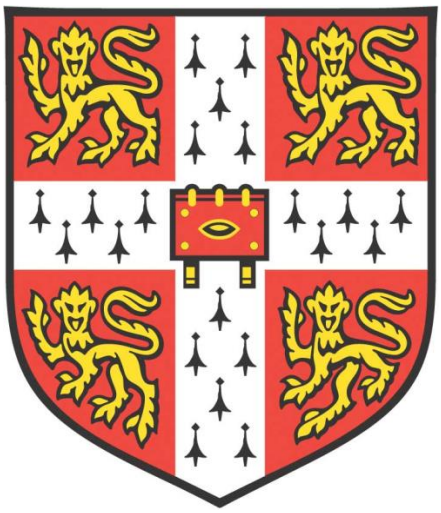
PREDICTIVE MODELS OF CELLULAR CYTOTOXICITY BASED ON CELL PAINTING READOUTS AND MOLECULAR FINGERPRINTS



Srijit Seal

Clare Hall

Centre for Molecular Informatics

Department of Chemistry

University of Cambridge

This dissertation is submitted for the degree of Master of Philosophy

August 2020

This work is dedicated to Prof John Varghese, Dr Rakhi Thareja

and Dr Violet Rajeshwari Macwan for their moral, continual and unrelenting support

for me to pursue my dreams.

“The true value lies in the small moments though - the moment you help someone, or are kind to someone; are completely immersed into the science you do; or the time you spend with your family; or doing arts, or sports, etc. Do not think only 'the goal' matters, after reaching one, there will always be the next one waiting anyway. It's about the moment just as much. And this moment is there, in front of you, every second of your life. You cannot saddle the horse that has already departed - or the one that didn't arrive yet.”

-A.B.

DECLARATION

This dissertation is the result of my work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Following the Department of Chemistry guidelines, this thesis does not exceed 15,000 words.

Srijit Seal

01/08/2020

SUMMARY

Cell morphology features, such as from the Cell Painting assay, are relatively easy to generate and represent versatile biological descriptors of the state of a biological system, and thereby compound response. In this work, we explore compound fingerprints and cell morphology features, separately and in combination, for the prediction of cytotoxicity- and proliferation-related in vitro assays endpoints. To this end, we selected from the MoleculeNet benchmark dataset 135 compounds annotated with both ten Toxcast endpoints, as well as Cell Painting readouts, where the relatively small size of the dataset in this exploratory study is due to the overlap of data annotations required. Using Cell Painting readouts, Morgan and ErG fingerprints, and their combinations, we trained Random Forest classification models for the endpoints considered using nested cross-validation. It was found that performance improved when both Cell Painting readouts and Morgan/ErG fingerprints were used in combination, over using either representation alone. When using leave one-cluster-out validation (with clusters defined based on physicochemical descriptors), Cell Painting features achieved higher average metric scores over all assays (Balanced Accuracy of 0.65, Matthews Correlation Coefficient of 0.28 and AUC-ROC of 0.71) and outperform models using ErG fingerprints (BA 0.55, MCC 0.09 and AUC-ROC 0.59) and Morgan fingerprints alone (BA 0.53, MCC 0.07 and AUC-ROC: 0.57). When using random shuffling splits over the folds, the combination of Cell Painting fingerprints with ErG and Morgan fingerprints improved balanced accuracy in 7 out of 10 assays (by 9.6% and 23.1% on average, respectively). Further, we determined the most contributing Cell Painting features in models and found that features involving correlation and granularity of cells and cytoplasm, as well as cell neighbours and cell radial distributions were contributing most significantly to performance, which is plausible given the endpoint considered. We conclude that cell morphological readouts

contain considerably distinct information from molecular fingerprints and can improve the performance of predictive cytotoxicity models, in particular in areas of novel structural space, while being generally interpretable.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr Andreas Bender, for his valuable guidance and supervision throughout this work and for supporting me throughout the course in every way. I would also like to thank all my colleagues at the Bender group for their love and encouragement in my studies. Thank you to Luis for studying alongside with learning various techniques in chemoinformatics. Thank you to Danilo and Hongbin for helping me understand machine learning and toxicity prediction using the same. I am extremely grateful to fellow residents at Elmside, especially, Anna, Peter, Hubert and Debora for making work from home one of the best things to happen. A big thanks to Srishti, Anmol, Vidhi, Aashi and Alkausil for the support throughout the tough times of this pandemic. I am greatly indebted to Dr Trudi Tate and my tutor Dr Wai Yi Feng for her valuable guidance not just in academics but also in making Clare Hall fell home and to Clare Hall for funding my academic studies here. Finally, I am very grateful to my family for their constant support in these years away from home.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ACEA, ACEA Biosciences; APR, Apredica; AUC-ROC, Area Under Curve- Receiver Operating Characteristic; AGP, Actin, Golgi, Plasma membrane; BA, Balanced Accuracy; BMF, Bayesian Matrix Factorisation; BSK, BioSeek; CP, Cell Painting; ELISA, Enzyme-Linked ImmunoSorbent Assay ; ErG, Extended reduced Graph; FN: False Negative; FP: False Positive; KS, Kolmogorov-Smirnov; MACCS, Molecular ACCess System ;MCC, Mathew's Correlation constant; MOA, Mechanism of Action; mRNA, Messenger RNA (mRNA); PCA, Principal component analysis; RF, Random Forest; RNA, Ri-bonucleic acid; SEN, Sensitivity; SMILES, Simplified Molecular-Input Line-Entry System ; SPE, Specificity; SRB, sulforhodamine B; SVM, Support Vector Machine ;TN: True Negative; TP: True Positive.

LIST OF APPENDICES

1 INTRODUCTION

The escalating cost of drug development and the slow rate of new drugs being developed is extremely concerning. A recent study in the Journal of Health Economics estimates that developing a new prescription medicine and subsequent approvals are estimated to cost £2.1 billion.[1] There is every possibility that a compound with the best therapeutic effect might fail to clear safety tests in later stages of the pipeline in drug discovery.[2] Therefore, predicting toxicity in early stages of the process is crucial. There is often a high degree of risk which comes with initial animal models, including but not limited to, irreversible toxicity to human organs which are determined at later stages in the pipeline. Animal models cannot predict human toxicity with high accuracy due to differential disease mechanisms in the two species.[3] The development of computational methods in predicting toxicity with high sensitivity and specificity is pivotal in achieving lower attrition rates in drug discovery. In such a scenario, it is crucial to augment experiments with novel computational models that achieve transfer learning from animal to human toxicity.

Machine learning, often seen as a subset of artificial intelligence, is the study of algorithms that give the ability to learn and improve from the experience. Toxicity, when

documented, can be fed into such models while we use novel techniques to relate them to various aspects of the compound such as structure or effects or the target. These models provide us with a comparatively fast, cost-effective and accurate way to predict toxicity effects.

1.1 Types of Toxicity

Toxicity can be classified into various types depending on adverse effects, target organs or even the method of study.

The adverse toxicology effects of a chemical within short periods are termed as acute toxicity. One part of assessing the safety of compounds is acute oral toxicity testing.[4] The lethal dose (LD_{50}) value is described as the concentration of the chemical at which 50% of the testing animals are killed within a 24-hour exposure (including oral, dermal, inhalation etc.). Predicting rodent toxicity is the first step towards transfer learning the idea for prediction of human toxicity and reducing and limiting the use of animals in experimental studies.

Drugs may also be studied according to their toxic effects on specific organs such as liver (biliary hyperplasia, fibrosis, and necrosis) or kidney (diabetic kidney diseases). When modelling, chronic toxicity or the effects of long-term exposure, can be much difficult to predict compared to acute toxicity.

Further, drug toxicity can be classified into two categories, intrinsic toxicity or idiosyncratic toxicity. Intrinsic toxicity is generally dose-dependent and related to the target of the drug whereas idiosyncratic toxicity is unpredictable, dose-independent and shows off-target effects. While dose-dependent intrinsic toxicity is predictable to a certain extent, it is the dose-independent idiosyncratic toxicity that is highly unpredictable and often unnoticeable in animal models.[5]

In vivo toxicology refers to the study of the toxicity of chemical substances in biological entities on whole including animals, humans and other living organisms while *in vitro* toxicology refers to the study of toxicity in non-whole animal models such as tissue slices, organs, primary and secondary cell cultures, cell lines or cell compartments such as mitochondria. Here the experimental setup is comparatively faster, better reproducible and most importantly, more reliable.

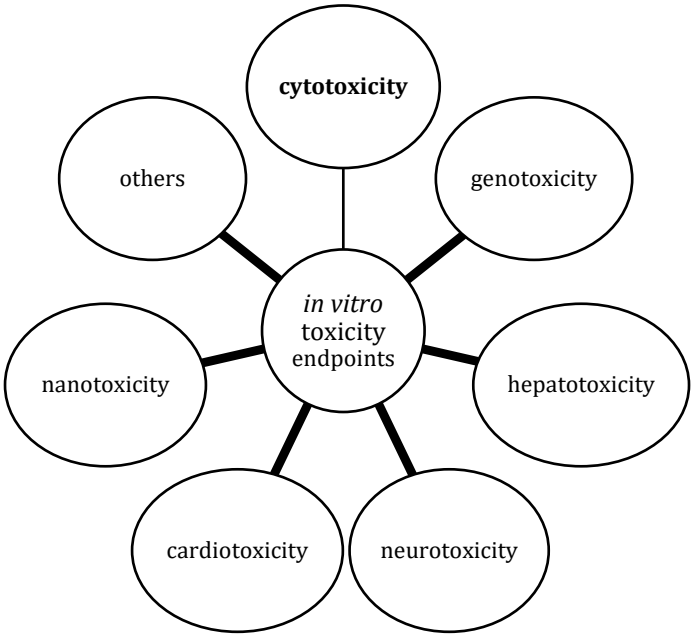


Figure 1.1: Types of *in vitro* toxicity endpoints

As shown in Figure 1.1, *in vitro* toxicity endpoints may be organ-specific; one of the most important organ-specific endpoints of *in vitro* toxicity is cytotoxicity, which is often the first screen for biological safety of a compound. Other forms of organ-specific targets include, but are not limited to, genotoxicity (damage to DNA), hepatotoxicity (damage to the liver), neurotoxicity, nephrotoxicity (damage to kidneys), cardiotoxicity, nanotoxicity (damage to the cellular and molecular level).[6] Overall, *in vitro* toxicology has gained vast attention as a rapidly evolving high throughput model. Data from such experiments over the years have been used to train models in supplementing clinical trials.

1.2 Computational approaches in toxicity prediction

Despite chemoinformatics being developed over the past five decades, there exists a huge scope for improvement when it comes to toxicity prediction. To explain the given complex mechanism behind drug toxicity, several reports have suggested using novel strategies. The National Toxicology Program (NTP) published its 2004 report aiming to support the evolution of toxicology observational science of disease-specific models to a predominantly predictive science based on several chemical and biological factors such as targets and mechanism of action.[7] In its report 'Toxicity Testing in the 21st Century: a Vision and a Strategy', the National Research Council (NRC) proposes using computational methods to make toxicity testing more relevant to humans by using a cell from humans, thus decreasing animals testing and making the process cheaper and faster.[8] Computational methods include studying chemical similarity, correlation measures, predicting protein-binding, using gene expression and metabolomics profiles or leveraging cell morphological changes concerning perturbations to understand better the mechanisms behind human toxicity. When predicting *in silico*, most machine learning models are not very useful when tested on new compounds outside its space of knowledge. Hence, there remains much to achieve in developing such models to be able to handle novel unseen data especially in predicting human toxicity, where the availability of training data is a challenge in the first place. To ensure better results from *in silico* models, it is necessary to understand the limits, strengths, application domain and interpretation of a model.[9] The choice of algorithm and customisation of the method to the problem is crucial. It is also important to account for other biological and chemical factors, such as dosage of chemicals and physiologically based pharmacokinetic properties to allow extrapolation of *in vivo* effects from *in vitro* effects.[10] The future of toxicity prediction will be largely determined by the evolution of machine learning and the availability of new data. New endpoints of toxicity need to be established to better

predict acute and chronic toxicity as well as the effects of low-dose repeated exposure. It should also be possible to determine a relationship between organ level toxicity with *in vitro* toxicity relevant to specific changes in cell morphology. Integrating genetic information could also help predict toxicity with greater accuracy. Development of evaluation methods is also critical to assess the models predicting toxicity. In such cases integrating different algorithms and strategies is crucial as is the use of novel fingerprints as input features that are relevant and meaningful in predicting the target outcome.

In summary, *in silico* is a strong promising tool in toxicity assessment and can provide a quantitative bridge towards determining human toxicity if they are improved and researched in the dawn of 21st-century toxicology.

1.3 Machine learning in toxicity prediction

Machine learning models allow prediction without explicit programming while for a simple problem it is also possible at times to understand how the prediction was made. Given that toxicity endpoints are complex, machine learning methods have been widely used in toxicity prediction over the years to learn and develop knowledge about toxicity, the final aim often being to speed-up the drug discovery process.[11]

1.3.1 Data and Endpoints

1.3.1.1 Endpoints

In order to model toxicity, endpoints must be established for a model which often relate to cytotoxicity, cardiotoxicity, hepatotoxicity, carcinogenicity among others. An endpoint can be related to different organisms, including humans although such data is limited. Hence there is often a tendency to look for models representing similar biological systems (such as rats, mouse, dogs, etc.). Toxicity is however multidimensional and can be measured in different ways and time points. While there exist multiple public datasets that correspond to different targets, there are also in-house datasets owned by the industry.

Some common databases were compiled and described by Toropov et. al in an attempt to help systematization of criteria and methods involved in chemoinformatics research.[12]

1.3.1.2 Chemical Data Representation

A compound in such a dataset would ideally be represented in a machine-readable format. The most commonly used among them is the Simplified Molecular Input Line Entry System (SMILES).[13] SMILES strings are linear annotations that can be written in several variations however canonicalization algorithms can be used to generate unique SMILES for each chemical structure as shown in Figure 1.2.

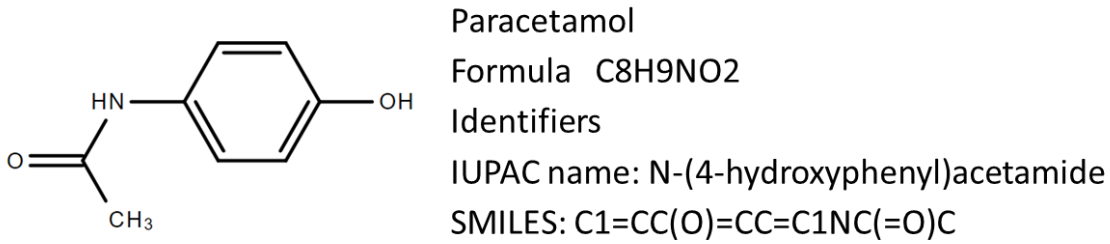


Figure 1.2: An example of a chemical structure, molecular formula, SMILES identifier of the common anti-inflammatory drug paracetamol.

1.3.2 Molecular Descriptors and Fingerprints

Conventionally, toxicity predictions have been based on structural information known about the compound of interest and its mechanism of action on the molecular target. However, in the past decade, many works have begun to revolutionise this thought. High-throughput image-based screens, for example, have shown to be a wide choice for drug screening[14] Further phenotypic profiling can be improved when combined with chemical, bioactivity, image-based screening profiles. With the correct machine learning algorithms, even seemingly unrelated image-based fingerprints could be repurposed to predict the biological activity of *in vitro* assays. The use of fingerprints as features to get the best and interpretable performance from a model remains a challenge. A compound

is after all unique to itself and there can be no one way of representing it either structurally or biologically or in some other way.

Each of the molecules in a dataset can be represented by molecular descriptors of physicochemical properties or molecular fingerprints. Here the aim remains to derive properties of compounds which can later be utilized to model the desired endpoint.

1.3.2.1 Physicochemical Properties

Molecular descriptors are mathematical representations of the molecular structure in the form of physicochemical properties that may be derived theoretically (such as number rotatable bonds) or from experimentation (dipole moment, log P). Among them, are descriptors such as topological polar surface area (TPSA) [15] which measures the compound's molecular surface area and the octanol-water partition coefficient (ClogP). The effect of compounds in humans and the environment may depend on such physicochemical properties which affect its absorption, distribution, metabolism, excretion and toxicity in the human body making them valuable as parameters.

1.3.2.2 Morgan Fingerprints

Morgan Fingerprints are usually one of the conventional 2D fingerprints used in chemoinformatics that represent molecules using a hashed binary bit strings that encode the chemical structure with each bit as either 1 or 0. The neighbourhood of each non-hydrogen atom in the molecule is extracted into multiple circular layers in a different iteration of the algorithm up to a given diameter. Using circular neighbourhoods, such fingerprints can then represent an infinite number of molecular structural features. Following Figure 1.3, the first step involves initialises the initial integer identifier of each atom. In the next iteration, this identifier is further updated with identifier values from the immediate neighbourhood environment. Each iteration henceforth moves further away and at the end of the loop, duplicate identifiers are removed.

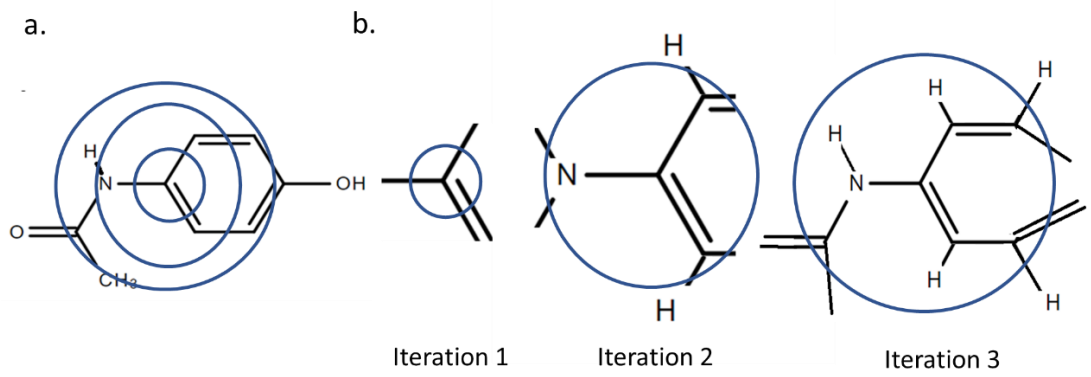


Figure 1.3: An illustrative example of circular atom environments within Paracetamol. a. Centre atom and atom environments at the second and third level, illustrated here via concentric circles. b. Updating identifier values by iteration on the information from atoms in the immediate neighbourhood.

Thus, ECFPs are not based on predefined structural keys, but they are derived from the molecule. They represent the molecule based on its circular neighbourhood and thus can be calculated rapidly. For example, ECFP_4 indicates that the number of iterations performed in generating the fingerprint is 2, half of the diameter of the largest feature in the fingerprint which in this case is 4.

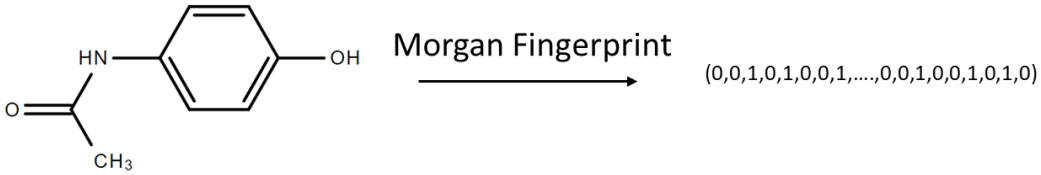


Figure 1.4: Basic Approach to Morgan Fingerprinting.

The feature vector is then hashed into integer values and collected into a list. This representation is then folded into a shorter bit string of a given length. This binary code is also known as an extended-connectivity fingerprint (ECFP).[16] In our work, we use binary encoded Morgan fingerprints which are ECFP_4 fingerprints with a bit length of 2,048.

1.3.2.3 ErG Fingerprints

ErG fingerprints or extended reduced graph (ErG) fingerprints are a 2D description of chemical entities based on the concepts enumerating reduced graphs. Reduced graphs are encoding techniques used to retain only the chemically relevant information about graph substructures. They use pharmacophore-type node descriptors to encode physicochemical and structural properties such as pharmacophore properties, size, and shape of the small molecules including graph topology, atomic charge and H-bond donor and acceptor sites.[17]

1.3.3 Statistical Data Analysis

Data evaluation can be undertaken using several univariate and multivariate analysis to provide some understanding of the distribution, central value, spread and correlation among the data points. Some examples have been discussed here.

1.3.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is one of the fundamental methods for application of multivariate data analysis which may be used in chemoinformatics for example to plot the chemical space. PCA was used in this study to cluster compounds based on physicochemical properties of the compounds. As an unsupervised method of learning, it is commonly used in dimensionality reduction on large datasets to project them into lower-dimensional variable space by calculating linear latent variables.[18] These PCA scores preserve the distances or similarities between different data points in higher dimensions and can be applied to any numeric data matrix. The lower-dimensional dataset is spanned by the orthogonal components of the higher dimensional space. The scaling of the higher dimensional space may in some cases help remove the unintentional larger contributions of some features to the variation. The direction along which the Euclidean distance between objects is best described is also the direction with the highest variance

of scores known as the first principal component or PC1. The second component PC2 is orthogonal to PC1 having the maximum possible variance of the score.

In other words, PCA involves the computation of linear, orthogonal combinations of the original high dimensional space. The components (PC) are *m* eigenvectors of a covariance matrix ($\underline{\underline{X}}^T\underline{\underline{X}}$), computed from the *n*×*m* matrix $\underline{\underline{X}}$, where each element x_{nm} is the n^{th} row value of the m^{th} column of the matrix. The eigenvalues thereby represent the independent contribution to variance associated with the original high dimensional space. The distribution of data by the original dataset can then be approximated using a scatter plot between the two PCs with the largest eigenvalues, namely the first two components, PC1 and PC2 in a plate or the PCA plot. The percentage of total variance retained by the two components is important analytics of any PCA plot. If this sum is around 70%, the PCA can be accounted for to give a good representation of the original multidimensional data.

1.3.3.2 Kolmogorov-Smirnov test

There are instances when one needs to understand the distribution of the population in their dataset. The Kolmogorov–Smirnov test (KS) on two samples may also be used to determine whether two probability distributions differ, form examples, two different classes, in a dataset. [19]

Suppose the first set of samples has a cardinality of *m* with observed cumulative distribution function *F(x)* and the second set, *n* and *G(x)* respectively. We define $D_{m,n}$ as

$$D_{m,n} = \max |F(x) - G(x)|$$

This is a two-sided test where the null hypothesis H_o states that two independent samples are drawn from the same continuous distribution. Using the KS test, we reject the null hypothesis at significance level α if $D_{m,n} > D_{m,n,\alpha}$ where $D_{m,n,\alpha}$ is the critical value.

If *m* and *n* are sufficiently large,

$$D_{m,\,n,\,\alpha} = c(\alpha)\sqrt{\frac{m+n}{mn}}$$

Where $c(\alpha)$ = inverse of the Kolmogorov distribution at α.

This means, the test checks if both the samples come from the same distribution irrespective of the fact it being normal or not and is thus devised to be sensitive against all types of differences between two distribution functions. The statistic value and the p-value are used to determine the relationship between the two distributions. If the KS statistic value is small or the p-value is high, we cannot reject the null hypothesis. This would imply that the distributions of the two samples are the indeed same. Therefore, the KS test is widely used as a general nonparametric method in comparing two samples.

1.3.3.3 T-test

The dependent t-test is designed to understand the statistical significance of two sets of data using a paired difference.[20] The two-sided t-test for related samples tests for the null hypothesis H_o states that two related or repeated samples have identical expected average values, that is if the averages of two groups of data are significantly different from each other. The t statistic is calculated as

$$t = \frac{\bar{X}_D - \mu_0}{s_D \,/\, \sqrt{n}}$$

Where $\bar{X}_D$ and s_D are the mean and standard deviation of the difference between all pairs. μ_0 is a constant. $\mu_0 = 0$ if we want to test if the mean of the difference is significantly different. *n* represents the number of pairs and thus the degree of freedom is *n-1*. The higher a t-statistic value, the groups are progressively different from each other. The p-value, on the other hand, is a percentage representing the probability that the results are by chance. If the p-value is larger, for example, greater than 0.05 or 0.10, then one cannot reject the null hypothesis of identical average scores. A lower p-value indicates the results did not occur by chance and is often associated with a large t-statistics. The t-statistic

value is, therefore, a ratio between the difference between the two groups of data and difference within the groups itself.

1.3.4 Models

Models can be trained on classification or regression tasks using a battery of different algorithms, some using machine learning, others using artificial neural networks (deep learning). In this work, we only used machine learning algorithm of Random Forest to build our classification models, however, to explain the working of a Random Forest model, we will also briefly describe a decision tree (DT).

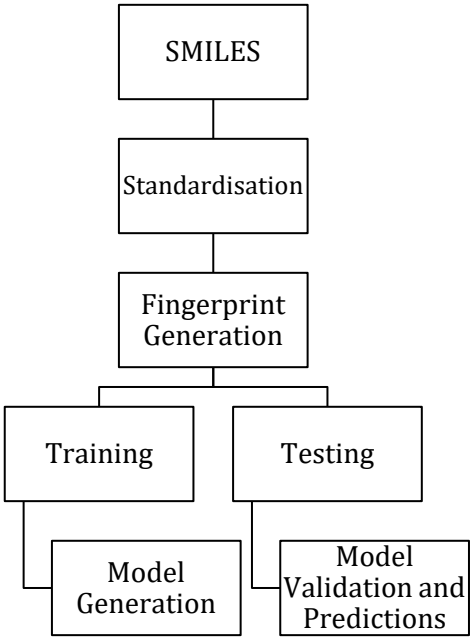


Figure 1.5: Basic prediction workflow employed in a machine learning model using chemical structure.

1.3.4.1 Decision Tree Classifier

A decision tree classifier[21] is one of the simplest classification techniques and is at the foundation of random forest algorithms. Decision tree uses a tree structure partition down the training data into smaller subsets based on one variable at a time in an iterative process.

Figure 1.6 explains the basic working of a decision tree. To classify a data point *y*, the sample moves down a tree and at each node, a feature of *y* is compared to a threshold value stored in the node. Depending on this outcome, *y* moves down to the left or right child. When the sample reaches a left at the end of the tree, it is classified to the class label assigned to this leaf.

