CONFERENCE ARTICLE

EFSA Journal

APPROVED: 29 May 2019

doi: 10.2903/j.efsa.2019.e170710

Predicting toxicity of chemicals: software beats animal testing

Thomas Hartung

Johns Hopkins University, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA; and University of Konstanz, CAAT-Europe, Konstanz, Germany

Abstract

We created earlier a large machine-readable database of 10,000 chemicals and 800,000 associated studies by natural language processing of the public parts of Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) registrations until December 2014. This database was used to assess the reproducibility of the six most frequently used Organisation for Economic Co-operation and Development (OECD) guideline tests. These tests consume 55% of all animals in safety testing in Europe, i.e. about 600,000 animals. With 350–750 chemicals with multiple results per test, reproducibility (balanced accuracy) was 81% and 69% of toxic substances were found again in a repeat experiment (sensitivity 69%). Inspired by the increasingly used read-across approach, we created a new type of QSAR, which is based on similarity of chemicals and not on chemical descriptors. A landscape of the chemical universe using 10 million structures was calculated, when based on Tanimoto indices similar chemicals are close and dissimilar chemicals far from each other. This allows placing any chemical of interest into the map and evaluating the information available for surrounding chemicals. In a data fusion approach, in which 74 different properties were taken into consideration, machine learning (random forest) allowed a fivefold cross-validation for 190,000 (non-) hazard labels of chemicals for which nine hazards were predicted. The balanced accuracy of this approach was 87% with a sensitivity of 89%. Each prediction comes with a certainty measure based on the homogeneity of data and distance of neighbours. Ongoing developments and future opportunities are discussed.

Keywords: alternatives to animal testing, computational toxicology, read-across, risk assessment

Correspondence: THartung@jhu.edu

Conflict of interest statement: The author consults Underwriters Laboratories (UL) on computational toxicology and receives shares from their respective sales. He also holds stock options and consults ToxTrack LLC.

Acknowledgements: This work was mainly executed by Dr Tom Luechtefeld during his PhD and within ToxTrack. The most valuable input of many colleagues inside and outside our team, especially from Dr Alexandra Maertens, is gratefully appreciated. 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).

Suggested citation: Hartung T, 2019. Conference article on predicting toxicity of chemicals: software beats animal testing. EFSA Journal 2019;17(S1):e170710, 8 pp. https://doi.org/10.2903/j.efsa.2019.e170710

ISSN: 1831-4732

The EFSA Journal is a publication of the European Food Safety Authority, an agency of the European Union.

Table of contents

1. Introduction

Big data and artificial intelligence are the most recent additions to the toolbox of toxicity assessments. The increasing availability of large, more-or-less curated data sets of toxicity data form the basis for this development. Prominent examples are PubChem, ChemBL and ToxRefDB. The strong publication bias for toxic results, which does not really reflect the chemical universe, is a dramatic shortcoming of these databases. A major step forward was the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation, which includes the publication of a robust summary of findings for each registration dossier by the European Chemical Agency (ECHA). These files, however, are not machine-readable. A download of the data in December 2014 has already found about 10,000 chemicals with 800,000 associated studies (Luechtefeld et al., 2016a). Using natural language processing, large parts of the data were extracted and transferred to a database that could be processed. The database showed a predominance of non-toxic substances: no Globally Harmonised System (GHS) classification was assigned to more than 20% of chemicals. This analysis might be biased, in part, by the predominance of high-production-volume chemicals registered until December 2014.

Although the data are public, ECHA has concerns about ownership and republication, and we have agreed to use the database only for our own research and collaborations and not to publish it until it is approved. ECHA, in turn, made a database publicly available, which included about 15,000 chemicals but appears to be redacted for about 70% of the information otherwise available on their website.

One impressive finding was the high number of repeat animal experiments. On average, every study was carried out three times, although this distribution was very skewed, with, for example, two chemicals tested in the rabbit Draize eye test more than 90 times (69 chemicals were tested more than 45 times) (Luechtefeld et al., 2016b). In particular, the acute and topical tests – often referred to as the toxicological six-pack – had many repeat studies, allowing a formal analysis of the reproducibility of these tests. Notably, these six tests consumed 55% of all animals in toxicology in Europe (European Commission, 2011). No test carried out a single time can be better than its reproducibility. It will actually be worse, as besides reproducibility the relevance of a test includes aspects such as strain differences, species differences, relevance of dose regimen, relevance of age, human diversity including susceptible subpopulations, etc. However, reproducibility is probably the only one of these shortcomings that can be easily addressed.

