Regulatory Toxicology and Pharmacology 144 (2023) 105486

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

DeepAmes: A deep learning-powered Ames test predictive model with potential for regulatory application

Ting Li [a], Zhichao Liu [a,1], Shraddha Thakkar [b], Ruth Roberts [c,d], Weida Tong [a,*]

[a] National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR, USA
[b] Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA
[c] ApconiX Ltd, Alderley Park, Alderley Edge, SK10 4TG, UK
[d] University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

ARTICLE INFO

Handling Editor: Dr. Martin Van den berg

Keywords:
Ames test
QSAR
Mutagenicity
Context of use
Applicability domain
Deep learning
Machine learning

ABSTRACT

The Ames assay is required by the regulatory agencies worldwide to assess the mutagenic potential risk of consumer products. As well as this *in vitro* assay, *in silico* approaches have been widely used to predict Ames test results as outlined in the International Council for Harmonization (ICH) guidelines. Building on this *in silico* approach, here we describe DeepAmes, a high performance and robust model developed with a novel deep learning (DL) approach for potential utility in regulatory science. DeepAmes was developed with a large and consistent Ames dataset (>10,000 compounds) and was compared with other five standard Machine Learning (ML) methods. Using a test set of 1,543 compounds, DeepAmes was the best performer in predicting the outcome of Ames assay. In addition, DeepAmes yielded the best and most stable performance up to when compounds were >30% outside of the applicability domain (AD). Regarding the potential for regulatory application, a revised version of DeepAmes with a much-improved sensitivity of 0.87 from 0.47. In conclusion, DeepAmes provides a DL-powered Ames test predictive model for predicting the results of Ames tests; with its defined AD and clear context of use, DeepAmes has potential for utility in regulatory application.

1. Introduction

Understanding the potential for compounds to cause DNA mutation is a key step in assessing the regulatory safety of consumer products since mutations resulting from chemical interactions are associated with cancer development (Benigni and Bossa, 2011; Cassano et al., 2014). Chemical-induced mutagenesis, such as frame-shifts and base-pair substitutions, can be detected by the Ames test that uses bacteria to determine if a chemical is likely to cause genetic mutations (Ames et al., 1975; McCann et al., 1975; Mortelmans and Zeiger, 2000). There are also many *in silico* methods such as Quantitative Structure-Activity Relationship (QSAR) models that have been developed to predict the outcome of the Ames assay results solely based on the chemical structure of the compound. These models play a role in the assessment of chemical safety and are cited in the International Council for Harmonization (ICH) M7 guideline regarding the use of *in silico* methods to assess the mutagenicity of impurities in pharmaceuticals (Honma et al., 2019), and some of which are utilized by the FDA.

Over the past decades, a variety of QSAR models have been developed to predict the outcome of the Ames assay (Greene et al., 1999; Hanser et al., 2014; Kasamatsu et al., 2021; Kazius et al., 2005; Klopman, 1984, 1992; Kumar et al., 2021; Lahl and Gundert-Remy, 2008; Mekenyan et al., 2004; Pavan and Worth, 2008; Roberts et al., 2000; Sanderson and Earnshaw, 1991; Schwab et al., 2016; Serafimova et al., 2007; Vian et al., 2019; Xu et al., 2012). For example, Xu et al. curated a comprehensive database of 6,786 diverse compounds from four published papers to develop predictive models (Xu et al., 2012). Kumar et al. curated 3,039 compounds from the literature and developed a mutagenicity prediction model using deep neural networks (DNNs) (Kumar et al., 2021). One drawback of these studies is that the datasets were often pooled from diverse sources, sometimes derived within different test conditions and guidelines. Indeed, estimated inter-laboratory reproducibility of Ames tests is around 85% (Kamber et al., 2009). Thus, without stringent inclusion/exclusion criteria, there may be inconsistency in the data used for model development.

A large dataset generated using consistent methodologies is

* Corresponding author.
E-mail address: Weida.Tong@fda.hhs.gov (W. Tong).
[1] Current affiliation: Integrative Toxicology, Nonclinical Drug Safety, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, U.S.A.

https://doi.org/10.1016/j.yrtph.2023.105486
Received 17 March 2023; Received in revised form 14 July 2023; Accepted 23 August 2023
Available online 25 August 2023

important in developing a reliable and robust predictive model for the Ames assay. Recently, the second Ames/QSAR international challenge project was launched, which included a proprietary Ames database of ~12,000 chemical compounds for developing predictive models. In addition to this, a test set of 1,589 chemical compounds were provided with blinded Ames test results. Five Ames strains, '*S. thyphimurium* TA100, TA98, TA1535, TA1537 and *E. coli* WP2 uvrA' were utilized. We participated in this international challenge project by providing results generated using in-house deep learning (DL)-based consensus approach, called DeepAmes. Most reported QSAR-based Ames test predictive models use a single machine learning (ML) or DL method (Kumar et al., 2021; Vian et al., 2019; Xu et al., 2012). However, ensemble methods or consensus approaches gained by combining predictions from multiple algorithms have been demonstrated to deliver improved predictive performance (Sagi and Rokach, 2018). Our in-house DeepAmes method, is a DL version of the consensus approach where the prediction results from multiple models are integrated with a DL architecture. The framework was successfully applied in classifying drug-induced liver injury (Li et al., 2020) and carcinogenicity (Li et al., 2021) potential risks for chemical compounds.

