S. Jannicke Moe ORCID iD: 0000-0002-3681-3551

EVALUATION OF A BAYESIAN NETWORK FOR PREDICTING ACUTE FISH TOXICITY FROM FISH EMBRYO TOXICITY DATA

Adam Lillicrap[1*], S. Jannicke Moe[1], Raoul Wolf[1], Kristin A. Connors[2], Jane M. Rawlings[2], Wayne G. Landis[3], Anders Madsen[4,5], Scott E. Belanger[2]

[1]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, Oslo, Norway

[2]Procter and Gamble, Cincinnati, OH, USA

[3]Western Washington University, USA

[4] Dept. of Computer Science, Aalborg University, 9220 Aalborg, Denmark

[5] HUGIN EXPERT A/S, 9000 Aalborg, Denmark

*corresponding author

Acknowledgements:

1) The authors declare no conflicts of interest
2) The authors acknowledge financial support from the NIVA internal strategic funding initiative (DIGISIS), grant/contract number 160016 for NIVA's basic funding (GB/SIS) from the Norwegian Research Council.
3) No other individuals or organizations were involved with this manuscript
4) Reference to the data supplier, Procter and Gamble, are referred to as [The company] to ensure anonymity.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ieam.4258.

Disclaimer: Not applicable

Data Accessibility Statement: Data in the database are available on the Wiley Online Library at DOI: 10.1002/etc.4351. All other data can be made available on request.

1. ABSTRACT

The use of fish embryo toxicity (FET) data for hazard assessments of chemicals, in place of acute fish toxicity (AFT) data, has long been the goal for many environmental scientists. The FET test was first proposed as a replacement to the standardized AFT test nearly 15 years ago, but as of now, it has still not been accepted as a standalone replacement by regulatory authorities such as the European Chemicals Agency (ECHA). However, the ECHA have indicated that FET data can be used in a weight of evidence (WoE) approach, if enough information is available to support the conclusions related to the hazard assessment. To determine how such a WoE approach could be applied in practice has been challenging. To provide a conclusive WoE for FET data, we have developed a Bayesian Network (BN) to incorporate multiple lines of evidence to predict AFT. There are four different lines of evidence in this BN model including: 1) physico-chemical properties; 2) AFT data from chemicals in a similar class or category; 3) ecotoxicity data from other trophic levels of organisms (e.g. daphnids and algae) and 4) measured FET data. The BN model was constructed from data obtained from a curated database [the company] and conditional probabilities assigned for the outcomes of each line of evidence. To evaluate the model, 20 data rich chemicals, containing a minimum of three AFT and FET test data points, were selected to ensure a suitable comparison could be performed. The results of the AFT predictions indicated that the BN model could accurately predict the toxicity interval for 80% of the chemicals evaluated. For the

remaining chemicals (20%), either daphnids or algae were the most sensitive test species, and for those chemicals, the daphnid or algal hazard data would have driven the environmental classification.

KEYWORDS: Fish Embryo Toxicity, Acute Fish Toxicity, Weight of Evidence, Bayesian Network, Hazard assessment.

2. INTRODUCTION

Many different chemical legislations (such as the European legislation for Registration, Evaluation, Authorisation and restrictions of Chemicals- REACH) place an emphasis on the need for acute fish toxicity (AFT) data for the hazard assessment of chemicals. These data are required in combination with acute toxicity data from species of different trophic levels (i.e., algae and invertebrates) and are the minimum data requirements to perform an environmental risk assessment. However, the use of vertebrate organisms for acute ecotoxicity assessments has been challenged for ethical reasons and due to certain legislations (such as EU 2010) that specify that the use of vertebrate organisms should be avoided wherever possible. REACH, consistent with the EU Directive on the Use of Animals for Scientific Experimentation, also specifies that the use of vertebrate organisms should be avoided wherever possible.

Russel and Burch (1959) first introduced the concept of humane experimental techniques for animal testing where they proposed the term “the 3Rs”. These 3Rs refer to Reduction (in numbers of animals), Refinement (of any procedure) and Replacement (of vertebrate organisms). More recently, the traditional “3Rs” have been expanded to include 3 additional Rs, namely that any alternative should be Reliable/Robust, Relevant and gain Regulatory acceptance (OECD 1996; Lillicrap et

al., 2016a)). Henceforth, the traditional term of "3Rs" has since been refined and replaced to be considered as the "6Rs" of (eco)toxicity testing (Lillicrap et al., 2016a). For over 20 years, scientists have been developing alternative approaches to predict acute fish toxicity without the need to use fish. These include, but are not limited to, the use of quantitative structure activity relationships (QSARs), *in vitro* test methods (for example, the use of fish cells to determine cytotoxicity; Fischer et al. (2019), the use of other organisms such as invertebrates and algae (Rawlings et al. 2019), and the use of non-protected life-stages, in certain geographical regions, such as fish embryos (Busquet et al. 2014).

