Environmental Sciences Europe

RESEARCH

Open Access

Towards a reliable prediction of the aquatic toxicity of dyes

Gisela de A. Umbuzeiro[1,2,3*], Anjaina F. Albuquerque[1], Francine I. Vacchi[1,3], Malgorzata Szymczyk[2], Xinyi Sui[2], Reza Aalizadeh[4], Peter C. von der Ohe[5], Nikolaos S. Thomaidis[4], Nelson R. Vinueza[2] and Harold S. Freeman[2]

Abstract

Background: The Max Weaver Dye Library (MWDL) from North Carolina State University is a repository of around 98,000 synthetic dyes. Historically, the uses for these dyes included the coloration of textiles, paper, packaging, cosmetic and household products. However, little is reported about their ecotoxicological properties. It is anticipated that prediction models could be used to help provide this type information. Thus, the purpose of this work was to determine whether a recently developed QSAR (quantitative structure–activity relationships) model, based on ACO-SVM techniques, would be suitable for this purpose.

Results: We selected a representative subset of the MWDL, composed of 15 dyes, for testing under controlled conditions. First, the molecular structure and purity of each dye was confirmed, followed by predictions of their solubility and pKa to set up the appropriate test conditions. Only ten of the 15 dyes showed acute toxicity in Daphnia, with EC_{50} values ranging from 0.35 to 2.95 mg L^{-1}. These values were then used to determine the ability of the ACO-SVM model to predict the aquatic toxicity. In this regard, we observed a good prediction capacity for the 10 dyes, with 90% of deviations within one order of magnitude. The reasons for this outcome were probably the high quality of the experimental data, the consideration of solubility limitations, as well as the high purity and confirmed chemical structures of the tested dyes. We were not able to verify the ability of the model to predict the toxicity of the remaining 5 dyes, because it was not possible to determine their EC_{50}.

Conclusions: We observed a good prediction capacity for the 10 of the 15 tested dyes of the MWDL, but more dyes should be tested to extend the existing training set with similar dyes, to obtain a reliable prediction model that is applicable to the full MWDL.

Keywords: Azo dyes, Anthraquinone dyes, ACO-SVM model, QSAR, MWDL, Daphnia

Background

Following the development of synthetic dyes during the period covering the mid-nineteenth and early twentieth centuries, when dyes were mainly used for textile coloration [1], the end of the twentieth century was marked by an emphasis on dye design for non-textile applications [2]. Consequently, dyes are nowadays used in almost all types of products on the market, including textiles, food, paper, plastics, packaging, biomaterials, lasers, diagnostic products, solar capture, household products and cosmetics. And there is still a search for new applications, especially in the medical arena.

The rapid development of new dye-based commercial products would benefit from the ability to screen large databases containing a wide variety of molecular structures. We believe that the Max Weaver Library (http://www.youtube.com/watch?time_continue=7&v=vIdB1aTx5cY) is such a database, as a repository of 98,000 physical dyes samples donated to the North Carolina State University in 2014 (Fig. 1). It was anticipated that this donation would lead to technological advances for the good of society. To help enable these advances, steps were taken to digitize the dye structures, together with

*Correspondence: giselau@ft.unicamp.br; giselau@unicamp.br
[1] School of Technology, UNICAMP, Campinas, SP, Brazil
Full list of author information is available at the end of the article

Fig. 1 An example of the physical dyes samples in the MDWL

their spectroscopic properties, and to make this information publicly available [3, 4].

Because unspent dyes from coloration processes can end up in freshwater and marine environments, their aquatic toxicity needs to be determined before introducing them to the marketplace (e.g., REACH, 2000). In cases where a lot of candidates are screened, prediction models such as QSARs (quantitative structure–activity relationships) can help in the identification of the less toxic ones.

Ecotoxicity predictions from chemical structures via QSAR models are often restricted by small or biased training sets (i.e., experimental bioassay results of well-known chemicals) as well as limited knowledge about all modes of action involved. Baseline toxicity is assumed to be the minimum toxicity of any neutral organic chemical, which is often associated with the phenomenon of narcosis, and is used as default model in these cases. On the contrary, reactive or specific modes of actions may result in excess toxicity, i.e., being more toxic than expected from narcosis alone. Narcosis toxicity can be predicted quantitatively with good accuracy from chemical structure for various aquatic species, but there is no general model available for predicting the toxicological potency across different modes of action with comparable quality [5].

Building non-generic QSAR models is a way to trade off between prediction accuracy and the application domain. For instance, the existing baseline QSARs sometimes underestimate the acute toxicity for compounds deviating from the octanol–water partition coefficient (log Kow) regression line. In these cases, the prediction accuracy can be enhanced by inclusion of the ionization potency of the chemical or the use of consensus log Kow values from various models. Recently, a QSAR study, based on a non-linear regression method (i.e., a support vector machine) [6], was developed to predict the acute toxicity to *Daphnia magna*. The model has good prediction accuracy for emerging compounds with a wide polarity range. It includes a defined applicability domain and has a rather low prediction error (89.7% of the test data set was predicted with less than a onefold logarithmic error).

In general, the current literature on experimental ecotoxicity values of dyes is rather scarce and many tests were performed in the late 70–80 s, e.g., [7, 8]. At that time, the confirmation of the chemical structures of dyes and information on their purity were often missing. Sometimes, the commercial dye, which usually contains several auxiliaries (e.g., surfactants), was tested and the results were reported as for the dye itself [9–13], confounding the test results.

The purpose of this work was, therefore, to verify whether the recently developed ACO-SVM QSAR model would be a good tool to correctly predict the acute ecotoxicity available from existing experimental data as well as for a newly tested subset of dyes from the MWDL.

Materials and methods

Literature toxicity data for model validation

As a first step, we collected acute toxicity data to the water flea *Daphnia magna* for 22 commercial colorants (dyes and pigments) that were available in the peer-reviewed literature to help validate the ACO-SVM model. However, the data found pertained to 3 water-insoluble organic pigments, 13 sparingly water-soluble disperse dyes, and 6 water soluble (2 FD&C and 4 acid dyes). Because the majority of the dyes in the MWDL belong to the class of disperse dyes, we focused our data collection on dyes of this class. Moreover, we included dyes that are commonly used in detergents and for which experimental toxicity data are available from REACH registration dossiers [14]. Experimental and predicted toxicity data were compiled together with their predicted and, if available, experimental water solubility (Table 1).

Selection of 15 dyes from the MWDL for toxicity testing

Initially, 15 dyes were selected for testing from a group of 200 dyes, previously defined as representative of the MWDL [3]. The selection was made based on or considering a visual inspection of the dye material and the quantity available. Due to limitations in sample quantities, it was important to define a strategy for the most comprehensive evaluation of the dyes, using a minimum amount of sample. For this study, 20 mg of each dye was taken from the library and used for chemical characterization and ecotoxicity testing.

