

Integrated in silico approaches for the prediction of Ames test mutagenicity

Sandeep Modi · Jin Li · Sophie Malcomber ·
Claire Moore · Andrew Scott · Andrew White ·
Paul Carmichael

Received: 16 November 2011 / Accepted: 9 August 2012 / Published online: 24 August 2012
© Springer Science+Business Media B.V. 2012

Abstract The bacterial reverse mutation assay (Ames test) is a biological assay used to assess the mutagenic potential of chemical compounds. In this paper approaches for the development of an in silico mutagenicity screening tool are described. Three individual in silico models, which cover both structure activity relationship methods (SARs) and quantitative structure activity relationship methods (QSARs), were built using three different modelling techniques: (1) an in-house alert model: which uses SAR approach where alerts are generated based on experts judgements; (2) a kNN approach (k-Nearest Neighbours), which is a QSAR model where a prediction is given based on outcomes of its k chemical neighbours; (3) a naive Bayesian model (NB), which is another QSAR model, where a prediction is derived using a Bayesian formula through preselected identified informative chemical features (e.g., physico-chemical, structural descriptors). These in silico models, were compared against two well-known alert models (DEREK and ToxTree) and also against three different consensus approaches (Categorical Bayesian Integration Approach (CBI), Partial Least Squares Discriminate Analysis (PLS-DA) and simple majority vote approach). By applying these integration methods on the validation sets it was shown that both integration models (PLS-DA and CBI) achieved better performance than any of the individual models or consensus obtained by simple

majority rule. In conclusion, the recommendation of this paper is that when obtaining consensus predictions for Ames mutagenicity, approaches like PLS-DA or CBI should be the first choice for the integration as compared to a simple majority vote approach.

Keywords Ames · QSAR · SAR · Admet · In silico models

Introduction

For the safety of new chemicals, an early alerting system for potential genotoxicity is very important. The bacterial reverse mutation assay (Ames test) to detect mutagenicity has widely been used as an early alerting system for potential genotoxicity. This assay was designed to detect and identify genetic damage caused by chemicals in bacterial cells [1–5].

In silico predictive models for genotoxicity fall into two principal categories: rule based (expert systems) and quantitative structure–activity relationship models (QSAR). The first approach is associated with the local reactivity of chemicals, i.e., reactivity of functional groups (also called structural alerts, pharmacophores, or toxicophores). The key step in the development of this approach is defining chemical categories for genotoxicity and defining the organic chemistry associated with the formation of a covalent bond between DNA and an exogenous chemical. In this approach, a well-defined reactive group which has the potential to interact with DNA is highlighted. Several rule based systems have been developed [6–9] which summarise the systematised relationships between chemical substructures and observed mutagenicity outcomes. This technique can be an invaluable tool in the in silico prediction of genotoxicity, as it is very simple and highlights the

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9595-5) contains supplementary material, which is available to authorized users.

S. Modi (✉) · J. Li · S. Malcomber · C. Moore · A. Scott ·
A. White · P. Carmichael
Safety & Environmental Assurance Centre, Unilever, Colworth
Science Park, Sharnbrook, Bedford MK44 1LQ, UK
e-mail: sandeep.modi@unilever.com

presence of certain substructures (toxicophores) within the molecule, which can be related to the Ames test outcome. This alert approach may also provide mechanistic understanding of the Ames test outcome. It should be noted that this approach is generally used to predict a chemical to be Ames positive. If no alert is present it does not mean that the chemical is predicted to be Ames negative [10].

Quantitative-structure activity-relationship (QSAR) models are widely used as a way of approximating a functional relationship between a set of descriptors and a given endpoint. Any QSAR model is an approximation to a relationship between biological activity and compound characteristics and can be viewed as a mathematical function. In contrast to structural-alert based model, one of the advantages with QSAR modelling approach is that it enables the prediction of negative chemicals as well as positive chemicals. Different QSAR and machine learning methods use different ways of deriving these approximations to provide information about the Ames outcome of the chemicals [11–23]. Several papers have been published regarding QSAR models for predictions of Ames test results using Partial Least Squares, Neural network, Random Forests, and Support Vector Machines [11–23]. It has also been noted that model interpretation plays an important role in acceptance of a model, without it the possibility of modifying compounds for designing out a given problem is restricted. Moreover, chemists often resort to model algorithms that are easy to understand, such as linear models. Nonlinear models are commonly viewed as hard to interpret, and these model algorithms or models are sometimes referred to as black box models. The mechanistic interpretation is also one of the principles from the Organisation for Economic Co-operation and Development (OECD), which should be followed whenever possible in order to achieve regulatory acceptance of the predictive chemistry tools. There has to be balance between model interpretation and model accuracy. Depending on endpoint probabilistic and statistic QSAR methods in many cases may provide superior results as compared to simple SAR methods. The modeling algorithm linking molecular descriptors to the output variable needs to be chosen so that it takes the complexity of the particular relationship into account, otherwise over-fitting (in case of too complex a modeling procedure used) or insufficient predictivity of a model (in case of too simple a modeling procedure used) may result.

Existing commercial tools suitable for predicting the outcome of the Ames test, such as DEREK for Windows [24–27], provide an easy interpretation. It is possible to derive structure–activity and/or mechanistic information from each of the predictions. The same is true for ToxTree, which is a flexible open source application tool by the Joint Research Council (JRC) [28].

A recent comprehensive review [29] of different *in silico* models and approaches for predictions of genotoxic

outcome, shows that most of the earlier approaches described for the prediction of Ames mutagenicity, produced good specificity and sensitivity values (prediction accuracy of up to 85 %). Depending on the descriptors and the statistical methods used, some of the models offer simple structure–activity information [16, 21, 25, 30], whilst others are harder to interpret due to the choice of chemical descriptors derived from structural information [12, 20].

In order to compare and integrate these different approaches for better consensus predictions for mutagenicity, a large and clearly defined benchmark data set is needed. This comprehensive database of chemicals can then not only be used for comparison of the different methodologies, but also to improve on existing QSAR models. In view of this, a new benchmark set of 6718 compounds was collected from public sources together with their Ames test results. Three in-house Ames models were then built using different approaches (1) a functional group/structural alert approach, which is associated with reactivity of the particular chemical fragment; (2) a kNN approach, which is based on majority rule of the experimental data of the nearest similar chemicals to the target molecule and (3) a naïve Bayesian classification model which is a typical QSAR approach in which qualitative relationships between a set of descriptors and a given endpoint are explored. Two external models (DEREK and ToxTree) were used for comparison and validation of each individual approach. These external approaches, use a set of rules derived from the collective expertise of toxicologists from academia, industry and government.

The same large validation data set was used throughout this study; therefore it was possible to perform a reasonable comparison of these different methods. This can also highlight the similarities and differences between these different approaches (SAR/QSAR). It is well known that each individual model has its own problems and pitfalls, therefore integration of all these above approaches was carried out to produce consensus predictions. By having large datasets it has been possible to try various integration strategies for producing consensus predictions, and also to compare them with each other. In the first integration approach all models were run in parallel and a final prediction was based on the majority rule by combining the output of different models. In this method, weak and strong models were treated equally and given equal votes, and correlation between different models was not taken into account. In the second method PLS-DA was performed using different individual model predictions as input and a consensus prediction was generated. A third method for integration used was based on Bayesian theory to generate an overall prediction linked with a probability of the result being positive or negative. The Bayesian integration

approach is capable of not only weighting each individual model according to its quality in terms of sensitivity and specificity of the models, but also incorporating the prevalence information via a prior as a parameter in its integration calculation. Validation of these integration approaches on an external set appeared to show that the accuracy of the consensus predictions was superior, and that the confidence of the prediction was better as compared to predictions based on an individual model.

Data and methods

The Ames mutagenicity dataset

In the Ames test, frame-shift mutations or base-pair deletions may be detected by exposure of histidine dependent strains of *Salmonella typhimurium* and/or *Escherichia coli* to a test compound [1–5]. When these strains are exposed to a mutagen, reverse mutations that restore the functional capability of the bacteria to synthesize histidine, enable bacterial colony growth on a medium deficient in histidine. These altered bacteria are referred to as “revertants”. Since many chemicals interact with genetic material only after metabolic activation by enzyme systems not available in the bacterial cell, the test compounds are, in many cases, additionally examined in the presence of a mammalian metabolising system, which contains liver microsomes (S9 mix). A compound is classified Ames positive if it significantly induces revertant colony growth in any of the usual five strains, tested either in the presence or absence of S9 mix. A compound is judged negative if it does not induce significant revertant colony growth in any strain either in the presence or absence of S9 mix. As a consequence of this definition, Ames negative compounds in any dataset which have not been tested in all recommended strains may turn out to cause reverse mutations (i.e. positive) when being examined in additional strains.

