Erschienen in: Toxicology Research ; 7 (2018), 5. - S. 732-744.
https://dx.doi.org/10.1039/C8TX00051D

Big-data and machine learning to revamp computational toxicology and its use in risk assessment

Thomas Luechtefeld,[a] Craig Rowlands[b] and Thomas Hartung *[a]

The creation of large toxicological databases and advances in machine-learning techniques have empowered computational approaches in toxicology. Work with these large databases based on regulatory data has allowed reproducibility assessment of animal models, which highlight weaknesses in traditional *in vivo* methods. This should lower the bars for the introduction of new approaches and represents a benchmark that is achievable for any alternative method validated against these methods. Quantitative Structure Activity Relationships (QSAR) models for skin sensitization, eye irritation, and other human health hazards based on these big databases, however, also have made apparent some of the challenges facing computational modeling, including validation challenges, model interpretation issues, and model selection issues. A first implementation of machine learning-based predictions termed REACH*across* achieved unprecedented sensitivities of >80% with specificities >70% in predicting the six most common acute and topical hazards covering about two thirds of the chemical universe. While this is awaiting formal validation, it demonstrates the new quality introduced by big data and modern data-mining technologies. The rapid increase in the diversity and number of computational models, as well as the data they are based on, create challenges and opportunities for the use of computational methods.

Introduction

The growth of chemical property data and machine learning advances in the past decade provide opportunities and pitfalls in the field of cheminformatics. They promise to complement the current slow, expensive and animal-consuming system with rapid automated estimates available at low costs. Machine learning methods can tailor predictions away from animals to human-relevant results. Successful cheminformatic applications will transform environmental health and the chemical industry by increasing our knowledge of chemical hazards/properties. However, misapplication of cheminformatic models may result in inappropriate use or regulation of dangerous chemicals. Environmental health must grapple with these changes by delivering new methods of high quality to properly validate such cheminformatic models and apply their predictions only where appropriate. These models can only little improve on the quality of input data and predictions will often fail where input data are scarce or lacking as well when the complexity of phenomena, data and our mechanistic understanding do not match.

An additional challenge for regulatory efforts is the large number of chemicals and products without robust safety testing data. Current approaches to health and environmental hazard identification rely primarily on costly animal testing, which is insufficient to address the growing number of untested chemical products.[1]

Regulatory efforts such as REACH in Europe and the reauthorized Toxic Substance Control Act (TSCA) in the US aim to modernize chemical regulation but face an enormous backlog of testing.[2–4] The regulated industries need ways to reduce the cost and accelerate the discovery of chemical properties/hazards. Animal tests are simply too expensive and too slow to address the current excess of untested chemicals. While cheminformatics models seem like an ideal solution to this problem, it is not clear exactly how they should be applied to evaluating untested chemicals.[5,6] The methods available are rather limited with respect to applicability domain and validation/regulatory acceptance status.[7,8] In this article, we discuss some of the environmental health challenges of cheminformatics and consider some of the solutions to these challenges.

Recent public policy changes have increased pressure on regulatory agencies to control risks associated with chemical

[a]*Center for Alternatives to Animal Testing at Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA.*
E-mail: THartun1@jhu.edu
[b]*Underwriters Laboratories (UL), UL Product Supply Chain Intelligence, 333 Pfingsten Road, Northbrook, IL 60062, USA*

exposures. Global reporting on health and environmental dangers associated with air pollution, oil spills, water contamination, and other public health problems has increased awareness of the diverse dangers associated with chemical exposures. Agencies are faced with public demand for scrutinization of all chemical exposures. This is an international process with new and emerging policies in the EU, USA, Canada, Turkey, Korea, Taiwan, China, India, and more to follow. Due to only most recent advances of cheminformatics, the foreseen contribution of models to these programs is very different, but the dimensions of the task, *i.e.* all together more than 100 000 chemicals on the market and in products (not even including natural and degradation products) to be comprehensively assessed and more than 1000 added per year (*i.e.* the number of pre-marketing notifications in the US) to be evaluated too, calls for supportive screening tools to focus testing resources on cases of uncertainty and probable risk.

These programs also generate tremendous amounts of quality-assured data, which – at least in case of REACH – are also made public, at least as robust summaries. This creates for the first time truly "Big Data" sets in toxicology, which can be mined with machine learning algorithms, often referred to now as Artificial Intelligence. Fig. 1 depicts the new opportunities this paradigm shift makes possible.

The validity of cheminformatic models must be determined before integration into regulatory or commercial decision-making toolboxes. Due to the commercial value of successful cheminformatic models, developers are incentivized to build models that appear to perform well. Thus, evaluators are challenged to build model validation methods that cannot be spoofed.

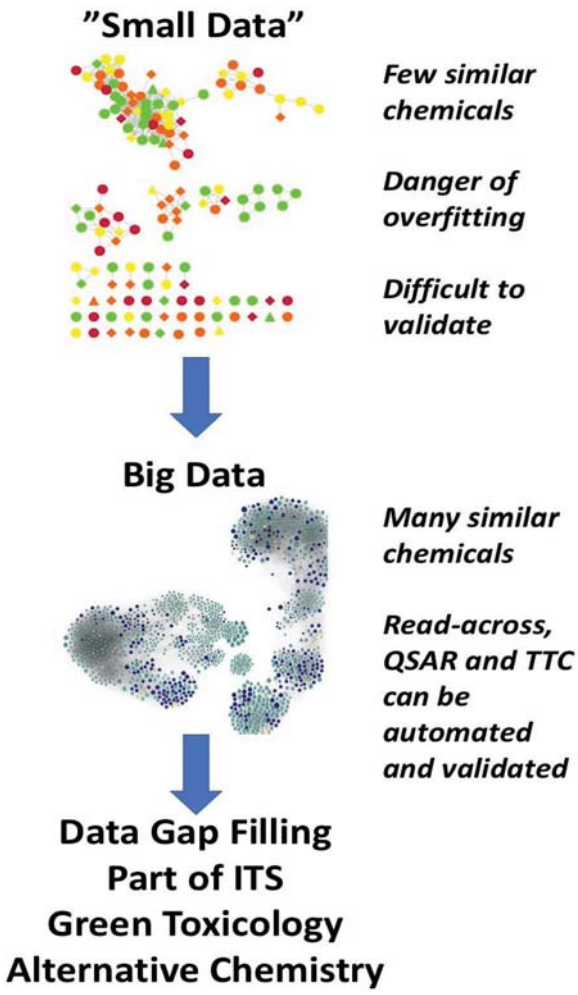

Fig. 1 Illustration of the novel opportunities resulting from the transition to Big Data in toxicology. The two graphs are taken from our earlier work on skin sensitization. The upper one modified from Luechtefeld *et al.* (2015) represents a dataset of 145 chemicals, which is a size typically used in most model developments up to now.[9] The lower one is taken from Luechtefeld *et al.* (2016) compiling about 3000 chemicals with almost 9000 animal studies.[10] ITS = Integrated Testing Strategies. Computational models can be integrated into broader testing strategies with access to sufficient training data. TTC = Threshold of toxicological concern. Large datasets allow for greater confidence in definition minimum levels of exposure, below which toxicity is not a concern for entire categories of chemicals.

