# On the Comparability of Studies in Predictive Ecotoxicology

Christoph Schür,[1,*] Kristin Schirmer,[1,2,†] and Marco Baity-Jesi[1,‡]

[1] *Eawag, Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland*

[2] *Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland*

We address several comparability issues in chemical hazard assessment through *in silico* models. We review the recent literature and show that studies predicting fish acute toxicity are not comparable based on the following criteria: (i) The datasets are different, representing different chemical and/or taxonomic spaces; (ii) Even if the dataset were the same, cleaning might be different; (iii) Even if the cleaning is the same, the data is split in a different manner. Additionally, the test set is often also used as a validation set. (iv) Even when the dataset and the splittings are the same, the metrics are not clearly defined or deceptively different (*e.g.* different definitions of $R^2$ give different results). As a short-term solution and as the only course of action ensuring the effective progression of the field as a whole, we call for the usage of benchmark datasets, which provide well-defined data cleaning, versioning and splittings. This should be complemented by a clear definition of the used metrics, and a transparent reporting, which includes code and the use of reporting/data sheets.

**Keywords: Machine learning; QSAR; QSTR; Reproducibility; Performance Metrics; Benchmark; Environmental Hazard Assessment**

## INTRODUCTION

The scientific community is making joint efforts to solve grand environmental challenges [1]. This includes those related to chemical pollution. The 350,000 chemicals and mixtures currently globally registered on the market [2] and the practical doubling of chemi-

* christoph.schuer@eawag.ch

† kristin.schirmer@eawag.ch

‡ marco.baityjesi@eawag.ch

cal industry production since the year 2000, lead to chemical pollution with unknown toxicity to be considered a planetary boundary threat [3–6]. Ensuring the safety of chemicals for humans and the environment is a Herculean task that requires new approaches for hazard assessment outside of classical animal experiments. The legislation for the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) by the European Commission requires invertebrate animal tests (*Daphnia* acute toxicity) for chemicals with a manufacturing or import quantity of $\geq 1$ tons per year and fish acute toxicity tests $\geq 10$ tons per year [7]. Mittal *et al.* estimated the global annual usage of fish and birds to range between 440,000 and 2.2 million individuals at a cost upwards of $39 million *p.a.* [8]. Hence, moving away from the reliance on animal experiments, both ethically and economically, is warranted. Using computational (*in silico*) methods, such as machine learning (ML), to make predictions of ecotoxicological outcomes for chemicals in various species, garners high hopes, but is a yet under-explored approach to environmental hazard assessment (EHA) [9–11]. Progression of the field of predictive ecotoxicology (and with it the adoption of machine learning methods) requires the scientific community to act collaboratively, with each work building on the findings of the previous [12]. This in turn is hindered by a lack of comparability between studies.

With the regulatory bodies slow to accept alternative methods and vague about what performance a method needs to achieve in order to be accepted into the regulatory process, there has not been much progress recently when it comes to predictive ecotoxicology [13–15]. Much of the literature concerned with predicting ecotoxicological outcomes is focused on quantitative structure-activity relationships (QSAR) models [16]. They are based on the empiric observation that similar compounds will elicit similar biological responses (including toxicity). They have a long-standing tradition in predictive chemistry and can be based on linear and non-linear relationships between chemical descriptors and biological responses. Recent years have seen the adoption of machine learning into QSAR modeling, enabling users to integrate non-chemical data into their predictions [17, 18]. Enhancing a model, beyond the limitation of using only chemical descriptors, ultimately blurs the lines between what can be considered a QSAR model and what goes beyond that. Hence, even though our point ultimately is on the adoption of better practices for comparability in the application of machine learning towards ecotoxicological hazard assessment, it is impossible to exclude

QSAR investigations from the scope of this study.

Here, we raise the point that, currently, the field of predictive ecotoxicology is lacking comparability. This, in turn, prevents the transfer of insights between studies and, as a consequence, meaningful progress of the field beyond individual studies. To reinforce this point, we screen a representative subset of available literature on fish acute toxicity prediction and evaluate their comparability based on five comparability criteria (similarity in dataset, dataset cleaning, splitting, performance metrics, and reporting), outlined in Fig. 1. Although the initial goal of this work consisted of finding the fraction of comparable studies, the collected data revealed a sobering truth. At the date of data screening, no two studies in our field can meaningfully and reasonably be compared to each other in a quantitative manner, since they do not fulfill the comparability criteria for ML-based predictive ecotoxicology. Our focus is only on comparability among studies; we do not address flaws in individual studies that could further decrease their usefulness (*e.g.* data leakage [19, 20]).

## MATERIALS & METHODS

### A. Scope & Study selection

We used GoogleScholar with different keyword combinations (*i.e.*, Machine learning, predictive toxicology, ecotoxicology, fish, acute, QSAR), and limited the search results to the period between 2013 and end of 2024. The resulting collection of papers was augmented through adding similar and younger works *via* ResearchRabbit. The resulting dataset is based on 50 papers, representing an overview of the state of the field. Studies could include QSARs making predictions solely based on chemical descriptors as well as studies employing ML methods and/or including a broader range of data types. We settled on acute fish toxicity because it is a vertebrate-based endpoint required by chemical regulation that has been the focus of many predictive ecotoxicology studies, likely due to high data availability and discussions calling for the need of its replacement [21–23]. Additionally, for the purpose of this analysis, the endpoint needed to be clearly defined (*e.g.*, through a test guideline) to try and compare related studies. Obviously, studies concerned with different taxonomic groups (*e.g.* fish *v.s.* daphnids) will not be based on comparable data. These selection criteria do not rule out studies where several taxonomic groups and/or species and/or effect levels (*e.g.*,

