*molecules*

MDPI

*Article*

# Predictive Capability of QSAR Models Based on the CompTox Zebrafish Embryo Assays: An Imbalanced Classification Problem

**Mario Lovrić [1,2], Olga Malev [2,3], Göran Klobučar [3], Roman Kern [1,4], Jay J. Liu [5,*] and Bono Lučić [2,*]**

1 Know-Center, Inffeldgasse 13, 8010 Graz, Austria; mlovric@know-center.at (M.L.); rkern@know-center.at (R.K.)
2 Ruđer Bošković Institute, P.O. Box 180, 10002 Zagreb, Croatia; olga.malev@irb.hr
3 Department of Biology, Faculty of Science, University of Zagreb, Rooseveltov Trg 6, 10000 Zagreb, Croatia; goran.klobucar@biol.pmf.hr
4 Institute of Interactive Systems and Data Science, TU Graz, Inffeldgasse 16c, 8010 Graz, Austria
5 Department of Chemical Engineering, Pukyong National University, Busan 608-739, Korea
* Correspondence: jayliu@pknu.ac.kr (J.J.L.); lucic@irb.hr (B.L.); Tel.: +82-51-629-6453 (J.J.L.); +385-1-456-1111 (B.L.)

**Abstract:** The CompTox Chemistry Dashboard (ToxCast) contains one of the largest public databases on Zebrafish (*Danio rerio*) developmental toxicity. The data consists of 19 toxicological endpoints on unique 1018 compounds measured in relatively low concentration ranges. The endpoints are related to developmental effects occurring in dechorionated zebrafish embryos for 120 hours post fertilization and monitored via gross malformations and mortality. We report the predictive capability of 209 quantitative structure–activity relationship (QSAR) models developed by machine learning methods using penalization techniques and diverse model quality metrics to cope with the imbalanced endpoints. All these QSAR models were generated to test how the imbalanced classification (toxic or non-toxic) endpoints could be predicted regardless which of three algorithms is used: logistic regression, multi-layer perceptron, or random forests. Additionally, QSAR toxicity models are developed starting from sets of classical molecular descriptors, structural fingerprints and their combinations. Only 8 out of 209 models passed the 0.20 Matthew's correlation coefficient value defined a priori as a threshold for acceptable model quality on the test sets. The best models were obtained for endpoints mortality (MORT), ActivityScore and JAW (deformation). The low predictability of the QSAR model developed from the zebrafish embryotoxicity data in the database is mainly due to a higher sensitivity of 19 measurements of endpoints carried out on dechorionated embryos at low concentrations.

**Keywords:** predictive QSAR; toxicity; ToxCast; zebrafish embryo; rdkit; structural descriptors; structural fingerprints; machine learning; imbalanced classification; aquatic toxicology



**Citation:** Lovrić, M.; Malev, O.; Klobučar, G.; Kern, R.; Liu, J.J.; Lučić, B. Predictive Capability of QSAR Models Based on the CompTox Zebrafish Embryo Assays: An Imbalanced Classification Problem. *Molecules* **2021**, *26*, 1617. https://doi.org/10.3390/molecules26061617

Academic Editor: Alla P. Toropova

Received: 5 February 2021
Accepted: 11 March 2021
Published: 15 March 2021



## 1. Introduction

The Computational Toxicology Chemistry Dashboard (CompTox) [1] provides data that have been modeled for determination of Mode of Action (MoA), hazard identification, compound screening, and prioritization, as well as risk assessment for roughly 8000 unique compounds [2]. To date, the CompTox effort has been successful in giving new perspectives to chemical characterization, toxicity testing, and exposure modeling. The CompTox database is a respectable source of toxicity data of chemicals created by the US Environmental Protection Agency (US EPA). It allows for a shift to simultaneous evaluation of numerous chemicals based on disruption of different biological target and signaling pathways using high-throughput screening data and computational approaches [3]. As a part of CompTox, a zebrafish (ZF) developmental toxicity assay has been used to evaluate potential ecotoxicological and harmful effects on humans' health. These animals are easy to

rear and maintain, they mature rapidly and are sufficiently small for sustaining testing in 96-well microtiter plates reducing the cost by means of small amounts of test chemicals [4]. Physiological, molecular, and functional features such as rapid development, optical transparency during the whole embryonic development, well characterized embryonic ontogenesis and ex utero development, availability of genomic data and ~70% of genetic similarity amongst humans and zebrafish make this model appropriate for evaluating a broad range of chemical-biological endpoints across vertebrate taxa [5,6]. Zebrafish assays have been subjected to numerous QSAR studies. Many of these focus on relatively small sets of molecules [7–10] and close-to-mechanistic models. QSARs developed on large data sets on zebrafish embryo toxicity are rare [11–13]. The ZF embryo assays consist of up to 1092 compounds (by ID) tested on diverse developmental malformations. Such data sets are valuable for creating models on broad chemical spaces and low concentrations, which are key for evaluating the risk due to many novel compounds present in mixture at nano-to-micro-concentration ranges in fish [14] and in river surface water and sediment [15], with often unknown MoA and synergistic actions.

Several reports present QSAR models on the subsets of the CompTox ZF library, namely the NHEERL_ZF_144hpf_TERATOSCORE assay, described originally in [4] that uses zebrafish embryos to screen 309 Phase 1 environmental compounds, which are mainly pesticides and antimicrobials. Models developed on the NHEERL_ZF assay are reported also in the literature [11,12]. However, both studies involved only one zebrafish toxicity endpoint (50% mortality data) and the models were developed and validated on data sets of less than 300 compounds from Padilla et al. [4]. They showed reasonable prediction quality for the self-defined toxicity cutoffs, having Matthew's correlation coefficients on the test sets of 0.89 ($n$ = 58) [11] and 0.77 ($n$ = 61) [12]. Furthermore, these studies present a high importance of the LogP descriptor in the models which should not be neglected. Nevertheless, this appears expected given that assays where the chorion (membrane around the embryo) was kept will show association of toxicity with hydrophobicity [5]. It should be noted that the chorion can serve as a permeability barrier to larger molecules [16] and their penetration can vary due to physiochemical properties of compounds, their cationic charge or electrostatic attraction between chemicals and the chorion [17]. Another set of 19 "Tanguay_ZF" CompTox assays [18] were conducted on dechorionated embryos that are reported to be more sensitive than the chorionated ones [19], and represent a bigger challenge in models as reported here and in [20]. It has been demonstrated that with the use of chorion-intact embryos the sensitivity to identify teratogens was higher while the specificity was lower compared to data obtained with dechorionated embryos [21]. The use of dechorionated embryos may involve different agreements between sensitivity and specificity [17], due to the higher mortality of embryos. Dechorionation of embryos is desirable as it removes a potential barrier to chemicals, thus allowing a more effective evaluation of toxicity mechanisms that underlie effects of chemical exposure [22].