A benchmark data set containing 6718 chemicals with Salmonella/Ames test data was selected for this study. The main source of data used was Leadscape (which covers the data from US Food and Drug Administration (FDA) [31]), Centre for Food Additives and Applied Nutrition (CFSAN), Food Additive Resource Management system (FARM), Chemical Carcinogenicity Research Information System (CCRIS) [32], the National Toxicology Programs (NTP) Genetic Toxicology database [33], the Tokyo-Eiken database [34] and historical Unilever data (Pre 2009). Models are clearly only as good as the data they are based on, therefore any conflicting data (3 %), duplicate structures and also a small number of extraordinary large (mw > 1000) and/or inorganic molecules were omitted from the dataset (0.2 %). Chemicals which have been only been examined only in 1 strain were also omitted from the dataset considered for this

study. It should be noted that percentage of mutants in this data set is slightly lower as compared to recent Sushko et al. [62] paper. This could be due to different data sources (e.g. we have included data from Leadscape and historical in-house data whereas Sushko et al. [62] paper includes data from Hansen et al. [51]). Moreover we have only included chemicals with at least two strains tested for this study. In a separate study we are in process of studying SAR for different strains and we hope to release more details about datasets including results about the strains and activation.

In our collective benchmark dataset, there are experimental data which has been generated during lead optimisation and number of chemicals are closely related analogues, therefore structural and chemical diversity amongst members of this large dataset was a key consideration for the choice of the chemical to be in the training set for the development of new in-house QSAR models for Ames genotoxicity. Figure 1 shows the workflow how different datasets used for this study were selected and designed. As shown only a small portion of training set of 1534 compounds (roughly one quarter) was used to build several in silico models and to explore SAR. All compounds were first clustered using Extended Connectivity Fingerprints (ECFP) in Pipeline Pilot using a Tanimoto similarity cut-off of 0.7 for each cluster, and only centroids of each cluster (380 clusters) with at least 5 members were chosen to be in the training set. The rest of the training set was chosen at random from singletons. A small test set (213 chemicals) was also chosen which was used for selecting/tuning any (Q)SAR approaches. The remaining data (4971 compounds) and an internal independent Unilever set (post 2009, 42 compounds) were used as an independent validation set. This validation data was not used in any of the model building, testing and tuning exercises, it was purely used as an independent validation set. Figure 2 shows the break down of number of compounds (mutagen and non-mutagen) in each set used for this study.

Performance criteria for models

To judge the accuracy and predictive power of each method, the following statistical parameters were calculated. These numbers can be easily calculated based on a confusion matrix generated by correlation between experimental and prediction data [35].

Sensitivity

This represents the proportion of compounds correctly predicted to be Ames positive relative to all compounds experimentally determined to be Ames positive $TP / (TP + FN)$.

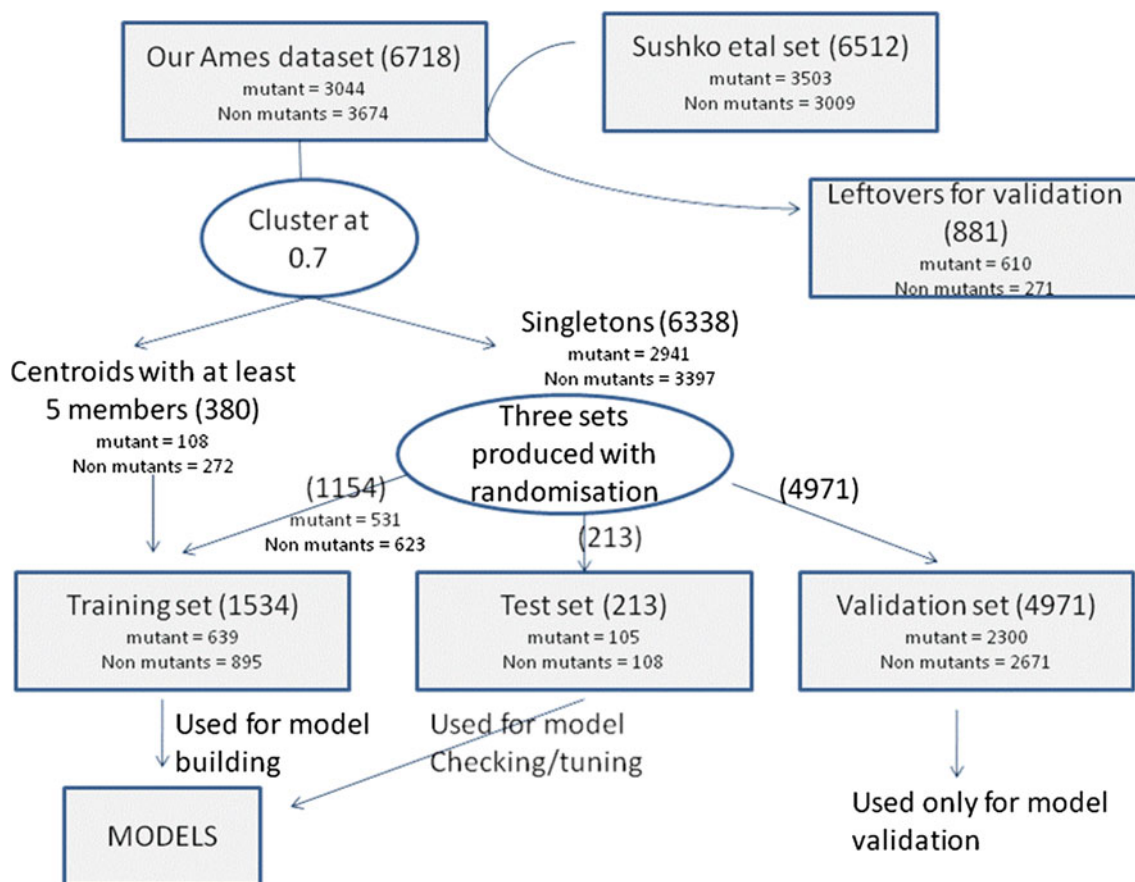
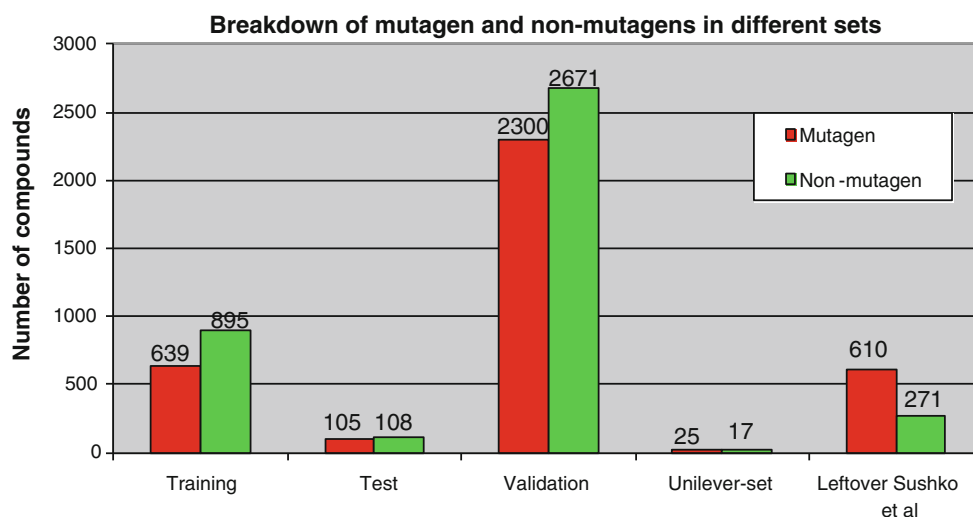


Fig. 1 Workflow to show how different datasets used for this study were selected and designed

Fig. 2 Shows the break down of number of compounds in each set



Specificity

This represents the proportion of compounds correctly predicted to be Ames negative relative to all compounds experimentally determined to be Ames negative $TN / (TN + FP)$.

Total accuracy

This represents the proportion of compounds correctly predicted to be Ames positive and negative relative to total number of predictions $(TP + TN) / (TP + FP + TN + FN)$.

Positive predictivity

This represents proportion of compounds correctly predicted to be Ames positive relative to all predictions categorized as Ames positive $TP/(TP + FP)$.

Negative predictivity

This represents proportion of compounds correctly predicted to be Ames negative relative to all predictions categorized as Ames negative $TN/(TN + FN)$.

In addition to these, the Cohen's kappa coefficient [36] was also calculated; this is a statistical measure of inter-rater agreement for categorical items (mutagen/non-mutagen). It is generally thought to be a more robust measure than simple percent agreement calculation since this takes into account the agreement occurring by chance, but there have been a few concerns over the conservative approach of kappa coefficient [37]. In view of this Matthews' correlation coefficient was also calculated [38] which is used as another measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. There is no single way to describe the confusion matrix of true and false positives and negatives by a single number, however the Matthews correlation coefficient is generally regarded as one of the best measures to describe the accuracy of two class models.

Results

Alert approaches

This approach, often referred to as a structural-alert approach is associated with the local reactivity of chemicals, i.e., reactivity of functional groups also called alerts, pharmacophores, or toxicophores. Due to its simplicity this is a well known and popular approach where well-defined reactive groups are highlighted and mechanistic reasons given for how the presence of particular functional groups/alerts within the chemical could be related to the Ames outcome. Rule based systems for mutagenicity have been built [6–9] where relationships between chemical substructures and observed Ames outcomes were studied. In a recent review [39] a detailed study was conducted to highlight 57 sub-structures (alerts) which could be associated with mutagenicity and genotoxic carcinogenicity. Although if no alert is present it does not mean that the chemical is predicted to be Ames negative, however, for the sake of a quantitative assessment of the usefulness of these alerts, and also to calculate various statistical parameters it is


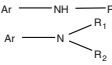


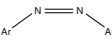




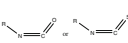

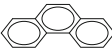
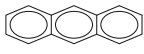
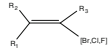



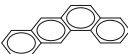
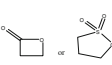

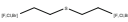

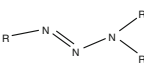
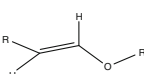
unavoidable to consider the absence of alerts as a negative prediction. The similar strategy approach has been taken by many studies in order to evaluate predictive power of these alerts and for comparing and improving the predictions by combining with other methods [41–44].