Table 1 Summary of the acute toxicity data reported in the literature, toxicity predictions using the ACO-SVM model, predicted intrinsic solubility and experimental water solubility for 22 commercial dyes and pigments

ID	Dye/pigment	CAS	Experimental EC_{50} (mg L^{-1})	Reference experimental toxicity	Predicted EC_{50} (mg L^{-1})	Similarity	Predicted intrinsic solubility (mg L^{-1})	Experimental water solubility (mg L^{-1})	Reference experimental solubility
DD001	C.I. Pigment Yellow 1	2512-29-0	> 100	[14]	3.6	0.421	1	0.013	[14]
DD002	C.I. Pigment Red 5	6410-41-9	> 100	[14]	0.1	0.409	1	0.0078	[14]
DD003	FD&C Yellow 5	1934-21-0	> 125	[14]	16.1	0.378	310	167,050	[14]
DD004	C.I. Acid Blue 3	3536-49-0	42,900	[14]	0.3	0.390	31	20,980	[14]
DD005	FD&C Blue No. 1	3844-45-9	> 100	[14]	0.4	0.417	1	611,000	[14]
DD006	C.I. Acid Red 52	3520-42-1	> 120	[14]	0.1	0.400	1	95,300	[14]
DD007	C.I. Acid Yellow 3 disodium salt	8004-92-0	> 100	[14]	1.6	0.394	42	200,000–500,000	[14]
DD008	C.I. Acid Blue 80	4474-24-2	> 67	[14]	1.3	0.430	1	10,950	[14]
DD009	C.I. Pigment Blue 16	574-93-6	> 500	[14]	0.1	0.447	1	Not available	
DD010	C.I. Disperse Blue 291	56548-64-2	> 0.02	[15]	0.3	0.392	1		
DD011	C.I. Disperse Blue 373	51868-46-3	> 0.005	[15]	0.3	0.392	1	< 0.052	[14]
DD012	C.I. Disperse Blue 79	12239-34-8	4.5	[29]	0.4	0.419	1	Not available	
DD013	C.I. Disperse Blue 79:1	3618-72-2	4.5	[29]	0.6	0.412	1	< 2	[14]
DD014	C.I. Disperse Orange 1	2581-69-3	10	[30]	0.2	0.508	1	0.00955	[31]
DD015	C.I. Disperse Orange 29	19800-42-1	70	[29]	2.6	0.447	1	< 0.04	[14]
DD016	C.I. Disperse Orange 30	5261-31-4	0.03	[32]	0.6	0.433	1	< 0.04	[14]
DD017	C.I. Disperse Red 1	2872-52-8	0.18	[33]	11.6	0.424	0.16	Not available	
DD018	C.I. Disperse Red 13	3180-81-2	0.0187	[34]	7.8	0.422	0.1		
DD019	C.I. Disperse Red 17	3179-89-3	98	[29]	29.8	0.439	91		
DD020	C.I. Disperse Red 73	16889-10-4	110	[29]	1.1	0.479	1		
DD021	C.I. Disperse Violet 31	6408-72-6	177.9	[32]	0.1	0.455	0.001	0.0026	[14]
DD022	C.I. Disperse Violet 93	52697-38-8	> 0.02	[15]	0.3	0.407	0.001	0.020	[14]

Chemical characterization of the dyes' samples

Each dye of the library is stored in a vial with a label containing its number and chemical formula (Fig. 1). As a quality control procedure, the molecular mass of each dye was confirmed, and the purity determined before acute toxicity testing. Purity analysis was performed on HPLC–MS systems from Thermo Fisher Scientific and Agilent Technology except for dyes 117 and 118, which were only performed in the Agilent instrument.

The exact mass of each dye was determined by an Agilent Technologies 1260 high-performance liquid chromatography (HPLC) system coupled with an Agilent 6520B Q-TOF high-resolution mass spectrometer. To achieve optimum HPLC separation, a gradient mobile phase

composed by water and acetonitrile was used. The proportion of acetonitrile started at 60% and increased to 95%. An Agilent ZORBAX SB-Aq (3.0 × 150 mm, 3.5 µm) reversed phase column was used as the stationary phase. The flow rate was set to 0.5 mL min^{-1} and the total runtime for each sample was 5 min. Ionization was performed via dual electrospray ionization (ESI) system and was carried out in both positive and negative modes with the following parameters: gas temperature 350 °C, drying gas 5 L min^{-1}, nebulizer 50 psi, V_{cap} voltage 3500 V and fragmentor voltage at 175 V. To improve mass accuracy, a solution of the mass reference mix obtained from Agilent was introduced via the secondary ESI needle.

The purity of each dye was checked by an Ultimate 3000 UHPLC system coupled with a Diode Array Detector and a Velos Pro ion trap mass spectrometer from Thermo Fisher Scientific using the same mobile phase and gradient applied to the mass determination. Ionization was performed via heated electrospray ionization (HESI) and was carried out in both positive and negative modes with the following parameters: heater temperature 60 °C, sheath gas flow rate 60 arbitrary unit (arb), auxiliary gas flow rate 20 arb, spray voltage +3 kV/−2.5 kV (positive/negative), capillary temperature 260 °C.

Acute toxicity testing

Stock solutions were prepared in dimethyl sulfoxide (DMSO, Sigma Aldrich, >99.5%) at the limit of solubility of each dye, if the predicted water solubility was low. The test solutions were then prepared in Daphnia media. DMSO was employed at a maximum of 0.1% (v/v) in Daphnia media and this same concentration of DMSO was used as the negative control of the tests [15]. Based on the outcomes of the first experiments, two dyes were re-tested by directly diluting them in Daphnia media for comparison purposes. In those cases, the negative controls consisted of the media itself.

Daphnia similis was chosen as test species, because of a long history of using in aquatic toxicity testing of various chemicals—including dyes and their effluents. Moreover, it is commonly used to conduct environmental in situ studies in water bodies composed of soft waters. Its sensitivity has been compared with *Daphnia magna*, in a study including metals, organics (herbicides, detergents, phenol) and industrial effluents, and the researchers found a 99% agreement in the responses of *D. similis* and *D. magna* [16]. *Daphnia similis* organisms were cultivated in our Laboratory of Ecotoxicology and Genotoxicity (LAEG). Cultures were maintained at 20 ± 2 °C, under a 16:8 h (light/dark) and fed daily with the green algae *Raphidocelis subcapitata*. Total media exchanges were performed three times a week. The sensitivity of the *D. similis* culture was monitored with sodium chloride (NaCl) as a reference substance. The laboratory participates routinely in interlaboratory trials.

