Computers in Biology and Medicine 178 (2024) 108731

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

DeepRA: A novel deep learning-read-across framework and its application in non-sugar sweeteners mutagenicity prediction

Tarapong Srisongkram *

Division of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Khon Kaen University, 40002, Thailand

ARTICLE INFO

Keywords:
Read-across
Convolutional neural network
Artificial sweeteners
Mutagenicity
Machine learning
Stacking ensemble learning
Deep learning

ABSTRACT

Non-sugar sweeteners (NSSs) or artificial sweeteners have long been used as food chemicals since World War II. NSSs, however, also raise a concern about their mutagenicity. Evaluating the mutagenic ability of NSSs is crucial for food safety; this step is needed for every new chemical registration in the food and pharmaceutical industries. A computational assessment provides less time, money, and involved animals than the *in vivo* experiments; thus, this study developed a novel computational method from an ensemble convolutional deep neural network and read-across algorithms, called DeepRA, to classify the mutagenicity of chemicals. The mutagenicity data were obtained from the curated Ames test data set. The DeepRA model was developed using both molecular descriptors and molecular fingerprints. The obtained DeepRA model provides accurate and reliable mutagenicity classification through an independent test set. This model was then used to examine the NSSs-related chemicals, enabling the evaluation of mutagenicity from the NSSs-like substances. Finally, this model was publicly available at https://github.com/taraponglab/deepra for further use in chemical regulation and risk assessment.

1. Introduction

Humanity has always preferred sweet taste, starting from infants to young adults [1,2]. The first recorded sweetener was honey; it has been used since ancient Greece and China [3]. Only then it was replaced by sucrose, a common table sugar, that originated from sugar cane. Furthermore, sugar cane has been substituted by beet sugar due to a shortage of sugar cane during World War II. Saccharin is the first artificial sweetener discovered in 1879; it is well-known for its bitter aftertaste, affordable price, and low-calorie sweetener [4]. Artificial sweeteners refer to non- or low-calorie non-sugar sweetener (NSS) substances that consist of two generations: first generation (i.e., saccharin and aspartame) and second generation (i.e., acesulfame-K, sucralose, alitame, and neotame). The difference between those NSSs established a wide range of chemical space properties of artificial sweeteners. Novel NSSs are developed to improve the bitter aftertaste of the first artificial sweeteners and support an obesity-profitable sugar--free market [5].

Saccharin, however, was found to be linked with bladder cancer risk according to the World Health Organization (WHO) recent meta-analysis [6]. Moreover, aspartame was also categorized as possibly carcinogenic to humans according to the International Agency for Research in Cancer (IRAC) [7]. These NSSs raise concerns about whether the NSS substances are safe for human consumption [8]. Moreover, the NSS substances emphasize the importance of mutagenicity evaluation, such as DNA-damaging screening assay, for qualitatively assessing the potential carcinogenicity of the chemicals [9]. New chemicals used in foods or pharmaceutical products are subjected to assess their mutagenicity effects; this is to limit the potential carcinogenicity risk from these particular products [10]. Therefore, evaluating the mutagenicity risk of NSS-like substances is necessary before utilizing these chemicals in the end consumer products.

Assessing mutagenicity recommended by the Organisation for Economic Cooperation Development (OECD) composts with two methods: 1) *in vitro* assays (i.e., bacteria reverse mutation test (Ames test) [11], mammalian chromosome aberration test [12], mammalian cell micronucleus test [13], mammalian cell gene mutation test using the Hprt-Xprt genes [14], and thymidine kinase gene [15]) and 2) *in vivo* methods (i.e., genotoxicity test, alkaline Comet test [16], transgenic rodent somatic and germ cell mutation assay [17], and rodent dominant lethal test [18]). The combination of those *in vitro* and *in vivo* assays can answer various aspects of genotoxicity; however, performing these experiments would be costly, time-consuming, and require a high number of animal testing. Therefore, computational methods were also

* Division of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen, 40002, Thailand.
E-mail address: tarasri@kku.ac.th.

https://doi.org/10.1016/j.compbiomed.2024.108731
Received 5 April 2024; Received in revised form 7 May 2024; Accepted 8 June 2024
Available online 12 June 2024

recommended to reduce *in vivo* testing, while enhancing the benefit of existing toxicity data from the literature. There are two methods recommended by the European Chemicals Agency (ECHA) and OECD organizations: read-across (RA) and quantitative structure-activity relationship (QSAR) [19–21].

The read-across (RA) approach is a chemical properties prediction technique that fundamentally relies on similar chemicals that share similar molecular features and should exert similar chemical properties [20]. To date, several regulatory agencies have employed the RA principle in their prediction systems such as the ECHA [19], the U.S. Environmental Protection Agency (EPA) [22], and the Japan Chemical Substances Control Law (CSCL) [23]. Generally, the RA framework is based on a qualitative expert-based system; however, it suffers from a prediction bias and high variability, leading to a low generalizability of the RA system [24]. Quantitative RA (qRA) is a systematic quantitative-based RA method to predict chemical properties. This approach is invented to overcome the prediction bias and variability from the expert-based systems by utilizing the molecular features similarity coefficient scores between a new predictor and source chemicals [24]. This algorithm can search for the most similar source chemicals to the new predictor, and then make an output decision based on the molecular properties of the most similar compounds [25]. The qRA models mostly rely on a weighted average of the similarity coefficient of source molecules [24–26]. This method can ensure that activity prediction does not come from a single compound.

The QSAR method is a gold standard computational approach used to predict data gaps in the chemical risk assessment [19,21]. This method relies on a relationship between chemical structure attributes and their activities. The prediction is made after the well-defined output and validated statistical model of the QSAR were developed [21]. The chemical attributes can be varied from a chemical representation such as a Simplified Molecular-Input Line Entry System (SMILES) to molecular descriptors or molecular fingerprints [27]. The mathematical algorithm of the QSAR method can also be varied from a simple multiple linear regression to sophisticated machine learning and deep learning techniques. Integrating both the qRA and QSAR methods into one framework, called quantitative read-across structure-activity relationship, provides advantages in prediction accuracy and model robustness compared to a single qRA or QSAR model [25,26,28]. The method of combining two or more algorithms into one predictor is well-known as a stacking ensemble technique [29].

An stacking ensemble is one of the popular machine learning techniques that combines multiple machine learning predictors into one estimator scheme [30]. This scheme normally composts with two feature layers: molecular features and model-level representation features [31,32]. The first layer predictor takes the molecular features as an input and produces a prediction output. This prediction output will be used as model-level representation features or predictive features in the second layer. The second layer also contains an estimator called a meta-learner. This meta-learner will weigh the stacking prediction outputs from the first layer and make a meta-decision based on those outputs [25,31,32]. Generally, the second layer will only contain the predictive features, however, in this study, we combined both molecular features and predictive features and predicted the outcome based on this new feature scheme. The QSAR that was used in this paper is a deep convolutional neural network (CNN) algorithm. This model was compared to a random forest (RF) estimator, which was frequently used as a benchmark predictor for stacking ensemble learning [25,31–33]. Notably, there is no evidence of stacking ensemble learning using a combination of CNN and qRA models, which prompted us to explore these model combinations in this study.

In this paper, we proposed a novel learning scheme to classify the mutagenicity of chemicals using a stacking ensemble CNN with qRA models, named as the DeepRA framework. We optimized the predictive performance of the new classifier with the three main molecular features: Mordred descriptors [34], extended-connectivity fingerprint (ECFP) [35], and RDKIT fingerprints. We utilized Matthew's correlation coefficient (MCC), balanced accuracy, precision, recall, and F1 score to evaluate the performance of the constructed models. Subsequently, we used this novel framework to predict the potential mutagenicity of non-sugar sweeteners from the sweetener database [36]. We also identified the boundary of application also named as the applicability domain of the model for the proposed use in the future. All the main ideas of this paper are illustrated in Fig. 1.

2. Results

2.1. Chemical distribution of curated mutagenic data set

We started by evaluating the chemical distribution of the curated mutagenic data set obtained from the TOXRIC database [37]. We found that this data set initially contained 7485 compounds with mutagenic toxicity values. We then removed the missing IUPAC name, missing canonical SMILES compounds, the inorganic, and the mixture compounds from the data set as they cannot be identified or computed the molecular fingerprints or descriptors, respectively. We then removed all duplicate entries as they can process different toxicity values, which are unreliable to be used in the model construction. The total of the chemicals in the data set after preprocessing was 6881 molecules. All of these can be divided into mutagen (n = 3781 or 55 %) and non-mutagen (n = 3100 or 45 %), respectively (Fig. 2A).

We further explore the basic chemical space of the mutagenic data set using drug-likeness criteria that predict the oral active molecules in Fig. 2B and C. A molecule that passes this criteria tends to be an oral active drug [38]. The drug-likeness criteria include any molecule that contains molecular weight (MW) less than 500 Da, LogP less than 5, hydrogen-bond acceptor (HBA) less than 10, and hydrogen-bond donor (HBD) less than 5. We found that the mutagenic data set passed the drug-likeness criteria at 6023 chemicals or 88 % of total data points, while only 858 chemicals, or 12 % of the total data points, failed the criteria. The 88 % of data points that passed the drug-like category dictates that this data set is suitable for measuring the mutagenicity of the drug-like compounds. We further evaluate how many mutagenic compounds pass the drug-likeness and can be orally active. We found that 3307 mutagens, or 87 % of the total mutagens, passed the drug-likeness criteria, while only 474 mutagens, or 13 % of the mutagens, failed the drug-likeness criteria. This similar pattern was also observed in the non-mutagens, where 2716 chemicals, or 88 % of non-mutagens, passed the drug-likeness criteria, while 384 chemicals, or 12 % of non-mutagens, failed the drug-likeness criteria. The similar drug-likeness distribution patterns between mutagens and non-mutagens indicate that these drug-likeness criteria cannot delineate the mutagenic and non-mutagenic characteristics. A sophisticated molecular feature is needed to evaluate this mutagenicity endpoint.