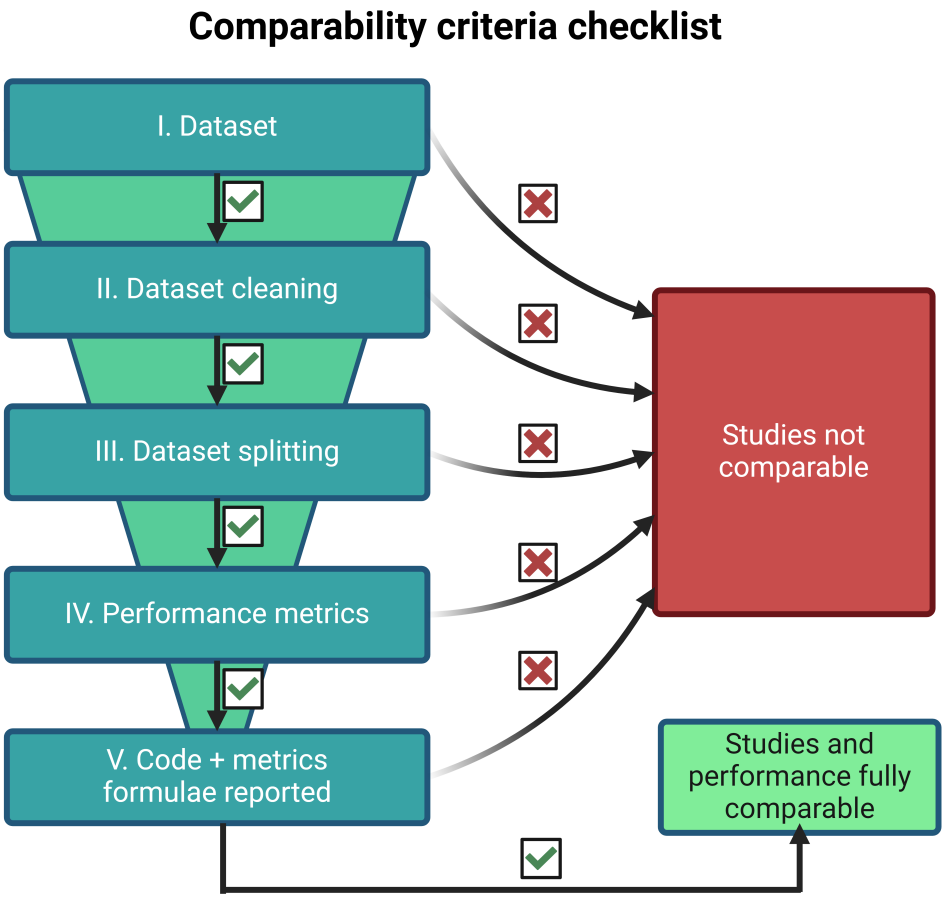


FIG. 1. Overview of the comparability levels the studies were judged across. These criteria are elaborated in Secs. I, II, III, IV and V, in the main text. We find that no two currently available studies on *in silico* methods in ecotoxicology meet all of these criteria at the same time.

acute, chronic, reproduction, growth) have been tested in parallel, but our analysis focuses exclusively on the data related to fish acute toxicity.

### B. Data processing and analysis

The information from the selected studies was extracted from the published PDF version of the publication into a table using ChatGPT 4o. The prompt used for this is given in Appendix A. The results were then manually checked for relevance to this work and relevant information. If necessary, the prompt processing was split into parts or additional prompts were given to extract further details or to clarify certain information. Likewise, manual checking of the extracted information against the PDF was performed when results were

unclear, and for randomly selected studies as a sanity check. Each study was processed individually in a separate prompt window to avoid confusion between the studies. The resulting output was manually evaluated, cleaned, and transferred to the table comprising the dataset our study is based on. Further cleaning, summarizing, and visualization of the data was conducted in R (Version 4.4.1) with RStudio (Version 2024.09.0+375). The code and data of the analysis is available on https://github.com/mbaityje/COMPARABILITY.

### *Limitations*

We acknowledge that the extraction of (sometimes complex) information using a large language model may produce errors. This is particularly likely in the case of the definitions of performance metrics, because they appear to not always be recognizable by the model, especially when not explicitly referred to in the text. The same is true where information is not phrased in an easily comprehensive way or implicitly. Manual checking of a subset of extracted information revealed that the error rate was not higher than what would be expected for a human. Nonetheless, even when expecting a certain error rate and thus consider our analysis more qualitative than strictly quantitative, the emerging patterns remain.

As for the number of considered studies, this could have been increased by extending our analysis to studies considering other taxa (*e.g.*, other common ecotoxicological test organisms, such as crustaceans or algae) or effects and endpoints. However, within ML applied to ecological hazard assessment, fish mortality is the most used test, *i.e.* the one with the highest probability of finding a couple of comparable studies. Thus, extending our work to less used datasets would not change its message.

## RESULTS & DISCUSSION

### I. Datasets

Studies that want to position themselves into a broader context need to be comparable with previous studies, so they should use the same datasets. This criterion comes with the caveat that what ultimately matters for comparability is the applicability domain (AD) of the dataset, *i.e.* the kind of species and chemicals, the range of chemical properties, and

outcome variables (toxicity) that is represented [24]. One cannot expect models to achieve quantitatively comparable performances across datasets that, for example, contain different chemical groups or only partially (if at all) overlapping ranges of molecular weight, octanol-water partition coefficient ($LogK_{ow}$ or $LogP$) or other descriptors. The same is true for data based on differing experimental designs, *i.e.* not derived from the same standardized test guideline, such as the OECD test guideline 203 for fish acute toxicity [25]. Fixed benchmark datasets are commonly used to exclude variability in used datasets as a driver of model performances and to ensure comparability. The 50 studies we include in our analysis contained relevant toxicity data originating from 82 sources (Fig. 2-A shows the datasets that were used at least three times). Many of these data sources, such as the US EPA ECOTOX database, which was the most commonly used single source (Figure 2-A), are updated frequently and are thus not a static resource that can be linked to without indicating the access date/version of the dataset. The dataset source is more clearly defined when it is associated with single scientific paper, which was only the case in 10% of cases. More often, datasets from different sources were combined, resulting in a new dataset (even when all sources were defined individually). In this case, the final dataset a study is based on heavily depends on data cleaning. For the majority of the screened studies ($> 70\%$) the concrete version of the used dataset was not reported (Figure 2-B).

**Take home message**

Studies that want to position themselves into a broader context need to be comparable with previous studies, so they should use the same datasets.

## II. Cleaning

Data cleaning (selecting or removing specific data points, dealing with duplicates, outliers, missing values, standardization of the data) affects the dataset composition and applicability domain and, thus, the questions a researcher is able to ask and the scope of the answer a model is able to give. Potential reasoning underlying these decisions can range from purely domain-specific, like limiting data to a range of a chemical property or species found in a certain geographic region, to considerations stemming from technical limitations (compatibility

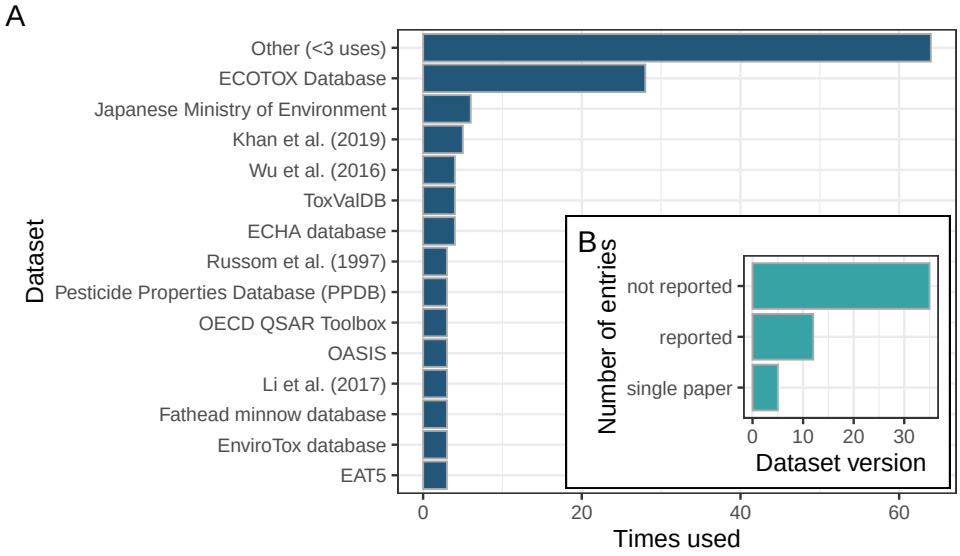


FIG. 2. A: Top distinct data sources used across studies in our analysis. Sources used less than three times are aggregated under "Other"; B: Number of times specific dataset versions were reported or single paper sources were used *v.s.* not reported.

of data with model types, constraints imposed by molecular representations). Often, cleaning is not reported transparently enough for the steps to be fully reproducible (Figure 3). Frequently, it is not even clear if data was aggregated using the geometric or arithmetic mean (Figure 3-A). Studies can only be considered comparable if the dataset and cleaning match and are reported thoroughly, ideally with accompanying code (see section 6). We acknowledge that not all datasets require extensive cleaning, especially those originating from single papers ( 10 % of the studies screened here) or that do not contain duplicates or missing values. These are also included into the "not reported" group.