Summary of our recent work and its status relative to real-world applications

Our recent work has contributed to the development of cheminformatics first by creating the – at the time – largest toxicity database, which included about 10 000 chemicals with 800 000 associated toxicological studies.[11] This was possible, by downloading and making machine-readable by natural language processing the emerging public database from the European REACH legislation hosted by the European Chemicals Agency (ECHA). Addressing some of the especially data-rich endpoints, *i.e.* acute oral toxicity, eye irritation and skin sensitization, the value of such data-sets was illustrated. This ranks from descriptive statistics on its contents, reproducibility analysis for multiple time tested compounds to QSAR.[10,12,13] The challenge of developing and exploiting such chemical databases has been reviewed recently.[14]

The availability of large chemical databases often allows identification of highly similar compounds. This allows inference of respective properties in a process that is called read-across.[15,16] Parallel to this development, we contributed to the generation of Good Read-Across Practices.[17,18] Large databases allow even automated types of read-across as explored by Shah *et al.* and in Luechtefeld *et al.*[14,19,20] The obvious emerging opportunities of such QSAR within integrated testing strategies (ITS), for Thresholds of Toxicological Concern (TTC) and Green Toxicology will be discussed later.[5,21–26]

The first publication based on the organized ECHA-REACH data describes how in December 2014 the European Chemicals Agency (ECHA) public dataset of *in vivo* and *in vitro* toxicity tests was converted into a structured, machine-readable and searchable database using language pattern matching and many database and web manipulation packages.[11] It con-

tains data for 9801 unique substances, 3609 unique study descriptions and 816 048 study documents. This allows exploring toxicological data on a scale far larger than previously possible. Data-extraction by natural language processing has obvious limitations, *i.e.* some data will be missed or extracted with mistakes. The size of the database, however, does not allow systematic human curation; some snap sampling suggests overall high quality, but possible impact on individual analyses cannot be excluded and might warrant specific checks.

While many positive responses resulted, when the creation of the database was announced at AAAS in 2016, some concerns about the proprietary nature of REACH dossiers remained at ECHA.[27,28] This has been sorted out and ECHA published an even larger database of about 15 000 chemicals in March 2017 and clarified that this data can be used for other chemicals representing legitimate access for structure–activity relationships. However, our processed machine-readable database cannot be made generally available beyond select collaborations. This is unfortunate as such a database should be jointly expanded and curated by the scientific community. A case is made that REACH should systematically open regulatory data for research purposes.

Substance similarity analysis was used to determine clustering of substances for hazards by mapping to PubChem. Similarity was measured using PubChem 2D conformational substructure fingerprints, which were compared *via* the Tanimoto metric. Tanimoto is the most commonly used but only one of many similarity metrics as discussed.[14] A comparison and possible combination of different metrics is a promising expansion of this work. Following K-Core filtration, the Blondel *et al.* clustering algorithm was used to identify chemical modules. This analysis illustrates that the chemical universe is not evenly distributed but clustered, which is helpful for finding similar structures.[29]

A first rough analysis addressed the prevalence of toxic properties. The Global Harmonized System of Classification and Labeling provides a valuable information source for hazard analysis. The most prevalent hazards are H317 "May cause an allergic skin reaction" with 20% and H318 "Causes serious eye damage" with 17% positive substances. These prevalences appear rather low, which might be biased though by the fact that mainly high-production volume chemicals were registered so far. Such prevalences obtained for all hazards here are key for the design of integrated testing strategies.[30] The data allowed estimation of animal use; it appears that the number of animals used was lower in the late 2000s relative to the 1990s, but a possible impact of REACH cannot yet be seen as the registration only foresees testing proposals no testing for the demanding endpoints at the registrations covered in this analysis.[31]

Direct comparison of the ECHA database to existing chemical databases shows overlap with other databases, which illustrates the value to support joint analyses, especially by providing test guideline study reference data from this database for those curating new approach method data. The database covers about 20% of substances in the high-throughput biological assay database Tox21 (1737 substances) and has a 917-substance overlap with the Comparative Toxicogenomics Database (~7% of CTD). The biological data available in these datasets combined with ECHA *in vivo* endpoints have enormous modeling potential. Until now, training sets for QSAR typically comprised maximally a few hundred chemicals.

The extracted ECHA dataset allows us to better understand the landscape of substances for a given hazard. Though biased by the production volume triggering the first two registration deadlines, the low prevalence of hazards in the extracted database challenges the famous notion of Paracelsus that "all things are poison". The database represented the largest machine-readable resource especially for *in vivo* toxicity data at the time. The very well-curated EPA database ToxRefDB, commonly used animal testing database, covers only 474 substances with multiple animal endpoints while our extraction covers over 9800.

The second paper in this series analyzed acute oral toxicity data, one of the most common tests, as it serves among others work place safety and transport requirement regulations.[12] The database included a total of 13 832 oral toxicity studies for 8568 substances (up to December 2014). Seventy-five percent of studies were from the retired OECD Test Guideline 401 (11% TG 420, 11% TG 423, and 1.5% TG 425). This suggests that most studies were carried out before 2001, which is in line with the fact that most chemicals are high-production volume chemicals, for which such data are typically available. It does not yet allow understanding of the time the adaption of the new-tiered testing strategies takes. Noteworthy, 76% of chemicals were not orally toxic in rats at the limit of 2 g kg^{-1}. Concordance across guidelines, evaluated by comparing LD_{50} values $\geq$2000 or <2000 mg kg^{-1} bodyweight from chemicals tested multiple times between different guidelines, was at least 75% and for their own repetition more than 90%. This is quite remarkable, as alternative methods to predict acute toxicity from cytotoxicity data achieved similar predictivities but were not found to be acceptable as a replacement.[32] This is another example of how we overestimate the reproducibility and concordance of animal tests.[33]

In 2009, Bulgheroni *et al.* created a simple model for predicting acute oral toxicity using no observed adverse effect levels (NOAEL) from 28-day repeated dose toxicity studies in rats.[34] This was reproduced here for 1625 substances. This analysis has in the meantime been repeated and expanded by ECHA and ECHA accepted formally this approach stating that "Ahead of the last REACH registration deadline, in 2018, ECHA estimates that registrants of about 550 substances can omit the *in vivo* acute oral toxicity study by using this adaptation".[35] Our analysis thus directly contributed to animal saving.

In 2014, Taylor *et al.* suggested no added value of the 90-day repeated dose oral toxicity test given the availability of a low 28-day study with some constraints.[36] We confirmed that the 28-day NOAEL is predictive (albeit imperfectly) of 90-day NOAELs, however, the suggested constraints did not affect predictivity. This is another example how a large database can

support the analysis of proposed changes to testing regimens, though here less satisfying. A lot of this, though, seems to be owed to the inconsistency of 28-day and 90-day studies in general (see Fig. 4 of the 2016 publication).[12]

The large dataset allowed the exploration of data modeling and prediction exercises. One thousand fifty nine substances with acute oral toxicity data (268 positives, 791 negatives, all Klimisch score 1) were used for modeling: first, instance-based learning was applied, *i.e.* the k-nearest neighbors (KNN) algorithm, which indicated that similarity-based approaches alone may be poor predictors of acute oral toxicity: evaluation of KNN on the global dataset results in a balanced accuracy of 71% with higher specificity than sensitivity (89% *versus* 54%) – a skew that is largely the consequence of the high prevalence of non-toxicants. Our use of the simplistic PubChem 2D fingerprint is one source for this poor predictive power. To illustrate the combination of a KNN feature with chemical descriptors, the Chemical Development Kit was used to generate 27 molecular descriptors and we built a simple decision tree using the KNN feature and the next most informative feature, as determined by the Ranker approach to feature importance. This decision tree illustrates how instance-based learning can be used in concert with supervised learning methods *via* feature generation. Although decision trees are fundamentally limited in their expressive power (they can only model endpoints *via* a conjunction of feature values), this model serves to demonstrate the possibility to improve sensitivity by fusing supervised and instance-based learning. This data can also be used in feed-forward artificial neural networks (multilayer perceptron or MLP) algorithms, which can build more expressive feature relationships for toxicity prediction than decision trees. However, MLP models are more difficult to visualize. Rather than only capturing conjunctions of feature values, multilayer perceptrons can model a wide variety of feature relationships. This "universal approximator" property is important for modeling potentially complex relationships of the sort chemical descriptors may have for predicting toxicological endpoints. One consequence of the increased expression of multilayer perceptrons is the risk of overfitting. When training and testing multilayer perceptron on the entire dataset the resulting sensitivity, specificity and balanced accuracy is 94%, 99.6% and 97%, respectively. To account for overfitting, we performed stratified 10-fold cross-validation. Noteworthy, in the multilayer perceptron model, KNN was the feature with highest information value. The stratified results show 71% sensitivity and 72% specificity for the multilayer perceptron, which might not sound exciting, but is at the level of concordance of different OECD protocols against each other. Obviously, it is difficult to improve much over such data inconsistencies, *i.e.* as often stated "trash in, trash out". The three modeling approaches demonstrate how predictions can be carried out. It is quite likely that additional descriptors – for example, corrosivity or sensitization – as well as a more detailed analysis of either ADMET or *in vitro* assays of biological activity – could substantially improve the models.