The fish embryo test (FET; ISO 2007; German DIN, 2001), was first considered as a promising alternative to the use of juvenile fish (48-hour Golden ide acute toxicity test) for assessing effluent toxicity by the German Federal Agency of the Environment. Subsequently, the method was submitted to the Organisation for Cooperation and Development (OECD) as a new test guideline for the purposes of replacing the fish acute toxicity test (AFT; OECD, 1992). After significant international validation efforts (Busquet et al., 2014), an extension of the test from 48 to 96 hours in duration, and omission of the term "replacement to fish acute toxicity" in the introduction of the test guideline, the fish embryo test was finally accepted as an OECD test guideline nearly 8 years later (OECD, 2013). However, its universal acceptance as a replacement to the acute fish toxicity test (OECD, 1992) has remained an issue because regulators, such as the European Chemicals Agency (ECHA) have not accepted it as a complete replacement (Sobanska et al., 2018). A conservative approach to the acceptance of the FET test has been argued due to the existence of some limitations (e.g., neurotoxic mode of action) and/or remaining uncertainties (e.g., deviation of some narcotic substances) regarding the FET test (Sobanska et al.,

2018). Furthermore, it was concluded that "the FET test alone is currently not sufficient to meet the essential information on AFT as required by the REACH regulation" (Sobanska et al., 2018). This is despite Sobanska et al. (2018) and other authors (Belanger et al., 2013) describing a near perfect 1:1 correlation between toxicity values (EC50 and LC50) from the FET and the AFT test. Nonetheless, Sobanska et al. (2018) stated that the FET test "may be used within weight-of-evidence approaches together with other independent, relevant, and reliable sources of information". The use of weight of evidence has been specified for REACH as an option to meet the information requirements of Annexes VII-X where: "Animal tests can be avoided if there is a weight of evidence which points to the likely properties of a substance […] if there is sufficient information from several independent sources leading to the conclusion that a substance has (or has not) a particular dangerous property, while the information from each single source alone is regarded insufficient to support this notion" (ECHA, 2016). Understanding how this weight of evidence (WoE) may work in practise, to enable FET data to be used in place of AFT data, has been a major challenge. Therefore, we have previously developed a Bayesian Network which incorporates FET data with multiple lines of evidence to provide a probabilistic estimate for acute fish toxicity (Moe et al., 2020). In contrast to Moe et al (2020), which describes both the model development and model training, the purpose of this paper is to evaluate the performance of the Bayesian Network for predicting AFT data from FET data along with other lines of evidence.

A Bayesian network (BN) is a probabilistic modelling methodology which is increasingly being used in ecological risk assessment (e.g. Landis et al. 2019) as well as more generally in environmental research (e.g. Moe et al. 2016; Barton et al. 2012). Bayesian approaches were used to integrate several biological lines of evidence to

predict human skin sensitization to chemicals and is accompanied by a now elucidated Adverse Outcome Pathway (Jaworska et al. 2015; OECD 2012 b). A BN can integrate large amounts of data and other information sources by using discrete probability distributions and predicts the probability of specified states of selected variables. With regards to the FET/AFT BN, a specified state refers to, for example, a given interval of LC_{50} values for AFT. The purpose of the proposed BN model is to integrate information from large ecotoxicological and physico-chemical datasets and apply it to predict fish acute toxicity from data on fish embryo toxicity tests, in combination with other relevant information in a weight of evidence approach. The following sections describes the model set up and the evaluation of the model with a suite of different chemicals. It is envisaged that the use of the BN model will eventually fulfil the requirements of the regulatory community to accept FET data in place of AFT data.

3. MATERIALS AND METHODS

3.1.Chemical database selection

A database of ecotoxicity and other physico- chemical properties for a range of different chemicals, was obtained from [the company]. This database contained data for 237 substances and the following toxicity data:

- Algae: 264 EC_{50} values (duration 72 or 96 hours, according to OECD, 2006). The data comprised seven algal species: the green algae *Chlorella pyrenoidosa*, *Chlorella vulgaris*, *Desmodesmus subspicatus* and *Pseudokirchneriella subcapitata*; the diatom *Skeletonema costatum*; and the cyanobacteria *Anabaena flos-aquae* and *Microcystis aeruginosa*.

- Daphnids: 1164 EC_{50} values (48 hours, according to OECD, 2004). The two species used were *Daphnia magna* and *D. pulex*.

- Juvenile fish: 1459 LC_{50} values (24, 48, 72, 91 or 96 hours, according to OECD, 1992). The data comprised five species: *Danio rerio* (zebrafish), *Lepomis macrochirus* (bluegill), *Oncorhynchus mykiss* (rainbow trout), *Oryzias latipes* (medaka) and *Pimephales promelas* (fathead minnow).

- Fish embryo: 541 LC_{50} values (48, 72, 96, 108 or 120 hours, according to OECD 2013). Data was available for four species: *D. rerio*, *O. latipes*, *P. promelas* and *Clarias gariepinus* (African sharptooth catfish).

The complete dataset (i.e., data available for all 237 chemicals) was used for training the BN model, but a subset of chemicals was used to evaluate how accurately the model could predict AFT. The criteria for selecting these chemicals were that they had either a minimum of 3 FET and AFT data points or a minimum of 1 FET and AFT data point and a molecular weight >600 g/mol. This resulted in 28 candidate chemicals. An additional exclusion criterion was applied to remove any chemicals that had an extremely large spread of AFT data (e.g. for cadmium the data varied by 5 orders of magnitude) or any chemicals that did not have any QSAR data. Even after the selection and exclusion criteria were applied, there were still some chemicals with quite a large spread of AFT data (see Figure 1). In comparison, the FET data were much less variable. The exclusion and selection criteria resulted in 20 prioritized chemicals (see Figure 1 and Table 1) which were subsequently used for the evaluation of the Bayesian network.

3.2.Description of the Bayesian network

The objective of this BN model is to predict the acute toxicity of a chemical to juvenile fish, corresponding to the interval of LC_{50} values from the AFT test, by integrating FET data with other relevant physico- chemical and toxicological information. The application of this model is comparable to a weight-of-evidence process as described by Suter et al. (2017) where the assignment of prior and conditional probabilities to different variables correspond to assigning weights to pieces of evidence. When predicting the AFT for a given chemical, the calculation of posterior probability of each toxicity interval corresponds to the weighing of the total evidence for each hypothesis. The BN was implemented in the software HUGIN Researcher version 8.7, developed by HUGIN EXPERT A/S (http://www.hugin.com). The online demonstration web interface was made available through the demonstration web site http://demo.hugin.com/FET. Parameterization and a description of the model construct is described in detail in Moe et al. (2020).