2.2. Molecular feature representation

We computed the molecular features of the mutagenic data set using both molecular descriptors and fingerprints. The Mordred descriptors containing 971 curated two-dimension (2D) molecular descriptors were used as molecular descriptors for this study. On the other hand, the ECFP fingerprints containing 4096 bits 8 radius circular and the RDKIT fingerprints containing 2048 bits were used as molecular fingerprints for this study. We then visualized the distributions of the three molecular features using the unsupervised t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. We found that the distribution between the mutagens and non-mutagens can be discriminated better by the ECFP and RDKIT fingerprints than the Mordred descriptors, as observed in the non-overlapping island distributions between orange and blue circles in ECFP and RDKIT, represent the unique chemical space between mutagen and non-mutagen, respectively (Fig. 3, arrow). However, the visualization of the molecular features is still not the definite

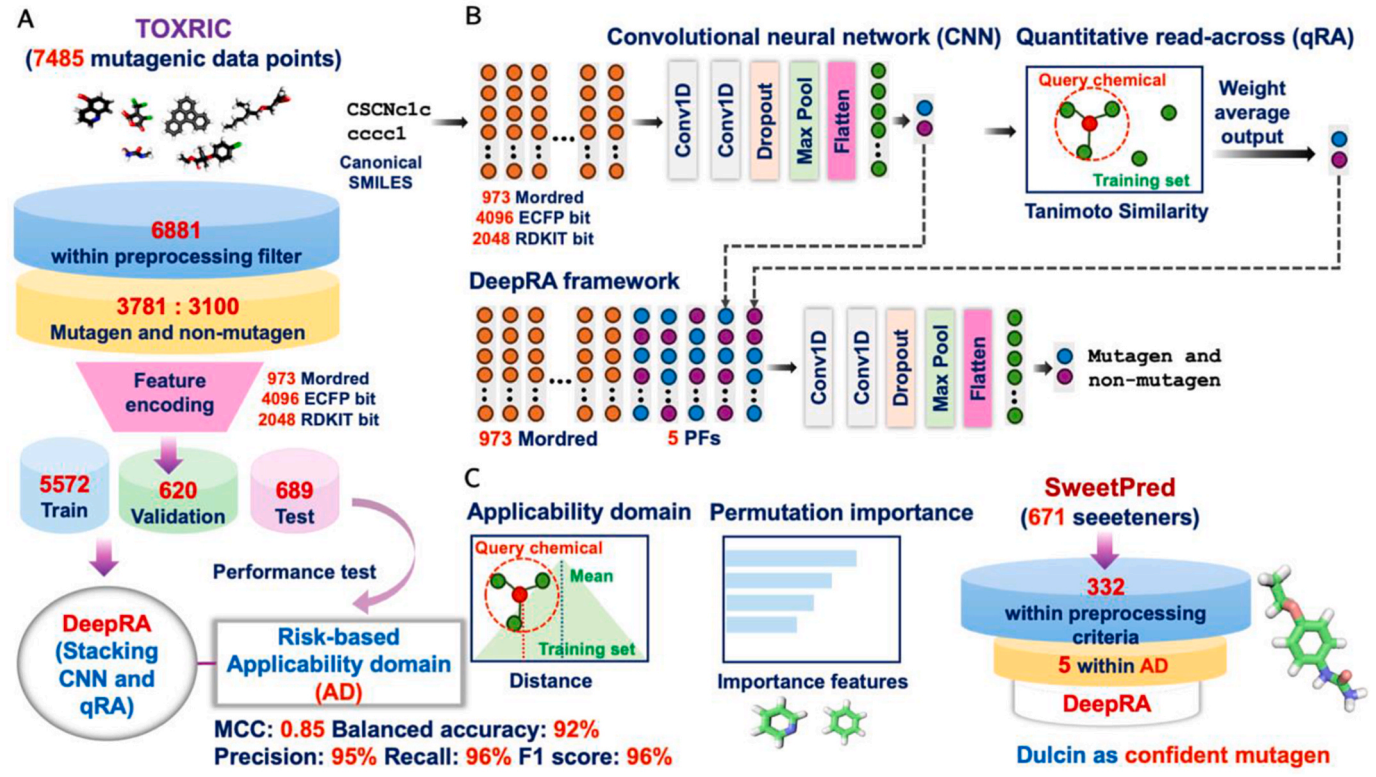

Fig. 1. Schematic diagram and main idea of this study. A) Flowchart of how to build DeepRA model. B) Convolutional neural network architecture (CNN), quantitative read-across (qRA), and tacking ensemble Deep learning with quantitative read-across (DeepRA) framework. C) Risk-based applicability domain, permutation importance, and example of DeepRA application for accessing mutagen of sweeteners.

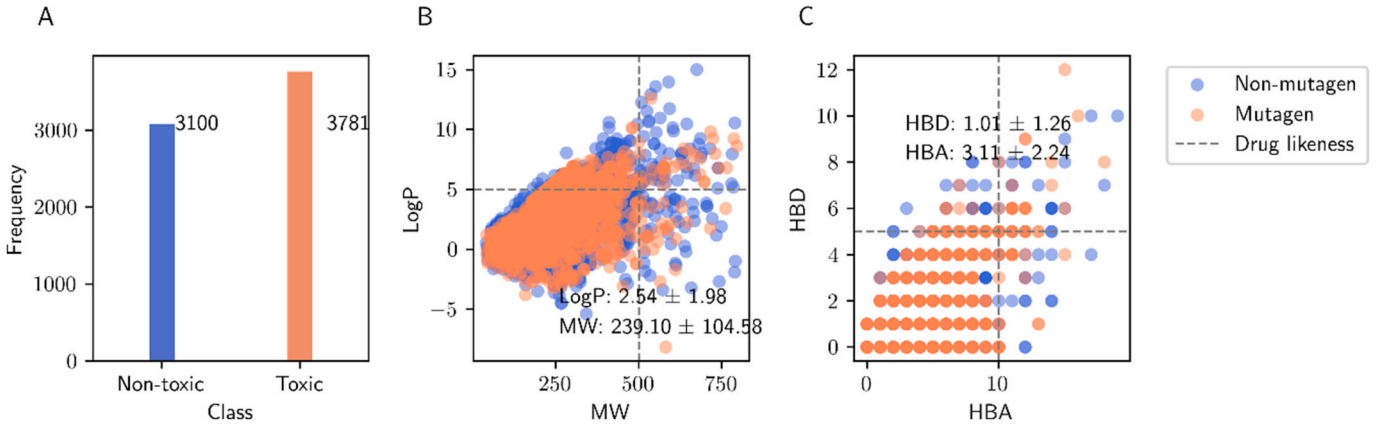

Fig. 2. Chemical distribution of the mutagenic data set. A) Class distribution between non-mutagenic and mutagenic chemicals, B) Distribution between molecular weight (MW) and LogP of the molecules in the mutagenic data set. C) Distribution between hydrogen-bond acceptor (HBA) and hydrogen-bond donor (HBD) of the molecules in the mutagenic data set. Blue and orange circles indicate the non-mutagen and mutagen, respectively. The dashed line represents the drug-likeness threshold.

criteria to omit the Mordred descriptors; the predictive performance of these three features should be a quantitatively assessed using the classification models, as done in the next section.

2.3. *Performance of machine learning models*

We split the mutagenic data set into training and independent test sets using a 9: 1 ratio, which results in 6192 and 689 chemicals, respectively. We subsequently split the training set into training (n = 5572) and validation (n = 620) with the same ratio as the previous split. We then built the RF model from the training set and optimized the model via the validation set. Subsequently, we evaluated the predictive performance of the model using the independent test set.

In the performance test using an independent test set, we found that the RF models had MCC values of 0.38, 0.00, and 0.32 for the Mordred, ECFP, and RKIT features, respectively (Fig. 4). The MCC values of the RF models were lower than 0.5 for all three features, indicating an unacceptably low positive correlation between the predicted and experiment values. The balanced accuracies of the RF models ranged from 0.50 to 0.68 (Fig. 4; Mordred-RF, ECFP-RF, and RDKIT-RF), also indicating poor predictive accuracy. These results suggest that the RF models are unreliable and cannot be used as prediction systems. Furthermore, we found that the precision, recall, and F1 scores ranged from 0.57 to 0.72, 0.76–1.00, and 0.72–0.74, respectively. Those results delineate that the

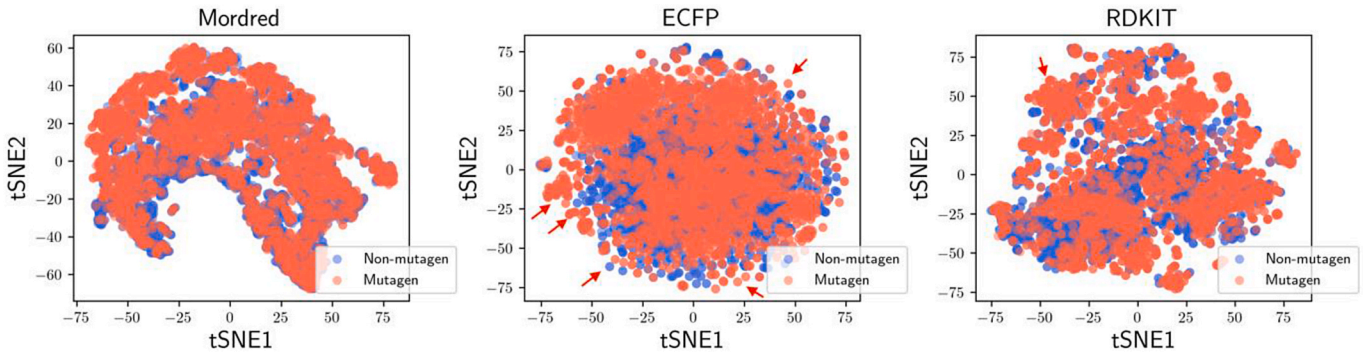

Fig. 3. Molecular features distribution of the mutagenic data set using A) Mordred descriptors, B) extended circular fingerprints (ECFP), and C) RDKIT fingerprints. Blue and orange circles indicate the non-mutagen and mutagen, respectively. Red arrows indicate the unique non-overlap island of chemicals in each group.