In-house alert approach

In order for these structural alerts to operate to their full potential it is vital that they develop in parallel with current knowledge and understanding. The steric and electronic environment surrounding a structural alert fragment can diminish or enhance its genotoxic potency. Expert systems are composed of structural rules derived from specific toxicological mechanisms or plausible modes of action of chemical agents in combination with pattern recognition routines to identify substructures associated with specific toxic effects [24–26]. One of the methods for assessing the performance of such structural alerts is to take these alerts through validation sets. In this paper 86 different structural alerts from different sources for Ames data [6–9, 39] and also for chemical reactivity [40] were collected and integrated. These were then encoded in SMARTS strings [SMARTS], and their frequency was measured in the training set used for this study. It should be noted that for some of the alerts effect of more specific alert was compared within more generalised alert (e.g. in training set 74 % of chemicals with any three membered ring containing N, O and S chemicals are Ames positives; whereas 82 % of chemicals with three membered ring containing only O (i.e. epoxide) are Ames positive). Any alerts which were present in less than 10 chemicals of the training set were not considered for any further analysis. For each particular alert, the percentage of chemicals with a positive Ames outcome was calculated, and only alerts which had more than 70 % of chemicals as Ames positive were selected. Table 1 lists 24 chemical alerts, which were selected based on the number of known examples and their ability to distinguish between Ames outcomes as described above. This is one of the advantages in this approach as not only is it able to highlight the alert/fragment in a chemical but it is also able to give the probability of the chemical to be Ames positive based on the training data. These probabilities can then be used as continuous input descriptors in one of the integration approaches discussed later. As it can be seen many of the alerts/fragments reported in earlier studies [6–9] have not been proven to be important in our training and validation sets, highlighting the need for continual refinement of structural alerts in light of new knowledge.

As discussed above, this approach only works for chemicals being Ames positives, Table 1 also shows the percentages of chemicals with positive Ames in case none

Table 1 The number of chemicals and fraction of Ames positive results for each selected alerts are shown below for the training, test and validation sets

Fragment	Name	Training set		Test/Validation set	
		Fraction of positives	Number of examples	Fraction of positives	Number of examples
	Aromatic amine	0.72	267	0.69	832
	Quat N	0.74	17	0.76	34
	SN2	0.74	51	0.72	167
	Alkyl chloride	0.74	105	0.78	356
	Aromatic-diazo	0.75	37	0.71	102
	Alkyl bromide	0.75	31	0.78	89
	Acyl chloride	0.75	10	0.79	31
	Thiazole	0.78	38	0.78	149
	Aromatic N-acyl	0.78	23	0.81	123
	Isocyanate and isothiocyanate	0.8	15	0.83	41
	Epoxide	0.82	42	0.88	92
	Tri polycyclic 1	0.84	48	0.83	203
	Tri polycyclic 2	0.86	49	0.88	190
	Monohalo alkene	0.87	23	0.89	142
	Aromatic nitro	0.91	136	0.89	173
	Nitroso	0.91	44	0.93	168
	N-nitroso	0.91	34	0.89	87
	Tetra polycyclic	0.94	15	0.96	45
	Propiolactone or propiosultones	0.94	32	0.92	67
	Aromatic_hydroxyl	0.93	14	0.86	14
	Sulphur mustard	1	10	1	12
	a,b-unsaturated	1	11	0.92	10
	Azide	1	11	0.95	20
	a,b-unsaturated aliphatic alkoxy	1	10	0.94	17
	No alert present	0.21	802	0.27	2,663

of the alert was present, as it can be seen if none of the alerts listed in Table 1 were present then almost 21 % of the chemicals in the training set and about 26 % of chemicals in test/validation sets were Ames positive. Table 2 shows the various statistics for training, test and validation sets for the prediction of Ames outcome based just on the in-house alert approach. Although sensitivity/selectivity for this approach are low, it still performs quite well as a first filter for removing genotoxic chemicals. It can also be seen from Fig. 3 that the number of true positives and true negatives is quite consistent in all the sets (around 70 %), except in the case of the test set, true positives are only 56 %; this could be due to low number of compounds in this set.

External Alert approaches

In order to compare and validate the in-house alert approach, DEREK and ToxTree were run with the same sets of chemicals and different statistical parameters were generated in a similar manner on the different datasets mentioned above. DEREK (deductive estimate of risk from

existing knowledge) uses a set of rules derived from the collective expertise of toxicologists from academia, industry and government [24–27]. DEREK does not provide a quantitative assessment of the probability of mutagenicity of a particular compound, but rather provides a series of structural alerts. The mutagenicity prediction by DEREK was considered positive if a structure triggered at least one mutagenicity alert (with either certain, probable or plausible reasoning). In order to evaluate the performance of DEREK, the absence of alert (i.e. “nothing to report” call) has been associated with a negative prediction, a similar approach have been taken in previous studies [41–44]. Table 2 shows the various statistical parameters for different sets (training, test, validation and Unilever sets) for the prediction of Ames outcome based just on DEREK software.

The hazard estimation software called ToxTree, which was developed by Ideconsult Ltd. (Sofia, Bulgaria) under the terms of a contract with the European Commission Joint Research Centre. This is capable of making structure-based predictions for a number of toxicological endpoints [28]. One of the modules developed as an extension to

Table 2 The various statistical parameters for different sets (training, test, validation and Unilever sets) for the prediction of Ames outcome based on different approaches

SET	Sensitivity	Specificity	Total accuracy	pos predictivity	neg predictivity	Kappa	MCC
In-house alerts							
Training	0.74	0.70	0.72	0.64	0.79	0.43	0.43
Test	0.56	0.74	0.65	0.68	0.63	0.30	0.32
Validation	0.71	0.70	0.70	0.67	0.74	0.40	0.40
Unilever	0.68	0.94	0.79	0.94	0.67	0.58	0.61
DEREK							
Training	0.72	0.82	0.78	0.74	0.81	0.55	0.55
Test	0.55	0.8	0.68	0.73	0.65	0.36	0.39
Validation	0.67	0.79	0.74	0.73	0.74	0.47	0.47
Unilever	0.65	0.83	0.74	0.85	0.64	0.48	0.48
ToxTree							
Training	0.79	0.73	0.76	0.68	0.83	0.51	0.51
Test	0.65	0.73	0.69	0.70	0.68	0.38	0.38
Validation	0.76	0.70	0.73	0.69	0.77	0.45	0.46
Unilever	0.80	0.76	0.78	0.83	0.72	0.56	0.55
kNN based on ECFP_6 fingerprints (atom type based, extended connectivity fingerprint, MaxDistance = 6)							
Training	0.85	0.87	0.86	0.82	0.89	0.72	0.71
Test	0.71	0.89	0.8	0.86	0.75	0.6	0.62
Validation	0.87	0.87	0.87	0.85	0.89	0.74	0.74
Unilever	0.88	0.94	0.9	0.96	0.84	0.81	0.8
Naive Bayesian classification model based on ECFP 4(atom type based, extended connectivity fingerprint, MaxDistance = 4)							
Training	0.85	0.92	0.89	0.89	0.90	0.78	0.79
Test	0.74	0.93	0.84	0.92	0.79	0.68	0.70
Validation	0.87	0.91	0.90	0.90	0.89	0.79	0.79
Unilever	0.92	0.88	0.91	0.92	0.88	0.81	0.81