The concept of evaluating the predictability or modelability, essentially based on distance and similarity measures, has been developed in several studies [23–26]. The distance or similarity between molecules was estimated from the predictions of activities done by previously developed QSAR models. For every compound in a data set, and on a given set of structural features, the Euclidean distance to its first nearest neighbor was calculated. After that, it is estimated whether its first nearest neighbor compound belongs to the same or to a different activity class, and the total number of those belonging to each class was counted. Then, the modelability index for classification QSAR endpoint is defined as the ratio of compounds having the first nearest neighbor in the same class to the total number of compounds in the data set [24]. Later, also analogous modelability index based on the Euclidean distance measured between compounds in feature space and activity prediction by classification QSAR models was introduced for classification endpoints [25]. An alternative and conceptually simpler method for estimating modelability is the one used by Thomas et al. [26] which is based only on the consideration of predictive capabilities of models comparing with the gain of the model over the level of random

(chance) accuracy. Namely, they considered two sets of features for 309 compounds and 84 classification algorithms to analyze modelability of 60 ToxCast phase I endpoints measured for approximately 300 compounds in the five-fold cross-validation procedure. The final result was that all endpoints are of low modelability. As the main evaluation metrics, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve was used, which has the value of 0.5 in case of random models. The approach applied in this study for analysis of modelability of 19 CompTox endpoints related to zebrafish toxicity is analogous to the one introduced and used by Thomas et al. [26] with the differences in validation methodology and the model evaluation metrics for classification models. Namely, we evaluated the real predictive capabilities of QSAR models on external (never seen) test data set, which is a more objective measure of real model's (endpoints') predictivity or modelability. All mentioned methodologies [23,24,26,27] have applied three-fold or five-fold cross-validation which is a less strict validation procedure than the validation on external test set involving 20% of chemical compounds of the complete data set, as used in this study.

Our aim is to assess the use of the "Tanguay_ZF" CompTox embryo assays for QSARs studies and the building of predictive models for future utilization of chemical compounds and risk assessment (ERA). Majority of published literature on use of QSAR in ERA relies heavily on acute lethal effects (LC50) predictions while sublethal effects are equally or even more important for toxicity assessment of polluted aquatic environment and especially fish organisms. It is therefore of great importance to assess the usefulness of such data on toxic effects on fish organism for possible use in risk assessment of polluted aquatic ecosystems. Additionally, the "Tanguay_ZF" set is conducted on whole organisms, thus the complexity of reactions to specific chemicals is included unlike the majority of the ToxCast data which are based on cell-based in vitro assays. To achieve this, we have tested three different classifiers and chemical representations in a large experiment matrix of 209 model runs (11 models $\times$ 19 targets) to obtain results independent of machine learning, chemical representation (i.e., descriptors and fingerprints) and model hyperparameters. The prediction results are reported on train and test sets by means of multiple evaluation metrics such as the Matthew's correlation coefficient (MCC) [28], Cohen's Kappa [29] and Real-Accuracy (RA) (previously named $\Delta Q_2$ [30]), which were selected due to their ability to capture performance in imbalanced datasets [31–33]. Additionally, the same methodology was applied on a 6–7 times larger set of toxicities from the Tox21 US EPA database. In such a way, the validity of the applied methodology was confirmed through obtained higher values of predictive quality parameters which are completely comparable to the corresponding results of other authors.

## 2. Results

The accuracies and qualities of obtained QSAR models are reported by means of MCC, Real-Accuracy, Balanced Accuracy (BA), and Accuracy across the 209 models. Interested readers are referred to the Supplementary Materials (Table S1) for a full confusion matrix of each model. All model combinations are indexed based on the options they include, i.e., algorithm, scaling, predictive data set, target, and feature selection (see Section 4). To understand the relationships between diverse metrics, we have correlated the values of different quality parameters obtained on the Test set for all models.

The values of the Pearson correlation coefficients are given in Table 1. The results show that MCC and Cohen's Kappa correlate almost perfectly (0.97), while both Kappa and RA correlate above 0.84 with MCC. BA shows negative correlation with all the three afore mentioned metrics ($<-0.19$) so does Accuracy ($<-0.18$).

**Table 1.** Pearson correlation coefficients between quality metrics obtained for the test set across the 209 models.

| | Real Accuracy | MCC | Cohen's Kappa | Accuracy | Balanced Accuracy |
|---|---|---|---|---|---|
| Real Accuracy | 1 | 0.84 | 0.86 | −0.39 | −0.28 |
| MCC | 0.84 | 1 | 0.97 | −0.24 | −0.19 |
| Cohen's Kappa | 0.86 | 0.97 | 1 | −0.18 | −0.21 |
| Accuracy | −0.39 | −0.24 | −0.18 | 1 | 0.59 |
| Balanced Accuracy | −0.28 | −0.19 | −0.21 | 0.59 | 1 |

The comparison of different model quality metrics is also shown in Figure 1. Both BA and Accuracy have high values even for models with MCC values close to 0, which are considered here as random models. Therefore, reporting the model quality only with the Accuracy or BA appears inappropriate. Since MCC highly correlates (0.99) with the Cohen's Kappa, we have transferred the categorization [29] which considers a Cohen's Kappa score below 0.20 as to "slight agreement" (just above random, which is 0), to results reported by MCC and defined a threshold of below 0.20 MCC as slight correlation (just above random).

Scatter plots of model quality parameters for all models (11 models for each of 19 endpoints) are presented in Figure 1. We aim to reveal the model's predictability independently of the set of options considered in the model optimization process. The distributions show that quality metrics of most of models on the endpoints perform in the region defined previously as almost random models (the MCC axis in Figure 1 does not pass MCC 0.4). The criteria being set, 26 out of 209 have MCC Test values above 0.20, while 19 models surpass the same threshold with MCC during cross-validation in model training (MCC CV) (see Table S1). There are eight models which satisfy both criteria (MCC CV and MCC Test > 0.20). Overall, four out of these models are related to the endpoint ActivityScore, three on MORT and one on JAW, and these models have Real-Accuracy values (%) on the test set between 3 and 9%, i.e., all being above the level of the random accuracy. The correlation between Cohen's Kappa and MCC is higher for positive values of both parameters (Figure 1). In addition, these two parameters are identical for models having FN = FP (see Table 3 for definitions).

Distributions of results of each endpoint are present in Figure 2 by boxplots. Endpoints such as embryo survival (MORT) and changes in developmental defects as they relate to the whole embryo (ActivityScore) are considered as apical (robust) endpoints which gather all exposure effects at organism level consequently increasing their relevance.

Even though most of the models in this study show a relatively low MCC, this is not uncommon in biological studies. A recent study by Idakwo et al. [34] on the Tox21 data set, which became a popular data set for many QSAR and machine learning experiments [35–37], shows that some of the toxicological endpoints even when conducted on cell lines can have even negative values for MCC. It is therefore not unexpected that whole organism toxicity at low concentration ranges is hard to model given the MCC metrics which is expected to be more sensitive considering other often employed metrics, such as accuracy, BA, or real accuracy.

