STopTox: An *in-silico* alternative to animal testing for acute Systemic and TOPical TOXicity

Joyce V.B. Borba[1,2], Vinicius M. Alves[1], Rodolpho C. Braga[3], Daniel Korn[1], Kirsten Overdahl[4], Arthur C. Silva[2], Steven U. S. Hall[2], Erik Overdahl[1], Nicole Kleinstreuer[5], Judy Strickland[6], David Allen[6], Carolina Horta Andrade[2], Eugene Muratov[1,7*], Alexander Tropsha[1,*].

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

[2]*Laboratory for Molecular Modeling and Drug Design, Federal University of Goias, Goiania, GO Brazil;*

[3]*InsilicAll, Sao Paulo, SP, Brazil*

[4]*Nicholas School of the Environment, Duke University, Durham, NC, USA;*

[5]*NTP ICVAM, NIEHS, RTP, NC, USA.*

[6]*Integrated Laboratory Systems, Contractor supporting the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM).*

[7]*Department of Pharmaceutical Sciences, Federal University of Paraiba, Joao Pessoa, PB, 58059, Brazil.*

* To whom correspondence and material requests should be addressed: 100K Beard Hall, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.

Abstract

Since 2009, animal testing for cosmetic products has been prohibited in Europe, and in 2016, US EPA announced their intent to modernize the so-called "6-pack" of acute toxicity tests (acute oral toxicity, acute dermal toxicity, acute inhalation toxicity, skin irritation and corrosion, eye irritation and corrosion, and skin sensitization) and expand acceptance of alternative methods to reduce animal testing of pesticides. We have compiled, curated, and integrated the largest publicly available dataset and developed an ensemble of QSAR models for all six endpoints. All models were validated according to the OECD QSAR principles and tested using newly identified data on compounds not included in the training sets. We have established a publicly accessible Systemic and Topical chemical Toxicity (STopTox) web portal (https://stoptox.mml.unc.edu/) integrating all developed models for “6-pack” assays. This portal can be used by scientists and regulators to identify putative toxicants or non-toxicants in chemical libraries of interest.

Introduction

Historically, animal testing has been required by regulatory agencies for hazard categorization, labeling, and packaging.[1] However, there have been multiple calls, especially, in the last two decades for reducing, refining, and replacing animal tests for hazard identification.[2] Still, the potential for chemicals, especially those used in cosmetic products to cause several types of acute toxicity in humans is still evaluated by the animal *in vivo* assays historically termed the "6-pack".[3] This battery of tests includes three topical (skin sensitization, skin irritation and corrosion, and eye irritation and corrosion) and three systemic (acute oral toxicity, acute inhalation toxicity, and acute dermal toxicity) endpoints.[3]

The European Union (EU) has maintained an animal testing ban on finished cosmetic products since 2004, and since 2009, the EU has maintained an animal testing ban on cosmetic ingredients. Since 2013, this ban has been applied irrespective of the availability of alternative non-animal tests.[4] In the United States, although animal testing is not prohibited, Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and its supporting center, the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), have been coordinating the development, validation, acceptance, and harmonization of alternative toxicological test methods throughout the U.S. Federal Government. Still, traditionally, all compounds registered under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), are required to be tested in the 6-pack assays. In 2015, 6-pack guidelines were revised to include *in vitro* and *ex vivo* test methods for classification and labeling for eye irritation – these new guidelines applied both to household antimicrobial cleaning products and to conventional pesticide products.[4] In 2018, the EPA's Draft Interim Science Policy allowed *in vitro* tests for skin sensitization as well; however, only pure substances may be used as test materials for submission purposes, and an Integrated Testing Strategy (either the "2 out of 3 Defined Approach" or the "Key Event 3/1 Sequential Testing Strategy") must be followed.[5] However, the European Commission reported 9.58 million uses of animals in research and testing at UE, and 23% of those animal uses were for regulatory purposes in 2017.[6]

Along with ethical concerns, the validity of animal testing has also come under question in recent years. Several studies have shown that animal-based assay outcomes do not always equate with human response[7,8] and that animal models are less reproducible than some alternative methods.[9] However, despite the international adoption of a variety of in vitro alternatives, their widespread implementation has been hampered by concerns over concordance with the traditional animal tests, and practical considerations as compared to the traditional animal tests.[10]

Computational approaches, such as structural alerts, read-across, and Quantitative Structure-Activity Relationship (QSAR) modeling, have earned broad acceptance as weight of evidence for assessing chemical toxicity.[11,12] Structural alerts are molecular substructures that are associated with a particular adverse outcome.[13] Read-across is a technique that proposes to identify potential hazards of untested compounds by associating them with structurally similar compounds that have been tested.[14] QSAR modeling is a computational approach that employs statistical or machine learning techniques to establish correlations between intrinsic chemical properties (chemical descriptors) and measured properties or toxicological effects.[15] QSAR modeling has been used extensively to model and predict chemical toxicity, and best practices for model development and validation have been developed to ensure their reliability.[16] Both structural alerts and read across approaches have been preferred by regulators due to ease of use, transparency, and mechanistic interpretability. However, there have been concerns that these tools often do not help with reliable assessment of whether the underlying compounds present true hazard to humans and the environments. For instance, we have previously demonstrated that alerts have a tendency to flag compounds as toxic even when the experimental evidence shows otherwise.[17]

In the last several years, both our[18–20] and other[21,22] groups have developed reliable computational models for predicting skin sensitization potential of chemicals. These and other models developed for one or more of the 6-pack endpoints are summarized in Table 1, which indicate that the development of reliable computational models for predicting the outcomes of all 6-pack tests is still a significant challenge. Herein, we have compiled, integrated, and curated the most comprehensive collection of experimental in vivo data on 6-pack endpoints. We especially emphasize, with vivid examples, the importance and impact of data curation on the rigor of the study design and reliability of its outcomes. We have developed and rigorously validated QSAR models for all 6-pack assays and demonstrated their utility in identifying potentially safe or unsafe chemicals in industrial products (Figure 1). We integrated these models into a software package called STopTox (Systemic and Topical chemical Toxicity (STopTox) and made it publicly available to the research community via a dedicated web portal (https://stoptox.mml.unc.edu/).

Results and Discussion

Data Curation and Cross-endpoint concordance analysis

Although predictive models have been developed and reported previously for subsets of the 6-pack endpoints (Table 1), many of these models did not fully comply with the model validation guidelines specified by the Organization for Economic Cooperation and Development (OECD)[23] and, most notably, lacked proper data curation. As can be seen in Figure 2, applying data curation protocols decreased the size of the available data by more than 50%. Our final database represented a sparse matrix containing 11,941 compounds with activity measurements for at least one of the 6-pack endpoints. Only 12 compounds were tested in all six endpoints. This low number can be due to waivers granted by regulators in an effort to decrease animal use in toxicity assessment.[24]

Figure 3 shows a heat map of the pairwise concordance and overlap (in parenthesis) between compounds tested in different endpoints. We have identified 328 compounds overlapping between acute oral toxicity and acute inhalation toxicity. These two endpoints showed the highest (80%) pairwise concordance among all six endpoints. A previous study[25] analyzed the binary concordance between the acute systemic toxicity tests in “pesticide” and “chemical” sets. The oral-inhalation overlap for “pesticides” was 328 compounds with concordance as low as 24%, while for “chemicals” (71 compounds) the concordance was 72%. The oral-dermal overlap for pesticides (307 compounds) demonstrated a concordance of 71%. The oral-dermal overlap for “chemicals” (1569 compounds) resulted in high concordance of 93%. Our concordance analysis also showed an oral-dermal concordance of 71% estimated on the overlap of 1308 compounds. Among compounds that were “Not Classified” in acute oral tests, 98% were also “Not Classified” in acute dermal tests. While the oral route is mandatory for acute toxicity testing under most regulatory frameworks for food-use pesticides, selection of a second route of exposure (dermal or inhalation) is based on the amount of a chemical being commercialized, taking into account the inherent acute toxicity of the chemical and the primary route of exposure when handling the material. [26] Computational models built on curated datasets and following the best practices for development and validation of QSAR models can support the decision of waiving dermal tests in case the substance has acute oral toxicity profile, since there is still a small possibility for the chemical to be toxic upon skin contact.[27]

Skin irritation and eye irritation showed the same binary outcomes in 60% of 380 overlapping compounds. A concordance of 67% has been previously observed for 205 formulations tested for both skin irritation and eye irritation.[28] This concordance was higher for "Not Classified" compounds, whereas 89% of formulations classified as non-irritant for eye irritation were also non-irritants for skin irritation. Chemicals that were considered corrosive or severely irritant to the skin or highly toxic (category 1 or 2) in the dermal test, do not need to be tested for eye irritation and corrosion, according to the OECD guidance on considerations for waiving mammalian acute toxicity tests.[24] Also, skin sensitization and skin irritation datasets showed the same binary outcomes in 72% of 237 overlapping compounds.