Eye irritation hazard, for which the rabbit Draize eye test still represents the reference method, was analyzed next.[13] Dossiers contained 9782 Draize eye studies on 3420 unique substances, indicating frequent retesting of substances. Two chemicals were found, which were tested more than 90 times in rabbit eyes, 69 were tested more than 45 times. This allowed assessment of the test's reproducibility based on all substances tested more than once. There was a 10% chance of a non-irritant and a 20% chance for mild-irritant evaluation after a prior severe-irritant result. The most reproducible outcomes were negative (94% reproducible) and severe eye irritant (73% reproducible). 34% of the substances were eye irritants, somewhat higher than suggested in an earlier analysis of the New Substances Database of the former ECB with 17.4% eye irritants showing differences in the type of substances registered between 1981 and 2008 and those under REACH in the initial phase, *i.e.*, predominantly high-production volume substances.[37] They also confirm the reproducibility issues already described by Weil and Scala in 1971; very often this problem has been belittled by stating that these studies were done before OECD guideline standardization and GLP.[26] They also confirm the assessments by Adriaens *et al.* about the test's reproducibility.[37] Their database includes fewer substances, but had access to the raw data, allowing intra-assay variability assessment. This demonstrates the extent to which access to the full REACH datasets could strengthen assessments.

To evaluate whether other GHS categorizations predict eye irritation, we built a dataset of 5629 substances (1931 "irritant" and 3698 "non-irritant"). The two best decision trees with up to three other GHS classifications resulted in balanced accuracies of 68% and 73%, *i.e.*, in the rank order of the Draize rabbit eye test reproducibility itself, but both use inhalation toxicity data ("May cause respiratory irritation"), which is not typically available. However, the approach shows that different toxic properties inform each other and that – as typically done – staying within data for a single hazard misses an opportunity to improve predictions.

Next, a dataset of 929 substances with at least one Draize eye study was mapped to PubChem to compute chemical similarity using 2D conformational fingerprints and Tanimoto similarity. Using a minimum similarity of 0.7 and simple classification by the closest chemical neighbor resulted in balanced accuracy from 73% over 737 substances to 100% at a threshold of 0.975 over 41 substances. Thus, the more similar chemical neighbors with data are available, the more precise the prediction, hinting to the value of big datasets. This represents a strong support of read-across and (Q)SAR approaches in this area.

The preliminary analysis and mining of the dataset shows that there is both considerable predictivity from chemical structure (our analysis based on the closest chemical neighbor with data) and biological activity (our analysis based on other GHS classifications). Neither alone has adequate accuracy to supplant the Draize eye test, although given the reproducibility problems of the assay, this statement might be contested. Here, no attempt was made to use the information from chemico-physical properties, dedicated *in vitro* assays for eye irritation, toxicokinetic information or biological profiling as

attempted in ToxCast or the Tox21 program, all of which could likely considerably boost the predictive value of the knowledge-base. Follow-up research should focus on the integration of external databases with the ECHA data to create stronger models for eye irritation. The relatively impressive predictive value of the naïve approaches attempted here, however, strongly supports read-across and *in silico* approaches.[15]

The public data in the database included 19 111 studies on skin sensitization, making it the largest repository of such data so far (1470 substances with mouse LLNA, 2787 with GPMT, 762 with both *in vivo* and *in vitro* and 139 with only *in vitro* data).[10] 21% were classified as sensitizers, considerably lower than expected (about 35% according to earlier sources). The extracted skin sensitization data was analyzed to identify relationships in skin sensitization guidelines, visualize structural relationships of sensitizers, and build models to predict sensitization. Altogether, reproducibility of and between different OECD test guidelines *in vivo* ranged from 77% to 95% with 89% for the LLNA to reproduce itself. This means we can expect no alternative method to be better than this in direct comparisons if used on the same sets of substances. It is important to note that the datasets used to evaluate the reproducibility between tests do not contain the same substances and for this reason percentage agreement should not be considered a direct comparison.

Structural alerts for skin sensitization identify substructures predictive for substance reactivity and sensitization proclivity. Distribution of structural alerts in chemical clusters (modules) revealed wide variation. Approximately 31% of mapped chemicals are Michael's acceptors but alone this does not imply skin sensitization. This opens up for more detailed analyses and *e.g.* fine-tuning of thresholds of toxicological concern (TTC).[22]

A chemical with molecular weight >500 Da is generally considered non-sensitizing owing to low bioavailability, but 49 sensitizing chemicals with a molecular weight >500 Da were found in line with parallel work by Fitzpatrick *et al.* (2017).[38] Again, this illustrates how big data can identify popular myths and quality-control rule-based models. A chemical similarity map was produced using PubChem's 2D Tanimoto similarity metric and Gephi force layout visualization. Nine clusters of chemicals were identified by Blondel's module recognition algorithm revealing wide module-dependent variation.[29] A simple sensitization model using molecular weight and five ToxTree structural alerts showed a balanced accuracy of 66% (specificity 80%, sensitivity 51%), demonstrating that structural alerts have information value.

A simple variant of k-nearest neighbors outperformed the ToxTree approach even at 75% similarity threshold (balanced accuracy 68%). At higher thresholds, the balanced accuracy increased (82% balanced accuracy at 0.95 threshold). Lower similarity thresholds decrease sensitivity faster than specificity. This analysis scopes the landscape of chemical skin sensitization, demonstrating the value of large public datasets for health hazard prediction. The prediction of a binary outcome (sensitizer *vs.* non-sensitizer) in this article was necessitated by failing to extract potency information where available in ECHA dossiers. The available *in vitro* data in the database have not been analyzed and exploited yet. The promising predictivity of rather naïve prediction models from chemical neighbors suggests that such advanced predictions could actually bring predictions into the range of *in vivo* reproducibility.[9]

Rarely, *in silico* approaches will satisfy regulatory information needs as a stand-alone approach, if not validated for large datasets. Increasingly, the need for systematic integration of different information sources as Integrated Testing Strategies (ITS) is recognized.[39,40] Supervised learning methods promise to improve ITS, but must be adjusted to handle high dimensionality and dose–response data. ITS approaches are currently fueled by the increasing mechanistic understanding of adverse outcome pathways (AOP) and the development of tests reflecting these mechanisms.[41] Simple approaches to combine skin sensitization data sets, such as weight of evidence, fail due to problems in information redundancy and high dimensionality.