2. Assessment of animal test reproducibility

We used the simple method of joint probability of agreement. Based on 350–700 chemicals per hazard, each tested multiple times, the probability that an Organisation for Economic Co-operation and Development (OECD) guideline animal test would output the same result in a repeat test was 78–96% balanced accuracy (BAC; sensitivity 50–87%) (Luechtefeld et al., 2018a,b). On average, BAC was 81%, which reflected the high prevalence of negative test results. A non-toxic substance will rarely suddenly become positive, and indeed sensitivity, which reflects whether a toxic finding can be reproduced, was on average only 69%.

These data confirm similar findings for smaller sets of chemicals. For example, a reproducibility of only 77% was found for skin sensitisation by Kolle et al. (2013). Hoffmann (2015) analysed the variability of the local lymph node assay (LLNA) test, using the NICEATM database with more than 60 repeat experiments, the false-positive rate ranged from 14 to 20% (false-negative rate 4–5%). With respect to the Draize eye test, Weil and Scala (1971) and Cormier et al. (1996) studied the variability of responses of nine substances tested in up to 24 laboratories showing enormous variability. This was also confirmed in the assessment by Adriaens et al. (2014) of the test's reproducibility.

Others have shown even more pronounced reproducibility issues for ecotoxicological tests (Hrovat et al., 2009). They analysed fish acute lethality (LC_{50}) values for 44 compounds (for which at least 10 data entries existed) that were extracted from the ECOTOX database, yielding 4,654 test records showing variability of test results exceeding several orders of magnitude. There is no indication that reproducibility is better for systemic toxicities (Smirnova et al., 2018): a concordance of 57% was found comparing 121 replicate rodent carcinogenicity assays (Gottmann et al., 2001). Our own analysis (Luechtefeld et al., 2016c) of 28-day vs 90-day no observed adverse effect levels (NOAELs) showed no correlation.

Some animal tests are carried out on two species, allowing interspecies variability to be addressed. We showed this earlier for skin sensitisation, for which guinea pigs and mice are commonly used. While both assays were about 90% reproducible, based on 655 chemicals, the two species were only

77% predictable of each other (Luechtefeld et al., 2016c). That is still better than mice and rats in cancer bioassays predicting each other in only 53% of cases (Basketter et al., 2012; Smirnova et al., 2018) or various species predicting each other's results in reproductive toxicity studies by about 60% (Hurtt et al., 2003; Bailey et al., 2005). Species concordance (310 chemicals) for non-neoplastic pathology between mouse and rat was 68% (Wang and Gray, 2015). Wang and Gray (2015) also showed interspecies concordance for 37 substances tested in the US National Toxicology Program (NTP) of mouse with rats at 57–89% (average 75%) in the short-term and 65–89% (average 80%) in long-term studies. The mouse-to-rat organ prediction in long-term studies aligned, on average, in 55% of cases; in short-term studies with an average of 45%; and for rat to mouse, the averages are 27% and 49%, respectively.

These reproducibility estimates provided benchmarks for what any method calibrated (to avoid the term 'validated') against mere test results of single-tested chemicals can achieve. Any better results represented either overfitting or carefully selected reference substances, representing more information than a single OECD guideline test.

3. Development of a predictive tool for chemical toxicity

After the 2016 study (Luechtefeld et al., 2016a), which was also accompanied by analyses of ocular, skin sensitisation and ocular toxicity (Luechtefeld et al., 2016b,c,d), the potential of such a database to develop a predictive tool for chemical toxicity was explored. Taking the example of skin sensitisation (Luechtefeld et al., 2016c), a rather simple approach asked how the closest chemical neighbour with data could substitute for the experimental result of a given chemical (k-nearest neighbour with n = 1). Requesting a minimum 75% similarity (Tanimoto index), 2,288 of 3,024 chemicals had a neighbour with data, and the result was the same in 80%. With increasing minimum similarity to 85, 90, and 95%, the number of chemicals with neighbours fell to 1,738, 1,189, and 525 chemicals, respectively, i.e. coverage dropped from 70% to 15%. At the same time, accuracy of prediction increased to 84, 85 and 92%. This means that this simple read-across was as good as mice and guinea pigs predicting each other in the respective guideline studies (already at 75% similarity) and outperformed the reproducibility of either assay at 95% similarity.

These findings encouraged the development of automated read-across (Hartung, 2016). They also showed that a certainty can be assigned to a prediction based on the similarity of its neighbour. This is an important advantage not offered by any other current predictive tool. Whether these public data represent a legitimate access to use for registration purposes is an important question. This has fortunately been clarified by ECHA, if used in an aggregated manner for predicting other substances: 'A registrant would need permission to use protected data to read-across from a single substance to the target substance. But they would not need this to make a quantitative structure–activity relationship (QSAR) prediction' (ECHA gives clarity on IP issues for QSAR predictions. Chemical Watch, 5 July 2017).

