Using Machine Learning to Predict Genotoxic Mode of Action from High Content Imaging Data

Master of Biomedical Research

Department of Surgery & Cancer

Faculty of Medicine

Imperial College London

2019-03-10

Principal supervisor: Dr. Timothy Ebbels

External supervisor: Dr. Ann Doherty and Dr. Ahlberg Helgee Ernst

Author: Xiaokai Cui (student of Data Science Stream)

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline. This project was undertaken with the cooperation of toxicology and machine learning experts of AstraZeneca and I fully acknowledge their contribution to my project. The experiments of biology and toxicology were designed and implemented by the Genetic Toxicology team lead by Dr. Ann Doherty at AstraZeneca. The datasets used in this project were prepared by Dr. Amy Wilson. The Venn-Abers predictor and the heatmap of clusters were two useful methods suggested by Dr. Ahlberg Helgee Ernst and Dr. Timothy Ebbels respectively, which contributed a lot to this thesis.

Acknowledgements

I would like to acknowledge and thank my supervisors Dr. Timothy Ebbels, Dr. Ann Doherty and Dr. Ahlberg Helgee Ernst. Their supervision is full of wisdom, kindness and patience, I learned much knowledge of data science, toxicology and many skills of communication and cooperation from them. Dr. Amy Wilson is the one who I can always seek help from, she introduced the toxicology side to me and helped me with preparing and checking experimental data, I cannot achieve anything without her help in this project. I also got the support from Dr. Joanne Elloway and Dr. Muireann Coen, it is my pleasure to have the opportunity of working with them. I got valuable advice from my supervisors and Amy when I wrote this thesis, their help means a lot to me and I cherish the experience of cooperating with them in the project.

Table of Contents

Abbreviations

4NQO:	*4-Nitroquinoline N-oxide*
γH2AX:	*the phosphorylated histone H2AX*
AD:	*the AstraZeneca dataset*
AUC:	*the areas under the curve*
DL:	*dose levels*
EC50:	*maximal effective concentration for cell number*
FIML:	*full information maximum likelihood*
G-features:	*4 features selected by experts for distinguishing compounds*
IVM:	*in vitro micronucleus*
MCCV:	*Monte Carlo cross-validation*
MEGA Screen:	*multi endpoint genotoxicity screens*
ML:	*machine learning*
MNO:	*“Micronuclei – Number of Objects”*
MOA:	*mechanism of action*
PCA:	*principal component analysis*
SMOTE:	*synthetic minority over-sampling technique*
t-SNE:	*t-Distributed Stochastic Neighbour Embedding*
VAP:	*Venn-Abers predictors*
VD:	*the validation dataset*

Abstract

Genotoxicants need to be detected before the clinical test period in order to avoid unnecessary time-consuming tests when developing new drugs, and it is also much more economical if potential genotoxic compounds can be excluded from the candidates of new drugs as soon as possible. A new version of in vitro micronucleus assays which is one of the standard methods to assess genotoxicity called MEGA Screen has been developed in AstraZeneca and it is used to distinguish different categories of genotoxicants from negative compounds. Although researchers have used the results generated from MEGA Screen for detecting genotoxicants for years, the data was not fully utilised and the efficiency of the detection was under development. This study aimed to find an approach equipped with data science technologies and toxicology knowledge to improve the efficiency and the performance when distinguishing genotoxicants from negatives in AstraZeneca. In this project, a creative data visualisation method has been developed to help researchers with a much more instinctive way to recognise and understand the information contained in the MEGA Screen data. A supervised machine learning method (Random Forest) and an unsupervised machine learning method (K-means) are creatively combined together to provide suggestion of feature selection to researchers. Several classification models are combined with Venn-Abers predictors to provide not only predictions which are as accurate as 91.9% but more importantly the confidence of predictions to researchers, so that the predictions can be used more flexible depending on different requirements. In summary, this research has developed several creative and valuable data science tools in the workflow of genotoxicants detection, illustrating a promising step towards applying data science approaches to routine drug safety assessment in pharmaceutical industry.

1. Introduction

1.1 Background

In the modern pharmaceutical industry, the Research and Development (R&D) of new medicines is one of the most challenging issues, not only because of the high cost of time and funding, but more importantly because of the uncertainty throughout the R&D process and the safety assurances which are required to avoid the fatal failure of drugs. In general, new drugs R&D can be divided into 2 periods, one is the researching period which includes target identification, biochemical mechanism, efficient molecules identification and animal (in vivo) tests for safety and efficacy, the other one is the clinical testing period which includes three phases focusing on from initial efficacy to long-term effects of drugs in humans (1).

In the clinical testing period, Phase II and III are the most vulnerable phrases which cause the majority of disruption in new medicine development (2). According to the research by Steven Paul, two thirds of compounds in Phase II fail entering next phase, and the corresponding ratio for the compounds in Phase III is 70% (3). Safety is the second and the third most important reasons for clinical testing failures of Phases III and II respectively compared to efficacy, commercial, DMPK (drug metabolism and pharmacokinetics) and Misc. (miscellaneous factors) (2). The failures in clinical period are much more expensive than the ones in researching period, thus the main aim is to make harmful drugs fail before clinical period in order to reduce cost.

1.1.1 Toxicology in pharmaceutical industry

Toxicology plays an important role in pharmaceutical industry, and the toxic potential of a candidate drug is tested in different ways through the lifecycle of drugs development. In spite of multiple toxicity assays which are applied in the process of pharmaceutical development, it is not enough to guarantee the completely safety of new drugs. For example, a number of drugs have been withdrawn from market and the clinical testing period because they would inhibit hERG (The human ether-a-go-go related gene) and lead to a potentially fatal arrhythmia (4).

Genetic toxicology is an important area of research within the pharmaceutical industry, genetic toxicology assays assess the ability of a compound to induce DNA damage in the form of mutations, deletions and changes in the number of chromosomes (5), which can all lead to cancer (6). In the pharmaceutical industry, the assays used to detect genotoxicity are comprehensive as there is no single experiment that can alone detect all of the classes of DNA damages discussed above and ensure the genetic safety of a compound. In general, both in vitro and in vivo tests are used for genotoxicity detection as they complement to each other. The 3-test battery is comprised of a bacterial reverse mutation test, an in vitro test with mammalian cells and an in vivo test for chromosomal damage using rodent hematopoietic cells (7). The work in this thesis is focused on this ‘second test’, the in vitro test with mammalian cells, at AstraZeneca the mammalian in vitro test utilised routinely is the in vitro micronucleus (IVM) assay.

1.1.2 Compounds with Genotoxicity

The IVM assay can be used to identify two types of compounds which induce different types of DNA damages, clastogens and aneugens. A clastogen is a genotoxicant which can cause chromosomes disruption or breakages, as well as complex chromosomes changes. For example, 4-Nitroquinoline 1-oxide(4-NQO) is a carcinogenic and clastogenic chemical which induces DNA lesions including single-strand breaks, pyrimidine-dimer formation and oxidized bases (8). In order to survive from massive mutagenic sources coming from endogenous and exogenous environment, healthy cells are able to repair many types of DNA damages including those induced by clastogens through different mechanism, for example removing harmful chemicals from DNA directly (9). An Aneugen causes a daughter cell to have an abnormal number of chromosomes and may lead birth defects, pregnancy wastage, and cancer (10). Thus, aneugens can be detected by an abnormal number of chromosomes or by the detection of the micronuclei (micronuclei are small nuclei formed during cell division when a chromosome or a fraction of a chromosome failed to incorporate into one of the daughter nuclei) containing whole chromosomes.

The IVM assay is the most common method for distinguishing aneugens from clastogens by assessing whether there is a kinetochore within a micronucleus by the method of anti-kinetochore antibody staining, however kinetochores may be damaged by the process of aneugenic activities which may lead errors to the detection (11). γH2AX (the phosphorylated histone H2AX) foci is a biomarker used for detecting the double-strand DNA damages induced by mechanisms such as ionizing radiation and clastogenic activities, because it related to the localization and the repair of DNA damage, the mechanism of action (MOA) can be visualised by immunofluorescence (12).

1.2 The development of IVM

The IVM assay is a set of tests which use micronucleus as a biomarker to detect genotoxicants of clastogens and aneugens (13). It is well accepted that the chromosome fragments (for example, stem from the mechanism of clastogenicity) or a whole chromosome which doesn't attach to the spindle (for example, due to the reason of aneugenicity) formed in the cell cycle of mitotic phase would be enclosed by a nuclear membrane then become micronuclei eventually (14). Due to the different mechanisms of generation of micronuclei, it is possible to assess the type of genotoxicity by observing the shape, size, number, the presence of kinetochores of micronuclei.

The version of IVM that is required for regulatory submission at this time is a manual version of the IVM assay, which is time consuming, requires a large amount of testing compounds, and is relatively hard to add other biomarkers which limits the features size of dataset and mostly binary output(positive, negative) which is fairly insufficient for further analysis (15) .

1.2.1 Multi Endpoint Genotoxicity Screen

After years of developing, leading pharmaceutical companies such as AstraZeneca have begun to adopt Multi endpoint genotoxicity screens (MEGA Screen) with Acoustic Droplet Ejection technology which lowers the amount of compound possible to be tested in IVM and significantly improves the efficiency of the experiments. Compared to manual IVM, MEGA Screen uses automatic robots to run the experiments and acquires high content confocal images using a CellVoyager™ C7000 (Yokogawa, Japan) followed by image analysis performed using Columbus™ software (PerkinElmer, UK). The data generated from Columbus is assessed by experts who have set thresholds and hyper-parameters to assess the genotoxic potential of a compound. A 384-well plate can be utilised with high accuracy in a short period of time in the experiments, and multiple biomarkers, for example, γH2AX can be test simultaneously. Besides that, the cost of MEGA Screen is relatively low because the required space and reagents of the experiments is reduced (15).

1.2.2 MEGA Screen Data Generation

Manifold data generated from MEGA Screen is the most significant advantage of this method, the large amount of data is the fundamental of further machine learning jobs. All the data comes from the analysis of high content images generated from MEGA Screen, the quality of the generated data is not only dependent on the rigour of biological part of the experiments, but also on the thresholds and hyper-parameters set by users during initial image analysis in the Columbus software whilst generating data from images of testing cells. In this project, I cooperated with experts of AstraZeneca who have years of experience of calibrating and adjusting these parameters to ensure the high-quality data.

The data coming from the MEGA Screen is firstly classified by cell cycles as sub G1, G1, S, G2, more than 4N, each phase of cell cycle have very similar features to be measured. The profile of nuclei and micronuclei, for example shape and number, is calculated as highly-precise continuous numeric features. Biomarkers and stains are also used to provide more comprehensive data for further analysis. For example, a kinetochore antibody is used to detect a kinetochore protein within the micronuclei, which is useful for distinguishing aneugens from clastogens. And two types of γH2AX staining are used in the experiments to describe the MOA with γH2AX foci providing early evidence of DNA double strand breaks which can be also detected by spot analysis in MEGA Screen and by pan-nuclear γH2AX staining being associated with early apoptosis or cell death. Another example of biomarker used in this project is measuring cellular viability by Hoechst staining via quantifying the intensity of DNA (15). The stains used in this project are shown in Fig. 1.

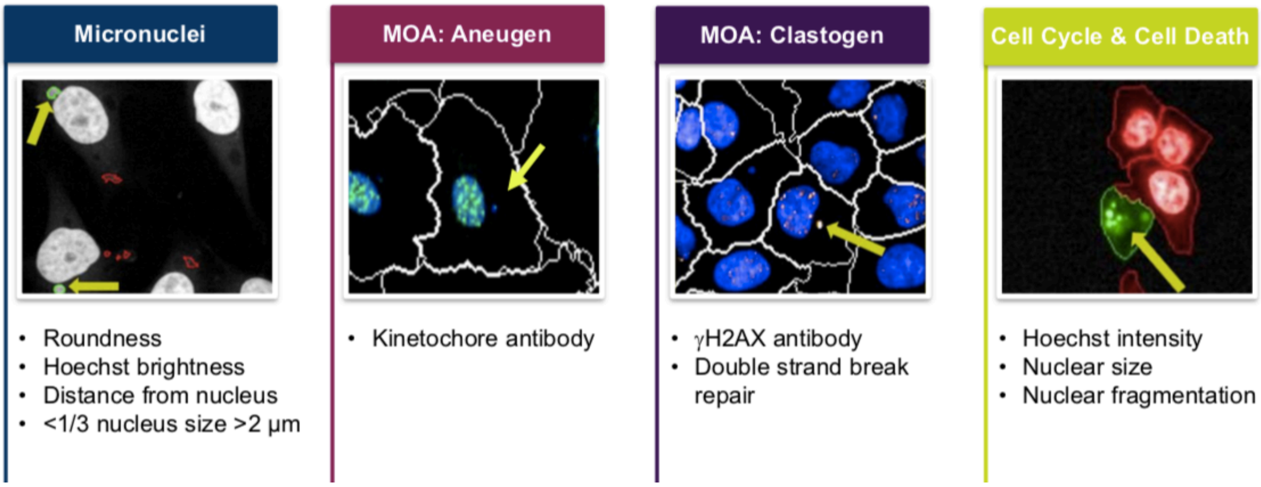


Figure 1: There are 4 channels used in MEGA Screen and in these channels certain stains were used for detecting micronuclei, kinetochore, DNA breaks and cell cycles respectively. The features generated in the experiments are all related to one of these channels. (29).

1.3 Existing Genotoxicity Problems which data science may contribute to

The MEGA Screen generates a huge amount of data which could benefit from computational analysis and machine learning jobs as it becomes more challenging for researchers to recognise and understand the dataset as its size increases. Also, it is very difficult and time-consuming for researchers to deal with the MEGA data even with data analysis tools such as SAS and Excel.

Random errors and systematic errors often exist in biological experiments. For example, there are random errors between the duplications of a single compound at the same concentration level in the same plate because, theoretically, identity of duplications is impossible to be achieved in the context of a cell-based assay system. Systematic errors are also difficult to avoid, for example the temperature and the batches of reagents both possibly impact assays reproducibility. In this project, the methods to detect these errors have been developed to help the following data analysis and classification models.

With the development of High content imaging analysis, the number of features has increased significantly and some of the features are highly correlated, which means that it is more difficult for researchers to choose the ideal set of features during their research. For example, there are 4 features selected by genetic toxicology staff to distinguish the categories and determine MOA of compounds. They are “Nuclei – Number of gH2AX foci – per Nucleus – Mean per Well”, “Nuclei – Number of Micronuclei – per Cell – Mean per

Well", "Nuclei – Nuclear Fragmentation – Mean per Well", "Nuclei – Number of Micronuclei with kinetochore – per Cell – Mean per Well". This set of features is denoted as G-features in this thesis, and a question is that if machine learning algorithms could suggest a better set of features for doing the same job.

