS statistic and negative training samples for the SVM. The profile_svmnormalvector script accepts the additional option `--rfe`, which indicates that recursive feature elimination should be used. The profile_gmm script accepts the additional option `--components` *NCOMPONENTS*, which specifies the number of mixture components. The profile_factoranalysis_mean script accepts the additional options `--factors` *NFACTORS*, which specifies the number of factors, and `--save-model` *FILENAME*, which indicates that the model should be saved to file.

Post-processing scripts: The script cpa.profiling.median_profiles groups a set of profiles according to a CPA filter given as an argument and computes the median profile in each group. The script expects three positional arguments: the properties file, the filename of the profiles to read (generated by one of the profiling scripts), and the name of the CPA filter to group by. The option `-o` *OUTPUT-FILENAME* causes the median profiles to be written to the specified file. As an example, the following command computes profiles for each compound-concentration as the median of the replicate samples:

```
python -m cpa.profiling.median_profiles \
  -o my_mean_profiles_per_treatment.txt \
  supplement.properties /path/to/cache \
  CompoundConcentration
```

The script cpa.profiling.leave_one_out performs nearest-neighbor classification with leave-one-out crossvalidation. It expects three positional arguments: the properties file, the path to the cache, and the name of a CPA filter that specifies the true class of each image. The option `-h` *HOLDOUT-GROUP* specifies a CPA group that defines which profiles to hold out when classifying a profile. As an example, the following command classifies per-treatment profiles, holding out all profiles treated with the same compound as the one being classified, even if those other profiles were treated with a different concentration, and prints the confusion matrix in sparse form to standard output:

```
python -m cpa.profiling.leave_one_out \
  -H Compound supplement.properties \
  my_mean_profiles_per_treatment.txt MOA
```

The scripts plot_profiles and plot_distances provide rudimentary visualization of profiles and the distances between them. To plot per-treatment profiles, group and label them by MOA, and output a PDF file:

```
python -m cpa.profiling.plot_profiles \
  -o profiles.pdf \
  supplement.properties \
  mean_profiles_per_treatment.txt MOA
```

To plot the distance matrix of the profiles, grouped and labeled by MOA, and output a PNG file:

```
python -m cpa.profiling.plot_distances \
  -o distance_matrix.png \
  supplement.properties \
  mean_profiles_per_treatment.txt MOA
```

Text S3

Computer programs

The file `reproduce.zip` contains four directories: `inputs` contains source data for Figure S 9; `src` contains the source code, `properties` contains the CPA properties file, and `outputs` contains the generated data files.

The GNU makefile `src/Makefile` encodes every step to recreate the results, so typing `make` in the `src` directory will recreate all the output files.

The scripts require CellProfiler Analyst (which include the implementations of the profiling method). In addition, most scripts require matplotlib 1.1.0, and calculate_aucs.py requires xalglib 3.4.0 (http://www.alglib.net/).

The scripts in the src directory are as follows:

– mean_confusion_fa.py computes the mean confusion of the 20 factor-analysis models.
– mean_confusion_gmm.py computes the mean confusion of the 20 Gaussian mixture models.
– pretty-confusion-matrix.py renders a confusion matrix as a PDF file.
– gmm_varycomponents_plot.py plots Figure S 5).
– fa_varyfactors_plot.py plots Figure 2.
– loo_confusion.py processes the data from Loo et al.[15] for use in Figure S 9.
– make_features_table.py generates Table S 1.
– make_ground_truth_table.py generates Table S 2.
– make_compounds_table.py generates Table S 3.
– make_aucs_pvalues_table.py generate Table S 10.
– calculate_aucs.py calculates AUCs and p-values for Table S 10.
– misclassified.py lists compound-concentrations that were misclassified.

Text S4

Classification bias from batch and plate effects

The samples were prepared and imaged in 10 batches. If the compounds with one mechanism of action were entirely in a different batch than other mechanisms, there would be a concern that they may be discernible due to differences between batches rather than biological differences[20]. In this experiment, however, all MOAs except two are split between two or more batches (Table S 9). The exceptions are cholesterol-lowering drugs and kinase inhibitors, which are represented only in one batch each, but because other mechanisms are also represented in these batches, there is no reason to believe that the classification is biased by batch effects.

Text S5

Previous comparisons by classification accuracy

Loo et al.[15] evaluated their method, which includes the SVM normal-vector method at its core, applying four label sets to a compound set different from ours. Based on the nearest-neighbor search results in their Supplementary Data 2, we calculated classification accuracies for their experiments, and found them to be in the range of 32–51%, depending on the marker set (Figure S 9). If the best marker set is chosen for each mechanistic class, the accuracy increases to 68%.

Kümmel et al.[26] assessed several dimensionality-reduction methods and well-summary approaches in a screening context: the methods were evaluated by their ability to distinguish cells treated with each of five compounds from mock-treated cells. Although the study did not measure ability to distinguish compounds from each other, it is interesting to note that the KS statistic was found to perform similarly to a simple median method. Factor analysis led to a decrease in accuracy, although they only measured 29 features of each cell, so compared to our experiment it is more likely for useful variation to be restricted to a single measurement, and therefore to be discarded as noise.

Text S6

Comparison by p-value of AUC

The papers that presented the KS, SVM, and GMM methods include comparisons to previous methods. Here, we describe the method of comparison and explain why we compare the methods using a different method. For the sake of completeness, we summarize the previously published comparison results and apply the same method to our dataset, using our implementations of the core parts of the profiling methods.

Method of comparison: In all three papers the performance of a profiling method is quantified in a screening context, using the distances between compounds as follows. If the profiling method worked as desired, the distances between compounds with the same MOA should be smaller than the distances between compounds with different MOA. Therefore, two sets of distances are calculated: pair-wise distance between members of a mechanistic group (the intra-set distance) and pair-wise distances between members of different mechanistic groups (the inter-set distances). To test whether the former are significantly larger than the latter, a Mann–Whitney U test is performed (Ott, R.L., Longecker, L. (2001) *An Introduction to Statistical Methods and Data Analysis*, 5th ed., ISBN 978-0-534-25122-2).

The Mann–Whitney U test assumes that the data are independent and drawn from distributions with identical shape. (The distributions can differ in position.) This assumption is problematic because it is unlikely that the intra-set and inter-set distance will have distributions of similar shape: the intra-set distances most likely form a narrow distribution, whereas the inter-set distances will be more dispersed because some mechanisms are more different than others. Say, for instance, that of the twelve MOAs, ten are very distinct and two are very similar according to the calculated profiles. The p-values for the two similar MOAs calculated using the Mann-Whitney U test are still very small, as the medians of the intra-set and inter-set distances are so different. It is, however, impossible to definitively classify a sample in either one of the two similar MOAs.

We chose instead to compare the methods by classification accuracy because that measure is more relevant for profiling experiments, which aim to distinguish multiple phenotypes. If two MOAs appear identical based on their profiles, their prediction accuracies will be around 50%.

Summary of previous comparison results: Perlman et al.[14] compare the Translation Invariant Similarity Score (TISS) method, which includes the KS method as its core, to the means method. The comparison takes only the total-intensity measurements into account, not the other measurements. The p-values are listed for nine mechanisms of action. The KS method has markedly lower p-values for five of the nine MOAs, and the mean method has markedly lower p-values for three of the nine MOAs.

Loo et al.[15] and Slack et al.[18] compare the SVM and GMM methods, respectively, to TISS. The SVM method using each of four markers sets separately is compared to TISS using all markers sets simultaneously. The SVM method appears to work better overall based on the p-values, but the combining of the marker sets for TISS makes a conclusive comparison impossible. The GMM method is compared to TISS using the same images but different features sets. This makes it difficult to tell if a change in p-value is caused by a different method or a different feature set. The GMM method has significantly lower p-values than TISS for three out of nine MOA; the inverse is true for one MOA. The remaining five MOAs have comparable p-values.

Comparison of core methods by p-value: We computed areas under the ROC for each of the five profiling methods (Table S 10). For factor analysis, the p-values are 0.0001 or below for every MOA. The other four methods yield slightly higher p-values for compounds that cause protein degradation. The means method yields higher p-values for actin disruptors. The GM method yields higher p-values for actin disruptors and cholesterol-lowering compounds.

Text S7

Most influential variables

In order to measure of how influential each variable is in distinguish each MOA from the rest (Table S 5), we first computed per-MOA profiles as element-wise medians of the profiles of the wells treated with compounds labeled as having that MOA. We then rank the variables, for each MOA profile, by the element-wise minimum distance to any other MOA profile.

Text S8

Number of factors

Why is 50 factors optimal in our experiment when Young et al.[19] use 6 factors? First of all, the number 6, which was suggested by the Kaiser criterion, is a somewhat arbitrary cutoff: it is simply the threshold where an additional factor would contain less information than an average feature. It is entirely possible that a few more factors would have led to more discriminative profiles, but in Young et al.'s paper, interpretability of the factors is important, so it makes sense to choose a relatively low number of factors.

Even if Young et al. were to choose their number of factors based on performance in an analytic task (such as classification), they could not have found more than 36 factors because they only measured 36 features. In constrast we measure 453 features. Although there are unquestionably redundancies in our feature set, it seems reasonable to assume that the rich feature set would allow us to detect additional effects on cells—effects that are too subtle/complex to be interpretable, but which nevertheless help in classification. It may also matter that we have an actin stain and a tubulin stain, and that we therefore are able to measure cellular morphology. (Young et al. only measured nuclear features.) A number of factors around 50 is consistent with our experience with training boosting classifiers for particular phenotypes in a screening context, where classifiers with 20 rules work well but adding up to about 50 rules helps[5].

SUPPLEMENTARY DATA

Data S1: `pipelines.zip` – CellProfiler pipelines for processing the images. There are two pipelines, one that computes illumination-correction functions per plate, and one that applies the illumination correction functions to images and analyzes them. The pipelines are provided in both human-readable and machine-readable format.

Data S2: `database.zip` – Measurements and metadata in CSV format, as well as a script for loading them into a MySQL database, as described (**Text S1**).

Data S3: `reproduce.zip` – Input data and source code for regenerating the figures, tables, and results (**Text S3**).

SUPPLEMENTARY FIGURES

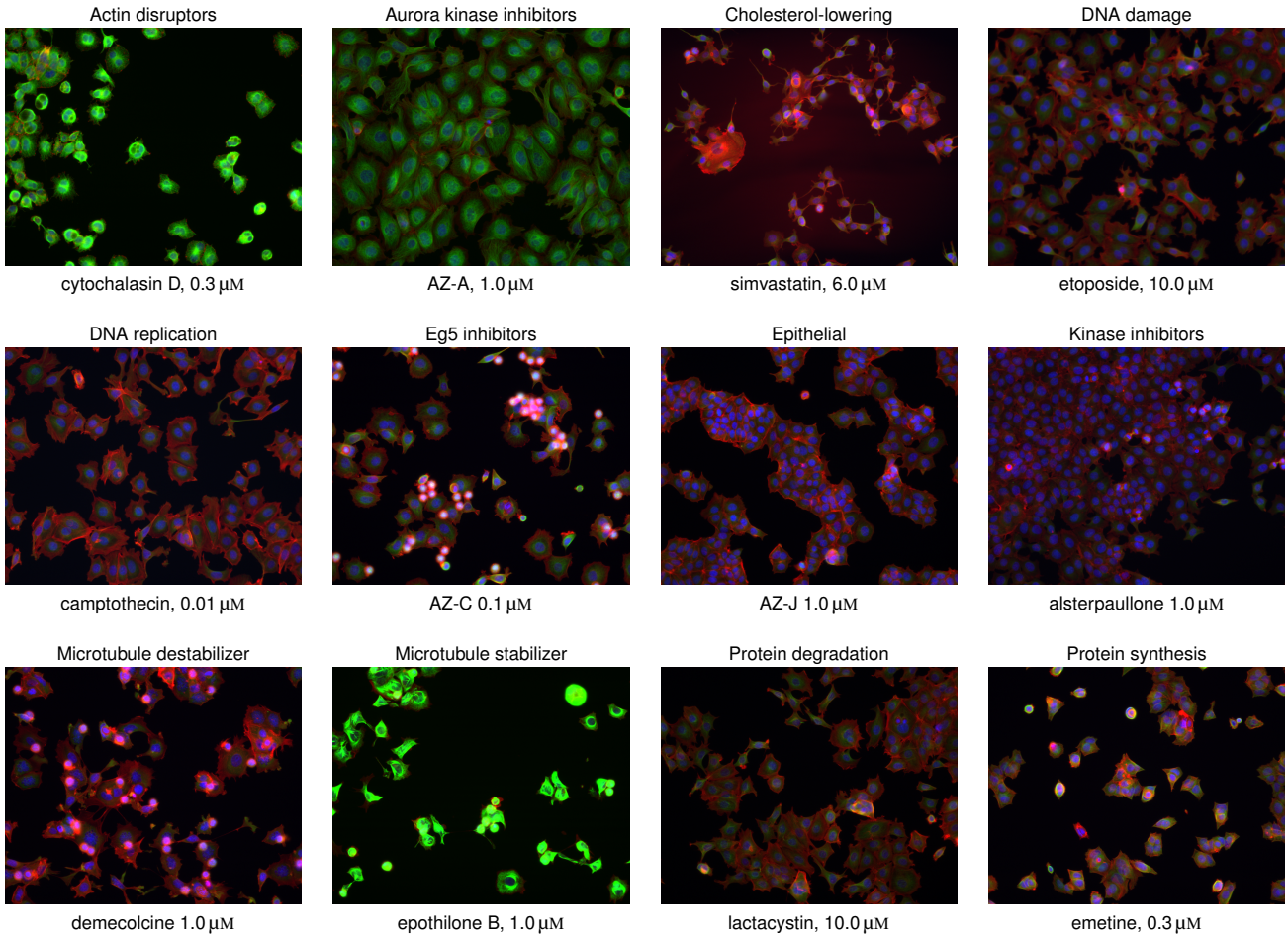


Figure S1: Images of one compound-concentration in each of the 12 MOAs spanned by our ground-truth set. Red color shows F-actin, green shows β-tubulin, and blue shows DNA.

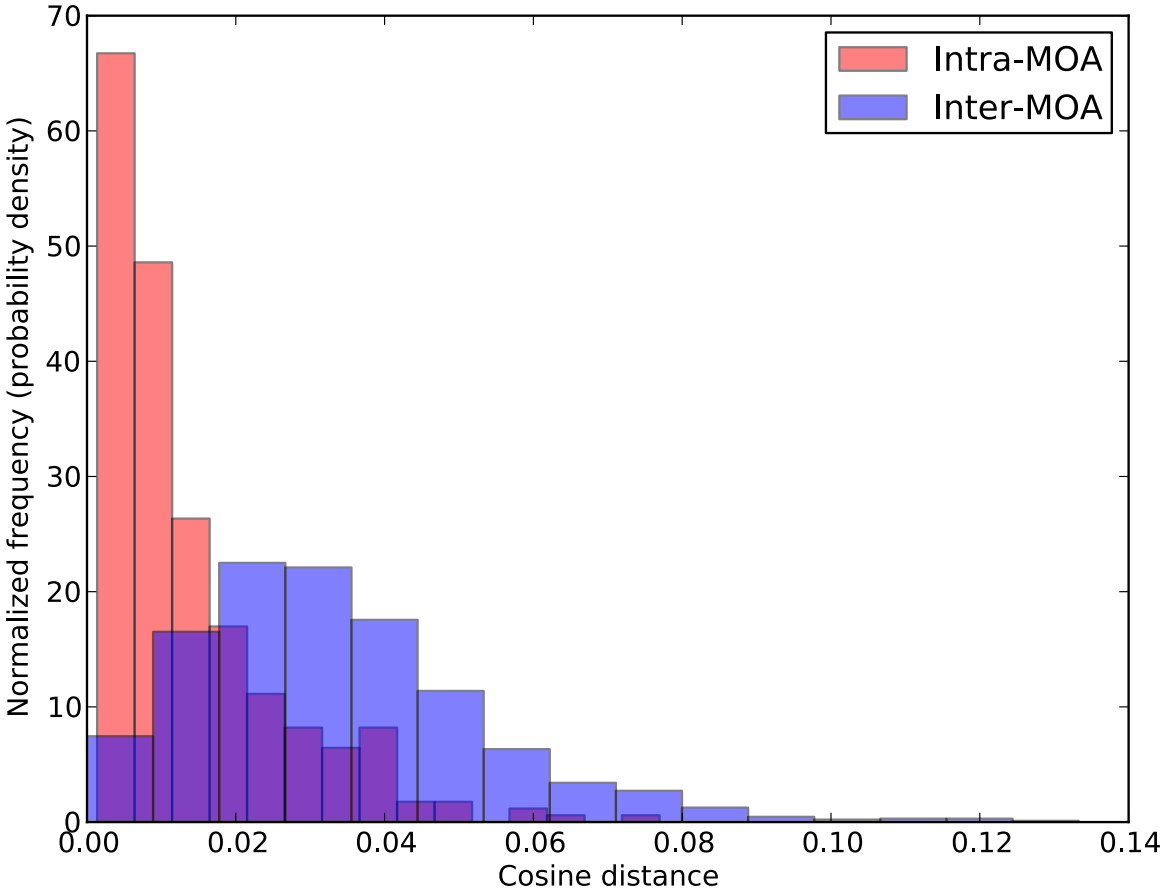


Figure S2: Histograms of distances between compound-concentrations that have the same MOA and between compound-concentrations that have different MOA. These distances are between profiles produced by the "means" method; the other methods produced histograms that were not dramatically different.

