Optimum concentration-response curve metrics for supervised selection of discriminative cellular phenotypic endpoints for chemical hazard assessment

James Alastair Miller, Lit-Hsin Loo*

Innovations in Food and Chemical Safety Programme and Bioinformatics Institute, Agency for Science, Technology, and Research, 30 Biopolis Street, #07-01 Matrix, Singapore 138671, Singapore

Keywords: Dose response curve analysis, in vitro cell models, high-content imaging
Supplementary material: Supplementary data will be made available online.

* To whom correspondence should be addressed: loolh@bii.a-star.edu.sg (ORCID:0000-0001-6303-9840)

ABSTRACT

High-content imaging (HCI) provides quantitative and information-rich measurements of chemical effects on human *in vitro* cell models. Identification of discriminative phenotypic endpoints from cellular features obtained from HCI is required for accurate assessments of potential chemical hazards. However, the use of suboptimal metrics to quantify the concentration response curves (CRC) of chemicals based on these features may obscure discriminative features, and lead to non-predictive endpoints and poor chemical classifications or hazard assessments. Here, we present a systematic and data-driven study on the performances of different CRC metrics in identifying image-based phenotypic features that can accurately classify the effects of reference chemicals with known *in vivo* toxicities. We studied four previous HCI *in vitro* nephro- or pulmono-toxicity datasets, which contain phenotypic feature measurements from different cell and feature types. Within a feature type, we found that efficacy metrics at higher chemical concentrations tend to give higher classification accuracy, whereas potency metrics do not have obvious trends across different response levels. Across different cell and feature types, efficacy metrics generally gave higher classification accuracy than potency metrics and area under the curve (AUC). Our results suggest that efficacy metrics, especially at higher concentrations, are more likely to help us to identify discriminative phenotypic endpoints. Therefore, HCI experiments for toxicological applications should include measurements at sufficiently high chemical concentrations, and efficacy metrics should always be analyzed. The identified features may be used as specific toxicity endpoints for further chemical hazard assessment.

INTRODUCTION

High-content imaging (HCI) is increasingly used to develop *in vitro* cell-based toxicity models (Slikker et al. 2018; Thomas et al. 2019), including those for nephrotoxicity (Su et al. 2016; Sjögren et al. 2018; van der Ven et al. 2020), pulmonary toxicity (Lee et al. 2018), hepatotoxicity (Wink et al. 2018), neurotoxicity (Delp et al. 2019), and cardiotoxicity (Grimm et al. 2017). The technology is especially useful when the modes of action of a chemical is unknown or involves multiple biological pathways, because different phenotypic features can be measured simultaneously from HCI images, including features of cellular morphology, intracellular organelle structures, and protein expression and spatial distribution patterns (Loo et al. 2007, 2009; Bougen-Zhukov et al. 2017). However, which of these features should be used as toxicity endpoints for specific adverse effects of interest or concern? Not all features are expected to provide the same discriminative information about these adverse effects. Cell death and other related cytotoxicity endpoints are common choices, however these features were previously found to be sensitive, but not specific for predicting *in vivo* toxicity (Lin and Will 2012; Lee et al. 2018). Other non-cell-death related phenotypic features may be more specific to the key cellular events associated to the adverse effects, and thus more accurate in distinguishing between toxic and non-toxic chemicals. Therefore, discriminative endpoints for different adverse effects are likely to be different, and thus have to be separately identified for each adverse effect.

Commonly used feature selection methods (Kohavi and John 1997) cannot be directly applied to HCI data generated from toxicological studies, because cellular responses in these studies are usually measured in multiple discrete chemical concentrations (Sirenko et al. 2015; Grimm et al. 2015; Su et al. 2016; Hafner et al. 2017; Lee et al. 2018) (**Fig. 1a**). To model the relationship between chemical concentrations and a phenotypic effect, a concentration-response curve (CRC) is usually first fitted based on the measured discrete feature values and then characterized by a CRC metric (**Fig. 1b**). To the best of our knowledge, there is no previous published work on which CRC metrics are appropriate or optimum for selecting discriminative phenotypic endpoints. This an important question because the use of non-optimum CRC metrics in HCI toxicological studies may lead to non-predictive endpoints and poor chemical classifications or hazard assessments.

Common CRC metrics include potency metrics that report the concentrations of chemicals required to produce a pre-defined effect, such as the half-maximal effective concentration (EC_{50}); and efficacy metrics that report the maximum effect levels of chemicals (E_{max}). In fact, for a CRC of a phenotypic feature, there are infinitely many possible potency- or efficacy-based metrics, such as the effective concentration at any Y% response level (i.e., EC_Y), or the response level at any X concentration level (i.e., $R_{[X]}$), that can be evaluated from the same curve. (For a log-logistic CRC model, E_{max} is equal to $R_{[\infty]}$.) The area under the curve (AUC) is another common CRC metric which combines elements of both potency and efficacy quantifications. In the literature, potency metrics are more commonly used than efficacy metrics for *in vitro* toxicological studies (O'Brien et al. 2006; Lin and Will 2012; Sirenko et al. 2015; Sjögren et al. 2018) (**Supplementary Fig. S1**). In the US EPA Toxicity Forecaster (ToxCast) Programme, a variant of EC_{50} called "activity concentration at 50% of maximal activity" (AC_{50}) is used to characterize and study the high-throughput toxicity screening data generated by the programme, some of which are based on HCI (Kleinstreuer et al. 2014; Paul Friedman et al. 2020). The wide adoption of potency metrics may be due to the fact that most traditional toxicological endpoints are dichotomous (or quantal) in nature, such as the percentages of cells, animals, or humans exhibiting phenotypes related to an adverse effect. Therefore, these endpoints usually have bounded and normalized dynamic ranges (and thus efficacy values). However, phenotypic features obtained from HCI studies are usually continuous in nature, thus having non-bounded, non-normalized, or even mixed-signed dynamic ranges. For example, a chemical may increase the intra-cellular level of a toxicity marker, while another chemical may reduce the level of the same marker (**Fig. 1b**). Similarly, a chemical may cause cell death and reduce cell size, while another chemical may create multi-nucleated cells and increase cell size. Therefore, it is not obvious which kind of metric should be used to quantify CRCs based on these types of phenotypic features, especially for the aim of selecting the most discriminative features to be used as toxicity endpoints.

Previous studies of high-throughput screening data based on cell viability or growth rates have found that potency metrics may yield unreliable results (Hafner et al. 2017), and efficacy metrics may reveal systematic variation in responses to perturbations that are not obvious under potency metrics

(Fallahi-Sichani et al. 2013). A recent HCI toxicological study evaluated three potency metrics when constructing a nephrotoxicity model (Sjögren et al. 2018). However, to the best of our knowledge, no previous systematic survey has been performed to understand how different types of CRC metrics may affect the identification of phenotypic features related to an adverse effect of interest. Most other HCI studies focused on the selection of the most informative phenotypic features (Su et al. 2016; Wink et al. 2018; Lee et al. 2018), or the most appropriate CRC fits for phenotypic features (Calhelha et al. 2017). A rigorous and systematic study will help to guide future analysis of HCI data for toxicological applications.

Our study aimed to answer three principal questions. First, how similar is the information that potency and efficacy metrics provide on the cellular effects of reference chemicals with or without known adverse effects? If the metrics are highly correlated, they would lead to the same discriminative endpoints and thus no further analysis is needed. Second, do potency or efficacy metrics help us to identify phenotypic features that yield more accurate classifiers, and are there different trends across different cell lines or feature types? In this study, supervised classification accuracy was used as a proxy indicator of discriminative features relevant to an adverse effect (**Fig. 1c**). Third, of the many possible potency- or efficacy-based metrics, what are the characteristics of metrics that produce the most accurate classifiers? By identifying optimum CRC metrics for discriminative feature selection, we can prevent informative phenotypic features from ending up 'hidden' behind obfuscating metrics, and recommend best practices for HCI data analysis that may be applicable to a broad range of feature types, datasets, and applications.

These optimum CRC metrics are not meant to replace other existing CRC metrics designed for estimating the points of departure (POD) of phenotypic endpoints, such as the benchmark dose or concentration (BMD or BMC) (Setzer and Hogan 2012; Gift et al. 2019). In fact, these metrics complement each other. BMC requires the selection of a benchmark response (BMR) level that is either generally considered to be "biologically significant" or, in cases where such a level is unknown or unclear, statistically different from the negative controls (Setzer and Hogan 2012). For most HCI features, the BMR levels are usually unknown, and highly dependent on the adverse effects of concern, feature types,

and biological or experimental variations in the collected data. For example, the same phenotypic endpoint may have different BMR levels for different adverse effects that may be associated with the endpoint. A uniform BMR threshold, such as one control standard deviation away from the mean, may not be suitable for all features under all circumstances. Therefore, to allow a systematic comparison of CRC metrics, our study only considered metrics that are fully defined based on raw or fitted values of CRCs, such as EC_Y and $R_{[X]}$. Once the discriminative endpoints for an adverse effect have been identified, the standard BMC or other relevant metrics may still be applied to the CRCs of these endpoints for determining BMRs or PODs, while taking into considerations of the various aforementioned factors or uncertainty specific to the selected endpoints.

MATERIALS AND METHODS

HCI datasets

We analyzed four previous HCI datasets (Su et al. 2016; Lee et al. 2018), representing two human lung cell types (a bronchial epithelial cell line, BEAS-2B; and an alveolar epithelial cell line, A549) or two human kidney cell types (a proximal tubule epithelial cell line, HK-2; and primary human proximal tubule cells, "HPTC") treated with 33 or 42 chemical compounds, respectively (**Fig. 1d**). The total number of chemicals assessed exceeds those of other contemporary HCI *in vitro* toxicological studies (Grimm et al. 2017; Sjögren et al. 2018; Delp et al. 2019), and their constituent moeities cover a broad area of toxicological relevance, as evinced by the spread of their chemical structure space coverage when compared against all the 8,795 chemicals from the United States Tox21 library (**Supplementary Fig. S2**) (Richard et al. 2016). The datasets contain phenotypic feature measurements obtained from a wide range of chemical concentrations: 0-2,000 μM for lung and 0-2,000 μg/ml for kidney cells. Each chemical was tested in seven discrete concentrations over these concentration ranges. Furthermore, the *in vivo* toxicity or non-toxicity of these chemicals are known and annotated based on expert review of the literature. For example, paraquat is annotated as pulmonotoxic due to studies reporting human *in vivo* pulmonary edema and fibrosis following accidental ingestion (Smith and Heath 1974; Dinis-Oliveira et al. 2008); ketoconazole is annotated as non-pulmonotoxic because clinical trials or post-marketing surveillance

reported liver damage in humans but no lung damage (Sugar et al. 1987); details of all annotation sources exist in the previous studies (Su et al. 2016; Lee et al. 2018). For the lung cell data sets, 13 chemicals were annotated as pulmonotoxic and the remaining 20 as non-pulmonotoxic; for the kidney cell data sets, 23 were annotated as nephrotoxic and 19 annotated as non-nephrotoxic. Finally, all four datasets are based on four similar fluorescent markers, namely (1) 4′,6-diamidino-2-phenylindole (DAPI) or Hoechst, staining DNA; (2) phalloidin, staining the cytoskeletal actin filaments; (3) antibodies specific to phosphorylated histone 2AX (hereafter referred to as "γH2AX"), which is implicated in DNA damage response (Rogakou et al. 1998); and (4) a whole cell stain for the full cellular region. These four datasets provide similar types of phenotypic features based on these four markers, allowing us to systematically compare the performances of CRC metrics for the same feature types across different cell lines, and determine the generality of the observed trends.

Phenotypic feature types

In HCI images, cells are located and oriented arbitrarily with respect to the field of view. The phenotypic features used to describe the cells must therefore be invariant under translations or rotations of the images. The four datasets that we used contain either 129 or 166 invariant phenotypic features (**Fig. 1d**). The complete list of phenotypic features used in these previous studies and their definitions can be found in **Supplementary Table S1** and **Supplementary Methods**. In the original studies, these features were measured using the cellXpress software (v1.4.3; Bioinformatics Institute, Singapore) (Laksameethanasan et al. 2013). The features can be divided into six types. Morphology features are cellular shape properties, such as cell size and aspect ratios, derived from the binary cellular or nuclear masks obtained from cell and nuclear segmentations, respectively. Aspect ratios are an example of a feature with an unbounded range; unlike cell size, there is no theoretical lower (or upper) limit to their divergence from a control population. Intensity features summarize the staining levels of fluorescent markers (**HCI Datasets**) at the whole-cell level or in different subcellular regions. Most intensity features have mixed-signed dynamic ranges (**Fig. 1b** and **e**), because a chemical may increase or decrease the expressions of the proteins or other biomolecules labelled by these markers. Intensity ratio features are

the ratios of the staining levels of pairs of different markers at the same subcellular regions, or the same markers at two different subcellular regions. Correlation features measure the spatial or pixel correlations between marker pairs in the obtained cellular images, indicating possible subcellular co-localizations of the markers. These can be assessed either as correlation coefficients (quantifying markers pairs' co-occurrence in the same areas) or as cross-correlations (quantifying marker pairs' co-occurrence in similar patterns). Texture features summarize a marker's spatial occurrence patterns as defined in Haralick's original paper (Haralick et al. 1973). Finally, cell count is the number of identified cells. The feature is usually expressed as a proportion of the cell count of a control experiment, and its range has a definite lower bound, i.e. 0 cells.

CRC fitting

Each of the HCI datasets included phenotypic feature values from four replicates. We first calculated the median response value from the four replicates, and then the $\log_2$-ratio of this median value with respect to the median value of the corresponding solvent controls. The resulting quantity represents an experimentally-determined average response to the chemical. Each chemical was tested in seven discrete concentrations (X), resulting in a set of seven averaged response values. In cases where chemicals induced high cell-death rates, reliable average response values cannot be computed. Therefore, treatments yielding median cell counts < 15% that of the solvent controls had their feature responses recorded as "NA".