The problem is further amplified when potency information (dose/response) of hazards would be estimated. Skin sensitization currently serves as the foster child for AOP and ITS development, as legislative pressures combined with a very good mechanistic understanding of contact dermatitis have led to test development and relatively large high-quality data sets. In work, which preceded the generation of the ECHA dataset, we curated such a dataset of 145 chemicals with various *in vitro* assay data.[9]

Recursive feature elimination involves first building a model on a large number of features, then eliminating features that do not impact model accuracy. Feature importance was calculated with the scikit-learns implementation of the Breiman random forest variable importance algorithm. We used this recursive variable selection algorithm to evaluate the information available through *in silico*, *in chemico* and *in vitro* assays. Chemical similarity alone could not cluster chemicals' potency, and *in vitro* models consistently ranked high in recursive feature elimination. This allows reducing the number of tests included in an ITS. Feature elimination can be useful for determining the mechanistic pathways behind toxicity. We saw the KeratinoSens and direct peptide reactivity assay (DPRA) as strong models for skin sensitization as they resulted in the top three features in all datasets. The chemical descriptors most heavily represented in variable importance included many descriptors of electrophilicity, molecular weight and descriptors related to the ability to penetrate skin. These descriptors make sense in the absence of skin permeability information provided from *in vitro* assays.

Next, we analyzed the data with a hidden Markov model, which allows us to enforce proper prediction series by encoding our knowledge of allowable toxicity transformations. It takes advantage of an intrinsic inter-relationship among the local lymph node assay classes, *i.e.* the monotonic connection between local lymph node assay and dose. The dose-informed random forest/hidden Markov model was superior to the dose-naive random forest model on all data sets. Although balanced accuracy improvement may seem small, this obscures the

actual improvement in misclassifications as the dose-informed hidden Markov model strongly reduced "false-negatives" (*i.e.* extreme sensitizers as non-sensitizer) on all data sets. This dose-informed approach already shows promising improvements in off-by-one accuracy and average class error. It can be improved by using descriptors with dose–response data and by accounting for dermal penetration data. Dermal penetration has the potential to change the proposed dose-informed model, by adjusting the "effective" dose of a given toxicant to account for penetration features.[42]

Some probabilistic models already exist for skin sensitization, including the Bayesian ITS.[43] The Bayesian ITS shows a remarkable balanced accuracy (94% on a small external test set). While this and some other approaches show high accuracies on test sets, some failed to use cross-validation. Here, for the first time, a successful cross-validation was included. However, the possibility remains that accuracy on test sets is not representative of the method. The Bayesian ITS method, while demonstrating strong accuracy and a valuable approach (Bayesian networks), also used the TIMES QSAR, which may cause problems due to peeking (*i.e.* that some chemicals were in the training set already). Our approach improves over existing models by incorporating structure/activity relationship data without using QSARs and by proposing a method to incorporate the range of dose–response data rather than summary statistics alone. Combining these approaches with new data sets will be interesting for future research.

Current computational approaches to support closing data-gaps

Computational models are primarily in the proposal process with many new models recently published. Some models have been applied to chemical regulation for chemical labeling and many authors suggest their use in integrated testing strategies, as we discussed earlier in a whitepaper and workshop report, for prioritization of chemical testing.[39,40] Models have long been in medical use for screening and drug design – practices that are directly relevant to potential uses in accelerating untested chemical evaluation.[44]

A most promising approach to integrated chemical testing uses Bayesian networks to (1) predict for chemical properties and (2) estimate the value of information provided by tests in the network. The second property allows for construction of testing policies that iteratively select the most informative test for individual chemicals. In this approach, a chemical with incomplete testing can be assessed and a probability/confidence of hazard determined. If confidence is too low, the model can be used to suggest new required testing. This potential application is far ahead of the traditional 'weight-of-evidence' approach.[45,46]

Future directions

Testing policies define, which test to do given the current state of chemical testing. Chemical prioritization is a 'testing policy', in which chemicals are tested according to their probability of hazard (as could be determined by a supervised learning model). Some proposals of 'Integrated Approaches to Testing and Assessment' (IATA), a term favored by OECD over Integrated Testing Strategies in recent years, describe paradigms where computational models can be integrated into testing requirements even combining them with the emerging Adverse Outcome Pathway concept of mechanistic toxicological knowledge.[41] These proposals are promising for future integration of computational models into regulatory decisions. However, there may be opportunities to use models already earlier and at a larger scale in regulatory testing.

In an ideal world, there should be a co-evolution of testing and test method development. Sometimes, testing specific chemicals improves our knowledge about a certain testing method more than the testing required of substances because of legislation. Here, programs such as the US National Toxicology Program could come into play, though their selection is driven more by data-gaps and concerns as to these substances. Validation of test methods is an example – though typically not for traditional animal tests – where systematic testing to evaluate a method is undertaken.[47] The field of reinforcement learning and optimal experimental design demonstrate that machine learning models can be used to construct powerful polices of much greater complexity than typical approaches. To optimize policy definitions for chemical testing a state S and an evaluation function $F: S \rightarrow R$ needs to be defined. The state should define all the variables of concern in chemical testing (enumerate tested chemicals, aggregate total costs and time, *etc.*). The evaluation function merges these variables into a numeric score. Balancing these variables is difficult, if money and time were not a concern then a policy may suggest testing all chemical sequentially but doing so may be prohibitively expensive or slow. Improved testing policies can direct the considerable growth rate of chemical databases to improve models faster, increase economic growth, and reduce harm. This could represent an example for the strategic development of safety sciences.[48]

When chemicals are tested, they provide data that can improve cheminformatic models. Thus an important consideration for testing policies is to maximize the rate of model improvement. When testing policies are allowed to maximize model training they can be used to address the causality weakness seen in correlative supervised learning algorithms. One simple approach to maximize model learning is to select chemicals with structures that are unlike chemicals in model training sets. This should be balanced against the economic benefits of such testing. This problem is very similar to the multi-armed bandit problem – where a gambler is given a choice of slot machines and a limited set of tries and must decide, which machines to play, which illustrates the value of exploration (trying different machines) *versus* exploitation (playing a profitable machine repeatedly).

Evaluating animal models

The potential risks associated with a compound define the testing requirements for that compound. For example, chemi-

cal tonnage and production environment are factors that determine testing requirements for a given compound. Required tests are selected to identify chemicals, which can be regulated to minimize the potential economic/public health damage. These tests must be evaluated to ensure that they can accurately categorize chemicals for regulation. This process is called validation, which is under constant evolution.[47,49,50]

Chemical properties are not always known, if they were then chemical testing would not be necessary. For example, one of the most tested hazard in humans is skin sensitization with variants of the patch-test: less than 1000 chemicals have been deliberately tested in humans to determine skin sensitization. Many of the health hazards have extremely little 'gold standard' human data due to the ethical impossibility of performing human tests. Conventional techniques of model validation are severely handicapped by the lack of target data.

Alternative models such as computational or *in vitro* models are frequently validated based on their ability to predict the results of animal models. When the animal models are invalid or irreproducible, they fail to act as valid testing data for alternative models.[33,51]

Current approaches

In our work, we have evaluated animal model tests *via* a pairwise method. When the same chemical has received the same animal test multiple times it can be used to determine the reproducibility of the test. Test outcomes are compared to each other for specific chemicals and the conditional probability of one test outcome given another is calculated by counting the number of pairs with the given outcomes and dividing by the total number of pairs. This approach gives an estimate of the reproducibility of an animal test. Our application of this approach to the Draize Eye Irritation Test demonstrates that some animal tests are not very reproducible.[13] Similar work expanding the approach to all acute and topic endpoints is currently in preparation for publication.

Other approaches, like most validation testing, still rely on manually selected reference data. This can inflate estimates of test validity *via* animal tests that are built to perform well on relatively small well-known reference examples.

