Machine Learning – Predicting Ames Mutagenicity of Small Molecules

Charmaine S. M. Chu 1, Jack D. Simpson, Paul M. O’Neill, Neil G. Berry 1,\*,

1 Department of Chemistry, University of Liverpool, Crown Street, Liverpool L69 7ZD;

\* Correspondence: ngberry@liverpool.ac.uk; Tel. : +44 151 794 3877

**Abstract:** In modern drug discovery, detection of a compound’s potential mutagenicity is crucial. However, the traditional method of mutagenicity detection using the Ames test is costly and time consuming as the compounds need to be synthesised and then tested and the results are not always accurate and reproducible. Therefore, it would be advantageous to develop robust *in silico* models which can accurately predict the mutagenicity of a compound prior to synthesis to overcome the inadequacies of the Ames test. After curation of a previously defined compound mutagenicity library, over 5000 molecules had their chemical fingerprints and molecular properties calculated. Using 8 classification modelling algorithms, including support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGB), a total of 112 predictive models have been constructed. Their performance has been assessed using 10-fold cross validation and a hold-out test set and some of the top performing models have been assessed using the y-randomisation approach. As a result, we have found SVM and XGB models to have good performance during the 10-fold cross validation (AUROC > 0.90, sensitivity > 0.85, specificity > 0.75, balanced accuracy > 0.80, Kappa > 0.65) and on the test set (AUROC > 0.65, sensitivity > 0.65, specificity > 0.60, balanced accuracy > 0.65, Kappa > 0.30). We have also identified molecular properties that are the most influential for mutagenicity prediction when combined with chemical molecular fingerprints. Using the Class A mutagenic compounds from the Ames/QSAR International Challenge Project, we were able to verify our models perform better, predicting more mutagens correctly then the StarDrop Ames mutagenicity prediction and TEST mutagenicity prediction.

**Keywords:** Machine learning; Ames; Toxicity; random forest; support vector machine; extreme gradient boosting

1. Introduction

Within the modern drug discovery field, the mutagenicity of a compound is a crucial property that can restrict the development of a particular compound series at all stages of drug development due to its close relationship with carcinogenicity [1,2]. In order to assist the early identification of potential mutagenic compounds and hence reduce the time and expense associated with hit to lead optimisation, *in silico* prediction of compound mutagenicity has attracted much attention from several research groups [3-5]. One of the most widely used assays for testing the mutagenicity of a compound is the Ames test, invented by Professor Bruce Ames in the early 1970s [6-9]. The Ames test contains a bacterial revertant mutation assay with a simulation of mammalian metabolism, which is highly sensitive for chemicals which can induce genetic damage and frameshift mutation in the environment [10]. However, there are limitations to mutagenicity detection using this approach, including identification of false-positives and false-negatives amongst the outputs [11], and the interlaboratory reproducibility rate is not 100% [12-14]. Nonetheless, it still serves well as a quick and cheap alternative to the standard carcinogen assays on test animals as there are no genotoxic causes of carcinogenicity involved in the Ames test.

As the Ames test can take considerable time and material to carry out especially where large compound libraries need to be analysed, the availability of a reliable *in silico* model would be advantageous. By using robust predictive *in silico* models, the number of Ames tests needed to be carried out can be reduced, thus reducing the time and resources needed. Over the past several decades, there have been many statistical models [15-17] and structural alert based models [3,18,19] published in literature alongside various commercial [4,20] and open source software packages [20] which attempt to address Ames mutagenicity prediction. In one of the works, the performance of commercial programmes for Ames prediction have been compared with statistical models [16]. It was found that the statistical models constructed within the work outperform the commercial programs when analysing the corresponding Receiver Operating Characteristic curve (ROC). However, the predictive accuracy and robustness of these models are not yet satisfactory as their application domain is bound by the database which the models were constructed from and none of them have over 95% accuracy [21-23]. Although for theoretical reasons, it is impossible to reach 100% accuracy with statistical model and with the imperfect reproducibility of the experimental Ames test itself, it is difficult to achieve models with accuracy over 95 %, it is still important to investigate in method to overcome this hurdle.

In order to try and overcome the above difficulties, C. *Xu et al.* constructed a large database based on five different sources containing more than 8300 compounds with experimentally derived mutagenicity [5]. Using this, the Xu group reported some predictive models using molecular fingerprints, a type of molecular descriptor widely used in similarity searching [24], virtual screening [25] and classification [26,27]. In addition, recently an Ames/QSAR International Challenge Project has been reported where 12 QSAR vendors across the world have worked in collaboration to test and improve their Ames QSAR tools using a database of over 12000 molecules established by the Division of Genetics and Mutagenesis, National Institute of Health Sciences of Japan [14].

We have taken the openly available mutagenicity database published by C. *Xu et al.* and produced a different range of models in order to try to identify improved models to those from Xu *et al.* [5]*.* By combining a range of physicochemical molecular properties and molecular fingerprints as compound descriptors and an alternative selection of modelling algorithms, we aimed to identify models that have excellent predictive ability and that can provide molecular insights to uncover aspects of the molecules that cause mutagenicity. Scheme 1 displays the overall approach, with further details and discussion presented below.

|  |
| --- |
|  |

**Scheme 1.** A flow chart of the process of searching for the best performing model and data set. Rounded rectangle: raw data; rectangle: processes; coloured background: Software used. Each step is further described in *2.1. Data Preparation*, *2.2. Data compilation*, *2.4. Data pre-processing*, *2.6. Model building and performance assessment* and *2.7. Model validation via y-randomisation*.

2. Materials and Methods

2.1. Data Preparation

The raw data used in this study was the C. Xu’s Ames data collection provided within ‘In silico Prediction of Chemical Ames mutagenicity’ as Supporting Information Table S1 [5]. The data collection consists of a training set (7617 molecules), an external validation set (731 molecules) and a balanced external set (234 molecules). The balanced external set is a subset of the external validation set which the number of mutagens and non-mutagens are balanced. We discovered that the library *Xu et al.* presented in their work contained duplicates, and therefore we curated the library.

The curation was achieved using Pipeline Pilot 2017 [28] via the following steps, as suggested by Tropsha [29] while taking into account the steps *Xu et al.* [5] followed. Firstly, any inorganic molecules, defined as those without carbon atoms within the structure, were removed. Secondly, each molecule was analysed and any molecules with unspecified stereochemistry were removed. Thirdly, the molecules were standardised using the InChI key [30]. Fourthly, any salt fragments, defined with the built-in salt fragment list in Pipeline Pilot, were removed. Finally, any duplicates across the data collection were identified and removed using the InChI key.

2.2. Data compilation

Continuing in Pipeline Pilot 2017 [28], the curated data collection was split into training and test sets by the following steps. Firstly, the data collection was clustered into 1000 clusters using extended connectivity fingerprints [31] (ECFP, diameter = 6). Secondly, where possible, a mutagenic and a non-mutagenic representative closest to the cluster centre were taken from each cluster as candidates for the test set. Thirdly, the representative molecules from any cluster with only one representative were put into the test set. Fourthly, the mutagenic representatives from the clusters with two representatives were put into the test set until the threshold of 500 molecules was reached. Fifthly, the non-mutagenic representatives from the rest of the clusters were put into the test set. Finally, any molecules not in the test set were put into the training set.

