RespiraTox – Development of a QSAR model to predict human respiratory irritants

Wehr, Matthias M.; Sarang, Satinder S.; Rooseboom, Martijn; Boogaard, Peter J.; Karwath, Andreas; Escher, Sylvia E.

DOI:
10.1016/j.yrtph.2021.105089

RespiraTox – Development of a QSAR Model to Predict Human Respiratory Irritants

M. Wehr[1,*], S. S. Sarang[2], M. Rooseboom[3], Peter J. Boogaard[3], A. Karwath[4], and S. E. Escher[1]

[1]Fraunhofer Institute for Toxicology and Experimental Medicine, ITEM, Hannover, Germany

[2]Shell International, Shell Health, Houston, TX, USA

[3]Shell International, Shell Health, The Hague, Netherlands

[4]University of Birmingham, Birmingham, United Kingdom

*Corresponding author

Abstract

Respiratory irritation is an important human health endpoint in chemical risk assessment. There are two established modes of action of respiratory irritation, 1) sensory irritation mediated by the interaction with sensory neurons, potentially stimulating trigeminal nerve, and 2) direct tissue irritation. The aim of our research was to, develop a QSAR method to predict human respiratory irritants, and to potentially reduce the reliance on animal testing for the identification of respiratory irritants. Compounds are classified as irritating based on combined evidence from different types of toxicological data, including inhalation studies with acute and repeated exposure.

The curated project database comprised 1997 organic substances, 1553 being classified as irritating and 444 as non-irritating. A comparison of machine learning approaches, including Logistic Regression (LR), Random Forests (RFs), and Gradient Boosted Decision Trees (GBTs), showed, the best classification was obtained by GBTs. The LR model resulted in an area under the curve (AUC) of 0.65, while the optimal performance for both RFs and GBTs gives an AUC of 0.71. In addition to the classification and the information on the applicability domain, the web-based tool provides a list of structurally similar analogues together with their experimental data to facilitate expert review for read-across purposes.

Key words: human respiratory irritation, QSAR, in silico, machine learning, read-across, web application

Glossary

CLP – Classification, Labelling and Packaging

GHS - Global Harmonized System of Classification and Labelling of Chemicals

STOT_SE - specific target organ toxicity- single exposure

OECD – Organisation for Economic Co-operation and Development

RD_{50} – respiratory depression 50%

OEL - occupational exposure levels

LOAEL – lowest observed adverse effect level

REL - reference exposure level

VOC – volatile organic compound

QSAR – quantitative structure activity relationship

NAM – new approach methodologies

ECHA – European Chemicals Agency

EPA – Environmental Protection Agency

EFSA – European Food Safety Authority

NTP – National Toxicology Program

Highlights

- QSAR model development
- Highly curated in-vivo database
- Predict respiratory irritation
- Alternative method for hazard assessment

Introduction

The human respiratory tract may be exposed to exogenous substances from occupational environments and consumer settings. Inhalation of certain chemicals may induce local damage such as respiratory irritation, acute and chronic inflammation or sensitization. If unresolved these health effects may lead to impaired respiratory function, irritation, inflammation, hyperplasia, or fibrosis. It has been shown that the respiratory tract is one of the most sensitive and frequently impacted target organs in inhalation animal studies (Escher et al., 2010). There are two well established modes of action that a respiratory toxicant may follow, the direct respiratory tissue irritation and sensory irritation pathway. Direct tissue irritation is characterized by inflammation at and surrounding the site of contact. Tissue irritation may result in some of the following cellular effects; cell membrane and cytoskeleton damage, damage to the mitochondria resulting in cell death, receptor activation, or cell signal pathways activation or dysregulation. Sensory irritation occurs when a chemical interacts with the sensory neurons, the neuronal cells become activated by the chemical and this results in the subsequent irritation pathway (Alarie et al., 2001). Sensory irritants may stimulate the trigeminal nerve endings resulting in combinations of symptoms including but not limited to, burning sensitization of the eyes, nose and or throat, as well coughing (Alarie et al., 2001). Sensory irritation is generally caused by chemicals at low concentrations and occurs rapidly upon stimulation of the sensory neurons (Brüning et al., 2014).

In Europe, the Classification, Labelling and Packaging (CLP) guidance (Guidance to Regulation (EC) No 1272/2008) is based on the UN Global Harmonized System of Classification and Labelling of Chemicals (GHS), which classifies a pulmonary irritant as human health hazard (Hazard code: H335 - May cause respiratory irritation). Compounds irritating the respiratory tract

are classified according to STOT_SE (specific target organ toxicity- single exposure) in category 3, if they have not already met the classification criteria of STOT_SE categories 1 and 2. Category 1 and 2 comprise compounds showing significant toxicity in humans and/or animals after single exposure and classifies compounds according to dose. A similar guidance value is not used for category 3. Therefore, if the in vivo data show clear evidence for respiratory tract irritation at any dose level then this could support classification to category 3.

Assessing whether a chemical will cause respiratory irritation is often determined by human experiences. Respiratory irritant effects in humans include symptoms of cough, pain, choking and breathing difficulties. Measurements of respiratory tract irritation in humans may include electrophysiological responses or biomarkers of inflammation in nasal or bronchioloalveolar lavage fluids. Rodent in vivo studies with acute and repeated exposure (OECD Test Guidelines 403, 412, 413, 436) might provide more evidence on the irritating properties, although both in vivo assays do not specifically address respiratory tract irritation. The animals are monitored for clinical signs of respiratory tract toxicity (e.g. dyspnoe, rhinitis) and evaluated for histopathological changes (e.g. hyperemia, edema, minimal inflammation, thickened mucous layer) in the lung.

Pulmonary function tests may be measured if a test chemical is known to have or is likely to have sensory irritant properties. In addition, respiratory rate may be measured to determine irritant effects of a test chemical (ASTM, 2004; Kane et al., 1979). The Alarie test measures the RD_{50} value, a concentration of the test chemical that causes a 50% decrease in rodent respiratory rate (Alarie, 1966). The RD_{50} concentration has been described as intolerable to humans as indicated by the ASTM method. The RD_{50} values are in part the basis for setting several occupational exposure levels (OELs) for example by ACGIH (ACGIH, 2020). Previous studies

have shown a good correlation between RD_{50} and OELs (Kane et al., 1979; Schaper, 1993). The RD_{50} values appear most useful when qualitative data are available indicating sensory irritation. A good correlation between RD_{50} values and lowest overserved effect levels (LOAELs) provides support for using RD_{50} values in determining guidance levels to protect the general public form sensory irritants (Kuwabara et al., 2007). Alaire suggested the use of QSARs capable of predicting RD_{50} values, LOAEL and reference exposure level (REL) values for regulatory purpose for reactive volatile organic compounds (VOCs) or non-reactive VOCs (Alarie, 2016; Federal Register, 2016). However, it is difficult to extrapolate rodent respiratory rate to humans, as rodents have different mechanisms to adapt to respiratory rate changes due to toxicants. Therefore, a combination of animal and human studies can be used as part of a weight of evidence evaluation to determine if a chemical is a human respiratory irritant (UN GHS, 2019).