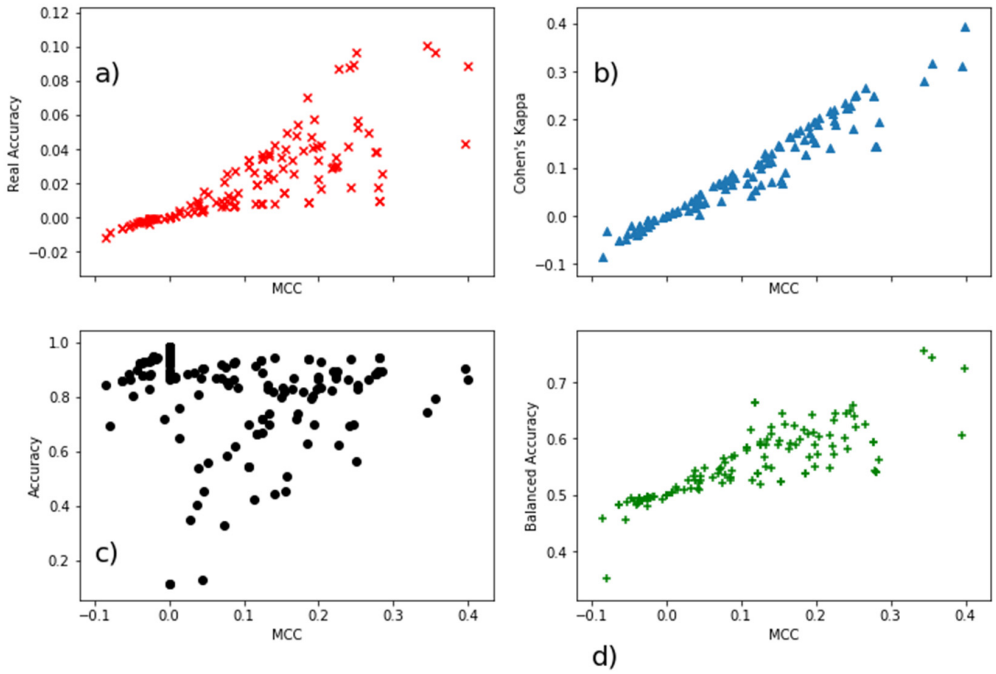
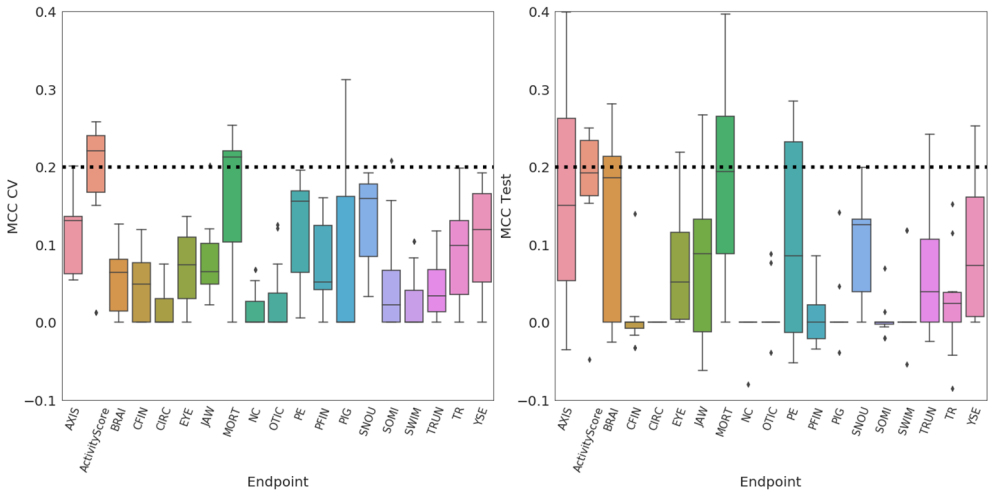



**Figure 1.** Scatter plots of values of four model quality parameters against MCC corresponding to 209 models on the respective test sets (**a**) Real Accuracy, (**b**) Cohen's Kappa, (**c**) Accuracy and (**d**) Balanced Accuracy.




(a) (b)

**Figure 2.** Boxplot diagrams of MCC CV values for the training set (**a**) and MCC Test values for the test set (**b**) for 209 models generated for 19 endpoints (on the X-axis). The threshold MCC value of 0.20 is marked by the dashed horizontal line. Median value of quality metrics for each endpoint is given by horizontal line in each box.

Out of the best models per endpoint which passed the threshold (JAW, MORT, ActivityScore) we have chosen one model each (with the highest average MCC CV and MCC Test) for interpretation. The best model for the endpoint JAW was trained on Morgan fingerprints as molecular features. It is a logistic regression model with a MCC CV of 0.20 and a MCC Test of 0.27. The average permutation importance for this model returned 10 fingerprints as the most important. The structural background (meaning) of 8 out of 10 most important fingerprints is illustrated in Figure 3. Fingerprints are bit-wise strings with zeroes and ones which are folded to a fixed length [38]. Even though they work well in

building QSAR models [39,40], the folding procedure can introduce bit collision [40,41] meaning that different sub-structural fragments can be assigned to the same position in the vector. As we observe this in our own work we followed recommendations to keep a longer vector and shorter radius [41]. Nevertheless, interpretation of fingerprints should be taken with caution. In our case 2 string positions out of the 10 are heavily colliding and not presented in Figure 3. The other two chosen models for MORT and ActivityScore are both trained on RDKit descriptors. The top 10 descriptors chosen by permutation importance for the two models are listed in Table S2. For MORT (MCC CV = 0.21, MCC Test = 0.40), which was based on the Random forests classifier, amongst the most important features appear physico-chemical descriptors such as MolLogP or total polar surface area (TPSA) describing solubility and permeability of compounds through cell membrane or the number of heavy atoms in a molecule. ActivityScore (MCC CV = 0.23, MCC Test = 0.25) models is also based on the random forests classifier and among the most important features (molecular descriptors) appear the total number of NO, pyridine, aryl-COO, alkyl-OH, aryl-OH, and C(O)O groups, as well as of H-bonds acceptors (see more in Table S2). A dominance of the Random forests algorithm is seen in our previous work as well [42,43].

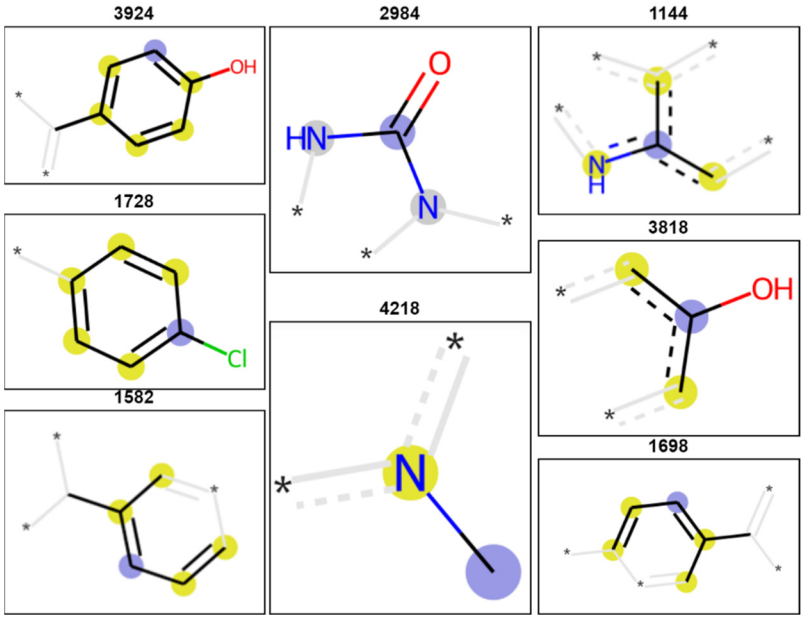



**Figure 3.** Structural fragments presented by fingerprints utilized in the final model for the JAW endpoint. The purple circle denotes the center of the fingerprint with a radius which involves atoms denoted by the yellow-colored circles. The asterisk denotes a continuation of the structure.