In this study, we compared DeepAmes with five standard ML methods using the same training and test sets. These ML methods were k-nearest neighborhood (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGBoost), covering a broad range of complexity (Wu et al., 2021). Early observations were that all the ML algorithms were capable of generating a statistically comparable model where there was a comparable goodness-of-fit, but they often displayed some significant differences when challenged with a blind test set. Therefore, our comparative analysis in this paper is specifically focused on the performance of the test set with two important measurements that are critical for a model to be used in regulatory applications; these are applicability domain (AD) and context-of-use.

We developed models using six algorithms on the training set of >10,000 compounds and the performance of these models on a test set of >1,500 compounds were compared. We compared the prediction performance in distinguishing mutagenic from non-mutagenic molecules, where the Matthews correlation coefficient (MCC) was applied since the dataset was unbalanced with 15/85 positives versus negatives; MCC is a more reliable metric than accuracy in binary classification evaluations for an unbalanced dataset (Chicco and Jurman, 2020). The comparative analysis specifically focused on AD and sensitivity to provide insights into context of specifically in regulatory applications. Our results demonstrate that DeepAmes can predict the results of Ames tests with a defined AD and clear context of use with potential for use in FDA regulatory decision making.

2. Materials and methods

2.1. Ames data set and its preparation

The Ames data set used in this study was provided by Division of Genetics and Mutagenesis, National Institute of Health Sciences of Japan (DGM/NIHS) and had three class of mutagenicity: strong mutagenic, mutagenic, and non-mutagenic. Strong mutagenic indicated that a chemical generally induced more than 1000 revertant colonies per milligram of at least one Ames test strain in the presence or absence of rat S9. Mutagenic indicated that a chemical induced at least a 2-fold increase in revertant colonies (but less than strong mutagenic compounds) compared to the negative control in at least one Ames strain in the presence or absence of rat S9. Non-mutagenic indicated that a chemical neither belonged to strong mutagenic nor mutagenic categories. This study was focused on binary classification. Thus, compounds were grouped with strong mutagenic and mutagenic categories as positives and the non-mutagenic category as negatives. Prior to model development, it was extremely important to clean the data to fit for QSAR analysis (Vian et al., 2019). In this study, data were cleaned by removing inorganic compounds, salts, and compounds with molecular weights greater than 700 (only focused on small molecules).

The training set consisted of 10,026 compounds with 1,480 positives and 8,546 negatives, yielding a positive prevalence of 14.8%. The test set consisted of 1,543 compounds and was provided blinded with no class label information until data were exposed to five standard ML methods. The test set was also preprocessed using the same approach as for the training set.

2.1.1. External validation set

In this study, we acquired a commonly known benchmark Ames dataset(Hansen et al., 2009) to serve as an external validation set. There were 175 compounds of this dataset that overlapped with our training set. Compared the mutagenicity call between the two datasets, we found that 25 of which had opposite mutagenicity call; 21 negatives in our dataset were called as positives by the benchmark dataset while only 4 compounds were another way around. The concordance between two dataset is 86%. These 175 compounds were subsequently removed. We also excluded additional 83 compounds with a molecular weight exceeding 700. As a result, the external validation set comprised 6,254 compounds, of which 3,412 were positive compounds while 2,842 were negative compounds.

2.2. Molecular descriptors

Mold2 (https://www.fda.gov/science-research/bioinformatics-tools/mold2) was used to calculate 777 chemical-physical 1D/2D descriptors, utilizing the compounds' structure description file (SDF) for the training, test sets and external validation set (Hong et al., 2008). Based on the training set, we first removed the descriptors with zero variance, and then if any two descriptors had a pairwise correlation coefficient greater than $|0.9|$, we kept one descriptor. In total, 381 descriptors were kept for model development.