It is difficult to build a model which is able to distinguish aneugens, clastogens and non-genotoxicants. The major reason is that the gap between them are usually too narrow to tell or even couldn't be defined based on the limited features of experimental data. There are compounds which act like both clastogens and aneugens, and there are also compounds standing in the borderline between genotoxicants and non-genotoxicants. It is also the reason why model compounds are much easier to be distinguished than newly developed compounds which are likely to be mixed-positive or borderline compounds. Besides that, there are many DNA analogues which can incorporate into DNA and make the images generated by MEGA Screen look similar to the ones of genotoxicants, so it is a particular difficulty in MEGA Screen which may not affect the results of the other assays. In addition, the confidential structure of AZ compounds is another limitation for building a good classifier.

1.4 Aim

To help researchers with increasing the amount of knowledge generated from MEGA Screen instantly by using data science approaches.

1. Clean the original dataset generated from MEGA Screen and visualise the numeric dataset in a more intuitive and understandable way.
2. Estimate the relationship between the features and the categories of compounds, for example clastogen, and recommend feature selections to researchers for detecting a certain category of compounds.
3. Classify compounds and estimate the genotoxicity risks as well as predict the types of genotoxicity.

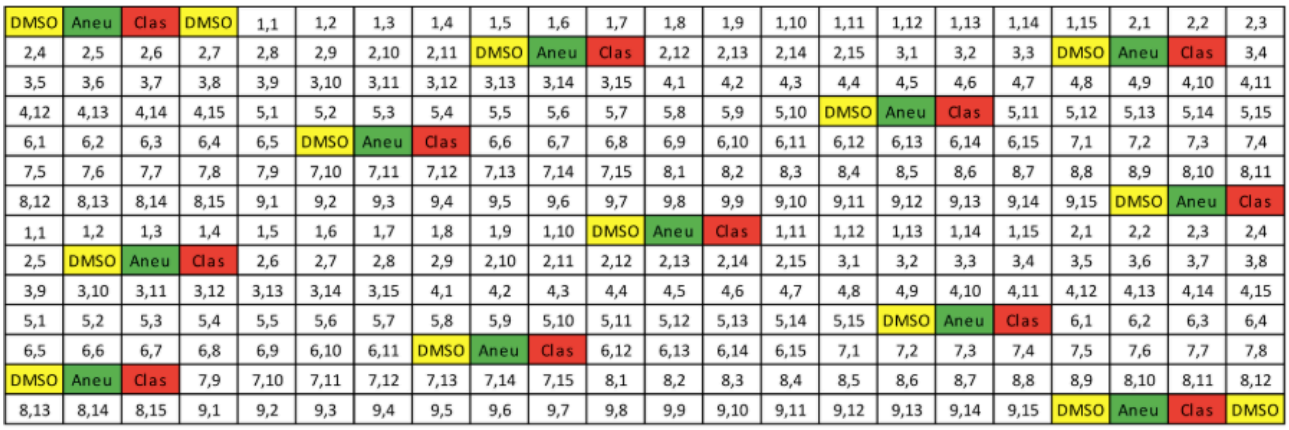
2. Methods

2.1.1 Experiments

Table 1: The concentration of each DL for tested compounds.

Concentration(mM)	DL	Concentration(mM)	DL	Concentration(mM)	DL
1E-07	1	3.17E-07	2	1E-06	3
3.33E-06	4	1E-05	5	3.17E-05	6
1E-04	7	3.33E-04	8	1E-03	9
3.17E-03	10	0.01	11	0.033	12
0.1	13	0.317	14	1	15

The experiments were designed and implemented by the experts in AstraZeneca based on the guideline (39) and their experience of toxicology and IVM assay. The cell line and the vehicle used in the experiments was p53 wildtype A549 human cell line and 2% DMSO v/v respectively. In detail, 750 cells per well of 384-well CellCarrier-384 Ultra microplates (PerkinElmer) were plated 24 hours before treatment. An Echo 555 Acoustic Dispenser (Labcyte) was used to generate 15 dose levels (DL) from 1nM to 1mM for each tested compound (The table of the concentration of each DL is shown in Table 1). In order to test the consistency, compounds were tested twice at each DL in different wells of a plate, the details of plate map are shown in Fig. 2. Compounds were incubated with cells for 24 hours followed by a wash step, then another 24 hours of recovery, allowing for the expression of whole chromosome loss.



DMSO	Aneu	Clas	DMSO	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	1,10	1,11	1,12	1,13	1,14	1,15	2,1	2,2	2,3
2,4	2,5	2,6	2,7	2,8	2,9	2,10	2,11	DMSO	Aneu	Clas	2,12	2,13	2,14	2,15	3,1	3,2	3,3	DMSO	Aneu	Clas	3,4
3,5	3,6	3,7	3,8	3,9	3,10	3,11	3,12	3,13	3,14	3,15	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	4,10	4,11
4,12	4,13	4,14	4,15	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	5,10	DMSO	Aneu	Clas	5,11	5,12	5,13	5,14	5,15
6,1	6,2	6,3	6,4	6,5	DMSO	Aneu	Clas	6,6	6,7	6,8	6,9	6,10	6,11	6,12	6,13	6,14	6,15	7,1	7,2	7,3	7,4
7,5	7,6	7,7	7,8	7,9	7,10	7,11	7,12	7,13	7,14	7,15	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	8,10	8,11
8,12	8,13	8,14	8,15	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	9,10	9,11	9,12	9,13	9,14	9,15	DMSO	Aneu	Clas
1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	1,10	DMSO	Aneu	Clas	1,11	1,12	1,13	1,14	1,15	2,1	2,2	2,3	2,4
2,5	DMSO	Aneu	Clas	2,6	2,7	2,8	2,9	2,10	2,11	2,12	2,13	2,14	2,15	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8
3,9	3,10	3,11	3,12	3,13	3,14	3,15	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	4,10	4,11	4,12	4,13	4,14	4,15
5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	5,10	5,11	5,12	5,13	5,14	5,15	DMSO	Aneu	Clas	6,1	6,2	6,3	6,4
6,5	6,6	6,7	6,8	6,9	6,10	6,11	DMSO	Aneu	Clas	6,12	6,13	6,14	6,15	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8
DMSO	Aneu	Clas	7,9	7,10	7,11	7,12	7,13	7,14	7,15	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	8,10	8,11	8,12
8,13	8,14	8,15	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	9,10	9,11	9,12	9,13	9,14	9,15	DMSO	Aneu	Clas	DMSO

Figure 2: The sample of plate map. DMSO is used as negative control in yellow wells. A model aneugen and a model clastogen are both used as positive controls in green and red wells separately. The other wells are used by at most 9 tested compounds with 15 DLs.

Cells were imaged on a Cell Voyager 7000 (Yokogawa Inc.) and 8 images were taken from the same location in each well. Image analysis was completed by modified Columbus™ (PerkinElmer) image analysis algorithms, the data generated from each of the 8 images of a well was combined together by averaging. Thus, each sample in the output dataset represented the average response of a single well.

Besides tested compounds, there were also positive controls (which are model aneugen and clastogen shown as “Aneu” and “Clas” in Fig. 2) and negative controls (which is DMSO shown as “DMSO” in Fig. 2) included in each plate not only for calibrating the algorithms for image analysis but also for setting the threshold of half maximal effective concentration for cell number (EC50) in each plate. EC50 is the concentration of a compound that induces a reduction to half the number of living cells compared to the in-plate negative control. At this concentration a compound can be represented effectively, because it is neither too low to affect cells nor too high to be distinguished from other compounds due to the massive death of cells. (15)

2.1.2 Dataset Description

There are two datasets generated from the experiments, one contains only well-defined compounds with known responses and smaller sample size, the other one contains newly developed compounds which were classified by experts in AstraZeneca at high accuracy while it also includes the same well-defined compounds used for checking the consistency of experiments. The smaller dataset with well-defined compounds is denoted as VD (the validation dataset), and the larger dataset with compounds newly developed by AstraZeneca is denoted as AD (the AstraZeneca dataset). The details of VD and AD can be found in Sec. 3.1.2.

VD is more suitable for understanding and interpreting features because all of the compounds in it are well-defined, thus it is also a good validation dataset for proving an effective classification model can be built to distinguish genotoxicants from negative compounds. Previous work showed that genotoxicants can be distinguished by machine learning (ML) algorithms trained using VD (15). In VD, experts labeled compounds into 4 categories, “aneugen”, “clastogen”, “negative” and “DNA analogue”. “Negative” and “DNA analogue” compounds may cause cytotoxicity cell death but not by a genotoxic mechanism, so they are considered as non-genotoxicants in this study. The reason that “DNA-analogue” was listed separately from negatives was because they are analogues of the DNA bases and they can incorporate into DNA by their MOA, and as a result, the images of compounds of this category generated from MEGA Screen may look similar to the ones of genotoxicants. VD was mainly used for feature description in this study, because it only contains well-defined compounds, which makes the results more reliable.

AD was mainly used for training and testing classification models not only because of the larger sample size, but more importantly because the more varied compounds compose a much more complex chemical space of both genotoxicants and negative compounds therefore the classification models trained using AD are more adaptive and effective from the industrial perspective and can represent the real-world challenge. In AD, there are 4 categories, “aneugen”, “clastogen”, “negative” and “positive mixed” which means a compound shows genotoxic features of both aneugen and clastogen. In this study, the main aim is to distinguish genotoxicants from negative compounds, therefore 3-category (aneugen, clastogen and negative) classifiers were built for this purpose and it was considered as a good case if a compound labelled “positive mixed” was predicted as either clastogen or aneugen. AD was split into a training and a testing dataset by Monte Carlo Cross-Validation (MCCV), in this case samples are randomly picked from AD, and 80% of them used for training while 20% of them used for testing.

2.2 Dataset Cleaning

Each dose was performed in duplicate in each plate, thus those duplications should be similar and if these duplications are far apart from each other, they are considered as “dirty” data and needed to be cleaned afterwards. The measurement of differences among duplications is called uncertainty of duplications in this project and denoted as UD. To calculate the UD of the **i**th element **e**$_i$ in the set of duplications which contain **n** elements and have an average as **a**, the equation can be written as:

$$UD_i = |\frac{e_i - a}{a(n-1)}| \qquad \text{(Equation 1)}$$

There are two main advantages of this statistic, the first one is that the range of UD is scaled to (0,1) and 0 represents the minimal difference while 1 represents the maximal difference. The other advantage is that the statistic is based on the distance between elements and their mean. For example, given 3 elements in a duplication set as 0.2, 0.5, 0.8, the values of UD for these elements are 0.3, 0, 0.3, the elements 0.2 and 0.8 have the same value of UD because they have the same distance from the mean. In this study a threshold of 0.8 for UD is set for distinguishing dirty data from valid data as default. The original values with a value of UD higher than the threshold would be assigned as missing values, in other words, they were denoted as dirty data and would be cleaned in the following step.

Plate-related patterns were found in the datasets too, which were due to experimental procedures, such as time or temperature change between runs, and needed to be excluded from the datasets. To achieve this, a method of data transformation was applied to original dataset by using DMSO which is the in-plate negative control in each plate as the baseline. Before applying the transformation to the datasets, 0 and missing values were replaced by 10^{-7} and 10^{-8} separately in order to make $mean_{(DMSO,plate_i,feature_j)}$ definitely larger than 0 as well as distinguish missing values from 0. For $plate_i$ (the i^{th} plate), given a value x of $feature_j$ (the j^{th} feature), the updated value of x is calculated as shown in Eq. 2:

$$updated\_x = \frac{x_{(plate_i,feature_j)}}{mean_{(DMSO,plate_i,feature_j)}} \qquad \text{(Equation 2)}$$

2.3 Barcode-like Images

Data visualisation is a good way of presenting numeric data, but it is very difficult to visualize a dataset with too many features, such as VD and AD, let alone the 15 DLs of each compound. Thus, general data visualisation methods may not, at least efficiently, work in this case, one alternative way is that use every value in a matrix as a pixel in images, so that we can transform a numeric matrix into an image. And one image can be plotted for each compound in a plate, in addition, each feature can represent one column in the image while each DL can represent one row in the image.

Readable images are usually comprised with certain patterns which can be identified by people. Pixels must be combined together by certain rules to show a pattern, in other words, pixels won’t show any pattern unless they are regularly rather than randomly arranged. Thus, in order to plot images of compounds which can show visible patterns hiding in the massive numeric dataset, the features and the DLs for each compound must be arranged by the same regular rule. In this project, one of the simplest rules is used, it is “the adjacent 2 rows or columns must be highly correlated to each other”. Thus, both the features and the DLs must be reordered based on the rule before plotting.

The features were reordered by hierarchical clustering which used the matrix of Pearson's correlation coefficients among features as the input data and used the Euclidean Distance as the metric. After the clustering, a list of features reordered by the inter-correlation was generated, in this list a random selected feature was adjacent to its most correlated feature due to the property of hierarchical clustering. Thus, this list of features generally obeyed the rule which was set previously.

There are 30 samples of a single compound within a plate, which is comprised by 2 duplications of each of the 15 DLs. In order to show the consistency in images, the duplications were left separately rather than combined by averaging them. These samples were divided into 2 groups, each group contained 15 samples with DLs from 1 (the lowest 1nM) to 15 (the highest 1mM). Then the first group was reordered by the DLs from 1 to 15 in ascending order and the second group was reordered by DLs from 15 to 1 in descending order. Finally, the two reordered groups were combined together, so that the DLs started from 1 to 15 and then ended from 15 to 1. The difference of DLs between 2 adjacent rows was not larger than 1, in other words, every 2 adjacent rows are highly similar and correlated.

Considering 30 rows of pixels are too thin for people to observe, the number of rows is required to be increased by applying an approach of interpolation. For every 2 adjacent original row vectors, generate one new row vector by averaging them then insert the new vector between them. After this operation, the number of rows increased to 59. And the operation was executed twice in total for getting enough rows, as a result, the number of rows became 117 which was nearly 3 times larger than the original number of rows.

The last step before plotting was the data transformation of the exclusion of plate-related patterns introduced in Sec. 2.2, then the values were scaled based on each feature. **max_i** and **min_i** are the maximum and minimum of the feature i, given a value **$x_{i,j}$** of feature i and sample j, the new value **$x_{i,j}'$** can be calculated by Eq. 3 as below, and the squared root was applied to the result in order to reduce the influence of the positive outliers.

$$x'_{i,j} = \sqrt{\frac{x_{i,j} - min_i}{max_i - min_i}} \qquad \text{(Equation 3)}$$

After the features and the DLs reordering, the row interpolation and the data transformation, each compound in a plate generated one 117 by 377 (377 features in VD) or 455 (455 features in AD) matrix. Each element in the matrix represent a pixel in the corresponding images, and the larger the value is, the brighter the pixel is. Images generated in this way look like barcodes, and interesting patterns can be found in the images of compounds which lead cell death or other unique responses in the experiments.

2.4.1 Relations between Categories and Features

There are more than 300 features in VD, the differences between features can be interpreted based on the trees generated from the random forest (RF) model. The interpretation can be used to distinguish and select features for certain purposes for which the RF model was built. For example, in this project, the relations between features and categories are worth exploring for the purpose of finding a better set of features instead of G-features. Thus, a RF model for distinguishing clastogens, aneugens and negatives was built using VD with plate-related patterns excluded, then the trees generated from the model were used for interpreting the relationship between features and the three categories.