Future directions

There are many methods for evaluating the repeatability of tests. ANOVA tests and mixed effects models try to capture differences between groups and the impact of random *versus* fixed effects. Mixed effects models are particularly interesting in the context of QSARs.[52] Mixed effects models let us define variables that we think may be affected by the properties of a chemical.

The pairwise approach used in our publications can be improved by a latent variable model. In the latent variable model, every test outcome is conditionally independent and identically distributed (c.i.i.d.) to every other test outcome given the chemical. Unlike the pairwise approach, which assumes that pairs of outcomes are independent of other test outcomes (a definitely incorrect assumption), this approach assumes that a given outcome only depends on specific effects given the same test chemical.

Unfortunately, building outcome distributions for each chemical requires testing the given chemical and does not allow to generalize over chemicals. So instead of building a repeatability model where test outcomes are c.i.i.d. given the chemical, we can only build a model where outcomes are c.i.i.d. due to unobservable property of the chemical. One could argue that this is exactly what a perfect test does; it measures some property of interest and is independent of all other properties. A test for skin sensitization should measure, whether a chemical is a sensitizer, and be independent of chemical solubility, boiling point, *etc*.

Latent models can account for test confounders by adding them as observable variables, on which the test outcomes depend. For example, tests using a solvent can only test chemicals that are soluble, if solubility is added as an observable variable, test outcomes should be less consistent for chemicals that are less soluble. These latent models are not new; they have been used in medical diagnosis, where a disease can be treated as a latent variable, upon which symptoms or diagnostic measures depend. Some approaches to estimate parameters for these distributions include expectation maximization and the concave convex procedure.[53–56]

QSAR validation sets

The validity of cheminformatic models must be determined before integration into regulatory or commercial uses.[57] The question of model accuracy is a tricky one but it is almost always evaluated on some set of test chemicals. Validation sets are used in model competitions and also to aid validation studies. Creation of a good validation set for a promising model selected is critical, as bad validation sets can result in inaccurate evaluations of models, which can be dangerous to public health and commercial applications. Validation datasets (selections of chemicals with known target values) are used to evaluate computational models. Model predictions are compared against known values and models that are correct more often are given better scores.

A good validation set should accurately represent the eventual goal of the evaluated model. If a model has a specific domain of applicability, then the validation set should be composed of chemicals in that domain. This is not as easy as it sounds: the first question is how to categorize chemicals? There are many such methods of categorization (some of which we reviewed in Luechtefeld and Hartung, 2017).[14] Chemicals can be clustered *via* distance metrics on chemical descriptors, or *via* expert defined chemical categories, or *via* their usage, or by many other approaches. How do we know what categories of chemicals to pick? How to handle multiple memberships? How do we handle model evaluation when categories do not have balanced representation?

Another major problem with validation sets is combatting overfitting. When modelers work with a training set, they train

models on the data set. If the entire validation set is revealed, models can 'overfit' the data. Evaluators must be careful to handle overfitting at the validation stage.

Size is the greatest challenge for construction of data sets. To build strong models and test them we need very large datasets (>10 000 well spread compounds to cover the entire chemical universe); until recently such datasets were not available for toxicology. Concerns over proprietary data is one reason for the small size of public toxicological test data. The testing for a validation set can be very expensive; owners of proprietary data may not be willing to add their data to a validation set due to the high cost of collection. However, also concerns of giving insights about products to competitors, liability cases and advocacy groups contribute here.

Current approaches

Different methods to validate machine learning models (cross validation, y-scrambling, *etc*.) all rely on validation sets.[58] At a high level, these methods split datasets into training and validation sets and/or modify dataset features to measure impacts on model outcomes. In our work and the work in many other publications, training and validation sets are created by random selection from some large data source for chemicals. This solution is not ideal as it is likely to have large imbalances in chemical categories. In some cases, groups will split datasets according to broad chemical categories; for toxicological testing it can for example be sensible to treat drugs separately from industrial compounds.

Future directions

The willingness to share data as well as legislative pressures to do so is increasing, especially as the different agencies worldwide realize their value. However, in order to foster data-sharing on a larger scale, some of the concerns associated with the publication of proprietary data have to be overcome. Data encryption schemes are a possible solution to problems with data sharing. New research indicates that some machine learning algorithms termed 'privacy preserving classifiers' can be built on encrypted data.[59] If this approach improves, then owners of large proprietary datasets will be able to share data without revealing testing results.

Sampling methods that seek to balance representation of different chemical groups can widen domains of applicability and ensure less biased accuracy measurements. These methods are not commonly used in the literature and rely on some definition of 'chemical grouping'.[36,60,61]

QSAR aspects of experimental design and causality

Most statistical models are based on observational data. While causal inference is an area of active research, many QSARs are difficult to interpret and cannot be used to define causal links. They represent correlation rather than causation. The lack of experimental design to test variable-causality leads to models that become biased due to coincidental variable relationships. This can be a tough problem for models that rely on large numbers of variables, particularly those that use variables that do not have easily explained relationships to the target, *i.e.* they are vulnerable to over-fitting. We have earlier proposed a concept of mechanistic validation;[62] this argument was largely targeted toward the validation of biological models with more objective tools such as systematic reviews, but it can be similarly used to argue for looking into the mechanistic basis of QSAR approaches[62–64] as attempted also in the discussions by Tollefsen *et al.*[41]

Chemical classification decision trees, for example, split a set of chemicals into different buckets. At each split a feature is selected based on data to divide chemicals. This feature selection rule is meant to divide chemicals into buckets with other chemicals sharing the same target property, but it has no means to determine causality. Even large datasets can result in non-causal but correlated features. This possibility makes regulators wary of using QSAR models (even when they perform well in testing challenges).

There is a possible danger of experimental research taking a back seat to modeling due to the speed and lower costs of QSAR-based chemical property determination. While this might be the ultimate goal of modeling efforts (to allow faster chemical labeling at lower cost), there is the danger that overfitted, non-causal models will be trusted prematurely.[8] In reverse, once strong models exist for a given endpoint, the problem is in costs and animal use associated with unnecessary testing, when delaying transition.

Current approaches

The OECD guidelines for QSARs concede that models may not have mechanistic interpretations.[25] They state that models should have "a mechanistic interpretation, if possible". The current state of software for cheminformatics and machine learning make it much easier to construct a model than it is to understand the causal role behind cheminformatic descriptors. For now, models with clear mechanistic descriptions are more likely to be accepted in regulatory applications. As more testing data is generated, models that are susceptible to overfitting on non-causal relationships will be proven untrustworthy.

Some models are more interpretable than others. Many QSARs rely on linear regression, some use decision trees. Both of these methods (particularly decision trees) are relatively easy to interpret. When decision trees are built from chemical testing data, they can even be used to optimize testing. Bayesian networks are particularly strong in their potential to build 'integrated testing strategies' due to their capacity to handle missing data and the ability to interpret networks built from relatively small numbers of features. Like decision trees, Bayesian networks can also be used to optimize testing.

Other models are much less interpretable, these include random forests, neural networks, and other complex models. Models built using these methods can be powerful predictors, but are also more prone to overfitting, especially if they are built on thousands of features.