2.3. Model development

Five ML methods were selected in this study based on their uniqueness in algorithm and explainability; these were KNN, LR, SVM, RF, and XGBoost. KNN is a non-parametric algorithm that classifies new compounds based on the labels of k nearest neighbors (Duda and Hart, 2006). LR outputs a probability value to indicate new compounds belonging to a certain class by a logistic function of multiple independent variables and a dependent variable (Cox, 1958). SVM uses kernel functions projecting data from low-dimensional space to high-dimensional space and then builds a hyperplane to classify new compounds into different categories (Noble, 2006). RF determines new compounds by majority voting of a large number of decision trees built with subsets of training samples and features (Svetnik et al., 2003). XGBoost makes decisions of new compounds' class through gradient boosted decision trees (Chen and Guestrin, 2016). We developed predictive models using each of these algorithms based on a well-established process. In addition, we also used five algorithms to generate a pool of models for the development of DeepAmes, also described below. The models from these five ML methods were developed with python using the package of Scikit-learn (Pedregosa et al., 2011). The DeepAmes model was also developed with python using both Scikit-learn and TensorFlow (Abadi et al., 2016). The code of the models described in this study are available at https://github.com/TingLi2016/DeepAmes.

2.3.1. DeepAmes

The DeepAmes model was developed based on a published method for the study of drug-induced liver injury (Li et al., 2020), which is a DL powered ensemble framework using model-level representation. The detail of the modeling strategy is described elsewhere (Li et al., 2020).

Briefly, the framework consisted of base classifiers and a meta classifier, where the outputs from base classifiers (model-level representation) were the input for the meta classifier. There were 100 base classifiers associated with each of five algorithms (i.e., KNN, LR, RF, SVM, XGBoost). For each algorithm, we took 8,355 compounds of the training set to develop 100 base classifiers, where a hyperparameter optimization using a grid search with a bootstrap aggregating strategy based on 80/20 split were employed (Breiman, 1996) to create 100 base classifiers with the optimal parameters. Thereafter, the 100 classifiers for each algorithm were ranked by MCC. To prevent overfitting or underfitting, we only kept the classifiers with their MCCs ranging from 5 to 95 percentile of the 500 models as the optimized base classifiers to generate the model-level representation. The value "5%" was chosen to align the common practice of using 5% as a threshold to define outliers as well as the statistical significance threshold (p-value) that often set at 0.05.

The generated model-level representation was fed into a three-layer neural network, which served as a meta-classifier to optimize the base classifiers' information for the Ames test prediction. The rest of 1,671 compounds of the training set were used to train the meta classifier. In the meta-classifier, the number of neurons in the input layer was the same as the number of selected classifiers, while the hidden layer, and output layer consisted of 32 and 1 neuron, respectively. The batch size of 32 was based on a study conducted by Masters et al. who demonstrated that a batch size of 32 consistently yields optimal performance in deep learning architecture (Masters and Luschi, 2018). Activation function was Rectified Linear Unit (ReLU) and the optimization function was stochastic gradient descent (SGD) with the learning rate of 0.1. The class weight of positive compounds was set to six, as the number of positive compounds is 1 over six to the number of negative compounds. The total training epochs was set to 100 and early stop was applied if the loss was not improved in five epochs.

2.3.2. Five ML models

For a fair comparison, we applied the same training set (i.e., 10,026 compounds) to develop five ML models from KNN, LR, RF, SVM, and XGBoost, respectively. Then, these models were evaluated on the same test set; of note, the test set was blinded when DeepAmes was developed but we had the label information when these five ML models were developed. We applied the same hyperparameter searching strategy mentioned above to obtain the optimal parameters based on which the final models were developed using the entire training set. The final optimized hyperparameters are presented in Supplementary Table S1.

2.4. Applicability domain (AD)

To assure the confidence of assessing a new compound, the Organization for Economic Cooperation and Development (OECD), which includes FDA representatives, proposed guidelines to include the applicability domain in QSAR model development (OECD, 2014), especially in support of regulatory applications. One of the recommended approaches to define applicability domain is a similarity-based approach using the Euclidean distance. In our analysis, the Euclidean distance between a pair of closest neighbors is calculated based on the 381 descriptors used for model development for every compound in the training set, from which the median value of Euclidean distances defines the boundary of the applicability domain. For a compound in the test set, we calculated its Euclidean distance to its closest neighbor from the training set, which was used to define how close the compound was from the test set to the applicability domain. If its distance was smaller than a defined applicability domain boundary, the compound was considered to be within the applicability domain; otherwise, the compound was considered outside of the applicability domain.

2.5. Statistical metrics

We provided seven statistical metrics to assess the model performance. As the compounds used for model development had an imbalanced distribution of positives (~15%) and negatives (~85%), we largely relied on the MCC to assess performance which had a range between −1 and 1; a high score towards 1 indicated that a binary classifier correctly predicted the majority of both positives and negatives (Chicco and Jurman, 2020; Chicco et al., 2021). The other six metrics included the area under the receiver operating characteristic (ROC) curve (range between 0 and 1 and the higher AUC a better prediction is), sensitivity, specificity, accuracy, balanced accuracy, and F1 score. The formulas used to calculate these metrics are listed as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (1)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

$$BA = \frac{sensitivity + specificity}{2} \quad (5)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (6)$$

where the TP (True Positive) indicates the Ames result of a chemical is positive in experiment and predicted as positive; TN (True Negative) indicates the Ames result as negative and predicted as negative as well; FP (False Positive) and FN (False Negative) measures the chemicals that are wrongly predicted by a model in terms of positive and negative, respectively.