QSAR modeling

Previously, we have built models to predict skin sensitization endpoints using a combination of animal,[29] OECD validated *in vitro* assays,[18] and human data.[18,20,30] In this study, we have developed consensus models for predicting outcomes of skin sensitization testing in animals as STopTox is intended as a reliable alternative to animal testing in the 6-pack assays. Therefore, all the models reported here were built using only data collected from animal tests that followed the OECD protocols. Statistical characteristics of QSAR models developed in this study are summarized in Table 2. All cross-validated models showed high predictive accuracy on independent external evaluation sets based on several metrics including correct classification rate (CCR), sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). The acute toxicity models showed CCRs ranging from 70% to 78%; SEs ranging from 67% to 79%; SPs ranging from 68% to 81%; PPVs ranging from 71% to 79%; and NPVs ranging from 70% to 78%.

Comparative assessment of new "6-pack" models versus alternative tools

Over the decades, many computational tools have been built to evaluate adverse health effects of chemical compounds from their chemical structure.[31,32] Predictive models, such as quantitative structure-activity relationship (QSAR), facilitate replacement and reduction of animal testing in toxicology and risk assessment; however, many subject-matter experts acknowledge that "these methods are not always reliable and must be assessed on their individual merit for the compound and context in question".[33] For this and other reasons catalogued by Dearden et al.[34], results derived from QSAR and other *in silico* models are usually met with caution. Model predictions are almost always used in combination with other evidence, or only for the purpose of the initial screening/ranking of compounds for further testing.[35]

The curation of the ECHA database proved to be an extremely laborious task and the most time-consuming part of this work. It is important to emphasize that much of 6 pack endpoint data included in the ECHA database should not be used for model development. As seen from the summary of data curation (Figure 2), the major reduction in the size of individual datasets used eventually for QSAR model development was due to a large fraction of inconsistent data in the original ECHA database which, upon careful inspection, were found to be reported as predictions made by QSAR models, or read across, or expert systems, or labeled as "not reliable".

Comparison to models developed for the same endpoints without rigorous data curation[36] suggests that our extensive data curation procedures resulted in the net decrease in both the dataset sizes and model performance. Indeed, we compared the models produced in this study to those reported by Luechtefeld et al.[36] who described the development of a suite of *in silico* models, termed read-across structure activity relationships (RASAR) for the 6-pack endpoints. Since the model predictions based by RASAR can only be accessed through a fee-based commercial platform (https://www.ulreachacross.com), we have performed an indirect comparison of statistics (see Table S1 in the supplementary materials). Our models showed on average, a 10% lower CCR. Furthermore, the amount of data reported in the study mentioned above[36] was, on average, five times larger than the size of the carefully curated dataset used in this study. Previously, we already expressed concerns that high accuracy of models as reported[36] could be the consequence of inadequate data curation leaving many duplicate compounds in the modeling datasets.[37] We posit that our results more accurately reflect the actual model performance for these endpoints since we have eliminated such confounders as duplicate entries or the use of predicted or "not reliable" values and conducted more rigorous validation procedures according to the established guidelines. We strongly suggest that this exercise reemphasizes the importance of proper data curation and cautions against overinterpreting results from models built on non-curated datasets.

STopTox usability and interpretation

The limitations of the experimental assay results used in model development define the usability and interpretation of STopTox. It is important to note that the binary compound annotations in all 6-pack assays as toxic or non-toxic are derived from dose-dependent testing conditions. For instance, the skin sensitization potencies for substances are based on a function of lymph node cell proliferation induced by the test chemical and expressed as a stimulation index (SI) relative to values obtained with concurrent controls. If SI ≥ 3, the substance is considered a sensitizer in the tested concentration. The skin irritation criteria are based on a single dose of a chemical applied to the skin of an experimental animal. The degree of irritation/corrosion is based on an erythema/edema scale system that is subjectively scored at specified observation intervals. Similarly, eye irritation is based on subjectively scoring lesions of conjunctiva, cornea, and iris, at specific intervals after application of a single dose of the test substance. If the effects are reversible after 21 days, the chemical is considered an irritant, but if it does not reverse (or if a severe score is noted at any timepoint), the chemical is labeled as corrosive. The tests for systemic endpoints in STopTox (Acute Dermal Toxicity, Acute Inhalation Toxicity, and Acute Oral Toxicity) rely on LD_{50} data, meaning the dose of a substance required to kill half of the experimental animals. The Globally Harmonized System of classification and labelling of chemicals (GHS)[38] threshold for Acute Oral and Dermal toxicity for chemicals is $LD_{50} < 2000$ mg/kg bodyweight. As for Acute Inhalation toxicity, the threshold varies with the form of tested substance (gases – $LC_{50} \leq 2500$ ppm, vapors – $LC_{50} \leq 10$ mg/L, and dusts/mists – $LC_{50} \leq 20$ mg/L).

It is essential to note that, if the model predicts a compound as toxic or non-toxic, such prediction should be considered only in the context of specific dose-dependent observation for each assay; obviously, increasing the dose of any compound in any assay could often lead to toxic effects. These considerations are often overlooked when making predictions or assertions concerning the expected chemical toxicity. The ultimate goal of any method for evaluating acute toxicity is to provide an accurate assessment of the potential risk of a chemical concerning human safety.[39] Therefore, we reinforce that the limitation of assays should influence the interpretation of the predictions made by the models and how these models can be used by toxicologists to help in decision making. Prediction with QSAR models implemented in STopTox (and, actually, with any models) do not take the dose into account; they merely state whether a chemical is predicted toxic or non-toxic in each assay. Thus, users interpreting these predictions should always be familiar with and keep in mind the underlying experimental conditions briefly discussed above under which compounds in the training sets have been annotated as toxic or non-toxic. Further, these models are limited to binary hazard-based predictions, rather than providing information on potency and GHS or EPA subcategorization. Thus, they are not directly applicable for many regulatory classifications and labeling requirements requiring higher level of granularity. However, these models are well suited to assist in hazard assessment and chemical screening/prioritization, and, because of high accuracy in terms of both sensitivity and specificity, they are especially useful in identifying non-toxic compounds (tested in the same conditions as those identified as toxic where additional subcategorization is indeed important).

Predictions of known toxicants

For additional external validation of our models, we have conducted a literature search for toxicants described in clinical studies or known toxicants for each endpoint that was absent in our database. A list of 45 potential skin sensitizers in cosmetic ingredients was compiled by the Norwegian Scientific Committee for Food Safety.[40] Eleven out of 45 compounds were absent in our skin sensitization dataset and eight of the eleven chemicals were correctly predicted as sensitizers by our skin sensitization model (SE = 72%).