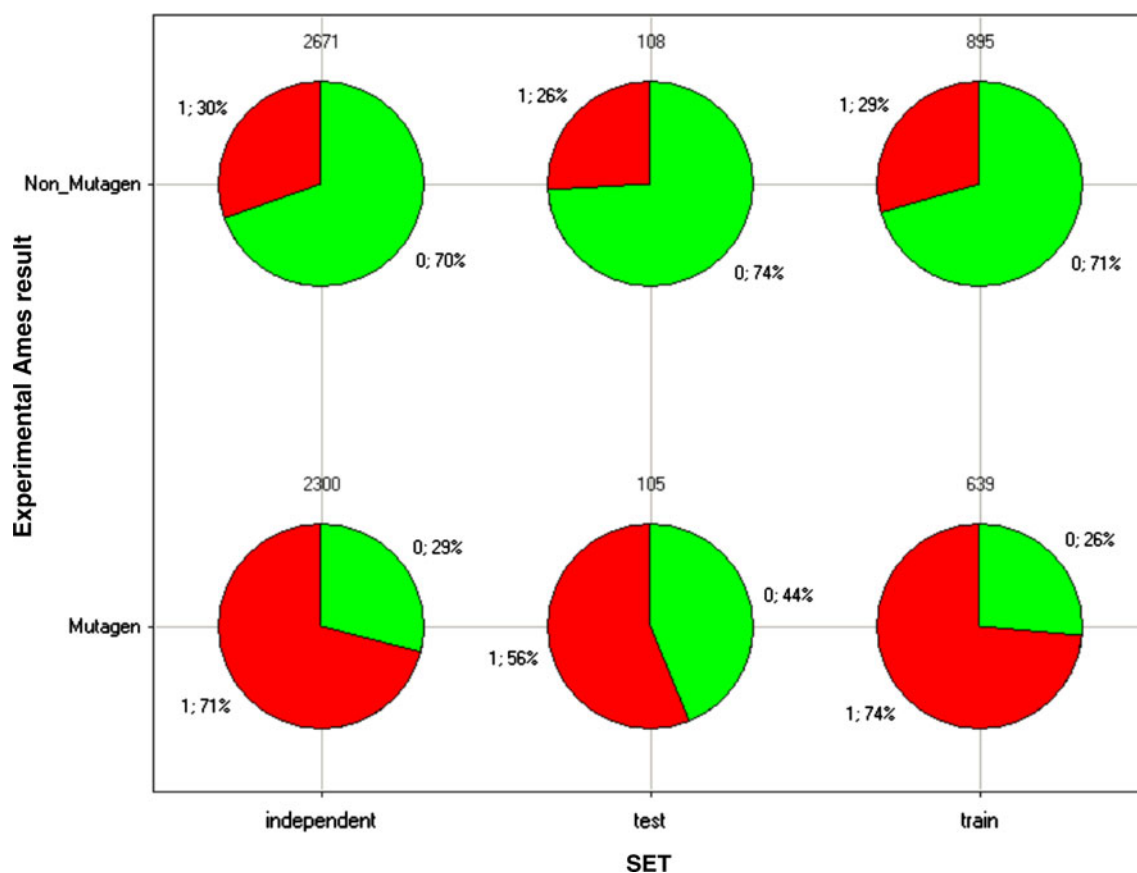


Fig. 3 This shows the number of true positives and true negatives for the different sets using in-house alert approach. Colours are based on the structural alert outcome (Red if any of the alert listed in Table 1 was present; green: no alert listed in table present was present)

ToxTree is aimed at the prediction of carcinogenicity and mutagenicity. This module encodes the Benigni/Bossa rule base for carcinogenicity and mutagenicity [45, 46]. This approach is very similar to the DEREK alert approach, but in addition to these alerts it also includes an additional QSAR analysis approach for certain chemical groups/classes (e.g. aromatic amines or unsaturated aldehydes). Based on these rules/QSAR, for each chemical it is possible to predict negative or positive outcome for Ames. Table 2, shows the various statistical parameters for different sets (training, test, validation and Unilever sets) for the prediction of Ames outcome based on ToxTree software.

kNN QSAR approach

The first in-house QSAR approach to be assessed was K-nearest Neighbours method (kNN). This is an advanced non-linear, non-parametric instance-based classification technique that assigns predictions based on a test compound's similarity to training compounds with known activities [47]. To judge the similarity of chemicals with others, several chemical fingerprinting methods in Pipeline

Pilot including Extended Connectivity Fingerprints (ECFP), Functional class connectivity fingerprints (FCFPs) and path fingerprints (FPFP) were assessed including the MDL Public keys (results not shown). The ECFP method (which is SciTegic's proprietary method for calculating structural fingerprints) was chosen as this method offered better predictivity for the training set and also it provides excellent characterization of chemicals by indexing the environments of every atom in a molecule by using up to four billion different structural features [48–50]. It has been reported that ECFP is an efficient and useful method for performing searching and clustering [50]. Several maximum path distances (such as number of bonds) were used for indexing an individual fragment. For this training set, ECFP_6 performed better than ECFP_2 and ECFP_4 (results not shown), and hence was chosen as the best option for this study. The value of k was optimised by taking 3, 5, 7 and 9 as the number of nearest molecules. An optimal value of “5” was identified such that each compound in the training set can be best predicted by the k-most similar compounds (out of the remaining training set). The resulting model consists of an activity relevant descriptor subspace, a defined k (in this case 5), and

training set compounds with known Ames outcomes used to populate this descriptor subspace. For the prediction for any target compound, the decision was based on the majority rule of at least 3 similar compounds from the training set using a minimum threshold of similarity as 0.6. This minimum threshold was selected based on number of nearest neighbours and on the performance of training and test sets using 0.5, 0.6, 0.7, 0.8 and 0.9 as minimum threshold (data not shown). It was not possible to make any predictions if the number of similar chemicals from the training set were less than 3 or there were equal numbers of mutagens and non-mutagens for a given target. In this approach for about 10 % of the chemicals, it was not possible to make any predictions because of reasons mentioned above. Table 2 lists different statistical parameters for the kNN approach based on ECFP₆ chemical fingerprinting and k as 5, for the different sets chosen in this study. Using this approach it is also possible to report the confidence/probability of a compound being mutagen or non-mutagen based on majority of its nearest neighbours (e.g. if for a target molecule, all 5 nearest neighbours are mutagens, then it is predicted to be a mutagen and the probability of this compound being a mutagen is 1.0; where as for another target molecule, out of 5 its nearest neighbours, only 3 were mutagens, then in this case it will still be predicted to be a mutagen but the probability of this compound being a mutagen is now 0.6). These probabilities/confidence can then be used for integration approaches which are discussed later.

Naïve Bayesian classification

This is the third in-house approach (second QSAR method), in which qualitative relationships between a set of descriptors and a given endpoint are explored. The Bayesian learner in Pipeline Pilot is a so-called naïve Bayesian classifier [51, 52]. The “naïve” term refers to the assumption that any particular feature contributes a specific amount to the likelihood of a sample being assigned to a given class, irrespective of the presence of any other features. In this naïve Bayesian classifier approach all the interaction effects of different fragments are ignored and each fragment is treated independently. In this example, this Bayesian classifier was combined with Pipeline Pilot’s chemical fingerprint technology. Several finger-printing methods again were explored and Extended Connectivity Fingerprints as developed by Sci-Tegic (especially ECFP₄) in combination with a few physchem properties (ALogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_AromaticRings; Num_RotatableBonds; Molecular_FractionalPolarSurface Area) yielded optimal Bayesian models for the current training data set. Similar to the above two methods (in-house alerts and kNN), it is also possible to get the probability of a

compound being a mutagen or non-mutagen using this Bayesian classification model approach.

Table 2 shows the various statistical parameters for different sets (training, test, validation and Unilever sets) for the prediction of Ames outcome based on naïve Bayesian classification QSAR approach using ECFP₄ fingerprints and physchem properties. It can be seen that both QSAR models (naïve Bayesian classification models and kNN model) perform quite well as compared to simple SAR alert approaches.

Integration of different in silico approaches

By definition, all models, including ADME-Tox models, are simulations of reality, and as such will never be completely accurate. However when multiple models are combined as a single consensus model, more accurate predictions can be achieved. Integration of SAR approaches with artificial intelligence systems for Ames in a consensus modelling manner has been shown to provide advantages over that of a single model [41, 53, 54]. It is well known, that each individual model has its own problems and pitfalls. By combining models for the same end point it is noted that the accuracy of the combined model usually becomes superior, and also the confidence of the prediction is accordingly better. The Toxicity Estimation Software Tool (T.E.S.T.), which has been developed by EPA allows users to easily estimate toxicity for different end points using variety of QSAR methodologies [53]. The consensus method for Ames in this was shown to achieve the best prediction accuracy (concordance) and prediction coverage. A statistical model for mutagenicity was developed and released as an open source software tool in the frame of the EU CAESAR project (<http://www.caesar-project.eu/>). Here the support vector machine (SVM) classification method was to develop a QSAR model and it was also investigated to see advantages of the combined use of their model with some SAs from the Benigni-Bossa rulebase (using ToxTree) [55].

For this study, three different ways were used to get a consensus prediction by integration of output from these individual models. Please note that to take into account of all the models for integration purposes a few compounds (~6 %; 85 chemicals from training set and 344 from validation set) were removed which had kNN prediction as “No-results”, due to either equal number of compounds as mutagens and non-mutagens in its nearest neighbours, or due to the fact that less than 3 nearest neighbours were selected.