3. Results

The performance of the six models on the test set of 1,543 compounds is plotted in Fig. 1 and summarized in Table 1. These models performed in an order of DeepAmes > XGBoost > LR > SVM > RF > KNN with respect to MCC which varied significantly with >70% increasefrom the lowest (KNN = 0.22) to the highest (DeepAmes = 0.38) MCC. The difference between the top performer (DeepAmes) and the second performer (XGBoost) was >10%.

DeepAmes and the other five ML methods were applied to the external validation set, consisting of 6,254 compounds, and the results were summarized in Table 2. In terms of MCC, the models ranked as follows: DeepAmes > LR > SVM > XGBoost > RF > KNN. DeepAmes had the best performence with the MCC that were the same between the test set and the external validation set. The DeepAmes predictions of the external validation set are included in the Supplementary Table S2.

3.1. Applicability domain (AD)

The training domain was defined by all the compounds used to develop the model while applicability domain (AD) defines reliability and confidence in predicting compounds which could be within the training domain or beyond. The further the expansion beyond the training domain, the better the model is in respect to AD.

In this study, we followed the OECD's guideline to calculate the AD. Specifically, the Euclidean distances of all the pairs of closest neighbors were first calculated for all the training compounds which ranged from 0.42 to 2.61 x10^{11}. The median value 62.70 of the Euclidean distances in the training domain was used as the threshold to define the applicability domain. Then, the Euclidean distances for the test set of 1,543 compounds to the training set was calculated which also exhibited a wide range varying between 2.81 and 8.73x10^{9}. There were 781 compounds

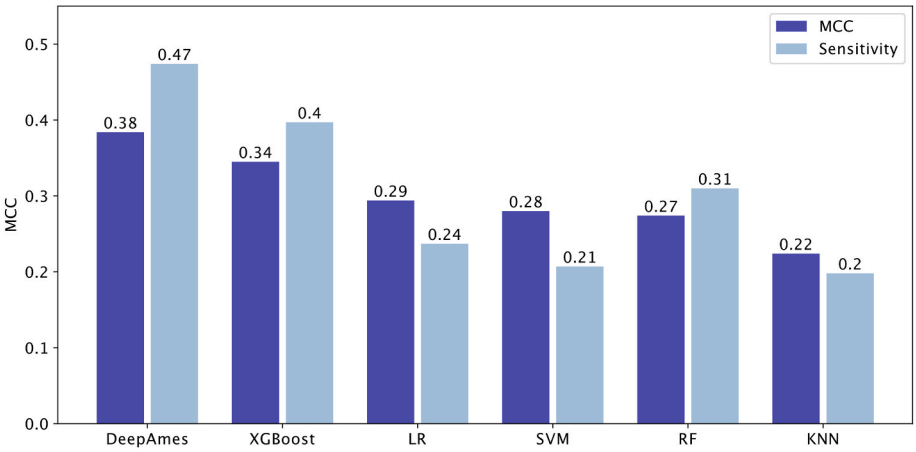

Fig. 1. MCC and Sensitivity of the six models on the test set.

Table 1
Six models performance on the test set.

Name	MCC	Accuracy	AUC	Sensitivity	Specificity	F1	BA
DeepAmes	0.38	0.84	0.74	0.47	0.91	0.48	0.69
XGBoost	0.34	0.84	0.79	0.40	0.92	0.43	0.66
LR	0.29	0.86	0.79	0.24	0.97	0.33	0.60
SVM	0.28	0.86	0.78	0.21	0.97	0.30	0.59
RF	0.27	0.84	0.75	0.31	0.93	0.36	0.62
KNN	0.22	0.84	0.66	0.20	0.96	0.28	0.58

Table 2
Six models performance on the external validation set.

Name	MCC	Accuracy	AUC	Sensitivity	Specificity	F1	BA
DeepAmes	0.38	0.68	0.73	0.61	0.78	0.68	0.69
LR	0.33	0.63	0.76	0.41	0.89	0.54	0.65
SVM	0.32	0.61	0.78	0.37	0.91	0.51	0.64
XGBoost	0.29	0.62	0.72	0.45	0.82	0.57	0.64
RF	0.20	0.57	0.66	0.33	0.85	0.45	0.59
KNN	0.16	0.53	0.63	0.22	0.90	0.34	0.56

inside the training domain while 762 compounds were outside of the domain defined by the threshold value of 62.70. The prediction performance of six models for the compounds that were both inside and outside of the domain are summarized in Supplementary Table S3 and the MCC performance is plotted in Fig. 2A. All the models performed better for the compounds that were within the AD compared to these outsides of the domain, among which DeepAmes yielded the highest MCC in both within and outside of the domain.