MS-222, a fish anesthetic commonly used in aquaculture,[41] and sodium lauryl sulfate,[42] a product commonly used in personal care products, are both known skin irritants that were not present in our skin irritation training data. Our models predicted sodium lauryl sulfate as a skin irritant and MS-222 as not classified (according to OECD protocols, chemicals not classified as skin irritants are considered “not classified”). There were three compounds that were not present in our eye irritation model: glutaraldehyde, glyphosate, and Paraquat (1,1'-Dimethyl-4,4'-bipyridinium dichloride). Our model predicted glutaraldehyde and glyphosate as eye irritants. Exposure of glutaraldehyde during cataract surgery led to the development of toxic eye anterior segment syndrome in six patients.[43] Ocular glyphosate exposure was reported to lead to the development of chemosis, heart palpitations, raised blood pressure, headache, and nausea.[44] In two cases of accidental eye exposure to Paraquat, eye damage was reported.[45]

Dichloromethane[46] and methanol[47] have been reported as systemic toxicants after dermal exposure and were not in the modeling set. Our acute dermal toxicity model correctly predicted both compounds as toxic after dermal exposure. Twenty chemicals commonly present in occupational inhalation accidents were compiled in a study.[48] There were five organic chemicals in this list that were absent in our acute inhalation dataset. All five compounds were correctly predicted by acute inhalation models. One clinical case of accidental oral exposure to the pyrethroid deltamethrin led to the poisoning of a 4-year old girl that consumed insecticidal chalk and was found unconscious 20 minutes after going outside to play.[49] Mephedrone, a psychoactive drug, has been proven toxic in a study reporting cases of acute toxicity related to self-reported use of mephedrone.[50] Our acute oral toxicity model predicted both compounds as toxic if swallowed. Together, our models correctly predicted 80% of the known toxicants compiled (Figure 4).

Figure 5 shows the predictions generated for N-phenyl-*p*-phenylenediamine, a known skin sensitizer that is usually added to temporary black henna tattoos leading to many cases of contact allergy.[51] We also generated maps showing the relative significance of fragment contributions, providing a graphical interpretation of developed models (Figure 5), where atoms and structural fragments enhancing toxicity are highlighted in pink and those decreasing toxicity are shown in green. These maps are generated for each six-pack endpoints independently.

Virtual screening of CosIng, REACH and all STopTox compounds

As a case study illustrating STopTox usability, we have applied our QSAR models to the European Commission Cosmetic ingredient database (CosIng), REACH, and to all STopTox compounds. The majority of compounds in each of these datasets were predicted as Not Classified by each individual model. In the CosIng dataset (n = 3,930 compounds), 1655 compounds were predicted as skin sensitizers, 1400 compounds were predicted as skin irritants, 2177 compounds were predicted as eye irritants, 551 compounds were predicted as toxic if swallowed, 407 compounds were predicted as toxic if inhaled, and 372 compounds were predicted as toxic after dermal exposure. Out of N=XX total compounds, there were 3398 compounds predicted as toxic in at least one endpoint and 1381 compounds predicted as Not Classified in all six endpoints. All compounds and corresponding predictions are listed in Supplementary Table S2.

In the REACH dataset (n = 10,465 compounds), 4,018 compounds were predicted as skin sensitizers, 2,445 compounds were predicted as skin irritants, 4,605 compounds were predicted as eye irritants, 2,679 compounds were predicted as toxic if swallowed, 2,139 compounds were predicted as toxic if inhaled, and 1,899 compounds were predicted as toxic after dermal exposure. There were 7,641 compounds predicted as toxic in at least one endpoint and 2,824 compounds predicted as Not Classified in all six endpoints. All compounds and corresponding predictions are listed in Supplementary Table S3.

In the STopTox dataset (n = 11,941 compounds), 4,792 compounds were predicted as skin sensitizers, 2,491 compounds were predicted as skin irritants, 4,766 compounds were predicted as eye irritants, 5,232 compounds were predicted as toxic if swallowed, 2,394 compounds were predicted as toxic if inhaled, and 2,902 compounds were predicted as toxic after dermal exposure. There were 7,641 compounds predicted as toxic in at least one endpoint and 2,824 compounds predicted as Not Classified in all six endpoints. All compounds and corresponding predictions are listed in Supplementary Table S4.

Model implementation

The STopTox web-based application (Figure 6) runs machine learning routines written in Python by using Flask v. 0.12.2, a small framework for creating web microframeworks at the backend. Models were developed using Scikit-Learn. Angular 4 and Typescript were used for the development of the frontend, and Docker and Docker-Compose for the orchestration of containers. The developed models and all datasets are publicly available at https://stoptox.mml.unc.edu/.

Conclusions

We have developed STopTox, a comprehensive collection of computational models that can be used as an alternative to *in vivo* 6-pack tests for predicting chemical toxicity hazard. Models were established following the best practices for the development and validation of QSAR models[16,23] using the largest publicly available and *carefully curated* datasets that we compiled for all 6-pack assays. To the best of our knowledge, STopTox is the first publicly available portal that enables accurate predictions of chemical hazards in all the 6-pack endpoints at once. Despite the limitations of these models with respect to potency classes, they are reliable for predicting chemicals that do not require classification, i.e., those expected to be nontoxic if tested following the same protocols used for compounds in the modeling set. We suggest that these models are valuable for both regulatory agencies and respective industries in helping them identify safer alternatives to chemicals of interest. We reinforce that, in order to build predictive models, it is not enough just to utilize adequate chemical descriptors and powerful statistical techniques;[52] we shall stress that, STopTox is the only 6-pack endpoint predictor in the public domain developed with extensively curated data. The STopTox web app provides users with the access to statistically significant and externally predictive QSAR models of acute toxicity tests. The web app can be used for rapid evaluation of acute toxicity hazards in chemical inventories. STopTox is freely available at https://stoptox.mml.unc.edu/. To the best of our knowledge, STopTox does not have analogs in terms of the level of data curation, validated statistical accuracy of constituting models, transparency of the data, modeling methods and software tools, and public accessibility.

Materials and Methods

Data collection

We compiled data from multiple publicly available databases, such as the REACH data, available at the ECHA (European Chemical Agency) dissemination portal, eChemPortal (https://www.echemportal.org/echemportal/), the Toxicology Data Network (TOXNET - https://toxnet.nlm.nih.gov/), the National Toxicology Program´s Integrated Chemical

Environmental (ICE) resource (https://ice.ntp.niehs.nih.gov/), the Environmental Protection Agency (EPA)´s database ToxValDB (https://comptox.epa.gov/), and from the scientific literature.[27,53–61] These data encompass animal sources of the experimental tests for the 6-pack endpoints: (i) skin sensitization; (ii) skin irritation and corrosion; (iii) eye irritation and corrosion; and acute systemic toxicity via (iv) dermal, (v) inhalation, and (vi) oral routes. Unfortunately, there were numerous problems with the collected raw data. For instance, many numerical data were represented as string variables, the units of measurements were not standardized through the datasets, and there were many “free text” data. Therefore, we extensively cleaned and standardized all the data and converted measurements to the same units in each dataset. We also used regex expressions to find important features for the database that were described in text format; this was key to endpoint classification into GHS hazard classes. Following this laborious data preparation and standardization, we performed both chemical and biological data curation. This attention to detailed data curation at different levels of the data preparation protocol is, unfortunately, uncommon in computational chemical toxicology, as we noted previously.[37]

Data curation

Datasets were thoroughly curated following the workflows developed by us earlier.[52] First, we excluded inconsistent data, which significantly decreased the size (number of data points) of our datasets (Figure 2). Data were categorized as inconsistent if they were generated not following the OECD protocols, if compounds were tested in few or only one concentration and could not be classified into GHS classes, labeled as non-experimental (*e.g.*, labeled as obtained using QSAR and/or read across predictions and/or weight of evidence decisions), data with measurements different from the standard protocols for the 6-pack endpoints (e.g., NOAEL, NOEL, LC_{10}, LC_0, *etc*.), data without identification number (CAS number, EC number), formulations, and complex mixtures. Then, we performed chemical structure curation and removed mixtures, inorganics, and organometallic compounds, cleaned and neutralized salts, normalized the specific chemotypes, and applied special treatment to chemicals with multiple replicated records as follows: (i) when replicated records presented the same binary outcome, only one record was kept; (ii) when majority of replicated records presented the same binary outcome and one had different binary outcome, only one record with the agreeing binary outcome was kept, (iii) when replicated records had different binary outcomes, all of them were removed. All the curated data will be available in Supplementary Table S5.