Affinity is the concept defined by myself in this project and used to describe the relationship between features and categories, if the compounds of a certain category generally have relatively large values of a feature at EC50, we can say this feature has a positive affinity with the certain category and the larger the values are, the stronger the affinity between the feature and the category is.

The affinity between a feature and 3 categories was calculated from all of the binary trees which contained the feature and generated from the RF model. The calculation of the affinity between the feature of “Micronuclei – Number of Objects” (MNO) and 3 categories was taken as an example. And there was one of the binary trees containing this feature shown as Fig. 3.

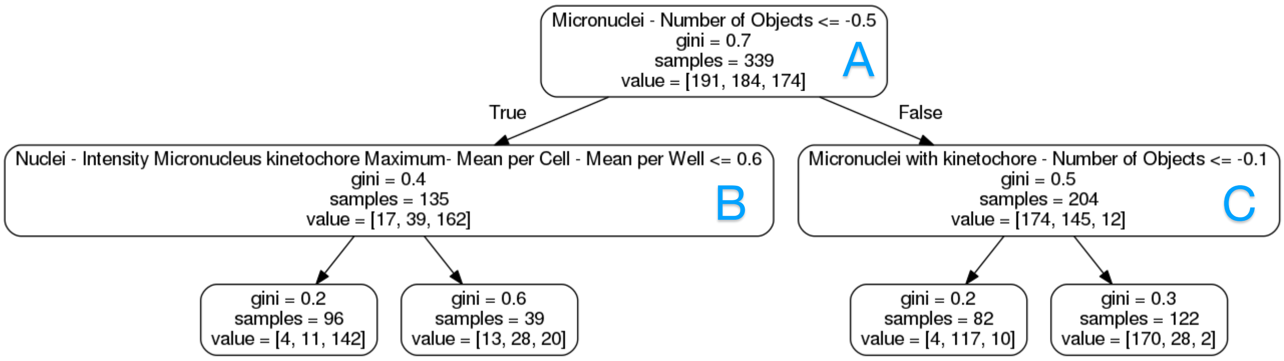


Figure 3: One decision tree generated from RF model. The name of features used for dividing nodes are shown in the top of each parent node. The value vector in each node shows the sample size in the sequence of "aneugen", "clastogen" and "negative".

MNO is in the node of A which has two children nodes of B and C, if a sample have a smaller value (smaller than -0.5 in this particular case) of MNO it will go into the right node C, otherwise the left node B. The “value” in each node is the vector which represent the sample size of each category in the order from “aneugen”, “clastogen” to “negative”. Thus, the differences of sample sizes between node C and node B can be calculated by using the “value” of node C subtract the “value” of node B, which is [174,145,12] – [17, 39, 162]. The result is [157, 106, -150], which illustrates that if a sample has a relatively large value of MNO (larger than -0.5), it is likely to be “aneugen”, “clastogen” or “negative” in the order of possibilities from high to low.

In this project, 100 binary trees generated by RF model were used to calculate affinity values. For a certain feature like MNO, the results calculated from all of the binary trees which contain the feature were summed together, and the vector generated from the summation, for example, was [1570, 1060, -1500]. Then it was transformed into a vector comprised by percentage elements as [38%, 26%, -36%] by using the original values divided by the summation of the absolute original values in the vector. As a result, 38%, 26% and -36% are the values of affinity between MNO and the categories of “aneugen”, “clastogen”, and “negative” separately.

The confidence which used for measuring the reliability of affinity was calculated by summing the absolute values in the integer result vector, and in this example the value of confidence is 4130. There are two requirements for a feature to get a high confidence. The first one is that the feature is required to be used to distinguish a large number of samples by the RF model. The second one is that its result vectors of affinity generated from different trees have similar directions, so that when these vectors are summed together the values of each element in these vectors are not offset against each other, which leads to a relatively large absolute summation which is used for calculating confidence.

2.4.2 Features Clustering

There are several hundred of features in VD and AD, so it is difficult for researchers to understand the interrelationship between these features. In this case, feature clustering is a good way to simplify the process of feature understanding. However, there are also difficulties of interpreting the result of clustering algorithms, for example K-means, due to two reasons. The first one is that clustering results are very hard to evaluate, as there is no common criterion to assess the clustering quality. It is also difficult to interpret the differences among groups clustered by unsupervised ML algorithm.

In this project, K-means algorithm was applied to scaled VD after plate-related pattern exclusion (the scaling method is shown in Eq. 4) with K equals to 8 which was selected by experts by comparing the clustering results of different values of K from 5 to 12 based on the rationality of taxonomy in toxicology.

$$updated\_x = \frac{x_{(plate_i,feature_j)} - min_{(plate_i,feature_j)}}{max_{(plate_i,feature_j)} - min_{(plate_i,feature_j)}} \qquad \text{Equation 4}$$

Then Pearson's correlation coefficients were calculated within each of 8 clusters, and one heatmap was plotted based on the correlation matrix of features in each cluster. In addition, the order of features in correlation matrix was determined by features hierarchical clustering (the same with the approach of reordering features in Sec. 2.3), so that similar features were close to each other. To interpret the taxonomy of the clustering results, affinity between categories and features generated from RF models was used to define the property of a certain cluster by averaging the values of affinity of all the features in the cluster. Moreover, affinity between categories and features can also be used to describe and select features for certain purpose. This process of feature description can provide researchers with a comprehensive perspective of numerous features generated in MEGA Screen, so that they can classify compounds better by using a specific set of features suggested by ML algorithm.

2.5 Data Pre-processing

In order to build a classification model, the process of data pre-processing is one of the most important parts which must not be overlooked. In this section, the approach of calculating EC50, the approach of balancing data and finally the reason and the solution of flattening data will be introduced. By pre-processing data, the most representative DL will be selected, and more samples and features will be generated to build a better classification model.

2.5.1 EC50 Selecting

Considering there are 15 DLs for each compound, and different compounds show observable toxicity at different DLs, it is crucial to select the most representative DL for each compounds, so that these compounds can be classified based on their most distinct responses. EC50 is a common effective concentration used in toxicology (20,21). There are more than 300 features generated in the project, each feature can generate one specific EC50 for each compound. In this project, the number of living cells was used for generating the EC50 values for all of the compounds, because it is the most general measurement for testing toxicity. The EC50s of cell number calculated in this way is denoted as EC50.

The feature "Whole cells - Number of Objects" is used for measuring the living cells for each sample, and the total number of living cells is calculated by averaging the values of all the negative controls from the corresponding plate. For each compound in a plate, compare its values of "Whole cells - Number of Objects" at the DLs from low to high with the half of the mean calculated from the in-plate negative controls. The lowest concentration at which the sample has less than half of the total living cells in negative controls is selected as the EC50 of this compound in the plate. If a compound has multiple EC50s in different plates, the average of its EC50s is used. The EC50s generated this way was manually checked for accuracy, and no abnormal case was found. If a compound, for example a negative compound, does not have the EC50 because it does not kill cells at any DL in the experiments, then the 2nd highest DL will be selected as its "EC50" for classification. The top 5 DLs were tested as the alternatives of EC50 when a compound does not have the EC50, as a result, the 2nd highest DL is the most effective alternative, using which the most precise classification model was built.

2.5.2 Data Balancing

Unbalanced datasets usually lead to the classification models which have poor recall of minority classes. In this project, the sample size of negative compounds is 2 - 3 times larger than the sizes of clastogens and aneugens. In order to improve the recall of clastogens and aneugens, balancing the training data is indispensable. Synthetic minority oversampling technique (SMOTE) is one of the most common oversampling methods and has proved effective for decision tree (DT) models and random forest (RF) models (22,23).

The main idea of SMOTE is that given a randomly selected minority sample and its k neighbors (k is a parameter can be set by researchers), then randomly select one point on the line between the selected sample and one of its neighbors which is also randomly selected, the vector of the selected point is the synthetic sample. However, this algorithm is not suitable for this project, because features in the original dataset are highly repetitive and inter-correlated, for example in this project the features measuring the same response in different stages of cell cycles may be highly correlated to each other, thus the distances calculated from these features are probably highly depended on a small set of inter-correlated features, which cannot represent the reasonable distance of samples in classification spaces.

Considering the difficulty of getting the reasonable distance between samples, I simplified the case by using 2 randomly selected samples of the same category to generate new samples. In this case, additional inaccurate information may be generated comparing to SMOTE. In order to expand the sample spaces of 2 minority classes, the mean of the 2 randomly selected samples was used instead of a random combination which is used in SMOTE. Although this operation may amplify the inaccurate information for classification models, it can improve the recall of genotoxicants as much as possible in a relatively appropriate way, so that experts would be able to de-risk potential genotoxicants as a genotoxicant with a predicted negative result may progress through the drug development pipeline, causing deleterious effects, for example increased cost of further testing. By applying this method, the sample size of 3 categories became equal in training dataset.

2.5.3 Data Flattening

The volume of data is always a significant factor in data science, and it is generally true that a better model can be trained just by adding more training data. There are two ways of increasing the data size, the first one is simply generating more samples, which is limited by the MEGA Screen experiments in this project. The other one is using more useful features. Considering only samples at EC50 were used to build training dataset, it is possible to use the features from the other DLs in order to increase the number of features.

According to this idea, I extracted the features from the other DLs, for example the top 5 levels of concentrations, and then combined them into the samples at EC50, so the feature size increased several times larger depending on how many other DLs were used. In order to find the best combination of features sourced from the other DLs, different combinations were tested (adding DLs into the combination beginning with a moderate DL until the precision of classification models does not increase) by building different classification models, and the combination which could train the best model won the test. It turned out that [EC50, EC50-1, EC50-2, the 10th DL, the 11th DL, the 12th DL, the 13th DL, the 14th DL] (the table of concentration of each DL is shown in Table 1) is the best combination based on the tests.

Therefore, I used the features sourced from 9 DLs rather than only EC50, which means that the flattened dataset has 9 times as many features as the EC50 selected dataset. More data was expected to train a more accurate classification model, and it was tested how much it could improve the precision by comparing with EC50 selected dataset in this project.

2.6 Classification algorithms

There are three measurements used to evaluate the performance of a classification model in this study, they are accuracy, precision and recall. Accuracy is used for measuring the overall reliability of the predictions and it is the ratio of correctly predicted observation to the total observations. Precision is the accuracy for one of the predicted categories and it is the ratio of correctly predicted observation in a certain category to the total observation predicted as the category. Recall is used for measuring the proportion of correctly predicted observation in a certain category and it is the ratio of correctly predicted observation in a certain category to the total observation in the category.

Building an effective classification model for genotoxicants is the major purpose of this project, 4 different algorithms were tested to provide more comprehensive classification models. The first one is DT (the tool used for building DT models is the scikit-learn package of python), which is one of the most interpretable models and can be used to get more understandable classification models, the maximal depth of DT models is set as 2 in order to make the results easier to read. The second one is RF (the tool used for building RF models is the scikit-learn package of python), which is one of the most commonly used models in biomedical research. The advantages of RF include that it is unlikely to overfitting, it easily handles massive features, making this model quite suitable for this project. In order to improve accuracy and avoid overfitting 100 trees with a maximal depth of 6 were used to build RF models in this project (31,32,33).

The third model is multi-label XGBoost (the tool used for building XGBoost models is the xgboost package of python), which is one of the most powerful classification models except deep learning models (24). In order to avoid overfitting, a common issue with XGBoost, 50 trees with maximal depth of 2 were used to build XGBoost model, and the learning rate was set as 0.1 which is also the default value. The fourth model is binary-label XGBoost, which was combined with Venn-Abers Predictors (VAP) to provide researchers with the confidence of predictions. VAP is much easier to be implemented on top of a binary-label classification model compared to multi-label classification model. The parameters of binary-label XGBoost model were set the same as the multi-label one mentioned above (25,34).

VAP is an algorithm which can provide valid probabilities of predictions. “Validity is the property of a predictor that outputs probability distributions that perform well against statistical tests based on subsequent observation of labels” (26). According to this property, the probabilities output by VAP are more reliable than using the scores output by classification models as probabilities directly. VAP can generate an upper and a lower bound (denoted as $\mathbf{p_u}$, $\mathbf{p_l}$ respectively) of the probability of a prediction, and a single-valued probability can be calculated by applying the Eq. 5. “**p**” calculated by this formula is the value which is used for measuring the confidence of predictions in this project (25).

$$p = \frac{p_u}{1 - p_l + p_u} \qquad \text{(Equation 5)}$$

Finally, in order to produce a better model, multi-label and binary-label XGBoost models were combined together with VAP for generating the confidence of prediction. In this combination, there were 3 binary-label classification models with VAP built in, they were used for distinguishing aneugens, clastogens and negatives separately. For every testing sample, the three models output the confidence that the sample is belonged to “aneugen”, “clastogen” and “negative” separately, for example, 0.01, 0.3, 0.8. In this case, the compound is more likely to be a negative, so its label is assigned as “negative”. In other cases, if the 3 values of confidence are all predicted lower than 0.1, which means the binary-label models estimated that the sample is not probably belonged to any of the 3 categories, then the predicted label calculated by multi-label XGBoost model will be used.

2.7 Dimension Reduction algorithms

Principal component analysis (PCA) is one of the dimension reduction algorithms commonly used in biomedical research. PCA uses orthogonal components, which can be much fewer than the feature size, to explain a certain proportion of variance of the original dataset, in other words it uses fewer components to represent a relatively larger number of features in order to reduce the dimension. In this study, the first and the second components of PCA was used for visualising the hidden patterns in numeric datasets, because components can be used to describe unique characters of samples (35).

Considering the feature sizes of both datasets are more than 300, it is impossible to visualise the chemical space in full dimension. In this study, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used for dimension reduction in order to visualise the chemical space in a 2-D plot. t-SNE, unlike PCA, can keep the original Euclidean-distance relation between samples as much as possible when mapping the high dimension feature space into a low dimension feature space, so it is the best way of describing and visualising the

Euclidean-distance relation between original samples. Perplexity is one of the key hyperparameters of tSNE, which determine how many neighbours and the distances between them will be considered when doing mapping. There is no common criterion for selecting the best perplexity, in this study the perplexity was chosen from many trials based on manual observation (31, 36).

3. Results

3.1 Data Exploration

The datasets used in this project are different from mature public datasets which are reliably labelled and commonly used in academic research. Besides inevitable experimental errors, many other factors, e.g. different batches of reagents, can affect the quality of experimental results in this project. Most importantly, the uncertainty of labels in AD makes our project even more difficult to output a rigorous and scientific dissertation, on the other hand provides an exciting opportunity to discover more effective and practical data processing work flow to allow the testing of increasingly large numbers of compounds. For these reasons, a thorough data exploration process is required prior to model building and feature description, by doing this the following work would be more reliable and effective.

There are two main parts of data exploration in this project, the first one is patterns detection which aims at finding regular patterns within the datasets and remove the ones caused by experimental error. The other one is data cleaning which is one of the most crucial procedure of data science.