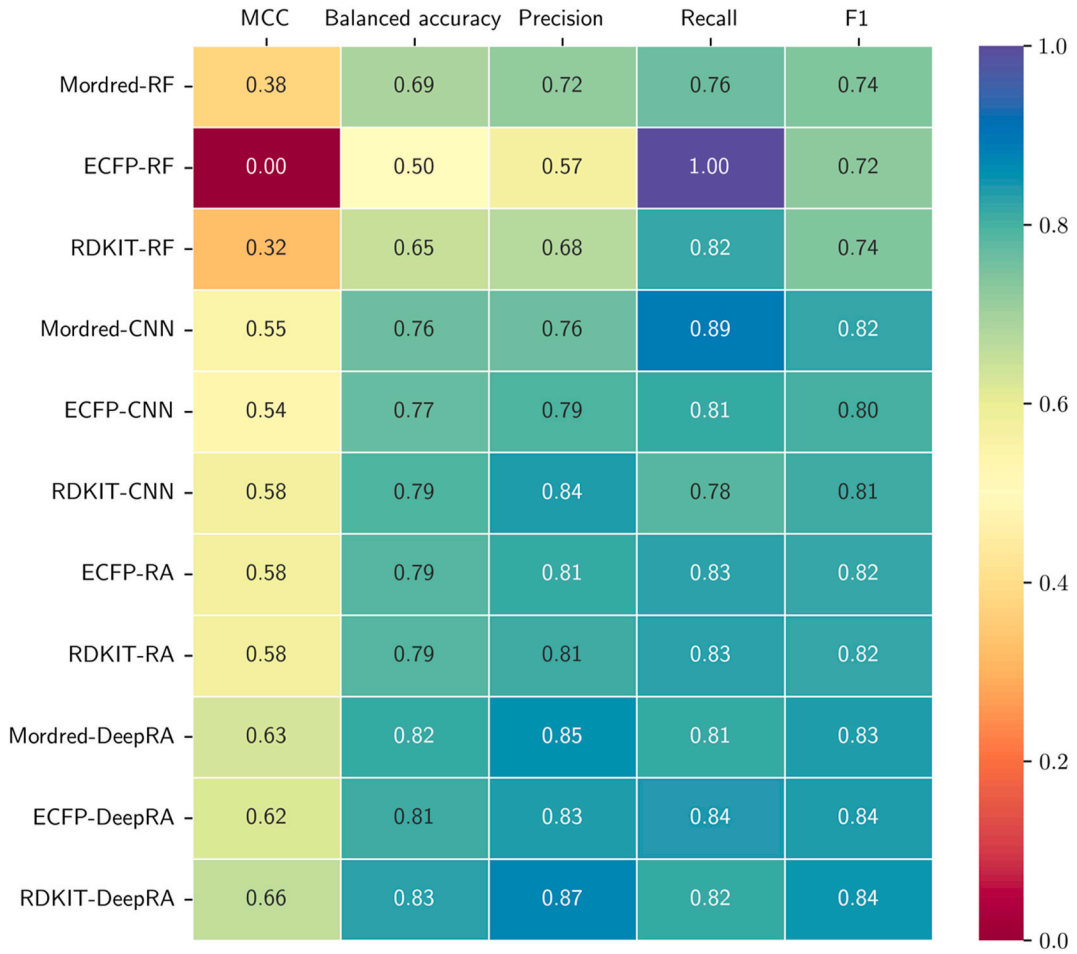

Fig. 4. Predictive performance of the random forest (RF), convolutional neural network (CNN), quantitative read-across, and DeepRA models with an independent test set. MCC: Matthew's correlation coefficient.

RF models had moderate to high precision, recall, and F1 metrics; however, the performance range between the lowest and highest precision values was high (0.57–0.72), suggesting unreliable prediction performance for both positive (mutagen) and negative (non-mutagen) chemicals, respectively. Based on these results, the RF models were omitted as baseline predictors in the DeepRA system.

2.4. *Performance of convolutional neural network (CNN) models*

We further constructed one dimension (1D)-CNN architecture to effectively predict the mutagenicity of the chemical compounds using all three molecular features (see Method and Fig. 1A) and evaluated their performance using an independent test set. We found that all three CNN models had MCC values greater than 0.5, indicating that the predicted values from these three models are moderately positively correlated to the experiment values (Fig. 4; Mordred-CNN, ECFP-CNN, and RDKIT-CNN). The obtained MCC values from the CNN models ranged from 0.54 to 0.58, which were greater than the MCC of the RF models in all molecular features, suggesting that the CNN models outperformed the RF models in all three features. Notably, the MCC of the RDKIT-based CNN was 0.58, which was greater than the Mordred and ECFP-based CNN models, respectively. These results inform that the prediction

results from the CNN models are acceptable and can be trusted. Moreover, the RDKIT-based CNN models may produce a greater predictive performance than the Mordred and ECFP-based CNN models.

We further explored other metrics of the CNN models and found that the CNN models contained a very high balanced accuracy of 0.76, 0.77, and 0.79 for the Mordred, ECFP, and RDKIT, respectively. These results showed similar performance trends compared to the MCC values. They also imply that the CNN models can accurately predict mutagen and non-mutagen with 76–79 % accuracy. The precision values of these three models also showed greater performance compared to the RF models, which ranged from 0.76 to 0.84, indicating that the model can capture the false positive or false mutagenic chemical from the independent test set. Moreover, the recall values of these three models were also greater than 0.78, especially for Mordred-based CNN, which possessed a very high recall value of 0.89, indicating that the CNN model can capture the false negative or false non-mutagenic chemicals from the independent test set. These predictive performances of the CNN models were confirmed with the F1 score. The F1 scores of all three models were 0.82, 0.80, and 0.81 for the Mordred, ECFP, and RDKIT, respectively. The high F1 score (>0.8) confirmed that all three models performed well in capturing both false positive and false negative instances. Overall, those three models could be further used in the DeepRA model construction.

2.5. Performance of quantitative read-across (qRA) models

We developed the qRA models using both ECFP and RDKIT fingerprints with the Tanimoto similarity coefficient as they can be computed using the binary Tanimoto similarity function (see Methods and Fig. 1B). We observed that the ECFP-based qRA and RDKIT-based qRA models produced similar performance values for the independent test set in all five metrics (Fig. 4; RA-ECFP and RA-RDKIT). The MCC, balanced accuracy, precision, recall, and F1 score values of both models were 0.58, 0.79, 0.81, 0.83, and 0.82, respectively (Fig. 4). Note that the ECFP-based qRA model can improve the performance of the MCC, balance accuracy, precision, recall, and F1 metrics compared to the ECFP-based CNN model at 2–4%. However, the RDKIT-based qRA model can only improve the performance of the recall and F1 scores compared to the RDKIT-based CNN model at 1–5%. With the performance improvements from the qRA models, we utilized both models in the construction of the DeepRA framework.

2.6. Performance of DeepRA models

We utilized the model outputs from the CNN and qRA models—called model-level representations or predictive features (PF)—and then combined them with the original molecular features (see Methods and Fig. 1C). This approach can be called two-step ensemble framework [39]. We evaluated the performance with the independent test set, and we found that the Mordred-based DeepRA produced the test MCC value of 0.63, which was greater than those of Mordred-CNN at 8 % (Fig. 4; DeepRA-Mordred, DeepRA-ECFP and DeepRA-RDKIT). The balanced accuracy and precision of the Mordred DeepRA model were 0.82, 0.85, and 0.82, respectively, which were increased from its baseline CNN by 6 % and 9 %, respectively. Only the recall was reduced from 0.89 to 0.81 in the DeepRA model, while the F1 scores were the same. These results indicate that the DeepRA tries to maximize the balanced accuracy and precision of the model, while compensating both recall and F1 metrics.

For the ECFP-based DeepRA, we found that the DeepRA model produced the test MCC value of 0.62, which was greater than the ECFP-CNN and ECFP-qRA models for 8 % and 4 %, respectively. The balanced accuracy of ECFP-DeepRA was 0.81, which was greater than the ECFP-CNN and ECFP-qRA models by 4 % and 2 %, respectively. The precision of ECFP-DeepRA was 0.83, which was greater than ECFP-CNN and ECFP-qRA models by 4 % and 2 %, respectively. The recall of ECFP-DeepRA was 0.84, which was greater than the ECFP-CNN and ECFP-qRA models by 3 % and 1 %, respectively. The F1 score of ECFP-DeepRA was 0.84, which was greater than ECFP-CNN and ECFP-qRA models by 4 % and 2 %, respectively. Overall, the ECFP-DeepRA improves the performance of the baseline models from 1 to 4 %. Notably, this improvement is lower than the improvement on the Mordred-DeepRA model compared to its baseline model.

The RDKIT-DeepRA model also showed the highest performance compared to the other two models. The MCC value of the DeepRA model was 0.66, which was greater than the CNN and qRA models by 8 %. Importantly, the MCC value of the RDKIT-DeepRA model was the highest compared to other DeepRA models. The balanced accuracy of the DeepRA model was 0.83, which was greater than the CNN and qRA models by 4 %. The precision of the DeepRA model was 0.87, which was greater than the CNN and qRA models by 3 % and 6 %, respectively. The recall of the DeepRA model was 0.82, which was greater than the CNN model by 4 %, but lower than the qRA by 1 %. The F1 score of the DeepRA model was 0.84, which was greater than the CNN and qRA models by 3 % and 2 %, respectively. Overall, the RDKIT-DeepRA model improves the performance of the baseline model from 1 to 8 % across those five metrics. To sum up, all three DeepRA models have improved their predictive performances compared to their baseline models. However, the three DeepRA models also have comparable performances across all five metrics, which make it difficult to select only one final model; thus, we further analyzed the applicability domain of these three models to select the highest performance after defining their applicability boundaries.

