Multi-Task Deep Neural Networks for Ames Mutagenicity Prediction

María Jimena Martínez,[*,†,@] María Virginia Sabando,[‡,¶,@] Axel J. Soto,[‡,¶] Carlos Roca,[§] Carlos Requena-Triguero,[§] Nuria E. Campillo,[§,‖,⊥] Juan A. Páez,[#] and Ignacio Ponzoni[‡,¶]

†*ISISTAN (CONICET - UNCPBA) Campus Universitario - Paraje Arroyo Seco, Tandil, Argentina.*

‡*Institute for Computer Science and Engineering, UNS-CONICET, Bahía Blanca, Argentina.*

¶*Department of Computer Science and Engineering, Universidad Nacional del Sur, Bahía Blanca, Argentina.*

§*CIB Margarita Salas (CSIC) Ramiro de Maeztu, 9. 28740, Madrid, Spain.*

‖*Instituto de Ciencias Matemáticas (CSIC), Nicolás Cabrera, no13-15, Campus de Cantoblanco, UAM, 28049- Madrid, Spain.*

⊥*CoFounder of AItenea Biotech.*

#*Instituto de Química Médica. Consejo Superior de Investigaciones Científicas (CSIC), Juan de la Cierva 3, 28006, Madrid, Spain.*

@*These authors contributed equally to this work.*

E-mail: mariajimena.martinez@isistan.unicen.edu.ar

Abstract

The Ames mutagenicity test constitutes the most frequently used assay to estimate the mutagenic potential of drug candidates. While this test employs experimental results using various strains of *Salmonella typhimurium*, the vast majority of the published *in silico* models for predicting mutagenicity do not take into account the test results of the individual experiments conducted for each strain. Instead, such QSAR models are generally trained employing overall labels (i.e. *mutagenic* and *non-mutagenic*). Recently, neural-based models combined with multi-task learning strategies have yielded interesting results in different domains, given their capabilities to model multi-target functions. In this scenario, we propose a novel neural-based QSAR model to predict mutagenicity that leverages experimental results from different strains involved in the Ames test by means of a multi-task learning approach. To the best of our knowledge, the modeling strategy hereby proposed has not been applied to model Ames mutagenicity previously. The results yielded by our model surpass those obtained by single-task modeling strategies, such as models that predict the overall Ames label or ensemble models built from individual strains. For reproducibility and accessibility purposes, all source code and datasets used in our experiments are publicly available.

Introduction

Genotoxicity is a destructive effect affecting the integrity of the genetic material of the cells and it is an essential requirement when analyzing the safety of drug candidates and industrial, chemical and environmental samples. The continuous discovery of new chemical compounds has led to strengthened regulatory measures to assure the safe use of new and existing substances. The first recommended approach to assess the genotoxic risk is the Ames test, which allows the assessment of the mutagenic potential of chemical compounds.[1,2] The test is used as a screening method to determine mutagenic potential of substances and also for regulatory purposes previous to registration and acceptance of these substances.

The Ames test is an *in vitro* model that consists in the detection of mutations in different

Salmonella typhimurium strains in the presence of the compound of interest. The principle of this test is to detect mutations that reverse the necessity of the bacteria to grow in presence of histidine and restore the functional capability of the bacteria to synthesize this essential amino acid, and thus be able to grow. The Organisation for Economic Co-operation and Development (OECD) Guidelines for the Testing of Chemicals[3] points out that at least five strains (*TA1535; TA1537* or *TA97a* or *TA97; TA98;* and *TA100*) have GC base pairs at the primary reversion site and it is known that they may not detect certain oxidizing mutagens, cross-linking agents and hydrazines. Such substances may be detected by *E. coli WP2* strains or *S. typhimurium TA102*, which have an AT base pair at the primary reversion site. Therefore, the use of different strains contributes to detect distinct types of mutagens.[1]

The development of *in silico* methods for predicting the result of the AMES toxicity test is an active field of research in computational toxicology[4–7] and several articles have reviewed the most relevant models and software tools for predicting mutagenicity under different datasets.[8–13] However, the impact that different strains used in the Ames test could have in the design of QSAR (Quantitative Structure Activity Relationship) methods has been scarcely studied. QSAR models found in the literature are trained using only overall values (i.e. toxic and non toxic classes) resulting from the Ames test, without considering the intermediate results achieved by the experiments individually conducted for each strain.

In addition, although the OECD guidelines define a minimum set of strains that must be present in the *in vivo* experiments of the Ames test, in practice it is common to find differences in the strain sets used in toxicity studies and in public datasets.[5,14,15] Furthermore, Williams et al.[16] showed evidence that supports the hypothesis that *S. typhimurium* strains *TA1535, TA1537, TA102,* and *E. coli* strain *WP2 uvrA* could be removed from the recommended set of strains in OECD TG471 with little, if any, loss of sensitivity for the detection of bacterial mutagens. This study further demonstrates that there is no absolute consensus about a unique way for conducting the Ames test. Therefore, the research question that

underpins this manuscript is whether it is possible to design *in silico* models for predicting mutagenicity by taking into account the unique contribution of each different strain and their complementarity. Furthermore, we wonder whether it is possible to attain more accurate and interpretable computational models for mutagenic toxicity tests by this means.

During the last decade, Deep Learning techniques have become the standard for a wide variety of tasks in the drug discovery pipeline. In particular, Deep Neural Networks (DNNs) are nowadays among the most used techniques for QSAR modeling.[17–19] Some of the reasons behind this phenomenon include the increasingly large volumes of molecular data available; the effectiveness of DNNs in learning complex, high-dimensional, nonlinear functions; and that recent advances on DNNs have made them less prone to overfit and more successful in complex predictive tasks. One particular technique, namely multi-task learning (MTL),[20] has shown potential in various domains and is being increasingly explored as a useful strategy to model multiple targets simultaneously. MTL could potentially yield QSAR models that predict an arbitrary combination of properties, consisting of regression or classification tasks. Although the idea behind MTL is not exclusive to Deep Learning, the neural-based approach for MTL allows to combine the information of the various properties being predicted during the learning process, thus potentially leveraging their complementarity while exploiting the information of each individual predictive task.

In this scenario, we propose the design of a QSAR model to predict mutagenicity based on the individual experiments with several of the most frequently used strains in the Ames test. This model is based on DNNs and follows an MTL strategy. This strategy predicts each individual strain as a separate target property while the information of all strains is jointly learned by the model. The outcomes of this model are afterwards combined by means of a consensus strategy to recreate the Ames test. Our hypothesis is that an MTL approach would prompt the model to leverage the complementarity of the different strains while exploiting their specific predictive traits. To the best of our knowledge, this is the first time that this approach has been applied to model Ames mutagenicity.