Over the past several years there has been an increase in the development of new approach methodologies (NAMs), which comprise human *in vitro* and *in silico* models. NAMs provide an opportunity for the screening of large numbers of chemicals and for prioritization of chemicals for which further assessment is required, which result in reduction on the reliance of animal testing. There are (Q)SAR models for skin- and eye irritation, skin sensitization, mutagenicity, and endocrine disruption related effects e.g. available in the OECD QSAR toolbox (https://qsartoolbox.org) or the DTU QSAR database (http://qsar.food.dtu.dk/index.html). These models are utilized in predicting human health hazards, data-gap filling, read-across approaches, and for screening of chemicals. The modelling of toxicity endpoints using QSAR´s helps regulators to conduct a scientifically defensible analysis in situations when other laboratory and computational methods either are missing or ineffective, and where prior knowledge is scarce (Demchuk et al., 2011). Regulatory authorities in Europe (ECHA, EFSA), and US (EPA, NTP)

125 are interested in receiving relevant data generate from validated NAMs for classification and
126 labelling, prioritization and risk assessment (Escher et al., 2019). In providing data from NAMs,
127 authorities will build trust in these methodologies eventually leading to more acceptance.
128 A limited number of in vitro models for predicting human respiratory irritation exist. The in vitro
129 models include single cell types or co-cultures of epithelial cell and lung specific immune cells.
130 These cells are grown in culture systems that closely resembles human physiological conditions
131 (Clippinger et al., 2018). Human primary bronchial epithelial cells or human derived cell lines
132 (NCI-H292) have been used to determine how a substance may impact the human respiratory
133 system. Human primary cells maintain a closer physiological response to that of cell lines, since
134 the primary cells express markers and functions as observed in vivo (Clippinger et al., 2018). Other
135 models include ex vivo human lung tissue, this model retains the structure of the lung, includes
136 the complement of lung cell types, and can be cultured for weeks for longer term study (Clippinger
137 et al., 2018). These in vitro models can be used to screen chemicals for potential respiratory irritant
138 hazard identification. Further refinement of these models will be required for regulator acceptance
139 or classification.

140 Currently, there are a limited number of QSAR models for the prediction of human respiratory
141 irritation, these models use molecular properties and RD50 values to predict threshold limit values
142 (TLV) for workspace exposure for sensory irritating volatile organic compounds (VOCs) (Alarie
143 et al., 1996; Gupta et al., 2015; Nielsen and Wolkoff, 2017). A model exists, which is derived from
144 a baseline narcosis model that compares vapour pressure to the logarithm of lethal concentration
145 at 50% mortality ($\log LC_{50}$) value in rodents for neutral organic substances (Veith et al., 2009).
146 Another QSAR model for inhalation toxicity of organic chemicals was based on octanol air
147 partition coefficients and molecular physicochemical descriptors (Raevsky et al., 2011).

The goal of this research was to develop a QSAR model that can predict human respiratory irritants and reduces the reliance on animal testing. The QSAR model was developed to predict human respiratory irritants from molecular physicochemical and structural information and follows the OECD QSAR principles. Compounds were classified as irritating to the respiratory tract distinguishing two observed in vivo effect types, namely irritation or damage reported in the tissues of the respiratory tract and respiratory distress which potentially results from sensory irritation. A workflow for training the machine learning models and predicting new chemicals properties using previously trained models is provided for reproducibility. The model is provided as web application and can be freely accessed via “respiratox.item.fraunhofer.de”.

Materials and Methods

QSAR model principles

The development of the QSAR model applied the five OECD QSAR principles to facilitate the consideration of the QSAR model for regulatory purposes. The principles considered include, 1) a defined endpoint, 2) an unambiguous algorithm, 3) a defined domain of applicability, 4) appropriate measures of goodness-of-fit, robustness and predictivity, 5) a mechanistic interpretation, if possible (OECD QSAR, 2007). In the following sections we address these principles with regards to the learnt model in more detail.

Data sources

Substances having a respiratory irritant potential are defined in several data repositories including those from regulatory and authoritative bodies (ECHA, ACGIH, US EPA). Respiratory irritants terms related to respiratory distress and pathological findings in respiratory tract tissue were retrieved from following sources: a) toxicological studies - 1) acute inhalation exposure toxicity studies from ECHA CHEM (https://echa.europa.eu/), 2) Fraunhofer AcuTox database which contains acute inhalation toxicity data originally extracted and further curated from CHEMID Plus (https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp) 3) repeated dose studies with inhalation exposure from the Fraunhofer RepDose database (https://repdose.item.fraunhofer.de/); b) classification information: 4) the endpoint conclusion on respiratory irritancy from ECHA CHEM (https://echa.europa.eu/), 5) the Hazardous Substance Database (HSDB; https://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm), 6) the harmonized classification and labelling inventory from ECHA (https://echa.europa.eu/de/regulations/clp/harmonised-classification-and-

labelling). The following selection criteria were used to compile the project dataset: for acute toxicological studies from ECHA CHEM only key studies in rodents or dogs were included, whereas weight of evidence data, and poor quality studies were removed, the same criteria apply to ECHA classification information. At the time of access ECHA CHEM listed about 19,000 registered substances including 4,400 experimental inhalation studies of which 800 were excluded based on the given criteria.

Classification of compounds as irritating or non-irritating

The toxicological data gathered form the above listed sources provide highly heterogeneous information, which differs with regard to the type of data (classification or observed effects) and also the level of detail. The acute and repeated dose studies for example provided toxicological effect descriptions with, in some cases, associated dose information, whereas the classification data e.g. ECHA C&L H provided GHS codes and thus categorical data only.

Studies from Fraunhofer AcuTox and RepDose databases are comprised of only high to acceptable quality studies with Klimisch rating 1 or 2.

In contrast to other data sources, RepDose is a relational database, which includes a curated vocabulary and ontology and is thus enabled to search for irritating compounds based on a predefined and controlled set of terms/organ systems (as listed in Supplement 1, Annex 4)(Bitsch et al., 2006). Here, the high quality studies in rodents from RepDose were included. For all data sources, the terms used to classify compounds as respiratory distressing or tissue irritating are listed in Supplement 1. This is of particular importance to better understand the classification arising from the toxicological data of the ECHA and AcuTox acute toxicity studies and the HSDB. This information was only available as free-text components in these data sources and a data

parsing approach was needed to extract relevant terms. The raw list of identified clinical terms discovered in ECHA CHEM studies alone was comprised of about 18,000 terms including different (miss-) spellings and styles. After consolidating similar terms the remaining 1,600 effect terms were organized into categories. The category for respiratory distress contained 234 terms; whereas 221 and 59 terms indicate tissue damage to nose and to lung respectively. As the occurrence of dose response data was sparse and could only be found in fully reported toxicological studies, dose-response relations were not considered in this work.

In summary, compounds were classified as respiratory distressing when clinical signs included laboured breathing or dyspnoea. The category tissue irritation was assigned to compounds, for which histopathological effects were present.

Molecular descriptors

To define individual substances three different descriptor information groups were employed. Group 1 consists of defined chemical structures retrieved from selected data sources; Fraunhofer *RepDose*, *ChemIDplus* (https://chem.nlm.nih.gov/chemidplus/), and *Chemicalbook* (https://www.chemicalbook.com). All structural information was converted to canonical SMILES. The chemical structures were used to calculate Extended Connectivity Fingerprints (ECFPs) using the Morgan method and converted to a binary fingerprint for each compound (Morgan, 1965; Rogers and Hahn, 2010). Group 2 comprises the information of the predicted physicochemical (PC) properties based on each compounds respective SMILES code. For the derivation of PC-parameters, we used a list of chemical toolkits, such as CDK, RDKit, EpiSuite, Vega QSAR, and the OECD QSAR Toolbox (Supplement 1, Annex 7). Additionally, we generated a third group, Group 3. This group consists of the combined set of both descriptor sets Group 1 and Group 2.

The final list of descriptors is given in the supplement (Supplement 1, Annex 8). The descriptors were used to calculate a correlation matrix. Descriptors with the maximum correlation of 1 or -1 to another descriptor were removed in order to reduce the effect of redundant information in the machine learning (Supplement 1, Annex 9).

CAS numbers and SMILES codes are used in conjunction to uniquely identify a given compound within our dataset. Duplicated structures could occur due to CAS identifiers being highly specific (e.g., they can distinguish the same compound in different purities). For this reason, we created a chemical curation workflow. We merged duplicate structures using their canonical SMILES codes and removed multi-constituents compounds. The result is a set of defined, unique, and mono-constituent compounds. The curation workflow is outlined in the results section.

