**Development of computational rules and strategies to improve (Q)SAR models of impurity assessment: mutagenicity.**

David Mandia

Bachelor of Science in Biochemistry

Supervisor: Dr Claire Ellison

School of Environment and life sciences

University of Salford

Contents

[Abstract 3](#_Toc133407476)

[Introduction 3](#_Toc133407477)

[Drug Development 3](#_Toc133407478)

[Mutagenicity Testing 4](#_Toc133407479)

[ICH M7 Legislations 6](#_Toc133407480)

[Rule Based and Statistical Based QSAR 7](#_Toc133407481)

[Deep Learning 9](#_Toc133407482)

[Aim of the Project 11](#_Toc133407483)

[Methods 11](#_Toc133407484)

[Data collection 11](#_Toc133407485)

[Data Cleaning and Processing 12](#_Toc133407486)

[Model Building 13](#_Toc133407487)

[Performance evaluation 14](#_Toc133407488)

[QSAR Comparison with DEREK 15](#_Toc133407489)

[Results 15](#_Toc133407490)

[Data input 18](#_Toc133407491)

[Network Settings 21](#_Toc133407492)

[External Validation 24](#_Toc133407493)

[Comparison with Lhasa Nexus 26](#_Toc133407494)

[Discussion 28](#_Toc133407495)

[Comparison with Other DNN models 29](#_Toc133407496)

[Issues with Prediction 31](#_Toc133407497)

[Issue with Data and Test 33](#_Toc133407498)

[Potential of deep neural networks 36](#_Toc133407499)

[Conclusion 37](#_Toc133407500)

[REFERENCES 37](#_Toc133407501)

## Abstract

All pharmaceutical drugs contain some unwanted and unavoidable compounds that are present at very low concentrations, called impurities. Unfortunately, some of these impurities are capable of mutating the sequence of nucleic acids. Not all impurities present in all drugs have been tested. Some are extremely difficult, if not impossible, to isolate and test. According to the International Council of Harmonization (ICH) M7 guidelines, if no other test result is available, results obtained computationally are now valid for legislative purposes and for drug approval. In this project, computational models based on deep neural network technology have been developed to predict the mutagenicity of impurities. Overall, the accuracy achieved was comparable to not only some commercially available software, but also to the reproducibility of the most common in vitro mutagenicity assay. This project helps to illustrate the potential of these technologies. While also, it indicates some of the limitations of these computational tools and the in vitro test data present.

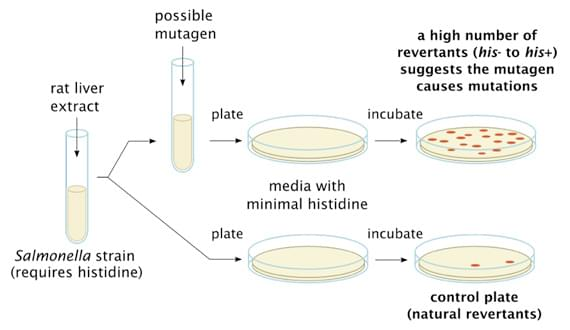
## Introduction

### Drug Development

Over the recent years, the cost to develop a novel drug has skyrocketed reaching up to 2.6 billion dollars. The prohibitive cost is caused by many factors involved in drug discovery. The most important factors are the amount of time needed to develop a novel drug and the attrition rate of newly developed pharmaceutical products. From its formulation to entering the market, a new drug takes up to 20 years due to the testing and clinical trials involved. (Mohs et al., 2017).The rate of attrition is defined as the number of drugs that are stopped from entering the market. Recently, they attrition rate has become extremely high. Nowadays, 19 out of 20 drug candidates are stopped from entering the market ( Xu et al., 2012). For instance, only 7 % of newly formulated oncology drugs go from phase 1 clinical trial to FDA (Food and Drug Administration) approval (Hay et al., 2014). The high degree of attrition is due to issues with pharmacokinetic profiles, and safety concerns. That has led to an imbalance between the risk and the benefit associated with the development of a novel drug. In addition to that, new regulations or further testing may cause some drugs to be withdrawn from the market post-approval. This is usually caused by major issues with the lack of efficacy of a drug, and or serious safety concerns (Siramshetty et al., 2015). That has prompted pharmaceutical companies to increase the level of controls and testing during non-clinical testing, where roughly 4 out 10 drugs are stopped (Waring et al., 2016).

### Mutagenicity Testing

Among the toxicity profiles, mutagenicity is an essential endpoint to consider in the chemical risk assessment of all drug components (Coppi et al., 2022). Mutagenicity is the ability of compounds to interact with nucleic acids changing their structure, but most importantly the sequence of nucleic acids (Jillella et al., 2020). The mutagenic effect has been shown to have a close relationship with carcinogenicity in rats. The reason being is that changes in the sequence of DNA lead to abnormal cellular growth, which eventually causes the formation of tumors (Pandit et al., 2022). In addition to that, the formation of cancerous tissue due to the presence of mutagenic compounds can occur at concentrations of compounds that are extremely low (Yang et al., 2017). The high risk associated with mutagenic compounds has led to the development of methods to assess this endpoint. Sometimes, the mutagenicity of a compound can be roughly estimated by investigating the tri-dimensional structure of the compound (Benigni et al., 2020). In fact, back in the 70s, Miller devised some rules to predict mutagenicity. According to Miller’s rules, most mutagens have low solubility in water, and are electrophilic. Some of the mutagens become electrophilic when they interact with nucleophilic groups, such as nucleic acids (Miller et al., 1977). However, To reliably class a compound as mutagen, an vitro testing must be carried out . The most common *in vitro* test for mutagenicity is the Ames Salmonella assay, also known as the microsome mutagenicity assay (in short Ames test). The Ames test uses modified histidine-dependent strains of bacteria, either Salmonella or E.coli strains. Each of these strains has been modified to carry various mutations in the histidine operon making the bacteria incapable of synthesizing histidine. In addition to that, each of these strains has been modified to be susceptible to certain types of mutations. These bacteria are placed on a plate with media containing a minimal amount of histidine. Some of these bacteria naturally reverse back to the histidine-independent wild type forming a small number of colonies (Ames, 1971). However, the amount of histidine present is not enough to allow cell proliferation and growth. Therefore, the number of grown colonies is low and constant. However, if a mutagenic compound is added to the media, The number of bacteria that reverse back to histidine independence is significantly higher forming a higher number of new colonies called “revertants” (Ames et al., 1975). When testing for mutagenicity, If the compound produces a significant number of revertant colonies when compared to the control, it is classed as positive. If no or minimal growth is observed, the compound is classed as negative. Some chemicals do not interact with nucleic acid, but when they are activated by metabolic enzymes can become potent mutagens. To simulate metabolic activation, liver microsome S9 is added to the media (Figure 1).



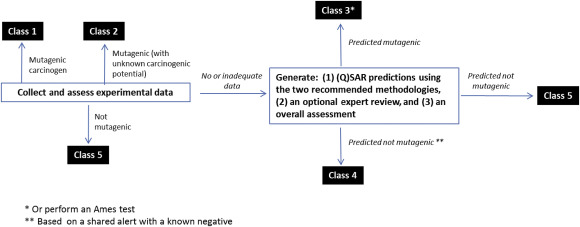
**Figure 1. The difference in bacteria growth between Ames test positives compared to Ames negative.** The possible mutagen is tested on plate containing histidine dependent bacteria (his-). High number of colonies suggest that the compound induced mutation that allowed the bacteria to revert back to wild type histidine independent (his+).

To assess for mutagenicity, a compound is tested on different plates with different strains of bacteria in the presence or not of S9 liver extract. If the compound is positive in at least one of the plates, it is classed a mutagenic. If the compound does not cause a significant amount of revertants in any of the strains is tested on, it is classed as non-mutagenic (Mortelmans et al., 2000). For legislative purposes, compounds are Ames-positive when the Ames test is performed on at least five different strains of bacteria, with or without the presence of the S9 liver extract, and significant increase of revertants is observed in at least one plate. Ames test results can be recorded as continuous or categorical. In the continuous result, the amount of revertants caused by the mutagen is related to its dose (revertants/mol). In the categorical results, the compound is assigned a binary result as either positive/mutagenic or negative/non-mutagenic. The categorical classification is the most used and relevant for QSAR mutagenicity predictions (Smith et al., 2018). Scientist just by studying chemical structures have been able to associate chemical properties to the Ames outcome, shedding light on the possible devise of rational rules for the prediction of this endpoint (Madia et al., 2020). The Ames assay is a cost-effective and rapid assay to assess the mutagenic potential of compounds. Moreover, it has been shown to have a very high correlation to the expensive and time-consuming 2-year test of carcinogenicity in rodents. Scientists have found than 80 % of Ames-positive compounds were shown to have carcinogenic effects in rats. (Hansen et al., 2009). Despite its great advantages, the Ames/Salmonella assay presents some limitations. Firstly, it involves bacteria, which are insensitive to some compounds, such as dioxin ( Hao et al., 2019). Secondly, its reproducibility rate of around 85 % combined with the fact that it needs time, manpower, and at least to 2gr of the compound to be carried out, has prompted researchers to develop new techniques to test the mutagenic effect of substances when they are present at very low levels, such as in impurities (Honma et al., 2020). Impurities are unintended substances found in drugs at minimal concentrations (Landry et al. ,2019). Unfortunately, the presence of impurities in the final product is impossible to avoid. Impurities come from the starting materials used, reagents, intermediates, by-products, or from the degradation of other substances (Hasselgren et al., 2020). Their minimal but unavoidable presence has induced pharmaceutical companies to develop computational methods to test the mutagenicity of impurities that are otherwise difficult to extract. Moreover, computational methods are also used to guide the synthetic process of drugs to avoid or limit the presence of genotoxic impurities (Hemmerich et al., 2020). *In silico* methods such as Quantitative structure activity relationship (QSAR) techniques predict an unknown activity endpoint of a compound. QSAR techniques are based on the principle that compounds with similar structures will have similar activities (Luechtefeld et al., 2018). Nowadays, computers are used in QSAR development to rapidly compute the relationship between chemical structures of compounds and their mutagenic potentials. These structure-activity relationships are then used to predict the mutagenicity of compounds for which Ames test result is not available. When it comes to the mutagenic potential of impurities, and the difficulty associated with their isolation along with the ever-increasing accuracy of these computation methods have allowed computational predictions to be no longer considered a screening tool only. Thus, in some circumstances QSARs can now be used for the legislative risk assessment of mutagenic impurities (Fournier et al., 2023).

### ICH M7 Legislations

The International Council for Harmonisation M7 guidelines state that, in the absence of toxicological in vivo or in vitro mutagenicity data, a result obtained computationally via QSARs models can be accepted for the human health assessment of mutagenic impurities (Hsu et al., 2016). Under ICH-M7 guidelines, mutagenic impurities are divided into five classes. Each of these classes have defined limits that the impurities concentration must be below of. The limits are based on the threshold of toxicological concern (TTC ) for each class. TTC is defined as the dose of compound causing a negligible risk when the compound is present below that concentration. In the instance of mutagenic compound, the threshold of toxicological concern is very conservative. The TTC for mutagenic compounds is associated with the dose capable of increasing the incidence of cancer in 1 out of 100.000 patients over a lifetime. ([ICH](https://ich.org/page/multidisciplinary-guidelines#7-2) website ). In the ICH M7 guidelines, the first class is made of those compounds that are known carcinogens. (Honma et al., 2019).

Compounds in the first class had a two-year in vivo carcinogenicity test on rodents performed and have been found to cause the formation of cancer. They have compound- specific limits depending on the carcinogenicity effect that these compounds have displayed in rodents. The higher the carcinogenicity power, the lower, therefore, the stricter the threshold of toxicological concern. The second class is made of those compounds that are known to be mutagens resulting positive in the Ames test. However, for various reasons, class 2 compounds have not had their carcinogenicity effect studied yet. Class 2 compounds are the compounds that have positive Ames test results in databases. Therefore, they are the compounds that are used to train computational models. The third and fourth category include those compounds for which no data is available. These compounds are actually the most abundant in the pharmaceutical world. Therefore, computational predictions via QSAR models can be used to assess their mutagenic effect (Sizochenko et al. 2019 ). Lastly, the fifth category includes those compounds for which there is no evidence, either experimental or computational, to suggest they could be active (Pandit et al., 2022). Thus, these last compounds are classed and found in databases as non-mutagens (Fig 2). Being non-mutagens, class 5 compounds presence in drugs as impurities is allowed with less strict TTC according to ICH Q3, the non-mutagenic impurities guidelines (Barber et al., 2016).

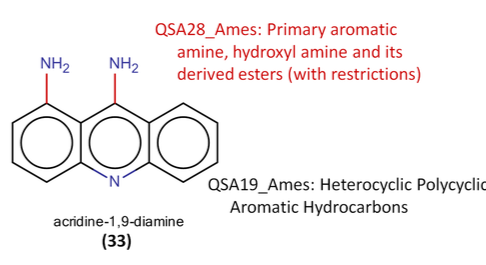


**Figure 2. Mutagenic impurities assessment.** Compounds are divided into classes depending on the carcinogenic/mutagenic present about them. Class 3 and 4 compounds have their mutagenic risk assessed via computational model. (Figure taken from Amberg et al, 2016)

ICH M7 guidelines state that for the Mutagenicity assessment, results predicted via *in silico* QSAR must be obtained via two prediction methodologies one rule-based and one statistical-based (Dongsheng et al., 2020 ). When both techniques predict a compound to be negative, it is sufficient to conclude that the compound is not of mutagenic concern and can be categorized as a compound belonging to class 5.