To select the feature on which to base the tree splits, a decision tree classifier chooses the features providing maximum information 'gain'. Mathematically, entropy may be calculated as

$$-\sum_{i}^{m} n_i \log_2(n_i)$$

where n_i is the faction of data points in class i and m is the total number of classes. Entropy is representative of the impurity of a subset of data. A homogenous dataset thus has an entropy of 0 and a randomly generated dataset, a maximum entropy of 1.

The information gain can then be equated to

$$information\ gain = entropy(parent) - [weighted\ average] \times entropy(children)$$

This way we deal with the noise before (parent) and after (children) the split of data, if the entropy decreases there is an overall information gain. A decision tree will, therefore, split using the feature having the highest information gain.

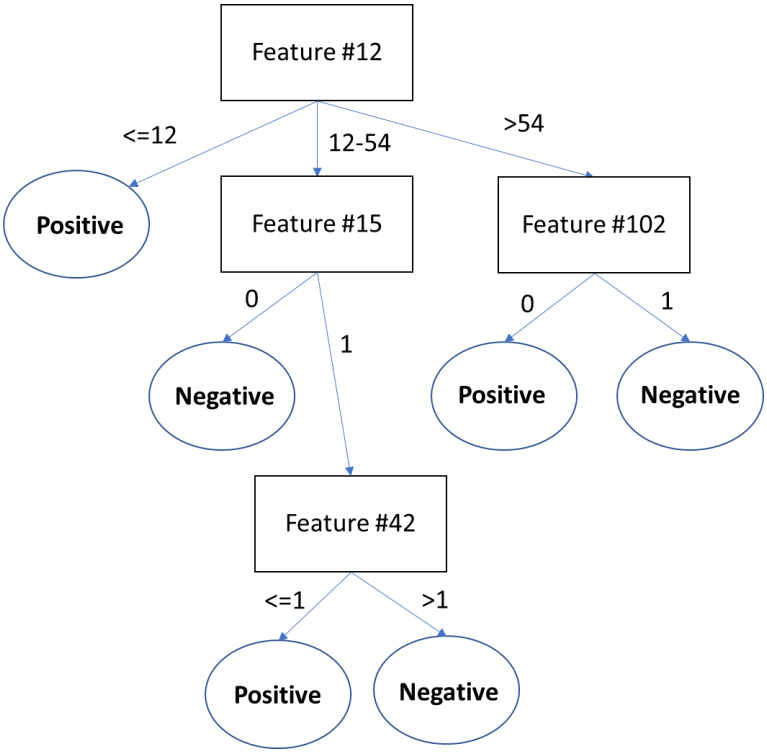


Figure 1.6: Basic working of a Decision Tree Classifier.

Decision trees are non-parametric and can model complex relations between endpoints and features without prior assumptions. They automatically choose the important features while modelling from all features provided in training which makes then to some extent, robust to variables which are less contribution or contain less information. In the same way, they are also robust to outliers and errors and are easily interpretable compared to other complex machine learning algorithms or deep neural networks.

1.3.4.2 Random Forest Classifier

Decision trees are prone to overfitting especially if too deep. The Random forest classifier is an ensemble of unpruned decision trees, each being trained independently using the training set and can be used in compound classification and modelling.[22] The ensemble method uses the concept of 'wisdom of the crowd' to make the final prediction based on majority voting of all the trees combined (see Figure 1.7). As decision trees are low in bias and usually have high variance, they are likely to benefit from ensembles. A training

sample is randomly selected with or without replacement (as determined by the bootstrap parameter) for each decision tree and grown with randomly selected features.

The trees are trained independently of each other, however not all data is used on all trees. Using a technique known as bagging, a different part of the dataset is used for different trees, with or without replacement of data. The correlation between trees is thus reduced. For predictions, a data point from the test set is passed through all trees independently. This results in individual predictions, where the simplest way is to assign the class that got the greatest number of votes from the individual trees.

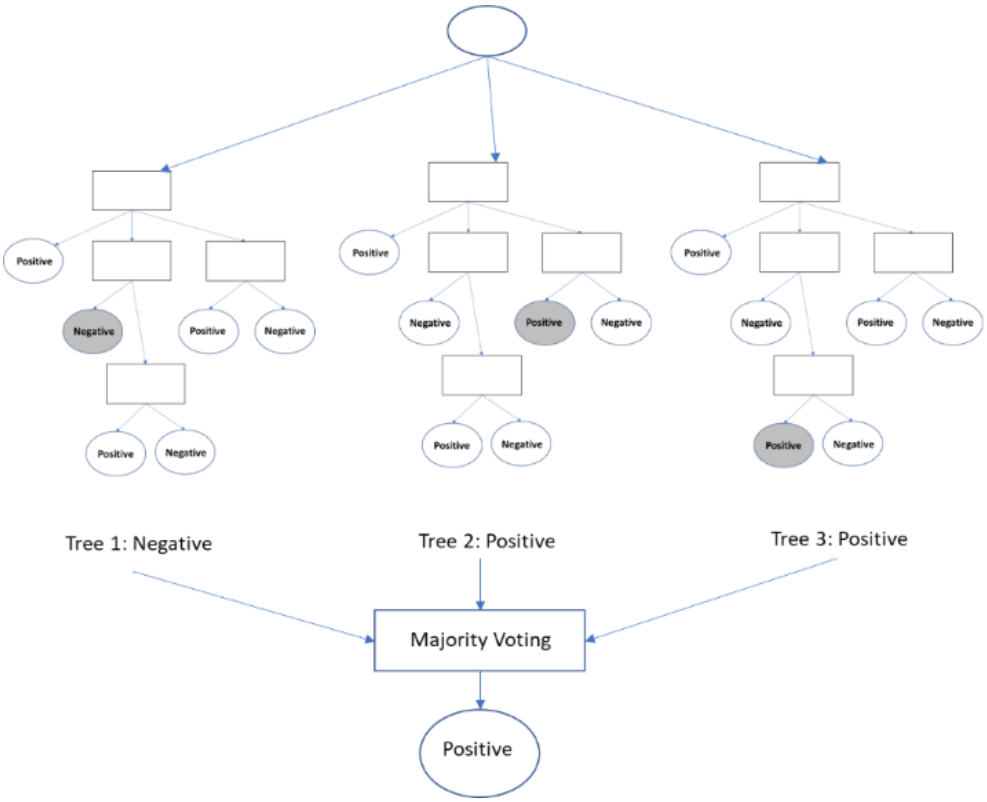


Figure 1.7: Visualization of a Random Forest Model Prediction. The training subset in each tree is generally two-third of the total training set used in a model. In the case of toxic/non-toxic classes, a compound will be labelled toxic if most of the individual decision trees predict the label as toxic.

.

The forest size can be adjusted to have a trade-off between accuracy and computational cost while the tree depth can be adjusted to keep a low value to reduce overfitting. The more the individual trees are decorrelated, the better the generalisation which can be achieved by bagging. In summary, random forests are fast in training and testing as the individual trees are independent of each other and can be run in parallel. Like with decision trees, these models can also be easily interpreted.

1.3.5 Cross-Validation Strategies

If a model were to learn and test on the same data, it would just repeat the predictions from the data it is already seen and thus have a perfect score. This case of overfitting would not allow the model to make useful predictions on unseen data. It is common in training models to split the entire data into training and test set. However, if only one such split is made, it would drastically limit the model performance of the nature of the random split, the performance would depend on the random seed used to split the data. To avoid this methodological mistake, one may use cross-validation. In *k*-fold cross-validation using shuffle stratified split, the training set is partitioned into *k* folds and a model is trained on *k-1* of these folds. The resulting model is then validated on the remaining fold. This process is repeated for each of the *k* folds and performance measures are the average of the values computed from each loop. A stratified shuffle split returns randomised folds with a percentage of samples of different classes nearly the same for each fold.

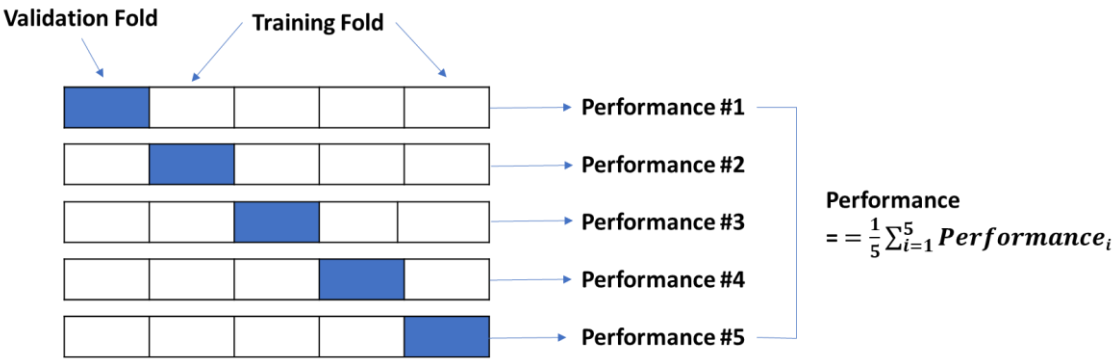


Figure 1.8: A 5-fold cross-validation diagram. The dataset was divided into five parts, and four of them are taken as training data in turn, and one as test data.

1.3.6 Nested Cross-Validation

Model validations allow us to assess the accuracy of a model. However just a train-test split is not always enough, and we often need a validation set. A held-out test set which is never used to adapt the models will help perform on unseen data – data that was neither used in training the model nor used in picking which model works best (see Figure 1.9). This is best realised using a nested-cross validation as used in this work. Nested cross-validation runs an inner and an outer loop.

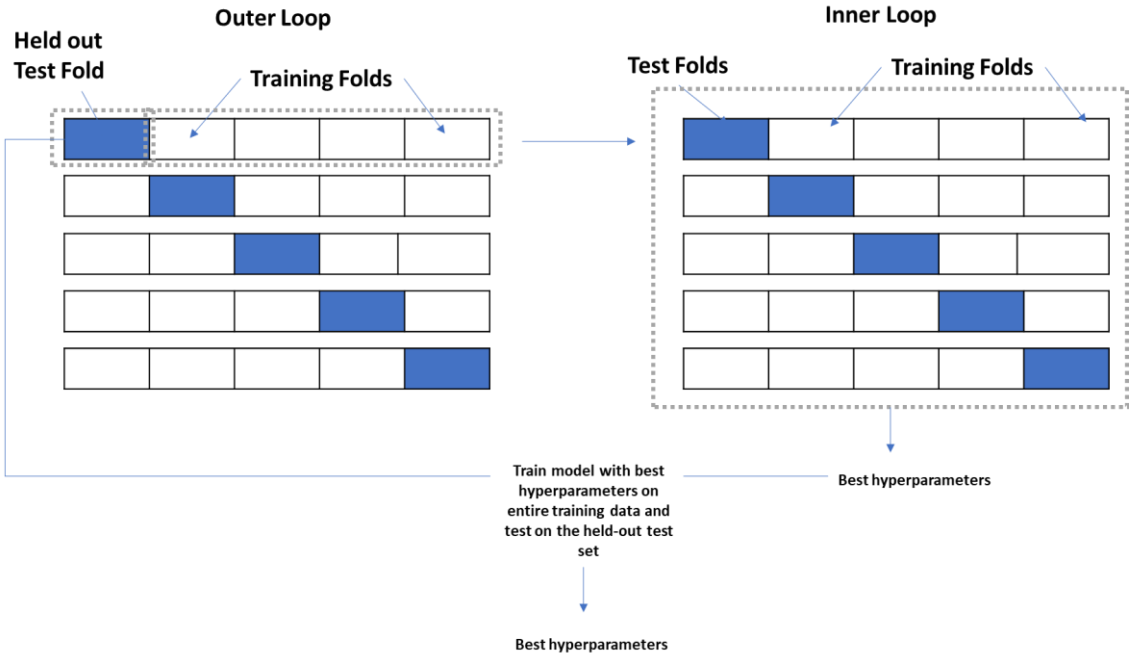


Figure 1.9: A nested cross-validation visualisation.

While the inner loop is used to train and tune the model as per the best results from different hyperparametric changes, the outer loop is used to test the winning model on the held-out data.[23]

1.3.7 Splitting Strategy

To avoid errors in modelling, it is essential to use a splitting strategy in each loop such that data is representative of the test data. This representation is generally in terms of its endpoints or feature space (in our case, the chemical space). Pure random shuffles do not guarantee us a proportional balance of classes or feature space distribution.

1.3.7.1 Random Stratified Shuffle

A stratified split ensures that the splits are proportional in the balance of classes. This ensures that the model training and testing are done on data having a similar distribution of classes. Ensuring representation of the actual data in both training and testing set provides greater precision and eliminates chances on unrepresented groups in either set.

1.3.7.2 Cluster-Based Splitting

After running a principal component analysis of the feature space, one can cluster the data points using k-means clustering. This algorithm selects collections of data that have certain similarities and groups them into a pre-defined number of clusters. When a cluster-based split in done, the held-out test set is least representative of the rest of the data on the features used to carry out the principal component analysis. In our study, this splitting was used to split the data into sets that are different from each other when considering certain physicochemical properties. Thus, learning from data points that are physicochemically different from the test set, we can see if our features can build a model that classify the endpoints with greater accuracy.

1.3.8 Evaluation Metrics

To evaluate classification model performances, we use the metrics such as of Sensitivity (SEN), Specificity (SPE), Balanced Accuracy (BA)[24], Area Under Curve- Receiver Operating Characteristic (AUC-ROC)[25]and Matthews Correlation constant (MCC)[26] as metrics. For the reader, we will first define the classification matrix before understanding what the metrics used in this study represent.

1.3.8.1 Classification Matrix

Several terminologies are defined to evaluate the prediction of a machine learning model. Among them, for binary classification they are:

- True Positive (TP): Where both prediction and true value are positive
- True Negative (TN): Where both prediction and true value are negative
- False Positive (FP): Where the prediction was positive although the true value was negative
- False Negative (FN): Where the prediction was negative although the true value was positive

A confusion matric simplifies the following as:

		Predicted	
		Negative	Positive
Actual	Negative	Number of true negatives (TN)	Number of false positives (FP)
	Positive	Number of false negatives (FN)	Number of true positives (TP)

Figure 1.10: The confusion matrix for a binary classifier

1.3.8.2 Sensitivity, Specificity and Balanced Accuracy

The ratio of actual positives correctly classifies is regarded as sensitivity (SEN) or true-positive rate (TPR) while the specificity (SPE) measures the proportion of correctly classifies negative instances. The false-positive rate (FPR) measures the ratio between false positives and the total number of actual negatives.

$$\text{SEN} = \frac{TP}{TP+FN} \text{ (also known as TPR)}$$

$$\text{SPE} = \frac{TN}{TN+FP}$$

$$\text{FPR} = \frac{FP}{FP+TN}$$

The balanced accuracy (BA) is defined as the average of the sensitivity and specificity.

$$\text{BA} = \frac{SEN+SPE}{2}$$

1.3.8.3 AUC-ROC

The AUC-ROC of a model is regarded as the probability that a randomly chosen positive example will be ranked higher by the model than a randomly chosen negative example.[27] Metrics such as sensitivity or specificity depend on a chose decision threshold that characterises each prediction into a class. When we compute the true positive rate and the false-positive rate for every possible decision cut off, we produce a receiver operating characteristic (ROC) curve.

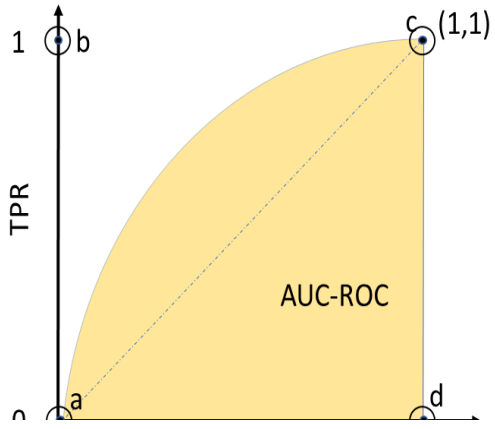


Figure 1.11: A ROC curve and points in the ROC space

Table 1.1 Important points and regions of an AUC-ROC curve

Point of interest	Coordinates	Comment
a	(0,0)	Cut-off=1 (All predictions are negative, no false-positive errors but also no true positives)
b	(0,1)	Best case scenario of perfect classification, TPR is maximal and FPR is minimal.
c	(1,1)	Cut-off=0 (All predictions are positive, no false-negative errors but also no true negative)
d	(1,0)	Worst case scenario, FPR is maximal and TPR is minimal
Dashed line		ROC curve for a random prediction, AUC-ROC=0.5
Shaded curve		The area under ROC curve (AUC-ROC)

This measure is thus independent of the decision threshold depending on which type of error is less hindering to the results, that is, sometimes sensitivity and specificity trade-offs are necessary. One point in the ROC space is better than another if the TPR is higher, FPR is lower, or both. ROC curves are useful for visualisation and evaluation as they provide more information than scalar metrics. It is considered harder to increase if the baseline model already performs well and thus will be a test of improvement in the model on varying fingerprints.[28] In practical, the increase of AUC-ROC score from 0.60 to 0.65 is the same as 0.80 to 0.85, however the latter is much difficult to achieve.

1.3.8.4 MCC

The MCC is used to evaluate the model's ability to predict for unbalanced classes.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The MCC shows a high score only if all the true positives, false negatives, true negatives, and false positives yield good results which can be used even if the size of both actives and in-actives differ in the dataset.

1.3.9 Feature Importance

It is ideal to be able to interpret a model rather than focus on the predictions alone. In those circumstances, knowing which features are the most helpful in predicting the endpoint may help interpret the model. In our work we use permutation importance, however, for comparison, we will also discuss in brief the Gini-index.

1.3.9.1 Gini Index

This is the most common mechanism in determining feature importance. The mean decrease in the impurity of a feature can be calculated by measuring its ability to reduce the uncertainty when building individual decision trees within the random forest.

Although this is a fast approach to average the scores over all individual trees, the Gini index tends to increase the importance of continuous or high-cardinality categorical variables. Additionally, Strobl et. al show that the variable importance measures using Gini-index are not reliable when features vary in their scale of measurement."[29] Hence in our work, we use permutation importance.

1.3.9.2 Permutation Importance

An alternative approach uses a baseline accuracy using the validation set through the Random Forest. Further, the column values of the features are permuted and then the test set is passed through the model again. The new accuracy is recorded. The permutation importance of this feature can now be regarded as the difference between the baseline importance and the drop in the accuracy metric after permuting values in the feature column.[30,31] The basic principle in working is that randomly permuting a feature column should break its relation with the endpoint column and if in reality the column was indeed associated to the endpoint, permuting will result in an increase in the error. Hence measuring the change would give us the increase in error, and the larger this increase the more important would be a variable (and vice versa).

Although this strategy is computationally expensive, it is considered more reliable. Hence, we use permutation importance in our work here.

1.4 Cytotoxicity assays as an endpoint

Predicting cytotoxicity remains one of the challenging steps in drug discovery. Cellular toxicity can be linked to a wide variety of mechanisms including cell death due to structural damage and cell stress leading to cell death.[32] Cytotoxicity and this is where it becomes particularly relevant in the context of drug discovery, can also be linked to organism level toxicity[33], such as hepatotoxicity.[34] Hence there remains a high possibility to use *in vit*ro models as biological descriptors into predicting *in vivo* toxicity in the future, at the least, in reducing compound attrition due to adverse effects found in later stages of the drug discovery pipeline.[35] Predicting cytotoxicity with high precision and accuracy holds great potential since cytotoxicity can not only influence the outcome of other biological endpoints but is also related to adverse endpoints thus benefiting drug discovery by not only reducing the time and resources required in high-throughput primary screens but also facilitate in interpreting the results for insight into underlying mechanisms.[36] From the practical angle, one would hope that cellular morphology is predictive of a wide range of efficacy and safety endpoints, given that in this case a single screen would be able to characterize a compound from multiple angles, and the current work hence explores one of the many possible avenues in this regard.

1.4.1 Computational Prediction of cytotoxicity

Previous models in cytotoxicity predictions have used many machine learning algorithms such as Random Forests[37,38] , Bayesian learning[39] and deep learning[40] based on physicochemical properties, molecular fingerprints and descriptors, as well as cell line descriptors of mRNA expression data.[41]

Table 1.2 Performance of several cytotoxicity prediction models reported in the literature

Algorithm	Method of Validation	Fingerprint	The ratio of Active: Inactive	Accuracy	SEN	SPE	AUC	Reference
RF	5-fold CV	MACCS	49:3262	0.76 ± 0.03 (Q)	0.82 ± 0.15	0.76 ± 0.03	0.85 ± 0.08	Yin et al.
SVM	Separate training and test data	4D-Fingerprints	57:1243	0.45 (Q)	0.42	0.65		Chang et al.
RF	5-fold CV	Cytotoxicity	92:257	0.67 ±0.06 (BA)	0.40 ±0.10	0.93 ±0.03		Allen et al.
Bayes	5-fold CV	Scitegic FCFP_6	varied		0.57 ± 0.04	0.71 ± 0.01		Langdon et al
Naïve-Bayes	5-fold CV	FeatMorgan (radius = 2A and 2048 bits) and ToxPrint	varied	0.69 (BA)	0.67	0.71	0.76	Schrey et al.
ANN	Separate training and test data	Atomic7 descriptors	3336 :4962 (train) 823 :1177 (test)	0.73 (Q)				Molnár et al.
RF	Separate training and test data	BCI fingerprints	276:499	0.68 (BA)	0.56	0.80	0.73	Guha and Schürer
RF	Aggregated Conformal Prediction	RDKit Physiochemical descriptors	48:3247	0.70 (BA)	0.74	0.65		Svensson et al
FNN		Morgan fingerprints	465:10000	0.69±0.01 (BA)	0.62±0.07	0.76±0.07		Webel et al.

AUC, area under the curve; ANN, artificial neural network; BA, balanced accuracy; CV, cross-validation; FNN, feedforward neural network; Q, overall predictive accuracy; RF, random forest; SEN, sensitivity; SPE, specificity; SVM, support vector machine.

Although the datasets differ which makes the direct comparison of performances difficult, the best models developed by Yin et al. using Random Forest ensembles on MACCS fingerprint achieved an area under the curve (AUC) of 0.85 in 5-fold cross-validation.[37] Chang et at. used a Support Vector Machine (SVM) model constructed from the full PubChem dataset AID 426 using over-sampling with active-to-inactive ratios of

1:1 for predicting the toxicity of combined PubChem datasets of AID364 and AID463 and achieved an accuracy of 0.45 using the best classifier.[42] Other studies such as those of Langdon et al.[39] and Schrey et al.[43] used Bayesian algorithms due to advantages over data imbalance. Schrey et al. used a Naïve-Bayes algorithm and 10-fold cross-validation using cytotoxicity assays on different types of immune cells in an attempt to predict immune cell cytotoxicity.[43] In recent years, models have started to use also other high-dimensional readouts for toxicity prediction, such as by combining chemical structural descriptors with gene expression data.[44]

1.5 Cell Painting

Biological endpoints, such as -omics data and high-dimensional biological data, have become easier to obtain and more applied in recent year, such as from the transcriptomics L1000 platform[45] as well as in the form of the Cell Painting[46] assay from the cell morphology side. The image-based profiling assay, Cell Painting from Broad Bioimage Benchmark Collection (BBBC) is one of the largest datasets comprising human osteosarcoma cells (USO2). The morphological profiling of cells subjected to perturbations collects a large number of geometric features which can then be further analysed and correlated to particular endpoints of interest. Given the general nature of the readout, this includes, for example, target identification, lead hopping, library enrichment, and functionally annotating genes (among others).[47]

1.5.1 Image-based profiling

The Cell Painting assay uses a mixture of six well characterised fluorescent dyes to stain specific major organelles and sub-compartments in human USO2 cells. Live-cell staining was performed in the recently published dataset with 30,616 different small molecules to characterise the effects on the nucleus, nucleoli and cytoplasmic RNA,

endoplasmic reticulum, Golgi apparatus and plasma membrane, and the actin cytoskeleton of human U2OS cells, with respect to DMSO control.[48]

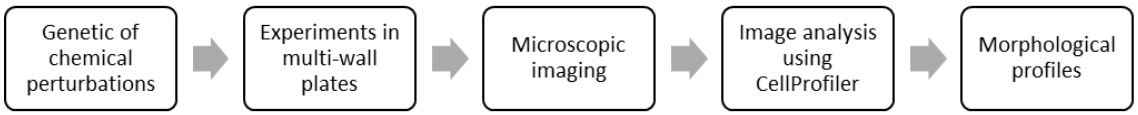


Figure 1.12: Schematic Representation of morphological profiling using an image-based assay to extract morphological features of each cell. The microscopic images were processed using a three-step pipeline workflow using open-source software Cell Profiler, the extracted features which can be used for classification comprise various cellular shape and adjacency statistics and intensity and texture statistics for each channel.

Of the 30,616 compounds, 10,080 came from the Molecular Libraries Small Molecule Repository (MLSMR), 2260 were drugs, natural products, small-molecule probes among the bioactive molecules identified by the Broad Institute, 269 were confirmed screening hits from Molecular Libraries Program (MLP) and 18,051 were diversity-oriented synthesised novel compounds.

The Cell Painting dataset consists of processed data from the images, the analysis was done using a three-step pipeline workflow using the open-source software CellProfiler.[49] The non-uniform light source in a microscope causes each image to have a non-homogeneous illumination. It is thus necessary to recover the true image from this distorted image. The first step involves estimating the heterogeneities in the microscope optics induced spatial fluorescence. Secondly, quality control pipeline labels images with aberrations such as focal blur, saturation artefacts etc. In the final step, illumination correction functions are used to correct each channel, identify the nuclei, cell body and cytoplasm. The quality control used supervised machine learning to identify images using the CellProfiler Analyst software package which additionally normalises the total number

of cells using median per-image cell count thus preventing any bias due to densely populated slides.[50] The phenotypic characteristics of each cell are measured in this feature extraction step. The extracted morphological features are then deposited in a database for interpreting and validation patterns in the profiles. This downstream analysis of these features then allows understanding biologically meaningful correlations.