**Take home message**

Data set version and cleaning must be the same in order to allow for comparability.

The cleaning should be thoroughly reported. Providing scripts that perform the data cleaning is desirable.

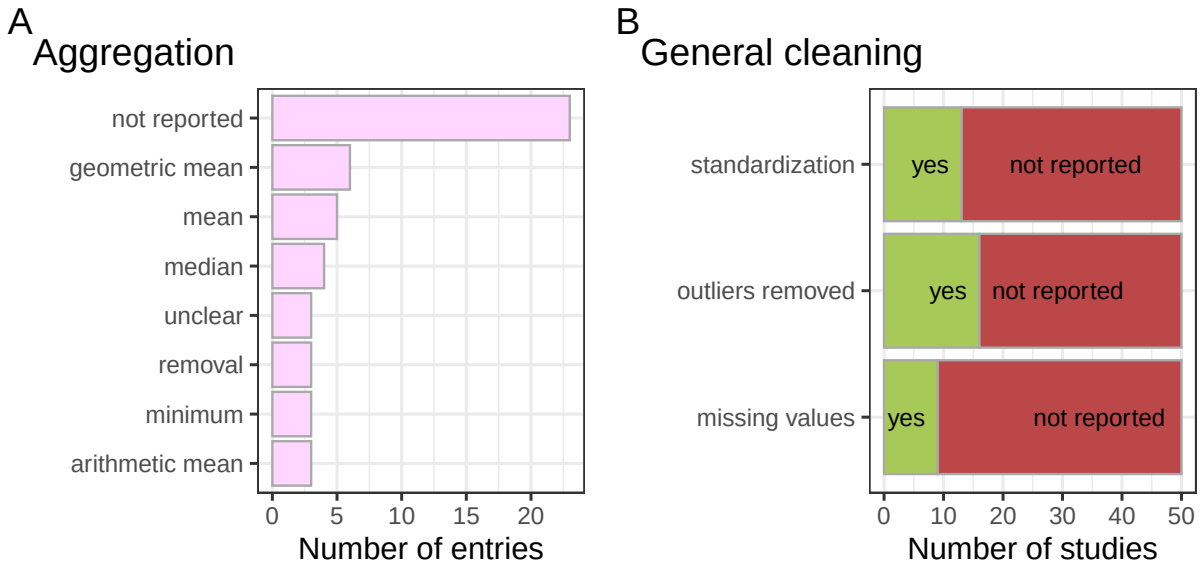


FIG. 3. Reporting of important cleaning steps across studies: A: Aggregation of duplicate values; B: Standardization of data points, treatment of missing values, and removal of outlier values.

## III. Splitting

Splitting is the process of dividing a dataset into the training, validation, and test set and is an integral part of assessing model performance. Model training and hyperparameter-tuning is performed on the training+validation set, while the test set is kept completely separate until the final phase of the study where model performance is put to the test on it. Several ways to achieve a train-test-split are available: Random splitting, stratified random splitting, or splitting according to other aspects of the data, like chemical scaffolds [26] (*i.e.* a shared molecular backbone of molecules). Random splitting is the most commonly used approach (Figure 4-A), and is done by randomly distributing data points between the train and test sets at a pre-determined ratio (often 80 % of the data points go to train and 20 % to test; Figure 4-B). As it was already highlighted in previous work, the way the data is split can artificially inflate performances [27, 28], so it is crucial to ensure that splittings are kept under control. This has recently been pointed out by Stock *et al.* (2023) for ecological applications of machine learning [20]. Gasser *et al.* compared model performances on the ecotoxicological benchmark dataset *ADORE* using the contained random and stratified splittings and found model performance to be more dependent on the

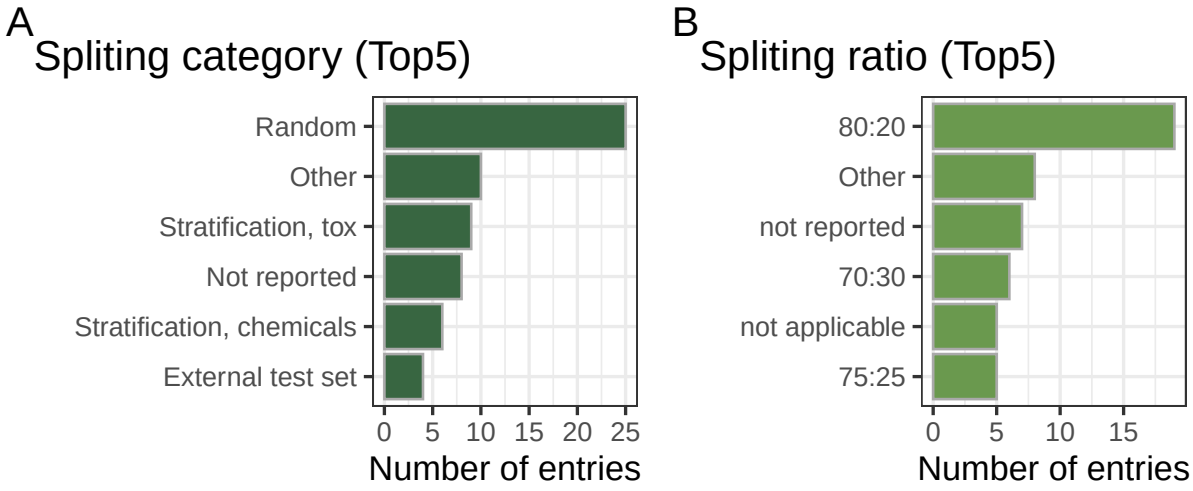


FIG. 4. Overview of the way in which training and testing dataset splitting was achieved (A) and the most often reported splitting ratios (B).

splitting approach than on molecular descriptors [17, 29]. This substantiates the importance of data splitting for assessing and comparing model performances.

**Take home message**

Within a study, even with the same exact dataset, comparisons can only be made if the splittings are exactly the same.