The BN model has four pathways (lines of evidence) for predicting the AFT of a given chemical (see Figure 2). These include data on; (1) physical and chemical properties, (2) toxicity data to fish (AFT) from chemicals in the same category, (3) toxicity to other species (algae and daphnids) and (4) fish embryo toxicity (FET) data. For the purposes of this paper, it was important that all substances that were evaluated had a minimum of 3 values of corresponding FET and AFT data to enable suitable comparisons between the predicted AFT values and the observed AFT data.

Within each pathway (lines of evidence), discrete nodes (e.g. limited membrane crossing) were assigned. Each node has a conditional probability table (CPT; for an example see Table 1) which is conditional on the parent node (i.e. the prior input data point such as hydrophobicity). A full description of each of the CPTs for each node have been previously described (see Moe et al., 2020). The CPT values were obtained by two main methods: counts of observations, reflecting the distributions within the database, and expert judgement by the authors. All four pathways were assigned the same weight when combined in the predicted toxicity nodes. The toxicity intervals, used in the BN model, were discretized to 5 toxicity levels: very low (>100 mg/L), low (5-100 mg/L), medium (0.5-5 mg/L), high (0.5-0.01 mg/L) and very high (<0.01 mg/L).

It was recognized that variability within FET and AFT data would need to be accounted for in the BN. In Busquet et al. (2014) interlaboratory coefficients of variability (CV) of the FET test was estimated to approximately 26% for all the substances evaluated in the international ring trial of the OECD test guideline 236. With regards to AFT, in an interlaboratory variability study (US EPA, 2001), the CV was estimated to 20%. Therefore, as a conservative approach, the prior probabilities of all toxicity data in the input nodes were set at levels corresponding to minimum 30% CV. The BN predicts the toxicity level to juvenile fish for each chemical by combining information from all four pathways and each of these pathways are described below.

3.2.1. Pathway 1, Toxicity to juvenile fish based on physico-chemical properties

Several descriptors of physico-chemical properties can be used to estimate whether a chemical is likely to partition through biological membranes and hence increase the

likelihood of cellular/molecular interactions with the substance. One of the simplest metrics to determine if the substance is bioavailable is the molecular weight (mw) or size of the molecule. These metrics have been included as one of the endpoints for PBT assessments within REACH, if used in a weight of evidence approach (ECHA, 2017). As indicated in Lillicrap et al. (2016b), if the molecule has a physical size >4.3 nm and the molecule is unlikely to fold along linear structures (thus altering the length) or has a size (based on maximum diameter or $Dmax_{average}$) larger than 1.7 nm then it is unlikely to pass across biological membranes. In addition, if the substance has a molecular weight above a certain size then it should not be bioavailable, and according to ECHA (2017) substances with an average maximum diameter of >1.7 nm and a molecular weight of >1100 g/mol or >700 g/mol should not be bioaccumulative or very bioaccumulative, respectively (i.e., above these sizes they are increasingly non-bioavailable). However, these cut-off criteria should be considered with caution since not all molecules behave the same way, and according to Arnot et al. (2010) some of these assumptions may not have accounted for biotransformation of the substance occurring (i.e., providing a false positive assumption for bioavailability or not). For the purposes of this node in pathway 1, we have chosen a molecular weight cut off from 600 g/mol to assume that substances with a mw >600 g/mol are less likely to cross biological membranes. This is in line with Brooke et al. (1986) who indicated bioaccumulation potential had an upper limit of 600 g/mol.

In line with molecular weight, a measure of substance hydrophobicity can also be used as an indicator for assuming limited membrane crossing potential. Hydrophobicity may be expressed by a substances solubility in water being very low or based on the octanol water partition coefficient (i.e., log K_{ow}) being very high. For the purpose of this second node of the first pathway, we have chosen hydrophobicity

to be based on log K_{ow} and have assigned a cut-off value of 5.5. This is in line with OECD TG 305 (OECD, 2012) where substances with a log K_{ow} >5.5 are recommended to be experimentally assessed using a dietary uptake exposure route, rather than an aqueous exposure route, due to reduced bioavailability. To account for the assumptions for potential membrane crossing, the conditional probability table (Table 1) in the BN model has 3 states: low probability of crossing (for large molecules); high probability (for small molecules) and a medium probability to account for uncertainty in hydrophobicity or molecular size as a cut off value (as detailed by Arnot et al., 2010). Other indicators, such as the Lipinski's rule of 5 (Lipinski et al., 1997), may be possible for accounting for limited uptake across biological membranes, but for the purposes of this BN, and to avoid too much complexity within the model, we have focused on only 2 molecular indicators in this pathway.

The third node of this first pathway is the use of QSAR data. The QSAR data are based on predicted values obtained from the USEPA Ecological Structure Activity Relationships (ECOSAR 1.11) Class Program (https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model) and the Danish QSAR Database http://qsar.food.dtu.dk. The Danish QSAR database differs from the USEPA ECOSAR model because it has 2 different models (Leadscope and SciQSAR) which use different descriptors. The Leadscope model is based on structural features and numeric molecular descriptors (e.g. aLogP, polar surface area, number of hydrogen bond donors, Lipinski score, number of rotational bonds, parent atom count, parent molecular weight and number of hydrogen bond acceptors), and the SciQSAR model is based on molecular descriptors (e.g. molecular connectivity

indices, molecular shape indices, topological indices, electrotopological (Atom E and HE-States) indices and electrotopological bond types indices).