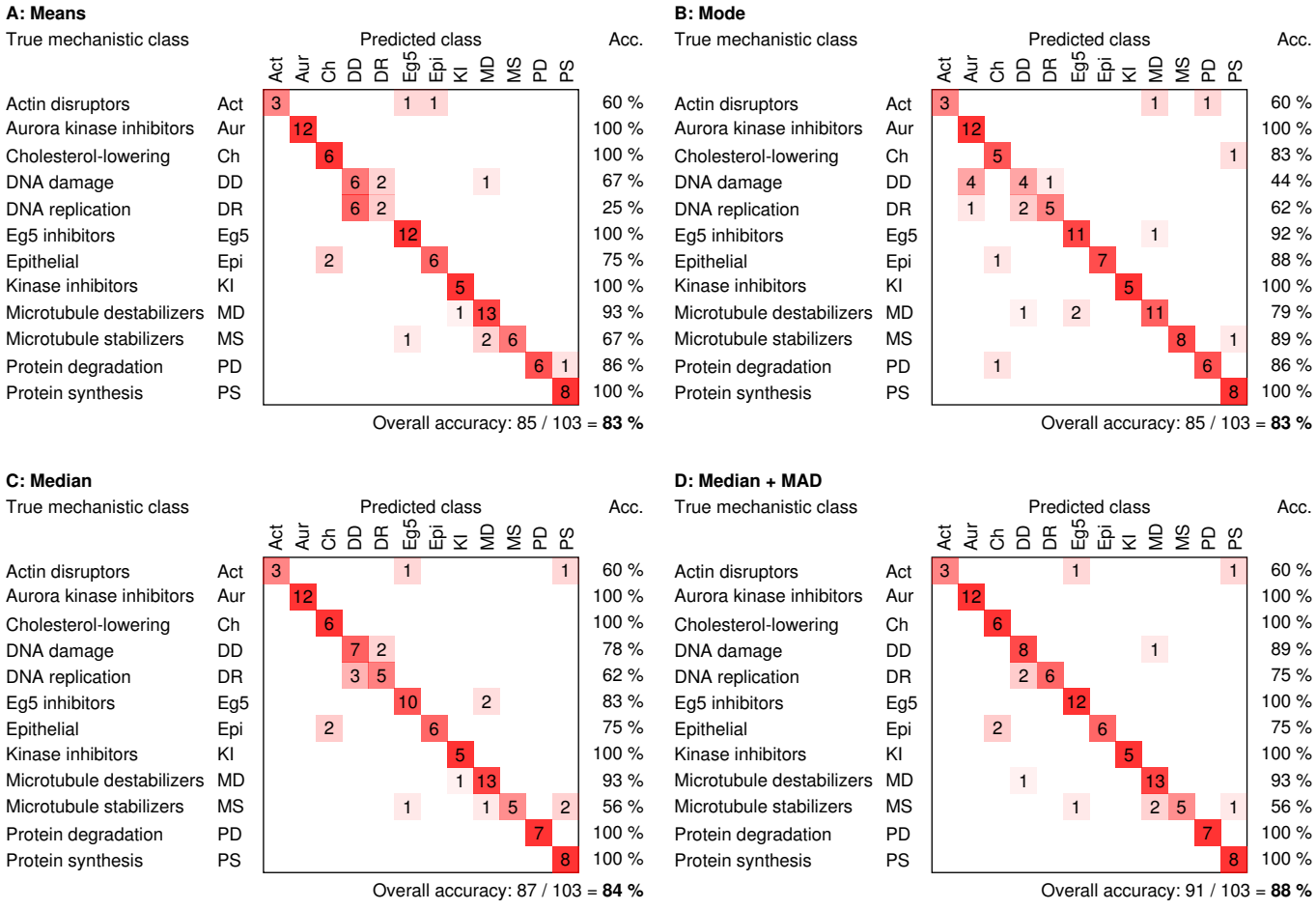


Figure S3: Confusion matrices for other population statistics.

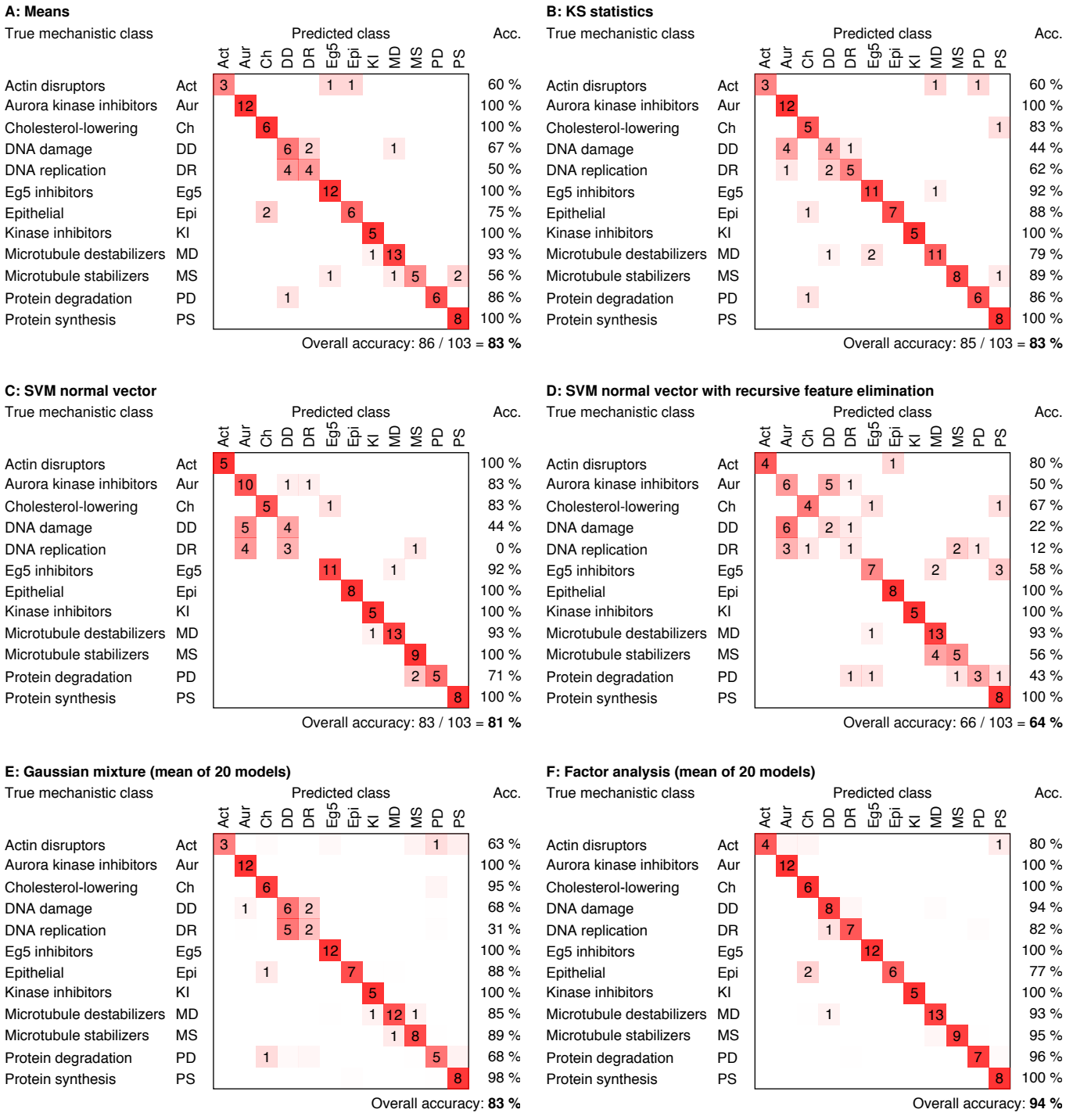


Figure S4: Confusion matrices for the five profiling methods, showing the number of compound-concentrations that were classified correctly (on the diagonals) and incorrectly (off the diagonals), the classification accuracies for each MOA (right columns), and overall classification accuracy (number of correctly classified compound-concentration divided by the total number of compound-concentrations). Panels E and F show average outcomes over 20 models; dimly colored squares without numbers indicate classification outcomes that occurred less than 0.5 times on average.

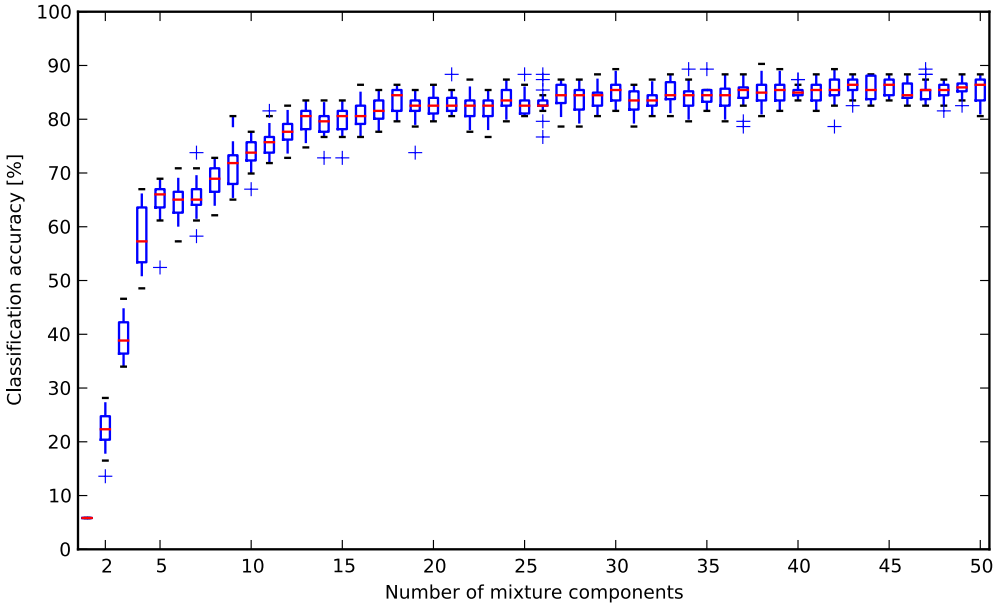


Figure S5: Distributions of classification accuracies for 20 runs of the Gaussian mixture method for each possible choice of the number of mixture components from 2 to 50.

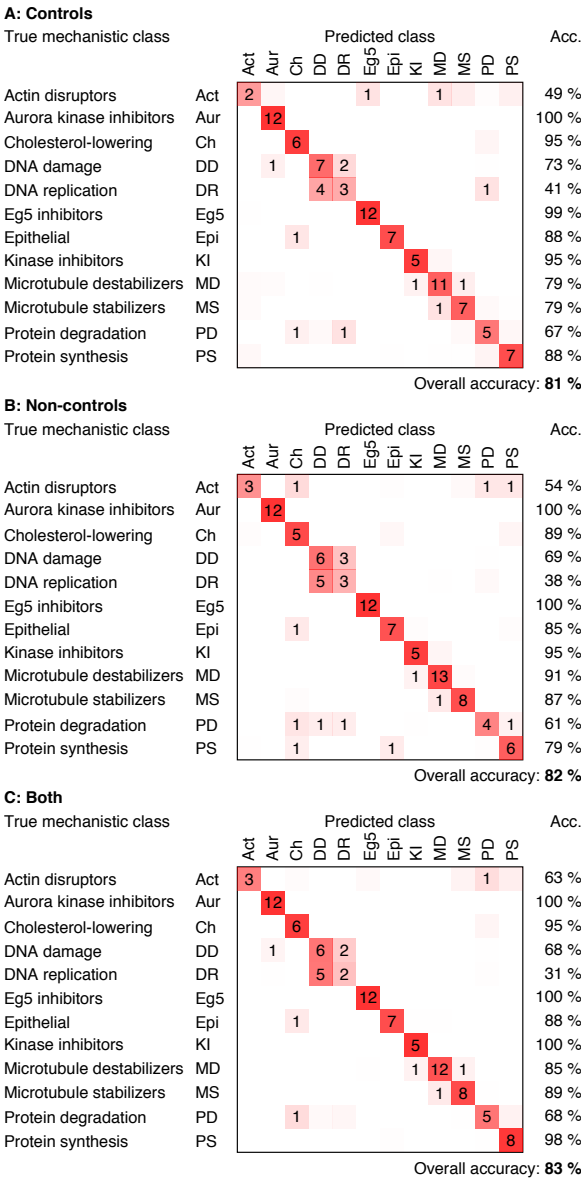


Figure S6: The GM method performs similarly whether the model is built from a subsample of control cells, a subsample of non-control (treated) cells, or a mixture of both.

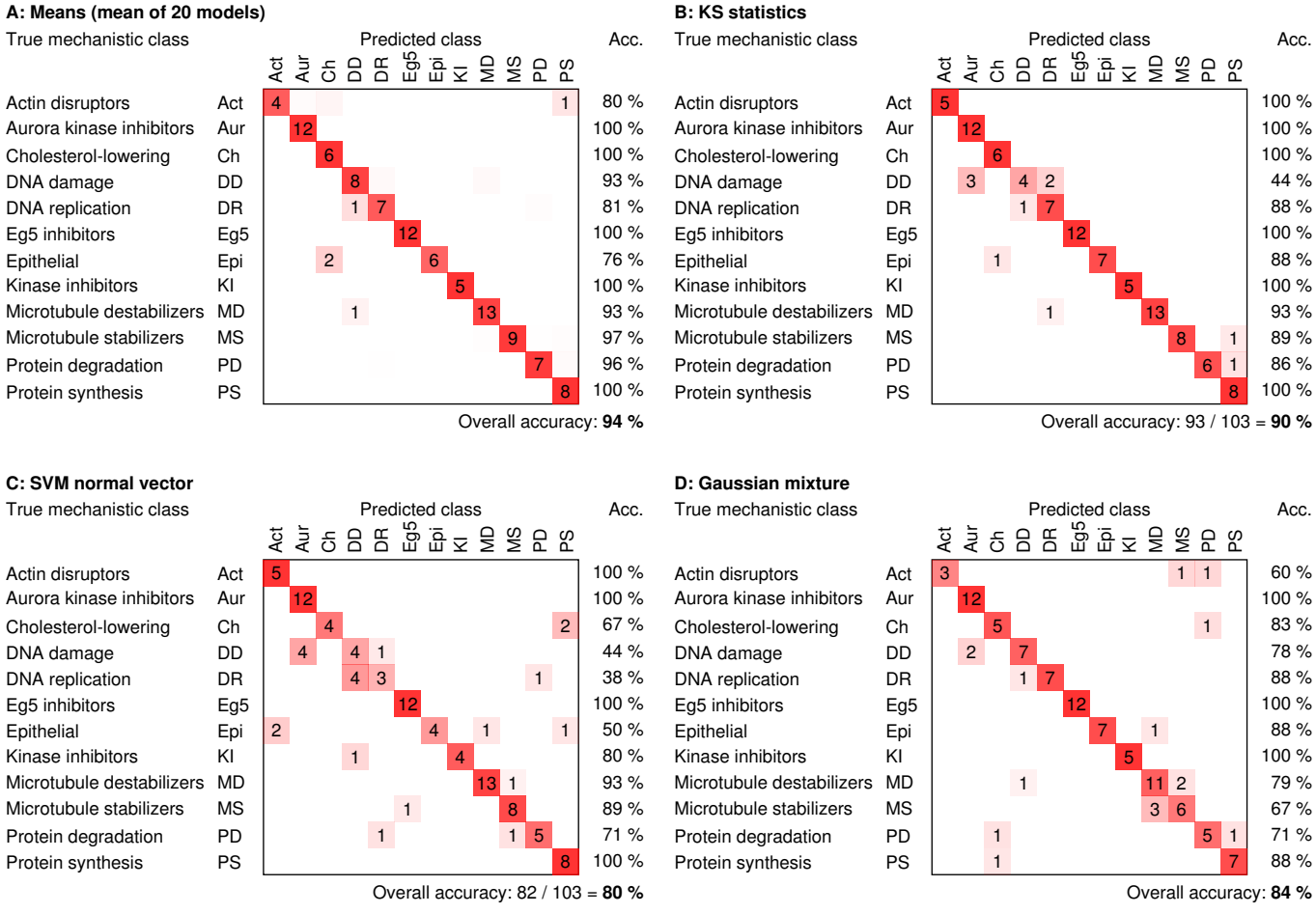


Figure S7: Confusion matrices for each method, using factor analysis as a preprocessing step.