Machine learning tools used for respiratory irritant selection and classification

To build our (Q)SAR models we employed Random Forests (RFs) and Gradient Boosted Decision Trees (GBTs) as underlying machine learning algorithms. Both approaches can address the imbalance in our training datasets. GBTs are similar to RFs as they build models containing a number of Decision Trees (DTs). In contrast to RFs where each tree is learnt on a random subset of the data and features - and the individual models are combined subsequently, GBTs induce DTs sequentially and re-weight misclassified examples using a boosting approach. The algorithms are tuned to avoid overfitting of the data. In comparison to logistic regression (LR), GBTs, and RFs are well suited to account for high variance in the training data. Furthermore, we trained a LR as a baseline model for performance comparison. The *k*-fold cross validation technique was used for internal validation during the training stage. Within each iteration the training data are normalized

independently, and normalization factors are applied to the test data in each fold. The parameters for the workflow optimizations are described in Table 1, and the workflow optimization is presented in Figure 1. We selected $k = 5$ resulting in an 80:20 split for training and test set in each training iteration. Choosing a higher k would result in smaller validation sets.

As performance measure we selected the Area Under ROC Curve (AUC), also known as c-statistic. The receiver operating characteristic (ROC) curve is drawn using a sliding threshold from 0 to 1 on the positive prediction probability.

Table 1: Workflow optimization, iterating over parameters to tune for best AUC.

Algorithm	Parameter	Range	Steps
Linear Regression (KNIME Core)	max. epochs	1-5000	1.5 (logarithmic)
	step size	1×10^{-1} - 1×10^{-4} or line search	1.0 (logarithmic)
	prior variance	1×10^{-1} - 1×10^{-8} or uniform	1.0 (logarithmic)
Gradient Boosted trees (KNIME Core)	num. models	50 – 200	50 (linear)
	max. levels	1-26	5 (linear)
	learning rate	1×10^{-1} – 3×10^{-1}	0.05 (linear)
Random Forest (KNIME Core)	num. models	50 – 200	50 (linear)
	max. levels	10 – 150	10 (linear)

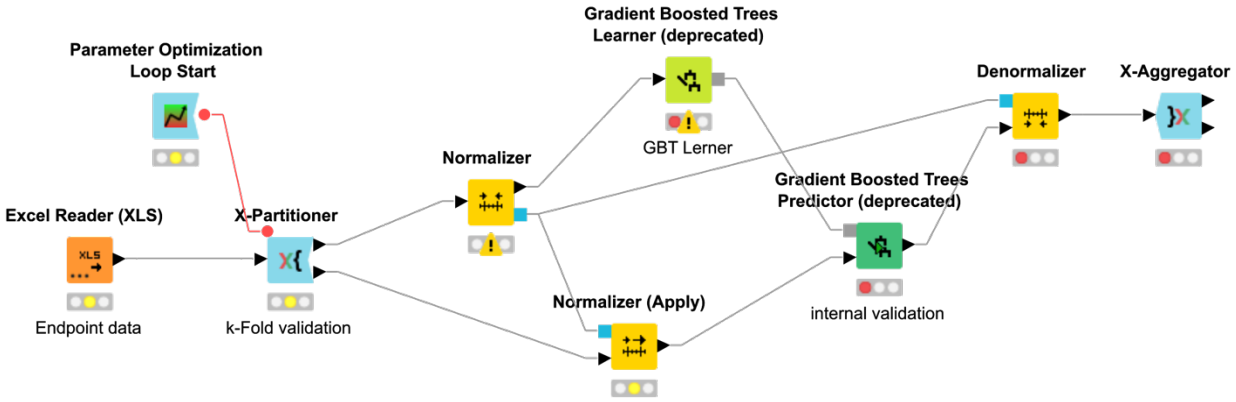


Figure 1: The workflow is optimized using a gradient boosted tree model. Data is prepared for training the model by normalizing, then the model is trained and tested using an independent test set. Then input data is split to 80% training and 20% test set using a stratified method. The varying parameters are selected to overall train different models optimizing for high Area Under Curve in each iteration.

Methods used for random experiments

To assess the predictive performance of each of the different algorithms and dataset/feature combinations, we performed a 5-fold stratified cross validation. This ensures that the same number of irritant and non-irritant compounds are selected for training and testing in each fold. For consistent results, we used the same random seed for the creation of these folds, such that all model performances can easily be compared.

Applicability domain

The Euclidean distance of a test compound to compounds in the training set is calculated using the molecular descriptors (see section molecular descriptors). The Euclidean distance is used to assess whether a test compound falls into the applicability domain for our model (Zhang et al., 2006). To visualize the distance between compounds in the training set the descriptors have been reduced

using PCA and plotted in 2D space (Supplement 2, Figure 1). The reduction of descriptors using the PCA was useful during the model development to assess the chemical space and compound distribution within the training set. The training set development and members of the training set are presented in the supplementary sections (Supplement 1, Annex 6 and Annex 8). We calculated the Applicability domain threshold (APD) using the pairwise Euclidean distances of the compounds as described in Melagraki et al. using the ENALOS domain similarity KNIME nodes for modelling (http://enalosplus.novamechanics.com/index.php/enalosplusnodes/modelling/). The default cutoff value of 0.5 for Z is used (Afantitis et al., 2011; Melagraki et al., 2010).

Web application

We incorporated the model in a designated web application, which can be accessed online (https://respiratox.item.fraunhofer.de). After registering and authentication, the web application offers to perform the toxicological analysis for a given compound. The compound has to be entered either as SMILES code or can be drawn in a graphical editor. The tool allows to specify a number of distinct fingerprint algorithms and distance definitions to be used for determining chemical similarity scores; in particular two molecular fingerprints (PubChem (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf) and ECFP (Rogers and Hahn, 2010)) and three algorithms (Tversky (Tversky, 1977), Tanimoto (Levandowsky and Winter, 1971) and Euclidean distance).

After submitting the compound and selection fingerprints and similarity calculation, the web application returns the toxicology prediction, including the applicability domain employed. Furthermore, the tool lists up to 100 chemically similar compounds from the internal training

database in descending order of similarity score. The lists includes per compound the name, CAS number and similarity score as well as a graphical representation, and the observed type of irritation (respiratory distress/tissue) and the data sources. The result table can also be downloaded as CSV/XLS.

Results

The originally collected 3137 compounds with defined endpoints were pre-processed according to the chemical curation workflow to ensure a homogenous dataset (Figure 2). First, a total of 45 compounds were excluded because of errors in the SMILES syntax. A further set of 799 inorganic compounds were removed, such that the resulting model is only applicable to organic compounds. Additionally, 180 disconnected or multi-constituent structures were disregarded. A further set of 82 structures were dropped, after merging duplicate structures. Finally, 34 compounds had to be excluded, as some feature calculations failed. The resulting dataset is comprised of 1997 unique organic compounds.

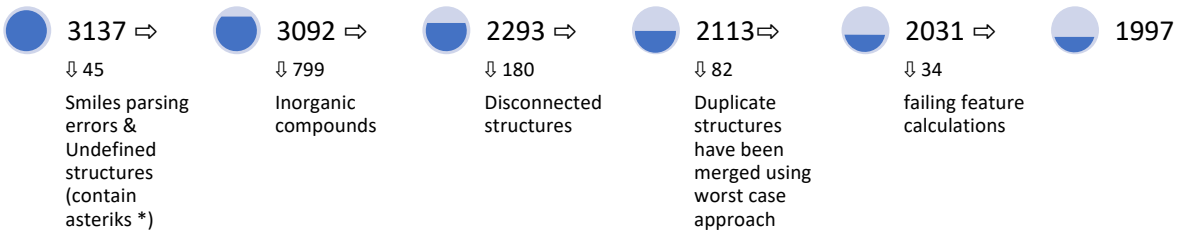


Figure 2: Chemical curation workflow to derive the RespiraTox project dataset. The workflow includes the following steps 1) parse SMILES codes to ensure valid structure definition 2) remove inorganic compounds 3) remove multi constituent compounds 4) merge compound information based on identical canonical SMILES code.

