Published in final edited form as:
Artif Intell Life Sci. 2021 December ; 1: . doi:10.1016/j.ailsci.2021.100028.

Novel computational models offer alternatives to animal testing for assessing eye irritation and corrosion potential of chemicals

Arthur C. Silva[a], **Joyce V.V.B. Borba**[a,b], **Vinicius M. Alves**[b], **Steven U.S. Hall**[a], **Nicholas Furnham**[c], **Nicole Kleinstreuer**[d], **Eugene Muratov**[b,e], **Alexander Tropsha**[b], **Carolina Horta Andrade**[a,*]

[a]LabMol-Laboratory for Molecular Modeling and Drug Design, Faculdade de Farmácia, Universidade Federal de Goiás-UFG, Goiânia, GO, Brazil

[b]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA

[c]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom

[d]National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, NIEHS, Durham, North Carolina 27560, USA

[e]Department of Pharmaceutical Sciences, Federal University of Paraiba, Joao Pessoa, PB 58059, Brazil

Abstract

Eye irritation and corrosion are fundamental considerations in developing chemicals to be used in or near the eye, from cleaning products to ophthalmic solutions. Unfortunately, animal testing is currently the standard method to identify compounds that cause eye irritation or corrosion. Yet, there is growing pressure on the part of regulatory agencies both in the USA and abroad to develop New Approach Methodologies (NAMs) that help reduce the need for animal testing and address unmet need to modernize safety evaluation of chemical hazards. In furthering the development and applications of computational NAMs in chemical safety assessment, in this study we have collected the largest expertly curated dataset of compounds tested for eye irritation and corrosion, and employed this data to build and validate binary and multi-classification Quantitative Structure-Activity Relationships (QSAR) models that can reliably assess eye irritation/corrosion potential of novel untested compounds. QSAR models were generated with Random Forest (RF) and Multi-Descriptor Read Across (MuDRA) machine learning (ML) methods, and validated using a 5-fold external cross-validation protocol. These models demonstrated high balanced accuracy (CCR of 0.68–0.88), sensitivity (SE of 0.61–0.84), positive predictive value (PPV of 0.65–0.90),

[*]Corresponding author. carolina@ufg.br (C.H. Andrade).

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. VM, EM, and AT are co-founders of Predictive, LLC, a UNC startup specializing in computational toxicology.

Supplementary materials
Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100028.

specificity (SP of 0.56–0.91), and negative predictive value (NPV of 0.68–0.85). Overall, MuDRA models outperformed RF models and were applied to predict compounds' irritation/corrosion potential from the Inactive Ingredient Database, which contains components present in FDA-approved drug products, and from the Cosmetic Ingredient Database, the European Commission source of information on cosmetic substances. All models built and validated in this study are publicly available at the STopTox web portal (https://stoptox.mml.unc.edu/). These models can be employed as reliable tools for identifying potential eye irritant/corrosive compounds

Introduction

Chemicals employed in cosmetics, drugs, pesticides, household products, among others, need to be classified appropriately according to their potential ocular toxicity to ensure safety [1]. Eye irritation or corrosion are characterized by cell membrane lysis, coagulation, saponification, and chemical reactivity [2]. All of these characteristics are mediated by contacts between a chemical and the eye surface (cornea and conjunctiva) [3].

The Draize test, published more than 70 years ago [4], relies on *in vivo* exposure to rabbits' eyes to classify chemicals according to their irritation/corrosion potential based on the damage caused within a well-defined timeframe [5]. However, this test relies upon qualitative scoring metrics of the severity and reversibility of highly subjective lesions, demonstrates poor reproducibility, and has questionable relevance to human exposure scenarios and human ocular biology [6]. Despite the scientific concern regarding the extrapolation of the observed results in rabbits to human eyes [7], the test is still used and recommended by the Organization for Economic Cooperation and Development (OECD).

The United Nations Globally Harmonized System (UN GHS) [8] proposes four categories to classify the chemicals: (*i*) Category 1 are compounds that cause irreversible eye effects within 21 days; (*ii*) Category 2A are compounds whose effects are reversible within 21 days; (*iii*) Category 2B are compounds whose effects are reversible within seven days; and (*iv*) No-Cat (NC) are compounds unable to cause eye corrosion or irritation.

Since the animal test ban in Europe for cosmetics ingredients in 2013, the development of alternative methods to substitute and reduce the number of animals in toxicological tests has become imperative [9]. The development of effective and efficient NAMs to animal testing [10] has been fueled in the last two decades by both public and political pressure [11] to employ the "Three Rs principles" to reduce, refine, and replace animal tests [12], and recent guidelines imposed by regulatory agencies create new demand for developing rapid, efficient alternative methods to animal testing [10]. Within this context, the 2018 ICCVAM strategic roadmap [13]called for the development of fit-for-purpose NAMs and the US EPA publicized its commitment to "eliminate all mammal study requests and funding by 2035" [14].

NAMs have been developed and made available for *in vitro* identification of ocular corrosives/severe irritants using alternative biological material including rabbit corneal cells (OECD Test Guideline 491) [15], isolated bovine corneas (OECD Test Guideline 437) [16], and a monolayer of Madin-Darby Canine Kidney (MDCK) cells (OECD Test Guideline 460)

[17,18]. Other three-dimensional human tissue models such as the Reconstructed Human Cornea-like Epithelium (RhCE) test (OECD Test Guideline 492) and the Vitrigel-Eye Irritancy test (OECD Test Guideline 494) are approved for use in a bottom-up approach identifying substances not classified for ocular irritation. These tests provide varying coverage of the biology relevant to eye irritation and corrosion when compared to human ocular anatomy and physiology [19].

Computational models provide a fast and low-cost solution to obtain reliable predictions for the endpoint of concern when generated on high-quality curated data and properly validated [10]. A major computational approach, named Quantitative Structure-Activity Relationship (QSAR) modeling, employs various statistical and artificial intelligence (AI) approaches, such as machine learning (ML) and deep learning (DL) to generate models that can accurately predict the outcome of testing new compounds in a specific assay, based on their molecular features. In recent years, the growth in publicly available data enabled the development of highly robust and predictive models [20,21]. However, modeling toxicity is a complex task as the underlying mechanisms are not always clear [3,21]. For this reason, QSAR models are highly dependent on the quality and volume of the data [12] in the training set, proper chemical and biological curation of primary data is critical [22,23], and failure to follow these practices question the trustworthiness of models [6].

Recently, there have been many attempts to model eye irritation endpoints with varying degrees of success (see Table 1). Though many of the models showed good overall accuracy, most models were not compliant with the OECD's guidelines for QSAR model development and validation [24], with models lacking the recommended use of an external set or Y-randomization [25-41], or not reporting the model applicability domain [28-41]. Many studies lack a rigorous curation and standardization of the chemicals used in the modeling, such as the study conducted by Verma et al. [25], resulting ultimately in unreliable predictions [42]. Additional problems include using unbalanced datasets, causing models to have an intrinsic bias toward the largest class [20,22]; and lack of model interpretation [20].[1] These limitations make it impossible to fairly compare those tools with other peer reviewed and public QSAR models.

Our team has extensively worked on the development of QSAR models for toxicity endpoints and developed web applications to disseminate the use of these models, such as Pred-hERG [43] and Pred-Skin [44]. Considering the lack of reliable models for eye irritation and corrosion, herein, we have collected, curated, and integrated the largest publicly available eye irritation and corrosion datasets, used it to build predictive and rigorously validated ML and instance-learner models, integrated these models into a software package called STopTox (Systemic and Topical chemical Toxicity), and made it publicly available (https://stoptox.mml.unc.edu/). We offer these models as reliable

[1] It is important to notice that there are some commercially available models such as ADC/Percepta (https://www.acdlabs.com/products/percepta/index.php) and Case Ultra (http://www.multicase.com/case-ultra) and freely available software tools such as Toxtree (http://toxtree.sourceforge.net/) and QSAR toolbox that do not fully disclose their parameters, as well as datasets and statistics (https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm#Guidance_Documents_and_Training_Materials_for_Using_the_Toolbox).

computational tools developed under the NAMs paradigm for evaluating chemical hazard potential for eye irritation/corrosion.