Acute toxicity tests were performed according to the guidelines in Test No 202: *Daphnia* sp. Acute Immobilisation Test of Organization for Economic Co-operation and Development [17] and ABNT/NBR 12713 [18]. Twenty neonates (<24 h old) of *D. similis* were placed in 4 replicates for each concentration (5 organisms/replicate). Negative and solvent controls were included and tested in parallel. Tests were performed at 21 ± 1 °C under a photoperiod of 16-h light and 8-h darkness without feeding. The percentage of immobilized organisms was recorded after 48 h.

First, the dyes were tested at the limit of water solubility in a single concentration experiment. This was done to preserve the limited quantity of dyes available in the library. In cases where no effect was observed, no further test was done. The dyes that showed more than 10% of immobile organisms were tested again in concentration–response experiments. The 50% effective concentration (EC_{50}) was calculated for each dye using a non-linear regression based on a logistic distribution of the responses, and the Hill 2 parameters function programmed in Origin (OriginLab, Northampton, MA). When necessary, experiments were repeated for confirmation (data not shown).

Solubility predictions

Solubility calculations were performed using the ALOGpS model [19]. The model was developed using 1291 compounds and provided a low prediction error (RMSE = 0.38). Thirty-eight different atom-type E-state molecular descriptors were used in the model development, which was based on an artificial neural network non-linear regression technique. The atom-type E-state molecular descriptors described information pertaining to the topological environment and the electronic interactions of an atom. The predicted aqueous solubility was expressed as log*S*, where *S* is the solubility in mol L^{-1} and converted into log*S* in mg L^{-1} when compared with the predicted and experimental EC_{50} values. The prediction of aqueous solubility was conducted online at (http://www.vcclab.org/lab/alogps/) [20, 21].

For the highly ionizable dyes, the intrinsic and pH-dependent aqueous solubility was calculated by Marvin Sketch [22]. The prediction was based on a fragment-based method that detects different structural fragments in the compound and assigns an intrinsic solubility contribution to them [23] or corrected solubility at given pH by Henderson–Hasselbalch equation. The contributions are then summed to derive the final intrinsic/pH-dependent solubility value.

QSAR model used for ecotoxicity prediction

The selected QSAR model was recently developed using ACO-SVM techniques [6], to predict the acute toxicity towards the standard test organism *Daphnia magna*. This model was built based on 1006 unique compounds and tested externally with an additional set of 327 compounds. Six molecular descriptors were used to model the toxicity of organic chemicals in the test set. Among the molecular features selected, there were three different measures of logP (i.e., AlogP, CrippenlogP and XlogP) that were found to increase the accuracy of the model in a consensus-like manner, highlighting the importance of this descriptor in predicting the toxicity of organic chemicals [6]. The other descriptors were Average centered Broto–Moreau autocorrelation (lag0) weighted by polarizabilities; Minimum atom-type E-State (centered on –OH); and Overall or summation solute hydrogen bond basicity. To apply this model, the chemical structures of all dyes were standardized by the Balloon program [24]. When generating 3D structures for dyes having multiple tautomeric forms, the tautomer with the lowest energy was used to calculate the six previously mentioned molecular descriptors using PADEL [25], as well as 1024 chemical fingerprints for derivation of the applicability domain [26]. All calculations related to QSAR modelling were performed in MATLAB v 8.5.

Application domain to verify the suitability of the toxicity model

Here, in addition to the effect of the model predictors described above, we have developed a new application domain framework based on the chemical similarity of the suspect dyes to the training set compounds. Chemical similarity is derived based on the presence or absence of 1024 chemical fingerprints in the molecules. The difference between two compounds is then calculated based on the Jaccard Index. The cross matrix of chemical similarity values of the Daphnia training set and the 15 dyes to be tested (1006×15) was derived with a *k* nn *(nearest neighbor)* value set to 3. The *k* value is the number of the most similar compounds to be used to calculate the average chemical structure similarity between the predicted dyes and the compounds of the training set.

The results of the chemical structure similarity approach (*y*-axis) were coupled to the Euclidean distance of a PCA [i.e., the first two principal components (PC1 and PC2)] of the model predictors and hat values to create a density plot (Fig. 2). This allows for comparisons of the molecule-to-molecule activity as well as their chemical structures. Depending on the diversity of the dataset, the acceptable thresholds for chemical structure similarity and Euclidean distance of PCA results can be adjusted. A value close to 1 would indicate that a compound is very similar to, or even part of

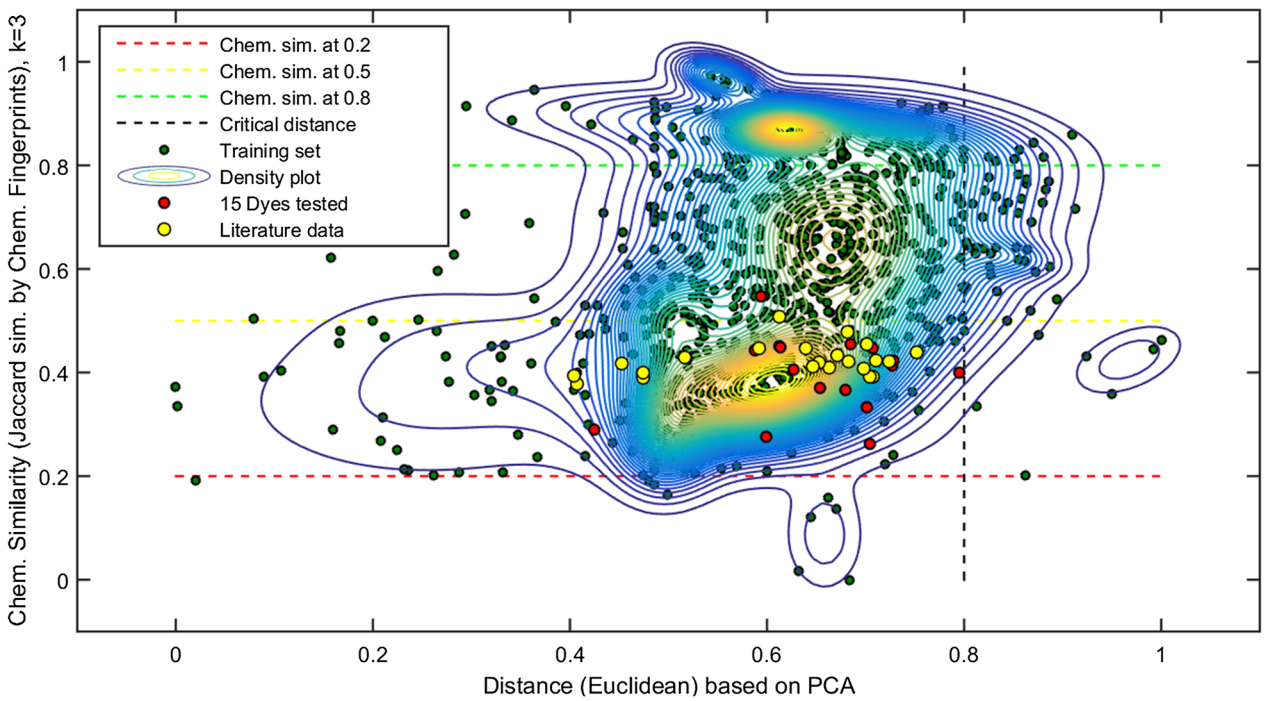


Fig. 2 Density plot of the 1006 training set compounds (green) together with literature data for 22 compounds (yellow dots) as well as for the 15 dyes of the MWDL (red dots)