For these 1997 compounds, respiratory irritants terms related to respiratory distress and tissue irritation of the respiratory tract were retrieved from various data sources, with most compounds having information about their respiratory irritation potential in multiple data sources (Supplement

2, Figure 2). The majority of the respiratory distress data was collected from the ECHA CHEM Acute toxicity studies and the HSDB.

In case of conflicting data, a worst-case approach was used to classify compounds. The final dataset comprised of 1553 irritating compounds; of which 739 are exclusively related to respiratory distress, 240 tissue irritants, while 574 compounds showed both distressing and tissue damaging properties and 444 are not observed to be irritating up the highest tested dose (Figure 4). In our dataset, respiratory distress is described by a variety of terms such as labored breathing or, dyspnoea (see Supplement 1, Annex 1). Out of 574 compounds, 94 had an RD50 above the arbitrary threshold of 10.000 ppm or TLV Basis indicating sensory irritation according to the 2015 update of 1993 Schaper database (Schaper, 1993). For the remaining 480 compounds the cause of distress is unknown and might originate from tissue damage (e.g. edema), CNS effects or stimulation of the trigeminal system. Hence, the final model will distinguish between any of these effects and no observed effects.

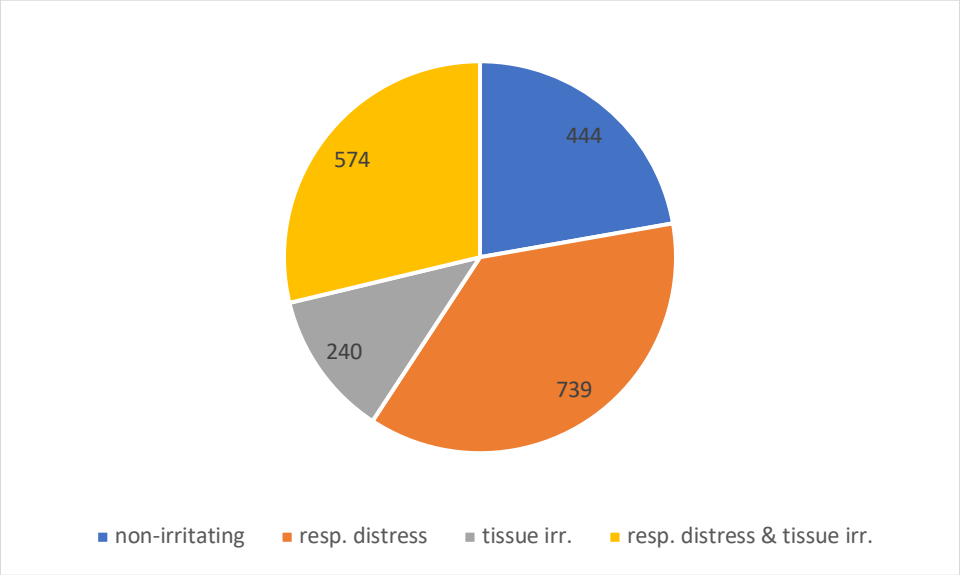


Figure 3: Overview on the number of compounds showing different types of respiratory irritation properties in the data set, namely non-irritating (blue); respiratory distress (orange); tissue irritation (grey) and tissue irritation together with respiratory distress (yellow).

Classification of compounds

The classification of compounds in this dataset might depend on data annotation richness. An annotation concordance analysis was performed to better understand the impact of the number of annotations from the data sources on the classification. Three sources include in vivo animal study data (ECHA CHEM Acute toxicity studies, FhG AcuTox and RepDose). Overall, 50 compounds were labelled as respiratory distressors and possess annotations in all three sources (Figure 5a). The concordance analysis revealed that 28 (56%) compounds are consistently classified as respiratory distressing (Figure 5a). A higher concordance is obtained for the 21 compounds, labelled as respiratory irritants.

When based on classification data (HSDB; ECHA Classification and Labelling and ECHA Irritation), 18 out of 21 (86%) showed a concordant annotation in all three sources.

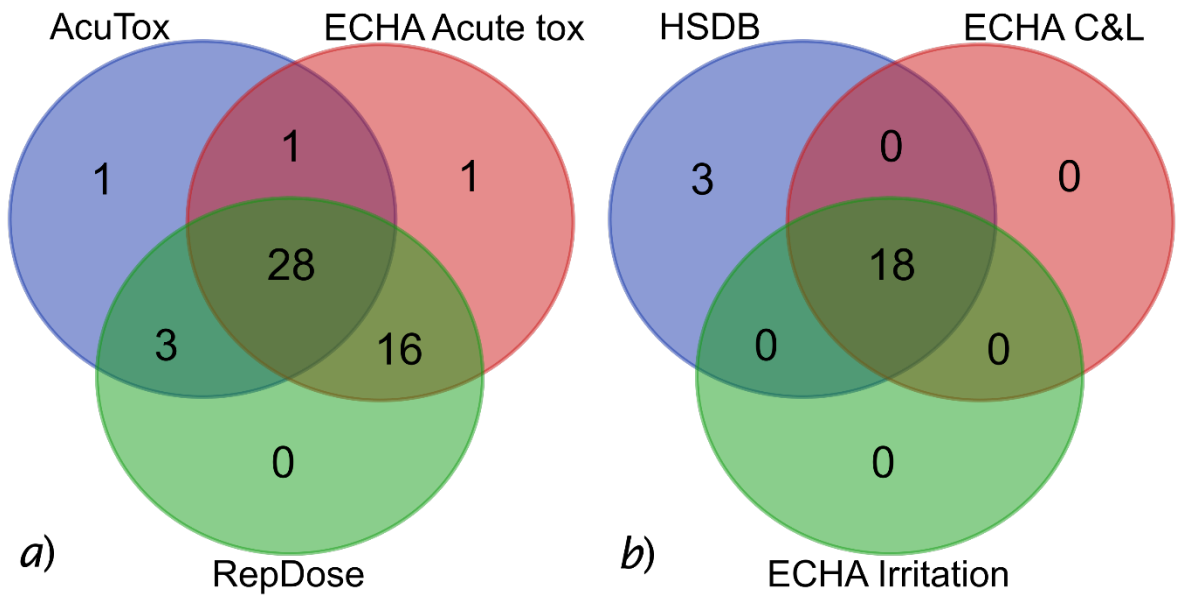


Figure 4: Concordance analysis for classification of irritating compounds: a) 50 compounds have in vivo animal data in all three sources; 28 were classified by all sources; for 16 others there were concordances from acute toxicity studies of the ECHA CHEM database and RepDose studies. b) For 21 compounds classification data in all three sources. 18 out of 21 were classified by all three sources as irritating. The 3 compounds labelled as irritating by HSDB have CNS effects in the ECHA C&L.

The low concordance between the in vivo animal data sources indicates that the applied classification inherit a certain amount of uncertainty. It is likely that a case by case evaluation of in vivo animal data, considering e.g. the differences in study design, dose spacing, dose selection, or severity of effect would result in different classifications. This weight of evidence approach was not possible here due to the number of compounds involved in this analysis.

For the non-irritating compounds, the concordance between the different sources is 100%, due to the applied worst case approach. Only when all sources agreed on the absence of irritating properties were these compounds labelled as non-irritant. This is one reason why a minority of

444 compounds possess a non-irritating label, resulting in an unbalanced data set where the majority of compounds are classified as irritating (Figure 6). Additionally, sources like the HSDB, ECHA C&L or ECHA Irritation mainly comprise of irritating compounds.