3.1.1 Patterns Detection

The MEGA Screen assays have been running for about 3 years, thus it is possible that the datasets contain some patterns related to the passage of time, for example the changing of laboratory conditions and the improvement of experimental skills can affect the quality of output. In this case, the plate ID is a good identification for time-related patterns detection, because the plate ID increases as time flows though the IDs which are denoted as discontinuous integers. PCA was used for detecting these patterns, and the figure of component 1 (27% variance was explained by it) and 2 (17% variance was explained by it) of PCA using scaled (using Eq. 4) VD was shown as Fig. 4:

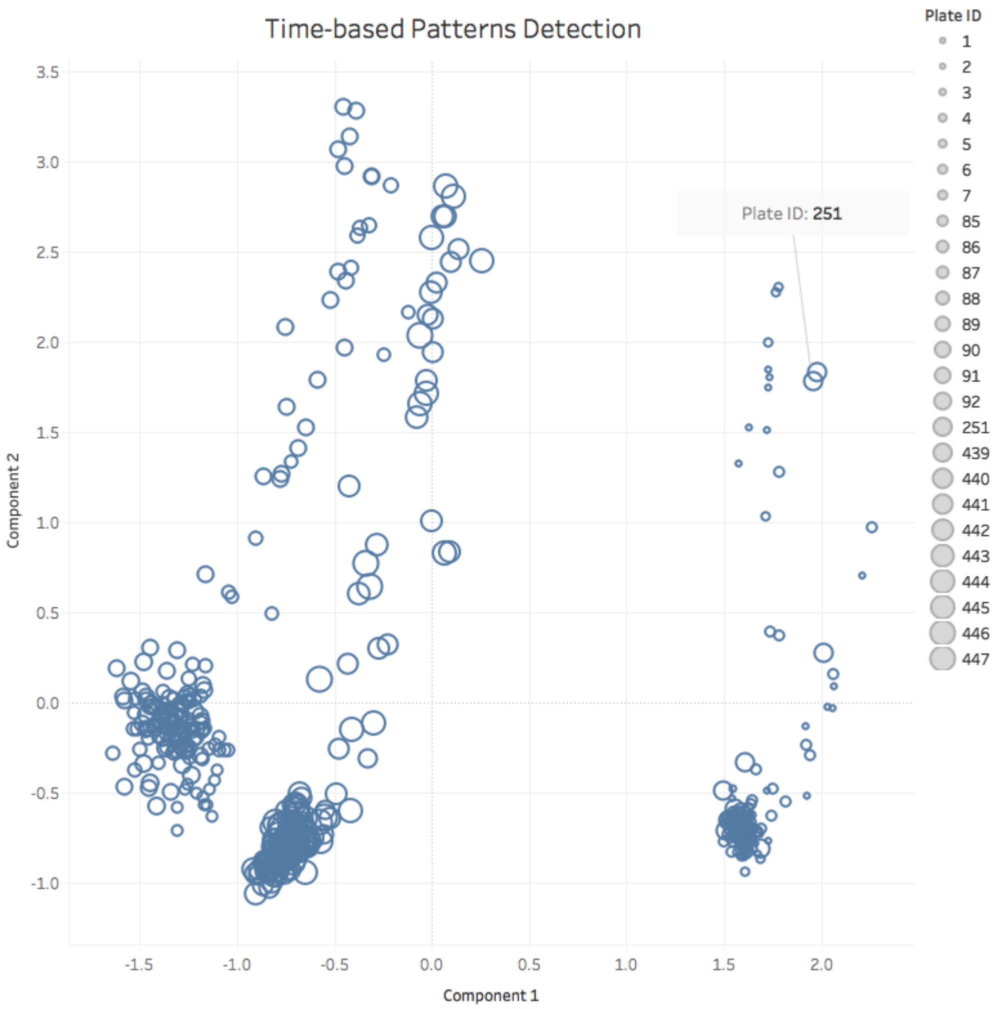


Figure 4: The plot of component 1,2 of PCA used for showing the plate-related patterns. The size of circle represents the plate ID of each sample (compound). There are 3 clusters shown in the plot, each of them is comprised by certain continuous plate ID.

Each circle in the figure was a sample in a certain plate, and these samples formed three main clusters in Fig. 4 which could be distinguished by 3 sets of plates as plates 1 to 7 with 251, plates 85 to 92 and plates 439 to 447. The 3 sets of plates represented 3 periods of time except plate 251 which proved to be label mistakes by experts of AstraZeneca and should be labelled as 8. It generally illustrated the fact that there were specific patterns related to time or sets of plates in the dataset. These patterns were originated from experimental variation that needed to be excluded from the dataset.

The Eq. 2 was applied to the whole dataset, and then the plate-related patterns were excluded from the dataset. After this process, PCA transformation was applied to the scaled updated dataset to check if these patterns had been removed. As a result shown in Fig. 5, the samples in different plates mixed together and there are two new clusters formed by different compounds. This is reasonable because different compounds are expected to act differently in the experiments. However, two reagent batches of 5-Fluorouracil were belonged to different clusters as Fig. 5 shows (just selected the compounds with multiple batches in this plot)

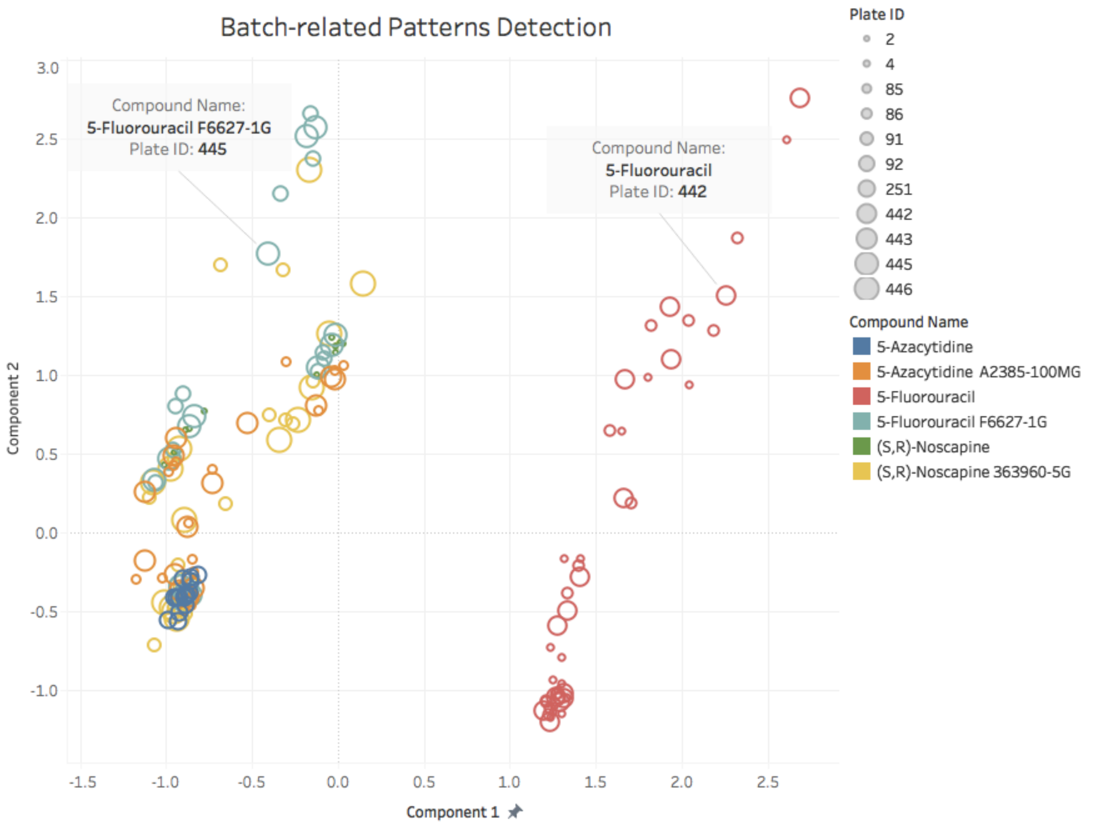


Figure 5: The plot of component 1 and 2 of PCA after plate-related pattern exclusion. Only compounds with multiple reagent batches are shown in the plot and separated by different colors. The size of circle represents the plate ID of each sample (compound). Two batches of 5-Fluorouracil are separated into 2 clusters, which shows the pattern related to reagent batches are hidden in the dataset.

This illustrated the fact that different batches of compound may not behave consistently, and batches of the same compound may behave differently at each time or dose. Thus, each batch of a certain compound was considered as an individual drug in the following ML jobs.

3.1.2 Data Cleaning

After data exploration and the removal of variation, the next step was sample cleaning. In this step, compounds which cannot dissolve in DMSO or have abnormal experimental data labelled as “retest required” and excluded from further ML jobs. This step was followed by feature cleaning, features containing only missing values and 0 or containing only 1 unique value were excluded from VD and AD. The table 2 and 3 shows the compound number, sample size and feature size of VD and AD. The compound number in AD is unclear because a marked ID was used for denoting each compound in each plate instead of the real names for the confidential purpose. It is impossible to identify the same compound in different plates in AD, so the real number of compounds in AD is unknown to me in this project.

Table 2: The description of VD.

VD	Compound Number	Sample Size	Feature Size
DNA Analogue	7	16	377
Aneugen	9	20	377
Clastogen	9	22	377
Other Negative	7	17	377

Table 3: The description of AD

AD-2017	Compound Number	Sample Size	Feature Size
Aneugen	<76	76	455
Clastogen	<89	89	455
Negative	<227	227	455
Positive Mixed	<88	88	455

The Fig. 6 and Fig. 7 shows the distributions of values of UD in VD and AD separately. There are 2.45% and 3.04% values exceeded the threshold of UD which was manually set as 0.8 in VD and AD separately. These values were set as missing values afterwards.

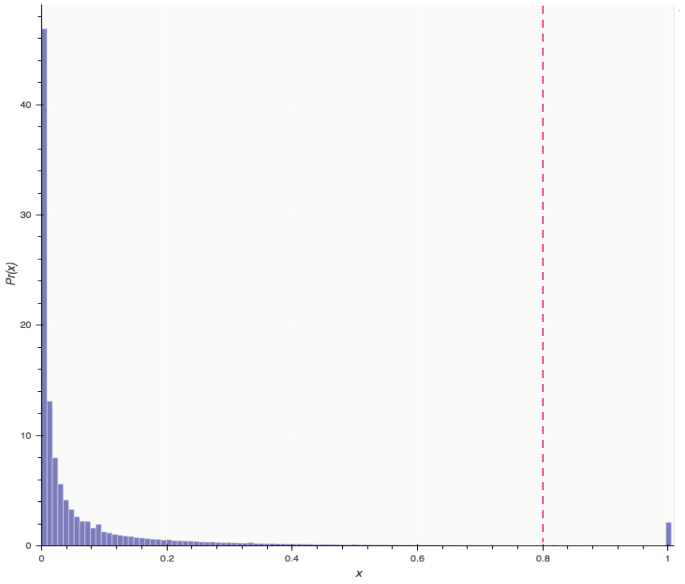


Figure 6: The distribution of values of UD in VD. The threshold is set as 0.8 for distinguishing "dirty" data. The values of UD for most of the "dirty" data are around 1(UD = 1 means 0 and non-0 values appear in duplications simultaneously).

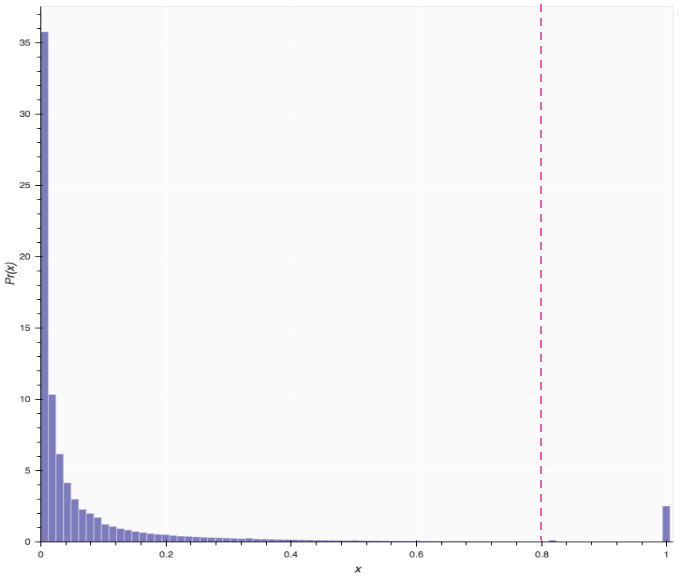


Figure 7: The distribution of values of UD in AD. The threshold is set as 0.8 for distinguishing "dirty" data. The values of UD for most of the "dirty" data are around 1(UD = 1 means 0 and non-0 values appear in duplications simultaneously).

3.2.1 Data Visualisation

People can perceive the information contained in pictures much better than in numbers, thus data visualisation can offer a shortcut to researchers and help them with understanding the information hiding in numeric dense data. The barcode-plotting approach developed in this project was used for data visualisation. Here two compounds were selected from each category in VD for exemplifying different cases. As all of these compounds were tested twice in different plates, both of them were shown for checking consistency.

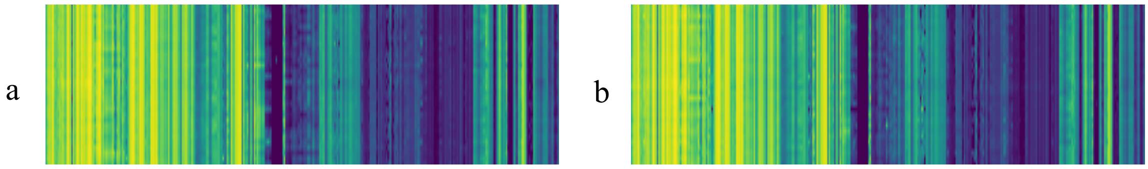


Figure 8: The images of Caffeine in plate 87 (a) and 440 (b)

Fig 8 are two images of Caffeine in plate 87 and 440, they are relatively clean barcode because Caffeine is in “negative” category, with no obvious cell death. Thus, it is hard to tell the differences among all of the DL.

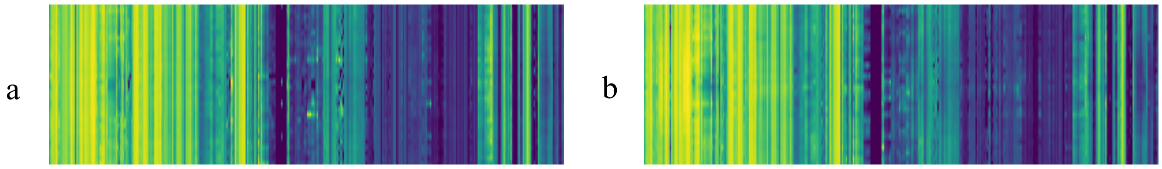


Figure 9: The images of Hydrocortisone in plate 5 (a) and 88 (b)

Fig 9 are two images of Hydrocortisone in plate 5 and 88, there is a dark area in the middle(high DLs) of both images compared to the top and bottom sides(low DLs), it shows the high DLs of Hydrocortisone have a negative impact on living cells though it is labelled as negative in VD, because the similar patterns can be found in images of genotoxicants.