Counts/majority rule

This is the simplest category method for integration, where consensus prediction is obtained simply based on the

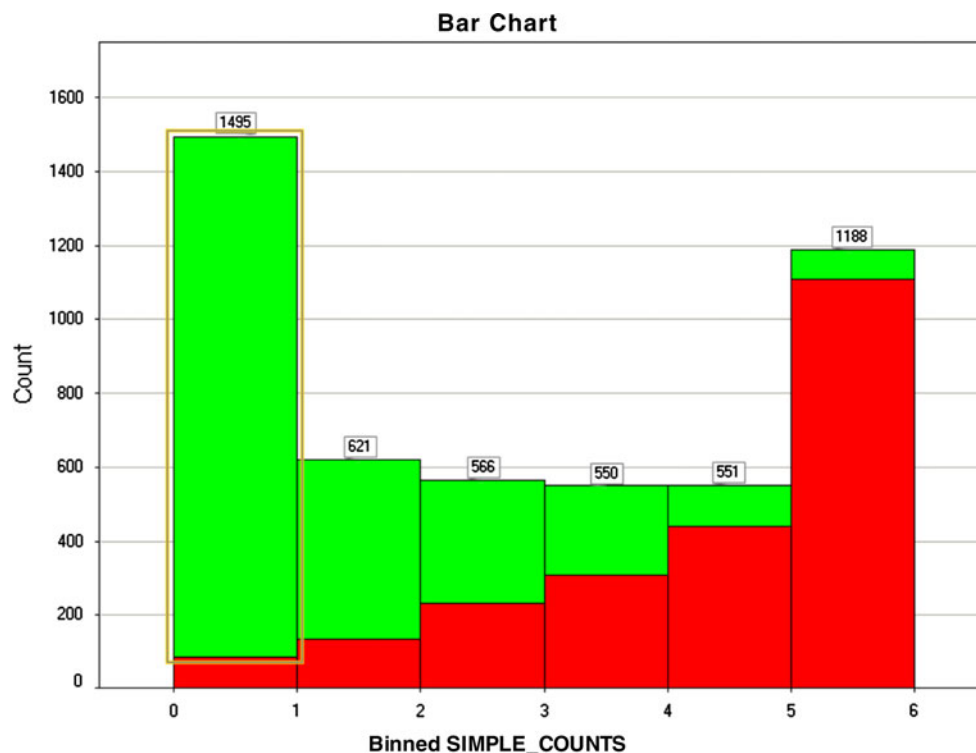
majority rule (i.e. counts of positive and negative predictions for all the models is performed for the same chemical and based on the majority vote a consensus prediction is reported). One of the main pitfalls of this method is that all models (weak or strong) are treated equally, and no consideration is given to correlations between these different models. It is also possible to obtain confidence in consensus predictions from this method of integration. Figure 4 shows the counts of positive or negative predictions using all the 5 different approaches discussed above (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian) for the validation set. As it can be seen from Fig. 4 if more and more models are consistent with each other, the overall predictivity/confidence becomes better.

Partial least square discriminant analysis (PLS-DA)

For the second integration method, the PLS-DA using SIMCA software was used to integrate individual models for producing the consensus predictions. In this approach, not only correlation between individual models is dealt with better but also different individual models are given different scores depending on their overall accuracy (i.e. strong models were given higher coefficients than the weaker models). This overcomes the problem of the first method which was simply counting the votes and each vote was given the same weighting irrespective the accuracy of the model. Another advantage of PLS integration approach over a first binary input integration method is that in the

case of a PLS integration approach it is possible to take continuous numbers as the input e.g. probabilities from individual models which highlights how strongly that individual model predicts a particular chemical is Ames positive or negative. This provides the opportunity to add confidence to consensus predictions. Therefore, in this PLS integration approach, probabilities and confidence levels, in addition to predicted class as the input were used for in-house alerts/kNN and Bayesian classification approaches, but for ToxTree and DEREK, only the binary classification was used as an input. Table 3 shows the performance of the PLS integration method on the training and test set. Tables 4, 5 shows the various statistical parameters for the reduced validation set (4627 chemicals) and Unilever set (42 chemicals) for the prediction of Ames outcome based on different individual models (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian) and also using three integration methods for consensus predictions. As discussed above, 344 chemicals were removed from validation set which had kNN prediction as “No-results”, due to this statistical parameters for the validation set in Tables 2, 4 and 5 are different except for the kNN models. The following equation shows the relationship between different individual models components and a PLS score. For DEREK and ToxTree, a score of 1 was given if there was alert found and 0, if no alert was found. For in-house methods, the probability output was used as described above in following equation. In case of chemicals without any structural alert in DEREK/ToxTree, predictions were

Fig. 4 Counts of models predicting to be mutagen out of 5 different approaches listed above (in-house alerts, kNN, Naïve Bayesian, DEREK and ToxTree) for the validation set. Colors are based on experimental Ames outcome (Red Mutagen; green Non-mutagen). As discussed above all the compounds with no predicted output from kNN approach are removed from this analysis



mainly influenced by QSAR techniques (kNN and Naïve Bayesian).

$$\begin{aligned} \text{PLS score} = & -0.821 + (0.099 * \text{DEREK}) \\ & + (0.015 * \text{ToxTree}) + (0.05 * \text{in-house alerts}) \\ & + (0.56 * \text{kNN}) + (1.51 * \text{Naive Bayesian}) \end{aligned} \quad (1)$$

If the PLS score was positive, the consensus prediction based on this approach was predicted to be a mutagen; and a non-mutagen in the case of a negative PLS score.

Bayesian integration methods

Bayesian method originated from a posthumous publication in 1763 by Thomas Bayes [56, 57]. The method starts with a suitable prior distribution, combines the prior distribution with the data available to produce a posterior distribution, and finally analyses and draws conclusions based on the posterior distribution.

For example, we wish to assess the probability of a hypothesis θ being true given available observed data X (also known as the evidence) and the prior distribution of the hypothesis, $p(\theta)$. From Bayes' theorem we can obtain the posterior distribution that is the updated probability of the hypothesis θ being true after taking into account both the prior and the evidence.

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \quad (2)$$

Where $p(X|\theta)$ is the likelihood of the data X given the hypothesis θ is fixed, e.g., being true. This method has been used historically for integration of several in vitro assays [56, 57].

In the context of the mutagenicity prediction, we wish to estimate the probability of a chemical being a mutagen given (1) the evidence of a test A, which yields a positive (+), and (2) a prior probability of the chemical being a mutagen, $P(\text{Mut})$. We can compute a posterior probability of the chemical to be mutagen by taking into account both the prior and the result of the test A using Eq. 2:

$$P(\text{Mut}/+) = \frac{P(\text{Mut})P(+/\text{Mut})}{P(\text{Mut})P(+/\text{Mut}) + P(\text{Non})P(+/\text{Non})} \quad (3)$$

where the “Mut” stands for mutagen while the “Non” stands for Non-mutagen.

Please note that $P(+/\text{Mut})$ in Eq. (2) is the probability that the test result is positive given that the tested chemical is a mutagen, which is actually one of the statistical measures of the performance of a binary classification test, i.e., sensitivity, the proportion of actual positives which are correctly identified as such. Meanwhile, given $P(+/\text{Non}) = 1 - P(-/\text{Non})$, $P(+/\text{Non})$ can be computed from $P(-/\text{Non})$, which is the probability that the test result is negative given that the tested chemical is a Non-mutagen. This is another statistical measure of the performance, i.e., specificity, the proportion of negatives which are correctly identified. Therefore the posterior probability of the chemical being a mutagen can be calculated if we know the values of three parameters, i.e., sensitivity, specificity of the test A and the prior probability of the chemical being a mutagen, $P(\text{Mut})$, the later could be potentially obtained from expert knowledge on structural or mechanism information of the chemical or alternatively from the knowledge of distribution of mutagens used in the chemical industry in the world that are closely similar to the chemical. $P(\text{Mut})$ can also be considered as the prevalence information. However, if no prior knowledge or intuition is available, a non-informative prior could be set at 0.5, which signifies no bias to either mutagen or non-mutagen. In this integration 0.5 will be taken as a prior.

Suppose we have a battery of K tests, whose qualitative (positive/negative) results for a given chemical are denoted by A_1, A_2, \dots, A_k . According to Bayes's theorem, the probability of the chemical being mutagen given A_1, A_2, \dots, A_k is

$$\begin{aligned} P(\text{Mut}/A_1, A_2, \dots, A_k) = \\ \frac{P(A_1, A_2, \dots, A_k|\text{Mut})P(\text{Mut})}{P(A_1, A_2, \dots, A_k|\text{Mut})P(\text{Mut}) + P(A_1, A_2, \dots, A_k|\text{Non})P(\text{Non})} \end{aligned} \quad (4)$$

and by invoking the conditionally independence of these tests, Eq. (4) can be transformed to

Table 3 The various statistical parameters for the training and test sets for the prediction of Ames outcome based just on two different integration methods for consensus predictions (PLS-integration and Bayesian classification integration)

SET	Sensitivity	Specificity	Total accuracy	pos predictivity	neg predictivity	Kappa	MCC
PLS integration							
Training	0.88	0.93	0.91	0.9	0.91	0.81	0.81
Test	0.78	0.93	0.86	0.92	0.81	0.72	0.74
Bayesian integration							
Training	0.87	0.92	0.9	0.88	0.91	0.79	0.79
Test	0.7	0.93	0.82	0.91	0.77	0.64	0.68

Table 4 The various statistical parameters for the validation sets (4627 chemicals, independent set, kNN with “No-results” removed) for the prediction of Ames outcome based just on different individual models (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian)