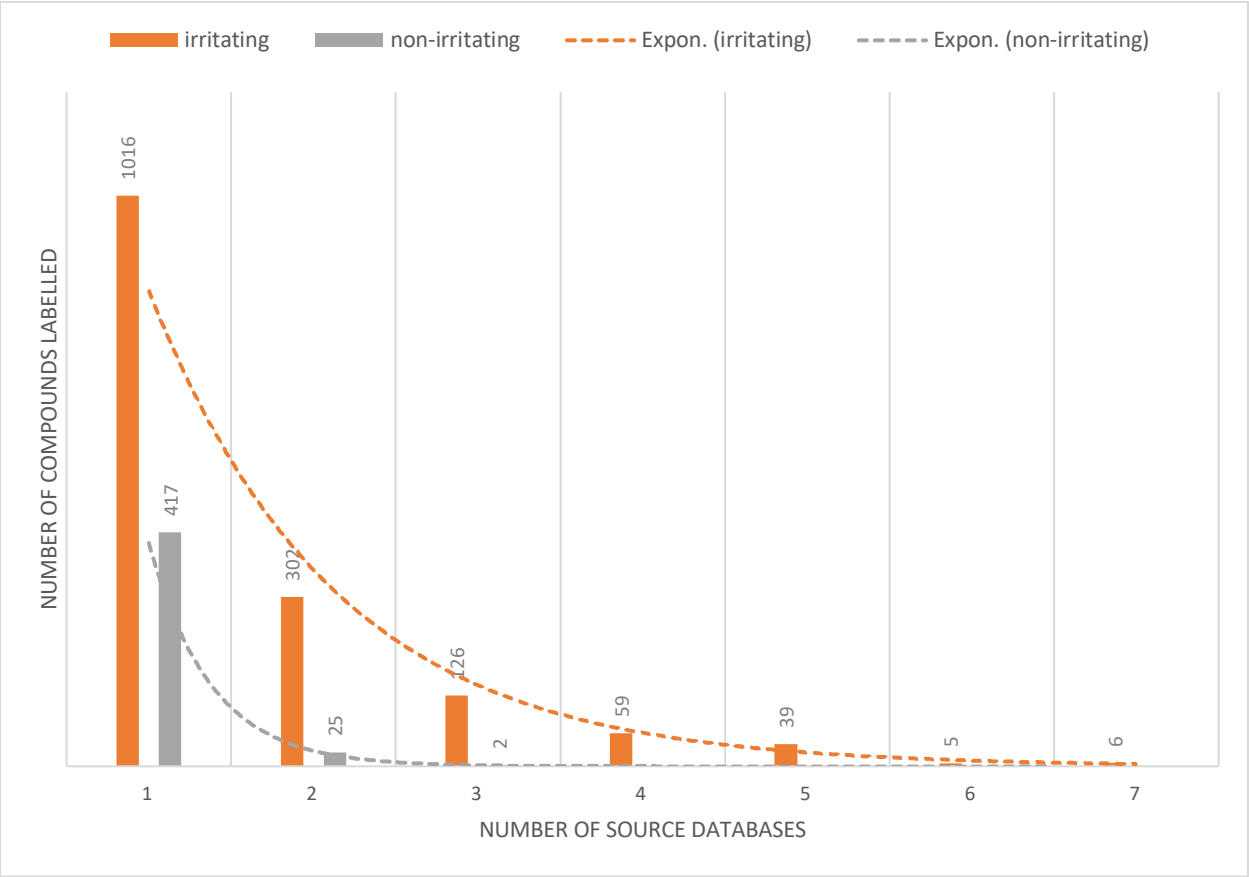


Figure 5: A decreasing number of compounds are classified based on multiple source databases. The decrease is more prominent for non-irritating compounds. At exactly three source databases only two are non-irritating compounds, while 126 are irritating compounds exist. From one to three source databases, the ratio of non-irritating vs. irritating decreases by an order of magnitude. Final classification labels used for training are presented in Supplement 1, Annex 6.

Modelling

The final model aims to accurately distinguish between compounds being irritating or non-irritating to the respiratory tract. Initially, individual models have been trained to classify

385 compounds by respiratory distress or damaging to tissue of the respiratory tract or nose. Using the
386 AUC metric as an indication of model performance, the individual models on respiratory distress
387 and tissue irritation did not perform as well as a the combined overall classification resulting from
388 a worst case approach. This might be due to difficulties distinguishing respiratory distress from
389 tissue damaging effects based on reported effect descriptions in the in vivo rodent studies in this
390 project. The observation of respiratory distress in these studies could potentially be secondary to
391 actual tissue damage.

392 We chose three different machine learning techniques. First we used logistic regression (LR) and
393 compared it to two other tree-based methods, namely Gradient boosted trees (GBTs) and Random
394 Forest (RF). In this project, the tree-based methods always outperformed logistic regression in all
395 experiments (Table 2). For each experiment, we evaluated a number of parameters for the machine
396 learning algorithm (see Table 1) and performed a 5-fold cross validation. Furthermore, we
397 evaluated different partitioning approaches for the fold selection: purely at random, balanced, or
398 stratified by endpoint. The stratified 5-fold approach, whereas each fold contains the same
399 proportion of irritating and non-irritation compounds as in the overall training set, was the
400 preferred choice. This is due to the fact that this approach does not produce test folds where either
401 label is missing completely.

402 Each experiment was evaluated using the AUC considering the probability of a compound being
403 predicted positive. In the ROC curve true positive rate and false positive rate are plotted for a
404 decision cut-off from 0 to 1.

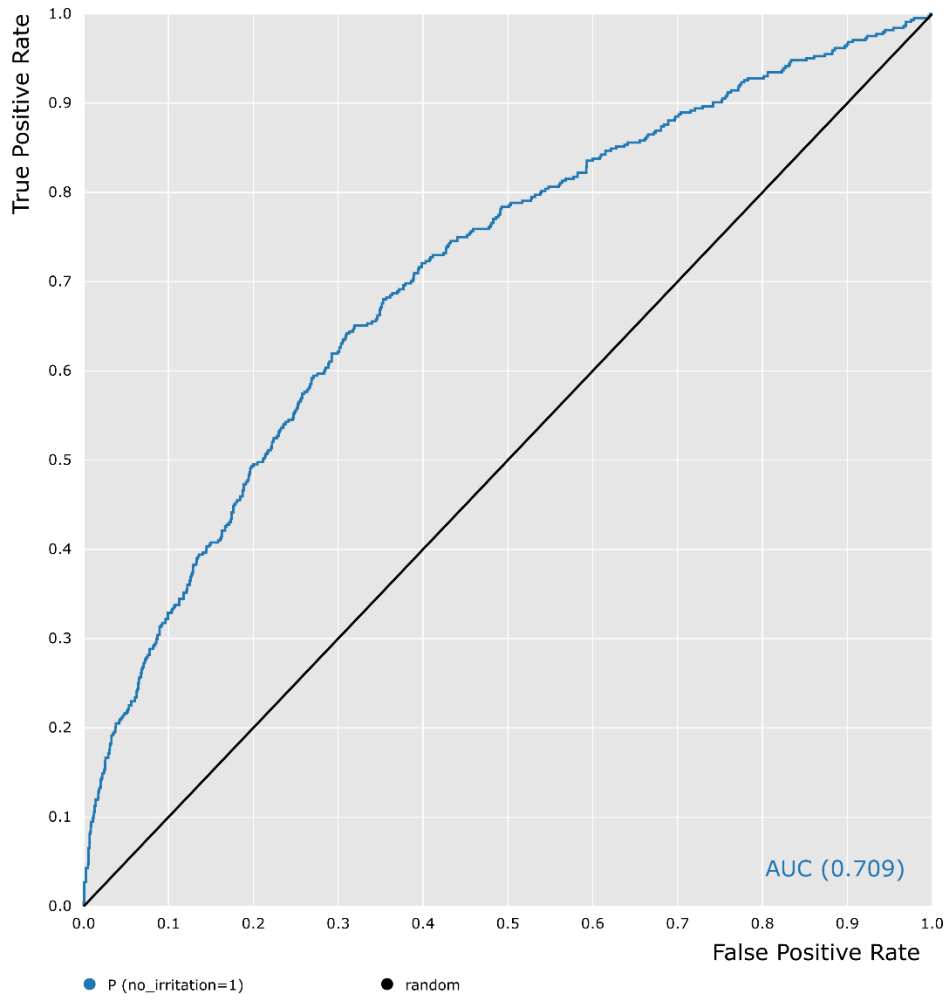


Figure 6: Receiver Operating Curve, for best performing model overall. The model used is GBTs, trained with 200 trees and a learning rate of 0.15. The AUC for this curve is given in the bottom right corner.