The primary research questions we aim to answer are whether it is possible to model the mutagenicity Ames test by means of a DNN-based model using an MTL approach, and whether such model entails any benefits with respect to traditional modeling strategies. The main challenge behind these questions is to effectively model the contribution of the different strains to an overall Ames result, while enriching their individual modeling process by a joint learning procedure.

As part of the experimental workflow designed in order to answer these questions, we propose the analysis of two different settings. First, we seek to analyze if a QSAR model using an MTL approach can outperform the performance of single-task QSAR models trained to predict the overall Ames test results. By this analysis, we aim to shed light onto the benefits of modeling Ames mutagenicity by using information of individual strains instead of aggregating them into a single overall value, which is the standard practice. Second, we examine the performance of the MTL approach against a QSAR model for Ames mutagenicity based on QSAR models developed for individual strains following a single-task learning strategy. This would give insight on whether the modeling process for each individual strain is influenced by the joint learning process, thus potentially impacting the Ames mutagenicity prediction. In the following sections we present a thorough review of the literature related to our research, as well as a detailed description of our experimental workflow and the results obtained.

Related work

The development of *in silico* models for predicting the Ames test has become an active research field during the last decades.[21] Ames mutagenicity QSAR models can be classified in two main groups: rule-based and statistics-based models.[10] The rule-based models qualitatively predict particular endpoints by matching identified molecular fragments of the unseen compounds to structural alerts, i.e., similar structures with well-known adverse effects

(e.g. mutagenicity). These rules can be obtained from scientific literature and human expert knowledge, known as human-based rules, or extracted from large collections of datasets, known as induction-based rules, or by a combination of both approaches.[7] Rule-based models usually lead to a binary output, because structural alerts are either present or absent in the compounds. Therefore, these rules only provide qualitative predictions of certain endpoints (e.g. *mutagenic* or *non-mutagenic*). In contrast, statistics-based QSAR models predict toxicity by analyzing statistical correlations with molecular descriptors. These methods use experimental data, such as bacterial mutagenicity, in order to train predictive models using machine learning methods.[22] Even though both methodologies have strengths and drawbacks, statistics-based models tend to yield better perfomances than rule-based models and, moreover, can compute predictions even when the mechanism of action is unknown.[7]

Several QSAR tools have been developed following these two approaches. Honma[10] presented a critical review about the most popular QSAR tools for predicting Ames mutagenicity. Most of these tools participated in the Ames/QSAR International Challenge Project, where 12 QSAR international vendors tested 17 QSAR tools in three phases conducted between 2014 and 2017. The final results of this competition were reported by Honma et al.[5]. In terms of performance, most tools achieved above 50% sensitivity (positive prediction among all Ames positives) and accuracy was as high as 80%, which is comparable to the interlaboratory reproducibility of Ames tests. In that competition, only one QSAR tool, known as *MUT_Risk*,[23] considered the individual contributions of different strains for predicting toxicity. *MUT_Risk* is explained by Honma[10] as an ADMET Risk™ score that uses ten models created from data on five individual strains of *S. typhimurium* or *E. coli* (strains *98, 100, 97+1537, 1535, 102+wp2*), with and without rat liver *S9* metabolic activation. For each positive classification produced by each of the five $\pm S9$ model pairs (i.e. *98* and *M98*), a point is added to a total score. A threshold value is set by the user to assess positive toxicity. In particular, they presented two threshold scenarios during the competition: *MUT_Risk-0* judges whether the chemical compounds are mutagenic when the score

is greater than 0, while *MUT_Risk-1* judges whether the compounds are mutagenic when the score is greater than 1. Flexible thresholds allow balance the tradeoff between sensitivity and specificity according to each application. Nevertheless, the performance of this approach in terms of balanced accuracy and Mathew's correlation coefficient was low in comparison with the other competing QSAR tools.[5]

Beyond the QSAR tools presented in the Ames/QSAR International Challenge Project, several efforts for improving the performance of QSAR models for toxicity emerged under the advent of Deep Learning.[24] The interest in these technologies has been increasing in the last few years since the availability of large and complex datasets for QSAR analysis. In particular, a variety of such models have been developed using an MTL approach, which we cover in this section.

Regarding toxicity prediction, Tang et al.[25] presented a brief review that included several MTL-based QSAR models trained to predict different types of toxicity. Mayr et al.[26] constructed an MTL-based toxicity prediction model using a large dataset from the 2014 Tox21 data challenge.[27] In this challenge, 12,707 chemicals were tested for 12 different toxicity effects, including stress response and nuclear receptor effects, where most of the compounds were labeled for several tasks. The authors compared the performance of single-task and multi-task DNN models on the Tox21 leaderboard set. Additionally, they also computed a linear SVM for every single task. Multi-task models achieved higher performance than single-task models and SVM models in 10 out of the 12 toxicity predictions. Nevertheless, for a strong imbalanced dataset that included only three positive compounds, both the single-task models and the multi-task models failed. This illustrated how a strong imbalanced data distribution may affect the performance of deep learning models.

Hughes et al.[28] proposed a multi-task learning model that can predict chemicals reactivity of molecules with glutathione (GSH), cyanides, deoxyribonucleic acid (DNA) and proteins. The identification of these kinds of chemical reactions plays a central role in the detection of common mechanisms underlying many types of drug toxicities. For building their

predictive model, they collected 1364 electrophilic molecules reactive with GSH, cyanides, DNA or proteins and 1439 nonreactive molecules from the Accelrys Metabolite Database (AMD).[29] Over 200 topological descriptors were employed to build the predictive model. The authors hypothesized that modeling several types of reactivity jointly in an MTL model would improve predictions on the smaller datasets. Indeed, the MTL models outperformed the individual modeling approaches at predicting cyanide and protein sites of reactivity. They concluded that the high performance attained by their MTL model for such tasks is likely due to the challenges these particular tasks entail, such as small and diverse datasets, therefore benefiting from an MTL approach.

Finally, Wu and Wei[30] evaluated multi-task models in a newly proposed set of molecular descriptors, namely the element specific topological descriptors (ESTDs). ESTDs are constructed via element specific persistent homology (ESPH) for quantitative toxicity analysis and prediction of small molecules. The authors experimented with multi-task DNNs, single-task DNNs, random forest, and gradient boosting decision trees, to construct topological learning strategies for predicting toxicity. Four benchmark toxicity datasets involving quantitative measurements were used to validate the proposed approaches. The MTL modeling strategy yielded the best performances, which the authors attributed to the inherent correlation among the different quantitative toxicity endpoints.