The major types of features profiled in Cell Painting assay involve shape, intensity-based, texture and microenvironment and context features. Shape features such as size, perimeter, area, roundness is computed on boundaries of nucleus, cells, cytoplasm or other sub-compartments. Intensity-based features are computed from actual intensity values per channel such as mean intensity and maximum intensity. These features are calculated on a single cell basis within each compartment within each compartment. Texture features are exclusively used for single-cell analysis and quantify the regularity of intensity in their images. Using a range of mathematical functions such as cosines and correlation matrices, these features can detect the periodic changes in the cell intensity. The microenvironment and context features include counts and spatial relationships among cells. Calculated over the field of view defined using the number of and distance to cells in a neighbourhood or the relative position to a cell colony, these features not only include segmented regions of nuclei and cells but also subcellular structures (for example distances between the nucleus and individual cytoplasmic vessels).

Based on this dataset, many protocols have been developed for using these image-based features.[51] Among them, one of the simplest ways suggested averaging the cellular features across each cell to produce a morphological profile for each sample (which of course would not consider readout distributions, and hence cellular subpopulations, in detail). When predicting the mechanism of action (MOA), Ljosa et al. used average measurements over the cells thus having single value for each feature and using supervised led to the prediction the MOA with an overall accuracy of 83%.[52] The

complete set of morphological features can be interpreted as a morphological fingerprint that is characteristic for the MOA (or even more generally the biological activity) of a certain small molecule, at least to the extent the activity is visible in morphological readout space.[53] Trapotsi et al. compared chemical (extended connectivity fingerprints) and cell morphology (Cell Painting) information for bioactivity prediction.[54] Their work using multitask Bayesian Matrix Factorisation (BMF) approach Macau could predict around 45% and 40% of 224 targets using ECFP and image data respectively with high AUC-ROC (>0.80) thus showing that the two descriptors contain partially complementary information. Hence, cell morphology holds the potential to provide information related to compound action to a wide variety of endpoints, be they efficacy- or safety-related, with the potential to enable compound profiling for a wide variety of purposes.

In recent years, many studies have used images and image-based morphological fingerprints for several machine learning tasks for diverse applications in the field of chemoinformatics. Hofmarcher et al. were able to achieve high predictive performances with AUC-ROC scores above 0.90 for biological assays using convolution neural networks trained on microscopic images of cells in the Cell Painting assay.[55] However, training convolution neural networks on images are computationally expensive. An alternative approach adopted by Simm et al. extracted information from microscopy-based screen specifically designed for glucocorticoid receptor nuclear translocation.[56] The repurposed data was used to train machine learning models to predict assay-specific biological activity. Their results indicate a 60-fold to a 250-fold increase of hit rates over the initial high-throughput screens for two drug discovery projects. This shows that extracted information in the form of image-based fingerprints could be repurposed to predict biological activity on a targeted basis. Gustafsdottir et al. presented using Cell Painting profiles to cluster compounds with similar annotated protein targets or chemical

structures.[57] The results established that even though the targets were not stained directly, the models were able to detect patterns in morphological labels related to the on-target activity. Persson et al. use multi-parametric imaging of cell health to model drug-induced liver injury and bile salt transport inhibition, both cause adverse effects in the human system but unfortunately cannot be readily detected in animal models in pre-clinical testing.[58] In such cases, high content screening in certain cellular assays shows better productivity than animal toxicity models.

1.5.2 Applications of morphological profiling with transcriptomic data

Transcriptomic data have been used before to annotate protein targets and chemical structures. Wawer et al. used both cell morphological profiles and gene expression to derive interpretable structure-activity relationship rules with several top-scoring rules were identified by both.[59] Further corresponding changes in morphology can also be linked to transcriptomic changes as demonstrated by Nassari et al. using a cell morphology enrichment analysis to identify sets of landmark genes associated with cell morphology changes in response to perturbations.[60] Rohban et al. validated how disease-associated alleles and genes can be annotated using morphological profiling of cDNA via a Cell Painting assay.[61] They showed that morphologic phenotypes are potent indicators of cell state by using novel subpopulation-based visualization methods and successfully grouping biologically meaningful clusters of genes consistent with known functionality. Wawer et al. demonstrated using small molecule profiling to enrich large compound libraries from diverse biological performance and having high rates of activity, in the process also showing that cell morphology and gene expression profiling are not redundant in contained information.[62] Lapins et al. compared the performances of machine learning models trained on gene expression, cell painting assay data and *in vitro* assays by chemical structure descriptors to predict mechanisms of action and drug

targets.[63] Their work shows for some targets, only one of either gene expression or cell morphology was able to model with AUC exceeding 0.70, thus suggesting that the two profiles contain distinct information. Hence, these studies not only validate the existence of meaningful information in image-based profiling studies but also their applicability to areas such as target identification and mechanism analysis to toxicity prediction.[64]

1.6 Aim of this research

Particularly what closely guided the motivation for this study was Martin et al. who demonstrated the application of using cell morphology and cell proliferation markers in identifying cytotoxic compounds.[65] They developed a staining and image analysis protocol which was applied to a novel chemical library consisted of 329 chemically diverse compounds. Using the changes observed in nuclear morphology, cell shape and proliferation, they were able to identify compounds having adverse cellular effects such as cell loss, sub-lethal alteration to cell morphology or cell proliferation with hit rates of 10.0% and 3.6% respectively. These rates although lower than other studies,[66] are however used non-liver derived cell line to examine cytotoxicity by high content imaging in an unenriched chemical library which is novel. Their study further suggests that cell morphology and, in particular, nuclear morphology can be used to identify adverse cellular effects, which we explore further here using Cell Painting readouts to predict cytotoxicity- and proliferation-related endpoints. Given the wide variety of mechanisms which can lead to cytotoxicity,[36] it seems plausible that single biological endpoints (for example, based on targets) have insufficient coverage to predict this endpoint, which of course can also be applied to other higher-level effects, such as adverse events on an organism level etc.

Hence we concluded for our work that modelling for *in vitro* cytotoxicity-related or reduced proliferation assays related to different cell lines, biological processes and

biological targets will allow us to assess the performance of the Cell Painting assay in detecting cytotoxic effects in a wider variety of cell lines and relate cell morphology changes and physicochemical properties of compounds with different cytotoxicity mechanisms. Efforts in this regard are mainly disadvantaged by lack of data and understanding of cell morphology features and its correlation with toxic endpoints.

In this proof-of-concept study, we hence investigated the use of Cell Painting assays in predicting the outcomes of cytotoxicity- and proliferation-related *in vitro* assays. Additionally, we also determined if the combination of Cell Painting with other physicochemical and structural fingerprints aid in predicting the outcomes of cytotoxicity- and proliferation-related *in vitro* assays, with the overall concept being generally applicable across endpoints related both to safety/toxicity and efficacy.

2 METHODS

This chapter introduces the datasets, procedures and workflows for the theoretical methods employed in our work. Our results and discussion on the same are elaborated in chapters 3 and 4.

2.1 Introduction

We introduced a workflow as a tool to use correlated Cell Painting morphological data as input features to predict cytotoxicity-related or reduced proliferation assays perturbated with small compounds and use machine learning algorithms to determine if Cell Painting assays and cytotoxicity are related, aiming to achieve reliable and high cytotoxicity predictive performances. Additionally, we also determined if the combination of Cell Painting with other physicochemical and structural fingerprints aid in predicting the outcomes of cytotoxicity-related or reduced proliferation *in vitro* assays.

2.2 Dataset

For this work, we use Cell Painting datasets and *in vitro,* cytotoxicity-related or reduced proliferation assays as described below the overlap of which was constructed as follows, which led to 135 unique compounds distributed in 10 assay endpoints.

2.2.1 Cytotoxicity and Proliferation Annotations

Endpoints annotations were obtained from the MoleculeNet Toxcast database.[67,68] The MoleculeNet database performed pre-processing of Toxcast *in vitro* data and includes qualitative binary results based on the original *in vitro* high-throughput readouts.[69] Cell death (and proliferation modulation) can be a consequence of various underlying mechanisms and in this proof-of-concept study, we attempt to study 10 cytotoxicity- and proliferation-related *in vitro* assays from 3 different technological platforms, namely ACEA Biosciences (ACEA), Apredica (APR) and BioSeek (BSK) which were meant to capture a range of different cell lines and biological processes in the assays (for details see Table 2.1).

Table 2.1 Cytotoxicity- and proliferation decrease related assay endpoints from ToxCast included in this study (see Table 2.2 for dataset sizes)

Assays	**Biological process**	**Tissue**	**Cell line/ Cell type**	**Assay Type (Assay endpoint)**
BSK_3C_SRB_down	Cytotoxicity SRB	Vascular	Umbilical vein endothelium	Protein content (cell death)
BSK_4H_SRB_down	Cytotoxicity SRB	Vascular	Umbilical vein endothelium	Protein content (cell death)
BSK_LPS_SRB_down	Cytotoxicity SRB	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells	Protein content (cell death)
ACEA_T47D_ 80hr_Negative	Proliferation decrease	Breast	T47D	Real-time cell growth kinetics (cell proliferation)
APR_HepG2_ CellLoss_72h_ dn	Proliferation decrease	Liver	HepG2	Cell number (cell death)
BSK_3C_ Proliferation_ down	Proliferation decrease	Vascular	Umbilical vein endothelium	Protein content (cell proliferation)
BSK_3C_Vis_down	Proliferation decrease	Vascular	Umbilical vein endothelium	Cell phenotype

				(cell morphology)
BSK_CASM3C_Proliferation_down	Proliferation decrease	Vascular	Umbilical vein endothelium and coronary artery smooth muscle cells	Protein content (cell proliferation)
BSK_hDFCGF_Proliferation_down	Proliferation decrease	Skin	Foreskin fibroblast	Protein content (cell proliferation)
BSK_SAg_Proliferation_down	Proliferation decrease	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells	Protein content (cell proliferation)

The cell-growth kinetics assay of the ACEA technological platform using the T47D cell line is sensitive to estrogen-receptor (ER) agonists, measuring cellular events further downstream than ER transcriptional activity assays, such as cell morphology, cell number or cell-cell adhesions using an impedance-based technique.[70] When analysed in the loss of signal direction, they can be utilized in the purpose of cell viability based on cell-growth kinetics.[71] The BSK platforms use multiplexed-readout assays. The assays selected in this study from BSK were run using various human primary cells including dermal fibroblasts (hDFCGF), endothelial cells (3C, SAg) and smooth muscle cells (CASM3C)[72] and measure endpoints including sulforhodamine B (SRB) staining for cell density and total protein and a categorical morphologic visualization score.[73] For example, the endpoint BSK_3C_Vis_down measures the decrease in visual activity in umbilical vein endothelial cells, thus aiming to characterize cellular health. Another two assays, the BSK_3C_Proliferation_down and the BSK_SAg_Proliferation_down are anti-proliferation assays measuring the level of downstream response proteins for cell proliferation in umbilical vein endothelium cells in an enzyme-linked immunosorbent assay (ELISA).[74,75] The selected high-content screening (HCS) multiparametric

cytotoxicity assay of APR at 72h time point evaluates the cellular markers of cell loss for chemicals in cell culture of human hepatocellular carcinoma cell-lines (Hep G2).[76] The selected assays thus correspond to cell death, cell proliferation, cell viability and cell morphology endpoints and cover a range of biological processes.[77]

2.2.2 Cell Painting

A dataset of morphological profiles of small molecule perturbations from the Cell Painting assay was used for this study.[78] The assay contains cell morphology readouts including intensity, texture and adjacency statistics (among others) generated using Cell Profiler. A description of feature nomenclature can be found in the Cell Profiler manual.[79] Using the protocol followed by Lapins et. al[63] the Cell Painting data was split into control samples and treated samples and sorted by plate number. Most compounds on the dataset had been used to treat cells multiple times, thus giving sets of values for each feature. On a plate-by-plate basis, each value of the features was centred by subtracting the mean of the control samples from the treated samples. Finally, the mean of all values of a particular feature was calculated thus giving us a single for each of the features. Since random forest models are tree-based model, they do not require feature scaling (feature normalization) but the absolute values for branching, hence no scaling or normalisation was attempted further.

2.3 Data Preparation

For both the ToxCast and Cell Painting datasets, molecular representations in SMILES format were standardized and canonicalized to the parent molecular form using the MolVS standardiser[80], an open-source toolkit based on the RDKit[81], including tautomer standardisation. For the complete set of standardization rules using the functions sanitization, normalisation, largest fragment chooser, charge neutralisation, tautomer enumeration and canonicalization implemented in the MolVS tool, the reader is referred

to the MolVS standardizer guide.[82] The overlap between the Cell Painting data and the MoleculeNet ToxCast was determined in terms of canonical SMILES. The final dataset comprised 135 unique compounds across 10 assay endpoints. Not all chemicals were present in all assays. In this study, endpoints selected were relatively balanced (between 27% and 49% actives as shown in Table 2.2, Figure 2.1) with all individual assays having more than 100 compounds except for one assay having 88 compounds.

Table 2.2 Description of the dataset.

Assay	Actives	In actives	Total Compounds	% of actives
BSK_3C_Proliferation_down	52	54	106	49.06%
APR_HepG2_CellLoss_72h_dn	43	45	88	48.86%
BSK_hDFCGF_Proliferation_down	50	56	106	47.17%
BSK_SAg_Proliferation_down	48	58	106	45.28%
BSK_3C_SRB_down	45	61	106	42.45%
BSK_3C_Vis_down	41	65	106	38.68%
BSK_LPS_SRB_down	39	67	106	36.79%
BSK_CASM3C_Proliferation_down	36	70	106	33.96%
BSK_4H_SRB_down	34	72	106	32.08%
ACEA_T47D_80hr_Negative	34	90	124	27.42%

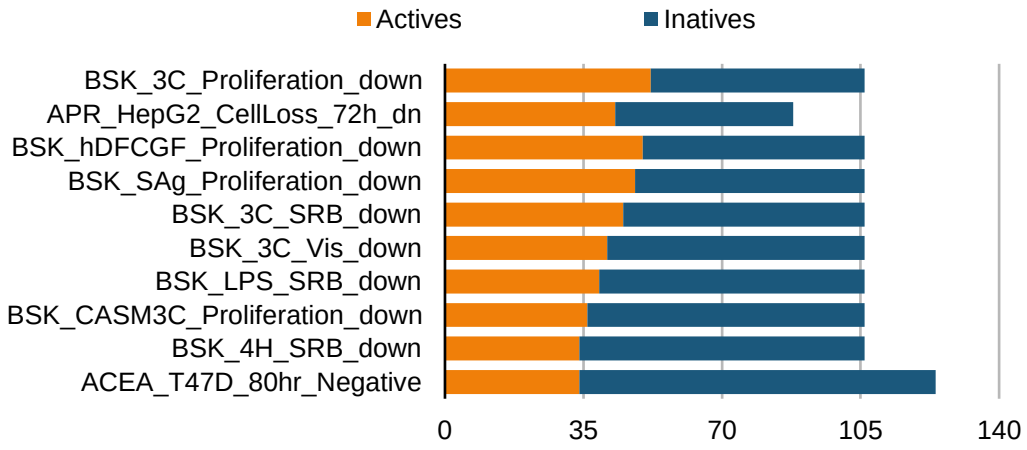


Figure 2.1: Distribution of active and inactive compounds among selected endpoints.

For the purpose of a proof-of-concept study, extrapolating dose/concentration as well as physiologically based pharmacokinetic (PBPK) modeling have not been accounted for in this study. In future, we hope to extrapolate the dose/concentration of Cell Painting assays to those of endpoints accounting them while interpreting our models. Machine learning is often highly dependent on the size and quality of the dataset, and the relatively smaller size of our assays is indeed a challenge to overcome in future. It should be acknowledged here that the small dataset size is indeed not ideal to evaluate prospective model performance; however, we have here performed a relative analysis of model performance, which we believe is still valid also on this dataset. Further limitations of this study have been discussed in our results and discussions section.

2.4 Descriptors and Fingerprints

2.4.1 Molecular and Morphological

We used in total, five different descriptors as input features (including combinations), (i) Morgan fingerprints, (ii) ErG fingerprints[17], (iii) Cell Painting after feature selection (described in sub-section '*Cell Painting Feature Selection'*), (iv) a combination of Morgan fingerprint and selected Cell Painting features and (V) a combination of ErG fingerprints and selected Cell Painting features.

Molecular fingerprints of (i) Morgan fingerprints of 2048 bit and radius 2 and (ii)extended reduced graph (ErG) fingerprints were calculated from the standardised SMILES using RDKit.[81] Morgan fingerprints belong to the group of circular fingerprints which have performed well in previous benchmarks[83] , while the ErG fingerprint uses a pharmacophore-type representation.

For cell morphological features, (iii) Cell Painting after feature selection (described in *'Cell Painting Feature Selection'*) were used.

Finally, Morgan and ErG fingerprints for each compound was combined with the selected Cell Painting features by appending them. Since the order of features does not affect random forest algorithms, such combination was used to create (iv) a combination of Morgan fingerprint and selected Cell Painting features and (V) a combination of ErG fingerprints and selected Cell Painting features.

2.4.2 Feature Selection for Cell Painting

In order to prevent overfitting, we implemented feature selection from all the cell painting features as a part of the hyper-parameter tuning process. This method used the Kolmogorov-Smirnov (KS) test[1919] as implemented in SciPy[84] to identify significant feature frequency differences between individual features and endpoint class and retained all features with p-values below 0.02. In each run of the outer loop of the nested cross-validation (described in *'Model training'*), feature selection was performed based on data in *k-1* folds that are used to optimise hyper-parameters of the model. Our feature selection did not use a preconceived understanding of cell viability-related mechanisms to select important features but is rather an entirely data-driven approach.

2.5 Model training

Random Forest (RF) classifiers were trained for each endpoint using nested cross-validation consisting of an outer and an inner parameter optimisation loop (see Figure 2.2 for a schematic representation). We used the python library scikit-learn [85] (sklearn.ensemble.RandomForestClassifier) for this purpose.

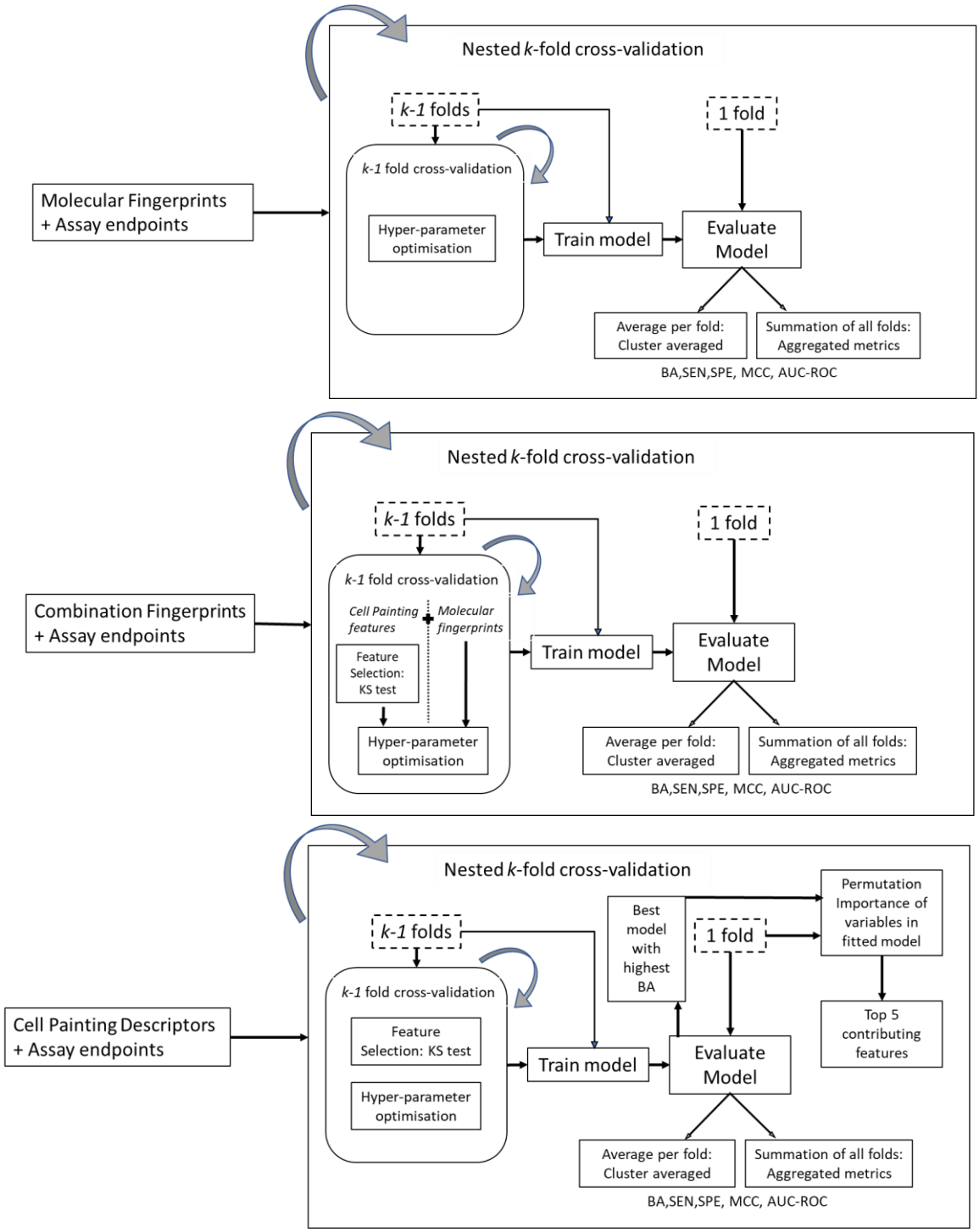


Figure 2.2: Schematic representation of nested cross-validation workflow, for three different inputs domains: molecular fingerprints, cell morphology and combination fingerprints. The models consist of an inner and an outer loop comprising the nested cross validations. The automatic feature selection takes place for only cell painting features inside the inner loop. The permutation importance is calculated for the model using cell painting data for the best performing folds.

Evaluation is done either on average of each held-out set or by aggregating results for all held-out set

For each outer loop iteration, data was split into *k* folds and one of the folds was reserved as a held-out test set. The remaining *k-1* folds combined as a training set were passed into the inner loop for hyper-parameter optimisation using grid search (from sklearn.model.selection, see Tables 2.3 for the hyperparameter space).

Inside this inner loop, each parameter was optimised using 5-fold cross-validation; for models using Cell Painting features, KS-test was used for feature selection (as described previously in 'Cell Painting Feature Selection'). Finally, the model was trained on the data in the entire of the inner fold (that is, the *k-1* folds) using the best parameters and tested on the held-out test set (that is, the k^{th} fold) which provided an unbiased evaluation of the model. This process was repeated *k* times, once for each iteration in the outer loop, until the entire data set was treated as a test set.

Table 2.3 Hyperparameter search spaces for the random forest model.

Algorithm	**Parameters Space**
RF	{'max_depth': [10, 15, 20], 'min_samples_leaf': [3, 6, 12, 15], 'min_samples_split': [6, 9, 12, 15], 'n_estimators': [100, 200, 300, 700], 'criterion': ['gini', 'entropy'], 'class weight': [None, 'balanced']}

We trained two types of models based on the splitting of data into *k* folds in the outer loop, (a) *random shuffling splits* (using 5-fold shuffle stratified splitting), or (b) *cluster-based splits (using* a group 5-fold splitting leaving out one cluster). (For one of the assays, ACEA_T47D_80hr_Negative assay, we used 4-fold splits in place of 5-folds for the outer loop to ensure folds having a minimum of 10 data points.)

The *random shuffling splits* do not disadvantage any fingerprint/descriptor in particular, and hence can be considered as a comparison for the models that are in this way unbiased; however, it does not necessarily require extrapolation to novel chemical space. This is the advantage of the *cluster-based splits* on the other hand; however, this type of split disadvantages ErG/Morgan fingerprints to a certain extent, as the chemical space that is used for clustering is also related to the chemical features used for classification. Hence, we have chosen to perform both types of the split in the current study to evaluate model performance.

For the outer loop on (a) *random shuffling splits*, a 5-fold shuffle stratified splitting of data into train/test ensures randomly splitting such that test sets contain the same distribution of classes, or as close as possible. In the other type of model, for a realistic validation, (b) *cluster-based splits*, where a group 5-fold splitting leaving out one cluster was used. For cluster-based compound splits for each assay, we calculated for each compound the following properties using RDKit[81]: molecular weight, topological polar surface area, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors and the partition coefficient between octanol and water (log P). The first two principal components PC1 and PC2 were calculated and normalised. Finally, we explicitly split our sample into 5 clusters using k-means clustering using the python module of scikit-learn[85] based on the two principal components. Hence, here 5 folds of the split were the 5-clusters.

The explained variance using Principal component analysis (PCA) and silhouette scores for clusters among selected endpoints are shown in Table 2.4. A higher the silhouette scores determine a better cluster distribution and the PC1 and PC2 explain around 72 % of the variability of our datasets. Hence the split based on physicochemical descriptors selected varying chemical space for the folds however was not disadvantage molecular fingerprints directly.

Table 2.4 Explained variance using PCA and silhouette scores for 5 clusters among selected endpoints.

Assay	PCA Explained Variance (%)	Mean Silhouette Score
BSK_3C_Proliferation_down	72.1	0.40
APR_HepG2_CellLoss_72h_dn	72.0	0.40
BSK_hDFCGF_Proliferation_down	72.1	0.40
BSK_SAg_Proliferation_down	72.1	0.40
BSK_3C_SRB_down	72.1	0.40
BSK_3C_Vis_down	72.1	0.40
BSK_LPS_SRB_down	72.1	0.40
BSK_CASM3C_Proliferation_down	72.1	0.40
BSK_4H_SRB_down	72.1	0.40
ACEA_T47D_80hr_Negative	72.7*	0.36*

(*For the assay, ACEA_T47D_80hr_Negative assay which was a real-time cell growth kinetics proliferation decrease assay for the human breast cell line, only 4 clusters were used based on PCA.)

2.6 Model Evaluation

For the first type of model using (a) *random shuffling splits*, an aggregated metrics was used where the results of all the held-out test sets were aggregated to give predictions of the entire dataset. The aggregated metrics is preferred here as the model had significant exposure to the entire chemical space in the training set. On the other hand, for the second type of model using (b) *cluster-based splits*, a cluster-averaged metrics were evaluated using the mean and standard deviation of a metric for each fold. Here the cluster-averaged metric is preferred over aggregated metrics since clusters are not of equal size but should be given equal weight in the performance evaluation.