The ROC curve and corresponding AUC are shown exemplary for the best performing model in Figure 7. The ROC curve was neither skewed towards high true positive rates or false positive rates.

The larges impact on performance was the used feature set, followed by the selected machine learning algorithm. Tuning the parameters marginally improved the model performances further. The AUC for each combination of machine learning model and feature set are given in Table 2.

The Gradient boosted trees based on physico-chemical properties derived the best AUC value of 0.709 (Table 2).

Table 2: Comparison of AUC per machine learning technique and feature set after tuning parameters for each technique. The Gradient Boosted Tree models performed best overall when only including physicochemical properties.

AUC	*Physico-chemical properties*	*ECFP Fingerprint (PubChem FP)*	*Combined set*
Logistic regression	0.647	0.57 (0.591)	0.587
Gradient boosted trees	**0.709**	0.579 (0.674)	0.677
Random Forest	0.695	0.591 (0.685)	0.625

The AUC improved selecting appropriate learning techniques and features in the training set, still the outcome is limited by the input training sets.

Further statistics for the selected final model are presented in Table 3. The table shows a high imbalance of sensitivity and specificity of the model. The results are very consistent throughout each fold of the k-Fold validation.

Table 3: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) denote the number of true or falsely predicted compounds of their respective class (actual positive / negative). Precision is the ratio of true positives to all positive predicted (TP+FP) compounds. Sensitivity is synonym to true positive rate and denotes the ratio of true positives to actual positives (TP+FN). Specificity or true negative rate is the ratio of true negative to actual negative (TN+FP). F-measure is the harmonic mean of sensitivity and specificity. Each line represents an iteration of the k-fold algorithm and shows the individual statistics as well as the overall statistics

Model	*TP*	*FP*	*TN*	*FN*	*Precision*	*Sensitivity*	*Specificity*	*F-measure*	*Fold*
Gradient boosted trees	1470	344	100	83	0.810	0.947	0.225	0.873	overall
	292	67	22	19	0.813	0.939	0.247	0.872	1
	296	70	19	15	0.809	0.952	0.213	0.874	2
	289	69	19	22	0.807	0.929	0.216	0.864	3
	295	70	19	15	0.808	0.952	0.213	0.874	4
	298	68	21	12	0.814	0.961	0.236	0.882	5

A high sensitivity indicates that all actual positive compounds are predicted positive, the probability of a positive compounds being predicted negative is low. The obtained relatively low specificity of the RepiraTox model indicates that the models tends to over predict, resulting in a relative high number of actually negative compounds that are predicted as false positives. Most machine learning models try to balance / optimize these values, while for some application one might be preferred over the other. The tendency to over predict might be related to the nature (regulatory purpose) of the repositories used for this project, since most compounds were positive for irritation, resulting in an unbalanced dataset with fewer negative than positive compounds.

Expert-review of model predictions

For the evaluation of a QSAR model outcome, an expert review is recommended to evaluate the obtained prediction with regard to its relevance and reliability (Barber et al., 2015). To facilitate this expert review, the RespiraTox web application (https://respiratox.item.fraunhofer.de) provides

the 100 structurally closest analogues in the trainings set, together with their individual classification data and the underlying data sources. The basis for calculation of similarity here can be chosen, e.g. by using either of the offered Fingerprint and any distance measure (refer to section Web application of Materials and Methods). Relevant analogue compounds for this query have been identified up to the 59th compound (*2-Isopropanol (CAS 67-63-0)*) of this list. In the following the prediction for the hypothetical query compound tert-butyl isopropyl ether (Figure 8) is described, illustrating the advances and current limitations of the RespiraTox model.

The RespiraTox model predict the query compound tert-butyl isopropyl ether to be positive for irritation with very high probability of 99.5%. Additionally the compound is well within the Applicability domain with a calculated APD of 1027.

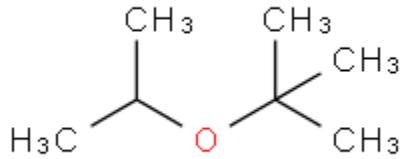


Figure 7: Structure of the query compound tert-butyl isopropyl ether; CAS 17348-59-3; DTXSID4051792; canonical smiles code CC(C)OC(C)(C)C). The query compound is characterized by a molecular weight of 116.2 g/mol; a boiling point of 87.6 °C; a logPow of 1.83; a vapour pressure of 97 hPa and a Henry constant of 1.60e-3 atm-m3/mole (source Comptox dashboard (https://comptox.epa.gov/dashboard))

The combination of the PubChem fingerprint and the Tversky algorithm yields 100 short-chain dialkyl ethers ordered in descending order of similarity score. Since the toxicity of the query compound is not known, the review might explore two different read-across hypothesis. The first read-across hypothesis assumes that toxicity of the query compound is based on its structural

features and biotransformation does not occur (Figure *8*), whereas the second read-across hypothesis assumes that the query compound undergoes hydrolysis in vivo and is cleaved to the two alcohols tert-butanol and isopropanol (Figure *9*).

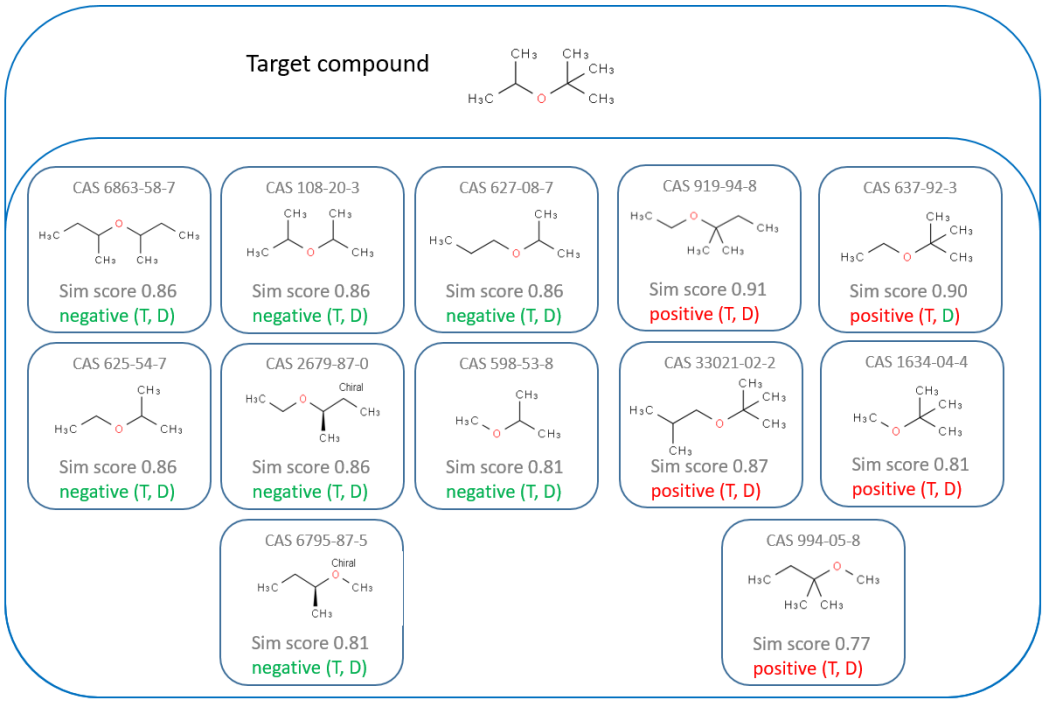
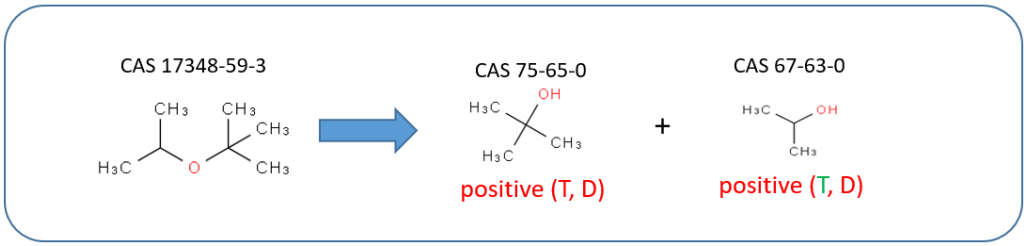