Datasets

Skin sensitization (dataset A)

Skin sensitization data were compiled from two sources: (i) National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods on behalf of ICCVAM (http://ntp.niehs.nih.gov/go/40500) and (ii) the REACH study results database publicly available at https://iuclid6.echa.europa.eu/reach-study-results. The original ICCVAM database included 1,060 chemical records with local lymph node assay (LLNA) data. After curation, 516 unique compounds (332 sensitizers and 184 non-sensitizers) were retained. The REACH database initially comprised 10,588 records for 9,801 chemicals. The REACH dataset is composed of many types of assays and study categories. *In vitro* and “weight of evidence” categories were discarded. Data from different OECD (Organization for Economic Co-operation and Development) skin sensitization assays (OECD guidelines 406, 411, 429 and 442B)[62–65] were available; only the data corresponding to LLNA assays (429 and 442B) were selected, resulting in 1,275 data points with LLNA records. After curation, 566 compounds (197 sensitizers and 369 non-sensitizers) were retained. Eventually, we merged the curated data from ICCVAM and REACH and examined the content of this combined data. There were 58 pairs of duplicate chemicals between these two datasets, and the sensitization potential of five of these pairs was different. These discordant records were removed, and only one record for each concordant pair of duplicates was kept. The merged dataset had 1,000 unique compounds (481 sensitizers and 519 non-sensitizers).

Skin irritation and corrosion (dataset B)

Experimental animal data on skin irritation and corrosion were retrieved from the REACH study results database (https://iuclid6.echa.europa.eu/reach-study-results). After removing inconsistent data, 1,631 out of original 5,274 data points were left. After removal of mixtures, inorganics, and neutralization and removal of Not Classified counter-ions, 1,326 records remained. Among 124 duplicate chemicals in the dataset, 95 were concordant and 29 were discordant. All the discordant replicates and one of each concordant replicate were removed. The final dataset has 1,012 unique chemical compounds including 40 corrosives, 277 irritants, and 695 non-irritants. As there were only few corrosive compounds in our dataset, we decided to merge the corrosive and irritant classes and model only irritant versus non-irritant compounds. We note that these models have limited regulatory value at the moment given the data available with respect to compounds predicted as toxic as regulators typically would like to see more granular measurement or prediction at the level of specific subcategories of toxicity. However, we shall highlight and emphasize that our models make accurate predictions of non-toxic compounds thereby helping both regulators and respective regulated industries to develop safer chemicals. Our resulting dataset contains 317 irritants versus 695 non-irritants. Because the dataset is imbalanced, we applied an undersampling technique where the majority class was sampled in a way to match the number of records of the minority class. This was done by searching for the compounds in the majority class that had higher similarity (Tanimoto coefficient) with compounds in the minority class. The balanced dataset consisted of 554 compounds (277 irritants and 277 non-irritants).

Eye irritation and corrosion (dataset C)

The eye irritation and corrosion dataset was retrieved from the REACH study results database (https://iuclid6.echa.europa.eu/reach-study-results) and the literature[53–61]. We first curated data from each source separately, then we merged the curated datasets and checked for overlapping compounds. After removing inconsistent data, 7,196 out of original 7,332 experimental animal data points for eye irritation and corrosion were left. After removal of mixtures, inorganics, and neutralization and removal of Not Classified counter-ions, 5,985 records remained. All the discordant chemical replicates and one of each concordant replicate were removed. The final dataset has 3,547 unique chemical compounds including 1,146 irritants, and 2,401 non-irritants. Because the dataset is imbalanced, we applied an undersampling technique where the majority class was sampled in a way to match the number of records of the minority class. This was done by searching for the compounds in the majority class that had higher similarity (Tanimoto coefficient) with compounds in the minority class. The balanced dataset consisted of 2,292 compounds (1,146 skin irritants and 1,146 skin non-irritants).

Acute dermal toxicity (dataset D)

The acute dermal toxicity dataset was retrieved from the REACH study results database (https://iuclid6.echa.europa.eu/reach-study-results), the publicly available database ToxValDB, and from a literature dataset.[27] After removing inconsistent data, 5,259 out of original 29,824 data points were left; the major reasons for compound removal were the presence of many compounds without a defined LD50. After removal of mixtures, inorganics, organometallic compounds, 4,601 records remained. Among 1,979 chemical replicates in the dataset, 1,836 were concordant and 143 were discordant. All the discordant replicates and one of each concordant replicate were removed. The final dataset has 2,622 unique chemical compounds including 382 dermally toxic compounds and 2,234 Not Classified compounds. Because the dataset is imbalanced, we applied an undersampling technique where the majority class was sampled in a way to match the number of records of the minority class. This was done by searching for the compounds in the majority class that had higher similarity (Tanimoto coefficient) with compounds in the minority class. The balanced dataset consisted of 764 compounds including 382 toxic compounds and 382 Not Classified compounds.

Acute inhalation toxicity (dataset E)

The acute inhalation toxicity dataset was retrieved from the REACH study results database (https://iuclid6.echa.europa.eu/reach-study-results) and from the publicly available database ToxValDB. After removing inconsistent data, 2,061 out of original 8,176 data points were left only. This dramatic reduction of the dataset was mainly due to the presence of many compounds without a defined LD50 and because of the absence of information regarding the exposure method used (gas, dust or mist), which is essential for GHS classification.   After removal of mixtures, inorganics, and neutralization and removal of Not Classified counter-ions, 1,637 records remained. Among 527 chemical replicates in the dataset, 501 were concordant and 26 were discordant. All the discordant replicates and one of each concordant replicate were removed. The final dataset has 681 unique chemical compounds including 345 toxic compounds and 336 Not Classified compounds.

Acute oral toxicity (dataset F)

The acute oral toxicity dataset was retrieved from NICEATM (The National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods) workshop for the Collaborative Acute Toxicity Modeling Suite (CATMoS)[66] project that our team was part of. After removing inconsistent data, 8,981 out of original 8,994 data points were left. After removal of mixtures, inorganics, and neutralization and removal of counter-ions, 8,979 records remained. A total of 406 chemical replicates were found in the dataset. All replicates with different toxicity calls were removed and one of the duplicative compounds with concordant toxicity calls was kept. The final dataset has 8,442 unique chemical compounds including 4,803 toxic compounds and 3,639 Not Classified compounds.

CosIng (Dataset g)

CosIng is the European Commission database for information on cosmetics substances and ingredients (https://ec.europa.eu/growth/sectors/cosmetics/cosing_en). This dataset contained 5166 chemical records with a defined chemical structure. After curation, 3,930 unique chemical substances were kept for prediction purposes.

REACH (Dataset e)

The REACH data come from registration dossiers submitted to ECHA by May 2019 (https://echa.europa.eu/information-on-chemicals/registered-substances). The database originally contains 20,000 substances, of which 15,438 are chemical records with a defined chemical structure. After curation, 10,465 unique chemical substances were kept for prediction purposes.

Cheminformatics approaches

Binary QSAR models were developed and rigorously validated according to the best practices of QSAR modeling[16]. For each dataset corresponding to each 6-pack endpoint, the three types of descriptors (Morgan[67], MACCS[68], and MORDRED[69]) in combination with random forest[70] algorithm were used. A consensus model for each endpoint was built by averaging the predicted values from each different descriptor model. The consensus model for each 6-pack endpoint considered majority rule (at least two out of three) for the final classification. The consensus AD (Consensus with Applicability Domain) model was developed in a similar way; however, only predictions from individual models when the queried compound was inside the applicability domain (*z*-cutoff method[15]) were considered. In every case, only the modeling set was used to develop the models, while the external sets were used for evaluation of their predictive power. Here we followed a true 5-fold external cross-validation procedure: the full set of compounds with known experimental activity is randomly divided into five subsets of equal size; then one of these subsets (20% of all compounds) is set aside as test set and the remaining four sets together form the modeling set (80% of the full set). This procedure is repeated five times allowing each of the five subsets to be used as a test set. Models are built using the modeling set only, and it is important to emphasize that the test set compounds are never employed either to build and/or select the models. Each modeling set is divided into many internal training and validation sets; then models are built using compounds of each training set and applied to test set compounds to assess their properties.[71] In addition, 10 rounds of Y-randomization were performed for each dataset to assure that the model performance were not due to chance correlations.