Materials and methods

Dataset overview

The publicly available data from the European Chemical Agency (ECHA) (https://echa.europa.eu/) used in this work was graciously provided by Thomas Luetchtefeld [51]. Additional eye irritation-related data were also extracted from multiple literature sources [25,29,35,40,45,52-55], curated (see next section), and integrated. The ECHA dataset was initially composed of 18,428 records for 9,801 chemicals and we compiled 2,769 additional records from the literature. In the ECHA dataset, 5,238 records with imputed eye irritation/corrosion data from QSAR models, weight-of-evidence or read-across were excluded, leaving 7,332 records. Chemicals with inconsistent hazard classification data ($n = 236$) were removed. Inorganics ($n = 330$) and mixtures ($n = 860$), totalizing 1,190 entries, were also removed from the dataset. The data collected from the literature had high overlap with the ECHA data presenting 2,438 duplicates. No discordant duplicates in terms of hazard characterization were found between the ECHA data and the literature. These duplicates were carefully analyzed and only one entry per compound was kept. Furthermore, only studies following the OECD Test Guideline 405 [56] (*in vivo* data) were kept, with 3,547 records remaining after the data curation process (Fig. 1).

The final (unbalanced) dataset was composed of 3,547 compounds, of which 2,401 were classified as non-irritant/non-corrosive, 937 were classified as irritant (categories 2A and 2B) of which 209 were classified as corrosive (category 1). The GHS classification for irritant/corrosive compounds was only available for 1,248 compounds of the dataset, where 209 compounds were classified as category 1 members, 166 were classified as category 2A, and 84 as category 2B, whereas 789 were classified as NC. These compounds classified under GHS system were used to generate multiclass models.

Binary QSAR models using the unbalanced data typically lead to biased models. To overcome this, the negative class in the unbalanced dataset was under-sampled to balance the data set. We used the smaller group of irritant compounds as probes to search for the most structurally similar non-irritants selecting half of the irritant group (469 compounds). The remaining 468 compounds were randomly chosen from the rest of the initial non-irritant class to maximize the chemical space coverage. This similarity-based selection procedure was carried out in KNIME using Tanimoto coefficient in two stages: (i) generate a similarity matrix of chemical space between all the pairs of compounds; and then (ii) choose 469 non-irritants with the largest Tanimoto similarity to the nearest irritant and 468 via random selection. Such procedures allowed us to create the most challenging training set with structurally similar irritants and non-irritants to achieve the most rigorous model capable of separating these two classes from each other and including a fraction of more diverse non-irritants to provide broader chemical space coverage. The final dataset consisted of 1,874 compounds (937 irritants and 937 non-irritants). The same approach was performed to balance the data for the generation of QSAR models to predict eye corrosion, *i.e.*, the NC class of compounds was under-sampled using both structural similarity and random sampling, leading to a balanced data of 418 compounds (209 corrosive and 209 non-corrosive).

Data curation

The compiled data was carefully curated and inspected according to protocols proposed by Fourches et al. [42,57]. Briefly, counter ions were stripped, mixtures and inorganics were removed, and specific chemotypes such as nitro groups and aromatic rings were standardized. Duplicates were identified, carefully analyzed, and only one entry was kept if biological responses were similar. The curation steps were implemented in the KNIME analytics platform (https://www.knime.com/) using in-house workflows. ISIDA Duplicates [58] was used to identify structural duplicates and ChemAxon Standardizer (v.16.5.16.0, ChemAxon, Budapest, Hungary, http://www.chemaxon.com) was used to standardize the chemical structures.

Cluster analysis

A 50 × 50 neuron self-organizing map (SOM) was generated using the open-source software Data Warrior (http://www.openmolecules.org/) [59] and employing SkelSpheres descriptors (http://www.openmolecules.org/help/similarity.html) [60]. Data Warrior software was used to cluster compounds that were colored according to their Global Harmonization System (GHS) [8] class, in order to provide an overview of the chemical space.

Molecular descriptors

We employed RDKit whole-molecule descriptors, Morgan, MACCS, and Dragon to develop QSAR and MuDRA models. SkelSpheres descriptors were calculated and used to cluster compounds in the SOM cluster analysis.

SkelSpheres

Skeleton Spheres descriptors [60] were calculated through the Osiris Data Warrior software (http://www.openmolecules.org/). SkelSphere is a 1,024 bin byte-vector descriptor that, despite being time- and memory-consuming, is more suitable than the other descriptors to perceive fine similarities. It also considers stereoisomers and has fewer hash collisions due to its higher resolution. The SkelSphere descriptor was calculated prior to the SOM generation to better understand and cluster the compounds of the modeling dataset and to visualize GHS classification labels.

RDKit molecular descriptors and fingerprints

In KNIME, a collection of 117 different RDKit molecular descriptors were calculated for the dataset followed by the removal of invariant descriptors and descriptors with a correlation higher than 0.9. MACCS structural keys [61] are implemented in the RDKit module available in the KNIME platform, as well as Morgan fingerprints [62]. RDKit provides 166 publicly available structural keys to represent molecules, and, for the Morgan fingerprint, it is possible to define the number of bits to encode the fingerprints as well as the radius, as the Morgan fingerprint is a circular fingerprint similar to ECFP and FCFP fingerprints family. For this study, Morgan fingerprints were generated using radius of 2 and 2048-bits length.

Dragon descriptors

Version 5.5 of Dragon software (Talete SRL, Milan, Italy) was used to generate all the 0D, 1D, and 2D descriptors provided by the software, totaling 2489 descriptors [63]. After the descriptors were calculated, invariant descriptors and descriptors with a correlation higher than 0.9 were also removed prior to the model generation step.

QSAR modeling

QSAR models for eye irritation and eye corrosion were generated employing a variety of chemical descriptors and algorithms. Binary and multiclass models were generated through the following steps: (i) data curation, preparation, and analysis; (ii) model generation and validation; (iii) model selection. To validate the method, we applied a 5-fold external cross-validation, where the curated dataset is divided into five equal-sized parts with an 80%/20% split between the modeling and test sets; this process is iteratively repeated until all parts of the dataset are used once as a test set. It is important to note that only the modeling set is used to generate the model; hence, during the 5-fold cross-validation procedure, compounds from the test set are not used in the generation of the models whatsoever and are solely reserved for the test set. Best models were carefully selected according to acceptable threshold values for all statistical metrics (for our purposes, this was set at 0.6). In addition, 10 rounds of Y-randomization were conducted to assess if the results were obtained by chance via annotating the statistical characteristics of the shuffled-labels models. Binary models were built for both corrosive and irritant classes of chemicals. Compounds classified as NC were used as non-corrosives and non-irritants as well.

Algorithms

Both RF [64] and MuDRA [65] algorithms were applied. RF is a well-known ensemble decision tree learning algorithm, while MuDRA is an instance-based learning process. MuDRA does not build an underlying model to make its predictions but performs an instantaneous classification of known irritant/corrosive and non-irritant/non-corrosive compounds based on their similarity range and nearest neighborhood. An in-depth explanation of how the MuDRA method can be applied can be found elsewhere [65]. Both methods are implemented in the KNIME analytics platform; RF is a built-in node provided by different developers, while MuDRA is implemented through the integration between KNIME platform and Python scripting language via built-in nodes for this purpose.