Although none of the published articles was aimed at predicting the Ames test using an MTL model, they allowed us to positively assess the capacity of MTL DNNs to model several toxicity tasks. In this scenario, we propose to explore an MTL strategy applied to a DNN to model the Ames mutagenicity test by means of integrating information of individual *S. typhimurium*.

Materials and methods

In this section we present a detailed description of our experimental setup and the data preprocessing stage. We describe the architecture of the model herein proposed, as well as its training and evaluation process. All data and source code used and developed for this paper can be found in the Supporting Information.

Data preprocessing

In order to conduct our experiments we used the ISSSTY v1-a dataset,[31] which contains publicly available data of *in vitro* mutagenicity in *Salmonella typhimurium* (Ames test) for 7367 compounds.[32] The dataset was collected and curated by the *Istituto Superiore di Sanita' (ISS)* and comprises information about the outcome of the Ames test in a wide variety of *S. typhimurium* strains with and without metabolic activation. In addition, it includes an *Overall* mutagenicity mark for each tested compound with regard to its outcome from all available strains. A compound is marked as *mutagenic* or *positive* when it exhibited positive results for at least one strain, regardless of the specific strain and regardless of whether or not it was under metabolic activation. A compound is marked as *non-mutagenic* or *negative* if two conditions are met: (i) no positive or equivocal results are obtained in any of the tested strains and (ii) the compound tested negative for at least one strain among *TA1535*, *TA100* and *TA97*, and at least one strain among *TA1538, TA98* and *TA1537*, with and without metabolic activation.[33] Compounds can also be marked as *equivocal*, if no strain yielded positive results and there is at least one equivocal result in any strains; or as *inconclusive*, if not enough experimental data is provided to support one of the previously described markings.

The initially retrieved 7367 compounds from ISSSTY dataset were screened and sanitized prior to their use for model generation. We first discarded those compounds having an *inconclusive Overall* marking in the dataset, since no information about their mutagenic-

ity was available for the subsequent Ames modeling task. From the remaining compounds, mixtures, polymeres and metal were excluded. The database was processed using LigPrep[34] software implemented on the Maestro Suite,[35] removing counterions, ionzing the ligands at pH 7.2 and stereoisomers (enantiomers R/S, diastereomers, cis/trans isomers) were considered. After this sanitization process, all SMILES strings were canonicalized using RDKit[36] and duplicates among the canonical SMILES were also removed. As a result of this sanitization process, we obtained 6445 compounds. Subsequently, we computed 0, 1, and 2D molecular descriptors for the compounds using Mordred[37] Those descriptors with more than 60% of missing values were eliminated, which yielded 1360 descriptors per compound. Then, all the remaining missing values were replaced by the mean value of the descriptor and constant values were removed.

After the sanitization process, we proceeded to analyze the labels in the ISSSTY dataset in order to prepare it for the modeling stage. We first aggregated all labels corresponding to variations of a same strain (i.e., different metabolic activations) into a single label for that strain. We computed labels for strains ***TA98, TA100, TA102, TA1535*** and ***TA1537***, following the standard OECD-five: *TA98, TA100, TA1535, TA1537* (or *TA97*) and *E. coli* (or *TA102*).[3] As a result, we obtained a dataset comprising 6445 compounds and five labels per compound, each corresponding to a different strain. Following the labeling criteria established at the creation of the ISSSTY dataset,[33] these labels could be either *positive, negative, equivocal* or *inconclusive*. It is worth noting that, at this point of the data preprocessing stage, those compounds with an *inconclusive* label in a certain strain would necessarily have a different label in at least one other strain.

Afterwards, we changed all *inconclusive* and *equivocal* labels throughout all strains to a new label: *undefined*. The reason behind this step is that, while *inconclusive* and *equivocal* labels do not provide meaningful information for predicting Ames mutagenicity, a compound might have different labels for each strain. Therefore, those compounds having an *undefined* label in one or more strains, but having either *positive* or *negative* labels in the remaining

strains, would still be considered in the modeling process. After this step, a few compounds having *undefined* labels for all strains in the dataset were removed, resulting in a final dataset of 5536 compounds for model generation.

Finally, we computed an *Overall* label which aggregates the labels of the five strains. The criteria used to compute this aggregation, henceforth dubbed *ground-truth labeling criteria*, is the following:

- a compound is labeled *positive* if any of the strains marks it as positive (mutagenic);
- *negative*, if *all* of the strains mark it as *negative* (non-mutagenic), and
- *undefined*, if none of the strains exhibit a *positive* label and at least one of them has an *undefined* label.

Table 1 summarizes the strain and labels distribution for our dataset.

Table 1: Summary of the contents of the dataset used in our experiments.

Strain	# Compounds (Positive / Negative)	Strain variants considered (TA)
TA98	4.854 (1.676 / 3.178)	98, 98_S9, 98(NR), 98(NR)_S9, 98(1,8-DNP6), 98(1,8-DNP6)_S9
TA100	5.366 (2.096 / 3.270)	100, 100_S9, 100(NR), 100(NR)_S9, 100(1,8-DNP6), 100(1,8-DNP6)_S9
TA102	975 (226 / 749)	102, 102_S9
TA1535	2.657 (436 / 2.221)	1535, 1535_S9
TA1537	2.229 (365 / 1.864)	1537, 1537_S9
Overall	3.334 (3103 / 231)	All strains

Model architecture

In order to answer our research questions, we designed a QSAR model that predicts Ames mutagenicity based on experimental data of individual *S. typhimurium* strains. As previously stated, our QSAR model is based on Deep Neural Networks (DNNs) and was developed following a multi-task learning (MTL) strategy. In our scenario, the purpose of the MTL model is to learn to predict the mutagenicity test results of each individual strain based on the labels computed for each strain in the dataset previously described, while at the same time it aims to jointly learning the traits that make such strains complementary or correlated

by means of a shared model architecture. To accomplish that, and as shown in Figure 1 (a), an MTL model learns from multiple targets at the same time.

In order to attain such a model, we treated the mutagenicity test results of each strain as a separate predictive target. We employed an MTL DNN architecture that roughly consists of two connected DNN cores: (i) a *shared core* consisting of one feed-forward DNN whose weights and activation functions are shared for all targets, and (ii) a *target-specific core* consisting of five individual sets of fully connected neural layers, one set for each target, where the targets are the strains involved in the modeling process: *TA98, TA100, TA102, TA1535* and *TA1537*. The outcome of the shared core feeds the target-specific core, as it can be seen in Figure 1 (a). The weights of the shared core are learned by iteratively optimizing all five predictive tasks at once, so the first layers of the architecture are trained by combining information of all five strains. The target-specific core comprises the output of the architecture, so our MTL model has five outputs, one per each target or strain.