Acknowledgements

This study was supported in part by NIH (grant 1R43ES032371) and CNPq (grant 400760/2014-2). JB thank CNPq and Science without Borders program for the financial support of her visit to the University of North Carolina at Chapel Hill. VA thanks the Lush Prize.

Author Contributions

Each author has contributed significantly to this work. JVBB, VMA, CHA, EM and AT conceived and designed the experiment. JVBB, KO, ACS, SUSH, and EO performed the computational experiments. JVBB, VMA, NK, JS, DA, CHA, EM, and AT analyzed the data. RB and DK implemented the models. JVBB, VMA, and EM wrote the paper. All authors read, edited and approved the final manuscript.

Competing Interests

AT and EM are co-founders of Predictive, LLC, that develops computational methodologies and software for toxicity prediction. All other authors declare no conflicts.

References

1. National Research Council (US) Commitee on Animals as Monitors of Environmental Hazards. *Animals as Sentinels of Environmental Health Hazards*. *The National Academies press* (National Academies Press, 1991). doi:10.17226/1351
2. Flecknell, P. Replacement, reduction and refinement. *ALTEX* **19**, 73–8 (2002).
3. U.S. Environmental Protection Agency. Test Guidelines for Pesticide Data Requirements.
4. European Commission. Ban on Animal Testing. *European Commission: Internal Market, Industry, Entrepreneurship and SMEs.*

5. Institute for In Vitro Sciences. Modernization of EPA '6-Pack' of Acute Toxicity Tests Continues. (2018).

6. Report from the Commission to the European Parliament and the Council. 2019 report on the statistics on the use of animals for scientific purposes in the Member States of the European Union in 2015-2017. (2020).

7. Hartung, T. Look back in anger - what clinical studies tell us about preclinical work. *ALTEX* **30**, 275–91 (2013).

8. Bailey, J., Thew, M. & Balls, M. An analysis of the use of animal models in predicting human toxicology and drug safety. *Altern. Lab. Anim.* **42**, 181–99 (2014).

9. Hartung, T. Making big sense from big data in toxicology by read-across. *ALTEX* **33**, 83–93 (2016).

10. US EPA struggles to replace animal tests for pesticide toxicity. *C&EN Glob. Enterp.* **97**, 24–25 (2019).

11. U.S. Environmental Protection Agency. Process for evaluation and implementing alternative approaches to traditional in vivo acute toxicity studies for FIFRA regulatory use. *Office of Pesticide Programs* (2016).

12. ECHA - European Chemicals Agency. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint specific guidance. (2017).

13. Blagg, J. Structural Alerts for Toxicity. in *Burger's Medicinal Chemistry and Drug Discovery* 301–334 (John Wiley & Sons, Inc., 2010). doi:10.1002/0471266949.bmc128

14. Ball, N. *et al.* Toward Good Read-Across Practice (GRAP) guidance. *ALTEX* **33**, 149–166 (2016).

15. Tropsha, A. & Golbraikh, A. Predictive QSAR modeling workflow, model applicability

domains, and virtual screening. *Curr. Pharm. Des.* **13**, 3494–3504 (2007).

16. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **29**, 476–488 (2010).

17. Alves, V. M. *et al.* Alarms about structural alerts. *Green Chem.* **18**, 4348–4360 (2016).

18. Alves, V. M. *et al.* A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustain. Chem. Eng.* **6**, 2845–2859 (2018).

19. Braga, R. C. *et al.* Pred-Skin: A Fast and Reliable Web Application to Assess Skin Sensitization Effect of Chemicals. *J. Chem. Inf. Model.* **57**, 1013–1017 (2017).

20. Borba, J. V. B. *et al.* Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* acs.chemrestox.0c00186 (2020). doi:10.1021/acs.chemrestox.0c00186

21. Roberts, D. W., Aptula, A. & Api, A. M. Structure–Potency Relationships for Epoxides in Allergic Contact Dermatitis. *Chem. Res. Toxicol.* **30**, 524–531 (2017).

22. Toropova, A. P. & Toropov, A. A. Hybrid optimal descriptors as a tool to predict skin sensitization in accordance to OECD principles. *Toxicol. Lett.* **275**, 57–66 (2017).

23. OECD. OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models. 1–2 (2004).

24. U.S. Environmental Protection Agency. Guidance for Waiving or Bridging of Mammalian Acute Toxicity Tests for Pesticides and Pesticide Products (Acute Oral, Acute Dermal, Acute Inhalation, Primary Eye, Primary Dermal, and Dermal Sensitization). *Office of Pesticide Programs* (2012).

25. Seidle, T. Examining the regulatory value of multi-route mammalian acute systemic toxicity studies. *ALTEX* **28**, 95–102 (2011).

26. Seidle, T. *et al.* Cross-Sector Review of Drivers and Available 3Rs Approaches for Acute Systemic Toxicity Testing. *Toxicol. Sci.* **116**, 382–396 (2010).

27. Creton, S. *et al.* Acute toxicity testing of chemicals—Opportunities to avoid redundant testing and use alternative approaches. *Crit. Rev. Toxicol.* **40**, 50–83 (2010).

28. Corvaro, M. *et al.* A retrospective analysis of in vivo eye irritation, skin irritation and skin sensitisation studies with agrochemical formulations: Setting the scene for development of alternative strategies. *Regul. Toxicol. Pharmacol.* **89**, 131–147 (2017).

29. Alves, V. M. *et al.* Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. *Toxicol. Appl. Pharmacol.* **284**, 273–280 (2015).

30. Alves, V. M. *et al.* QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. *Green Chem.* **18**, 6501–6515 (2016).

31. Barratt, M. D. Prediction of toxicity from chemical structure. *Cell Biol. Toxicol.* **16**, 1–13 (2000).

32. Myatt, G. J. *et al.* In silico toxicology protocols. *Regul. Toxicol. Pharmacol.* **96**, 1–17 (2018).

33. Cronin, M. T. D. *et al.* In Silico Prediction of Organ Level Toxicity: Linking Chemistry to Adverse Effects. *Toxicol. Res.* **33**, 173–182 (2017).

34. Dearden, J. C., Cronin, M. T. D. & Kaiser, K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **20**, 241–266 (2009).

35. Hansson, S. O. & Rudén, C. Priority Setting in the REACH System. *Toxicol. Sci.* **90**, 304–308 (2006).

36. 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.* **165**, 198–212 (2018).

37. Alves, V. M. *et al.* Oy Vey! A Comment on "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships Outperforming Animal Test Reproducibility." *Toxicol. Sci.* **167**, 3–4 (2019).

38. United Nations iLibrary | Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Available at: https://www.un-ilibrary.org/transportation-and-public-safety/globally-harmonized-system-of-classification-and-labelling-of-chemicals-ghs_a6a925c3-en. (Accessed: 12th November 2020)

39. Basketter, D. A., White, I. R., McFadden, J. P. & Kimber, I. Skin sensitization: Implications for integration of clinical data into hazard identification and risk assessment. *Hum. Exp. Toxicol.* **34**, 1222–1230 (2015).

40. Paulsen, J. E. *et al. Sensitisation caused by exposure to cosmetic products. Opinion of the Panel on Food Additives, Flavourings, Processing Aids, Materials in Contact with Food and Cosmetics of the Norwegian Scientific Committee for Food Safety.* (2009).