Statistical evaluation of models

The predictive power of both binary and multiclassification models was performed based on the output of the models during their respective validation processes. As described above, the 5-fold external cross-validation procedure was chosen to validate the models in this study. Hence, the statistical analysis is based on the collected results of predictions made in each fold of the cross-validation approach. For the multiclassification models, the same metrics were calculated, but considering the confusion matrix and comparing each class against all. The statistical metrics and the respective formulas are described below.

$$Se = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$CCR = \frac{Se + Sp}{2}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

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

$$Coverage = \frac{Reliable\ predictions}{Total\ predictions}$$

Here, TP and TN are true positives and negatives, respectively; FP and FN are false positives and negatives, respectively. *Se* stands for the sensitivity of the models, which is the correct identification of positive samples, while *Sp* is the measure of the specificity of the models, evaluating the ability of the model to identify negative samples correctly. CCR stands for the correct classification rate, and are calculated as the arithmetic mean of *Se* and *Sp*. PPV means positive predicted value, while NPV means negative predicted value; these metrics evaluate the probability of certainty of a positive or negative prediction, respectively. F_1 score, the harmonic mean of PPV and Se (aka precision and recall), evaluates the ability of the model to identify each instance correctly within the data. MCC encompasses the Mathews Correlation Coefficient and has been largely used as a goodness-of-fit in machine learning modeling tasks. MCC ranges from −1 to 1, being 0 equal to a random prediction.

The Coverage was calculated based on what we defined as reliable predictions, which means a prediction of a sample laying inside the applicability domain (AD) of the model, calculated through the formula below.

$$D_T = \bar{y} + Z\sigma$$

where D_T is a distance threshold, $\bar{y}$ is the average Euclidean distance of the k nearest neighbors of each compound of the training set, σ represents the standard deviation of the Euclidean distances and Z is an arbitrary parameter to control the level of significance. We set the default value of 0.5 for Z.

Multiclass modeling

To build multiclass models, three classes were considered based on GHS classification: corrosive, irritant (comprised by classes 2A and 2B), and NC. The binary statistical metrics described above were computed for each of the three classes and averaged to report overall performance for the multiclass models.

Virtual screening of CosIng and inactive ingredients database

The best models were applied to virtually screen the Cosmetics & Ingredients Substances Database (CosIng) [66], a European database of information about cosmetics and their ingredients. After curation, 4,780 compounds from CosIng were screened using our best performing models to identify compounds with the potential to cause eye irritation/corrosion.

The FDA inactive ingredients database (IID) set of compounds is freely available at https://www.fda.gov/drugs/drug-approvals-and-databases/inactive-ingredients-database-download. We also retrieved the data and curated it following the protocol described above. We applied the best models reported in this study to predict the ocular toxicity potential of the final IID set, composed of 4,673 inactive ingredients for pharmaceutical products.

Dissemination

All workflows used in this work are available in the supplementary material for those who want to build models to other endpoints as well as for instructions about how to implement MuDRA along with Python and KNIME set-ups.

Results and discussion

Cluster analysis

The SOM approach is an unsupervised classification technique that maps compounds to visualize their structural similarity. The structural map (Fig. 2) is colored by the three classes as defined by the GHS hazard classification system used to develop multiclass models. The highlighted compounds show that small structural differences can be observed in pairs of compounds belonging to distinct classes. Analyzing the background (shown in green in Fig. 2), the surrounding compounds are similar to each other (Tc = 0.85); major structural changes in the scaffold can also be observed (shown in yellow in Fig.

2), and structures with high dissimilarity are highlighted (shown in blue in Fig. 2). As seen in Fig. 2, the overall similarity between non-irritant and irritant compounds can be found across the whole map, sharing regions of the map in a non-compartmentalized way. This indicates that there are many activity cliffs in this dataset (as highlighted in Fig. 2). This type of dataset represents a challenge and although both RF and MuDRA are used to generate predictive models, they predict activity cliffs differently. As a prime example, 2,3-dihydro-1,2-benzothiazol-3-one (158 in Fig. 2) and 2,3-dihydro-1H-isoindole-1,3-dione (1044 in Fig. 2) are respectively within corrosive and NC categories, but share the same structural region in the SOM. Another example is 3-amino-4-chlorobenzene-1-sulfonic acid (200 in Fig. 2) and 3,4-dimethylbenzene-1-sulfonic acid (1031 in Fig. 2) pair, both in the same region of the SOM but respectively categorized as corrosive and NC. The same can be observed for the aliphatic compounds 1-chloro-2-[2-(2-*chloroethoxy*)*ethoxy*]ethane (148 in Fig. 2) and 1-chloro-2-[(2-*chloroethoxy*)*methoxy*]ethane (800 in Fig. 2), grouped together within the same region of the structural map and respectively classified as corrosive and NC. However, regions of chemical space are clearly enriched for particular categories, lending support to the application of QSAR modeling approaches while highlighting the necessity of nonlinear AI methods to identify the complex feature combinations that will discriminate categories.

QSAR modeling

Binary models—In this study, we built five binary models for eye irritation and five binary models for eye corrosion. For each endpoint, we built RF models using four molecular descriptors described in the methods section as well as one MuDRA model. The AD of each model was calculated, with an exception for the MuDRA method as it is an instance-based modeling approach.

As one can see, models generated using the RF method for both endpoints presented similar metrics. However, they were outperformed by Models 5 and 10, generated using the MuDRA method, which was in agreement with the advanced performance of MuDRA in comparison with other QSAR methods as reported by Alves and colleagues previously [65]. The binary models for eye corrosion showed higher predictivity. This could be because the eye corrosion dataset is smaller as compared to the eye irritation dataset.

Obtaining high PPV values is crucial when dealing with toxicological endpoints as they indicate the ability of models to accurately predict toxic compounds. For eye irritation, we can see that PPV values ranged from 0.70 to 0.90 (with the lowest value from Model 2 and the highest value from Model 5). For eye corrosion, PPV values ranged from 0.65 to 0.88 (with the lowest value from Model 6 and the highest value from Model 9). Overall, PPV values were above acceptable thresholds and reached high values (0.9), meaning that a prediction made by the two best models generated in this study for both eye irritation and corrosion would be correct with more than 85% certainty.

Likewise, high NPV values are equally important as they provide the certainty of the prediction made by the model regarding the nonirritant/non-corrosive classes. Classifying a molecule correctly as nonirritant/non-corrosive is very important as an incorrect prediction could lead to eyes being damaged. NPV values for the models built in our study ranged from 0.77 to 0.85 for eye irritation and from 0.68 to 0.83 for eye corrosion. This shows that negative predictions made using our best models have at least 83% certainty.

In analyzing sensitivity and specificity, other studies have reported that specificity values are usually higher than sensitivity values [45,67]. Here the sensitivity values' range was 0.77–0.89 for eye irritation and 0.61–0.84 for eye corrosion, while specificity values' range was 0.56–0.86 for eye irritation and 0.71–0.91 for eye corrosion. In our study, all models for eye irritation showed sensitivity values higher than specificity. For the eye corrosion models, the same pattern was observed only for Model 7 and Model 8, otherwise specificity value was higher than the sensitivity value.

An additional cluster analysis was conducted to further investigate the better performance of MuDRA when compared to the models generated using RF for eye irritation and eye corrosion endpoints. In this analysis, it was noticed that all compounds belonging to the biggest cluster of compounds of eye irritation dataset (six irritants and 11 non-irritants) were correctly predicted by MuDRA. Meanwhile, from those 17 compounds within eye irritation dataset, the models built with RF algorithm combined with one type of molecular descriptor (Dragon, MACCS, Morgan, or RDKit) mis-predicted on average 7 of them (see Supplementary File 9).

Fig. 3 compares 1,646 correct predictions made by MuDRA, 1,446 correct predictions made by RF_Dragon (Model 3), 1,436 correct predictions made by RF_MACCS (Model 4), 1,420 correct predictions made by RF_Morgan (Model 2), and 1,440 correct predictions made by RF_RDKit (Model 1). It shows that MuDRA was able to correctly predict 198 compounds that the other models were not. It is important to note that 1,081 correct predictions were shared by all models. This reinforces the importance of data curation process as well as the use of best practices for QSAR modeling. On the other hand, when the overlap between all mis-predicted compounds was checked, it was noticed that 38 compounds (25 irritants and 13 non-irritants – see File S9) mis-predicted by all models were predicted correctly only by MuDRA (Fig. 4).