## IV. Metrics

Model performance is assessed through performance metrics, depending on the modeling task and usually on a hold-out test set after training a model on the training set. Several kinds of metrics are commonly used, even though their naming and definition is not always made clear within the studies. Figure 5-A highlights the most commonly reported performance metrics (metrics that were not part of the 10 most common are summarized under "Other"). Reported designations for similar metrics differed between studies, hence they were manually summarized under broad umbrella terms. Sensible performance metrics differ between modeling tasks, *i.e.* regression modeling *vs.* classification. The most common task was regression followed by binary classification, followed by classification with different

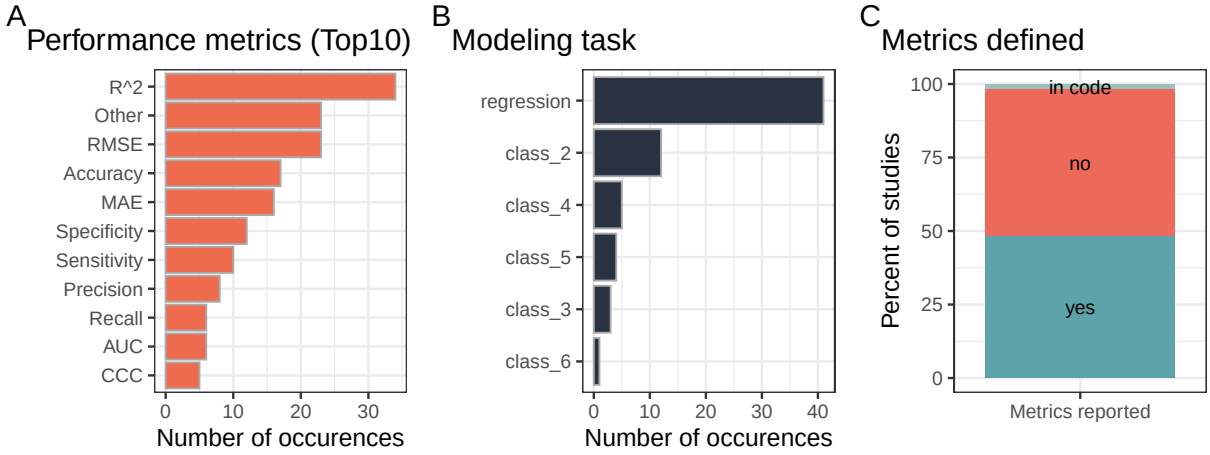


FIG. 5. A: Ten most commonly reported performance metrics (less-commonly used metrics are summarized under "Other"); B: Reported modeling task (class_n refers to classification task with n number of classes); C: Are the metrics clearly defined/reported, either through equations or by referencing specific functions from software packages. RMSE = Root Mean Square Error; MAE = Mean Absolute Error; AUC = Area Under the Curve; CCC = Concordance Correlation Coefficient.

numbers of classes (Figure 5-B). Lastly, we checked how many of the studies defined their metrics, either directly through equations or by referencing a specific function from a software package that was used to calculate (Figure 5-C). Here, roughly 50% of the studies did not adequately define the performance metrics. Concurrently, only stating the name of a metric is often not sufficient to adequately identify them, as we argue in the following.

## Performance metrics

We now highlight that metrics that appear to be the same can sometimes be different. Therefore, one needs to 1) accurately report the formula used to calculate the performances and 2) possibly adhere to the same exact measures. As an example, we will take the $R^2$ metric (often called *coefficient of determination*), which is both the most widely used (see Fig. 5–A), and also problematic as far as comparability is concerned [30–33]. The general expression of the $R^2$ is

$$R^2 = 1 - \frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{true}})^2}{\sum_i (y^{\text{average}} - y_i^{\text{true}})^2}\,, \tag{1}$$

where $y_i^{\text{pred}}$ are the model predictions about each test data point $i$, $y_i^{\text{true}}$ are the ground truths (or measured values), and $y^{\text{average}}$ is the average value of the ground truths. If the $R^2$

is close to 1, the predictions are good, if it is around 0, the predictions are equally as good as taking random points around the average. If they are negative, the model is worse than just taking the average.

The $R^2$ metric simplifies in different ways for specific kinds of models, usually depending on whether these have intercepts or whether they are generalized linear models [30, 33]. For this reason, some softwares and communities that are focused on specific kinds of models, may use specific definitions of the $R^2$ which are equivalent to Eq. (1) only under specific conditions. One example is that, for linear problems, the $R^2$ reduces to the squared Pearson correlation coefficient, $r^2$ [34, 35]. Ref. [30] collects previous literature and shows that 8 different definitions of $R^2$ are used. Through some simple examples, it further shows that each of these metrics can give rise to different values. This multiplicity of ways to calculate the $R^2$ can lead to conflicts in the results, when they are used under conditions that do not guarantee their equivalence to Eq. (1).

A further ambiguity arises when splitting in train/test sets [33].[1] Which data split is $y^{\text{average}}$ evaluated on? Training set, test set, or entire dataset? Ref. [33] states that Eq. (1) "should all relate to test data, not training data", but it also recognizes that others advocate for calculating $y^{\text{average}}$ on the training set. From the point of view of comparability, all strategies are acceptable, as long as it is clearly stated how $y^{\text{average}}$ was calculated, and comparisons are only performed among studies which adopt the same definition. We also highlight that the Mean Square Error (MSE= $\frac{1}{N}\sum_i^N [y_i^{\text{pred}} - y_i^{\text{true}}]^2$) measures similar deviations as the $R^2$, without requiring to calculate $y^{\text{average}}$, thus being less ambiguous. However, as we will see in the next section, even the MSE, as any other performance metric, can be ambiguously defined.

**Take home message**

The exact definition of the used metric should be provided. Some metrics depend on the splittings too, so their definition should also be specified in relation with the splittings.

[1] Our argument also applies to validation splitting, but for the sake of clarity we omit it without loss of generality.

### Relevance of repeated entries, micro and macro averaging for performance metrics

Ecotoxicological experiments are often performed more than once. While, in many cases, one can aggregate these values by *e.g.* taking the median or geometric mean, it is often desirable not to perform any aggregation. This is for example the case of Bayesian approaches, where it is assumed that the output does not necessarily converge to a single value, but it could rather only be well-represented through a distribution of points [17]. Regardless of the reason, if repeated entries are not aggregated, one must be especially careful to, besides not including the same entry in both training and test set, specify how the given metric deals with repeated entries.

As an example, let us take a metric where every single datapoint has the same weight (these are usually called micro-averaged, in opposition to macro-averaged metrics where sets of datapoints are coarsened according to some rule). If 50% of the points in the test set are related to glyphosate tested on one or more species, then models with good performance on glyphosate will be advantaged with respect to others. This can be undesirable, since it may be by pure chance that glyphosate appears in the test set. It can also be desirable, since perhaps we want to use our model in a geographical region where glyphosate is very present. In general, when micro-averaged metrics are used, since those are agnostic about repetitions, every repetition in the test set gives more weight to that chemical. This is equivalent to making the hidden assumption that the chemicals that are tested more often are more relevant for applications, which is not always reasonable. One example of different ways to define a metric which is apparently always the same appears in Ref. [17], where four different definitions of the root mean square error (RMSE) are explicitly provided. These different definitions of the RMSE (micro-averaged, micro average chemicals and macro average taxa, macro average chemicals and micro average taxa, macro-averaging taxa and chemicals) represent different desires at the level of interpretation of the model performance.

**Take home message**

The metrics should take into account repeated entries, by explicitly specifying how they are dealt with.