We fit three different CRC models to the set of average response values for each feature and chemical (**Fig. 1b**):

Model A: $$f(x) = \alpha - \frac{\alpha}{1+exp(\beta(\log x - \log \gamma))} \quad (1)$$

Model B: $$f(x) = \frac{\alpha'}{1+exp\left(\beta'(\log x - \log \gamma')\right)} \quad (2)$$

and

Model C: $$f(x) = 0 \quad (3)$$

where α,β,γ,α',β', and γ' are all empirical parameters determined via least-squares-error minimization. The average responses for the cell count feature were not $\log_2$-transformed prior to curve-fitting, which necessitates a modification to Model C:

$$\text{Model C':} \quad f(x) = 1 \tag{4}$$

For each chemical and phenotypic feature combination, we selected the best fitted CRC model according to the Akiake Information Criterion:

$$\text{AIC} = 2\left(D - log\left(\frac{\sum(\varepsilon_i)^2}{m}\right)\right) \tag{5}$$

where D is the number of degrees of freedom in the model, ε_i is the residual error for data point i, and m is the number of data points (which in our case will usually be equal to the number of experimentally tested concentrations). The model with the lowest AIC is interpreted as exhibiting the best compromise between model complexity and goodness of fit.

Area under the curve

For consistency, the concentration range for which the area under the curve is computed must remain the same for all CRCs (Pozdeyev et al. 2016), so in our quantification this metric is defined by an area bound by the chords [X] = 31 μM; [X] = 2,000 μM; the CRC; and the response value at control (i.e. *f(x)* = 0 for Models A, B, C; *f(x)* = 1 for Model C'). For a Model B CRC, the AUC is then the area 'above' the curve; in these cases we assign the AUC a negative valence in order that the metric can retain the directionality information contained in its efficacy component. The AUC values were computed in $\log_{10}$-space of concentration and $\log_2$-space of response (except for the cell count feature, where we used linear space for response) via the trapezium rule at seven log-equidistant concentration intervals (Huang and Pang 2012).

Quality control

Some highly cytotoxic chemicals may yield multiple NA values such that there are fewer than four concentrations with finite response values. In these cases, Model A and Model B CRCs cannot be fit, so

we designated the chemical as “No Cell” (or NC). No further phenotypic analysis is performed for these chemicals.

Noisy experimental results may yield CRCs which extrapolate extremely large EC_Y values. Thus, we limit all EC_Y values to a maximum of 99,999 μM. Across all datasets, approximately 14% of Model A or Model B CRCs gave EC_{50} values that hit this limit. Analogous limiting conditions were unnecessary for the $R_{[X]}$ values because they are constrained to the experimental test concentrations ≤2,000 μM. Model C and Model C’ describe constant responses that are invariant to the tested chemical concentrations. We assign their potency metric values (EC_{10} to EC_{90}) to the same maximum limit of 99,999 μM. Across all datasets, approximately 38% of CRC were Model C or C’.

Supervised toxicity classification

For each phenotypic feature, the CRC metric values for all the chemicals were linearly normalized to a [-1, 1] range before a linear L2-regularized L2-loss support vector machine was trained to classify the data (Fan et al. 2008). We used linear SVM classifiers because they produce continuous decision boundaries that are easier to interpret in a biological context than the discontinuous decision boundaries that may be produced by more complex kernels.

For each dataset, phenotypic feature, and CRC metric combination, we trained a two-stage cascade classifier to assess the chemicals according to two annotated classes: “positive” for chemicals in the nephrotoxic and pulmonotoxic classes, “negative” for those in the non-nephrotoxic and non-pulmonotoxic classes. The first stage of the cascade assigns all the NC chemicals to “positive”. In the second stage, we used a stratified 10-fold cross validation procedure (Su et al. 2016) to assemble training and test datasets from the CRC metric values for the remaining chemicals. The proportion of annotated toxic chemicals that are correctly assigned to the positive class gives a classifier’s sensitivity; the proportion correctly assigned to the negative class gives a classifier’s specificity. The mean of specificity and sensitivity gives the balanced accuracy of toxicity classification (BAC).

Chemical Structure Space

Chemical structure space comparison with the Tox21 chemicals database (U.S. EPA 2013) involved visualization via t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton 2008). SMILES were used to generate a chemical space distance matrix using the `smiles2sdf()` and `sdf2ap()` routines from the "chemmineR" library (v3.36.0). Approximately 2% of the Tox21 chemicals had invalid SMILES (e.g. non-stoichiometric, polymeric); after data reduction there were 8,599 chemicals which could be incorporated into the t-SNE plot. For plotting we used the `Rtnse()` routine from the "Rtsne" library (v0.15) with perplexity = 6 and theta = 0.4, other settings default (**Supplementary Fig. S2**).

Analysis software

All analyses were conducted under the R environment (v.3.6.3). We used the `drm()` and `predict()` functions of the "drc" package (v 3.0.1) for model fitting and CRC-metric evaluations. The fitting procedures used are identical to those used by the authors of the original study for the BEAS-2B and A549 datasets (Lee et al. 2018). The original study for the HK-2 and HPTC datasets used a slightly different procedure (Su et al. 2016) to fit the CRCs, so the fitted metric values take slightly different values in the current study.

To assemble the SVM classifiers we used the `LiblineaR()` function of the "LIBLINEAR" package (v.2.10-8) (Fan et al. 2008), maintaining the default values for all the parameters except `cost`, which describes the penalty applied to misclassifications far from the decision boundary. During each fold of the cross validation, we automatically determine the optimum `cost` value using a grid search of 10^0, 10^1, 10^2, 10^3, 10^4, and 10^5. Computations for the 95th-percentile values (**Results**) were performed using the `quantile` function from the "stats" package (v.3.6.3) with `type` = 4 and other default parameters.

RESULTS

Phenotypic features are mostly mixed signed

We found that most of the phenotypic features (64%) from the four HCI datasets have "mixed signs", where at least 10% of the tested chemicals give Model A CRCs (i.e., increased response relative to controls) and at least 10% of other tested chemicals give Model B CRCs (i.e., decreased response) (**Fig. 1e**). Interestingly, the proportions of mixed-signed features are similar across the four datasets, despite the different tested chemicals. The direction of a CRC is likely to be indicative of the mechanism of action of the associated chemical; and a potency metric, such as EC_{50}, does not capture this information. Therefore, the existence of many mixed-signed features within these HCI datasets leads us to suspect that supervised feature selection based on potency metrics may not be ideal.

Most potency and efficacy metrics provide non-redundant information

We then determined to what extent potency and efficacy metrics may convey the same or redundant information about the cellular effects of a chemical. We also considered AUC, which contains information from both potency and efficacy metrics. For each best-fitted CRC model, we extracted 17 CRC metric values**.** They include the AUC; seven efficacy metrics, $R_{[31]}$, $R_{[62]}$, …, $R_{[2,000]}$, which report the response of the CRC at 31, 62, …, and 2,000 μM, respectively; nine potency metrics, EC_{10}, EC_{20}, …, EC_{90}, which report the concentrations required to elicit 10, 20, …, and 90% of the maximum response value of the CRC (**Fig. 2a** and **Methods**). The correlation of AUC values with those of other metrics is inevitable as area must increase with the height (efficacy) and width (potency) of the CRC. For the other CRC metrics, the relationship is less obvious. When comparing potency to efficacy metrics, redundancy is indicated by a negative correlation, because lower EC_Y values represent stronger potency, while lower $R_{[X]}$ values represent weaker efficacy. Mixed-signed phenotypic features (**Fig. 1b** and **e**) complicate the comparison because a strong efficacy is represented by the magnitude of the response, but not by the sign. Therefore we computed the Kendall's correlation coefficients (τ) between absolute response values $|R_{[X]}|$ and EC_Y for each feature. Metric-pairs with EC_Y values obtained from the constant responses (Models C and C') or extrapolated substantially beyond the measured data ranges were excluded from

this analysis. Overall, we found that the mean correlation coefficients between most of the evaluated potency and efficacy metrics have low to moderately negative values (τ = 0 to -0.50) (**Fig. 2b**). The global minima (τ = -0.534) occurs at $|R_{[125]}|$ and EC_{10}. There are moderately negative correlations between $|R_{[X]}|$ at low-to-intermediate concentrations (X) and EC_Y at low effect levels (Y). Similar trends were observed when each of the datasets were analyzed individually (**Supplementary Fig. S3**).

To better understand the observed weak correlations, we compared the values of $R_{[2,000]}$ and EC_{50} evaluated from the same CRCs for all the phenotypic features from the BEAS-2B dataset (**Fig. 2c**). $R_{[2,000]}$ was used to develop the predictive pulmonary toxicity models in the original study of the dataset (Lee et al. 2018). We found that features with low $R_{[2,000]}$ magnitudes across most of the chemicals ("low-effect features" in **Fig. 2c**) have varying EC_{50} values. Furthermore, for those features with increased $R_{[2,000]}$ values induced by certain chemicals, we often did not observe corresponding systematic changes in their EC_{50} values under the same chemicals (**Fig. 2d**). Different chemicals clusters can be identified with similar phenotypic responses across sets of features, and analogously we observe different phenotypic feature clusters with similar response values across sets of chemicals, possibly indicating shared mechanisms of action of these chemicals (**Fig 2c**). For example, towards the top of the left dendrogram there is a dendrite or cluster composed almost exclusively of actin-related intensity features (e.g. the mean actin intensity over the whole-cell region), all giving a similar profile of responses across all chemicals. And at the bottom of the dendrogram we identified a cluster of "low-effect features", predominantly texture and correlation feature types (e.g. the spatial correlation coefficient of γH2AX and actin intensities at the whole-cell region), which are collectively inactive for all of the lung datasets' chemicals. Most other features do not form clear clusters, suggesting they are not strongly correlated. Thus, a diverse group of features were being studied in our work. Interestingly, for the BEAS-2B dataset, most of the high-effect features are intensity or morphology features (**Fig. 2c** and **d**). Our results suggest that most of the tested potency and efficacy metrics convey non-redundant information, and one type of metric cannot be used to infer the value of the other. Therefore, using different CRC metrics as classifier inputs is likely to result in supervised chemical classification with dramatically different accuracies, and

thus different final endpoints being selected. This affirms the importance of identifying the most appropriate CRC metric before performing feature selection to identify a discriminative endpoint.

Efficacy metrics are more likely to yield top-performing optimal classifiers

Known toxic vs non-toxic chemicals might be better distinguished by the magnitude of their elicited biological responses (i.e. an efficacy metric), or by the concentration at which they elicit a response (i.e. a potency metric), or some hybrid of the two (i.e. the AUC metric). To determine which, we built 17 support vector machine (SVM) classifiers (Cortes and Vapnik 1995) per feature, one for each of the metrics, and estimated their balanced accuracies using a cross-validation procedure (**Fig. 3a**). The optimal CRC classifier for a feature is the one that yields the SVM with the highest BAC value. Features that are not informative for the specific adverse effects of interest will also have "optimal" classifiers, but such classifiers are liable to have low BAC values at ~50-60%. The identities of the metrics that contribute to such classifiers are not useful for our study, as we are interested only in the phenotypes which might be ranked highly by a feature selection method. Therefore, we categorized the results according to either feature sources (BEAS-2B, A549, HK-2, or HPTC datasets) or types (intensity, intensity ratio, correlation, texture, morphology features, or cell count), and only considered "top-performing" features with optimal classifier BACs in the top decile of all optimal classifiers associated to each feature category. Overall, we found that efficacy metrics consistently give the largest proportions of optimal toxicity classifiers for top-performing features in all categories (**Fig. 3b**). For efficacy metrics, we found that $R_{[2,000]}$ was usually over-represented (>1/17 metrics = 5.88%) and contributed to >29% of top-performing features' classifiers in all of the categories, except cell count. Most potency metrics were under-represented (<5.88%), and even taking all nine together they contributed to only ≤25% of top-performing features' classifiers for all feature types except cell count, despite constituting 53% (9/17) of the metrics. After $R_{[2,000]}$ and $R_{[1,000]}$, AUC was the third best-performing metric overall, providing the best BAC for 11.3% of top-performing features' classifiers. These results suggest that, for a feature of any type, using an efficacy metric for classifier training is more likely to yield a top-performing toxicity classifier than a potency metric. Therefore, if we do not possess any other requirement or prior knowledge about a feature's optimal CRC

metric, we should default to feature selection based on efficacy or AUC metrics, especially efficacy metrics at high concentration values.

Top efficacy-based classifiers are more accurate than top potency-based classifiers

To identify the CRC metric that is more likely to select the feature with the highest accuracy among all features from a given feature type, we first determined the median BAC value amongst all the top-performing features' optimal classifiers (equivalent to the 95^{th}-percentile BAC value amongst all the optimal classifiers) trained on a specific metric but based on different features from the same feature type. Then, the analysis was repeated for all the metrics, and the metric that provided the top-performing features' classifier with the maximum median BAC was identified. For all intensity features, we found that the maximum median BACs are associated to top-performing features' classifiers trained on efficacy metrics in three of the four datasets (**Fig. 4a**). Then, we repeated the same analysis for all the six feature types. In many cases even the top-performing results for a metric give BACs in the range of 50-60%, implying that these metrics are poorly suited for toxicity discrimination and should be avoided when building a classifier. Meanwhile, the metrics that yield the globally optimal top-performing features' classifiers across all the feature types are consistent across all datasets, namely $R_{[2,000]}$ (**Fig. 4b**). Meanwhile the feature types that yield the globally optimal top-performing features are not consistent: intensity ratio features for BEAS-2B (BAC = 81.7%), pixel correlation features for A549 (81.2%), texture features for HK-2 (75.6%), and intensity ratio features for HPTC (74.7%) (**Fig. 4b**). Our results show that best efficacy-metric-based classifiers tend to have higher performances than the best potency-metric-based classifiers. Efficacy metrics at high concentration levels usually select features that provide globally optimum toxicity classifiers. Classifiers based on the AUC metric broadly perform better than those based on potency metrics and low-concentration efficacy metrics, but show lower BACs for top-performing features than high-concentration efficacy metrics. These trends are applicable to all the tested datasets and feature types.

The evaluation concentration of efficacy metrics correlates with accuracy

If high-concentration efficacy metrics generally identify more highly discriminative features, is this phenomenon due to a positive association between supervised classifier performance and the concentration at which an efficacy metric is evaluated? We found that the BACs of top-performing optimal classifiers based on efficacy metrics ($R_{[X]}$) show moderate to strong rank correlation to the concentrations (X) at which the metrics were evaluated, whereas the BACs of top-performing optimal classifiers based on potency metrics (EC_Y) show no or very little rank correlation to the effect levels (Y) at which the metrics were evaluated (**Fig. 4c** and **d**). Similar trends hold across all four data sets. Our results suggest that, for classifiers based on potency metrics, we cannot find general trends to guide the selection of optimum effect levels for these metrics. Therefore, one would need to compute and compare the performances of classifiers based on multiple potency metrics at different effect levels in order to identify the most discriminative features during feature selection. However, for classifiers based on efficacy metrics, higher concentration levels generally yield higher BACs, and thus should always be included in the analysis. Importantly, this also suggests that experiments at sufficiently high concentration levels will need to be performed to allow the training of highly accurate classifiers.