Similarity-based models are an interesting variant. Similarity of chemicals is the basis for read-across, but can also be used to build QSAR based on similarity.[10,20] Thus, the principles of read-across and QSAR-type computing are merged here. Similarity-based models can seem deceptively interpretable. When two compounds are very 'similar', their chemical structures are likely to look the same to human eyes. This enables applications where human users evaluate the computationally identified or generated analogues to a given target compound. In collaboration with Underwriter Laboratories (UL), our approaches have been expanded to a web-based automated rad-across-based QSAR named REACH*across*™ (https://www.ulreachacross.com). By combining several public databases, the underlying dataset included in March 2017 70 + million structures, of which 300 000 had biological data and 20 000 animal data. Table 1 shows the number of chemicals with data (column 2) ranging from four thousand (mutagenicity) to fifteen thousand (eye irritation). The predictive capacity for these acute and topical endpoints was examined by a leave-one-out cross-validation (unpublished, Table 1). This means that for each and every of the several thousand chemicals with animal results (column 2) a prediction based on similar chemicals was made, presuming that no data were available; then the prediction was compared with the actual data. The sensitivity (ability to identify a toxicant) and specificity (correctness of these results) show that more than 80% of toxic chemicals were found with still reasonable specificities of 54–71%. Such predictions were possible for 64–83% of the chemicals (labeled coverage). This suggests that for two thirds of chemicals predictions can be made with accuracies on par with animal studies.

Still, these promising approaches should be used carefully, as it is easy to build similarity-based methods that hide much of the complexity in their underlying features. They require thus the same rigorous validation approach as other QSAR.[65] The respective validation with the US Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and National Toxicology Program's Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) has been initiated by FDA in 2017.

Future directions

As the barriers to model creation lower, researchers will integrate computational models into their experimental work.

Table 1 Sensitivities (Se) and specificities (Sp) for 6 health hazard models built from thousands of classification and labelling results stored on the ECHA database

Endpoint	Tested	Se	Sp	Coverage
Skin sensitization	5136	83%	55%	83%
Eye Irritation	15 214	83%	54%	79%
Acute oral	12 342	82%	71%	77%
Mutagenicity	4077	80%	58%	81%
Skin irritation/corrosion	14 718	88%	57%	64%
Acute dermal	6732	89%	70%	59%

Adversarial modeling is the practice of designing competing models. One model predicts outcomes from input data, and another model constructs input data that will trick the first model. In hazard prediction, adversarial models could be used to physically test computational models and search for causal links.

New methods for aiding in the interpretability of models are also emerging. LSTMV (http://lstm.seas.harvard.edu/) is a visual tool for investigating the hidden variables in recurrent neural networks. These tools may aid researchers in evaluating models. They may also aid researchers in discovery of new biochemical relationships.

In the future, the boundary between 'wet lab' work and computational models is likely to reduce. Pharmaceutical efforts to integrate computational screening into drug design/testing demonstrate this possibility. There is no reason why similar approaches cannot be used to predict chemical properties. If computational models take on greater industrial and public health roles, it will become more important to construct robust testing methods. An interesting OECD QSAR guideline could require that computational models be physically 'testable'.

The QSAR zoo

There are a huge number of computational models for chemical properties – casually speaking a zoo of QSAR. The Danish QSAR database reports ~300 distinct QSAR authors, many of which have published multiple QSARs. Even after deciding to accept/use computational models in a commercial or regulatory workflow, users still need to decide on which specific models to employ. Selecting models is not as simple as asking, which model has the best evaluation. Models may have varying domains of applicability, susceptibility to overfitting, expense in feature generation, interpretability and other differentiators. All of these factors must be considered when selecting a model. An important consideration when deciding on a QSAR is the ease of use and the development support. QSARs with supporting computational packages or programming interfaces are easier to integrate into new workflows.

Machine learning is an extremely active field with advancements made constantly and large numbers of packages supported by giant corporations such as Microsoft, Google, and Amazon. This progress bodes well for the eventual widespread use of modeling in chemistry, but it causes a selection problem. Given a thousand models for a given hazard, which one should be selected?

Current approaches

The simplest approach to model selection is *via* some form of model scoring (we and others have outlined several validation methods). In some cases, high-scoring models are combined into ensemble models. Ensemble methods combine multiple classifiers into one. The simplest ensemble method simply takes a majority vote from a set of models. Much more

complex ensemble models exist that can account for differences in domain of applicability. These ensembles, however, can be prone to the causality/overfitting issues mentioned earlier.

Unfortunately, there is no unifying interface for cheminformatic models. Many different softwares exist for creating models, and packages exist for creating new models in every popular programming language. Opentox.net, for example, aims to provide an 'interoperable predictive toxicology framework' with an application programming interface (API). API refers in computer programming to a set of subroutine definitions, protocols, and tools for building application software allowing remote use of certain computation services. This would allow accessing many different models remotely. There are several similar efforts. One hurdle for these efforts is resistance from model developers, who may seek to provide their own portal for customer use.

Future directions

The use of supervised learning algorithms in QSAR design is a narrow subset of the possibilities available with new machine learning and computer science research. We have already discussed the potential of optimizing experimental design, adversarial models, and aids to model interpretation. These advances also have the potential to aid in simplifying the QSAR zoo problem.

Multi-label learning

Recent advances in machine learning have resulted in models that can handle missing data and model multiple targets at once (multi-label learning, in case of toxicology for example multiple-hazard learning). These models can sometimes outperform single-label models by increasing the available data for training and by transferring concepts applicable to one label to predictions on another label. Multi-label models have the potential to simplify the QSAR space. Rather than having a model for every chemical property, a single model can predict many different chemical properties. In toxicology, many hazards are interrelated; thus, they can inform each others' predictions. For example, a skin irritant is likely also an eye irritant, which means that information on both labels synergizes. So, the prediction of one hazard (label) informs other labels for the same and similar chemicals can improve predictions.

Biological features

Biological data created from chemical experiments can provide valuable features to machine learning models. Chemical similarity or 'read-across' can be performed on biological features for finding chemicals with 'similar' biological activity.[17,66]

Biological features built from high throughput screening have also proven effective for supervised learning models. These models have demonstrated strength in hepatotoxicity, estrogen receptor binding, and broadly in animal toxicant endpoints.[67–70]

Unifying interfaces

Proprietary models are resistant to unifying interfaces due to problems of incentive. When regulatory agencies 'accept' a computational model they inadvertently give the developer a commercial edge over other developers. The recent explosion of distributed applications built on blockchain technology may be an interesting solution to this problem. The iota data marketplace (https://data.iota.org/) is one such decentralized market, where users can purchase data directly from data producers without paying a fee to any central distributor. Similar projects are used to share compute power, and even machine learning predictions in distributed marketplaces.

It may be possible to build such a decentralized marketplace on QSAR models thus providing a unifying interface for chemical property predictions. Such a marketplace could validate models and provide financial incentives commensurate to model value without requiring a centralized provider.

Other non-regulatory uses of cheminformatics

In product development, regulatory toxicology due to its costs and animal use is usually only initiated at advanced stages, when marketing is likely. However, toxic properties then come as a bad surprise challenging enormous investments. Given the limited reliability of many tools of this safety assessment, enormous losses of investments or delays (time to market) in investigating these toxicities are the only options, in order not to risk consumer health. The solution to this is to consider toxicological information earlier in the product development process. This is known as Frontloading of Toxicology in the pharmaceutical sector and as Green Toxicology in the chemical one.[21,23,24]

In silico and *in vitro* approaches are especially suited here, as they give typically such information faster and cheaper than animal studies and have no ethical concerns. *In silico* approaches can even be applied before a chemist synthesizes a new molecule. By excluding upfront, or making less likely, toxic liabilities in the process, consumer and worker safety are improved, and animals are exposed to less harmful substances in the regulatory part of safety testing. The obvious emerging opportunities of such QSAR within testing strategies and for Thresholds of Toxicological Concern (TTC) further strengthens this approach.[9,22,26] The possibly lower predictivity of these methods and lack of validation is less of concern at such early stages. What is needed, is simply a change in workflows and decision processes, *i.e.* to involve toxicological considerations earlier in product development, not in developing new science.