Figure 8: Read-across hypothesis 1 – tert-butyl isopropyl ether causes respiratory irritation, biotransformation is not involved. The experimental data of the structurally closest analogues to the query compound are used for expert review. In the display of structural analogues the labels for Tissue irritation (T) and respiratory Distress (D) are colored in red for positive in regard to this endpoint and green for negative respectively.

The 12 structurally closest similar analogues comprise of five dialkyl ethers, which contain one tert-alkyl carbon atom next to the oxygen atom (termed tert-alkyl ether in the following) and seven secondary alkyl ether, e.g. diisopropyl ether (CAS 108-20-3; Figure 9). A high concordance of the experimental data is observed for both classes. All tert-alkyl ethers have experimental data that show irritation in the respiratory tract, with tert-butyl ethyl ether (CAS 637-92-3) showing only tissue irritation. The seven secondary dialkyl ethers are all negative for respiratory irritation based on the training set data of the RespiraTox model.

A more detailed investigation of the observed effects in the underlying toxicological data (Supplement 1, Annex 5) reveals, that tert-alkyl ether are weak respiratory irritants, showing lung foci and tissue irritation at high doses in acute toxicity studies provided by the ECHA CHEM DB. Supporting data are obtained from the HSDB, which classifies the analogue Methyl-tertiary-butyl ether (CAS 1634-04-4), as respiratory irritant. 2-Methoxy-2-Methylbutane (CAS 994-05-8) is reported to affect lungs, thorax and to cause pulmonary emboli. The secondary ethers do not have data in the ECHA CHEM database, nor the HSDB. They are consistently reported as general anesthetic in the Fraunhofer AcuTox database. One source of uncertainty is the difference of data richness for the different analogues, another source is the missing dose response information, as the dose at which the respiratory effects occurred is not clearly stated. As the query compound contains a tert-butyl group, it might be assumed that the obtained prediction is valid based on the positive majority vote from these five most similar tert-alkyl ethers.

The second read-across hypothesis is based on the activity of the two alcohols which result from ether hydrolysis (Figure 10). Cleavage of the ether bond of the query compound results in tert-butyl alcohol (CAS 75-65-0) and 2-Isopropanol (CAS 67-63-0). Both compounds are part of the RespiraTox database and their experimental data classify both as respiratory irritants. For tert-butyl alcohol, the acute toxicity studies agree on tissue irritation in terms of focal areas of redness in the lung/pulmonary emboli and respiratory distress characterized by dyspnoe and change in breathing rates, whereas isopropanol showed narcotic effects accompanying with laboured breathing and mild irritation reported in humans according to the HSDB. Based on this hypothesis the query compound tert-butyl alcohol would also be classified as irritating to the respiratory tract. One source of uncertainty are the missing data on kinetics.



507

508 Figure 9: Read-across hypothesis 2 – formation of the two metabolites tert-butyl alcohol (CAS 75-
509 65-0) and 2-Isopropanol (CAS 67-63-0) explain the toxicological effects of the query compound
510 tert-butyl isopropyl ether. The experimental data of the structurally closest analogues to the query
511 compound are used for expert review. Labels for endpoints as Fig. 9

512

Discussion

The RespiraTox project developed a QSAR model for identifying potential human respiratory irritants using a novel in silico strategy. The goal of our strategy was to develop a reliable model based on experimental and classification data from multiple sources, and machine learning techniques to increase the confidence in the model. To date, there are several QSAR models for related human health endpoints; e.g. like skin and eye irritation. However, there are no QSAR models that predict human respiratory irritation. One reason might be, that respiratory irritants are not classified based on a specific in vivo outcome, but are usually classified within a weight of evidence approach taking into account several types of information, ranging from observation in humans to evidence from in vivo animal studies with acute or repeated inhalation exposure.

QSAR models are utilized in predicting human health hazards, data-gap filling, read-across approaches, and for screening of chemical libraries. Regulatory authorities such as the US Environmental Protection Agency's are following a directive to reduce the reliance on animal testing by 2035 (USEPA, 2020), and this has increased the need to develop QSARs for chemical hazard identification. In addition, Europe regulatory authorities (ECHA) are increasingly accepting data generated from validated QSARs for human health classification and labelling, prioritization, and risk assessment.

To develop the current respiratory irritation QSAR, we mined four human health data repositories. Most of the endpoint information is based on in vivo animal data. A case by case weight of evidence was not possible to classify compounds as irritating or not-irritating with the amount of data, but data driven approach using clearly defined criteria. By combination of different evidences using a worst case approach, the dataset may be skewed towards over representing active compounds. This imbalance, is resulting in higher uncertainty given for example by the low

specificity of the final model. The worst case approach effects the character of this model making it more useful for screening / alerts as seen by the high sensitivity. The relatively low specificity, on the other hand, indicates that the model will classify many compounds as false positive.

Because almost 50% of the endpoint annotations stem from classification databases, dose information, and severity of effects could not be used for informing the endpoint. Although, such a qualitative approach is in line with the principles outlined in the STOT-SE Cat 3 classification, a next step for model refinement is the consideration of dose-response data in a full quantitative or semi-quantitative manner to better distinguish potent from less potent compounds/compound classes, severe from slight respiratory irritants. The consideration of dose response and severity poses an additional challenge compared to endpoints for which there is one specific in vivo endpoint test, such as the OECD 405 in vivo test to assess eye irritation.

Compounds with curated endpoint annotations where filtered for defined structural properties to ensure a homogenous dataset for training. The model trained in this research is only applicable to organic compounds due to the selection criteria used. The resulting test set comprised 1997 unique organic compounds. The model doesn't seem to hugely improve upon tuning the parameters, it is mainly driven by the input dataset and the learning techniques used (compare: logistic regression / tree based learners).