2.7. Applicability domain (AD) analysis

We further evaluate the applicability domain (AD) of the model in which we used the distance distribution to evaluate the abnormality of the new predictor compared to the average and standard deviation of the chemical's distance in the training set [25,40]. The concept of applicability domain (AD) is crucial when using computational models such as QSAR and deep learning. The AD defined as "the response and chemical structure space in which the model makes predictions with a given reliability." ([21]). This response and chemical structure space is based on the training data used to develop the model. Moreover, the AD is very important for QSAR or deep learning because these two models are inductive models, meaning that they generalize their output from the training data to make predictions on new compounds. This generalization ability is limited with the chemical structure used to train the model; the predictions also become increasingly unreliable if the new compounds distinct significantly from the training data. Therefore, we evaluate the performance of the model by varying the k nearest neighbor (kNN) from k equal to three to ten, which is used for measuring the distance between the prediction and each k nearest training data point. Only the within- AD from the independent test set was used to evaluate the performance; the out-of-domain chemicals were tracked and removed before evaluating the model performance. For Mordred-DeepRA, we found that at the k value equal to four, the model produced the highest prediction performance of 0.85, 0.92, 0.95, 0.96, and 0.96 for the MCC, balanced accuracy, precision, recall, and F1 score, respectively (Fig. 5A). This performance demonstrates significant improvement over the non-AD measurement, with increases of 22 %, 10 %, 10 %, 15 %, and 13 % for the MCC, balanced accuracy, precision, recall, and F1 score, respectively. This result implies that at k equal to four, we can ensure that the DeepRA can produce 92 % accurate prediction, 95 % accurately predict false positive, 96 % accurately predict false negative and 85 % of predicted results are correlated to the experimental values.

Conversely, we observed that the ECFP-DeepRA does not have a similar performance gain after removing the out-of-AD from the independent test set like the one from the Mordred-DeepRA (Fig. 5B). For the ECFP-DeepRA, we found that at the k equal to three, the model produced the highest prediction performance of 0.65, 0.82, 0.83, 0.79, and 0.81

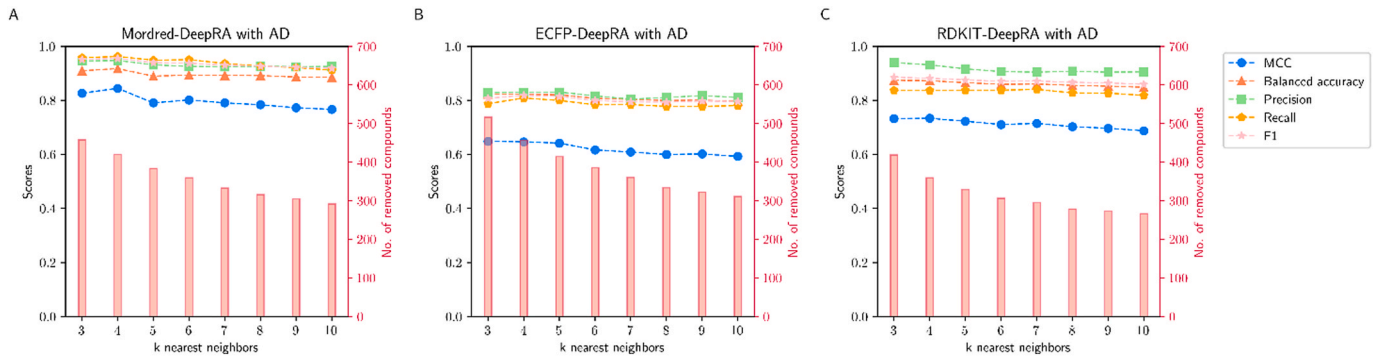

Fig. 5. Performance evaluation of the DeepRA models with only within applicability domain independent test set. A) Mordred-DeepRA, B) ECFP-DeepRA, and C) RDKIT-DeepRA models. Blue circles, orange triangles, green squares, yellow pentagons, pink stars, and bar graphs indicate Matthew's correlation coefficient (MCC), balanced accuracy, precision, recall, F1 score, and number of removed compounds, respectively.

for the MCC, balanced accuracy, precision, recall, and F1 score, respectively (Fig. 5B). This model performance has been changed by +3 %, +1 %, 0 %, −4%, and −3% for the MCC, balanced accuracy, precision, recall, and F1 score, respectively. Moreover, at the k equal to three, the number of removing chemicals was 517 (Fig. 5B bar chart), which was significantly greater than the chemicals that were removed from the k equal to four of the Mordred-DeepRA (n = 420, Fig. 5A bar chart). These two results indicate that the ECFP-DeepRA with applicability domain does not improve the performance of the ECFP-DeepRA model even though it removed a greater number of compounds compared to the other model. Thus, the ECFP-DeepRA cannot be used further in the mutagenicity prediction.

On the other hand, the performance gains were observed in the RDKIT-DeepRA with AD model like the Mordred-DeepRA model. At k equal to three, the RDKIT-DeepRA produced the highest prediction performance of 0.73, 0.88, 0.94, 0.84, and 0.89 for the MCC, balanced accuracy, precision, recall, and F1 score, respectively (Fig. 5C). The RDKIT-DeepRA model slightly improved over the non-AD test set, with increases of 7 %, 5 %, 7 %, 2 %, and 5 % for the MCC, balanced accuracy, precision, recall, and F1 score, respectively. This result suggests that the AD can improve the RDKIT-DeepRA prediction metrics. Notably, the performance gain of the RDKIT-DeepRA was lower than the Mordred-DeepRA, even though it removed a similar number of chemicals (at k = 3, n = 419, Fig. 5C bar chart) compared to the Mordred-DeepRA (at k = 4, n = 420). Moreover, the Mordred-DeepRA with the AD boundary exhibits the highest prediction accuracy, precision, recall, F1, and MCC values, indicating that the Mordred-DeepRA framework is the most suitable for predicting the mutagenicity of a chemical.

Caution should be carefully investigated when defining the AD boundary as it can provide a very high performance, but it also removes the out-of-AD chemicals from the data set. For example, at k equal to four of the Mordred-DeepRA model, the number of compounds that were removed from the independent data set was 420 chemicals or 60 % of the test set (Fig. 8A bar chart). On the other hand, at k equal to ten, the number of compounds that were removed from the test data set was only 292 or 42 % of the test set (Fig. 8A bar chart). The accuracies between these two k values were 0.92 and 0.89, which is only a 3 % difference. The high accuracy model (k = 4, balanced accuracy ≥0.92), contains a very promising predictive performance for predicting mutagenic compounds; however, it undeniably has a narrow chemical space for model application. On the other hand, a medium-high performance (k = 10, balanced accuracy ≥0.89), contains a relatively promising predictive performance compared to the top performance model (k = 4) with a larger chemical space for model application (40 % and 58 % remaining chemicals). This result suggests that the more restricted AD, the greater the performance. Additionally, the more restricted AD may not proportionally improve the performance of the model like the number of compounds that were removed. Thus, it indicates that defining the AD boundary needs to be compensated between accuracy and the area of application of the model.

The larger chemical space also came with an error rate of 3 % higher than the most accurate Mordred-DeepRA with the AD (k = 4) model. Thus, to select the suitable k value, one thing to consider is the risk of the false positive and false negative that can occur if the model predicts the chemical incorrectly. In our case, we opted to predict a chemical that produces genotoxicity or mutagen. This mutagen is very harmful to human health; therefore, the false negative, which means the chemical does contain mutagen, but the model predicts it as non-mutagen is very dangerous for the outcome compared to the false positive (falsely define non-mutagen as mutagen). So, increasing the accuracy to correctly predict the mutagen is necessary compared to increasing the model boundary, but followed by a high error rate. Thus, in this study, we select the k value of four from the Mordred-DeepRA for the prediction of the mutagen of non-sugar sweetener to ensure that the non-mutagen (non-toxic) and mutagen (toxic) are correctly predicted with the highest confidence that we can obtain from the DeepRA model.

2.8. Interpretation of DeepRA using permutation importance

We further analyze the importance feature of the DeepRA model using permutation importance. This technique relies on randomly shuffling the values of molecular features and predicting the model accuracy. If the accuracy is decreased after shuffling the value of the feature; it indicates that this feature is important for the accurate prediction of the model [41]. We computed the permutation features from all features of the DeepRA (Mordred-based DeepRA) and used the balanced accuracy metric as the method to determine the importance score of the model. The important score was computed based on the deviation between the permuted accuracy and the original accuracy of the DeepRA model (see Methods). We found that the permutation importance algorithm identified twelve important features including 1) CNN-ECFP, 2) RA_ECFP, 3) RA-RDKIT, 4) CNN-Mordred, 5) n6aHRing, 6) mZagreb2, 7) JGI8, 9) JGI7, 9) MID_O, 10) nFRing, 11) TopoPSA, and 12) CNN-RDKIT (Fig. 6A). The top four important features of the DeepRA model were the predictive features from the ECFP-based CNN, ECFP-based RA, RDKIT-based RA, and Mordred-based CNN, suggesting that the decision-making of this Mordred-DeepRA mostly contributed by those four predictive features. Moreover, the most important molecular descriptors from were 6-membered aromatic hetero ring count (n6aHRing). This ring contains 6-membered hetero atoms. This 6-membered aromatic hetero ring is generally found in nitrogen-or oxygen-containing 6-membered aromatic rings such as pyridine, pyrimidine, or oxazine rings. The mZagreb2 is the secondary importance molecular descriptor, which corresponds to the topological index based on graph theory. The JGI8 and JGI7 are the 8-ordered and 7-ordered mean topological charge transfer between a pair of atoms,