### Rule Based and Statistical Based QSAR

The rule-based technology is based on structural alerts (SA). SAs are important tools for the prediction of toxicity, because they are derived directly from mechanistic knowledge. Structural alerts identify key substructures related to toxicity. They are also used to help pharmaceutical companies to avoid potential toxic compounds in the very early stages of drug design( Tcheremenskaia et al., 2019). By using structural alerts, Rule based QSARs mimic the decision-making ability of humans by reviewing existing knowledge on mechanistical approaches, physiochemical properties, specificity of structural alerts, and existing Ames test data of analogs. It has been widely applied not only in drug discovery but also in other fields such as cosmetic research and environmental protection ( Zhang et al., 2017). Structural alerts (SA) methods are based on the ability of the model to identify substructures ( Figure 2 ). This feature has allowed them to be recognized as an interpretable and user-friendly technique (Yang et al., 2018 ). Currently, the most widely used commercial tools for the mutagenic test using a rule-based system is Derek (Lhasa, UK).



**Figure 2. A structural alert is highlighted by a rule based QSAR model.** The compound acridine-1,9-diamine was predicted to be a mutagen due to the presence of the primary aromatic amine alerting structure (Hsu et al., 2016)

The other methodology to predict mutagenicity is via a statistical learning approach. Models developed using these methods relate the response to chemical predictors. In models predicting mutagenicity, the response is the Ames test outcome. Predictors are ways to describe a chemical structure (Wichard, 2017). Predictors used are chemical descriptors and fingerprints. Chemical descriptors-based models are built using a table of data containing physical-chemical properties of each compound in the training dataset. Fingerprints are a bitstring representation of the molecule’s geometry (Fan et al., 2018). Statistical models can be designed using chemical descriptors, or fingerprints, or both as chemical predictors. Statistical methods are often referred to as machine learning (ML) methods. Machine learning is a subset of artificial intelligence that consists of techniques that enable machines to improve at tasks using experience learnt from the training data. These methods are used to find the mathematical relationships between a predictor and a response (Ding et al., 2017).

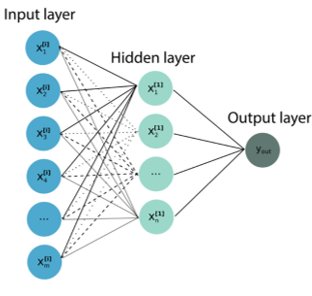
Machine learning methods are divided into supervised or unsupervised learning. Supervised learning is the best method to be considered for the learning process involved in statistical methods with mutagenicity data. In supervised methods, the learning process involves training the model with a dataset containing the label (in mutagenicity model, the label is the Ames test result ), and the molecular descriptors. Machine learning QSAR models mathematically describe the relationship between the descriptors and the response (Winkler et al., 2016).

Once trained, the machine classifies molecules as mutagenic or non-mutagenic based on the structural and physiochemical properties of the molecules. Several types of algorithms are used for the prediction of the mutagenicity of a compound (Pandit et al., 2018). Some of the most widely used machine learning methods are support vector machine (SVM), random forest (RF), k-nearest neighbour (kNN) (Basant et al., 2017).

These methods differ from each other based on the algorithm that is used during the learning process. Support vector machines (SVM) use data to build a hyperplane. The aim of the hyperplane is to separate the data points achieving the largest distance possible between compounds that belong to different classes. Random forest (RF) is a decision tree algorithm, it combines random features into many decision trees that are utilized as a voting system for the final prediction. In QSAR models, random forest algorithms displayed great performance in mutagenicity predictions. Chu et al combined with molecular descriptors and fingerprints into final decision-making tree obtaining high accuracy when compared to other models (Chu et al., 2018 ). K-nearest Neighbour (KNN) is a simple algorithm that classifies data points by the similarities with their neighbours (Yang et al., 2018). KNN has been used extensively in QSAR models. The high performance achieved by QSAR methods using this algorithm is due to its unique ability to classify data points according to their nearest neighbours (Helma et al., 2021). Likewise, most chemicals behave similarly to compounds that are structurally similar to them (Xu et al., 2012).

### Deep Learning

Most recently, with the advances in technology and in computational power, statistical or machine QSAR methods are being developed using the latest technology in Artificial Intelligence (AI), and in Deep Learning (DL). Deep learning is a subset of machine learning based on an artificial neural network, a series of connected neurons. In Machine learning, a neuron is a self-learning unit that computes an output given an input. In neural networks, multiple neurons are stacked together forming a layer. Then, a deep neural network is constructed when multiple layers of neurons are placed one after another (Mayr et al., 2016). A deep neural network attempts to mimic the decision-making ability of the human brain by transforming randomly sampled useful input data into information. The extracted information is then passed to the next layers where they are modified to eventually produce some predictions on the tasks given (Lenselink et al. 2017). Deep neural networks are made of three layers. The first layer is called the input layer (I). This is where the Input data is provided to the neural network, and where the features are extracted from the raw data. Then, information is transferred to the next layer called the hidden layer (K). The hidden layer is made of multiple layers (Hi) of various number of neurons that modify the input data through the use of mathematical functions called activation functions (Kumar et al., 2021). The activation functions assign a degree of importance, called weights, to each of the extracted features. Finally, the information is passed to the output layer (O), where the final classification results is provided (Figure 3).



**Figure 3. Schematic representation of a deep neural network model**. Features are extracted at the input layer, and following math function transformations are eventually passed to the output layer, where model provides the final prediction.

During training, the model updates the weights and the bias to achieve a loss function as low as possible. The loss function is defined as the difference between the predictions and the actual values called targets. This is a pivotal step of deep learning techniques, and it is known as backpropagations (LeCun et al., 2015). In chemistry, Neural networks have been shown to be capable of assigning the weights of the neurons to specific structural alerts. That allows scientists to outpoint the relationships between chemical toxicities and the structural alerts present in the molecules, in a manner similar to rule-based QSAR methods (Wu et al., 2021).

Also, In most machine learning techniques, the performance reaches a plateau as the dataset size increase. Whereas, deep learning models can be trained on a large volume of data of different types without compromising their performances (Martinez et al., 2022) (Ghasemi et al., 2018 ). This is a promising technology for the development of QSAR models showing even greater accuracy than other machine learning methods. In fact, Kumar et al developed a deep learning model that accurately predicted the mutagenicity of 83% of the tested compound. This is a remarkable degree of accruacy, given the proximity to the Ames test reproducibility rate of 85 % (Kumar et al., 2021).

### Aim of the Project

The aim of this project was to develop a computational model capable of predicting the mutagenicity of compounds found in impurities. Initially, mutagenicity data will be collected from publicly available datasets. Successively, the data will be utilized to train a deep neural network. Once the model is fully built and trained, its performance has been evaluated using metrics, and its predictions were compared to the commerically available software Lhasa Nexus. According to the ICH M7 guidelines, the results obtained from our model combined with the results from another rule-based model could potentially be used for the health assessment of mutagenicity in impurities, predictions that are now acceptable for the regulations and drug approval. This project has been an opportunity to shed some light on the ability of neural networks and artificial intelligence to predict mutagenicity. With the improvement of the data quality and the development of better and more powerful AI tools, these predictions have the potential to improve to the point where in vitro and in vivo testing could become obsolete.

## Methods

### Data collection

The data to build and test the model has been collected from various publicly available databases. In these datasets, two essential features are a SMILE (Simplified Molecular Input Line Entry Specification) of the molecule and its mutagenicity value expressed in Ames assay result. A Smiles, or simplified molecular-input line-entry system, is a string representation of the structure of a molecule. An Ames assay result is a binary result, which can be either positive or negative. Each of these datapoints contained the SMILE of the molecule, and the mutagenicity expressed in Ames assay results. The mutagenicity result is also referred as endpoint. The endpoint was expressed in “0” and “1”, where “0” indicates a negative result, and “1” a positive Ames result.

Following data cleaning, the final dataset was made of more than 8000 compounds from a total of 8 databases (Table 1).

Table 1. **Summary of datasets**. The table contains the source of the dataset, and the number of compound in it.

|  |  |
| --- | --- |
| **Database** | **Number of compounds** |
| Hansen et al. 2009 | 7092 |
| Helma et al., 2021 | 8310 |
| Chu et al. 2021 | 5396 |
| Fan et al., 2019 | 577 |
| Data.europa.eu | 786 |
| Benigni et al., 2021 | 227 |
| Pradeep et al., 2021 | 2383 |
| Patel et al. 2018 | 286 |
| Fan et al. External Validation | 60 |
| **Total** | **25,117** |

### Data Cleaning and Processing

The data has been collected from several databases. Some of those databases on the other hand also collected their data from other existing databases. Therefore, there was a high number of duplicates. Duplicate have been removed from the final dataset using the “Remove Duplicates” function in Microsoft Excel. Rows containing the same CAS number, “SMILE”, and “Label” combination were removed. This process removed roughly 10000 compounds. However, removal by selecting only “SMILE” duplicates removed roughly 12000 compounds. That indicates that in some instances the same compounds had two different labels associated with them. Which led to some data quality issues (more on that in the Discussion section). Moreover, molecules made of only two atoms (excluding hydrogens) have been excluded. That has been done by creating a new “length” columns using the “len()” function, and smiles made of a length equal to or below 2 have been removed. In QSAR models found in the literature, it is also common practice to remove charged molecules. Charged molecules have been removed through the “search function”, two charged molecules were found and excluded. This procedure avoids the inclusion of molecules that have structures that have activities that are too unpredictable. Their activities are hard to predict, despite having similar structure to other molecules found in the database. In the literature, this issue is known as “activity cliff”( more on that in the discussion section)

The final database was made of 8.000 compounds. However, an 8,000-compound database was time-consuming to run. Therefore, the model was trained in a database containing 600 compounds that were also randomly extracted from the main dataset using the “Partitioning” nodes in Knime. This allowed rapid model training and tuning. Once finally optimized the model was tested on the database containing 8,000 compounds.

### Model Building

Following data collection and cleaning, the model has been built using the latest version of the free KNIME platform ([KNIME | Open for Innovation)](https://www.knime.com/). The Knime software is a coding-less data analysis tool that allows users with minimal data science experience to design a workflow for computational models. To work on chemical structure, RDKit extension has been downloaded. RDKit permits the extraction of chemical descriptors and fingerprints starting from the chemical structure expressed as SMILES strings.

These features are then used to train the models and predict mutagenicity activity. The two types of descriptors input for this model were fingerprints or chemical descriptors. Fingerprints are bit-like string representation of molecule made of “0 “and “1”. There are various types of fingerprints with different lengths depending on the need.

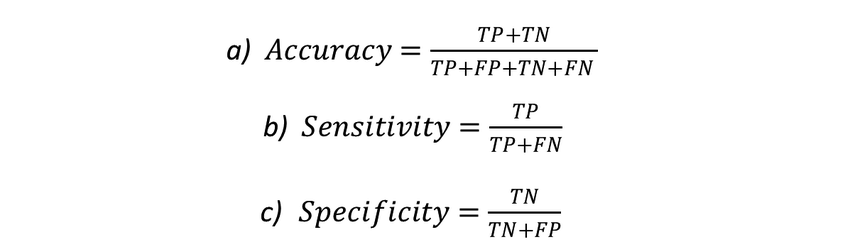
Chemical descriptors are numerical measurements of chemical-physical properties. Some of the most relevant ones include solubility, molecular weight, number of rings and so on. There were a total of 130 descriptors. However, some of them were found to be repeated in a very similar manner, such as “Molecular weight” and “Exact Molecular Weight”. Also, other descriptors were found to have a minimal impact on the prediction quality. Therefore, to allow faster and more accurate predictions, some of these descriptors were removed ( More on the choice and impact of various descriptors on the Result:Data input section).

The model was then designed using the Keras extension in Knime. The Keras extension allows a rapid and easy built of a deep neural network model without the need to code. Keras extension included different nodes for the setup of the whole workflow for neural network, from input, to learning, and to predicti. These nodes were inserted in a fully connected workflow and tune to achieve optimal performance.

Data-points were extracted from a Comma Separated Values file (.csv) using the “File Reader” nodes. SMILES strings were converted from “string” type to “chemical SMILE” type to extract the features needed using the “Molecular Type Cast” node.

### Performance evaluation

Every model requires some metrics to evaluate and to compare its ability to predict certain endpoints against other available models. The main metrics are sensitivity, specificity, and accuracy. Sensitivity is calculated by dividing the number of true positives divided by the number of true positives plus the false negative. It is a measurement for the ability of a model to predict positive samples. Specificity is sensitivity’s analogs; it is the measurement of the ability to predict negatives. It is calculated by dividing the number of true negatives by the total number of actual negatives present, including those incorrectly classified by the model as positive. Accuracy is the sum of all correct predictions, both true positives and true negatives, divided by the total number of predictions (Figure 4). Accuracy was used to compare the model against the other QSAR models available. In addition to that, it was used to study and tune different features based on their impact on the overall performance within our model.