the training set; while a hypothetical value of 0 would indicate that the new compound does not share a single identical fragment with the training set compounds. We found empirically that values below 50% similarity have significantly higher uncertainty in the model predictions (data not shown), and thus suggest this as a threshold for the suitability of a model to derive predictions with acceptable uncertainty. All calculations related to derive the applicability domain were performed in MATLAB v 8.5.

OTrAMS to verify experimental data

In addition to the density plot, the method "OTrAMS" [27] was used to accept/reject the prediction results when compared to the experimental EC_{50} values. To better compare the toxicity of the various dyes, all measured EC_{50} (mg L^{-1}) values were converted into molar units and the inverse logarithm of the EC_{50} [pEC_{50} (mol L^{-1})] was used [28]. Derivation of pEC_{50} values would enable the direct comparison of experimental and predicted values in the residual plot (in logarithmic scale). The variability among experimental data can often exceed half a log unit, and hence, the QSAR value with its reported prediction error should preferably not be outside of the error of the experimental measurement [28]. A wide acceptance threshold is used here (± 1 log unit) because of the assumption that the dyes have diverse chemical structures and hence, the predication error would be higher.

OTrAMs basically couples three applicability domain approaches in a single 3D bubble plot. In this plot, the *z*-axis shows the Standardized Residuals (SR) (calculated from the predicted and experimental EC_{50} values), the *y*-axis shows the normalized mean distance (i.e., whether the training set compounds are representative of the suspect compound in terms of model predictors) and the *x*-axis relates to the experimental value (i.e., minimum and maximum acute toxicity value in the training set). The bubble size is proportional to the William hat value (i.e., leverage), which shows the individual compounds that are affected dominantly by their diverse molecular descriptor values. Each compound is also coded with a color representing the SR values (green (less than $-1.0 \leq SR \leq 1.0$), yellow ($1.0 < SR \leq 2.0$ or $-2.0 \leq SR < -1.0$), purple ($2.0 < SR \leq 3.0$ or $-3.0 \leq SR < -2.0$) and red ($SR > 3.0$ or $SR < -3.0$)). Since the SRs include the effect of similarity of compounds (based on the molecular descriptors used to model the EC_{50} values) in the error calculation, it can be used to study the origin of the errors between experimental and predicted EC_{50} values. More details about OTrAMS can be found in [27].

Results and discussion

Comparing literature data with predictions

Table 1 shows the experimental toxicity values and the respective predictions from the ACO-SVM model for the dyes that we retrieved data from the literature. Except for two cases, the predicted values were off by more than one order of magnitude of the empirical data.

Table 1 also presents the similarity of the dyes with the training set of the ACO-SVM model, predicted intrinsic solubility and experimental water solubility, when available. Please note that only for 7 dyes, experimental toxicity data were consistent with their experimental water solubility (i.e., DD010, DD011, DD16, DD017, DD018, DD019 and DD022).

For the other dyes, data were inconclusive, mainly for two reasons: some were reported as "non-toxic" (i.e., > values) at concentrations much lower than their actual solubility (e.g., DD03, DD005, DD006 and DD07). Or the opposite, some dyes have EC_{50} values > 100 mg L^{-1}, while their reported solubility is only in the low μg L^{-1} range (e.g., DD001, DD002, DD020 and DD021). These values represent limit tests that are required for the classification and labelling of chemicals [35]. If no effect is observed up to these rather high concentrations, the chemical is classified as "non-toxic". In our case, however, such high concentrations are unlikely to be reached, given the poor water solubility of these dyes (Table 1). Hence, we assume that in these cases, the dyes just precipitated and indeed no toxic effect was observed. However, this rather relates to experimental shortcomings than to real "non-toxicity", as these chemicals would often be expected to bioaccumulate rather quickly. In those cases, the use of passive dosing devices could be useful to evaluate the acute toxicity of poorly water-soluble dyes [36]. Impurities possibly present in the testing material could also have affected the experimental results and be one of the causes of the observed deviations from the experimental and predicted EC_{50}s.

Another reason for the observed discrepancy could be that the selected model is not suitable for these compounds. Figure 2 shows the density plot of the compounds of the training set from the daphnia model as compared to the 22 dyes from the literature. According to the new application domain, predictions would be accepted when the dyes are similar enough to the compounds in the training set, i.e., having a mean chemical similarity (i.e., to the three most similar compounds) above 0.5 and a Euclidean distance in the PCA below 80%. However, all mean chemical similarities were clearly below the threshold of 50%, which was set as a minimum for highly accurate predictions. Figure 3 illustrates how far the predicted data actually are from the experimental ones. Moreover, dyes DD005 and DD009 have bigger

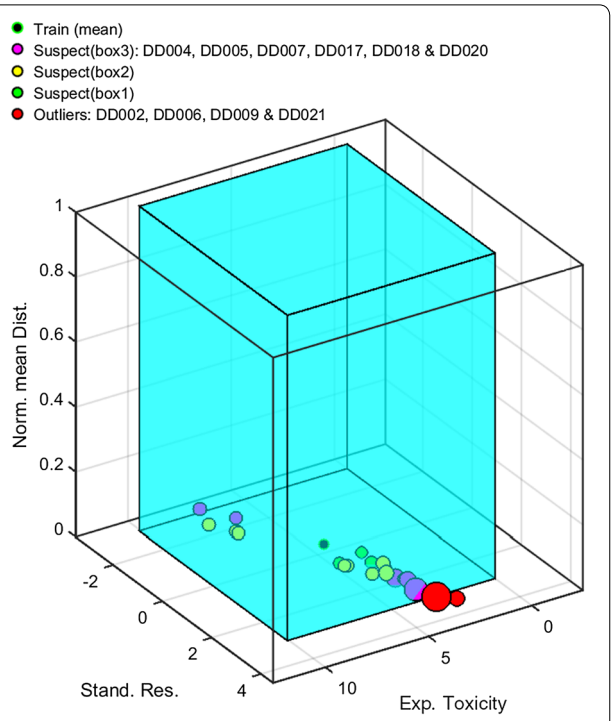


Fig. 3 OTrAMS plot of the experimental and predicted toxicity from the literature data

bubbles which indicate that these dyes also have molecular descriptor values outside the training set domain.