Fitted efficacy metrics provide more accurate classifications than raw feature averages

All the 17 metrics discussed so far are derived from CRCs fitted from data points measured at up to seven concentrations. If efficacy metrics for high concentrations tend to yield optimal classifiers, is it necessary to experimentally measure the feature values at low and/or intermediate concentrations? To investigate, we trained additional classifiers based on the averaged raw feature values at 2000 μM without any CRC fitting ("$Avr_{[2,000]}$" metric). For the HK-2 and HPTC nephrotoxicity datasets, $Avr_{[2,000]}$ data were not available so they were not used for this analysis. For the A549 and BEAS-2B pulmonotoxicity datasets, four of the 33 chemicals have no data at the highest concentration (due to solubility issues), so to facilitate a fair comparison we retrained and compared both the $R_{[2,000]}$ and $Avr_{[2,000]}$ classifiers on datasets of only 29 chemicals.

For both the BEAS-2B and A549 datasets, we identified the features that contribute to the five highest BAC results for $R_{[2,000]}$, and compared these BACs with the BACs of classifiers trained on the same features but based on $Avr_{[2,000]}$. We found that most of the top features see their BAC decrease, some by >10%, when $Avr_{[2,000]}$ was used (**Fig. 5a**). To better understand the cause of the decrease, we investigated in more detail the discrete experimentally measured values of one of these features, namely "the ratio between total γH2AX intensity at the chromosomal region over the whole-cell region", before CRC fitting. In BEAS-2B cells treated with nickel sulfate, this feature shows a near-monotonic increase from 31 to 1,000 μM, followed by an abrupt drop at 2,000 μM (**Fig 5b**), which may be an experimental artifact. However, $R_{[2,000]}$, which is based on a fitted CRC, is much closer to $Avr_{[500]}$ and $Avr_{[1,000]}$ than $Avr_{[2,000]}$ is. Using the supervised classifier trained on this feature, nickel sulfide is incorrectly classified as negative for pulmonotoxicity when $Avr_{[2,000]}$ is used, but is classified correctly as positive when $R_{[2,000]}$ is used. This example illustrates how CRC-fitted feature values are less susceptible to experimental outliers. Therefore, features described by fitted efficacy metrics at high concentrations should not be replaced by averaged raw feature values from the same concentrations. Measurements at multiple concentrations are still required to get a robust fit of the features' CRC, and more accurate estimations of the response values at high concentrations.

DISCUSSION

Our investigation on the optimum CRC metrics for supervised selection of discriminative phenotypic features for chemical effect assessment has shown that efficacy metrics ($R_{[X]}$) consistently provide classifiers with higher toxicity classification accuracy than potency metrics (EC_Y) (**Fig. 3b** and **4b**). For efficacy metrics, we also found that there are positive correlations between classification accuracy and the concentrations at which the metrics are determined. AUC contributes to more accurate classifiers than potency metrics and low-concentration efficacy metrics, but is not as accurate as the high-concentration efficacy metrics. We suspect that the inclusion of potency information in AUC does more harm than good for toxicity classification. These findings are consistent across different data sets and feature types.

Several factors may contribute to the positive correlations. First, most of the CRCs are fitted by log-logistic functions, which have very small response values ($R_{[X]} \approx 0$) at low concentrations. Thus, a low-concentration-based classifier is unlikely to be able to make clear distinction of these response values, which may lead to lower classification accuracies. Second, higher concentrations of chemical are likely to lead to larger magnitudes of phenotypic response, in turn improving the signal-to-noise ratio and consistency of phenotypic readouts, leading to higher classification accuracies. Third, chemicals may induce phenotypic changes that are more consistent to their adverse effects at higher concentrations. Regardless of the underlying reasons, our results suggest that efficacy metrics, especially at higher concentration values, provide the most useful information for the purpose of supervised selection of discriminative phenotypic endpoints for chemical hazard assessment. Cytotoxicity at high concentrations is unlikely to be a major reason for the performance of high-concentration efficacy metrics. The BEAS-2B dataset gives broadly the best BAC classifiers, but has very few toxic chemicals which induce cytotoxicity at the highest concentrations (Lee et al. 2018). Instead, we suspect that the main reason may be due to the differences between *in vitro* and *in vivo* toxicokinetics and microenvironments, such that higher *in vitro* concentrations may be needed to activate the biological pathways leading to the adverse effects.

The measurement of high-concentration responses poses several practical challenges. Chemicals may be insoluble or form aggregates at high nominal concentrations, making it difficult to experimentally achieve the desired actual concentrations. Furthermore, chemicals may be cytotoxic at high concentrations, to the extent that there are too few viable cells left to accurately perform phenotypic profiling. Possible solutions to the problems may include the use of solvents with higher solubility limits, or shorter exposure times for cells with the chemicals. Despite the difficulties in assessing chemicals at high concentrations, our results agree with several previous HCI studies that use measurements at similarly high concentration values (e.g. 3 mM or higher, or ~30 to 100× the human efficacious maximum serum concentrations, C_{max}) (O'Brien et al. 2006; Xu et al. 2008; Lin and Will 2012). Therefore, measurements at high concentrations are still recommended. The need of measuring high-concentration responses in *in vitro* cell-based toxicity models has been recently re-identified (Sjögren et al. 2018) and debated (Sjögren and Hornberg 2019; Zink 2019). Our study provides data-driven justifications for using such measurements in HCI.

As both efficacy and potency metric types are derived from the same CRCs, they may convey correlated information, but we found that the magnitude of the rank correlation (τ) between any pair of potency and efficacy metrics was always less than 0.6 (**Fig. 2b** and **Supplementary Fig. S3**). These correlations are largest between EC_{10} and $|R_{[X]}|$ evaluated at intermediate concentration levels. Qualitatively, EC_{10} may be used as an estimate of the concentration at which a CRC starts to deviate from the controls' response. If the EC_{10} for the CRC is higher than the concentration (X) at which an efficacy metric ($R_{[X]}$) is evaluated, then the curve has not deviated much from the controls at X and $R_{[X]}$ is most likely close to zero at X or lower concentrations. This would likely lead to a negative correlation coefficient between EC_{10} and $R_{[X]}$. This relationship also explains why the magnitude of the correlation coefficient decreases as the response percentile (Y) at which a potency metric is evaluated increases: the higher the percentile, the less probable it is that $R_{[X]} \approx 0$ for $X < EC_Y$. Our results show that most of the tested potency and efficacy metrics convey non-redundant information, and thus are likely to result in the selections of phenotypic features with very different classification accuracy levels.

Our results have important implications for the design of future HCI-based toxicological studies.

From a data processing perspective, among the top-performing features we found a strong correlation between classifier BAC and the concentration at which an efficacy metric is evaluated, so efficacy metrics at high concentrations are more likely to yield the most discriminative endpoints. Conversely, classifiers based on potency or low concentration efficacy metrics were found to give lower BAC and so should be avoided. Furthermore, no correlation was found between classifier BAC and the effect percentile at which a potency metric is evaluated, so finding optimal features based on chemical potency would require testing a range of potency metrics. From an experimental design perspective, BAC may be improved by including measurements at high concentrations. However, lower-concentration measurements should not be discarded, because efficacy metrics derived from CRCs fitted from multiple-concentration measurements yield more accurate classifiers than those derived from single high-concentration measurements. Our results may be broadly applicable to other cellular phenotypic datasets and the identification of optimum features for other adverse effects.

REFERENCES

Bougen-Zhukov N, Loh SY, Lee HK, Loo L-H (2017) Large-scale image-based screening and profiling of cellular phenotypes. Cytometry A 91:115–125. https://doi.org/10.1002/cyto.a.22909

Calhelha RC, Martinez MA, Prieto MA, Ferreira ICFR (2017) Mathematical models of cytotoxic effects in endpoint tumor cell line assays: critical assessment of the application of a single parametric value as a standard criterion to quantify the dose-response effects and new unexplored proposal formats. Analyst 142:4124–4141. https://doi.org/10.1039/c7an00782e

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

Delp J, Funke M, Rudolf F, et al (2019) Development of a neurotoxicity assay that is tuned to detect mitochondrial toxicants. Arch Toxicol 93:1585–1608. https://doi.org/10.1007/s00204-019-02473-y

Dinis-Oliveira RJ, Duarte JA, Sánchez-Navarro A, et al (2008) Paraquat poisonings: mechanisms of lung toxicity, clinical features, and treatment. Crit Rev Toxicol 38:13–71. https://doi.org/10.1080/10408440701669959

Fallahi-Sichani M, Honarnejad S, Heiser LM, et al (2013) Metrics other than potency reveal systematic variation in responses to cancer drugs. Nat Chem Biol 9:708-+. https://doi.org/10.1038/NCHEMBIO.1337

Fan R-E, Chang K-W, Hsieh C-J, et al (2008) LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res 9:1871–1874

Gift J, Davis JA, Blessinger T (2019) Benchmark Dose Software (BMDS), US EPA, https://www.epa.gov/sites/production/files/2018-19/documents/bmds_3.0_user_guide.pdf, accessed January 2019

Grimm FA, Iwata Y, Sirenko O, et al (2015) High-Content Assay Multiplexing for Toxicity Screening in Induced Pluripotent Stem Cell-Derived Cardiomyocytes and Hepatocytes. Assay Drug Dev Technol 13:529–546. https://doi.org/10.1089/adt.2015.659

Grimm FA, Sirenko O, Ryan KR, Iwata Y (2017) In vitro cardiotoxicity assessment of environmental chemicals using a organotypic human induced pluripotent stem cell-derived model. Toxicol Appl Pharmacol 322:. https://doi.org/10.1016/j.taap.2017.20.020

Hafner M, Niepel M, Sorger PK (2017) Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. Nat Biotechnol 35:500–502

Haralick R, Shanmugam K, Dinstein I (1973) Textural Features for Image Classification. IEEE Trans Syst Man Cybern SMC3:610–621. https://doi.org/10.1109/TSMC.1973.4309314

Huang S, Pang L (2012) Comparing statistical methods for quantifying drug sensitivity based on in vitro dose-response assays. Assay Drug Dev Technol 10:88–96. https://doi.org/10.1089/adt.2011.0388

Kleinstreuer NC, Yang J, Berg EL, et al (2014) Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. Nat Biotechnol 32:583–591. https://doi.org/10.1038/nbt.2914

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Laksameethanasan D, Tan RZ, Toh GW-L, Loo L-H (2013) cellXpress: a fast and user-friendly software platform for profiling cellular phenotypes. BMC Bioinformatics 14:S4. https://doi.org/10.1186/1471-2105-14-S16-S4

Lee J-YJ, Miller JA, Basu S, et al (2018) Building predictive in vitro pulmonary toxicity assays using high-throughput imaging and artificial intelligence. Arch Toxicol 92:2055–2075. https://doi.org/10.1007/s00204-018-2213-0

Lin Z, Will Y (2012) Evaluation of Drugs With Specific Organ Toxicities in Organ-Specific Cell Lines. Toxicol Sci 126:114–127. https://doi.org/10.1093/toxsci/kfr339

Loo L-H, Lin H-J, Steininger RJ, et al (2009) An approach for extensibly profiling the molecular states of cellular subpopulations. Nat Methods 6:759–765. https://doi.org/10.1038/nmeth.1375

Loo L-H, Wu LF, Altschuler SJ (2007) Image-based multivariate profiling of drug responses from single cells. Nat Methods 4:445–453. https://doi.org/10.1038/nmeth1032

O'Brien PJ, Irwin W, Diaz D, et al (2006) High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. Arch Toxicol 80:580–604. https://doi.org/10.1007/s00204-006-0091-3

Paul Friedman K, Gagne M, Loo L-H, et al (2020) Utility of In Vitro Bioactivity as a Lower Bound Estimate of In Vivo Adverse Effect Levels and in Risk-Based Prioritization. Toxicol Sci 173:202–225. https://doi.org/10.1093/toxsci/kfz201

Pozdeyev N, Yoo M, Mackie R, et al (2016) Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. Oncotarget 7:51619–51625. https://doi.org/10.18632/oncotarget.10010

Richard AM, Judson RS, Houck KA, et al (2016) ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol 29:1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135

Rogakou EP, Pilch DR, Orr AH, et al (1998) DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem 273:5858–5868. https://doi.org/10.1074/jbc.273.10.5858

Setzer RW, Hogan K (2012) Benchmark Dose Technical Guidance, US EPA, https://www.epa.gov/risk/benchmark-dose-technical-guidance, accessed January 2019

Sirenko O, Mitlo T, Hesley J, et al (2015) High-Content Assays for Characterizing the Viability and Morphology of 3D Cancer Spheroid Cultures. ASSAY Drug Dev Technol 13:402–414. https://doi.org/10.1089/adt.2015.655

Sjögren A-K, Breitholtz K, Ahlberg E, et al (2018) A novel multi-parametric high content screening assay in ciPTEC-OAT1 to predict drug-induced nephrotoxicity during drug discovery. Arch Toxicol 92:3175–3190. https://doi.org/10.1007/s00204-018-2284-y

Sjögren A-K, Hornberg JJ (2019) Compound selection and annotation to validate the predictivity of in vitro toxicity assays for use in drug discovery, in response to Commentary by Dr. Zink (Zink, D. Arch Toxicol (2018)). Arch Toxicol 93:225–226. https://doi.org/10.1007/s00204-018-2359-9

Slikker W, de Souza Lima TA, Archella D, et al (2018) Emerging technologies for food and drug safety. Regul Toxicol Pharmacol 98:115–128. https://doi.org/10.1016/j.yrtph.2018.07.013

Smith P, Heath D (1974) Paraquat lung: a reappraisal. Thorax 29:643–653. https://doi.org/10.1136/thx.29.6.643

Su R, Xiong S, Zink D, Loo L-H (2016) High-throughput imaging-based nephrotoxicity prediction for xenobiotics with diverse chemical structures. Arch Toxicol 90:2793–2808. https://doi.org/10.1007/s00204-015-1638-y

Sugar AM, Alsip SG, Galgiani JN, et al (1987) Pharmacology and toxicity of high-dose ketoconazole. Antimicrob Agents Chemother 31:1874–1878. https://doi.org/10.1128/aac.31.12.1874