A stratified random shuffle on 5-fold split does not disadvantage any model or fingerprint and hence can be considered as a true comparison for the models. However, the cluster-based splits disadvantage ErG/Morgan fingerprints as the chemical space are distanced from each other for these splits.

2.7 Evaluation Metrics

To evaluate our models, we used the Sensitivity (SEN), Specificity (SPE), Balanced Accuracy (BA), Mathew's Correlation constant (MCC) and Area Under Curve- Receiver Operating Characteristic (AUC-ROC) as metrics.

BA and MCC evaluate the model's prediction ability with a different focus on aspects of the confusion matrix, with the MCC considering all four quadrants in the same way, and which hence can be seen, from this angle, as an 'unbiased' performance metric, while the AUC-ROC[2727] is a global performance measure, which may not be required when for example, predicting the toxicity of molecules. In any case, performance measures all have their characteristics, and hence we here aim to provide a wide set of them to evaluate models.

2.7.1 Variable importance of Cell Painting feature

We further evaluated the feature importance from the models using *random shuffling splits and aggregated metrics* and Cell Painting as input features.

Our interpretation of feature importance using Cell Painting fingerprints used permutation importance[30,31] rather than the popularly used Gini-index, given that the latter tend to inflate the importance of continuous or high-cardinality categorical variables.[86] In addition, features that seem relevant to the training set but are not on the held-out set cause a model to overfit. The permutation importance can be used on a held-out test set and measures the drop in overall accuracies when a column is permuted. For each iteration of the nested cross-validation, we used 10-fold permutation importance using the held-out test set and the fitter hyper-parameter optimised random forest model. This enabled to highlight most-contributing features to the generalization power of the random forest model. Thus, for each endpoint, we determine the top 5 features from the best performing held-out test set based on balanced accuracy and having the positive permutation scores. It is customary to note that the permutation test was hence used in conjunction with the

feature selection method (KS test to select relevant features initially to select for model training). The permutation importance hence provided a finer tuning to detect important features, that is, while the KS test selected features from the training data of the model, the permutation importance was evaluated using this model for data in the held-out test set.

The Python scikit-learn[85] library contains implementations of random forest models as well as hyperparameter optimisation, model training, model evaluation and permutation feature importance used in our work.

3 RESULTS AND DISCUSSIONS

This chapter discusses the results obtained in this study.

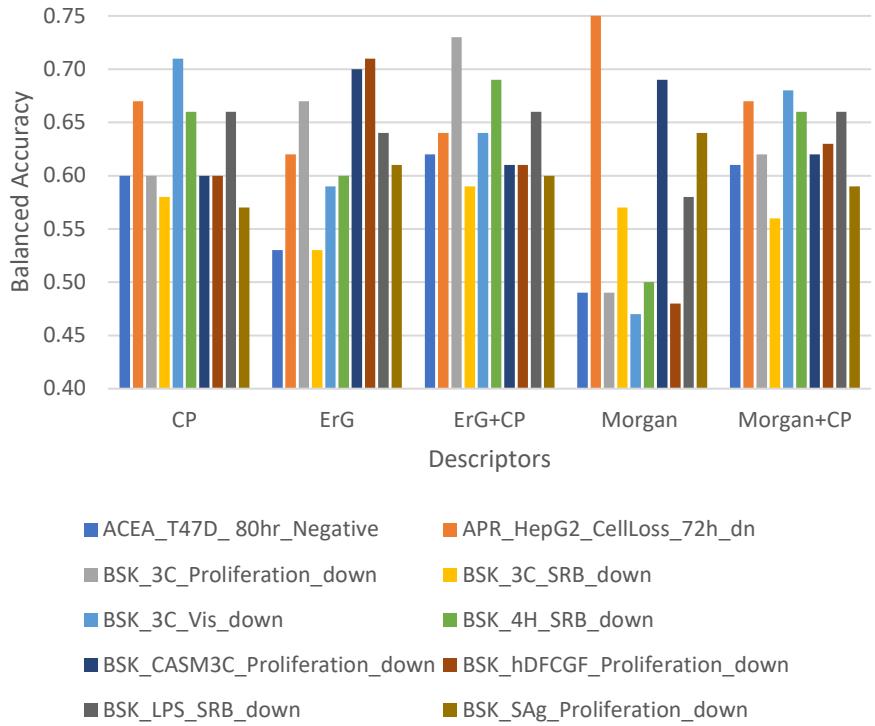
3.1 Introduction

For each of the 10 cytotoxicity- and proliferation-related *in vitro* assays endpoints, we constructed models using Cell Painting and two types of molecular fingerprints and their combinations and evaluated the metrics for the 5-fold and leave one cluster out methods, the results of which are shown in Table 3.1 and Figure 3.1-3.3 (for further details see Appendix A and B).

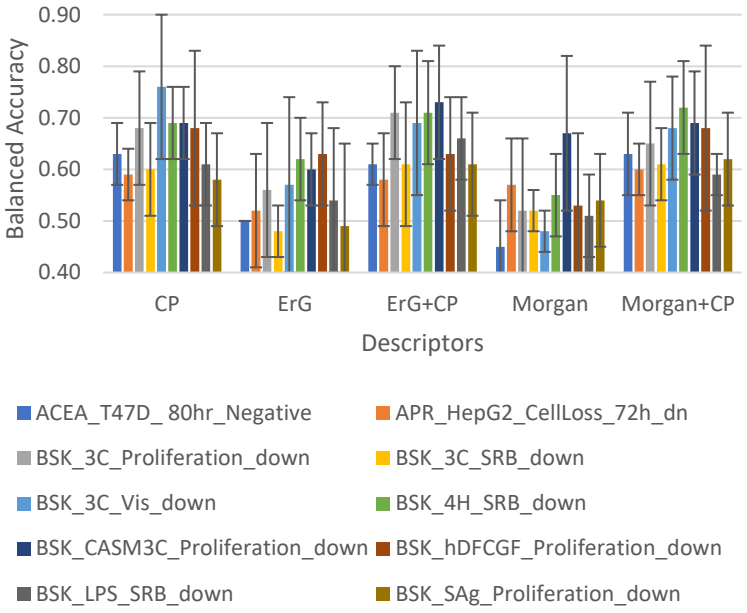
Table 3.1 Performance metrics for each cytotoxicity assay for different fingerprints.

Target	Fingerprint	5-fold outer split (aggregated metrics)			Leave one cluster split (cluster averaged)		
		BA	MCC	AUC-ROC	BA	MCC	AUC-ROC
ACEA_T47D_80hr_Negative	CP	0.60	0.295	0.63	0.63 ± 0.06	0.28 ± 0.14	0.74 ± 0.13
	ErG	0.53	0.105	0.59	0.50 ± 0.00	0.01 ± 0.01	0.47 ± 0.11
	ErG+CP	0.62	0.352	0.65	0.61 ± 0.04	0.26 ± 0.08	0.76 ± 0.13
	Morgan	0.49	-0.079	0.61	0.45 ± 0.09	-0.05 ± 0.09	0.46 ± 0.07
	Morgan+CP	0.61	0.303	0.65	0.63 ± 0.08	0.28 ± 0.18	0.74 ± 0.14
APR_HepG2_CellLoss_72h_dn	CP	0.67	0.361	0.68	0.59 ± 0.05	0.17 ± 0.08	0.67 ± 0.13
	ErG	0.62	0.249	0.67	0.52 ± 0.11	0.03 ± 0.24	0.56 ± 0.16
	ErG+CP	0.64	0.293	0.71	0.58 ± 0.09	0.15 ± 0.17	0.69 ± 0.11
	Morgan	0.75	0.506	0.83	0.57 ± 0.09	0.12 ± 0.14	0.65 ± 0.12

	Morgan+CP	0.67	0.374	0.67	0.60 ± 0.05	0.17 ± 0.07	0.68 ± 0.11
BSK_3C_ Proliferation _down	CP	0.60	0.196	0.64	0.68 ± 0.11	0.28 ± 0.15	0.75 ± 0.12
	ErG	0.67	0.339	0.75	0.56 ± 0.13	0.12 ± 0.21	0.62 ± 0.18
	ErG+CP	0.73	0.458	0.79	0.71 ± 0.09	0.35 ± 0.10	0.83 ± 0.11
	Morgan	0.64	0.285	0.70	0.54 ± 0.09	0.07 ± 0.17	0.60 ± 0.12
	Morgan+CP	0.62	0.237	0.68	0.65 ± 0.12	0.22 ± 0.18	0.73 ± 0.11
BSK_3C_SRB_ down	CP	0.58	0.163	0.57	0.60 ± 0.09	0.18 ± 0.15	0.67 ± 0.12
	ErG	0.53	0.057	0.61	0.48 ± 0.05	-0.04 ± 0.09	0.53 ± 0.12
	ErG+CP	0.59	0.187	0.63	0.61 ± 0.12	0.20 ± 0.21	0.69 ± 0.10
	Morgan	0.49	-0.028	0.53	0.52 ± 0.14	0.06 ± 0.26	0.51 ± 0.17
	Morgan+CP	0.56	0.135	0.56	0.61 ± 0.07	0.20 ± 0.13	0.66 ± 0.06
BSK_3C_Vis_ down	CP	0.71	0.447	0.73	0.76 ± 0.14	0.49 ± 0.25	0.74 ± 0.12
	ErG	0.59	0.214	0.67	0.57 ± 0.17	0.11 ± 0.37	0.55 ± 0.12
	ErG+CP	0.64	0.315	0.71	0.69 ± 0.14	0.34 ± 0.24	0.73 ± 0.14
	Morgan	0.57	0.152	0.63	0.52 ± 0.04	0.02 ± 0.08	0.53 ± 0.10
	Morgan+CP	0.68	0.403	0.70	0.68 ± 0.10	0.34 ± 0.17	0.70 ± 0.12
BSK_4H_SRB_ down	CP	0.66	0.392	0.70	0.69 ± 0.07	0.40 ± 0.11	0.74 ± 0.09
	ErG	0.60	0.269	0.70	0.62 ± 0.08	0.22 ± 0.12	0.67 ± 0.08
	ErG+CP	0.69	0.443	0.74	0.71 ± 0.10	0.43 ± 0.16	0.78 ± 0.12
	Morgan	0.47	-0.083	0.56	0.48 ± 0.04	-0.03 ± 0.06	0.59 ± 0.13
	Morgan+CP	0.66	0.389	0.72	0.72 ± 0.09	0.49 ± 0.14	0.72 ± 0.11
BSK_CASM3C_ Proliferation_ down	CP	0.60	0.231	0.63	0.69 ± 0.07	0.37 ± 0.13	0.77 ± 0.12
	ErG	0.70	0.403	0.81	0.60 ± 0.07	0.16 ± 0.11	0.75 ± 0.11
	ErG+CP	0.61	0.251	0.74	0.73 ± 0.11	0.44 ± 0.15	0.81 ± 0.10
	Morgan	0.50	0.011	0.54	0.55 ± 0.08	0.10 ± 0.15	0.50 ± 0.17
	Morgan+CP	0.62	0.272	0.65	0.69 ± 0.10	0.43 ± 0.21	0.76 ± 0.12
BSK_hDFCGF_ Proliferation_ down	CP	0.60	0.209	0.63	0.68 ± 0.15	0.25 ± 0.22	0.71 ± 0.13
	ErG	0.71	0.436	0.76	0.63 ± 0.10	0.15 ± 0.16	0.58 ± 0.08
	ErG+CP	0.61	0.211	0.68	0.63 ± 0.11	0.17 ± 0.14	0.69 ± 0.14
	Morgan	0.69	0.376	0.73	0.67 ± 0.15	0.22 ± 0.21	0.77 ± 0.17
	Morgan+CP	0.63	0.267	0.66	0.68 ± 0.16	0.26 ± 0.24	0.68 ± 0.18
BSK_LPS_SRB_ down	CP	0.66	0.346	0.71	0.61 ± 0.08	0.25 ± 0.21	0.70 ± 0.06
	ErG	0.64	0.290	0.69	0.54 ± 0.14	0.07 ± 0.30	0.62 ± 0.15
	ErG+CP	0.66	0.332	0.73	0.66 ± 0.08	0.31 ± 0.16	0.70 ± 0.09
	Morgan	0.48	-0.047	0.58	0.53 ± 0.14	0.03 ± 0.22	0.55 ± 0.19
	Morgan+CP	0.66	0.342	0.71	0.59 ± 0.04	0.23 ± 0.15	0.72 ± 0.06
BSK_SAg_ Proliferation_ down	CP	0.57	0.148	0.58	0.58 ± 0.09	0.16 ± 0.15	0.65 ± 0.15
	ErG	0.61	0.213	0.63	0.49 ± 0.16	0.00 ± 0.28	0.58 ± 0.19
	ErG+CP	0.60	0.209	0.63	0.61 ± 0.10	0.22 ± 0.19	0.71 ± 0.12
	Morgan	0.58	0.162	0.60	0.51 ± 0.08	0.01 ± 0.15	0.51 ± 0.15
	Morgan+CP	0.59	0.146	0.59	0.62 ± 0.09	0.23 ± 0.15	0.70 ± 0.10



(A) Aggregated Metrics.



(B) Cluster-averaged Metrics.

Figure 3.1 Balanced Accuracies on using different fingerprints and combinations across all endpoints. (A) Aggregated Metric. The combination of Cell Painting with ErG and Morgan fingerprints improved the balanced accuracies in 7 out of 10

assays (by 9.6% and 23.1% on average respectively). (B) Cluster-averaged Metrics (mean performance over clusters and error bars show standard deviations).Cell Painting fingerprints improved balanced accuracy when compared to using ErG and Morgan fingerprints in all 10 assays by 16.8% and 15.9% on average respectively.

3.2 Model Evaluation

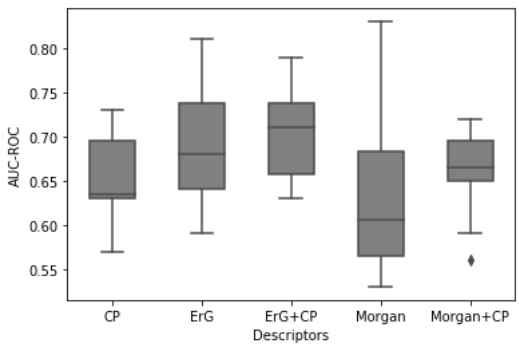
3.2.1 Random shuffling splits and aggregated metrics

When using *random shuffling splits and aggregated metrics,* using Cell Painting fingerprints alone performed with nearly equal success to the ErG and Morgan fingerprint models with an even better performance recorded in many assays (7/10). The mean increase in such cases of Cell Painting over Morgan fingerprints (BA: +0.14, MCC: +0.32 and AUC-ROC: +0.09) and to using Cell Painting over ErG fingerprints (BA: +0.06, MCC: +0.14 and AUC-ROC: +0.03) shows that the improvement on using combination fingerprints indeed comes from the advantage of adding Cell Painting features over other molecular structure-based fingerprints. For one of the assays, namely the BSK_3C_Vis_down, we expected Cell Painting features to work best, given the endpoint itself quantifies changes cell morphology in umbilical vein endothelium, a human vascular primary cell. Although Cell Painting features and endpoint in question are obtained on different cell lines, indeed, Cell Painting features alone yielded the best metric scores (BA: 0.71, AUC-ROC: 0.73, MCC: 0.45) compared to other fingerprints.

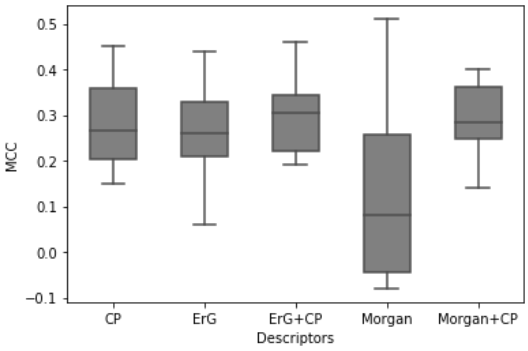
Another interesting observation involves the combination of Cell Painting fingerprints with ErG and Morgan fingerprints that improved the balanced accuracies in 7 out of 10 assays (by 9.6% and 23.1% on average respectively) compared to using ErG and Morgan fingerprints alone. Using Cell Painting features in particular, either alone or in combination with ErG/Morgan fingerprints, exhibited higher mean performance in *all*

assays then fingerprints alone showing that cell morphology contains relevant information when modelling cytotoxicity- and proliferation decrease assays. (see Figure 3.2 and Table 3.1 and Appendix A for individual assay results). In 6 out of 10 assays the Cell Painting features, either alone or in combination fingerprints, achieved highest BA scores, highest AUC-ROC scores in 7 out of 10, and the best MCC scores in 6 out of 10. K-Fold splitting aggregated metrics indicate that in all assays but one, a combination fingerprint of Cell Painting with either Morgan/ErG fingerprint improved the balanced accuracy in comparison to using only Morgan or ErG fingerprint. The mean increase in assays where improvement was observed on using Morgan and Cell Painting over Morgan fingerprints was greater (BA: +0.11, MCC: +0.32 and AUC-ROC: +0.09) compared to the increase when using ErG and Cell Painting over ErG fingerprints (BA: +0.06, MCC: +0.12 and AUC-ROC: +0.04). This shows that cell morphology provided better complimentary and relevant information to Morgan fingerprints (circular fingerprints) than ErG fingerprints (pharmacophore fingerprints). ErG fingerprints may perform better alone than structural information in Morgan fingerprints as pharmacophore fingerprints contain the relevant molecular properties that are able to better model cytotoxicity, therefore the complementary information of Cell Painting shows further improvement in combination with Morgan fingerprints than ErG fingerprints.

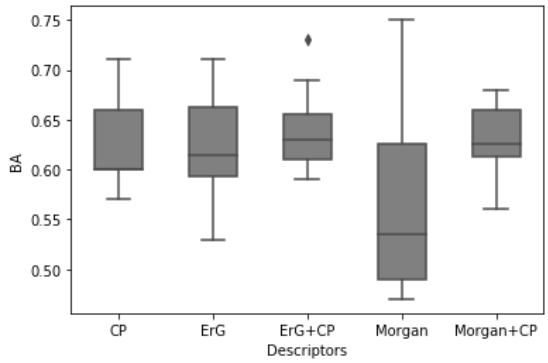
The highest increase in BA, MCC and AUC-ROC were recorded as 0.19, 0.47 and 0.16 respectively, when using Morgan and Cell Painting over Morgan fingerprints alone for the assay BSK_4H_SRB_down, an assay related to cell death. This indicating that the said biological process could be modelled with structural and cell morphological features in combination which seem to contribute distinct information. Overall, we can hence conclude that Cell Painting features contain information with respect to the endpoints considered here.



(A) AUC-ROC



(B) MCC



(C) BA

Figure 3.2: Mean aggregated performance using different compound representations and combinations there-of across all endpoints. (A) AUC-ROC (Area under the curve: Receiver operating characteristic), (B) MCC (Matthews correlation coefficient), and (C) BA (Balanced Accuracy) scores. Combination fingerprints are seen to have better performances over the use of Morgan or ErG fingerprints alone.

3.2.2 Cluster-based splits and cluster-averaged metrics

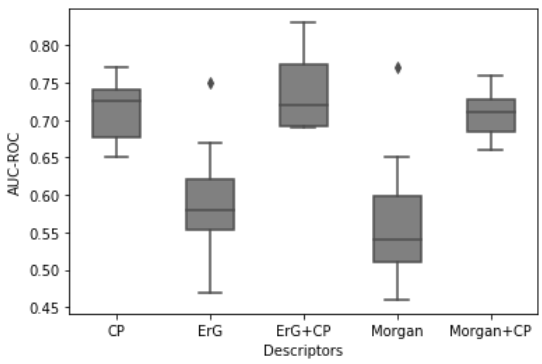
When using *cluster-based splits and cluster-averaged metrics*, most ErG and Morgan fingerprint models (for ErG 8 out of 10, and Morgan fingerprints 9 out of 10) performed poorly with BA <= 0.60 (see Figure 3.3, and Table 3.2 and Appendix B for individual assays). The size of the standard deviation indicated the performance varies considerably

depending on the different chemical spaces in testing and training set as well as the type of the assay, which is also likely related to the overall relatively small size of the dataset used. For no assay endpoints were the ErG or Morgan fingerprints performing best using any of the three metrics. Interestingly, while ErG and Morgan fingerprints alone struggled to model the data (average BA of 0.55 and 0.53 respectively for all assays), with scores equivalent to random models, most models using only Cell Painting fingerprints could perform with nearly equal success to combination fingerprints (average BA: 0.65, MCC: 0.28 and AUC-ROC: 0. 71, for all assays). Cell Painting fingerprints improved balanced accuracy when compared to ErG and Morgan fingerprints in all 10 assays, and by 16.8% and 15.9% on average, respectively. Once again, the model for BSK_3C_Vis_down endpoint assays could achieve success using just Cell Painting fingerprints with a BA of 0.76, a significant 46.1% improvement over using Morgan fingerprints indicating morphological information was far relevant in modelling cell death markers.

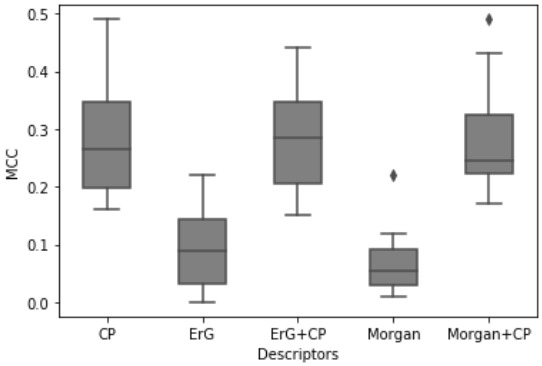
Combinations of ErG and Cell Painting or Morgan fingerprint and Cell Painting models performed with a higher BA, MCC and AUC-ROC across all cases, when compared to using ErG and Morgan fingerprints alone, with, the mean increase in using ErG fingerprints (BA: +0.11, MCC: +0.20 and AUC-ROC: +0.15) was similar to that of Morgan fingerprints (BA: +0.11, MCC: +0.21 and AUC-ROC: +0.14). For most assays, combination fingerprints of ErG and Cell Painting features and Morgan fingerprints with Cell Painting features achieved the best BA scores and MCC scores in 9 out of 10 cases, and the best AUC-ROC scores in 8 out of 10 cases, hence further strengthening the relevance of morphological features in predicting the endpoints in this study.

Overall, models that included Cell Painting fingerprints, either alone or in combination with ErG/Morgan fingerprints, showed the best comparative performance across in all assay endpoints when leaving one cluster-based method was implemented. It can be observed that while ErG and Morgan fingerprints failed with scores similar to random

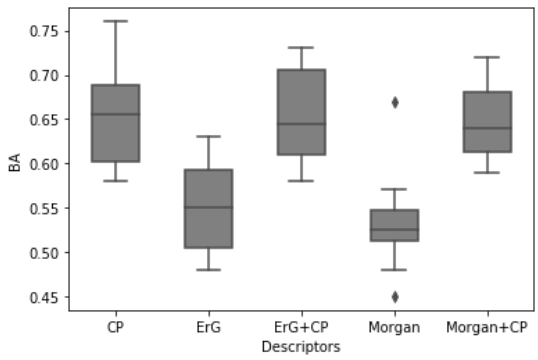
predictions, the combination fingerprints, as well as Cell Painting features by themselves, were able to model these endpoints much better. In all assays that showed improvements, the increase in BA from using ErG or Morgan fingerprint to combination fingerprints were recorded as 21.2% and 22.3% respectively. It can be inferred from the above discussions that Cell Painting can better model cytotoxicity- and proliferation-related *in vitro* assays endpoints also in a cluster-based split scenario, which requires the extrapolation to novel chemical space.



(A) AUC-ROC



(B) MCC



(C) BA scores

Figure 3.3: Mean cluster averaged performance using fingerprints and combinations across all endpoints. (A) AUC-ROC (Area under the curve: Receiver operating characteristic), (B) MCC (Matthews correlation coefficient), and (C) BA

(Balanced Accuracy) scores. Combination fingerprints have increased performance compared to Morgan or ErG fingerprints alone.

3.3 Significance of results using T-Test

Next, we compared the performance of the different descriptors on the cluster-averaged metrics using a two-sided T-test[2020] for the null hypothesis that two related samples have an identical average expected values, and found that in 7 out of 10 assays, using a combination fingerprint of ErG and Cell Painting, and Morgan fingerprints and Cell Painting features, respectively, improved the model performances compared to when using ErG and Morgan fingerprint alone. Results tabulated in Table 3.2 shows for all metrics and fingerprints, where the T-test p-value < 0.05 shows an improvement in performance is observed that is statistically significant when using Cell Painting fingerprint in addition to the respective other fingerprints. In multiple assays, the mean importance scores lay between 0.02 and 0.05 while features corresponded to cell compartment and granularity feature group. Overall, relatively large p-values generally are also due to the small size of the dataset used in this work.

Table 3.2 Differences in cluster-averaged performance of models trained/tested on same data but using two different descriptor sets. The p-value is calculated using two-sided T-test where null hypothesis is 2 related or repeated samples have identical average (expected) values.