### V. Providing the code and dataset

Since some metrics can be calculated in different ways and metric names are not always consistent or unique, it is possible that the software used to implement metrics sometimes use faulty formulas for a given use case. The best way to allow for comparability in this case is to allow people to recalculate the metrics themselves. This is achieved through maximal transparency, *i.e.*, making FAIR (Findable, Accessible, Interoperable, Reusable) models with all the code associated with a publication available [12]. Code availability was recently highlighted by Gundersen et al. (2024) to be the most important factor towards successful replication of a ML-based study, irrespective of the quality of documentation of said code [36]. Likewise, Cronin *et al.* (2023) argue that following the FAIR principle could positively affect regulatory acceptance of models [12]. Figure 6-A shows the fraction of screened papers where the corresponding code was made available. Fewer than 50% of studies make their code freely accessible. Reproducibility of published work requires the use of the exact dataset, hence the value provided by freely available benchmark datasets that are published independently of associated works [37]. Figure 6-B indicates that roughly 60 % of the screened papers made their data available or linked to an unambiguous source where the dataset was available. However, if given, the data is often not provided in an easily machine-readable format (*e.g.*, comma-separated values (*csv*)) but rather as a table as part of the supplementary information (often in a word document) that requires further extraction/cleaning.

**Take home message**

A usable code should be provided.

Data sheets with all the necessary information should be easily accessible.

## SUGGESTIONS TOWARDS ACHIEVING COMPARABILITY

Machine-learning-based research in general faces fundamental issues in terms of implementation, reproducibility, and reporting [12, 19, 20, 38–42]. Several researchers addressed these shortcomings and suggested checklist-based solutions to harmonize model implemen-

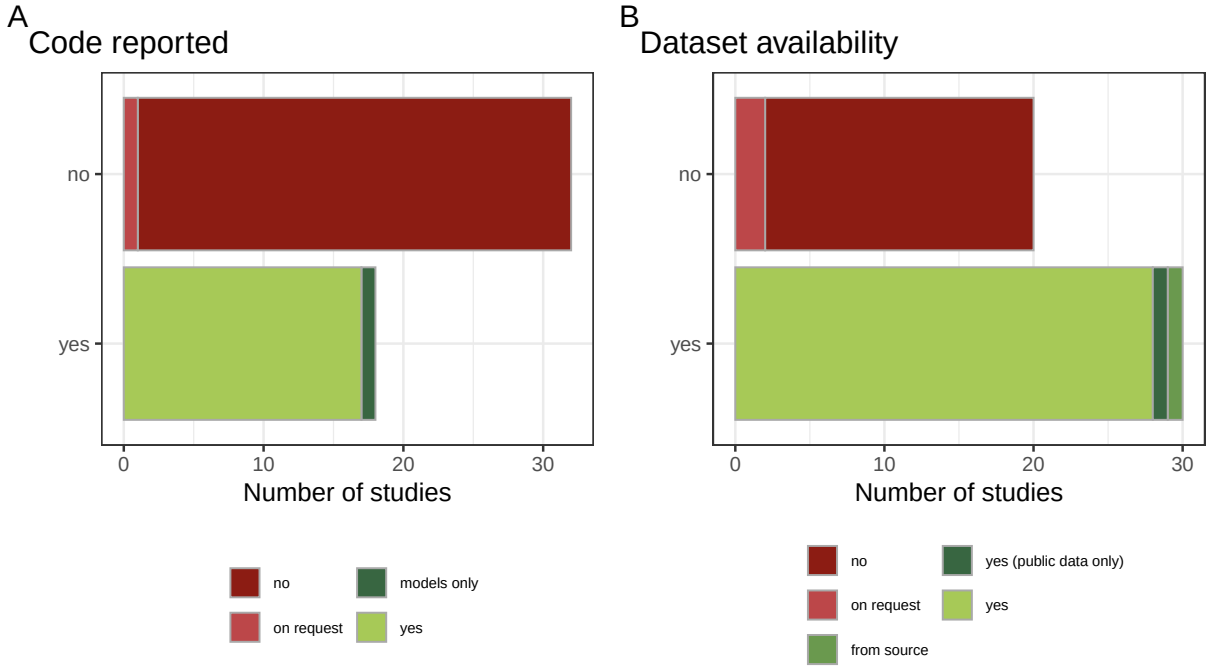


FIG. 6. A: Number of studies that made their code available; B: Number of studies that made their dataset available.

tation and reporting [12, 43–46]. These issues are not unique to any field of ML-based research, but different branches have been quicker or slower in the adoption of best practices. In the field of predictive ecotoxicology, we see that progress is already hindered by the lack of comparability among available studies. Model performances can only reasonably be compared to each other when they were obtained on the same dataset, covering a similar range of species and chemical space. This can be addressed through the use of *benchmark datasets*: curated, freely accessible, and commonly accepted datasets that can form the basis of comparability. Benchmark datasets constituted the basis for AI revolutions in diverse fields, from computer vision [47–49] to hydrology [50, 51], and are arguably a fundamental element needed for an AI revolution in predictive ecotoxicology [52]. Such datasets include documented and explained cleaning, fixed splitting *etc.*, and more importantly, they are "frozen", meaning that everyone using them trains and tests their models in the exact same conditions. Performances on these datasets can then be tracked through centralized documentation databases, such as Papers with Code: *"a free and open resource with Machine Learning papers, code, datasets, methods and evaluation tables"*. Once typical performances

on a specific benchmark data set are established, it becomes straightforward to know how good performances need to be in order to be considered cutting-edge.

Standardized datasets at the same time solve several fundamental problems for different stakeholders: Authors spend less time curating and cleaning data as well as documenting these steps. In turn, authors lacking strong domain knowledge are enabled to pursue relevant problems of more fringe fields of research (think: ecotoxicology *v.s.* human-centerered toxicology). Documentation is further streamlined through the use of standardized reporting sheets, such as REFORMS [45] and EMBRACE [46], and adhering to best practices for the implementation of machine learning and QSARs into ecotoxicology [12, 43, 53, 54]. This also reduces potential frustration about the time investment to find out specific information on model development or the data from a study, if it is given at all. This enables the interested audience to immediately know where to find the desired information for a study and other researchers to know how their model performs compared to others. Oftentimes, the desired information is provided, but hidden inside the text. As editors and reviewers, members of the community can ask researchers to adhere to comparability and reporting standards, in turn reducing time devoted to review work. Special attention needs to be given to definitions and reporting of performance metrics, ideally in concert with code documentation. Communally working on improving the standards of model development and transparency will ultimately progress the field as a whole, getting ML closer to helping progress environmental hazard assessment in ways many are already envisioning it to do.

To provide a solution to this need within the realm of predictive ecotoxicology and to reduce the barrier of entry for researchers lacking the necessary biological/ecotoxicological expertise to compile and clean such data themselves, we recently published the *ADORE* dataset to serve as a benchmark for machine learning in ecotoxicology [29], as have others [55]. This dataset should serve as one point for comparison between methods addressing similar questions, though it must not be the only one for two reasons. First, different benchmark datasets might highlight different kinds of modeling needs (*e.g.*, regression modeling based on many featuresc̃itegasser *v.s.* recommender systems based on few [56]). It would then be insightful to assess whether a method fulfills specific modeling needs, leaving others behind. Second, the usage of benchmark datasets implies that the test set is always the same throughout several studies. However, there is a risk that new methods become tailored

to a specific benchmark dataset [57]. To overcome this, new benchmark datasets should be proposed every few years, in order to avoid that the developed methods are too influenced by the specific choice of a dataset.