Furthermore, many industries do not have the profit margins allowing safety testing on par with drug or pesticide companies. *In silico* methods can help to focus resources on the more problematic compounds and hazards. Similarly, within the supply chain, those obtaining many chemicals or products containing many substances such as retailers or big consumer product companies, get more control over what they are using with ease. Also, the regulatory agencies can benefit from cheminformatics approaches: they can check the plausibility of submitted findings or where the burden of proof of possible problems is with the regulators, they can employ cheminformatics to produce evidence of such concerns. This shows how cheminformatics is possibly transforming the world of consumer products and their safety control. Reliable predictions are key – thus better data, better algorithms and rigorous validation represent the door-openers for this transformative change.

Conclusions

In this article, we explored computational approaches to chemical hazard assessment and examined real-world data in four earlier publications examining the REACH registration public data repository – the world's largest regulatory toxicology database. Rapid developments in the regulatory and industrial use of computational models promises transformative change in chemical synthesis, manufacturing, and regulation.

A first implementation in UL's REACH*across* tool validated by cross-validation with thousands of chemicals suggests unprecedented predictive capacity, but this has to be taken with a grain of salt until peer-reviewed publication and a formal independent validation.

There are many hurdles to the broad use of QSAR for toxicity prediction. However, it appears that a critical mass of data has been reached, which combined with cutting-edge machine-learning deserves a validation challenge. The available data are constantly growing – our own data-base included 10 000, ECHA published early 2017 data for 15 000 and their current website includes about 20 000 chemicals. Within short time, in May 2018, they might rise to 40–60 000 chemicals with respective data.

This work has contributed to making data available and demonstrating their scientific usefulness. This illustrates the path forward for cheminformatics in toxicology. Their practical usefulness will have to be shown, but the sector is benefitting from the fast advances in data warehousing and machine learning allowing the deployment of artificial intelligence also in the safety sciences.

Conflicts of interest

Craig Rowland is an employee of Underwriters Laboratories (UL). The other authors consult UL on computational toxicology, especially read-across, and have a share of their respective sales. Tom Luechtefeld has created ToxTrack LLC to develop such computational tools.

Acknowledgements

Thomas Luechtefeld was supported by an NIEHS training grant (T32 ES007141). This work was supported by the EU-ToxRisk project (An Integrated European "Flagship" Program Driving Mechanism-Based Toxicity Testing and Risk Assessment for the 21st Century) funded by the European Commission under the Horizon 2020 program (Grant Agreement No. 681002). Editing of the manuscript by Sean Doughty is gratefully appreciated.

Notes and references

1 T. Hartung, *Int. J. Risk Assess. Manage.*, 2017, **20**, 21–45.
2 T. Hartung and C. Rovida, *Nature*, 2009, **460**, 1080.
3 E. K. Silbergeld, D. Mandrioli and C. F. Cranor, *Annu. Rev. Public Health*, 2015, **36**, 175–191.
4 C. Rovida and T. Hartung, *ALTEX*, 2009, **26**, 187–208.
5 G. Patlewicz, G. Helman, P. Pradeep and I. Shah, *Comput. Toxicol.*, 2017, **3**, 1–18.
6 A. M. Richard, R. S. Judson, K. A. Houck, C. M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M. T. Martin, J. F. Wambaugh, T. B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A. J. Williams, S. B. Little, K. M. Crofton and R. S. Thomas, *Chem. Res. Toxicol.*, 2016, **29**, 438–451.
7 M. T. D. Cronin, J. S. Jaworska, J. D. Walker, M. H. I. Comber, C. D. Watts and A. P. Worth, *Environ. Health Perspect.*, 2003, **111**, 1391–1401.
8 T. Hartung and S. Hoffmann, *ALTEX*, 2009, **26**, 155–166.
9 T. Luechtefeld, A. M. Maertens, M. James, T. Hartung, A. Kleensang and V. Sá-Rocha, *J. Appl. Toxicol.*, 2015, **35**, 1361–1371.
10 T. Luechtefeld, A. Maertens, D. P. Russo, C. Rovida, H. Zhu and T. Hartung, *ALTEX*, 2016, **33**, 135–148.
11 T. Luechtefeld, A. Maertens, D. P. Russo, C. Rovida, H. Zhu and T. Hartung, *ALTEX*, 2016, **33**, 95–109.
12 T. Luechtefeld, A. Maertens, D. P. Russo, C. Rovida, H. Zhu and T. Hartung, *ALTEX*, 2016, **33**, 111–122.
13 T. Luechtefeld, A. Maertens, D. P. Russo, C. Rovida, H. Zhu and T. Hartung, *ALTEX*, 2016, **33**, 123–134.
14 T. Luechtefeld and T. Hartung, *ALTEX*, 2017, **34**, 459–478.
15 G. Patlewicz, N. Ball, R. A. Becker, E. D. Booth, M. T. Cronin, D. Kroese, D. Steup, B. van Ravenzwaay and T. Hartung, *ALTEX*, 2014, **31**, 387–396.
16 E. Berggren, P. Amcoff, R. Benigni, K. Blackburn, E. Carney, M. Cronin, H. Deluyker, F. Gautier, R. S. Judson, G. E. Kass, D. Keller, D. Knight, W. Lilienblum, C. Mahony, I. Rusyn, T. Schultz, M. Schwarz, G. Schuurmann, A. White,