Method	Sensitivity	Specificity	Total accuracy	pos predictivity	Neg predictivity	kappa	MCC
DEREK	0.69	0.80	0.75	0.74	0.75	0.49	0.50
TOXTREE	0.78	0.71	0.74	0.70	0.79	0.49	0.49
In-house Alerts	0.72	0.70	0.71	0.67	0.75	0.42	0.42
kNN	0.87	0.87	0.87	0.85	0.89	0.74	0.74
NB	0.89	0.92	0.91	0.90	0.91	0.81	0.81
Simple counts integration (5 models Mutagen)	0.52	0.97	0.76	0.93	0.70	0.50	0.60
Simple_Counts_Integration (4 models Mutagen)	0.71	0.93	0.83	0.89	0.79	0.64	0.68
Simple counts integration (3 models Mutagen)	0.83	0.84	0.84	0.82	0.85	0.67	0.67
Simple_Counts_Integration (2 models Mutagen)	0.92	0.72	0.82	0.74	0.92	0.64	0.66
Simple counts integration (1 models Mutagen)	0.97	0.54	0.74	0.64	0.96	0.49	0.60
PLS-integration	0.90	0.92	0.91	0.91	0.92	0.83	0.83
Bayesian integration	0.89	0.89	0.89	0.88	0.90	0.78	0.78

Table 5 The various statistical parameters for internal Unilever sets (42 chemicals) for the prediction of Ames outcome based just on different individual models (in-house alerts, DEREK, ToxTree, kNN

and also using three integration methods for consensus predictions (simple counts, PLS-integration and Bayesian classification integration)

and Naïve Bayesian) and also using two integration methods for consensus predictions (PLS-integration and Bayesian classification integration)

Method	Sensitivity	Specificity	Total accuracy	pos predictivity	Neg predictivity	kappa	MCC
DEREK	0.68	0.83	0.74	0.85	0.64	0.48	0.48
TOXTREE	0.80	0.76	0.78	0.83	0.72	0.56	0.55
In-house alerts	0.68	0.94	0.79	0.94	0.67	0.58	0.61
kNN	0.88	0.94	0.90	0.96	0.84	0.81	0.80
NB	0.92	0.88	0.91	0.92	0.88	0.80	0.80
Simple counts integration (5 models Mutagen)	0.60	1.00	0.76	1.00	0.63	0.55	0.63
Simple counts integration (4 models Mutagen)	0.72	0.94	0.81	0.95	0.70	0.63	0.64
Simple counts integration (3 models Mutagen)	0.84	0.82	0.83	0.87	0.78	0.66	0.65
Simple counts integration (2 models Mutagen)	na	na	na	na	na	na	na
Simple counts integration (1 models Mutagen)	1.00	0.76	0.90	0.86	1.00	0.79	0.86
PLS-integration	0.88	0.94	0.90	0.96	0.84	0.81	0.80
Bayesian integration	0.88	0.94	0.90	0.96	0.84	0.81	0.80

$$P(Mut/A_1, A_2, \dots, A_k) =$$

$$\frac{\left[\prod_{i=1}^k P(A_i|Mut) \right] P(Mut)}{\left[\prod_{i=1}^k P(A_i|Mut) \right] P(Mut) + \left[\prod_{i=1}^k P(A_i|Non) \right] P(Non)} \quad (5)$$

The Bayesian integration formula in Eq. 5 shows that the accuracy of integration is determined not only by the quality of the estimations (i.e., sensitivity and specificity) of each individual test constituting the test battery, but also by the accuracy of the prevalence information concerning the chemical. The prevalence information is represented by

the distribution of mutagens in a specified group of chemicals to which the chemical under study belongs.

The calculations produce two theoretical measures of sensitivity and specificity of a test battery. When applied to a given dataset, empirical performance metrics of sensitivity and specificity would be slightly different from two theoretical measures

Table 6 lists the overall prediction conclusions for all 32 combination scenarios from all possible predictions of 5 models. For each scenario (No. i), it also includes 4 corresponding conditional probabilities, where $P(Mut/No. i)$, $P(Non/No. i)$ were used to determine an overall conclusion whilst $P(No. i/Mut)$, $P(No. i/Non)$ were used to calculate the sensitivity and specificity of the integration approach.

Our calculations here use the values of sensitivities and specificities of 5 methods based on the training dataset and a prior probability of 0.5. The performance of the integration is theoretically better than any individual performance in terms of sensitivity and specificity, whose results are shown in Fig. 5. Theoretically, it achieves a sensitivity of 94.45 % and a specificity of 96.40 %.

To understand the performance of the integration in real-world situations, this approach was applied to make predictions for the chemicals in the training dataset and in the independent dataset. Table 3 shows the performance of the *Bayesian* integration method on the training and test set. For the training set, an actual sensitivity of 87 % and a specificity of 91 % with an overall accuracy of 90 % was achieved. For the large independent dataset, an actual sensitivity of 88.5 % and a specificity of 89.5 % with an overall accuracy of 89 % was achieved. For comparison, the performance of all 5 individual methods is also shown in Tables 4, 5 for external independent (Table 4) and internal Unilever sets (Table 5). Theoretically, the Bayesian integration method performs best through integration of those individual methods with a good trade-off between sensitivity and specificity. In reality, it achieves the performance at a similar level as the best individual method of Naïve Bayesian among all available 5 methods.

Discussion

The need to assess the ability of a chemical to act as a genotoxin is one of the primary requirements in regulatory toxicology and is essential information when performing risk assessments for the consumer safety. In silico models developed using different modelling techniques (especially QSAR methods) are more likely to perform well for chemicals in a certain chemical space, rather than in the whole chemical space [58, 59]. More to the point, QSARs are unlikely to predict outside the chemical space of the training set because they rely on similarity measures, whereas structural alerts rely on the identification of fragments, which can be present on molecules that have low similarity to the training set. In all respects, both QSAR and structural alert approaches are complementary approaches, each with their advantages and limitations, which is why a consensus approach, in theory, should offer better predictability as compared to individual models (SAR or QSAR). It should be noted that issue of applicability domain applies to all (Q)SAR methods, and does not really depend on the method and descriptors used for building the models.

In this paper several QSAR/SAR approaches were assessed, and using a consistent large validation data set, a comparison of different methods was performed. Three different

methods were used for producing consensus predictions using five different individual methods. As can be seen from Table 4 that consensus predictions made by PLS-DA and Bayesian classification integration methods for integration of all five methods offers better predictivity and confidence as compared to simple voting integration. Figure 5 shows the comparison of different individual models and integration methods on a validation set. As it can be seen from Fig. 5 that both of QSAR methods (kNN/NB) outperforms other three SAR methods (DEREK, ToxTree or in-house alerts). This also shows that consensus Ames predictions made by the integration of five individual models (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian) can offer superiority over predictions based on single models (especially SAR alert methods) and this increases the overall confidence of the predictions. It should be noted that experimental repeatability of the Ames test is known to be around 85 % [60] or 94 % [62], and our methods are predicting around 90 %, which is not too far away from the experimental repeatability. One of the reason for such high accuracy could simply be due to the fact that that our datasets have many similar molecule clusters (discussed in data/methods), even though our training set was smaller and consists of only one quarter of the whole dataset, still there is possibility of models to learn dataset. It could be argued that our validation set is not really ‘independent’ as it has been taken from the same pool of chemicals as training set. Recently it has been reported [61] that QSAR models often appears to have superior results because it is often validated with data from the same pool of compounds that the training set was derived and may therefore not be considered an effective external validation. To test this we also used additional data from Sushko et al. [62] as shown in Fig. 1 to verify these models. Table 7 shows the various statistical parameters for the leftover chemicals from Sushko et al. which were not part of our set (881 chemicals) for the prediction of Ames outcome based on different individual models (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian) and also using three integration methods for consensus predictions. As it can be seen from Table 7 that the consensus predictions made by PLS-DA and Bayesian classification integration methods for integration of all five methods offers better predictivity and confidence as compared to any of the individual models. As discussed above there needs to be balance between predictivity and interpretability. It could be urged that although these integrated methods work better from the predictivity point of view but they are less transparent as compared to individual models. Simple transparent models can be of very much of help when we need guidance over the synthesis of new chemicals for designing out particular toxicity, where as an accurate model can help in prioritisation of existing chemicals for testing. The majority vote and Bayesian integration methods are categorical methods for integration, and the latter method appears

Table 6 Combination scenarios of 5 in silico methods and the overall conclusions for each scenario of the test battery