The target values to be predicted by the MTL model correspond to the labels of each of the five strains being considered, which can be *positive (1), negative (0)* or *undefined (-1)*. The *undefined* labels were masked during the training process; as a result, a compound having an *undefined* label in a certain strain would have no impact in the computation of the loss function during the training process of that target. The outputs of the MTL model are finally aggregated through a consensus strategy that computes the Ames test prediction. This resulting model, depicted in Figure 1(a), is hereafter called $\boldsymbol{MTL\text{-}DNN_{Cons}}$. The *Overall* labels were used to evaluate the performance of $MTL\text{-}DNN_{Cons}$, which were computed by means of the ground-truth labeling criteria previously described. Finally, as a means to establish a fair comparison with traditional approaches to model Ames mutagenicity, we also developed a single-task DNN-based model, as shown in Figure 1(b), that is trained to predict such *Overall* labels, namely $\boldsymbol{STL\text{-}DNN_{Overall}}$, and a consensus model based in single-task DNN-based models for the individual strains, namely $\boldsymbol{STL\text{-}DNN_{Cons}}$ depicted in Figure 1(c). All our models were built and developed using Keras and Tensorflow,[38]

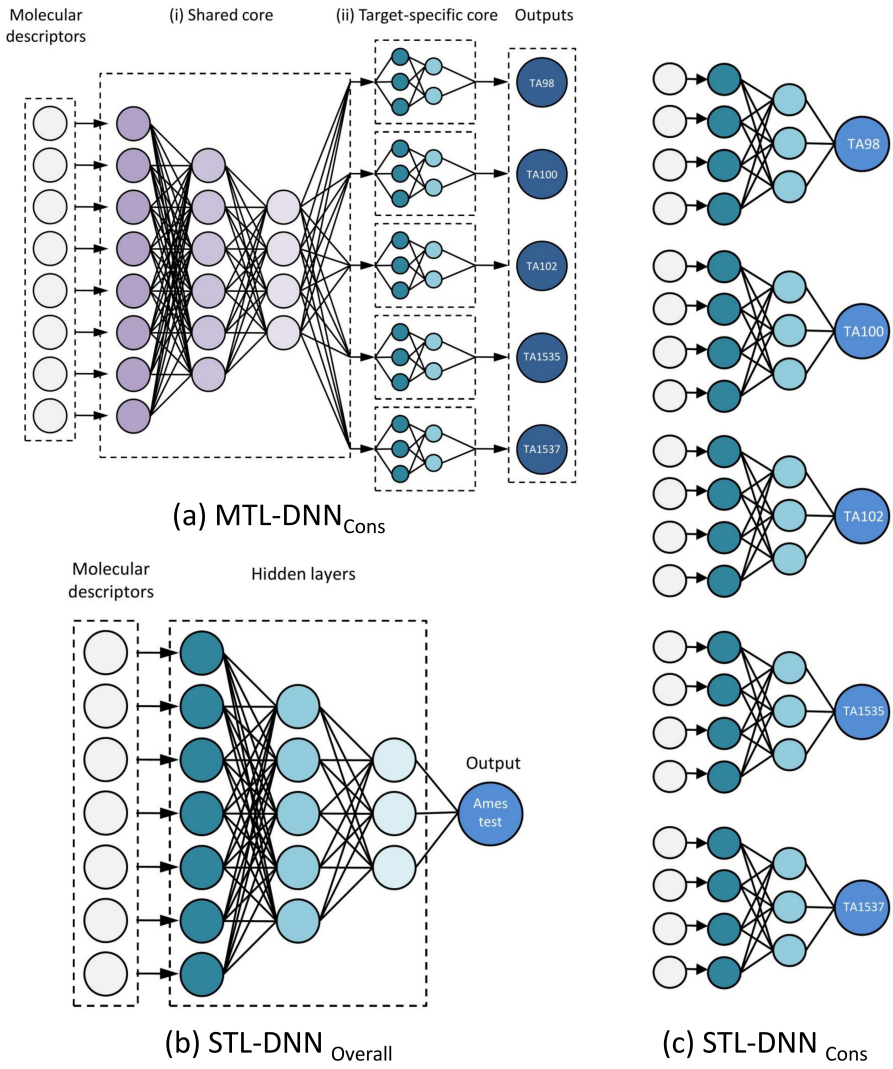

Figure 1: Overview of the three Ames models that we compare in this work. (a) Our newly proposed MTL architecture. Each target corresponds to a strain (*TA98, TA100, TA102, TA1535* and *TA1537*). The architecture consists of two connected DNN cores: (i) a *shared core* for all strains and (ii) a *target-specific core* with individual sets of fully-connected layers for each strain. (b) Single-task architecture for overall Ames mutagenicity test prediction. (c) Single-task architectures to model each strain experiment.

whereas the performance analysis was conducted using Scikit-Learn.[39]

Experimental design

Our experimental workflow consisted of several steps to ensure reproducibility and fairness during model generation and evaluation. An overview of our experimental design can be

found in Figure 2. The first step in our experimental workflow consisted of a hyperparameter *grid search*, i.e., a preliminary search in which we tried different hyperparameters in order to select those that yielded the best performance. This process was carried out both for our newly proposed *MTL-DNN*$_{Cons}$ model as well as for the *STL-DNN*$_{Overall}$ and *STL-DNN*$_{Cons}$ models. Since the grid search stage is supposed to help find the best hyperparameter combination and, therefore, can potentially entail a large number of experiments, we conducted such a search by splitting the dataset in fixed stratified partitions: 70% for training, 10% for internal validation and 20% for external validation, which was preserved until the final evaluation stages. We employed different random initialization seeds for each run in the grid search. The grid search stage corresponds to Figure 2(a).

Among the hyperparameters tested during the grid search, we randomly varied the $L2$ regularization coefficient $\lambda \in \{0.001, 0.005, 0.01\}$, the architecture of the *shared core* in terms of number of units per layer $n \in \{(100, 50, 10, 5), (100, 50, 20, 10), (200, 100, 50, 10), (200, 100, 20, 10), (200, 100, 10, 5)\}$ and the architecture of the *target-specific core* in terms of its number of layers $s \in \{0, 1, 2\}$, where the number of units per layer matched the number of units in the last s layers of the *shared core*. We also tested the impact of using weighed cost functions during training, such that the model loss was adjusted with a weight computed based on the class imbalance. Since the class imbalance scenarios might be different for each strain, different weights were applied to each set of fully-connected layers in the *target-specific core*. This technique, however, did not yield better results than applying no weighed cost function in any of the models. Other hyperparameters, such as the activation functions, the batch size and the learning rate, were varied during preliminary experiments but they showed negligible effects on the learning process, and thus were fixed further on the grid search.