J. Burton, A. M. Lostia, S. Munn and A. Worth, *Environ. Health Perspect.*, 2015, **123**, 1232–1240.
17 N. Ball, M. T. Cronin, J. Shen, K. Blackburn, E. D. Booth, M. Bouhifd, E. Donley, L. Egnash, C. Hastings, D. R. Juberg, A. Kleensang, N. Kleinstreuer, E. D. Kroese, A. C. Lee, T. Luechtefeld, A. Maertens, S. Marty, J. M. Naciff, J. Palmer, D. Pamies, M. Penman, A. N. Richarz, D. P. Russo, S. B. Stuard, G. Patlewicz, B. van Ravenzwaay, S. Wu, H. Zhu and T. Hartung, *ALTEX*, 2016, **33**, 149–166.
18 D. P. Russo, M. T. Kim, W. Wang, D. Pinolini, S. Shende, J. Strickland, T. Hartung and H. Zhu, *Bioinformatics*, 2017, **33**, 464–466.
19 I. Shah, J. Liu, R. S. Judson, R. S. Thomas and G. Patlewicz, *Regul. Toxicol. Pharmacol.*, 2016, **79**, 12–24.
20 T. Hartung, *ALTEX*, 2016, **33**, 83–93.
21 S. E. Crawford, T. Hartung, H. Hollert, B. Mathes, B. van Ravenzwaay, T. Steger-Hartmann, C. Studer and H. F. Krug, *Environ. Sci. Eur.*, 2017, **29**, 16, DOI: 10.1186/s12302-017-0115-z.
22 T. Hartung, *ALTEX*, 2017, **34**, 331–351.
23 A. Maertens, N. Anastas, P. J. Spencer, M. Stephens, A. Goldberg and T. Hartung, *ALTEX*, 2014, **31**, 243–249.
24 A. Maertens and T. Hartung, *Toxicol. Sci.*, 2018, **161**, 285–289.
25 OECD, *ENV/JM/MONO(2004)24*, 2004, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2004)24.
26 B. van Ravenzwaay, X. Jiang, T. Luechtefeld and T. Hartung, *Regul. Toxicol. Pharmacol.*, 2017, **88**, 157–172.
27 N. Gilbert, *Nature online*, 2016, https://www.nature.com/news/legal-tussle-delays-launch-of-huge-toxicity-database-1.19365.
28 T. Rabesandratana, *Science*, 2016, **351**, 651.
29 V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, *J. Stat. Mech.: Theory Exp.*, 2008, **10**, P10008, DOI: 10.1088/1742-5468/2008/10/P10008.
30 S. Hoffmann and T. Hartung, *Toxicol. Sci.*, 2005, **85**, 422–428.
31 T. Hartung, *ALTEX*, 2010, **27**, 3–14.
32 A. Kinsner-Ovaskainen, A. Bulgheroni, T. Hartung and P. Prieto, *Toxicol. in Vitro*, 2009, **23**, 1535–1540.
33 T. Hartung, *ALTEX*, 2017, **34**, 193–200.
34 A. Bulgheroni, A. Kinsner-Ovaskainen, S. Hoffmann, T. Hartung and P. Prieto, *Regul. Toxicol. Pharmacol.*, 2009, **53**, 16–19.
35 A. Gissi, K. Louekari, L. Hoffstadt, N. Bornatowicz and A. M. Aparicio, *ALTEX*, 2017, **34**, 353–361.
36 K. Taylor, D. J. Andrew and L. Rego, *Regul. Toxicol. Pharmacol.*, 2014, **69**, 320–332.
37 E. Adriaens, J. Barroso, C. Eskes, S. Hoffmann, P. McNamee, N. Alepee, S. Bessou-Touya, A. De Smedt, B. De Wever, U. Pfannenbecker, M. Tailhardat and V. Zuang, *Arch. Toxicol.*, 2014, **88**, 701–723.
38 J. M. Fitzpatrick, D. W. Roberts and G. Patlewicz, *J. Appl. Toxicol.*, 2017, **37**, 105–116.
39 T. Hartung, T. Luechtefeld, A. Maertens and A. Kleensang, *ALTEX*, 2013, **30**, 3–18.
40 C. Rovida, N. Alepee, A. M. Api, D. A. Basketter, F. Y. Bois, F. Caloni, E. Corsini, M. Daneshian, C. Eskes, J. Ezendam, H. Fuchs, P. Hayden, C. Hegele-Hartung, S. Hoffmann, B. Hubesch, M. N. Jacobs, J. Jaworska, A. Kleensang, N. Kleinstreuer, J. Lalko, R. Landsiedel, F. Lebreux, T. Luechtefeld, M. Locatelli, A. Mehling, A. Natsch, J. W. Pitchford, D. Prater, P. Prieto, A. Schepky, G. Schuurmann, L. Smirnova, C. Toole, E. van Vliet, D. Weisensee and T. Hartung, *ALTEX*, 2015, **32**, 25–40.
41 K. E. Tollefsen, S. Scholz, M. T. Cronin, S. W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth and G. Patlewicz, *Regul. Toxicol. Pharmacol.*, 2014, **70**, 629–640.
42 D. Basketter, C. Pease, G. Kasting, I. Kimber, S. Casati, M. Cronin, W. Diembeck, F. Gerberick, J. Hadgraft, T. Hartung, J. P. Marty, E. Nikolaidis, G. Patlewicz, D. Roberts, E. Roggen, C. Rovida and J. van de Sandt, *ATLA, Altern. Lab. Anim.*, 2007, **35**, 137–154.
43 J. Jaworska, Y. Dancik, P. Kern, F. Gerberick and A. Natsch, *J. Appl. Toxicol.*, 2013, **33**, 1353–1364.
44 K. Roy, *Advances in QSAR Modeling*, Springer International Publishing, 2017.
45 J. S. Jaworska, A. Natsch, C. Ryan, J. Strickland, T. Ashikaga and M. Miyazawa, *Arch. Toxicol.*, 2015, **89**, 2355–2383.
46 I. Linkov, O. Massey, J. Keisler, I. Rusyn and T. Hartung, *ALTEX*, 2015, **32**, 3–8.
47 M. Leist, N. Hasiwa, M. Daneshian and T. Hartung, *Toxicol. Res.*, 2012, **1**, 8–22.
48 F. Busquet and T. Hartung, *ALTEX*, 2017, **34**, 3–21.
49 T. Hartung, *ALTEX*, 2007, **24**, 67–80.
50 T. Hartung, *ALTEX*, 2010, **27**, 253–263.
51 T. Hartung, *ALTEX*, 2013, **30**, 275–291.
52 J. Pinheiro, D. Bates, S. DebRoy and D. Sarkar, *R package version 3.1-137*, 2017, https://CRAN.R-project.org/package=nlme.
53 P. Felzenszwalb, D. McAllester and D. Ramanan, *cs.brown.edu*, 2008, 1–8, http://cs.brown.edu/people/pfelzens/papers/lsvm-pami.pdf.
54 M. P. Kumar, B. Packer and D. Koller, *ai.stanford.edu*, 2010, https://ai.stanford.edu/~koller/Papers/Kumar+al:NIPS10.pdf.
55 C. N. J. Yu and T. Joachims, Proceedings of the International Conference on Machine Learning, 2009, 1169–1176.
56 A. L. Yuille and A. Rangarajan, *Neural Comput.*, 2003, **15**, 915–936.
57 A. Tropsha, *Mol. Inf.*, 2010, **29**, 476–488.
58 L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361–1375.
59 R. Agrawal and R. Srikant, *ACM Sigmod Record*, 2000, **29**, 439–450.
60 G. Patlewicz, G. Helman, P. Pradeep and I. Shaha, *Comput. Toxicol.*, 2017, **3**, 1–18.

61 A. P. Worth, A. Bassan, J. De Bruijn, A. Gallegos Saliner, T. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska and S. Eisenreich, *SAR QSAR Environ. Res.*, 2007, **18**, 111–125.
62 T. Hartung, S. Hoffmann and M. Stephens, *ALTEX*, 2013, **30**, 119–130.
63 S. Hoffmann, R. B. M. de Vries, M. L. Stephens, N. B. Beck, H. Dirven, J. R. Fowle 3rd, J. E. Goodman, T. Hartung, I. Kimber, M. M. Lalu, K. Thayer, P. Whaley, D. Wikoff and K. Tsaioun, *Arch. Toxicol.*, 2017, **91**, 2551–2575.
64 M. L. Stephens, K. Betts, N. B. Beck, V. Cogliano, K. Dickersin, S. Fitzpatrick, J. Freeman, G. Gray, T. Hartung, J. McPartland, A. A. Rooney, R. W. Scherer, D. Verloo and S. Hoffmann, *Toxicol. Sci.*, 2016, **152**, 10–16.
65 T. Hartung, S. Bremer, S. Casati, S. Coecke, R. Corvi, S. Fortaner, L. Gribaldo, M. Halder, S. Hoffmann, A. J. Roi, P. Prieto, E. Sabbioni, L. Scott, A. Worth and V. Zuang, *ATLA, Altern. Lab. Anim.*, 2004, **32**, 467–472.
66 H. Zhu, M. Bouhifd, E. Donley, L. Egnash, N. Kleinstreuer, E. D. Kroese, Z. Liu, T. Luechtefeld, J. Palmer, D. Pamies, J. Shen, V. Strauss, S. Wu and T. Hartung, *ALTEX*, 2016, **33**, 167–182.
67 J. Zhang, J. H. Hsieh and H. Zhu, *PLoS One*, 2014, **9**, e99863.
68 K. Ribay, M. T. Kim, W. Wang, D. Pinolini and H. Zhu, *Front. Environ. Sci.*, 2016, **4**, 12, DOI: 10.3389/fenvs.2016.00012.
69 H. Zhu, J. Zhang, M. T. Kim, A. Boison, A. Sedykh and K. Moran, *Chem. Res. Toxicol.*, 2014, **27**, 1643–1651.
70 M. T. Kim, R. Huang, A. Sedykh, W. Wang, M. Xia and H. Zhu, *Environ. Health Perspect.*, 2016, **124**, 634–641.