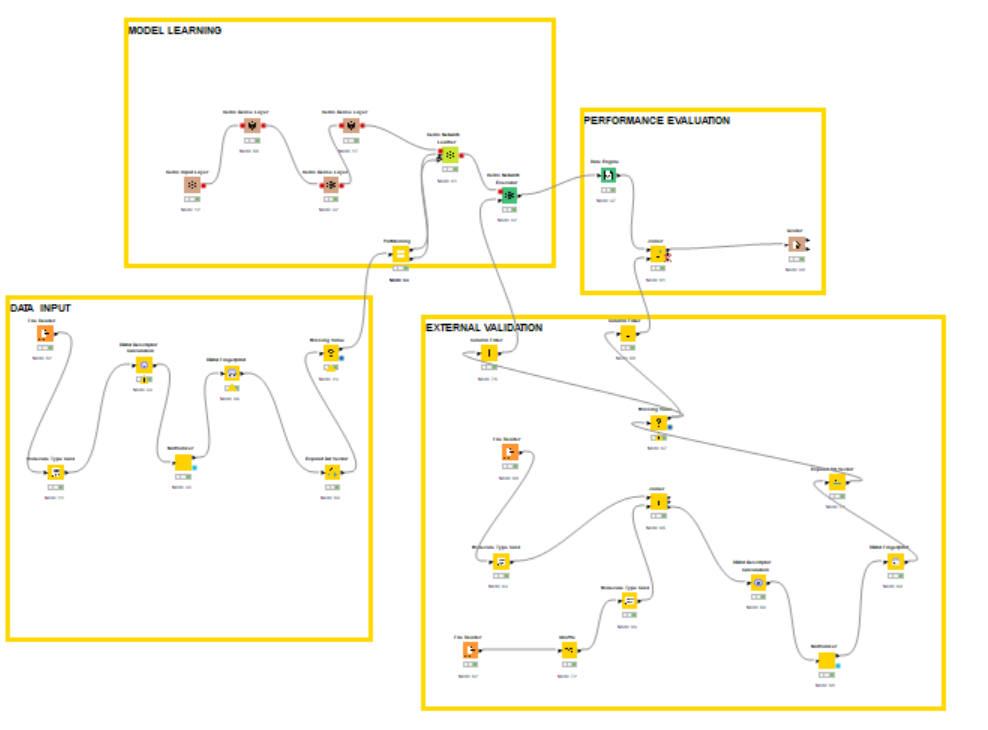
**Figure 4. Metrics for the model performance evaluation.** Mathematical formula to measure Accuracy, sensitivity, and Specificity. TP stands for True positive, the compounds predicted to be positive and are actual positive. TN (True negative ) are those compounds predicted to be negative and actual negative. FP (False positive) are those compounds that the model incorrectly classify them as positive, but are negative. FN ( False negative) are predicted to be negative, when they are actually positive. These metrics are not for QSAR models only. In fact, they are essential and universal for all binary classification models.

### QSAR Comparison with DEREK

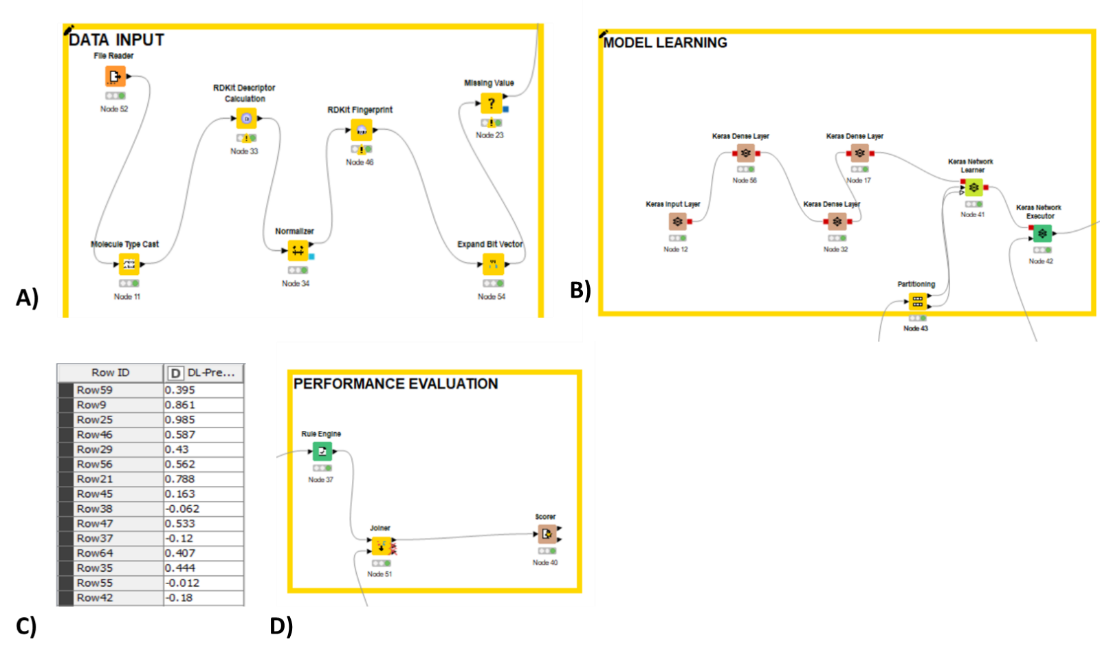
Once fully built and optimized, the model’s performance was compared against the Lhasa Nexus package using an external validation dataset. The Lhasa package required molecules to be in the Structure Data Format (.sdf). Therefore, the structures expressed in SMILES have been converted using the OpenBabel converter function ([OPENBABEL](http://www.cheminfo.org/Chemistry/Cheminformatics/FormatConverter/index.html) ). The structures have been loaded into the Nexus and the batch was set to run and predict the mutagenicity in bacteria (The Ames test ) for the external validation dataset. Within the Nexus package, mutagenicity was predicted using the statistical QSAR model Sarah, the rule-based model DEREK, and the combination of both. The output has then been compared with the label, and the accuracy compared to our model.

## Results

A general deep neural network that predicts endpoint from chemical structures was built by consulting literature and KNIME documentations. Once designed, the models had to be adapted to the mutagenicity predictions. The models built were improved until their predictive ability was considered acceptable. An acceptable performance was defined as an accuracy comparable to the reproducibility of the in vitro mutagenicity Ames test. The acceptable accuracy was of around 70/80 %. Several models were built on Knime. The models were optimized to present a high predictive ability. Considerations on the model features and the parameters are described below. Among all models built, there are some steps that they all share. These steps are referred to as data input, model learning, validation (external or internal), and performance evaluation (Figure 5 ). All models regardless of the input or the parameters setting had a similar workflow. The data input was split into three parts using the “Partitioning node”. The first split was a 70/30, 30% of the data was used to test the predictive ability of the model. to the model for “internal validation”. The remaining 70% of the data was partitioned again with an 80/20 split, where 20 % of the data was used for internal validation. This validation is considered internal as the data came from the same dataset. Nevertheless, the data used to internally validate the model is not part of the data that is used to train the model(Fig. 6 A). The remaining 80 % of the data from the second partitioning were supplied to the model for training. The input data was then utilized for training by the “Keras Network Learner”. “Keras Network Learner” node uses mathematical function to relate the molecular descriptor values to the mutagenicity label. The weight assigned to each of these functions then will be used to determine the mutagenicity of molecules with unknown labels during the validation process. The “Keras network Executer” adopts these functions to predict mutagenicity( Fig. 6B), it assigns a numerical mutagenicity value called “DL Pred” (from Deep Learning prediction) to each molecule (Fig 6C ). This value can be any number. However, since the mutagenicity is expressed as either positive (“1”) or negative (“0”). It must be rounded to either 0 or 1 by the “Rule Engine" node. Finally, the prediction and the label from the validation dataset are joined and the prediction are compared using the “Scorer” node. It compares the target “Label” column of the original dataset to the predictions made by the model. Its output is a table containing some statistical metrics to determine the performance of the model (See Fig 6D) .



**Figure 5 . A general fully-connected workflow of the Deep neural network models designed on KNIME.** The figure shows the final model workflow divided into the four steps of computational prediction.



**Figure 6. More detailed representations of the section of the Deep neural network models.** The datapoint are extracted from the model, and features calculation is performed. The data is then partitioned into training and validation (A). The Deep neural network model designed using Keras nodes on KNIME (B). The mutagenicity prediction provided by the model in numerical value (C). The “Rule Engine” converts the numerical mutagenicity resultt into a binary positive or negative prediction. The predictions are then compared to the label by the Scorer node (D).

The accuracy was highly dependent on settings such as the type of data input, shape of the network, and parameters.

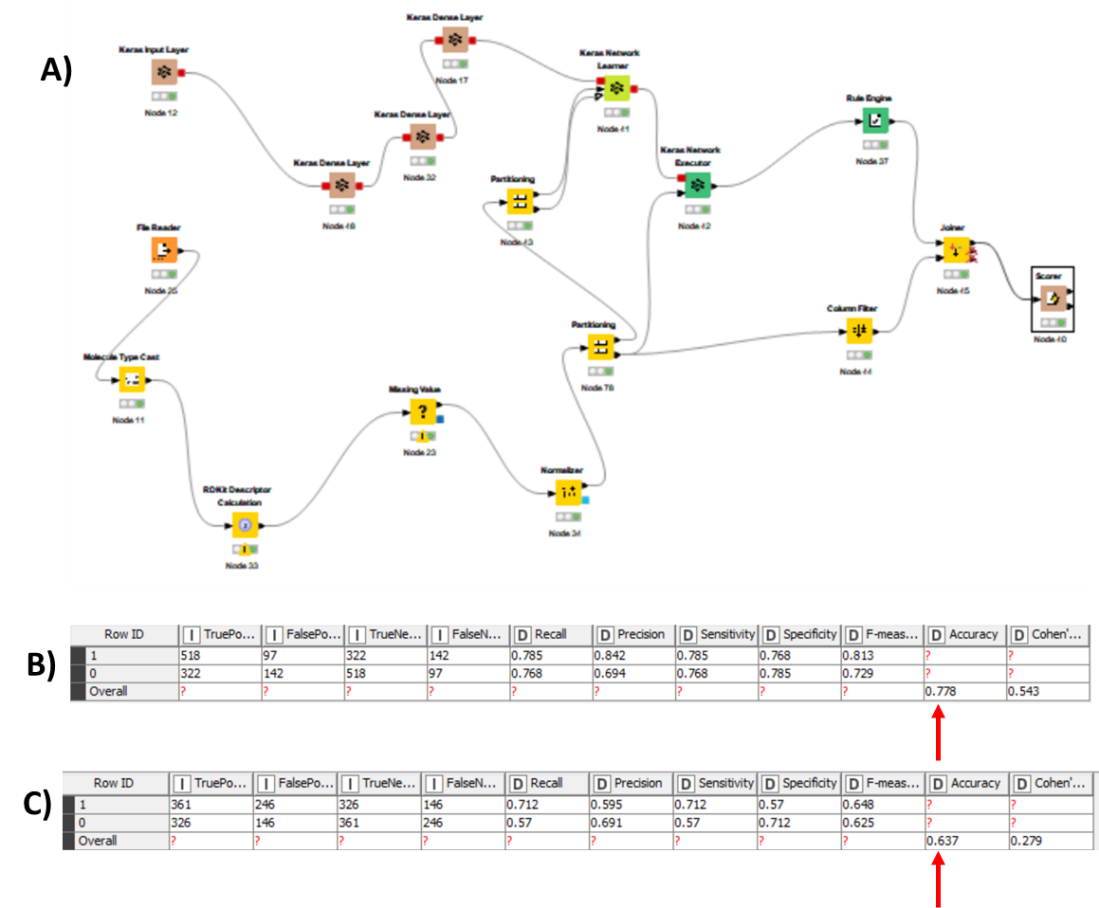
### Data input

This is the features that was shown to have the greatest impact on the accuracy. Since QSAR models require the best possible description of a 3-D structure to be able to predict the activity associated with that structure. During the designing of the models, several types of input were tried, and their performance evaluated.

#### Molecular descriptor only

The most accurate descriptors-only models were built on Knime with the help of the RDKit nodes. A total of 128 molecular descriptors were calculated from the SMILE of each structure using the “RDKit Descriptor Calculator”. Some molecules had missing values in some values. If one of the descriptors was missing for a given molecule, the whole row had to be removed to avoid errors during the training of the model. All values were then normalized to be within 0 and 1. At this point, the data processing was complete, and the input data was given to the model for training and validation (Fig 7A).

In this instance, the best-achieving model had an accuracy of 77.8% (Fig 7B). A list of 42 molecular descriptors called Molecular Quantum Number (MQN) was removed from the input. MQNs are algorithms used to represent molecular properties (Nguyen et al., 2009). The removal of MQNs descriptor not only reduced the running time but also increased the accuracy. In fact, keeping all other variables the same the accuracy in the model containing the MQN was lower the model with the MQNs removed, 63.7 % (Fig 7C) against 77.8% (Fig 7B) The MQN were removed by excluding them from the list of “Input Data” in the “Keras Network Learner” node.