As a result of the grid search stage, we selected the hyperparameter combination that yielded the top results. This resulted in one *MTL-DNN* model and one *STL-DNN* model. Full details on the parameterization of these models can be found in the Supporting Informa-

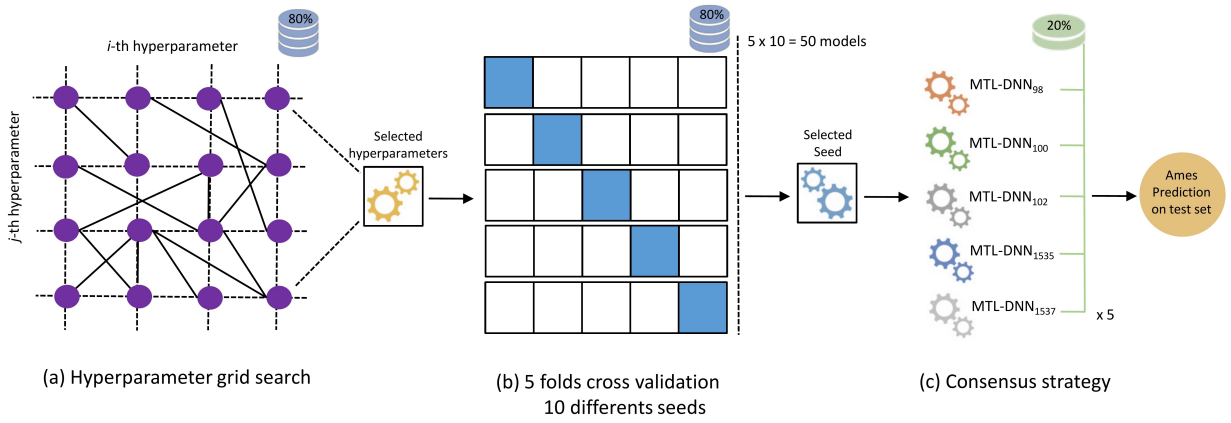

Figure 2: Overview of the experimental design. Stratified fixed partitions were used and 20% of the data was held out for external validation (a) Random grid search was used to evaluate the different combinations of hyperparameters. As a result, a single combination of hyperparameters values was selected. (b) A repeated 5-fold cross validation was performed with 10 random seeds, where the best performing seed is chosen. This yields five models corresponding to each iteration of the cross-validation process (c) Each of the five MTL models were evaluated with the external validation data partition. The predictions for each strain (*MTL-DNN$_{98}$, MTL-DNN$_{100}$, MTL-DNN$_{102}$, MTL-DNN$_{1535}$, and MTL-DNN$_{1537}$*) were obtained and aggregated through a consensus strategy to obtain the final predictive value of the Ames Mutagenicity test.

tion. From these parameterizations, we conducted a *five-fold cross validation* training stage, by merging the train and the internal validation splits to compute the stratified folds. We used the same five folds for each model and we conducted ten replications of each experiment using different random initialization seeds, in order to ensure that the performance observed was not bound to inherent variance in the data partitions or to the initialization of the model weights. This stage is illustrated in Figure 2(b). The five-fold cross validation stage yielded $5 \times 10 = 50$ trained models for each architecture. We then computed the average performance on the internal validation of each model along with their 95% confidence intervals for each of the ten different random initialization seeds. We selected the five instances corresponding to the best performing random initialization for each model. Finally, we computed the outcomes of the external validation split of the dataset on these instances.

While the *STL-DNN* model yielded one outcome per compound, the *MTL-DNN* model yielded five different outcomes, one per strain. For this reason, the final stage of our ex-

perimental workflow is the *consensus evaluation* in the *MTL-DNN$_{Cons}$* model. Based on the five outcomes of the MTL model, as shown in Figure 2(a), we computed the predicted consensus value for each compound in the external validation set, following the same criteria used to compute the *Overall* labels. The final evaluation was performed by comparing such predictions, as well as the *MTL-DNN$_{Cons}$* model predictions, with the *Overall* labels.

In addition to this experimental workflow, we built and trained a set of five single-task DNN models, one per each strain. Such models were afterwards combined by means of a consensus strategy, yielding the *STL-DNN$_{Cons}$* model. For the development of these models, we followed all steps described previously, including the hyperparameter grid search and the five-fold cross validation stage, and trained them using the individual strain labels as ground truth. The purpose of these models was to provide a means to evaluate the impact of combining the information of individual strains by means of a multi-task learning approach compared to multiple single-task learning models for the same strains.

Results and discussion

In this section, we present a discussion of the results obtained from our experimental design. We analyze different comparison scenarios of the models shown in Figure 1, to answer the research questions proposed in this work.

The performance of our models was evaluated by seven metrics: *Sensitivity (Sn), Specificity (Sp), Precision, Accuracy (Acc), Balanced Accuracy (BAcc), F1 score and H1 score*. *Sn* and *Sp* measure the ability of the model to detect *mutagenic* and *non-mutagenic compounds*, respectively. *Precision* indicates the proportion of mutagenic compounds that were correctly predicted. *Acc* measures the percentage of accurate predictions. *BAcc* is the arithmetic mean of *Sn* and *Sp*. *F1 score* is the harmonic mean of *Sn* and *Precision*, and *H1 score* is the harmonic mean of *Sn* and *Sp*. We focus on *F1 score* and *H1 score* for model selection.

As explained in Figure 2(b), five models were selected from the five-fold cross validation

stage corresponding to the best performing random initialization. The performances reported in this section correspond to the average results of those five models on the external validation set and their 95% confidence intervals. For evaluation, we did not take into account those compounds having an *undefined* label in either the MTL consensus model or the *Overall* model for the evaluation. Tables with the results for each fold can be found in the Supporting Information.

We first focused on evaluating whether an MTL QSAR model could better capture the contributions of the different strains than a QSAR model trained to predict the overall Ames test results. In addition, a question that naturally arises is whether an MTL model could more efficiently harness shared information between strains to model the Ames test compared to modeling each strain individually with a single-task architecture. For this analysis, we compare the performances of our newly proposed *MTL-DNN$_{Cons}$* model (Figure 1(a)) with the performances of the *STL-DNN$_{Overall}$* and *STL-DNN$_{Cons}$* models (Figures 1(b) and (c), respectively).