Moreover, we observed that MuDRA was more accurate than RF models when making predictions for certain chemical classes, such as long chain hydrocarbons and fatty acid derivatives (Fig. 4), such as ethyl tetradecanoate, 1,6-dioctyl-hexanedioate, 2-methylpropyl octadecenoate, and 2-[2-(*nonanoyloxy*)*ethoxy*]ethyl nonanoate. However, as this cluster was composed by only 17 compounds, this is not enough to assure MuDRA superiority over RF models. Overall, MuDRA uses a broader descriptor space, which is able to capture more rigorously the structural differences between compounds, to identify the nearest neighbor, read-across it, and then return a more accurate prediction.

Multiclass models—Using the data and the GHS labeling system, multiclassification models were generated. We used three classes based on the GHS classification: corrosive, irritant (comprised by classes 2A and 2B), and NC. Table 3 shows the overall statistical metrics for all multiclass models built in this study, averaged across the binary metrics for each class performance.

Model 15, generated using the MuDRA method, outperformed the other RF models in all statistical metrics except sensitivity. Thus, the majority of generated models using RF were above the acceptable threshold. It is important to note that all metrics shown in Table 2 were calculated using each class's mean. The statistical characteristics showed that all models performed poorly when classifying compounds on GHS classes 2A and 2B This class has also been shown to have the lowest reproducibility when analyzing replicate animal tests, demonstrating the potential unreliability of classifications that are based solely on one result. However, the MuDRA method was able to handle the complexity of the data better by exploring the neighborhood of each compound and classifying them based on the nearest neighbor compound.

We have shown that MuDRA models were the best performing models in this work. As an external evaluation of model performance, we have predicted a set of 118 compounds extracted from Yamaguchi and colleagues' study [68]. After dataset curation and preparation to remove compounds that were also present in our dataset, 107 compounds could be predicted. All 73 corrosive compounds and the 38 irritant compounds in the dataset were correctly classified, as well as the remaining 6 NC compounds. To further validate our approach, we made predictions of three compounds found in the literature and not included in our modeling set, that had been reported as capable of triggering moderate to serious issues in human eyes [69]. The compounds are glutaraldehyde [70], Paraquat [71,72], and glyphosate [73]. All three were correctly predicted as being irritants. This reinforces the predictive power of the MuDRA approach and its applicability to important toxicological endpoints such as eye irritation/corrosion.

Virtual screening—As a further application of our models, we have retrieved and carefully curated 4780 compounds from the Cosing database and predicted their effects on eye corrosion / irritation using the MuDRA models; complete details of the results are available in the supplementary material. In summary, our prediction identified 2003 compounds with the potential to cause eye irritation. We also predicted the effects on eye irritation and corrosion of the Inactive Ingredients Database (IID) containing 4673 inactive ingredients using MuDRA based model. The subset of compounds used in the ophthalmic route of administration had 181 entries consisting of 76 unique ingredients. Among them, 24 were predicted as potential eye irritants and 12 as corrosive, where most of these are reported as a component of formulations such as ointments, solutions, suspensions, and eye drop products. The list of compounds predicted by our models as eye irritants and corrosion is available in the Supplementary Materials of the paper.

Conclusions

Eye irritation and corrosion are important toxicological endpoints for assessing chemical safety in humans and animals and respective tests are mandated by many regulatory agencies for the approval of a variety of products. The standard animal test for the evaluation of this endpoint is still the *in vivo* rabbit Draize test, a method developed decades ago and considered cruel, unreliable, and with questionable biological relevance to human exposure scenarios. Therefore, we aimed to develop predictive computational models using thoroughly curated data that could serve as NAMs for predicting eye irritation/corrosion potential of chemicals. Data curation is an extremely important factor in the development of robust and predictive ML models and a considerable amount of time was devoted to curating the ECHA dataset to ensure high quality training/test data and to optimize the predictive power of the models generated. All the curated data and developed models are available in KNIME workflows within the Supplementary Materials. These models presented high statistical characteristics. We have applied our models to predict a large publicly available cosmetics dataset (CosIng) as well as an Inactive Ingredient Dataset of chemicals commonly found in cosmetics and drugs. From CosIng database, 2003 compounds were predicted to cause damage to the eyes as corrosive/irritants; on the other hand, among 76 unique compounds from the Inactive Ingredients Dataset related to the ophthalmic route, 12 were predicted as corrosive, and 24 were predicted as irritants. The predictions for these chemicals are publicly available in the Supplementary Materials that accompanies this publication. Moreover, the models generated here are publicly available at the STopTox web portal (https://stoptox.mml.unc.edu/). These models can be employed as reliable alternatives to animal testing for identifying potential eye irritant/corrosive compounds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by CNPq (grant 400760/2014–2), FAPEG (grant 201310267001095). AT, ENM, and VMA acknowledge the financial support from NIH (grant 1R43ES032371). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (scholarships for ACS and JVVBB).

References

[1]. Verstraelen S, Van Rompay AR. CON4EI: development of serious eye damage and eye irritation testing strategies with respect to the requirements of the UN GHS/EU CLP hazard categories. Toxicol Vitr 2018;49(March 2017):2–5. doi:10.1016/j.tiv.2017.06.011.

[2]. Scott L, Eskes C, Hoffmann S, Adriaens E, Alepée N, Bufo M, Clothier R, Facchini D, Faller C, Guest R, Harbell J, Hartung T, Kamp H, Le Varlet B, Meloni M, McNamee P, Osborne R, Pape W, Pfannenbecker U, Prinsen M, Seaman C, Spielmann H, Stokes W, Trouba K, Van den Berghe C, Van Goethem F, Vassallo M, Vinardell P, Zuang V. A proposed eye irritation testing strategy to reduce and replace *in vivo* studies using bottom-up and top-down approaches. Toxicol Vitr 2010;24(1):1–9. doi:10.1016/j.tiv.2009.05.019.

[3]. Meek (Bette) ME. AOPs in hazard characterization for human health. Curr Opin Toxicol 2017;3:80–6. doi:10.1016/j.cotox.2017.06.002.

[4]. Draize JH, Woodard G, Calvery HO. Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. J Pharmacol Exp Ther 1944;82(3) 377 LP–390.

[5]. Wilhelmus KR. The draize eye test. Surv Ophthalmol 2001;45(6):493–515. doi:10.1016/S0039-6257(01)00211-9. [PubMed: 11425356]

[6]. Alves VM, Borba J, Capuzzi SJ, Muratov E, Andrade CH, Rusyn I, Tropsha A. Oy Vey! A comment on ‘machine learning of toxicological big data enables read-across structure activity relationships outperforming animal test reproducibility. Toxicol Sci 2019;167(1):227–38. doi:10.1093/toxsci/kfy286. [PubMed: 30215777]

[7]. Verma RP, Matthews EJ. Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models (QSAR-21):

part I: irritation potential. Regul Toxicol Pharmacol 2015;71(2):318–30. doi:10.1016/j.yrtph.2014.11.011. [PubMed: 25497990]

[8]. Globally harmonized system of classification and labelling of chemicals (GHS); Globally harmonized system of classification and labelling of chemicals (GHS); UN, 2019. doi:10.18356/f8fbb7cb-en.

[9]. da Silva ACG, Chialchia AR, de Ávila RI, Valadares MC. Mechanistic-based non-animal assessment of eye toxicity: inflammatory profile of human keratinocytes cells after exposure to eye damage/irritant agents. Chem Biol Interact 2018;292(February):1–8. doi:10.1016/j.cbi.2018.06.031. [PubMed: 29953848]

[10]. Alves VM, Capuzzi SJ, Braga RC, Borba JVB, Silva AC, Luechtefeld T, Hartung T, Andrade CH, Muratov EN, Tropsha A. A perspective and a new integrated computational strategy for skin sensitization assessment. ACS Sustain Chem Eng 2018;6(3):2845–59. doi:10.1021/acssuschemeng.7b04220.