41. Park, I.-S. The Anesthetic Effects of Clove Oil and MS-222 on Far Eastern Catfish, Silurus asotus. *Dev. Reprod.* **23**, 183–191 (2019).

42. De Jongh, C. M. *et al.* Stratum corneum cytokines and skin irritation response to sodium lauryl sulfate. *Contact Dermatitis* **54**, 325–333 (2006).

43. 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.* **32**, 1696–1701 (2006).

44. Bradberry, S. M., Proudfoot, A. T. & Vale, J. A. Glyphosate Poisoning. *Toxicol. Rev.* **23**, 159–167 (2004).

45. Joyce, M. Ocular damage caused by Paraquat. *Br. J. Ophthalmol.* **53**, 688–690 (1969).

46. Pacheco, C., Magalhães, R., Fonseca, M., Silveira, P. & Brandão, I. Accidental intoxication by dichloromethane at work place: Clinical case and literature review. *J. Acute Med.* **6**, 43–45 (2016).

47. Kahn, A. & Blum, D. Methyl alcohol poisoning in an 8-month-old boy: An unusual route of intoxication. *J. Pediatr.* **94**, 841–843 (1979).

48. Miller, K. & Chang, A. Acute inhalation injury. *Emerg. Med. Clin. North Am.* **21**, 533–557 (2003).

49. O'Malley, M. Clinical evaluation of pesticide exposure and poisonings. *Lancet* **349**, 1161–1166 (1997).

50. Wood, D. M. *et al.* Case series of individuals with analytically confirmed acute mephedrone toxicity. *Clin. Toxicol.* **48**, 924–927 (2010).

51. Panfili, E., Esposito, S. & Di Cara, G. Temporary Black Henna Tattoos and Sensitization to para-Phenylenediamine (PPD): Two Paediatric Case Reports and a Review of the Literature. *Int. J. Environ. Res. Public Health* **14**, 421 (2017).

52. Fourches, D., Muratov, E. & Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **56**, 1243–1252 (2016).

53. Verheyen, G. R., 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.* **265**, 47–52 (2017).

54. Geerts, L. *et al.* CON4EI: Evaluation of QSAR models for hazard identification and

labelling of eye irritating chemicals. *Toxicol. Vitr.* 0–1 (2017). doi:10.1016/j.tiv.2017.09.004

55. Barroso, J. *et al.* 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.* **91**, 521–547 (2017).

56. Basant, N., Gupta, S. & Singh, K. P. A three-tier QSAR modeling strategy for estimating eye irritation potential of diverse chemicals in rabbit for regulatory purposes. *Regul. Toxicol. Pharmacol.* **77**, 282–291 (2016).

57. Verma, R. P. & Matthews, E. J. An in silico expert system for the identification of eye irritants. *SAR QSAR Environ. Res.* **26**, 383–395 (2015).

58. 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.* **68**, 1555–1572 (2006).

59. Barratt, M. A quantitative structure-activity relationship for the eye irritation potential of neutral organic chemicals. *Toxicol. Lett.* **80**, 69–74 (1995).

60. Barratt, M. D. QSARS for the eye irritation potential of neutral organic chemicals. *Toxicol. Vitr.* **11**, 1–8 (1997).

61. Adriaens, E. *et al.* CON4EI: Selection of the reference chemicals for hazard identification and labelling of eye irritating chemicals. *Toxicol. Vitr.* **44**, 44–48 (2017).

62. OECD. Test No. 406: Skin Sensitisation. (1992). doi:10.1787/9789264070660-en

63. OECD. Test No. 411: Subchronic Dermal Toxicity: 90-day Study. (1981). doi:10.1787/9789264070769-en

64. OECD. *Test No. 429: Skin Sensitisation*. (OECD Publishing, 2010). doi:10.1787/9789264071100-en

65. OECD. *Test No. 442B: Skin Sensitization*. **412**, (OECD, 2018).

66. Kleinstreuer, N. C. *et al.* Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation. *Comput. Toxicol.* **8**, 21–24 (2018).

67. Aslett, M. *et al.* TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38**, D457–D462 (2010).

68. Anderson, S. Graphical representation of molecules and substructure-search queries in MACCStm. *J. Mol. Graph.* **2**, 83–90 (1984).

69. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).

70. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

71. Zhu, H. *et al.* Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **48**, 766–784 (2008).

Figure Legends:

Figure 1. General workflow of STopTox.

Figure 2. Summary of data curation.

Figure 3. Cross-endpoint pairwise concordance of binary outcomes and number of overlapping compounds (in parentheses).

Figure 4. Toxicants identified in the literature that were absent in our modeling data. StopTox predictions for a given endpoint are listed below each structure. “Not classified” compounds were predicted as non-toxic.

Figure 5. Predicted probability maps and predictions of each model for N-phenyl-*p*-phenylenediamine. The predicted probability of a toxic effect is accompanied by the colored map of the atomic contributions to toxicity. Red: predicted to increase toxicity; Green: predicted to decrease toxicity.

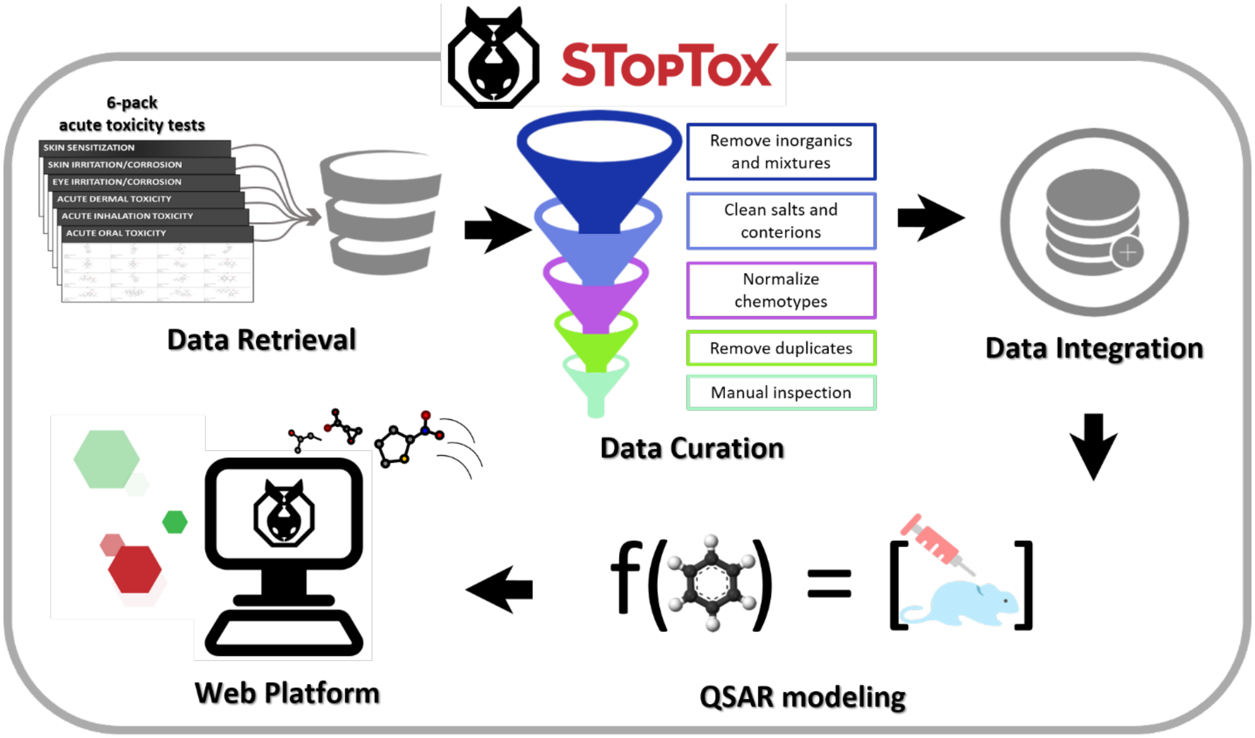
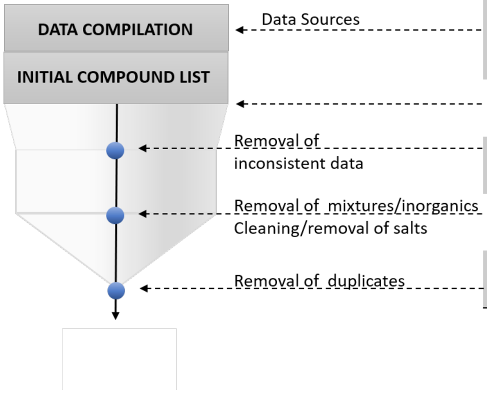


Figure 1.