Thomas RS, Bahadori T, Buckley TJ, et al (2019) The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. Toxicol Sci 169:317–332. https://doi.org/10.1093/toxsci/kfz058

U.S. EPA (2013) ToxCast Data Generation: Chemical Lists, ToxCast_Generic_Chemicals_2013_12_10.xlsx. U.S. EPA, https://www.epa.gov/chemical-research/toxcast-data-generation-chemical-lists, accessed July 2016

van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. J Mach Learn Res 9:2579–2605

van der Ven LTM, Rorije E, Sprong RC, et al (2020) A Case Study with Triazole Fungicides to Explore Practical Application of Next-Generation Hazard Assessment Methods for Human Health. Chem Res Toxicol 33:834–848. https://doi.org/10.1021/acs.chemrestox.9b00484

Wink S, Hiemstra SW, Huppelschoten S, et al (2018) Dynamic imaging of adaptive stress response pathway activation for prediction of drug induced liver injury. Arch Toxicol 92:1797–1814. https://doi.org/10.1007/s00204-018-2178-z

Xu JJ, Henstock PV, Dunn MC, et al (2008) Cellular imaging predictions of clinical drug-induced liver injury. Toxicol Sci 105:97–105. https://doi.org/10.1093/toxsci/kfn109

Zink D (2019) Comment on Sjögren et al. (2018) A novel multi-parametric high-content screening assay in ciPTEC-OAT1 to predict drug-induced nephrotoxicity in drug discovery. Arch Toxicol 92(10):3175–3190. Arch Toxicol 93:221–223. https://doi.org/10.1007/s00204-018-2327-4

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

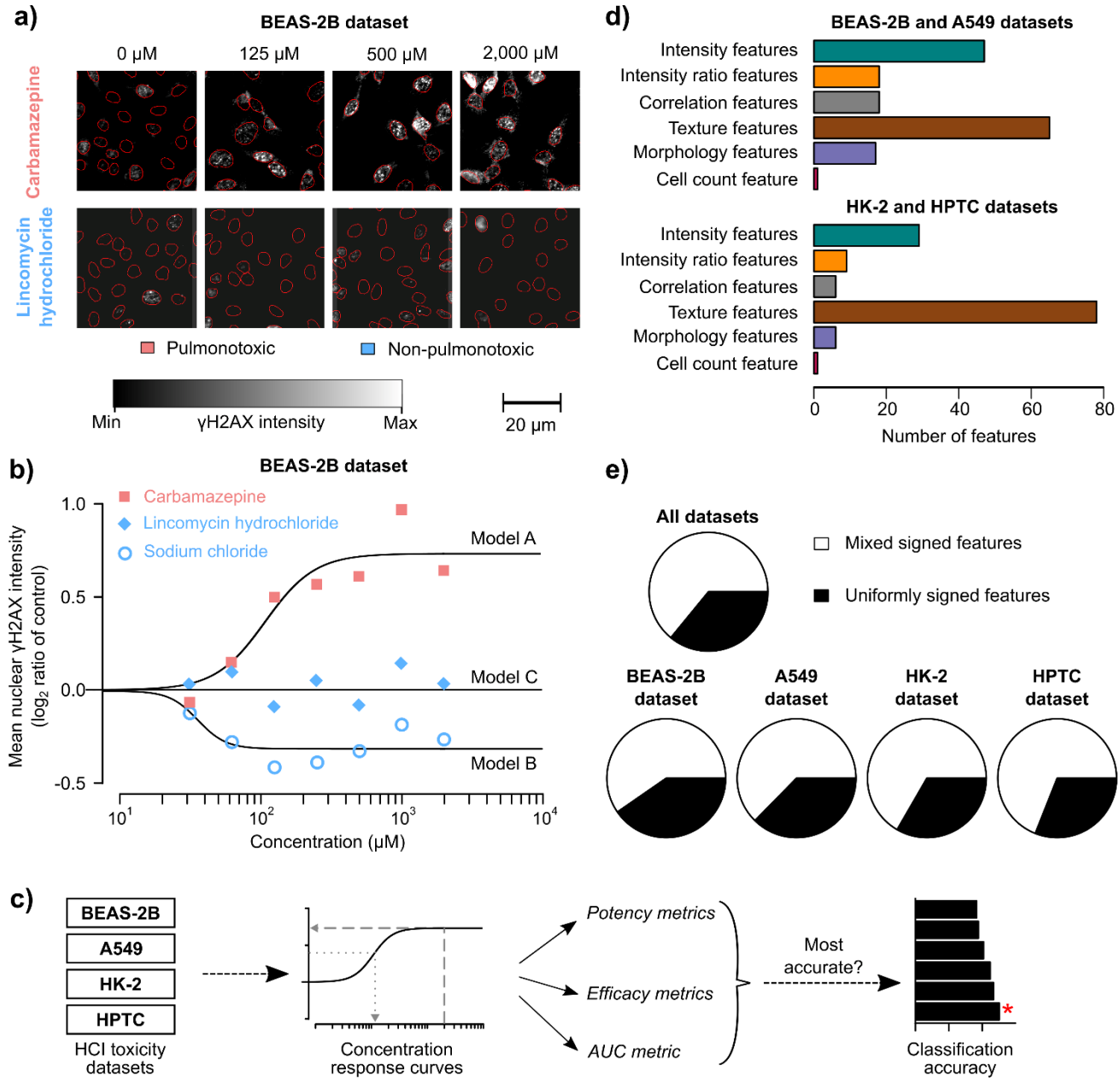
We thank members of the Loo Lab for support and discussions, and the National Supercomputing Centre Singapore (NSCC) for providing access to the high-performance computing system. The work was supported by a grant from the Biomedical Research Council (BMRC) Industry Alignment Fund - Pre-positioning Programme (H18/01/a0/B14), Agency for Science, Technology and Research (A*STAR), Singapore.

AUTHOR CONTRIBUTIONS

LLH conceived the study. JAM performed the computational analyses. JAM and LLH wrote the manuscript.

FIGURES

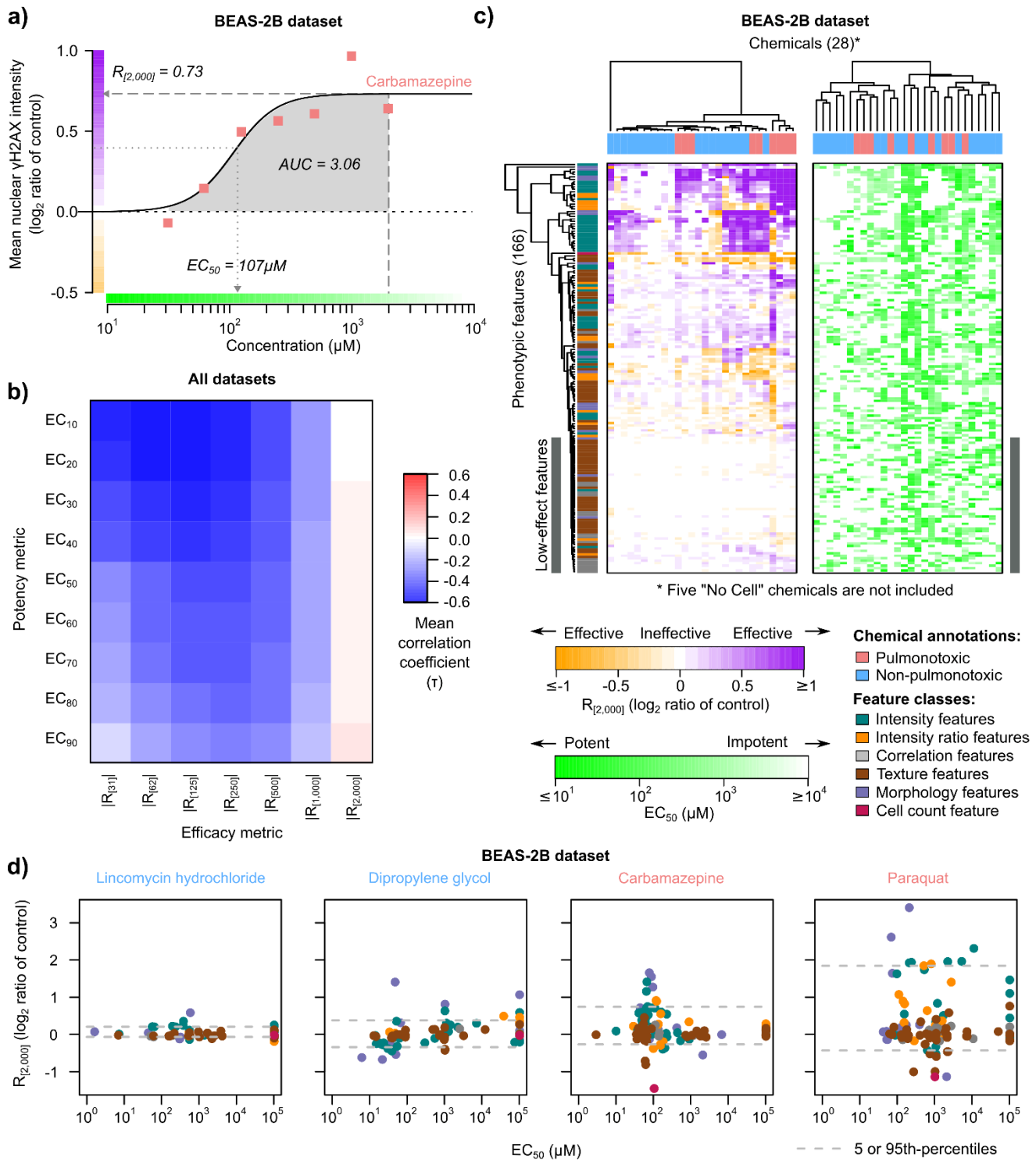
Fig. 1: Phenotypic features are mostly mixed signed



a) Exemplary immunofluorescence images showing γH2AX stains in BEAS-2B cells treated with increasing concentrations of carbamazepine (top) and lincomycin hydrochloride (bottom) (red lines = boundaries of automatically segmented nuclear regions.) **b)** Examples of three different fitted concentration response curve (CRC) models (Models A, B, and C) obtained from BEAS-2B cells treated with carbamazepine, sodium chloride, and lincomycin hydrochloride, respectively

(squares/circles/diamonds = medians of the measured feature values obtained from 4 replicates; curves = fitted CRCs based on the median values.) **c)** Schematic showing the study workflow: *in vitro* cellular response data from four previous HCI toxicological datasets were used to fit CRCs, and derive CRC metrics. Classifiers trained with these CRC metrics were evaluated for classification accuracy based on the known *in vivo* toxicities of the reference chemicals in these datasets, permitting the identification of the most discriminative features (red asterisk). **d)** Bar charts showing the numbers of different phenotypic features in the four HCI datasets that we used. **e)** Pie charts showing the proportion of mixed- or uniformly-signed features in the four datasets. “Mixed-signed” features are those with >10% of the tested chemicals with Model-A CRCs and >10% of the tested chemicals with Model-B CRCs. All other features are “uniformly-signed”.

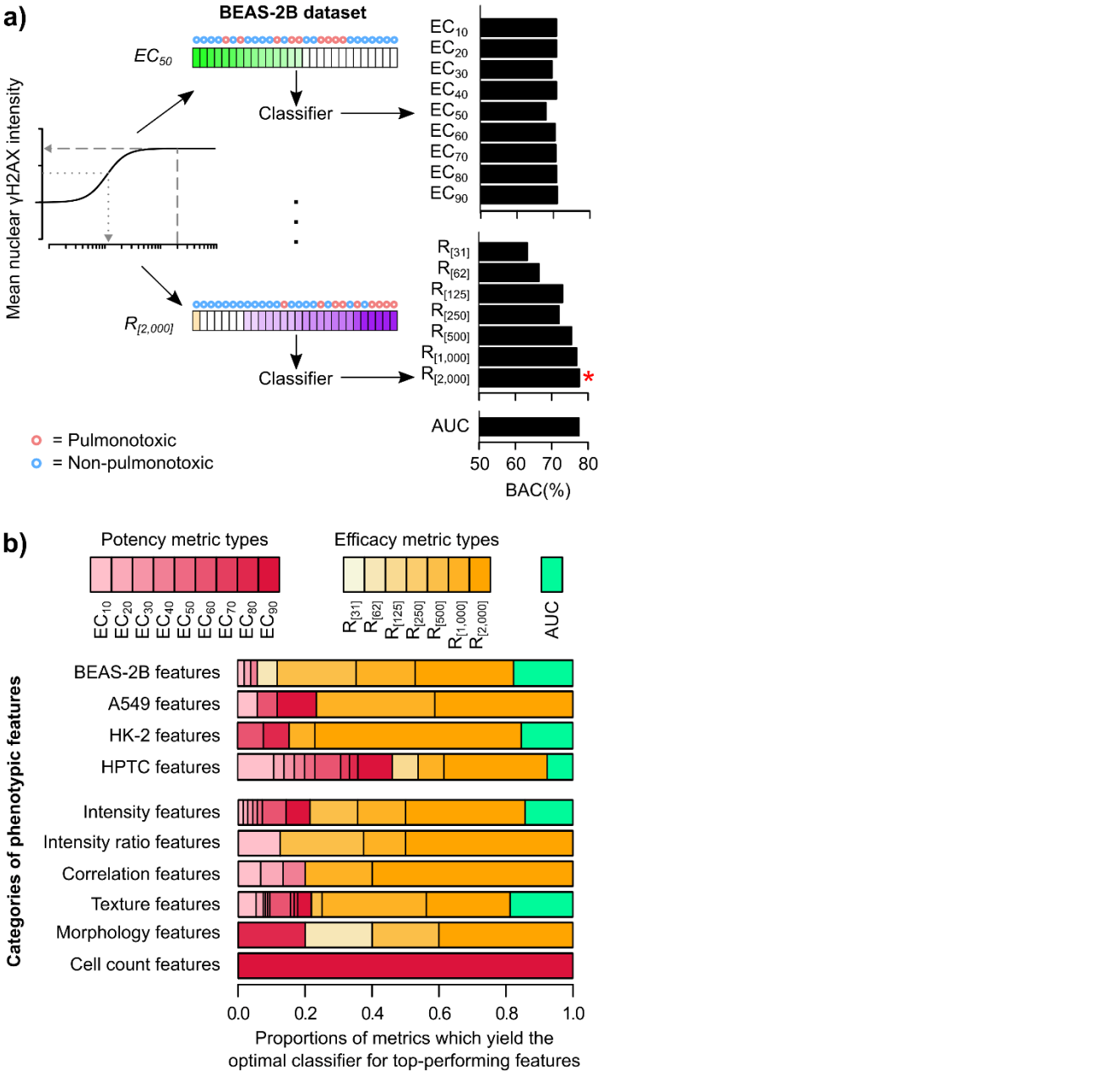
Fig. 2: Potency and efficacy metrics convey non-redundant information



a) Schematic showing an example of how an efficacy, potency, or AUC metric is determined from a CRC (squares = means of the measured feature values obtained from 4 replicates; curve = fitted CRC based on the median values; EC_{50} = concentration at which the fitted response achieves 50% of its maximum

level; $R_{[2,000]}$ = value of the fitted response at 2,000 μM). **b)** Heatmap showing the mean Kendall's rank correlation coefficients (τ) between potency and efficacy metrics, averaged over all the measured phenotypic features from all four datasets. Extremely large potency values (**Methods**) are not included. The heatmaps for the individual datasets are shown in **Supplementary Fig. S3**. **c)** Heatmaps showing the $R_{[2,000]}$ (left) and the EC_{50} (right) values of the 28 chemicals (columns) based on 166 phenotypic features (rows) from the BEAS-2B dataset. (Row dendrogram = a hierarchical clustering of the $R_{[2,000]}$ values; column dendrograms = hierarchical clusterings of the $R_{[2,000]}$ or EC_{50} values.) **d)** Scatter plots showing the $R_{[2,000]}$ and EC_{50} values (points) of four exemplary chemicals with increasing maximum $R_{[2,000]}$ values (left to right) from the BEAS-2B dataset. The points are color-coded according to the types of features on which the underlying CRCs are based, as in **Fig. 2c** (dash lines = 5 or 95^{th}-percentiles of all the $R_{[2,000]}$ values).