Target	Metric	Fingerprint	Metric Score using fingerprint	Comparison Fingerprint	Metric Score using comparison fingerprint	T-test	
						T-statistic value	P-value
ACEA_T47D_80hr_Negative	AUC-ROC	ErG+CP	0.76 ± 0.13	Morgan	0.46 ± 0.07	4.18	0.025
	AUC-ROC	Morgan+CP	0.74 ± 0.14	Morgan	0.46 ± 0.07	3.61	0.036
	BA	ErG+CP	0.61 ± 0.04	ErG	0.50 ± 0.00	4.15	0.025
	MCC	ErG+CP	0.26 ± 0.08	ErG	0.01 ± 0.01	5.61	0.011
	MCC	ErG+CP	0.26 ± 0.08	Morgan	-0.05 ± 0.09	6.03	0.009

	MCC	Morgan+CP	0.28 ± 0.18	Morgan	-0.05 ± 0.09	3.58	0.037
	SEN	ErG+CP	0.28 ± 0.13	Morgan	0.00 ± 0.00	3.69	0.035
	SEN	Morgan+CP	0.33 ± 0.17	Morgan	0.00 ± 0.00	3.35	0.044
BSK_3C_ Proliferation_ down	AUC-ROC	ErG+CP	0.83 ± 0.11	Morgan	0.60 ± 0.12	3.04	0.038
	AUC-ROC	Morgan+CP	0.73 ± 0.11	Morgan	0.60 ± 0.12	3.14	0.035
	BA	ErG+CP	0.71 ± 0.09	Morgan	0.54 ± 0.09	3.05	0.038
BSK_3C_ SRB_down	AUC-ROC	ErG+CP	0.69 ± 0.10	ErG	0.53 ± 0.12	3.36	0.028
BSK_3C_ Vis_down	AUC-ROC	ErG+CP	0.73 ± 0.14	ErG	0.55 ± 0.12	3.05	0.038
	BA	Morgan+CP	0.68 ± 0.10	Morgan	0.52 ± 0.04	3.55	0.024
	MCC	Morgan+CP	0.34 ± 0.17	Morgan	0.02 ± 0.08	3.77	0.020
BSK_4H_ SRB_down	AUC-ROC	ErG+CP	0.78 ± 0.12	ErG	0.67 ± 0.08	3.30	0.030
	BA	ErG+CP	0.71 ± 0.10	Morgan	0.48 ± 0.04	3.61	0.023
	BA	Morgan+CP	0.72 ± 0.09	Morgan	0.48 ± 0.04	4.30	0.013
	MCC	ErG+CP	0.43 ± 0.16	ErG	0.22 ± 0.12	3.33	0.029
	MCC	ErG+CP	0.43 ± 0.16	Morgan	-0.03 ± 0.06	4.62	0.010
	MCC	Morgan+CP	0.49 ± 0.14	Morgan	-0.03 ± 0.06	6.75	0.003
	SEN	ErG+CP	0.53 ± 0.25	Morgan	0.10 ± 0.20	3.53	0.024
	SEN	Morgan+CP	0.53 ± 0.25	Morgan	0.10 ± 0.20	3.53	0.024
BSK_ CASM3C_ Proliferation_ down	AUC-ROC	ErG+CP	0.81 ± 0.10	Morgan	0.50 ± 0.17	2.91	0.044
	BA	ErG+CP	0.73 ± 0.11	ErG	0.60 ± 0.07	3.25	0.031
	MCC	ErG+CP	0.44 ± 0.15	ErG	0.16 ± 0.11	4.10	0.015
	MCC	ErG+CP	0.44 ± 0.15	Morgan	0.10 ± 0.15	2.89	0.044
	SPE	ErG+CP	0.91 ± 0.09	ErG	0.70 ± 0.10	7.45	0.002
	SEN	ErG+CP	0.55 ± 0.30	Morgan	0.22 ± 0.19	3.04	0.039
BSK_SAg_ Proliferation _down	AUC-ROC	ErG+CP	0.71 ± 0.12	Morgan	0.51 ± 0.15	3.67	0.021
	AUC-ROC	Morgan+CP	0.70 ± 0.10	Morgan	0.51 ± 0.15	3.68	0.021
	BA	ErG+CP	0.61 ± 0.10	Morgan	0.51 ± 0.08	3.13	0.035
	BA	Morgan+CP	0.62 ± 0.09	Morgan	0.51 ± 0.08	3.61	0.023
	MCC	ErG+CP	0.22 ± 0.19	Morgan	0.01 ± 0.15	3.60	0.023
	MCC	Morgan+CP	0.23 ± 0.15	Morgan	0.01 ± 0.15	3.67	0.021

(For the ACEA_T47D_80hr_Negative assay, only 4 clusters were used based on PCA and hence a group 4-fold was used in the outer loop for this assay.)

3.4 Cell Painting feature interpretation

We next tried to understand better which Cell Painting features contributed to model performance, both to ensure interpretability and confidence into the resulting models. Using permutation importance on nested cross-validation as described in the *methods*

section (*variable importance of Cell Painting features*), top 5 contributing features from the best model are shown in Table 3.3. (for the distribution of features type among the endpoints see Appendix C)

Table 3.3 Feature Importance scores in using Cell Painting fingerprints. The mean feature importance value is calculated using 10-fold permutation importance from the best performing held-out test set based on balanced accuracy and having a positive permutation score.

Endpoint	Sl no	Features	Mean Feature Importance	Compartment	Feature Group
ACEA_T47D_80hr_Negative	1a	Cytoplasm_Intensity_MassDisplacement_Mito	0.036	Cytoplasm	Intensity
	1b	Nuclei_Granularity_8_Mito	0.016	Nuclei	Granularity
	1c	Cells_Granularity_8_Mito	0.016	Cells	Granularity
	1d	Nuclei_Texture_DifferenceVariance_ER_5_0	0.016	Nuclei	Texture
	1e	Cytoplasm_Intensity_MADIntensity_Mito	0.016	Cytoplasm	Intensity
APR_HepG2_CellLoss_72h_dn	2a	Nuclei_Texture_InverseDifferenceMoment_ER_3_0	0.050	Nuclei	Texture
	2b	Cytoplasm_Correlation_K_DNA_Mito	0.050	Cytoplasm	Correlation
	2c	Nuclei_Granularity_13_Mito	0.044	Nuclei	Granularity
	2d	Cells_Granularity_13_Mito	0.039	Cells	Granularity
	2e	Cytoplasm_Granularity_13_Mito	0.039	Cytoplasm	Granularity
BSK_3C_Proliferation_down	3a	Cells_Neighbors_NumberOfNeighbors_Adjacent	0.048	Cells	Neighbours
	3b	Cells_Correlation_Costes_Mito_AGP	0.048	Cells	Correlation
	3c	Cytoplasm_Correlation_Costes_Mito_AGP	0.048	Cytoplasm	Correlation
	3d	Cells_Granularity_13_AGP	0.029	Cells	Granularity
	3e	Cells_Neighbors_PercentTouching_Adjacent	0.024	Cells	Neighbours
BSK_3C_SRB_down	4a	Cells_Neighbors_PercentTouching_Adjacent	0.032	Cells	Neighbours
	4b	Cells_RadialDistribution_FracAtD_ER_4of4	0.032	Cells	Radial Distribution
	4c	Nuclei_Granularity_13_ER	0.023	Nuclei	Granularity
	4d	Cells_Texture_SumVariance_ER_3_0	0.018	Cells	Texture
	4e	Cells_RadialDistribution_FracAtD_DNA_1of4	0.009	Cells	Radial Distribution
BSK_3C_Vis_down	5a	Cells_Granularity_8_Mito	0.145	Cells	Granularity
	5b	Cytoplasm_Granularity_8_Mito	0.132	Cytoplasm	Granularity
	5c	Nuclei_Granularity_8_Mito	0.105	Nuclei	Granularity
	5d	Nuclei_Granularity_7_ER	0.082	Nuclei	Granularity
	5e	Cells_Neighbors_NumberOfNeighbors_5	0.077	Cells	Neighbours
	6a	Nuclei_Correlation_Costes_DNA_AGP	0.024	Nuclei	Correlation

BSK_4H_SRB_down	6b	Cells_Granularity_8_Mito	0.024	Cells	Granularity
	6c	Cytoplasm_Granularity_8_Mito	0.024	Cytoplasm	Granularity
	6d	Cells_RadialDistribution_FracAtD_ER_4of4	0.019	Cells	Radial Distribution
	6e	Cells_Neighbors_PercentTouching_5	0.019	Cells	Neighbours
BSK_CASM3C_Proliferation_down*	7a	Cells_Granularity_7_RNA	0.014	Cells	Granularity
	7b	Cytoplasm_Correlation_K_DNA_Mito	0.014	Cytoplasm	Correlation
	7c	Nuclei_Correlation_K_DNA_Mito	0.010	Nuclei	Correlation
	7d	Cells_Neighbors_NumberOfNeighbors_Adjacent	0.009	Cells	Neighbours
BSK_hDFCGF_Proliferation_down	8a	Cytoplasm_Correlation_K_DNA_Mito	0.081	Cytoplasm	Correlation
	8b	Cells_Correlation_Costes_DNA_AGP	0.043	Cells	Correlation
	8c	Cytoplasm_Texture_AngularSecondMoment_Mito_5_0	0.038	Cytoplasm	Texture
	8d	Cytoplasm_Texture_Entropy_ER_10_0	0.033	Cytoplasm	Texture
	8e	Cells_Granularity_12_RNA	0.029	Cells	Granularity
BSK_LPS_SRB_down	9a	Cells_RadialDistribution_FracAtD_ER_4of4	0.067	Cells	Radial Distribution
	9b	Nuclei_Correlation_K_ER_Mito	0.048	Nuclei	Correlation
	9c	Cytoplasm_Intensity_IntegratedIntensityEdge_Mito	0.048	Cytoplasm	Intensity
	9d	Cytoplasm_Correlation_K_Mito_RNA	0.038	Cytoplasm	Correlation
	9e	Cytoplasm_Intensity_MADIntensity_Mito	0.038	Cytoplasm	Intensity
BSK_SAg_Proliferation_down	10a	Cells_Correlation_RWC_Mito_ER	0.071	Cells	Correlation
	10b	Cytoplasm_Correlation_Correlation_Mito_RNA	0.043	Cytoplasm	Correlation
	10c	Cytoplasm_Granularity_8_Mito	0.033	Cytoplasm	Granularity
	10d	Cells_Granularity_8_Mito	0.029	Cells	Granularity
	10e	Cytoplasm_Correlation_Costes_AGP_ER	0.029	Cytoplasm	Correlation

*Only 4 important features could be determined across the two best-performing folds for endpoint BSK_CASM3C_Proliferation_down

It needs to be kept in mind that Cell Painting features are highly correlated, so there is a gradual scale as to their importance for an endpoint, and not a clear-cut selection of only individual features are possible. To interpret the relevance of Cell Painting profiles, we divide our assays into two categories: proliferation decrease and cytotoxicity SRB assays.

3.4.1 Proliferation decrease endpoints

For four assays that measure downregulation of proliferation (namely BSK 3C Proliferation down, BSK CASM3C Proliferation down, BSK hDFCGF Proliferation down, BSK Sag Proliferation down) features from the cytoplasm and cell correlations with mitochondria (Mito), actin, Golgi, plasma membrane (AGP), nucleus (DNA), among others, were found to be most important for model performance (features 3b, 3c, 7b, 7c, 8a, 8b, 10a, 10b, 10e in Table 3.3).

Our features 3b, 3c, 8b from Table 3.3 determine the correlation between the cytoplasm and cell compartments with Golgi apparatus and could reflect changes from Golgi apparatus to the cytoplasmic and nuclear location of cells, which may be due to proliferation mediation.[87]

Features 7b, 8a, 10b, 10e and the feature 7c from Table 3.3 relate to cytoplasm and nucleus image correlations with sub-compartments respectively; variable importance indicates that these features had relevant information about proliferation endpoints while KS-test ensures their distribution was varying as per activity. One could argue it represents likely the depletion of krit1 and subsequent icap1α depletion from cytoplasm and nuclei as shown by Zhang et. al. that Krev interaction trapped 1 (krit1) and integrin cytoplasmic domain-associated protein-1 α (icap1α) in β1-integrin mediated cell proliferation.[88] It was reported "…krit1 co-localizes with icap1α in both the nucleus and the cytoplasm; however, most of icap1α is found in the nucleus and most of krit1 is found in the cytoplasm at steady state. On depletion of krit1, icap1α decreases in the cytoplasm and is no longer detected in the nucleus.".

Another contributing feature that was found to be important for predicting these four assay endpoints was cell granularity of RNA, AGP and mitochondria (3d, 7a, 8e, 10c, 10d in Table 3.3). The measure of granularity could be related to cell adhesion and cell spreading as image granularity measures how well the texture of images fit structure

elements of increasing size. A varying distribution of this for active and inactive for an endpoint could signify that cell adhesion and cell spreading has enabled the fits to differ. In such a case, increased integrins that control complex cell functions could be the cause for general defects in cell adhesion and cell spreading.[89]

Overall although direct mechanistic relations cannot be drawn, the novelty lies in the fact that possibility of such information in morphological features (which is absent over structural fingerprints) can be presented as complementary (or perhaps even better than structural/pharmacophore fingerprints) which is why our models using only Cell Painting features seem to predict our endpoints well even when extended to novel chemical space using *cluster-based splits and cluster-averaged metrics*.

3.4.2 Cytotoxicity SRB endpoints

For three assays (namely, BSK 3C SRB down, BSK 4H SRB down, BSK LPS SRB down) that measured cytotoxicity-related endpoints in different cell types using sulforhodamine B (SRB) staining for total protein content features related to cellular neighbourhood and radial distributions as well as nuclei correlation with sub-compartments were found to be most important for model performance.

The selected features (6a, 9b in Table 3.3) for nucleus morphological correlations with different organelles or cellular components are consistent with the fact that significant changes in nucleus morphology indeed plays a role in cytotoxicity as described by Wu et al. in their work on the biophysical assessment of single-cell cytotoxicity using mechanical maps.[90]

Further, all three assay endpoints selected as important contributing features (4b, 4e, 6d, 9a in Table 3.3) were stemming from the cell radial distributions, which is a measure of the radial distribution within the cell from the centre of the cell with bins from 1 (innermost) to 4 (outermost). The radial distribution function is thus interesting as

changes in its distribution for endpoints can be relevant to morphological transition including cell protrusion and elasticity, cell-substate adhesion. One could argue that the cell radial distribution feature signifies the change in certain bins are more significant and thus alteration of the distribution could signify changes in cell orientation and elasticity in the cell as shown by Wu et al. in measurements on single endothelial cells revealed that "..region of cell nucleus possesses a lower elasticity comparing with cytoplasm, indicating heterogeneity in adhesion behaviour and elasticity over the whole cell surface".[90]

Additionally, two of the three assays selected features (4a, 6e in Table 3.3) of cell neighbours having high permutation importance scores. These features calculate the percentage of a cell's edge pixels that touch another cell after they have been expanded to a specified distance. This indicates that neighbouring cells appear to have an integral role in cell death which reinforces recent experimental evidence.[91]

Hence, while overall the interpretation of features is not generally mechanistically possible, features selected are still plausible for the given endpoints considered, which provides further evidence that a meaningful model has been obtained.

4 CONCLUSIONS AND FUTURE WORK

The thesis explored the utility of Cell Painting readouts in containing information that helps predict cytotoxicity, which could be used in future. Chapter 1 gives a general introduction to toxicity predictions and machine learning methods. This chapter also explained the various works using cell painting assays and other related morphological fingerprints in predicting cytotoxicity-related or reduced proliferation assay endpoints. chapter 2 presented an overview of the methods, including datasets, model generation and evaluation and other aspects of the work. Chapter 3 discusses the results and inferences drawn from the results. In this chapter, we shall discuss the important conclusions of this work and point at the possibilities in future.

4.1 Conclusions

In this work, we explored the utility of Cell Painting readouts to predict cytotoxicity- and proliferation-related assay endpoints. To this end, we applied Random Forest algorithm to generate classification models for 10 such endpoints from the ToxCast assay using Cell Painting annotations, ErG fingerprints, Morgan fingerprints and a combination of Cell Painting features with either of the latter two.

We found that the Cell Painting features led to well-performing models by themselves, and generally improved performance of models using fingerprints when used in combination by up to 20%. This was in particular also true for *cluster-based splits and cluster-averaged metrics* employed here, where the model was required to extrapolate to novel chemical space. From the practical side, this could be particularly relevant for

incompletely defined structures and mixtures of compounds (also very complex ones, where the investigation of constitutive components can be impossible). Features selected were generally plausible (albeit difficult to interpret in detail), and hence we can conclude that Cell Painting readouts contain a predictive value for the endpoints considered in this work, which in combination with the easy of their generation might render them a good choice for compound profiling, also beyond the endpoints considered in this study.

4.2 Limitations of this study

The main limitations of the current study are as follows: Firstly, the dataset used covers a limited number of data points (and area of chemical space covered) due to the required overlap of the Cell Painting assay with assay annotations. Secondly, the binarization of data (here endpoint labels, into toxic and non-toxic) always leads to a loss of information[92] related to activity and dose, and hence for quantitative predictions with *in vivo* relevance, this would need to be considered. Thirdly (related to the previous point), the pharmacokinetic behaviour of a compound, which is crucial to understand it's *in vivo* behaviour, is not considered here either.[93] In the absence of such data, the cell-based assay and the way it is used here may be useful for hazard identification, but not by itself as an indicator of quantitative risk.[94] Fourthly, predicting cytotoxicity-related assays is generally a challenge due to multiple mechanistic pathways that are highly dynamic and dependent on dose, and they also depend in particular on the biological setup (such as cell line) considered. Finally, *in vitro* may be completely irrelevant to target organ exposure *in vivo* (among other possible factors). As a remedy, it would be possible to employ e.g. physiologically based biokinetic modelling to move towards *in vivo* toxicity predictions.[95]

4.3 Future Work

In future, there is every possibility to replace, or at the least, reinforce high throughput assays with cost-effect image-based techniques and artificial intelligence. Cell

morphology provides us with a better picture of understanding cellular mechanisms and can act as a fingerprinting method for compounds as shown in this work. This methodology could then be used directly in the early stages of the drug discovery pipeline. Different data techniques may be used in building such fingerprints, mean, median or the distribution of the cell morphology over all cells, which will be investigated in future work. Such profiling may also be used in key target identification, toxicity evaluation or small molecule mechanism of action study. The link between cell morphology and gene expression may also benefit our studies when used in combination.

In summary, the thesis has investigated the use of cell morphological features with machine learning as a novel approach to predicting drug-induced *in vitro* cytotoxicity-related or reduced proliferation assay outcomes and we anticipate the results of this study will help impact the utility of the Cell Painting assay and provide a proof-of-concept in using similar fingerprints in toxicity evaluation.

5 REFERENCES

(1) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. J. Health Econ. 2016, 47, 20–33.

(2) Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. Drug Discovery Today. 2005, 1421–1433.

(3) Shanks, N.; Greek, R.; Greek, J. Are Animal Models Predictive for Humans? Philosophy, Ethics, and Humanities in Medicine. 2009, 4, 2.

(4) Walum, E. Acute Oral Toxicity. In Environmental Health Perspectives; 1998; Vol. 106, 497–503.

(5) Iasella, C. J.; Johnson, H. J.; Dunn, M. A. Adverse Drug Reactions: Type A (Intrinsic) or Type B (Idiosyncratic). Clinics in Liver Disease. 2017.

(6) Singh, S.; Khanna, V. K.; Pant, A. B. Development of In Vitro Toxicology: A Historic Story. In In Vitro Toxicology; 2018; 1–19.

(7) National Toxicology Program. A National Toxicology Program for the 21 St Century, a Roadmap for the Future; 2004.

(8) Toxicity Testing in the 21st Century: A Vision and a Strategy; 2007.

(9) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2016, 6, 147–172.

(10) Cronin, M. T. D., & Yoon, M. Chapter 5.3 - Computational Methods to Predict Toxicity. In M. Balls, R. Combes, & A. B. T.-T. H. of A. T. M. in T. Worth (Eds.), History of Toxicology and Environmental Health. Academic Press. 2019, 287–300.

(11) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of Machine Learning Methods in Drug Toxicity Prediction. Curr. Top. Med. Chem. 2018, 18 (12), 987–997.

(12) Toropov, A. A.; Toropova, A. P.; Raska, I.; Leszczynska, D.; Leszczynski, J. Comprehension of Drug Toxicity: Software and Databases. Comput. Biol. Med. 2014, 45 (1), 20–25.

(13) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. J. Chem. Inf. Comput. Sci. 1989, 29 (2), 97–101.

(14) Carpenter, A. E. Image-Based Chemical Screening. Nature Chemical Biology. 2007, 461–465.

(15) Prasanna, S.; Doerksen, R. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. Curr. Med. Chem. 2008, 16 (1), 21–41.

(16) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50 (5), 742–754.

(17) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. In Journal of Chemical Information and Modeling; 2006; Vol. 46, 208–220.

(18) Ringnér, M. What Is Principal Component Analysis? Nature Biotechnology. 2008, 303–304.

(19) Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 1951, 46 (253), 68–78.

(20) Conway, F.; Hugill, M. Advanced Statistics. Math. Gaz. 1987, 71 (455), 76.

(21) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. Classification and Regression Trees; 2017.

(22) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 2003, 43 (6), 1947–1958.

(23) Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv 2018.

(24) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings - International Conference on Pattern Recognition; 2010; 3121–3124.

(25) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 1982, 143 (1), 29–36.

(26) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. BBA - Protein Struct. 1975, 405 (2), 442–451.

(27) Fawcett, T. An Introduction to ROC Analysis. Pattern Recognit. Lett. 2006, 27 (8), 861–874.

(28) Pencina, M. J.; D'Agostino, R. B.; Massaro, J. M. Understanding Increments in Model Performance Metrics. Lifetime Data Anal. 2013, 19 (2), 202–218.

(29) Strobl, C.; Boulesteix, A. L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 2007, 8, 25.

(30) Breiman, L. Random Forests. Machine Learning 2001, 45 (1), 5–32.

(31) Breiman, L. Statistical Modeling: The Two Cultures. Statistical Science 2001, 16 (3), 199–215.

(32) Galluzzi, L.; Bravo-San Pedro, J. M.; Vitale, I.; Aaronson, S. A.; Abrams, J. M.; Adam, D.; Alnemri, E. S.; Altucci, L.; Andrews, D.; Annicchiarico-Petruzzelli, M.; Baehrecke, E. H.; Bazan, N. G.; Bertrand, M. J.; Bianchi, K.; Blagosklonny, M. v.; Blomgren, K.; Borner, C.; Bredesen, D. E.; Brenner, C.; Campanella, M.; Candi, E.; Cecconi, F.; Chan, F. K.; Chandel, N. S.; Cheng, E. H.; Chipuk, J. E.; Cidlowski, J. A.; Ciechanover, A.; Dawson, T. M.; Dawson, V. L.; de Laurenzi, V.; de Maria, R.; Debatin, K. M.; di Daniele, N.; Dixit, V. M.; Dynlacht, B. D.; El-Deiry, W. S.; Fimia, G. M.; Flavell, R. A.; Fulda, S.; Garrido, C.; Gougeon, M. L.; Green, D. R.; Gronemeyer, H.; Hajnoczky, G.; Hardwick, J. M.; Hengartner, M. O.; Ichijo, H.; Joseph, B.; Jost, P. J.; Kaufmann, T.; Kepp, O.; Klionsky, D. J.; Knight, R. A.; Kumar, S.; Lemasters, J. J.; Levine, B.; Linkermann, A.; Lipton, S. A.; Lockshin, R. A.; LópezOtín, C.; Lugli, E.; Madeo, F.; Malorni, W.; Marine, J. C.; Martin, S. J.; Martinou, J. C.; Medema, J. P.; Meier, P.; Melino, S.; Mizushima, N.; Moll, U.; Muñoz-Pinedo, C.; Nuñez, G.; Oberst, A.; Panaretakis, T.; Penninger, J. M.; Peter, M. E.; Piacentini, M.; Pinton, P.; Prehn, J. H.; Puthalakath, H.; Rabinovich, G. A.; Ravichandran, K. S.; Rizzuto, R.; Rodrigues, C. M.; Rubinsztein, D. C.; Rudel, T.; Shi, Y.; Simon, H. U.; Stockwell, B. R.; Szabadkai, G.; Tait, S. W.; Tang, H. L.; Tavernarakis, N.; Tsujimoto, Y.; vanden Berghe, T.; Vandenabeele, P.; Villunger, A.; Wagner, E. F.; Walczak, H.; White, E.; Wood, W. G.; Yuan, J.; Zakeri, Z.; Zhivotovsky, B.; Melino, G.; Kroemer, G. Essential versus Accessory Aspects of Cell Death: Recommendations of the NCCD 2015. Cell Death and Differentiation. 2015, 58–73.

(33) Allen, C. H. G.; Koutsoukas, A.; Cortés-Ciriano, I.; Murrell, D. S.; Malliavin, T. E.; Glen, R. C.; Bender, A. Improving the Prediction of Organism-Level Toxicity through Integration of Chemical, Protein Target and Cytotoxicity QHTS Data. Toxicology Research 2016, 5 (3), 883–894.

(34) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in Vitro Bioactivity and Chemical Structure. Chemical Research in Toxicology 2015, 28 (4), 738–751.

(35) Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., & Tropsha, A. (2011). Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. Environmental Health Perspectives, 119(3), 364–370.

(36) Mervin, L. H.; Cao, Q.; Barrett, I. P.; Firth, M. A.; Murray, D.; McWilliams, L.; Haddrick, M.; Wigglesworth, M.; Engkvist, O.; Bender, A. Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection. ACS Chem. Biol. 2016, 11 (11), 3007–3023.

(37) Yin, Z.; Ai, H.; Zhang, L.; Ren, G.; Wang, Y.; Zhao, Q.; Liu, H. Predicting the Cytotoxicity of Chemicals Using Ensemble Learning Methods and Molecular Fingerprints. Journal of Applied Toxicology. 2019, 1366–1377.

(38) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. Toxicology Research 2017, 6 (1), 73–80.

(39) Langdon, S. R.; Mulgrew, J.; Paolini, G. v.; van Hoorn, W. P. Predicting Cytotoxicity from Heterogeneous Data Sources with Bayesian Learning. Journal of Cheminformatics 2010, 2 (1), 11.

(40) Webel, H. E.; Kimber, T. B.; Radetzki, S.; Neuenschwander, M.; Nazaré, M.; Volkamer, A. Revealing Cytotoxic Substructures in Molecules Using Deep Learning. Journal of Computer-Aided Molecular Design 2020, 34 (7), 731–746.

(41) Nakano, T.; Brown, J. B. Prediction of Compound Cytotoxicity Based on Compound Structures and Cell Line Molecular Characteristics. Journal of Computer Aided Chemistry 2020, 21 (0), 1–10.

(42) Chang, C. Y.; Hsu, M. T.; Esposito, E. X.; Tseng, Y. J. Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods. Journal of Chemical Information and Modeling 2013, 53 (4), 958–971.

(43) Schrey, A. K.; Nickel-Seeber, J.; Drwal, M. N.; Zwicker, P.; Schultze, N.; Haertel, B.; Preissner, R. Computational Prediction of Immune Cell Cytotoxicity. Food and Chemical Toxicology 2017, 107, 150–166.

(44) Wu, Y.; Wang, G. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. International Journal of Molecular Sciences 2018, 19 (8), 2358.

(45) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W. N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. v.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 2017, 171 (6), 1437-1452.e17.