Next, molecular properties (MP) (AlogP, molecular weight, number of atoms, number of hydrogen acceptors, number of hydrogen donors, number of rotatable bonds, number of rings, number of aromatic rings, molecular surface area, molecular polar surface area, molecular polar solvent-accessible surface area, molecular solubility, logD), extended connectivity fingerprints [31] (ECFP, diameter = 4, 2048 bits) and functional-class fingerprints [31] (FCFP, diameter = 4, 2048 bits) for both the training and test set were calculated. Various combinations of these generated predictors sets were exported as separate .csv files for each of the training and test sets, resulting in seven predictor data sets:

1. ECFP: 2048 bits of extended connectivity fingerprints only,
2. FCFP: 2048 bits of functional-class fingerprints only,
3. MP: 13 molecular properties,
4. ECFP+FCFP: combination of 2048 bits of extended connectivity fingerprints and 2048 bits of functional-class fingerprints,
5. MP+ECFP: combination of 13 molecular properties and 2048 bits of extended connectivity fingerprints,
6. MP+FCFP: combination of 13 molecular properties and 2048 bits of functional-class fingerprints,
7. MP+ECFP+FCFP: combination of 13 molecular properties, 2048 bits of extended connectivity fingerprints and 2048 bits of functional-class fingerprints.

Using the training set of each predictor data set, an applicability domain was defined by tracking the property range and analysing the optimum prediction space[32]. The molecules from the test set were then filtered for model applicability by analysing them against the defined applicability domain. Any molecules which did not pass the model applicability filter were swapped for the next molecule closest to the cluster centre within the same cluster with the same mutagenicity label from the training set. The applicability domain was then defined again via the same process until no more swaps could be made. Any molecules which did not pass the model applicability filter were then removed from the test set and placed back into the training set. The training and test sets were then exported as .sdf files (File S1 – S2) and .csv files for further calculation.

2.3. R and caret

R is an openly available programming language and software environment for statistical computing and graphics [33]. For the purpose of this study, aside from the “base” package, a user-created package “caret” (classification and regression training) was used extensively as it contains a number of very useful functions for:

* Balanced training/test set splitting – training sets used to build the predictive model and test sets to test the model’s action are required for validation of the constructed model predictivity,
* Predictor pre-processing – processing of the predictors into a format suitable for the modelling algorithm,
* Model tuning using resampling – training set is resampled and cross validated for each kernel parameter to find the optimal value(s) to reduce overfitting [34],
* Variable importance estimation – estimation of the importance of the variables using different methods depending on the modelling algorithm,

and much more [35]. It also allows various model training using unified syntax by calling on the relevant function from other packages and returns the result in a uniform fashion, allowing easy comparison of various models trained with the same data set.

2.4. Data pre-processing

For each of the seven predictor data sets (see *2.2. Data compilation*), the training and test set were imported into R and the column containing the mutagenicity label was converted into factors.

Using the training set, near zero variances predictors, predictors with low frequency ratio for the most common value over the second most common value were removed. As a result, two variations of each predictor set are selected using the following rules:

* Variation 1: the full set of filtered predictors where predictors with frequency ratio above #observation/10 were removed
* Variation 2: the reduced set of filtered predictors where predictors with frequency ratio above #observation/100 were removed

Predictors with pair-wise absolute correlations over 0.9 were identified with caret:::findCorrelation. For each predictor pair, the average correlation with the rest of the predictors were calculated and the predictor with the higher average correlation was removed to give the final version of the two variations of each predictor set.

2.5. Algorithm selection

One linear (PLSDA), three non-linear (MDA, SVM and KNN) and four tree/rule-based (C5, RF, GBM and XGB) model algorithms were chosen to investigate their performance on the seven data sets, covering the range of simpler and more interpretable models to potentially more robust but more complex models for identifying the optimally performing methods.

2.5.1. Partial least squares discriminant analysis (PLSDA)

Partial least squares discriminant analysis (PLSDA) is an application of the linear regression model partial least squares (PLS) to a classification problem. The algorithm involves the search of latent variables, composed of a defined number of predictor components, which reduces the predictor space dimension and optimises correlation with the categorical response represented as dummy variables (0s and 1s) [34]. New samples are predicted as a number for dummy variables of each class and the class with the largest predicted value is the predicted class [34]. The default number of components for tuning within “caret” are 1, 2 and 3.

2.5.2. Mixture discriminant analysis (MDA)

Mixture discriminant analysis (MDA) is a non-linear classification model based on linear discriminant analysis (LDA). Within MDA, each class is represented by multiple multivariate normal distributions, and this number can be controlled [34]. A single multivariate normal distribution is then generated from the multiple multivariate normal distribution by creating a per-class mixture and the class of a new sample is then determined by the position of it within the multivariate normal distribution for each class [34]. The default number of multiple multivariate normal distribution for tuning within “caret” are 2, 3 and 4.

2.5.3. Support vector machines (SVM)

Support vector machines (SVM) were originally developed by Vapnik for classification problems which seeks to find the optimal hyperplane within the predictor space which can separate the classes [36]. The optimal hyperplane would have the maximum margin and the training points defining this margin are called support vectors, which are used to find the class probability for the new samples [34,36]. For a data set that is not completely separable, SVM utilises different kernel functions to map the input space into a high dimensional feature space, allowing calculation without needing to transform the predictors [34,36]. When there are data points that lie on the wrong side of the hyperplane or within the margin, the margin is penalised by adding a cost [34]. Within this study, we used radial basis function as the kernel function and the kernel parameter *σ* was derived as the mean of the 0.1 and 0.9 quantile of the squared Euclidean distance of the predictors within “caret” [37,38]. The default cost for tuning within “caret” were 0.25, 0.5 and 1.

2.5.4. K-nearest neighbours (KNN)

K-nearest neighbours (KNN) simply predicts a new sample using the K-closest samples from the training set. Due to this nature, KNN construction is solely based on the observations from the training set and therefore cannot be summarised by a model clearly [34]. In classification, the class probability of the new sample is the proportion of K-closest samples’ observed in each class [34]. The default K for tuning within “caret” were 5, 7 and 9.

2.5.5. C5.0

C5.0 (C5) is an advanced version of Quinlan’s C4.5 classification model [34]. Although there is little literature on the improvements C5 contain, the author of the “caret” package unravelled the program source code and found that the improvements within C5 leads to generation of smaller and simpler trees than C4.5, including carrying out a final global pruning procedure that attempts to remove the sub-trees until the error rate exceeds that when there was no pruning [34]. C5.0 is the only model used that does not contain any parameters for tuning.