As a consequence, we could not conclude why the model predictions were inaccurate, i.e., whether this was because of the low similarity level of the tested compounds or because of data inconsistencies. Therefore, we concluded that testing additional dyes from the MWDL would provide us with a set of empirical data that could be used to better verify the suitability of the existing model and to decide whether an extension of the model domain will be needed.

Characterization of the molecular structures and purity of the 15 selected dyes

The molecular structures of the tested dyes and their purities are presented in Fig. 4. The molecular structures generally agree with information on the MWDL original vials, except for dye numbers 70 and 117, which had different molecular structures. The respective information was updated in the library accordingly. This finding highlighted the importance of the HR-MS confirmation step before doing any predictions or even testing. The purity of 12 of the 15 selected dyes was greater than (90%) (Fig. 4). Dye 5 had the lowest value (79%), followed by dye 145 (86%) and dye 72 (87%) (Fig. 4). Detailed analytical information can be found in Additional file 1: Table S1.

Experimental toxicity of the 15 dyes

Only ten of the 15 dyes showed acute toxicities with more than 10% of immobilized organisms under the testing conditions (Additional file 1: Table S2). Concentration–response experiments were performed with acute EC_{50} values (Table 2, Additional file 1: Tables S2–S12; Figs. S1–S10) ranging from 0.35 to 2.95 mg L^{-1}. Three dyes (i.e., dyes 9, 70 and 83) with EC_{50} between 1 and 10 mg L^{-1} were, therefore, classified as category II in the GHS system [37]. The other seven dyes were classified as category I ($EC_{50} < 1$ mg L^{-1}) (Table 2). For those 10 dyes, the observed EC_{50} were below or in the range of the predicted solubility, which was not the case for the remaining 5 dyes.

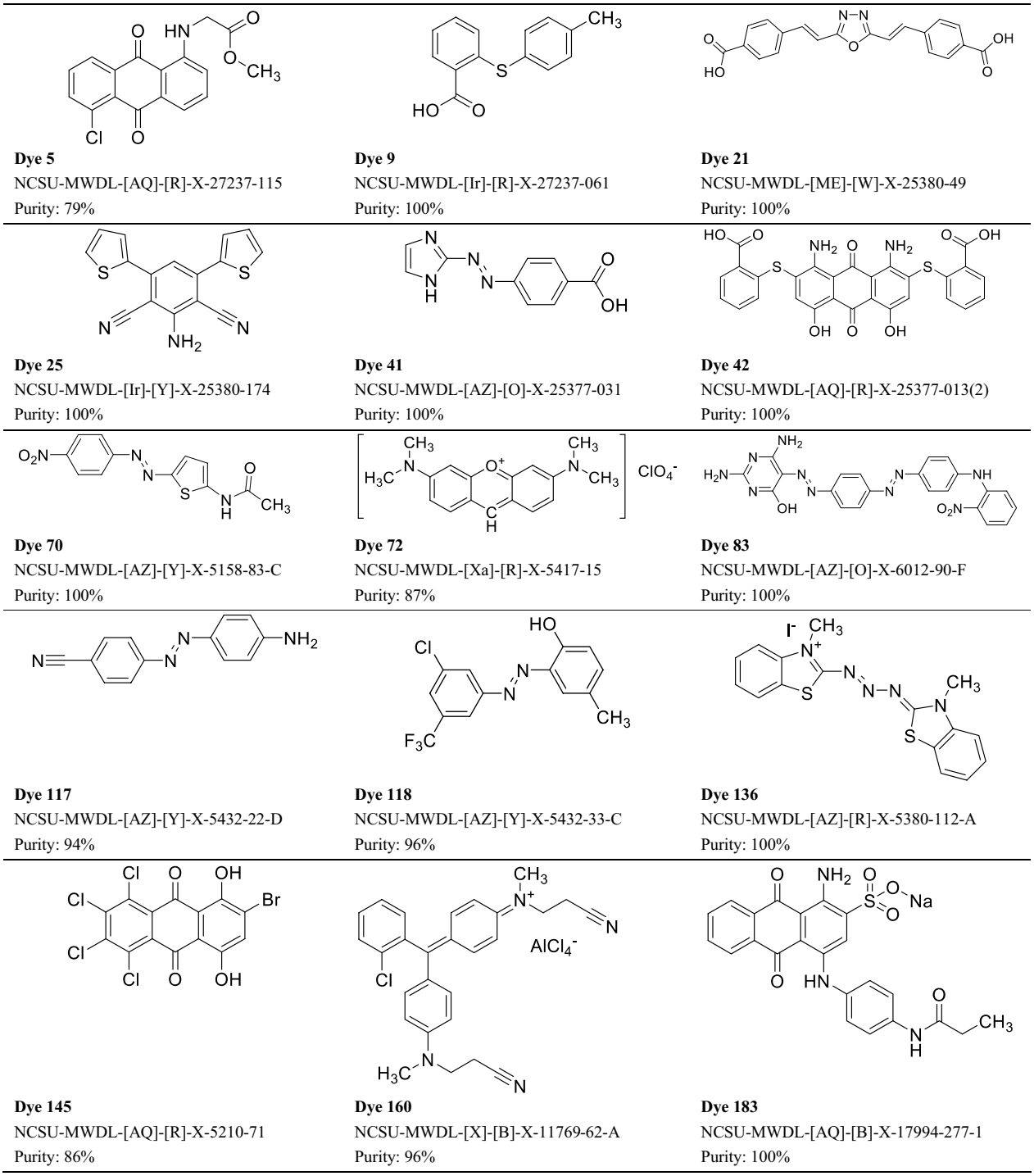
In fact, dye 21 was tested up to the maximum solubility in DMSO at a concentration of 1.3 mg L^{-1}. However, because its water solubility was predicted to be 20 mg L^{-1}, we also prepared a solution directly in Daphnia media. We observed 20% of immobility at 10 mg L^{-1}, but at 20 mg L^{-1}, precipitation occurred without toxic effect (Table 3). Although toxicity was observed for this dye, it was not possible to determine a reliable EC_{50}.

For dye 25, the predicted toxicity was 0.17 mg L^{-1} (Table 2), but no toxicity was observed when we tested the dye even at higher concentrations than the predicted water solubility (Additional file 1: Table S2).