(46) Bray, M. A.; Gustafsdottir, S. M.; Rohban, M. H.; Singh, S.; Ljosa, V.; Sokolnicki, K. L.; Bittker, J. A.; Bodycombe, N. E.; Dančík, V.; Hasaka, T. P.; Hon, C. S.; Kemp, M. M.; Li, K.; Walpita, D.; Wawer, M. J.; Golub, T. R.; Schreiber, S. L.; Clemons, P. A.; Shamji, A. F.; Carpenter, A. E. A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay. GigaScience. Oxford University Press December 1, 2017, 1–5.

(47) Caicedo, J. C.; Singh, S.; Carpenter, A. E. Applications in Image-Based Profiling of Perturbations. Current Opinion in Biotechnology. 2016, 134–142.

(48) Bray, M. A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; KostAlimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes. Nature Protocols 2016, 11 (9), 1757–1774.

(49) Carpenter, A. E.; Jones, T. R.; Lamprecht, M. R.; Clarke, C.; Kang, I. H.; Friman, O.; Guertin, D. A.; Chang, J. H.; Lindquist, R. A.; Moffat, J.; Golland, P.; Sabatini, D. M. CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. Genome Biology 2006, 7 (10).

(50) Jones, T. R.; Kang, I. H.; Wheeler, D. B.; Lindquist, R. A.; Papallo, A.; Sabatini, D. M.; Golland, P.; Carpenter, A. E. CellProfiler Analyst: Data Exploration and Analysis Software for Complex Image-Based Screens. BMC Bioinformatics 2008, 9, 482.

(51) Caicedo, J. C.; Cooper, S.; Heigwer, F.; Warchal, S.; Qiu, P.; Molnar, C.; Vasilevich, A. S.; Barry, J. D.; Bansal, H. S.; Kraus, O.; Wawer, M.; Paavolainen, L.; Herrmann, M. D.; Rohban, M.; Hung, J.; Hennig, H.; Concannon, J.; Smith, I.; Clemons, P. A.; Singh, S.; Rees, P.; Horvath, P.; Linington, R. G.; Carpenter, A. E. Data-Analysis Strategies for Image-Based Cell Profiling. Nature Methods 2017, 14 (9), 849–863.

(52) Ljosa, V.; Caie, P. D.; ter Horst, R.; Sokolnicki, K. L.; Jenkins, E. L.; Daya, S.; Roberts, M. E.; Jones, T. R.; Singh, S.; Genovesio, A.; Clemons, P. A.; Carragher, N. O.; Carpenter, A. E. Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment. Journal of Biomolecular Screening 2013, 18 (10), 1321–1329.

(53) Aulner, N.; Danckaert, A.; Ihm, J. E.; Shum, D.; Shorte, S. L. Next-Generation Phenotypic Screening in Early Drug Discovery for Infectious Diseases. Trends in Parasitology. 2019, 559–570.

(54) Trapotsi, M. A. ; Barrett, I. ; Mervin, L. ; Afzal, A. ; Sturm, N. ; Engkvist, O. ; Bender, A. Multitask Bioactivity Predictions Using Structural Chemical and Cell Morphology Information. 2020. 10.26434/chemrxiv.12571241.

(55) Hofmarcher, M.; Rumetshofer, E.; Clevert, D. A.; Hochreiter, S.; Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. Journal of Chemical Information and Modeling 2019, 59 (3), 1163–1171.

(56) Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijnsters, P.; Velter, I.; Vapirev, A.; Singh, S.; Carpenter, A. E.; Wuyts, R.; Hochreiter, S.; Moreau, Y.; Ceulemans, H. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chemical Biology 2018, 25 (5), 611-618.e3

(57) Gustafsdottir, S.; Ljosa, V.; Sokolnicki, K.; Walpita, D.; Kemp, M.; Petri Seiler, K.; Carrel, H.; Golub, T.; Schreiber, S.; Clemons, P.; Carpenter, A.; Shamji, A. Multiplex Cytological Profiling Assay to Measure Diverse Cellular States. PLoS ONE 2013, 8 (12), e80999.

(58) Persson, M.; Løye, A. F.; Jacquet, M.; Mow, N. S.; Thougaard, A. V.; Mow, T.; Hornberg, J. J. High-Content Analysis/Screening for Predictive Toxicology: Application to Hepatotoxicity and Genotoxicity. Basic Clin. Pharmacol. Toxicol. 2014, 115 (1), 18–23.

(59) Wawer, M. J.; Jaramillo, D. E.; Dančík, V.; Fass, D. M.; Haggarty, S. J.; Shamji, A. F.; Wagner, B. K.; Schreiber, S. L.; Clemons, P. A. Automated Structure-Activity Relationship Mining: Connecting Chemical Structure to Biological Profiles. J. Biomol. Screen. 2014, 19 (5), 738–748.

(60) Nassiri, I.; McCall, M. N. Systematic Exploration of Cell Morphological Phenotypes Associated with a Transcriptomic Query. Nucleic Acids Res. 2018, 46 (19), e116.

(61) Rohban, M. H.; Singh, S.; Wu, X.; Berthet, J. B.; Bray, M. A.; Shrestha, Y.; Varelas, X.; Boehm, J. S.; Carpenter, A. E. Systematic Morphological Profiling of Human Gene and Allele Function via Cell Painting. eLife 2017, 6, e24060.

(62) Wawer, M. J.; Li, K.; Gustafsdottir, S. M.; Ljosa, V.; Bodycombe, N. E.; Marton, M. A.; Sokolnicki, K. L.; Bray, M. A.; Kemp, M. M.; Winchester, E.; Taylor, B.; Grant, G. B.; Hon, C. S. Y.; Duvall, J. R.; Wilson, J. A.; Bittker, J. A.; Dančík, V.; Narayan, R.; Subramanian, A.; Winckler, W.; Golub, T. R.; Carpenter, A. E.; Shamji, A. F.; Schreiber, S. L.; Clemons, P. A. Toward Performance-Diverse Small-Molecule Libraries for Cell-Based Phenotypic Screening Using Multiplexed High-Dimensional Profiling. Proceedings of the National Academy of Sciences of the United States of America 2014, 111 (30), 10911–10916.

(63) Lapins, M.; Spjuth, O. Evaluation of Gene Expression and Phenotypic Profiling Data as Quantitative Descriptors for Predicting Drug Targets and Mechanisms of Action. bioRxiv 2019, 580654.

(64) Scheeder, C.; Heigwer, F.; Boutros, M. Machine Learning and Image-Based Profiling in Drug Discovery. Current Opinion in Systems Biology 2018, 10, 43–52.

(65) Martin, H. L.; Adams, M.; Higgins, J.; Bond, J.; Morrison, E. E.; Bell, S. M.; Warriner, S.; Nelson, A.; Tomlinson, D. C. High-Content, High-Throughput Screening for the Identification of Cytotoxic Compounds Based on Cell Morphology and Cell Proliferation Markers. PLoS ONE 2014, 9 (2), e88338.

(66) O'Brien, P. J.; Irwin, W.; Diaz, D.; Howard-Cofield, E.; Krejsa, C. M.; Slaughter, M. R.; Gao, B.; Kaludercic, N.; Angeline, A.; Bernardi, P.; Brain, P.; Hougham, C. High Concordance of Drug-Induced Human Hepatotoxicity with in Vitro Cytotoxicity Measured in a Novel Cell-Based Model Using High Content Screening. Arch. Toxicol. 2006, 80 (9), 580–604.

(67) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. Chemical Science 2018, 9 (2), 513–530.

(68) Datasets http://moleculenet.ai/datasets-1 (accessed Jul 22, 2020).

(69) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chemical Research in Toxicology 2016, 29 (8), 1225–1251.

(70) Phuong, J. Masters Dissertation, Structured Application of Biological Ontologies to Annotate High-Throughput Screening Assays and Their Targets of Activity, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2014

(71) Rotroff, D. M.; Dix, D. J.; Houck, K. A.; Kavlock, R. J.; Knudsen, T. B.; Martin, M. T.; Reif, D. M.; Richard, A. M.; Sipes, N. S.; Abassi, Y. A.; Jin, C.; Stampfl, M.; Judson, R. S. Real-Time Growth Kinetics Measuring Hormone Mimicry for ToxCast Chemicals in T-47D Human Ductal Carcinoma Cells. Chemical Research in Toxicology 2013, 26 (7), 1097–1107.

(72) Kunkel, E. J.; Dea, M.; Ebens, A.; Hytopoulos, E.; Melrose, J.; Nguyen, D.; Ota, K. S.; Plavec, I.; Wang, Y.; Watson, S. R.; Butcher, E. C.; Berg, E. L. An Integrative Biology Approach for Analysis of Drug Action in Models of Human Vascular Inflammation. The FASEB Journal 2004, 18 (11), 1279–1281.

(73) Houck, K. A.; Dix, D. J.; Judson, R. S.; Kavlock, R. J.; Yang, J.; Berg, E. L. Profiling Bioactivity of the ToxCast Chemical Library Using BioMAP Primary Human Cell Systems. Journal of Biomolecular Screening 2009, 14 (9), 1054–1066.

(74) Shah, F.; Greene, N. Analysis of Pfizer Compounds in EPA's ToxCast Chemicals-Assay Space. Chemical Research in Toxicology 2014, 27 (1), 86–98.

(75) Kunkel, E. J.; Plavec, I.; Nguyen, D.; Melrose, J.; Rosler, E. S.; Kao, L. T.; Wang, Y.; Hytopou-los, E.; Bishop, A. C.; Bateman, R.; Shokat, K. M.; Butcher, E. C.; Berg, E. L. Rapid Structure-Activity and Selectivity Analysis of Kinase Inhibitors by BioMAP Analysis in Complex Human Primary Cell-Based Models. Assay and Drug Development Technologies. 2004, 431–441.

(76) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstreuer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; Richard, A.; Rotroff, D.; Sipes, N.; Dix, D. Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Man-agement. Chemical Research in Toxicology. 2012, 1287–1302.

(77) Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wam-baugh, J.; Setzer, R. W.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Klein-streuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S. Analysis of the Effects of Cell Stress and Cytotoxicity on in Vitro Assay Activity across a Diverse Chemical and Assay Space. Toxicological Sciences 2016, 152 (2), 323–339.

(78) GigaDB Dataset - DOI 10.5524/100351 - Supporting data for "A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay", http://gigadb.org/dataset/100351 (accessed Jul 22, 2020).

(79) Manuals | CellProfiler https://cellprofiler.org/manuals (accessed Jul 22, 2020).

(80) MolVS · PyPI https://pypi.org/project/MolVS/ (accessed Jul 22, 2020).

(81) RDKit http://www.rdkit.org/ (accessed Jul 22, 2020).

(82) Standardization — MolVS 0.1.1 documentation https://molvs.readthedocs.io/en/latest/guide/standardize.html (accessed Jul 22, 2020).

(83) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. J. Chem. Inf. Model. 2009, 49 (1), 108–119.

(84) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. pietro; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. v.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G. L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020, 17 (3), 261–272.

(85) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.

(86) Gomez-Raḿırez; Angel, M.; Jaime, G.; Avilavillanueva, M. Selecting the Most Important Self-Assessed Features for Predicting Conversion to Mild Cognitive Impairment with Random Forest and Permutation-Based Methods. bioRxiv 2019, 785519.

(87) Herbert, S. P.; Ponnambalam, S.; Walker, J. H. Cytosolic Phospholipase A2-α Mediates Endothelial Cell Proliferation and Is Inactivated by Association with the Golgi Apparatus. Molecular Biology of the Cell 2005, 16 (8), 3800–3809.

(88) Zhang, J.; Basu, S.; Rigamonti, D.; Dietz, H. C.; Clatterbuck, R. E. KRIT1 Modulates B1-Integrin-Mediated Endothelial Cell Proliferation. Neurosurgery 2008, 63 (3), 571–578.

(89) DraHeim, K. M.; Huet-Calderwood, C.; Simon, B.; Calderwood, X. D. A. Nuclear Localization of Integrin Cytoplasmic Domainassociated Protein-1 (ICAP1) Influences B1 Integrin Activation and Recruits Krev/Interaction Trapped-1 (KRIT1) to the Nucleus. Journal of Biological Chemistry 2017, 292 (5), 1884–1898.

(90) Wu, Y.; Yu, T.; Gilbertson, T. A.; Zhou, A.; Xu, H.; Nguyen, K. T. Biophysical Assessment of Single Cell Cytotoxicity: Diesel Exhaust Particle-Treated Human Aortic Endothelial Cells. PLoS ONE 2012, 7 (5), e36885.

(91) Eroglu, M.; Derry, W. B. Your Neighbours Matter-Non-Autonomous Control of Apoptosis in Development and Disease. Cell Death and Differentiation. 2016, 1110–1118.

(92) Altman, D. G.; Royston, P. The Cost of Dichotomising Continuous Variables. British Medical Journal. 2006, 1080.

(93) Bois, F. Y. How to Integrate in Vitro PK/PD Information for Toxicity Prediction; 2009.

(94) National Academies of Sciences, E. and M. Application of Modern Toxicology Approaches for Predicting Acute Toxicity for Chemical Defense; National Academies Press, 2015.

(95) Blaauboer, B. J.; Hermens, J.; van Eijkeren, J. Estimating Acute Toxicity Based on in Vitro Cytotoxicity: Role of Biokinetic Modelling. Altex -Heidelberg- 2006, 1 (23), 250–253.

6 APPENDICES

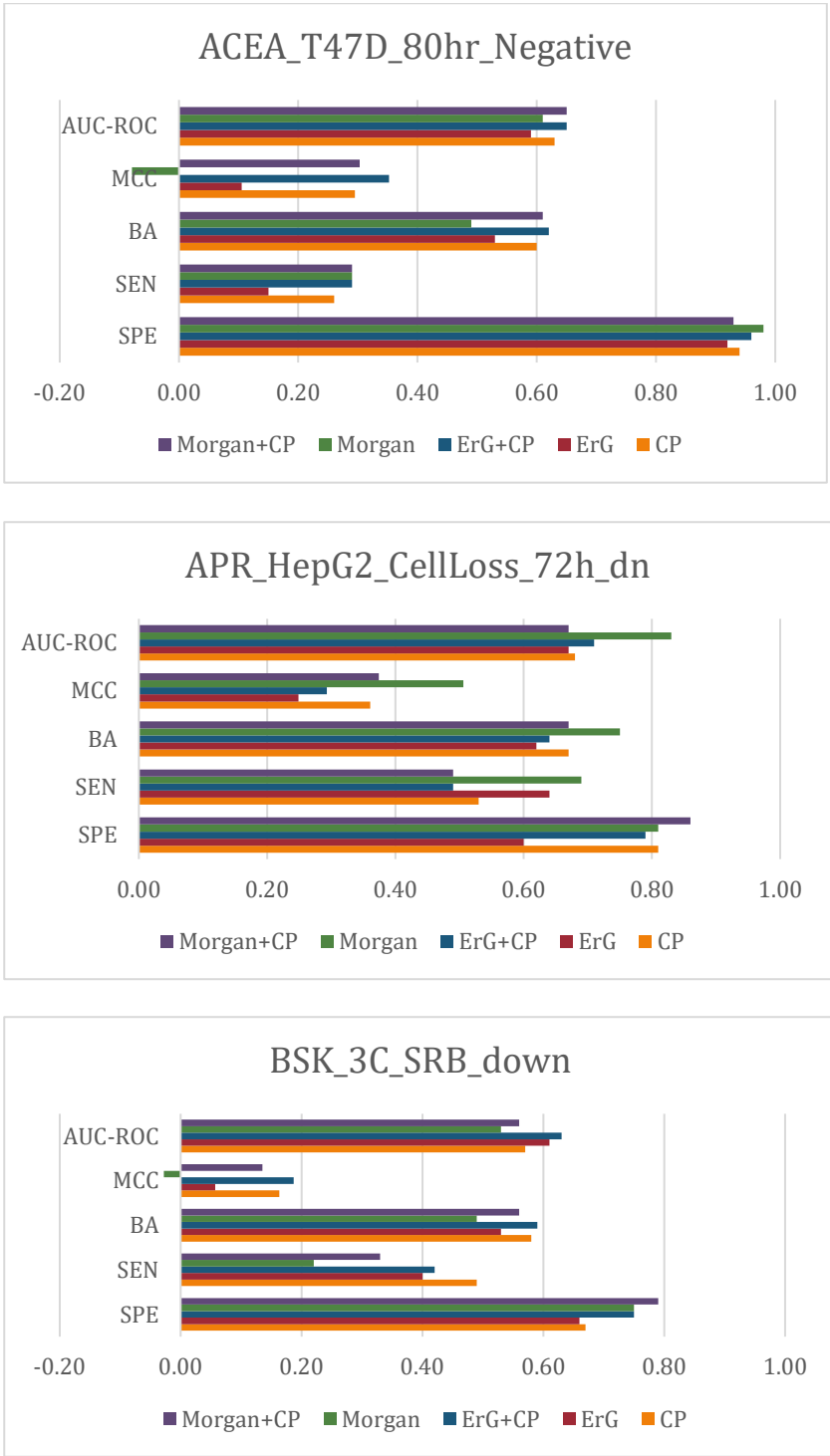
APPENDIX A

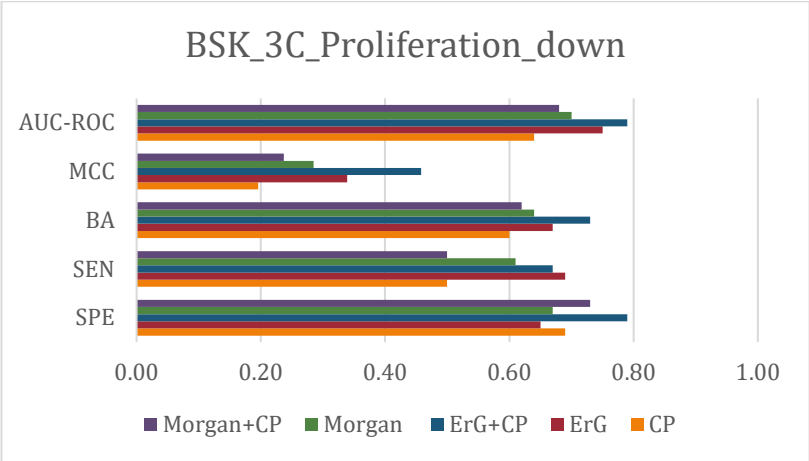
Table A1. Aggregated metrics for random forest model using shuffle stratified splitting.

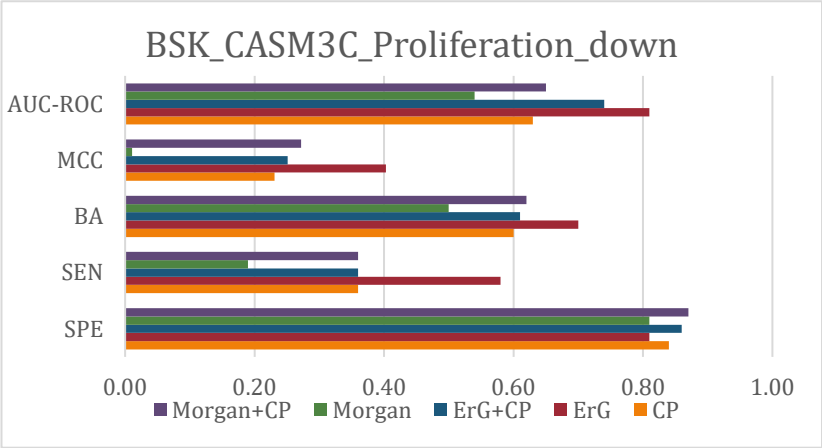
Target	Fingerprint	SPE	SEN	BA	MCC	AUC-ROC
ACEA_T47D_80hr_Negative	CP	0.94	0.26	0.60	0.30	0.63
	ErG	0.92	0.15	0.53	0.11	0.59
	ErG+CP	0.96	0.29	0.62	0.35	0.65
	Morgan	0.98	0.29	0.49	-0.08	0.61
	Morgan+CP	0.93	0.29	0.61	0.30	0.65
APR_HepG2_CellLoss_72h_dn	CP	0.81	0.53	0.67	0.36	0.68
	ErG	0.60	0.64	0.62	0.25	0.67
	ErG+CP	0.79	0.49	0.64	0.29	0.71
	Morgan	0.81	0.69	0.75	0.51	0.83
	Morgan+CP	0.86	0.49	0.67	0.37	0.67
BSK_3C_Proliferation_down	CP	0.69	0.50	0.60	0.20	0.64
	ErG	0.65	0.69	0.67	0.34	0.75
	ErG+CP	0.79	0.67	0.73	0.46	0.79
	Morgan	0.67	0.61	0.64	0.29	0.70
	Morgan+CP	0.73	0.50	0.62	0.24	0.68
BSK_3C_SRB_down	CP	0.67	0.49	0.58	0.16	0.57
	ErG	0.66	0.40	0.53	0.06	0.61
	ErG+CP	0.75	0.42	0.59	0.19	0.63
	Morgan	0.75	0.22	0.49	-0.03	0.53
	Morgan+CP	0.79	0.33	0.56	0.14	0.56
BSK_3C_Vis_down	CP	0.88	0.54	0.71	0.45	0.73
	ErG	0.86	0.32	0.59	0.21	0.67
	ErG+CP	0.85	0.44	0.64	0.32	0.71
	Morgan	0.82	0.32	0.57	0.15	0.63
	Morgan+CP	0.89	0.46	0.68	0.40	0.70
BSK_4H_SRB_down	CP	0.92	0.41	0.66	0.39	0.70

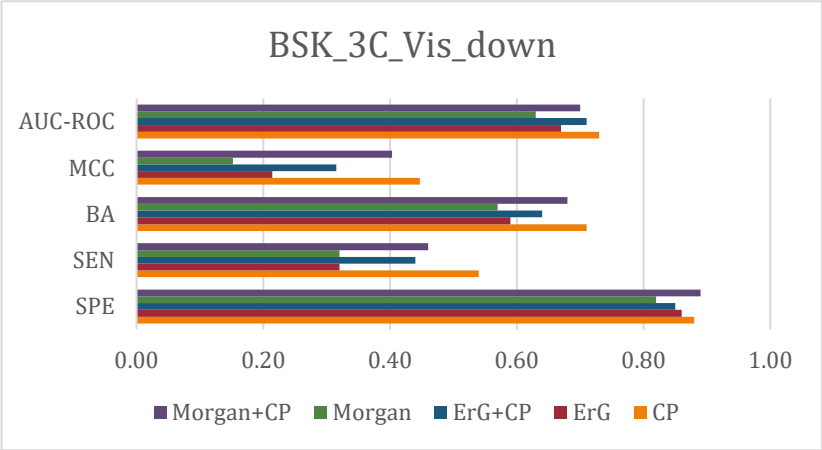
	ErG	0.93	0.26	0.60	0.27	0.70
	ErG+CP	0.93	0.44	0.69	0.44	0.74
	Morgan	0.89	0.06	0.47	-0.08	0.56
	Morgan+CP	0.93	0.38	0.66	0.39	0.72
BSK_CASM3C_Proliferation_down	CP	0.84	0.36	0.60	0.23	0.63
	ErG	0.81	0.58	0.70	0.40	0.81
	ErG+CP	0.86	0.36	0.61	0.25	0.74
	Morgan	0.81	0.19	0.50	0.01	0.54
	Morgan+CP	0.87	0.36	0.62	0.27	0.65
BSK_hDFCGF_Proliferation_down	CP	0.62	0.59	0.60	0.21	0.63
	ErG	0.58	0.84	0.71	0.44	0.76
	ErG+CP	0.64	0.57	0.61	0.21	0.68
	Morgan	0.68	0.70	0.69	0.38	0.73
	Morgan+CP	0.66	0.61	0.63	0.27	0.66
BSK_LPS_SRB_down	CP	0.84	0.49	0.66	0.35	0.71
	ErG	0.79	0.49	0.64	0.29	0.69
	ErG+CP	0.81	0.51	0.66	0.33	0.73
	Morgan	0.87	0.10	0.48	-0.05	0.58
	Morgan+CP	0.85	0.46	0.66	0.34	0.71
BSK_SAg_Proliferation_down	CP	0.72	0.42	0.57	0.15	0.58
	ErG	0.69	0.52	0.61	0.21	0.63
	ErG+CP	0.74	0.46	0.60	0.21	0.63
	Morgan	0.62	0.54	0.58	0.16	0.60
	Morgan+CP	0.74	0.40	0.59	0.15	0.59

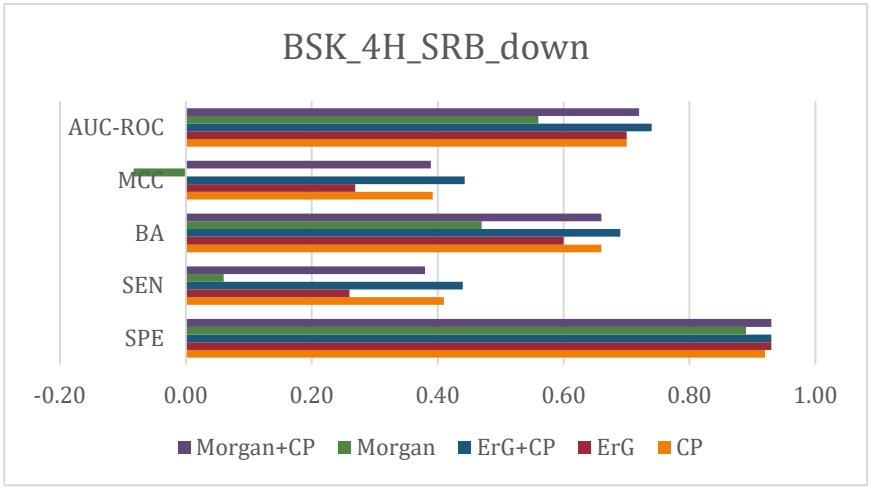
Figure A1. Aggregated metrics for random forest model using shuffle stratified splitting.