Combination No. i	Possible test results					Conclusions or results of test battery	P(Mut/No. i)	P(Non/No. i)	P(No. i/Mut)	P(No. i/Non)
	Derek	ToxTree	Alerts	NB	kNN					
1	+	+	+	+	+	+	0.9996	0.0004	0.5905	0.0000
2	+	+	+	+	–	+	0.9854	0.0146	0.0656	0.0001
3	+	+	+	–	+	+	0.9687	0.0313	0.0656	0.0001
4	+	+	+	–	–	–	0.4464	0.5536	0.0073	0.0008
5	+	+	–	+	+	+	0.9972	0.0028	0.0656	0.0001
6	+	+	–	+	–	+	0.9025	0.0975	0.0073	0.0008
7	+	+	–	–	+	+	0.8097	0.1903	0.0073	0.0008
8	+	+	–	–	–	–	0.0997	0.9003	0.0008	0.0073
9	+	–	+	+	+	+	0.9957	0.0043	0.0656	0.0001
10	+	–	+	+	–	+	0.8572	0.1428	0.0073	0.0008
11	+	–	+	–	+	+	0.7340	0.2660	0.0073	0.0008
12	+	–	+	–	–	–	0.0670	0.9330	0.0008	0.0073
13	+	–	–	+	+	+	0.9694	0.0306	0.0073	0.0008
14	+	–	–	+	–	–	0.4518	0.5482	0.0008	0.0073
15	+	–	–	–	+	–	0.2748	0.7252	0.0008	0.0073
16	+	–	–	–	–	–	0.0098	0.9902	0.0001	0.0656
17	–	+	+	+	+	+	0.9952	0.0048	0.0656	0.0001
18	–	+	+	+	–	+	0.8427	0.1573	0.0073	0.0008
19	–	+	+	–	+	+	0.7112	0.2888	0.0073	0.0008
20	–	+	+	–	–	–	0.0602	0.9398	0.0008	0.0073
21	–	+	–	+	+	+	0.9658	0.0342	0.0073	0.0008
22	–	+	–	+	–	–	0.4238	0.5762	0.0008	0.0073
23	–	+	–	–	+	–	0.2527	0.7473	0.0008	0.0073
24	–	+	–	–	–	–	0.0087	0.9913	0.0001	0.0656
25	–	–	+	+	+	+	0.9483	0.0517	0.0073	0.0008
26	–	–	+	+	–	–	0.3231	0.6769	0.0008	0.0073
27	–	–	+	–	+	–	0.1799	0.8201	0.0008	0.0073
28	–	–	+	–	–	–	0.0057	0.9943	0.0001	0.0656
29	–	–	–	+	+	+	0.7157	0.2843	0.0008	0.0073
30	–	–	–	+	–	–	0.0615	0.9385	0.0001	0.0656
31	–	–	–	–	+	–	0.0292	0.9708	0.0001	0.0656
32	–	–	–	–	–	–	0.0008	0.9992	0.0000	0.5905

Results are computed using the sensitivities and specificities of 5 models calculated over the training dataset and using 0.5 as a prior probability. A theoretical sensitivity and specificity of the battery calculated using Eqs. 3 and 4 are 94.45 and 96.40 % respectively

to perform better than just simply majority vote. This Bayesian method works better as compared to a simple majority vote method as it is able to give weighting to individual methods based on their prediction performance, as well as by quantitatively handling conflict information in a consistent mathematical manner. The PLS integration method performs the best in all the sets tested in this study, as it is able to take into account the probability/confidence from each different model. The correlation between different models was expected as the same training set was used for the different in-house models. Moreover, SAR approaches (in-house, DEREK and ToxTree) have many similar alerts.

Having these correlated models in simple counts consensus modelling approach can usually biased predictions as equal weights are given in simple count consensus method. Integration using the PLS integration approach correlation between these different individual models was taken into account by the PLS-DA method and is able to give an overall score, which can be used as a reflection of the confidence in the Ames predictions. It should be noted SAR methods usually work on one side e.g. in our case if no alert is present it does not mean that the chemical is predicted to be Ames negative. By doing integration of these SAR methods with (Q)SAR method (kNN/Bayesian) predictivity towards

Fig. 5 This shows the various statistical parameters for the validation sets (4627 chemicals) for the prediction of Ames outcome based just on different individual models (in-house alerts, DEREK, ToxTree, kNN and Naïve Bayesian) and also using three integration methods (simple counts, PLS-integration and Bayesian integration) for consensus prediction. Solid outside *light blue line*, shows the scenario of perfect model and we would like to be as close as possible to this

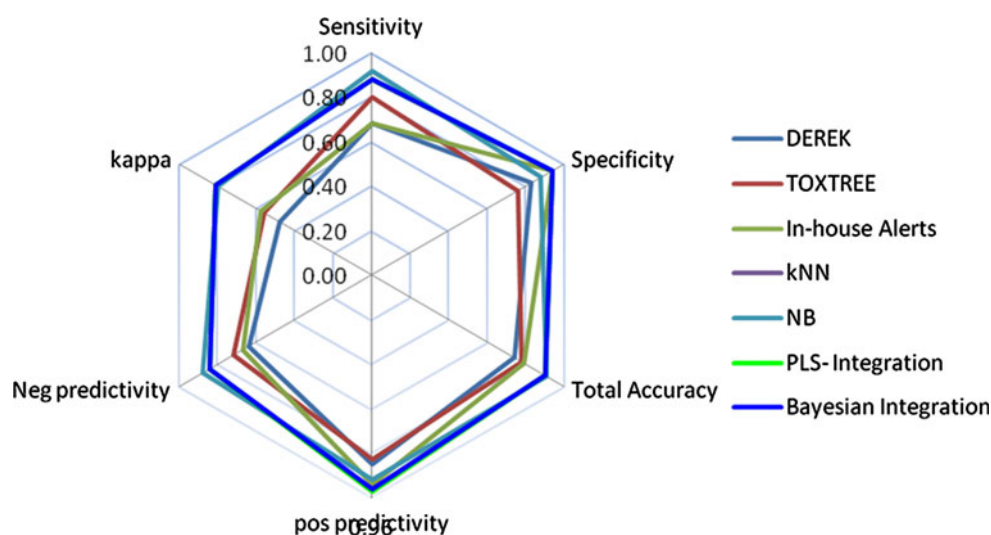


Table 7 The various statistical parameters for leftover chemicals from Sushko et al. [62] which were not part of our original dataset (881 chemicals) for the prediction of Ames outcome based just on different individual models (in-house alerts, DEREK, ToxTree, kNN

and Naïve Bayesian) and also using two integration methods for consensus predictions (PLS-integration and Bayesian classification integration)

Method	Sensitivity	Specificity	Total accuracy	pos predictivity	Neg predictivity	kappa	MCC
DEREK	0.81	0.83	0.81	0.91	0.66	0.59	0.57
TOXTREE	0.85	0.74	0.82	0.88	0.69	0.59	0.57
In-house Alerts	0.80	0.73	0.78	0.87	0.62	0.50	0.48
kNN	0.82	0.79	0.81	0.90	0.66	0.58	0.55
NB	0.77	0.93	0.82	0.96	0.64	0.62	0.60
Simple counts integration (5 models Mutagen)	0.47	1.00	0.63	1.00	0.46	0.35	0.45
Simple counts integration (4 models Mutagen)	0.74	0.98	0.82	0.99	0.63	0.63	0.62
Simple counts integration (3 models Mutagen)	0.89	0.92	0.90	0.96	0.79	0.77	0.75
Simple counts integration (2 models Mutagen)	0.95	0.72	0.88	0.88	0.86	0.7	0.74
Simple counts integration (1 models Mutagen)	0.99	0.41	0.81	0.79	0.93	0.47	0.73
PLS-integration	0.88	0.95	0.90	0.98	0.78	0.78	0.75
Bayesian integration	0.87	0.93	0.89	0.96	0.76	0.75	0.72

positive or negative AMES is improved where (Q)SAR models will dominate the consensus predictions especially for the negative predictions in the case of PLS and Bayesian integration. These in silico consensus model predictions for Ames can be used as an alerting system for potential genotoxicity and contribute to the toxicity risk assessment of chemicals. Nevertheless, validation of these integration approaches on these external sets appeared to show that the accuracy of the consensus predictions (especially with PLS-DA and CBI methods) was superior as compared to individual models, and that the confidence of the prediction was also better as compared to predictions based on an individual model. Consensus approach by simple count doesn't offer better predictivity as discussed above as weaker or strong

models are given equal weightings. Although all of the validation sets performs reasonably well as compared to training or test set, however by looking at high performance of kNN method in these datasets (Tables 4, 5, 7), it still could be argued that the performance estimate could be somehow overoptimistic for true "unknown" chemical, when it is from a different chemical class/space. Our independent Unilever set is ideal for the evaluation of these models, but care should be taken due to small size of this dataset (42 chemicals). There is no doubt in silico modelling for Ames will continue to be improved, thanks to the increasing availability of good quality toxicity data and more advanced computational techniques. New data will be able to highlight the underlying reasons for outliers and will be able to refine and define new rules.

Acknowledgments Development of the dataset for this work was performed with help from Leadscape. The authors wish to acknowledge Dr Glenn Myatt (Leadscape) for his constant and generous help during the course of this work.