Skin Sensitization	Skin Irritation/ Corrosion	Eye Irritation/ Corrosion	Acute Dermal Toxicity	Acute Inhalation Toxicity	Acute Oral Toxicity
ECHA/ ICCVAM	ECHA	ECHA/ Literature	ECHA/ ToxValDB/ Creton et al.	ECHA/ ToxValDB	NICEATM
10,861	5,274	7,332	29,824	8,176	8,994
2,347	1,631	7,196	5,259	2,061	8,987
1,110	1,326	5,985	4,601	1,637	8,952
1,000	1,012	3,547	2,622	681	8,495

Figure 2.

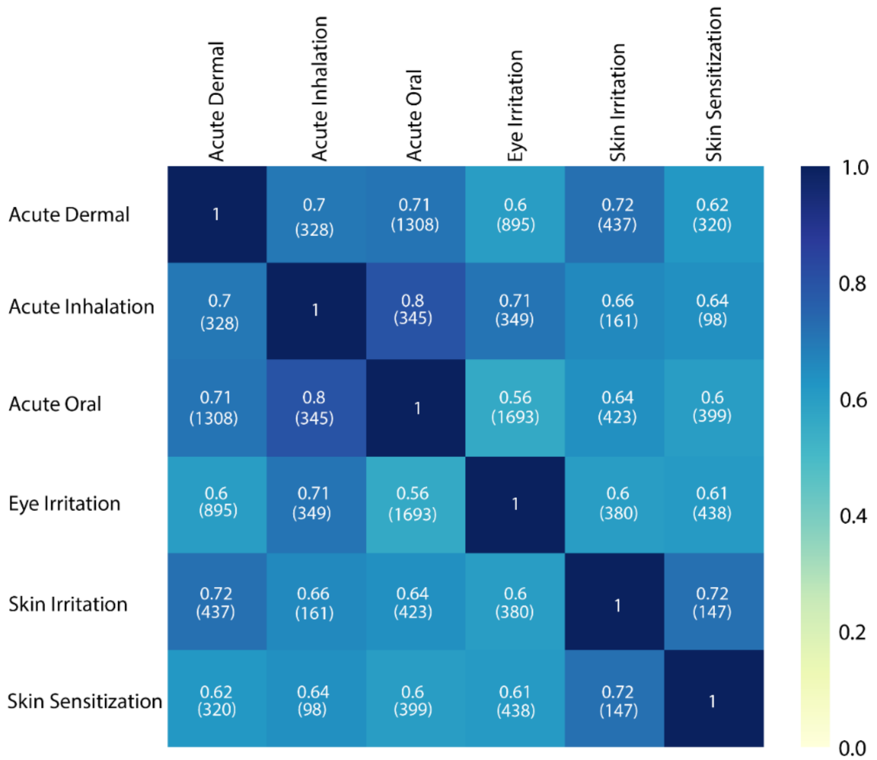

Acute Dermal
Acute Inhalation
Acute Oral
Eye Irritation
Skin Irritation
Skin Sensitization
Acute Dermal 1 0.7 (328) 0.71 (1308) 0.6 (895) 0.72 (437) 0.62 (320)
Acute Inhalation 0.7 (328) 1 0.8 (345) 0.71 (349) 0.66 (161) 0.64 (98)
Acute Oral 0.71 (1308) 0.8 (345) 1 0.56 (1693) 0.64 (423) 0.6 (399)
Eye Irritation 0.6 (895) 0.71 (349) 0.56 (1693) 1 0.6 (380) 0.61 (438)
Skin Irritation 0.72 (437) 0.66 (161) 0.64 (423) 0.6 (380) 1 0.72 (147)
Skin Sensitization 0.62 (320) 0.64 (98) 0.6 (399) 0.61 (438) 0.72 (147) 1
1.0
0.8
0.6
0.4
0.2
0.0

Figure 3.