Dye 41 presented the highest predicted water solubility (440 mg L^{-1}) and it is also highly ionizable (Additional file 1: Fig. S12). However no toxicity was observed when the dye was tested in DMSO at 12.6 mg L^{-1} (Additional file 1: Table S1). Therefore, we performed a test with higher concentrations, diluting the dye directly in Daphnia media as we did for dye 21. Negative results were obtained until 20 mg L^{-1} (Table 4), but at 40 mg L^{-1}, 100% of the organisms were immobile. However, the pH dropped (5.10), which was also observed in the higher concentrations (Table 4). This dye is a weak acid (Additional file 1: Fig. S12), which would be consistent with the reduced pH observed at the higher concentrations. However, the dye still precipitated at the two highest concentrations (Table 4). Therefore, it was again not possible to determine a reliable EC_{50} for this dye. The tests could be repeated, adjusting the pH, in buffered Daphnia media.

Dye 42 was also tested at higher concentrations (6.4 mg L^{-1}) than the predicted water solubility when DMSO was used to prepare the dye solution (Table 1). However, no toxic effects were obtained. We tried to prepare higher concentrations to verify if any toxic effect would occur, but this time, the pH increased at unacceptable levels; so, no further ecotoxicity tests were performed.

Both dyes 41 and 42 are examples of how important it is to test the dyes in Daphnia media after adjusting the pH. However, a protocol for testing with buffered



Dye 5
NCSU-MWDL-[AQ]-[R]-X-27237-115
Purity: 79%

Dye 9
NCSU-MWDL-[Ir]-[R]-X-27237-061
Purity: 100%

Dye 21
NCSU-MWDL-[ME]-[W]-X-25380-49
Purity: 100%

Dye 25
NCSU-MWDL-[Ir]-[Y]-X-25380-174
Purity: 100%

Dye 41
NCSU-MWDL-[AZ]-[O]-X-25377-031
Purity: 100%

Dye 42
NCSU-MWDL-[AQ]-[R]-X-25377-013(2)
Purity: 100%

Dye 70
NCSU-MWDL-[AZ]-[Y]-X-5158-83-C
Purity: 100%

Dye 72
NCSU-MWDL-[Xa]-[R]-X-5417-15
Purity: 87%

Dye 83
NCSU-MWDL-[AZ]-[O]-X-6012-90-F
Purity: 100%

Dye 117
NCSU-MWDL-[AZ]-[Y]-X-5432-22-D
Purity: 94%

Dye 118
NCSU-MWDL-[AZ]-[Y]-X-5432-33-C
Purity: 96%

Dye 136
NCSU-MWDL-[AZ]-[R]-X-5380-112-A
Purity: 100%

Dye 145
NCSU-MWDL-[AQ]-[R]-X-5210-71
Purity: 86%

Dye 160
NCSU-MWDL-[X]-[B]-X-11769-62-A
Purity: 96%

Dye 183
NCSU-MWDL-[AQ]-[B]-X-17994-277-1
Purity: 100%

Fig. 4 Structure, designation, and purity level for MWDL dye evaluated in this study

Daphnia media still needs to be developed in our laboratory. A priori pKa predictions (Additional file 1: Fig. S12) can, therefore, be very helpful to define appropriate testing conditions for these dye in future studies.

Comparing experimental toxicity results with predictions

We used the residual plot analysis instead of a correlation approach to compare the predictions with the experimental values, because the experimental gradient was rather narrow (i.e., about two orders of magnitude). Nine

Table 2 Acute toxicity data for *Daphnia similis*, together with the predicted values, chemical similarity with the training set and their predicted solubility

Dye	NCSU code	Result	EC_{50} (mg L^{-1})	Confidence interval	ACO-SVM EC_{50} (mg L^{-1})	Similarity	Intrinsic solubility (mg L^{-1})
5	NCSU-MWDL-[AQ]-[R]-X-27237-115	Positive	0.94	0.63–0.99	14.07	0.427	10
9	NCSU-MWDL-[Ir]-[R]-X-27237-061	Positive	1.85	1.28–1.98	3.09	0.420	1
21	NCSU-MWDL-[ME]-[W]-X-25380-49	Negative	–	–	26.44	0.428	20
25	NCSU-MWDL-[Ir]-[Y]-X-25380-174	Negative	–	–	0.17	0.373	1
41	NCSU-MWDL-[AZ]-[O]-X-25377-031	Negative	–	–	30.60	0.399	440
42	NCSU-MWDL-[AQ]-[R]-X-25377-013(2)	Negative	–	–	5.82	0.447	1
70	NCSU-MWDL-[AZ]-[Y]-X-5158-83-C	Positive	2.95	1.79–3.99	2.92	0.333	10
72	NCSU-MWDL-[Xa]-[R]-X-5417-15	Positive	0.40	0.40–0.41	2.82	0.471	30
83	NCSU-MWDL-[AZ]-[O]-X-6012-90-F	Positive	1.04	0.74–1.41	2.48	0.364	1
117	NCSU-MWDL-[AZ]-[Y]-X-5432-22-D	Positive	0.86	0.59–1.26	1.04	0.466	140
118	NCSU-MWDL-[AZ]-[Y]-X-5432-33-C	Positive	0.73	0.40–1.33	0.40	0.447	1
136	NCSU-MWDL-[AZ]-[R]-X-5380-112-A	Positive	0.35	0.33–0.36	2.02	0.404	1
145	NCSU-MWDL-[AQ]-[R]-X-5210-71	Positive	0.57	0.33–0.94	0.77	0.547	1
160	NCSU-MWDL-[X]-[B]-X-11769-62-A	Positive	0.79	0.55–0.81	1.34	0.456	1
183	NCSU-MWDL-[AQ]-[B]-X-17994-277-1	Negative	–	–	2.68	0.435	1

Table 3 Toxicity data for dye 21 diluted in Daphnia media

Concentration (mg L^{-1})	Total immobilized organisms	%	pH
Control	0/20	0	6.75
1.25	0/20	0	
2.5	1/20	5	
5	2/10	10	
10	4/20	20[b]	
20[a]	1/20	5	6.33

[a] Dye precipitated, [b] more than 10% indicate toxicity

Table 4 Toxicity data for dye 41 diluted in Daphnia media

Concentration (mg L^{-1})	Total	%	pH
Control	0/20	0	6.75
5	2/20	10	6.57
10	2/20	10	6.44
20	6/20	5	6.23
40	20/20	100[b]	5.10
60[a]	20/20	100	4.72
80[a]	20/20	100	4.60