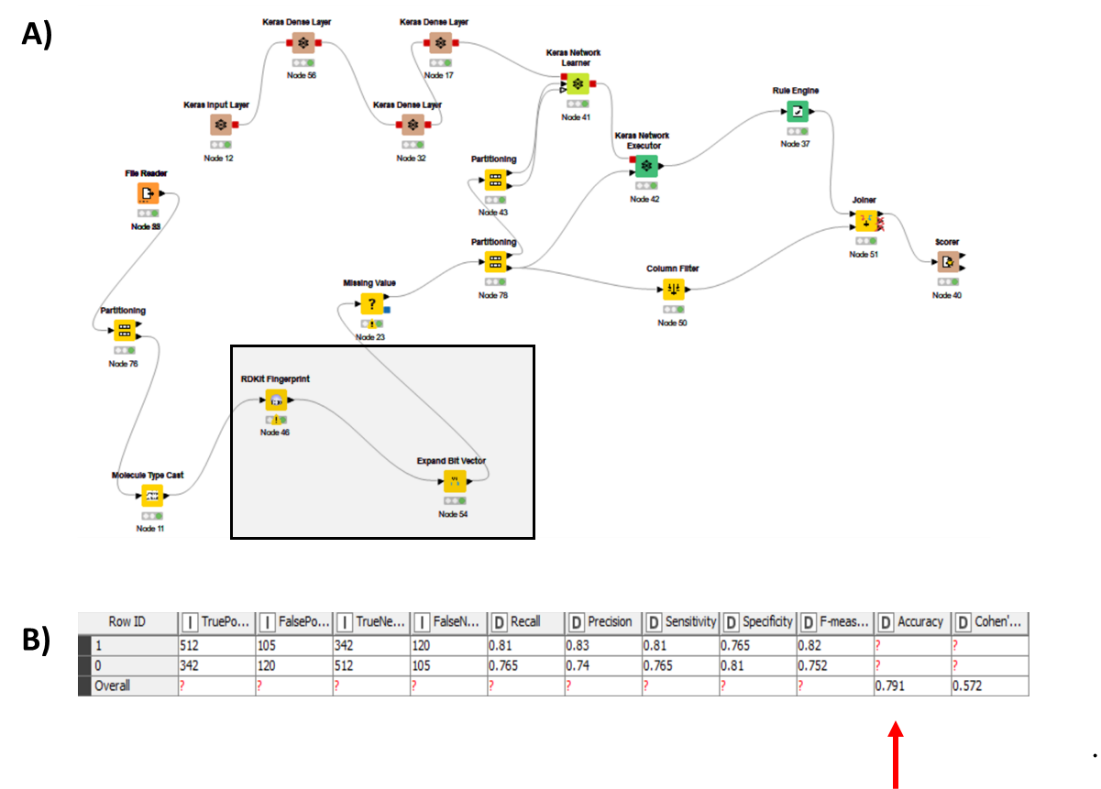


**Figure 7. A molecular descriptors-only neural network model.** A fully-connected deep neural network model trained on physico-chemical properties (A). The best accuracy achieved by models having molecular descriptors as input (B), compared to when all RDKit descriptors such as MQNs were included (C). The accuracies of both models are pointed out by the red arrows.

#### Molecular fingerprints

As mentioned, fingerprints are strings made of “0” and “1”. There are different types of fingerprints of various lengths. The best-achieving model was the one built on MACCS fingerprints, 166 bit-long strings. Shorter fingerprints were not enough informative to allow a decision to be made. In fact, the accuracy was found to be slightly above a coin-toss prediction, with the highest accuracy being around 60%. On the other end, longer fingerprints were too information-dense, and the model was incapable of extracting the most contributing features resulting in an accuracy of around 45 %, even below a random guess prediction. In addition to that, longer fingerprints had strings of up to 1048 bits long, which tremendously slowed down the learning process.

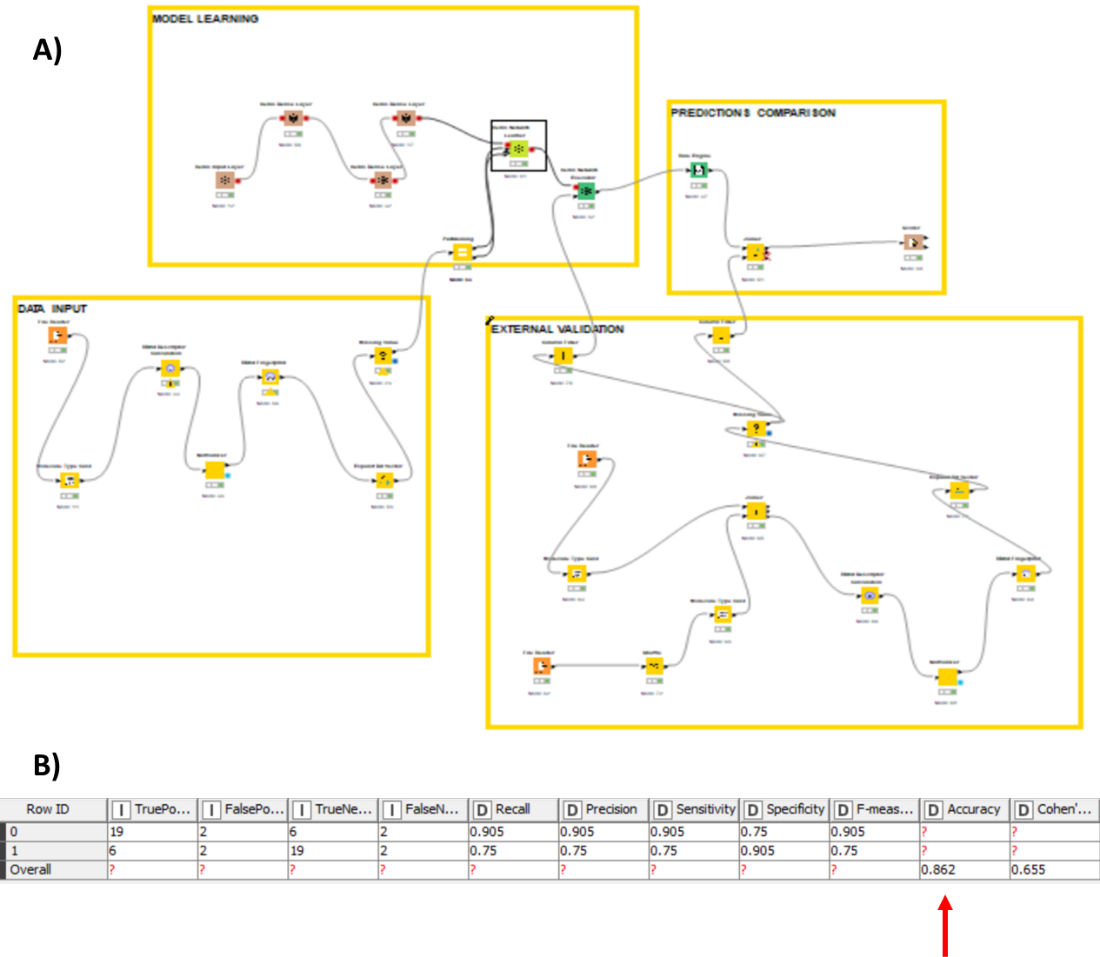
Fingerprints were obtained from the SMILES using the “RDKit Fingerprint” node. The type of fingerprint was chosen from the drop-down menu in the node configuration. Once generated, the fingerprint is converted in tabular format by the “Expand Bit Vector” node, which takes the fingerprint string and places every bit in a separate column (Fig 8A). In the model, the 166-bit-long MACCS fingerprint is separated into 166 columns containing either “1” or “0”. As with the descriptors only model, the data was then fetched into the model. The best-achieving fingerprints-only model had an accuracy of roughly 79.1% (Fig 8B). However, fingerprints models required a high amount of work and consideration with many trial-and-error tuning attempts just to achieve acceptable accuracy. Unlike, descriptors-only model that resulted in an acceptable ability to predict even with no tuning at all.



**Figure 8.** A fully-connected deep neural network model trained on MACCS fingerprints as descriptors. The workflow of this model with the nodes that are nodes necessary to process molecular fingerprints highlighted by the grey box (A). The Accuracy matrix generated by the best achieving model, with an accuracy of 79.1% (B) The accuracy of the models is pointed out by the red arrow.

#### **Fingerprints and molecular descriptors combined**

The final model was built using molecular descriptors without the MQNs properties combined with MACCS fingerprints. This type of model was found to be the most accurate. It was built by combining the RDKit descriptors along with the RDK Fingerprints nodes in a single workflow. The table made of the two combined was fetched into the “Keras Network Learner” node, and used to train the model (Fig 9A). Once trained the model was tested and its statistical metrics measured. The fingerprint with molecular descriptors combined model achieved an accuracy of 86 % (Fig 9B)even during external validation.



**Figure 9.** A fully-connected deep neural network model trained on MACCS fingerprints and physico-chemical properties combined as descriptors. The fully connected workflow of the neural network model trainied on combined input ( A). Accuracy of the model in external validation (B) The accuracy of the models is pointed out by the red arrow.

### Network Settings

#### Shape of the network

In this process the highest possible accuracy was achieved via trial and error, where two models presenting only one different parameter were compared, the model with the better accuracy was kept, duplicated, and compared again with a different parameter.This process was repeated till accuracy was acceptable enough, and it did not improve any further.

The neural network was built using the KERAS package in KNIME. Deep neural network is made of three layers, An input layer, a hidden layer, and an output layer. In the KNIME, the input layer is referred to as the node “Keras Input Layer”. It does not require any implementations as its only parameter is the number of neurons, which is determined by the number of features/columns in the training dataset. The hidden layer is represented as “Keras Dense Layer” nodes . However, the last “Keras Dense Layer” node is part of the output layer. Number of neurons in the output layer is defined based on the type of classification. In Ames prediction, the result is a binary classification. Therefore, the number of neurons in the output layer was only 1.

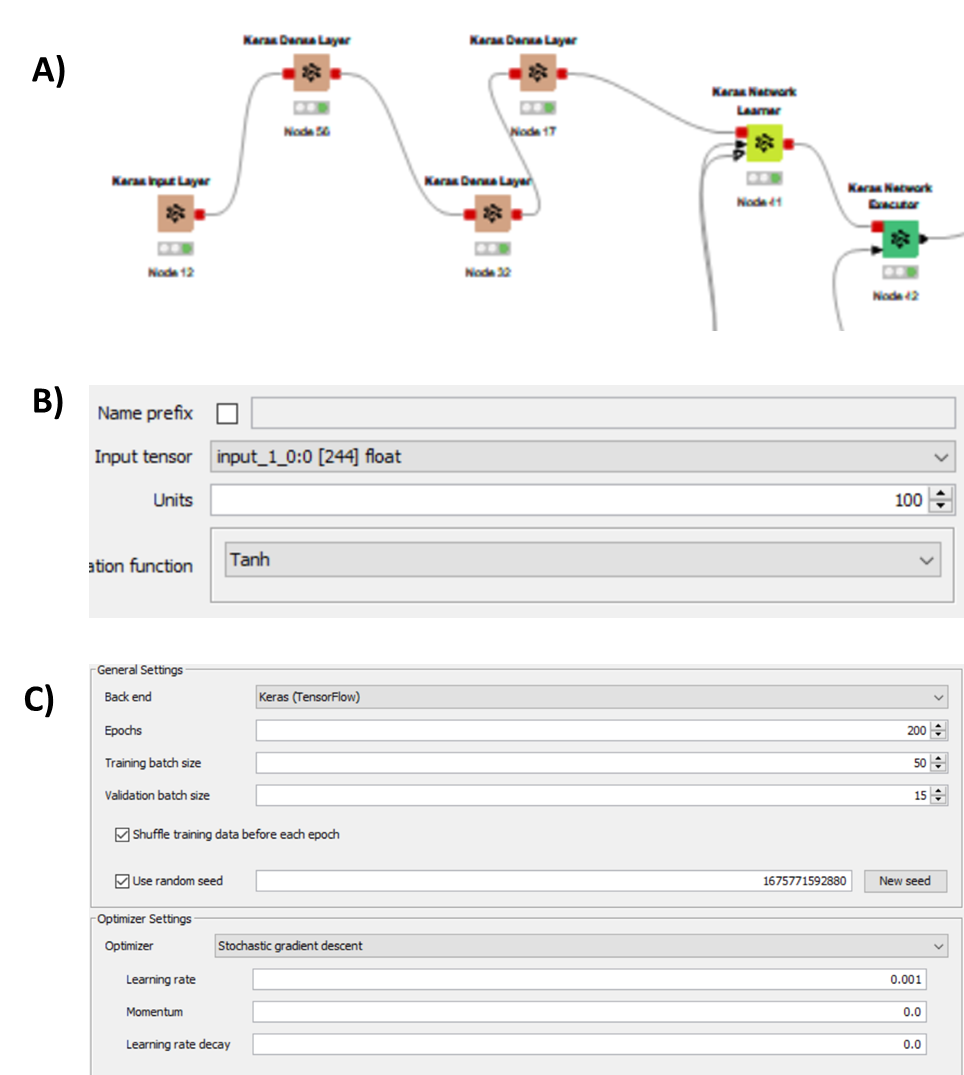
In this project, the number of hidden layers was increased or reduced in attempt to improve the performaces by adding or removing “Keras Dense Layer” nodes. However, in most instances accuracy was found to be unsatisfactory. Accuracies in deeper or shallower neural networks, networks presenting more or less than two hidden layers, were below 60 %. The predictive ability of these models was unacceptable, even when other features were optimized. Therefore, any other development was conducted on neural networks made of 2 hidden layers (Fig 10A). These last models were the best performing models, with accuracies that were very close to 70/80 % regardless of the input or network optimization. Once the optimal depth of the network is established, the width of the network is tuned. The width of the neural network is referred to as the number of neurons in each layer. It is considered as one of the main features of any neural network. The input and the output layer have their number of neurons set. The number of neurons in the input layer was changed based on the number of features present in the training data, the number of columns in the training data. Nevertheless, it remained set once the features of the input were established

The first hidden layer was made of 100 neurons (Fig 10B). The number of neurons in the second layer was reduced to 50. In the second layer the features were already partly extracted and their importance in the final decision already partly assessed. Neural networks presenting more neurons had acceptable performances. However, they were found to drastically decrease their accuracies when the number of compounds in the training dataset was increased. Models made of neural networks with less neurons produced predictive accuracies that were independent of the size of the training dataset. Yet, the accuracy was still below 60 % and deemed not accurate enough.