2.5.6. Random forest (RF)

Random forest (RF) is a robust tree modelling algorithm used widely in statistical analysis and predictive modelling, and can produce accurate models without the need for pre-processing the input data. RF is essentially an ensemble of multiple classification trees, where each tree is built using a bootstrap sample of the input data and a random subset of the top k predictors at each split of the tree, specified by mtry [34]. Each of the classification trees produces a prediction for the class of a new sample as a vote, and the proportion of votes in each class across all the trees within the ensemble is the predicted probability for each class [34]. Due to this ensemble nature of RF, it is difficult to gain an understanding of the relationships between the predictors and the observations; however, it is possible to quantify the impact of the predictors within the ensemble on prediction using the improvement criteria aggregated across the ensemble [34]. The 3 mtry values for tuning each model were calculated within “caret” via caret:::var\_seq using the number of predictors.

2.5.7. Stochastic gradient boosting (GBM)

Stochastic gradient boosting model (GBM) is a type of boosted tree model where for a given loss function (e.g. area under the Receiver Operating Characteristic (AUROC) curve) and classification tree with depth restriction as a weak learner, the algorithm seeks to find an additive model that minimises the loss function with multiple iterations [34]. GBM holds similarity to RF where both are ensembles of tree models, however, the tree models within GBM are built using a randomly selected subset of the input data at each iteration, dependent on previously computed trees. The tree models within GBM also have depth restriction and unequal contribution to the final model, which differ from RF [34]. GBM can have the tendency to over-fit as, although with restriction, regression trees seek to find the optimum model for the given data, and therefore can be low in prediction for new data. As a countermeasure a regularisation parameter, shrinkage, is used to constrain the learning progress [34]. Within “caret”, the default interaction depth and number of trees for tuning were 1, 2, 3 and 50, 100, 150 respectively while shrinkage was kept at a constant value of 0.1.

2.5.8. Extreme gradient boosting (XGB)

Extreme gradient boosting (XGB) is an alternative version of GBM which contains more regularisation parameters within the algorithm to aid the control of over-fitting [34,39], which results in possibly better performance. Within “caret”, the default *eta* was 0.3 or 0.4, maximum depth was 1, 2 or 3, fraction of sampling columns for each tree was 0.6 or 0.8, fraction of sample for each tree was 0.50, 0.75 or 1.00 and number of rounds was 50, 100 or 150 while gamma and minimum child weight was kept at a constant value of 0 and 1 respectively.

2.6. Model building and performance assessment

Model construction using caret:::train() with the default tuning parameters outlined above and 10-fold cross validation on the training set for each of the predictor data sets (see *2.2. Data compilation*) proceeded. For PLSDA, SVM and KNN, the predictors were centred and scaled within caret:::train() using the preProcess option. The constructed models were then tested using the test set.

The performance of classification models was assessed using a variety of metrics: Sensitivity (1) and specificity (2) which describes how much of each class are correctly predicted [34], accuracy (Acc) which describes the overall rate of true predictions for all observations [34], balanced accuracy (BalAcc) which describes accuracy with data skewness considerations [40], area under the Receiver Operating Characteristic curve (AUROC) which can assess how much better the model prediction is over a random guess [34,41] and Kappa. Many of these metrics are based on a combination of aspects of a confusion matrix for the model (Table 1) and for all of these metrics, a higher value indicates better performance.

**Table 1**. An outline of a confusion matrix for a classification model where Class A is chosen as the positive class and Class B is the negative class.

|  |  |  |
| --- | --- | --- |
|  | Observation |  |
| Class A(positive) | Class B(negative) |
| Prediction | Class A(positive) | True Positive (TP) | False Positive (FP) | Positive |
| Class B(negative) | False Negative (FN) | True Negative (TN) | Negative |
|  | True | False | Total |

|  |  |
| --- | --- |
| Sensitivity = (TP) / True | (1) |
| Specificity = (TN) / False | (2) |
| Accuracy = (TP + TN) / Total | (3) |
| Balanced Accuracy = (Sensitivity + Specificity) / 2 | (4) |

Kappa (5) assesses the accuracy aspect of a model with class distribution considerations which can have a value between -1 to 1. However, normally Kappa of a model is a value between 0 and 1, where it is commonly interpreted as follows:

* < 0.20 = poor agreement
* 0.20 – 0.40 = fair agreement
* 0.40 – 0.60 = moderate agreement
* 0.60 – 0.80 = good agreement
* 0.80 – 1.00 = very good agreement

Models with moderate agreement or above are usually considered to be good. However, Kappa is prone to error induced by prevalence [42,43] of the data and therefore as a safety net, Kappa should be considered together with accuracy such that models with high accuracy and Kappa values are ones that are truly in good agreement [42].

|  |  |
| --- | --- |
| Kappa = ((TP + TN) - (True × Positive + False × Negative)) / (1 + (True × Positive + False × Negative)) | (5) |

Matthews correlation coefficient (MCC) was also another metrics considered for analysing the performance of the models based on the confusion metrics. However, although MCC is considered more informative than Kappa in binary classification problems[44], as our predictive models only consisted of Kappa > 0 (see *3.2. Performance of optimal models identified by 10-fold cross-validation*, *3.3. Performance of the models on test set*, Table S1 and S2), and MCC and Kappa generate similar and concordant scores above this threshold, MCC was not added to the assessment.

Across all models constructed, the number of times a modelling algorithm produced a model with Kappa > 0.45 on the test set were counted and the top 3 (SVM, RF and XGB) were selected for validation. The threshold was chosen to be 0.45 as this can still be considered to fall within the moderately accurate classification. Details of the results are described within *3.4. Performance of model validation via y-randomisation*.

2.7. Model validation via y-randomisation and model robustness

Validation of the SVM, RF and XGB models using all the predictor data sets apart from the MP data set (for reason explained within *3.4. Performance of model validation via y-randomisation*) was carried out by random shuffling of the observations before training using base:::sample(). This was repeated three times for SVM, RF and XGB to test the validity of original models. Model robustness (6) is calculated subsequently [45].

|  |  |
| --- | --- |
| Z = (Kappa original training – Average(Kappa y-randomised training)) / SD(Kappa y-randomised training) | (6) |

If the original model was valid, the overall performance of y-randomised models should be greatly reduced in comparison, with an expected measure of performance being close to random. This can be observed by a high Z score, with Z > 3 considered as significant[46]. The variable importance of the predictors of the original models were also calculated using caret:::varImp.

2.8. Model comparison with commercial product and alternative data set

In addition to the model validation process, predictions of the Ames mutagenicity of the curated data was calculated using the commercial mutagenicity categorical model in StarDrop [47] and open source Toxicity Estimation Software Tool (TEST) [48] for comparison. The StarDrop mutagenicity categorical model is based on a range of decision tree algorithms which employs the C4.5 algorithm introduced by Quinlan[47,49]. TEST predicts Ames mutagenicity via the weighted average of predictions form several different cluster models and the estimation based on the three nearest chemical neighbor in the training set of TEST to the predicting molecule[48]. Both of these models were constructed using the same database[16]. A confusion matrix was created for the training set and test set individually.