To construct the human respiratory irritation QSAR models we employed RFs and GBTs as underlying machine learning algorithms. Both approaches addressed the imbalance in our training datasets. GBTs are similar to RFs as they build models containing a number of Decision Trees (DTs). To assess the performance of each of the different algorithm and dataset setups, we performed a cross validation. Irritant and non-irritant compounds were selected for training and testing. The area under the ROC curve (AUC) for LR using the combined feature set was 0.65,

559 performance for GBTs was 0.71. The applicability domain was determined by features with the
560 highest impact on the final model.
561
562 For illustration of the entire model, the RespiraTox QSAR model was applied exemplarily to tert-
563 butyl isopropyl ether. The QSAR model predicted tert-butyl isopropyl ether to be a respiratory
564 irritant.
565 To further evaluate the obtained prediction, the RespiraTox model provides the structurally closest
566 neighbors to the query compounds and their underlying dataset. The RespiraTox training set
567 contains multiple structural analogues which could be used to read- information across in a many-
568 to one manner. Based on a read-across hypothesis, the user is able to analyze the activity of the
569 most relevant neighboring compounds to gain more confidence in the obtained prediction. Two
570 read-across hypothesis were explored exemplarily, i) considering structurally closest compounds
571 that differed only by the side chain length of the ether, or ii) assuming that the metabolites formed
572 by hydrolysis are the drivers of respiratory irritation. In both cases, the prediction aligns with the
573 experimental data of the analogues/metabolites.
574
575 ***Conclusion***
576 The QSAR model was developed to predict human respiratory irritants from molecular
577 physicochemical and structural information and follows the OECD QSAR principles. The tool
578 informs the investigator whether the prediction is reliable and if the test structure is within the
579 applicability domain of the training dataset. As with other QSAR models generated by machine
580 learning it is important to acknowledge that the classifiers maintain their levels of accuracy for
581 molecules structurally similar to the chemicals used during training of the model. It is an integral

part that investigators use applicability domain in their analysis. A list of structurally similar neighbours within the database is presented for further consideration by the investigator. The result of the prediction can be downloaded in a common spreadsheet format for documentation, and as a starting point for expert review, and read across.

Acknowledgements

The authors would like to acknowledge the NC3Rs CRACK-IT program for funding this research, and Shell for sponsoring the research. AK was partially supported by the Medical Research Council grant MR/S003991/1. The authors would like to thank Yves C. Alarie for providing Threshold Limit Value datasets and for his helpful discussions on the topic.

References

ACGIH, 2020. TLVs and BEIs: Based on the Documentation of the Threshold Limit Values for Chemical Substances and Physical Agents & Biological Exposure Indices. American Conference of Governmental Industrial Hygienists, Cincinnati, OH.

Afantitis, A., et al., 2011. Ligand - based virtual screening procedure for the prediction and the identification of novel β-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks. European Journal of Medicinal Chemistry. 46**,** 497-508.

Alarie, Y., 1966. Irritating Properties of Airborne Materials to the Upper Respiratory Tract. Archives of Environmental Health: An International Journal. 13**,** 433-449.

Alarie, Y., QSARs to Replace RD50 for Inhalation of Volatile Organic Chemicals. 2016.

Alarie, Y., et al., 2001. Animal Bioassays for Evaluation of Indoor Air Quality. in: Spengler, J. D., et al., (Eds.), Indoor Air Quality Handbook. McGraw-Hill Education, New York, pp. 23.1-23.49.

Alarie, Y., et al., 1996. Estimating the sensory irritating potency of airborne nonreactive volatile organic chemicals and their mixtures. SAR and QSAR in Environmental Research. 5**,** 151-165.

ASTM, Standard Test Method for Estimating Sensory Irritancy of Airborne Chemicals. Vol. ASTM E981 - 04. ASTM International, West Conshohocken, PA, 2004.

Barber, C., et al., 2015. Establishing best practise in the application of expert review of mutagenicity under ICH M7. Regulatory Toxicology and Pharmacology. 73**,** 367-377.

Bitsch, A., et al., 2006. REPDOSE: A database on repeated dose toxicity studies of commercial chemicals—A multifunctional tool. Regulatory Toxicology and Pharmacology. 46**,** 202-210.

Brüning, T., et al., 2014. Sensory irritation as a basis for setting occupational exposure limits. Archives of Toxicology. 88**,** 1855-1879.

Clippinger, A. J., et al., 2018. Alternative approaches for acute inhalation toxicity testing to address global regulatory and non-regulatory data requirements: An international workshop report. Toxicology in Vitro. 48**,** 53-70.

Demchuk, E., et al., 2011. SAR/QSAR methods in public health practice. Toxicology and Applied Pharmacology. 254**,** 192-197.

Escher, S. E., et al., 2019. Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project. Archives of Toxicology. 93**,** 3643-3667.

Escher, S. E., et al., 2010. Evaluation of inhalation TTC values with the database RepDose. Regulatory Toxicology and Pharmacology. 58**,** 259-274.

Federal Register, Request for Data and Information on Technologies Used To Identify Substances With the Potential To Cause Acute Systemic Toxicity. Vol. 81 FR 46696. National Institutes of Health,, 2016.

Gupta, S., et al., 2015. Estimating sensory irritation potency of volatile organic chemicals using QSARs based on decision tree methods for regulatory purpose. Ecotoxicology. 24**,** 873-886.

Kane, L. E., et al., 1979. A short-term test to predict acceptable levels of exposure to airborne sensory irritants. Am Ind Hyg Assoc J. 40**,** 207-29.

Kuwabara, Y., et al., 2007. Evaluation and application of the RD50 for determining acceptable exposure levels of airborne sensory irritants for the general public. Environmental health perspectives. 115**,** 1609-1616.
Levandowsky, M., Winter, D., 1971. Distance between Sets. Nature. 234**,** 34-35.
Luechtefeld, T., et al., 2018. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. Toxicology Research. 7**,** 732-744.
Melagraki, G., et al., 2010. In Silico Exploration for Identifying Structure–Activity Relationship of MEK Inhibition and Oral Bioavailability for Isothiazole Derivatives. Chemical Biology & Drug Design. 76**,** 397-406.
Morgan, H. L., 1965. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. Journal of Chemical Documentation. 5**,** 107-113.
Nielsen, G. D., Wolkoff, P., 2017. Evaluation of airborne sensory irritants for setting exposure limits or guidelines: a systematic approach. Regulatory Toxicology and Pharmacology. 90**,** 308-317.
OECD QSAR, Guidance document on the validation of (quantitative)structure-activity relationships [(Q)SAR]. 2007.
Raevsky, O., et al., 2011. Prediction of acute rodent toxicity on the basis of chemical structure and physicochemical similarity. molecular Informatics. 30**,** 267-275.
Rogers, D., Hahn, M., 2010. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 50**,** 742-754.
Schaper, M., 1993. Development of a database for sensory irritants and its use in establishing occupational exposure limits. American Industrial Hygiene Association Journal. 54**,** 488-544.
Tversky, A., 1977. Features of similarity. Psychological Review. 84**,** 327-352.
UN GHS, Globally Harmonized System of Classification and Labelling of Chemicals (GHS), 8th edition. 2019.
USEPA, New approach methods work plan: Reducing use of animals in chemical testing. Washington, DC, 2020.
Veith, G., et al., 2009. A baseline inhalation toxicity model for narcosis in mammals. SAR and QSAR in Environmental Research. 20**,** 567-578.
Zhang, S., et al., 2006. A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. Journal of Chemical Information and Modeling. 46**,** 1984-1995.

Web references

Experimental Animal Studies

Acute Toxcity Inhalation Studies – ECHA (echa – acute tox) - https://echa.europa.eu/ Accessed: 22.02.2018 (19.767 registered substances)

Repeated Dose Inhalation Studies – Fraunhofer (RepDose) - https://repdose.item.fraunhofer.de/

4_Table 2_2015 List of TLVs for chemicals with URT irr basis with no RD50 values (alarie essay) - http://www.pitt.edu/~rd50/

Irritation Endpoints – ECHA (echa –irritation) - https://echa.europa.eu/ Accessed: Sep. 2018

Acute Tox Inhalation Studies – Fraunhofer (AcuTox) - CHEMID Accessed: 2016

Classification sources

Hazardous Substance Inventory – Toxnet (HSDB) - https://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm

C&L Harmonized – ECHA (echa – C&L H) - https://echa.europa.eu/de/regulations/clp/harmonised-classification-and-labelling

OCED Test Guideline - https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2009)28/rev1&doclanguage=en

EPA NAM - https://www.epa.gov/sites/production/files/2020-06/documents/epa_nam_work_plan.pdf