We calculated the difference in prediction accuracy (i.e., MCC) between within and outside of the domain and normalized by (divided by) the within domain accuracy (i.e., MCC) as a composite score to assess a model's overall performance when the training domain is defined as the AD. The lower composite score indicates a better performance of a model when considering the compounds were both within and outside of the AD. As depicted in Fig. 2A, DeepAmes yielded the best performance with the lowest composite score.

As generally expected, a model's performance would decrease if a test compound furthered away from the training domain. This distance-dependent drop in prediction accuracy measures a model's AD with respect to its potential to provide accurate prediction for the compounds outside of the training domain. We conducted a comparative analysis of applicability domain for six models by evaluating their performance on the compounds away from the training domain in every 5% incremental degree. As summarized in Fig. 2B and Supplementary Table S3, the MCC measures of all six models were generally decreased as the compounds furthering away from the training domain. The biggest drop was found on the compounds beyond 45% away from the training domain. DeepAmes held the highest MCC until 30% beyond the training domain, while both KNN and SVM models hold a bit longer in applicability domain (35% and 40%, respectively).

3.2. *DeepAmes model's sensitivity*

The training set had an imbalanced ratio of positive versus negative (i.e., ~15/85). As expected, the sensitivity of all six models was relatively low as presented in Fig. 1 for the test set. The rank by sensitivity followed a trend of DeepAmes > XGBoost > RF > SVM > LR > KNN, which was in opposite to the modeling complexity of the algorithm. DeepAmes yielded the highest sensitivity, way above the prevalence of the training set. More specifically, DeepAmes achieved the highest sensitivity, with over 100%, and 15% increase compared to the lowest sensitive model (KNN) and the second largest sensitive model (XGBoost), respectively (Fig. 1).

To explore different weight configurations and assess how they influenced the model's performance (Ho and Wookey, 2019), as such that a specific weight can be adopted for specific application, such as regulatory application, we revised the DeepAmes model using the same framework of original approach. Specifically, we penalized with a higher "weight" on a wrong prediction for the positive class compared to the negative one. Since the number of negative compounds was six times

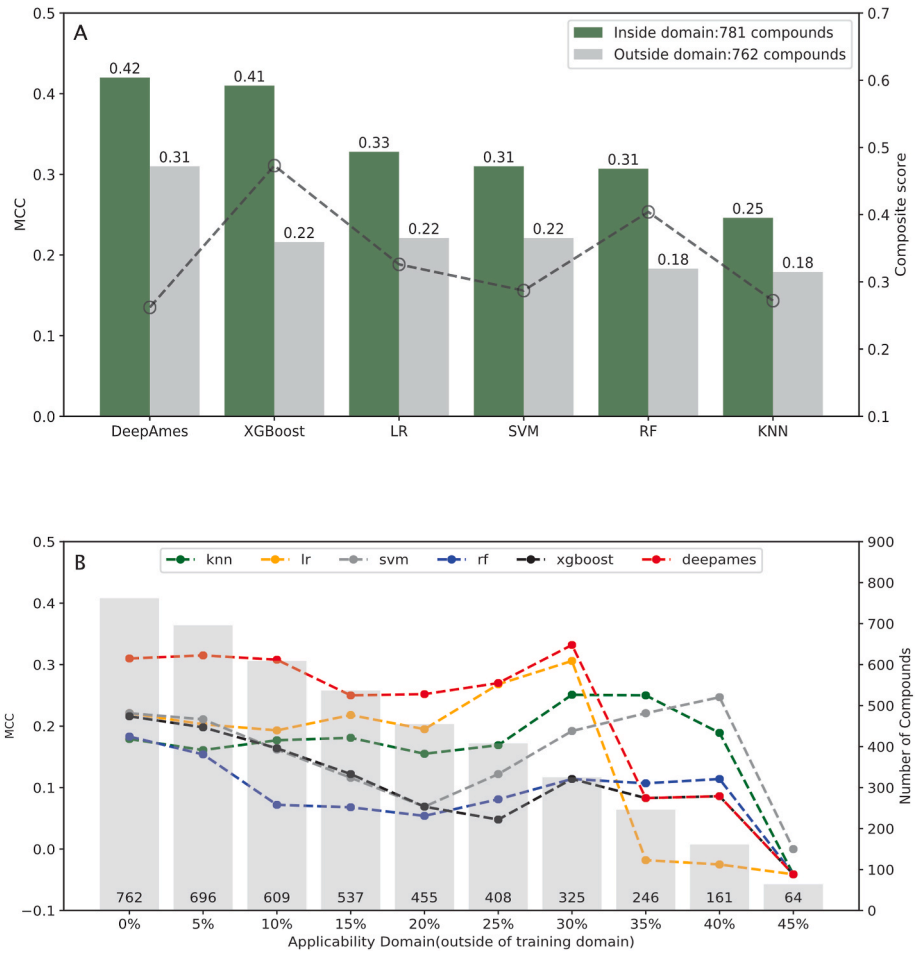

Fig. 2. **A**: The MCC comparison between within and outside of the training domain on the test set. **B**: The MCC distribution of the six models on the compounds away from the training domain in every 5% incremental degree.