#### Parameters tuning

In general, all models need some parameters adjustment. This procedure allows the model to adapt to the training mode, predict the specific endpoint, while obtaining the best performance possible. This process was carried out manually. Similarly, to the procedure for the shape of the network tuning, the model was duplicated, a single parameter was amended, and its impact on the performance evaluated. One of them major parameter involved with the “Keras Dense Layer” node was the activation function. It is defined as the type of mathematical operation that the layer applies on the data. Models with activation function “Tanh” performed the best (Fig 10 B). Although, activation functions had negligible effect on the accuracy . The difference between “Tanh” activation function and the worst performing activation function, which was “sigmoid”, were between 0.5 and 8 % depending on the model.

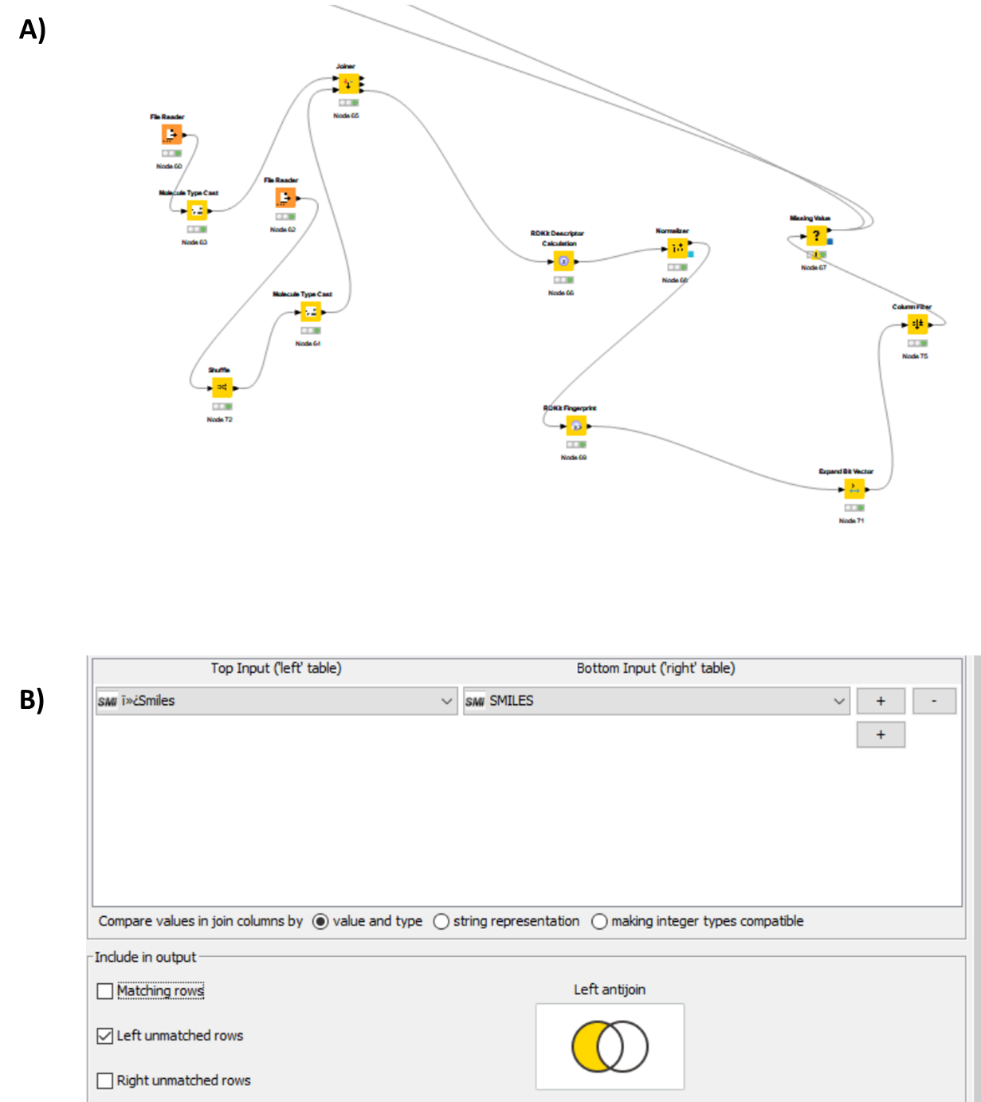
The “Keras Network Learner” was also configured to achieve a high predictive ability. The number of epochs, defined as the times the whole database is run before predictions are generated, was set to 200 (Fig 10 C). The number of epochs had a more relevant impact on the run-time rather than on the accuracy. Nevertheless, an ephemeral number of epochs reduced the final accuracy of about 20/30 % in some instances. In the same node, the Optimizer also had a considerable impact on the performance. The optimizer is defined as the type of operation that the model performs to amend the weight assigned to each chemical feature. These weights are amended to achieve the lowest loss between the predicted values and the targets. The best achieved optimizer was the “stochastic gradient descent” with a learning rate of 0.001 (Fig 10C). Where the weights are amended gradually after each epoch based on their impact on the loss.



**Figure 10. Tuning of the network depth and width along parameters optimization.** The final model had 2 layers in the hidden layer section. The figure shows 3 Keras dense layer. Since, the last Keras Dense layer is made of one neuron and is part of the output layer (A). Configuration of the first Keras Dense layer node. The width of the first dense layer was set to 100 with Tanh used as activation function (B) (The second dense layer had the same activation function, but 50 neurons). Parameters such as the number of Epochs and optimizer are set in the Keras Network Learner node configuration (C).

### External Validation

Once the models were fully built and optimized to have the best predictive ability, they have been tested on a separate dataset. Fan et al., 2018 contained a dataset with 60 compounds that the group used for external validation, this database was loaded into Knime for external validation of the models built (Fig 11A). External validation was conducted on the best achieving models. The external validation and the training dataset were combined using the “Joiner” node. The joining operation occurred via rows having an identical SMILES column. From the output, only the “Left Join” results were taken. The left join is made of those datapoints that are present in the left database (the external validation), but not in the right database (training dataset ) (Fig 11B). A total of 30 compounds were part of the left join and were used for external validations. The data was then processed to have the molecular descriptors and fingerprints combined. A new node was added called “Column filter” to remove the column called “Classification” present only in the external validation dataset (Fig 11A). The column classification contained the Ames result expressed as “positive” or “negative”, instead of “1” or “0”. The presence of this column trained the model to compute mutagenicity with a column that already had the target to predict. In fact, without the removal of that column all data points the accuracy reached 100%. However, the “Classification” did not appear on the data input settings in “Keras Network Learner” node as the node was configured with a dataset that did not contain the “Classification” column. Therefore, the classification column in the external validation set was still part of the training process. That led to incorrect external validation prediction with accuracy of 1 ( 100%). Once the “Classification” column was removed, the model computed the mutagenicity on the external validation dataset. The predictions obtained were then compared to the targets, and accuracy was determined. In all models, the accuracy was reduced during external validation. The descriptor-only model accuracy went down to 64%. The fingerprints-only went down to 73 %. Nevertheless, the accuracy in the best performing model was still acceptable. The final accuracy in the model that combined fingerprints with molecular descriptors was of 86% (Fig 9B).



**Figure 11. General external validation workflow.** Features are extracted and fetched to the model for prediction (A). Left join operation were performed on the training and external validation to ensure that the validation data is not part of the training datasets, the left join operation was done on the “Smiles” column, the only columns that two separate database have in common to describe a molecule (B). Cas compounds registration number was not available for part of the training dataset.

### Comparison with Lhasa Nexus

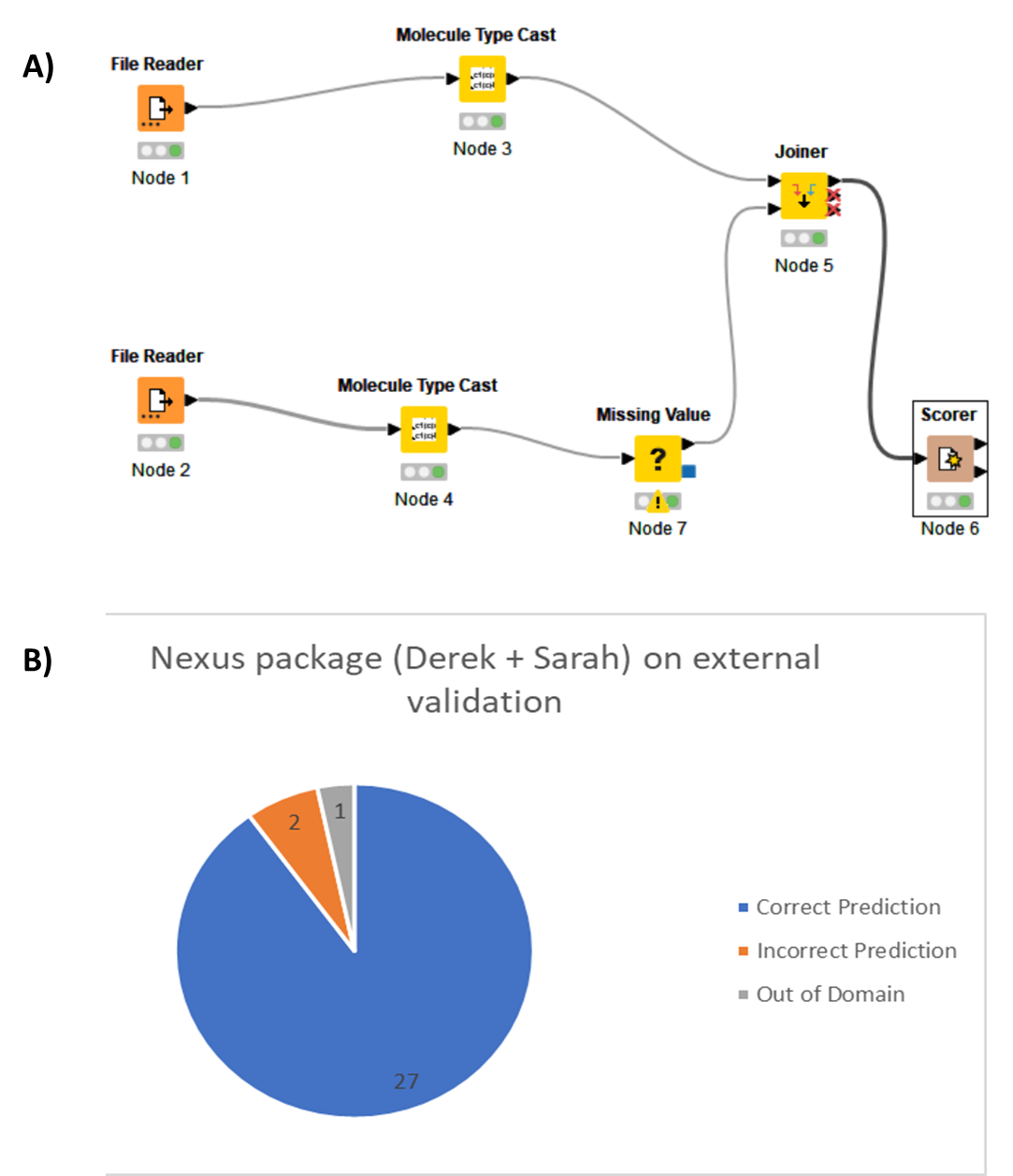
The model was then compared to the Lhasa Nexus package. The Nexus package includes both a rule based QSAR called Derek, and a statistical model called Sarah.The structures in .sdf format were loaded into Nexus, where the ICH M7 batch prediction was run. Results obtained from the DEREK and SARAH run are combined together in a column called “Final Call” (Table 2). The predictions were then loaded into KNIME, the “SMILE” column was used in the “Joiner” to combine the predictions from the Nexus package, and the designed Deep Neural Network model. Predictions from the “Final Call” were compared to the labels, the known Ames results (Fig 12A). Nexus’ accuracy was calculated to 90%, with 2 incorrect predictions and one “out of domain prediction” out of a total of 30 compounds (Fig 12B).