In order to compare the performance of the top performing models, the StarDrop Ames mutagenicity category model and TEST, mutagenic predictions were made for the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project[14]. The SMILES of the Class A mutagenic chemicals were extracted from the relevant PDF files, and any incomplete SMILES were removed as were molecules without a provided CAS number. The molecules were then processed and compiled as described in 2.1. Data preparation and 2.2. Data compilation. The mutagenicity of the molecules was predicted using the top performing models, the StarDrop Ames mutagenicity categorical model and TEST. As only mutagenic chemicals are available in this dataset (non-mutagens were not published), the performance of these models was compared using sensitivity only.

2.9. OECD QSAR guidelines

In this study, we mapped onto the following recommendation from the OECD QSAR Guidelines for structure-activity modelling (Table 2).

Table 2. OECD QSPR Guidelines and how this work maps onto them

|  |  |
| --- | --- |
| Recommendation | This work |
| A defined end point | Class of mutagen and non-mutagen as provided in the original data collection |
| An unambiguous algorithm | Methods fully described in experimental, results and discussion section |
| A defined domain of applicability | Domain of applicability is defined as described in experimental section |
| Appropriate measures of goodness-of-fit, robustness and predictivity | Measures described and justified in experimental section |
| A mechanistic interpretation, if possible | Interpretation suggested in results and discussion sections |

3. Results

3.1. Data set generation

From the C. Xu’s Ames Data collection [5], a total of 5395 unique molecules were identified. These 5395 molecules were split into a training set of 4402 molecules (2549 mutagens and 1853 non-mutagens) and a test set of 993 molecules (498 mutagens and 495 non-mutagens). A total of 4109 predictors (13 molecular properties and 2 × 2048 bits fingerprint from ECFP and FCFP) were calculated for each molecule (Table 3). Using the training set, near zero variance and highly correlated predictors were removed to give the final number of predictors for each data set as shown in Table 3.

**Table 3.** The number of predictors generated within Pipeline pilot and after data pre-processing

|  |  |  |  |
| --- | --- | --- | --- |
| **Data set** | **Original number of predictors** | **Variation 1#** | **Variation 2#** |
| ECFP | 2048 | 1425 | 196 |
| FCFP | 2048 | 788 | 138 |
| MP | 13 | 8 | 8 |
| ECFP+FCFP | 4096 | 2200 | 337 |
| MP+ECFP | 2061 | 1433 | 204 |
| MP+FCFP | 2061 | 796 | 146 |
| MP+ECFP+FCFP | 4109 | 2208 | 345 |

# See 2.4. Data pre-processing for detail

3.2. Performance of optimal models identified by 10-fold cross validation

Within this study, a total of 112 binary classification models were constructed using the combination of eight algorithms with the two variations of predictor set for each of the seven predictor data sets (see *2.2. Data compilation*). All the models found showed good performance of AUROC > 0.7, specificity > 0.5, sensitivity > 0.7, balanced accuracy > 0.6 and accuracy > 0.6. 45 out of the 112 constructed models also have Kappa > 0.7 and accuracy > 0.85, which shows they are in good agreement. On the other hand, most models constructed with the variation 1 of the predictor sets are seen to have a slightly better performance than the corresponding model built with the variation 2 of the predictor sets (differences: AUROC +0.022 ± 0.033, specificity +0.025 ± 0.045, sensitivity +0.025 ± 0.044, balanced accuracy +0.025 ± 0.041, accuracy +0.025 ± 0.042, Kappa +0.051 ± 0.084). The detailed performances of these models are given in Table S1.

3.3. Performance of the models on test set

All of the 112 constructed models were assessed using the test set, with the detailed performance of these models given in Table S2. In comparison to the training set, the performance on the test set was seen to decrease for each model, especially in Kappa (-0.285 ± 0.116), sensitivity (-0.221 ± 0.052) and AUROC (-0.149 ± 0.043). Although the Kappa metric for these models dropped notably in general, 44 of the 112 models still had Kappa > 0.45 and accuracy > 0.7 on the test set (Table 4 and S2). In particular, most models constructed with the MP descriptors had very poor performance where specificity was less than 0.5, sensitivity less than 0.5 and balanced accuracy between 0.5 – 0.7. Again, the models constructed with the variation 1 of the predictor sets are seen to have approximately the same performance generally (AUROC +0.002 ± 0.027, specificity -0.005 ± 0.046, sensitivity +0.010 ± 0.038, balanced accuracy +0.002 ± 0.024, accuracy +0.002 ± 0.024, Kappa +0.004 ± 0.048) as the corresponding model built with the variation 2 of the predictor sets. With the decrease in performance on the test set in comparison to the training set, concerns of possible overfitting was considered. However, this possibility was reduced via two approaches taken throughout the model construction proves. First, the training and test split was carried out in a fashion which allows the test set to cover all chemical structural domain the entire database covers. Secondly, the integrated model tuning during the 10-fold cross validation at the model construction phase which identifies the model kernel parameters with the best predictive performance and least overfitting.

**Table 4.** The number of times a modelling algorithm produced a model with Kappa > 0.45 on the test set

|  |  |
| --- | --- |
| **Modelling algorithm** | **Number of models with Kappa > 0.45 on the test set** |
| PLSDA | 3 |
| MDA | 2 |
| SVM | 11 |
| KNN | 0 |
| RF | 11 |
| C5 | 0 |
| GBM | 6 |
| XGB | 11 |

3.4. Performance of model validation via y-randomisation and model robustness

Y-randomisation was performed on the models with Kappa > 0.45 on the test set (Table 4. (SVM, RF and XGB)). Kappa was selected as the focusing analysis metric as the accuracy of the models with Kappa > 0.45 also had moderate to good performance when looking at the other metrics (Table S1 and S2). As the MP data set has been identified to have poor performance when used to construct models alone, y-randomisation was not performed for the models which were constructed using only the MP descriptors. As expected, the performance of the y-randomised models for SVM, RF and XGB was reduced greatly for both the 10-fold cross validation and test set, verifying that the performance of the models is much better than random (Figure 1, S1 – S12). This is supported by the high Z scores of the models (Table 5). It is to note that some SVM models have Z score of infinity due to all predictions of the y-randomised models during training were in the positive class, leading to the denominator of Z score, standard deviation of the training y-randomised Kappa, being zero.

|  |
| --- |
| (**a**) |
| (b) |

**Figure 1.** Kappa of the selected classification models 1 using (a) the variation 1 of the predictor sets; (b) variation 2 of the predictor sets. 1 ECFP extended connectivity fingerprint, FCFP functional class fingerprint, MP molecular properties, RF random forest, SVM support vector machine, XGB extreme gradient boosting. 2 The average test set Kappa of the repeats.