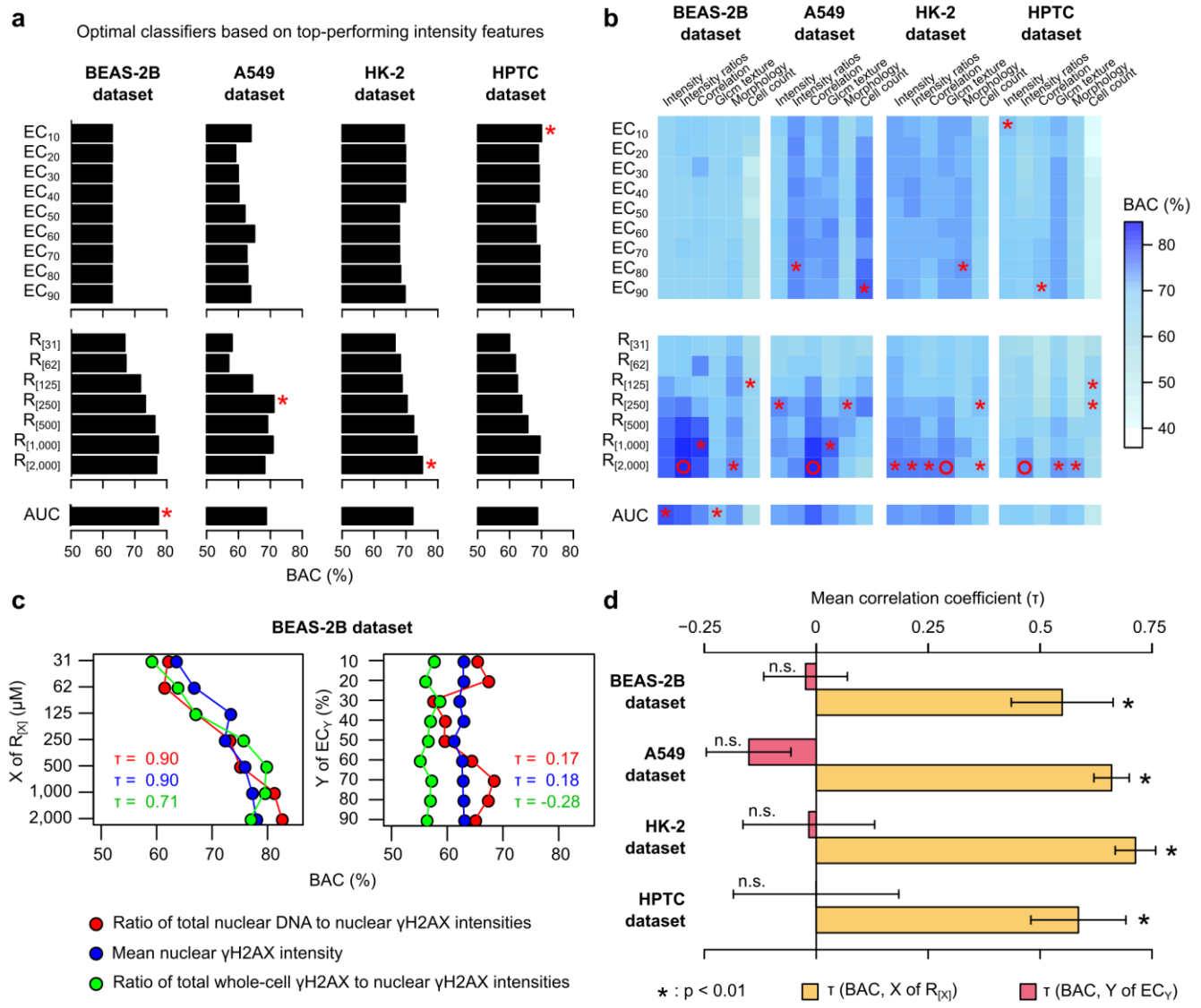
Fig. 3: Efficacy metrics are more likely to yield top-performing optimal classifiers



a) Schematic showing an example of how the balanced accuracies (BAC) of classifiers trained on different CRC metrics based on the same phenotypic feature (namely, mean nuclear γH2AX intensity) are determined using a 10-fold cross validation procedure (**Methods**). The annotations of the chemicals (red = toxic, blue = non-toxic) are used to determine the BAC values. In this example, $R_{[2,000]}$ (*) provides the optimal classifier for the shown feature. **b)** Stacked barcharts showing the distributions of CRC metrics (pinks = potency metrics; oranges = efficacy metrics; turquoise = AUC) that maximise classification

accuracy for different categories of top-performing phenotypic features. For each feature category, these classifiers have BAC values within the top decile (or 90th percentile) among all the optimal classifiers based on each feature from that category. The number of features for each category is not equal, and thus the number of top-performing optimal classifiers is also not equal.

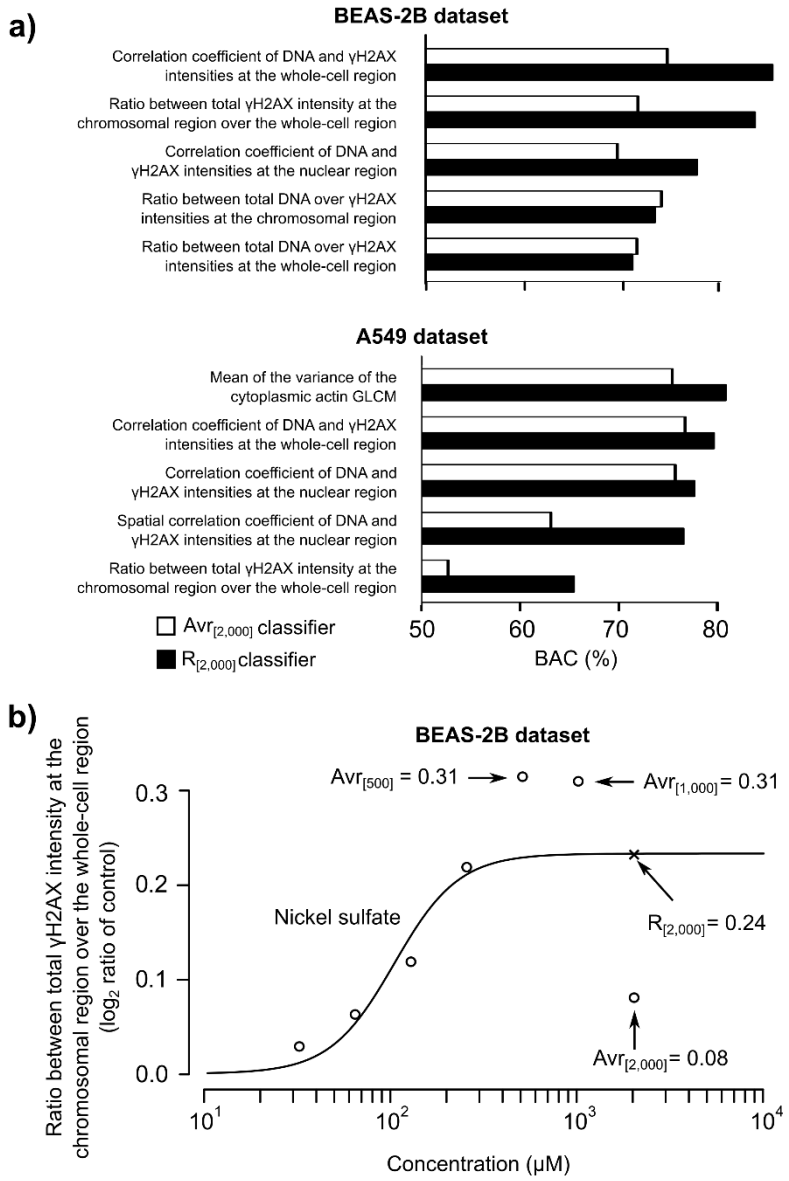
Fig. 4: Top efficacy-based classifiers are more accurate than top potency-based classifiers



a) Barcharts showing the median BACs of all the top-performing features' optimal classifiers based on intensity features, and categorized according to the CRC metrics used (* = metrics that provide classifiers with the highest median BACs). **b)** Heatmaps showing the median BACs of all the top-performing features' optimal classifiers based on different feature types, and categorized according to the CRC metrics used (* = metrics that provide classifiers with the highest median BACs across all the metrics based on specific feature types; O = metrics that provide classifiers with the globally highest median BACs for the dataset across all the metrics and feature types). **c)** Plots showing the relationships between the BACs of classifiers trained on potency ($R_{[X]}$) or efficacy (EC_Y) metrics and the concentration (X) or effect-percentile (Y) in which the metrics are determined, respectively. The relationships for three

examples of features that provide top-performing classifiers from the BEAS-2B dataset are shown, and quantified using the Kendall's correlation coefficients (τ). **d)** Bar charts showing the mean τ over all the features with top-performing classifiers for different datasets (means ± standard error; significance computed via two-tailed one-sample t-test, null hypothesis = τ is zero).

Fig. 5: Fitted efficacy metrics provide more accurate classifications than raw feature averages



a) Barcharts showing the BACs of classifiers trained on five phenotypic features that provide classifiers with the highest BAC values at 2,000 μM (black bars = classifiers based on the fitted feature values, i.e., $R_{[2,000]}$; white bars = classifiers based on the average raw feature values, i.e., $Avr_{[2,000]}$). The results shown are for the reduced set of 29 chemicals which had both $R_{[2,000]}$ and $Avr_{[2,000]}$ data. **b)** Example demonstrating how different feature values may be returned by $Avr_{[x]}$ and $R_{[x]}$ from the same CRC (circles = medians of the measured raw feature values obtained from 4 replicates, also correspond to $Avr_{[x]}$ values; curve = fitted CRC based on the median values; cross = $R_{[2,000]}$ value evaluated from the fitted CRC.)

SUPPLEMENTARY MATERIALS

Optimum concentration-response-curve metrics for supervised selection of discriminative cellular phenotypic endpoints for chemical hazard assessment

James Alastair Miller, Lit-Hsin Loo*

Innovations in Food and Chemical Safety Programme and Bioinformatics Institute, Agency for Science, Technology, and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix, Singapore 138671, Singapore

* To whom correspondence should be addressed: loolh@bii.a-star.edu.sg (ORCID:0000-0001-6303-9840)

Supplementary Methods

Definitions of GLCM features

A Grey-Level Co-occurrence Matrix (GLCM) describes how often a pixel with intensity level *j* occurs adjacent to a pixel with intensity level *i*.

$i, j = \text{Intensity levels}$ $\quad d = \text{Pixel separation distance}$

$L = \text{Total number of intensity levels}$ $\quad \theta = \text{Pixel separation direction}$

$M(i, j, d, \theta) = \text{GLCM matrix}$

Important derived properties of a GLCM for a given direction and distance include:

The probability distribution matrix of a co-occurrence matrix $M(i, j, d, \theta)$ is used in the computation of many other properties, and is given by:

$$p(i, j, d, \theta) = \frac{M(i, j, d, \theta)}{\sum_{i}^{L}\sum_{j}^{L} M(i, j, d, \theta)}$$

With this we can compute other intermediate properties:

Mean of the probability distribution matrix $p(i, j, d, \theta) = \mu$

Entropy of $p(i, j, d, \theta)$ = $H = -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i, j, d, \theta)\log 2[p(i, j, d, \theta)]$

Marginal row probabilities $p_x(i) = \sum_{j=1}^{L} p(i, j, d, \theta)$

Mean of $p_x(i) = \mu_x$, standard deviation of $p_x(i) = \sigma_x$, row entropy $HX = -\sum_{i=1}^{L} p_x(i)\log 2[p_x(i)]$

Marginal column probabilities $p_y(i) = \sum_{i=1}^{L} p(i, j, d, \theta)$

Mean of $p_y(i) = \mu_y$, standard deviation of $p_y(i) = \sigma_y$, col entropy $HY = -\sum_{i=1}^{L} p_y(i)\log 2[p_y(i)]$

$p_{x+y}(k) = \sum_{j=1}^{L}\sum_{i=1}^{L} p(i, j, d, \theta)$ where $k = i + j = 2,3,...,2L$

$p_{x-y}(k) = \sum_{j=1}^{L}\sum_{i=1}^{L} p(i,j,d,\theta)$ where $k = |i-j| = 0,1,...,L-1$; the variance of this quantity is the difference variance of the GLCM.

$$HXY1 = -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j,d,\theta)\log\left[p_x(i)p_y(j)\right]$$

$$HXY2 = -\sum_{i=1}^{L}\sum_{j=1}^{L} p_x(i)p_y(j)\log\left[p_x(i)p_y(j)\right]$$

And from these we can compute features used in the study (**Supplementary Table S1**):

GLCM contrast = $\sum_{i}^{L}\sum_{j}^{L}|i-j|p(i,j,d,\theta)$

GLCM correlation = $\dfrac{\sum_{i}^{L}\sum_{j}^{L} ij\,p(i,j,d,\theta) - \mu_i(i)\mu_j(j)}{\sigma_x(i)\sigma_y(j)}$

GLCM difference entropy = $\sum_{i=0}^{L-1} p_{x-y}(i)\log 2\left(p_{x-y}(i)\right)$

GLCM sum entropy = $SE = -\sum_{i=2}^{2L} p_{x+y}(i)\log 2\left(p_{x+y}(i)\right)$

GLCM sum variance = $\sum_{i=2}^{2L}(i-SE)^2\,p_{x+y}(i)$

Sum average of the GLCM = $\sum_{k=2}^{2L} ip_{x+y}(i)$

GLCM inverse difference moment = $\sum_{i}^{L}\sum_{j}^{L}\dfrac{p(i,j,d,\theta)}{1+|i-j|^2}$

GLCM informational measure of correlation 1 = $\dfrac{H - HXY1}{\max(HX, HY)}$

GLCM informational measure of correlation 2 = $\sqrt{1-\exp\left(-2\left[HXY2-H\right]\right)}$

Angular second moment of the GLCM = $\sum_{i}^{L}\sum_{j}^{L} p(i,j,d,\theta)^2$

For more GLCM derivations see (Haralick et al., 1973; Zhao et al., 2016).