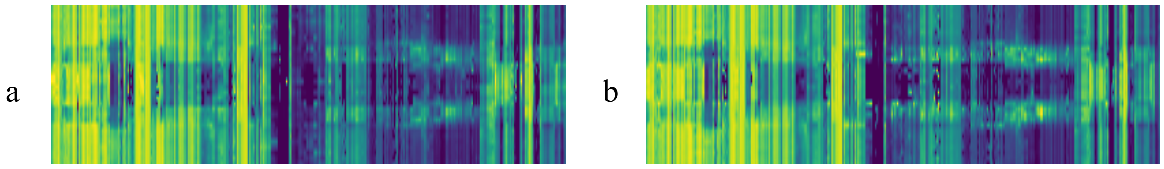


Figure 10: The images of 4-Nitroquinoline N-oxide in plate 86 (a) and 444 (b)

Fig. 10 are two images of 4-Nitroquinoline N-oxide (4NQO) in plate 86 and 444, 4NQO is a clastogen, and the signals in both images are much stronger than those of negative compounds. Two brighter lines in the upper and lower middle (moderate DLs) combined with a darker line in the central middle (high DLs) show that 4-Nitroquinoline N-oxide may have very different MOA at different DLs.

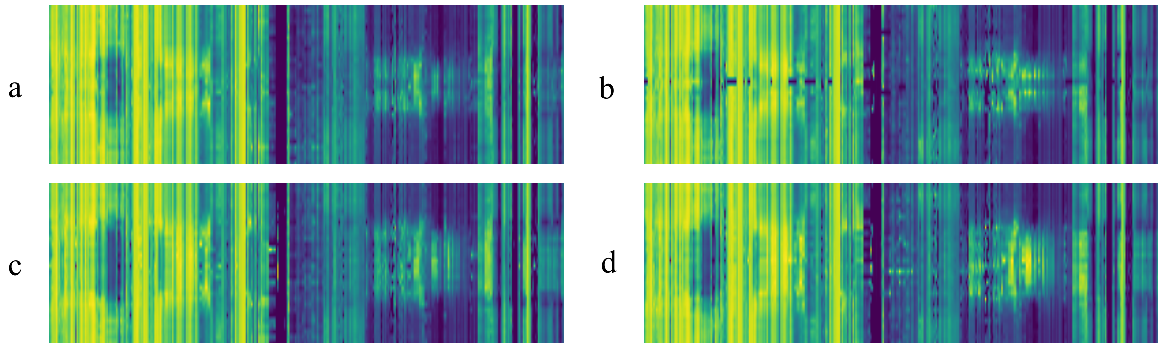


Figure 11: The images of Aphidicolin in plate 91 (a) and 439 (b) and Cytarabine in plate 86 (c) and 445 (d)

Fig. 11-a and Fig. 11-b are two images of Aphidicolin in plate 91 and 439. Although it is a clastogen as 4-Nitroquinoline N-oxide, the patterns showed in these images are much different from the ones in Fig. 10-a and Fig. 10-b. On the other hand, it is clear that similar patterns can be found in Fig. 11-c and Fig. 11-d, which are images of Cytarabine labelled as “analogue” in plate 86 and 445. This similarity demonstrates the difficulty of distinguishing a certain kind of genotoxicants from its similar “analogue”.

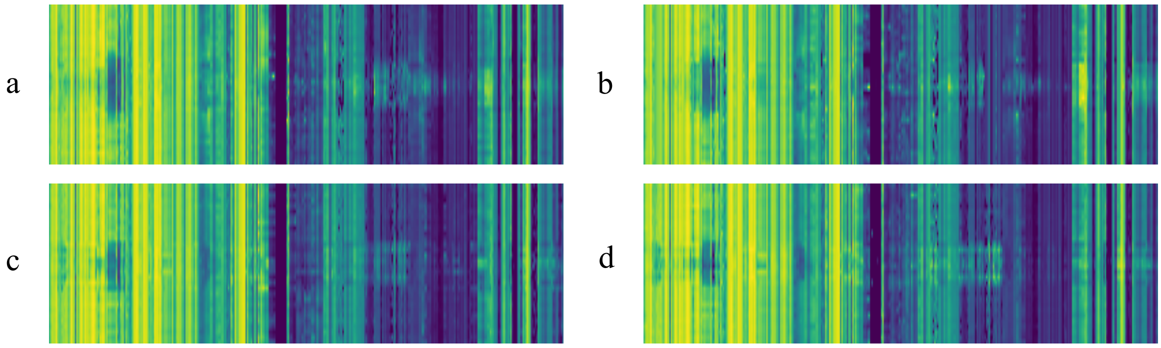


Figure 12: The images of 5-Azacytidine in plate 251 (a) and 443 (b) and (S,R)-Noscapine in plate 91 (c) and 446 (d)

Meanwhile, similar relation is revealed between an aneugen and an analogue as well by comparing Fig. 12-a and Fig.12-b with Fig. 12-c and Fig. 12-d, they are the images of 5-Azacytidine (labelled “analogue”) and (S,R)-Noscapine(labelled “aneugen”) respectively.

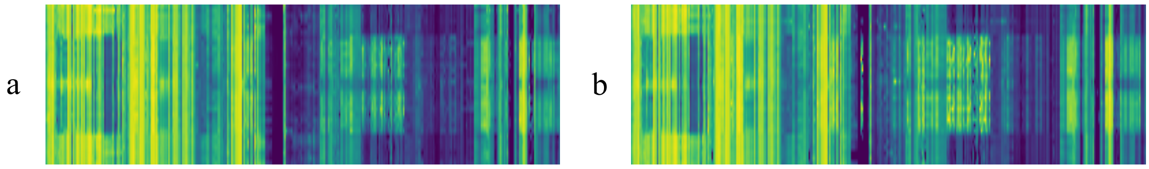


Figure 13: The images of Colchicine in plate 91 (a) and 444 (b)

Fig. 13 are two images of Colchicine in plate 91 and 444, which show one of the common patterns for aneugens, for example, a similar wide and consistent pattern can be also found in the images of Palclitaxel.

3.2.2 Dirty data detection

In this project, the quality of experiments was influenced by many factors, for example the precision of automatic equipment and manual mistakes. Sometimes a certain set of DLs of some compounds were not well controlled and may lead to the requirement of re-testing. Thus, the checking of data quality is indispensable before further analysis. With the help of barcode-like images, the quality checking can be much easier than looking into details in datasets.

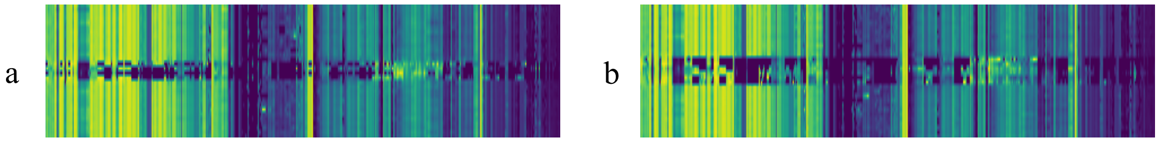


Figure 14: The images of two examples of abnormal responses at high DLs with upheavals in the middle.

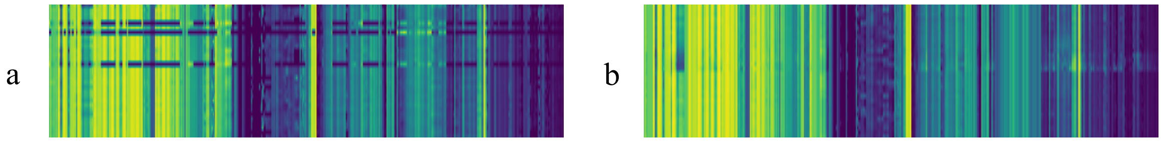


Figure 15: The image of an example of false doses (a), and the image of an example of missing doses (b).

Fig. 14 and Fig. 15 shows typical examples of flawed data which may be generated from defective experiments. Although not every type of flawed data can be detected by this visualisation way, it can provide researchers with a much easier way to pick out most of error data generated from defective experiments.

3.3.1 Affinity between Categories and Features

The affinity between categories and features was calculated based on the model of RF, which can be used for profiling features from the aspect of distinguishing different categories of compounds. For example, features which can distinguish clastogens from aneugens can be identified by plotting Fig. 16.

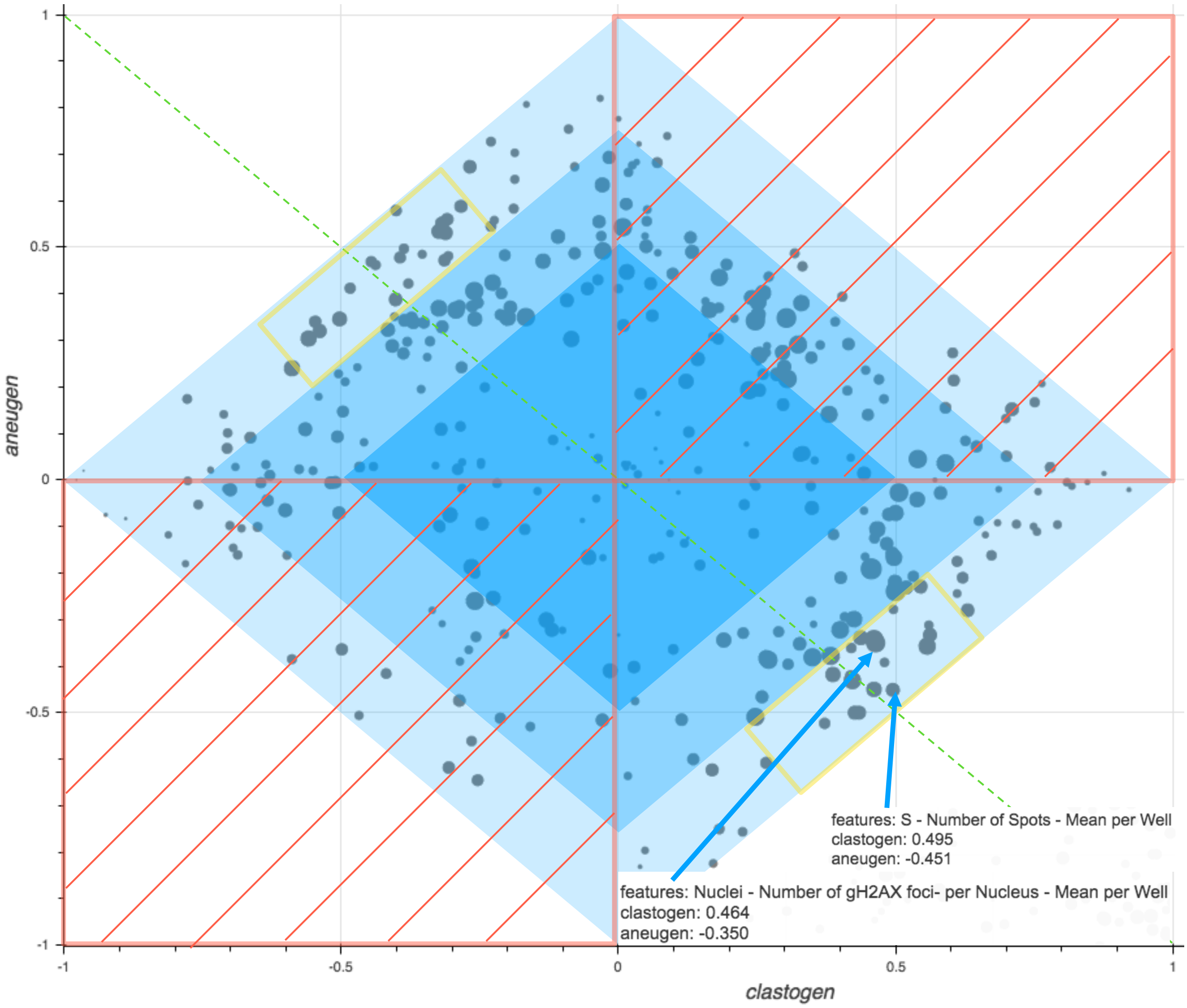


Figure 16: The plot of affinity between features and categories of "aneugen" and "clastogen" using the affinity between "clastogen" and features as x axis and the affinity between "aneugen" and features as y axis. The features shown in the first and the third quadrants (framed by red rectangle) are excluded because of the purpose of analysis. There are three areas colored with different shades of blue in the figure, the lightest, moderate, and richest blue area shows the features with highest, moderate and lowest distinguishing ability for either or both of the two classes. The size of circle represents the confidence of affinity.

In this case, there are 4 criteria of selecting suitable features to distinguish clastogens from aneugens. The first criterion is that features should have opposite affinity of clastogens and aneugens, thus the features shown in the first and the third quadrants (framed by red rectangle) are excluded. The second and the third criteria restrict the features to those with higher distinguishing ability (lightest blue area) for both (closed to the green diagonal) classes, and these selected features are shown within the yellow rectangle. The last criterion is that features should have high confidence of affinity, so that the result is more reliable.

By combining these criteria, features in the yellow rectangle with larger size were selected as the ones which can distinguish clastogens from aneugens. For example, the feature of "Nuclei – Number of gH2AX foci- per Nucleus – Mean per Well" was selected by this way, which was also a G-feature for classifying genotoxicants. This consistency shows that the affinity between categories and features is a reasonable measurement of features for building classification models.