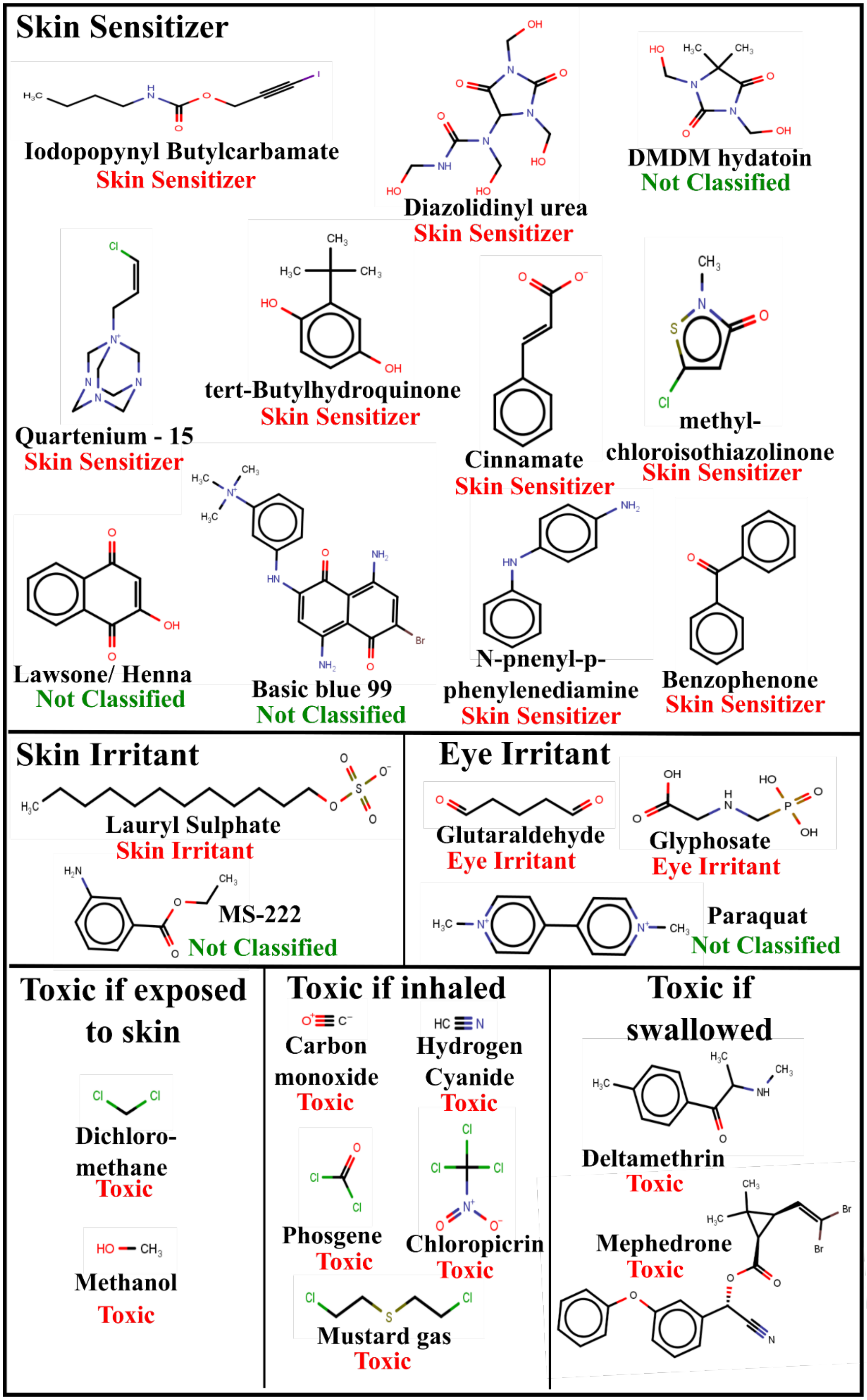

Skin Sensitizer
Iodopopynyl Butylcarbamate
Skin Sensitizer
Diazolidinyl urea
Skin Sensitizer
DMDM hydatoin
Not Classified
Quartenium - 15
Skin Sensitizer
tert-Butylhydroquinone
Skin Sensitizer
Cinnamate
Skin Sensitizer
methyl-chloroisothiazolinone
Skin Sensitizer
Lawsone/ Henna
Not Classified
Basic blue 99
Not Classified
N-pnenyl-p-phenylenediamine
Skin Sensitizer
Benzophenone
Skin Sensitizer
Skin Irritant
Lauryl Sulphate
Skin Irritant
MS-222
Not Classified
Eye Irritant
Glutaraldehyde
Eye Irritant
Glyphosate
Eye Irritant
Paraquat
Not Classified
Toxic if exposed to skin
Dichloro-methane
Toxic
Methanol
Toxic
Toxic if inhaled
Carbon monoxide
Toxic
Hydrogen Cyanide
Toxic
Phosgene
Toxic
Chloropicrin
Toxic
Mustard gas
Toxic
Toxic if swallowed
Deltamethrin
Toxic
Mephedrone
Toxic

Figure 4.

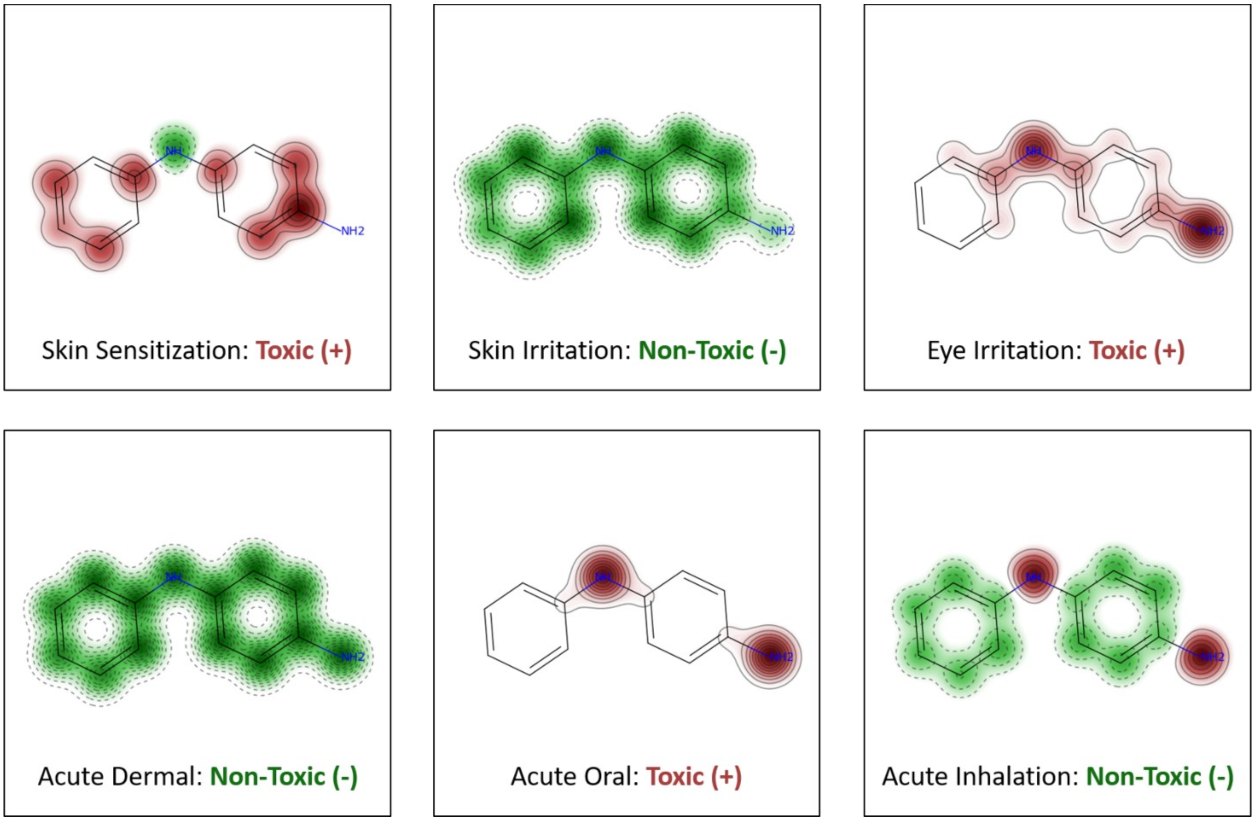

NH
NH2
Skin Sensitization: Toxic (+)
NH
NH2
Skin Irritation: Non-Toxic (-)
NH
NH2
Eye Irritation: Toxic (+)
NH
NH2
Acute Dermal: Non-Toxic (-)
NH
NH2
Acute Oral: Toxic (+)
NH
NH2
Acute Inhalation: Non-Toxic (-)

Figure 5.

Table 1. Computational software covering 6-pack endpoints.

Software name	Endpoints	Computational Approach	License	Access
Danish QSAR database	Acute oral, skin irritation and skin sensitization	Consensus model from ACDLabs, Leadscope, CASE Ultra, and SciQSAR	Free	http://qsar.food.dtu.dk/
T.E.S.T.	Acute oral	QSAR	Free	https://www.epa.gov/chemical-research/users-guide-test-version-42-toxicity-estimation-software-tool-program-estimate
TOPKAT	Acute oral, Acute inhalation, eye irritation, skin irritation, and skin sensitization	QSAR	Commercial	https://www.epa.gov/chemical-research/users-guide-test-version-42-toxicity-estimation-software-tool-program-estimate
ACD/ Percepta	Acute oral, eye irritation, and skin irritation	QSAR	Commercial	https://www.acdlabs.com/products/percepta/index.php
CASE Ultra	Acute oral, acute inhalation, eye irritation, skin irritation, and skin sensitization	Structural Alerts	Commercial	http://www.multicase.com/case-ultra
ToxTree	Eye irritation, skin irritation, and skin sensitization	Structural Alerts	Free	http://toxtree.sourceforge.net/
DEREK NEXUS	Irritation, skin sensitization	Structural Alerts	Commercial	https://www.lhasalimited.org/products/derek-nexus.htm#
OECD QSAR Toolbox	Eye irritation, skin irritation, and skin sensitization	QSAR and Read-Across	Free	https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm#Guidance_Documents_and_Training_Materials_for_Using_the_Toolbox
VEGA	Skin	Read-Across	Free	https://www.vegahub.eu/

	sensitization			
CATMoS (OPERA)	Acute Oral	Consensus model	Free	https://github.com/NIEHS/OPERA
PredSkin	Skin sensitization	QSAR	Free	http://predskin.labmol.com.br/
REACHAcross (RASAR)	All 6-pack	Read-Across and QSAR	Commercial	https://www.ulreachacross.com/

Table 2. Statistical characteristics of QSAR models for 6-pack animal tests.

Endpoint	**CCR**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**# compounds**
Skin Sensitization	0.70	0.66	0.75	0.71	0.75	1,000
Skin irritation/corrosion	0.72	0.77	0.66	0.69	0.74	1,012
Eye irritation/corrosion	0.72	0.72	0.71	0.71	0.71	3,547
Acute dermal	0.76	0.74	0.78	0.77	0.75	2,622
Acute inhalation	0.74	0.69	0.8	0.77	0.72	681
Acute oral	0.77	0.85	0.70	0.79	0.78	8,465

CCR = Correct Classification Rate; PPV = Positive Predictive Value; NPV = Negative Predictive Value