Literature search parameters for metric types

To estimate the prevalence with which different metric types are used in contemporary research, we performed web-based literature searches within the Web of Science citation indexing service. All searches were performed with Web of Science's "Advanced Search" function on 14th October 2019 (www.webofknowledge.com). First, we searched for recent papers with abstracts referencing "concentration response curves" or related terms as a baseline using the Boolean search parameters:

```
TS=("Dose response profile" OR "Concentration response profile" OR "Dose response curve" OR "Concentration response curve" OR "Dose response profiles" OR "Concentration response profiles" OR "Dose response curves" OR "Concentration response curves")

AND

PY=(2000-2019)
```

Then we searched for papers with abstracts referencing both CRCs and terms related to potency or efficacy metrics, as listed in the *IUPAC Glossary of Terms Used in Toxicology* (Duffus et al., n.d.):

```
TS=("Dose response profile" OR "Concentration response profile" OR "Dose response curve" OR "Concentration response curve" OR "Dose response profiles" OR "Concentration response profiles" OR "Dose response curves" OR "Concentration response curves")

AND
PY=(2000-2019)
AND
TS=("effective concentration" OR ECn OR EC*0 OR AC*0 OR "effective dose" OR ED*0 OR "inhibitory concentration" OR IC*0 OR ICn OR "inhibitory dose" OR ID*0 OR IDn OR
```

"lethal concentration" OR LCmin OR LC*0 OR "lethal dose" OR LDmin OR LD*0 OR "observed effect level" LOEL OR NOEL "observed adverse effect level" OR LOAEL OR NOAEL OR "No effect level" OR "No effect dose" OR "No effect concentration" OR "NEL" OR "No response level" OR "No response dose" OR "No response concentration" OR "adverse response level" OR "adverse response dose" OR "adverse response concentration" OR "SNARL" OR "Maximum allowable concentration" OR "Maximum allowable dose" OR "Maximum contaminant level" OR "Maximum exposure limit" OR "Maximum permissible concentration" OR "Maximum permissible dose" OR "Maximum tolerable concentration" OR "Maximum tolerable dose" OR "Maximum tolerable exposure" OR "Median concentration narcotic" OR "MCn" OR "Median dose narcotic" OR "Mdn" OR “potenc*” OR “potent” OR “Benchmark dose” OR “BMD” OR “Benchmark concentration” OR “BMC”)

TS=("Dose response profile" OR "Concentration response profile" OR "Dose response curve" OR "Concentration response curve" OR "Dose response profiles" OR "Concentration response profiles" OR "Dose response curves" OR "Concentration response curves")

AND

PY=(2000-2019)

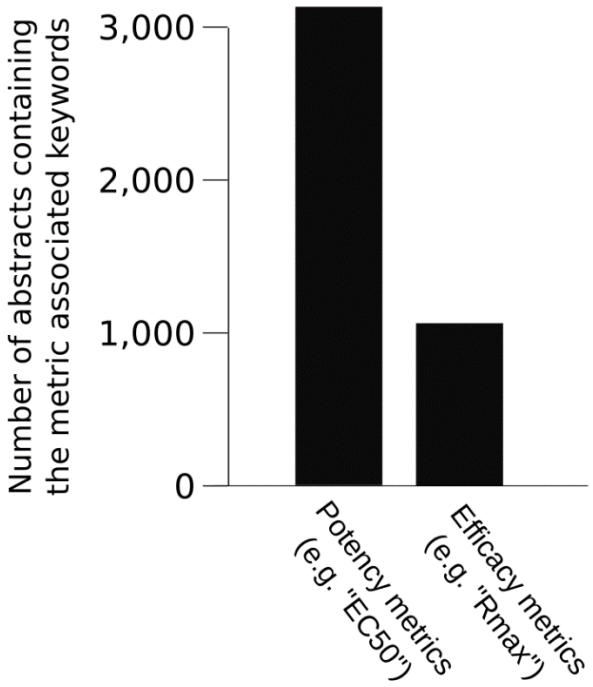
AND

TS=(“efficac*” OR “R max” OR “R*0” OR “Emax”)

There were 10,276 hits for the generic “concentration response curve” search, 3,122 hits for potency-metric-related terms, and 1,043 hits for efficacy-metric-related terms (**Supplementary Fig. S1**).

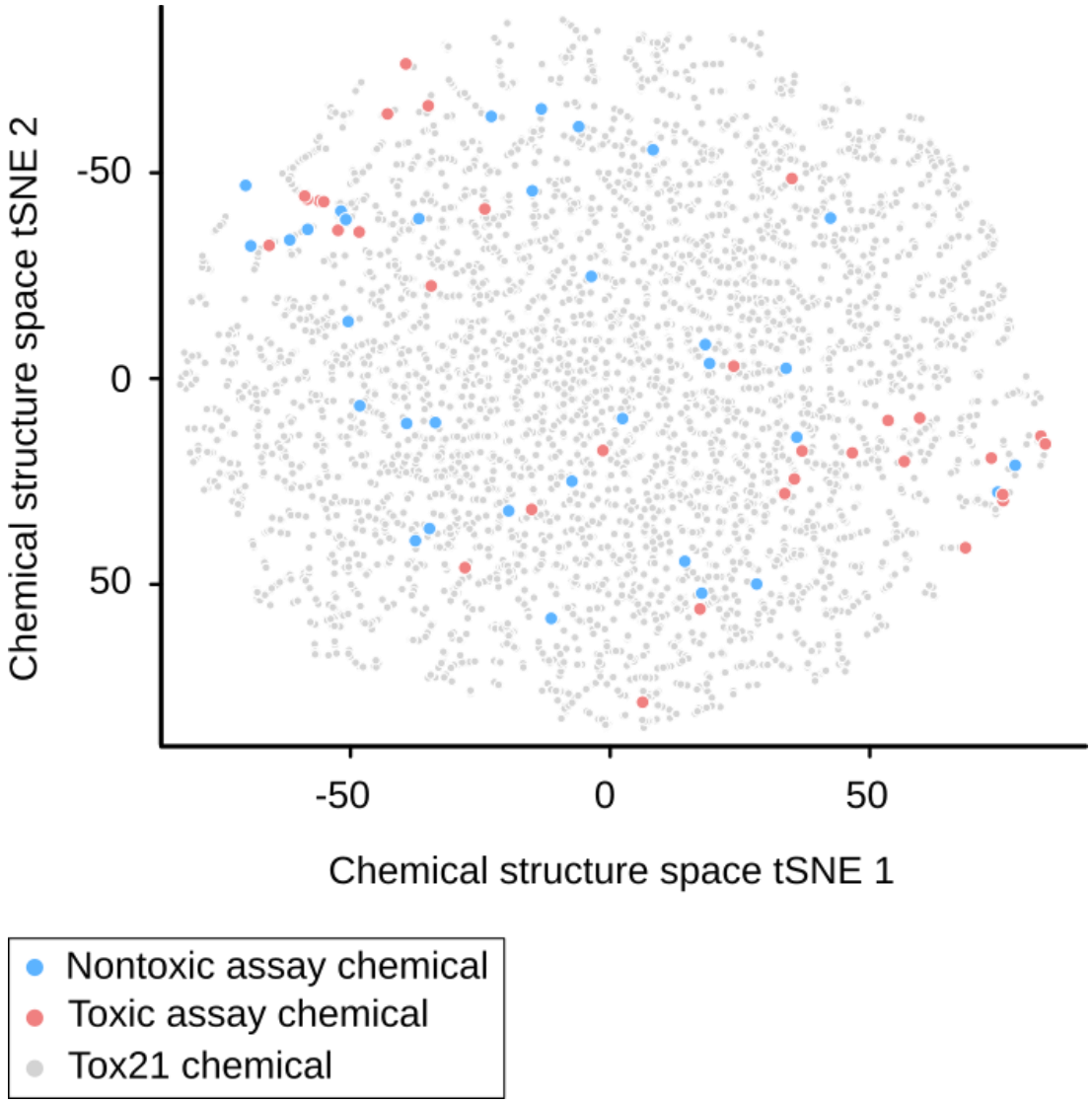
There are certain limitations to this search, most obviously that it only covers academic research papers or other resources that happen to have been included in Web of Science. Also the search parameters above will count papers which include both potency-metric-related and efficacy-metric-related terms in both subsets.

Supplementary Figure S1. Literature occurrence of metric types in the Web of Science database (2000-2019)



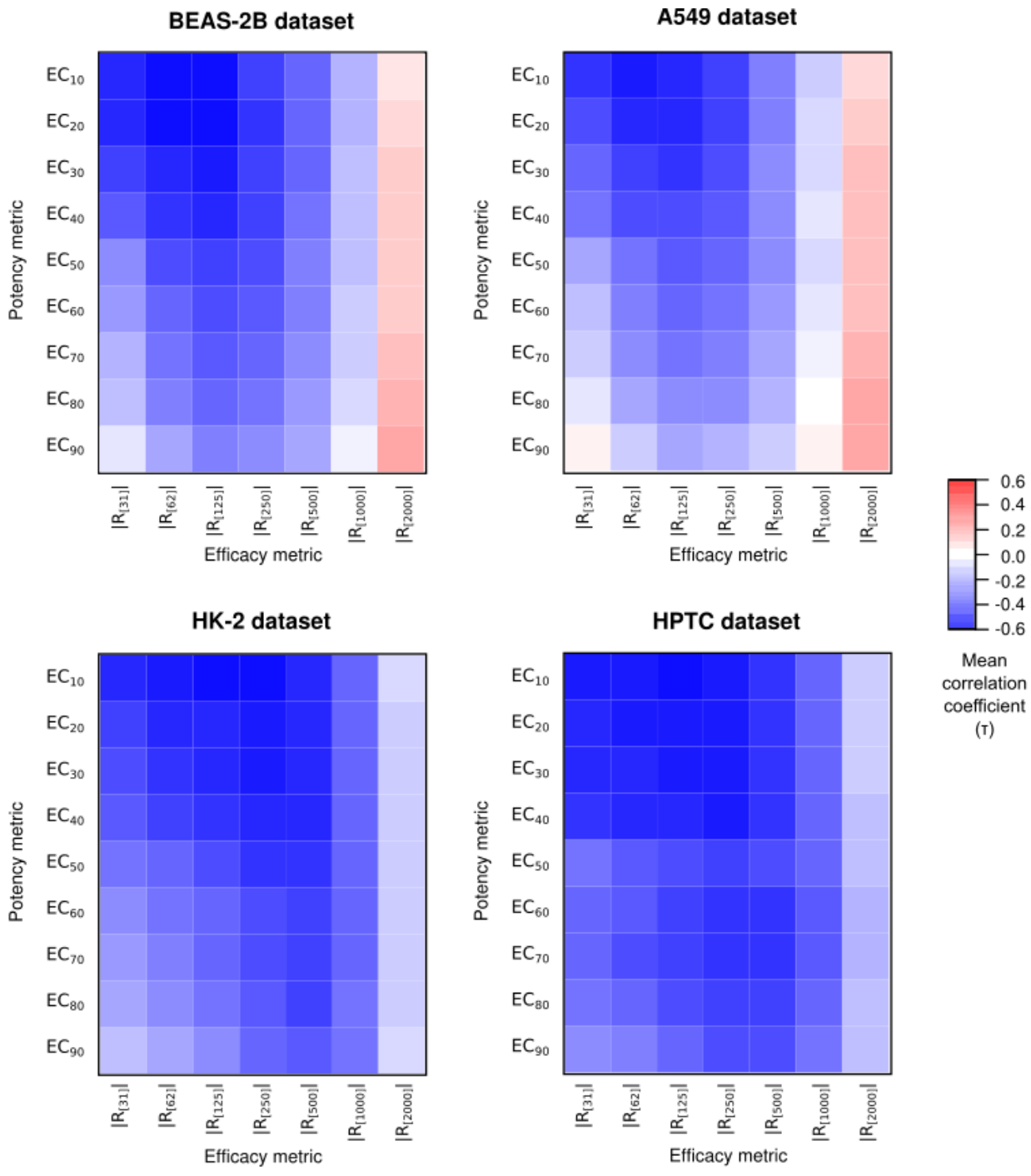
Supplementary Figure S2. t-SNE of study chemicals within Tox21

Visualisation of the studies' assay chemicals (red = pulmono-/nephro-toxic annotation, blue = non-pulmono-/nephro-toxic annotation) and the chemicals of the U.S. EPA's Tox21 chemical database (U.S. EPA, 2013) via t-distributed stochastic neighbour embedding (van der Maaten and Hinton, 2008). The Tox21 database contained 8,795 chemical entries, of which 8,599 had valid SMILES which could be incorporated into the t-SNE plot.



Supplementary Figure S3. Kendall's τ between metric pairs by dataset

Heatmap showing the Kendall's rank correlation coefficients (τ) between potency and efficacy CRC metrics for all the measured phenotypic features from all four datasets. Potency metrics with extremely large or NA values (**Methods**) are excluded. See **Fig. 2b** for the coefficients over all datasets taken together.