All data passed the pre-established criteria that was assigned as follows: for ECOSAR 1.11 class-specific QSAR models were used to predict acute fish (96 hours) if the equations passed set acceptability criteria ($R_2 \geq 0.6$, $n \geq 4$). If class-specific models were not available or did not meet these criteria, results from the Neutral Organic QSAR model were accepted. For the Danish QSAR database, extracted predicted values for Leadscope and SciQSAR fish 96-hour LC_{50} values (*Pimephales promelas* - Fathead minnow) were used. Data were not used if the Domain listed as "OUT" in the Danish QSAR model. Solubility restrictions were applied and predicted values were excluded if they were 10x the solubility values listed in the database. All QSAR fish LC50 results were than averaged for each chemical.

3.2.2. Pathway 2, toxicity to juvenile fish predicted by read across from chemical analogues with a similar mode of action based on structural alerts

Of the extensive list of existing substances, there are already significant amounts of data that have been generated for determining the acute toxicity to fish. These chemicals can be categorized into functional groups which may be considered similar, and it is possible to "read across" from one substance to another in the same category. Read across is already being used (and sometimes possibly misused) by registrants within REACH and other chemical legislations. However, it is quite reasonable to assume that extrapolating data from one chemical to another can either be logical or illogical, and for this reason, expert judgement is needed when applying read across. To avoid incorrect extrapolations being made, reliability frameworks (e.g. ECHA, 2017) should be applied when assessing read across predictions.

The data used to develop this BN model included substances with different modes of action. Mode of action classes based on structural alerts were assigned by ECOSAR (v1.11), or by expert judgment if the chemical was outside of the domain. These included; nonpolar narcosis, polar narcosis, uncoupler of oxidative phosphorylation, alkylation / arylation based reactivity, carbonyl reactivity (aldehyde eq. #3), ester narcosis, OP mediated AChE inhibition, hydrazine based reactivity, pyridinium compounds, carbamate mediated AChE inhibition, acrylate toxicity, diester toxicity, neurotoxicant: cyclodiene-type, neurotoxicant: pyrethroid, carbonyl reactivity (aldehyde eq. #1), quinoline reactivity and respiratory blocker. Where a chemical fitted a similar class of mode of action based on structural alerts, and corresponding AFT data were available, the AFT data were extracted and incorporated into the BN model (see Moe et al., 2020 for a full description of parameterization).

3.2.3. Pathway 3, toxicity to juvenile fish based on other species

The third pathway of the BN model incorporates ecotoxicity data from additional trophic levels. Environmental risk assessments incorporate hazard data from algae, daphnids and fish. For that reason, we have included results from the algal inhibition assay (according to OECD TG 201) and the *Daphnia magna* immobility assay (OECD TG 202). These trophic levels have been shown to be more sensitive than the AFT (Hutchinson et al., 2003, Jeram et al., 2005) and FET (Rawlings et al., 2019) between 75-80% of the time. As hazard classification and environmental risk assessment is based on the most sensitive taxa, algae and daphnids tend to routinely drive these assessments.

In pathway 3, we have assumed that the probability of high toxicity to fish is conditional on the chemical not having a species-specific mode of action (MoA). For instance, if the chemical results in considerably different EC_{50} values for algae vs. daphnids, this suggests that there is potentially a species-specific MoA affecting either algae or daphnids. To define what constitutes a difference, the ratio between the algae and invertebrate data are first calculated. Cut-off values for the ratio between the daphnid and algae data, as described in ECETOC (2005), have been applied. For example, if the ratio of the two EC_{50} values is between 0.5 and 2, it can be assumed that these data are similar, and there is no specific MoA affecting either species. In these instances, it is reasonable that these EC_{50} values can be used towards extrapolating an LC_{50} value for fish. Therefore, the CPT converts the EC_{50} values from algae and daphnids to LC_{50} values for fish with high precision (a narrow distribution). Conversely, a ratio >2 or <0.5 indicates that one of the species is more sensitive than the other, to a degree that suggests a species-specific MoA (e.g. herbicidal or insecticidal). In these cases, the probability that these EC_{50} data can be extrapolated to AFT LC_{50} data is low, and the CPT therefore has a wide distribution (high uncertainty) in these cases. A full overview of the different conditional probabilities used, and a justification for each selection, is detailed in Moe et al. (2020) and an example of the prior probability distributions applied for both algae and daphnids are included in the supplementary information (SI).

3.2.4. Pathway 4, Fish Embryo Toxicity based on experimental data

The fourth pathway is the use of fish embryo toxicity data. These data need to be derived using the OECD 236 fish embryo acute toxicity test (FET). The data are then

discretized, as previously described, into 5 levels: very low (>100 mg/L), low (5-100 mg/L), medium (0.5-5 mg/L), high (0.5-0.01 mg/L) and very high (<0.01 mg/L).

3.2.5. Predicted toxicity to juvenile fish based on all 4 nodes

All four pathways of the BN model feed into the final node which is an estimate of the predicted AFT based on molecular properties, fish data from other chemicals in the same category, ecotoxicity data from other species and FET data. The final probability node is discretized into the same 5 categories for toxicity assignment as described previously.

4. RESULTS AND DISCUSSIONS

The results from the BN predictions of the 20 prioritized chemicals are shown in Table 2. These results were compared to the actual measured AFT data to evaluate the accuracy of the predictions from the BN. The predicted toxicity interval for juvenile fish, based on the 4 different pathways, are also presented in Table 2. For the 20 chemicals evaluated herein, the BN model correctly predicted the toxicity interval for 80% of the substances. An example of the BN model, showing the distributions for each node for the chemical tetradecyl sulfate (chemical category anionic surfactant), is shown in the SI. AFT data obtained from the database (based on three observations) for tetradecyl sulfate indicated that it would be classified as causing medium toxicity (i.e., LC_{50} 0.5-5 mg/L) to juvenile fish. The node "Toxicity to embryo (level)" was almost 100% high toxicity (i.e., 0.5-0.01 mg/L), while the predictions from the other lines of evidence were centered around low-to-medium toxicity. The resulting predicted toxicity to juvenile fish had the highest probability (41%) of the medium state (i.e., LC_{50} 0.5-5 mg/L), which was consistent with the measured AFT data. In

this case, using FET data alone would have overestimated the risk to juvenile fish, whilst FET data, in combination with the other lines of evidence, resulted in a more accurate prediction on toxicity to juvenile fish.