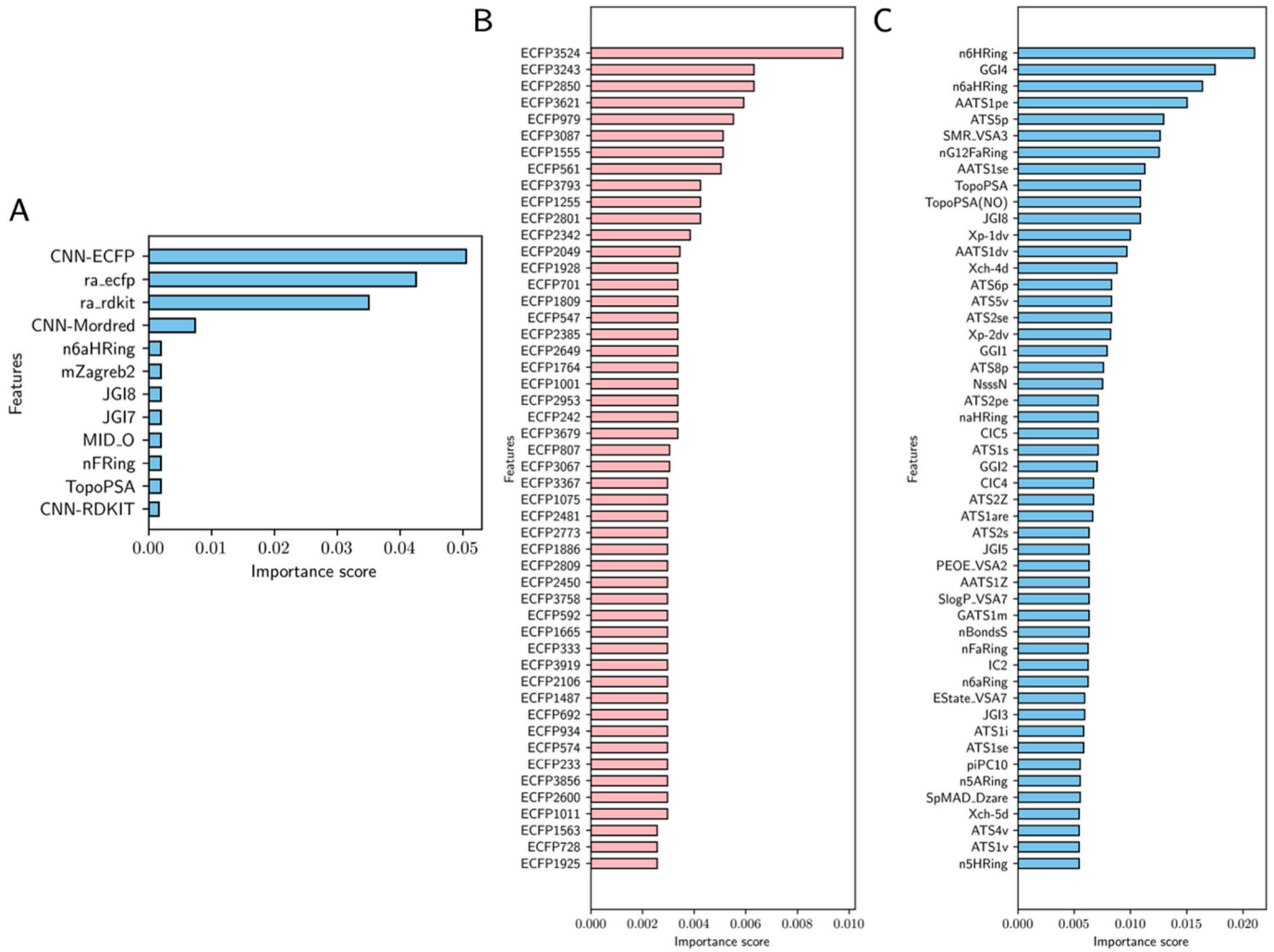

Fig. 6. Feature importance of the DeepRA model identified by permutation importance. A) Feature importance of the Mordred-based DeepRA. B) Feature importance of the ECFP-based CNN. C) Feature importance of the Mordred-based CNN.

respectively. The MID_O is the molecular identifier specifically associated with oxygen atoms within a molecule. The nFRing is a fuse ring count descriptor that measures the number of fused rings in the molecule. The TopoPSA is the topological polar surface area that counts only nitrogen and oxygen atoms. Based on the meaning of these descriptors, we found that the mutagenicity is determined by those 6-membered hetero or fused rings with oxygen and nitrogen atoms. Interestingly, the last important feature identified by permutation importance is the RDKIT-based CNN. This means all the five predictive features are important to the DeepRA system but on different importance scores.

We also further performed the permutation importance of the ECFP-CNN and Mordred-CNN as the top four predictive features of the DeepRA model (Fig. 6B and C). It should be noted that the top two and three predictive features (i.e., ECFP-qRA and RDKIT-qRA models) cannot be computed by permutation importance as the framework computed the fingerprints per instant input and then further used in the qRA algorithm, but the permutation feature required the fingerprint features of the whole data set first prior to begin shuffling each feature. Therefore, the permutation importance was demonstrated only CNN-based models. We found that the top chemical features of ECFP3524, which correspond to the chemical structures that contain aromatic ring, 6-membered ring, oxygen atom, and rings identified by MACCS key (Supplementary Materials). These identified importance substructures were similar to the importance substructure of the Mordred-CNN model, where the top three key features are the number of 6-membered hetero rings (n6HRing), topological charge at 4 distances (GGI4), and the 6-membered aromatic hetero ring count (n6aHRing). Thus, these results confirmed that the 6-membered hetero rings and 6-membered hetero aromatic ring are significant features that correspond to the mutagenic activity.

2.9. Prediction of mutagenicity of non-sugar sweeteners

To explore the mutagenic activity of the NSSs data set, we first explored the chemical space of the NSSs data set. Then we evaluate the applicability domain of the NSS chemicals. After that, we performed the prediction experiments using the Mordred-based DeepRA model. We found that the NSS has MW and LogP of 370.34 ± 226.57 and -0.50 ± 2.32, respectively (Fig. 7A), while the HBA and HBD of the NSS were 7.39 ± 5.15 and 4.47 ± 3.41, respectively (Fig. 7B). We found that the sweetness (logarithm of sweetness or LogSw) can also be discriminated against MW and LogP, but not in HBA and HBD (Fig. 7A and B). We also found that the sweetness was distributed into several clusters in all Mordred, ECFP, and RDKIT (Fig. 7C-E), suggesting that these three features can be used to learn the sweetness in further study.

After that, we implemented the DeepRA model to evaluate the mutagenic of the NSSs. We found that out of 332 NSS chemicals, 56 NSSs were classified as mutagenicity (DeepRA = 1), while 276 NSSs were classified as non-mutagenicity (DeepRA = 0). The common NSSs such as dulcin, acesulfame potassium, lactose, saccharin, fucose, lactulose,

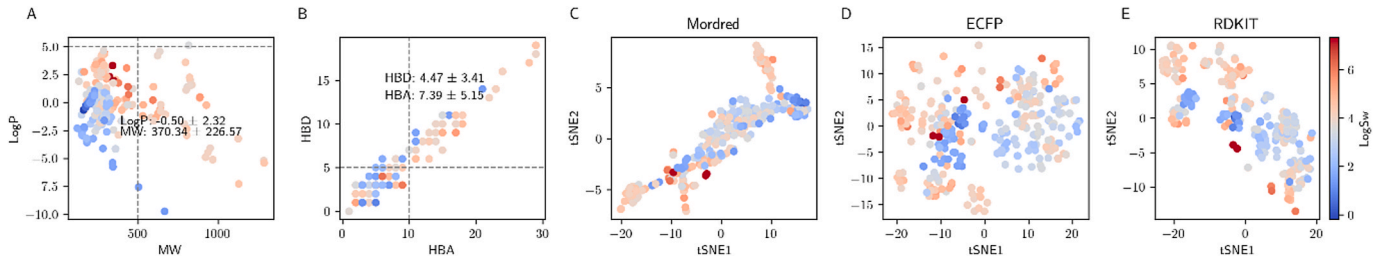

Fig. 7. Chemical spaces of non-sugar sweeteners (NSS). A) Distribution of molecular weight (MW) and LogP. B) Distribution of hydrogen bonding acceptor (HBA) and hydrogen bonding donor (HBD). C) Distribution of Mordred descriptors. D) Distribution of extended circular fingerprints (ECFP), and E) Distribution of RDKIT fingerprints. The dashed line represents the drug-likeness threshold. Circle colors represent the logarithm of sweetness (LogSw).

A

Name	DeepRA	class	AD_status	Remark
dulcin	0.997	1	within_AD	Confident mutagen
acesulfame potassium	0.453	0	outside_AD	Not confident
lactose	0.325	0	outside_AD	Not confident
saccharin	0.052	0	outside_AD	Not confident
fucose	0.029	0	outside_AD	Not confident
lactulose	0.026	0	outside_AD	Not confident
sucralose	0.018	0	outside_AD	Not confident
stevioside	0.000	0	outside_AD	Not confident

B

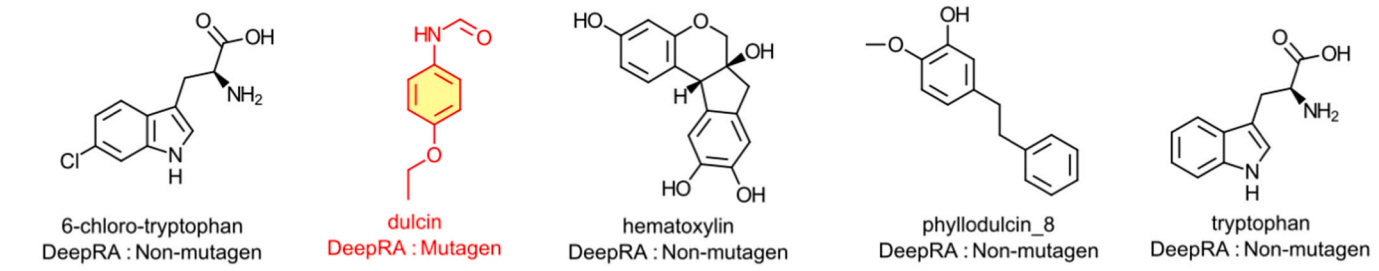

Fig. 8. Mutagenic prediction of non-sugar-sweeteners (NSS) from the DeepRA model. A) example of common NSSs and B) Chemicals that are within the application domain were. The red chemical structure indicates the chemical that was predicted as a mutagen.