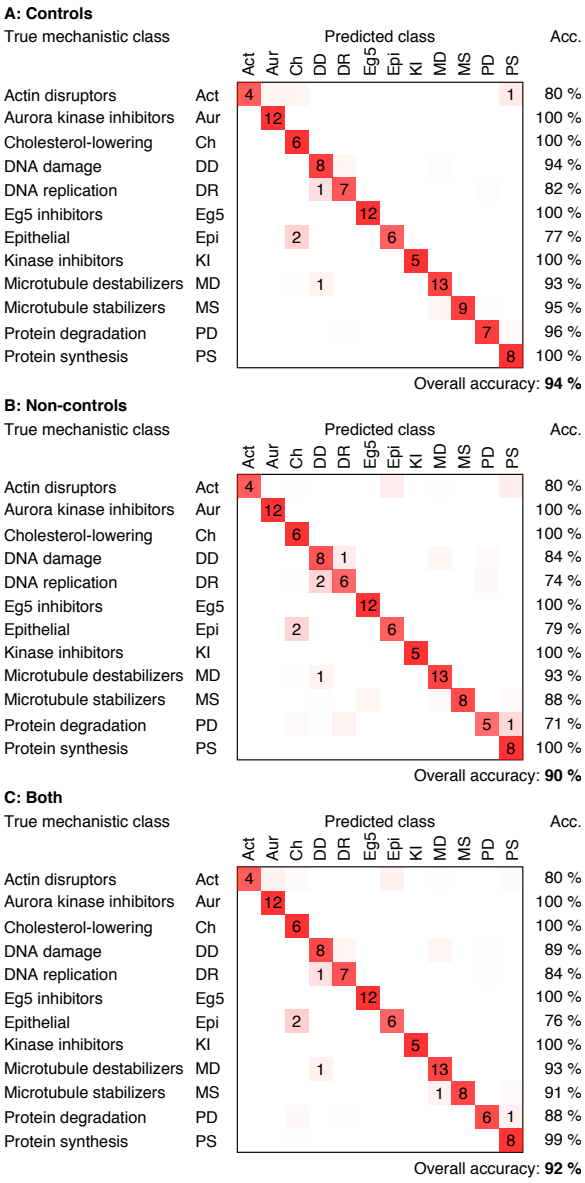


Figure S8: The FA method performs similarly whether the model is built from a subsample of control cells, a subsample of non-control (treated) cells, or a mixture of both.

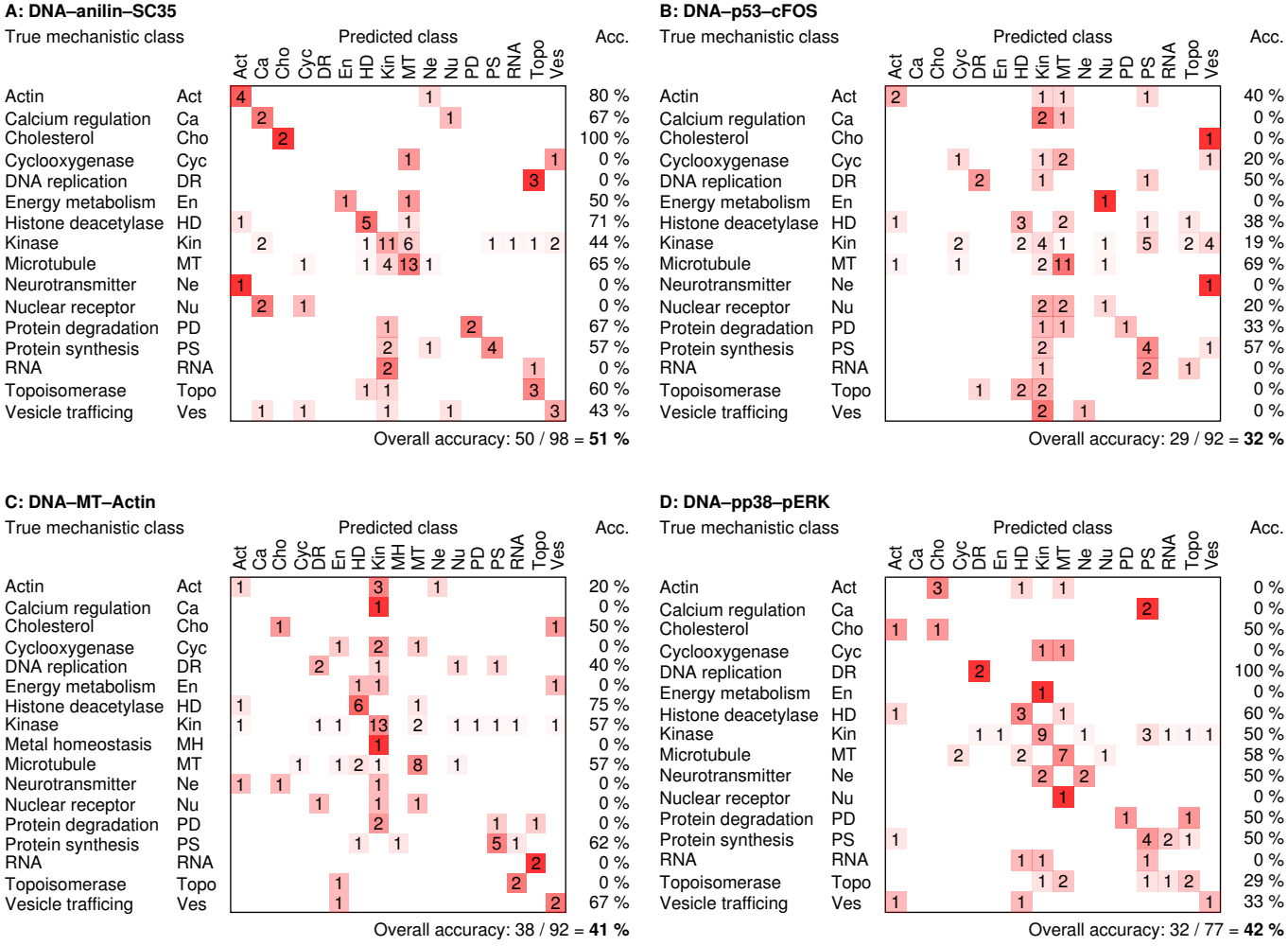


Figure S9: Confusion matrices from Loo et al.'s experiment, compiled from information in their Supplementary Data 2[15]

SUPPLEMENTARY TABLES

Table S1: The 453 measurements made by CellProfiler for each cell. (See the CellProfiler manual for descriptions of each.)

Image feature name
TableNumber
ImageNumber
ObjectNumber
Nuclei_Children_Cells_Count
Nuclei_Children_Cytoplasm_Count
Nuclei_AreaShape_Area
Nuclei_AreaShape_Eccentricity
Nuclei_AreaShape_Solidity
Nuclei_AreaShape_Extent
Nuclei_AreaShape_EulerNumber
Nuclei_AreaShape_Perimeter
Nuclei_AreaShape_FormFactor
Nuclei_AreaShape_MajorAxisLength
Nuclei_AreaShape_MinorAxisLength
Nuclei_AreaShape_Orientation
Nuclei_Zernike_0_0
Nuclei_Zernike_1_1
Nuclei_Zernike_2_0
Nuclei_Zernike_2_2
Nuclei_Zernike_3_1
Nuclei_Zernike_3_3
Nuclei_Zernike_4_0
Nuclei_Zernike_4_2
Nuclei_Zernike_4_4
Nuclei_Zernike_5_1
Nuclei_Zernike_5_3
Nuclei_Zernike_5_5
Nuclei_Zernike_6_0
Nuclei_Zernike_6_2
Nuclei_Zernike_6_4
Nuclei_Zernike_6_6
Nuclei_Zernike_7_1
Nuclei_Zernike_7_3
Nuclei_Zernike_7_5
Nuclei_Zernike_7_7
Nuclei_Zernike_8_0
Nuclei_Zernike_8_2
Nuclei_Zernike_8_4
Nuclei_Zernike_8_6
Nuclei_Zernike_8_8
Nuclei_Zernike_9_1
Nuclei_Zernike_9_3
Nuclei_Zernike_9_5
Nuclei_Zernike_9_7
Nuclei_Zernike_9_9
Nuclei_Intensity_IntegratedIntensity_CorrDAPI
Nuclei_Intensity_MeanIntensity_CorrDAPI
Nuclei_Intensity_StdIntensity_CorrDAPI
Nuclei_Intensity_MinIntensity_CorrDAPI
Nuclei_Intensity_MaxIntensity_CorrDAPI

Continues...

Table S1 continued

Image feature name
Nuclei_Intensity_IntegratedIntensityEdge_CorrDAPI
Nuclei_Intensity_MeanIntensityEdge_CorrDAPI
Nuclei_Intensity_StdIntensityEdge_CorrDAPI
Nuclei_Intensity_MinIntensityEdge_CorrDAPI
Nuclei_Intensity_MaxIntensityEdge_CorrDAPI
Nuclei_Intensity_MassDisplacement_CorrDAPI
Nuclei_Intensity_LowerQuartileIntensity_CorrDAPI
Nuclei_Intensity_MedianIntensity_CorrDAPI
Nuclei_Intensity_UpperQuartileIntensity_CorrDAPI
Nuclei_Intensity_IntegratedIntensity_CorrActin
Nuclei_Intensity_MeanIntensity_CorrActin
Nuclei_Intensity_StdIntensity_CorrActin
Nuclei_Intensity_MinIntensity_CorrActin
Nuclei_Intensity_MaxIntensity_CorrActin
Nuclei_Intensity_IntegratedIntensityEdge_CorrActin
Nuclei_Intensity_MeanIntensityEdge_CorrActin
Nuclei_Intensity_StdIntensityEdge_CorrActin
Nuclei_Intensity_MinIntensityEdge_CorrActin
Nuclei_Intensity_MaxIntensityEdge_CorrActin
Nuclei_Intensity_MassDisplacement_CorrActin
Nuclei_Intensity_LowerQuartileIntensity_CorrActin
Nuclei_Intensity_MedianIntensity_CorrActin
Nuclei_Intensity_UpperQuartileIntensity_CorrActin
Nuclei_Intensity_IntegratedIntensity_CorrTub
Nuclei_Intensity_MeanIntensity_CorrTub
Nuclei_Intensity_StdIntensity_CorrTub
Nuclei_Intensity_MinIntensity_CorrTub
Nuclei_Intensity_MaxIntensity_CorrTub
Nuclei_Intensity_IntegratedIntensityEdge_CorrTub
Nuclei_Intensity_MeanIntensityEdge_CorrTub
Nuclei_Intensity_StdIntensityEdge_CorrTub
Nuclei_Intensity_MinIntensityEdge_CorrTub
Nuclei_Intensity_MaxIntensityEdge_CorrTub
Nuclei_Intensity_MassDisplacement_CorrTub
Nuclei_Intensity_LowerQuartileIntensity_CorrTub
Nuclei_Intensity_MedianIntensity_CorrTub
Nuclei_Intensity_UpperQuartileIntensity_CorrTub
Nuclei_Neighbors_NumberOfNeighbors_10
Nuclei_Neighbors_PercentTouching_10
Nuclei_Neighbors_FirstClosestXVector_10
Nuclei_Neighbors_FirstClosestYVector_10
Nuclei_Neighbors_SecondClosestXVector_10
Nuclei_Neighbors_SecondClosestYVector_10
Nuclei_Neighbors_AngleBetweenNeighbors_10
Nuclei_Neighbors_NumberOfNeighbors_20
Nuclei_Neighbors_PercentTouching_20
Nuclei_Neighbors_FirstClosestXVector_20
Nuclei_Neighbors_FirstClosestYVector_20
Nuclei_Neighbors_SecondClosestXVector_20
Nuclei_Neighbors_SecondClosestYVector_20
Nuclei_Neighbors_AngleBetweenNeighbors_20
Nuclei_Texture_AngularSecondMoment_CorrDAPI_10
Nuclei_Texture_Contrast_CorrDAPI_10
Nuclei_Texture_Correlation_CorrDAPI_10

Continues...

Table S1 continued

Image feature name
Nuclei_Texture_Variance_CorrDAPI_10
Nuclei_Texture_InverseDifferenceMoment_CorrDAPI_10
Nuclei_Texture_SumAverage_CorrDAPI_10
Nuclei_Texture_SumVariance_CorrDAPI_10
Nuclei_Texture_SumEntropy_CorrDAPI_10
Nuclei_Texture_Entropy_CorrDAPI_10
Nuclei_Texture_DifferenceVariance_CorrDAPI_10
Nuclei_Texture_DifferenceEntropy_CorrDAPI_10
Nuclei_Texture_InfoMeas1_CorrDAPI_10
Nuclei_Texture_InfoMeas2_CorrDAPI_10
Nuclei_Texture_GaborX_CorrDAPI_10
Nuclei_Texture_GaborY_CorrDAPI_10
Nuclei_Texture_AngularSecondMoment_CorrActin_10
Nuclei_Texture_Contrast_CorrActin_10
Nuclei_Texture_Correlation_CorrActin_10
Nuclei_Texture_Variance_CorrActin_10
Nuclei_Texture_InverseDifferenceMoment_CorrActin_10
Nuclei_Texture_SumAverage_CorrActin_10
Nuclei_Texture_SumVariance_CorrActin_10
Nuclei_Texture_SumEntropy_CorrActin_10
Nuclei_Texture_Entropy_CorrActin_10
Nuclei_Texture_DifferenceVariance_CorrActin_10
Nuclei_Texture_DifferenceEntropy_CorrActin_10
Nuclei_Texture_InfoMeas1_CorrActin_10
Nuclei_Texture_InfoMeas2_CorrActin_10
Nuclei_Texture_GaborX_CorrActin_10
Nuclei_Texture_GaborY_CorrActin_10
Nuclei_Texture_AngularSecondMoment_CorrTub_10
Nuclei_Texture_Contrast_CorrTub_10
Nuclei_Texture_Correlation_CorrTub_10
Nuclei_Texture_Variance_CorrTub_10
Nuclei_Texture_InverseDifferenceMoment_CorrTub_10
Nuclei_Texture_SumAverage_CorrTub_10
Nuclei_Texture_SumVariance_CorrTub_10
Nuclei_Texture_SumEntropy_CorrTub_10
Nuclei_Texture_Entropy_CorrTub_10
Nuclei_Texture_DifferenceVariance_CorrTub_10
Nuclei_Texture_DifferenceEntropy_CorrTub_10
Nuclei_Texture_InfoMeas1_CorrTub_10
Nuclei_Texture_InfoMeas2_CorrTub_10
Nuclei_Texture_GaborX_CorrTub_10
Nuclei_Texture_GaborY_CorrTub_10
Nuclei_Texture_AngularSecondMoment_CorrDAPI_3
Nuclei_Texture_Contrast_CorrDAPI_3
Nuclei_Texture_Correlation_CorrDAPI_3
Nuclei_Texture_Variance_CorrDAPI_3
Nuclei_Texture_InverseDifferenceMoment_CorrDAPI_3
Nuclei_Texture_SumAverage_CorrDAPI_3
Nuclei_Texture_SumVariance_CorrDAPI_3
Nuclei_Texture_SumEntropy_CorrDAPI_3
Nuclei_Texture_Entropy_CorrDAPI_3
Nuclei_Texture_DifferenceVariance_CorrDAPI_3
Nuclei_Texture_DifferenceEntropy_CorrDAPI_3
Nuclei_Texture_InfoMeas1_CorrDAPI_3

Continues...