For the remaining 20% of the substances, where there was an incorrect prediction, either the daphnid or algae data were more sensitive in all cases. For the 3 substances, 2,4-dichlorophenol, 4-chlorophenol and malathion, daphnids were most sensitive. These compounds are classified, according to the Verhaar classification system, as having a Polar narcosis MoA for the 2 phenols, and organophosphate (OP) mediated Acetylcholinesterase (Ache) inhibition for malathion (Verhaar et al., 1992). Being that malathion is an insecticide, it is not surprising that daphnids were the most sensitive species. For the remaining substance Naphthalene, which is a neutral organic causing nonpolar narcosis, algal growth inhibition was the most sensitive endpoint. For these 4 chemicals, where there was an incorrect prediction of the AFT, either daphnids or algae would have driven the environmental classification.

It is important to note that it was not possible, within this present study, to consult each study record for each data point to perform a reliability evaluation. Therefore, the comparisons between the BN predictions and the AFT data herein assume that all data were reliable. However, it is possible that some data may not be reliable, or that the data could be questionable. As an example, AFT data generated from different laboratories, and from different species of fish, can vary by multiple orders of magnitude (Belanger et al., 2013). In the initial exclusion exercise to identify the prioritized chemicals used to validate the BN, the AFT data for some substances, such as cadmium, varied by more than 5 orders of magnitude and were excluded from the evaluations. The reason for such high variability may be due to reasons such as the

applicability of the test, differences between species sensitivities or, more likely, poor experimental data. It may also be due to other confounding factors such as when AFT data is based on nominal versus measured concentrations. To illustrate this, Jonker et al. (2018) performed an international ring trial on passive sampling and attributed the large inter laboratory variability that was observed to analytical chemistry. Only when 1 single laboratory performed the analytical chemistry for the ring trial it was possible to eliminate the large source of variability. Furthermore, the appropriateness of the AFT test design (OECD, 1992) may also be questionable. For instance, for hydrophobic substances (i.e., with a log K_{ow} >5.5) a 96-hour duration may not be sufficiently long enough to ensure that equilibrium within the fish and the exposure media has been achieved. This means that for certain hydrophobic substances, the AFT test may be underestimating the actual toxicity since a critical body burden has not been achieved within the fish. Similarly, certain embryo toxic substances such as triazoles and glycol ethers, that cause growth retardation and malformations in zebrafish embryos (Hermsen et al., 2011), will not elicit the same response in juvenile fish. Henceforth, since the AFT is a relatively crude test (i.e., do the fish live or die), or significantly underestimates the toxicity due to the exposure duration being too short, or there is a species specific MoA (e.g. herbicide or insecticide) the other base line toxicity assays (e.g. algae and invertebrates) are often more sensitive than fish. To illustrate this point, where the BN did not correctly predict the AFT (20% of the chemicals), it was inconsequential since algae or daphnids were more sensitive in all cases. At this stage, it should be noted that the FET test is also not without its limitations. For example, it has been shown that the FET test exhibits a weak response to substances that have a neurotoxic mode of action (Klüver et al., 2015). However, Klüver et al. (2015) recommended that substances with a neurotoxic mode of action

could be identified using behavioral endpoints such as embryonic locomotion. In the future, it might be possible that more sensitive endpoints, such as behavior, may/should be incorporated into the OECD test guideline 236 to account for specific modes of action such as neurotoxicity.

Clearly, data from only one ecotoxicity test (e.g. algal toxicity, daphnid immobilization or AFT) is insufficient to provide adequate information to perform a hazard and risk assessment or for the purposes of chemical classification. Therefore, it is imperative that all environmental data be incorporated to provide better confidence in performing environmental risk assessments and for classification purposes. To this end, the requirement to develop a weight of evidence approach has been welcomed by the authors of this paper as we believe that our Bayesian Network model is timely, and of greater importance than to simply predict acute fish toxicity data from fish embryo toxicity data. Moreover, the use of our BN model will help to improve future environmental risk assessments of chemicals *per se*. However, the model that has been developed thus far, is based on currently regulatory accepted methods and approaches, and there are many other lines of evidence that could be incorporated to increase the predictive power of the model. One example would be the use of cytotoxicity data from the RT-gill cytotoxicity assay, which has recently been accepted as an International standard (ISO, 2019). Another line of evidence would be to incorporate information related to metabolism and neurotoxicity to inform those substances which might require metabolic activation, or which have a specific mode of action causing toxicity to older life stages (e.g. juvenile) of fish. Furthermore, data from other sources of information (e.g. other species, chronic toxicity data or human safety information) or other (Q)SAR models could also refine the current model. With these additional pieces of information, it is envisaged that the BN model could

improve the weight of evidence to more accurately predict AFT data, enabling FET data to be submitted in place of AFT data for regulatory requirements such as REACH.