Table 2: Average results on the external validation set for *MTL-DNN$_{Cons}$*, *STL-DNN$_{Overall}$* and *STL-DNN$_{Cons}$*, along with their corresponding confidence intervals at 95%. These results were computed by evaluating the external validation set on the five trained trials resulting from the five-fold cross validation stage of our experimental workflow. As it can be seen from the best results highlighted in **bold**, our proposed model significantly surpass single-task learning strategies.

	Sp	Sn	Precision	Acc	BAcc	F1 score	H1 score
MTL-DNN$_{\mathbf{Cons}}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.92 \pm 0.02}$
STL-DNN$_{\mathrm{Overall}}$	0.43 ± 0.06	0.99 ± 0.00	0.96 ± 0.00	0.95 ± 0.00	0.71 ± 0.03	0.98 ± 0.00	0.60 ± 0.06
STL-DNN$_{\mathrm{Cons}}$	0.72 ± 0.04	0.99 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.86 ± 0.02	0.99 ± 0.00	0.84 ± 0.02

The performances reported in Table 2 show that the *MTL-DNN$_{Cons}$* model surpassed the results yielded by the *STL-DNN$_{Overall}$* model and the *STL-DNN$_{Cons}$* model. In general, it can be seen that the three models present high *Sn* and *Precision* values, thus correctly detecting a high proportion of the mutagenic compounds. However, a significant difference is observed in the *Sp* metric, i.e., the ability of the models to detect non-mutagenic compounds. In this sense, the consensus model *MTL-DNN$_{Cons}$* yielded the highest *Sp* value of 0.86. In

contrast, the *STL-DNN*$_{Overall}$ model obtained an average *Sp* value of 0.46. Considering that the mutagenic compounds constitute the majority class, this phenomenon might indicate that the strong class imbalance scenario presented by the dataset constitutes an additional challenge for the learning process of the *STL-DNN*$_{Overall}$ model, which is based on a single overall value.

STL-DNN$_{Cons}$ also exhibits lower performance than *MTL-DNN*$_{Cons}$ in terms of *Sp*, with an average value of 0.72. In this sense, it can be seen that the consensus from individual strain models is leaving out relevant information for detecting non-mutagenic compounds that can be effectively captured in an MTL learning approach. It is worth noticing that the *STL-DNN*$_{Cons}$ model outperformed the *STL-DNN*$_{Overall}$ model, suggesting that taking into consideration the information provided by the individual strains might have a positive impact in the performance in contrast to modeling an overall value. Since *BAcc* and *H1 score* evaluate the ability of the model to properly perform in imbalanced scenarios, low *Sp* values have an impact on those metrics as well.

We also analyzed the performances of the individual models per strain shown in Figure 1(c) in order to determine whether an MTL approach would potentially entail any benefits with respect to modeling Ames mutagenicity by means of individual models for each strain. In this sense, we conducted a comparison of the results obtained by the *MTL-DNN* and *STL-DDN* architectures on each strain. As it can be seen in Table 3, there are no significant differences in the performance yielded by the outputs of the *MTL-DNN* architecture and the individual *STL-DNN* models.

It is interesting to note that despite there are no major differences in the prediction performances for the methods and strains reported in Table 3, the MTL architecture (*MTL-DNN*$_{Cons}$) exhibits a much higher overall prediction performance than the STL counterpart (Table 2). This is specially the case for specificity (*Sp*), which indicates a high false positive ratio when *STL-DNN*$_{Cons}$ is used. From the *ground-truth labeling criteria* used in the consensus, an underperforming model associated to a strain may lead to false positives in the

Table 3: Prediction metrics obtained by the *MTL-DNN* and *STL-DDN* architectures on each of the five strains involved in the modeling process. In most cases, there are no significant differences between the outcomes of the *MTL-DNN* model and the individual *STL-DNN* models.

		Sp	Sn	Precision	Acc	BAcc	F1 score	H1 score
TA98	$STL\text{-}DNN_{98}$	0.85 ± 0.01	0.82 ± 0.01	0.78 ± 0.01	0.84 ± 0.01	0.83 ± 0.01	0.80 ± 0.01	0.83 ± 0.01
	$MTL\text{-}DNN_{98}$	0.85 ± 0.01	0.81 ± 0.01	0.78 ± 0.01	0.84 ± 0.00	0.83 ± 0.00	0.80 ± 0.00	0.83 ± 0.00
TA100	$STL\text{-}DNN_{100}$	0.83 ± 0.02	0.76 ± 0.04	0.78 ± 0.01	0.80 ± 0.01	0.79 ± 0.01	0.77 ± 0.01	0.79 ± 0.01
	$MTL\text{-}DNN_{100}$	0.81 ± 0.02	0.77 ± 0.01	0.77 ± 0.01	0.79 ± 0.00	0.79 ± 0.00	0.77 ± 0.00	0.79 ± 0.00
TA102	$STL\text{-}DNN_{102}$	0.82 ± 0.03	0.53 ± 0.04	0.52 ± 0.02	0.75 ± 0.01	0.68 ± 0.01	0.52 ± 0.02	0.64 ± 0.03
	$MTL\text{-}DNN_{102}$	0.79 ± 0.03	0.55 ± 0.04	0.49 ± 0.04	0.73 ± 0.02	0.67 ± 0.02	0.52 ± 0.03	0.65 ± 0.03
TA1535	$STL\text{-}DNN_{1535}$	0.94 ± 0.02	0.57 ± 0.02	0.66 ± 0.06	0.88 ± 0.01	0.76 ± 0.01	0.61 ± 0.03	0.71 ± 0.01
	$MTL\text{-}DNN_{1535}$	0.93 ± 0.01	0.59 ± 0.02	0.61 ± 0.04	0.87 ± 0.01	0.76 ± 0.02	0.60 ± 0.03	0.72 ± 0.02
TA1537	$STL\text{-}DNN_{1537}$	0.94 ± 0.01	0.79 ± 0.01	0.71 ± 0.03	0.91 ± 0.01	0.86 ± 0.01	0.75 ± 0.01	0.85 ± 0.01
	$MTL\text{-}DNN_{1537}$	0.93 ± 0.01	0.78 ± 0.02	0.68 ± 0.02	0.90 ± 0.01	0.85 ± 0.01	0.73 ± 0.01	0.85 ± 0.01

overall predicted label for the Ames test. Strains that present class imbalance or lack of data are likely to negatively affect the predictive performance. For instance, this is the case of strain *TA102*, which has the lowest number of labeled compounds of the dataset (Table 1). However, an MTL approach is more robust to scenarios with little data or class imbalance, as the shared core is trained jointly from data coming from multiple strains.