of the positive compounds, we varied the weight by multiplying the prediction value with a number between seven and 18. We noticed that, out of 12 models with a weighted penalty, 11 achieved a better sensitivity compared to the original mode (Fig. 3 and Supplementary Table S4). One of the revised DeepAmes models achieved a sensitivity as high as 0.87 with weight = 16, which was an 85.1% ((0.87–0.47)/0.47 * 100%) improvement compared to the original model sensitivity (0.47).

4. Discussion

We developed DeepAmes to predict mutagenicity using a large and consistent dataset. We compared DeepAmes with five standard ML methods covering a broad range of algorithmic complexity and explainability. The comparative analysis was specifically conducted on a test set of over 1,500 compounds that was blinded when DeepAmes was developed. DeepAmes achieved the highest performance in prediction

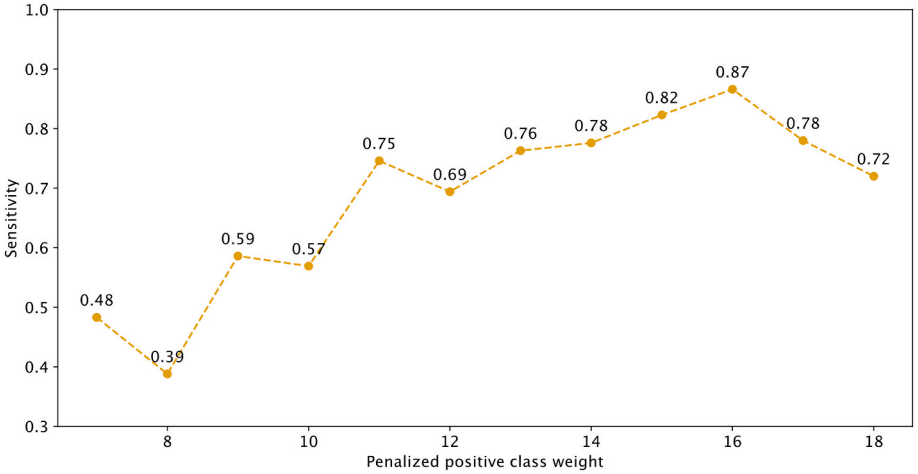

Fig. 3. The plot of Sensitivity of DeepAmes against the positive class penalized weight on the test set.

accuracy (MCC) with superior applicability domain. The key reason for this comparative analysis was to identify a model with potential for FDA regulatory application, where context-of-use is of concern. In the drug review process, context-of-use focuses on sensitivity (the rate of false negatives) since these are of more concern; false positives could be eliminated by downstream analysis via experimental methods. DeepAmes exhibited superior performance in sensitivity compared with other models.

In addition to the test set, we further evaluated the DeepAmes model with an external dataset and compared its performance against five other conventional ML methods. DeepAmes achieved the highest MCC among all six methods and comparable to the results from the test set. It is worthwhile to emphasize that there were 175 compounds overlapped between this external dataset and our training set, and the experimental concordance between two datasets for mutagenicity call was 86%. With respect to this fact, the DeepAmes performance actually was higher than what was summarized in Table 2. More specifically, compared to the mutagenicity call between two datasets, we found that 25 of which had opposite mutagenicity call; 21 negatives in our dataset were called as positives by the benchmark dataset while only 4 compounds were another way around. That indicates that our model prediction tended to generate more false negatives (i.e., lower sensitivity) as judged by the benchmark dataset. In other words, the actual sensitivity could be higher with using the mutagenicity call by the Japanese method as a reference. This also explained DeepAmes yielded the highest specificity but lowest sensitivity in comparison to the original study, which utilized seven methods in a 5-fold cross-validation.

The size and the quality of the dataset used in a QSAR study could have a significant impact on the performance of developed models (Cherkasov et al., 2014; Rácz et al., 2021). A larger and high-quality dataset can provide more useful information for the model to learn from, which may lead to better predictions. In this study, we used a large dataset, consisting of more than 10,000 chemical compounds where Ames test results were generated under a consistent test guideline. In comparison to the studies with small datasets ranging from hundreds to thousands (Honma et al., 2020; Kasamatsu et al., 2021; Vian et al., 2019; Xu et al., 2012), this dataset could provide a better assessment of algorithmic superiority of various methods. Moreover, this dataset was collected under a standard test guideline, giving advantages over datasets of similar size but collected by combining multiple resources tested under different guidelines (Hung and Gini, 2021; Landry et al., 2019). For example, Landry et al. curated a large data set of 13,514 compounds from various resources, including FDA approval packages and other regulatory authorities, online repositories of genetic toxicology data (e.g., NTP, EPA GENE-TOX, and CCRIS), data sharing efforts, published literature, and MultiCASE and Leadscope internal databases (Landry et al., 2019).