5. CONCLUSIONS

A Bayesian Network has been developed and is able to predict acute fish toxicity from fish embryo toxicity data, in combination with other lines of evidence. For a subset of chemicals, the BN was able to accurately predict the toxicity interval for 80% of the chemicals evaluated. In these cases, the BN demonstrated that a sufficient weight of evidence could justify the use of FET data in place of AFT data. For the remaining 20% of the chemicals, where an incorrect prediction was made, either the daphnid or algae data were more sensitive, and these data would have driven any environmental classification. Nonetheless, the current BN model could benefit from additional lines of evidence to be included, and a larger database of chemicals to further train the model to reduce any uncertainties in the predictions. It is recommended that BN models should be used more often for determining weight of evidence and we encourage further dialogue with the scientific and regulatory community to advance the acceptance of such models to replace the use of AFT tests in the future.

6. REFERENCES

Arnot JA, Arnot MI, Mackay D, Couillard Y, MacDonald D, Bonnell M, Doyle P (2010). Molecular Size Cutoff Criteria for Screening Bioaccumulation Potential: Fact or Fiction? Integr Environ Assess Manag, 6 (2) pp. 210–224.

Barton DN, Kuikka S, Varis O, Uusitalo L, Henriksen HJ, Borsuk M, de la Hera A, Farmani R, Johnson S, Linnell JD (2012). Bayesian networks in environmental and resource management. Integr Environ Assess Manag, 8, 418-429.

Belanger SE, Rawlings JM, Carr GJ (2013). Use of fish embryo toxicity tests for the prediction of acute fish toxicity to chemicals. Enviro Toxicol Chem, 32:1768-1783.

Brooke DN, Dobbs AJ, Williams N (1986). Octanol: Water partition coefficients (P): Measurement, estimation, and interpretation, particularly for chemicals with P > 105. Ecotoxicol Environ Saf, 11, 251-260.

Busquet F, Strecker R, Rawlings JM, Belanger SE, Braunbeck T, Carr GJ, Cenijn P, Fochtman P, Gourmelon A, Hubler N, Kleensang A, Knobel M, Kussatz C, Legler J, Lillicrap A, Martinez-Jeronimo F, Polleichtner C, Rzodeczko H, Salinas E, Schneider KE, Scholz S, van den Brandhof E-J, van der Ven LTM, Walter-Rohde S, Weigt S, Witters H, Halder M (2014). OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. Reg Tox Pharm, 69:496-511.

ECETOC (2005). Alternative Testing Approaches in Environmental Safety Assessment. Technical report No. 97, ISSN-0773-8072-97.

European Union (2010). Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official J Eur Union L276:34–79.

ECHA (2016). Practical guide: How to use alternatives to animal testing to fulfil your information requirements for REACH registration. Version 2.0 – July 2016,

https://echa.europa.eu/documents/10162/13655/practical_guide_how_to_use_alternatives_en.pdf/148b30c7-c186-463c-a898-522a888a4404 (accessed 28.08.2019).

ECHA (2017). Read-Across Assessment Framework (RAAF). ECHA reference number ECHA-17-R-01-EN, ISBN 978-92-9495-758-0. DOI: 10.2823/619212.

Fischer M, Belanger SE, Berckmans P, Bernhard MJ, Bláha L, Coman Schmid DE, Dyer SD, Haupt T, Hermens JLM, Hultman MT, Laue H, Lillicrap A, Mlnaříková M, Natsch A, Novák J, Sinnige TL, Tollefsen KE, von Niederhäusern V, Witters H, Županič A, Schirmer K (2019). Repeatability and Reproducibility of the RTgill-W1 Cell Line Assay for Predicting Fish Acute Toxicity. Toxicol Sci, 169, 353-364.

German DIN (2001). German standard methods for the examination of water, waste water and sludge — Subanimal testing (group T) — Part 6: Toxicity to fish. Determination of the non-acute-poisonous effect of waste water to fish eggs by dilution limits (T 6). DIN 38415-6.

Hermsen SAB, van den Brandhof E-J, van der Ven LTM, Piersma AH (2011). Relative embryotoxicity of two classes of chemicals in a modified zebrafish embryotoxicity test and comparison with their in vivo potencies. Toxicol in Vitro, 25, 745-753.

Hutchinson TH, Barrett S, Buzby M, Constable D, Hartmann A, Hayes E, Huggett D, Lange R, Lillicrap AD, Straub JO, Thompson RS. (2003). A Strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. Enviro Toxicol Chem, 22:3031-3036.

International Standards Organization (ISO) (2007). Water Quality-Determination of the Acute Toxicity of Wastewater to Zebrafish Eggs (Danio rerio). ISO 15088:2007(E), Geneva, Switzerland.

International Standards Organization (ISO) (2019). Water quality -- Determination of acute toxicity of water samples and chemicals to a fish gill cell line (RTgill-W1). ISO 21115:2019, Geneva, Switzerland.

Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, Miyazawa M. (2015). Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol, 89:2355–2383.

Jeram S, Riego Sintes JM, Halder M, Fenatanes JB, Sokull-Kluttgen B, Hutchinson TH. (2005). A strategy to reduce the use of fish in acute ecotoxicity testing of new chemical subtances in the European Union. Reg Tox Pharm, 42:218-224.

Jonker MTO, van der Heijden SA, Adelman D, Apell JN, Burgess RM, Choi Y, Fernandez LA, Flavetta GM, Ghosh U, Gschwend PM, Hale SE, Jalalizadeh M, Khairy M, Lampi MA, Lao W, Lohmann R, Lydy MJ, Maruya KA, Nutile SA, Oen AMP, Rakowska MI, Reible D, Rusina TP, Smedes F, Wu Y (2018). Advancing the Use of Passive Sampling in Risk Assessment and Management of Sediments Contaminated with Hydrophobic Organic Chemicals: Results of an International Ex Situ Passive Sampling Interlaboratory Comparison. Environ Sci Technol, 52, 3574-3582.

Klimisch HJ, Andreae M, Tillmann U (1997). A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. Reg Tox Pharm, 25, 1-5.