Additionally, we conducted an analysis of the results reported by the participants in the Ames/QSAR International Challenge Project.[5,10] We compiled such results in a supplementary table named Table S1 (see Supporting Information section), that shows the performance metrics corresponding to the final phase of the competition (Phase III) for each of the QSAR tools.[5] The *F1* and *H1* scores were calculated from the information available in that publication, since they were not provided in the original paper. Since the datasets and the setup of the experimental evaluation applied in this work are not the same as those used in the competition, performances shown in the Table S1 are not directly comparable with the values reported in Table 2. In general, the competing models correctly classify non-mutagenic compounds in a proportion similar to that obtained by the $MTL\text{-}DNN_{Cons}$ model (*Sp*). However, a significant decline is observed in the abilities to detect mutagenic compounds (*Sn*) for most of the models. Such low performance in terms of *Sn* is also related to the differences observed in the remaining metrics. It is noteworthy that many of the

competing models display a bias towards one of the two classes (positive/negative), whereas our proposed model *MTL-DNN$_{Cons}$* attains good results for both classes even in a highly imbalanced scenario. For this reason, and despite the differences in terms of experimental design and data partitions between the competing models in Table S1 and ours, we argue that the multi-task learning approach adopted in the development of our proposal favors a balanced and accurate prediction of both mutagenic and not mutagenic compounds.

To the best of our knowledge, this is the first work that presents an MTL approach to model Ames mutagenicity. The results obtained show that the modeling of Ames mutagenicity applying a consensus strategy from an MTL model surpasses the modeling strategies from overall labels commonly found in the literature. In addition, our approach also outperforms the consensus strategy based on individual single-task models by strains. These results further confirm the feasibility of applying MTL approaches for the test Ames modeling.

Conclusions

The Ames test is one of the most popularly used methods to detect mutagenicity. Currently, the development of *in silico* models for the prediction of mutagenicity is an active research field. The QSAR models found in the literature generally use overall labels (*mutagenic* and *non-mutagenic*) without considering intermediate results obtained individually for each strain. These models usually exhibit imbalanced performance issues in terms of sensitivity (*Sn*), i.e., the true positive rate, and specificity (*Sn*), i.e., the true negative rate. Having QSAR models that predict mutagenicity with both high sensitivity and specificity is of utmost importance for drug discovery, and food and environmental regulations, considering the time, costs and risks involved in such processes.

We proposed a novel model for the Ames test using a deep learning MTL approach using experimental information from five strains: *TA98, TA100, TA102, TA1535,* and *TA1537*. This approach allows each strain to be predicted separately while the information shared by

all strains is learned jointly by the model. Consequently, the Ames test prediction is obtained by aggregating the model outputs corresponding to each strain. To the best of our knowledge, this MTL approach has not been previously applied to Ames test modeling. We contrasted the results obtained with a single-task model that predicts mutagenicity using overall labels and with individual single-task models for each strain and their ensemble model.

The results obtained by our MTL model surpass those obtained by the single-task models, i.e., those that predict the *Overall* label of the Ames test, and those that model mutagenicity by an ensemble of individual strain models. Our MTL approach presents balanced values of *Sn* and *Sp*, which means that it is able to accurately detect both mutagenic and non-mutagenic compounds. These results support our hypothesis that multi-task learning is beneficial for QSAR modeling given that it allows learning in an environment with little information and class imbalance without negatively affecting the prediction performance. Finally, it is worth emphasizing once again that multi-task modeling, compared to approaches that only infer the overall value, entails the additional benefit of providing a prediction for each strain, which favors the interpretability of mutagenicity prediction of a compound. All data and scripts are made publicly available in order to enable reuse and reproducibility of our experiments.

Supporting Information Available

As a means to ensure the reproducibility of our experimental workflow, the source code and a link to the dataset can be found in our GitHub repository: `https://github.com/VirginiaSabando/MTL_DNN_Ames`. We also provide all the intermediate results of our experimental workflow, including the grid selection stage and the five-fold cross validation stage results, in the supplementary file **Model_selection.xlsx**. Finally, we provide an appendix (**Supporting_information.pdf**), which contains parameterization details of the models and the results of the final phase of the competition (Table S1) for each of the QSAR tools

that participated in the Ames/QSAR International Challenge Project.[5,10]

Author contributions

M.J.M. and M.V.S. designed and developed the multi-task learning architecture, and carried out all the computational experiments. C.R. and C.R.-T. designed and conducted the data sanitization stage. M.J.M., M.V.S., A.J.S. and I.P. were responsible for the conceptualization and the idea behind modeling each strain separately. N.E.C. and J.A.P. established the rationale and theoretical background on Ames toxicity for the development of a QSAR method based on individual strains. M.J.M., M.V.S. and I.P. wrote most of the article. All authors provided feedback, participated substantively in revision, and approved the final version of the article.

Acknowledgement

This work was partially supported by the Argentinean National Council of Scientific and Technological Research (CONICET for its acronym in Spanish) [grant PIP 112-2017-0100829], by the National Agency for the Promotion of Research, Technological Development and Innovation of Argentina (AGENCIA I+D+i in Spanish), through the Fund for Scientific and Technological Research (FONCyT for its acronym in Spanish) [grant PICT-2019-03350], by the Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina [grant PGI 24/N042], and *Ministerio de Economía, Industria y Competitividad, Gobierno de España under Grant RTI2018-096100B-100*.

References

(1) Ames, B. N.; Lee, F. D.; Durston, W. E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National*

Academy of Sciences **1973**, *70*, 782–786.

(2) Maron, D. M.; Ames, B. N. Revised methods for the Salmonella mutagenicity test. *Mutation Research/Environmental Mutagenesis and Related Subjects* **1983**, *113*, 173–215.

(3) OECD, T. N. 471: bacterial reverse mutation test. *OECD Guidelines for the Testing of Chemicals, Section* **1997**, *4*.

(4) Gini, G.; Zanoli, F.; Gamba, A.; Raitano, G.; Benfenati, E. Could deep learning in neural networks improve the QSAR models? *SAR and QSAR in Environmental Research* **2019**, *30*, 617–642.

(5) Honma, M.; Kitazawa, A.; Cayley, A.; Williams, R. V.; Barber, C.; Hanser, T.; Saiakhov, R.; Chakravarti, S.; Myatt, G. J.; Cross, K. P., et al. Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis* **2019**, *34*, 3–16.