Supplementary Table S1: Descriptions of all phenotypic features

Name in cellXpress syntax	Study	Feature type	Description
area:mask:cell_region	Both	Morphology	Cell area
area:mask:dna_region	Both	Morphology	Nuclear area
ccoeff_normed:DNA-Actin:cell_region	Both	Pixel correlations	Correlation coefficient of DNA and actin intensities at the whole-cell region
ccoeff_normed:DNA-gH2AX:cell_region	Both	Pixel correlations	Correlation coefficient of DNA and γH2AX intensities at the whole-cell region
ccoeff_normed:gH2AX-Actin:cell_region	Both	Pixel correlations	Correlation coefficient of γH2AX and actin intensities at the whole-cell region
ccorr_normed:DNA-Actin:cell_region	Both	Pixel correlations	Spatial correlation coefficient of DNA and actin intensities at the whole-cell region
ccorr_normed:DNA-gH2AX:cell_region	Both	Pixel correlations	Spatial correlation coefficient of DNA and γH2AX intensities at the whole-cell region
ccorr_normed:gH2AX-Actin:cell_region	Both	Pixel correlations	Spatial correlation coefficient of γH2AX and actin intensities at the whole-cell region
cellcount	Both	Cell count	Number of cells
cv_intensity:Actin:cell_region	Both	Intensity	Coefficient of variation of actin intensity at the whole-cell region
cv_intensity:DNA:dna_region	Both	Intensity	Coefficient of variation of DNA intensity at the nuclear region
cv_intensity:gH2AX:cell_region	Both	Intensity	Coefficient of variation of γH2AX intensity at the whole-cell region
glcm_asm_mean:Actin:cell_region	Both	Glcm textures	Mean angular second moment of the whole-cell actin GLCM
glcm_asm_mean:DNA:dna_region	Both	Glcm textures	Mean angular second moment of the nuclear DNA GLCM
glcm_contrast_mean:Actin:cell_region	Both	Glcm textures	Mean contrast of the whole-cell actin GLCM
glcm_contrast_mean:DNA:dna_region	Both	Glcm textures	Mean contrast of the nuclear DNA GLCM
glcm_corr_mean:Actin:cell_region	Both	Glcm textures	Mean correlation of the whole-cell actin GLCM
glcm_corr_mean:DNA:dna_region	Both	Glcm textures	Mean correlation of the nuclear DNA GLCM
glcm_diff_ent_mean:Actin:cell_region	Both	Glcm textures	Mean difference entropy of the whole-cell actin GLCM
glcm_diff_ent_mean:DNA:dna_region	Both	Glcm textures	Mean difference entropy of the nuclear DNA GLCM
glcm_diff_var_mean:Actin:cell_region	Both	Glcm textures	Mean difference variance of the whole-cell actin GLCM
glcm_diff_var_mean:DNA:dna_region	Both	Glcm textures	Mean difference variance of the nuclear DNA GLCM

glcm_ent_mean:Actin:cell_region	Both	Glcm textures	Mean entropy of the whole-cell actin GLCM
glcm_ent_mean:DNA:dna_region	Both	Glcm textures	Mean entropy of the nuclear DNA GLCM
glcm_idm_mean:Actin:cell_region	Both	Glcm textures	Mean inverse difference moment of the whole-cell actin GLCM
glcm_idm_mean:DNA:dna_region	Both	Glcm textures	Mean inverse difference moment of the nuclear DNA GLCM
glcm_info_corr1_mean:Actin:cell_region	Both	Glcm textures	Mean information measure of correlation 1 of the whole-cell actin GLCM
glcm_info_corr1_mean:DNA:dna_region	Both	Glcm textures	Mean information measure of correlation 1 of the nuclear DNA GLCM
glcm_info_corr2_mean:Actin:cell_region	Both	Glcm textures	Mean information measure of correlation 2 of the whole-cell actin GLCM
glcm_info_corr2_mean:DNA:dna_region	Both	Glcm textures	Mean information measure of correlation 2 of the nuclear DNA GLCM
glcm_sum_ave_mean:Actin:cell_region	Both	Glcm textures	Mean sum average of the whole-cell actin GLCM
glcm_sum_ave_mean:DNA:dna_region	Both	Glcm textures	Mean sum average of the nuclear DNA GLCM
glcm_sum_ent_mean:Actin:cell_region	Both	Glcm textures	Mean sum entropy of the whole-cell actin GLCM
glcm_sum_ent_mean:DNA:dna_region	Both	Glcm textures	Mean sum entropy of the nuclear DNA GLCM
glcm_sum_var_mean:Actin:cell_region	Both	Glcm textures	Mean of the sum variance of the whole-cell actin GLCM
glcm_sum_var_mean:DNA:dna_region	Both	Glcm textures	Mean of the sum variance of the nuclear DNA GLCM
glcm_var_mean:Actin:cell_region	Both	Glcm textures	Mean of the variance of the whole-cell actin GLCM
glcm_var_mean:DNA:dna_region	Both	Glcm textures	Mean of the variance of the nuclear DNA GLCM
mean_intensity:Actin:cell_region	Both	Intensity	Mean actin intensity at the whole-cell region
mean_intensity:Actin:dna_region	Both	Intensity	Mean actin intensity at the nuclear region
mean_intensity:Actin:nondna_inner	Both	Intensity	Mean actin intensity at the inner cytoplasmic region
mean_intensity:Actin:nondna_outer	Both	Intensity	Mean actin intensity at the outer cytoplasmic region
mean_intensity:Actin:nondna_peridna	Both	Intensity	Mean actin intensity at the pericellular region
mean_intensity:Actin:nondna_region	Both	Intensity	Mean actin intensity at the cytoplasmic region
mean_intensity:DNA:dna_region	Both	Intensity	Mean DNA intensity at the nuclear region
mean_intensity:gH2AX:cell_region	Both	Intensity	Mean γH2AX intensity at the whole-cell region
mean_intensity:gH2AX:dna_region	Both	Intensity	Mean γH2AX intensity at the nuclear region
roundness:mask:cell_region	Both	Morphology	Cell roundness
roundness:mask:dna_region	Both	Morphology	Nuclear roundness

total_intensity_ratio:Actin-Actin:dna_region-cell_region	Both	Intensity ratios	Ratio between total actin intensity at the nuclear region over the whole-cell region
total_intensity_ratio:Actin-Actin:nondna_outer-cell_region	Both	Intensity ratios	Ratio between total actin intensity at the outer cytoplasmic region over the whole-cell region
total_intensity_ratio:Actin-Actin:nondna_peridna-cell_region	Both	Intensity ratios	Ratio between total actin intensity at the pericellular region over the whole-cell region
total_intensity_ratio:DNA-Actin:cell_region-cell_region	Both	Intensity ratios	Ratio between total DNA over actin intensities at the whole-cell region
total_intensity_ratio:gH2AX-Actin:cell_region-cell_region	Both	Intensity ratios	Ratio between total γH2AX over actin intensities at the whole-cell region
total_intensity_ratio:gH2AX-DNA:cell_region-cell_region	Both	Intensity ratios	Ratio between total γH2AX over DNA intensities at the whole-cell region
total_intensity_ratio:gH2AX-gH2AX:dna_region-cell_region	Both	Intensity ratios	Ratio between total γH2AX intensity at the nuclear region over the whole-cell region
total_intensity:Actin:cell_region	Both	Intensity	Total actin intensity at the whole-cell region
total_intensity:Actin:dna_region	Both	Intensity	Total actin intensity at the nuclear region
total_intensity:Actin:nondna_inner	Both	Intensity	Total actin intensity at the inner cytoplasmic region
total_intensity:Actin:nondna_outer	Both	Intensity	Total actin intensity at the outer cytoplasmic region
total_intensity:Actin:nondna_peridna	Both	Intensity	Total actin intensity at the pericellular region
total_intensity:Actin:nondna_region	Both	Intensity	Total actin intensity at the cytoplasmic region
total_intensity:DNA:dna_region	Both	Intensity	Total DNA intensity at the nuclear region
total_intensity:gH2AX:cell_region	Both	Intensity	Total γH2AX intensity at the whole-cell region
total_intensity:gH2AX:dna_region	Both	Intensity	Total γH2AX intensity at the nuclear region
glcm_asm_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean angular second moment of the whole-cell γH2AX GLCM
glcm_asm_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the angular second moment of the whole-cell actin GLCM
glcm_asm_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the angular second moment of the nuclear DNA GLCM
glcm_asm_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the angular second moment of the whole-cell γH2AX GLCM
glcm_contrast_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean contrast of the whole-cell γH2AX GLCM
glcm_contrast_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the contrast of the whole-cell actin GLCM

glcm_contrast_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the contrast of the nuclear DNA GLCM
glcm_contrast_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the contrast of the whole-cell γH2AX GLCM
glcm_corr_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean correlation of the whole-cell γH2AX GLCM
glcm_corr_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the correlation of the whole-cell actin GLCM
glcm_corr_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the correlation of the nuclear DNA GLCM
glcm_corr_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the correlation of the whole-cell γH2AX GLCM
glcm_diff_ent_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean difference entropy of the whole-cell γH2AX GLCM
glcm_diff_ent_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the difference entropy of the whole-cell actin GLCM
glcm_diff_ent_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the difference entropy of the nuclear DNA GLCM
glcm_diff_ent_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the difference entropy of the whole-cell γH2AX GLCM
glcm_diff_var_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean difference variance of the whole-cell γH2AX GLCM
glcm_diff_var_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the difference variance of the whole-cell actin GLCM
glcm_diff_var_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the difference variance of the nuclear DNA GLCM
glcm_diff_var_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the difference variance of the whole-cell γH2AX GLCM
glcm_ent_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean entropy of the whole-cell γH2AX GLCM
glcm_ent_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the entropy of the whole-cell actin GLCM
glcm_ent_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the entropy of the nuclear DNA GLCM
glcm_ent_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the entropy of the whole-cell γH2AX GLCM
glcm_idm_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean inverse difference moment of the whole-cell γH2AX GLCM

glcm_idm_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the inverse difference moment of the whole-cell actin GLCM
glcm_idm_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the inverse difference moment of the nuclear DNA GLCM
glcm_idm_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the inverse difference moment of the whole-cell γH2AX GLCM
glcm_info_corr1_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean information measure of correlation 1 of the whole-cell γH2AX GLCM
glcm_info_corr1_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 1 of the whole-cell actin GLCM
glcm_info_corr1_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 1 of the nuclear DNA GLCM
glcm_info_corr1_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 1 of the whole-cell γH2AX GLCM
glcm_info_corr2_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean information measure of correlation 2 of the whole-cell γH2AX GLCM
glcm_info_corr2_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 2 of the whole-cell actin GLCM
glcm_info_corr2_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 2 of the nuclear DNA GLCM
glcm_info_corr2_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the information measure of correlation 2 of the whole-cell γH2AX GLCM
glcm_sum_ave_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean sum average of the whole-cell γH2AX GLCM
glcm_sum_ave_std:Actin:cell_region	Kidney	Glcm textures	Mean sum average of the whole-cell actin GLCM
glcm_sum_ave_std:DNA:dna_region	Kidney	Glcm textures	Mean sum average of the nuclear DNA GLCM
glcm_sum_ave_std:gH2AX:cell_region	Kidney	Glcm textures	Mean sum average of the whole-cell γH2AX GLCM
glcm_sum_ent_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean sum entropy of the whole-cell γH2AX GLCM
glcm_sum_ent_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the sum entropy of the whole-cell actin GLCM
glcm_sum_ent_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the sum entropy of the nuclear DNA GLCM
glcm_sum_ent_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the sum entropy of the whole-cell γH2AX GLCM
glcm_sum_var_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean of the sum variance of the whole-cell γH2AX GLCM
glcm_sum_var_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the sum variance of the whole-cell actin GLCM

glcm_sum_var_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the sum variance of the nuclear DNA GLCM
glcm_sum_var_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the sum variance of the whole-cell γH2AX GLCM
glcm_var_mean:gH2AX:cell_region	Kidney	Glcm textures	Mean of the variance of the whole-cell γH2AX GLCM
glcm_var_std:Actin:cell_region	Kidney	Glcm textures	Standard deviation in the variance of the whole-cell actin GLCM
glcm_var_std:DNA:dna_region	Kidney	Glcm textures	Standard deviation in the variance of the nuclear DNA GLCM
glcm_var_std:gH2AX:cell_region	Kidney	Glcm textures	Standard deviation in the variance of the whole-cell γH2AX GLCM
mean_intensity:gH2AX:nondna_inner	Kidney	Intensity	Mean γH2AX intensity at the inner cytoplasmic region
mean_intensity:gH2AX:nondna_outer	Kidney	Intensity	Mean γH2AX intensity at the outer cytoplasmic region
mean_intensity:gH2AX:nondna_peridna	Kidney	Intensity	Mean γH2AX intensity at the pericellular region
mean_intensity:gH2AX:nondna_region	Kidney	Intensity	Mean γH2AX intensity at the cytoplasmic region
solidity:mask:cell_region	Kidney	Morphology	Cell solidity
solidity:mask:dna_region	Kidney	Morphology	Nuclear solidity
total_intensity_ratio:gH2AX-gH2AX:nondna_outer-cell_region	Kidney	Intensity ratios	Ratio between total γH2AX intensity at the outer cytoplasmic region over the whole-cell region
total_intensity_ratio:gH2AX-gH2AX:nondna_peridna-cell_region	Kidney	Intensity ratios	Ratio between total γH2AX intensity at the pericellular region over the whole-cell region
total_intensity:gH2AX:nondna_inner	Kidney	Intensity	Total γH2AX intensity at the inner cytoplasmic region
total_intensity:gH2AX:nondna_outer	Kidney	Intensity	Total γH2AX intensity at the outer cytoplasmic region
total_intensity:gH2AX:nondna_peridna	Kidney	Intensity	Total γH2AX intensity at the pericellular region
total_intensity:gH2AX:nondna_region	Kidney	Intensity	Total γH2AX intensity at the cytoplasmic region
aspect_ratio:mask:cell_region	Lung	Morphology	Cell aspect ratio
aspect_ratio:mask:dna_region	Lung	Morphology	Nuclear aspect ratio
ccoeff_normed:DNA-Actin:dna_chromosome	Lung	Pixel correlations	Correlation coefficient of DNA and actin intensities at the chromosomal region
ccoeff_normed:DNA-Actin:dna_region	Lung	Pixel correlations	Correlation coefficient of DNA and actin intensities at the nuclear region
ccoeff_normed:DNA-gH2AX:dna_chromosome	Lung	Pixel correlations	Correlation coefficient of DNA and γH2AX intensities at the chromosomal region