**Table 5.** Z scores of the selected classification models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor data set** | **Predictor variation**# | **Model** | **Z Score** | **Predictor data set** | **Predictor variation**# | **Model** | **Z Score** |
| ECFP | 1 | SVM | Inf | ECFP | 2 | SVM | Inf |
| ECFP | 1 | RF | 54.96 | ECFP | 2 | RF | 93.75 |
| ECFP | 1 | XGB | 69.88 | ECFP | 2 | XGB | 23.08 |
| FCFP | 1 | SVM | Inf | FCFP | 2 | SVM | 144.64 |
| FCFP | 1 | RF | 124.43 | FCFP | 2 | RF | 43.67 |
| FCFP | 1 | XGB | 42.22 | FCFP | 2 | XGB | 25.62 |
| ECFP+FCFP | 1 | SVM | Inf | ECFP+FCFP | 2 | SVM | Inf |
| ECFP+FCFP | 1 | RF | 54.24 | ECFP+FCFP | 2 | RF | 39.90 |
| ECFP+FCFP | 1 | XGB | 11.29 | ECFP+FCFP | 2 | XGB | 71.20 |
| MP+ECFP | 1 | SVM | Inf | MP+ECFP | 2 | SVM | 104.18 |
| MP+ECFP | 1 | RF | 142.39 | MP+ECFP | 2 | RF | 176.00 |
| MP+ECFP | 1 | XGB | 22.12 | MP+ECFP | 2 | XGB | 93.11 |
| MP+FCFP | 1 | SVM | Inf | MP+FCFP | 2 | SVM | 209.77 |
| MP+FCFP | 1 | RF | 148.82 | MP+FCFP | 2 | RF | 148.34 |
| MP+FCFP | 1 | XGB | 40.25 | MP+FCFP | 2 | XGB | 38.50 |
| MP+ECFP+FCFP | 1 | SVM | Inf | MP+ECFP+FCFP | 2 | SVM | 22.78 |
| MP+ECFP+FCFP | 1 | RF | 50.91 | MP+ECFP+FCFP | 2 | RF | 124.69 |
| MP+ECFP+FCFP | 1 | XGB | 101.95 | MP+ECFP+FCFP | 2 | XGB | 140.75 |

# See 2.4. Data pre-processing for detail Inf - infinity

3.5. Variable importance of selected models

The variable importance of the models that successfully passed the y-randomisation validation assessment was calculated and is summarised in Table 6. From this, we can conclude MP predictors are often important in contributing toward constructing a good predictive model, followed by ECFP and closely FCFP. However, we must bear in mind that the MP predictors were never used alone in the models analysed as MP predictors alone produces poor predictive models. This suggests MP predictor require additional structural information from the fingerprint predictors to provide vital information for a good predictive model. A detailed summary of the average variable importance per predictor is given in Table S3.

**Table 6.** The overall importance of each predictor type within the SVM, RF and XGB models

|  |  |  |
| --- | --- | --- |
| **Predictor type** | **Number of models present in** | **Overall importance 1** ± SD |
| MP2 | 6 | 51.8 ± 12.0 |
| ECFP | 8 | 21.6 ± 1.7 |
| FCFP | 8 | 21.8 ± 2.6 |

1 sum of average importance (Table S3) divided by total number of predictors within each predictor type
2individual predictors are shown in Table 9

3.6. Performance comparion of top performing models with commercial product

When comparing the average Kappa of the selected classification models verified via y-randomisation and Z score with the Kappa from StarDrop and TEST (**Table 7**), it is clear that there is a clear drop in performance for the selected classification models between the training set and test set. The decrease in performance also existed for TEST, however this is not as prominent. On the other hand, StarDrop shows comparable performance across the training and test set. The classification models are constructed using the training set and assessed using the test set, while 4657 of the 6512 molecules involved in the overall construction process of the StarDrop model and TEST were included in the training and test set. One note needed to be taken is that for across the training and test set, 202 and 102 molecules did not have a prediction (predict result = N/A).

**Table 7.** Kappa of the selected classification models and StarDrop Ames mutagenicity category model on the training and test set

|  |  |  |
| --- | --- | --- |
| **Model** | **Training set** | **Test set** |
| SVM1 | 0.76 ± 0.06 | 0.48 ± 0.02 |
| RF1 | 0.98 ± 0.03 | 0.52 ± 0.04 |
| XGB1 | 0.76 ± 0.02 | 0.49 ± 0.03 |
| StarDrop | 0.69 | 0.62 |
| TEST | 0.70 (0.62)2 | 0.57 (0.41) 2 |

1average Kappa (Table S1, S2) of the selected classification models and their standard deviation

2Kappa within the bracket calculated for the unpredicted molecules counted as falsely predicted (i.e. unpredicted mutagens as false negative and unpredicted non-mutagens as false positive)

Therefore, the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project[14] was used as a fairer comparison between the selected classification models, the StarDrop model and TEST. The Class A mutagenic chemicals were published much more recently and therefore are more unlikely to be involved in the construction process of the StarDrop model and TEST. After data curation and cross-checking against the molecules involved in model construction of our models, StarDrop’s model and TEST, 508 mutagens were extracted from Ames/QSAR International Challenge Project[14] and used for comparison. From **Table 8**, we can note that our selected classification models are marginally better performing than both the StarDrop model and TEST. Here again, TEST failed to predict 71 molecules of the extracted Class A mutagenic chemical. However, we have to bear in mind that this only captures the ability of the models to predict mutagens correctly as no non-mutagens are involved in this analysis.

**Table 8.** Sensitivity of the selected classification models and StarDrop Ames mutagenicity category model on the Class A mutagenic chemicals from the Ames/QSAR International Challenge Project[14]

|  |  |
| --- | --- |
| **Model** | **Sensitivity** |
| SVM1 | 0.62 ± 0.01 |
| RF1 | 0.62 ± 0.02 |
| XGB1 | 0.61 ± 0.02 |
| StarDrop | 0.56 |
| TEST | 0.57 (0.49) 2 |

1average sensitivity (Table S1, S2) of the selected classification models and their standard deviation

2Sensitivity within the bracket calculated for the unpredicted molecules counted as falsely predicted (i.e. false negatives)

4. Discussion

4.1. Comparison of different predictor sets used in model building

From the results of the 10-fold cross validation and external validations (Figure 1, Table S1 and S2), we can conclude that the MP predictors are not sufficient alone to construct a good model using most of the algorithms. This is expected as the number of MP predictors is very small and the distribution of all 8 molecular properties, including Lipinski’s Rule of Five parameters [50,51] such as number of hydrogen bond donors and LogD, overlap greatly (Figure 2), so a simple set of descriptors would not be able to distinguish mutagen from non-mutagen in general. On the other hand, when the MP predictors are used in combination with fingerprint predictors, they can have a relatively high variable importance (Table 6 and S3). In particular, the molecular solubility and molecular surface area have high average variable importance of 69.6 ± 31.2 and 62.11 ± 25.2 respectively whereas the lowest average importance of the MP predictors is the number of hydrogen bond donors (30.8 ± 26.6) (Table 9). This suggest that although the MP predictors alone are not sufficient, they can still contribute towards the construction of a good model.