Our parallel work on read-across (Patlewicz et al., 2014; Ball et al., 2016; Zhu et al., 2016) prompted the development of a read-across-based structure–activity relationship (RASAR). The RASAR models are not traditional QSARs, in which a highly curated, small training data set is used to predict a single property based on chemical descriptors (i.e. classifications per hazard). The method is based as a first step for creating a map of the chemical universe (forced layout graph). This is based on chemical similarity, in which similar chemicals are close and different ones are far from each other. Similarity is assessed in two steps, i.e. first a fingerprinting of the molecules (typically substructures present) and then a metric to express similarity. The most commonly used metric is the so-called Tanimoto index, which expresses, how many of all substructures in the two molecules are shared. Tanimoto makes use of about 900 such substructures. There is a variety of other metrics that could be used, e.g. Dice, Cosine, Sokal, Kulczynski, McConnaughey, Asymmetric, Braun Blanket, etc. Our initial exploration suggests that no single metric alone performs significantly better than Tanimoto, but that a combination of such metrics could improve the approach.

Any chemical of interest can then in a second step be placed into this similarity map. To our surprise, extending the k-nearest neighbour approach to more than one neighbour (i.e. informing the prediction by more surrounding chemicals) did not improve predictions to major extent. However, an approach in which the distance to the closest negative and to the closest positive neighbour were used, predictions improved. This means, however, setting two thresholds for a valid prediction, i.e. minimum similarity to the positive and minimum similarity to the closest negative neighbour. This creates a grey zone in which these distances are too similar to classify. Setting these thresholds

directly affects sensitivity (% of toxic substances found), specificity (% of real positives among toxic calls), and coverage (percentage of chemicals for which a prediction is possible). We chose a sensitivity of at least 80% for each hazard and balanced the resulting specificity (above 50%) and coverage (aiming at two-thirds of chemicals). Teaming up with Underwriters Laboratories (UL), a global safety certification company headquartered in Northbrook, Illinois, this was implemented as a web-based prediction tool (now discontinued) called REACH*across*™ (Luechtefeld and Hartung, 2017). This extended database included 58,000 known hazards associated with a chemical, up to 15,000 for each of the six hazards covered (skin sensitisation, eye irritation, skin irritation/corrosion, acute oral and dermal toxicity, and mutagenicity). This allowed a leave-one-out cross-validation, i.e. for each and every chemical with a given hazard information a prediction was made and then compared with the actual result. A prediction was possible for 42,500 cases (73% coverage) with on average 84% sensitivity and 61% specificity. This puts the method roughly on par with the reproducibility of the respective animal tests as discussed above.

This 'simple' RASAR only made use of the information of neighbours on the hazard to be predicted. This means, it uses what we know from neighbours about eye irritation to predict the eye irritation of the chemical of interest. So, one model per hazard is developed. There is an enormous wealth in what else we know about the chemical itself and its neighbours, however. This is called transfer learning or data fusion.

We next developed a data fusion RASAR. The published model (Luechtefeld et al., 2018b) uses data on 100,000 chemical structures, calculates 5 billion similarities, and simultaneously makes 190,000 predictions for nine hazards of toxic properties of chemicals (the six endpoints above, plus acute inhalation as well as acute and chronic fish toxicity). The database represents a combination of reliable data sources (PubChem, ECHA, ICE, etc.). It still uses the most common Chemical Similarity (PubChem2d) with the Tanimoto (Jaccard) metric. The network features use proximity to positive and negative neighbours. Data fusion makes use of the other 74 toxicity, biological and chemophysical endpoints. This means that a 222-dimensional vector is used for the prediction. Machine Learning (logistic regression, random forest) gives probabilistic hazard estimates. Computing Clusters (Apache Spark pipeline) allow this massive scale computing. In a similar cross-validation as carried out for the simple RASAR, the 190,000 predictions of chemical classifications were 87% correct. Markedly, our model provided 100% coverage for nine hazards with high, balanced accuracies throughout.

4. Conclusions

For the six most used toxicity tests (which use 55% of animals in toxicology in Europe, almost 600,000 animals per year), animal repeat tests showed 81% (balanced) accuracy. The machine learning approach achieved 87% (balanced) accuracy, as shown for 4 to 48,000 chemicals with animal data-based classifications. The comparison was even more impressive when looking for the identification of toxic substances (sensitivity), in which the repeat animal test succeeded in 69% of cases and the machine learning in 89%.