3.3.2 Feature Description

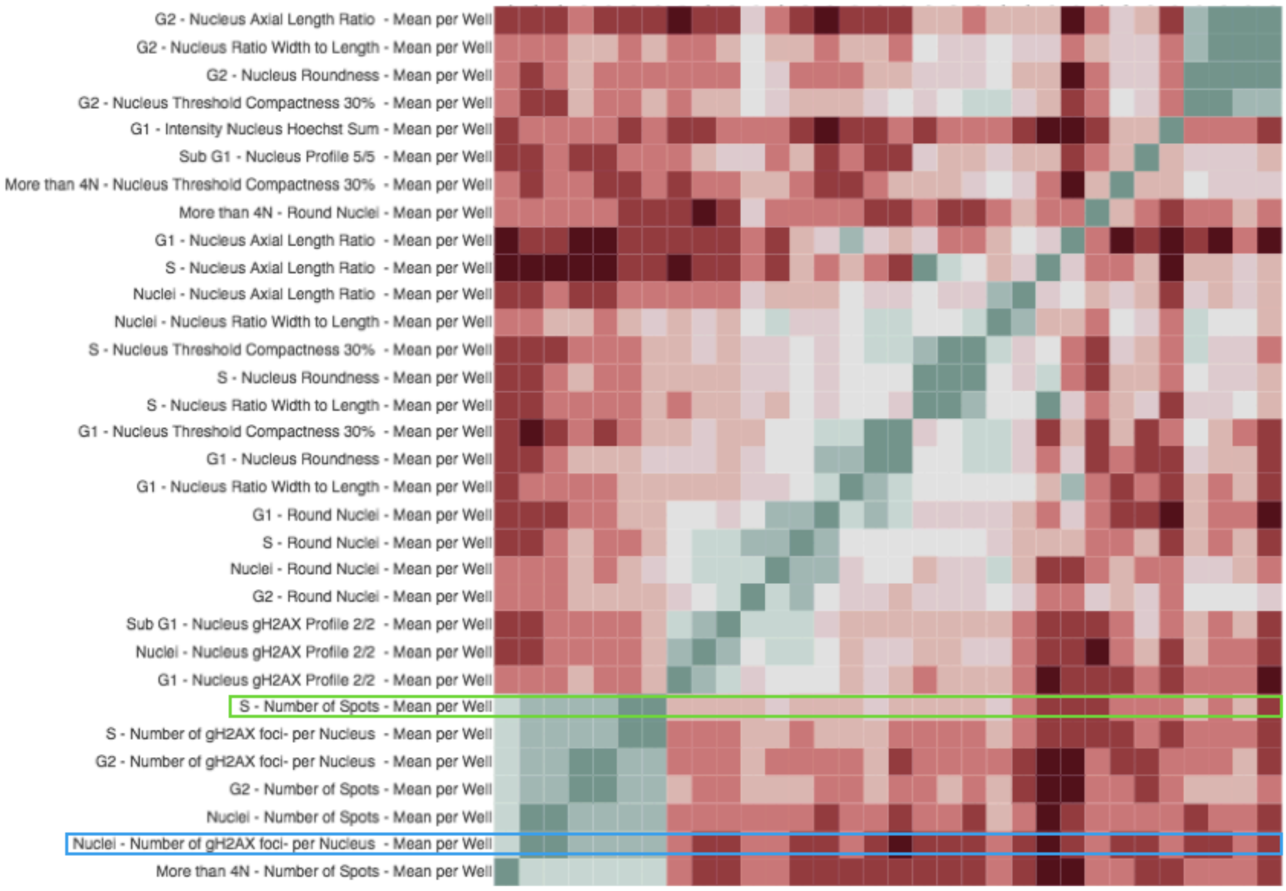


Figure 17: The heatmap of one cluster generated from K-means algorithm used Pearson's correlation coefficients as values. The feature framed by blue color is one of the G-features, and the feature framed by green color is the one suggested to replace the former one by ML algorithm.

Fig. 17 is a heatmap formed by a group of features clustered by K-means algorithm. The rows and columns of this heatmap are composed by the same set of features ordered bottom-up and left-to-right respectively. The color from red to green represents the correlation between a pair of features from low to high. The cluster was interpreted and titled as "against-aneugen", because most of the features in this cluster have negative affinity with aneugens while some of them have higher affinity with clastogens and others have higher affinity with negatives. The affinity vector of this cluster was calculated by averaging the affinity vectors of the features in this cluster, it is (-0.24, 0.13, 0.17) in the order of "aneugen", "clastogen" and "negative". The quality of clustering result can be evaluated by the proportion of green pixels in the heatmap, the more the green pixels are, the better the cluster is. Although the cluster shown in Fig. 17 does not seem to be one of best classified group of the K-means result due to too many red pixels in the heatmap, some useful information can be still inferred from the sub-clusters within it.

In Fig. 17, highly correlated features compose several sub-clusters which are shown as several green squares in different sizes. These sub-clusters are better inter-correlated than the main cluster, and features can be selected as the alternatives of another feature in the same sub-clusters for certain purposes. For example, "S – Number of Spots – Mean per Well" (framed by the green rectangle in Fig. 17, denote as F1) is a similar feature of "Nuclei – Number of gH2AX foci- per Nucleus – Mean per Well"(framed by the blue rectangle in Fig. 17, denote as F2) which is one of the G-features for distinguishing clastogens from aneugens. And F1 was shown as a better feature for distinguishing clastogens from aneugens based on the affinity between categories and features ((-0.451, 0.495, -0.054) against (-0.350, 0.464, -0.186), referring Fig. 16), so F2 was replaced by F1 in the G-features to build a better classification model.

And another G-feature of "Nuclei – Nuclear Fragmentation – Mean per Well" was also replaced by "Nuclei – Nucleus Width [μm] – Mean per Well" in the same way. Then, two set of features: original G-features and updated G-features were tested by building DT models using plate-related patterns excluded AD without "positive mixed" category for 50 times each in MCCV with 20% data used for testing separately. The models trained by the original and updated feature sets achieved the average accuracy of 80.5% and 82.1% respectively, it then had proved that the updated "Experts selecting features" could build a better DT model than the original G-features (in one tail independent-samples t-test, $t = 2.02$, $p<0.05$). The test showed that a probably better set of features could be selected by this process of feature description, and it can efficiently help researchers with understanding and selecting features among hundreds of them for certain purposes.

3.4.1 Classification Models

There are 5 steps of data pre-processing in order to generate the input dataset for building classification models listed as below:

Step 1: The category of “positive mixed” and samples of controls (both positive and negative controls) were excluded from AD, and then compounds which cannot dissolve in DMSO or have abnormal experimental data were also excluded, this step cannot be skipped.

Step 2: Clean the dataset generated in step 1 by the approach described in Sec. 2.2, this step can be skipped.

Step 3: Exclude the plate-related patterns from the dataset generated in step 2 by applying the Eq. 2, this step cannot be skipped.

Step 4-1: Flatten the dataset generated in step 3 by the approach described in Sec. 2.5.3, either step 4-1 or step 4-2 must be applied before moving to the next step.

Step 4-2: Select samples at the EC50 levels for each compound in the dataset generated in step 3 and then output the selected data by averaging the duplications of each compound in each plate, either step 4-1 or step 4-2 must be applied before moving to the next step.

Step 5: Balance the dataset generated in step 4-1 or step 4-2 by the approach described in Sec. 2.5.2, this step can be skipped.

There were three supervised ML algorithms (DT, RF and multi-label XGBoost) used to test different training dataset. The combinations of models and training datasets were tested for 50 times each by MCCV with 20% data used for testing separately. Overall accuracy with the recall and precision of 3 categories were calculated by averaging 50 times of tests, the results are shown in Table 4.

Table 4: The performance of the combinations of algorithms and datasets, these combinations are separated into 2 parts by the green line. The highlighted blue model is the best one in the first part denoted as model-A, and the highlighted red model is the best on in the second part denoted as model-B.

Models	Step 2	Step 4	Step 5	Mean of Accuracy	STD of Accuracy	Aneugen Recall	Aneugen Precision	Clastogen Recall	Clastogen Precision	Negative Recall	Negative Precision
DT	**Cleaned**	**EC50**	**Skipped**	80.6%	3.4%	72.4%	76.1%	69.2%	63.2%	87.6%	89.4%
XGBoost	**Cleaned**	**EC50**	**Skipped**	86.9%	3.5%	86.8%	90.8%	70.4%	79.0%	93.2%	88.3%
RF	**Cleaned**	**EC50**	**Skipped**	86.3%	3.4%	83.2%	89.0%	66.9%	83.4%	94.6%	86.3%
DT	**Skipped**	**EC50**	**Skipped**	82.6%	3.8%	79.5%	86.8%	70.7%	68.8%	88.2%	86.7%
XGBoost	**Skipped**	**EC50**	**Skipped**	85.8%	3.4%	86.8%	91.9%	66.7%	77.8%	92.3%	86.1%
RF	**Skipped**	**EC50**	**Skipped**	86.4%	3.3%	85.1%	90.0%	70.0%	80.2%	93.0%	87.2%
DT	**Cleaned**	**EC50**	**Balanced**	81.9%	5.0%	78.7%	80.4%	69.9%	66.4%	87.5%	88.7%
XGBoost	**Cleaned**	**EC50**	**Balanced**	86.4%	3.7%	86.9%	88.8%	74.7%	75.1%	90.7%	89.9%
RF	**Cleaned**	**EC50**	**Balanced**	85.6%	3.6%	83.3%	89.9%	76.6%	72.8%	89.8%	89.4%

DT	Skipped	EC50	Balanced	82.9%	4.7%	77.6%	91.2%	69.5%	66.9%	89.8%	86.8%
XGBoost	Skipped	EC50	Balanced	86.5%	3.4%	87.9%	91.1%	73.2%	74.0%	90.3%	88.9%
RF	Skipped	EC50	Balanced	86.9%	3.9%	85.7%	93.6%	75.6%	76.3%	91.6%	88.8%
RF	Cleaned	Flattened	Balanced	87.5%	3.6%	84.4%	92.0%	75.5%	77.8%	93.0%	89.6%
XGBoost	Cleaned	Flattened	Balanced	88.7%	3.4%	87.6%	90.4%	75.2%	82.7%	94.2%	90.2%
RF	Skipped	Flattened	Balanced	88.2%	3.1%	87.2%	92.9%	77.1%	78.8%	92.8%	90.2%
XGBoost	Skipped	Flattened	Balanced	88.3%	2.8%	86.7%	91.1%	77.8%	80.9%	92.7%	90.0%

The tests were divided into 2 parts by the green line, EC50 selected dataset was used in the first part to test the efficiency of different data pre-processing, because in this case the cost of time for building models was much less than using flattened dataset which was 8 times larger than EC50 selected data. Then flattened data was used in the second part to build the best classification model by applying the experience gained from the tests of first part.

The Table 4 shows, in general, there is no significant difference of accuracy whether the duplications errors were cleaned from the dataset by applying the approach described in Sec. 2.2. And the balanced data cannot affect the accuracy significantly either by compared to unbalanced data, but the recall of clastogen could be improved from 69.0% (average recall of the models trained by balanced data) to 73.3% (average recall of the models trained by unbalanced data) in the tests of the first part. So balanced data was going to be applied to the tests of the second part while cleaned data would not be applied in the following steps due to the negligible contribution.

The best model using the flattened dataset (model-B, highlighted as red in Table 4) achieved a precision of 88.7% which was 1.8% higher than the best one using the EC50 dataset (model-A, highlighted as blue in Table 4). And the areas under the curve (AUC) of model-B (Fig. 18-b) were also better than the one of model-A (Fig. 18-a). In conclusion, it showed the flattened dataset could train a better model than EC50 selected dataset which was the most widely-used format of dataset in genotoxicant classification problems. Based on previous tests, data balancing and concentration flattening are two useful ways of data pre-processing for building classification models in this case.

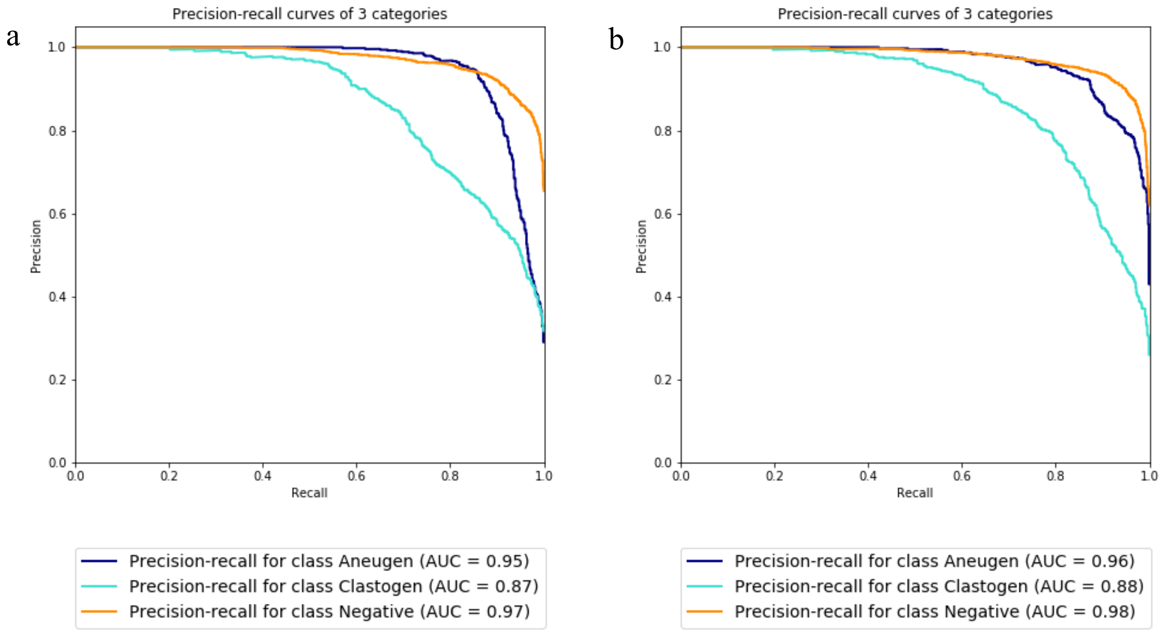


Figure 18: The AUC curve of model-A (a) and model-B (b)

3.4.2 Confidence of Predictions

Although the best classification models achieved a good result of both accuracy and AUC, there were still many samples predicted as wrong labels. Considering that bad cases of predictions are inevitable, it is certainly better to show the probability of every predictions which represent the confidence that the prediction is correct. However, the scores generated from classification models, for example RF models, cannot be directly used as probabilities, they have to be transformed to get close to the valid probabilities (to achieve the property of validity). To achieve it, VAP which can be applied easily on the top of any binary-label classification models was used in this project. In order to make the classification models more suitable for applying VAP, the best multi-label XGBoost model was replaced by three independent binary-label XGBoost models which can classify aneugens, clastogens and negatives from the others separately (27).

The prediction scores of binary-label XGBoost model ranges from 0 to 1 which is also the range of the VAP transformed probabilities, however the specific values were changed after the transformation. Due to the changes, the predicted labels may change too considering the prediction scores and the transformed probabilities share the same classifying threshold as 0.5 (values larger than 0.5 will be predicted as label 1, otherwise as label 0).

Thus, the accuracy of the original and VAP transformed models was compared in the following step.

Table 5: The performance of binary-label XGBoost models with VAP

Binary-label Model	Mean of Accuracy	Std. of Accuracy	Accuracy with VAP	Recall of the category with VAP	Precision of the category with VAP	AUC of the category with VAP
Model of aneugen	96.5%	2.0%	96.7%	94.7%	88.8%	0.97
Model of clastogen	91.7%	3.1%	91.8%	76.5%	85.0%	0.88
Model of negative	91.9%	2.5%	92.2%	94.4%	92.2%	0.97

Table 5 is the result of three binary-label XGBoost models tested 50 times each in MCCV with 20% data used for testing. The training dataset was the same one used for building the best multi-label XGBoost model. It is clear that VAP can slightly improve the accuracy of binary classification models in this case. The performance of three models was similar to the one of the best multi-label XGBoost model based on AUC.

Considering the VAP can improve accuracy as well as show the confidence of predictions, it is a good way to combine these binary-label classification models with VAP and build a better model with the help of multi-label XGBoost model if necessary. The main idea of the combination is that if any of the three binary-label models is confident enough for a prediction then its result will be used, otherwise the prediction of multi-label model will be used. The details are shown in Sec. 2.6. The combined model had been trained and tested using the balanced and flattened AD for 50 times by the method of MCCV with 20% data used for testing.