|  |
| --- |
|  |

**Figure 2.** Distribution of the 8 retained predictors on the training set from the MP predictor data set.

**Table 9.** The overall importance the MP predictors within the SVM, RF and XGB models

|  |  |  |
| --- | --- | --- |
| **Predictors** | **Average importance** | **SD** |
| LogD | 59.8 | 20.8 |
| Molecular polar solvent accessible surface area | 52.4 | 25.1 |
| Molecular solubility | 69.6 | 31.2 |
| Molecular surface area | 62.1 | 25.2 |
| Number of aromatic rings | 37.5 | 32.2 |
| Number of hydrogen bond donors | 30.8 | 26.6 |
| Number of rings | 54.4 | 44.3 |
| Number of rotatable bonds | 47.8 | 32.4 |

During the 10-fold cross validation, the ECFP predictors are seen to have a slightly better performance (+0.01 ±0.13) than the FCFP predictors in four out of six performance metrics on average when using the same algorithm, whether alone or in combination with the MP predictors. ECFP is a circular fingerprint which represent molecular structures by circular atom neighbourhoods, defined by the fingerprint diameter, while FCFP is a variation of ECFP which generalises the atoms of a molecule by their functional classes [31]. As the fingerprints are generated using the circular atom neighbourhoods, it is able to capture any novel substructures.

For 16 out of 32 models, FCFP is seen to have a slightly better performance (+0.01 ±0.15) than ECFC in four out of six performance metrics on average when using the same algorithm, whether alone or in combination with the MP predictors.

From the model performance (Table S1 and S2), variable importance (Table 6) and the above analysis, we can rate the predictor set as MP < <ECFP ≤ FCFP for the amount of information they each contain that is crucial to mutagenicity prediction (Table 10). It was also noted that by combining the predictor sets, better performing models were constructed. Together with the fact that the variation 1 of the predictor sets results in models with slightly better performance, we can conclude that increased numbers of predictors allow the generation of models which perform better.

**Table 10.** Summary of predictor type performance

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor type** | **Number of models with Kappa > 0.7 in 10-fold cross validation** | **Number of models with Kappa > 0.45 on either or both validation sets** | **Overall importance 1** ± SD |
| MP | 24 | 21 | 51.8 ± 12.0 |
| ECFP | 30 | 31 | 21.6 ± 1.7 |
| FCFP | 28 | 32 | 21.8 ± 2.6 |

Colour code: green highest, orange middle, red lowest. 1 sum of average importance (Table S3) divided by total number of predictors within each predictor type on models selected for y-randomisation validation

4.2. Comparison of different algorithms used for predicting mutagenicity

Out of the eight algorithms (PLSDA, MDA, SVM, KNN, C5, RF, GBM and XGB) used within this study, RF was seen to have the best performance with the test set across all data sets (ECFP, FCFP, MP, ECFP+FCFP, MP+ECFP, MP+FCFP, MP+ECFP+FCFP) when using the variation 1 of the predictor sets and in four out of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.70, sensitivity > 0.65, specificity > 0.65, balanced accuracy > 0.65, Kappa > 0.45).

RF is a robust tree modelling algorithm used widely in statistical analysis and predictive modelling. RF is an ensemble of multiple classification trees, and due to this nature, it can be difficult to gain an understanding of the relationships between the predictors and the observations. However, it is possible to quantify the impact of the predictors within the ensemble on prediction using the improvement criteria aggregated across the ensemble [34]. As mentioned previously, due to the poor performance of the MP predictors alone, the models with MP predictors only were excluded from the analysis. Upon inspection of the variable importance of the nine best performing models (Table 11 and S3), we can identify that one to five predictors with importance of over 70. It is clear that when MP predictors are included (Model No. 5 – 9), molecular surface area, number of rings and logD are seen to have high importance in all cases, closely followed by molecular solubility. It is to note that in these models, the molecular fingerprints generally do not have importance over 70. Although the molecular fingerprints do not have high importance, as mentioned previously, the MP predictors alone do not give good models. Therefore, even though the importance is not high, the molecular fingerprints must hold some importance in the models.

**Table 11.** The variable importance of the selected best performing RF models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model No.** | **Predictor data set** | **Predictor variation**# | **Important predictors** |
| 1 | ECFP | 1 | ECFP bit: 925 |
| 2 | ECFP | 2 | ECFP bit: 1069 |
| 3 | FCFP | 1 | FCFP bits: 466, 870 and 1966 |
| 4 | ECFP+FCFP | 1 | ECFP bit: 925 |
| 5 | MP+ECFP | 1 | ECFP bit: 925, molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD |
| 6 | MP+ECFP | 2 | Molecular solubility, molecular surface area, number of rings and logD |
| 7 | MP+FCFP | 1 | Molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD |
| 8 | MP+ECFP+FCFP | 1 | Molecular solubility, molecular surface area, molecular polar surface area, number of rings and logD |
| 9 | MP+ECFP+FCFP | 2 | Molecular surface area, number of rings and logD |

# See 2.4. Data pre-processing for detail

Aside from the best performing RF models, the performance of the other algorithms was looked at in attempt to aid the understanding from the RF models. XGB was found to have the best performance with the test set across all data sets (ECFP, FCFP, ECFP+FCFP, MP+ECFP, MP+FCFP, MP+ECFP+FCFP) when using the variation 1 of the predictor sets and in three out of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.65, sensitivity > 0.65, specificity > 0.60, balanced accuracy > 0.65, Kappa > 0.45). On the other hand, SVM was found to have the best performance in four of seven data sets when using the variation 2 of the predictor sets (AUROC > 0.75, sensitivity > 0.65, specificity > 0.75, balanced accuracy > 0.70, Kappa > 0.45). Although the performance of these models does not look as good as the 10-fold cross validation training data and are not to the desired level (e.g. AUROC > 0.9), they are still good models as all the metrics are closely matching and the Kappa signifies fair to moderate agreement.

SVM is an algorithm widely used for predictive modelling as it has shown to have great capability of fitting non-linear relationships within the pharmaceutical industry [52-54]. However, as the predictors within SVM models are transformed by the radial basis kernel function, it is the type of model where the analysis of predictor – observation relationship can only be achieved via the number of times a predictor appears in the model. Therefore, the use of SVM models to identify the direct predictor – observation relationship poses some difficulty. Nonetheless, upon close inspection of the variable importance of the four top performing SVM models (Table 12 and S3), we can identify that 7 to 65 predictors which have importance of over 70. In particular, number of rotatable bonds and molecular solubility have high importance in both of the models which used MP predictors (Model No. 13 and 14), whereas only >20% of the used fingerprint predictors in such models have an importance of over 70.