In the meantime, the model has been implemented with customisable user interfaces as the UL Cheminformatics Tool Kit and is available for beta-testing. As a new feature, it now also predicts the different GHS classes as a measure of potency. Notably, this feature has not yet been validated. The tool kit allows installation behind a firewall, combination with proprietary data and customised user interfaces. Features such as running lists of chemicals, chemical design, one-on-one comparisons for alternative chemistry, and identification of candidate alternative chemicals are currently implemented.

Evaluation and validation with various data sets in agencies and industry is on the way. These efforts are integrated into the EUToxRisk flagship project (Daneshian et al., 2016) and attempts to further the regulatory acceptance of read-across (Chesnut et al., 2018). The data fusion RASAR demonstrated the potential of big data and machine learning to reduce animal testing for acute and topical endpoints.

Besides REACH and other global legislation for new and existing chemicals, the RASAR approach lends itself to prioritisation, classification and labelling, emergency risk assessments, and risk assessment for new products allowing frontloading of risk assessments (i.e. Green Toxicology) (Maertens et al., 2014; Crawford et al., 2017; Maertens and Hartung, 2018). Another interesting use is for determining impurities, for example, in pharmaceuticals; these can be easily detected but the effort to synthesise large quantities and test them is usually not possible. These examples might suffice to illustrate possible uses, which we have only started to explore with the advent of artificial intelligence and big data in safety sciences.

References

Adriaens E, Barroso J, Eskes C, Hoffmann S, McNamee P, Alepee N, Besssou-Touva S, De Smedt A, De Wever B, Pfannenbecker U, Talhardat M and Zuang V, 2014. Retrospective analysis of the Draize test for serious eye damage/eye irritation: importance of understanding the *in vivo* endpoints under UN GHS/EU CLP for the development and evaluation of *in vitro* test methods. Archives of Toxicology, 88, 701–723. https://doi.org/10.1007/s00204-013-1156-8

Bailey J, Knight A and Balcombe J, 2005. The future of teratology research is *in vitro*. Biogenic Amines, 19, 97–145. https://doi.org/10.1163/1569391053722755

Ball N, Cronin MTD, Shen J, Adenuga MD, Blackburn K, Booth ED, Bouhifd M, Donley E, Egnash L, Freeman JJ, Hastings C, Juberg DR, Kleensang A, Kleinstreuer N, Kroese D, Lee AC, Luechtefeld T, Maertens A, Marty S, Naciff JM, Palmer J, Pamies D, Penman M, Richarz A-N, Russo DP, Stuard SB, Patlewicz G, van Ravenzwaay B, Wu S, Zhu H and Hartung T, 2016. Toward Good Read-Across Practice (GRAP) guidance. Altex, 33, 149–166. https://doi.org/10.14573/altex.1601251

Basketter DA, Clewell H, Kimber I, Rossi A, Blaaboer B, Burrier R, Daneshian M, Eskes C, Goldberg A, Hasiwa N, Hoffmann S, Jaworska J, Knudsen TB, Landsiedel R, Leist M, Locke P, Maxwell G, McKim J, McVey EA, Ouédraogo G, Patlewicz G, Pelkonen O, Roggen E, Rovida C, Ruhdel I, Schwarz M, Schepky A, Schoeters G, Skinner N, Trentz K, Turner M, Vanparys P, Yager J, Zurlo J and Hartung T, 2012. A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. Altex, 29, 3–89.

Chesnut M, Yamada T, Adams T, Knight D, Kleinstreuer N, Kass G, Luechtefeld T, Hartung T and Maertens A, 2018. Regulatory acceptance of read-across: report from an international satellite meeting at the 56th Annual Meeting of the Society of Toxicology. Altex, 35, 413–419.

Cormier EM, Parker RD, Henson C, Cruse LW, Merritt AK, Bruce RD and Osborne R, 1996. Determination of the intra- and interlaboratory reproducibility of the low volume eye test and its statistical relationship to the Draize eye test. Regulatory Toxicology and Pharmacology, 23, 156–161.

Crawford SE, Hartung T, Hollert H, Mathes B, van Ravenzwaay B, Steger-Hartman T, Studer C and Krug HF, 2017. Green toxicology: a strategy for sustainable chemical and material development. Environmental Sciences Europe, 29, 16. https://doi.org/10.1186/s12302-017-0115-z

Daneshian M, Kamp H, Hengstler J, Leist M and van de Water B, 2016. Highlight report: Launch of a large integrated European in vitro toxicology project: EU-ToxRisk. Archives of Toxicology, 90, 1021. https://doi.org/10.1007/s00204-016-1698-7

Gottmann E, Kramer S, Pfahringer B and Helma C, 2001. Data quality in predictive toxicology: Reproducibility of rodent carcinogenicity experiments. Environmental Health Perspectives, 109, 509–514. https://doi.org//10.1289/ehp.01109509

Hartung T, 2016. Making big sense from big data in toxicology by read-across. ALTEX - Alternatives to Animal Experimentation, 33, 83–93. https://doi.org/10.14573/altex.1603091

Hoffmann S, 2015. LLNA variability: an essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. Altex, 32, 379–383.