Applicability domain was explored for all these six models using an OECD recommended method. We found that all the models had better predictive ability on the compounds inside the applicability domain than outside of the applicability domain, as expected. However, when we investigated the distance-dependent drop in prediction accuracy (i.e., MCC) to assess prediction accuracy for compounds outside of the training domain, the biggest drop was found for all the models was 45% from the AD. This is consistent with the common understanding that predictive ability decreases as predicted compounds are more dissimilar to the compounds in the training set. DeepAmes presented the highest MCC for the compounds and remained robust until test compounds were >30% different from the AD.

We also demonstrated that models can be developed based on a specific application. For example, we revised DeepAmes by considering content-of-use (e.g., high sensitivity) in regulatory application. By weighting to compensate the preponderance of the positive, we were able to increase the sensitivity to 0.87. This is quite a notable increase in sensitivity considering that the prevalence of positives in the data set used for developing and testing the model was less than 15%. This approach should be equally applicable to develop models that meet specification of content-of-use such as improving specificity instead of sensitivity.

Data preprocessing (such as removing inorganic compounds) is a required step before developing QSAR models(Fourches et al., 2010). We investigated the chemical spatial distribution of molecule weight (MW), and LogP on both the training set and test set. The MW of most compounds is between 100 and 600 and LogP is mainly between −2 and 7 (Supplementary Fig. S1). Therefore, our results should be interpreted in this context. In addition, all of the Ames data used in this study was collected under the same test guideline, so the reported predictive power is only directly applicable within this guideline. Caution should be exercised when making predictions for datasets conducted outside of this such as under OECD TG471, the guideline most commonly used today.

This study applied six ML algorithms, including KNN, LR, SVM, RF, XGBoost and DeepAmes. The architecture of these six methods varied from simple to complex, with a decrease of explainability (Wu et al., 2021). All those six models were evaluated on the test set. These models performed in an order of DeepAmes > XGBoost > LR > SVM > RF > KNN with respect to MCC. KNN is an intuitive method but yields the lowest MCC. Meanwhile, DeepAmes was the most complicated method but achieved the highest MCC. The results imply that there is a challenge to balance explainability with predictive performance in a single modeling approach, which requires further investigation and study to improve the explainability of DeepAmes.

In summary, our comparative analysis showed that DeepAmes, and particularly the revised DeepAmes models could be a valuable tool during regulatory review for the potential mutagenicity of drugs, drug impurities and food additives together with environmental, and industrial chemicals.

Disclaimer

This manuscript reflects the views of the authors and does not necessarily reflect those of the Food and Drug Administration (FDA). Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

Funding

This research is in part funded by the FY2022 CDER Safety Research Interest Group (SRIG) grant as a collaborative initiative between FDA's Center of Drug Evaluation and Research's Office of Computational Sciences and National Center for Toxicological Research's Division of Bioinformatics and Biostatistics.

CRediT authorship contribution statement

Ting Li: collected the datasets, performed the model development and computational analysis, wrote and revised the manuscript. **Zhichao Liu:** wrote and revised the manuscript. **Shraddha Thakkar:** wrote and revised the manuscript. **Ruth Roberts:** wrote and revised the manuscript. **Weida Tong:** conceived and designed the study. All authors edited and approved the final manuscript.

Declaration of competing interest

RR is co-founder and co-director of ApconiX, an integrated toxicology and ion channel company that provides expert advice on non-clinical aspects of drug discovery and drug development to academia, industry, and not-for-profit organizations.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

The authors do not have permission to share data.

Acknowledgments

We thank the Division of Genetics and Mutagenesis, National Institute of Health Sciences of Japan for providing the dataset for this study. RR is grateful to the contract program with NCTR for the support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2023.105486.