Klüver N, König M, Ortmann J, Massei R, Paschke A, Kühne R, Scholz S (2015). Fish Embryo Toxicity Test: Identification of compounds with weak toxicity and analysis of behavioral effects to improve prediction of acute toxicity for neurotoxic compounds. Environ Sci Technol, 49, 7002-7011.

Landis WG, hu VR, Graham SE, Harris MJ, Markiewicz AJ, Mitchell CJ, von Stacelburg KE, Stark JD (2019). The integration of chlorpyrifos acetylcholinesterase inhibition, water temperature and dissolved oxygen concentration into a regional scale multiple stressor risk assessment estimating risk to Chinook salmon in four rivers in Washington State, USA. Integr Environ Assess Manag, 16: 28-42.

Lillicrap A, Belanger S, Burden N, Pasquier DD, Embry MR, Halder M, Lampi MA, Lee L, Norberg-King T, Rattner BA, Schir mer K and Thomas P (2016a). Alternative approaches to vertebrate ecotoxicity tests in the 21st century: A review of developments over the last 2 decades and current status. Environ Toxicol Chem, 35: 2637-2646.

Lillicrap A, Springer T, Tyler CR (2016b). A tiered assessment strategy for more effective evaluation of bioaccumulation of chemicals in fish. Reg Tox Pharm, 75, 20-26.

Moe SJ, Haande S, Couture RM (2016). Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. Ecological Modelling, 337, 330-347.

Moe, SJ, Belanger, SE, Connors KA, Landis WG, Madsen AL, Rawlings JM, Wolf R, Lillicrap A (2020). Detailed description of a Bayesian network model developed for predicting toxicity of chemicals to fish: towards a probabilistic weight-of-evidence approach. Environ Modell Software, doi.org/10.1016/j.envsoft.2020.104655.

OECD (1992). OECD guidelines for testing of chemicals. 203: Fish, acute toxicity test. Organisation for Economic Co-operation and Development, Paris, France.

OECD (1996). Final report of the OECD workshop on harmonization of validation and acceptance criteria for alternative toxicological test methods. Organisation for Economic Co-operation and Development, Paris, France.

OECD (2004). OECD guidelines for testing of chemicals. 202: Daphnia sp. Acute immobilization test. Organisation for Economic Co-operation and Development, Paris, France.

OECD (2006). OECD guidelines for testing of chemicals. 201: Freshwater alga and cyanobacteria growth inhibition test. OECD Test Guidelines for the Testing of Chemicals. Paris, France.

OECD (2012). OECD Guidelines for Testing of Chemicals. 305: Bioaccumulation in fish: Aqueous and dietary exposure. Organisation for Economic Co-operation and Development, Paris, France.

OECD (2012b). The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. Part 1: scientific assessment. OECD series on testing and assessment no. 168. OECD Publishing, Paris.

OECD (2013). OECD guidelines for testing of chemicals. 236. Fish Embryo Acute Toxicity (FET) test. Organisation for Economic Co-operation and Development, Paris, France.

Rawlings JM, Belanger SE, Connors KA, Carr GJ. (2019). Fish embryo tests and acute fish toxicity tests are interchangeable in the application of the threshold approach. Enviro Toxicol Chem, 38(3):671-681.

Russell WMS, Burch RL. 1959. The Principles of Humane Experimental Technique. Muethen and Company, London, UK.

Sobanska M, Scholz S, Nyman A, Cesnaitis R, Gutierrez Alonso S, Klüver N, Kühne R, Tyle H, de Knecht J, Dang Z, Lundbergh I, Carlon C, De Coen W (2018). Applicability of the fish embryo acute toxicity (FET) test (OECD 236) in the regulatory context of Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH). Environ Toxicol Chem, 37: 657-670.

Suter G, Cormier S, Barron M (2017). A weight of evidence framework for environmental assessments: Inferring qualities. Integr Environ Assess Manag, 13(6) 1038-1044.

US EPA (2001). Final Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods, vol. 1. EPA 821-B-01-004. Office of Water, U.S. Environmental Protection Agency, Washington, D.C.

Verhaar, H.J.M., van Leeuwen, C.J., Hermens, J.L.M. (1992). Classifying environmental pollutants. 1. Structure-activity relationships for prediction of aquatic toxicity. Chemosphere, 25, 471-491.

Figure 1 Overview of the data spread for the AFT data and the FET data used in the validation of the BN. Numbers in parenthesis, after each chemical name, are the number of data points (i.e., individual study data)

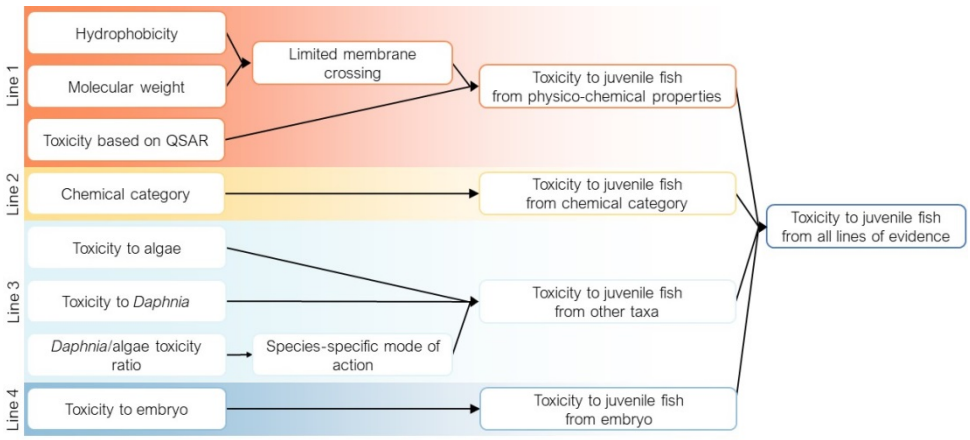

Figure 2. Simplified illustration of the BN model with the 4 pathways (lines of evidence) contributing to predicting the AFT in a WoE approach