**Table 12.** The variable importance of the selected best performing SVM models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model No.** | **Predictor data set** | **Predictor variation**# | **Important predictors** |
| 10 | ECFP | 2 | 7 ECFP bits |
| 11 | FCFP | 2 | 10 FCFP bits |
| 12 | MP+FCFP | 2 | 27 FCFP bits, molecular solubility, number of rotatable bonds |
| 13 | MP+ECFP+FCFP | 2 | 36 ECFP bits, 27 FCFP bits, molecular solubility, number of rotatable bonds |

# See 2.4. Data pre-processing for detail

XGB is a relatively new modelling algorithms which holds great potential in tackling machine learning problems [39]. As a tree-based algorithm, it would be possible to derive some rules for the underlying predictor – observation relationship; however, this is not a straightforward task. Upon inspection of the variable importance of the eight top performing XGB models (Table 13 and S3), we can identify that number of rings and number of aromatic rings both have importance of 100 in two out of the 10 models (Model No. 20 and 21), whereas for molecular fingerprints, ECFP bit: 925 and FCFP bit: 870 have both made their appearance in four and three of the analysing models respectively, with importance over 70.

**Table 13.** The variable importance of the selected best performing XGB models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model No.** | **Predictor data set** | **Predictor variation**# | **Important predictors** |
| 14 | ECFP | 1 | ECFP bit: 925 |
| 15 | FCFP | 1 | FCFP bit: 870, 1226 and 1966  |
| 16 | ECFP+FCFP | 1 | ECFP bit: 925 |
| 17 | ECFP+FCFP | 2 | FCFP bit: 870 and 2007 |
| 18 | MP+ECFP | 1 | ECFP bit: 925 and number of rings |
| 19 | MP+ECFP | 2 | Number of rings |
| 20 | MP+FCFP | 1 | FCFP bit: 870 and number of aromatic rings |
| 21 | MP+ECFP+FCFP | 1 | ECFP bit: 925 and number of aromatic rings |

# See 2.4. Data pre-processing for detail

Nonetheless, as the performance of the constructed XGB models do not match the desired values, especially when looking at Kappa (Kappa > 0.6), we decided not to attempt to examine further the predictor – observation relationship.

4.3. Ames prediction of Class A mutagenic chemicals

From **Table 7**, we can see that our constructed models have a better chance in correctly predicting an Ames mutagenic compound than the StarDrop model and TEST. This then brought our interest to if our models can correctly predict the mutagenicity of the Class A mutagenic chemical which one or less models within the Ames/QSAR International Challenge Project[14] correctly predicted. Out of the 36 curated Class A mutagenic chemical which were correctly predicted by one or fewer models within the Ames/QSAR International Challenge Project, our models can correctly predict the mutagenicity of 28 of them (**Table S4**). In comparison, the StarDrop model and TEST can only each correctly predict six compounds. Again, here TEST fails to produce prediction for six of the molecules. This further shows the capability of our models to correctly predict mutagens, possibly due to the different chemical space the training set covers in comparison to the StarDrop model and models within the Ames/QSAR International Challenge Project. However, as the specificity of models are also important due to the possibility of predicting a strong drug candidate incorrectly as mutagen, similar analysis using non-mutagens will be necessary.

5. Conclusions

Within this work, we attempted to build models with comparable performance with the models *Xu et al.* presented using a different range of descriptors and modelling algorithms. During the process, we discovered that the library *Xu et al.* presented in their work contained duplicates, and therefore we curated the library identifying 5395 unique molecules. As this resulting library is different to the library *Xu et al.* used for their work, comparison of model performance against their work was not carried out. After constructing 112 models using eight different algorithms, we discovered SVM, RF and XGB to have the best performance. The RF models had the best performance across most data sets during the 10-fold cross validation training (AUROC > 0.95, sensitivity > 0.90, specificity > 0.95, balanced accuracy > 0.95, Kappa > 0.90) and on the test set (AUROC > 0.75, sensitivity > 0.75, specificity > 0.65, balanced accuracy > 0.70, Kappa > 0.40); the SVM and XGB models had good performance during the 10-fold cross validation (AUROC > 0.90, sensitivity > 0.85, specificity > 0.75, balanced accuracy > 0.80, Kappa > 0.65) and on the test set (AUROC > 0.65, sensitivity > 0.65, specificity > 0.60, balanced accuracy > 0.65, Kappa > 0.30).

With the RF, SVM and XGB models, we discovered that the MP descriptors showed the highest importance when used in combination with molecular fingerprints. Such descriptors are logD, molecular solubility, molecular surface area, number of aromatic rings, number of rings and number of rotatable bonds, with their importance differing with different modelling algorithms.

By comparing the performance of our top performing models against the StarDrop Ames mutagenicity prediction model and TEST using the Class A mutagenic compounds from the Ames/QSAR International Challenge Project, we found our models to be better performing in predicting mutagens correctly. We also discovered that our models were able to predict some of the Class A mutagenic compounds where one or less models were able to predict correctly in the Ames/QSAR International Challenge Project. Attempts to improve the performance and robustness of the models would involve searching for more experimental Ames data and extend the variation of modelling algorithms used, as well as the range of molecular descriptors calculated.

**Supplementary Materials:** The following are available online at www.mdpi.com/xxx/s1, Table S1: 10-fold cross validation performance, Table S2: External validation sets performance, File S1: SD file of the training set, File S2: SD file of the test set, File S3: (Figure S1 – S12) ROC curve comparison of original models and y-randomised models, Table S3: The average variable importance for each predictor used within the models selected for y-randomisation variation, Table S4: The prediction probability of Class A mutagenic compounds with one or less correct prediction within the Ames/QSAR International Challenge Project.

**Author Contributions:** Conceptualization, C. C. and N. G. B.; Methodology, C. C. and N. G. B.; Formal Analysis, C. C. ; Data Curation, C. C. ; Writing-Review & Editing, C. C. , J. D. S. P. O. N. and N.G.B. ; Supervision, N. G. B. and P. O. N.; Funding Acquisition, N. B..

**Funding:** The research was funded by EPSRC DTA award with a CASE contribution from Unilever.

**Acknowledgments:** We gratefully acknowledge the financial contribution from Unilever for the studentship for C. C.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Custer, L.L.; Sweder, K.S. The role of genetic toxicology in drug discovery and optimization. *Curr. Drug Metab.* **2008**, *9*, 978-985.

2. Kramer, J.A.; Sagartz, J.E.; Morris, D.L. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nature reviews. Drug discovery* **2007**, *6*, 636-649.

3. Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* **2005**, *48*, 312-320.

4. White, A.C.; Mueller, R.A.; Gallavan, R.H.; Aaron, S.; Wilson, A.G.E. A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **2003**, *539*, 77-89.

5. Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico prediction of chemical ames mutagenicity. *Journal of Chemical Information and Modeling* **2012**, *52*, 2840-2847.

6. Ames, B.N.; Durston, W.E.; Yamasaki, E.; Lee, F.D. Carcinogens are mutagens: A simple test system combining liver homogenates for activation and bacteria for detection. *Proceedings of the National Academy of Sciences of the United States of America* **1973**, *70*, 2281-2285.

7. Ames, B.N.; Gurney, E.G.; Miller, J.A.; Bartsch, H. Carcinogens as frameshift mutagens: Metabolites and derivatives of 2-acetylaminofluorene and other aromatic amine carcinogens. *Proceedings of the National Academy of Sciences* **1972**, *69*, 3128-3132.