[11]. European Parliament, C. of the E.U. Regulation (EC) No 1223/2009 of the European parliament and of the council of 30 November 2009 on cosmetic products.

[12]. Alves VM, Auerbach SS, Kleinstreuer N, Rooney JP, Muratov EN, Rusyn I, Tropsha A, Schmitt C. Curated data in-trustworthy in silico models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing. Altern Lab Anim 2021;49(3):73–82. doi:10.1177/02611929211029635. [PubMed: 34233495]

[13]. ICCVAM. A strategic roadmap for establishing new approaches to evaluate the safety of chemicals and medical products in the United States https://ntp.niehs.nih.gov/pubhealth/evalatm/natl-strategy/index.html (accessed Jan 27, 2021).

[14]. US Environmental Protection Agency. EPA directive to prioritize efforts to reduce animal testing https://www.epa.gov/sites/production/files/2019-09/documents/image2019-09-09-231249.pdf (accessed Jun 15, 2021).

[15]. Test no. 491: short time exposure *in vitro* test method for identifying i) chemicals inducing serious eye damage and ii) chemicals not requiring classification for eye irritation or serious eye damage; oecd guidelines for the testing of chemicals, section 4; OECD, 2018. doi:10.1787/9789264242432-en.

[16]. Test no. 437: bovine corneal opacity and permeability test method for identifying i) chemicals inducing serious eye damage and ii) chemicals not requiring classification for eye irritation or serious eye damage; oecd guidelines for the testing of chemicals, section 4; OECD, 2017. doi:10.1787/9789264203846-en.

[17]. Test no. 460: fluorescein leakage test method for identifying ocular corrosives and severe irritants; oecd guidelines for the testing of chemicals, section 4; OECD, 2017. doi:10.1787/9789264185401-en.

[18]. Wilson SL, Ahearne M, Hopkinson A. An overview of current techniques for ocular toxicity testing. Toxicology 2015;327:32–46. doi:10.1016/j.tox.2014.11.003. [PubMed: 25445805]

[19]. Clippinger AJ, Raabe HA, Allen DG, Choksi NY, van der Zalm AJ, Kleinstreuer NC, Barroso J, Lowit AB. Human-relevant approaches to assess eye corrosion/irritation potential of agrochemical formulations. Cutan Ocul Toxicol 2021;40(2):145–67. doi:10.1080/15569527.2021.1910291. [PubMed: 33830843]

[20]. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: where have you been? Where are you going to? J Med Chem 2014. doi:10.1021/jm4004285.

[21]. Gleeson MP, Modi S, Bender A, Robinson RLM, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC. The challenges involved in modeling toxicity data in silico: a review. Curr Pharm Des 2012;18(9):1266–91. [PubMed: 22316153]

[22]. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform 2010;29(6–7):476–88. [PubMed: 27463326]

[23]. Zhu H. From QSAR to QSIIR: searching for enhanced computational toxicology models. Methods Mol Biol 2013;930:53–65. doi:10.1007/978-1-62703-059-5_3. [PubMed: 23086837]

[24]. OECD principles for the validation, for regulatory purposes, of (Quantitative) structure-activity relationship models.

[25]. Verma RP, Matthews EJ. An *in silico* expert system for the identification of eye irritants. SAR QSAR Environ Res 2015;26(5):383–95. doi:10.1080/1062936X.2015.1039578. [PubMed: 25967253]

[26]. Liew CY, Yap CW. QSAR and predictors of eye and skin effects. Mol Inform 2013;32(3):281–90. doi:10.1002/minf.201200119. [PubMed: 27481523]

[27]. Wang Q, Li X, Yang H, Cai Y, Wang Y, Wang Z, Li W, Tang Y, Liu G. *In silico* prediction of serious eye irritation or corrosion potential of chemicals. RSC Adv 2017. doi:10.1039/c6ra25267b.

[28]. Abbasitabar F, Zare-Shahabadi V. *In silico* prediction of toxicity of phenols to tetrahymena pyriformis by using genetic algorithm and decision tree-based modeling approach. Chemosphere 2017;172:249–59. doi:10.1016/j.chemosphere.2016.12.095. [PubMed: 28081509]

[29]. Geerts L, Adriaens E, Alépée N, Guest R, Willoughby JA, Kandarova H, Drzewiecka A, Fochtman P, Verstraelen S, Van Rompay AR. CON4EI: evaluation of QSAR models for hazard identification and labelling of eye irritating chemicals. Toxicol Vitr 2017. doi:10.1016/j.tiv.2017.09.004.

[30]. Bhhatarai B, Wilson DM, Parks AK, Carney EW, Spencer PJ. Evaluation of TOPKAT, toxtree, and derek nexus *in silico* models for ocular irritation and development of a knowledge-based framework to improve the prediction of severe irritation. Chem Res Toxicol 2016. doi:10.1021/acs.chemrestox.5b00531.

[31]. Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H, Hartung T. Analysis of draize eye irritation testing and its prediction by mining publicly available 2008-2014 reach data. ALTEX 2016;33(2):123–34. doi:10.14573/altex.1510053. [PubMed: 26863293]

[32]. Luechtefeld T, Marsh D, Rowlands C, Hartung T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) Outperforming animal test reproducibility. Toxicol Sci 2018(No. August):1–15. doi:10.1093/toxsci/kfy152.

[33]. Verma RP, Matthews EJ. An *in silico* expert system for the identification of eye irritants. SAR QSAR Environ Res 2015;26(5):383–95. doi:10.1080/1062936X.2015.1039578. [PubMed: 25967253]

[34]. Worth AP, Cronin MTD. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. J Mol Struct Theochem 2003. doi:10.1016/S0166-1280(02)00622-X.

[35]. Cruz-Monteagudo M, González-Díaz H, Borges F, González-Díaz Y. Simple stochastic fingerprints towards mathematical modeling in biology and medicine. 3. Ocular irritability classification model. Bull Math Biol 2006;68(7):1555–72. doi:10.1007/s11538-006-9083-y. [PubMed: 16865609]

[36]. Solimeo R, Zhang J, Kim M, Sedykh A, Zhu H. Predicting chemical ocular toxicity using a combinatorial QSAR approach. Chem Res Toxicol 2012. doi:10.1021/tx300393v.

[37]. Patlewicz G, Rodford R, Walker JD. Quantitative structure-activity relationships for predicting skin and eye irritation. Environ Toxicol Chem 2003;22(8):1862–9. [PubMed: 12924585]

[38]. Sugai S, Murata K, Kitagaki T, Tomita I. Studies on eye irritation caused by chemicals in rabbits-1. A quantitative structure-activity relationships approach to primary eye irritation of chemicals in rabbits. J Toxicol Sci 1990;15(4):245–62. doi:10.1248/cpb.37.3229. [PubMed: 2082022]

[39]. Cronin MTD, Basketter DA, York M. A quantitative structure-activity relationship (QSAR) investigation of a draize eye irritation database. Toxicol Vitr 1994;8(1):21–8. doi:10.1016/0887-2333(94)90204-6.

[40]. Barratt MD. QSARS for the eye irritation potential of neutral organic chemicals. Toxicol Vitr 1997;11(1–2):1–8. doi:10.1016/S0887-2333(96)00063-X.

[41]. Abraham MH, Kumarsingh R, Cometto-Muniz JE, Cain WS. A quantitative structure–activity relationship (QSAR) for a draize eye irritation database. Toxicol Vitr 1998;12(3):201–7. doi:10.1016/S0887-2333(97)00117-3.