Other biologically relevant effects (in specific sublethal endpoints) are less often identified at low or very low concentrations being highly specific and focused on targeted changes which consequently reduce their statistical significance, as confirmed in our analyses.

Using the same methodology, we performed the modelling as the one applied on 19 CompTox endpoints presented above on four selected Tox21 endpoints related to cell-lines toxicity of 6000–7500 compounds. For each endpoint we developed model combinations based on three algorithms (RF, MLP, and LR) using structural fingerprints (FP and DS) as features. Then, two additional models for each endpoint were developed by RF algorithm on data set of physico-chemical descriptors (DS) with and without feature selection. The data set was randomly split into the training (80% compounds) and external test set (20% compounds). The obtained results are summarized in Table S3. All developed models for four Tox21 endpoints and data set of compounds have significantly higher values of MCC (training set: 0.53–0.90; test set: 0.31–0.71) and BA (training set: 0.72–0.97; test set: 0.60–0.80) both for the training and for the test set than in the cases of modeling of 19

endpoints from the CompTox data set of compounds (MCC Test between -0.085 and 0.4). It should be emphasized here that the evaluation metrics values obtained on the Tox21 set were calculated on a 6–7 times larger data set (thus being of higher significance) than the corresponding values for the CompTox set.

However, there is a need to test the quality of obtained Tox21 models in comparison with to other QSAR models performed on the same data set. In 2014, Tox21 Data Challenge [44] was organized in the prediction of 12 Tox21 endpoints. A total of four endpoints we selected for this part of modeling in order to verify the correctness of modelling methodology and the usefulness of sets of structural features (DS and FP) applied to CompTox data are from the Tox21 set of data. We compare our results with four participants [31,33,34,37] on that Tox21 challenge (Table S3), and among them are the results of the winning solutions [37], as well as the second-ranked group [33]. We were able to reproduce two metrics calculated and used in the display of results on the CompTox set, i.e., MCC and BA calculated for the test set. The data set used for training and validation of models was not standardized. Different groups designed and applied different procedures for standardization and cleaning chemical structures in the Tox21 database. Because of that, the training sets used were different in size and the compounds involved. All models developed during the competition were evaluated on the same external set of 647 chemical compounds (approximately 10 % of the data set). In order to have a more robust external (never seen) test set, we decided in our approach to take 80% of data for the training set, and the rest for the test set. (1200–1500 compounds). Thus, our external set for validation of predictive abilities of models is twice as large as in the models with which they are compared [31,33,34,37]. This means that, with a similar value of individual evaluation metrics, the reliability of the parameter related to our method is higher than the method with which we compare here. By MCC Test values our results for four Tox21 endpoints are 0.71, 0.63, 0.37, and 0.57 (Table S3), what is higher than the corresponding values obtained by Abdelaziz et al. 0.25, 0.08, 0.36, and 0.59 (respectively) [33], which are second the best overall results on the Tox21 Data Challenge. Moreover, MCC Test values obtained in this study are noticeably higher than in the study by Idakwo et al. [34] (0.29, 016, 0.62, and 0.55) and for endpoint no. 3 in Uesawa et al. [45] being 0.5 and 0.48 for two cases of dichotomization of toxicity of endpoint SR-MMP (Stress response panel - mitochondrial membrane potential). The comparison of our results with the corresponding models developed by other methodologies [31,33,34,37] by the BA gives analogous results (Table S3). Our results are 0.79, 0.73, 0.70, and 0.77 being completely comparable with the results obtained by the Tox21 Data Challenge winner [31] which are 0.74, 0.65, 0.73, and 0.9, respectively. An important characteristic of models from the study by Abdelaziz et al. [33] given in Table S3 is that, for each property, the best result is selected among 1023 models developed. In the modeling, 10 data sets of molecular descriptors are calculated and used together with other modeling options used in optimization of associative neural networks (ASNN) which were used as the algorithm in ref. [31].

Based on the results of this comparative analysis we can conclude that the methodology applied and sets of structural features calculated and used in modeling 19 CompTox endpoints are correct/valuable and correctly applied in modeling. Therefore, it seems correct to conclude that, within the methodology used and the set of structural features, the CompTox set is poorly modelable set of compounds and endpoints.

## 3. Discussion

Our results show that only three endpoints/targets (ActivityScore, MORT, JAW) can be modeled with a reasonable quality (reported with MCC > 0.20), thus promising that predictions could be above the random correlation level. In addition, the correctness of applied modeling methodology is confirmed by the comparative analysis with other studies in modeling four larger sets of Tox21 endpoints related to cell-toxicities of chemicals. An absolute value of correlation coefficient in the range 0.0 to 0.19 is characterized by many researchers as very weak, then as weak (0.2–0.39), moderate (0.4–0.59), strong (0.6–0.79), and

very strong (0.8–1.0) [46,47]. Thus, because MCC is a variant of the correlation coefficient customized for classification variables, if MCC is in the range 0.0 to 0.19, such a correlation can be considered as close to random. However, if variables in correlation are medium to large like in the case of training and test sets analyzed in this study, then even the lower value of the correlation coefficient can be significant. The presented results show also that Accuracy and BA do not suffice to report classification results in imbalanced scenarios and the quality metrics such as the Cohen's Kappa Score, MCC, and RA must be employed when reporting results on imbalanced sets of data related to toxicity [27]. The recently proposed parameter RA gives important information about the real contribution introduced by the models which are above the random accuracy level. Given as percentage, it shows the percentage contribution of the model to the total accuracy (Accuracy). One can see that all models having MCC > 0.2 have RA > 0. The models developed by [11,12] show comparingly higher MCC values on the ZF developmental endpoint provided by Padilla et al. [4], i.e., the TERATOSCORE at 144 hpf. The results are difficult to compare not only due to the endpoints being different and a lower number of compounds, but also possibly due to the different paradigms applied in data splitting. The mentioned papers utilized techniques such as diversity picking and Kennard–Stone. In our previous work, we suggest that the use informative splitting instead of random splitting can lead to optimistic generalization on external sets [43].

Only few QSAR methods are available to evaluate developmental toxicity [11,12,48–50] and the general lack of quantitative models further justifies our scope in assessing the quality and applicability of CompTox ZF model to predict outcomes on developmental endpoints. The main advantage of models such as those presented is the rapid and simultaneous toxicity prediction of numerous chemicals based on their action on development even though its applicability domain is restricted to only organic compounds. Furthermore, zebrafish dechorionated embryos (CompTox ZF; [18]) are more sensitive to chemicals exposure in comparison to chorionated embryos. The presence [51–53] or absence [18,21,54,55] of the chorion is important because it acts as a moderator of chemicals' contact to embryos and their biological response. Chorion removal increases embryos sensitivity, which is an important trait for chemicals hazard identification using this assay [21]. However, previous reports confirm that the CompTox Zebrafish embryo assay might be difficult to model due to the embryo over-sensitivity that induces high control mortality [20] which is also confirmed by our results highlighting higher model quality only for apical endpoints that sum all negative events such as: Mortality and ActivityScore. Even though dechorination is desirable and promoted, it should be considered that the process of chorion enzymatic removal with pronase probably poses additional stress for ZF embryos [56]. The effect of chorion on developmental toxicity in ZF embryos has previously been investigated following chemical exposure reporting only the effects on phenotypic mortality and morphological traits [21,57]. Recently, researchers found [22] that chorion removal increases embryonic toxicity at the phenotypic level in zebrafish embryos exposed to chemicals adding potential negative effects of dechorination. Taken together, our results suggest that, similar to conclusions by [18] and [58], CompTox ZF assay using embryos survival and overall developmental gross malformations as apical endpoints could help the identification and prioritization of chemicals for more specific, targeted, and MoA-driven testing using ZF embryos as designated model organism.