Table S1 continued

Image feature name
Nuclei_Texture_InfoMeas2_CorrDAPI_3
Nuclei_Texture_GaborX_CorrDAPI_3
Nuclei_Texture_GaborY_CorrDAPI_3
Nuclei_Texture_AngularSecondMoment_CorrActin_3
Nuclei_Texture_Contrast_CorrActin_3
Nuclei_Texture_Correlation_CorrActin_3
Nuclei_Texture_Variance_CorrActin_3
Nuclei_Texture_InverseDifferenceMoment_CorrActin_3
Nuclei_Texture_SumAverage_CorrActin_3
Nuclei_Texture_SumVariance_CorrActin_3
Nuclei_Texture_SumEntropy_CorrActin_3
Nuclei_Texture_Entropy_CorrActin_3
Nuclei_Texture_DifferenceVariance_CorrActin_3
Nuclei_Texture_DifferenceEntropy_CorrActin_3
Nuclei_Texture_InfoMeas1_CorrActin_3
Nuclei_Texture_InfoMeas2_CorrActin_3
Nuclei_Texture_GaborX_CorrActin_3
Nuclei_Texture_GaborY_CorrActin_3
Nuclei_Texture_AngularSecondMoment_CorrTub_3
Nuclei_Texture_Contrast_CorrTub_3
Nuclei_Texture_Correlation_CorrTub_3
Nuclei_Texture_Variance_CorrTub_3
Nuclei_Texture_InverseDifferenceMoment_CorrTub_3
Nuclei_Texture_SumAverage_CorrTub_3
Nuclei_Texture_SumVariance_CorrTub_3
Nuclei_Texture_SumEntropy_CorrTub_3
Nuclei_Texture_Entropy_CorrTub_3
Nuclei_Texture_DifferenceVariance_CorrTub_3
Nuclei_Texture_DifferenceEntropy_CorrTub_3
Nuclei_Texture_InfoMeas1_CorrTub_3
Nuclei_Texture_InfoMeas2_CorrTub_3
Nuclei_Texture_GaborX_CorrTub_3
Nuclei_Texture_GaborY_CorrTub_3
Cells_Parent_Nuclei
Cells_Children_Cytoplasm_Count
Cells_AreaShape_Area
Cells_AreaShape_Eccentricity
Cells_AreaShape_Solidity
Cells_AreaShape_Extent
Cells_AreaShape_EulerNumber
Cells_AreaShape_Perimeter
Cells_AreaShape_FormFactor
Cells_AreaShape_MajorAxisLength
Cells_AreaShape_MinorAxisLength
Cells_AreaShape_Orientation
Cells_Zernike_0_0
Cells_Zernike_1_1
Cells_Zernike_2_0
Cells_Zernike_2_2
Cells_Zernike_3_1
Cells_Zernike_3_3
Cells_Zernike_4_0
Cells_Zernike_4_2
Cells_Zernike_4_4

Continues...

Table S1 continued

Image feature name
Cells_Zernike_5_1
Cells_Zernike_5_3
Cells_Zernike_5_5
Cells_Zernike_6_0
Cells_Zernike_6_2
Cells_Zernike_6_4
Cells_Zernike_6_6
Cells_Zernike_7_1
Cells_Zernike_7_3
Cells_Zernike_7_5
Cells_Zernike_7_7
Cells_Zernike_8_0
Cells_Zernike_8_2
Cells_Zernike_8_4
Cells_Zernike_8_6
Cells_Zernike_8_8
Cells_Zernike_9_1
Cells_Zernike_9_3
Cells_Zernike_9_5
Cells_Zernike_9_7
Cells_Zernike_9_9
Cells_Intensity_IntegratedIntensity_CorrActin
Cells_Intensity_MeanIntensity_CorrActin
Cells_Intensity_StdIntensity_CorrActin
Cells_Intensity_MinIntensity_CorrActin
Cells_Intensity_MaxIntensity_CorrActin
Cells_Intensity_IntegratedIntensityEdge_CorrActin
Cells_Intensity_MeanIntensityEdge_CorrActin
Cells_Intensity_StdIntensityEdge_CorrActin
Cells_Intensity_MinIntensityEdge_CorrActin
Cells_Intensity_MaxIntensityEdge_CorrActin
Cells_Intensity_MassDisplacement_CorrActin
Cells_Intensity_LowerQuartileIntensity_CorrActin
Cells_Intensity_MedianIntensity_CorrActin
Cells_Intensity_UpperQuartileIntensity_CorrActin
Cells_Intensity_IntegratedIntensity_CorrTub
Cells_Intensity_MeanIntensity_CorrTub
Cells_Intensity_StdIntensity_CorrTub
Cells_Intensity_MinIntensity_CorrTub
Cells_Intensity_MaxIntensity_CorrTub
Cells_Intensity_IntegratedIntensityEdge_CorrTub
Cells_Intensity_MeanIntensityEdge_CorrTub
Cells_Intensity_StdIntensityEdge_CorrTub
Cells_Intensity_MinIntensityEdge_CorrTub
Cells_Intensity_MaxIntensityEdge_CorrTub
Cells_Intensity_MassDisplacement_CorrTub
Cells_Intensity_LowerQuartileIntensity_CorrTub
Cells_Intensity_MedianIntensity_CorrTub
Cells_Intensity_UpperQuartileIntensity_CorrTub
Cells_Neighbors_NumberOfNeighbors_3
Cells_Neighbors_PercentTouching_3
Cells_Neighbors_FirstClosestXVector_3
Cells_Neighbors_FirstClosestYVector_3
Cells_Neighbors_SecondClosestXVector_3

Continues...

Table S1 continued

Image feature name
Cells_Neighbors_SecondClosestYVector_3
Cells_Neighbors_AngleBetweenNeighbors_3
Cells_Neighbors_NumberOfNeighbors_10
Cells_Neighbors_PercentTouching_10
Cells_Neighbors_FirstClosestXVector_10
Cells_Neighbors_FirstClosestYVector_10
Cells_Neighbors_SecondClosestXVector_10
Cells_Neighbors_SecondClosestYVector_10
Cells_Neighbors_AngleBetweenNeighbors_10
Cells_Texture_AngularSecondMoment_CorrActin_10
Cells_Texture_Contrast_CorrActin_10
Cells_Texture_Correlation_CorrActin_10
Cells_Texture_Variance_CorrActin_10
Cells_Texture_InverseDifferenceMoment_CorrActin_10
Cells_Texture_SumAverage_CorrActin_10
Cells_Texture_SumVariance_CorrActin_10
Cells_Texture_SumEntropy_CorrActin_10
Cells_Texture_Entropy_CorrActin_10
Cells_Texture_DifferenceVariance_CorrActin_10
Cells_Texture_DifferenceEntropy_CorrActin_10
Cells_Texture_InfoMeas1_CorrActin_10
Cells_Texture_InfoMeas2_CorrActin_10
Cells_Texture_GaborX_CorrActin_10
Cells_Texture_GaborY_CorrActin_10
Cells_Texture_AngularSecondMoment_CorrTub_10
Cells_Texture_Contrast_CorrTub_10
Cells_Texture_Correlation_CorrTub_10
Cells_Texture_Variance_CorrTub_10
Cells_Texture_InverseDifferenceMoment_CorrTub_10
Cells_Texture_SumAverage_CorrTub_10
Cells_Texture_SumVariance_CorrTub_10
Cells_Texture_SumEntropy_CorrTub_10
Cells_Texture_Entropy_CorrTub_10
Cells_Texture_DifferenceVariance_CorrTub_10
Cells_Texture_DifferenceEntropy_CorrTub_10
Cells_Texture_InfoMeas1_CorrTub_10
Cells_Texture_InfoMeas2_CorrTub_10
Cells_Texture_GaborX_CorrTub_10
Cells_Texture_GaborY_CorrTub_10
Cells_Texture_AngularSecondMoment_CorrActin_3
Cells_Texture_Contrast_CorrActin_3
Cells_Texture_Correlation_CorrActin_3
Cells_Texture_Variance_CorrActin_3
Cells_Texture_InverseDifferenceMoment_CorrActin_3
Cells_Texture_SumAverage_CorrActin_3
Cells_Texture_SumVariance_CorrActin_3
Cells_Texture_SumEntropy_CorrActin_3
Cells_Texture_Entropy_CorrActin_3
Cells_Texture_DifferenceVariance_CorrActin_3
Cells_Texture_DifferenceEntropy_CorrActin_3
Cells_Texture_InfoMeas1_CorrActin_3
Cells_Texture_InfoMeas2_CorrActin_3
Cells_Texture_GaborX_CorrActin_3
Cells_Texture_GaborY_CorrActin_3

Continues...

Table S1 continued

Image feature name
Cells_Texture_AngularSecondMoment_CorrTub_3
Cells_Texture_Contrast_CorrTub_3
Cells_Texture_Correlation_CorrTub_3
Cells_Texture_Variance_CorrTub_3
Cells_Texture_InverseDifferenceMoment_CorrTub_3
Cells_Texture_SumAverage_CorrTub_3
Cells_Texture_SumVariance_CorrTub_3
Cells_Texture_SumEntropy_CorrTub_3
Cells_Texture_Entropy_CorrTub_3
Cells_Texture_DifferenceVariance_CorrTub_3
Cells_Texture_DifferenceEntropy_CorrTub_3
Cells_Texture_InfoMeas1_CorrTub_3
Cells_Texture_InfoMeas2_CorrTub_3
Cells_Texture_GaborX_CorrTub_3
Cells_Texture_GaborY_CorrTub_3
Cytoplasm_Parent_Cells
Cytoplasm_Parent_Nuclei
Cytoplasm_AreaShape_Area
Cytoplasm_AreaShape_Eccentricity
Cytoplasm_AreaShape_Solidity
Cytoplasm_AreaShape_Extent
Cytoplasm_AreaShape_EulerNumber
Cytoplasm_AreaShape_Perimeter
Cytoplasm_AreaShape_FormFactor
Cytoplasm_AreaShape_MajorAxisLength
Cytoplasm_AreaShape_MinorAxisLength
Cytoplasm_AreaShape_Orientation
Cytoplasm_Zernike_0_0
Cytoplasm_Zernike_1_1
Cytoplasm_Zernike_2_0
Cytoplasm_Zernike_2_2
Cytoplasm_Zernike_3_1
Cytoplasm_Zernike_3_3
Cytoplasm_Zernike_4_0
Cytoplasm_Zernike_4_2
Cytoplasm_Zernike_4_4
Cytoplasm_Zernike_5_1
Cytoplasm_Zernike_5_3
Cytoplasm_Zernike_5_5
Cytoplasm_Zernike_6_0
Cytoplasm_Zernike_6_2
Cytoplasm_Zernike_6_4
Cytoplasm_Zernike_6_6
Cytoplasm_Zernike_7_1
Cytoplasm_Zernike_7_3
Cytoplasm_Zernike_7_5
Cytoplasm_Zernike_7_7
Cytoplasm_Zernike_8_0
Cytoplasm_Zernike_8_2
Cytoplasm_Zernike_8_4
Cytoplasm_Zernike_8_6
Cytoplasm_Zernike_8_8
Cytoplasm_Zernike_9_1
Cytoplasm_Zernike_9_3

Continues...

Table S1 continued

Image feature name
Cytoplasm_Zernike_9_5
Cytoplasm_Zernike_9_7
Cytoplasm_Zernike_9_9
Cytoplasm_Intensity_IntegratedIntensity_CorrActin
Cytoplasm_Intensity_MeanIntensity_CorrActin
Cytoplasm_Intensity_StdIntensity_CorrActin
Cytoplasm_Intensity_MinIntensity_CorrActin
Cytoplasm_Intensity_MaxIntensity_CorrActin
Cytoplasm_Intensity_IntegratedIntensityEdge_CorrActin
Cytoplasm_Intensity_MeanIntensityEdge_CorrActin
Cytoplasm_Intensity_StdIntensityEdge_CorrActin
Cytoplasm_Intensity_MinIntensityEdge_CorrActin
Cytoplasm_Intensity_MaxIntensityEdge_CorrActin
Cytoplasm_Intensity_MassDisplacement_CorrActin
Cytoplasm_Intensity_LowerQuartileIntensity_CorrActin
Cytoplasm_Intensity_MedianIntensity_CorrActin
Cytoplasm_Intensity_UpperQuartileIntensity_CorrActin
Cytoplasm_Intensity_IntegratedIntensity_CorrTub
Cytoplasm_Intensity_MeanIntensity_CorrTub
Cytoplasm_Intensity_StdIntensity_CorrTub
Cytoplasm_Intensity_MinIntensity_CorrTub
Cytoplasm_Intensity_MaxIntensity_CorrTub
Cytoplasm_Intensity_IntegratedIntensityEdge_CorrTub
Cytoplasm_Intensity_MeanIntensityEdge_CorrTub
Cytoplasm_Intensity_StdIntensityEdge_CorrTub
Cytoplasm_Intensity_MinIntensityEdge_CorrTub
Cytoplasm_Intensity_MaxIntensityEdge_CorrTub
Cytoplasm_Intensity_MassDisplacement_CorrTub
Cytoplasm_Intensity_LowerQuartileIntensity_CorrTub
Cytoplasm_Intensity_MedianIntensity_CorrTub
Cytoplasm_Intensity_UpperQuartileIntensity_CorrTub
Cytoplasm_Texture_AngularSecondMoment_CorrActin_10
Cytoplasm_Texture_Contrast_CorrActin_10
Cytoplasm_Texture_Correlation_CorrActin_10
Cytoplasm_Texture_Variance_CorrActin_10
Cytoplasm_Texture_InverseDifferenceMoment_CorrActin_10
Cytoplasm_Texture_SumAverage_CorrActin_10
Cytoplasm_Texture_SumVariance_CorrActin_10
Cytoplasm_Texture_SumEntropy_CorrActin_10
Cytoplasm_Texture_Entropy_CorrActin_10
Cytoplasm_Texture_DifferenceVariance_CorrActin_10
Cytoplasm_Texture_DifferenceEntropy_CorrActin_10
Cytoplasm_Texture_InfoMeas1_CorrActin_10
Cytoplasm_Texture_InfoMeas2_CorrActin_10
Cytoplasm_Texture_GaborX_CorrActin_10
Cytoplasm_Texture_GaborY_CorrActin_10
Cytoplasm_Texture_AngularSecondMoment_CorrTub_10
Cytoplasm_Texture_Contrast_CorrTub_10
Cytoplasm_Texture_Correlation_CorrTub_10
Cytoplasm_Texture_Variance_CorrTub_10
Cytoplasm_Texture_InverseDifferenceMoment_CorrTub_10
Cytoplasm_Texture_SumAverage_CorrTub_10
Cytoplasm_Texture_SumVariance_CorrTub_10
Cytoplasm_Texture_SumEntropy_CorrTub_10

Continues...