Hrovat M, Segner H and Jeram S, 2009. Variability of *in vivo* fish acute toxicity data. Regulatory Toxicology and Pharmacology, 54, 294–300. https://doi.org/10.1016/j.yrtph.2009.05.013

Hurtt ME, Cappon GD and Browning A, 2003. Proposal for a tiered approach to developmental toxicity testing for veterinary pharmaceutical products for food-producing animals. Food and Chemical Toxicology, 41, 611–619.

Kolle SN, Basketter DA, Casati S, Stokes WS, Strickland J, van Ravenzwaay B, Vohr HW and Landsiedel R, 2013. Performance standards and alternative assays: Practical insights from skin sensitization. Regulatory Toxicology and Pharmacology, 65, 278–285. https://doi.org/10.1016/j.yrtph.2012.12.006

Luechtefeld T and Hartung T, 2017. Computational approaches to chemical hazard assessment. Altex, 34, 459–478.

Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H and Hartung T, 2016a. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008–2014. Altex, 33, 95–109. https://doi.org/10.14573/altex.1510052

Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H and Hartung T, 2016b. Analysis of public oral toxicity data from REACH registrations 2008–2014. Altex, 33, 111–122. https://doi.org/10.14573/altex.1510054

Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H and Hartung T, 2016c. Analysis of publically available skin sensitization data from REACH registrations 2008–2014. Altex, 33, 135–148. https://doi.org/10.14573/altex.1510055

Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H and Hartung T, 2016d. Analysis of Draize eye irritation testing and its prediction by mining publicly available 2008–2014 REACH data. Altex, 33, 123–134. https://doi.org/10.14573/altex.1510053

Luechtefeld T, Rowlands C and Hartung T, 2018a. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. Toxicological Research, 7, 732–744. https://doi.org/10.1039/C8TX00051D

Luechtefeld T, Marsh D, Rowlands C and Hartung T, 2018b. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicological Sciences, 165, 198–212. https://doi.org/10.1093/toxsci/kfy152

Maertens A and Hartung T, 2018. Green toxicology – know early about and avoid toxic product liabilities. Toxicological Sciences, 161, 285–289. https://doi.org/10.1093/toxsci/kfx243
Maertens A, Anastas N, Spencer PJ, Stephens M, Goldberg A and Hartung T, 2014. Green toxicology. Altex, 31, 243–249.
Patlewicz G, Ball N, Becker RA, Booth ED, Cronin MTD, Kroese D, Steup D, van Ravenzwaay B and Hartung T, 2014. Read-across approaches – misconceptions, promises and challenges ahead. ALTEX – Alternatives to Animal Experimentation, 31, 387–396.
Smirnova L, Kleinstreuer N, Corvi R, Levchenko A, Fitzpatrick SC and Hartung T, 2018. 3S – Systematic, systemic, and systems biology and toxicology. Altex, 35, 139–162. https://doi.org/10.14573/altex.1804051
Wang B and Gray G, 2015. Concordance of noncarcinogenic endpoints in rodent chemical bioassays. Risk Analysis, 35, 1154–1166. https://doi.org/10.1111/risa.12314
Weil CS and Scala RA, 1971. Study of intra- and interlaboratory variability in the results of rabbit eye and skin irritation test. Toxicology and Applied Pharmacology, 19, 276–360.
Zhu H, Bouhifd M, Donley E, Egnash L, Kleinstreuer N, Kroese E, Liu Z, Luechtefeld T, Palmer J, Pamies D, Shen J, Strauss V, Wu S and Hartung T, 2016. Supporting read-across using biological data. ALTEX - Alternatives to Animal Experimentation, 33, 167–182. https://doi.org/10.14573/altex.1601252

Abbreviations

BAC	balanced accuracy
ECHA	European Chemical Agency
GHS	Globally Harmonised System
LC_{50}	lethal concentration, median
LLNA	local lymph node assay
NOAEL	no observed adverse effect level
NTP	National Toxicology Program
OECD	Organisation for Economic Co-operation and Development
QSAR	quantitative structure–activity relationship
RASAR	read-across-based structure–activity relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
UL	Underwriters Laboratories