Letter to the Editor

Oy Vey! A Comment on “Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships Outperforming Animal Test Reproducibility”

A study by (Luechtefeld et al., 2018) described the development of a suite of *in silico* models, termed read-across structure activity relationships (RASAR), that have “balanced accuracies in the 80%–95% range across 9 health hazards with no constraints on tested compounds.” This work can be considered groundbreaking for apparently exceeding the most optimistic expectations for quantitative structure-activity relationship (QSAR) modeling accuracy, especially without restrictions on model applicability domains. Predictive *in silico* models have been facilitating replacement and reduction of animal testing in toxicology (Wold et al., 1985); however, it is also recognized that “these methods are not always reliable and must be assessed on their individual merit for the compound and context in question” (Cronin et al., 2017). It is widely acknowledged that QSAR and other *in silico* models should be subject to rigorous testing and validation (Dearden et al., 2009; Fourches et al., 2010, 2015; OECD, 2014; Tropsha, 2010). Thus, we were curious to understand what technological advances have enabled the RASAR models to achieve accuracy that, for the first time in the history of QSAR, was “outperforming animal tests reproducibility” (Luechtefeld et al., 2018).

With this letter, we aim to understand the genesis of the outstanding result presented by (Luechtefeld et al., 2018). Upon close examination of the methods and conclusions, we believe there is an urgent need to highlight several issues with this study in the context of known pitfalls in QSAR modeling and best practices for model validation, as detailed in Dearden et al. (2009) and (Tropsha, 2010). Our interest is magnified by the fact that the extensive publicity associated with this publication and commercial availability of RASAR as part of the Underwriters Laboratories Cheminformatics Suite (Luechtefeld et al., 2018) may encourage the regulatory bodies to accept RASAR as a decision support tool without extensive validation.

DATA COLLECTION, CURATION, INTEGRATION, QUALITY, AND REPRODUCIBILITY

Failure to take account of data heterogeneity (Dearden, et al., 2009). The authors stated that the chemical hazard labels in their dataset are derived from “OECD guideline studies, read across studies, QSAR studies and other information” (Luechtefeld et al., 2018). Thus, it appears that predicted data were used for model development. We reason that models predicting compound categories that have been predicted by similar models likely suffer from inflated accuracy. We posit that further discussion is needed on what data were used. We submit that only experimental results should be considered for modeling and only after careful data curation.

Poor transferability of QSARs/QSPRs (Dearden, et al., 2009). QSAR models should be transferable/available to other groups to use and/or reproduce. Unfortunately, neither the exact modeling dataset, nor the descriptors used by (Luechtefeld et al., 2018) have been made publicly available, precluding independent evaluation of the quality of the data and models. The only access to the predictions is through a paid transaction (<https://www.ulreachacross.com>).

Use of inadequate data/Replication of compounds in a dataset (Dearden et al., 2009). The dataset used by (Luechtefeld et al., 2018) was not publicly available at the time of this letter’s submission; therefore, we spot-checked ECHA databases (ECHA, 2018a,b). We found that compounds are often annotated as “not reliable” (eg, o-xylene, 3, 5-dimethylpyridine, and N, N'-diacetylhydrazine). Information is available in the abovementioned databases to explain lack of reliability; therefore, it is important for the authors (Luechtefeld et al., 2018) to clarify whether compounds classified as “unreliable” have been used for modeling.

A related concern is the lack of clarity regarding possible presence of duplicates in both training and test sets, a well-known pitfall that must be addressed in the data curation phase (Dearden et al., 2009; Fourches et al., 2010, 2015). Although the authors acknowledge that duplicates are present in the database, they did not explain a procedure for curation and standardization of the dataset.

DATA MODELING AND MODEL VALIDATION

Misuse/misinterpretation of statistics/Over-fitting of data/Failure to validate a QSPR correctly (Dearden et al., 2009). The authors discuss the accuracy of their predictions in the context of the reproducibility of the data used to generate models. The authors state that “For the 6 tests often referred to as “toxicological 6-pack” a reproducibility sensitivity of on average 70% was found (Table 2); the Simple RASAR matched this with on average the same 70%; by data fusion, 89% average sensitivity was achieved clearly outperforming the respective animal test” (Luechtefeld et al., 2018). This statement raises several concerns because the predictivity of a QSAR model may not be greater than the experimental uncertainty. It is our opinion that model accuracy and experimental reproducibility cannot be equated or compared. A recent publication (Helma et al., 2018) has addressed this very issue in the context of read-across concluding that “with missing information about the variability of experimental toxicity data it is hard to judge the performance of predictive models objectively.”