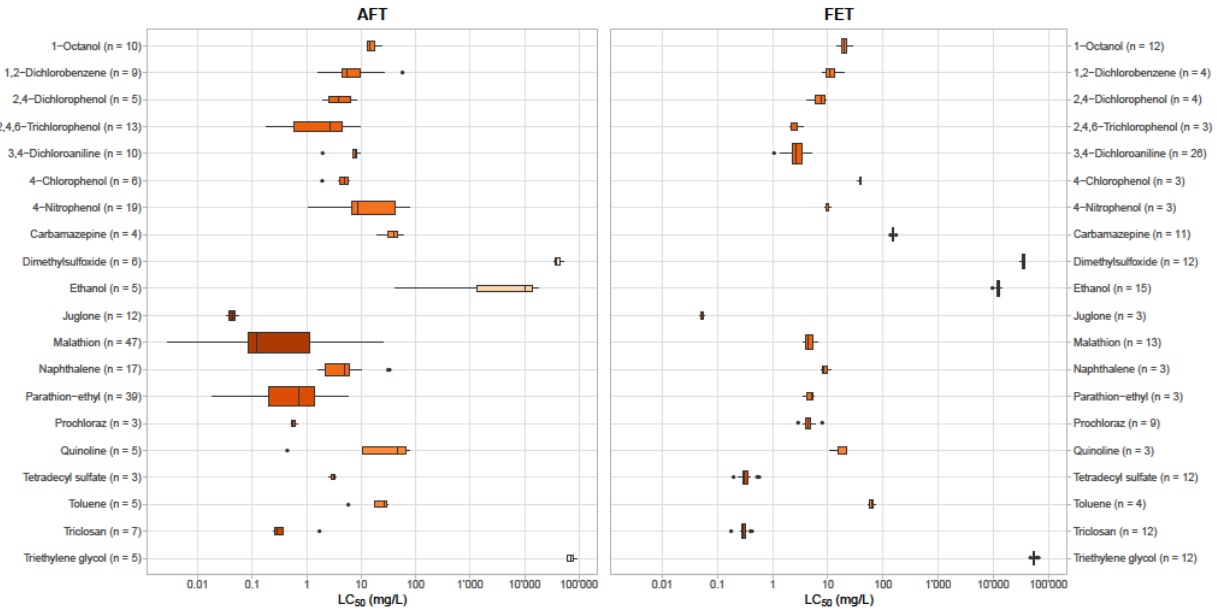

Table 1. Example of conditional probability table (CPT) for the hydrophobicity and molecular weight nodes in pathway 1. It illustrates the approach used for parameterisation of the CPTs for the node " Limited membrane crossing": the probability of a substance crossing a biological membrane based on its physical properties (i.e., molecular weight and hydrophobicity). The probabilities are based on expert judgement.

Hydrophobicity (log K_{ow})	Molecular weight (g/mol)	Probability for membrane crossing		
		low	medium	high
< 5.5	< 600	0%	25%	75%
< 5.5	> 600	25%	50%	25%
> 5.5	< 600	25%	50%	25%
> 5.5	> 600	75%	25%	0%

Table 2. Summary of BN model predictions for the 20 selected chemicals. The observed juvenile and embryo are data obtained from the database. For the chemicals that there was an incorrect prediction, the most sensitive species is identified in parenthesis.

Chemical	**Chemical Class (ECOSAR 1.11)**	**Verhaar classification**	**Observed Juv.**	**Obs. Emb.**	**BN Prediction**	**Correct prediction**

1,2-Dichlorobenzene	Neutral organic	Nonpolar narcosis	low	low	low	Y
1-Octanol	Neutral organic	Nonpolar narcosis	low	low	low	Y
2,4,6-Trichlorophenol	Phenol	Polar Narcosis	medium	medium	medium	Y
2,4-Dichlorophenol	Phenol	Polar Narcosis	medium	low	low	N (D)
3,4-Dichloroaniline	Aniline	Polar Narcosis	low	medium	low	Y
4-Chlorophenol	Phenol	Polar Narcosis	medium	low	low	N (D)
4-Nitrophenol	Phenol	Polar Narcosis	low	low	low	Y
Carbamazepine	Substituted urea	Nonpolar narcosis	low	low	low	Y
Dimethylsulfoxide	Neutral organic	Nonpolar narcosis	very low	very low	very low	Y
Ethanol	Neutral organic	Nonpolar narcosis	very low	very low	very low	Y
Juglone	Quinone	Alkylation / arylation based reactivity	high	high	high	Y
Malathion	Esters (dithiophosphates)	OP mediated AChE inhibition	high	medium	medium	N (D)
Naphthalene	Neutral organic	Nonpolar narcosis	medium	low	low	N (A)
Parathion-ethyl	Esters (monothiphosphates)	OP mediated AChE inhibition	medium	low	medium	Y
Prochloraz	Imidazole	Pyridinium compounds	medium	medium	medium	Y
Quinoline	Neutral organic	Quinoline	low	low	low	Y

		reactivity				
Tetradecyl sulfate	Anionic surfactant	Nonpolar narcosis	medium	high	mediu m	Y
Toluene	Neutral organic	Nonpolar narcosis	low	low	low	Y
Triclosan	Phenol	Polar Narcosis	high	high	high	Y
Triethylene glycol	Neutral organic	Nonpolar narcosis	very low	very low	very low	Y

A- algae, D- daphnid, Emb- embryo, Juv- juvenile