sucralose, and sativoside were showed in Fig. 8A. We found that dulcin has probability to become mutagen higher than 0.9. Acesulfame potassium, lactose, saccharin, fucose, lactulose, sucralose, and sativoside have probability to become mutagen lower than 0.5. Noted that only dulcin from the common NSSs was predicted to be within applicability domain (within-AD), meaning that it can be mutagen at accuracy rate at 92 %, the rest of NSSs may get accuracy rate of 82 % because they are out-of-domain. We further analysis the within-AD NSSs and found that there are another four NSSs that were classified as within-AD: 6-chlorotryptophan, hematoxylin, phyllodulcin 8, and tryptophan. Thus, the total NSSs that were classified as within-AD are five chemicals as shown in Fig. 8B. With these five chemicals, we found that only dulcin was predicted to be a mutagenic compound. This dulcin sweetener was found to be linked with liver cancer and therefore banned as a food additive worldwide [42]. The amino acid tryptophan and its derivative have no report on mutagenicity. The phyllodulcin is a natural sweetener found in *Hydrangea macrophylla*. Phyllodulcin 8 is the derivative flavonoid opening structure of phyllodulcin. This structure does not report DNA mutagenicity. Lastly, hematoxylin, which is a common dye used in medical laboratories, exhibits no effect on mutagenicity.

3. Discussion

This study provides the comprehensive development frameworks of the stacking ensemble model that combined two different layers of features called DeepRA: 1) molecular descriptors and 2) model-based features from the convolutional neural network and quantitative read-across models. This framework lets the algorithm learn from different knowledge-based features, guiding the meta-decision to be more optimistic and not rely on a single decision model. The knowledge of the meta-learner relied on both molecular descriptors and molecular fingerprints of the chemical structures. The molecular descriptors were used in the model's construction process because they can be generalizable between diverse functional groups or substructures, which makes it perform well in the prediction process as we demonstrated in this paper. On the other hand, the path-based molecular fingerprints (i.e., ECFP and RDKIT) were utilized as they are not predefined substructures and can identify the connectivity pattern of chemical substructure based on the certain distance (radius) defined by the researcher. This algorithm also allows the fingerprints to capture local chemical spaces and could be flexible based on the diverse training set.

The predictive performance on the independent test set of the DeepRA model is genuinely higher than the other benchmark mutagenic prediction models [43,44]. For example, DeepAmes model was constructed by using ensemble learning of an artificial neural network of model-level representation. Still, this model contained the MCC of 0.38, balanced accuracy of 0.69, sensitivity (recall) of 0.47, and F1 of 0.48 [43], which are lower than the performance obtained from the DeepRA model. Even though the DeepAmes has the model-level representation, it does not rely on the convolutional neural network and the molecular features. The input of that system is only molecular descriptors without molecular fingerprints as we developed in this study. Since our DeepRA model is based on the combination of model-level and molecular descriptors-level for the meta-learner, it could also enhance the performance of the model compared to the previous report [43]. Another benchmark on a combination of the machine learning and quantitative read-across model was shown in the previous study [44]. This study utilized the combination of the read-across and support vector machine model with the performance of the model were 0.521 MCC, 0.761 accuracy, 0.786 precision, 0.764 sensitivity (recall), and 0.775 F1 score. Compared with our DeepRA model, both non-defined applicability domain and defined applicability domain DeepRA models contained higher predictive performance in all performance metrics, suggesting that the CNN of the molecular descriptors and the model-level representation potentially enhance the predictive performance of the mutagenicity compared to the previous studies.

The CNN architecture has been used in various tasks of computer vision, natural language preprocessing, pattern recognition, as well as in bioinformatics. The CNN model can be used to annotated the bacterial type IV secretion system effectors (T4SEs) [45] and prediction of RNAs and RNA-associated interaction [46]. The sequence-based CNN model with multiple scale protein representation can also be used to annotate protein function, which can improve protein stability and accuracy of the model [47,48], which mean the CNN frameworks are useful and can give an accurate results. This because, the CNN architecture relies on its convolutional ability to extract input features; it takes advantage of local spatial coherence that is suitable for extracting relevant features with fewer weights as some features are shared (coherence together) [49]. This is very beneficial to the chemical structure features, as molecular features are often shared or have coherent ability. In this study, we stacked two convolutional layers to extract more complex and abstract features, enabling the model to recognize intricate patterns of molecular features. The dropout layer was used to improve the generalization prediction of the model by preventing overfitting of the model. The max pooling layer followed by the flattened and fully connected neuron layers would also preprocess the extracted features into the 0 to 1 with rectified linear unit (ReLU) function before deciding on class probability. Another deep learning methods that can be used for stacking ensemble learning are transformer-based framework [50]. This model enables the researcher to predict protein-protein interaction with sequence based deep learning model. For the qRA model, we utilized the weighted average of the Tanimoto similarity coefficient of the top k highest similarities between the new predictor and the chemicals in the training set. This qRA method was optimized by our previous study [25]. These qRA models can outperform the machine learning models and are present in the top three important features of the DeepRA, suggesting that this method is valuable for the accurate classification of the predictive model. Further study should try to incorporate this algorithm in the predictive framework as it may improve the predictive performance of the model. It should be noted that the performance of a predictive model can be enhanced by incorporating different input feature types, using different decision algorithms, stacking different level features, and employing the optimized applicability boundary, as demonstrated by our study.

One thing to consider when building a predictive model is the trading-off between the model's accuracy and its usability domain. In this study, the very high accuracy DeepRA model (balanced accuracy $\geq$0.92) can be obtained via the very restricted boundary of the applicability domain; only forty percent of the independent test set remained in the applicability domain. Conversely, the medium to high accurate model (balanced accuracy $\geq$0.89) can also be obtained via the loosely restricted boundary of the applicability domain; more than fifty-eight percent of independent data remained in the applicability domain. To design between the model's performance and the usability of the model, one can assess the consequences or risk of the false prediction and the availability of chemical resources. Especially for food toxicity prediction, the risk of false prediction can be significant. If a model incorrectly predicts that a food chemical is non-mutagenic, when it is mutagenic; it could lead to serious health issues, loss of product, and loss of consumer trust. On the contrary, if a model incorrectly predicts that a non-mutagenic food chemical is mutagenic; it could lead to unnecessary food chemical recalls but does not affect human health. In such critical applications, having a higher accuracy prediction would be preferable to minimize the risk of health risks from food additives. However, this could limit the chemical's applicability or confidence boundary of the predictive model. In this study, we used a very high restrict chemical's applicability domain, resulting in high balanced accuracy, which discarded 60 % of the independent test set out as outside AD. This could limit the usability of this model as observed in AD measurement in the sweetener data set.

4. Conclusion

In conclusion, this study demonstrates the proof-of-concept to build a stacking two-level feature, including molecular descriptors and CNN-based and qRA-based models-level representations, then employing the CNN for the meta-decision called DeepRA model. This DeepRA model takes advantage of 1) deep convolutional neural network and quantitative read-across models and 2) the stacking ensemble learning of two-level features, which can outperform its baseline and other mutagenic prediction models. The applicability domain of the DeepRA model was also defined based on risk assessment and it can improve the accuracy of the model on the prediction of unseen chemicals up to 0.92 or 92 percent. The DeepRA with applicability domain framework was then used to predict 332 non-sugar sweeteners, and correctly predict dulcin as a mutagenic agent, which was confirmed by a previous report. Moreover, this framework was produced based on the guidelines for the development of the QSAR model; therefore, this framework has implications to be used in the chemical legislation processes. To further use this algorithm, we hosted the DeepRA framework in the GitHub repository (https://github.com/taraponglab/deepra.)

5. Methods

5.1. Data set preparation

Ames data set was obtained from previous literature and the Toxicity TOXRIC database [37,51]. The initial data set contains 7485 chemicals that were preprocessed by removing the missing SMILES, toxicity values, IUPAC name, inorganics, mixtures, and duplicated entries. The final mutagenicity data set for developing the model was 6881 chemicals. Then, those data points were split into training and test sets with a 9:1 ratio, respectively. Subsequently, the training set was split into a validation set with a 9:1 ratio, before fine-tuning the model, while the independent test set was solely used to evaluate the predictive performance of the constructed models. The total data points of training, validation, and test were 5572, 620, and 689, respectively. The NSS data set was obtained from an extensive 671 NSS-related chemicals [36]. The NSS data set was then curated by removing the data points that missing SMILES, missing chemical names, containing inorganics, containing mixtures, or containing duplicated entries. Finally, the NSS data set was comprised of 332 chemicals with SMILES, name, and experimental log sweetness (LogSw) values.

5.2. Molecular features encoding

We utilized a unique isometric canonical SMILES to compute all molecular features used in this study. The drug-likeness was computed by using RDKIT software (http://www.rdkit.org/). These molecular features consist of one molecular descriptor system and two molecular fingerprint systems: Mordred descriptors, eight radius ECFP, and RDKIT fingerprints. The Mordred descriptors initially contained 1613 2D physicochemical-based descriptors, then the constant descriptors were removed, resulting in 973 molecular descriptors. The ECFP and RDKITs fingerprints contain binary bits (i.e., presence (=1) and absence (=0) of chemical substructures) of 4096 bits and 2048 bits, respectively.