Nexus Package Accuracy = 90 %

Deep Neural Network Accuracy = 86%

**Table 2. Example of Lhasa Nexus ICH M7 batch run results.** The table contains the chemical structure as SMILES, and the predictions from the rule based QSAR DEREK, and the statistical QSAR Sarah. The overall results are in the final call column.

|  |  |  |  |
| --- | --- | --- | --- |
| **SMILES** | **Derek Prediction** | **Sarah Prediction** | **Final Call** |
| NC=1C=CC(Cl)=CC=1 | INACTIVE: No misclassified or unclassified features | POSITIVE - 100% | 1 |
| CC(=O)NN | PLAUSIBLE: Alert033 - Hydrazine or monoacyl- or monosulphonyl-hydrazine | POSITIVE - 100% | 1 |
| COC(=O)C=1C=C(C=CC=1N)C=2C=CC(N)=C(C=2)C(=O)OC | PLAUSIBLE: Alert352 - Aromatic amine or amide | POSITIVE - 100% | 1 |
| NC=2C=CC(CC=1C=CC(N)=CC=1)=CC=2 | PLAUSIBLE: Alert351 - Aromatic amine or amide | POSITIVE - 100% | 1 |
| CCCCNC(=O)N1C(NC(=O)OC)=NC=2C=CC=CC1=2 | INACTIVE: No misclassified or unclassified features | NEGATIVE - 44% | Conflicted |
| CN1C(N)=NC=2N=CC(=CC1=2)C=3C=CC=CC=3 | PLAUSIBLE: Alert354 - Aromatic amine or amide | POSITIVE - 100% | 1 |
| NC=1C=CC=C2C=CC=CC=12 | PLAUSIBLE: Alert354 - Aromatic amine or amide | POSITIVE - 100% | 1 |
| NC=1C=CC(=CC=1)C=2C=CC(=CC=2)C=3C=CC(N)=CC=3 | PLAUSIBLE: Alert351 - Aromatic amine or amide | POSITIVE - 100% | 1 |
| CS(=O)(=O)OCCNP1(=O)OCCCN1CCCl | PLAUSIBLE: Alert069 - Nitrogen or sulphur mustard\\Alert027 - Alkylating agent | POSITIVE - 39% | 1 |



**Figure 12. Comparison with the commercial Lhasa Nexus software on the external validation datasets.** Knime workflow for the comparison between the software predictions and the labels (A). The pie chart of the overall performance of the commercial software Lhasa nexus (B)

## Discussion

### Comparison with Other DNN models

During this project various models were built on Knime. The model was built and optimized through trial-and-error, and by consulting the literature when possible. In our models, the molecules were described using chemical descriptors and molecular fingerprints or both. Chemical descriptors and fingerprints were computed by Knime using the SMILES representation of molecules. Although easily storable and interpretable, SMILES strings are not designed to display molecular similarity, with instances where similar molecules presented different SMILES (Hung et al., 2021). Out of all descriptors, 66 were used to build the deep neural network model. However, A better selection of the chemical descriptors to use could be optimized, allowing the model to achieve even greater performance in terms of efficiency. Some descriptors are more relevant for their impact on the ability of a compound to cause nucleic acid sequence alterations. In fact, the solubility (ClogP, the partition of a molecule between octanol and water) , the molecular weight, the number of rings, and the molecular surface have been shown to be twice as important as other less relevant descriptors such as hydrogen bond donor (Chu et al., 2021).

The rest of the input neurons were then used to train the model on molecular fingerprints that were extracted from the SMILES strings. The combination of the chemical descriptors and molecular fingerprints in a single model has previuosly been shown to achieve better performances compared to when the two types of descriptors were used separately (Chakravarti et al., 2018).

In our project we tried to combine fingerprints and molecular descriptors in a Deep neural network. The extracted data was used to train a neural network that was optimized to have a high number of neurons with only two dense layers. This allows rapid training and effective features extraction. As observed by Kumar et al., who also built a deep neural network model with two dense layers and a high number of neurons and obtain the best level of performance in terms of accuracy (Kumar et al., 2021).

The developed fingerprints and molecular descriptors combined deep neural model performed well in the external validation dataset. The accuracy was around 86%. Recently, Deep neural network models have achieved exceptional performances, with some instances performing better than traditional machine learning models. Kumar et al. Designed a DNN model that achieved an outstanding accuracy of 83% with an external validation dataset of over 1000 compounds compared to our deep neural network which had only 30 compounds. The number of datapoints in the external validation dataset is a key factor to consider for QSAR models performance and their usability in real-world settings Their model, like ours, also presented a high number of neurons and two dense layers in the hidden layers. In addition to that, they also noticed reduced accuracy during the external validation. Their accuracy as well was lower when the number of datapoints in the training dataset was increased. This suggests a possible overfitting of the training data in neural network models. (Kumar et al., 2021). In fact, for applicability and rapidity purposes the training process was assessed using accuracy as the main index. However, using accuracy may lead to the detection of features that are too specific to certain compounds (Sharma et al., 2021). These features are then not represented in the external validation, decreasing the model's predictive ability. This issue is known in QSAR models as the detection of features with high accuracy but low significance of these detected features in real-world data settings (Yang et al., 2022).

In addition to that, the preference for accuracy as a main metric may cause the design of model with high accuracy, but with unsatisfactory other metrics. This is especially the case with models having low sensitivity. Models with high sensitivity are usually called conservative models. These models tend to assign mutagenicity to some dubious molecules. This increases the level of false positive, reducing the accuracy. However, predicted mutagens tend to have their mutagenic potential tested in vitro, or at the very least reviewed by scientists. These equivocal, or positive calls are then reviewed. This leads to increase drugs safety from the risk of mutation-causing compounds. Therefore, conservative models are preferred for the legislation and the risk assessment of mutagenic impurities (Fournier et al., 2023).

In fact, on an analysis of over 17,000 compounds, Benigni et al. Revealed a false negative rate of around 23 %. These false negative compounds are molecules that would otherwise proceed with no further testing. Which suggests the implementation of more conservative QSARs for the correct classification of non-mutagens. (Benigni et al., 2019).

### Issues with Prediction

Some QSAR models may present different efficiency as well. Efficiency in QSAR models is defined as the percentage of single class prediction, where a compound is predicted to be either positive or negative.

The designed deep learning model presented an efficiency of 100 %. This is the case because the training was done on compounds with binary results, being either positive or negative. These databases do not consider the mutagenic strength of compounds (revertants / mg of substance). More importantly, the “Rule Engine” node was set to assign only “0” and “1” to the mutagenic values. That led to clear and unequivocal calls by the model. However, the presence of equivocal is sometimes beneficial for a model. Since equivocal calls are caused by the compounds being outside of the model's domain of applicability. A high degree similarity between the query molecule and the training set suggests that the model was trained with compounds that are somehow similar, and therefore they should behave similarly to the test compound (Sushko et al., 2010). The developed deep neural network model does not implement such a feature. Therefore, it does not have a defined applicability domain. Nevertheless, chemical similarity could be easily computed using the Tanimoto coefficient on chemical fingerprints (Norinder et al., 2018). Which could be considered as a possible improvement for our models.

Regardless of chemical similarity, what happens in the chemistry world differs from what QSAR models predict to happen. Statistical QSAR models predict activity based on the similarity between compounds. However, there is a phenomenon known as “activity cliff.” According to this phenomenon compounds presenting a high degree of similarity may still present totally different activities. These differences may involve toxicities outcome, which could then lead to errors in QSAR predictions. Two compounds that are almost identical from a structure point of view, may still have different Ames test results. If the model is trained on one of the compounds, it may then incorrectly predict the activity of the second compound due to their high degree of similarity. (Gini et al., 2019)

These compounds are structurally similar. Nevertheless, they present diverging biological activities. A model trained with the non-mutagenic compound may call the mutagenic compound as negative, leading to a false negative call. For example, it may happen when dealing with radical compounds predictions. In the case of radical compounds, the activity cliff is cliff issue is easily understandable and amendable. However, the issue also involves compounds that do not present such a distinguishable difference Guo et al., 2019 ).

Even when trained with many similar molecules, the model may still call compound incorrectly. The model may call some compounds as positive when they are not, these compounds are known as false positives. The issue with false positive predictions is more common than expected. A group of positive compounds in the training set may share a particular substructure or feature. The model may infer that they are positive because of that feature being present. The model will then predict compounds in the external validation containing those features as positive when they may not be. This increase the number of false positive predictions (Yang et al. ,2017). However, the compounds in the training set presenting that particular feature might have been positive due to the presence of multiple features that were different and independent of each other. The fact that they shared that feature, it trained the model to recognize that substructure/feature as a positive inducing feature. Compounds having that substructure and being positive was just a coincidence, and they were positives for other alerting structure that they did not share between each other, leading to incorrect predictions. (Benigni et al., 2021).

A model or even two models of distinct types can both erroneously class a compound as non-mutagen. However, inconclusive explanation of the call may induce scientists to analyze the compound and investigate for evidence of the prediction. If necessary and possible, an in vitro Ames test will be carried out on the test compound. This is a process called “expert review”. Conflictual calls, compounds with no analog, or even incorrect predictions indicate how expert review is still required and necessary in these matters. During the expert review, researchers study structural analogs and possible mechanisms of action through the use of publications and supporting documents to determine the mutagenic potential of a compound. Expert reviews may refute and overrule QSAR models predictions (Fukuchi et al., 2019 ).. In fact, in a study expert review changes the outcome of around 13% of predicted compounds, resolved 72 % of the equivocal predictions, and assign a mutagenicity outcome to an outstanding 95% % of the compounds that the models were unable to generate a prediction ( out-of-domain compounds). These numbers show the importance of researchers in the final decision, regardless of the performance of the model used. Expert review is advised by the ICH M7 guidelines, and when combined with computational predictions leads to an increased confidence in mutagenicity predictions. (Landry et al., 2019).

### Issue with Data and Test

The quality of a model predictions, how its performance can be compared to other models relies on the type and quality of the datasets used. Some classes of compounds are more difficult to predict than others. Ammines represent 14 % of the drugs present in the pharmaceutical drug market. Their mutagenic potential is associated with their activation through oxygenases, and followed detoxification. These steps can cause the formation of the nitrenium ion, a very reactive toxic species that interacts with DNA, and causes mutations to its sequence. Yet, the in-silico prediction of the nitrenium ionformations remains quite challenging. Therefore, invitro and mammalian cells assays are still being required, especially given the presence of this class of compounds in drugs (Patel et al., 2018). In silico models predicts aromatic amines as positive. Many QSAR models, perhaps to generate a conservative prediction, are incapable of assigning negative mutagenicity to this class of compound. Even when substructures in the mutagenicity inducing para substitution are considered, the level of accuracy remains low. The ratio of aromatic amines in the training datasets is key criteria to consider. Because their absence generates more accurate predictions, but it greatly reduces the model applicability with real world data. Nevertheless, their presence will reduce the level of accuracy and specificity. (Kuhnke et al., 2019)

Another essential dataset ratio to consider is the number of positives and negatives. Around 10 –15 % of compounds are mutagenic. However, when trained with datasets presenting that real-world setting ratio, the accuracy of the predictions greatly reduces. With a negative-to-positive ratio of 90:10, even a sensitivity of 40 % would still make a good model. The models are not trained on a number of positive compounds that is sufficiently high allow the models to learn about positive inducing alerting features or properties. In these regards, a neural network managed to achieve an accuracy of 77 % and a sensitivity of 55%, even when trained with data that resemble real world data. (Honma et al., 2020). Datasets presenting an imbalanced ratio between positive and negative are called skewed datasets. Skewed test datasets lead to an increased probability of certain predictions. Which can then affect the accuracy of the predictions of the model. When training set and external validation sets are similar, the model performs better. When models have not been trained on a sufficiently high number of a class of compound, accuracy can be increased by preventing predictions with low reliability, or by incorporating new data in the learning process. However, limiting the coverage of chemical space will greatly affect the usability of the model in the real world (Cavley et al., 2019). When selecting a training dataset, another issue to consider is the mutagenic strength of the test compounds. Newly designed databases that divide the mutagenic compounds into strongly mutagenic and not strongly mutagenic have been excluded from our training set. In fact, QSARs models trained on datasets that excluded the class made of “not strongly mutagenic” compounds (compounds showing less than 1000 revertants /mg of compounds) have more accurate and valid predictions due to the inconsistency surrounding this type mutagenic classification. Over the recent years, this has been an issue that led data validity and consistency concerns among the databases. (Norinder et al., 2018)

Any model can be as good as the data to build it is. Therefore, it is essential to control, and quality check the result of the Ames tests performed. Unfortunately, this is not always the case. Some compounds used during the Ames assay may artificially induce false positive results as well. By-products of the test are formed from the interaction of the compounds used for the test and the compound to test itself. This phenomenon has been noticed in acyl halides; compounds usually classed as positive. However, Amberg et al found that in many instances these compounds were not mutagenic. They were positive due to their interaction with a vehicle for test compounds known as DMSO, leading to a false positive result. Many of these compounds were then found negative when a different agent was used for the Ames assay (Amberg et al., 2016). In addition to that, chemists have also suggested that some compounds caused revertants in the Ames test due to the presence of impurities rather than the compounds themselves, shedding light on the matter of the quality of the data. Where a test that is used to assess the mutagenic potential of impurities may contain impurities themselves (Honma et al. ,2018). Even when the Ames test is performed correctly, there may still be issues with the classification and assignment of the mutagenic result.

Weak positive responses could be judged as negative. In fact, test compounds resulting as positive are supposed to have twice as much revertant colonies present as the control (2-fold rule). Therefore, a negative result is also fruit of the researcher judgement. According to ICH M7 conservative regulations, if only one positive result is present among the results from the various strains of Ames assay, the compound is conservatively assigned an overall positive result without consideration on the mutagenic mechanism. These results could be incorrect and may cause noise for the prediction of QSAR models (Honma et al., 2020). Even when the test is performed correctly and the result is not ambiguous, there are still unpredictability related to the pathway that compounds go through once inside mammals. Some compounds may be potential mutagenic, but still result as Ames negatives. This is the case when a compound requires specific enzymatic activation (Petkov et al., 2021). Ames assay imitate the metabolic activation via the usage of S9 liver extract. However, some enzymes present in humans that are not found in the S9 extract activate compounds to the point that they become mutagenic. These compounds will pass the health risk assessment as non-mutagenic impurities but may still cause mutations in humans (Cayley et al., 2019). Despite 80 % Ames positive being correlated with carcinogenicity in rodents, there is still no correlation between the number of revertants caused by compounds and their activity in terms of carcinogens. In fact, compounds causing a number of revertants of above 1000 revertants / mg have not displayed a carcinogenic activity in rodent as high. That is also the reason why mutagenicity is a positive or negative result in the ICH M7 legislations, and QSAR models are built using categorical values. (Honma et al., 2019) . Some QSARs models have achieved accuracies that are as high as the the reproducibility rate of the Ames test itself. Which indicates the need for future improvements in the amount and reliability of the Ames tests results.