Table 6: The recall and precision of the combined model

Combined Model	Recall	Precision
Aneugen	93.9%	91.5%
Clastogen	77.6%	90.2%
Negative	96.8%	92.7%

The recall and precision of the combined model are listed in table 6, and it is clear that each of the measurement is better than or similar to the best one of previous models. The average accuracy of the combined model was 91.9% which is 3.2% higher than the one of the best previous model, and the standard deviation is 2.7% which is lower than any one of previous multi-label models. The average confidence of correctly predicted cases and wrongly predicted cases was 83.0% and 56.2% respectively. 9.8% testing samples were assigned with confidence no larger than 50%, and 69.7% of them were predicted correctly. Meanwhile, 43.8% testing samples were assigned with confidence at least 90%, and 99.6% of them were predicted correctly. In conclusion, the combination model achieved amazing accuracy when predicted the newly developed compounds, which is even higher than the one used for predicting model compounds in recent research (18). Furthermore, the probabilities are good reference for researchers to decide whether the results of classification models can be used confidently.

4. Discussion

4.1 Potential Problems in Data Cleaning

In order to improve the quality of the datasets and remove the inconsistencies, the data cleaning approach introduced in Sec. 2.2 was applied before building classification models. However, as the results showed in Table 4, the cleaned data could not train a better classification model than uncleaned data, which means the data cleaning approach might not work well as expected. The data cleaning approach mentioned here is the one used for removing the inconsistencies of duplications, the other one used for removing the plate-related patterns had already proved effective and will not be discussed in this section.

In general, data cleaning can be divided into 2 steps, they are “dirty” data definition and transformation. The aim of definition step is to identify as much “dirty” data as possible meanwhile make sure valid data is not considered as “dirty” data by mistakes. This aim seems similar to the one of classification models, which use “precision” and “recall” as the measurements to show the performance of models. While “dirty” data is not labelled, these measurements for “dirty” data cannot be calculated. In this case, the criteria of defining “dirty” data vary in different cases, domain knowledge and statistical methods are the common methods used for “dirty” data definition.

In this project, the “dirty” data was measured by UD, the threshold set for distinguishing “dirty” data is 80% maximal UD which was decided subjectively. There are several drawbacks of the method, the first one is that 80% maximal UD is a relatively conservative threshold in order to make sure a high precision of “dirty” data and save most of the potential useful information in the dataset. As a result, only 3.04% data is defined as “dirty” data, and the tiny proportion of data might not affect the classification results much whether or not it was cleaned.

Besides the subjectively set threshold, UD may be limited for defining “dirty” data in duplications alone. For example, the majority of “dirty” data excluded in this way was assigned 1 for UD, which means, for a 2-duplication case, the two values are 0 and a value larger than 0, for example, 0 and 100. But if the two values are 0 and 1 rather than 0 and 100, they are probably valid data. By applying the data cleaning approach, the data in both cases are considered as “dirty” data, which may cause mistakes. In order to optimize the approach, the absolute values shall be considered together with the UD which represent the ratio of values in duplications.

Abnormal data was also existing in the dimension of DL, for example a genotoxicant killed more cells at a certain lower DL than some of the higher DLs. It happened may because of an error dose or stochastic factors if those DLs are lower than the effective concentration at which toxic responses can be first observed. Although serious dose errors were all excluded from the dataset by the help of data visualisation introduced in Sec. 3.2.2, there are still some “dirty” data remaining. Thus, it is better to devise an algorithm to exclude this kind of “dirty” data.

After identifying the “dirty” data, it was replaced by “missing values”, then the first step was over. In the second step of data cleaning, these missing values were transformed into estimated values, in this project, they were replaced by 10^{-8}. The reason why they were replaced by a tiny value close to 0 is that as most of “dirty” data has the values of UD is 1, which means there is at least one 0 existing in the duplications (according to the UD equation), so that 0 is the most reasonable representative value for these “dirty” data compared to other values. In order to distinguish the replaced missing values from 0 which would be assigned as 10^{-7} when excluding the plate-related patterns, they were assigned with 10^{-8}. 10^{-7} and 10^{-8} are small enough to be distinguished from the minimal clean data which is about 10^{-5}. Besides that, there are also missing values in the original dataset, and they were sourced from undetectable cases which were also reasonable to be replaced by 0.

This method of dirty data transformation has drawbacks too, although 0 may be the best common value for replacing missing values, there are even better alternatives of 0 in different cases. For example, the method of mean substitution which use the mean of duplications instead of 0 as the replacing value. But this method is very sensitive to outliers and attenuate variance. Another more advanced alternative is full information maximum likelihood (FIML), which use the expectation-maximization algorithm to find the estimation of missing values. This high computational-cost algorithm had proved effectively even when dealing with the datasets containing a high proportion of missing values (28,41).

4.2 Wrongly Predicted Compounds

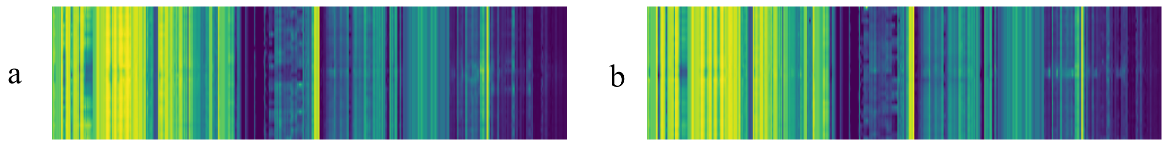
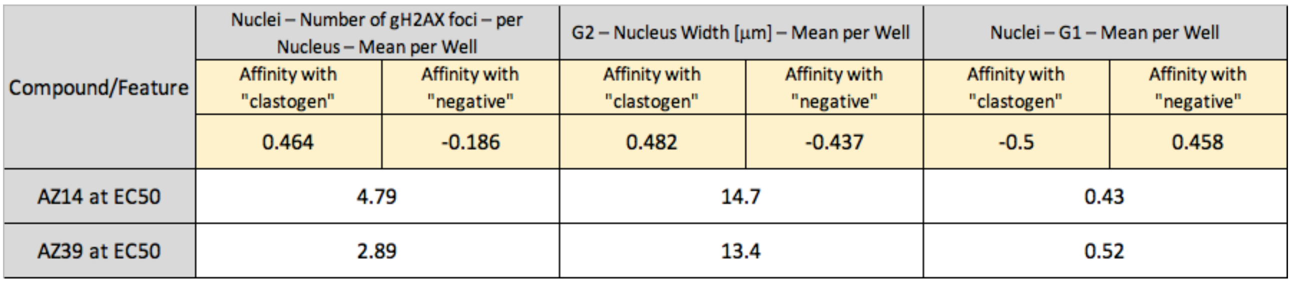


Figure 19: The images of AZ14 (a) and AZ 39 (b).

There were 392 compounds in the AD after data cleaning without the category of “positive mixed”, and 12 of them had never been predicted correctly while 324 of them had never been predicted wrongly. The other 56 compounds had been predicted either correctly or wrongly depending on the iterations of MCCV. The barcode-like images of all the wrongly predicted 12 compounds were checked for finding clues of why they were difficult to classify correctly. It turned out that some of them were really confusing based on the patterns in images. For example, AZ14 and AZ39 are two of the ten never predicted correctly compounds. The pattern of AZ14 (Fig. 19-a) looks similar to the ones of AZ39 (Fig. 19-b), and the former looks more obvious than the latter. In this case, AZ14 is more likely to be a genotoxicant than AZ39 as expected, but the reality is opposite as AZ14 was labelled “negative” and AZ39 was labelled “clastogen”. By comparing the barcode-like images, it is understandable when ML algorithms predict AZ14 and AZ39 as “clastogen” and “negative” separately.

However, images can only provide general information of tested compounds, so I chose three representative features to describe these two compounds in detail to see if it is still reasonable for ML algorithms to predict them wrongly. The first feature was chosen from G-features, which was “Nuclei – Number of gH2AX foci – per Nucleus – Mean per Well”. This feature is the best one in four G-features for distinguishing clastogens form negatives based on domain knowledge. The other two features were selected based on the way that the example showed in Sec. 3.3.1. The second feature “G2 – Nucleus Width [µm] – Mean per Well” has positive affinity with “clastogen” and negative affinity with “negative”, and the third feature “Nuclei – G1 – Mean per Well” has opposite affinity with categories of “clastogen” and “negative” compared to the second one. The compounds AZ14 and AZ39 at their EC50 were profiled by the three features as Table 7 shows:

Table 7: The values of AZ14(negative compound, predicted as clastogen) and AZ39(clastogen, predicted as negative compound) in 3 representative features for distinguishing clastogens from negatives.



Compound/Feature	Nuclei – Number of gH2AX foci – per Nucleus – Mean per Well		G2 – Nucleus Width [µm] – Mean per Well		Nuclei – G1 – Mean per Well	
	Affinity with "clastogen"	Affinity with "negative"	Affinity with "clastogen"	Affinity with "negative"	Affinity with "clastogen"	Affinity with "negative"
	0.464	-0.186	0.482	-0.437	-0.5	0.458
AZ14 at EC50	4.79		14.7		0.43	
AZ39 at EC50	2.89		13.4		0.52	

It is clear that AZ14 has higher values in the two features which have positive affinity with “clastogen” and AZ39 has a higher value in the feature which has positive affinity with “negative”. According to these features, AZ14 is more likely to be a clastogen compared to AZ39. Experts in AstraZeneca classified the compounds in AD based on a comprehensive analysis of concentration curves of representative features at the concentration below EC50, at these doses AZ14 and AZ39 were classified as “negative” and “clastogen” respectively.

Of course, it is possible that a compound behaves like “clastogen” at higher DLs above the EC50 while it behaves like a “negative” compound at lower DLs, and vice versa. It is therefore reasonable that the ML algorithms that consider 9 doses including those relatively high doses may misclassify compounds such as AZ14 and AZ39 (if only use lower doses, AZ14 and AZ39 may be classified correctly while the overall accuracy drops). It would be beneficial to further analyse, test and research these types of compounds due to the interesting behaviour and MOA.

12 compounds which were never predicted correctly were either “clastogen” but predicted as “negative” or “negative” but predicted as “clastogen”, because some of the clastogens were difficult to be distinguished from negatives based on MEGA Screen, and vice versa. Thus, the other features which can contribute to distinguish clastogens from negatives are better to be introduced into the MEGA Screen, otherwise the recall of “clastogen” which was the weakness of the best classification model may not be improved.

4.3.1 The Model Compounds vs The Real-world Compounds

Building classification models for compounds in either VD or AD are two completely different cases. The huge gap between the two cases is inevitable because the chemical space of AD, like the space of real-world compounds, is much more complex than the one of VD which only contains model compounds. In order to visualise the difference between the chemical spaces of VD and AD, t-SNE was used for dimension reduction then a 2-D figure was plotted based on the Euclidean-distance relation between compounds in each dataset. The difference can be observed by comparing the 2 figures.

Fig. 20 and Fig 21 are the t-SNE plots of VD and AD respectively, it is clear that compounds of different categories in Fig. 20 are separated better than those in Fig. 21 because many compounds mixed together in the middle in Fig. 21. Similar compounds are close to each other in t-SNE plots, for example Cytarabine and Aphidicolin are similar based on the barcode-like images of Fig. 11 and they are close to each other in the plot Fig. 20, and other similar examples can be found in the plots too. Thus, the t-SNE plots can reasonably reflect the distance relation between compounds, and the chemical space of AD is much more complex than the one of VD. In conclusion, it is much difficult to classify compounds in AD than those in VD.

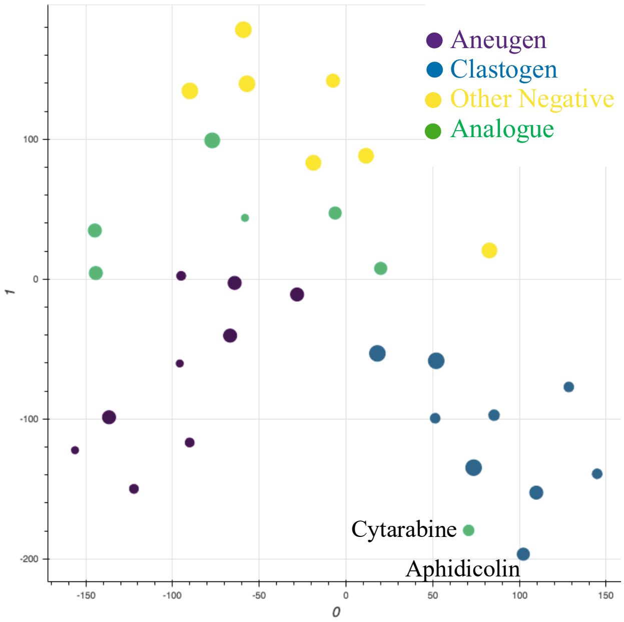


Figure 20: tSNE plot of VD with perplexity as 12.

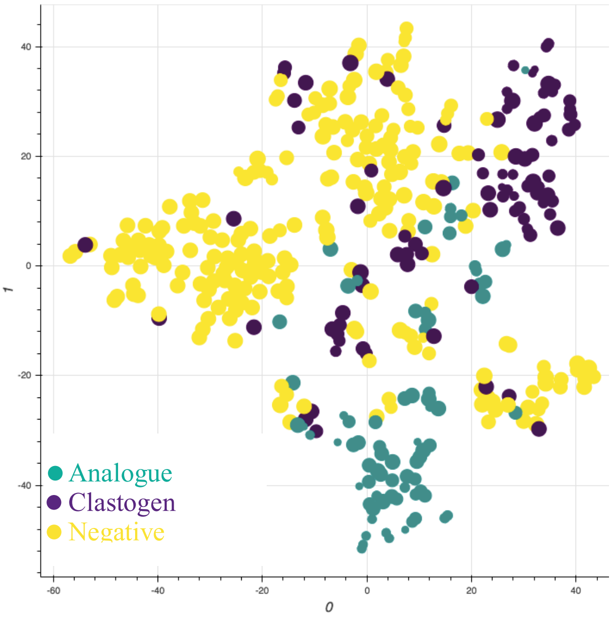


Figure 21: tSNE plot of AD with perplexity as 15.

4.3.2 Existing Data Science Research Related to Genotoxicity

The computational prediction of genotoxicity has been developed for over three decades, researchers are mainly focused on building predictive models via data coming from Ames Salmonella assay which is one of the approaches of bacterial reverse mutation test for genotoxicity at the early stage (16). Zhang et al. used random forest model to predict mutagenicity (positive or negative in Ames test) and achieved an error percentage of 15%-16% (17). In 2016, Bryce et al. applied logistic regression model to a dataset (contains 31 clastogens, 14 aneugens and 39 non-genotoxicants) originated from IVM assay and get a precision of 91% in cross validation (18). However, 67 and 17 compounds used for training and testing separately in Bryce's work were nearly all model compounds which were well defined and much easier to be distinguished by classification models. In contrast, the precision of the model for VD in this project was 95%, 3% and 4% higher than the one of the best model for AD and the one of Bryce's model respectively.