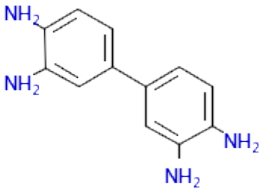
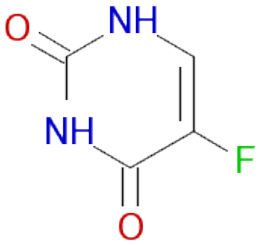
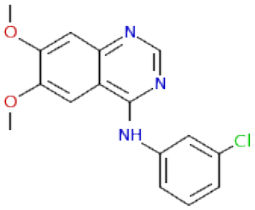
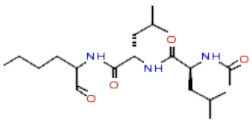
Table S1 continued

Image feature name
Cytoplasm_Texture_Entropy_CorrTub_10
Cytoplasm_Texture_DifferenceVariance_CorrTub_10
Cytoplasm_Texture_DifferenceEntropy_CorrTub_10
Cytoplasm_Texture_InfoMeas1_CorrTub_10
Cytoplasm_Texture_InfoMeas2_CorrTub_10
Cytoplasm_Texture_GaborX_CorrTub_10
Cytoplasm_Texture_GaborY_CorrTub_10
Cytoplasm_Texture_AngularSecondMoment_CorrActin_3
Cytoplasm_Texture_Contrast_CorrActin_3
Cytoplasm_Texture_Correlation_CorrActin_3
Cytoplasm_Texture_Variance_CorrActin_3
Cytoplasm_Texture_InverseDifferenceMoment_CorrActin_3
Cytoplasm_Texture_SumAverage_CorrActin_3
Cytoplasm_Texture_SumVariance_CorrActin_3
Cytoplasm_Texture_SumEntropy_CorrActin_3
Cytoplasm_Texture_Entropy_CorrActin_3
Cytoplasm_Texture_DifferenceVariance_CorrActin_3
Cytoplasm_Texture_DifferenceEntropy_CorrActin_3
Cytoplasm_Texture_InfoMeas1_CorrActin_3
Cytoplasm_Texture_InfoMeas2_CorrActin_3
Cytoplasm_Texture_GaborX_CorrActin_3
Cytoplasm_Texture_GaborY_CorrActin_3
Cytoplasm_Texture_AngularSecondMoment_CorrTub_3
Cytoplasm_Texture_Contrast_CorrTub_3
Cytoplasm_Texture_Correlation_CorrTub_3
Cytoplasm_Texture_Variance_CorrTub_3
Cytoplasm_Texture_InverseDifferenceMoment_CorrTub_3
Cytoplasm_Texture_SumAverage_CorrTub_3
Cytoplasm_Texture_SumVariance_CorrTub_3
Cytoplasm_Texture_SumEntropy_CorrTub_3
Cytoplasm_Texture_Entropy_CorrTub_3
Cytoplasm_Texture_DifferenceVariance_CorrTub_3
Cytoplasm_Texture_DifferenceEntropy_CorrTub_3
Cytoplasm_Texture_InfoMeas1_CorrTub_3
Cytoplasm_Texture_InfoMeas2_CorrTub_3
Cytoplasm_Texture_GaborX_CorrTub_3
Cytoplasm_Texture_GaborY_CorrTub_3

Table S2: The ground-truth set used for comparing profiling algorithms

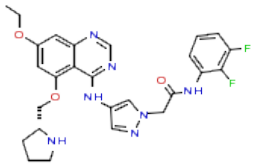
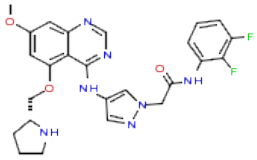
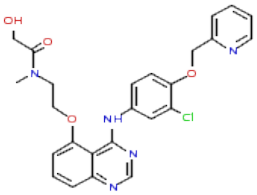
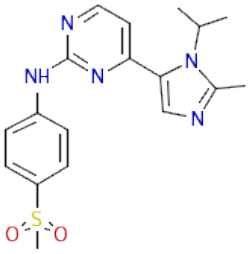
Mechanism of action	Compound	Concentrations [μM]
Actin disruptors	cytochalasin B	10.0, 30.0
	cytochalasin D	0.3
	latrunculin B	1.0, 3.0
Aurora kinase inhibitors	AZ-A	0.1, 0.3, 1.0, 3.0, 10.0, 30.0
	AZ258	0.1, 0.3, 1.0
	AZ841	0.1, 0.3, 1.0
Cholesterol-lowering	mevinolin/lovastatin	1.5, 5.0, 15.0
	simvastatin	2.0, 6.0, 20.0
DNA damage	chlorambucil	10.0
	cisplatin	10.0
	etoposide	1.0, 3.0, 10.0
	mitomycin C	0.1, 0.3, 1.0, 3.0
DNA replication	camptothecin	0.003, 0.01, 0.03
	floxuridine	10.0, 30.0
	methotrexate	10.0
	mitoxantrone	0.003, 0.01
Eg5 inhibitors	AZ-C	0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0
	AZ138	0.03, 0.1, 0.3, 1.0, 3.0
Epithelial	AZ-J	1.0, 3.0, 10.0
	AZ-U	1.0, 3.0, 10.0
	PP-2	3.0, 10.0
Kinase inhibitors	PD-169316	3.0, 10.0
	alsterpaullone	1.0, 3.0
	bryostatin	0.3
Microtubule destabilizers	colchicine	0.03
	demecolcine	0.3, 1.0, 3.0, 10.0
	nocodazole	1.0, 3.0
	vincristine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0
Microtubule stabilizers	docetaxel	0.03, 0.1, 0.3
	epothilone B	0.1, 0.3, 1.0
	taxol	0.3, 1.0, 3.0
Protein degradation	ALLN	3.0, 100.0
	MG-132	0.1, 3.0
	lactacystin	10.0
	proteasome inhibitor I	0.1, 3.0
Protein synthesis	anisomycin	0.3, 1.0
	cyclohexamide	5.0, 15.0, 50.0
	emetine	0.1, 0.3, 1.0

Table S3: The compounds and concentrations with which cells were treated

Compound name	Concentrations [μM]	Structure
3,3'-diaminobenzidine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
5-fluorouracil	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
AG-1478	0.003, 0.006, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6, 1.0, 2.0, 3.0, 6.0, 10.0, 20.0	
ALLN	0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 10000.0	 Chiral
AZ-A	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-B	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-C	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0	(*not disclosed*)

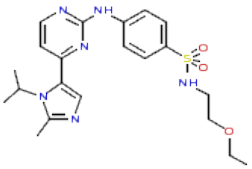
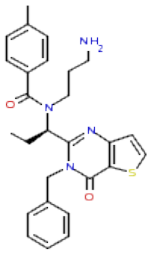
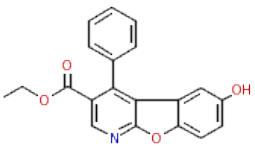
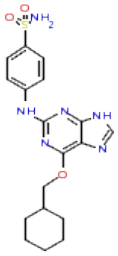
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
AZ-D (AZ841)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	Chiral 
AZ-E (AZ258)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	Chiral 
AZ-F (AZ701)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
AZ-G (AZ235)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
AZ-H	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-I	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-J	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	(*not disclosed*)
AZ-K	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	(*not disclosed*)
AZ-L	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-M	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)

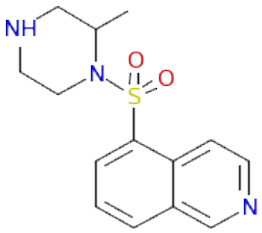
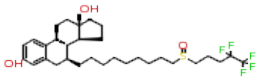
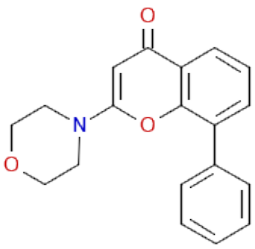
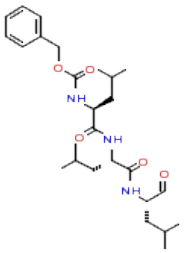
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
AZ-N	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	(*not disclosed*)
AZ-O	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
AZ-P (AZ970)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
AZ-Q (AZ138)	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	Chiral 
AZ-U	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
Cdk1 inhibitor III	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
Cdk1/2 inhibitor (NU6102)	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	

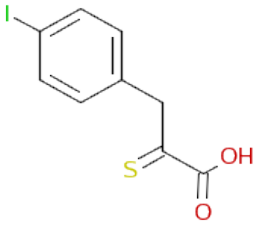
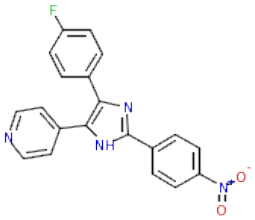
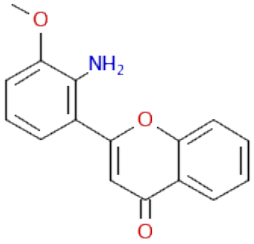
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
H-7	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	 Chiral
ICI-182,780	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
LY-294002	0.0006, 0.002, 0.006, 0.02, 0.06, 0.2, 0.6, 2.0	
MG-132	0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0	 Chiral

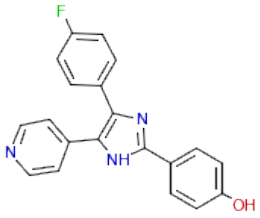
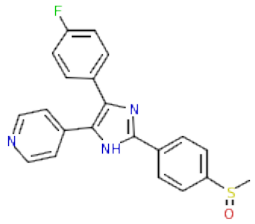
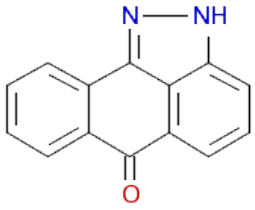
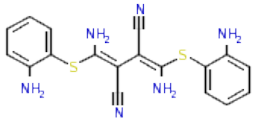
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
PD-150606	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	
PD-169316	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
PD-98059	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
PP-2	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	

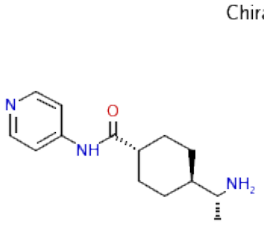
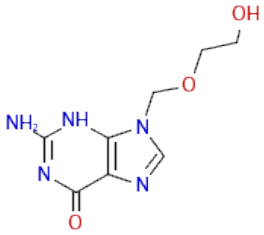
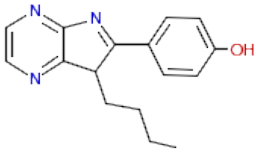
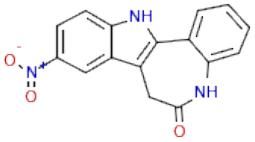
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
SB-202190	0.006, 0.02, 0.06, 0.2, 0.6, 2.0, 6.0, 20.0	
SB-203580	0.003, 0.006, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6, 1.0, 2.0, 3.0, 6.0, 10.0, 20.0	
SP-600125	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0, 10000.0	
TKK	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	(*not disclosed*)
UNKNOWN	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	(*not disclosed*)
UO-126	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	

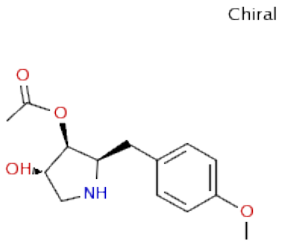
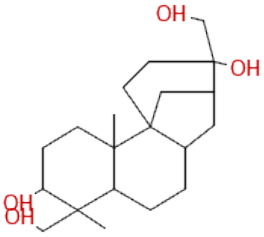
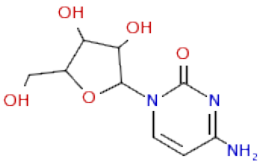
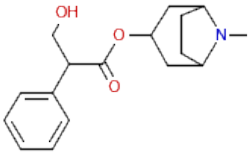
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
Y-27632	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
acyclovir	0.0015, 0.005, 0.015, 0.05, 0.15, 0.5, 1.5, 5.0	
adenine arabinofuranoside	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	(*not disclosed*)
aloisine A	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
alsterpaullone	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	

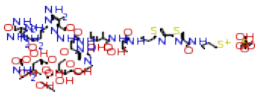
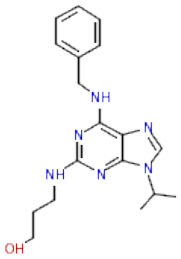
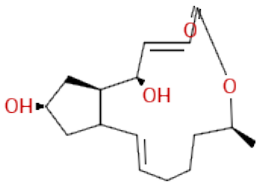
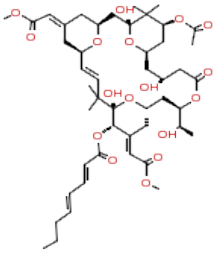
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
anisomycin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
aphidicolin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
arabinofuranosylcytosine	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
atropine	0.0015, 0.005, 0.015, 0.05, 0.15, 0.5, 1.5, 5.0	

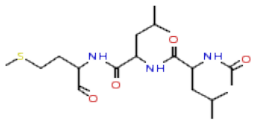
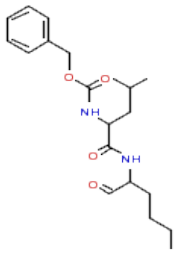
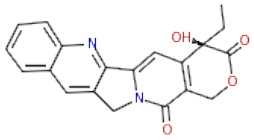
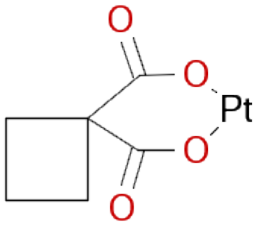
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
bleomycin	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0, 10000.0	Chiral 
bohemine	0.003, 0.006, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6, 1.0, 2.0, 3.0, 6.0, 10.0, 20.0	
brefeldin A	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	Chiral 
bryostatin	6e-05, 0.0002, 0.0006, 0.002, 0.003, 0.006, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 1.0, 3.0, 10.0	Chiral 

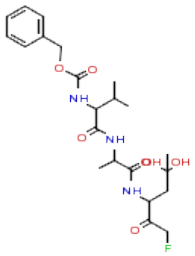
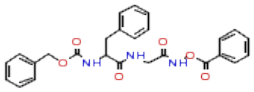
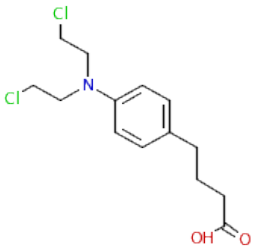
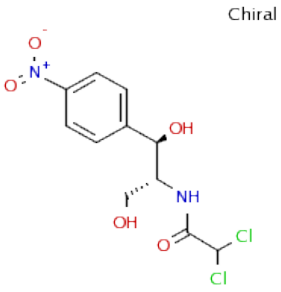
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
calpain inhibitor 2 (ALLM)	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	
calpeptin	0.006, 0.015, 0.02, 0.05, 0.06, 0.15, 0.2, 0.5, 0.6, 1.5, 2.0, 5.0, 6.0, 15.0, 20.0, 50.0	 Chiral
camptothecin	0.0015, 0.003, 0.005, 0.01, 0.015, 0.03, 0.05, 0.1, 0.15, 0.3, 0.5, 1.0, 1.5, 3.0, 5.0, 10.0	
carboplatin	0.006, 0.02, 0.06, 0.2, 0.6, 2.0, 6.0, 20.0	

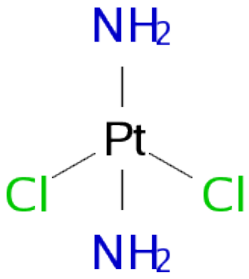
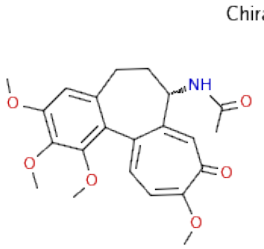
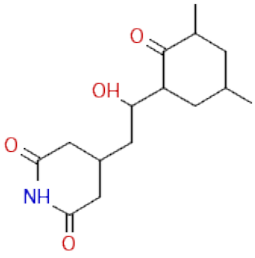
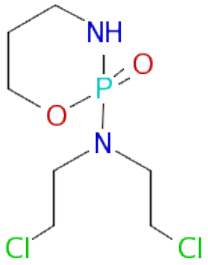
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
caspase inhibitor 1 (ZVAD)	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	
cathepsin inhibitor I	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
chlorambucil	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
chloramphenicol	0.006, 0.02, 0.06, 0.2, 0.6, 2.0, 6.0, 20.0	

