Science of the Total Environment 648 (2019) 80–89

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Prediction of bioconcentration factors in fish and invertebrates using machine learning

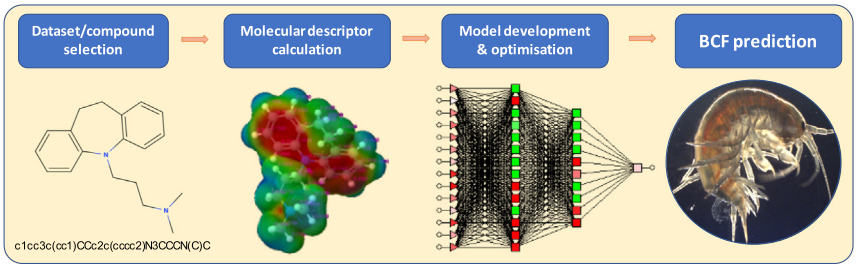
Thomas H. Miller [a,*], Matteo D. Gallidabino [b], James R. MacRae [c], Stewart F. Owen [d], Nicolas R. Bury [e,f], Leon P. Barron [a,*]

[a] Department of Analytical, Environmental & Forensic Sciences, School of Population Health & Environmental Sciences, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, UK
[b] Department of Applied Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK
[c] Metabolomics Laboratory, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK
[d] AstraZeneca, Global Environment, Alderley Park, Macclesfield, Cheshire SK10 4TF, UK
[e] Division of Diabetes and Nutritional Sciences, Faculty of Life Sciences and Medicine, King's College London, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK
[f] Faculty of Science, Health and Technology, University of Suffolk, James Hehir Building, University Avenue, Ipswich, Suffolk IP3 0FS, UK

HIGHLIGHTS

- Evaluation of 24 models to predict bioconcentration factors in fish is presented.
- Machine learning showed good predictive performance.
- First machine learning application to predict bioconcentration in invertebrates
- Cross-species modelling is limited by case similarity and biological variability.
- TPSA, LogD, and Mw were important descriptors for modelling accumulation processes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 July 2018
Received in revised form 8 August 2018
Accepted 9 August 2018
Available online 10 August 2018

Editor: D. Barcelo

Keywords:
Modelling
PBT
Pharmaceutical
Bioconcentration
BCF
Machine learning

ABSTRACT

The application of machine learning has recently gained interest from ecotoxicological fields for its ability to model and predict chemical and/or biological processes, such as the prediction of bioconcentration. However, comparison of different models and the prediction of bioconcentration in invertebrates has not been previously evaluated. A comparison of 24 linear and machine learning models is presented herein for the prediction of bioconcentration in fish and important factors that influenced accumulation identified. R^2 and root mean square error (RMSE) for the test data (n = 110 cases) ranged from 0.23–0.73 and 0.34–1.20, respectively. Model performance was critically assessed with neural networks and tree-based learners showing the best performance. An optimised 4-layer multi-layer perceptron (14 descriptors) was selected for further testing. The model was applied for cross-species prediction of bioconcentration in a freshwater invertebrate, *Gammarus pulex*. The model for *G. pulex* showed good performance with R^2 of 0.99 and 0.93 for the verification and test data, respectively. Important molecular descriptors determined to influence bioconcentration were molecular mass (MW), octanol-water distribution coefficient (logD), topological polar surface area (TPSA) and number of nitrogen atoms (nN) among others. Modelling of hazard criteria such as PBT, showed potential to replace the need for animal testing. However, the use of machine learning models in the regulatory context has been minimal to date and is critically discussed herein. The movement away from experimental estimations of accumulation to in silico modelling

* Corresponding authors.
E-mail addresses: thomas.miller@kcl.ac.uk (T.H. Miller), leon.barron@kcl.ac.uk (L.P. Barron).

https://doi.org/10.1016/j.scitotenv.2018.08.122

would enable rapid prioritisation of contaminants that may pose a risk to environmental health and the food chain.

1. Introduction

Both terrestrial and aquatic environments experience pollution from a wide range of chemical contaminants. The presence of these contaminants is a cause for concern as they may elicit adverse effects to environmental and public health. Bioaccumulation of chemicals is critically important for understanding the risk of chemicals in the environment. The complexity of confounding factors that affect uptake make simple relationships that can confidently predict the accumulation elusive; but it may not have to be that way.

Live animal exposure studies are currently the norm, using many hundreds of fish for each assessment (Rovida and Hartung, 2009). Across the European Union (EU), various guidelines have been established for industry to minimise the risk posed by their chemical products. For pharmaceuticals in the EU this is regulated by the European Medicines Agency (EMA) and for other chemicals substances the regulations are outlined by the Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) (European Commission, 2006; Euorpean Medicines Agency, 2006). According to REACH, any manufacturer of a chemical that exceeds quantities of 10 t per annum must submit a chemical safety assessment (CSA). For environmental risk assessment, part of the CSA includes persistence, bioaccumulation and toxicity (PBT) assessments. Alternatively, for pharmaceuticals environmental risk assessment (ERA) follows an initial screening (Phase I) where physico-chemical properties of the compound are determined (e.g. logP) and the expected exposure is estimated. The Phase I exposure estimation is calculated as the predicted environmental concentration (PEC). If the PEC is >0.01 μg L^{-1} then the pharmaceutical must undergo further testing to assess environmental fate and toxicity. However, it should be noted that substances with a logP >4.5, will trigger a PBT assessment (following REACH guidelines) regardless of the Phase I PEC.