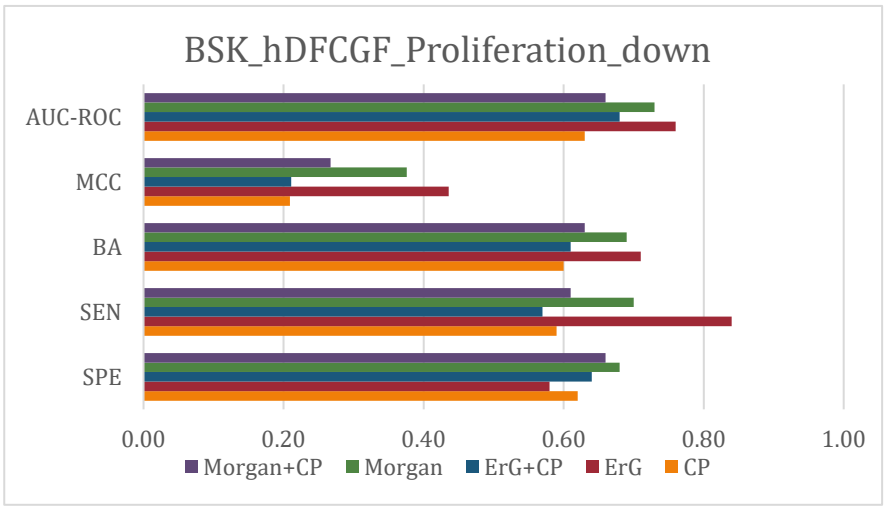


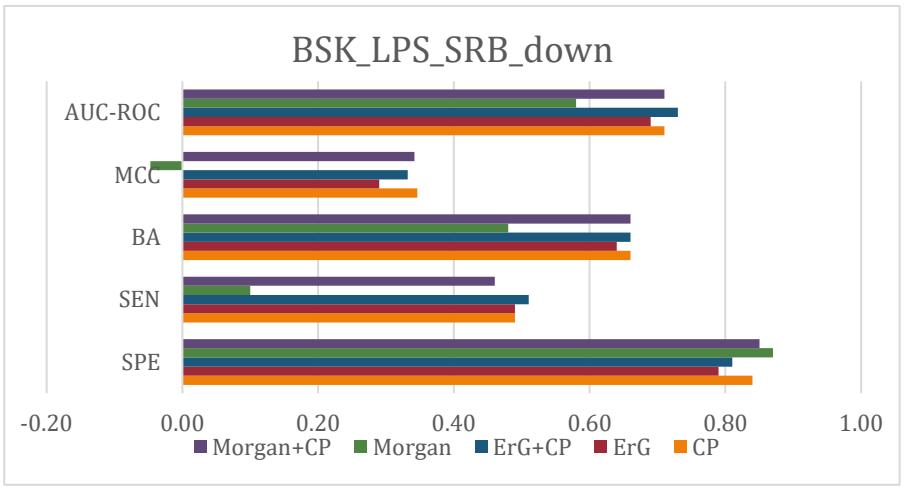

BSK_3C_Proliferation_down
AUC-ROC
MCC
BA
SEN
SPE
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP


BSK_CASM3C_Proliferation_down
AUC-ROC
MCC
BA
SEN
SPE
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP


BSK_3C_Vis_down
AUC-ROC
MCC
BA
SEN
SPE
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP


BSK_4H_SRB_down
AUC-ROC
MCC
BA
SEN
SPE
-0.20
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP


BSK_hDFCGF_Proliferation_down
AUC-ROC
MCC
BA
SEN
SPE
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP


BSK_LPS_SRB_down
AUC-ROC
MCC
BA
SEN
SPE
-0.20
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP

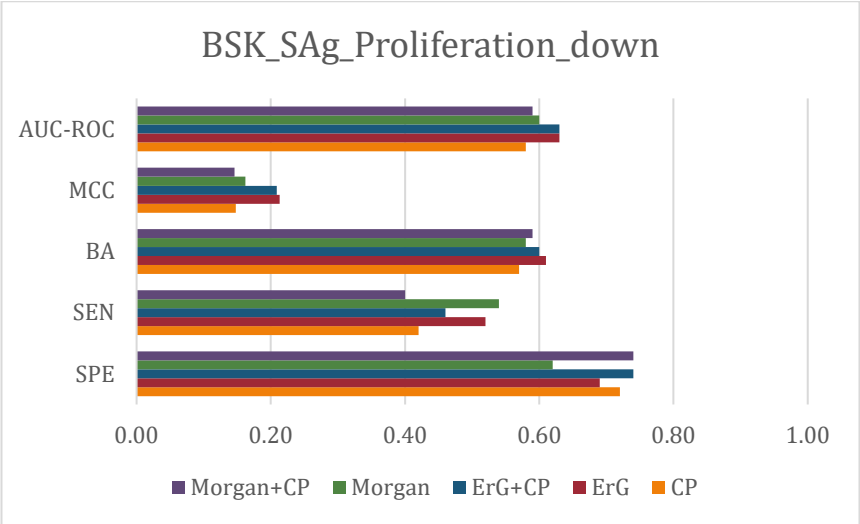

BSK_SAg_Proliferation_down
AUC-ROC
MCC
BA
SEN
SPE
0.00
0.20
0.40
0.60
0.80
1.00
Morgan+CP
Morgan
ErG+CP
ErG
CP

Table A2. Cluster-averaged metrics for random forest model using leave out one cluster splitting

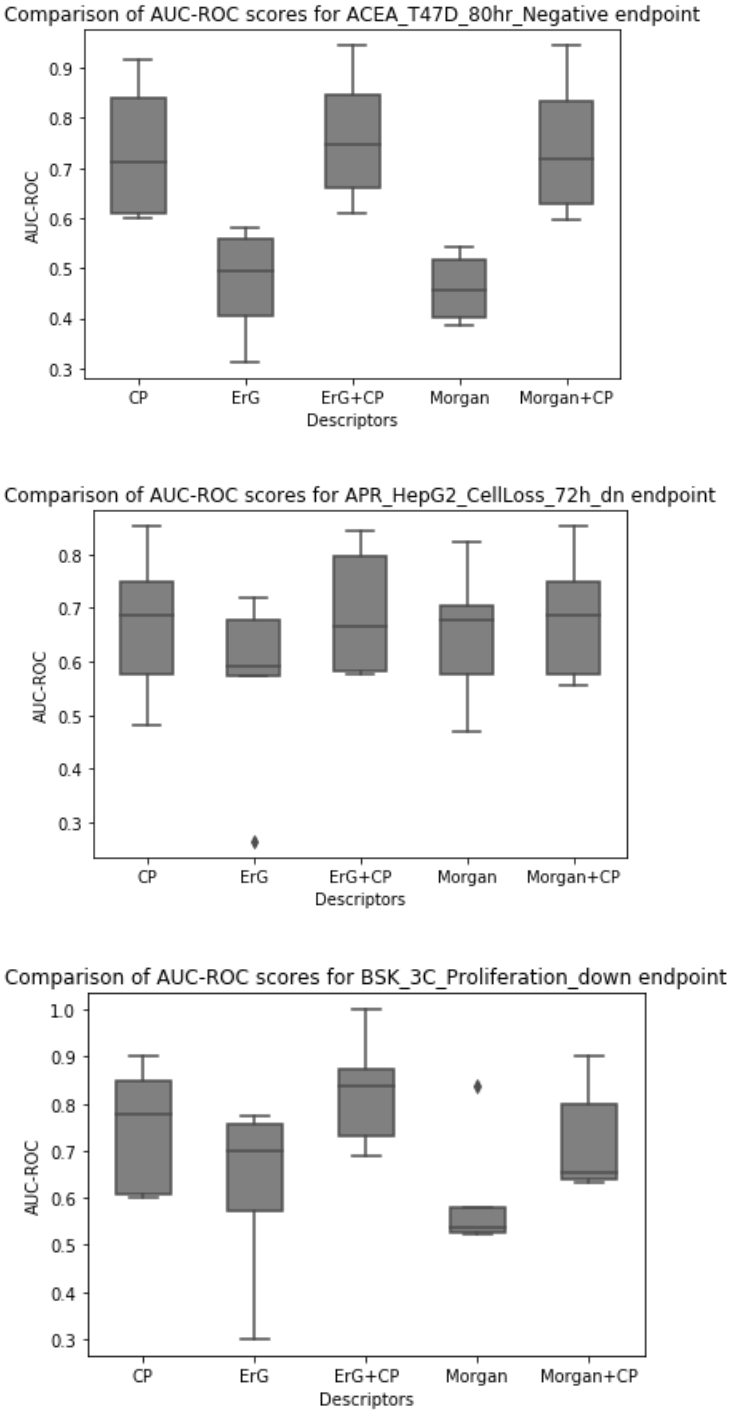
Target	Fingerprint	SPE	SEN	BA	MCC	AUC-ROC
ACEA_T47D_80hr_Negative	CP	0.93 ± 0.06	0.33 ± 0.13	0.63 ± 0.06	0.28 ± 0.14	0.74 ± 0.13
	ErG	0.92 ± 0.11	0.09 ± 0.12	0.5 ± 0.0	0.01 ± 0.01	0.47 ± 0.11
	ErG+CP	0.94 ± 0.06	0.28 ± 0.13	0.61 ± 0.04	0.26 ± 0.08	0.76 ± 0.13
	Morgan	0.9 ± 0.18	0.0 ± 0.0	0.45 ± 0.09	-0.05 ± 0.09	0.46 ± 0.07
	Morgan+CP	0.92 ± 0.08	0.33 ± 0.17	0.63 ± 0.08	0.28 ± 0.18	0.74 ± 0.14
APR_HepG2_CellLoss_72h_dn	CP	0.65 ± 0.22	0.54 ± 0.17	0.59 ± 0.05	0.17 ± 0.08	0.67 ± 0.13
	ErG	0.52 ± 0.25	0.52 ± 0.34	0.52 ± 0.11	0.03 ± 0.24	0.56 ± 0.16
	ErG+CP	0.61 ± 0.24	0.56 ± 0.2	0.58 ± 0.09	0.15 ± 0.17	0.69 ± 0.11
	Morgan	0.65 ± 0.12	0.49 ± 0.21	0.57 ± 0.09	0.12 ± 0.14	0.65 ± 0.12
	Morgan+CP	0.7 ± 0.17	0.49 ± 0.2	0.6 ± 0.05	0.17 ± 0.07	0.68 ± 0.11
BSK_3C_Proliferation_down	CP	0.73 ± 0.24	0.64 ± 0.14	0.68 ± 0.11	0.28 ± 0.15	0.75 ± 0.12
	ErG	0.54 ± 0.28	0.58 ± 0.29	0.56 ± 0.13	0.12 ± 0.21	0.62 ± 0.18
	ErG+CP	0.73 ± 0.27	0.7 ± 0.13	0.71 ± 0.09	0.35 ± 0.1	0.83 ± 0.11
	Morgan	0.52 ± 0.27	0.57 ± 0.24	0.54 ± 0.09	0.07 ± 0.17	0.6 ± 0.12
	Morgan+CP	0.75 ± 0.23	0.54 ± 0.12	0.65 ± 0.12	0.22 ± 0.18	0.73 ± 0.11
BSK_3C_SRB_down	CP	0.72 ± 0.16	0.47 ± 0.21	0.6 ± 0.09	0.18 ± 0.15	0.67 ± 0.12
	ErG	0.73 ± 0.3	0.23 ± 0.34	0.48 ± 0.05	-0.04 ± 0.09	0.53 ± 0.12
	ErG+CP	0.69 ± 0.27	0.52 ± 0.24	0.61 ± 0.12	0.2 ± 0.21	0.69 ± 0.1
	Morgan	0.62 ± 0.22	0.43 ± 0.32	0.52 ± 0.14	0.06 ± 0.26	0.51 ± 0.17
	Morgan+CP	0.75 ± 0.15	0.47 ± 0.21	0.61 ± 0.07	0.2 ± 0.13	0.66 ± 0.06
BSK_3C_Vis_down	CP	0.83 ± 0.07	0.7 ± 0.24	0.76 ± 0.14	0.49 ± 0.25	0.74 ± 0.12
	ErG	0.71 ± 0.3	0.43 ± 0.4	0.57 ± 0.17	0.11 ± 0.37	0.55 ± 0.12
	ErG+CP	0.76 ± 0.08	0.62 ± 0.26	0.69 ± 0.14	0.34 ± 0.24	0.73 ± 0.14
	Morgan	0.72 ± 0.23	0.31 ± 0.27	0.52 ± 0.04	0.02 ± 0.08	0.53 ± 0.1
	Morgan+CP	0.81 ± 0.04	0.54 ± 0.18	0.68 ± 0.1	0.34 ± 0.17	0.7 ± 0.12
	CP	0.85 ± 0.12	0.53 ± 0.25	0.69 ± 0.07	0.4 ± 0.11	0.74 ± 0.09

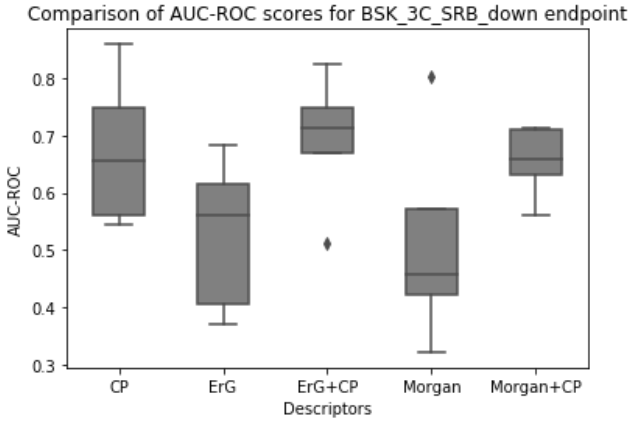
BSK_4H_ SRB_down	ErG	0.71 ± 0.2	0.54 ± 0.32	0.62 ± 0.08	0.22 ± 0.12	0.67 ± 0.08
	ErG+CP	0.89 ± 0.07	0.53 ± 0.25	0.71 ± 0.1	0.43 ± 0.16	0.78 ± 0.12
	Morgan	0.86 ± 0.27	0.1 ± 0.2	0.48 ± 0.04	-0.03 ± 0.06	0.59 ± 0.13
	Morgan+CP	0.92 ± 0.07	0.53 ± 0.25	0.72 ± 0.09	0.49 ± 0.14	0.72 ± 0.11
BSK_CASM3C_ Proliferation_ down	CP	0.9 ± 0.1	0.48 ± 0.19	0.69 ± 0.07	0.37 ± 0.13	0.77 ± 0.12
	ErG	0.7 ± 0.1	0.5 ± 0.24	0.6 ± 0.07	0.16 ± 0.11	0.75 ± 0.11
	ErG+CP	0.91 ± 0.09	0.55 ± 0.3	0.73 ± 0.11	0.44 ± 0.15	0.81 ± 0.1
	Morgan	0.88 ± 0.12	0.22 ± 0.19	0.55 ± 0.08	0.1 ± 0.15	0.5 ± 0.17
	Morgan+CP	0.94 ± 0.1	0.45 ± 0.2	0.69 ± 0.1	0.43 ± 0.21	0.76 ± 0.12
BSK_hDFCGF_ Proliferation_ down	CP	0.71 ± 0.27	0.65 ± 0.27	0.68 ± 0.15	0.25 ± 0.22	0.71 ± 0.13
	ErG	0.61 ± 0.37	0.65 ± 0.18	0.63 ± 0.1	0.15 ± 0.16	0.58 ± 0.08
	ErG+CP	0.64 ± 0.37	0.62 ± 0.28	0.63 ± 0.11	0.17 ± 0.14	0.69 ± 0.14
	Morgan	0.79 ± 0.18	0.55 ± 0.2	0.67 ± 0.15	0.22 ± 0.21	0.77 ± 0.17
	Morgan+CP	0.7 ± 0.28	0.66 ± 0.31	0.68 ± 0.16	0.26 ± 0.24	0.68 ± 0.18
BSK_LPS_SRB_ down	CP	0.82 ± 0.16	0.41 ± 0.13	0.61 ± 0.08	0.25 ± 0.21	0.7 ± 0.06
	ErG	0.58 ± 0.36	0.49 ± 0.16	0.54 ± 0.14	0.07 ± 0.3	0.62 ± 0.15
	ErG+CP	0.79 ± 0.17	0.52 ± 0.23	0.66 ± 0.08	0.31 ± 0.16	0.7 ± 0.09
	Morgan	0.75 ± 0.21	0.31 ± 0.37	0.53 ± 0.14	0.03 ± 0.22	0.55 ± 0.19
	Morgan+CP	0.81 ± 0.18	0.38 ± 0.15	0.59 ± 0.04	0.23 ± 0.15	0.72 ± 0.06
BSK_SAg_ Proliferation_ down	CP	0.68 ± 0.17	0.49 ± 0.23	0.58 ± 0.09	0.16 ± 0.15	0.65 ± 0.15
	ErG	0.49 ± 0.37	0.5 ± 0.17	0.49 ± 0.16	0.0 ± 0.28	0.58 ± 0.19
	ErG+CP	0.63 ± 0.24	0.59 ± 0.31	0.61 ± 0.1	0.22 ± 0.19	0.71 ± 0.12
	Morgan	0.6 ± 0.12	0.41 ± 0.14	0.51 ± 0.08	0.01 ± 0.15	0.51 ± 0.15
	Morgan+CP	0.72 ± 0.17	0.52 ± 0.23	0.62 ± 0.09	0.23 ± 0.15	0.7 ± 0.1

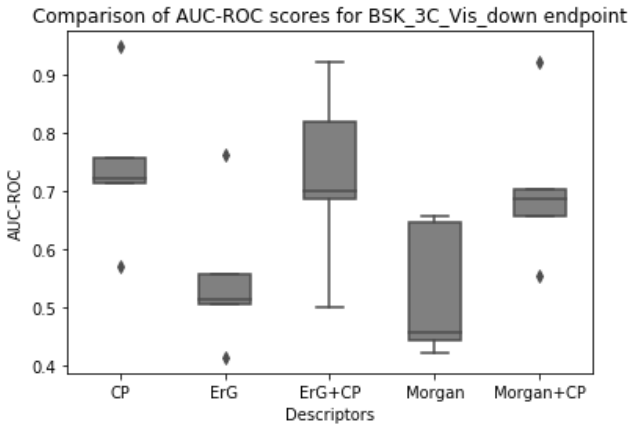
(For the ACEA_T47D_80hr_Negative assay, only 4 clusters were used based on PCA and hence a group 4-fold was used in the outer loop for this assay.)

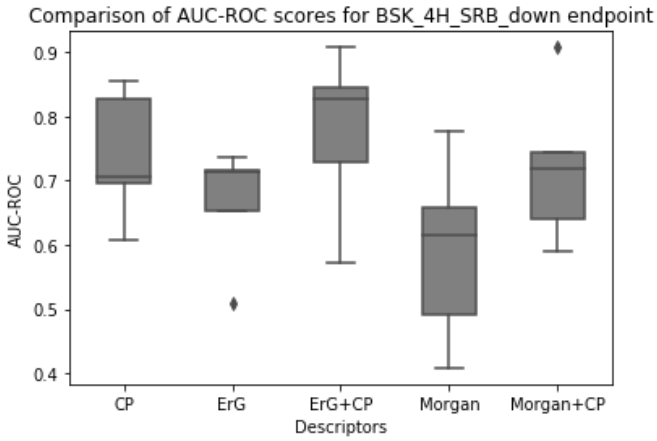
APPENDIX B

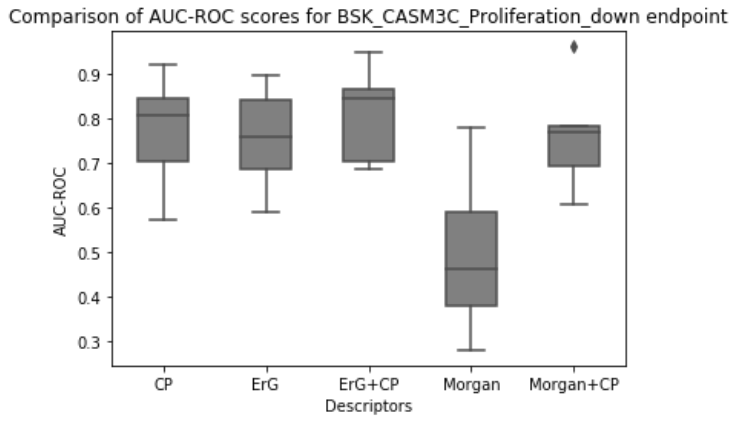
Figure B1. AUC-ROC Performance of Cluster Averaged Models

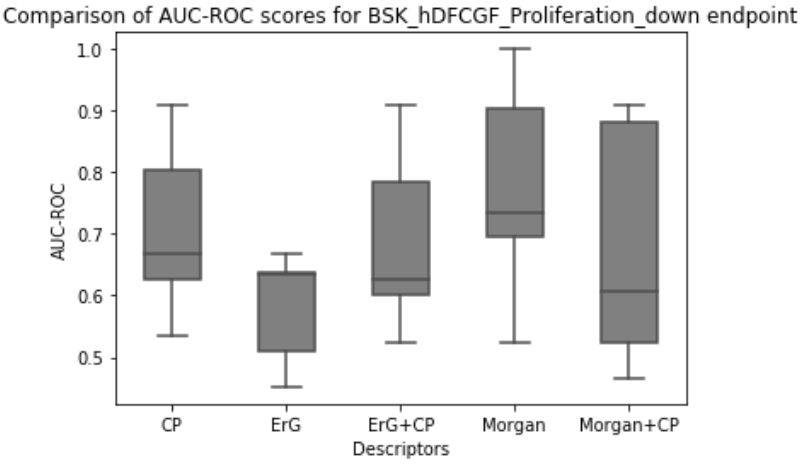


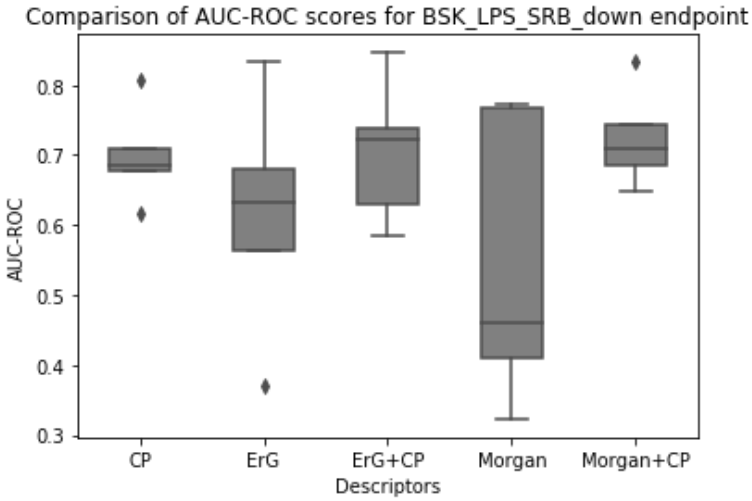

Comparison of AUC-ROC scores for BSK_3C_SRB_down endpoint
AUC-ROC
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of AUC-ROC scores for BSK_3C_Vis_down endpoint
AUC-ROC
0.9
0.8
0.7
0.6
0.5
0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of AUC-ROC scores for BSK_4H_SRB_down endpoint
AUC-ROC
0.9
0.8
0.7
0.6
0.5
0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of AUC-ROC scores for BSK_CASM3C_Proliferation_down endpoint
AUC-ROC
0.9
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of AUC-ROC scores for BSK_hDFCGF_Proliferation_down endpoint
AUC-ROC
1.0
0.9
0.8
0.7
0.6
0.5
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of AUC-ROC scores for BSK_LPS_SRB_down endpoint
AUC-ROC
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

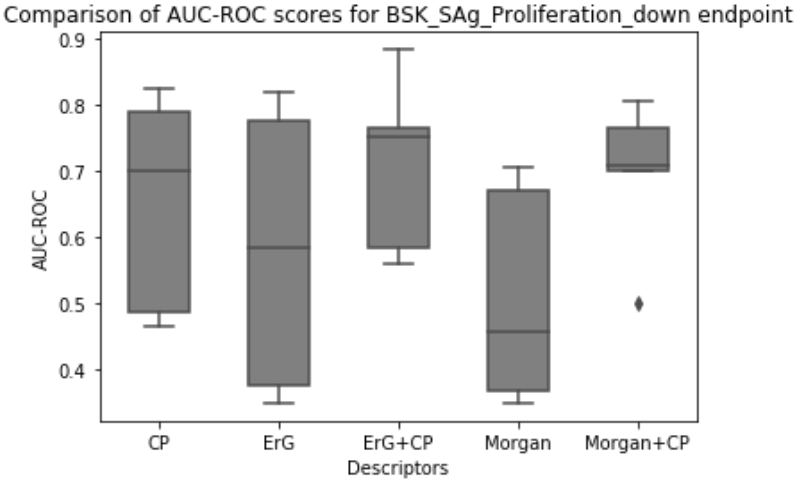
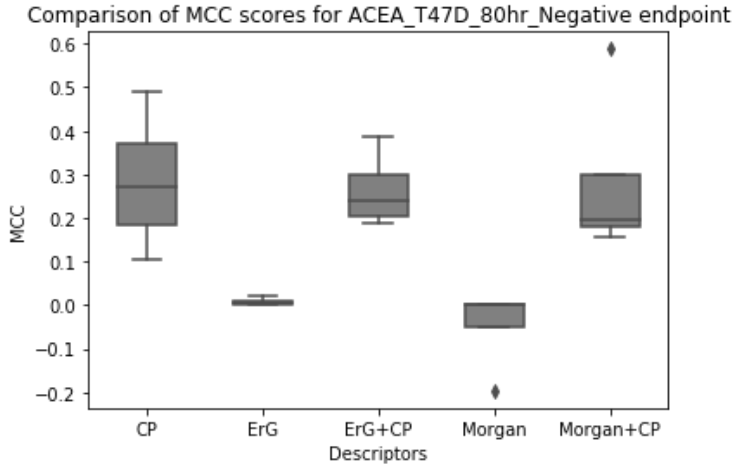
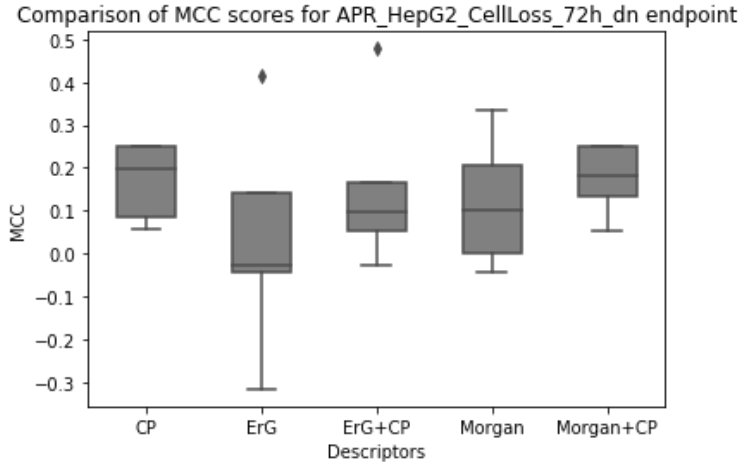
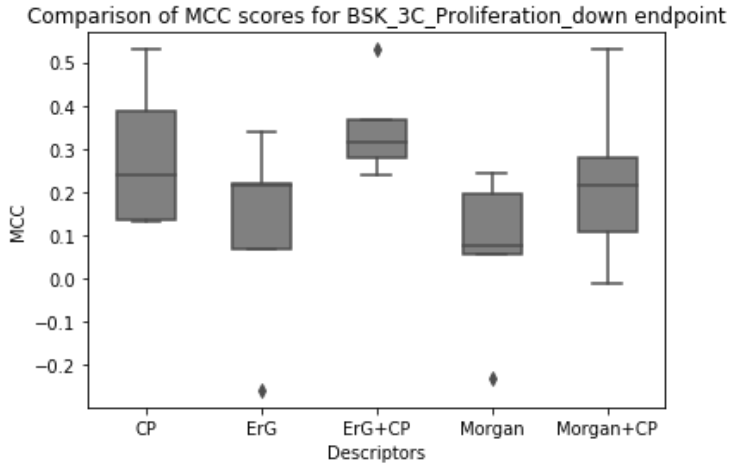
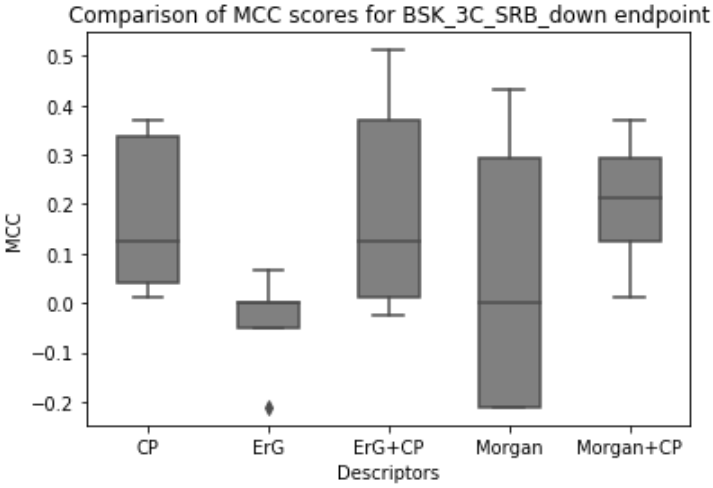