[42]. Fourches D, Muratov E, Tropsha ATrust, Verify but. On the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 2010;50(7):1189–204. doi:10.1021/ci100176x. [PubMed: 20572635]

[43]. Braga RC, Alves VM, Silva MFB, Muratov E, Fourches D, Lião LM, Tropsha A, Andrade CH. Pred-HERG: a novel web-accessible computational tool for predicting cardiac toxicity. Mol Inform 2015;34(10):698–701. doi:10.1002/minf.201500040. [PubMed: 27490970]

[44]. Braga RC, Alves VM, Muratov EN, Strickland J, Kleinstreuer N, Trospsha A, Andrade CH. Pred-Skin: a fast and reliable web application to assess skin sensitization effect of chemicals. J Chem Inf Model 2017. doi:10.1021/acs.jcim.7b00194.

[45]. Basant N, Gupta S, Singh KP. A Three-tier QSAR modeling strategy for estimating eye irritation potential of diverse chemicals in rabbit for regulatory purposes. Regul Toxicol Pharmacol 2016;77:282–91. doi:10.1016/j.yrtph.2016.03.014. [PubMed: 27018829]

[46]. PaDEL-DDPredictor. Eye/Skin Corrosion (version 20110805) http://www.yapcwsoft.com/dd/padelddpredictor/models/toxicity/eyeskincorrosion/20110805/.

[47]. Lu J, Zhang P, Zou XW, Zhao XQ, Cheng KG, Zhao YL, Bi Y, Zheng MY, Luo XM. *In silico* prediction of chemical toxicity profile using local lazy learning. Comb Chem High Throughput Screen 2017;20(4). doi:10.2174/1386207320666170217151826.

[48]. Verma RP, Matthews EJ. Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models (QSAR-21): part I: irritation potential. Regul Toxicol Pharmacol 2015;71(2):318–30. doi:10.1016/j.yrtph.2014.11.011. [PubMed: 25497990]

[49]. Verma RP, Matthews EJ. Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models (QSAR-21): part II: corrosion potential. Regul Toxicol Pharmacol 2015;71(2):331–6. doi:10.1016/j.yrtph.2014.12.004. [PubMed: 25510831]

[50]. Patlewicz GY. Rodford RA, Ellis G, Barratt MD. A QSAR model for the eye irritation of cationic surfactants. Toxicol *In Vitro* 2000;14(1):79–84. [PubMed: 10699364]

[51]. Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H, Hartung T. Analysis of publically available skin sensitization data from REACH registrations 2008-2014. ALTEX 2016;33(2):135–48. doi:10.14573/altex.1510055. [PubMed: 26863411]

[52]. Verheyen GR, Braeken E, Van Deun K, Van Miert S. Evaluation of existing (Q)SAR models for skin and eye irritation and corrosion to use for REACH registration. Toxicol Lett 2017;265:47–52. doi:10.1016/j.toxlet.2016.11.007. [PubMed: 27865849]

[53]. Adriaens E, Alépée N, Kandarova H, Drzewieckac A, Gruszka K, Guest R, Willoughby JA, Verstraelen S, Van Rompay AR. CON4EI: selection of the reference chemicals for hazard identification and labelling of eye irritating chemicals. Toxicol Vitr 2017;44(April):44–8. doi:10.1016/j.tiv.2017.06.001.

[54]. Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P. Cosmetics europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the draize eye test ref. Arch Toxicol 2017;91(2):521–47. doi:10.1007/s00204-016-1679-x. [PubMed: 26997338]

[55]. Barratt MD. A quantitative structure-activity relationship for the eye irritation potential of neutral organic chemicals. Toxicol Lett 1995. doi:10.1016/0378-4274(95)03338-L.

[56]. Test no. 405: acute eye irritation/corrosion; oecd guidelines for the testing of chemicals, section 4. OECD; 2017. doi:101787/9789264185333-en.

[57]. Fourches D, Muratov E, Tropsha A. Trust, but verify ii: a practical guide to chemogenomics data curation. J Chem Inf Model 2016;56(7):1243–52. doi:10.1021/acs.jcim.6b00129. [PubMed: 27280890]

[58]. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov’ev V, Hoonakker F, Tetko I, Marcou G. ISIDA-platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 2008;4(3):191–8. doi:10.2174/157340908785747465.

[59]. Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 2015;55(2):460–73. doi:10.1021/ci500588j. [PubMed: 25558886]

[60]. Boss C, Hazemann J, Kimmerlin T, von Korff M, Lüthi U, Peter O, Sander T, Siegrist R. The screening compound collection: a key asset for drug discovery. Chim Int J Chem 2017;71(10):667–77. doi:10.2533/chimia.2017.667.

[61]. Anderson S. Graphical representation of molecules and substructure-search queries in MACCStm. J Mol Graph 1984;2(3):83–90. doi:10.1016/0263-7855(84)80060-0.

[62]. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 1965;5(2):107–13. doi:10.1021/c160017a018.

[63]. Todeschini R. Methods and principles in medicinal chemistry. Handbook of molecular descriptors. Todeschini R, Consonni V, editors. Weinheim, Germany: Wiley-VCH Verlag GmbH; 2000. doi:10.1002/9783527613106.

[64]. Breiman LEO. Random forests. Mach Learn 2001;45:5–32. doi:10.1023/A:1010933404324.

[65]. Alves VM, Golbraikh A, Capuzzi SJ, Liu K, Lam WI, Korn DR, Pozefsky D, Andrade CH, Muratov EN, Tropsha A. Multi-zcross (MuDRA): a simple and transparent approach for developing accurate quantitative structure–activity relationship models. J Chem Inf Model 2018;58(6):1214–23. doi:10.1021/acs.jcim.8b00124. [PubMed: 29809005]

[66]. European Commission. Cosmetic ingredient database http://ec.europa.eu/growth/tools-databases/cosing/index.cfm?fuseaction=search.results.

[67]. Geerts L, Adriaens E, Alépée N, Guest R, Willoughby JA, Kandarova H, Drzewiecka A, Fochtman P, Verstraelen S, Van Rompay AR. CON4EI: evaluation of QSAR models for hazard identification and labelling of eye irritating chemicals. Toxicol Vitr 2017(No. April):0–1. doi:10.1016/j.tiv.2017.09.004.

[68]. Yamaguchi H, Kojima H, Takezawa T. Predictive performance of the vitrigel-eye irritancy test method using 118 chemicals. J Appl Toxicol 2016;36(8):1025–37. doi:10.1002/jat.3254. [PubMed: 26472347]

[69]. Jaga K, Dharmani C. Ocular toxicity from pesticide exposure: a recent review. Environ Health Prev Med 2006;11(3):102–7. doi:10.1265/ehpm.11.102 [PubMed: 21432383]

[70]. Ünal M, Yücel İ, Akar Y, Öner A, Altın M. Outbreak of toxic anterior segment syndrome associated with glutaraldehyde after cataract surgery. J Cataract Refract Surg 2006;32(10):1696–701. doi:10.1016/j.jcrs.2006.05.008. [PubMed: 17010870]

[71]. Joyce M. Ocular damage caused by paraquat. Br J Ophthalmol 1969;53(10):688–90. doi:10.1136/bjo.53.10.688. [PubMed: 5347169]

[72]. McKeag D. The ocular surface toxicity of paraquat. Br J Ophthalmol 2002;86(3):350–1. doi:10.1136/bjo.86.3.350. [PubMed: 11864897]

[73]. Bradberry SM, Proudfoot AT, Vale JA. Glyphosate poisoning. Toxicol Rev 2004;23(3):159–67. [PubMed: 15862083]