ccoeff_normed:DNA-gH2AX:dna_region	Lung	Pixel correlations	Correlation coefficient of DNA and γH2AX intensities at the nuclear region
ccoeff_normed:gH2AX-Actin:dna_chromosome	Lung	Pixel correlations	Correlation coefficient of γH2AX and actin intensities at the chromosomal region
ccoeff_normed:gH2AX-Actin:dna_region	Lung	Pixel correlations	Correlation coefficient of γH2AX and actin intensities at the nuclear region
ccorr_normed:DNA-Actin:dna_chromosome	Lung	Pixel correlations	Spatial correlation coefficient of DNA and actin intensities at the chromosomal region
ccorr_normed:DNA-Actin:dna_region	Lung	Pixel correlations	Spatial correlation coefficient of DNA and actin intensities at the nuclear region
ccorr_normed:DNA-gH2AX:dna_chromosome	Lung	Pixel correlations	Spatial correlation coefficient of DNA and γH2AX intensities at the chromosomal region
ccorr_normed:DNA-gH2AX:dna_region	Lung	Pixel correlations	Spatial correlation coefficient of DNA and γH2AX intensities at the nuclear region
ccorr_normed:gH2AX-Actin:dna_chromosome	Lung	Pixel correlations	Spatial correlation coefficient of γH2AX and actin intensities at the chromosomal region
ccorr_normed:gH2AX-Actin:dna_region	Lung	Pixel correlations	Spatial correlation coefficient of γH2AX and actin intensities at the nuclear region
cv_intensity:Actin:dna_chromosome	Lung	Intensity	Coefficient of variation of actin intensity at the chromosomal region
cv_intensity:Actin:dna_region	Lung	Intensity	Coefficient of variation of actin intensity at the nuclear region
cv_intensity:Actin:nondna_inner	Lung	Intensity	Coefficient of variation of actin intensity at the inner cytoplasmic region
cv_intensity:Actin:nondna_outer	Lung	Intensity	Coefficient of variation of actin intensity at the outer cytoplasmic region
cv_intensity:Actin:nondna_peridna	Lung	Intensity	Coefficient of variation of actin intensity at the pericellular region
cv_intensity:Actin:nondna_region	Lung	Intensity	Coefficient of variation of actin intensity at the cytoplasmic region
cv_intensity:DNA:dna_chromosome	Lung	Intensity	Coefficient of variation of DNA intensity at the chromosomal region
cv_intensity:gH2AX:dna_chromosome	Lung	Intensity	Coefficient of variation of γH2AX intensity at the chromosomal region
cv_intensity:gH2AX:dna_region	Lung	Intensity	Coefficient of variation of γH2AX intensity at the nuclear region
fraction_obj_intensity:Actin:dna_chromosome-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the chromosomal region over the whole-cell region
fraction_obj_intensity:Actin:dna_region-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the nuclear region over the whole-cell region
fraction_obj_intensity:Actin:nondna_inner-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the inner cytoplasmic region over the whole-cell region
fraction_obj_intensity:Actin:nondna_outer-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the outer cytoplasmic region over the whole-cell region

fraction_obj_intensity:Actin:nondna_peridna-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the pericellular region over the whole-cell region
fraction_obj_intensity:Actin:nondna_region-Actin_object	Lung	Intensity ratios	Fraction of total actin object intensity at the cytoplasmic region over the whole-cell region
fraction_obj_intensity:DNA:dna_chromosome-DNA_object	Lung	Intensity ratios	Fraction of total DNA object intensity at the chromosomal region over the whole-cell region
fraction_obj_intensity:DNA:dna_region-DNA_object	Lung	Intensity ratios	Fraction of total DNA object intensity at the nuclear region over the whole-cell region
fraction_obj_intensity:DNA:nondna_region-DNA_object	Lung	Intensity ratios	Fraction of total DNA object intensity at the cytoplasmic region over the whole-cell region
fraction_obj_intensity:gH2AX:dna_chromosome-gH2AX_object	Lung	Intensity ratios	Fraction of total γH2AX object intensity at the chromosomal region over the whole-cell region
fraction_obj_intensity:gH2AX:dna_region-gH2AX_object	Lung	Intensity ratios	Fraction of total γH2AX object intensity at the nuclear region over the whole-cell region
glcm_asm_mean:Actin:dna_region	Lung	Glcm textures	Mean angular second moment of the nuclear actin GLCM
glcm_asm_mean:Actin:nondna_region	Lung	Glcm textures	Mean angular second moment of the cytoplasmic actin GLCM
glcm_asm_mean:gH2AX:dna_region	Lung	Glcm textures	Mean angular second moment of the nuclear γH2AX GLCM
glcm_contrast_mean:Actin:dna_region	Lung	Glcm textures	Mean contrast of the nuclear actin GLCM
glcm_contrast_mean:Actin:nondna_region	Lung	Glcm textures	Mean contrast of the cytoplasmic actin GLCM
glcm_contrast_mean:gH2AX:dna_region	Lung	Glcm textures	Mean contrast of the nuclear γH2AX GLCM
glcm_corr_mean:Actin:dna_region	Lung	Glcm textures	Mean correlation of the nuclear actin GLCM
glcm_corr_mean:Actin:nondna_region	Lung	Glcm textures	Mean correlation of the cytoplasmic actin GLCM
glcm_corr_mean:gH2AX:dna_region	Lung	Glcm textures	Mean correlation of the nuclear γH2AX GLCM
glcm_diff_ent_mean:Actin:dna_region	Lung	Glcm textures	Mean difference entropy of the nuclear actin GLCM
glcm_diff_ent_mean:Actin:nondna_region	Lung	Glcm textures	Mean difference entropy of the cytoplasmic actin GLCM

glcm_diff_ent_mean:gH2AX:dna_region	Lung	Glcm textures	Mean difference entropy of the nuclear γH2AX GLCM
glcm_diff_var_mean:Actin:dna_region	Lung	Glcm textures	Mean difference variance of the nuclear actin GLCM
glcm_diff_var_mean:Actin:nondna_region	Lung	Glcm textures	Mean difference variance of the cytoplasmic actin GLCM
glcm_diff_var_mean:gH2AX:dna_region	Lung	Glcm textures	Mean difference variance of the nuclear γH2AX GLCM
glcm_ent_mean:Actin:dna_region	Lung	Glcm textures	Mean entropy of the nuclear actin GLCM
glcm_ent_mean:Actin:nondna_region	Lung	Glcm textures	Mean entropy of the cytoplasmic actin GLCM
glcm_ent_mean:gH2AX:dna_region	Lung	Glcm textures	Mean entropy of the nuclear γH2AX GLCM
glcm_idm_mean:Actin:dna_region	Lung	Glcm textures	Mean inverse difference moment of the nuclear actin GLCM
glcm_idm_mean:Actin:nondna_region	Lung	Glcm textures	Mean inverse difference moment of the cytoplasmic actin GLCM
glcm_idm_mean:gH2AX:dna_region	Lung	Glcm textures	Mean inverse difference moment of the nuclear γH2AX GLCM
glcm_info_corr1_mean:Actin:dna_region	Lung	Glcm textures	Mean information measure of correlation 1 of the nuclear actin GLCM
glcm_info_corr1_mean:Actin:nondna_region	Lung	Glcm textures	Mean information measure of correlation 1 of the cytoplasmic actin GLCM
glcm_info_corr1_mean:gH2AX:dna_region	Lung	Glcm textures	Mean information measure of correlation 1 of the nuclear γH2AX GLCM
glcm_info_corr2_mean:Actin:dna_region	Lung	Glcm textures	Mean information measure of correlation 2 of the nuclear actin GLCM
glcm_info_corr2_mean:Actin:nondna_region	Lung	Glcm textures	Mean information measure of correlation 2 of the cytoplasmic actin GLCM
glcm_info_corr2_mean:gH2AX:dna_region	Lung	Glcm textures	Mean information measure of correlation 2 of the nuclear γH2AX GLCM
glcm_sum_ave_mean:Actin:dna_region	Lung	Glcm textures	Mean sum average of the nuclear actin GLCM
glcm_sum_ave_mean:Actin:nondna_region	Lung	Glcm textures	Mean sum average of the cytoplasmic actin GLCM
glcm_sum_ave_mean:gH2AX:dna_region	Lung	Glcm textures	Mean sum average of the nuclear γH2AX GLCM
glcm_sum_ent_mean:Actin:dna_region	Lung	Glcm textures	Mean sum entropy of the nuclear actin GLCM
glcm_sum_ent_mean:Actin:nondna_region	Lung	Glcm textures	Mean sum entropy of the cytoplasmic actin GLCM
glcm_sum_ent_mean:gH2AX:dna_region	Lung	Glcm textures	Mean sum entropy of the nuclear γH2AX GLCM
glcm_sum_var_mean:Actin:dna_region	Lung	Glcm textures	Mean of the sum variance of the nuclear actin GLCM

glcm_sum_var_mean:Actin:nondna_region	Lung	Glcm textures	Mean of the sum variance of the cytoplasmic actin GLCM
glcm_sum_var_mean:gH2AX:dna_region	Lung	Glcm textures	Mean of the sum variance of the nuclear γH2AX GLCM
glcm_var_mean:Actin:dna_region	Lung	Glcm textures	Mean of the variance of the nuclear actin GLCM
glcm_var_mean:Actin:nondna_region	Lung	Glcm textures	Mean of the variance of the cytoplasmic actin GLCM
glcm_var_mean:gH2AX:dna_region	Lung	Glcm textures	Mean of the variance of the nuclear γH2AX GLCM
mean_intensity:Actin:dna_chromosome	Lung	Intensity	Mean actin intensity at the chromosomal region
mean_intensity:DNA:dna_chromosome	Lung	Intensity	Mean DNA intensity at the chromosomal region
mean_intensity:gH2AX:dna_chromosome	Lung	Intensity	Mean γH2AX intensity at the chromosomal region
obj_mean_total_area:mask:Actin_object	Lung	Morphology	Mean total area of actin objects
obj_mean_total_area:mask:DNA_object	Lung	Morphology	Mean total area of DNA objects
obj_mean_total_area:mask:gH2AX_object	Lung	Morphology	Mean total area of γH2AX objects
obj_number:mask:Actin_object	Lung	Morphology	Number of actin objects
obj_number:mask:DNA_object	Lung	Morphology	Number of DNA objects
obj_number:mask:gH2AX_object	Lung	Morphology	Number of γH2AX objects
obj_stddev_total_area:mask:Actin_object	Lung	Morphology	Standard deviation in the total area of actin objects
obj_stddev_total_area:mask:DNA_object	Lung	Morphology	Standard deviation in the total area of DNA objects
obj_stddev_total_area:mask:gH2AX_object	Lung	Morphology	Standard deviation in the total area of γH2AX objects
perimeter:mask:cell_region	Lung	Morphology	Cell perimeter length
perimeter:mask:dna_region	Lung	Morphology	Nuclear perimeter length
total_intensity_ratio:Actin-Actin:dna_chromosome-cell_region	Lung	Intensity ratios	Ratio between total actin intensity at the chromosomal region over the whole-cell region
total_intensity_ratio:Actin-Actin:dna_chromosome-dna_region	Lung	Intensity ratios	Ratio between total actin intensity at the chromosomal region over the nuclear region
total_intensity_ratio:Actin-Actin:nondna_inner-cell_region	Lung	Intensity ratios	Ratio between total actin intensity at the inner cytoplasmic region over the whole-cell region
total_intensity_ratio:DNA-Actin:dna_chromosome-dna_chromosome	Lung	Intensity ratios	Ratio between total DNA over actin intensities at the chromosomal region
total_intensity_ratio:DNA-Actin:dna_region-dna_region	Lung	Intensity ratios	Ratio between total DNA over actin intensities at the nuclear region
total_intensity_ratio:gH2AX-Actin:dna_chromosome-dna_chromosome	Lung	Intensity ratios	Ratio between total γH2AX over actin intensities at the chromosomal region
total_intensity_ratio:gH2AX-Actin:dna_region-dna_region	Lung	Intensity ratios	Ratio between total γH2AX over actin intensities at the nuclear region

total_intensity_ratio:gH2AX-DNA:dna_chromosome-dna_chromosome	Lung	Intensity ratios	Ratio between total γH2AX over DNA intensities at the chromosomal region
total_intensity_ratio:gH2AX-DNA:dna_region-dna_region	Lung	Intensity ratios	Ratio between total γH2AX over DNA intensities at the nuclear region
total_intensity_ratio:gH2AX-gH2AX:dna_chromosome-cell_region	Lung	Intensity ratios	Ratio between total γH2AX intensity at the chromosomal region over the whole-cell region
total_intensity_ratio:gH2AX-gH2AX:dna_chromosome-dna_region	Lung	Intensity ratios	Ratio between total γH2AX intensity at the chromosomal region over the nuclear region
total_intensity:Actin:dna_chromosome	Lung	Intensity	Total actin intensity at the chromosomal region
total_intensity:DNA:dna_chromosome	Lung	Intensity	Total DNA intensity at the chromosomal region
total_intensity:gH2AX:dna_chromosome	Lung	Intensity	Total γH2AX intensity at the chromosomal region

1

References

Duffus, J. H., Norberg, M., Templeton, D. M. 1954- et al. IUPAC Glossary of Terms Used In Toxicology, 2nd Edition - toxicokinetics, toxicology, chemistry, risk assessment, hazard assessment, chemical terminology. Available at: https://sis.nlm.nih.gov/enviro/iupacglossary/frontmatter.html [Accessed April 27, 2018].

Haralick, R., Shanmugam, K. and Dinstein, I. (1973). Textural Features for Image Classification. *Ieee Trans Syst Man Cybern SMC3*, 610–621. https://doi.org/10.1109/TSMC.1973.4309314.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *J Mach Learn Res 9*, 2579–2605

U.S. EPA (2013). *ToxCast Data Generation: Chemical Lists, ToxCast_Generic_Chemicals_2013_12_10.xlsx*. U.S. EPA. Available at: https://www.epa.gov/chemical-research/toxcast-data-generation-chemical-lists [Accessed July 5, 2016].

Zhao, B., Tan, Y., Tsai, W.-Y. et al. (2016). Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep 6*, 23428. https://doi.org/10.1038/srep23428.