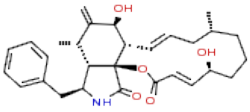
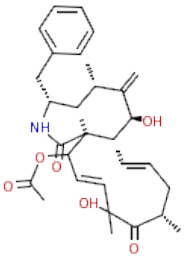
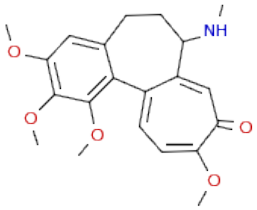
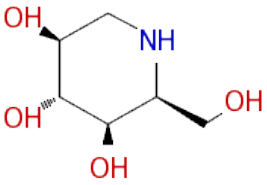
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
cisplatin	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
colchicine	1e-05, 3e-05, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0	
cyclohexamide	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	
cyclophosphamide	0.0015, 0.003, 0.005, 0.01, 0.015, 0.03, 0.05, 0.1, 0.15, 0.3, 0.5, 1.0, 1.5, 3.0, 5.0, 10.0	

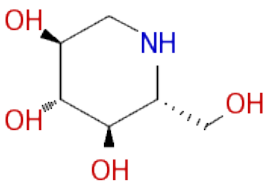
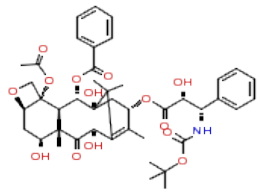
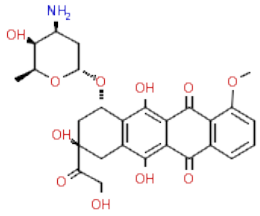
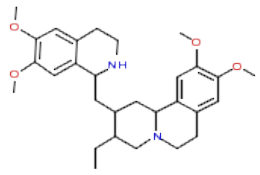
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
cytochalasin B	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	Chiral 
cytochalasin D	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 10000.0	Chiral 
demecolcine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
deoxymannojirimycin	0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 300.0, 1000.0	Chiral 

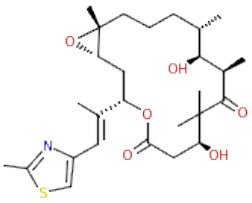
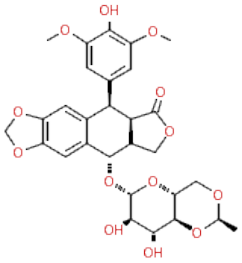
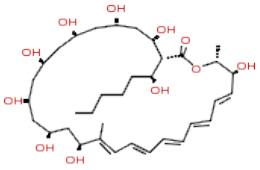
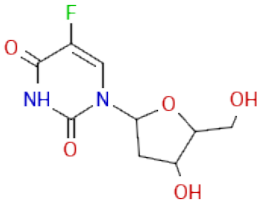
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
deoxynojirimycin	0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 300.0, 1000.0	Chiral 
docetaxel	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0	Chiral 
doxorubicin	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	Chiral 
emetine	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	

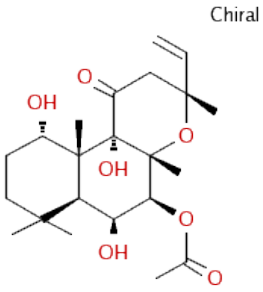
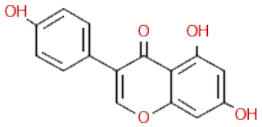
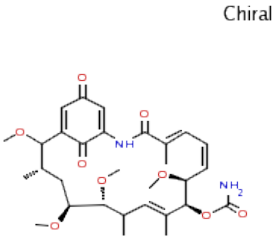
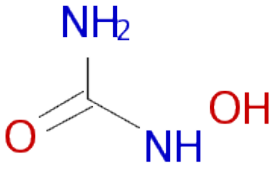
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [µM]	Structure
epothilone B	3e-05, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0	Chiral 
etoposide	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	Chiral 
filipin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	Chiral 
floxuridine	0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0	

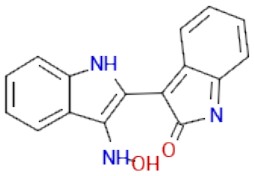
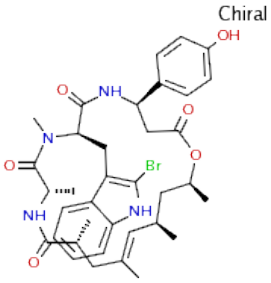
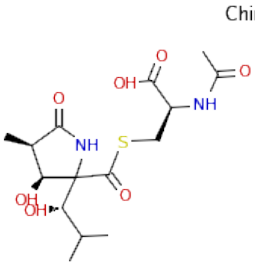
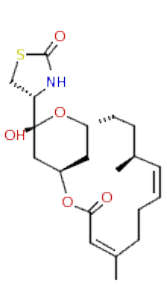
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
forskolin	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	
genistein	0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 10000.0	
herbimycin A	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
hydroxyurea	0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 300.0, 1000.0, 10000.0	

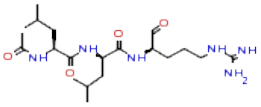
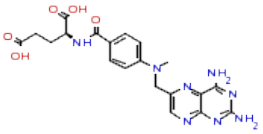
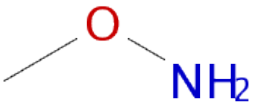
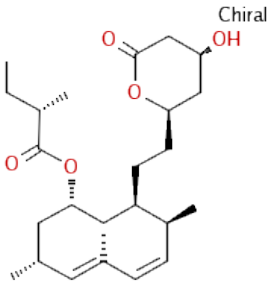
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
indirubin monoxime	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
jasplakinolide	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0	
lactacystin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 10000.0	
latrunculin B	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0	

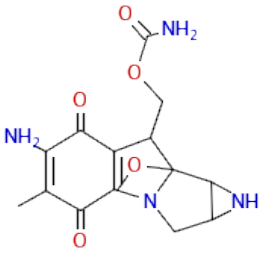
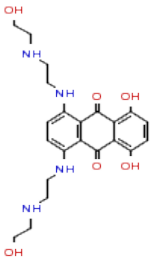
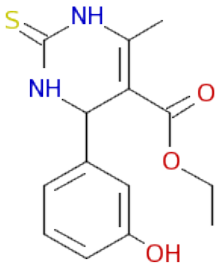
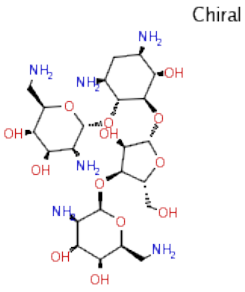
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
leupeptin	0.006, 0.015, 0.02, 0.05, 0.06, 0.15, 0.2, 0.5, 0.6, 1.5, 2.0, 5.0, 6.0, 15.0, 20.0, 50.0	Chiral 
methotrexate	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 10000.0	Chiral 
methoxylamine	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	
mevinolin/lovastatin	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	Chiral 

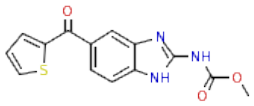
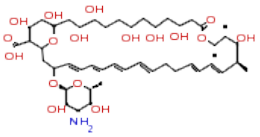
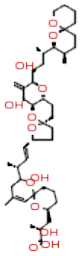
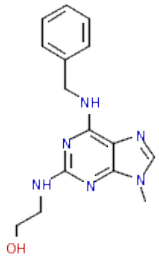
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
mitomycin C	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
mitoxantrone	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
monastrol	0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0	
neomycin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	 Chiral

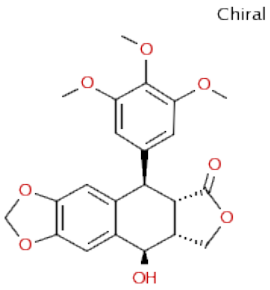
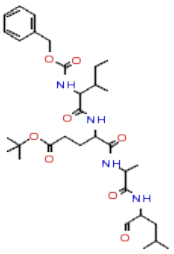
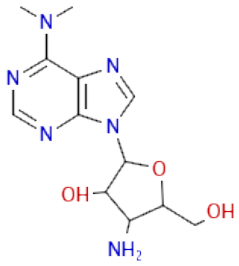
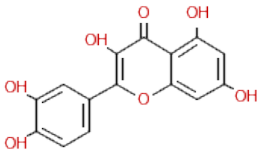
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
nocodazole	0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
nystatin	0.003, 0.006, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6, 1.0, 2.0, 3.0, 6.0, 10.0, 20.0	Chiral 
okadaic acid	6e-05, 0.0002, 0.0006, 0.002, 0.006, 0.02, 0.06, 0.2	Chiral 
olomoucine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	

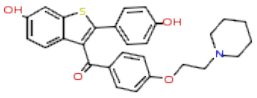
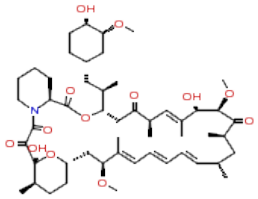
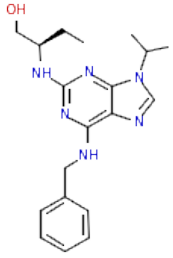
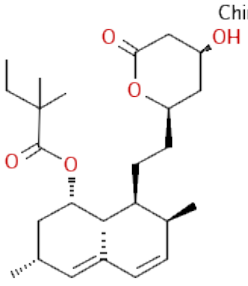
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
podophyllotoxin	3e-06, 1e-05, 3e-05, 0.0001, 0.0003, 0.001, 0.003, 0.01	
proteasome inhibitor I	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	
puromycin	0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0	
quercetin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	

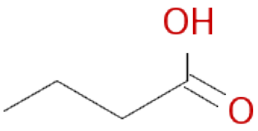
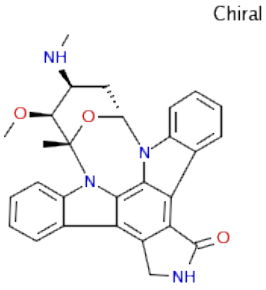
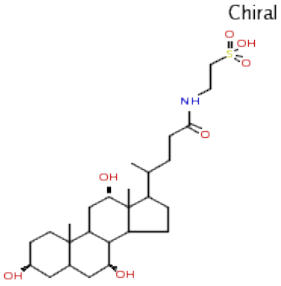
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
raloxifene	0.0015, 0.005, 0.015, 0.05, 0.15, 0.5, 1.5, 5.0	
rapamycin	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 10000.0	Chiral 
roscovitine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	Chiral 
simvastatin	0.006, 0.02, 0.06, 0.2, 0.6, 2.0, 6.0, 20.0	Chiral 

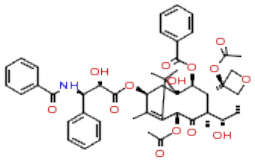
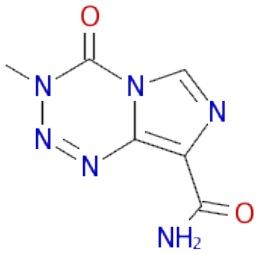
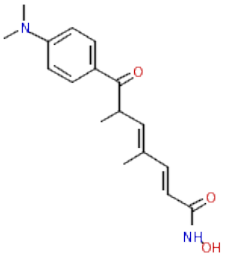
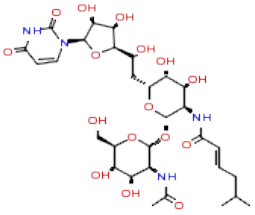
Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [µM]	Structure
sodium butyrate	0.083, 0.25, 0.83, 2.5, 8.3, 25.0, 83.0, 250.0	
sodium fluoride	0.15, 0.5, 1.5, 5.0, 15.0, 50.0, 150.0, 500.0, 10000.0	
staurosporine	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0	
taurocholate	0.0083, 0.025, 0.083, 0.25, 0.83, 2.5, 8.3, 25.0	

Continues on the next page

Table S3: *continued from the previous page*

Compound name	Concentrations [μM]	Structure
taxol	0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0	Chiral 
temozolomide	0.006, 0.02, 0.06, 0.2, 0.6, 2.0, 6.0, 20.0	
trichostatin	0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 10000.0	
tunicamycin	0.015, 0.05, 0.15, 0.5, 1.5, 5.0, 15.0, 50.0	Chiral 

Continues on the next page

Table S3: *continued from the previous page*

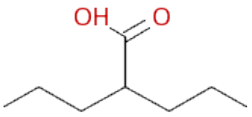
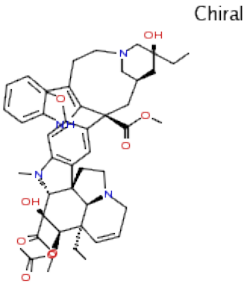
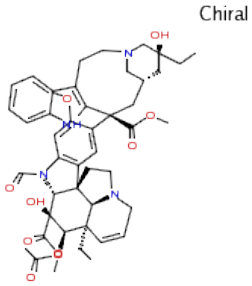
Compound name	Concentrations [μM]	Structure
valproic acid	0.15, 0.5, 1.5, 5.0, 15.0, 50.0, 150.0, 500.0	
vinblastine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 10000.0	
vincristine	0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	

Table S4: Sizes of the phenotypic subpopulations visible to the eye in four compounds in different mechanistic classes, based on inspection of three fields of view in each of two replicate samples for each compound.

			Prevalence of phenotype [%]							
Mechanism	Compound	Concentration [μM]	Mitotic	Monoaster	Fragmented nucleus	Diffuse tubulin	Control	Large cell, large nucleus	Large cell, multiple nuclei	Large nucleus, multiple micronuclei
Microtubule stabilizer	taxol	1.0	27	3	0	0	0	0	2	24
Microtubule destabilizers	demecolcine	1.0	44	0	27	16	12	0	0	0
Aurora kinase inhibitors	AZ-A	1.0	1	0	3	0	11	35	50	0
Eg5 inhibitors	AZ-C	0.1	13	13	7	9	11	4	2	4

Table S5: The 15 most important image features for distinguishing each mechanism of action using the means method.

Mechanism of action	Rank	Score	Image feature
Actin disruptors	1	0.281215	Cytoplasm_Texture_GaborY_CorrActin_3
	2	0.267669	Cytoplasm_Texture_GaborX_CorrActin_3
	3	0.208613	Cells_Texture_GaborY_CorrActin_3
	4	0.187357	Cells_Texture_GaborX_CorrActin_3
	5	0.167390	Nuclei_Texture_SumVariance_CorrDAPI_3
	6	0.159650	Cytoplasm_Intensity_UpperQuartileIntensity_CorrTub
	7	0.156146	Cells_Intensity_MedianIntensity_CorrTub
	8	0.153290	Nuclei_Texture_Variance_CorrDAPI_3
	9	0.152532	Nuclei_Texture_GaborY_CorrActin_3
	10	0.147827	Nuclei_Texture_SumVariance_CorrDAPI_10
	11	0.138384	Nuclei_Texture_GaborX_CorrActin_3
	12	0.124821	Cytoplasm_Intensity_StdIntensity_CorrTub
	13	0.114793	Nuclei_Texture_SumEntropy_CorrDAPI_10
	14	0.113721	Nuclei_Intensity_LowerQuartileIntensity_CorrTub
	15	0.104418	Cytoplasm_Intensity_MeanIntensity_CorrTub
Aurora kinase inhibitors	1	0.179038	Cells_Intensity_IntegratedIntensityEdge_CorrTub
	2	0.164493	Cells_Intensity_IntegratedIntensity_CorrTub
	3	0.152114	Cytoplasm_Intensity_IntegratedIntensity_CorrTub
	4	0.119348	Nuclei_Texture_DifferenceEntropy_CorrDAPI_10
	5	0.097981	Cytoplasm_Intensity_IntegratedIntensityEdge_CorrTub
	6	0.096588	Cytoplasm_AreaShape_Area
	7	0.088599	Nuclei_Texture_DifferenceVariance_CorrDAPI_10
	8	0.087314	Nuclei_Intensity_MeanIntensityEdge_CorrActin
	9	0.074868	Cells_AreaShape_Area
	10	0.070440	Nuclei_Texture_GaborY_CorrTub_3
	11	0.068690	Cytoplasm_Texture_Correlation_CorrTub_10
	12	0.065225	Cytoplasm_AreaShape_Perimeter
	13	0.065225	Cells_AreaShape_Perimeter
	14	0.063790	Nuclei_Intensity_UpperQuartileIntensity_CorrActin
	15	0.063738	Cells_Intensity_UpperQuartileIntensity_CorrActin
Cholesterol-lowering	1	0.093230	Nuclei_Intensity_IntegratedIntensityEdge_CorrDAPI
	2	0.084346	Cells_AreaShape_Extent
	3	0.081998	Cells_Texture_GaborX_CorrTub_3
	4	0.080526	Nuclei_Intensity_MinIntensityEdge_CorrDAPI
	5	0.080094	Nuclei_Intensity_MinIntensity_CorrDAPI
	6	0.074491	Nuclei_Texture_DifferenceEntropy_CorrDAPI_10
	7	0.072067	Nuclei_Texture_GaborX_CorrTub_10
	8	0.069402	Nuclei_Texture_GaborY_CorrTub_10