## 4. Materials and Methods

### *4.1. Data Set and Chemical Representation*

The data was obtained for 1092 compounds (by ID) from the US EPA CompTox Chemicals Dashboard [1]. The zebrafish embryos were assessed for 18 endpoints including yolk sac edema (YSE) and pericardial edema (PE); body axis (AXIS), trunk length (TRUN), caudal fin (CFIN), pectoral fin (PFIN), pigmentation (PIG), and somite (SOMI) deformities; eye (EYE), snout (SNOU), jaw (JAW), and otolith (OTIC) malformations; gross brain development (BRAIN); notochord (NC) and circulatory (CIRC) deformities; swim bladder

presence and inflation (SWIM); touch-responses (TR) and ActivityScore, which represents a cumulative score in the database 18 above mentioned endpoints. The description of the final data set is provided in Table 2.

**Table 2.** Data set overview sorted by the number of active compounds per endpoint. All endpoints are binary variables having only values 1 or 0 (active or inactive). The number of missing data in each endpoint is given in the last column ("missing").

| Endpoint | Negative (0) | Positive (1) | Missing Values |
|---|---|---|---|
| AXIS | 882 | 108 | 28 |
| ActivityScore | 812 | 187 | 19 |
| BRAI | 930 | 60 | 28 |
| CFIN | 942 | 48 | 28 |
| CIRC | 972 | 18 | 28 |
| EYE | 913 | 77 | 28 |
| JAW | 881 | 109 | 28 |
| MORT | 884 | 115 | 19 |
| NC | 977 | 13 | 28 |
| OTIC | 949 | 41 | 28 |
| PE | 874 | 116 | 28 |
| PFIN | 936 | 54 | 28 |
| PIG | 945 | 45 | 28 |
| SNOU | 883 | 107 | 28 |
| SOMI | 952 | 38 | 28 |
| SWIM | 958 | 32 | 28 |
| TRUN | 934 | 56 | 28 |
| TR | 912 | 78 | 28 |
| YSE | 867 | 123 | 28 |

The data set is heavily imbalanced with 13 to 187 active compounds per endpoint, in contrast to 812 to 977 inactive compounds per endpoint.

The data was indexed by the DTXSID and was crosschecked with the SMILES structure mappings. At first, we removed structures which did not have valid SMILES or IDs (15 compounds). Validity of SMILES was checked by the possibility to convert structures to the MOL format [59]. Furthermore, we removed duplicates by ID (19 c.) or SMILES (26 c.). We removed inorganic compounds (7 c.) and metal-containing compounds (7 c.). In the final data set compound were standardized by means of the ChemAxon Standardizer (Marvin/JChemv20.9.0, ChemAxon, Budapest, Hungary). The procedure is inspired by [60] to keep the active part of the compound. The processed data set consisting of 1018 compounds is given [61]. Molecular descriptors (2D, 3D) (DESC) and Morgan fingerprints (FP) for the predictive tasks were calculated for the 1018 structures by means of the RDKit library [62]. The fingerprint vector length was set to 5120 bits and radius to 2, i.e., the distance of 2 bonds in atom neighborhood are considered.

### *4.2. Machine Learning Methods*

We employed three different classifiers in our work: Logistic Regression (LR) [63], Multi-Layer Perceptron (MLP) [64] and Random forests classifiers (RF) [65]. Logistic regression is a classification algorithm (prediction of a binary variable) which is mainly applied in linearly separable problem, even in a multidimension setting. The regression coefficients defining the boundaries of the target classes in feature space are learned from the data and penalized in this work ($L^1$-norm penalty). The hyperparameters to be optimized are usually the regression coefficients (weights, bias) and the penalty. RF is an ensemble classifier. Ensemble classification algorithms are following a paradigm where multiple "weak classifiers" are trained and aggregated to improve the prediction capabilities and lower the prediction error. The weak learners here are decision trees and the aggregation is conducted by means of bootstrapping (each tree trained on a part of data and subset of features) and final voting. RF is considered a non-linear method. The

hyperparameter for RF can be large and complex. Commonly optimized hyperparameters are tree depth, number of trees, class-weights, and the number of features utilized. The MLP is a fully-connected neural network. Neural networks machine learning algorithm where multiple learners are connected in layers. The learners (neurons) learn parameters (weights, bias) from the data and are "activated" by means of a non-linear function such as the sigmoid function. Hyperparameters which are commonly optimized in MLP are the number of layers, penalty function, learning rate and activation function.

The models were trained using the library scikit-learn [66] based on our previous work [42]. Since the endpoints data are imbalanced, we employed penalization and optimization techniques of the model hyperparameters to improve classification outcomes. The data were randomly split into the training (80%) and test set (20%).

To unbias models for misclassification of the minor class (i.e., toxic compounds) during model training we employed the Matthews correlation coefficient (MCC) [28,34] as a scoring function during the model optimization. MCC is defined by Equation (1), where TP, TN, FN, FP are the elements of confusion matrix given in Table 3.

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}} \quad (1)$$

**Table 3.** Elements of confusion matrix.

| | **Positive (Model) (1)** | **Negative (Model) (0)** |
|---|---|---|
| Positive (Experimental) (1) | TP | FN |
| Negative (Experimental) (0) | FP | TN |

Furthermore, we have also utilized $\Delta Q_2$ [30] expressed, which is named here Real-Accuracy (RA) and defined by Equations (2) and (3):

$$\mathrm{RA} = \mathrm{Accuracy} - \mathrm{Random\ accuracy} \quad (2)$$

$$\mathrm{RA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} - \frac{(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TP} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2} \quad (3)$$

Balanced accuracy (BA) is given by Equation (4).

$$\mathrm{BA} = \frac{\mathrm{TP} * (\mathrm{TN} + \mathrm{FP}) + \mathrm{TN} * (\mathrm{TP} + \mathrm{FN})}{2 * (\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{TP})} \quad (4)$$

BA is a highly popular classification model quality measure used in QSAR studies [67]. In addition, the Cohen Kappa score [29] and other classification metrics for results evaluation [68] are also reported in Supplementary Materials (Table S1, since highly imbalanced sets (models) tend to be randomly classified [30]).

### *4.3. Modelling*

The training set was used for Bayesian hyperparameter optimization (BO) [43,69] by means of ten-fold cross-validation (10 × CV). The hyperparameter spaces are given in Appendix A for the three algorithms. BO was utilized with MCC as a loss function. BO aims to construct a posterior distribution of functions (Gaussian process) that best describes the loss function. With a growing number of observations, the posterior distribution becomes narrower, and the algorithm becomes more certain of which regions in the hyperparameter space are worth exploring further. In the process of parameter optimization, the model is continuously re-trained within the train, and the MCC results obtained by each parameter combination are evaluated. Finally, the optimal hyperparameter combination is obtained when a stopping criterion is reached (predefined number of iterations which is 20 in this work).