We also posit that more explanation may be warranted regarding an assertion that “reproducibility evaluations used all chemicals with multiple results” (Luechtefeld et al., 2018). To assess the reproducibility of duplicate chemical tests in animal models, a conditional probability was calculated: “Sensitivity and specificity are estimated by the conditional probability that a test is positive given that its paired test was positive (sensitivity) or that a test is negative given that its paired test is negative (specificity)” (Luechtefeld et al., 2018). We reason that this is not an appropriate statistical

method because respective toxicological studies are independent variables. The authors should comment on this concern as it directly questions the approach used to assess the accuracy of model predictions.

Also, there are 2 additional considerations that pertain to the modeling and model validation. First, it is unclear whether the authors have removed descriptors with missing values. Imputation is often performed for the datasets with missing variables; however, it is not clear if this was the case. Second, Y-randomization is a standard protocol used in the validation of QSAR models wherein models are built from randomly permuted data to test their validity and guard against chance descriptor correlations (Tropsha, 2010). It is not clear if this procedure was done by (Luechtefeld et al., 2018).

In conclusion, we are in full support of developing and using computational models as an alternative to animal testing (Bell et al., 2017; Hartung and Hoffmann, 2009). It is exciting that the issue of alternatives to animal testing receives high level of attention in both print and social media. However, because of heightened attention to the reproducibility in biomedical research (Collins and Tabak, 2014; Miller, 2014), it is important to ensure that bold claims are well justified, supported by carefully vetted data, and follow best scientific practices. Our conclusion is that it is difficult to accept RASAR model accuracy as stated in (Luechtefeld et al., 2018). Accurate prediction of adverse chemical effects is a critically important challenge; therefore, we hope that the authors would clarify ambiguities of their study highlighted in this letter.

SUPPLEMENTARY DATA

Supplementary data are available at Toxicological Sciences online.

Letter to the Editor

Missing the Difference Between Big Data and Artificial Intelligence in RASAR Versus Traditional QSAR

The letter to the editor by Alves et al. (2018a) was a surprise to the authors of "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility" (Luechtefeld et al., 2018a). The letter challenges the approach as one would challenge a traditional QSAR, by which it ignores many attributes and consequences of the RASARs construction and performance as an implementation of big data and artificial intelligence (machine learning) (Hartung, 2016; Luechtefeld and Hartung, 2017).

To state it simply: the RASAR models are not traditional QSARs, wherein a highly curated, small training dataset is used to predict a single property based on chemical descriptors, ie, classifications per hazard. The published model 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: 87% are correct, which should raise the question what we got right, not what we got wrong?

FUNDING

National Institute of Environmental Health Sciences, (Grant/Award Number: P42 ES027704).

REFERENCES

Available as a Supplementary File.

Vinicius M. Alves,* Joyce Borba,*[†] Stephen J. Capuzzi,* Eugene Muratov,*[†] Carolina H. Andrade,[†] Ivan Rusyn,^{§,1} and Alexander Tropsha*¹

*UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599; [†]Faculty of Pharmacy, Federal University of Goias, Goiania, Goias 74605-170, Brazil; [‡]Odessa National Polytechnic University, Odessa 65000, Ukraine; and [§]College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843

¹To whom correspondence should be addressed at College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, 4458 TAMU, College Station, TX 77843. E-mail: irusyn@tamu.edu and UNC Eshelman School of Pharmacy, University of North Carolina, 100K Beard Hall, Chapel Hill, NC 27599. E-mail: alex_tropsha@unc.edu.

doi: 10.1093/toxsci/kfy286

Advance Access Publication Date: November 30, 2018

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society of Toxicology. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Regulators are already considering how best to integrate these new kinds of machine learning models into their decisions. We refer also to our attempts to carry out now an external validation by the United States validation body Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM). Notably, the corresponding author held earlier responsibilities for such validations in Europe including *in silico* approaches (Worth et al., 2004) and has continuously contributed to QSAR validation (Hartung et al., 2004).

WHAT IS SIGNIFICANT ABOUT THE ACCURACIES DEMONSTRATED IN OUR MODEL?

The RASAR model is exciting because of the ability for one "composite" model to estimate many endpoints and integrate datasets. Similar work in machine learning has demonstrated the benefit of simultaneously modeling multiple endpoints via "transfer learning". In practice, this may enable modeling of data-sparse endpoints (a critical need in toxicology).

None of our individual published metrics far exceed previous demonstrations. The Alves et al. letter authors themselves published models with comparable accuracies (95% correlation on aquatic endpoints, 82% balanced accuracy for skin sensitization, 88%–92% accuracy on Ames test). The recent NTP acute oral toxicity event received an NIH National Center for Advancing