Regardless of the scope of the individual datasets, ML learning studies should report their results on as many benchmark datasets as possible. This proves the robustness of the results, and allows to track the progress of the state of the art. Furthermore, the usage of benchmark datasets must be complemented with a precise reporting of the used performance metrics, and reproducible code.

## CONCLUSION

Our analysis of current and past literature on the prediction of acute ecotoxicological outcomes in fish highlights that the usual way of describing and reporting modeling are insufficient to allow to compare model performances on equal grounds to find out which methods are the most effective. To address this, we propose a list of criteria through which one can assess if two studies are comparable to each other. Our examination spans key factors such as dataset selection, modeling tasks, and performance metrics, emphasizing the need for greater consistency and transparency. Concerningly, our findings indicate that no existing studies fully meet these criteria, underscoring a persistent gap in the field. While these concerns should not be new, they remain a significant barrier to progress — one we highlight through a representative bibliographic analysis of supposedly comparable studies. We believe that establishing clearer standards for comparability is a necessary step for advancing machine learning applications in predictive (eco)toxicology and fostering collaborative progress in the field.

## ACKNOWLEDGMENTS

We are thankful to L. Gasser from the Swiss Data Science Center (SDSC) for inspiring conversations. This work was made possible through the SDSC grant “Enhancing Toxicological Testing through Machine Learning” (project NoC20-04) and partly carried out in the framework of the European Partnership for the Assessment of Risks from Chemicals (PARC; Grant Agreement No. 101057014). Figure 1 was created with BioRender.com.

## AUTHOR CONTRIBUTIONS

CS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; KS: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing; MBJ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – original draft. All authors reviewed and approved the final manuscript.

## DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the authors used *ChatGPT 4o* in order to extract information from published PDF versions of the papers screened for analysis. Further information is given in section B. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. No generative AI or AI-assisted technology was used for generating the text of this manuscript.

## REFERENCES

[1] A. E. Urai and C. Kelly, Rethinking academia in a time of climate crisis, eLife **12**, e84991 (2023).

[2] Z. Wang, G. W. Walker, D. C. G. Muir, and K. Nagatani-Yoshida, Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories, Environmental Science & Technology **54**, 2575 (2020).

[3] L. M. Persson, M. Breitholtz, I. T. Cousins, C. A. de Wit, M. MacLeod, and M. S. McLachlan, Confronting unknown planetary boundary threats from chemical pollution, Environmental Science & Technology **47**, 12619 (2013).

[4] L. Persson, B. M. Carney Almroth, C. D. Collins, S. Cornell, C. A. de Wit, M. L. Diamond, P. Fantke, M. Hassellöv, M. MacLeod, M. W. Ryberg, P. Søgaard Jørgensen, P. Villarrubia-Gómez, Z. Wang, and M. Z. Hauschild, Outside the Safe Operating Space of the Planetary Boundary for Novel Entities, Environmental Science & Technology , acs.est.1c04158 (2022).

[5] W. Brack, D. Barcelo Culleres, A. B. A. Boxall, H. Budzinski, S. Castiglioni, A. Covaci, V. Dulio, B. I. Escher, P. Fantke, F. Kandie, D. Fatta-Kassinos, F. J. Hernández, K. Hilscherová, J. Hollender, H. Hollert, A. Jahnke, B. Kasprzyk-Hordern, S. J. Khan, A. Kortenkamp, K. Kümmerer, B. Lalonde, M. H. Lamoree, Y. Levi, P. A. Lara Martín, C. C. Montagner, C. Mougin, T. Msagati, J. Oehlmann, L. Posthuma, M. Reid, M. Reinhard, S. D. Richardson, P. Rostkowski, E. Schymanski, F. Schneider, J. Slobodnik, Y. Shibata, S. A. Snyder, F. Fabriz Sodré, I. Teodorovic, K. V. Thomas, G. A. Umbuzeiro, P. H. Viet, K. G. Yew-Hoong, X. Zhang, and E. Zuccato, One planet: One health. A call to support the initiative on a global science–policy body on chemicals and waste, Environmental Sciences Europe **34**, 21 (2022).

[6] K. Groh, C. vom Berg, K. Schirmer, and A. Tlili, Anthropogenic Chemicals As Underestimated Drivers of Biodiversity Loss: Scientific and Societal Implications, Environmental Science & Technology , acs.est.1c08399 (2022).

[7] EC – European Commission, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC (2006).

[8] K. Mittal, D. Crump, J. A. Head, M. Hecker, G. Hickey, S. Maguire, N. Hogan, J. Xia, and N. Basu, *Resource Requirements for Ecotoxicity Testing: A Comparison of Traditional and New Approach Methods*, Preprint (Pharmacology and Toxicology, 2022).

[9] N. Kleinstreuer and T. Hartung, Artificial intelligence (AI)—it's the end of the tox as we know it (and I feel fine)*, Archives of Toxicology 10.1007/s00204-023-03666-2 (2024).

[10] K. Von Borries, H. Holmquist, M. Kosnik, K. V. Beckwith, O. Jolliet, J. M. Goodman, and P. Fantke, Potential for Machine Learning to Address Data Gaps in Human Toxicity and

Ecotoxicity Characterization, Environmental Science & Technology **57**, 18259 (2023).

[11] L. Posthuma, M. Bloor, B. Campos, K. Groh, A. Leopold, H. Sanderson, H. Schreiber, C. Schür, and P. Thomas, Green Swans countering chemical pollution, Integrated Environmental Assessment and Management **20**, 888 (2024).

[12] M. T. Cronin, S. J. Belfield, K. A. Briggs, S. J. Enoch, J. W. Firman, M. Frericks, C. Garrard, P. H. Maccallum, J. C. Madden, M. Pastor, F. Sanz, I. Soininen, and D. Sousoni, Making in silico predictive models for toxicology FAIR, Regulatory Toxicology and Pharmacology **140**, 105385 (2023).

[13] F. Pistollato, F. Madia, R. Corvi, S. Munn, E. Grignard, A. Paini, A. Worth, A. Bal-Price, P. Prieto, S. Casati, E. Berggren, S. K. Bopp, and V. Zuang, Current EU regulatory requirements for the assessment of chemicals and cosmetic products: Challenges and opportunities for introducing new approach methodologies, Archives of Toxicology **95**, 1867 (2021).

[14] G. P. Daston, C. Mahony, R. S. Thomas, and M. Vinken, Assessing Safety Without Animal Testing: The Road Ahead, Toxicological Sciences **187**, 214 (2022).

[15] P. N. H. Wassenaar, J. Minnema, J. Vriend, W. J. G. M. Peijnenburg, J. L. A. Pennings, and A. Kienhuis, The role of trust in the use of artificial intelligence for chemical risk assessment, Regulatory Toxicology and Pharmacology , 105589 (2024).

[16] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, and A. Tropsha, QSAR without borders, Chemical Society Reviews **49**, 3525 (2020).