(6) Benigni, R.; Bassan, A.; Pavan, M. In silico models for genotoxicity and drug regulation. *Expert Opinion on Drug Metabolism & Toxicology* **2020**, *16*, 651–662.

(7) Herrmann, K.; Holzwarth, A.; Rime, S.; Fischer, B. C.; Kneuer, C. (Q) SAR tools for the prediction of mutagenic properties: Are they ready for application in pesticide regulation? *Pest Management Science* **2020**, *76*, 3316–3325.

(8) Cassano, A.; Raitano, G.; Mombelli, E.; Fernández, A.; Cester, J.; Roncaglioni, A.; Benfenati, E. Evaluation of QSAR models for the prediction of ames genotoxicity: a retrospective exercise on the chemical substances registered under the EU REACH regulation. *Journal of Environmental Science and Health, Part C* **2014**, *32*, 273–298.

(9) Benfenati, E.; Golbamaki, A.; Raitano, G.; Roncaglioni, A.; Manganelli, S.; Lemke, F.; Norinder, U.; Lo Piparo, E.; Honma, M.; Manganaro, A., et al. A large comparison of integrated SAR/QSAR models of the Ames test for mutagenicity. *SAR and QSAR in Environmental Research* **2018**, *29*, 591–611.

(10) Honma, M. An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship. *Genes and Environment* **2020**, *42*, 1–13.

(11) Carnesecchi, E.; Raitano, G.; Gamba, A.; Benfenati, E.; Roncaglioni, A. Evaluation of non-commercial models for genotoxicity and carcinogenicity in the assessment of EFSA's databases. *SAR and QSAR in Environmental Research* **2020**, *31*, 33–48.

(12) Tintó-Moliner, A.; Martin, M. Quantitative weight of evidence method for combining predictions of quantitative structure-activity relationship models. *SAR and QSAR in Environmental Research* **2020**, *31*, 261–279.

(13) Kasamatsu, T.; Kitazawa, A.; Tajima, S.; Kaneko, M.; Sugiyama, K.-i.; Yamada, M.; Yasui, M.; Masumura, K.; Horibata, K.; Honma, M. Development of a new quantitative structure–activity relationship model for predicting Ames mutagenicity of food flavor chemicals using StarDrop™ auto-Modeller™. *Genes and Environment* **2021**, *43*, 1–17.

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

(15) Thorne, D.; Kilford, J.; Hollings, M.; Dalrymple, A.; Ballantyne, M.; Meredith, C.; Dillon, D. The mutagenic assessment of mainstream cigarette smoke using the Ames assay: A multi-strain approach. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **2015**, *782*, 9–17.

(16) Williams, R. V.; DeMarini, D. M.; Stankowski Jr, L. F.; Escobar, P. A.; Zeiger, E.; Howe, J.; Elespuru, R.; Cross, K. P. Are all bacterial strains required by OECD mu-

tagenicity test guideline TG471 needed? *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **2019**, *848*, 503081.

(17) Ghasemi, F.; Mehridehnavi, A.; Perez-Garrido, A.; Perez-Sanchez, H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug discovery today* **2018**, *23*, 1784–1790.

(18) Hu, S.; Chen, P.; Gu, P.; Wang, B. A deep learning-based chemical system for QSAR prediction. *IEEE Journal of Biomedical and Health Informatics* **2020**, *24*, 3020–3028.

(19) Xu, Y. *Artificial Intelligence in Drug Design*; Springer, 2022; pp 233–260.

(20) Zhang, Y.; Yang, Q. An overview of multi-task learning. *National Science Review* **2018**, *5*, 30–43.

(21) Chu, C. S.; Simpson, J. D.; O'Neill, P. M.; Berry, N. G. Machine learning–Predicting Ames mutagenicity of small molecules. *Journal of Molecular Graphics and Modelling* **2021**, *109*, 108011.

(22) Hemmerich, J.; Ecker, G. F. In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2020**, *10*, e1475.

(23) Simulations Plus, Inc., MUT_Risk-8.5. 2018; `https://www.simulations-plus.com/?s=MUT\_Risk`, Accessed 2022-03-20.

(24) Pérez Santín, E.; Rodríguez Solana, R.; González García, M.; García Suárez, M. D. M.; Blanco Díaz, G. D.; Cima Cabal, M. D.; Moreno Rojas, J. M.; López Sánchez, J. I. Toxicity prediction based on artificial intelligence: A multidisciplinary overview. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, e1516.

(25) Tang, W.; Chen, J.; Wang, Z.; Xie, H.; Hong, H. Deep learning for predicting toxicity

of chemicals: A mini review. *Journal of Environmental Science and Health, Part C* **2018**, *36*, 252–271.

(26) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016**, *3*, 80.

(27) Huang, R.; Xia, M.; Nguyen, D., et al. Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. Front Environ Sci 5: 3. *Front. Environ. Sci* **2017**, *5*, 5.

(28) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. Modeling reactivity to biological macromolecules with a deep multitask network. *ACS central science* **2016**, *2*, 529–537.

(29) Dassault Systèmes, Biovia Bioactivity Databases Datasheet. 2020; `https://www.simulations-plus.com/?s=MUT\_Risk`, Accessed 2022-04-08.

(30) Wu, K.; Wei, G.-W. Quantitative toxicity prediction using topology based multitask deep neural networks. *Journal of chemical information and modeling* **2018**, *58*, 520–531.

(31) Benigni, R.; Battistelli, C. L.; Bossa, C.; Tcheremenskaia, O.; Crettaz, P. New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity. *Mutagenesis* **2013**, *28*, 401–409.

(32) Istituto Superiore di Sanità, ISSTOX Chemical Toxicity Databases. 2019; `https://www.iss.it/isstox`, Accessed 2022-03-22.

(33) Benigni, R. Towards quantitative read across: Prediction of Ames mutagenicity in a large database. *Regulatory Toxicology and Pharmacology* **2019**, *108*, 104434.

(34) Schrödinger, LLC, LigPrep. Schrödinger Release 2022-1. Accessed 2022-04-22.

(35) Schrödinger, LLC, Maestro. Schrödinger Release 2022-1. Accessed 2022-04-22.

(36) Greg Landrum, RDKit: Open-Source Cheminformatics Software. `http://www.rdkit.org`, Accessed 2021-08-05.

(37) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of cheminformatics* **2018**, *10*, 1–14.

(38) Chollet, F., et al. Keras. 2015; `https://github.com/fchollet/keras`, Accessed 2022-04-25.

(39) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

Graphical TOC Entry