### Potential of deep neural networks

Deep neural networks is a particular statistical QSAR. Its ability to extract features and relate them to activities is similar to what happened with toxicity fragments in rule-based model. These extracted features can be fetched to other machine learning techniques. This step of deep learning features extraction and statistical prediction via the usage of other machine learning methods has reached an accuracy of up to 90%, showing the possibility of combining this method with other machine learning techniques (Li et al., 2021). Deep neural networks have shown great accuracy even with troublesome databases containing hard-to-classify amines. A simple neural network made of one layer designed with selective focus on the parts of the input managed to achieve an accuracy of 80 % in the external validation set made of 130 amines. Even though, the group did not use a separate database for the external validations, which increases the risk of overfitting of the data (Chakravarti et al., 2019). Martinez et al designed a deep neural network model capble of predicting in which strains of bacteria the compound cuased the mutation.. Given the fact that different strains are susceptible to different mutations, that allows researchers to predict what kind of mutations are associated to certain compounds. However, this group as well has performed the external validation step on a database that highly resembles the training dataset and not real-world data (Martinez et al., 2022). This leads to some consideration about this technology for the QSAR predictions of mutagenicity, where high performance is usually linked to overfitting of the data. In summary, pharmaceutical companies have estimated that around 10 to 15% of the newly designed chemicals are mutagenic. This is an outstandingly high number, which indicates the need for improvements in the current testing and prediction of this endpoint. The advancement in technology combined with improvement in the data quality may allow computational tools to reach an accuracy high enough that in vitro and in vivo testing will not be necessary.

## Conclusion

This experiment helps to shed light on the potential of deep neural network models in the predictions of mutagenicity in impurities. A codeless deep neural model was designed using chemical descriptors and fingerprints combined. The model accurately predicted the mutagenicity of 86 % of the compound in the external validation dataset. Nevertheless, the advancement of computational methods is strictly related to the improvement of the quality and quantity of the mutagenicity results.