5.3. Random forest (RF)

The RF models were used as a baseline comparison machine learning model. The optimization was done using grid search validation. The parameters that were used in grid search optimization included 2–6 maximum features, 1–4 maximum depth of the tree, and 100–300 number of trees (T). The number of maximum features determines the maximum features to be used in each random decision tree (R_T). The maximum depth determines the maximum decision layer made by each T, and the number of trees determines how many trees are used for producing an average decision from the RF algorithm. The RF algorithm can be simply described as Eq (1).

$$\mathrm{RF}(x) = \frac{1}{T}\sum_{t=1}^{T} R_T(x) \tag{1}$$

5.4. Convolutional neural network (CNN)

The CNN architecture utilized in this study is as follows: one input layer, two stacking CNN layers with 32 kernel with ReLU function, one dropout layer with a 20 percent dropout rate compared to the total parameters received from the CNN layer, one two-size maximum pooling layer, one flatten layer, one fully connected neural network with 32 neurons containing ReLU function, and the output layer with one neuron containing sigmoidal function. The probability output of the last layer higher than 0.5 was set as a toxic or mutagenic chemical, whereas the probability output lower than 0.5 was set as a nontoxic or non-mutagenic chemical. The CNN model was trained and optimized using adaptive moment estimation (ADAM) and loss function with accuracy metrics. The CNN models were trained and validated with 50 epochs before evaluating their performances with the independent test set. All data were transformed into a one-dimension matrix before input into the system.

5.5. Quantitative read-across (qRA)

The qRA algorithm was constructed based on the idea that similar substructures of the chemicals will have similar bioactivity properties [25]. The Tanimoto similarity coefficient (Tc) was used to determine the similarity coefficient between a query predictor and the sources, which are training data points. The Tc value between a pair of the query and the source chemical was calculated using Eq (2).

$$\mathrm{Tc}(q,s) = \frac{q \cap s}{q \cup s} \tag{2}$$

where $\mathrm{Tc}(q,s)$ is the Tanimoto similarity coefficient between the query (q) and the source (s) chemicals. The $q \cap s$ and $q \cup s$ are the intersection and union between the molecular fingerprints of the query and source chemicals, respectively. The $\mathrm{Tc}(q,s)$ values are ranged from zero to one, where zero is the lowest similarity and one is the highest similarity between the pair. The $\mathrm{Tc}(q,s)$ values are further utilized in the similarity coefficient weight average of the top highest similarity to predict average mutagenic output ($\widehat{Y}(q)$) from the $\mathrm{Tc}(q,s)$ and it mutagenic value (Y_s). The number of top highest similarities can be defined by the researcher. In our study, we used the top three highest similarity coefficients to produce the average mutagenic output ($\widehat{Y}(q)$). The calculation of similarity coefficient weight average output can be computed by Eq (3).

$$\widehat{Y}(q) = \frac{\sum_{i=1}^{N} Tc(q,s) \times Y_s}{\sum_{i=1}^{N} Tc(q,s)} \tag{3}$$

5.6. Stacking DeepRA

The stacking DeepRA models were constructed by using the predictive outputs from the CNN-based models (i.e., Mordred-CNN, ECFP-CNN, and RDKIT-CNN) and the qRA models (i.e., ECFP-RA and RDKIT-RA). The predictive outputs can be called predictive features (PF). For the basic stacking ensemble learning, only PF was used in the stacking ensemble learning using CNN architecture described in section 5.4. For the two-level stacking DeepRA model, we utilized the molecular features (i.e., Mordred, ECFP, and RDKIT) and stacked with all five PF in the same one-dimension matrix. Then the inputs were fed into the CNN model and trained with the same condition as the CNN model section 5.4, except the training epoch that was only trained for 15 epochs to prevent the overfitting of the model. The PF features and DeepRA features were described in Eqs (4)–(7).

$$\mathrm{PF} = \{\widehat{Y}(q)_{\text{Mordred-CNN}}, \widehat{Y}(q)_{\text{ECFP-CNN}}, \widehat{Y}(q)_{\text{RDKIT-CNN}}, \widehat{Y}(q)_{\text{ECFP-RA}}, \widehat{Y}(q)_{\text{RDKIT-RA}}\} \tag{4}$$

$$\text{Mordred DeepRA} = \{\mathrm{Mord}_1, \ldots, \mathrm{Mord}_{973}\} + \mathrm{PF} \tag{5}$$

$$\text{ECFP DeepRA} = \{\mathrm{ECFP}_1, \ldots, \mathrm{ECFP}_{4096}\} + \mathrm{PF} \tag{6}$$

$$\text{RDKIT DeepRA} = \{\mathrm{RDKIT}_1, \ldots, \mathrm{RDKIT}_{2028}\} + \mathrm{PF} \tag{7}$$

5.7. Evaluation metrics

The predictive performances were evaluated using five metrics: MCC, balanced accuracy, precision, recall, and F1 score. The MCC was used instead of the area under the curve of receiver operating characteristic (AUROC) as the main evaluation metric as it is very sensitive to the unbalanced data set [52]. The balanced accuracy was used as it can minimize classification error better than the normal accuracy metric [53]. Precision and recall were used to detect false positives and false negatives, respectively. The F1 score is a harmonic mean between precision and recall, which symmetrically represents both precision and recall in one metric. Those five metrics were calculated by Eqs (8)–(12).

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN})}} \tag{8}$$

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right) \tag{9}$$

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{10}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$\text{F1 score} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP}+\mathrm{FN})} \tag{12}$$

where the TP, TN, FP, and FN are true positive (mutagenic), true negative (non-mutagenic), false positive (predicted to be mutagenic but

it is non-mutagenic), and false negative (predicted to be non-mutagenic, but it is mutagenic), respectively. The result of MCC values can be varied by −1 to 1, where minus one indicates that the prediction results are inversely correlated to the experiment values, zero indicates that the prediction is non-correlated to the experiment values, and one indicates that the prediction results are positively correlated to the experiment values. The balanced accuracy can be varied from zero to one, where zero is the lowest accuracy and one is the highest accuracy of the model. The precision, recall, and F1 score can be varied from zero to one like the balanced accuracy, where zero is the model containing the highest false positive, highest false negative, and the highest combination of false positive and false negative, respectively. The value one is the lowest false positive, lowest false negative, and the lowest combination of the highest false positive and false negative, respectively.

5.8. Applicability domain (AD)

The concept of the applicability domain refers to the specific range or scope within which a quantitative structure-activity relationship (QSAR) model can reliably predict its outcome [21]. This study utilized the Euclidean distance-based k nearest neighbors (kNN) between the new prediction and the closest k neighbors from the training set [40]. The AD was performed after the PFs were constructed. The AD criteria can be described by the following Eq (13) and (14).

$$\text{For within-domain; } D_i < D_k + \sigma \times Z \tag{13}$$

$$\text{For out-of-domain; } D_i \geq D_k + \sigma \times Z \tag{14}$$

where D_i is the Euclidean distance between the new prediction and the closest k neighbor. The D_k and σ are the average and standard deviation of the Euclidean distance between the whole training set, respectively. The Z score is an empirical parameter to control the significant level of the AD model. This value was set at 0.5.

5.9. Permutation importance

Permutation importance is the computational scheme that measures the importance of the machine learning features. The importance of features can be identified by first measuring the performance of the original model. In our study, we used balanced accuracy as the performance metric. Then, take one feature at a time and shuffle its values across all data points. This method randomly scrambles the values of a single feature, while keeping all other features unchanged. After permuting the values, the model performance was measured again and compared with the balanced accuracy of the original model. If the performance of the model is reduced, meaning that the single feature is important for making the model perform well. These features were obtained and represented in the result section without using an importance score cut-off. The importance score can be computed by Eq (15).

$$\begin{aligned}\text{Importance score} = &\text{Balanced accuracy (original)} \\ &\text{- Balanced accuracy (permuted)}\end{aligned} \tag{15}$$

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

All data used in this study were provided in Supporting materials. The developed software and model are made available on the GitHub repository: https://github.com/taraponglab/deepra.

CRediT authorship contribution statement

Tarapong Srisongkram: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108731.