Comparison of AUC-ROC scores for BSK_SAg_Proliferation_down endpoint
AUC-ROC
0.9
0.8
0.7
0.6
0.5
0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

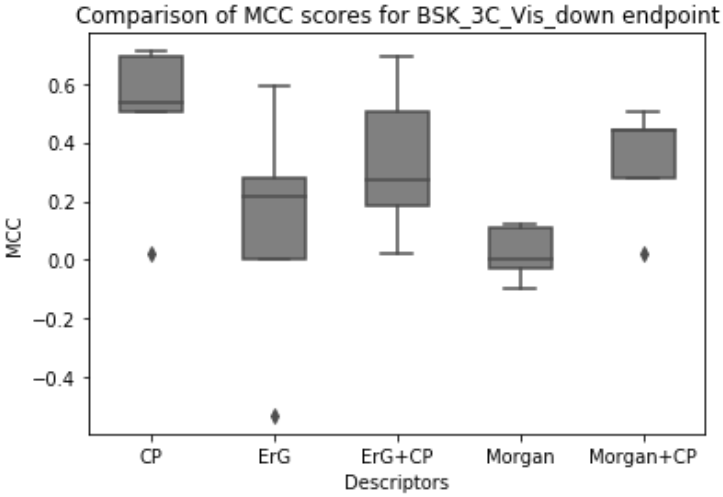
Figure B2. MCC Performance of Cluster Averaged Models

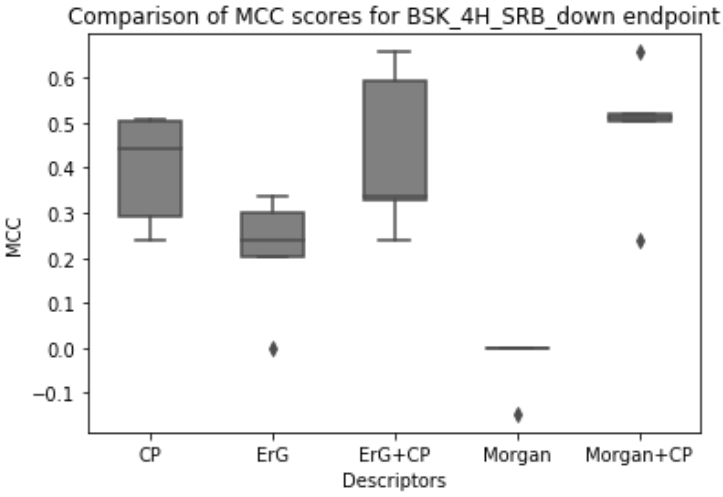


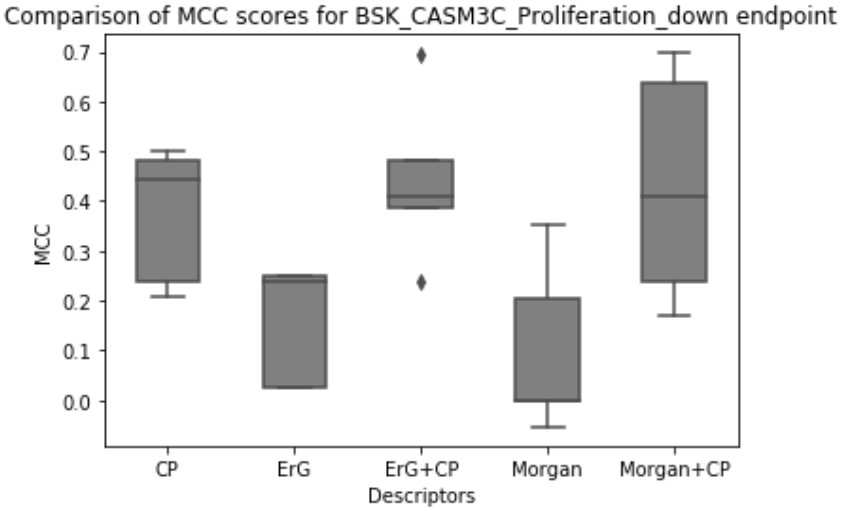


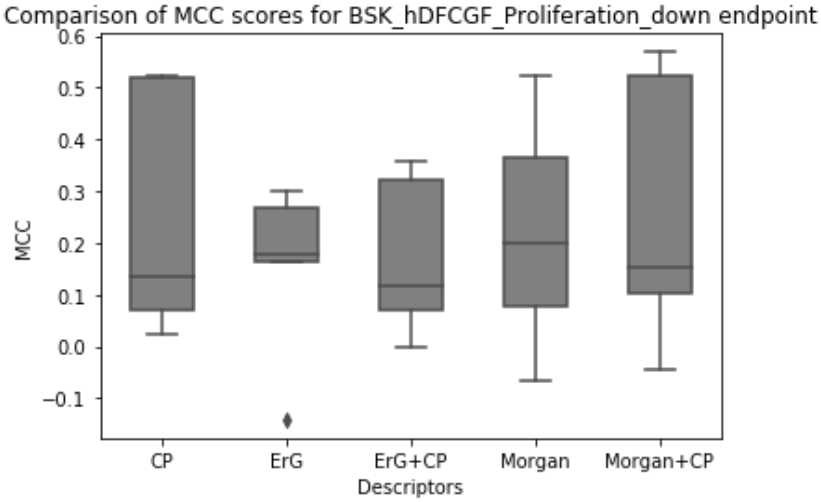


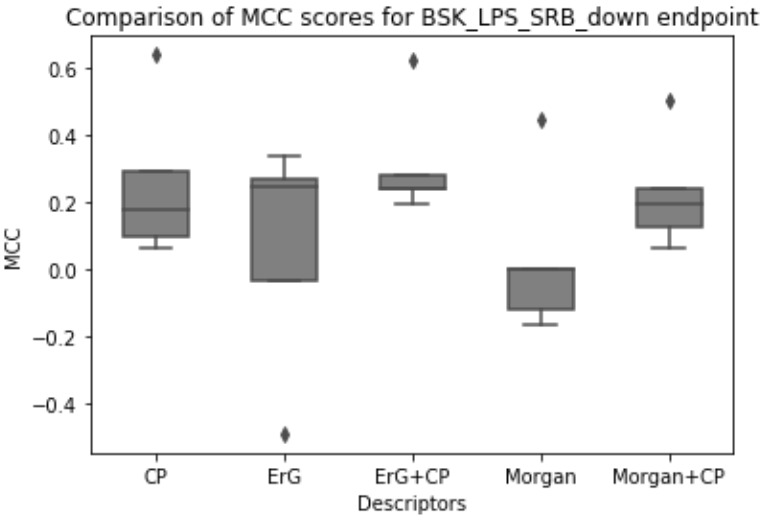

Comparison of MCC scores for BSK_3C_SRB_down endpoint
MCC
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of MCC scores for BSK_3C_Vis_down endpoint
MCC
0.6
0.4
0.2
0.0
−0.2
−0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of MCC scores for BSK_4H_SRB_down endpoint
MCC
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of MCC scores for BSK_CASM3C_Proliferation_down endpoint
MCC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of MCC scores for BSK_hDFCGF_Proliferation_down endpoint
MCC
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of MCC scores for BSK_LPS_SRB_down endpoint
MCC
0.6
0.4
0.2
0.0
−0.2
−0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

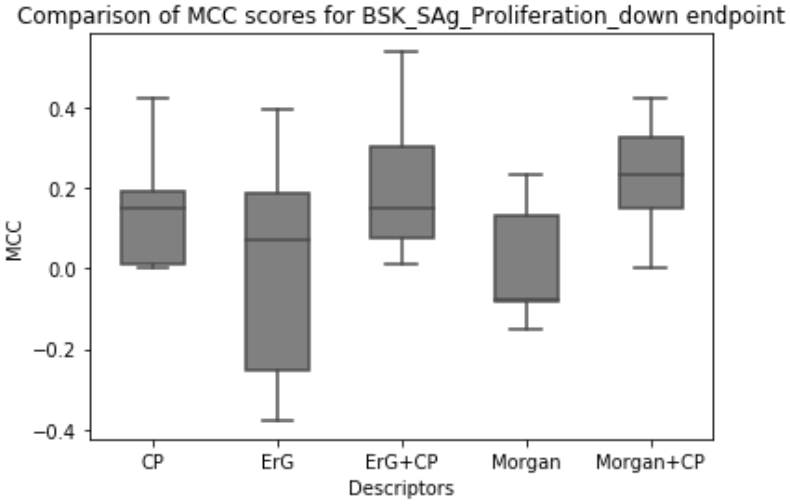
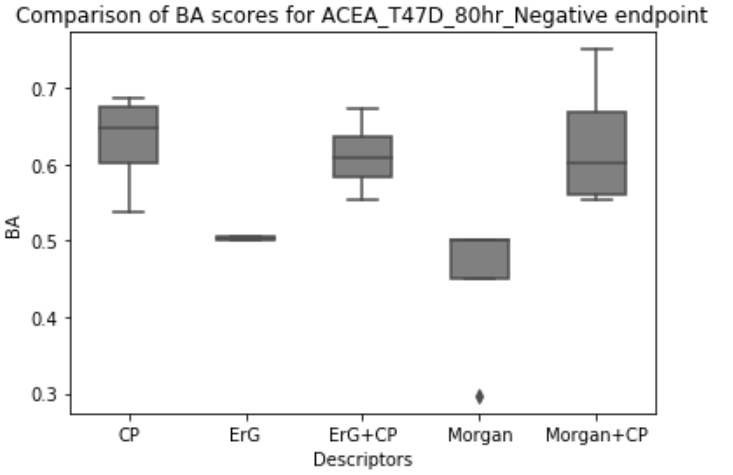
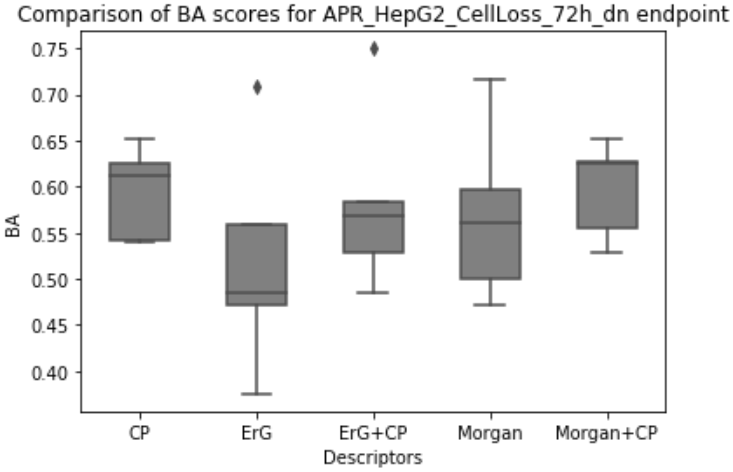
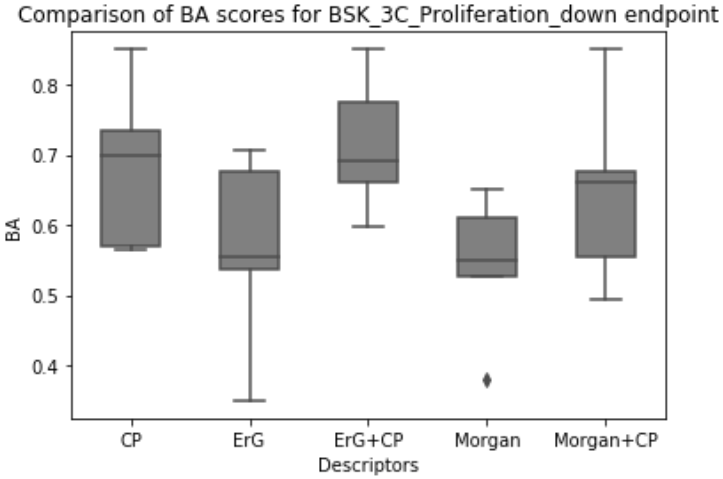
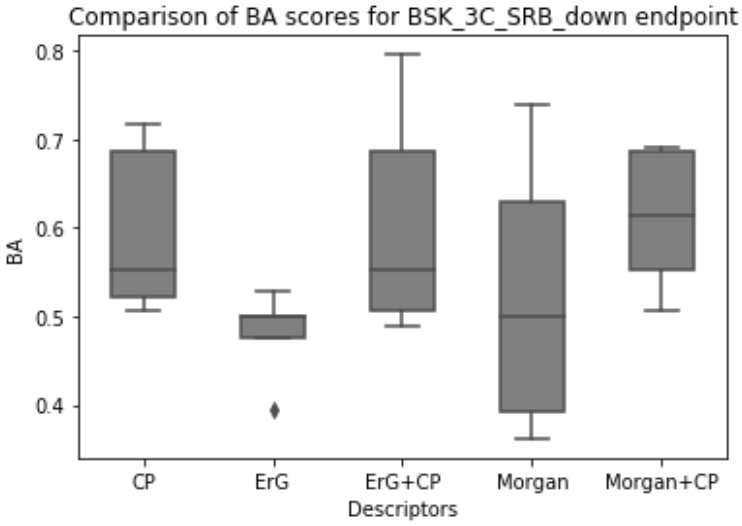

Comparison of MCC scores for BSK_SAg_Proliferation_down endpoint
MCC
0.4
0.2
0.0
−0.2
−0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

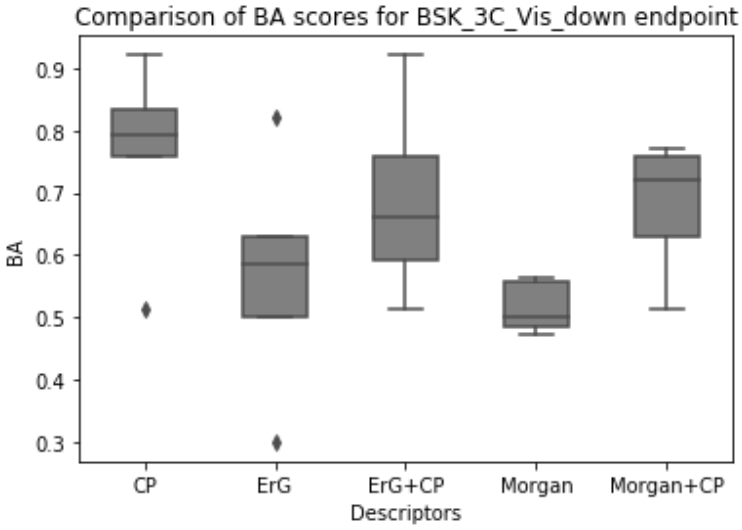
Figure B3. BA Performance of Cluster Averaged Models

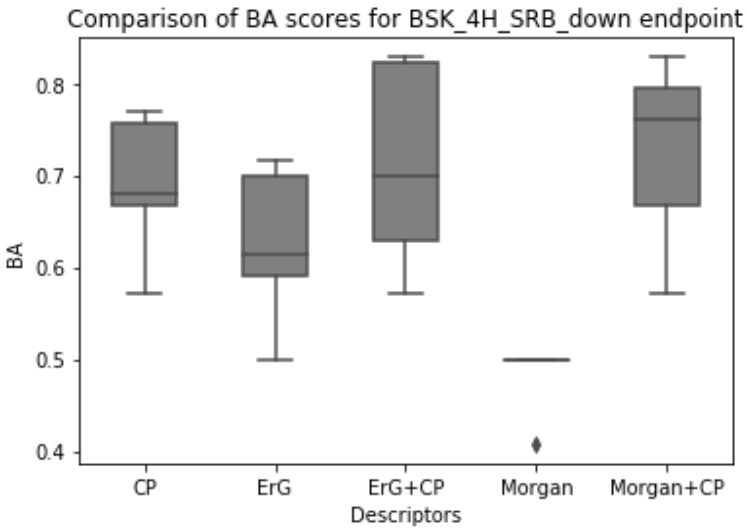


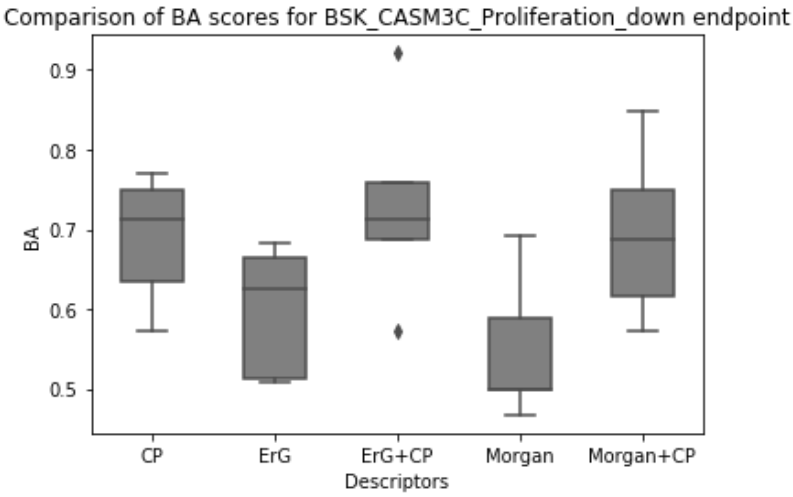


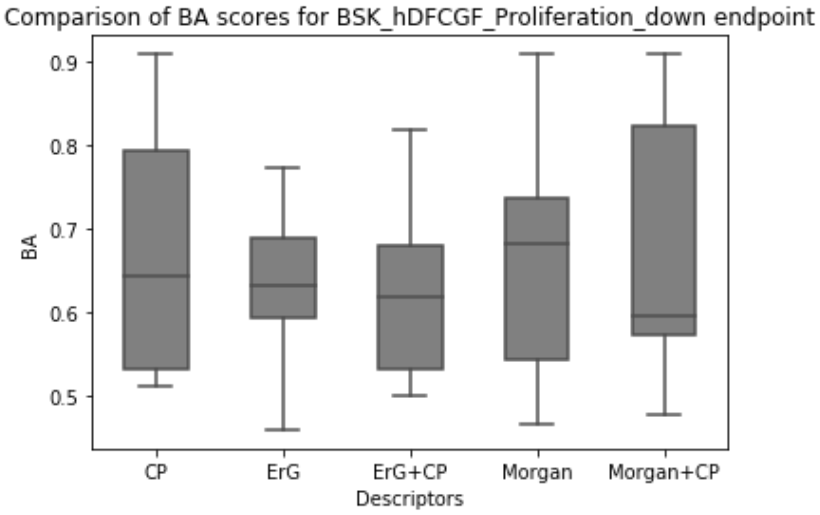


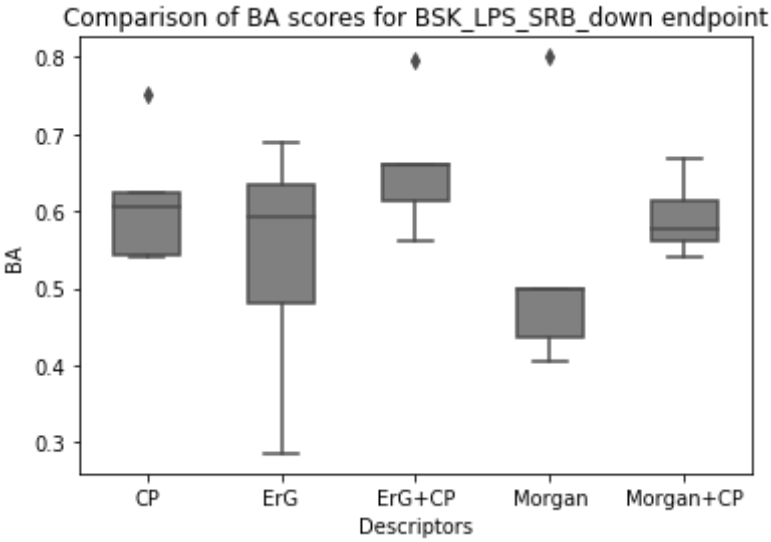

Comparison of BA scores for BSK_3C_SRB_down endpoint
BA
0.8
0.7
0.6
0.5
0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

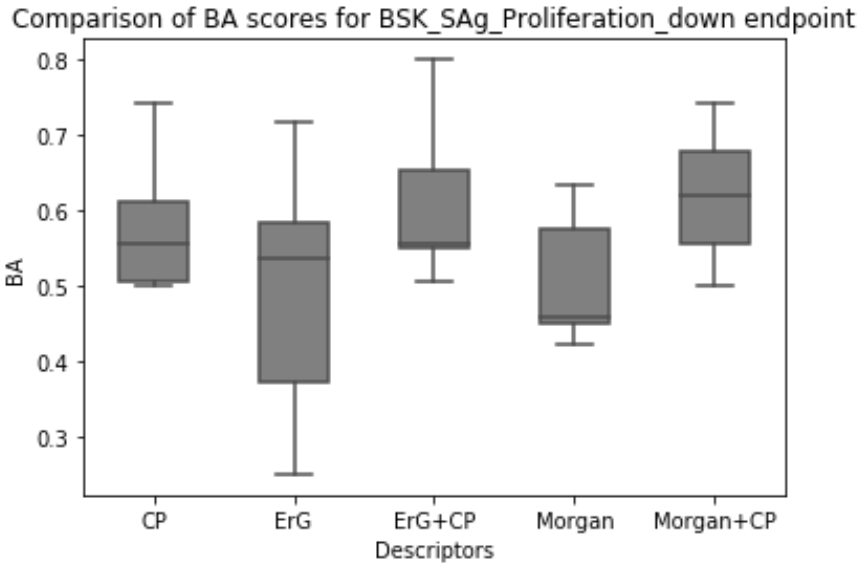

Comparison of BA scores for BSK_3C_Vis_down endpoint
BA
0.9
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of BA scores for BSK_4H_SRB_down endpoint
BA
0.8
0.7
0.6
0.5
0.4
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of BA scores for BSK_CASM3C_Proliferation_down endpoint
BA
0.9
0.8
0.7
0.6
0.5
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of BA scores for BSK_hDFCGF_Proliferation_down endpoint
BA
0.9
0.8
0.7
0.6
0.5
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of BA scores for BSK_LPS_SRB_down endpoint
BA
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors


Comparison of BA scores for BSK_SAg_Proliferation_down endpoint
BA
0.8
0.7
0.6
0.5
0.4
0.3
CP
ErG
ErG+CP
Morgan
Morgan+CP
Descriptors

APPENDIX C

Table C1. Distribution of five top contributing feature using 10-fold permutation importance having a positive permutation score for all assay endpoints in this work; the number of features for each assay belonging to a given compartment and feature group.

Assay	Biological process/ Assay Design/ Biological target (biological target)	Cytoplasm				Nuclei			Cells				
		Granularity	Intensity	Texture	Correlation	Granularity	Texture	Correlation	Granularity	Texture	Correlation	Neighbours	Radial
BSK_3C_SRB_down	Cytotoxicity SRB Protein content (cell death)					1				1		1	2
BSK_4H_SRB_down		1						1	1			1	1
BSK_LPS_SRB_down			2		1			1					1
ACEA_T47D_80hr_Negative	Proliferation decrease Real-time cell growth kinetics		2			1	1		1				
APR_HepG2_CellLoss_72h_dn	Proliferation decrease Cell number (cell death)	1			1	1	1		1				
BSK_3C_Vis_down	Proliferation decrease Cell phenotype (cell morphology)	1				2			1			1	
BSK_3C_Proliferation_down	Proliferation decrease Protein content (cell proliferation)				1				1		1	2	
BSK_CASM3C_Proliferation_down*					1			1	1			1	
BSK_hDFCGF_Proliferation_down				2	1				1		1		
BSK_SAg_Proliferation_down		1			2				1		1		

*Only 4 important features could be determined across the two best-performing folds for endpoint BSK_CASM3C_Proliferation_down