## REFERENCES

1. [*https://ich.org/page/multidisciplinary-guidelines#7-2*](https://ich.org/page/multidisciplinary-guidelines#7-2) *(Retrieved on 5th January 2023)*
2. *data.europa.eu. Data.europa.eu. Retrieved March 23, 2023,from* [*https://data.europa.eu/data/datasets/database-pesticide-genotoxicity-endpoints?locale=en*](https://data.europa.eu/data/datasets/database-pesticide-genotoxicity-endpoints?locale=en)
3. *ASSESSMENT AND CONTROL OF DNA REACTIVE (MUTAGENIC) IMPURITIES IN PHARMACEUTICALS TO LIMIT POTENTIAL CARCINOGENIC RISK M7(R1). (2017).* [*https://database.ich.org/sites/default/files/M7\_R1\_Guideline.pdf*](https://database.ich.org/sites/default/files/M7_R1_Guideline.pdf) *- Accessed on 26/01/2023*
4. *Amberg, A., Beilke, L., Bercu, J., Bower, D., Brigo, A., Cross, K. P., Custer, L., Dobo, K., Dowdy, E., Ford, K. A., Glowienke, S., Van Gompel, J., Harvey, J., Hasselgren, C., Honma, M., Jolly, R., Kemper, R., Kenyon, M., Kruhlak, N., & Leavitt, P. (2016). Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses. Regulatory Toxicology and Pharmacology, 77, 13–24.* [*https://doi.org/10.1016/j.yrtph.2016.02.004*](https://doi.org/10.1016/j.yrtph.2016.02.004)
5. *Ames, B. N. (1971). The Detection of Chemical Mutagens with Enteric Bacteria. Chemical Mutagens, 267–282.* [*https://doi.org/10.1007/978-1-4615-8966-2\_9*](https://doi.org/10.1007/978-1-4615-8966-2_9)
6. *Ames, B. N., McCann, J., & Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. Mutat. Res.; (Netherlands), 31.* [*https://www.osti.gov/etdeweb/biblio/6500670*](https://www.osti.gov/etdeweb/biblio/6500670 )
7. *Barber, C., Cayley, A., Hanser, T., Harding, A., Heghes, C., Vessey, J. D., Werner, S., Weiner, S. K., Wichard, J., Giddings, A., Glowienke, S., Parenty, A., Brigo, A., Spirkl, H.-P., Amberg, A., Kemper, R., & Greene, N. (2016). Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained. Regulatory Toxicology and Pharmacology, 76, 7–20.* [*https://doi.org/10.1016/j.yrtph.2015.12.006*](https://doi.org/10.1016/j.yrtph.2015.12.006 )
8. *Basant, N., & Gupta, S. (2017). QSAR modeling for predicting mutagenic toxicity of diverse chemicals for regulatory purposes. Environmental Science and Pollution Research, 24(16), 14430–14444.* [*https://doi.org/10.1007/s11356-017-8903-y*](https://doi.org/10.1007/s11356-017-8903-y)
9. *Benigni, R. (2019). Towards quantitative read across: Prediction of Ames mutagenicity in a large database. Regulatory Toxicology and Pharmacology, 108, 104434.* [*https://doi.org/10.1016/j.yrtph.2019.104434*](https://doi.org/10.1016/j.yrtph.2019.104434)
10. *Benigni, R., Bassan, A., & Pavan, M. (2020). In silico models for genotoxicity and drug regulation. Expert Opinion on Drug Metabolism & Toxicology, 16(8), 651–662.* [*https://doi.org/10.1080/17425255.2020.1785428*](https://doi.org/10.1080/17425255.2020.1785428)
11. *Benigni, R. (2021). In silico assessment of genotoxicity. Combinations of sensitive structural alerts minimize false negative predictions for all genotoxicity endpoints and can single out chemicals for which experimentation can be avoided. Regulatory Toxicology and Pharmacology, 126, p.105042. doi:10.1016/j.yrtph.2021.105042.*
12. *Cayley, A., Fowkes, A., & Williams, R. V. (2019). Important considerations for the validation of QSAR models for in vitro mutagenicity. Mutagenesis, 34(1), 25–32.* [*https://doi.org/10.1093/mutage/gey034*](https://doi.org/10.1093/mutage/gey034)
13. *Chakravarti, S. K., & Saiakhov, R. D. (2018). Computing similarity between structural environments of mutagenicity alerts. Mutagenesis, 34(1), 55–65.* [*https://doi.org/10.1093/mutage/gey032*](https://doi.org/10.1093/mutage/gey032)
14. *Chakravarti, S. K., & Alla, S. R. M. (2019). Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. Frontiers in Artificial Intelligence, 2.* [*https://doi.org/10.3389/frai.2019.00017*](https://doi.org/10.3389/frai.2019.00017)
15. *Chu, C. S. M., Simpson, J. D., O’Neill, P. M., & Berry, N. G. (2021). Machine learning – Predicting Ames mutagenicity of small molecules. Journal of Molecular Graphics and Modelling, 109, 108011.* [*https://doi.org/10.1016/j.jmgm.2021.108011*](https://doi.org/10.1016/j.jmgm.2021.108011)
16. *Coppi, A., Davies, R., Wegesser, T., Ishida, K., Karmel, J., Han, J., Aiello, F., Xie, Y., Corbett, M. T., Parsons, A. T., Monticello, T. M., & Minocherhomji, S. (2022). Characterization of false positive, contaminant-driven mutagenicity in impurities associated with the sotorasib drug substance. Regulatory Toxicology and Pharmacology, 131, 105162.* [*https://doi.org/10.1016/j.yrtph.2022.105162*](https://doi.org/10.1016/j.yrtph.2022.105162)
17. *Ding, Y.-L., Lyu, Y.-C., & Leong, M. K. (2017). In silico prediction of the mutagenicity of nitroaromatic compounds using a novel two-QSAR approach. Toxicology in Vitro, 40, 102–114.* [*https://doi.org/10.1016/j.tiv.2016.12.013*](https://doi.org/10.1016/j.tiv.2016.12.013)
18. *Dongsheng, L., Wei, C., Dongkuan, X., Wenchao, Y., Bo, Z., Haifeng, C., & Xiang, Z. (2020). Parameterized Explainer for Graph Neural Network. Advances in Neural Information Processing Systems, 33.* [*https://proceedings.neurips.cc/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html*](https://proceedings.neurips.cc/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html)
19. *Fan, D., Yang, H., Li, F., Sun, L., Di, P., Li, W., Tang, Y., & Liu, G. (2018). In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. Toxicology Research, 7(2), 211–220.* [*https://doi.org/10.1039/c7tx00259a*](https://doi.org/10.1039/c7tx00259a)
20. *Fournier, M., Vroland, C., Megy, S., Aguero, S., Chemelle, J., Defoort, B., Jacob, G., & Terreux, R. (2023). In silico genotoxicity prediction by similarity search and machine learning algorithm: optimization and validation of the method for High Energetic Materials. Propellants, Explosives, Pyrotechnics.* [*https://doi.org/10.1002/prep.202200259*](https://doi.org/10.1002/prep.202200259 )
21. *Fukuchi, J., Airi Kitazawa, Hirabayashi, K., & Honma, M. (2019). A practice of expert review by read-across using QSAR Toolbox. Mutagenesis.* [*https://doi.org/10.1093/mutage/gey046*](https://doi.org/10.1093/mutage/gey046)
22. *Ghasemi, F., Mehridehnavi, A., Fassihi, A., & Pérez-Sánchez, H. (2018). Deep neural network in QSAR studies using deep belief network. Applied Soft Computing, 62, 251–258.* [*https://doi.org/10.1016/j.asoc.2017.09.040*](https://doi.org/10.1016/j.asoc.2017.09.040 )
23. *Gini, G., Zanoli, F., Gamba, A., Raitano, G., & Benfenati, E. (2019). Could deep learning in neural networks improve the QSAR models? SAR and QSAR in Environmental Research, 30(9), 617–642. https://doi.org/10.1080/1062936x.2019.1650827*
24. *Guo, Y., Zhao, L., Zhang, X., & Zhu, H. (2019). Using a hybrid read-across method to evaluate chemical toxicity based on chemical structure and biological data. Ecotoxicology and environmental safety, 178, 178-187.*[*https://doi.org/10.1016/j.ecoenv.2019.04.019*](https://doi.org/10.1016/j.ecoenv.2019.04.019)*.(*[*https://www.sciencedirect.com/science/article/pii/S0147651319304324*](https://www.sciencedirect.com/science/article/pii/S0147651319304324)*)*
25. *Goh, G.B., Siegel, C.M., Vishnu, A., Hodas, N.O., & Baker, N. (2017). Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. ArXiv, abs/1706.06689*
26. *Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., & Müller, K. R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. Journal of chemical information and modeling, 49(9), 2077–2081. https://doi.org/10.1021/ci900161g*
27. *Hao, Y., Sun, G., Fan, T., Sun, X., Liu, Y., Zhang, N., Zhao, L., Zhong, R., & Peng, Y. (2019). Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods. Ecotoxicology and Environmental Safety, 186, 109822.* [*https://doi.org/10.1016/j.ecoenv.2019.109822*](https://doi.org/10.1016/j.ecoenv.2019.109822)
28. *Hasselgren, C., Bercu, J., Cayley, A., Cross, K., Glowienke, S., Kruhlak, N., Muster, W., Nicolette, J., Reddy, M. V., Saiakhov, R., & Dobo, K. (2020). Management of pharmaceutical ICH M7 (Q)SAR predictions – The impact of model updates. Regulatory Toxicology and Pharmacology, 118, 104807.* [*https://doi.org/10.1016/j.yrtph.2020.104807*](https://doi.org/10.1016/j.yrtph.2020.104807)
29. *Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. Nature Biotechnology, 32(1), 43+. https://link.gale.com/apps/doc/A356268116/AONE?u=salcal2&sid=bookmark-AONE&xid=ca3db96d*
30. *Helma, C., Schöning, V., Drewe, J., & Boss, P. (2021). A Comparison of Nine Machine Learning Mutagenicity Models and Their Application for Predicting Pyrrolizidine Alkaloids. Frontiers in Pharmacology, 12.* [*https://doi.org/10.3389/fphar.2021.708050*](https://doi.org/10.3389/fphar.2021.708050)
31. *Hemmerich, J., & Ecker, G. F. (2020). In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. WIREs Computational Molecular Science, 10(4).* [*https://doi.org/10.1002/wcms.1475*](https://doi.org/10.1002/wcms.1475)
32. *Honma, M. (2020). An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship. Genes and Environment, 42(1).* [*https://doi.org/10.1186/s41021-020-00163-1*](https://doi.org/10.1186/s41021-020-00163-1)
33. *Honma, M., Kitazawa, A., Cayley, A., Williams, R. V., Barber, C., Hanser, T., Saiakhov, R., Chakravarti, S., Myatt, G. J., Cross, K. P., Benfenati, E., Raitano, G., Mekenyan, O., Petkov, P., Bossa, C., Benigni, R., Battistelli, C. L., Giuliani, A., Tcheremenskaia, O., DeMeo, C., … Rathman, J. (2019). Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis, 34(1), 3–16.* [*https://doi.org/10.1093/mutage/gey031*](https://doi.org/10.1093/mutage/gey031)
34. *Honma, M., Kitazawa, A., Kasamatsu, T., & Sugiyama, K. (2020). Screening for Ames mutagenicity of food flavor chemicals by (quantitative) structure-activity relationship. Genes and Environment, 42, 32.* [*https://doi.org/10.1186/s41021-020-00171-1‌*](https://doi.org/10.1186/s41021-020-00171-1%E2%80%8C)
35. *Hsu, K.-H., Su, B.-H., Tu, Y.-S., Lin, O. A., & Tseng, Y. J. (2016). Mutagenicity in a Molecule: Identification of Core Structural Features of Mutagenicity Using a Scaffold Analysis. PLOS ONE, 11(2), e0148900.* [*https://doi.org/10.1371/journal.pone.0148900*](https://doi.org/10.1371/journal.pone.0148900)
36. *Hung, C., & Gini, G. (2021). QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Molecular Diversity.* [*https://doi.org/10.1007/s11030-021-10250-2*](https://doi.org/10.1007/s11030-021-10250-2)
37. *Jillella, G. K., Khan, K., & Roy, K. (2020). Application of QSARs in identification of mutagenicity mechanisms of nitro and amino aromatic compounds against Salmonella typhimurium species. Toxicology in Vitro, 65, 104768.* [*https://doi.org/10.1016/j.tiv.2020.104768*](https://doi.org/10.1016/j.tiv.2020.104768)
38. *Kumar, R., Khan, F.U., Sharma, A., Siddiqui, M.H., Aziz, I.B., Kamal, M.A., Ashraf, G.M., Alghamdi, B.S. and Uddin, Md.S. (2021). A deep neural network–based approach for prediction of mutagenicity of compounds. Environmental Science and Pollution Research, 28(34), pp.47641–47650. doi:10.1007/s11356-021-14028-9.*
39. *Kuhnke, L., ter Laak, A. and Göller, A.H. (2019). Mechanistic Reactivity Descriptors for the Prediction of Ames Mutagenicity of Primary Aromatic Amines. Journal of Chemical Information and Modeling, 59(2), pp.668–672. doi:10.1021/acs.jcim.8b00758.*
40. *Landry, C., Kim, M. T., Kruhlak, N. L., Cross, K. P., Saiakhov, R., Chakravarti, S., & Stavitskaya, L. (2019). Transitioning to composite bacterial mutagenicity models in ICH M7 (Q)SAR analyses. Regulatory Toxicology and Pharmacology, 109, 104488.* [*https://doi.org/10.1016/j.yrtph.2019.104488*](https://doi.org/10.1016/j.yrtph.2019.104488)
41. *LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444.* [*https://doi.org/10.1038/nature14539*](https://doi.org/10.1038/nature14539)
42. *Lenselink, E. B., ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., IJzerman, A. P., & van Westen, G. J. P. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of Cheminformatics, 9(1).* [*https://doi.org/10.1186/s13321-017-0232-0*](https://doi.org/10.1186/s13321-017-0232-0)
43. *Luechtefeld, T., Marsh, D., Rowlands, C., & Hartung, T. (2018). Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. Toxicological Sciences, 165(1), 198–212.* [*https://doi.org/10.1093/toxsci/kfy152‌*](https://doi.org/10.1093/toxsci/kfy152%E2%80%8C)
44. *Li, S., Zhang, L., Feng, H., Meng, J., Xie, D., Yi, L., Arkin, I.T. and Liu, H. (2021). MutagenPred-GCNNs: A Graph Convolutional Neural Network-Based Classification Model for Mutagenicity Prediction with Data-Driven Molecular Fingerprints. Interdisciplinary Sciences: Computational Life Sciences, 13(1), pp.25–33. doi:10.1007/s12539-020-00407-2.*
45. *Madia, F., Kirkland, D., Morita, T., White, P., Asturiol, D., & Corvi, R. (2020). EURL ECVAM Genotoxicity and Carcinogenicity Database of Substances Eliciting Negative Results in the Ames Test: Construction of the Database. Mutation Research. Genetic Toxicology and Environmental Mutagenesis, 854-855.* [*https://doi.org/10.1016/j.mrgentox.2020.503199*](https://doi.org/10.1016/j.mrgentox.2020.503199)
46. *Martínez, M. J., Sabando, M. V., Soto, A. J., Roca, C., Requena-Triguero, C., Campillo, N. E., Páez, J. A., & Ponzoni, I. (2022). Multi-Task Deep Neural Networks for Ames Mutagenicity Prediction. Chemrxiv.org.* [*https://doi.org/10.26434/chemrxiv-2022-852tf*](https://doi.org/10.26434/chemrxiv-2022-852tf)
47. *Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science, 3.* [*https://doi.org/10.3389/fenvs.2015.00080*](https://doi.org/10.3389/fenvs.2015.00080 )
48. *Miller A, Miller E C. Ultimate chemical carcinogen as reactive mutagenic electorophiles. In Origin of Human Cancer, Hiatt, H. H.; Watson, J. D.; Winsten, J. A., Eds. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1977; pp 605–627.*
49. *Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. Alzheimer’s & Dementia: Translational Research & Clinical Interventions, 3(4), 651–657. ncbi.* [*https://doi.org/10.1016/j.trci.2017.10.005*](https://doi.org/10.1016/j.trci.2017.10.005)
50. *Mortelmans, K., & Zeiger, E. (2000). The Ames Salmonella/microsome mutagenicity assay. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 455(1-2), 29–60.* [*https://doi.org/10.1016/s0027-5107(00)00064-6*](https://doi.org/10.1016/s0027-5107(00)00064-6)
51. *Nguyen, Kong T., Blum, Lorenz C., van Deursen, R., & Reymond, J.-L. (2009). Classification of Organic Molecules by Molecular Quantum Numbers. ChemMedChem, 4(11), 1803–1805.* [*https://doi.org/10.1002/cmdc.200900317*](https://doi.org/10.1002/cmdc.200900317 )
52. *Norinder, U., Ahlberg, E., & Carlsson, L. (2018). Predicting Ames Mutagenicity Using Conformal Prediction in the Ames/QSAR International Challenge Project. Mutagenesis, 34(1), 33–40.* [*https://doi.org/10.1093/mutage/gey038*](https://doi.org/10.1093/mutage/gey038 )
53. *Pandit, S., Dhawan, A. and Parthasarathi, R. (2018). Chapter Eight - Emerging Computational Methods for Predicting Chemically Induced Mutagenicity. [online] ScienceDirect. Available at:* [*https://www.sciencedirect.com/science/article/pii/B9780128092521000080*](https://www.sciencedirect.com/science/article/pii/B9780128092521000080) *[Accessed 31 Oct. 2022].*
54. *Pradeep, P., Judson, R., DeMarini, D. M., Keshava, N., Martin, T. M., Dean, J., Gibbons, C. F., Simha, A., Warren, S. H., Gwinn, M. R., & Patlewicz, G. (2021). An evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset. Computational Toxicology, 18, 100167.* [*https://doi.org/10.1016/j.comtox.2021.100167*](https://doi.org/10.1016/j.comtox.2021.100167)
55. *Patel, M., Kranz, M., Munoz-Muriedas, J., Harvey, J. S., Giddings, A., Swallow, S., Fellows, M., Naven, R., Werner, A.-L., Yeo, D. J., Bringezu, F., Wichard, J., Sutter, A., Glowienke, S., Whitehead, L., Selby, M., Reuberson, J., Atienzar, F., Gerets, H., & Kenyon, M. O. (2018). A pharma-wide approach to address the genotoxicity prediction of primary aromatic amines. Computational Toxicology, 7, 27–35.* [*https://doi.org/10.1016/j.comtox.2018.06.002*](https://doi.org/10.1016/j.comtox.2018.06.002)
56. *Petkov, P.I., Ivanova, H., Schultz, T.W. and Mekenyan, O.G. (2021). Criteria for assessing the reliability of toxicity predictions: I. TIMES Ames mutagenicity model. Computational Toxicology, 17, p.100143. doi:10.1016/j.comtox.2020.100143.*
57. *Sharma, A., Kumar, R., Shabnam Ranjta, & Pritish Kumar Varadwaj. (2021). SMILES to Smell: Decoding the Structure–Odor Relationship of Chemical Compounds Using the Deep Neural Network Approach. Journal of Chemical Information and Modeling, 61(2), 676–688.* [*https://doi.org/10.1021/acs.jcim.0c01288*](https://doi.org/10.1021/acs.jcim.0c01288)
58. *Smith, C. J., Perfetti, T. A., Ko, G. M., & Garg, R. (2018). Ames mutagenicity, structural alerts of carcinogenicity, Hansch QSAR parameters (ClogP, CMR, MgVol), tumor site concordance/multiplicity, and tumorigenicity rank in NTP 2-year rodent studies. Toxicology Research and Application, 2, 239784731875932.* [*https://doi.org/10.1177/2397847318759327*](https://doi.org/10.1177/2397847318759327)
59. *Siramshetty, V. B., Nickel, J., Omieczynski, C., Gohlke, B.-O., Drwal, M. N., & Preissner, R. (2015). WITHDRAWN—a resource for withdrawn and discontinued drugs. Nucleic Acids Research, 44(D1), D1080–D1086.* [*https://doi.org/10.1093/nar/gkv1192*](https://doi.org/10.1093/nar/gkv1192 )
60. *Sushko, I., Sergii Novotarskyi, Körner, R., Anil Kumar Pandey, Artem Cherkasov, Li, J., Gramatica, P., Hansen, K., Schroeter, T., Klaus-Robert Müller, Xi, L., Liu, H., Yao, X.-S., Öberg, T., Farhad Hormozdiari, Dao, P., S. Cenk Sahinalp, Todeschini, R., Polishchuk, P. G., & Artemenko, A. (2010). Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. Journal of Chemical Information and Modeling, 50(12), 2094–2111. https://doi.org/10.1021/ci100253r*
61. *‌ Tcheremenskaia, O., Battistelli, C. L., Giuliani, A., Benigni, R., & Bossa, C. (2019). In silico approaches for prediction of genotoxic and carcinogenic potential of cosmetic ingredients. Computational Toxicology, 11, 91–100.* [*https://doi.org/10.1016/j.comtox.2019.03.005*](https://doi.org/10.1016/j.comtox.2019.03.005)
62. *Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., Pairaudeau, G., Pennie, W. D., Pickett, S. D., Wang, J., Wallace, O., & Weir, A. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nature Reviews. Drug Discovery, 14(7), 475–486.* [*https://doi.org/10.1038/nrd4609*](https://doi.org/10.1038/nrd4609 )
63. *Wichard, J. D. (2017). In silico prediction of genotoxicity. Food and Chemical Toxicology, 106, 595–599. https://doi.org/10.1016/j.fct.2016.12.013*
64. *Winkler, D. A., & Le, T. C. (2016). Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. Molecular Informatics, 36(1-2), 1600118.* [*https://doi.org/10.1002/minf.201600118*](https://doi.org/10.1002/minf.201600118)
65. *Wu, Z., Jiang, D., Wang, J., Hsieh, C.-Y., Cao, D., & Hou, T. (2021). Mining Toxicity Information from Large Amounts of Toxicity Data. Journal of Medicinal Chemistry, 64(10), 6924–6936.* [*https://doi.org/10.1021/acs.jmedchem.1c00421*](https://doi.org/10.1021/acs.jmedchem.1c00421)
66. *Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., & Tang, Y. (2012). In silico Prediction of Chemical Ames Mutagenicity. Journal of Chemical Information and Modeling, 52(11), 2840–2847.* [*https://doi.org/10.1021/ci300400a*](https://doi.org/10.1021/ci300400a)
67. *Hongbin Yang, Jie Li, Zengrui Wu, Weihua Li, Guixia Liu, and Yun Tang,Chemical Research in Toxicology 2017 30 (6), 1355-1364,DOI: 10.1021/acs.chemrestox.7b00083,* [*https://pubs-acs-org.salford.idm.oclc.org/doi/full/10.1021/acs.chemrestox.7b00083*](https://pubs-acs-org.salford.idm.oclc.org/doi/full/10.1021/acs.chemrestox.7b00083)
68. *Yang, H., Sun, L., Li, W., Liu, G., & Tang, Y. (2018). In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Frontiers in chemistry, 6, 30,*[*https://doi.org/10.3389/fchem.2018.00030*](https://doi.org/10.3389/fchem.2018.00030)
69. *Yang, Y., Wu, Z., Yao, X., Kang, Y., Hou, T., Hsieh, C.-Y., & Liu, H. (2022). Exploring Low-Toxicity Chemical Space with Deep Learning for Molecular Generation. Journal of Chemical Information and Modeling, 62(13), 3191–3199.* [*https://doi.org/10.1021/acs.jcim.2c00671*](https://doi.org/10.1021/acs.jcim.2c00671)
70. *Zhang, H., Kang, Y.-L., Zhu, Y.-Y., Zhao, K.-X., Liang, J.-Y., Ding, L., Zhang, T.-G., & Zhang, J. (2017). Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity. Toxicology in Vitro, 41, 56–63.* [*https://doi.org/10.1016/j.tiv.2017.02.016*](https://doi.org/10.1016/j.tiv.2017.02.016)