References

[1] A.W. Logue, M.E. Smith, Predictors of food preferences in adult humans, Appetite 7 (1986) 109–125, https://doi.org/10.1016/S0195-6663(86)80012-5.
[2] T.R. Maone, R.D. Mattes, J.C. Bernbaum, G.K. Beauchamp, A new method for delivering a taste without fluids to preterm and term infants, Dev. Psychobiol. 23 (1990) 179–191, https://doi.org/10.1002/dev.420230208.
[3] G. Bright, others, Low-calorie sweeteners-from molecules to mass markets, Low-Calorie Sweeteners-from Molecules to Mass Markets (1999) 3–9.
[4] K. Kaur, S. Srivastava, Artificial sugar saccharin and its derivatives: role as a catalyst, RSC Adv. 10 (2020) 36571–36608.
[5] M.R. Weihrauch, V. Diehl, Artificial sweeteners—do they bear a carcinogenic risk? Ann. Oncol. 15 (2004) 1460–1465, https://doi.org/10.1093/annonc/mdh256.
[6] M. Rios-Leyvraz, J. Montez, W.H. Organization, Health Effects of the Use of Non-sugar Sweeteners: a Systematic Review and Meta-Analysis, 2022.
[7] E. Riboli, F.A. Beland, D.W. Lachenmeier, M.M. Marques, D.H. Phillips, E. Schernhammer, A. Afghan, R. Assunção, G. Caderni, J.C. Corton, G. de A. Umbuzeiro, D. de Jong, M. Deschasaux-Tanguy, A. Hodge, J. Ishihara, D. D. Levy, D. Mandrioli, M.L. McCullough, S.A. McNaughton, T. Morita, A.P. Nugent, K. Ogawa, A.R. Pandiri, C.M. Sergi, M. Touvier, L. Zhang, L. Benbrahim-Tallaa, S. Chittiboyina, D. Cuomo, N.L. DeBono, C. Debras, A. de Conti, F.E. Ghissassi, E. Fontvieille, R. Harewood, J. Kaldor, H. Mattock, E. Pasqual, G. Rigutto, H. Simba, E. Suonio, S. Viegas, R. Wedekind, M.K. Schubauer-Berigan, F. Madia, Carcinogenicity of aspartame, methyleugenol, and isoeugenol, Lancet Oncol. 24 (2023) 848–850, https://doi.org/10.1016/S1470-2045(23)00341-8.
[8] S. Pavanello, A. Moretto, C. La Vecchia, G. Alicandro, Non-sugar sweeteners and cancer: toxicological and epidemiological evidence, Regul. Toxicol. Pharmacol. 139 (2023) 105369, https://doi.org/10.1016/j.yrtph.2023.105369.
[9] G.R. Mohn, On the correlation between mutagenicity and carcinogenicity, in: Genetic Origins of Tumor Cells, Springer, 1980, pp. 11–24.
[10] ICH Harmonised Guideline, Assessment and control of dna reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk M7, in: International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), 2014. Geneva.
[11] OECD, Test No. 471: Bacterial Reverse Mutation Test, 2020, https://doi.org/10.1787/9789264071247-en.
[12] OECD, Test No. 473: in vitro mammalian chromosomal aberration test. https://doi.org/10.1787/9789264264649-en, 2016.
[13] OECD, Test No. 487: in vitro mammalian cell micronucleus test. https://doi.org/10.1787/9789264264861-en, 2023.
[14] OECD, Test No. 476: in vitro mammalian cell gene mutation tests using the Hprt and xprt genes. https://doi.org/10.1787/9789264264809-en, 2016.
[15] OECD, Test No. 490: in Vitro Mammalian Cell Gene Mutation Tests Using the Thymidine Kinase Gene, 2016, https://doi.org/10.1787/9789264264908-en.
[16] OECD, Test No. 489: in Vivo Mammalian Alkaline Comet Assay, 2016, https://doi.org/10.1787/9789264264885-en.
[17] OECD, Test No. 488: Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays, 2022, https://doi.org/10.1787/9789264203907-en.
[18] OECD, Test No. 478: rodent dominant lethal test. https://doi.org/10.1787/9789264264823-en, 2016.
[19] E. Echa, Guidance on Information Requirements and Chemical Safety Assessment: Chapter R. 6: QSARs and Grouping of Chemicals, Helsinki, Finland, 2008.
[20] European Chemicals Agency, Read-Across Assessment Framework (RAAF), Publications Office, LU, 2017. https://data.europa.eu/doi/10.2823/619212. (Accessed 16 July 2023).
[21] OECD, Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. https://doi.org/10.1787/9789264085442-en, 2014.
[22] G. Patlewicz, L.E. Lizarraga, D. Rua, D.G. Allen, A.B. Daniel, S.C. Fitzpatrick, N. Garcia-Reyero, J. Gordon, P. Hakkinen, A.S. Howard, A. Karmaus, J. Matheson, M. Mumtaz, A.-N. Richarz, P. Ruiz, L. Scarano, T. Yamada, N. Kleinstreuer, Exploring current read-across applications and needs among selected U.S. Federal

Agencies, Regul. Toxicol. Pharmacol. 106 (2019) 197–209, https://doi.org/10.1016/j.yrtph.2019.05.011.
[23] B.-M. Lee, S.H. Lee, T. Yamada, S. Park, Y. Wang, K.-B. Kim, S. Kwon, Read-across approaches: current applications and regulatory acceptance in Korea, Japan, and China, J. Toxicol. Environ. Health 85 (2022) 184–197, https://doi.org/10.1080/15287394.2021.1992323.
[24] G. Patlewicz, I. Shah, Towards systematic read-across using Generalised Read-Across (GenRA), Computational Toxicology 25 (2023) 100258, https://doi.org/10.1016/j.comtox.2022.100258.
[25] T. Srisongkram, Ensemble quantitative read-across structure–activity relationship algorithm for predicting skin cytotoxicity, Chem. Res. Toxicol. (2023), https://doi.org/10.1021/acs.chemrestox.3c00238 acs.chemrestox.3c00238.
[26] A. Banerjee, K. Roy, On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points, Chem. Res. Toxicol. 36 (2023) 446–464, https://doi.org/10.1021/acs.chemrestox.2c00374.
[27] T. Stepišnik, B. Škrlj, J. Wicker, D. Kocev, A comprehensive comparison of molecular feature representations for use in predictive modeling, Comput. Biol. Med. 130 (2021) 104197, https://doi.org/10.1016/j.compbiomed.2020.104197.
[28] A. Banerjee, K. Roy, Prediction-inspired intelligent training for the development of classification read-across structure-activity relationship (c-RASAR) models for organic skin sensitizers: assessment of classification error rate from novel similarity coefficients, Chem. Res. Toxicol. 36 (2023) 1518–1531, https://doi.org/10.1021/acs.chemrestox.3c00155.
[29] D.H. Wolpert, Stacked generalization, Neural Network. 5 (1992) 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1.
[30] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: opportunities and challenges, Journal of King Saud University - Computer and Information Sciences 35 (2023) 757–774, https://doi.org/10.1016/j.jksuci.2023.01.014.
[31] T. Srisongkram, N.F. Syahid, D. Tookkane, N. Weerapreeyakul, P. Puthongking, Stacked ensemble learning on HaCaT cytotoxicity for skin irritation prediction: a case study on dipterocarpol, Food Chem. Toxicol. 181 (2023) 114115, https://doi.org/10.1016/j.fct.2023.114115.
[32] N.F. Syahid, N. Weerapreeyakul, T. Srisongkram, StackBRAF: a large-scale stacking ensemble learning for braf affinity prediction, ACS Omega (2023), https://doi.org/10.1021/acsomega.3c01641.
[33] N. Schaduangrat, N. Anuwongcharoen, M.A. Moni, P. Lio, P. Charoenkwan, W. Shoombuatong, StackPR is a new computational approach for large-scale identification of progesterone receptor antagonists using the stacking strategy, Sci. Rep. 12 (2022) 16435, https://doi.org/10.1038/s41598-022-20143-5.
[34] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J Cheminform 10 (2018) 4, https://doi.org/10.1186/s13321-018-0258-y.
[35] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754, https://doi.org/10.1021/ci100050t.
[36] M. Goel, A. Sharma, A.S. Chilwal, S. Kumari, A. Kumar, G. Bagler, Machine learning models to predict sweetness of molecules, Comput. Biol. Med. 152 (2023) 106441, https://doi.org/10.1016/j.compbiomed.2022.106441.
[37] L. Wu, B. Yan, J. Han, R. Li, J. Xiao, S. He, X. Bo, TOXRIC: a comprehensive database of toxicological data and benchmarks, Nucleic Acids Res. 51 (2023) D1432–D1445, https://doi.org/10.1093/nar/gkac1074.
[38] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.
[39] W. Shoombuatong, N. Homdee, N. Schaduangrat, P. Chumnanpuen, Leveraging a meta-learning approach to advance the accuracy of Nav blocking peptides prediction, Sci. Rep. 14 (2024) 4463, https://doi.org/10.1038/s41598-024-55160-z.
[40] T. Srisongkram, D. Tookkane, Insights into the structure-activity relationship of pyrimidine-sulfonamide analogues for targeting BRAF V600E protein, Biophys. Chem. 307 (2024) 107179, https://doi.org/10.1016/j.bpc.2024.107179.
[41] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347, https://doi.org/10.1093/bioinformatics/btq134.
[42] Y. Uesawa, A.G. Staines, D. Lockley, K. Mohri, B. Burchell, Identification of the human liver UDP-glucuronosyltransferase involved in the metabolism of p-ethoxyphenylurea (dulcin), Arch. Toxicol. 81 (2007) 163–168, https://doi.org/10.1007/s00204-006-0138-5.
[43] T. Li, Z. Liu, S. Thakkar, R. Roberts, W. Tong, DeepAmes: a deep learning-powered Ames test predictive model with potential for regulatory application, Regul. Toxicol. Pharmacol. 144 (2023) 105486, https://doi.org/10.1016/j.yrtph.2023.105486.
[44] S.K. Pandey, K. Roy, Development of a read-across-derived classification model for the predictions of mutagenicity data and its comparison with traditional QSAR models and expert systems, Toxicology 500 (2023) 153676, https://doi.org/10.1016/j.tox.2023.153676.
[45] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings Bioinf. 21 (2020) 1825–1836, https://doi.org/10.1093/bib/bbz120.
[46] Y. Wang, Z. Pan, M. Mou, W. Xia, H. Zhang, H. Zhang, J. Liu, L. Zheng, Y. Luo, H. Zheng, X. Yu, X. Lian, Z. Zeng, Z. Li, B. Zhang, M. Zheng, H. Li, T. Hou, F. Zhu, A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder, Nucleic Acids Res. 51 (2023), https://doi.org/10.1093/nar/gkad929 e110–e110.
[47] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (2020) 1437–1447, https://doi.org/10.1093/bib/bbz081.
[48] L. Zheng, S. Shi, M. Lu, P. Fang, Z. Pan, H. Zhang, Z. Zhou, H. Zhang, M. Mou, S. Huang, L. Tao, W. Xia, H. Li, Z. Zeng, S. Zhang, Y. Chen, Z. Li, F. Zhu, AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding, Genome Biol. 25 (2024) 41, https://doi.org/10.1186/s13059-024-03166-1.
[49] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J Big Data 8 (2021) 53, https://doi.org/10.1186/s40537-021-00444-8.
[50] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research 6 (2023), https://doi.org/10.34133/research.0240, 0240.
[51] Z. Wu, D. Jiang, J. Wang, C.-Y. Hsieh, D. Cao, T. Hou, Mining toxicity information from large amounts of toxicity data, J. Med. Chem. 64 (2021) 6924–6936, https://doi.org/10.1021/acs.jmedchem.1c00421.
[52] D. Chicco, G. Jurman, The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification, BioData Min. 16 (2023) 1–23.
[53] P. Thölke, Y.-J. Mantilla-Ramos, H. Abdelhedi, C. Maschke, A. Dehgan, Y. Harel, A. Kemtur, L. Mekki Berrada, M. Sahraoui, T. Young, A. Bellemare Pépin, C. El Khantour, M. Landry, A. Pascarella, V. Hadid, E. Combrisson, J. O'Byrne, K. Jerbi, Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data, Neuroimage 277 (2023) 120253, https://doi.org/10.1016/j.neuroimage.2023.120253.