Continues on the next page

Table S5: *continued from the previous page*

Mechanism of action	Rank	Score	Image feature
	9	0.066591	Cells_Texture_Correlation_CorrTub_10
	10	0.058949	Cells_Zernike_3_1
	11	0.056188	Cytoplasm_Texture_InfoMeas2_CorrTub_3
	12	0.053279	Nuclei_Texture_Contrast_CorrDAPI_10
	13	0.052195	Nuclei_Texture_InverseDifferenceMoment_CorrDAPI_10
	14	0.050724	Cytoplasm_Texture_InfoMeas1_CorrTub_3
	15	0.050655	Cytoplasm_Texture_GaborX_CorrTub_3
DNA damage	1	0.045092	Cytoplasm_Intensity_IntegratedIntensity_CorrTub
	2	0.044668	Cytoplasm_Intensity_MeanIntensity_CorrTub
	3	0.043116	Cells_Intensity_MeanIntensityEdge_CorrTub
	4	0.040310	Nuclei_Intensity_MaxIntensity_CorrActin
	5	0.039393	Cells_Neighbors_PercentTouching_10
	6	0.039011	Cytoplasm_Intensity_MedianIntensity_CorrTub
	7	0.038127	Nuclei_Neighbors_PercentTouching_10
	8	0.035094	Nuclei_Intensity_MeanIntensity_CorrActin
	9	0.034896	Cytoplasm_AreaShape_FormFactor
	10	0.034740	Nuclei_Neighbors_PercentTouching_20
	11	0.033824	Cytoplasm_Intensity_StdIntensityEdge_CorrActin
	12	0.033655	Cells_Intensity_MedianIntensity_CorrTub
	13	0.032402	Cells_Intensity_StdIntensityEdge_CorrTub
	14	0.031690	Cytoplasm_Intensity_StdIntensity_CorrActin
	15	0.031184	Nuclei_Intensity_StdIntensityEdge_CorrDAPI
DNA replication	1	0.079445	Nuclei_Neighbors_PercentTouching_20
	2	0.060039	Nuclei_Neighbors_NumberOfNeighbors_20
	3	0.045092	Cytoplasm_Intensity_IntegratedIntensity_CorrTub
	4	0.044645	Cells_Intensity_UpperQuartileIntensity_CorrTub
	5	0.042304	Nuclei_Texture_SumEntropy_CorrDAPI_3
	6	0.040094	Cytoplasm_Texture_GaborY_CorrActin_3
	7	0.039256	Nuclei_Neighbors_PercentTouching_10
	8	0.037857	Nuclei_Texture_InfoMeas1_CorrDAPI_10
	9	0.037525	Nuclei_Texture_InfoMeas1_CorrDAPI_3
	10	0.035017	Cytoplasm_Intensity_UpperQuartileIntensity_CorrTub
	11	0.033534	Cytoplasm_AreaShape_Solidity
	12	0.032936	Cytoplasm_Intensity_MeanIntensityEdge_CorrTub
	13	0.032238	Cytoplasm_Intensity_IntegratedIntensity_CorrActin
	14	0.032039	Nuclei_Intensity_MedianIntensity_CorrTub
	15	0.031169	Cells_AreaShape_Area
Eg5 inhibitors	1	0.136123	Cytoplasm_Texture_GaborX_CorrTub_10
	2	0.134797	Nuclei_Intensity_StdIntensity_CorrTub
	3	0.133356	Cytoplasm_Texture_GaborY_CorrTub_10
	4	0.101140	Cells_Intensity_StdIntensity_CorrTub
	5	0.095494	Nuclei_Intensity_MedianIntensity_CorrActin
	6	0.092331	Nuclei_Intensity_MeanIntensity_CorrActin
	7	0.088028	Cells_Texture_InfoMeas1_CorrTub_3
	8	0.087820	Nuclei_Intensity_LowerQuartileIntensity_CorrActin
	9	0.082726	Cells_Texture_SumVariance_CorrTub_10
	10	0.080209	Cells_Texture_InfoMeas2_CorrTub_3
	11	0.077260	Nuclei_Intensity_UpperQuartileIntensity_CorrActin
	12	0.075767	Nuclei_Intensity_MeanIntensityEdge_CorrActin
	13	0.075408	Nuclei_Texture_SumEntropy_CorrTub_10
	14	0.072718	Cells_Intensity_StdIntensity_CorrActin

Continues on the next page

Table S5: *continued from the previous page*

Mechanism of action	Rank	Score	Image feature
	15	0.071903	Cells_Intensity_UpperQuartileIntensity_CorrActin
Epithelial	1	0.225988	Nuclei_Neighbors_NumberOfNeighbors_20
	2	0.204335	Nuclei_Neighbors_NumberOfNeighbors_10
	3	0.153913	Nuclei_Neighbors_PercentTouching_20
	4	0.135766	Nuclei_Neighbors_PercentTouching_10
	5	0.096608	Cytoplasm_Intensity_MinIntensity_CorrTub
	6	0.091836	Cells_Intensity_MinIntensityEdge_CorrTub
	7	0.087701	Cells_Intensity_MinIntensity_CorrTub
	8	0.085939	Cytoplasm_Intensity_MinIntensityEdge_CorrTub
	9	0.081036	Nuclei_Intensity_IntegratedIntensityEdge_CorrTub
	10	0.066969	Nuclei_Intensity_IntegratedIntensity_CorrTub
	11	0.060913	Cells_AreaShape_FormFactor
	12	0.055912	Cytoplasm_Texture_Contrast_CorrTub_3
	13	0.053228	Nuclei_Texture_Contrast_CorrDAPI_10
	14	0.052187	Cells_Intensity_IntegratedIntensity_CorrTub
	15	0.048274	Cytoplasm_AreaShape_MinorAxisLength
Kinase inhibitors	1	0.134838	Nuclei_Intensity_IntegratedIntensityEdge_CorrTub
	2	0.131200	Cytoplasm_Intensity_IntegratedIntensityEdge_CorrTub
	3	0.130932	Cytoplasm_AreaShape_FormFactor
	4	0.109810	Cells_Intensity_IntegratedIntensityEdge_CorrTub
	5	0.098938	Nuclei_Intensity_MeanIntensity_CorrDAPI
	6	0.098835	Nuclei_Intensity_IntegratedIntensity_CorrDAPI
	7	0.093230	Nuclei_Intensity_IntegratedIntensityEdge_CorrDAPI
	8	0.085660	Nuclei_Intensity_UpperQuartileIntensity_CorrDAPI
	9	0.084234	Nuclei_Texture_InfoMeas1_CorrDAPI_3
	10	0.077387	Nuclei_Intensity_MedianIntensity_CorrDAPI
	11	0.073766	Cytoplasm_AreaShape_Area
	12	0.069723	Nuclei_Intensity_MaxIntensity_CorrDAPI
	13	0.069517	Cytoplasm_Texture_InfoMeas1_CorrActin_10
	14	0.066969	Nuclei_Intensity_IntegratedIntensity_CorrTub
	15	0.066534	Cytoplasm_Texture_GaborX_CorrTub_3
Microtubule destabilizers	1	0.066222	Nuclei_Texture_AngularSecondMoment_CorrTub_10
	2	0.050640	Nuclei_Texture_AngularSecondMoment_CorrTub_3
	3	0.050038	Nuclei_Texture_SumVariance_CorrTub_3
	4	0.047789	Nuclei_Texture_Correlation_CorrDAPI_3
	5	0.047526	Nuclei_Intensity_IntegratedIntensityEdge_CorrActin
	6	0.045564	Nuclei_Texture_InfoMeas1_CorrDAPI_3
	7	0.043635	Cells_Texture_Variance_CorrTub_10
	8	0.039665	Nuclei_Texture_InfoMeas2_CorrDAPI_3
	9	0.036823	Nuclei_Texture_SumVariance_CorrActin_10
	10	0.035998	Nuclei_Intensity_StdIntensityEdge_CorrActin
	11	0.035751	Nuclei_Texture_SumVariance_CorrTub_10
	12	0.035645	Cells_Intensity_MeanIntensityEdge_CorrActin
	13	0.035200	Cells_Intensity_StdIntensityEdge_CorrActin
	14	0.034137	Nuclei_Texture_Variance_CorrTub_3
	15	0.034126	Cells_Texture_InfoMeas1_CorrActin_3
Microtubule stabilizers	1	0.266498	Nuclei_Intensity_StdIntensityEdge_CorrTub
	2	0.230081	Cytoplasm_Texture_AngularSecondMoment_CorrTub_3
	3	0.198321	Cells_Texture_GaborY_CorrTub_10
	4	0.191167	Cytoplasm_Texture_SumEntropy_CorrTub_3

Continues on the next page

Table S5: *continued from the previous page*

Mechanism of action	Rank	Score	Image feature
	5	0.190461	Cytoplasm_Texture_AngularSecondMoment_CorrTub_10
	6	0.189933	Cytoplasm_Intensity_StdIntensityEdge_CorrTub
	7	0.185869	Cells_Texture_GaborX_CorrTub_10
	8	0.178724	Cells_Texture_SumAverage_CorrTub_3
	9	0.174168	Cells_Texture_SumAverage_CorrTub_10
	10	0.160962	Cytoplasm_Texture_GaborY_CorrTub_10
	11	0.154275	Nuclei_Intensity_StdIntensity_CorrTub
	12	0.153313	Cytoplasm_Texture_SumAverage_CorrTub_3
	13	0.150168	Cells_Intensity_StdIntensity_CorrTub
	14	0.150092	Cytoplasm_Texture_SumEntropy_CorrTub_10
	15	0.149339	Cytoplasm_Texture_SumAverage_CorrTub_10
Protein degradation	1	0.084139	Cytoplasm_Texture_GaborX_CorrTub_10
	2	0.083899	Cytoplasm_Zernike_3_1
	3	0.081998	Cytoplasm_Intensity_IntegratedIntensity_CorrActin
	4	0.081266	Cytoplasm_Texture_GaborY_CorrTub_10
	5	0.075760	Cytoplasm_AreaShape_Area
	6	0.071945	Cytoplasm_AreaShape_Solidity
	7	0.067088	Cytoplasm_Texture_GaborX_CorrActin_10
	8	0.064361	Cytoplasm_Zernike_0_0
	9	0.056766	Cytoplasm_Texture_GaborY_CorrActin_10
	10	0.056530	Cells_Texture_GaborX_CorrActin_10
	11	0.046923	Cells_Texture_GaborY_CorrActin_10
	12	0.045643	Cells_Texture_Contrast_CorrActin_10
	13	0.045052	Cytoplasm_AreaShape_Extent
	14	0.044919	Cells_Texture_GaborX_CorrTub_10
	15	0.040845	Cells_Texture_DifferenceVariance_CorrActin_10
Protein synthesis	1	0.092005	Cytoplasm_Texture_Correlation_CorrActin_3
	2	0.086052	Cells_Texture_Correlation_CorrActin_3
	3	0.068384	Cytoplasm_Texture_DifferenceVariance_CorrActin_3
	4	0.066499	Cells_Texture_DifferenceVariance_CorrActin_3
	5	0.064255	Cytoplasm_Texture_Contrast_CorrTub_3
	6	0.053710	Cytoplasm_Texture_SumAverage_CorrActin_3
	7	0.053177	Cytoplasm_Texture_Contrast_CorrActin_3
	8	0.052644	Cells_Texture_InfoMeas1_CorrActin_3
	9	0.051165	Cells_Texture_SumVariance_CorrTub_3
	10	0.050724	Cytoplasm_Texture_InfoMeas1_CorrTub_3
	11	0.050487	Cytoplasm_Texture_GaborX_CorrTub_3
	12	0.048193	Cells_Texture_SumVariance_CorrTub_10
	13	0.045031	Cells_Texture_Variance_CorrTub_10
	14	0.041479	Nuclei_Intensity_MinIntensityEdge_CorrActin
	15	0.038174	Cells_Intensity_LowerQuartileIntensity_CorrActin

Table S6: The number of times each feature was selected by the SVMRFE method. Only the 20 most selected features are shown.

Feature	Number of times selected
Cells_Neighbors_NumberOfNeighbors_10	171
Cytoplasm_Texture_GaborY_CorrTub_10	138
Cytoplasm_Texture_GaborX_CorrTub_10	133
Cells_Neighbors_PercentTouching_3	115
Cells_Neighbors_NumberOfNeighbors_3	112
Nuclei_Intensity_LowerQuartileIntensity_CorrDAPI	105
Cells_Intensity_IntegratedIntensityEdge_CorrTub	85
Cytoplasm_Texture_GaborX_CorrActin_10	79
Nuclei_Intensity_MinIntensityEdge_CorrDAPI	78
Nuclei_AreaShape_FormFactor	77
Cells_Neighbors_PercentTouching_10	74
Nuclei_Intensity_MaxIntensity_CorrTub	74
Nuclei_Intensity_MedianIntensity_CorrDAPI	74
Nuclei_Texture_SumAverage_CorrDAPI_3	73
Nuclei_Texture_SumAverage_CorrDAPI_10	72
Nuclei_Intensity_MinIntensity_CorrDAPI	69
Cells_Intensity_MeanIntensityEdge_CorrTub	68
Cytoplasm_Texture_GaborY_CorrActin_10	67
Cells_Intensity_StdIntensityEdge_CorrTub	65
Nuclei_Texture_SumVariance_CorrDAPI_3	61

Table S7: Accuracies for all combinations of dimensionality-reducing preprocessing method and profiling method

	Profiling method			
Dimensionality reduction method	Means	KS	SVM	GM
None	83%	83%	81%	81%
Factor analysis	94%	90%	80%	84%
PCA	81%	90%	73%	73%
Factor-analysis–based feature selection	67%	65%	68%	75%

Table S8: The 15 features most heavily loaded onto each factor in the 50-factor model

Factor number	Measurements
1	Nuclei_Texture_GaborY_CorrDAPI_10 Nuclei_Intensity_StdIntensityEdge_CorrActin Nuclei_Texture_Variance_CorrDAPI_3 Nuclei_Texture_Entropy_CorrActin_10 Nuclei_Texture_GaborX_CorrDAPI_10
2	Nuclei_Texture_DifferenceEntropy_CorrDAPI_10 Nuclei_Texture_Contrast_CorrDAPI_10 Cytoplasm_Texture_InfoMeas2_CorrTub_3 Nuclei_Texture_DifferenceVariance_CorrDAPI_10 Cytoplasm_Texture_SumVariance_CorrTub_3

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
3	Cells_Zernike_1_1 Cells_Zernike_0_0 Cells_Neighbors_FirstClosestYVector_10 Cells_Neighbors_FirstClosestYVector_3 Cytoplasm_Zernike_1_1
4	Nuclei_Texture_AngularSecondMoment_CorrTub_3 Nuclei_Texture_Entropy_CorrTub_3 Nuclei_Intensity_MinIntensity_CorrTub Nuclei_Intensity_MinIntensityEdge_CorrTub Cells_Texture_AngularSecondMoment_CorrTub_3
5	Nuclei_Texture_SumEntropy_CorrTub_10 Nuclei_Texture_SumAverage_CorrTub_10 Cells_Texture_SumAverage_CorrTub_10 Nuclei_Texture_SumAverage_CorrTub_3 Nuclei_Texture_Entropy_CorrTub_10
6	Cells_Intensity_IntegratedIntensityEdge_CorrTub Cytoplasm_Texture_InfoMeas2_CorrActin_3 Nuclei_Texture_SumEntropy_CorrActin_3 Nuclei_Texture_DifferenceEntropy_CorrTub_3 Cytoplasm_Texture_InfoMeas1_CorrActin_3
7	Nuclei_Zernike_2_0 Nuclei_Zernike_0_0 Nuclei_AreaShape_Eccentricity Nuclei_Zernike_1_1 Nuclei_Zernike_2_2
8	Nuclei_Texture_DifferenceEntropy_CorrTub_10 Nuclei_Texture_Entropy_CorrTub_10 Nuclei_Texture_SumEntropy_CorrTub_3 Nuclei_Texture_DifferenceEntropy_CorrDAPI_10 Nuclei_Texture_Contrast_CorrDAPI_10
9	Cells_AreaShape_FormFactor Cytoplasm_Texture_InverseDifferenceMoment_CorrActin_10 Cells_Neighbors_PercentTouching_3 Cytoplasm_AreaShape_Perimeter Cells_AreaShape_Perimeter
10	Cells_Intensity_MeanIntensityEdge_CorrTub Cytoplasm_Intensity_MeanIntensityEdge_CorrTub Cytoplasm_Intensity_MeanIntensity_CorrTub Cytoplasm_Intensity_LowerQuartileIntensity_CorrTub Cytoplasm_Intensity_MedianIntensity_CorrTub