[17] L. Gasser, C. Schür, F. Perez-Cruz, K. Schirmer, and M. Baity Jesi, Machine learning-based prediction of fish acute mortality: Implementation, interpretation, and regulatory relevance, Environmental Science: Advances , 10.1039.D4VA00072B (2024).

[18] J. P. Zubrod, N. Galic, M. Vaugeois, and D. A. Dreier, Physiological variables in machine learning QSARs allow for both cross-chemical and cross-species predictions, Ecotoxicology and Environmental Safety **263**, 115250 (2023).

[19] S. Kapoor and A. Narayanan, Leakage and the Reproducibility Crisis in ML-based Science (2022), arXiv:2207.07048 [cs, stat].

[20] A. Stock, E. J. Gregr, and K. M. A. Chan, Data leakage jeopardizes ecological applications of machine learning, Nature Ecology & Evolution **7**, 1743 (2023).

[21] N. Burden, R. Benstead, K. Benyon, M. Clook, C. C. Green, C. C. Green, J. Handley, N. Harper, S. K. Maynard, C. Mead, A. Pearson, K. Ryder, D. Sheahan, R. van Egmond, J. R. Wheeler, and T. H. Hutchinson, Key Opportunities to Replace, Reduce, and Refine Regulatory Fish Acute Toxicity Tests., Environmental Toxicology and Chemistry **39**, 2076 (2020).

[22] I. Katsiadaki, T. Ellis, L. Andersen, P. Antczak, E. Blaker, N. Burden, T. Fisher, C. Green, B. Labram, A. Pearson, K. Petersen, D. Pickford, C. Ramsden, A. Rønneseth, K. Ryder, D. Sacker, C. Stevens, H. Watanabe, H. Yamamoto, F. Sewell, P. Hawkins, H. Rufli, R. D. Handy, S. K. Maynard, and M. N. Jacobs, Dying for change: A roadmap to refine the fish acute toxicity test after 40 years of applying a lethal endpoint, Ecotoxicology and Environmental Safety **223**, 112585 (2021).

[23] C. Schür, M. Paparella, C. Faßbender, G. Stoddart, M. Baity Jesi, and K. Schirmer, Daphnids can safeguard the use of alternative bioassays to the acute fish toxicity test: A focus on neurotoxicity, Environmental Toxicology and Chemistry , vgaf014 (2025).

[24] T. Hanser, C. Barber, S. Guesné, J. F. Marchaland, and S. Werner, Applicability Domain: Towards a More Formal Framework to Express the Applicability of a Model and the Confidence in Individual Predictions, in *Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science*, Challenges and Advances in Computational Chemistry and Physics, edited by H. Hong (Springer International Publishing, Cham, 2019) pp. 215–232.

[25] OECD, *Test No. 203: Fish, Acute Toxicity Test* (Organisation for Economic Co-operation and Development, Paris, 2019).

[26] G. W. Bemis and M. A. Murcko, The Properties of Known Drugs. 1. Molecular Frameworks, Journal of Medicinal Chemistry **39**, 2887 (1996).

[27] S. Kapoor and A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns , 100804 (2023).

[28] F. Pezzicoli, V. Ros, F. Landes, and M. Baity-Jesi, Class imbalance in anomaly detection: Learning from an exactly solvable model, arXiv:2501.11638 (AISTATS 2025).

[29] C. Schür, L. Gasser, F. Perez-Cruz, K. Schirmer, and M. Baity-Jesi, A benchmark dataset for machine learning in ecotoxicology, Scientific Data **10**, 718 (2023).

[30] T. O. Kvålseth, Cautionary note about $R^2$, The American Statistician **39**, 279 (1985).

[31] M. S. Lewis-Beck and A. Skalaban, The r-squared: Some straight talk, Political Analysis **2**, 153 (1990).

[32] A. G. Asuero, A. Sayago, and A. González, The correlation coefficient: An overview, Critical reviews in analytical chemistry **36**, 41 (2006).

[33] D. L. Alexander, A. Tropsha, and D. A. Winkler, Beware of $r^2$: simple, unambiguous assessment of the prediction accuracy of qsar and qspr models, Journal of chemical information and modeling **55**, 1316 (2015).

[34] S. Weisberg, Applied linear regression (2005).

[35] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An introduction to statistical learning*, Vol. 112 (Springer, 2013).

[36] O. E. Gundersen, O. Cappelen, M. Mølnå, and N. G. Nilsen, The Unreasonable Effectiveness of Open Science in AI: A Replication Study (2024), arXiv:2412.17859 [cs].

[37] M. Matveieva and P. Polishchuk, Benchmarks for interpretation of QSAR models, Journal of Cheminformatics **13**, 41 (2021).

[38] R. D. Peng, Reproducible research in computational science, Science **334**, 1226 (2011).

[39] O. E. Gundersen, Y. Gil, and D. W. Aha, On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications, AI Magazine **39**, 56 (2018).

[40] O. E. Gundersen, The fundamental principles of reproducibility, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **379**, 20200210 (2021).

[41] O. E. Gundersen, K. Coakley, and C. Kirkpatrick, Sources of Irreproducibility in Machine Learning: A Review (2022), arXiv:2204.07610 [cs].

[42] J. Achar, M. T. D. Cronin, J. W. Firman, and G. Öberg, A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment, Archives of Toxicology 10.1007/s00204-024-03721-6 (2024).

[43] M. T. Cronin, A.-N. Richarz, and T. W. Schultz, Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction, Regulatory Toxicology and Pharmacology **106**, 90 (2019).

[44] S. J. Belfield, M. T. Cronin, S. J. Enoch, and J. W. Firman, Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs), PLOS ONE **18**, e0282924 (2023).

[45] S. Kapoor, E. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik, P. Nanayakkara, R. A. Poldrack, I. D. Raji, M. Roberts, M. J. Salganik, M. Serra-Garcia, B. M. Stewart, G. Vandewiele, and A. Narayanan, REFORMS: Reporting Standards for Machine Learning Based Science (2023), arXiv:2308.07832 [cs, stat].

[46] J.-J. Zhu, A. B. Boehm, and Z. J. Ren, Environmental Machine Learning, Baseline Reporting, and Comprehensive Evaluation: The EMBRACE Checklist, Environmental Science & Technology 10.1021/acs.est.4c09611 (2024).

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *2009 IEEE conference on computer vision and pattern recognition* (Ieee, 2009) pp. 248–255.

[48] L. Deng, The MNIST database of handwritten digit images for machine learning research, IEEE Signal Processing Magazine **29**, 141 (2012).

[49] A. Krizhevsky, V. Nair, and G. Hinton, (2013).

[50] N. Addor, A. J. Newman, N. Mizukami, and M. P. Clark, The CAMELS data set: Catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences **21**, 5293 (2017).

[51] F. Kratzert, G. Nearing, N. Addor, T. Erickson, M. Gauch, O. Gilon, L. Gudmundsson, A. Hassidim, D. Klotz, S. Nevo, G. Shalev, and Y. Matias, Caravan - A global community dataset for large-sample hydrology, Scientific Data **10**, 61 (2023).