[a] Dye precipitated, [b] values greater than 10% indicate toxicity

of the 10 dyes with EC_{50} data (i.e., 9, 41, 70, 72, 83, 117, 118, 136, 145 and 160) were predicted with acceptable accuracy, i.e., with a prediction error within ± 1 log unit (Fig. 5). Only dye number 5 had a higher error. According to Additional file 1: Fig. S1, the concentration–response

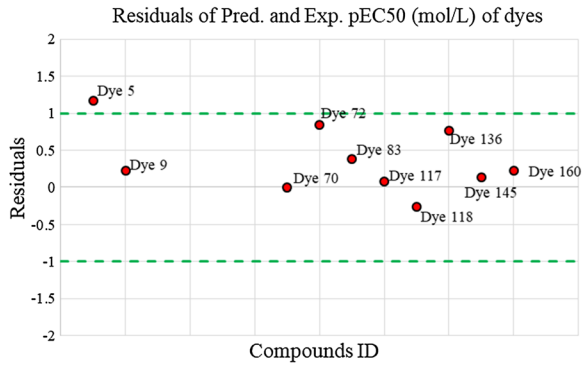


Fig. 5 Residual plot (error) for the 10 acutely toxic dyes (logarithmic scale)

curve for dye 5 shows a much higher toxicity value (EC_{50} = 0.94 mg L^{-1}) than predicted (14.07 mg L^{-1}) (Table 2). However, the predicted toxicity was in fact even higher than the predicted solubility, and precipitation started to occur at 5 mg L^{-1}.

We further investigated this dye to find the origin of the larger prediction error. As the dye has a moderate ionization potency and the major chemical macrospecies from pH 2 up to pH 8 is the neutral form (Additional file 1: Fig. S11), pH was not an issue. However, this was the dye with the lowest purity level (79%), and therefore, we could not rule out that some of the impurities that might be better soluble in DMSO could have been responsible for the observed toxicity. This highlights the importance of

choosing dyes with high purity. Our suggestion is to use purities higher than 90% in further studies to minimize their possible inference.

Although there was a rather low chemical similarity between each of the 10 dyes and the training set compounds of the model, the predicted EC_{50} values were still reasonably accurate. This could be an indication that the predictor space of the model (i.e., the PCA axis) was well covered. Therefore, we believe that the model is generally capable of predicting the toxicity of dyes, at least with medium accuracy, if they are located below 80% distance of the PCA axis (Fig. 2). However, with regard to a more general applicability of the model, there is a need to extend the existing model domain to dyes with higher structural similarity to enable a proper read-across approach and to have an overall higher accuracy of the estimated EC_{50} values.

Selection of additional dyes for future testing and model extension

A similarity analysis of the currently available digitalized dataset of the MWDL (around 3000 dyes) will be conducted with the 10 dyes that provided toxic effects to *Daphnia similis* in this study. Also, if needed, a manual search will be performed in the actual MWDL, because the dyes that will be tested should have a similarity of > 80% to the ten already tested dyes, as well as among themselves, to create a stable model extension. For that purpose, their purity will first be determined, and a confirmation of their molecular structures will be performed before testing or modelling. Only dyes with appropriate quality, i.e., with confirmed structure and showing at least 90% purity will be selected. Prediction of solubility and pKa will help to define the best strategy for testing in relation to the selection of solvents, the maximum concentrations to be tested and the need of buffer solutions to optimize testing conditions. Only then, the experimental toxicity data of those new dyes can be used to extend the training set of the ACO-SVM model.

Conclusion

We concluded that the confirmation of the molecular structure and purity of a dye is required to obtain reliable toxicity results. Solubility issues and the pKa should be taken into account before designing the toxicity experiments, e.g., by selecting the appropriate solvent, defining the maximum concentrations and the use of buffer solutions for testing. The ACO-SVM model used here was able to predict the toxicity of 10 dyes of the MWDL with good accuracy, but there is still a need for more dye compounds of higher similarity with the already tested dyes to extend the existent training set of the ACO-SVM model. Therefore, the next steps will be to select a new set of dyes to obtain additional toxicity data values, hopefully resulting in a prediction model that is applicable to the whole MWDL.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12302-019-0258-1.

Additional file 1. Additional tables and figures.

Abbreviations

ACO-SVM: Ant Colony Optimization-Support Vector Machine; DMSO: dimethyl sulfoxide; EC_{50}: effective concentration 50%; HR-MS: high-resolution mass spectrometry; MWDL: Max Weaver Dye Library; PCA: principal component analysis; QSAR: quantitative structure–activity relationship; SR: standardized residuals.

Acknowledgements

Nothing to declare.

Authors' contributions

GAU, NV, PVO and HSF—design the study, results interpretation and discussion, manuscript writing; AFA, FIV, XS and MS—laboratory experiments, data interpretation, manuscript writing; RA and NST—computational data, data interpretation, manuscript writing. All authors read and approved the final manuscript.

Funding

Fundação de Amparo à Pesquisa do Estado de São Paulo FAPESP Grant # 2017/19599-0 for GAU. CAPES for FIV fellowship.

Availability of data and materials

Additional material presents raw data and also laboratory records are available for verification, if required.

Ethics approval and consent to participate

No ethics approval or consent to participate required for the conducted study.

Consent for publication

The author and all the co-authors agreed with the publication of the article in ESEU.

Competing interests

The author declare that they have no competing interests.

Author details

[1] School of Technology, UNICAMP, Campinas, SP, Brazil. [2] Wilson College of Textiles, North Carolina State University, Raleigh, NC, USA. [3] Biology Institute, UNICAMP, Campinas, SP, Brazil. [4] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece. [5] Amalex Environmental Solutions, Leipzig, Germany.

Received: 6 June 2019 Accepted: 7 September 2019
Published online: 08 October 2019

References

1. Zollinger H (2003) Color chemistry: syntheses, properties, and applications of organic dyes and pigments. Wiley, New York
2. Freeman HS, Peters AT (2000) Colorants for non-textile applications. Elsevier, New York
3. Kuenemann MA, Szymczyk M, Chen Y et al (2017) Weaver's historic accessible collection of synthetic dyes: a cheminformatics analysis. Chem Sci 8:4334–4339. https://doi.org/10.1039/C7SC00567A