For dealing with data imbalance we employed two strategies: (1) we have changed the default scoring method during cross-validation which is accuracy [66] to MCC. This means that the model penalizes cross-validation with a more sensitive metric towards imbalance. In our preliminary studies this showed a significant performance improvement comparing to default metrics; (2) we used weighting (class weights) in logistic regression and random forest. The weight ratios which are parts of the hyperparameter tuning (see Appendix A) ranged during the cross-validate hyperparameter optimization from 1:1–1:30 (rare class being 30). In our prior experiment under- and over-sampling, commonly applied in imbalanced settings, also in our previous work [42,70] did not show any results in the models trained here. The final model, described in Section 4.2, have weight ratios range from 1:8.0 (JAW) to 1:14.4 (ActivityScore).

Feature selection was performed by means of stepwise post-hoc permutation importance which showed beneficial properties independent on the basic modeling algorithm applied [43]. The permutation importance was conducted 10x per model to return an average weight. We refer to the set of options applied and considered in the model optimization, i.e., modeling algorithm used (classifier), feature selection, chemical representation and scaling as to "model combinations" which are shown in Table 4.

**Table 4.** Model combinations presented in this work.

| Classifier | Feature Set | * Scaling | ** Feat. Sel. | Endpoints |
|---|---|---|---|---|
| Logistic regression | Fingerprints | No | No | 19 |
| Multilayer perceptron | Fingerprints | No | No | 19 |
| Random forest | Descriptors | No | No | 19 |
| Random forest | Descriptors | No | Yes | 19 |
| Random forest | Fingerprints | No | No | 19 |
| Logistic regression | Descriptors | Yes | No | 19 |
| Logistic regression | Descriptors | Yes | Yes | 19 |
| Multilayer perceptron | Descriptors | Yes | No | 19 |
| Multilayer perceptron | Descriptors | Yes | Yes | 19 |
| Random forest | Descriptors | Yes | No | 19 |
| Random forest | Descriptors | Yes | Yes | 19 |

* Scaling = standardization of features by removing the mean and scaling to unit variance, ** Feat. Sel. = Feature Selection.

This gives 11 distinct model combinations for each of the 19 endpoints, which finally yields 209 separate machine learning models.

## 5. Conclusions

Our research provides insight into the CompTox Zebrafish embryo assays, one of the largest publicly available and most diverse data set on zebrafish aquatic toxicity. We showed that even though there are unique 1018 compounds available, the endpoints are not easy to model by the given chemical features which were utilized here (RDKit physicochemical descriptors and Morgan fingerprints) by using three commonly used classification algorithms (Multilayer perceptron, Random forests, and Logistic regression). Molecular features used here are often used in QSAR modeling and they were calculated by the open software (RDKit). Moreover, we used three open-software classification algorithms for development of QSAR models and, consequently, this methodology can be reproduced by other authors on these data, but also applied in modeling on other problems and sets of chemicals especially in ERA of polluted aquatic ecosystems. In comparison with other methods on four Tox21 data sets and endpoints comparable (and in some cases even better)

models are obtained by the methodology described in this study and applied in modeling of CompTox endpoints, thus confirming its validity. Only three out of 19 endpoints show presence of models above of "slight agreement/correlation" space defined by means of a Matthew's correlation coefficient values (training CV and test set) above 0.20. These three models are ActivityScore, mortality (MORT), and jaw deformation (JAW). We suggest that amongst the limitations might be the experimental methods since ActivityScore and MORT are endpoints that can be predicted somewhat better by developed QSAR models as well as stress-inducing chorion removal. Other endpoints are more difficult to observe even with high-throughput screening and, consequently, it is harder to obtain a good prediction for them by developed QSAR models on the full data sets.

**Supplementary Materials:** A table with full classification results is available in Table S1, more details about the molecular features involved in the best selected models are given in Table S2. Information about QSAR models developed on Tox21 data set and details of comparison with other models developed on selected Tox21 endpoints are given in Table S3.

**Author Contributions:** M.L.—data preparation, concept, machine learning, writing; O.M.—writing, interpretation; R.K.—machine learning, funding, reviewing; G.K.—supervision, writing, reviewing, interpretation; J.J.L.—machine learning, reviewing, funding; B.L.—supervision, writing, machine learning, interpretation. All authors have read and agreed to the published version of the manuscript.

**Funding:** J.J.L. contribution was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (2019R1A2C2084709). The work of B.L. is supported by the Croatian Government and the European Union through the Programme KK.01.1.1.01—The Scientific Centre of Excellence for Marine Bioprospecting—BioProCro.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The processed data used in this study is available online at [61].

**Conflicts of Interest:** The authors declare no conflict of interest.

**Sample Availability:** Not applicable.

## Appendix A

The hyperparameters for the optimization procedure using Scikit-learn [66] are given here:

Random forests: ('max_depth': (5, 10), 'n_estimators': (70, 300), 'max_samples': (0.35, 0.5), n_samples_split': (5, 10), 'class_weight_ratio': (5, 30));

Logistic regression: ('l1_ratio': (−4, 0), 'C': (−2, 0), 'class_weight_ratio': (1, 30));

Multilayer perceptron: ('hidden layer 1': (100, 500), 'hidden layer 2': (20, 100), 'hidden layer 3': (5, 10), 'alpha': (−5, −2)).

## References

1. Williams, A.J.; Grulke, C.M.; Edwards, J.; McEachran, A.D.; Mansouri, K.; Baker, N.C.; Patlewicz, G.; Shah, I.; Wambaugh, J.F.; Judson, R.S.; et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *J. Cheminform.* **2017**, *9*, 1–27. [CrossRef] [PubMed]
2. Morger, A.; Mathea, M.; Achenbach, J.H.; Wolf, A.; Buesen, R.; Schleifer, K.J.; Landsiedel, R.; Volkamer, A. KnowTox: Pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *J. Cheminform.* **2020**, *12*, 1–17. [CrossRef]
3. Thomas, R.S.; Bahadori, T.; Buckley, T.J.; Cowden, J.; Dionisio, K.L.; Frithsen, J.B.; Grulke, C.M.; Maureen, R.; Harrill, J.A.; Higuchi, M.; et al. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. *Toxicol. Sci.* **2020**, *169*, 1–29. [CrossRef]
4. Padilla, S.; Corum, D.; Padnos, B.; Hunter, D.L.; Beam, A.; Houck, K.A.; Sipes, N.; Kleinstreuer, N.; Knudsen, T.; Dix, D.J.; et al. Zebrafish developmental screening of the ToxCastTM Phase I chemical library. *Reprod. Toxicol.* **2012**, *33*, 174–187. [CrossRef]
5. Noyes, P.D.; Garcia, G.R.; Tanguay, R.L. Zebrafish as an: In vivo model for sustainable chemical design. *Green Chem.* **2016**, *18*, 6410–6430. [CrossRef] [PubMed]