8. Ames, B.N.; Lee, F.D.; Durston, W.E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National Academy of Sciences of the United States of America* **1973**, *70*, 782-786.

9. McCann, J.; Spingarn, N.E.; Kobori, J.; Ames, B.N. Detection of carcinogens as mutagens: Bacterial tester strains with r factor plasmids. *Proceedings of the National Academy of Sciences of the United States of America* **1975**, *72*, 979-983.

10. Mortelmans, K.; Zeiger, E. The ames salmonella/microsome mutagenicity assay. *Mutation research* **2000**, *455*, 29-60.

11. Walmsley, R.M.; Billinton, N. How accurate is in vitro prediction of carcinogenicity? *British Journal of Pharmacology* **2011**, *162*, 1250-1258.

12. Benigni, R.; Giuliani, A. Computer‐assisted analysis of interlaboratory ames test variability. *Journal of Toxicology and Environmental Health* **1988**, *25*, 135-148.

13. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Kovalishyn, V.V.; Prokopenko, V.V.; Tetko, I.V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics* **2010**, *24*, 202-208.

14. Honma, M.; Kitazawa, A.; Cayley, A.; Williams, R.V.; Barber, C.; Hanser, T.; Saiakhov, R.; Chakravarti, S.; Myatt, G.J.; Cross, K.P.*, et al.* Improvement of quantitative structure-activity relationship (qsar) tools for predicting ames mutagenicity: Outcomes of the ames/qsar international challenge project. *Mutagenesis* **2019**, *34*, 3-16.

15. Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P.W.; Tang, Y. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *Journal of chemical information and modeling* **2011**, *51*, 2482-2495.

16. Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling* **2009**, *49*, 2077-2081.

17. Votano, J.R.; Parham, M.; Hall, L.H.; Kier, L.B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus qsar models for the prediction of ames genotoxicity. *Mutagenesis* **2004**, *19*, 365-377.

18. Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central journal* **2010**, *4 Suppl 1*, S2.

19. Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics (Oxford, England)* **2006**, *22*, 2099-2106.

20. Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. Comparative evaluation of in silico systems for ames test mutagenicity prediction: Scope and limitations. *Chemical research in toxicology* **2011**, *24*, 843-854.

21. Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* **2005**, *48*, 312-320.

22. Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central journal* **2010**, *4*, S2.

23. Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics (Oxford, England)* **2006**, *22*, 2099-2106.

24. Willett, P.; Barnard, J.M.; Downs, G.M. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983-996.

25. Ewing, T.; Baber, J.C.; Feher, M. Novel 2d fingerprints for ligand-based virtual screening. *Journal of chemical information and modeling* **2006**, *46*, 2423-2431.

26. Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. Adme evaluation in drug discovery. 10. Predictions of p-glycoprotein inhibitors using recursive partitioning and naive bayesian classification techniques. *Molecular Pharmaceutics* **2011**, *8*, 889-900.

27. Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. Admet evaluation in drug discovery. 12. Development of binary classification models for prediction of herg potassium channel blockage. *Molecular Pharmaceutics* **2012**, *9*, 996-1010.

28. Inc., A.S. *Chemistry collection: Basic chemistry user guide, pipeline pilot release 16.5.0.143*, Accelrys Software Inc.: San Diego, 2016.

29. Tropsha, A. *Best practices for qsar model development, validation, and exploitation*. 2010; Vol. 29.

30. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. Inchi, the iupac international chemical identifier. *Journal of Cheminformatics* **2015**, *7*, 23.

31. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742-754.

32. Inc., A.S. *Pipeline pilot statistics and data modelling collection 2016, pipeline pilot release 16.5.0.143*, Accelrys Software Inc.: San Diego, 2016.

33. R: What is r? <https://www.r-project.org/about.html> (07/07/2017),

34. Kuhn, M.; Johnson, K. *Applied predictive modeling*. Springer: 2013; Vol. 26.

35. Kuhn, M. The caret package. <http://topepo.github.io/caret/index.html> (08/11/2016),

36. Pirhadi, S.; Shiri, F.; Ghasemi, J.B. Multivariate statistical analysis methods in qsar. *RSC Advances* **2015**, *5*, 104635-104665.

37. Karatzoglou, A. Sigest - hyperparameter estimation for the gaussian radial basis kernel. <https://www.rdocumentation.org/packages/kernlab/versions/0.9-26/topics/sigest> (05/09/2018),

38. Kuhn, M. Svmradial.R. <https://github.com/topepo/caret/blob/master/models/files/svmRadial.R> (05/09/2018),

39. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM: San Francisco, California, USA, 2016; pp 785-794.

40. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. In *The balanced accuracy and its posterior distribution*, 2010 20th International Conference on Pattern Recognition, 23-26 Aug. 2010, 2010; pp 3121-3124.

41. Hallinan, J. Assessing and compairing classifier performance with roc curves. <http://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/> (07/07/2017),

42. Carpenter, B. High kappa values are not necessary for high quality corpora. <https://lingpipe-blog.com/2012/10/02/high-kappa-not-necessary-high-quality-corpora/> (07/07/2017),

43. Eugenio, B.D.; Glass, M. The kappa statistic: A second look. *Computational Linguistics* **2004**, *30*, 95-101.

44. Chicco, D.; Warrens, M.J.; Jurman, G. The matthews correlation coefficient (mcc) is more informative than cohen’s kappa and brier score in binary classification assessment. *IEEE Access* **2021**, *9*, 78368-78381.

45. Luo, M.; Wang, X.S.; Roth, B.L.; Golbraikh, A.; Tropsha, A. Application of quantitative structure–activity relationship models of 5-ht1a receptor binding to virtual screening identifies novel and potent 5-ht1a ligands. *Journal of Chemical Information and Modeling* **2014**, *54*, 634-647.

46. Rodgers, A.D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure−activity relationship method. *Chemical research in toxicology* **2010**, *23*, 724-732.

47. *Stardrop*, Optibrium Ltd: United Kingdom, 2011.

48. *User's fuide for t.E.S.T. (toxicity estimation software tool)*, 5.1; U.S. Environmental Protection Agency: United States, 2020.

49. *Toxicity prediction in stardrop*, Optibrium Ltd: United Kingdom, 2011.

50. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1pii of original article: S0169-409x(96)00423-1. The article was originally published in advanced drug delivery reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46*, 3-26.

51. Lipinski, C.A. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337-341.

52. Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N.P.E.; Olsen, L.; Jorgensen, F.S. Classification of cytochrome p450 1a2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37*, 658-664.

53. Zhao, C.Y.; Zhang, H.X.; Zhang, X.Y.; Zhang, R.S.; Luan, F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Prediction of milk/plasma drug concentration (mip) ratio using support vector machine (svm) method. *Pharm. Res.* **2006**, *23*, 41-48.

54. Zhou, S.; Li, G.B.; Huang, L.Y.; Xie, H.Z.; Zhao, Y.L.; Chen, Y.Z.; Li, L.L.; Yang, S.Y. A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (svm) method. *Comput. Biol. Med.* **2014**, *51*, 122-127.