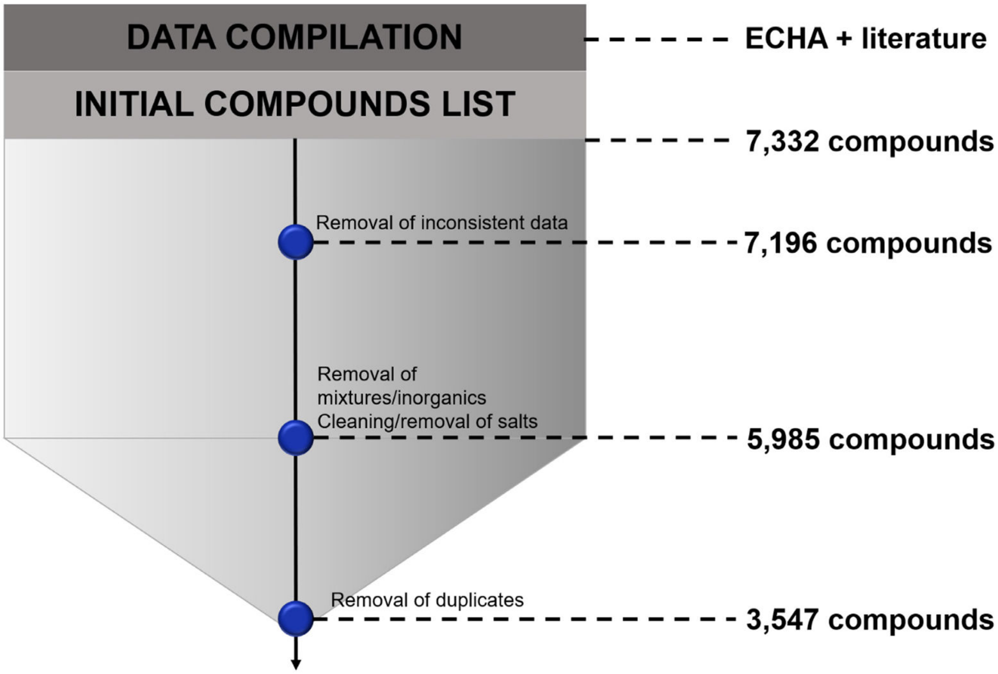


Fig. 1.

Data compilation and curation workflow.

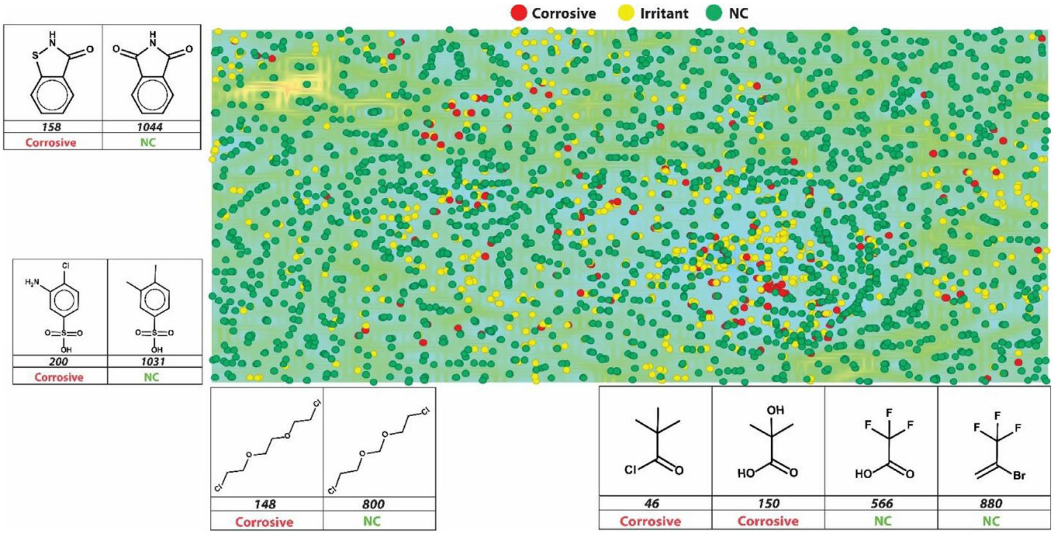


Fig. 2.

Graphical representation of a self-organized map for the chemical space covered by modeling set chemicals. Red circles represent corrosives, yellow circles represent irritants, and green circles represent NC class. Blue-green regions show compounds that share structural similarities compared to their neighbors, and yellow-orange-red regions represent an abrupt change in the chemical structure of the compounds compared to their neighbors. The dataset is notably complex; there are similar compounds belonging to different classes, which makes the construction of multiclassification models a challenge.

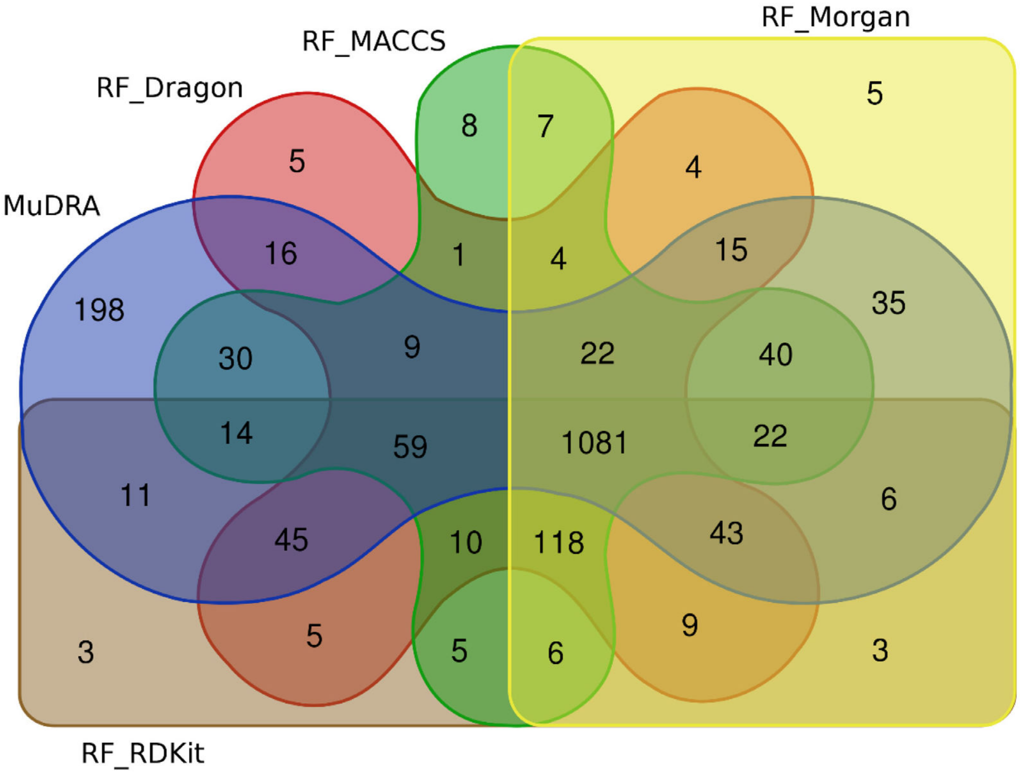
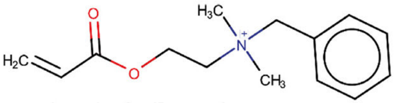
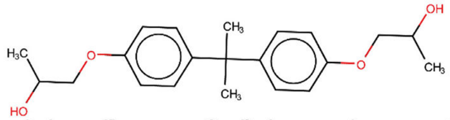


Fig. 3.
Venn diagram showing the overlap between correct predictions done by all models for the eye irritation dataset.



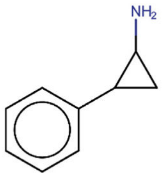
Benzyldimethyl[2-(prop-2-enoyloxy)ethyl]azanium

Experimental outcome	MuDRA	RF_MACCS	RF_Morgan	RF_Dragon	RF_RDKit
Irritant	Irritant	Non-irritant	Non-irritant	Non-irritant	Non-irritant



1-(4-{2-[4-(2-hydroxypropoxy)phenyl]propan-2-yl}phenoxy)propan-2-ol

Experimental outcome	MuDRA	RF_MACCS	RF_Morgan	RF_Dragon	RF_RDKit
Irritant	Irritant	Non-irritant	Non-irritant	Non-irritant	Non-irritant



2-phenylcyclopropan-1-amine

Experimental outcome	MuDRA	RF_MACCS	RF_Morgan	RF_Dragon	RF_RDKit
Non-irritant	Non-irritant	Irritant	Irritant	Irritant	Irritant

Fig. 4.