References

- Ames BN, Lee FD, Durston WE (1973) An improved bacterial test system for the detection and classification of mutagens and carcinogens. *PNAS* 70:782–786
- Mortelmans K, Zeiger E (2000) The Ames Salmonella/microsome mutagenicity assay. *Mutat Res* 455:29–60
- Yan Z, Caldwell G (eds) (2004) Improvement of the Ames test using human liver S9 preparation. Optimization in drug discovery: in vitro methods. Methods in pharmacology and toxicology. Humana Press
- Ames B, McCann J, Yamasaki E (1975) Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. *Mutat Res* 31:347–360
- Ames BN (1984) The detection of environmental mutagens and potential carcinogens. *Cancer* 53:2030–2040
- Miller JA, Miller EC (1977) Ultimate chemical carcinogen as reactive mutagenic electrophiles. In: Hiatt HH, Watson JD, Winsten JA (eds) Origin of human cancers. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 605–627
- Ashby J, Tennant RW (1988) Chemical-structure, salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the United-States NCI/NTP. *Mutat Res* 204:17–115
- Ashby J, Tennant RW (1991) Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the US NTP. *Mutat Res* 257:229–306
- Ashby J (1985) Fundamental structural alerts to potential carcinogenicity or non-carcinogenicity. *Environ Mutagen* 7:919–921
- Cariello NF, Wilson JD, Britt BH, Wedd DJ, Burlinson B, Gombar V (2002) Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. *Mutagenesis* 17:321–329
- Benfenati E, Benigni R, Demarini DM, Helma C, Kirkland D, Martin TM, Mazzatorta G, Ouedraogo-Arras G, Richard AM, Schilter B, Schoonen WGEJ, Snyder RD, Yang C (2009) Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives. *J Environ Sci Health C* 27:57–90
- Brinn M, Walsh P, Payne M, Bott B (1992) Neural network classification of mutagens using structural fragment data. *SAR QSAR Environ Res* 1:169–211
- Basak SC, Mills D, Gute BD, Hawkins DM (2003) Predicting mutagenicity of congeneric and diverse sets of chemicals using computed molecular descriptors: a hierarchical approach. In: Benigni R (ed) Quantitative structure-activity relationship (QSAR) models of chemical mutagens and carcinogens. CRC Press, Boca Raton, pp 207–234
- Klopman G, Chakravarti SK, Harris H, Ivanov J, Saiakhov RD (2003) In silico screening of high production volume chemicals for mutagenicity using the MCASE QSAR expert system. *SAR QSAR Environ Res* 14(2):165–180
- Klopman G, Chakravarti SK, Harris H, Ivanov J, Saiakhov RD (2003) In silico screening of high production volume chemicals for mutagenicity using the MCASE QSAR expert system. *SAR QSAR Environ Res* 14:165–180
- Serafimova R, Todorov M, Pavlov T, Kotov S, Jacob E, Aptula A, Mekenyan O (2007) Identification of the structural requirements for mutagenicity, by incorporating molecular flexibility and metabolic activation of chemicals. II. General Ames mutagenicity model. *Chem Res Toxicol* 20:662–676
- Carlsson L, Helgee EA, Boyer S (2009) Interpretation of non-linear QSAR models applied to Ames mutagenicity data. *J Chem Inform Model* 49:2551–2558
- Didziapetris R, Lanevskij K, Japertas P (2008) Trainable QSAR model of Ames genotoxicity. Abstracts of papers, 236th ACS national meeting, Philadelphia, United States, August 17–21, TOXI-088
- Langham JJ, Jain AN (2008) Accurate and interpretable computational modelling of chemical mutagenicity. *J Chem Inform Model* 48:1833–1839
- Maran U, Sild S (2003) QSAR modeling of genotoxicity on non-congeneric sets of organic compounds. *Artif Intell Rev* 20: 13–38
- Helma C, Cramer T, Kramer S, De Raedt L (2004) Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comp Sci* 44:1402–1411
- Contrera JF, Matthews EJ, Kruhlak NL, Benz RD (2005) In silico screening of chemicals for bacterial mutagenicity using electrophilological E-state indices and MDL QSAR software. *Regulat Pharmacol Toxicol* 43:313–323
- Mekenyan O, Dimitrov S, Serafimova R, Thompson ED, Kotov S, Dimitrova N, Walker JD (2004) Identification of the structural requirements for mutagenicity by incorporating molecular flexibility and metabolic activation of chemicals I: TA100 model. *Chem Res Toxicol* 17:753–766
- Sanderson DM, Earnshaw CG (1991) Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum Exp Toxicol* 10:261–273
- Greene N, Judson PN, Langowski JJ, Marchant CA (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR QSAR Environ Res* 10:299–314
- Judson PN (2006) Using computer reasoning about qualitative and quantitative information to predict metabolism and toxicity. In: Testa B, Kramer SD, Wunderli-Allespach H, Volkens G (eds) Pharmacokinetic profiling in drug research: biological, physicochemical, and computational strategies. Wiley, New York, pp 183–215
- Ridings JE, Barratt MD, Cary R, Earnshaw CG, Eggington CE, Ellis MK, Judson PN, Langowski JJ, Marchant CA, Payne MP, Watson WP, Yih TD (1996) Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* 106:267–279
- Benigni R, Bossa C, Jeliazkova N, Netzeva T, Worth A (2008) The Benigni/Bossa rulebase for mutagenicity and carcinogenicity: a module of ToxTree. Assessed on 25th Aug 11 <http://toxtree.sourceforge.net/carc.html>
- Serafimova R, Gatnik MF, Worth A (2010) Review of QSAR models and software tools for predicting genotoxicity and carcinogenicity, JRC Scientific and technical report (2010) Assessed on 5th Aug 10. http://ecb.jrc.ec.europa.eu/DOCUMENTS/QSAR/EUR_24427_EN.pdf
- Kazius J, McGuire R, Bursi R (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 48: 312–320
- FDA CDER FOI site: <http://www.accessdata.fda.gov/scripts/cder/drugsatfda>. Accessed June 5, 2010
- CCRIS: <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS> Accessed June 5, 2010
- Tennant RW (1991) The genetic toxicity database of the National Toxicology Program: evaluation of the relationships between

- genetic toxicity and carcinogenicity. *Environ Health Perspect* 96:47–51
34. Tokyo-Eiken (2007) Tokyo Metropolitan Institute of Public Health. Mutagenicity of food additives. <http://www.tokyo-eiken.go.jp/henigen/index.htm>. Accessed June 5, 2010
 35. Bercu JP, Stuart MM, Deahl JT, Gombar VK, Callis CM, van Lier R (2010) In silico approaches to predicting cancer potency for risk assessment of genotoxic impurities in drug substances. *Regul Toxicol Pharmacol* 57:300–306
 36. Cohen J (1968) Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–220
 37. Strijbos J, Martens R, Prins F, Jochems W (2006) Content analysis: what are they talking about? *Comput Educ* 46:29–48
 38. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405:442–451
 39. Enoch SJ, Cronin MTD (2010) A review of the electrophilic reaction chemistry involved in covalent DNA binding. *Crit Rev Toxicol* 40:728–748
 40. Enoch SJ, Madden JC, Cronin MTD (2008) Identification of mechanism of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ Res* 19:555–578
 41. Hillebrecht A, Muster W, Brigo A, Kansy M, Weiser T, Singer T (2011) Comparative evaluation of in silico systems for Ames test mutagenicity prediction: scope and limitations. *Chem Res Toxicol* 24:843–854
 42. Snyder RD, Smith MD (2005) Computational prediction of genotoxicity: room for improvement. *Drug Discov Today Biosilico* 10:1119–1124
 43. Marchant CA, Briggs KA, Long A (2008) In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows Meteor, and Vitic. *Toxicol Mech Methods* 18: 177–187
 44. White AC, Mueller RA, Gallavan R, Aaron S, Wilson AGE (2003) A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutat Res* 539:77–89
 45. Benigni R, Bossa C (2006) Structural alerts of mutagens and carcinogens. *Curr Comput Aid Drug Des* 2:169–176
 46. Benigni R, Bossa C (2006) Structure-activity models of chemical carcinogens: state of the art, and new directions. *Ann Ist Super Sanità* 42:118–126
 47. Zheng W, Tropsha A (2000) Novel variable selection quantitative structure property relationship approach based on k-nearest neighbor principle. *J Chem Inf Comput Sci* 40:185–194
 48. Hassan M, Brown RD, Varma-O'Brian S, Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* 10:283–299
 49. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW (2006) Enrichment of high-throughput screening data with increasing levels of noise using support-vector machines, recursive partitioning, and laplacian-modified Naïve Bayesian classifiers. *J Chem Inf Model* 46:193–200
 50. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754
 51. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller KR (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *J Chem Inf Model* 49:2077–2081
 52. Sun HA (2005) Naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem* 48:4031–4039
 53. <http://www.epa.gov/nrmrl/std/cppb/qsar/testuserguide.pdf>. Accessed 25th Aug 2011)
 54. Mazzatorta P, Tran L, Schilter B, Grigorov M (2007) Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of Ames test mutagenicity. *J Chem Inform Model* 47:34–38
 55. Ferrari T, Gini G (2010) An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem Central J* 2010, 4(Suppl 1):S2 (29 July 2010)–<http://www.journal.chemistrycentral.com/content/4/S1/S2>)
 56. Rosenkranz HS, Klopman G, Chankong V, Pet-Edwards J, Haimes YY (1984) Prediction of environmental carcinogens: a strategy for the mid-1980s. *Environ Mutagen* 6:231–258
 57. Chankong V, Haimes YY, Rosenkranz HS, Pet-Edwards J (1985) The carcinogenicity prediction and battery selection (CPBS) method: a bayesian approach. *Mutat Res* 153:135–166
 58. Modi S, Hughes M, Garrow A, White A (2012) The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug Discov Today* 17:135–142
 59. Gleeson PM, Modi S, Bender A, Marchese-Robinson RL, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC (2012) *Curr Pharm Des* 18:1266–1291
 60. Piegorsch WW, Zeiger E (1991) Measuring intra-assay agreement for the Ames Salmonella assay. In: Rienhoff O, Lindberg DAB (eds) *Statistical methods in toxicology*. Springer, Heidelberg, pp 35–41
 61. Naven RT, Louise-May S, Greene N (2010) Expert Opin Drug Metab Toxicol 6:797–807
 62. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A et al (2010) Applicability domains for classification problems: benchmarking of distance to models for ames mutagenicity set. *J Chem Inf Model* 50:2094–2111