[52] L. D. Burgoon, F. M. Kluxen, A. Hüser, and M. Frericks, The Database Makes the Poison: How the Selection of Datasets in QSAR Models Impacts Toxicant Prediction of Higher Tier Endpoints, Regulatory Toxicology and Pharmacology , 105663 (2024).

[53] T. H. Miller, M. D. Gallidabino, J. I. MacRae, C. Hogstrand, N. R. Bury, L. P. Barron, J. R. Snape, and S. F. Owen, Machine Learning for Environmental Toxicology: A Call for Integration and Innovation, Environmental Science & Technology **52**, 12953 (2018).

[54] S. J. Belfield, S. J. Enoch, J. W. Firman, J. C. Madden, T. W. Schultz, and M. T. Cronin, Determination of "fitness-for-purpose" of quantitative structure-activity relationship (QSAR) models to predict (eco-)toxicological endpoints for regulatory use, Regulatory Toxicology and Pharmacology **123**, 104956 (2021).

[55] P. Svedberg, P. A. Inostroza, M. Gustavsson, E. Kristiansson, F. Spilsbury, and T. Backhaus, Dataset on aquatic ecotoxicity predictions of 2697 chemicals, using three quantitative structure-activity relationship platforms, Data in Brief **51**, 109719 (2023).

[56] M. Viljanen, J. Minnema, P. Wassenaar, E. Rorije, and W. Peijnenburg, What is the ecotoxicity of a given chemical for a given aquatic species? Predicting interactions between species and chemicals using recommender system techniques, SAR and QSAR in Environmental Research **34**, 765 (2023).

[57] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, Do cifar-10 classifiers generalize to cifar-10?, arXiv:1806.00451 (2018).

## Appendix A: ChatGPT prompt

The following prompt was given to ChatGPT to extract the desired information from the PDF version of the studies.

```
Please answer a couple of questions about this scientific article.
    Please do so based on the whole document, not only parts. Do
   it step by step and make the output a table. Include the
   authors, paper title, publication year, and the doi in the
   table. Please answer each of the questions step by step. If
   nothing related to the question is included in the article
   state ‘‘not reported”.

What is the source of the toxicity data? Is there information
   about which database version was used or the date of retrieval
   for the database? Was there more than one data source that was
   used (a combination of several databases or scientific papers
   as sources)? Please specifically list each of the sources
   individually.

What kinds of chemicals are included in the dataset? This could be
    something like organic chemicals, pesticides or some other
   description of the chemical range.

What kind of species or taxonomic groups are included in the
   dataset? How many different species are in the dataset? Was
   there some specific selection criterium that was followed?

Was there a selection of specific experimental parameters, like
   limiting the data to include only specific experimental
   durations (e.g., 24 hours, 48 hours etc.)?
```

Was the data aggregated in any way? Example here include calculation of mean or medians for multiple values or just keeping multiple values as data points.

Was there any mention of missing values and how they were handled? This could include removing them or some form of imputation. If this is the case, please quote the text mentioning this.

Was the data standardized in any way before use? This could mean that it was reduced to minimum or maximum values. If this is the case, please quote the text mentioning this.

Is there a mention of outliers and how they were handled (for example removed according to a specific criterion)?

What was the modeling task? Classification, regression or both? If it was classification, how many classes were used?

What was the effect type that was modeled? This could be something like mortality, lethality, immobilization, intoxication, growth, behavioral alterations etc.

What was the endpoint that was modeled? This could be something like lethal concentration 50 (LC50), effective concentration (EC50), no observed effect concentration (NOEC), lowest observed effect concentration (LOEC).

When splitting the data points in training, validation and test set, did the authors separate the data in groups based on the values of some target variable or were they distributed

```
    randomly? Splitting can be by random, by chemical, by species
    or by some other parameter.

What was the ratio of splitting between either train/training and
    test/testing data? Was there a validation set as well? 80:20 is
     very common, but other ratios are possible.

Is the code for the performed modeling available, either on github
    , on request, through a repository, in the supplementary
    information or not at all?

What re the performance metrices used to evaluate model
    performance? This could be something like R^2, r^2 (note that
    there is a difference between a metric with an upper-case and a
     lower-case r), RMSE, MAE, Accuracy, F1 Score, Specificity,
    Sensitivity. There can be multiple different ones at the same
    time.

Are equations/formulas given for how to calculate the performance
    metrics or are the metrics otherwise defined?

Is the dataset used in the study provided through a repository or
    as part of the supplementary information?

Please summarize all the answers to the questions in a table.
```

## Appendix B: Example ChatGPT output

This section presents an example of the output ChatGPT produced from the prompt in Sec. A for the PDF of Gasser et al. (2024).

| **Question** | **Answer** |
|---|---|
| **Authors** | Lilian Gasser, Christoph Schür, Fernando Perez-Cruz, Kristin Schirmer, Marco Baity-Jesi |
| **Paper Title** | Machine learning-based prediction of fish acute mortality: implementation, interpretation, and regulatory relevance |
| **Publication Year** | 2024 |
| **DOI** | 10.1039/d4va00072b |
| **What is the source of the toxicity data?** | ADORE dataset based on the ECOTOX database, harmonized and pre-filtered. Additional data from various sources were included. No version/date was reported. |
| **What kinds of chemicals are included in the dataset?** | Primarily organic chemicals. |
| **What kind of species or taxonomic groups are included?** | Fish (140 species), crustaceans, and algae. No specific species selection criteria were elaborated. |
| **Was there a selection of specific experimental parameters?** | Yes, data includes specific durations: 24, 48, 72, and 96 hours, transformed for modeling. |
| **Was the data aggregated in any way?** | The dataset was filtered and harmonized, but individual data points were retained without averaging across experiments. |
| **Was there any mention of missing values and how they were handled?** | Not explicitly mentioned. |
| **Was the data standardized before use? If yes, please quote the text.** | Yes, continuous features were standardized based on the training data. Quote: "The continuous variables were standardized based on the training data." |

| | |
|---|---|
| **Is there a mention of outliers and how they were handled?** | Not explicitly mentioned. |
| **What was the modeling task? Classification, regression, or both?** | Regression. |
| **What was the effect type that was modeled?** | Mortality. |
| **What was the endpoint that was modeled?** | Lethal concentration 50 (LC50). |
| **When splitting data into training, validation, and test sets, was the separation random or based on criteria?** | Two splits: random sampling (totally random) and stratification by chemical compound (split by occurrence of chemical compounds). |
| **What was the ratio of splitting between training and test data? Was there a validation set?** | Training and test sets followed an 83:17 ratio. Validation was conducted using 5-fold cross-validation on the training data. |
| **Is the code for the performed modeling available?** | Yes, the code is available through a public repository. Quote: "All code was written in Python 3 using established ML libraries. It is available in a public repository." |
| **What performance metrics were used to evaluate the model performance?** | RMSE (root mean square error), MAE (mean absolute error), $R^2$ (coefficient of determination). |
| **Are equations/formulas given for how to calculate the performance metrics?** | Yes, equations for metrics like RMSE, MAE, and $R^2$ are provided in the appendix. |

| | |
|---|---|
| **Is the dataset used in the study provided through a repository or supplementary information?** | The dataset (ADORE) is a benchmark dataset, but no specific link to a repository was mentioned for direct access in this paper. |