4. Williams TN, Van Den Driessche GA, Valery ARB et al (2018) Toward the rational design of sustainable hair dyes using cheminformatics approaches: step 2. Identification of hair dye substance database analogs in the max weaver dye library. ACS Sustain Chem Eng 6:14248–14256. https://doi.org/10.1021/acssuschemeng.8b02882
5. Kühne R, Ebert R-U, von der Ohe PC et al (2013) Read-across prediction of the acute toxicity of organic compounds toward the water flea *Daphnia magna*. Mol Inform 32:108–120. https://doi.org/10.1002/minf.201200085
6. Aalizadeh R, von der Ohe PC, Thomaidis NS (2017) Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization-Support Vector Machine QSTR models. Environ Sci Process Impacts 19:438–448. https://doi.org/10.1039/C6EM00679E
7. Little LW, Lamb JC, Chillingworth MA, Durkin WB (1974) Acute toxicity of selected commercial dyes to the fathead minnow and evaluation of biological treatment for reduction of toxicity. In: Proceedings of the 29th industrial waste conference. Purdue University Libraries, pp 524–534
8. Anliker R, Clarke EA, Moser P (1981) Use of the partition coefficient as an indicator of bioaccumulation tendency of dyestuffs in fish. Chemosphere 10:263–274. https://doi.org/10.1016/0045-6535(81)90026-6
9. Novotný Dias N, Kapanen A et al (2006) Comparative use of bacterial, algal and protozoan tests to study toxicity of azo- and anthraquinone dyes. Chemosphere 63:1436–1442. https://doi.org/10.1016/j.chemosphere.2005.10.002
10. Verma Y (2008) Acute toxicity assessment of textile dyes and textile and dye industrial effluents using *Daphnia magna* bioassay. Toxicol Ind Health 24:491–500. https://doi.org/10.1177/0748233708095769
11. Vinitnantharat S, Chartthe W, Pinisakul A (2008) Toxicity of reactive red 141 and basic red 14 to algae and waterfleas. Water Sci Technol 58:1193–1198. https://doi.org/10.2166/wst.2008.476
12. Darsana R, Chandrasehar G, Deepa V et al (2015) acute toxicity assessment of reactive red 120 to certain aquatic organisms. Bull Environ Contam Toxicol 95:582–587. https://doi.org/10.1007/s00128-015-1636-z
13. Wong CK, Liu XJ, Lee AOK, Wong PK (2006) Effect of azo dyes on survivorship, oxygen consumption rate, and filtration rate of the freshwater *Cladoceran moina* macrocopa. Hum Ecol Risk Assess An Int J 12:289–300. https://doi.org/10.1080/10807030500531604
14. European Chemicals Agency (2019) European Chemicals Agency. Information on Chemicals. Registered substances. https://echa.europa.eu/information-on-chemicals/registered-substances. Accessed 25 May 2019
15. Umbuzeiro GA, Szymczyk M, Li M et al (2017) Purification and characterization of three commercial phenylazoaniline disperse dyes. Color Technol 133:513–518. https://doi.org/10.1111/cote.12307
16. Buratini SV, Bertoletti E, Zagatto PA (2004) Evaluation of *Daphnia similis* as a test species in ecotoxicological assays. Bull Environ Contam Toxicol 73:878–882. https://doi.org/10.1007/s00128-004-0508-8
17. OECD (2004) Test No. 202: *Daphnia* sp. Acute immobilisation test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris. https://doi.org/10.1787/9789264069947-en
18. ABNT (2016) ABNT NBR 12713—Ecotoxicologia aquática—toxicidade aguda—Método de ensaio com *Daphnia* spp (Crustacea, Cladocera). ABNT, Rio de Janeiro
19. Tetko IV, Tanchuk VY, Kasheva TN, Villa AEP (2001) Estimation of aqueous solubility of chemical compounds using E-state indices. J Chem Inf Comput Sci 41:1488–1493. https://doi.org/10.1021/ci000392t
20. VCCLAB (2005) Virtual Computational Chemistry Laboratory. http://www.vcclab.org. Accessed 20 May 2019
21. Tetko IV, Gasteiger J, Todeschini R et al (2005) Virtual computational chemistry laboratory—design and description. J Comput Aided Mol Des 19:453–463. https://doi.org/10.1007/s10822-005-8694-y
22. ChemAxon (2019) Marvin 6.3.1, 2014. Calculator Plugins. Toolkit for structure property prediction and calculation
23. Hou TJ, Xia K, Zhang W, Xu XJ (2004) ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. J Chem Inf Comput Sci 44:266–275. https://doi.org/10.1021/ci034184n
24. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47:2462–2474. https://doi.org/10.1021/ci6005646
25. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/jcc.21707
26. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
27. Aalizadeh R, Thomaidis NS, Bletsou AA, Gago-Ferrero P (2016) Quantitative structure–retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples. J Chem Inf Model 56:1384–1398. https://doi.org/10.1021/acs.jcim.5b00752
28. Cherkasov A, Muratov EN, Fourches D et al (2014) QSAR modeling: where have you been? where are you going to? J Med Chem 57:4977–5010. https://doi.org/10.1021/jm4004285
29. Environment Canada (2017) Screening Assessment Aromatic Azo and Benzidine-based Substance Grouping Certain Azo Disperse Dyes. http://www.ec.gc.ca/ese-ees/. Accessed 13 May 2019
30. Ferraz ERA, Grando MD, Oliveira DP (2011) The azo dye Disperse Orange 1 induces DNA damage and cytotoxic effects but does not cause ecotoxic effects in *Daphnia similis* and *Vibrio fischeri*. J Hazard Mater 192:628–633. https://doi.org/10.1016/j.jhazmat.2011.05.063
31. U.S. Environmental Protection Agency (2019) Benzenamine, 4-[(4-nitrophenyl)azo]-*N*-phenyl. In: U.S. Environ. Prot. Agency. Chem. Dashboard. comptox.epa.gov/dashboard/DTXSID7062536%0A
32. Wang H, Ii L, Wu G, Wei Y (2014) Single and joint acute toxicity of disperse violet HFRL and disperse orange S-4RL to *Daphnia magna*. J Environ Health 31:483–485
33. Vacchi FI, von der Ohe PC, de Albuquerque AF et al (2016) Occurrence and risk assessment of an azo dye—the case of Disperse Red 1. Chemosphere 156:95–100. https://doi.org/10.1016/j.chemosphere.2016.04.121
34. Ferraz ERA, Umbuzeiro GA, De-Almeida G et al (2011) Differential toxicity of Disperse Red 1 and Disperse Red 13 in the Ames test, HepG2 cytotoxicity assay, and Daphnia acute toxicity test. Environ Toxicol 26:489–497. https://doi.org/10.1002/tox.20576
35. European Parliament and Council (2008) Regulation on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006
36. Brack W, Ait-Aissa S, Burgess RM et al (2016) Effect-directed analysis supporting monitoring of aquatic environments—an in-depth overview. Sci Total Environ 544:1073–1118. https://doi.org/10.1016/j.scitotenv.2015.11.102
37. United Nations (2017) Globally harmonised system for classification and labelling of chemicals (GHS): seventh revised edition, UN, New York. https://doi.org/10.18356/e9e7b6dc-en

Publisher's Note