For PBT assessments, existing available screening data and prior assessment information are used to determine whether a chemical is bioaccumulative (B) or very bioaccumulative (vB) by estimation of a bioconcentration factor (BCF) or bioaccumulation factor (BAF). Currently, pharmaceuticals are not restricted or replaced as would normally be defined under REACH. Furthermore, whilst PBT assessments are implemented, the persistence and bioaccumulation outcome of these assessments are not taken into consideration for authorisation purposes, as no legal provisions specifically cover persistent, bioaccumulative and toxic substances for pharmaceuticals (European Medicines Agency, 2016).

Laboratory testing for PBT brings with it a significant level of planning, quality control and cost (Rovida and Hartung, 2009). Therefore, *in silico* methodologies to predict BCF or BAF offers a potential advantage to more intelligently use data to characterise potential exposure and risk. Quantitative Structure Activity Relationships (QSARs) are becoming increasingly popular within ecotoxicological fields as they represent, perhaps, the only realistically feasible scenario to assess the environmental risk of the several thousand chemicals that are available on the market (Gissi et al., 2013). In addition, such models can be used to ethically reduce or replace animal testing and falls under the replacement, reduction and refinement (3Rs) framework (de Wolf et al., 2007). Further, effective *in silico* models could also be utilised to help shape future drugs in terms of 'green by design' ambitions (Lockwood and Saïdi, 2017).

More recently, more complex machine learning-based QSAR models involving artificial neural networks (ANNs), tree-based learners or support vector machines (SVMs) have been used to model BCF in fish (Fatemi et al., 2003; Lombardo et al., 2010; Zhao et al., 2008; Strempel et al., 2013). However, several variations of machine learning-type models exist and wider applications of such models for bioaccumulation prediction have not yet been evaluated to identify any added benefits. Furthermore, current QSAR models have only been applied to modelling fish bioaccumulation data and do not incorporate pharmaceutical data. The potential for application to other taxa such as invertebrates is also non-existent, mainly due to a shortage of available data.

The aim of this work was to develop and critically evaluate several machine learning-based modelling tools for prediction of bioconcentration factor (BCF) in both a fish (*Cyprinus carpio*) and an invertebrate species (*Gammarus pulex*) for the first time. An open access fish BCF dataset was used in the first instance to build and compare 24 different models for 352 different compounds. Subsequently, the best model was applied to both a set of fish and invertebrate BCF data to assess its potential for cross-species prediction. The invertebrate dataset also contained mainly pharmaceuticals. In parallel, independent models were developed ab initio on a smaller set of invertebrate BCF data alone to assess the degree of commonality with the model developed on fish BCF data. Finally, the importance of molecular descriptors to understand the potential for a chemical to accumulate in biota was assessed. The use of such rapid and flexible modelling approaches is now critical to support the 3Rs, aid greener design and to help meet the demand for PBT assessments of potentially large numbers of compounds, which could be expanded to new and emerging environmental contaminants across different species.

2. Materials and methods

2.1. Dataset generation and pre-processing

Bioconcentration factors were collated from the European Chemical Industry Council Long-range Research Initiative (Cefic LRI) project EC07 in collaboration with European Academy for Standardisation e.V (EURAS) which established the BCF gold standard database across multiple fish species and is freely available at http://ambit.sourceforge.net/euras/. BCFs were down-selected to reduce variability between different species and experimental conditions within the database. The BCF data used herein were specific to *C. carpio* and were included by the Chemicals Inspection and Testing Institute (Institute, 1992). Out of all BCF data, this sub-selection resulted in the largest dataset with a single fish species (n = 352) for modelling purposes. The reported BCFs represented whole-body values only and included pigments, pesticides, fungicides, herbicides, insecticides, polyaromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs), organochlorines, nitroaromatics, alkylphenols, aromatic hydrocarbons, organosulfurs and organotins. Approximately 36% of the dataset contained ionisable compounds (estimated from ACD labs, Percepta software). The invertebrate BCF dataset (n = 34) was collated from literature reported data (Ashauer et al., 2006; Ashauer et al., 2010; Meredith-Williams et al., 2012; Miller et al., 2017; Miller et al., 2016a) for the benthic freshwater organism, *G. pulex*. This species was selected as there was a relatively large amount of BCF data available when compared with other invertebrate species. For these, BCF data were only available for pharmaceuticals and pesticides and, again, represented whole-body values.

Simplified molecular input line entry system (SMILES) strings were generated for each compound using Chemspider (Royal Society of Chemistry, UK). Molecular descriptors were generated from SMILES strings using Parameter Client (Virtual Computational Chemistry Laboratory, Munich, Germany), and ACD Labs Percepta (Advanced Chemistry Development Laboratories, ON, Canada). Approximately 450