Some existing commercial software can also be used for predicting genotoxicity. For example, Derek Nexus is an expert knowledge-based software which gives predictions for a variety of toxicological endpoints using the structures of compounds, another example is Leadscope statistical quantitative structure-activity relationship models which can generate toxicity predictions. These commercial tools have pros and cons, they showed strength in distinguishing model compounds while their performance became much worse when coping with pharmaceutical companies' proprietary datasets which contained unique and newly developed compounds (19).

The research based on real-world compounds is much less, however Kumar, et al. built an SVM model which used the data of both in vivo and in vitro assays to classify 38 newly reported genotoxic compounds at the precision of 81.6% (37). In 2018, a related research project was done by Fan, et al. who built a classification model based on the fingerprint and chemical descriptors of 641 compounds to predict genotoxicants and achieved the best accuracy of 88.9% for 5-fold cross validation and 93.8% for predicting 65 model compounds (38). Comparing with these binary-label classification models developed recently by other researchers, the combined 3-label classification model developed in this project achieved even higher accuracy can contribute more to the pharmaceutical industry.

4.3 Applications in AstraZeneca

The work of this project is highly appreciated by experts of AstraZeneca, and there is a considerable opportunity to apply the approaches developed in this project to the practical industrial environment. The most valuable one may be the barcode-like images plotted for data visualisation, which can provide a creative general description of a tested compound in seconds. The images can be used for detecting abnormal experimental data, comparing the differences between two compounds and providing general information for classifying. As long as this feature is integrated into the MEGA Screen data analysis system, it will save plenty of time for researchers by showing the general features of tested compounds to them.

There are many antibodies/stains shown in Fig. 22, these antibodies/stains cannot be used in MEGA Screen for now because of the limitation of 4 channels of MEGA Screen. By doing feature description shown in Sec. 3.3.2, it is possible to find a better set of representative features instead of using all of them. In this case, the features generated by some in use antibodies/stains may be replaced by others, thus it is possible to save a few channels by excluding some antibodies/stains in order to introduce potential more useful antibodies/stains to MEGA Screen (29).

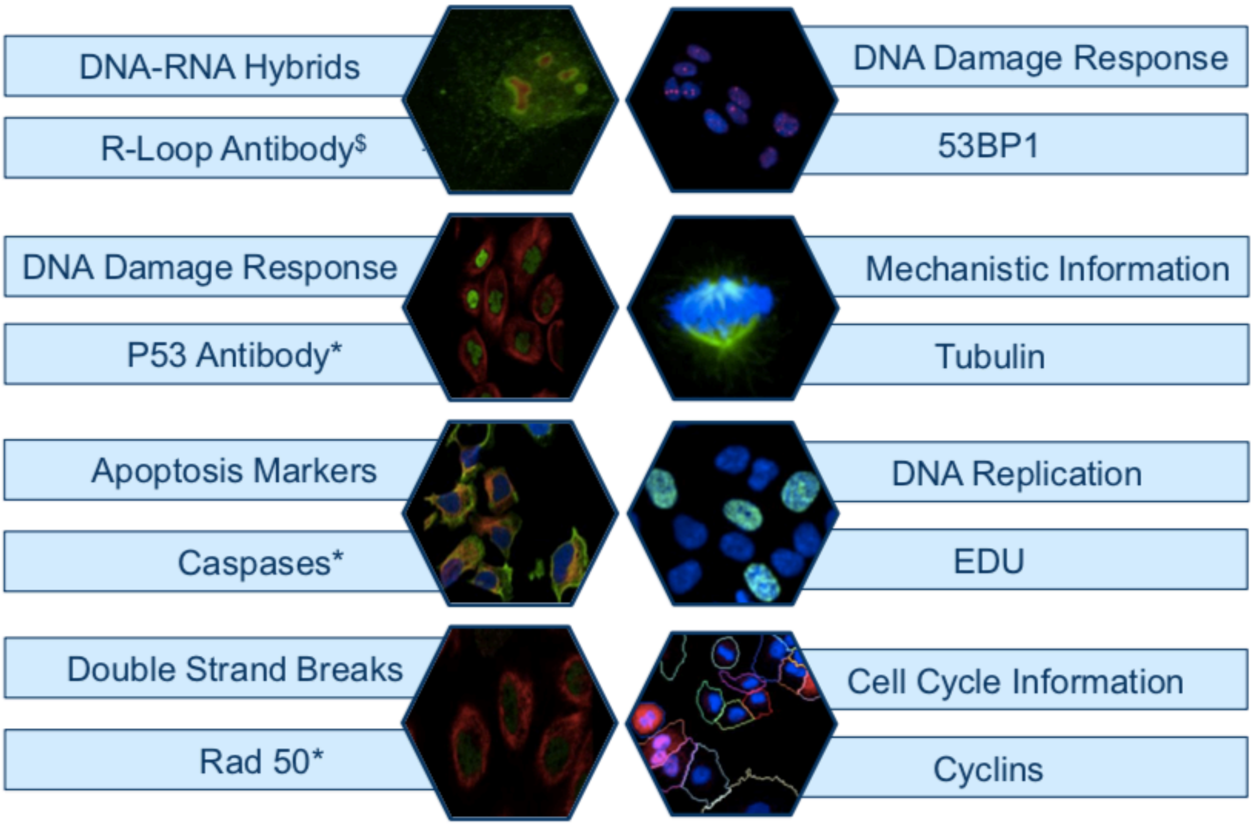


*Figure 22: The potential antibodies/stains and corresponding generated information for MEGA Screen. $ Makharashvili et al, 2018 (40) * Human protein Atlas. (29)*

The combined classification model can provide not only prediction results but also the confidence of predictions to researchers, and the researchers can decide which results are deserved to be referred when detecting genotoxicants. For example, if the predictions with confidence higher than 90% are decided to be referred, the researchers can have an overall precision of more than 99% in more than one third tested compounds. Which means at most about one third of the cost of time and money for detecting genotoxicants from the data of MEGA Screen can be saved by the help of the combined classification model.

4.4 Future Work

The barcode-like images generated in this project are all 1-layer images, which means there are 2 more layers can be used for visualising more data sources or comparing different compounds in a single image (the three layers are the red, green and blue layer). There is potential space for the development of the barcode-like images and showing more information based on the requirement and the demand of researchers.

The data cleaning method used in this project did not performed well in building classification models as the discussion in Sec. 4.1 shows. Many other more advanced data cleaning methods, for example FIML, are worth being tested in future.

In this project, the only data source is the MEGA Screen, which limited the performance of classification models. If the other data sources, for example the structure of compounds (27) and the result of manual IVM assays, can be imported to build classification models, the result will be improved probably. Besides that, convolutional neural network (CNN) is a potential more powerful model which can distinguish the categories of “aneugen”, “clastogen” and “negative” based on not only images generated from MEGA Screen but also the structures of compound. As the CNN model is one of the most powerful classification models for now, it may provide a better classifying tool than the combined classification model developed in this project (30).

References

1. Banerjee T, Siebert R. Dynamic impact of uncertainty on R&D cooperation formation and research performance: Evidence from the bio-pharmaceutical industry. *Research Policy.* 2017; 46 (7): 1255-1271. Available from: https://doi.org/10.1016/j.respol.2017.05.009
2. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today.* 2012; 17 (19-20): 1088-1102. Available from: doi: 10.1016/j.drudis.2012.05.007
3. Dunwiddie CT, Munos BH, Lindborg SR, Paul SM, Schacht AL, Mytelka DS, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery.* 2010; 9 (3): 203-214. Available from: doi: 10.1038/nrd3078
4. W Haverkamp, G Breithardt, A J Camm, M J Janse, M R Rosen, C Antzelevitch, et al. *The potential for QT prolongation and proarrhythmia by non-antiarrhythmic drugs: clinical and regulatory implications. Report on a Policy Conference of the European Society of Cardiology.* England: 2000. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10924311
5. Preston R, Hoffmann G. *Casarett & Doull's Essentials of Toxicology.* Chapter 9: Genetic Toxicology. United States of America: The McGraw-Hill Companies, Inc.; 2015.
6. Brookes P, Lawley P. Evidence for the Binding of Polynuclear Aromatic Hydrocarbons to the Nucleic Acids of Mouse Skin Relation between Carcinogenic Power of Hydrocarbons and their Binding to Deoxyribonucleic Acid. *Nature.* 1964; 202 781-784.
7. European Medicines Agency. *Genotoxicity: a standard battery for genotoxicity testing of pharmaceuticals.* the UK: EMEA; 2006.
8. Elizabeth Mambo, Xiangqun Gao, Yoram Cohen, Zhongmin Guo, Paul Talalay and David Sidransky. Electrophile and Oxidant Damage of Mitochondrial DNA Leading to Rapid Evolution of Homoplasmic Mutations. *Proceedings of the National Academy of Sciences of the United States of America.* 2003; 100 (4): 1838-1843. Available from: doi: 10.1073/pnas.0437910100
9. Fry RC, Begley TJ and Samson LD. Genome-wide responses to DNA-damaging agents. *Annual Review of Microbiology.* 2005; 59 (1): 357-377. Available from: doi: 10.1146/annurev.micro.59.031805.133658
10. Mishima M. Chromosomal aberrations clastogens vs aneugens. *Frontiers in bioscience (Scholar edition).* 2017; 9 (1): 1-16. Available from: doi: 10.2741/s468
11. E M Parry, J M Parry, C Corso, A Doherty, F Haddad, T F Hermine, et al. Detection and characterization of mechanisms of action of aneugenic chemicals. England: 2002. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12435848.
12. LINDA J. KUO, LI-XI YANG. γ-H2AX - A Novel Biomarker for DNA Double-strand Breaks. *In Vivo.* 2008; 22 (3): 305.
13. Doherty A, Bryce S and Bryce J. *Genetic Toxicology Testing: A Laboratory Manual.* Chapter 6: The In Vitro Micronucleus Assay 164. Elsevier Inc; 2016.
14. Naven RT, Louise-May S and Greene N. The computational prediction of genotoxicity. *Expert Opinion on Drug Metabolism & Toxicology.* 2010; 6 (7): 797-807. Available from: doi: 10.1517/17425255.2010.495118
15. Elloway J, Ling S, Stott J, Wilson A and Doherty A. Quantitative multi-parametric analysis of genotoxicants (unpublished).

16. Haworth, S. , Lawlor, T. , Mortelmans, K. , Speck, W. and Zeiger, E. (1983), Salmonella mutagenicity test results for 250 chemicals. Environ. mutagen, 5: 3-49. doi:10.1002/em.2860050703
17. Zhang Q, Aires-de-Sousa J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors *Journal of Chemical Information and Modelling* 2007 *47* (1), 1-8 doi: 10.1021/ci050520j
18. Bryce SM, Bernacki DT, Bemis JC,Dertinger SD. 2016. Genotoxic mode of action predictions from a multiplexed flow cytometric assay and a machine learning approach. *Environ Mol Mutagen*57:171–189.
19. Naven RT, Greene N and Williams RV. Latest advances in computational genotoxicity prediction. *Expert Opinion on Drug Metabolism & Toxicology.* 2012; 8 (12): 1579-1587. Available from: doi: 10.1517/17425255.2012.724059
20. Dragoeva A, Koleva V, Hasanova N and Slanev S. Cytotoxic and Genotoxic Effects of Diphenyl-ether Herbicide GOAL (Oxyfluorfen) using the Allium cepa test. *Research Journal of Mutagenesis.* 2012; 2 (1): 1-9.
21. Radka Zounková, Pavel Odráska, Lenka Dolezalová and Klára Hilscherová. Ecotoxicity and genotoxicity assessment of cytostatic pharmaceuticals. *Environmental toxicology and chemistry.* 2007; 26 (10): 2208-2214. Available from: doi: 10.1897/07-137R.1
22. Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.* 2002; 16 321-357. Available from: doi: 10.1613/jair.953
23. Chen C, Liaw Andy and Breiman L. Using Random Forest to Learn Imbalanced Data. 2004, available from: https://statistics.berkeley.edu/tech-reports/666
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*: ACM; Aug 13, 2016. pp. 785-794.
25. Vovk V, Petej I. *Venn-Abers predictors.* 2012. Available from: https://ui.adsabs.harvard.edu/\#abs/2012arXiv1211.0025V
26. Paolo Toccaceli (joint work with Alex Gammerman, Ilia Nouretdinov). Tutorial on Venn-ABERS prediction*. 6th Symposium on Conformal and Probabilistic Prediction with Applications. 2017.* Available from: http://www.clrc.rhul.ac.uk/copa2017/presentations/VennTutorialCOPA2017.pdf
27. Ahlberg E, Buendia R and Carlsson L. Using Venn Abers Predictors to assess Cardio Vascular Risk. *Proceedings of Machine Learning Research.* 2018; 91 1-15.
28. Hirose K, Kim S, Kano Y, Imada M, Yoshida M and Matsuo M. Full information maximum likelihood estimation in factor analysis with a lot of missing values*. Journal of Statistical Computation and Simulation,* 2016. 86 (1): 91-104
29. Amy Wilson. Development, Validation and Implementation of a High Content Genotoxicity Assessment. 2018.(unpublished)
30. Wallach I, Dzamba M and Heifets A. *Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. CoRR, abs/1510.02855, 2015.*
31. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.
32. Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning.* 1994; 16 (3): 235-240. Available from: doi: 10.1007/BF00993309
33. Breiman L. Random Forests. *Machine Learning.* 2001; 45 (1): 5-32. Available from: doi: 1010933404324

34. Paolo Toccaceli, VennABERS in python, 2017. Available from: https://github.com/ptocca/VennABERS
35. Trygg J, Holmes E and Lundstedt T. Chemometrics in Metabonomics. *Journal of proteome research.* 2007; 6 (2): 469-479. Available from: doi: 10.1021/pr060594q
36. van der Maaten, L. J. P, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research.* 2008; 9 (Nov): 2579-2605.
37. Kumar P, Ma X, Liu X, Jia J, Bucong H, Xue Y, et al. Effect of training data size and noise level on support vector machines virtual screening of genotoxic compounds from large compound libraries. *Journal of Computer-Aided Molecular Design.* 2011; 25 (5): 455-467. Available from: doi: 10.1007/s10822-011-9431-3
38. Defang Fan, Hongbin Yang, Fuxing Li, Lixia Sun, Peiwen Di, Weihua Li, Yun Tang, Guixia Liu. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicology Research (Camb)* 2018 Mar 1; 7(2): 211–220. Published online 2017 Dec 15. doi: 10.1039/c7tx00259a
39. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). *Guidance for Industry-S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use.* ICH; 2012.
40. Makharashvili N, Arora S, Yin Y, Fu Q, Wen X, Lee J, et al. Sae2/CtIP prevents R-loop accumulation in eukaryotic cells. *eLife. 7 (2018), p. e42733* Available from: doi: 10.7554/eLife.42733
41. Acock, A. C. (2005). Working with missing values. Journal of Marriage and Family, 67, 1012–1028. Available from: https://doi.org/10.1111/j.1741-3737.2005.00191.x