6. Pham, D.H.; De Roo, B.; Nguyen, X.B.; Vervaele, M.; Kecskés, A.; Ny, A.; Copmans, D.; Vriens, H.; Locquet, J.P.; Hoet, P.; et al. Use of Zebrafish Larvae as a Multi-Endpoint Platform to Characterize the Toxicity Profile of Silica Nanoparticles. *Sci. Rep.* **2016**, *6*, 1–13. [CrossRef]
7. Ducharme, N.A.; Peterson, L.E.; Benfenati, E.; Reif, D.; McCollum, C.W.; Gustafsson, J.Å.; Bondesson, M. Meta-analysis of toxicity and teratogenicity of 133 chemicals from zebrafish developmental toxicity studies. *Reprod. Toxicol.* **2013**, *41*, 98–108. [CrossRef]
8. Klüver, N.; Vogs, C.; Altenburger, R.; Escher, B.I.; Scholz, S. Development of a general baseline toxicity QSAR model for the fish embryo acute toxicity test. *Chemosphere* **2016**, *164*, 164–173. [CrossRef] [PubMed]
9. Liu, T.; Yan, F.; Jia, Q.; Wang, Q. Norm index-based QSAR models for acute toxicity of organic compounds toward zebrafish embryo. *Ecotoxicol. Environ. Saf.* **2020**, *203*, 110946. [CrossRef]
10. Qiao, K.; Fu, W.; Jiang, Y.; Chen, L.; Li, S.; Ye, Q.; Gui, W. QSAR models for the acute toxicity of 1,2,4-triazole fungicides to zebrafish (Danio rerio) embryos. *Environ. Pollut.* **2020**, *265*, 114837. [CrossRef]
11. Ghorbanzadeh, M.; Zhang, J.; Andersson, P.L. Binary classification model to predict developmental toxicity of industrial chemicals in zebrafish. *J. Chemom.* **2016**, *30*, 298–307. [CrossRef]
12. Lavado, G.J.; Gadaleta, D.; Toma, C.; Golbamaki, A.; Toropov, A.A.; Toropova, A.P.; Marzo, M.; Baderna, D.; Arning, J.; Benfenati, E. Zebrafish AC50 modelling: (Q)SAR models to predict developmental toxicity in zebrafish embryo. *Ecotoxicol. Environ. Saf.* **2020**, *202*, 110936. [CrossRef] [PubMed]
13. Toropov, A.A.; Toropova, A.P.; Benfenati, E. The index of ideality of correlation: QSAR model of acute toxicity for zebrafish (Danio rerio) embryo. *Int. J. Environ. Res.* **2019**, *13*, 387–394. [CrossRef]
14. Malev, O.; Lovrić, M.; Stipaničev, D.; Repec, S.; Martinović-Weigelt, D.; Zanella, D.; Ivanković, T.; Đuretec, V.S.; Barišić, J.; Li, M.; et al. Toxicity prediction and effect characterization of 90 pharmaceuticals and illicit drugs measured in plasma of fish from a major European river (Sava, Croatia). *Environ. Pollut.* **2020**, 115162. [CrossRef]
15. Babić, S.; Barišić, J.; Stipaničev, D.; Repec, S.; Lovrić, M.; Malev, O.; Martinović-Weigelt, D.; Čož-Rakovac, R.; Klobučar, G. Assessment of river sediment toxicity: Combining empirical zebrafish embryotoxicity testing with in silico toxicity characterization. *Sci. Total Environ.* **2018**, *643*, 435–450. [CrossRef] [PubMed]
16. Henn, K.; Braunbeck, T. Dechorionation as a tool to improve the fish embryo toxicity test (FET) with the zebrafish (Danio rerio). *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **2011**, *153*, 91–98. [CrossRef] [PubMed]
17. Nishimura, Y.; Inoue, A.; Sasagawa, S.; Koiwa, J.; Kawaguchi, K.; Kawase, R.; Maruyama, T.; Kim, S.; Tanaka, T. Using zebrafish in systems toxicology for developmental toxicity testing. *Congenit. Anom.* **2016**, *56*, 18–27. [CrossRef]
18. Truong, L.; Reif, D.M.; Mary, L.S.; Geier, M.C.; Truong, H.D.; Tanguay, R.L. Multidimensional in vivo hazard assessment using zebrafish. *Toxicol. Sci.* **2014**, *137*, 212–233. [CrossRef] [PubMed]
19. Villalobos, S.A.; Hamm, J.T.; Teh, S.J.; Hinton, D.E. Thiobencarb-induced embryotoxicity in medaka (Oryzias latipes): Stage-specific toxicity and the protective role of chorion. *Aquat. Toxicol.* **2000**, *48*, 309–326. [CrossRef]
20. Scholz, S.; Klüver, N.; Kühne, R. *Analysis of the Relevance and Adequateness of Using Fish Embryo Acute Toxicity (FET) Test Guidance (OECD 236) to Fulfil the Information Requirements and Addressing Concerns under REACH*; European Chemicals Agency: Helsinki, Finland, 2016.
21. Panzica-Kelly, J.M.; Zhang, C.X.; Augustine-Rauch, K.A. Optimization and performance assessment of the chorion-off [Dechorinated] Zebrafish Developmental toxicity assay. *Toxicol. Sci.* **2015**, *146*, 127–134. [CrossRef] [PubMed]
22. Tran, C.M.; Lee, H.; Lee, B.; Ra, J.S.; Kim, K.T. Effects of the chorion on the developmental toxicity of organophosphate esters in zebrafish embryos. *J. Hazard. Mater.* **2021**, *401*, 123389. [CrossRef]
23. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1–4. [CrossRef]
24. Marcou, G.; Horvath, D.; Varnek, A. Kernel Target Alignment Parameter: A New Modelability Measure for Regression Tasks. *J. Chem. Inf. Model.* **2016**, *56*, 6–11. [CrossRef]
25. Ruiz, I.L.; Gómez-Nieto, M.Á. Study of the Applicability Domain of the QSAR Classification Models by Means of the Rivality and Modelability Indexes. *Molecules* **2018**, *23*, 2756. [CrossRef]
26. Thomas, R.S.; Black, M.B.; Li, L.; Healy, E.; Chu, T.M.; Bao, W.; Andersen, M.E.; Wolfinger, R.D. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicol. Sci.* **2012**, *128*, 398–417. [CrossRef] [PubMed]
27. Ruiz, I.L.; Gómez-Nieto, M.Á. Study of Data Set Modelability: Modelability, Rivality, and Weighted Modelability Indexes. *J. Chem. Inf. Model.* **2018**, *58*, 1798–1814. [CrossRef] [PubMed]
28. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*. [CrossRef] [PubMed]
29. Czodrowski, P. Count on kappa. *J. Comput. Aided. Mol. Des.* **2014**, *28*, 1049–1055. [CrossRef]
30. Lučić, B.; Batista, J.; Bojović, V.; Lovrić, M.; Sović Kržić, A.; Bešlo, D.; Nadramija, D.; Vikić-Topić, D. Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges. *Croat. Chem. Acta* **2019**, *92*. [CrossRef]
31. Kurosaki, K.; Wu, R.; Uesawa, Y. A toxicity prediction tool for potential agonist/antagonist activities in molecular initiating events based on chemical structures. *Int. J. Mol. Sci.* **2020**, *21*, 7853. [CrossRef]
32. Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26*, 1111. [CrossRef]

33. Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.W.; Tetko, I.V. Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. *Front. Environ. Sci.* **2016**, *4*, 1–12. [CrossRef]
34. Idakwo, G.; Thangapandian, S.; Luttrell, J.; Li, Y.; Wang, N.; Zhou, Z.; Hong, H.; Yang, B.; Zhang, C.; Gong, P. Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J. Cheminform.* **2020**, *12*, 1–19. [CrossRef]
35. Hemmerich, J.; Asilar, E.; Ecker, G.F. Conformational Oversampling as Data Augmentation for Molecules. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions*; Tetko, I., Kůrková, V., Karpov, P., Theis, F., Eds.; Springer: Cham, Switzerland; New York, NY, USA, 2019; Volume 11731. [CrossRef]
36. Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; Leblanc, E.; Rennie, P.S.; Welch, W.J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *J. Chem. Inf. Model.* **2018**, *58*, 1533–1543. [CrossRef] [PubMed]
37. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*. [CrossRef]
38. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
39. Kausar, S.; Falcao, A.O. Analysis and comparison of vector space and metric space representations in QSAR modeling. *Molecules* **2019**, *24*, 1698. [CrossRef] [PubMed]
40. Gütlein, M.; Kramer, S. Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. *J. Cheminform.* **2016**, *8*, 1–16. [CrossRef] [PubMed]
41. Landrum, G. RDKit: Colliding Bits III. Available online: http://rdkit.blogspot.com/2016/02/colliding-bits-iii.html (accessed on 23 December 2019).
42. Žuvela, P.; Lovric, M.; Yousefian-Jazi, A.; Liu, J.J. Ensemble Learning Approaches to Data Imbalance and Competing Objectives in Design of an Industrial Machine Vision System. *Ind. Eng. Chem. Res.* **2020**, *59*, 4636–4645. [CrossRef]
43. Lovrić, M.; Pavlović, K.; Žuvela, P.; Spataru, A.; Lučić, B.; Kern, R.; Wong, M.W. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity or predictive ability? *chemrxiv* **2020**. [CrossRef]
44. Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S.A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Front. Environ. Sci.* **2016**, *3*, 85. [CrossRef]
45. Matsuzaka, Y.; Uesawa, Y. Molecular Image-Based Prediction Models of Nuclear Receptor Agonists and Antagonists Using the DeepSnap-Deep Learning Approach with the Tox21 10K Library. *Molecules* **2020**, *25*, 2764. [CrossRef]
46. Wang, Z.; Boulanger, L.; Berger, D.; Gaudreau, P.; Marrie, R.A.; Potter, B.; Wister, A.; Wolfson, C.; Lefebvre, G.; Sylvestre, M.P.; et al. Development and internal validation of a multimorbidity index that predicts healthcare utilisation using the Canadian Longitudinal Study on Aging. *BMJ Open* **2020**, *10*, 1–9. [CrossRef] [PubMed]
47. Correlation and regression. Available online: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression (accessed on 3 March 2021).
48. Hulzebos, E.; Sijm, D.; Traas, T.; Posthumus, R.; Maslankiewicz, L. Validity and validation of expert (Q)SAR systems. *SAR QSAR Environ. Res.* **2005**, *16*, 385–401. [CrossRef] [PubMed]
49. Patlewicz, G.; Ball, N.; Booth, E.D.; Hulzebos, E.; Zvinavashe, E.; Hennes, C. Use of category approaches, read-across and (Q)SAR: General considerations. *Regul. Toxicol. Pharmacol.* **2013**, *67*, 1–12. [CrossRef]
50. Lo Piparo, E.; Worth, A. Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity. *JRC Rep. EUR* **2010**, 24522. [CrossRef]
51. Han, J.; Wang, Q.; Wang, X.; Li, Y.; Wen, S.; Liu, S.; Ying, G.; Guo, Y.; Zhou, B. The synthetic progestin megestrol acetate adversely affects zebrafish reproduction. *Aquat. Toxicol.* **2014**, *150*, 66–72. [CrossRef]
52. McGee, S.P.; Cooper, E.M.; Stapleton, H.M.; Volz, D.C. Early zebrafish embryogenesis is susceptible to developmental TDCPP exposure. *Environ. Health Perspect.* **2012**, *120*, 1585–1591. [CrossRef]
53. Wang, Q.; Liang, K.; Liu, J.; Yang, L.; Guo, Y.; Liu, C.; Zhou, B. Exposure of zebrafish embryos/larvae to TDCPP alters concentrations of thyroid hormones and transcriptions of genes involved in the hypothalamic-pituitary-thyroid axis. *Aquat. Toxicol.* **2013**, *126*, 207–213. [CrossRef]
54. Noyes, P.D.; Haggard, D.E.; Gonnerman, G.D.; Tanguay, R.L. Advanced morphological - behavioral test platform reveals neurodevelopmental defects in embryonic zebrafish exposed to comprehensive suite of halogenated and organophosphate flame retardants. *Toxicol. Sci.* **2015**, *145*, 177–195. [CrossRef]
55. Wilson, L.B.; Truong, L.; Simonich, M.T.; Tanguay, R.L. Systematic Assessment of Exposure Variations on Observed Bioactivity in Zebrafish Chemical Screening. *Toxics* **2020**, *8*, 87. [CrossRef]
56. Mandrell, D.; Truong, L.; Jephson, C.; Sarker, M.R.; Moore, A.; Lang, C.; Simonich, M.T.; Tanguay, R.L. Automated zebrafish chorion removal and single embryo placement: Optimizing Throughput of zebrafish developmental toxicity screens. *J. Lab. Autom.* **2012**, *17*, 66–74. [CrossRef] [PubMed]
57. Kim, K.-T.; Tanguay, R.L. The role of chorion on toxicity of silver nanoparticles in the embryonic zebrafish assay. *Environ. Health Toxicol.* **2014**, *29*, e2014021. [CrossRef]
58. Volz, D.C.; Hipszer, R.A.; Leet, J.K.; Raftery, T.D. Leveraging Embryonic Zebrafish to Prioritize ToxCast Testing. *Environ. Sci. Technol. Lett.* **2015**, *2*, 171–176. [CrossRef]

59. Lovrić, M.; Molero, J.M.; Kern, R. PySpark and RDKit: Moving towards Big Data in Cheminformatics. *Mol. Inform.* **2019**, *38*. [CrossRef] [PubMed]
60. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204. [CrossRef] [PubMed]
61. Lovrić, M. CompTox Zebrafish Developmental Toxicity Processed Data. 2021. Available online: https://zenodo.org/record/4400418#.YE619J0zaUk (accessed on 25 January 2021).
62. Landrum, G. RDKit: Open-Source Cheminformatics Software. Available online: http://rdkit.org/ (accessed on 25 January 2021).
63. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
64. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **1991**, *2*, 183–197. [CrossRef]
65. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
66. Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830. [CrossRef]
67. Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A.M.; Alberga, D.; Alves, V.M.; Andersson, P.L.; Andrade, C.H.; Bai, F.; Balabin, I.; Ballabio, D.; et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ. Health Perspect.* **2020**, *128*, 027002. [CrossRef]
68. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
69. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012, Lake Tahoe, NV, USA, 3–6 December 2012*; Advances in Neural Information Processing Systems: Lake Tahoe, NV, USA, 2012; Volume 4, pp. 2951–2959.
70. Lovric, M.; Banic, I.; Lacic, E.; Kern, R.; Pavlovic, K.; Turkalj, M. Predicting treatment outcomes using explainable machine learning in children with asthma. *Authorea Prepr.* **2020**. [CrossRef]