Example of compounds correctly predicted only by MuDRA.

Table 1

Previously published QSAR models of eye irritation.

Author	Curation	Cross-validation	Y-rand or external set	AD	Number of compounds	Metrics	AI/Discriminant method	Descriptor	Year	Model availability
Basant et al. [45]	Yes	Yes	Yes	Yes	107	Training: 77–94% Test set: 72–87%	CT, RT	Padel	2016	Unavailable
Verma et al. [25]	No	No	External set only	Yes	816 training 86 test	CCR = 72.3%	DT	Molecular weight, logP, melting point, aqueous solubility, lipid solubility	2015	Unavailable
Liew et al. [26]	Yes	Yes	External set only	Yes	2108 split in multiple categories	Training: CCR = 65–100% Test: CCR = 41–69%	SVM, kNN	Padel	2013	Publicly available [46]
Wang et al. [27]	Yes	Yes	External set only	Yes	6015 training 1504 test	CCR = 0.92–95%	ANN, kNN, NB, SVM	Atom pair, estate fingerprint, CDK fingerprints, Klekota–Roth fingerprint, MACCS fingerprint, Pubchem fingerprint and substructure fingerprint	2017	Unavailable
Jing Lu [47]	No	No	External set only	No	1845 training 496 test	CCR = 68%	Read Across	Codessa	2017	Unavailable
Geerts et al. [29]	No	No	No	No	80	CCR = 60–80%	Third-part software	None	2018	Unavailable
Bhhatarai et al. [30]	No	No	No	No	1644	CCR = 74–80%	Third-part software	None	2016	Unavailable
Luechtefeld et al. [31]	No	No	External set only	No	929	DT, kNN CCR = 73%–100%		Pubchem2d fingerprint	2016	Unavailable
Luechtefeld et al. [32]	No	Yes	External set only	No	15,760	CCR = 98%	Read Across	Pubchem2d fingerprint	2018	Unavailable
Verma et al. [48,49]	No	No	External set only	No	2928	Training: CCR = 85% Test: CCR = 83%	ANN	ADMET predictor	2015	Unavailable
Worth and Cronin [34]	No	Yes	No	No	119	CCR = 60–73%	LDA, CT, LR	Molecular weight	2003	Unavailable
Cruz-Monteagudo et al. [35]	No	LOO	No	No	46	Acc = 80.43%	LDA	LogP	2006	Unavailable
Solimeo et al. [36]	Yes	Yes	No	No	75	CCR = 82–92%	RF, kNN	Dragon, MOE	2012	Available by request *
Patlewicz et al. [50]	No	No	No	No	29	R [2]= 0.702	ANN	Logcmc, logP, molvol, mas $n-mas	2000	Unavailable
Sugai et al. [38]	No	LOO	No	No	138	Acc = 86.3%, Validation = 74%	ALS	Physico-chemical descriptors	1990	Unavailable

Author	Curation	Cross-validation	Y-rand or external set	AD	Number of compounds	Metrics	AI/Discriminant method	Descriptor	Year	Model availability
Cronin et al. [39]	No	No	No	No	53	R [2] = 0.80	LDA, LR	ClogP, kappa indices, molecular connectivity indices	1994	Unavailable
Barratt et al. [40]	No	No	No	No	46	N/A	PCA	ClogP, mol. vol., Dipole moment,	1995	Unavailable
Abraham et al. [41]	No	No	No	No	91	R^2 = 0.94	LR	Liquid vapor pressure, mr, π^2, $\Sigma\alpha$, $\Sigma\beta$, liquid hexadecane partition	1998	Unavailable

CT = Classification Trees; RT = Regression Trees; SVM = support vector machines; kNN = k -Nearest Neighbor; ANN = Artificial Neural Networks; NB = Naïve Bayes; LDA = Linear Discrimination Analysis; LR = Linear Regression; RF = Random Forest; PCA = Principal Component Analysis; ALS = Adaptative Least Squares LOO = Leave One Out; CCR = Correct Classification Rate; Acc = Accuracy; N/A = not applicable; R^2 = Correlation coefficient.

[*] Compounds must be sent to the authors to be predicted.

Table 2

Statistical characteristics of binary QSAR models for eye irritation and eye corrosion assessed by 5-fold cross-validation.

Binary models for eye irritation generated with RF algorithm

Model	Descriptor	CCR	Se	PPV	Sp	NPV	F1	MCC	Coverage
1	RDKit	0.76	0.77	0.76	0.76	0.77	0.77	0.53	1
	RDKit-AD	0.77	0.78	0.77	0.76	0.77	0.77	0.53	0.96
2	Morgan	0.77	0.84	0.73	0.69	0.81	0.78	0.53	1
	Morgan-AD	0.72	0.88	0.70	0.56	0.81	0.78	0.47	0.91
3	Dragon	0.77	0.78	0.77	0.77	0.78	0.78	0.55	1
	Dragon-AD	0.77	0.79	0.76	0.75	0.78	0.77	0.54	0.97
4	MACCS	0.77	0.80	0.75	0.73	0.79	0.77	0.53	1
	MACCS-AD	0.76	0.81	0.74	0.71	0.79	0.77	0.52	0,99

Binary model for eye irritation generated with MuDRA algorithm

Model	Descriptor	CCR	Se	PPV	Sp	NPV	F1	MCC	Coverage
5	Multi	0.88	0.89	0.90	0.86	0.85	0.90	0.76	1

Binary models for eye corrosion generated with RF algorithm

Model	Descriptor	CCR	Se	PPV	Sp	NPV	F1	MCC	Coverage
6	RDKit	0.70	0.70	0.71	0.71	0.70	0.70	0.41	1
	RDKit-AD	0.75	0.61	0.86	0.89	0.68	0.72	0.52	0.88
7	Morgan	0.68	0.76	0.65	0.59	0.71	0.70	0.36	1
	Morgan-AD	0.75	0.76	0.81	0.75	0.69	0.78	0.5	0.85
8	Dragon	0.72	0.73	0.72	0.71	0.73	0.72	0.44	1
	Dragon-AD	0.76	0.67	0.84	0.86	0.69	0.75	0.54	0.92
9	MACCS	0.76	0.73	0.78	0.79	0.74	0.75	0.52	1
	MACCS-AD	0.77	0.64	0.88	0.91	0.71	0.74	0.57	0.98

Binary model for eye corrosion generated with MuDRA algorithm

Model	Descriptor	CCR	Se	PPV	Sp	NPV	F1	MCC	Coverage
10	Multi	0.85	0.84	0.86	0.86	0.83	0.85	0.69	1

Table 3

Statistical characteristics of multiclass QSAR models for eye irritation and eye corrosion.

				Multiclass models generated with RF modeling method					
Model	**Descriptor**	**CCR**	**Se**	**PPV**	**Sp**	**NPV**	**F1**	**MCC**	**Coverage**
11	RDKit	0.62	0.61	0.62	0.63	0.62	0.51	0.21	1
	RDKit-AD	0.60	0.60	0.60	0.60	0.61	0.52	0.22	1
12	Dragon	0.63	0.66	0.63	0.61	0.64	0.52	0.31	1
	Dragon-AD	0.60	0.61	0.61	0.58	0.58	0.52	0.31	1
13	MACCS	0.65	0.63	0.65	0.66	0.64	0.55	0.38	1
	MACCS-AD	0.64	0.67	0.64	0.62	0.64	0.56	0.39	1
14	Morgan	0.63	0.67	0.63	0.60	0.64	0.50	0.39	1
	Morgan-AD	0.60	0.71	0.62	0.49	0.60	0.51	0.39	1
					Multiclass model generated with MuDRA modeling method				
Model	**Descriptor**	**CCR**	**Se**	**PPV**	**Sp**	**NPV**	**F1**	**MCC**	**Coverage**
15	Multi	0.74	0.60	0.84	0.87	0.89	0.62	0.57	1