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
11	Nuclei_Texture_Entropy_CorrTub_10 Nuclei_Texture_Entropy_CorrTub_3 Cytoplasm_AreaShape_MinorAxisLength Nuclei_Texture_AngularSecondMoment_CorrTub_3 Cells_AreaShape_MinorAxisLength
12	Cells_Neighbors_SecondClosestYVector_3 Cells_Neighbors_SecondClosestYVector_10 Cytoplasm_Texture_InfoMeas2_CorrTub_3 Nuclei_Texture_InfoMeas2_CorrTub_3 Nuclei_Texture_InfoMeas1_CorrTub_3
13	Nuclei_Texture_DifferenceEntropy_CorrActin_3 Nuclei_Texture_Contrast_CorrActin_3 Nuclei_Texture_InverseDifferenceMoment_CorrDAPI_3 Nuclei_Texture_InverseDifferenceMoment_CorrActin_3 Nuclei_Texture_DifferenceVariance_CorrActin_3
14	Cells_Texture_Correlation_CorrTub_3 Cytoplasm_Texture_Correlation_CorrTub_3 Cells_Texture_InfoMeas2_CorrTub_3 Cells_Texture_SumEntropy_CorrTub_10 Cells_Texture_SumEntropy_CorrTub_3
15	Cells_Texture_Variance_CorrActin_3 Nuclei_Texture_Variance_CorrActin_3 Cells_Texture_Entropy_CorrActin_3 Cells_Texture_Variance_CorrActin_10 Nuclei_Texture_SumVariance_CorrActin_3
16	Cytoplasm_Texture_Correlation_CorrActin_3 Cytoplasm_Texture_InfoMeas2_CorrActin_3 Nuclei_Texture_SumVariance_CorrTub_3 Nuclei_Texture_Variance_CorrTub_3 Cells_Texture_SumEntropy_CorrActin_3
17	Nuclei_Texture_Entropy_CorrActin_3 Nuclei_Texture_AngularSecondMoment_CorrActin_3 Nuclei_Texture_Entropy_CorrActin_10 Nuclei_Texture_AngularSecondMoment_CorrActin_10 Nuclei_Texture_SumEntropy_CorrActin_3
18	Cells_Texture_Entropy_CorrTub_3 Cytoplasm_Texture_Entropy_CorrTub_3 Cells_Texture_DifferenceEntropy_CorrTub_3 Cells_Texture_AngularSecondMoment_CorrTub_3 Cells_Texture_InverseDifferenceMoment_CorrTub_3

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
19	Cells_Intensity_MinIntensityEdge_CorrActin Cytoplasm_Intensity_MinIntensityEdge_CorrActin Cytoplasm_Intensity_MinIntensity_CorrActin Cells_Intensity_MinIntensity_CorrActin Cells_Intensity_MeanIntensityEdge_CorrActin
20	Cells_Texture_Correlation_CorrTub_3 Cells_Texture_InfoMeas1_CorrTub_3 Cytoplasm_Texture_Correlation_CorrActin_3 Cytoplasm_Texture_Correlation_CorrTub_3 Cells_Texture_InfoMeas2_CorrActin_3
21	Cytoplasm_Texture_SumEntropy_CorrActin_10 Cells_Texture_InverseDifferenceMoment_CorrTub_10 Cytoplasm_Texture_Entropy_CorrActin_10 Cytoplasm_Texture_SumEntropy_CorrActin_3 Cytoplasm_Texture_AngularSecondMoment_CorrActin_10
22	Cytoplasm_Intensity_StdIntensityEdge_CorrActin Cells_Intensity_StdIntensityEdge_CorrActin Cells_Intensity_StdIntensity_CorrActin Cytoplasm_Intensity_StdIntensity_CorrActin Nuclei_Intensity_StdIntensityEdge_CorrActin
23	Nuclei_Intensity_UpperQuartileIntensity_CorrActin Nuclei_Intensity_IntegratedIntensity_CorrActin Nuclei_Texture_SumAverage_CorrActin_10 Nuclei_Intensity_MedianIntensity_CorrActin Nuclei_Texture_SumAverage_CorrActin_3
24	Nuclei_Intensity_IntegratedIntensityEdge_CorrTub Nuclei_Intensity_MeanIntensityEdge_CorrTub Nuclei_Intensity_MaxIntensityEdge_CorrTub Nuclei_Texture_Correlation_CorrActin_10 Cytoplasm_Intensity_StdIntensityEdge_CorrTub
25	Cytoplasm_Zernike_0_0 Cytoplasm_Zernike_1_1 Cytoplasm_Intensity_IntegratedIntensity_CorrTub Cells_AreaShape_MinorAxisLength Cytoplasm_Intensity_IntegratedIntensity_CorrActin
26	Nuclei_Intensity_IntegratedIntensityEdge_CorrTub Nuclei_Intensity_MaxIntensityEdge_CorrTub Cytoplasm_Intensity_MaxIntensityEdge_CorrTub Nuclei_Texture_SumEntropy_CorrDAPI_10 Nuclei_Texture_SumVariance_CorrDAPI_10

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
27	Cytoplasm_Intensity_IntegratedIntensityEdge_CorrActin Cells_Intensity_IntegratedIntensityEdge_CorrActin Nuclei_Intensity_MeanIntensityEdge_CorrActin Nuclei_Intensity_MinIntensityEdge_CorrActin Nuclei_Intensity_MinIntensity_CorrActin
28	Nuclei_Texture_AngularSecondMoment_CorrDAPI_3 Nuclei_Texture_Entropy_CorrDAPI_3 Cells_Texture_SumVariance_CorrTub_10 Nuclei_Texture_InverseDifferenceMoment_CorrDAPI_3 Cells_Texture_SumVariance_CorrTub_3
29	Nuclei_Texture_DifferenceEntropy_CorrTub_3 Nuclei_Texture_DifferenceVariance_CorrTub_3 Nuclei_Texture_Contrast_CorrTub_3 Cells_Texture_SumVariance_CorrActin_10 Nuclei_Texture_InverseDifferenceMoment_CorrTub_3
30	Cells_Intensity_MeanIntensityEdge_CorrActin Cells_Intensity_MaxIntensityEdge_CorrActin Cytoplasm_Intensity_MaxIntensity_CorrActin Cytoplasm_Intensity_MeanIntensityEdge_CorrActin Cytoplasm_Intensity_MeanIntensity_CorrActin
31	Cells_Intensity_MaxIntensityEdge_CorrTub Cells_Intensity_MeanIntensityEdge_CorrTub Cytoplasm_Texture_Correlation_CorrActin_3 Cytoplasm_Intensity_MeanIntensityEdge_CorrTub Cytoplasm_Intensity_MaxIntensity_CorrTub
32	Cytoplasm_Texture_Entropy_CorrActin_3 Cytoplasm_Texture_InverseDifferenceMoment_CorrActin_3 Cytoplasm_Texture_AngularSecondMoment_CorrActin_3 Cells_Texture_InverseDifferenceMoment_CorrActin_3 Cytoplasm_Texture_DifferenceEntropy_CorrActin_3
33	Nuclei_Texture_Correlation_CorrActin_3 Nuclei_Texture_InfoMeas2_CorrActin_3 Nuclei_Texture_InfoMeas1_CorrActin_3 Nuclei_Texture_Correlation_CorrActin_10 Cytoplasm_Intensity_MeanIntensity_CorrTub
34	Nuclei_Intensity_IntegratedIntensity_CorrActin Nuclei_Intensity_IntegratedIntensityEdge_CorrActin Cells_Neighbors_FirstClosestYVector_10 Cells_Neighbors_FirstClosestYVector_3 Cells_Neighbors_FirstClosestXVector_10

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
35	Cytoplasm_Intensity_StdIntensity_CorrTub Cells_Texture_DifferenceVariance_CorrTub_10 Cytoplasm_AreaShape_Eccentricity Cells_AreaShape_Eccentricity Cytoplasm_Zernike_2_2
36	Cells_Intensity_LowerQuartileIntensity_CorrActin Cells_Intensity_MedianIntensity_CorrActin Cells_Intensity_MeanIntensity_CorrActin Cytoplasm_Intensity_MedianIntensity_CorrActin Cytoplasm_Intensity_MeanIntensity_CorrActin
37	Nuclei_Intensity_IntegratedIntensityEdge_CorrDAPI Nuclei_AreaShape_Solidity Nuclei_AreaShape_FormFactor Nuclei_AreaShape_Extent Nuclei_Intensity_MeanIntensityEdge_CorrDAPI
38	Cytoplasm_AreaShape_Solidity Cytoplasm_AreaShape_Extent Cytoplasm_Texture_Contrast_CorrActin_3 Cytoplasm_Intensity_MinIntensity_CorrTub Cells_Intensity_MinIntensity_CorrTub
39	Nuclei_Texture_SumAverage_CorrDAPI_10 Nuclei_Intensity_StdIntensity_CorrTub Nuclei_Texture_SumAverage_CorrDAPI_3 Nuclei_Texture_GaborX_CorrTub_10 Cells_Intensity_MaxIntensity_CorrTub
40	Nuclei_Texture_Contrast_CorrDAPI_10 Nuclei_Texture_DifferenceVariance_CorrDAPI_10 Cells_Texture_DifferenceVariance_CorrActin_10 Nuclei_Texture_SumVariance_CorrDAPI_3 Cells_Texture_Contrast_CorrActin_10
41	Cytoplasm_Texture_SumAverage_CorrTub_10 Cytoplasm_Texture_SumAverage_CorrTub_3 Cells_Texture_SumAverage_CorrTub_3 Cells_Texture_SumAverage_CorrTub_10 Nuclei_Texture_SumAverage_CorrTub_10
42	Cytoplasm_Texture_Contrast_CorrTub_10 Cytoplasm_Texture_DifferenceVariance_CorrTub_10 Cytoplasm_Texture_DifferenceEntropy_CorrTub_3 Cytoplasm_Intensity_UpperQuartileIntensity_CorrTub Cells_Texture_Contrast_CorrTub_3

Continues on the next page

Table S8: *continued from the previous page*

Factor number	Measurements
43	Cytoplasm_Intensity_IntegratedIntensityEdge_CorrTub Cells_Intensity_IntegratedIntensityEdge_CorrTub Nuclei_Texture_Entropy_CorrActin_3 Nuclei_Intensity_IntegratedIntensityEdge_CorrTub Nuclei_Texture_SumEntropy_CorrActin_10
44	Nuclei_AreaShape_Perimeter Nuclei_AreaShape_Area Nuclei_AreaShape_MajorAxisLength Nuclei_Intensity_IntegratedIntensity_CorrDAPI Nuclei_AreaShape_MinorAxisLength
45	Cells_Zernike_3_1 Cytoplasm_AreaShape_Eccentricity Cells_AreaShape_Eccentricity Cells_Zernike_0_0 Cells_AreaShape_Solidity
46	Nuclei_AreaShape_Eccentricity Nuclei_Intensity_IntegratedIntensityEdge_CorrDAPI Nuclei_Zernike_0_0 Nuclei_Intensity_MeanIntensityEdge_CorrDAPI Nuclei_Zernike_2_0
47	Cells_Intensity_MinIntensity_CorrActin Cytoplasm_Intensity_MinIntensity_CorrActin Cytoplasm_Intensity_MinIntensityEdge_CorrActin Cells_Intensity_MinIntensityEdge_CorrActin Nuclei_Neighbors_PercentTouching_10
48	Cells_Texture_SumAverage_CorrActin_3 Cells_Texture_SumAverage_CorrActin_10 Nuclei_Texture_Variance_CorrDAPI_3 Nuclei_Neighbors_SecondClosestXVector_20 Nuclei_Neighbors_SecondClosestXVector_10
49	Cells_Texture_DifferenceVariance_CorrTub_3 Cells_Texture_Contrast_CorrTub_3 Cytoplasm_Texture_DifferenceVariance_CorrTub_3 Cytoplasm_Texture_Contrast_CorrTub_3 Cells_Texture_DifferenceEntropy_CorrTub_3
50	Cells_Texture_Correlation_CorrActin_10 Cells_Texture_InfoMeas2_CorrActin_3 Cells_Texture_Correlation_CorrActin_3 Cells_Texture_InfoMeas2_CorrActin_10 Nuclei_Texture_Contrast_CorrActin_3

Table S9: Distribution of mechanisms of action across batches

Mechanism of action	Number of compounds	Batches
Actin disruptors	3	02, 01
Aurora kinase inhibitors	3	01, 04, 03
Cholesterol-lowering	2	09
DMSO	1	02, 10, 03, 01, 06, 08, 07, 09, 04, 05
DNA damage	4	04, 03
DNA replication	4	08, 02, 09
Eg5 inhibitors	2	03, 10
Epithelial	3	05, 08, 10
Kinase inhibitors	3	07
Microtubule destabilizers	4	01, 03
Microtubule stabilizers	3	07, 01
Protein degradation	4	07, 02, 06
Protein synthesis	3	04, 03

Table S10: AUCs and p-values for the five profiling methods

	Profiling method									
	Means		KS statistic		SVM normalvector		Gaussian mixture		Factor analysis	
Mechanism	AUC	p-value	AUC	p-value	AUC	p-value	AUC	p-value	AUC	p-value
Actin disruptors	0.674	0.0295	0.914	≤ 0.0001	0.998	≤ 0.0001	0.689	0.0196	0.979	≤ 0.0001
Aurora kinase inhibitors	1.000	≤ 0.0001	1.000	≤ 0.0001	0.988	≤ 0.0001	1.000	≤ 0.0001	1.000	≤ 0.0001
Cholesterol-lowering	0.879	≤ 0.0001	0.966	≤ 0.0001	0.923	≤ 0.0001	0.741	0.0005	0.965	≤ 0.0001
DNA damage	0.834	≤ 0.0001	0.742	≤ 0.0001	0.942	≤ 0.0001	0.868	≤ 0.0001	0.930	≤ 0.0001
DNA replication	0.838	≤ 0.0001	0.708	≤ 0.0001	0.816	≤ 0.0001	0.834	≤ 0.0001	0.903	≤ 0.0001
Eg5 inhibitors	0.987	≤ 0.0001	0.994	≤ 0.0001	0.984	≤ 0.0001	0.995	≤ 0.0001	1.000	≤ 0.0001
Epithelial	0.973	≤ 0.0001	0.966	≤ 0.0001	0.924	≤ 0.0001	0.966	≤ 0.0001	0.978	≤ 0.0001
Kinase inhibitors	0.999	≤ 0.0001	0.995	≤ 0.0001	0.993	≤ 0.0001	0.994	≤ 0.0001	1.000	≤ 0.0001
Microtubule destabilizers	0.852	≤ 0.0001	0.793	≤ 0.0001	0.881	≤ 0.0001	0.854	≤ 0.0001	0.865	≤ 0.0001
Microtubule stabilizers	0.781	≤ 0.0001	0.804	≤ 0.0001	0.981	≤ 0.0001	0.955	≤ 0.0001	0.998	≤ 0.0001
Protein degradation	0.722	0.0002	0.715	0.0003	0.716	0.0003	0.701	0.0007	0.816	≤ 0.0001
Protein synthesis	0.976	≤ 0.0001	0.985	≤ 0.0001	0.997	≤ 0.0001	0.975	≤ 0.0001	0.997	≤ 0.0001