References

Abadi, M., et al., 2016. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems arXiv preprint arXiv:1603.04467.
Ames, B.N., et al., 1975. Methods for Detecting Carcinogens and Mutagens with the Salmonella/mammalian-microsome Mutagenicity Test. Mutat. Res., p. 31 (Netherlands).
Benigni, R., Bossa, C., 2011. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. Chem. Rev. 111, 2507–2536.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Cassano, A., et al., 2014. Evaluation of QSAR models for the prediction of ames genotoxicity: a retrospective exercise on the chemical substances registered under the EU REACH regulation. J. Environ. Sci. Health Part C 32, 273–298.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Cherkasov, A., et al., 2014. QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57, 4977–5010.
Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 21, 1–13.
Chicco, D., et al., 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 14, 1–22.
Cox, D.R., 1958. The regression analysis of binary sequences. J. Roy. Stat. Soc. B 20, 215–232.
Duda, R.O., Hart, P.E., 2006. Pattern Classification. John Wiley & Sons.
Fourches, D., et al., 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model. 50, 1189.
Greene, N., et al., 1999. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ. Res. 10, 299–314.
Hansen, K., et al., 2009. Benchmark data set for in silico prediction of Ames mutagenicity. J. Chem. Inf. Model. 49, 2077–2081.
Hanser, T., et al., 2014. Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. J. Cheminf. 6, 1–21.
Ho, Y., Wookey, S., 2019. The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. IEEE Access 8, 4806–4813.
Hong, H., et al., 2008. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model. 48, 1337–1344.
Honma, M., et al., 2019. Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis 34, 3–16.
Honma, M., et al., 2020. Screening for Ames mutagenicity of food flavor chemicals by (quantitative) structure-activity relationship. Gene Environ. 42, 1–6.
Hung, C., Gini, G., 2021. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Mol. Divers. 25, 1283–1299.
Kamber, M., et al., 2009. Comparison of the Ames II and traditional Ames test responses with respect to mutagenicity, strain specificities, need for metabolism and correlation with rodent carcinogenicity. Mutagenesis 24, 359–366.
Kasamatsu, T., et al., 2021. Development of a new quantitative structure–activity relationship model for predicting Ames mutagenicity of food flavor chemicals using StarDrop™ auto-Modeller. Gene Environ. 43, 1–17.
Kazius, J., et al., 2005. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48, 312–320.
Klopman, G., 1984. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J. Am. Chem. Soc. 106, 7315–7321.
Klopman, G., 1992. MULTICASE 1. A hierarchical computer automated structure evaluation program. Quant. Struct.-Act. Relat. 11, 176–184.
Kumar, R., et al., 2021. A deep neural network–based approach for prediction of mutagenicity of compounds. Environ. Sci. Pollut. Control Ser. 28, 47641–47650.
Lahl, U., Gundert-Remy, U., 2008. The use of (Q) SAR methods in the context of REACH. Toxicol. Mech. Methods 18, 149–158.
Landry, C., et al., 2019. Transitioning to composite bacterial mutagenicity models in ICH M7 (Q) SAR analyses. Regul. Toxicol. Pharmacol. 109, 104488.
Li, T., et al., 2020. DeepDILI: deep learning-powered drug-induced liver injury prediction using model-level representation. Chem. Res. Toxicol. 34, 550–565.
Li, T., et al., 2021. DeepCarc: deep learning-powered carcinogenicity prediction using model-level representation. Front Artif Intell 4, 757780.
Masters, D., Luschi, C., 2018. Revisiting Small Batch Training for Deep Neural Networks, 07612 arXiv preprint arXiv:1804.
McCann, J., et al., 1975. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. Proc. Natl. Acad. Sci. USA 72, 5135–5139.
Mekenyan, O., et al., 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.
Mortelmans, K., Zeiger, E., 2000. The Ames Salmonella/microsome mutagenicity assay. Mutat. Res.Fundam. Mol. Mech. Mutagen. 455, 29–60.
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567.
OECD, 2014. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models.
Pavan, M., Worth, A., 2008. Publicly-accessible QSAR software tools developed by the joint research centre. SAR QSAR Environ. Res. 19, 785–799.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Rácz, A., et al., 2021. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules 26, 1111.
Roberts, G., et al., 2000. LeadScope: software for exploring large sets of screening data. J. Chem. Inf. Comput. Sci. 40, 1302–1314.
Sagi, O., Rokach, L., 2018. Ensemble learning: a survey. Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 8, e1249.
Sanderson, D., Earnshaw, C., 1991. Computer prediction of possible toxic action from chemical structure; the DEREK system. Hum. Exp. Toxicol. 10, 261–273.
Schwab, C., et al., 2016. A reliable workflow for in silico assessment of genetic toxicity and application to pharmaceutical genotoxic impurities. Elsevier Ireland ltd Elsevier house, brookvale plaza, east park shannon, co … Toxicol. Lett. 258. S59-S59.
Serafimova, R., et al., 2007. Identification of the structural requirements for mutagencitiy, by incorporating molecular flexibility and metabolic activation of chemicals. II. General Ames mutagenicity model. Chem. Res. Toxicol. 20, 662–676.
Svetnik, V., et al., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.
Vian, M., et al., 2019. In silico model for mutagenicity (Ames test), taking into account metabolism. Mutagenesis 34, 41–48.
Wu, L., et al., 2021. Trade-off predictivity and explainability for machine-learning powered predictive toxicology: an in-depth investigation with Tox21 data sets. Chem. Res. Toxicol. 34, 541–549.
Xu, C., et al., 2012. In silico prediction of chemical Ames mutagenicity. J. Chem. Inf. Model. 52, 2840–2847.